EXPLORING STRATEGIES COMPUT`ER SCIENCE EDUCATORS NEED TO USE TO PREPARE MACHINE-LEARNING DATASETS FOR PREDICTING DROPOUT RATES IN MOOCS

A Dissertation Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Computer Science

By

Houssen Nafed

Colorado Technical University

August 2019

Committee

Dr. Mohamed Abdalla Lofty, Ph.D., Chair

Dr. Samuel Sambasivam, Ph.D., Committee Member

Dr. Bhanu Kapoor, Ph.D., Committee Member

August 9, 2019
Date Approved

Abstract

Different strategies to prepare machine learning datasets for predicting dropout rates of students in massive open online courses have not been established. The goal of this exploratory qualitative study was to explore the different strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in a computer science or technology massive open online courses (MOOCs). There was a critical need to examine the effectiveness and shortcomings of MOOCs and student retention in these courses. The central research question was, "What are the different strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses." A sample of 25 participants from LinkedIn machine learning groups, who have experience in computer science, machine learning, and MOOCs, responded to the study online questionnaire. The data analysis resulted in the emergence of the following three major themes, (a) the predictive models or algorithms used to predict dropout rates, (b) the elements of the MOOCs experience, and (c) the data elements needed in the machine-learning datasets.

Keywords: Machine learning, datasets, predicting, Massive Open Online Courses, dropout rates of students, computer science educators.

Dedication

I dedicate this dissertation to my wife, my kids, my parents, and my uncle. Also, to all my family and my relatives that I did not see them since 2012.

Acknowledgements

While all the faculty members and my cohort contributed to my success, I would like to thank the following professors for their help in making my dream of a doctorate come true. My deep and sincere acknowledgment goes to Dr. Mohamed Abdalla Lofty, also to Dr. Trish Elley Without their support, I would never have completed this dissertation.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

Many universities around the world are interested in adopting new learning environments, such as massive open online courses (MOOCs), for the required curriculum in various fields (Ewais & Samra, 2017) Those universities also are increasingly using machine learning to investigate academic areas, such as MOOC dropout rates, to predict educational trends in enrollment data (Zheng & Yin, 2015). The many possibilities for MOOC educational environments must be investigated to understand student dropout rates, especially as these rates reach new highs (Yang, Sinha, Adamson, & Rosé, 2013). Due to an increasing number of MOOC platform users, and the massive amount of personal and course data, it became necessary to collect educational data to analyze and improve the quality of education in the MOOC courses (Zheng & Yin, 2015). This extensive data is now known as "big data." The goal of this study was to explore the different strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in a computer science or technology massive open online course. There is a critical need to examine the effectiveness and shortcomings of MOOCs and student retention in these courses. Also, computer science educators are seeking to identify the type of models and algorithms in machine learning for predicting dropout rates of students. The types of data within prediction datasets used in machine learning algorithms to predict student dropout rates are MOOCs needed to be identified. This chapter contains the following sections: background of the study, research problem, the purpose of the study, research question, conceptual framework, theoretical perspectives, assumptions, biases, significance of the study, delimitations, definition of terms, and a general overview of the research. Current studies discussed different predictive models or algorithms in machine learning without focusing on the essential algorithms that should be used to solve the

problem of MOOC dropout. This gap in scholarly literature was noticed by Umer, Susnjak, Mathrani, and Suriadi (2017) while comparing machine-learning algorithms to evaluate which algorithm outperformed other algorithms. Therefore, this study's goal was to explore which predictive models or algorithms are currently used by computer science educators for predicting dropout rates of students in MOOCs.

TOPIC OVERVIEW/BACKGROUND

Computer science and information technology online education is one of the fastest growing forms in the field of technology education since 1999 (Nicholson, 2007). There has been an evolution in e-learning in recent years that led to the emergence of a modern educational phenomenon such as the universal use of MOOCs to increase the effectiveness of online courses (Jordan, 2014). Learners in these online courses are unlimited in number, and the MOOCs are open to anyone to participate. MOOCs started in 2008 by George Siemens and David Cormier at the University of Manitoba, Canada (Cormier, 2010). Recent studies indicated the needed attention to study student motivation in MOOCs, but still, there is concern regarding the dropout rates in these MOOCs, where learners do not finish their MOOC courses (Zheng, Rosson, Shih, & Carroll, 2015)

The MOOC platforms can generate different types of user data with can be given to machine-learning predictive models to help educators predict categories, such as discussion forums participation levels, and monitor the movement of participants in the MOOCs every week (Xing, Chen, Stein, & Marcinkowski, 2016). There is extensive interest in studying the dropout rate of students from MOOCs as well as identifying and classifying students regarding withdrawals rates, continuation rates, and incompletions (Zheng et al, 2015) Previous studies focused on how to create appropriate prediction models that will predict the enrollment of

students in educational MOOCs (Conijn, Van den Beemt, & Cuijpers, 2018). The predictive analytics are interpreted by using the Pipelines Model as a design, that helps researchers in computer science provide the most accurate interpretation of the dropout rates (Nagrecha, Dillon, & Chawla, 2017). The Pipelines Model approach in dropout prediction on MOOCs helps to get a simple idea on past student behavior to predict future results (Nagrecha et al., 2017).

In recent years, studies have increasingly used predictive models to understand the pattern and type of data in analyzing learning in MOOCs (Khalil, 2018). For example, machine-learning algorithms have an important role in analyzing student data based on the weekly history of student's behavior. By using machine learning algorithms, student behavior can be observed in the later phases of courses and over time to predict the dropout problem better than the traditional method without using prediction big data analytics (Kloft, Stiehler, Zheng, & Pinkwart,2014)

In 2015, the number of worldwide participants learning through MOOC platforms increased reaching 35 million learners from different countries. These learners had access to 4,200 training courses organized by 500 universities (Sunar, White, Abdullah, & Davis, 2017). MOOC course designers orchestrate content to give participants an opportunity to learn by varying the content of the course with lectures, videos, readings, quizzes, and discussions (Sunar et al., 2017). Research showed that by enrolling students from around the world, participation for online students is much lower than actual classroom students (Maitland & Obeysekare, 2015). Most studies associated with student completion rates in MOOCs showed that the future of MOOCs depends specifically on giving participants an opportunity to share their views and reflect through discussions and comments (Rodriguez, 2012). Studies have confirmed that when

interaction increases among students using MOOCs, the dropout rate is lower than classroom based courses when classroom interactions between students are infrequent (Sunar et al., 2017).

The open environment for learning through MOOCs has continuously increased its material offerings, which lead to attracting many learners with different educational backgrounds (Khalil, 2018). Many MOOCs in multiple disciplines are offered from high-quality universities all over the world (Khalil, 2018). Moreover, users from different educational backgrounds can enroll in these courses online. The content of these courses has a role in attracting many learners to the university's educational platform and provides learners with an opportunity to explore a variety of topics (de Freitas, Morgan, & Gibson, 2015) However, students may not benefit from the MOOC courses if the level is inappropriate for the learner, and the content is incompatible with their learning outcomes (Pilli & Admiraal, 2017).

Wang and Baker (2015) stated that they might be able to understand students' goals in MOOCs by investigating how students use MOOCs, including MOOC features such as video watching and discussion forums that help motivate students to learn across these platforms, and modifying them according to the learner's desire. Ewais and Samra (2017) study explored the principles used for delivering adaptive courses based on intended learning outcomes in terms of learner's satisfaction and reduction of dropout rates. Corrin, de Barba, and Bakharia (2017) study focused on understanding the methods of tracking learners based on their behavior. Instructors sought to improve the educational experience of MOOCs by providing recommendations to the learners based on their learning styles. MOOCs has attracted researchers and opened a discussion on the impact of this modern educational phenomenon and the problems online education faces (Alumu & Thiagarajan, 2016). Khalil (2018) studied the expansion of the educational process through MOOCs and the interaction of students within these educational

courses, hoping that the students who finish their courses successfully to adopt MOOCs in the future. (Guo & Reinecke, 2014) investigated the MOOCs dropout rates and the heterogeneity of learners across the platform to increase interaction on the selected platform.

Because these educational MOOCs are open to learners who want to learn, it is essential to consider the characteristics of the learners who participate in MOOCs. It is necessary for universities to understand the drop rates of students and student retention in MOOCs (Hmedna, El Mezouary, Baz, & Mammass, 2017). Hmedna et al., (2016) focused on the use of neural networks within big data to determine learning patterns for learners in MOOCs. The purpose of their study was to increase student satisfaction and interaction across the online courses in MOOCs, and in doing so, provided models for addressing the dropout problem. Maintaining student participation will have a broad impact on learning across educational platforms (Joseph, 2017). Understanding students' interaction across educational platforms helps to characterize student learning patterns that will help reduce dropout rates and require less teacher intervention Ramesh, Goldwasser, Huang, Daume, and Getoor, (2014). Qiu et al. (2016), indicated that constructing a probabilistic model to modify students' behaviors provide teachers with a probability model to help students interact with courses through educational platforms on MOOCs.

To address the dropout rate in online courses using MOOCs there is a need to develop mechanisms in machine-learning that can predict the dropout of students using a thorough understanding of the types of data and algorithms that help accurate predictions for dropout rates of students (Xing & Du, 2018). He, Bailey, Rubinstein, and Zahang (2015) proposed a machine learning framework using support vector machine (SVM) algorithm to predict dropout rates of

students in MOOCs from clickstream data. Kloft et al.(2014) indicated that the support vector machine algorithm helped diagnose the problem of MOOCs dropout rates.

Problem Statement

Wang, Yu, and Miao (2017) indicated that the strategies computer science educators need to use to prepare machine-learning datasets for predicting dropout rates of students in massive open online courses have not been established. Many students around the world use MOOCs for learning and to exchange knowledge or experiences (Alumu & Thiagarajan, 2016). MOOC courses are high-quality and are offered for free by the best universities in the world and are avaliable to everyone, even in developing countries (Welsh & Dragusin, 2013). The high enrollment of students in MOOCs is misleading. Less than half of the learners enrolled in MOOCs actively engage in their courses, while the other learners either drop the course or do not participate, which contributes to the high dropout rate of students (Hone & El Said, 2016). There are no exploratory studies of a similar scope within big data analytics and machine-learning that identified the different prediction strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in MOOCs (Xing et al., 2016). The problem addressed in this study was to identify the strategies computer science educators need to use to prepare machine-learning datasets for predicting dropout rates of students in massive open online courses.

The machine-learning challenge remains unclear because computer science educators do not have the machine-learning datasets necessary to predict the dropout rates of students in MOOCs (Kloft et al., 2014; Xing, Chen, Stein, & Marcinkowski, 2016). The problem addressed in this study was within the field of educational technology and machine-learning. Computer science educators continue to have difficulties with their ability to provide the needed

functionality within MOOCs. Furthermore, the complexity of growing technology demands is increasing the number of student dropouts in MOOCs.

Purpose Statement

The purpose of this qualitative exploratory study was to identify the predictive models or algorithms used to predict dropout rates of students in MOOCs, the MOOC experiences, and the data elements in the machine-learning datasets for predicting the dropout rates of students in MOOCs. This study also highlighted the importance of big data and machine-learning in the field of educational technology, such as MOOC online courses. With the increasing amount of data in all fields that led to the usage of data warehouse storage, there is a need for developing techniques, algorithms, and models within machine-learning to perform in-depth data analysis to get an accurate prediction on different problems such as dropout students in MOOCs. Identifying the required data elements within MOOCs will aid in preparing machine learning datasets that help address the problem of student dropout rates. The population of this study was computer science educators who have successfully addressed the different strategies needed to prepare machine-learning datasets for predicting the dropout rates of students in MOOCs. These computer science educators were recruited from LinkedIn. LinkedIn is a social networking website designed to provide services to business professionals and allows the sharing of professional and personal information with users, and keeps an online list providing professional online contacts (Skeels & Grudin, 2009). This population was appropriate because these computer science educators were capable of providing their personal views on the strategies needed to be used to prepare machine-learning datasets for predicting the dropout rates of students in MOOCs. In this dissertation, the researcher sought to capitalize on the

twenty-five participants' expertise and competencies in the field of information technology and machine-learning in terms of the courses in MOOCs.

Research Question

This exploratory research focused on the following question: "What are the different strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses?."

Propositions

Several propositions based on existing literature underlie this study. First, there was a critical need to examine the effectiveness and shortcomings of MOOCs and student retention in these courses. Secondly, computer science educators are seeking to identify the type of models and algorithms in machine learning for predicting dropout rates of students. Thirdly, computer science educators need to understand what types of data elements that should be included in the data sets used by these algorithms or predictive models. Finally, computer science educators need a clear understanding of how to increase the interactions in MOOCs. Due to these critical concepts, there is a need to address the problem of dropout rates of students in MOOCs and understand the strategies computer science need to use to prepare machine learning datasets (Wang, Yu, & Miao, 2017). The MOOC platform presents many challenges for future work, such as the high dropout rate of students as well as ineffective assessments of the performance of students in courses on different MOOC platforms. These challenges require a clear understanding of the nature of the MOOCs performance, and future research should explore the position of MOOCs and how they can be supported to be sustained (Chen, 2014). That led to the current study of exploring the strategies computer science educators need to use to predict dropout rates of students in MOOCs

Conceptual Framework

Computational learning theory is a branch of computer science in the field of artificial intelligence, machine learning, and statistics, which helps in determining the bounds of theories, algorithms, and human ability to analyze data and identify patterns and rules from the data (Angluin, 1992). This concept helps to understand the boundaries of the use of machine learning in computer science, and to identify the possibilities available to use the different types of machine learning algorithms or predictive models (Mitchell, 1997). Based on the previous theory, the performance of algorithms and data elements have become necessary to know the features for each algorithm.

The gap in scholarly literature noticed while investigating different techniques or algorithms for predicting dropout rates of students in MOOCs, that there was not an optimal algorithm. The machine learning techniques and algorithms were unclear and were not able to make predictions of student dropout in MOOCs. It was important to explore the type of algorithms to identify which type of algorithms outperform the others. In order to improve the MOOC experience, there is a need to identify the type of algorithms, understand the features for each algorithm and identify the elements of the data sets.

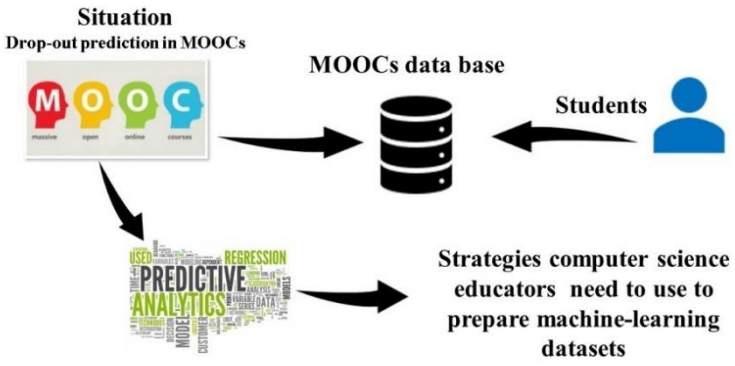

Figure 1: Conceptual framework of exploratory qualitative research study

The increase in the volume of data that have been collected by different MOOC platforms are not enough to improve the performance of students' in MOOCs (Adamopoulos, 2013). The log data from the forums in courses and all the activity data are essential to analyze the issues in MOOCs (Elo et al., 2014). That has led to the need to understand the many types of data, algorithms, and the MOOCs experience that help to address the leakage of students in the MOOC platforms (Clow, 2013). The comparisons between machine learning algorithms need more exploration to find the best performing algorithm. Umer, Susnjak, Mathrani, and Suriadi (2017) used different machine-learning algorithms in their experiments to identify which type of algorithms outperformed the others.

The literature review led to identifying a gap in knowledge, which is the need for exploring the strategies computer science educators need to use to prepare machine-learning datasets for predicting the causes of dropout from MOOCs. Based on the research question, the conceptual framework in this study will be restricted by four concepts: strategies of machine-learning, MOOCs, computer science educators, and the data elements in MOOCs datasets.

The exploratory research in this study can potentially affect future MOOCs and give computer science educators clear vision to find the types of data associated with learner's performance in MOOCs. Identifying these data elements will help them predict student dropout rates. The results of this study may be useful to computer science educators through the use of machine-learning, which can help them identify the various categorical predictive models and algorithms used to predict the student dropout problem.

Assumptions / Biases

In qualitative research the assumption is that the sample in the study represents the population that the researcher wishes to make inferences about (Simon & Goes, 2013).

Qualitative research is more helpful for exploring institutional phenomena, explaining participants' thoughts and opinions, and generating conditional ideas and theories that direct concern to appropriate situations(Hathaway, 1995). Those assumptions became the study endeavor, from the methodology used to the nature of the questions directed(Kivunja & Kuyini, 2017) This study was concerned with the machine learning, MOOCs, and datasets needed by computer science educators to address the dropout rates of students in MOOCs. The researcher assumed that the participant's response was based on their experiences and backgrounds to address dropout rates students in MOOCs.

Reducing bias in qualitative research ensures that the sample of the study meets the researcher's objectives (Smith & Noble, 2014).By focusing on participants' responses in the study before the questionnaire questions data was analyzed participants may not remember and report events and their responses accurately (Smith & Noble, 2014). To avoid this bias, I took special care to find the target sample related to the research topic and machine learning to achieve the objectives of the study by accurately searching for a convenience sample from the LinkedIn website. The subjects sampled must be able to inform important facets and perspectives related to the phenomenon being studied (Sargeant, 2012). The researcher focused on the academic backgrounds of the participants to achieve the goals of the study. In order to reduce bias and establish the validity of the research study, the researcher met two participants who met the criteria to fill out the questionnaire. According to their feedback some survey questions were revised. Three experts in machine learning revised the survey questions before starting to collect data.

Significance of the Study

Researchers are particularly interested in understanding why students are dropping out of MOOCs (Kolowich, 2013). They are trying to determine who registers on MOOC platforms in the classes where the percentage of dropout rates reaches at least 90% (Onah, Sinclair, & Boyatt, 2014; Rivard, 2013). The high dropout rate is a primary concern particularly to those who have spent time and effort and did not complete their studies, and the educators would have also spent their time to help the students, evaluating assignment and giving feedback (Gütl, Rizzardini, Chang, & Morales, 2014). Belanger's and Thornton's (2013) study indicated that approximately 12,000 students from more than 100 countries enrolled at Duke University via the Coursera platform. During the first week, approximately 8,000 of these learners logged in. This implied that 4000 (33.3%) students dropped out with out even trying the course. Educause (2013) also observed a difference between the registered students and the actual students who attended the courses. Wen, Yang, and Rose (2014) focused on the students participating in the discussion forum. Nearly 60,000 students participated in this study but the actual students who continued their courses were only about 2,500 students, or approximately 5%. To best address the student dropout phenomenon at different stages of the course, an instructor should focus on the learners who are at-risk by using association analysis in data mining (Srilekshmi, Sindhumol, Chatterjee, & Bijlani, 2016). With high dropout rates observed in many current large-scale online courses, there is an interest to understand the mechanisms of datasets that are able to predict student dropout**.** The significance of this study is that it explores the strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses. Identifying the data elements in the prediction data sets and the best prediction algorithms is essential to predict the dropout rates of students in MOOCs.

The results of this study will help MOOCs designers to improve the content and quality of courses in MOOCs thus enhancing the student experience. This study is mainly based on three principles: MOOCs experience, predictive algorithms of machine learning, and data sets. The three principals will provide computer science educators with a clear understanding of how to increase the interactions in MOOCs. Also, to understand the types of algorithms and datasets needed for predicting drop rates of students.

Delimitations

The delimitations help define the parameters of a future research study by identifying the scope of individuals or data collection (Creswell, & Creswell 2017). A delimitation in this study was that the participants were from LinkedIn, and they had to have at least one year of experience in machine learning. Also, the participants in this study had to answer three prequalifying questions to qualify before starting the questionnaire. These qualifying questions were used to determine if the participant had at least one-year experience in computer science, one-year experience in machine learning, and one-year experience in MOOCs. If the participant did not meet one of three criteria, the system stopped him/her from continuing to begin the survey.

Limitations

Limitations are to identify potential weaknesses in the research study that may be out the control of the researcher (Simon & Goes, 2013). This research study was subject to several limitations, which must be acknowledged. First, the research was limited to a sample of 25 participants from specific machine learning groups in the LinkedIn community. A second limitation is the inability to generalize the results of this study due to the small sample size and the different characteristics of the population, as well as the participants were from different

geographic regions. Thirdly, the answers of participants were not optimal; there were variances in the responses due to the level of experience of the participants. Finally, the duration of the survey was open presented a limitation of time. The study questionnaire was available for a limited time. Although a longer period of data collection may have presented more reliable results, there were time constraints that required the researcher to limit the data collection period.

Following Brown (2009), the researcher ensured the data instrument had reliability and conducted a pilot test and had two participants that met the criteria to be able to answer the questions in the questionnaire. Then additional fine-tuning of the questionnaire occurred upon completion of the pilot test. The survey monkey questionnaire contained three prequalifying questions before the participants were allowed to start the survey. The three prequalifying questions were placed to ensure that every participant have a background in machine learning, computer science, and MOOCs. Some study participants did not meet the three pre-qualifying criteria but all had a computer science background.

Definition of Terms

This section contains a summarization of definitions of the most important terms used in this research.

Machine Learning (ML): Machine-learning is a branch of artificial intelligence in computer science, and an algorithm that helps programming applications to obtain accurate outcomes to use data for predicting better in the future (Michalski, Carbonell, & Mitchell, 2013).

Big data analytics (BDA): Big data analytics is a term used to describe the volume of data. The data types include structured or unstructured data to obtain better decisions and strategies business in the future (Xu, Frankwick, & Ramirez, 2016).

Information technology (IT): Information technology is a term related to computing technology such as the data, software, networking, hardware, and people work with these technologies (Melville, Kraemer, & Gurbaxani, 2004).

Massive open online courses (MOOC): Massive open online courses, is a relatively new education method which appeared in 2008. Massive Open Online Courses (MOOCs) are free online courses open for anyone to register. MOOCs present an affordable and manageable method to learn new skills, advance your job and give quality educational experiences in a glob (Chen & Zhang, 2017).

General Overview of the Research Design

Qualitative research is an umbrella term for a group of activities that aim to explore interpretations, experiences, produce of the social world, and human understanding (Hammersley, 2012). The use of an exploratory design in qualitative research is appropriate for the researcher to learn from the participants of the study based on their experiences in the field. To do this, the researcher will ask participants questions and obtain narrative data related to the research topic. This type of research design focuses on problems not previously researched and considered nascent. This study used a qualitative approach to interpret the themes related to the three principals: MOOCs experience, algorithms or predictive model, and datasets in MOOCs. The researcher analyzed the data using Nvivo12 software to find the themes that answer the research question. The study data were collected using open-ended questions through an online questionnaire. The questionnaire consisted of one fixed alternative question for the participant to understand the use of the word “strategy” in this study and the rest of the questions was open ended questions to answer the research question.

Exploratory research aims to explore research questions and obtain new information (Mack, 2005). This type of research will not offer final solutions to existing problems; instead, it will help study a problem that has not yet been clearly defined. Also, it will identify opportunities that other studies may contribute to the solutions (Zikmund, Babin, Carr, & Griffin, 2013). The exploratory research method was appropriate for this study to help explore the strategies of machine learning that computer science educators need to use to make predictions and prepare machine-learning database to address the problem of MOOC student dropout rates. The findings of this study will help computer science educators who have experience in information technology, computer science, machine learning, and MOOCs to explore the algorithms, data elements, and improve the quality and contents of courses for predicting the dropout rates of students in MOOCs.

The selection criteria for participants in the research study required that they should have at least one year of experience in computer science, one-year experience in machine-learning, and one year experience in MOOCs. This research study had essential steps that were performed during the data collected stage .The exploratory approach of data analysis included the following steps: (a) collecting the data from the participants; (b) organizing the data from participants' answers; (c) coding the responses; (d) extracting themes from the data and finding specific labels; (e) separating the different themes in groups and finding similarities and differences, and establishing data relationships that helped understand the study themes (Scheider, Ostermann, & Adams, 2017)

The online questionnaire was conducted to obtain valuable data regarding the participants' experiences with machine-learning in computer science and their relevant views, impressions, and thoughts. The methodology used in this study was selected to explore the

participants' views and interpretations of the questionnaire (Boynton & Greenhalgh, 2004). Using structured and unstructured questions help the researchers to get data quality and clear answers. This method is more ethical to get responses from participants (O'Cathain & Thomas, 2004).

The previous studies indicated that there is no similarity to the current study (Gardner, Brooks, Andres, & Baker, 2018). This exploratory approach was the appropriate method to add related current study that contributes to clarifying other strategies computer educators need to explore

Summary of Chapter One

This chapter contained a discussion of the research topic and the background of the research with an explanation, problem statement, and a purpose statement. In this chapter the researcher addressed the research question with a definition of terms and explained the conceptual framework; theoretical perspectives; assumptions; biases; the significance of the study; and determines the limit of study in delimitations.

Organization of Dissertation

This dissertation is organized into five chapters. Chapter 1 contains an overview of the research topic and identifies the research method to help answer the research question. Chapter 2 provides a review of previous studies related to the research topic. It ends by identifying the conceptual framework used to research the study topic. Chapter 3 focuses on the research design; population, study sample; and instrumentation. It highlights the procedures and practical steps to collecting data to ensure validity and reliability of data, as well as, the specific steps the researcher took to protect the participants from harm. In Chapter four, the presentation and discussion of findings are provided. Finally, chapter 5 provides as explanation of the findings and conclusions.

CHAPTER TWO

Chapter two provides a comprehensive literature review of previous studies that covered the use of machine-learning datasets for predicting student retention. In addition, topics of interest to data scientists which include the many techniques and models that were used to predict and explore why students drop out from MOOC platforms were researched. The literature review indicated that computer science educators continue to have difficulties with providing the needed machine-learning datasets for predicting dropout rates in massive open online courses (MOOCs) due to the lack of understanding on how to employ the necessary prediction models (Bates, 2018). The challenge has remained to determine the strategies to reduce the dropout rates which can be predicted by the use of machine learning and elements of data.

Current studies discussed different predictive models or algorithms in machine learning without focusing on the essential algorithms that should be used to solve the problem of MOOC dropout. This gap in scholarly literature was noticed by Umer, Susnjak, Mathrani, and Suriadi (2017) while comparing machine-learning algorithms to evaluate which type of algorithm outperformed other algorithms. Therefore, this study's goal was to explore which predictive models or algorithms are currently used by computer science educators for predicting dropout rates of students in MOOCs. Finally, this chapter provides the body of literature and how it relates to the research question.

History of Machine-Learning Datasets for Predicting

This section includes the history of machine-learning and present examples of how data experts use machine-learning tools to analyze data in different fields. This section also provides ideas about why machine-learning is essential for this study.

The literature review showed that there is a relationship between machine-learning and big data. The concept of machine-learning began in the early 1950s (Hidalgo, 2018). In 1959,

Arthur Samuel defined machine-learning as how computers can learn, by the use of programming languages, and conduct data training and testing to predict outcomes without human supervision (Bhavsar, Safro, Bouaynaya, Polikar, & Dera, 2017). Jordan and Mitchell (2015) defined machine-learning as a computer program that can learn from experiments.

Machine-learning is one of the branches of artificial intelligence that train predictive models using test data to predict an outcome (Kotsiantis, Zaharakis, & Pintelas, 2007). This machine learning system can produce positive results based on previous learning to predict future learning (Holzinger, 2016). Many versatile machine-learning algorithms can be used in different fields to obtain predictive models that help make the right decision (Obermeyer & Emanuel, 2016). Using machine-learning analysis with different algorithms helps to increase the predictive model's efficiency by performing many activities that can be performed by the user and the variables related to the prediction model can be identified (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015). These algorithms can improve themselves by the results provided in the testing and by extracting valuable information and predicting them through observations and interpretation of the datasets (Nasrabadi, 2007).

Machine-learning is a topic developed from the study of pattern recognition and computational learning theory in artificial intelligence (Marr, 2016). The ideology of machine-learning is to reduce the use of human interaction and give the artificial intelligence full ability to process information independently. The scientific community uses machine-learning and different algorithms for predicting models and algorithms to get valuable information (Siegel, 2013). In recent years, machine-learning and predictive models have been popularly used in different fields.

The study areas that machine-learning and predictive models benefit include robotics, graphical models recommender systems, natural language processing, computer vision, and computational intelligence (Jordan & Mitchell, 2015). Machine-learning has many stages in respect to the form of the models. These stages consists of a training stage, a validation stage, and a testing stage (Sapp, 2017). Many algorithms in machine-learning primarily depend on the three stages includes: training, validation, and testing of data for predicting. Each algorithm has its method in the prediction process to obtain valuable information (Snoek, Larochelle, & Adams, 2012).

Machine-learning is widely used in many software applications to improve the work of experiments where the program can understand user behavior to obtain user-desired predictions (Sheth, 2013). Machine-learning applications have been used in many fields, for example, recognizing handwriting for preventing fraud on bank checks as well as speech recognition on different devices. In addition, companies rely on the use of machine-learning to improve their relationships with their customers and to improve their business (Nasrabadi, 2007). With increasing applications of machine-learning, there is also the interest of machine learning in financial institutions to assist in the analysis of past transactions to predict risk situations that could be potentially be damaging for the bank (Moin & Ahmed, 2012).

Machine-learning in computer science is an algorithm. The algorithm is a set of instructions that can be implemented to obtain appropriate outputs and get accurate results for predicting (Kotsiantis et al., 2007). For example, a programmer can develop an algorithm after identifying a particular problem to find solutions or find an alternative algorithm that works effectively (Haykin, 2009). With the increasing use of databases on computers and the Internet,

there is a need to use machine learning to extract data and algorithms automatically to obtain information that helps to capture predictive models (Márquez, Cano, Romero, & Ventura, 2013)

Existing computing devices have digital data. The customer recorded data in business, commerce, and education provides personal information digitally that can be used in machine-learning for predicting (Witten, Frank, Hall, & Pal, 2016). The machine-learning prediction process may not explain the outcomes one hundred percent, but a predictive model can help interpret the results by extracting patterns of data to help understand the process to predict the future (Rasmussen, 2004). The application of machine-learning on large databases is called data mining. There is a similarity between machine-learning processes and data mining processes, both of them require a search in that data to extract patterns and valuable information (Witten et al., 2016).

Data mining is the process of classifying, within machine-learning, sets to know patterns and build relations to solve problems during data analysis and allow companies to forecast future goals (Jackson, 2002). Data mining is an active branch of machine-learning that helps to establish relationships by extracting those relationships through the use of the fields in the relational database (Cheng, Tan, Gao, & Scripps, 2006). Many data mining techniques meet the needs of users such as association rules, clustering, and classification analysis (Agrawal, Imieliński, & Swami, 1993).

Data mining helps to explore hidden patterns in big data sets and to identify relationships between elements of data that can then be analyzed and extracted for valuable information (Sotiris, & Kanellopoulos, 2006). One of the most common data prediction techniques used in data mining is the market basket. The market basket model of data mining helps to determine

customer purchase behavior by reviewing items that are frequently purchased together (Berry & Linoff, 1997).

Data mining and machine-learning have become important in educational research and are used in an attempt to understand education data and improve the educational experience (Mackness, Mak, & Williams, 2010). MOOC is a revolution in the field of educational technology and has been associated with the area of science in education focusing on the challenges and problems facing the educational process through the MOOC platforms (Kop, Fournier, & Mak, 2011). Accelerated growth in machine-learning in the MOOC platforms encourages interested researchers to focus on data analysis methods to help address the dropout problem (Cook, 2016).

The Influence of Machine-Learning Datasets for Predicting

Due to the importance of machine-learning for predicting in many fields, predictive models play an effective role and have been used recently to support getting an appropriate decision for future (Libbrecht & Noble, 2015). Data analysis plays an essential role in many companies where data can be analyzed using machine-learning and big data in distributed computing platforms (Fisher, DeLine, Czerwinski, & Drucker, 2012). The most suitable platforms for data analysis in machine-learning and big data are Apache Spark and MapReduce Tools (Dean & Ghemawat, 2008). Mustafa, Elghandou, and Ismail (2018) reported the importance of using applications of machine-learning. They provided a different platform that predicts, with high precision, the execution time of Structured Query Language (SQL) inquiries and machine-learning applications executed by Spark. The results of their experiments to predict the execution of Spark jobs by using the proposed platform with accuracy greater than 90% for SQL queries and greater than 75% for machine-learning jobs.

In the health field, machine-learning techniques were used in Pima Indian diabetes datasets to improve trends and detect patterns with risk factors using R data manipulation tool. Kaur and Kumari (2018) utilized machine-learning algorithms to get accurate results for diabetes predictions. One of the effective algorithms of machine-learning is a neural network which was first proposed in the 1950s as an approximation of human neural cells. A neural network works with data training and testing for the results after training (Haykin, 2009). This algorithm has many extensive applications, for example, being used in materials science and engineering which helps to understand the properties of materials and how it can be improved to obtain predictive information (Cassar, de Carvalho, & Zanotto, 2018). (Feng, Zhou, & Dong, 2019) used neural network regression with a small dataset that included 487 data points in materials science to predict solidification defects. They found that a pre-trained and fine-tuned neural network algorithm shows better generalization performance to get predictive information for the properties of the material defects (Feng et al., 2019).

Wildfires are considered a worldwide disaster affecting many aspects of life such as the economy, natural environments, and health. Datasets can be provided to an artificial intelligence (AI) prediction model the can help us better understand the many factors that cause wildfires (Yaakob, Mustapha, Nuruddin, & Sitanggang, 2011). To solve this problem, research efforts had been conducted in order to observe, prevent and predict wildfires using different artificial intelligence techniques and strategies such as machine-learning, big data, and remote sensing (Dixon, Goodrich, & Cooke, 2008). Sayad, Mousannif, and Moatassime (2019) used a data mining and machine-learning algorithm utilizing satellite images to monitor wildfires, which provided a huge amount of data to predict the occurrence of wildfires and to help with preparing for such disasters.

In the oil and gas industry, the transient multi-phases wellbore flow is a challenging physical problem (Makarova, Mikhailov, & Shako, 2014). Such flows can happen through the startup process when the well is open and oil can flow to the surface for the first time. In this process, the fluids utilized to drill and finish the well are being displaced by the hydrocarbons beginning the well formation (Theuveny et al., 2013). Spesivtsev et al. (2018) developed a predictive model for multiphase wellbore flows using an artificial neural network algorithm to obtain the results. The results of their study showed the accurate predictive model as a simulator for transient wellbore flows mainly the bottomhole pressure.

In education technology such as e-learning systems (MOOCs) have advantages for learning but suffer from high student dropout rates. MOOCs generate huge amounts of data that raise significant challenges for the educational technology field to predict dropout rates of students (Kotsiantis, 2009). Burgos et al. (2018) study suggested using the logistic regression algorithm in machine-learning to experiment with a group of over 100 students from many distance learning courses. This predictive model helped to reduce the students dropout rate by 14% by proving students tutors (Burgos et al., 2018).

Massive Open Online Course (MOOCs)

MOOCs are a relatively new educational technology with a history dating back only to 2008 (Anderson; Fini,2009) A MOOC is a massive online course deliverd using the Internet. Many computer science classes are offered free via MOOC platforms from the best universities and experts around the world (Bonk, Lee, Reeves, & Reynolds, 2015). The strategic objective of MOOCs is to open up education to the public as much as possible (Zheng, Chen, & Burgos, 2018). It is an essential tool for achieving the sustainable development plan for the future, which is set in the different MOOC educational platforms such as Coursera and Udacity in the United

States, Europe, and many universites in developing countries (Castillo, Lee, Zahra, & Wagner, 2015).

The growing interest in MOOCs recently attracted the attention of higher learning institutions. Significant investments have been made by higher learning institutions to replace traditional courses with MOOCs (Yuan & Powell, 2013). Companies such as Microsoft, Google, and Dell Technologies are looking to create a modern learning environment using MOOCs to provide free educational and training courses for all learners without limits. There is an increase in academic partnerships between educational institutions and companies (Dyumin & Andrianova, 2016). The growing interest in the MOOC platforms and the widespread involvement of students who take these courses has led to significant interests in making MOOCs work more effectively (Chen, Feng, Zhao, Jiang, & Yu, 2014)

The most successful MOOC initiatives in the United States are Coursera, Udacity, and edX (Shafiq, Ashraf, Mahajan, & Qadri, 2017). These and other platforms receive extensive media attention (Pappano, 2012). MOOCs play an active role and revolutionized the field of education technology and provided opportunities for modern, advanced and open learning. This, in turn, allowed thousands of students from different countries around the world to access free, high-quality educational platforms (Anderson, 2013). China's interest in MOOCs was first observed in April of 2015. Since then, there has been a widespread interest in Chinese higher education to implement a comprehensive deployment of open courses online. So far, hundreds of education platforms have been built in China attracting many learners to take a variety of courses. In the United States and Canada, a study was conducted in 2014 under the same interest using more than 400 universities and 1,600 courses with 15 million students registered with MOOC (Yuan, Powell, & Olivier, 2014)

India has active participation in this modern education technology and has the second largest user base in the world (Kaveri, Gunasekar, Gupta, & Pratap, 2015). Approximately 800 thousand Indian users registered in the Coursera platform, accounting for 8% of the total number of users (Kaveri, Gunasekar, Gupta, & Pratap, 2016). The improved user engagement in MOOCs has motivated instructors and the academic community to take a solemn expression at the promises that the MOOC phenomenon drives with it. The followers of MOOCs find the technology engaging and will make a change in the use of education technology (Pappano, 2012)

In Europe, there are initiatives for MOOC platforms such as Futurelearn and Iversity. Many courses in Europe are offered on MOOC platforms through the Open University in the UK, a leading university in open education (Baturay, 2015). Iversity is a useful learning platform in Germany and has the potential to offer the advantages of the European Credit Transfer System for training and education (Cormier, 2010). Education and Training European Commission (ECTS) awards are provided, after the examinations of trainees, to obtain jobs. Institutions offer this educational feature and there are real efforts to expand this possibility further during platforms of education (Baturay, 2015). There have been few initiatives for MOOC courses production in Turkey (Baturay, 2015).

According to researchers in Thailand, Japan, and Malaysia, who discuss platform adoption and how to prepare MOOC educational platforms for Thailand and neighboring countries, many people in developing countries are interested in this modern educational phenomenon (Thaipisutikul & Tuarob, 2017). The goal of MOOCs in Thailand is to build a partnership network with universities for sharing human resources and educational resources to minimize investment. Thailand Cyber University (TCU) is an example of a collaboration project between 47 universities which give a sharable degree (Thaimoocs, 2017).

In 2017, there were about 94,000 MOOCs offered by more than 800 higher education schools helping 81 million students all over the globe (Shah, 2018). The "MOOC tsunami" phenomenon, due to the ability of this offering mode to break out of the geographic and time barriers, has come to be defined as “redefining” the field of higher education (Joseph & Nath, 2013). With the extensive use of MOOC platforms, there is a focus on the assessment of these platforms to identify challenges using assessments (Xiao, Qiu, & Cheng, 2019)

MOOC Student Dropout Rates

Researchers are particularly interested in understanding why students are dropping out of MOOCs (Kolowich, 2013). They are trying to determine who registers on MOOC platforms in the classes where the percentage of dropout rates reaches at least 90% (Onah, Sinclair, & Boyatt, 2014; Rivard, 2013). The high dropout rate is a primary concern particularly to those who have spent time and effort and did not complete their studies, and the educators would have also spent their time to help the students, evaluating assignment and giving feedback (Gütl, Rizzardini, Chang, & Morales, 2014). Researchers have considered the weaknesses of MOOCs compared to traditional courses and have conducted many analyses to discover the real causes of massive dropout rates of MOOC students in attempts to reduce attrition rates (Sachdeva, Singh, & Sharma, 2015).

Harju, Leppänen, and Virtanen (2018) highlighted the problem of dropout students in MOOCs and that it should be determined whether it is accurate. They discussed the causes behind the high student dropout rates and how the interaction limitations of MOOCs contributed to the low achievement rate (Harju, Leppänen, & Virtanen, 2018). The main goals of massive open online courses (MOOC) are to support knowledge through free high-quality learning to create new knowledge through various users’ interactions with the provided platform, and to enable research on education (Nawrot & Doucet, 2014). Student's skills and the quality of the MOOC also play a

role in the high dropout rate problem. Students who need the essential competencies, even if they study in a well-designed MOOC, will probably drop out during the course. Furthermore, learners with high skills in an ill-structured MOOC will likely fail to complete the course (Abeer & Miri, 2014).

Balakrishnan and Coetzee (2013) exploratory study focused on the behavior of students who dropped out of the university. Yang, Sinha, Adamson, and Rose (2013) explored student dropout behavior in MOOCs using a case study involving a recent Coursera platform to develop a survival model that allowed measuring the importance of factors extracted from that data on the student dropout rate. Wen, Yang, and Rose (2014) focused on the students participating in the discussion forum. Nearly 60,000 students participated in this study but the actual students who continued their courses were only about 2,500 students, or approximately 5%. Wen, Yang, and Rose (2014) relied on two types of analysis methods: traditional discussion forums and Mixed Membership Stochastic Block model (MMSB), in which data analysis is treated as a separate network. This model can track the way learners move between sub-communities during their participation (Rosé et al., 2014).

Researchers face many challenges that make it difficult to obtain accurate results from students who drop out of the MOOCs (Bates, 2018). Implementing a visual comparative analysis helped understand students' behaviors in a comparative study of two courses on a MOOC platform (Rivard, 2013). The method of a visual comparative analysis helped to determine the dropout rate among students. Attention to students' attitudes toward MOOCs is one of the most important predictors of success in the MOOC platform. The percentage of student dropout differs according to the structure and type of the course. Instructors encourage and make

recommendations to students to help them interact with the platforms of MOOCs (Schaffer et al., 2016).

Assessing students is difficult due to the growing number of learners around the world (Hew, 2016). Recent studies used an assessment planner on the MOOC platform such as ODALA. This type of planner aims at reducing learners dropout rates (Lynda & Dahmani, 2016). Lynda's and Dahmani's (2016) study explored adaptive planning for various educational activities by identifying and profiling learners across MOOC platforms. Breslow's et al (2013) study involved conducting a large-scale analysis of video courses in computer science courses to determine factors contributing to the student dropout rate. Four courses on the edX platform experienced high dropout rates when students watched longer videos or re-watched sessions. This experiment produced positive results that determined the contributing factors to dropout rates in video courses (Kim et al., 2014).

Belanger's and Thornton's (2013) study indicated that approximately 12,000 students from more than 100 countries enrolled at Duke University via the Coursera platform. During the first week, approximately 8,000 of these learners logged in. This implied that 4000 (33.3%) students dropped out with out even trying the course. Educause (2013) also observed a difference between the registered students and the actual students who attended the courses. Researchers need to study the reasons why some students become uncomfortable in participating in MOOCs and to help instructors rethink at least some of the elements of their courses and to know the actual reasons for the high dropout rates of students.

These studies show that student enrollment is misleading in terms of dropout rates because many students do not attempt the course at all. The dropout rate across MOOC platforms is high (Onah, Sinclair, & Boyatt, 2014). When the incompletion rate is high, this presents a

significant obstacle and challenge to the transformative potential of MOOCs (Khalil, & Ebner, 2014). Many theories exist to define those low MOOC completion rates. For example, adult learners may find MOOCs challenging because the courses are extensive, meaning that one course can include hundreds of thousands of students (Schulze, Leigh, Sparks, & Spinello, 2017).

Because of these massive course sizes, the idea of MOOC may not allow a single educator to direct or help all of the learners (Bryant, 2015). Cost may also be prohibitive for some students to enroll in MOOCs. Some MOOC providers, such as the University of British Columbia, pay a small group of educational partners to provide aids that can facilitate the MOOC discussion forums (Engle, 2014). As a result, these aids and methods compatible with MOOCs led to the achievement of the course outcomes (Bates, 2018)

Big Data and Analytics

Big data analytics is examining big and varied data sets to explore information including hidden patterns and to determine unknown correlations to make informed decisions (Bhadani & Jothimani, 2016). Big data introduces many challenges and opportunities for companies and institutions to extract valuable information since the data must be analyzed and processed (Srinivasa & Bhatnagar, 2012). The analysis of data in big data plays an important role in improving processes and functions; the benefits can be demonstrated by aggregating both internal and external data (Maltby, 2011).

Today, there is extensive interest in big data around the world (Bhadani & Jothimani, 2016). Many companies and organizations rely on big data. They analyze big data and use their findings to make the right decisions. The strategies and processes of organizations and companies that gain a greater understanding of big data experience positive and productive impacts (Hagen et al., 2013).

Ghofrani, He, Goverde, and Liu (2018) indicated considerable interest in big data analysis (BDA) particularly in areas such as engineering and railway transportation. Machine-learning and big data analysis process data sets to extract the knowledge and valuable information contained in the data. That helps to build and study algorithms that can provide information about data (Kohavi, 1998). The purpose of these algorithms is to help develop a model and identify inputs to make decisions and appropriate predictions for future work (Nasrabadi, 2007).

Data mining is the process of classifying within big data sets to identify patterns and build relations to solve problems during data analysis and allow companies to forecast future aims (Jackson, 2002). Data mining is an active branch of big data and helps to establish relationships or extract patterns and the relationships between fields in a relational database (Cheng et al., 2006). Many data mining techniques meet the needs of users such as association rules, clustering, and classification analysis (Agrawal et al., 1993).

Data mining helps to explore hidden patterns in big data sets and to identify relationships between elements of data that can then be analyzed, and valuable information extracted (Kotsiantis & Kanellopoulos, 2006). One of the most common data prediction techniques used in data mining is the market basket. The market basket model of data mining helps to determine customer purchase behavior by reviewing items that are frequently purchased together (Berry & Linoff, 1997).

The Role of Big Data and Analytics in MOOCs

Data has an active role in many areas especially in teaching and learning (Joseph 2017). This growing interest makes focusing on the educational process in MOOCs within the domain of big data important (Misra, 2018). MOOCs need to develop market strategies to attract higher numbers of students. Big data can help in market analytics in an academic area (García &

Secades, 2013). Big data has become common in recent years, which refers to datasets whose size is beyond the understanding of standard database software applications to dealing with, store, analyze, and manage the data (Manyika et al., 2011).

By using big data tools and predictive modeling techniques, MOOC platforms can provide the essential information the academic institution needs to improve customer experience and the overall experience of the platform (Manyika et al., 2011). This type of predictive modeling is required to know about the students' participation completion rate on the MOOC platforms (Yuan et al., 2014). Considering the rise of MOOC platforms around the world, there is increasing interest to adopt this phenomenon in educational institutions.

Although traditional education cannot be replaced entirely by MOOCs, the use of big data analysis (BDA) can provide new and improved modes of learning on MOOC platforms (Morabito, 2015). Big data, in the field of education, is an unexplored research area. Big data resources in the MOOC are mainly composed of three parts: course materials, user behavior, and MOOC users. Understanding the patterns from the gathered MOOC platforms data can improve the MOOC learning process (Wu & Chen, 2015).

There are many personal user data on the MOOC platform and that data is associated with each user. Because of the significant number of MOOC records, valuable information content is stored in the databases of the MOOC servers. The collection of such data leads to the use term "big data" (Zheng & Yin,2015). MOOC platforms provide more choices and opportunities to learners than traditional educational systems. However, some students cannot take advantage of these new learning models; but, with big data technologies, the learning efficiency of students can be significantly improved (Chen, Mao, Zhang, & Leung, 2014) Although much research has been conducted about MOOCs, it has had an insignificant effect on

shaping an education that uses big data technologies. Thus, MOOC data proves to be ineffective and inadequate (Reich, 2015). With the increasing number of MOOC platforms, data types are increasing, and there is a demand for big data analytics

The Coursera platform should encourage retention of students enrolled during the course by collecting more data which can be used by analysts (Ghobrial, 2014). This will help researchers to leverage big data to study student behavior and students' interactions with the MOOC platforms (Baker, Evans, Greenberg, & Dee, 2014). Moreover, analysts can use the vast amounts of data such as age, gender, and most viewed pages to understand MOOC student behaviour (Liyanagunawardena, Adams, & Williams, 2013). Increasing participants on MOOC platforms have aroused the interest of many researchers and universities to analyze the collected MOOC data (Wong, 2016). Many universities, including the University of California, intend to prepare future innovations and conduct much research in the field of big data to study and analyze information to meet the needs of users (Goral, 2013).

Recent studies referred to machine-learning methods such as regression, decision trees, and other prediction models to make decisions and predict dropout rates by using elements of data from two Coursera MOOCs (Chaturvedi, Goldwasser, & Daumé, 2014). Data analysis and machine-learning helps researchers understand learners and provides context to improve the learning environment (Siemens & Long, 2011). Big data on MOOC platforms requires sophisticated BDA. BDA has become essential for online services to extract the patterns in the data (Ghoting et al., 2011).

Analyzing Big Data to Predict Student Dropouts in MOOC

Data mining and big data have become important in educational research and are used in attempt to understand educational data and improve the educational experience (Mackness et al., 2010). MOOCs are a revolution in the field of education technology and have been associated

with the area of science education, focusing on the challenges and problems facing the educational process through MOOC platforms (Kop et al., 2011). Accelerated growth in big data in the MOOC platforms encourages interested researchers to focus on data analysis methods to help address the dropout problem (Cook, 2016).

One of the most critical challenges MOOC platforms face is the high student dropout rate; most students who enroll during courses do not finish the session (De Waard et al., 2011) Researchers have identified the reasons for the failure or incomplete courses—each reason has different meanings and context (Kaur & Kaur, 2015) Taking advantage of big data is in the common interest of researchers. The use of a massive online database on the internet to predict the successes and failures of students during the educational process on the MOOC platform can provide valuable information (Baker, 2010). Recently, there is extraordinary interest in new technologies in the education field and within big data that can help students to continue their studies through these platforms (Campbell, DeBlois, & Oblinger, 2007). The strong relationship between big data and the challenges facing MOOCs, specifically the student dropout rate, are effective and can contribute to improving the MOOC platforms (Yu & Wu, 2015).

The Elements of Data Used by Algorithms to Predict Drop Out

The different structure of classes on the MOOC platforms presents different types of data, such as online learning behavior, discussion postings, and videos watching (Wang & Baker, 2015). These types of courses may lead to student dropout. The collected MOOC data contains information related to students' participation across the MOOC platforms, which helps researchers explore students' performance (Abubakar & Ahmad, 2017).

The data elements used by algorithms or predictive models to study the dropout rates of students in MOOCs are very important to increase the performance of these algorithms((Bates, 2018). For example, studying the behavior of the students in a certain period of time can help

evaluate the educational process by focusing on the type of teaching and course presentation (Dekker, Pechenizkiy, & Vleeshouwers, 2009). Studies suggest predictive models on big data that focus on student learning behavior over a period of month or more during the educational process in a MOOC can predict dropout rates (Gardner & Brooks, 2018). Gardner and Brooks (2018) noted that the dropout rate relates mostly to student characteristics such as marital status, number of dependents, active duty students, the elderly, and gender (Niemi & Gitin, 2012).

During the MOOC platform data analysis, it is necessary to extract datasets that have attribute variables for each student from the completed curriculum that could be applied to predictive models by implementing machine-learning and multiple algorithms to predict the student dropout rate (Márquez et al., 2013). One of the standard methods to extract the pattern of data is a focus on behavioral data in the MOOC database that can help to predict the student success in the courses (Kizilcec, Piech, & Schneider, 2013). Also, it is possible to combine different student data such as demographics, behavior, and enrollment in data mining to implement an unsupervised machine-learning algorithm technique (Sinha, Li, Jermann, & Dillenbourg, 2014).

Wu and Zheng (2016) focused on the extraction of descriptive student information from training courses and course registration records in the edX platform, as well as user behavior while considering the privacy of data for students (Liang, Yang, Wu, Li, & Zheng, 2016). Focusing on different usage procedures in a MOOC helps to improve the learning process and increases motivation through the use of different methods and various use-cases in MOOCs (Kloft et al., 2014). The diversity of different data sources including assignment grades, demographic, and clickstreams play a decisive role in obtaining information on the student dropout phenomenon (Sinha, 2014). Clickstream and video data as well as student behavior, like

video interaction, can be used to determine the dropout rate among students (Nagrecha et al., 2017). The MOOC platforms often contain videos to facilitate the learning that can help predict the level of students' interaction based on the value of the content of these videos. These course offerings on the Coursera platform provide the types of data needed by a prediction model (Brinton, Buccapatnam, Chiang, & Poor, 2016).

In recent years, researchers have shown interest in finding ways to use student data by applying supervised learning methods to dropout prediction models using general features extracted from student behavior logs (Li et al., 2016). Given the dropout rates for students across the MOOC platforms, many researchers have analyzed stream server logs on the MOOC platforms associated with viewing the video lectures, studying the material, and completing various quizzes and homework-based assessments to predict students who will dropout from the course (Jiang, Williams, Schenke, Warschauer, & O'dowd, 2014). Ren, Rangwala, and Johri (2016) study provided predictive models to determine the future performance of students based on a personalized linear multi-regression (PLMR) approach while the student completes graded activities on the platform. Previous studies also suggested that an approach in the context of prediction depends on graded activities within courses offered from the university's online platforms. Also, the data extracted from the Moodle learning management system via click-stream server logs allow observing students in real time (Ren, Rangwala, & Johri, 2016).

The quality of the prediction models gives more significant opportunities for application on the different data from the MOOC platforms. Machine-learning models use the extracted data from learner activity logs from different courses through MOOC platforms such as wiki course, video, and the course forum (Sharkey & Sanders, 2014). Activity logs can be separated into different time periods, week or day, so that they can be represented in a predictive model, such as

the time series model (Taylor, Veeramachaneni, & O'Reilly, 2014). Because track data is typically logs and counters, we need to use a time series model to predict participation records in a MOOC (Crossley, Dascalu, McNamara, Baker, & Trausan-Matu, 2017).

When classifying students' activities on a MOOC as sequential data, it is easier to represent the sequence structure. The features extracted for students through MOOC platforms depend on the sequence of short activities rather than on individual activities. This helps in predicting the time series model (Brinton et al., 2016). Different predictive models can be applied from classical machine-learning models to obtain the features extracted from the students' log such as decision tree and logistic regression (Crossley, Paquette, Dascalu, McNamara, & Baker, 2016).

These types of machine-learning models are commonly used as prediction models. When analyzing data in the same course but in real time, these models are not suitable (Chaplot, Rhim, & Kim, 2015). With the increasing diversity of predictions models and elements of data in the MOOC platforms, a study is necessary to use DropoutSeer technology, which uses visual analytics in big data to understand reasons for dropout based on different data categories such as most viewed pages, forum posts, and assignment records (Chen et al., 2016)

Boyer and Veeramachaneni (2015) indicated that data recorded while learners are interacting with the MOOC platform provide a different possibility to build predictive models. To address this challenge, they designed a set of processes that take advantage of knowledge from both previous courses and previous weeks of the same course to make real-time predictions on learners behavior. Kloft er al (2014) proposed a machine learning method based on support vector machines (SVM) for predicting dropout between MOOC course weeks. This predictive model can present an approach that works on most viewed data. Also, this algorithm can take the

weekly history of student data into account and thus is able to notice changes in student behavior over time.

Data analysts can perform different predictive models or algorithms on big datasets generated by the educational technology system, such as big data delivered via the Coursera MOOC platform. Also, using elements of data on whether learners completed the course. Furthermore, data analysts can download data from the same course system after the course concludes (Wang & Baker, 2015) With the increasing size of data in education technology, it become necessary to use big data tools. Laveti, Kuppili, Ch, Pal, and Babu, (2017) indicated that to perform learning analytics in MOOCs, the researchers need to develop a workflow using Apache Spark, a scalable in-memory computing framework. The data from the edX MOOC platform has been used for experiments. The data used contained information on students from 39 courses. Laveti, Kuppili, Ch, Pal, and Babu, (2017) compared various machine learning algorithms where they found that the Baseline model was the best predictive model to predict student dropout rates.

Helpjour-seeking behaviors of learners in MOOCs are one of the data elements used for predicting dropout rate. The key help-seeking mechanisms within MOOC environments, such as discussion forums, and other elements of data should be explored (Kennedy, Coffrin, De Barba, & Corrin, 2015). Halawa, Greene, and Mitchell (2014) indicated that they use unique student-level administrative data from different courses across a range of controls all fielded on one widely used MOOC platform, Coursera. This method will help to understand the learning process in MOOCs at the student level.

Analyzing video lectures, weekly quizzes, and peer assessments data from a specific course using unsupervised learning methods can help to determine students dropout rates and

understand that completion rate problem in MOOCs (Ye & Biswas, 2014). Because of the gradual nature of the decline in the use of the MOOC, there have been an interest, in recent years, to build predictive models using a variety of data elements such as most viewed pages, discussion forum data and quiz scores to find the best designs to address the problem of student dropout (Xing & Du, 2018). Balakrishnan, and Coetzee (2013) indicated that using different data sets, such as lectures, each broken up into several 10-20 minute videos and contain ungraded multiple-choice practice problems, four homework assignments, each containing differing amount of programming problems, and 4 graded multiple-choice quizzes, in Hidden Markov Models help predict student retention as well as infer general patterns of behavior between those students that complete the course. That led different studies to investigate the different prediction algorithms and datasets to improve the performance of the prediction models that help the effectiveness of courses across the MOOC programs (Brooks, Thompson, & Teasley, 2015).

To best address the student dropout phenomenon at different stages of the course, an instructor should focus on the learners who are at-risk by using association analysis in data mining (Srilekshmi, Sindhumol, Chatterjee, & Bijlani, 2016). Older methods like surveys, interviews, focus groups, and observations for data collection do not accurately analyze student dropout, and these methods are time-consuming (Xing, Kim, & Goggins, 2015). There is now a possibility of identifying students at risk. Due to the high dropout rates, there is an urgent need to use big data tools , learning analysis, and educational data mining in order to quantify students' different behavioral characteristics(Goggins, Xing, Chen, Chen, & Wadholm, 2015).

Machine learning techniques, such as analytical learning and educational data mining, can make accurate predictions that help analyze data and identify low levels of student interaction with the MOOC courses (Xing et al., 2015). The different techniques of learning

analytics and educational machine-learning can analyze the low-level trace data regarding students' interactions within a course and with other students (Xing & Goggins, 2015). We need to indicate a preliminary expectation of identifying students who drop out of MOOC, including an ability to meet the requirements of interventions from educators that can be implemented early in an effective course (He, Bailey, Rubinstein, & Zhang, 2015).

With high dropout rates observed in many current large-scale online courses, there is an interest to understand the mechanisms of datasets that are able to predict student dropout. These datasets, such as most viewed pages and weekly history of student data, become increasingly important in the ability to notice changes in student behavior over time (Kloft et al., 2014). Wan, Yang ,and Rose (2014) used sentiment analysis to explore mining collective sentiment from MOOC forum posts in order to monitor students' trending opinions towards the course by using a survival modeling technique to study various factors' that impact attrition over the course weeks. They observed a significant correlation between the sentiment expressed in the course forum posts and the number of students who drop the course (Wan, Yang, & Rose, 2014).

To predict student dropout in MOOCs by using a different dataset, which was used previously and classified whether the elements of data indicates dropout or not, can help to understand the learning process and evaluate the models or algorithms to obtain better performance (Xing, Guo, Petakovic, & Goggins, 2015). Brinton et al (2016) indicated that the performance of the machine learning algorithmes was mainly based on increasing the elements of data in onder to get accurate predictions compared to other algorithms.

The role of learning analytics is to analyze and collect the data about the learners and their contexts to understand the environment of learning by using education data mining for the purpose of optimizing learning. (Marie, Mingyu, & Barbara, 2012). Amnueypornsakul, Bhat, and

Chinprutthiwong (2014) indicated that the diversity of data, such as discussion forum and most viewed pages data streams, enable mining of student behavior in MOOCs. The prediction process on student data in the MOOC platforms can happen by reviewing common features that are related to the quiz attempt/submission. Those features that capture the interaction with various course components are found to be reasonable predictors of attrition in a given week. For example, researchers noted that students' action during the first week of a course can be used to predict their subsequent performance (Gitinabard, Khoshnevisan, Lynch, & Wang, 2018).

Using another machine-learning algorithm , such as logistic regression, has the ability to predict the probability of students earning certificates for completion of the MOOC thus predicting dropout rates of students (Jiang et al., 2014). Strategies of machine-learning focus on the process of automatically discovering useful patterns in large quantities of data (Witten et al., 2016). Fei and Yeung (2015) indicated that using features that reflect student activities from different datasets in MOOCs, such as the courses materials including quizzes and the lecture videos, and using different temporal modules such as a recurrent neural network (RNN) model with long short-term memory (LSTM) will help predict dropout rates of students in MOOCs. By comparing all the temporal models, the results showed that the LSTM network outperformed other models for prediction.

Accordingly. Ramesh, Goldwasser, Huang , Daume, and Getoor (2014) regarded one week as a time unit and combine all student activities within a week for each student to predict dropout rates by using machine-learning models such as probabilistic model connecting student behavior and class performance. In the era of MOOCs, there is an increase in student dropout rates. Since MOOC data can be collected efficiently, data relevant to the dropouts depends on objective data, not self-reports Breslow et al (2013) used panel data methods to analyze the

relationship between performance on each assignment and the student's subsequent performance on the following assignment. The elements of data, such as most viewed pages (clickstream), can help to create a picture of performance over the entire class. Rodrigues et al. (2016) analyzed a 5100 learner MOOC course participation levels data in activities and interactions in a virtual forum using the hierarchical grouping method and the non-hierarchical grouping method (k-means). The research results showed detailed information of three different groups that describe behavioral characteristics in relation to activities and interactions via a forum for predicting dropout rates of students in MOOCs (Rodrigues et al., 2016)

Learning analytics and educational data mining are two important fields that help improve the teaching and the learning experience. Liang et al. (2016) used data collected from 39 courses from the XuetangX platform, which is based on the open-source Edx platform, to predict drop-out rates in MOOCs by using different machine learning algorithms such as SVM, Logistics Regression, Random Forest and Gradient Boosting Decision Tree. The results showed that by using Gradient Boosting Decision Tree (GBDT) model the predication accuracy reached 88% (Liang et al., 2016). Chen and Zhang (2017) indicated that using statistical analysis on students' behavioral data will help predict dropout rates students in MOOCs. The results showed that the dropout prediction system achieves high effectiveness at finding dropout students. The system was inspired by a statistical study of correlations between student behavioral data and course dropout.

Conceptual Framework

Machine-learning and big data models proposed several data categories of MOOC platform data that can be used to predict student dropout rates (Xing & Du, 2018). For example, MOOC forum data, from various courses across the MOOC platforms, were used to mine

collective forum posts to monitor students' trending towards the courses (Wen, Yang, & Rose, 2014). However, researchers are interested in exploring other non-standard data categories and non-scalar features in the future to predict the dropout rates (Khalifa et al., 2016). This study focused on exploring the strategies of machine learning within MOOCs, such as elements of datasets used by algorithms or predictive models to predict the dropout rates of students enrolled in MOOCs. The use of machine-learning models and elements of datasets to predict the problem of student dropout within the MOOC platforms is essential. The analysis of MOOC data is useful to improve learners interaction within the MOOC platform. For example, a linguistic analysis of the MOOC forum data can explore important indicators for foretelling dropout of students (Kloft et al., 2014). Therefore, it is essential to exploit the strategies of machine-learning such as algorithms and predictive models. Also, datasets elements can enable computer science educators predict dropout rate of students in MOOCs (Mduma, Kalegele, & Machuve, 2019).

The research at hand is a qualitative exploratory study that aimed to understand the data elements related to students in MOOCs and the different techniques, models, and algorithms in machine-learning that computer science educators need to use for predicting dropout rates of students. The conceptual framework of the study is illustrated in Figure 1 in chapter one. Based on the research question, the conceptual framework in this study was restricted by four concepts: strategies of machine-learning, MOOCs, computer science educators, and the data elements in the MOOC datasets.

In recent years, MOOC platforms have emerged as an alternative to the local community to learn by using modern technology and increase learner's general knowledge (Conole, 2016). However, a new problem emerged in MOOCs, which is student attrition and the dropout of students. That led Alario, Estvez, Prez, Kloos, and Fernndez (2017) to criticize the lack of

educational value in MOOCs, and the high dropout rates, which in many cases are over 90-95% of enrollees.

The nature of exploratory research requires a focus on detailed data from students (Stebbins, 2001). Elements of data in MOOCs will help to use the bests algorithms of machine-learning to reduce dropout of students in MOOC (Kloft et al., 2014). Future research mainly focus on increasing the interaction in MOOCs such as instructors feedback, course design, and optimize the prediction model's performance by using different elements of data for predicting MOOC student dropout (Xing & Du, 2018).

In recent years, there has been growing interest in the use of MOOC and educational technology online that attracted many learners around the world (Cohen & Nachmias, 2006).To provide a practical learning experience, this requires timely observation of the learning process. Despite the popularity of this modern educational phenomenon, the MOOC educational platform courses show higher dropout rates of students than conventional courses. Although many thousands of participants enroll in these courses, the completion rate for most courses is below 13%. (Onah et al., 2014). Umer, Susnjak, Mathrani, and Suriadi (2017) showed that logistic regression outperformed K nearest neighbor, random forest, and Naïve Bayes machine-learning algorithms with the highest accuracy.

In recent years, there has been an increasing interest in the use of tools, techniques, algorithms, and predictive models in machine-learning to determine the best performer of these technologies to address the student dropout rates in MOOCs (Liang et al., 2016). For example, learning analytics (LA) and educational data mining (EDM) were used to determine what type students learning activity data need to be analyzed to get valuable information. Many machine-learning algorithms have been implemented such as logistic regression, support vector machine

(SVM), random forest, and gradient boosting decision tree (GBDT). The results showed that the GBDT machine-learning algorithm has the highest accuracy and can be used by the instructors to get an idea and vision into the behavior of students that are expected to drop out from the MOOC courses (Kashyap & Nayak, 2018).

Machine-learning techniques have extensively been applied in the field of educational technology in MOOCs. Hong, Wei, and Yang (2017) used multinomial logistic regression (MLR), support vector machine (SVM), and random forest (RF) algorithms to predict student dropout. Combining all these techniques, the study results showed that student dropout can be predicted with a 97% accuracy. The variety of different prediction models in machine learning have an effective role for predicting dropout rates of students in MOOCs. For example, Moreno, Alario, Muñoz, Estévez, and Kloos(2018) implemented supervised learning and unsupervised learning algorithms by using datasets such as forum messages in MOOCs that can be analyzed to detect patterns and learners' behaviors. Dalipi, Imran, and Kastrati (2018) focused on the problem of student dropout in MOOCs by providing an overview of the types of predictive models and compared them. Predictive machine-learning models included logistic regression, deep neural network, support vector machine, hidden Markov models, recurrent neural network, natural language process technique, decision trees, survival analysis, and Bayesian network .The results showed that the Logistic regression algorithm has been the most frequently used technique

Computational learning theory

Computational learning theory is a branch of computer science in the field of artificial intelligence, machine learning, and statistics, which helps in determining the bounds of theories, algorithms, and human ability to analyze data and identify patterns and rules from the data (Angluin, 1992). This concept helps to understand the boundaries of the use of machine learning

in computer science, and to identify the possibilities available to use the different types of machine learning algorithms or predictive models (Mitchell, 1997). In this study, most of the algorithms and predictive models in machine-learning that can be used to address the problem of student dropout rates in MOOCs were identified. This study explored the types of machine learning algorithms or predictive models and the date elements of the datasets that need to be used to address the student dropout rates problem in MOOCs.

Summary of Literature Review

Student dropout rates in MOOCs is the fundamental problem that should be addressed within education technology by using strategies of machine learning and datasets. Understanding the data elements of the datasets needed by computer science educators in machine-learning was important. Also, exploring the types of machine-learning algorithms, models, and techniques that assist in the prediction process of student dropout rates in MOOC is important.

This exploratory study included the experiences of members of machine learning groups from LinkedIn that have experience in machine learning, computer science, and MOOCs. By considering the available body of literature and identifying previous studies, the knowledge gap identified in this chapter can be reduced. According to (Kloft et al., 2014) despite there are detected elements of data, such as the most viewed pages, when combined with forum data achieved a 15% increase in prediction accuracy of student dropout rates in MOOCs, still there is a need to explore more data elements such as country, operating system, browser, and other types of datasets. Additionally, there is a need to explore different machine-learning techniques and prediction models that involve more elements of data such as students' background information, prior experience in online learning, and test grades. These insights can help improve prediction models or algorithms, elements of datasets, and the MOOC experience for predicting dropout

rates of students (Xing et al., 2016). Chapter three presents the research tradition and design for the study and includes the population and sampling technique, instrumentation, and research reliability and validity. Additional information includes in chapter three include ethical considerations, data collection, and data analysis.

CHAPTER THREE

The problem addressed in this study was the strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses which have not been established (Kloft et al., 2014). The main concern was that computer science educators continue to have difficulties with their ability to provide the needed functionality within a MOOC to improve the performance of students to finish their courses (Adamopoulos, 2013). Computer science educators continue to have difficulties with providing the needed machine-learning datasets for predicting dropout rates in MOOCs due to the lack of understanding of how to employ the necessary prediction models (Bates, 2018). The challenge remained in how to determine the strategies to reduce the dropout rates which can be predicted using machine learning and elements of data. This study used an online questionnaire to learn how and why computer science professionals use machine-learning, educational technology elements and MOOCs for predicting student dropout rates in MOOCs. Furthermore, participants shared their insights about the strategies computer science educators need to use to prepare machine-learning database for predicting student dropout rates in MOOCs.

The purpose of this qualitative exploratory study was to explore the strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in MOOCs. This type of research does not intend to offer final solutions to existing problems but helps to study a problem that has not been clearly defined yet and identifies the expectations that other researchers need to contribute to the solution (Zikmund et al., 2013). The researcher selected this research design because it can depict the type of data required for computer science educators to predict student dropout rates. The qualitative study explored what types of algorithms and predictive models are needed within MOOCs to predict

student dropout rates. Also, this study helps to understand the nature of the predictive models and algorithm types used in the MOOC platforms. Understanding the nature of the predictive models and algorithm types used in MOOCs will facilitate the creation of machine-learning models that predict the probability of student dropout and improve the learning process in the MOOC platforms.

This chapter discusses the research traditions of the study. Specifically, this chapter describes the researcher's selected design, creation of the questionnaire questions, the study's theoretical basis, and the instruments of the research used in the collection of data and eventual analysis.

Research Tradition

A research methodology is used to find an appropriate framework that will help to answer the research questions of the research study (Teherani, Martimianakis, Stenfors-Hayes, Wadhwa, & Varpio, 2015). The researcher used exploratory qualitative research because there is a specific and obvious problem that needs to be better understood. This type of research is usually conducted to study an issue that is not clearly defined (Creswell, 2014).

There are no exploratory studies of a similar scope within big data analytics and machine-learning in MOOCs because the prediction strategies computer science educators need to use to prepare machine-learning datasets for predicting the student dropout rates in MOOCs are still undefined (Xing et al., 2016). The qualitative methodology also enables a deeper understanding of social relationships and humanitarian evaluation (Creswell, & Creswell 2017). The research in this study involves posing questions and analyzing the responses based on an inductive building of different concepts into general themes (Creswell, 2014). In qualitative research, researchers seek to understand the context of information and data from the participants and thus collect data

through a questionnaire. Other information is obtained through the researcher's experiences and background related to his concentration; this helps the researcher interpret the data. In chapter two the researcher identified the dropout rate problem among students enrolled in the MOOC platforms. The results of this study will help computer science educators explore the elements of data that can be used by machine learning algorithms or predictive models for predicting student dropout rate in MOOCs.

The qualitative methodology was appropriate for the research study because Northcraft (2017) reported that exploratory research can use open-ended survey questions. These types of questions can extract meaningful responses from the participant. The responses to the questions can become rich and explanatory to meet the needs of the research problem. Also, a qualitative research method helps the researcher to get the flexibility to investigate participant's responses (Jamshed, 2014). The qualitative data in this study included three fixed alternative questions, and the remaining questionnaire questions were open-ended. The study participants were from LinkedIn who are experts in the field of computer science, machine-learning, and MOOCs.

The quantitative methodology was not used for this study. Quantitative research uses a structured method such closed ended-questions, and analyzing the variables (Mack, 2014). This study used open-ended questions. Also, the mixed methods methodology was not used for this research study because it compares qualitative and quantitative data. This type of mixed methodology is useful when the researcher needs a deep understanding to observe the differences between qualitative and quantitative findings (Wisdom & Creswell, 2013).

A research design is used to help the researcher obtain the appropriate answers to the research questions (Creswell, & Creswell 2017). It explains the complete method that a study used, including the particular sections of the study, in an understandable and coherent design.

Using this approach, it ensures the research study successfully discusses the problem of the research. It also builds the framework for the gathering, analysis, and measurement of data (Yilmaz, 2013).

Based on the selection of the qualitative methodology, an exploratory design approach was used because there is a need to obtain many insights and learn further information about the research problem (Creswell, 2014). This type of research does not intend to offer final solutions to existing problems but helps to study a problem that has not been clearly defined yet and identifies the expectations that other researchers need to contribute to the solution (Zikmund et al., 2013). The exploratory research design helps to explore research questions and provides an opportunity to conduct many research studies in the future (Penwarden, 2014).

An exploratory qualitative approach was appropriate for this research study because this type of research seeks to explore a particular situation to evaluate the outcomes that answer the research questions (Baxter & Jack, 2008). Exploratory research gives a general view of the situation related to the research topic with the need to conduct many research studies in the future (Labaree, 2013). The exploratory research also gives preliminary and general outcomes of the research study (Yilmaz, 2013). The samples used in exploratory research are mostly limited and relatively small, so a generalization of the sample is limited.

The researcher considered these design options: case study, ethnography, phenomenology, and grounded theory. The case study design was not used for this research study because it is used r to study many cases comprehensively. It also uses different data sources to study different cases (Baxter & Jack, 2008). Case study design helps the researcher to explore organizations or individuals and find the relationships, study the communities, and programs (Yin, 2003).The ethnography design was not used in the research study because this

approach primarily focuses on describing and interpreting culture-sharing groups (Eriksson & Kovalainen, 2015). Also this method draws from sociology and anthropology (Morgan-Trimmer & Wood, 2016). The forms of data collection in this method include primarily interviews and observations, but sometimes include other sources of data (Creswell, & Creswell 2017).

The phenomenological design was not be used in the research study because this type of method draws from philosophy and philosophy education, and primarily focus on significant elements and textual and structural descriptions, and description of the essence of information (Creswell, 2014). Phenomenology research method describes a real phenomenon, and that could be the situation, events, and concepts or experiences. All these types of phenomenology need to be fully understood to conduct this type of research study (Padilla-Díaz, 2015).

The grounded theory design was not be used in the research study because this type of method focuses on the generation or the discovery of the theory (Glaser, & Strauss,2017). The grounded theory uses a systematic set of procedures to develop a specific phenomenon. Also to identify the key elements of that phenomenon to find the required results and the relationships of those elements to the context and process of the experiment (Charmaz, & Belgrave,2007)

Research Question

The research question for this study was, "What are the different strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses?"

Research Design

Exploratory research does not intend to offer final solutions to existing problems but helps to study a problem that is not clearly defined, and identifies the expectations that other researchers need to contribute to the solutions (Zikmund et al., 2013). LinkedIn is a social

networking website designed to provide services to business professionals and allows the sharing of professional and personal information with users, and keeps an online list providing professional online contacts (Skeels & Grudin, 2009). This study was conducted using LinkedIn contacts. The population for this study was LinkedIn professionals from LinkedIn groups, which are related to machine-learning in computer science. The LinkedIn groups were the Deep Learning, AI, Machine Learning & Machine Intelligent group, KD Nuggets Machine Learning, Data Science, Data Mining, Big Data, AI group, the Machine Learning and Data Science group, and the Artificial Intelligence, Machine Learning, Deep Learning group.

After considering the three qualitative design options, the researcher considered the requirements for this research study. To meet the study objectives and answer the research question, the researcher chose the exploratory design option. This design allowed me to explore strategies computer science educators need to use to prepare machine learning datasets

Sampling and Population

The population in scientific research is defined as the broader group of people onto which the researcher can generalize the results of the study (Silverman, 2016). The population of this study was computer science professionals who have successfully addressed the strategies used to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses. These computer science professionals were recruited from LinkedIn. The size of the LinkedIn groups, which form the population for this study was about 94 thousand machine-learning professionals. This population was appropriate because these computer science professionals within the LinkedIn machine-learning communities would be capable of providing views of the strategies computer scientists need to use to prepare machine-learning

datasets for predicting the dropout rates of students in massive open online courses (Khalil, 2018).

The sample size of a research study is defined as the number of individuals who will participate in the research study (Martínez-Mesa, González-Chica, Bastos, Bonamigo, & Duquia, 2014). The sample size for this exploratory research study was 25 computer science professionals within machine-learning (Guest, Bunce, & Johnson, 2006). The sample size was appropriate for this study because Irfan (2018) and Krone (2018) conducted related research studies using a sample size of 8 and 10 were used to answer the research question. The selection criteria for participants in the research study had at least one year of experience in computer science, one year in machine-learning, and one year in MOOC.

Sampling Procedure

A sampling procedure is a method of choosing a subset from a population to participate in the research study (Padilla-Díaz, 2015). Useful sampling approaches utilized in qualitative researchers intend to ensure the highest quality of essential and relevant information (Johnson & Christensen, 2008). To obtain a clearer idea of the research outcomes and why exploratory research helps in achieving a contextual understanding from interviews with participants it will require a deep understanding of sample of the study (Mayer, 2015).

After IRB approved the study, potential participants were contacted using emails obtained from their profiles from LinkedIn. Participants were chosen who can best answer the research questions and enhance the understanding of the research topic (Sargeant, 2012). The selection criteria for the participants in this research study were that each participant had at least one year of experience in computer science, one year in machine learning, and one year in MOOCs. Additionally, there were no constraints on participant demographics because the

LinkedIn community is open to the members who have experiences and backgrounds related to their interests. This study was limited to professionals who have their profiles on LinkedIn. The participants were selected according to the accessibility to the researcher and the relevance of the participants to the questionnaire questions of this study.

Potential participants were contacted by email requesting their participation. When a participant agreed to participate in this research study, the questionnaire link was emailed to the participant (Mayhew, 2017). After the participant clicked on the questionnaire link, they had to go over the informed consent form and consent to participate or decline to participate before proceeding (see Appendix B). All participants in the study had the right to withdraw from the study at any time. Also, according to Survey Monkey's IRB guidelines, its site requires for researchers to allow the participants not to respond to a question and be able to proceed to the next question (see Appendix D). The design of the study questionnaire included being respectful of the privacy of the participants in order to reduce the risks for the participants taking part in the study. Also, in Survey Monkey's IRB Guidelines, the researcher cannot enable IP tracking and enable SSL encryption (see Appendix D).

All documents, questionnaire through Survey Monkey, and additional notes remained in the researcher's possession (Creswell, 2014). In this study, a code name provided for each participant was used to maintain privacy. Only the researcher has known the true identity of the participants in the study. Moreover, all notes containing sensitive information will be destroyed by the researcher after five years, in accordance with Colorado Technical University policy.

Instrumentation

For qualitative research, the researcher is an implicit part of the research and must collect valid and reliable data (Suter, 2012). Qualitative researchers often serve as an instrument

(Creswell, 2014). Qualitative research use open-ended questions. The researcher took many steps to design the questionnaire questions to best answer the research question. These steps included ensuring participants' experiences and their relevance to the research study (Yin, 2015). This study used three fixed alternative questions and eleven open-ended questions designed to obtain a clear understanding of the research question of the research study.

For qualitative research studies, the researcher's part of the inquiry is an essential aspect that involved gathering valid and reliable data (Anney, 2014). Qualitative researchers frequently work as a key instrument (Yilmaz, 2013). The qualitative design of this study emphasized the precise and thorough use of questionnaire questions responses to explore the research question (Lewis, 2015). The researcher was involved with the primary instrument for collecting descriptive data (Creswell, & Creswell 2017).

The questionnaire instrument used explored the strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses (Dörnyei & Taguchi, 2009). Open-ended questions encouraged the participants to present descriptions of their professional experiences and give responses to use their own words (Harvey, 2011). Also, the use of a questionnaire was appropriate when the research question and questionnaire questions are clear, and the responses of participants are related to the research question (McGuirk & O'Neill, 2016). The questionnaire questions used for this study are presented in Appendix C. The researcher also met three participants who met the selection criteria and revised the questionnaire questions before starting to collect data.

After the participant clicked on the questionnaire link, they had to go over the informed consent form and consent to participate or decline to participate before proceeding (see Appendix B). All participants in the study had the right to withdraw from the study at any time.

Also, according to Survey Monkey's IRB guidelines, its site requires for researchers to allow the participants not to respond to a question and be able to proceed to the next question (see Appendix D). The design of the study questionnaire included being respectful of the privacy of the participants in order to reduce the risks for the participants taking part in the study. Also, in Survey Monkey's IRB Guidelines the researcher cannot enable IP tracking and enable SSL encryption (see Appendix D). The labeling of the captured data was used to ensure obtaining reliable information and understanding the response of the participant and respect the privacy of the participant.

Validity

The validity is one of the most important strengths of qualitative research, and it will ensure the accuracy of the results in the investigation from the perspective of the researcher (Leung, 2015). Validity is important for a qualitative study because it reflects the researcher's ability to obtain accurate results and convince readers of the validity of these details. A variety of strategies were followed to validate data such as verifying the identity of participants, using external auditing, clarifying the bias that the researcher brings to study, and clarifying negative information (Creswell, & Creswell 2017)

In other words, validity is the criterion for measuring the research study (Creswell, & Creswell 2017). There are two types of validity, external and internal; both play a role in determining accurate results. External validity aids in generalizing the results to the target population and the internal validity is the validation of the testing or exploratory process itself (Creswell, 2014). Validity was demonstrated by the researcher and ensured the validity of data because it is essential in measuring the research study, and the results can be meaningless if not accurately measured. The results, if inaccurately measured, cannot be used to answer the

research question research (Thomas & Magilvy, 2011). To ensure the validity of the study, the researcher was focused on the examination of the member checking and external auditors and work on the insights needed by the study (Daytner, 2006).

A pilot study or testing is essential in the survey of study to improve questionnaire questions and designate the people who will revise the questionnaire questions (well, & Creswell 2017). Also, a pilot study can be the pre-testing while collecting the data in qualitative exploratory research, that included the first stage before the questionnaire questions which is a pilot includes using in-depth interviews establish the issues to be addressed in a large-scale questionnaire survey (Van Teijlingen & Hundley, 2002).

Use peer debriefing in a qualitative study to improve the accuracy of the research. This process includes finding (a peer debriefer) who reviews the questionnaire questions so that the questionnaire questions in the research will give deep understanding with people other than the participants of the study to adds validity to the research study (Creswell, & Creswell 2017). Getting more information regarding questionnaire questions from peer debriefing gives an opportunity to understand search questions with different insights to achieve the goal of the research (Guba, 1981). To improve the quality of outcomes in qualitative research, peer feedback such as faculty members and postgraduate dissertation committee have an active role in obtaining accurate inquiries that support qualitative researcher (Anney, 2014). To reduce bias and establish the validity of the research study. The researcher met with three participants who met the participant criteria to fill out the questionnaire for the pilot study and revised the questionnaire questions before starting to collect data.

Dependability is defined as the ability of a different researcher to iterate the similar way that the study of this research used (Anney, 2014). Dependability is important for a qualitative

study because it mainly depends on the results of the research study and how to benefit from it by other researchers where other researchers can benefit from the research study for similar interpretations and results (Bradshaw & Stratford, 2010). For this research study, dependability was addressed by doing a reiterative study that used some of the strategies computer scientists need to do practical machine-learning algorithm by using the programming language to predict dropout rates in massive open online courses.

Credibility involves organizing the acceptability and believability of the findings of the qualitative study from the perspective of the participant in the research. Also requires the researcher to unquestionably associate the study 's outcomes with certainty to explain the precision of the research study's results (Nordhagen, Calverley, Foulds, O'Keefe, & Wang, 2014). The importance of using credibility in qualitative research lies to describe or understand the phenomena of interest from the participant's views; the participants are the accurate assistants of the credibility of the findings (Creswell, & Creswell 2017). To confirm the credibility of the research study, the researcher had reviewed the questionnaire of the study to evaluate the correctness of the responses given to each questionnaire question (Creswell, 2014). Debriefing or a peer review is the evaluation of the research and data process by someone who is familiar with the questionnaire questions to explore the results in the study (Bowen, 2008). Also, the credibility is part of the study, which focus on the trust of cooperation between the external auditor and the qualitative researcher, which supports the qualitative research by a discussion of the research and data process by someone who is familiar with the study to add credibility to the study (Creswell & Miller, 2000)

Transferability is the possibility of applying a particular research study in a similar environment where there is a similarity in the aspects of implementation (Lincoln & Guba,

2012). In qualitative research, transferability is primarily the responsibility of the one performing the generalizing. Also, qualitative research could enhance transferability by performing a general description of the study background and the assumptions that will be essential to the research (Perrier et al., 2014). Transferability for this research study required the researcher to study the geographic and demographic attributes of the populations of study to see if it can be implemented in different places.

Confirmability is the possibility of auditing a research study that relies on data collection where other auditors can verify the research study to be in the audit trail (Whitt & Kuh, 1989). Confirmability is important in a qualitative study because that will help the researcher work on validating data by focusing on different data during the data collection stage such as field observations, participant responses, and questionnaire questions (Anney, 2014). To accomplish confirmability in this research study, the researcher addressed the comprehensive view about the focus of the research the study and ensured getting new backgrounds and understandings from the participant's responses to the questionnaire questions. Furthermore, the researched also ensured to document ideas for investigating and re-review the data in this research study.

Reliability

Reliability relates to the ability of a measuring instrument to measure the same ability or quality in whole items on investigations and an analysis (Golafshani, 2003). Reliability refers to the consistency of the researcher's approach when compared with different research conducted and other projects (Bernard, Wutich, & Ryan, 2016). Reliability is the consistency with which researchers measure the results of an instrument by proposing minimizing imprecision in the analysis process (Knapp, 2015). For this reason, the target sample was clearly identified to the researcher. The participant's relationship to the research topic and the survey questions asked in

the interview was logical, and helpful to answer the research question (Creswell, & Creswell 2017).

Triangulation is a process that helps to enhance the credibility of the current study's results as well as to confirm the reliability of the data and connected the findings (Mayer, 2015). The reliability of the collected data that was obtained by asking open-ended questions increases consistency by grouping data into themes, ideas, thoughts, and groups (Suter, 2012). In summary, triangulation is to verify the validity of data and interpret it in a better way. It is essential that the researcher does the Survey Monkey questionnaire to capture all information and verify the data's validity, reliability, and relevance to the research topic. Utilizing a questionnaire after Survey Monkey helped the researcher to improve the accuracy of the collected data (Brace, 2018). Also, a triangulation method had used by using an online questionnaire evaluation from the participants to help identify themes. Also, conducting member checking enhances the reliability and validity of the data (Lincoln & Guba, 2012).

Member checking was conducted by the researcher to get informant respondent validation after the SurveyMonkey questionnaire process to ensure the validity of the study. After reviewing the data from the questionnaire responses, all online questionnaire model, and similarity of data were analyzed and reviewed by the researcher to ensure the most reliable and accurate account of what has transpired. The researcher focused on member checking, and that provided descriptive validation on the findings of the study (Creswell, 2014).

Using triangulation contributes to reliability and confidence by determining and describing gathered information to obtain similar and parallel data from which the researcher can benefit (Holtzhausen, 2001). Through triangulation, the researcher identified categories and themes using multiple resources. This process aided in interpreting the findings of the research

by the confirmation of using many procedures and managing the data in one study (Suter, 2012). Triangulation is a data analysis technique used in qualitative case studies to check and restrict reliability in a qualitative study, and that requires analyzing the research question from many thoughts to perform regularly in methods or data sources (Shenton, 2004). Triangulation is the process of using various measures to measure and set reliability in qualitative research that requires research question from many thoughts to achieve expected data source or another method (Heale & Forbes, 2013). For this research study, triangulation was accomplished by checking the open-ended question questionnaire. The survey questions were in the unstructured and structured format to help participants answer their questions with open answers and to express their views freely (Mayhew, 2017).

For ensuring internal validity of the research the essential strategies were used such as triangulation, and pilot study (Creswell, 2014). A pilot study was conducted by the researcher to remove bias from the questionnaire as well as establish a triangulation. The researcher met three participants who met the participant criteria to fill out the questionnaire and then revised the questionnaire questions data before starting to collect data. Qualitative research methods, along with the triangulation by comparison with another a questionnaire in a different were used in this study to determine the strategies computer science educators need to use to prepare machine learning datasets for a prediction dropout rate of students. This study used triangulation to increase credibility and the strength of results (Byrne, 2001). Also, the significance of the questionnaire design (Hassan, Schattner, & Mazza, 2006). For the research study, a pilot study was performed by requesting three participants to answer the questionnaire questions that align with the research study research question. Based on the participants' responses, further-fine tuning of questions that were required to ensure the questions align with the research question.

The researcher focused on some qualified experts with considering the selection criteria for participants in the research study was at least one year of experience in computer science, one year in machine-learning, and one year in MOOCs. The researcher found out what answers are expected, and the questionnaire questions data were clear to the participants. The measurement instrument in the research study was the questionnaire, and that was required a pilot study to identify any problem or flaw in the measuring instrument. Therefore, questionnaire questions were applicable to ensure reliability and validity in the research study (Srinivasan & Lohith, 2017).

Data Collection

The research question guides the data collection process necessary to capture the needed information for the research study (Bell, Bryman, & Harley, 2018). The research question was: "What are the different strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses?" The data collection technique was selected to answer the research question. The research questions were structured and open-ended questionnaire questions because it encourages participants to provide clear descriptions of their experiences (Harvey, 2011). Also, an open-ended questionnaire question was appropriate for the research question and questionnaire questions where the responses of participants were related to the research question (McGuirk & O'Neill, 2016). The methods of data collection should answer the research question of the study and capture answers to all the questionnaire questions the researcher asks the study's participants. (Most et al., 2003).

In this exploratory research, email was used to communicate with potential participants to confirm their willingness to participate and coordinate their availability. Participants received the

researcher's email and phone number. Communication via email and phone were used with participants until the questionnaire questions were completed. The participants of the of this study were computer science professionals who have at least one year of experience in computer science, one year experience in machine- learning, and one year experience in MOOCs. Twenty-five participants from the machine-learning professional's community on LinkedIn responded to the study questionnaire. The researcher has filled out an online questionnaire through Survey Monkey (See Appendix C). Survey Monkey, an internet survey tool, was the primary source for volunteers to submit their answers to the survey questions (Brace, 2018)

The researcher created a questionnaire and wrote the questionnaire questions of the research study. Also, the online survey was posted after identifying the participants in the study that met the criteria of participation in the study. The "Letter of Permission to use Site" to use Survey Monkey software for this research was obtained by contacting the Company and requesting their permission to use their software for this study (see Appendix A). The researcher explained the informed consent form to participants to know their rights for participation, including the right of privacy, before the questionnaire questions begin, the participants signed on informed consent the where each participant reviewed and read the informed consent before answering the questionnaire questions. The researcher began the coordination procedures for SurveyMonkey with the participants and started collecting data. The data collection included the following general process: (a) email LinkedIn Group of machine learning professionals (b) sign an electronic informed consent form (c) pre-qualifying questions to meet the participant criteria, (d) complete online Survey Monkey questionnaire

The researcher contacted the participants and explained the rationale and objectives of the study to ensure that all participants had a desire to participate in the study.

Moreover, the researcher selected 25 participants for this study. All 25 participants completed the Survey Monkey questionnaire. All the responses from the participants were analyzed, and responses were coded, and then themes were developed.

The collected data that was stored included personal information of participants as well as from the pilot study, the survey monkey questionnaire, and the participant's notes where stored on the researcher's personal laptop by using a secure password. The researcher will destroy the data and all documents after five years, in accordance with the 1974 Privacy Act, as well as the policies and guidelines of the Colorado Technical University and the IRB (Hanus & Relyea, 1975). In summary, data collected were occurred using Survey Monkey questionnaire and evaluated the responses of study participants.

Data Analysis

An exploratory qualitative methodology was selected over other qualitative designs because the focus of this research was to categorize and interpret themes (Creswell, & Creswell 2017). Exploratory research was selected because the focus of this research study was to interpret and describe themes for conceptual data analysis (Creswell, 2014). The methodology in this study was selected to explore the participants' views and interpretations of the questionnaire (Boynton & Greenhalgh, 2004). There are encouragement and support for researchers to use structured questionnaires when using open questions because this leads to data quality and helps in analyzing data and how to use it. This method is more ethical to get responses from participants (O'Cathain & Thomas, 2004). The researcher appropriately analyzed the experiences of the 25 participants from LinkedIn who have experience in machine learning. This exploratory study entailed obtaining details about the research topic by asking the participants questions to build clear and understandable information according to what the researcher has asked using

SurveyMonkey, and that met the objectives of the research which answered the research questions (Creswell, & Creswell 2017).

Qualitative data analysis methods are conceptual and relational (Suter, 2012). Conceptual data analysis involves establishes the presence of themes. Relational data analysis begins with the identification of present concepts and continues by looking for semantic relationships (Carley, 2014). Semantic relationships are set using thematic sections. Thematic units are high-level ideas interpreted from basic themes and patterns established in the qualitative data (Javadi & Zarea, 2016). The data analysis process involves the emergence of themes from the questionnaire questions transcripts.

Data analysis starts by preparing the collected information followed by data perusal, classification, and synthesis (Onwuegbuzie, Leech, & Collins, 2012). The data analysis approach for exploratory research in this study included the following: (a) compiling the data from the participants; (b) organizing the data by each participant, (c) coding the data (i.e., organizing the data by recognized categories), (d) identifying themes (i.e., the label attached to each recognized category), and (e) establishing data relationships (i.e., recognizing similarities and differences in the themes in order to condense or separate themed categories, as appropriate) (Irfan, 2018). Once this process was completed, the established themes categories were the findings of this study. The data analysis stage included the responses from the questionnaire questions and themes derived from the respondents. The initial stage of the data collection began by gathering the data. Next, the researcher reviewed the data with synthesis, perusal, and data classification to focus on the themes related to the research questions (Creswell, 2014). The questionnaire questions addressed the participants different individual experiences and opinions to explore the

strategies computer science educators need to use to prepare a machine-learning database to predict dropout rate of students in MOOCs.

For this study, an open-ended questionnaire served as the source information to identify themes, patterns, categorize, and coding. Coding is a way of indexing or categorizing the text in order to establish a framework of thematic ideas about it (Marshall & Rossman, 2014)
The coding was classified based on related words and phrases that the participants shared during the questionnaire questions into units of sense or themes. The researcher used the categorization process to provide a collection of themes. The technique that was used to translate data terms into the themes was NVivo software that analyzed the data to save time and to find more accurate findings. NVivo is software used in qualitative research and the analysis of unstructured data such as video, audio, or image data. NVivo is one of the most efficient applications that helps to analyze themes of data. In addition, NVivo software helps to improve accurate results in qualitative research (Zamawe, 2015).

NVivo software was appropriate because it met the requirements of the research topic for transcripts and survey questionnaire for data collection (Sotiriadou, Brouwers, & Le, 2014). After collecting the responses, the researcher used NVivo to analyses the responses to find themes in the SurveyMonkey collected data by exploring the SurveyMonkey of the study. The researcher obtained important findings and better recommendations to answer the research question. The researcher took essential steps to start the coding process using Nvivo12 included the following general process: (a) collected the data for analyzing the responses, (b) exported from SurveyMonkey to excel, (c) prepared an excel file, (d) scrubbed the data (e) imported the survey results into NVivo software for coding.

Themes were identified after obtaining responses from participants; this is considered to be one of the primary tasks when conducting qualitative research. These themes provided a clear and detailed understanding of the different views and opinions of the participants (Wildemuth, 2016). The researcher read the responses of the participants and compared them with the pilot test questionnaire questions for accuracy. The researcher also analyzed the data to obtain the essential themes. The combination of themes defined was based on the research question, and that established the research findings.

In data analysis, the possibly of data loss in the places the data was stored is a potential problem. Therefore, the researcher avoided losing or damaging the data and using reliable devices such as a laptop (Irfan, 2018). To avoid the loss of data, the researcher stored all questionnaire data in a secure place and in different places to secure the data.

Ethical Considerations

The ethical principles applied throughout the research process involved informing the participants of their rights. To create a safe environment before the participants had access to the questionnaire, they had to go over the informed consent form (see Appendix B), which explained the participant's right to end the questionnaire questions without providing a reason for stopping (Irfan, 2018). Every participant from the study as well as the pilot study signed an informed consent form before their questionnaire questions begin.

To ensure the highest level of ethical research, principles of the *Belmont Report* protocol was stated. The *Belmont Report* principles essentially focused on the efficiency and quality of study topics (Bromley, Mikesell, Jones, & Khodyakov, 2015). The vulnerable research population must be protected from possible exploitation (Rogers & Lange, 2013). Also, the three

principles of the *Belmont Report* protocol (i.e., autonomy, beneficence, and justice) was maintained (Owonikoko, 2013)

Researchers must ensure no harm comes to participants due to participation in a study (Miller, Birch, Mauthner, & Jessop, 2012). Risks must also be minimized to participants. The researcher must also avoid manifestations of deception and fraud as well as other types of unethical behaviors as recommended in the American Educational Research Association's Code of Ethics (G. Anderson & Arsenault, 2005). A consent form was completed by each participant in this study before the data collection began; this form is illustrated in Appendix A. The researcher will destroy the data and all documents after five years, in accordance with the 1974 Privacy Act, as well as the policies and guidelines of the Colorado Technical University and IRB (Hanus & Relyea, 1975).

To ensure awareness of the risks and benefits of the research study, each participant included in the pilot test was required to sign the informed consent form (see Appendix B). The consent form included (a) the purpose of the study; (b) the involvement of participates; (c) participation procedures; (d) the benefits of the research; (e) the risks of taking part; (f) costs and compensation; (g) confidentiality; (h) voluntary nature of participating; and (i) the rights of the participant to withdraw (Irfan, 2018).

Biases could happen due to preexisting information and experience with the topic (Creswell, 2014). Bias was mitigated by conducting an online Survey Monkey questionnaire, focusing only on the responses of participants and triangulation analysis of the responses from the formal pilot test (Creswell, & Creswell 2017). The researcher focused on participants who are experts and qualified with the considered specific selection criteria which were at least one year

experience in computer science, one year experience in MOOCs, and one year experience in machine-learning.

Summary of Chapter Three

The exploratory research design helped to answer the research questions (Penwarden, 2014). An exploratory qualitative approach was appropriate for this research study because this type of research explored a particular situation to evaluate the responses to the research questionnaire questions (Baxter & Jack, 2008). Data was collected from computer science professionals who have machine learning experience from LinkedIn groups using a SurveyMonkey online questionnaire. The participants in this research study are experts and qualified in the areas of machine- learning, computer science, education technology, and MOOC and were capable of answering the research questionnaire. Data analysis followed the general approach that was described by the researcher and included themes derived from the expected results (Suter, 2012). The study presented derived patterns and themes are representing the data that were collected to interpret the findings of this research study.

This chapter provided a discussion of the research design and the requirements of the research methodology. It also highlighted how the study was conducted, including data analysis and data collection methods. The study used an exploratory design and qualitative approach to obtain accurate and essential information from the participants. Also, the answers from that participants contained different opinions and experiences, in which the researcher explored the strategies needed by computer science educators to prepare machine-learning datasets to predict student dropout rate in MOOCs.

CHAPTER FOUR

This qualitative research study explored the research question, “What are the different strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses.” The purpose of this research study inquiry was to explore strategies computer science educators can use to prepare machine learning datasets and identify the different machine learning algorithms that can be used to predict dropout rates of students in massive open online courses thus helping student retention in MOOCs. The primary issue was that some educators in computer science in MOOCs might not have strategies necessary for predicting dropout rates. The questionnaire questions asked the study participants if they were pre-qualified to participate regarding their machine learning experience, types of modules or algorithms in machine learning they are familiar with, and their interaction with MOOCs. A presentation of the data collected from the survey Monkey, data analysis, and participants’ findings is presented in this chapter. Chapter four includes the results of data collection and provides the participants’ demographics, emerging themes, categories, the research findings.

Participant Demographics

The participants in this research study included computer scientists from the LinkedIn community website from different machine learning groups such as the "machine learning and data science group" and "KDnuggets machine learning, data science, data mining, big data, AI group." The participants had to answer three conditional questions to qualify before starting the questionnaire. These qualifying questions helped to identify if the participant had at least one-year experience in computer science, one-year experience in machine learning, and one-year experience in MOOCs. If the participant did not meet one of three criteria, the system stopped

him/her from continuing to begin the survey. Each member was initially contacted through LinkedIn and invited to participate by email. Once the participant accepted to be part of the survey, I sent out emails asking them to answer the survey. Twenty-five participants responded to the survey. The participants have different experiences in computer science. All the participants in this research met the criteria of computer science experience. Table 1 presents a summary of participant's years of experience in computer science.

Table 1

Years of experience in computer science

Years of Experience in Computer science	Number of Participants	Percentage of Participants Years of Experience in Computer science
15	2	8 %
14	1	4 %
13	1	4 %
12	5	20 %
10	3	12 %
9	2	8 %
8	2	8 %
7	3	12 %
6	3	12 %
5	2	8 %
5	1	4 %

All participants had experience in machine learning except for two. Table 2 presents a summary of the participants' machine learning experience

Table 2

The summary of the participant's machine learning experience

Experience in Machine learning	Number of Participants	Percentage of Participants
Yes	23	92 %
No	2	8 %

Also, there are differences in the levels of education among the participants. Some had Bachelors, Masters, and Ph.D.'s. Fourteen participants have a master's degree in Computer Science, eight participants have a Ph.D., and three participants have a bachelor's degree. Table 3 presents a summary of the participant's levels of education experience.

Table 3

The summary of the participant's levels of education.

Degree	Number of Participants	Percentage of Participants
Master	14	52.17%
PHD	8	34.78%
Bachelor	3	13.04%

All participants have experience in the MOOCs except for seven participants.

Table 4 presents a summary of the participant MOOCs experience.

Table 4

The summary of the participants MOOCs experiences

MOOCs experience	Number of Participants	Percentage of Participants
Yes	18	72 %
No	7	28 %

The participants in this study were from different geographic regions. A total of 18 participants were from the USA, and the other seven participants were from different countries such as Turkey, Spain, Australia, Serbia, Malaysia, and Libya. Table 5 presents a summary of the participant regions

Table 5

The regions of the participants

Regions of Participants	Number of Participants	Percentage of Participants
USA	18	72 %
Serbia	1	4 %
Malaysia	2	8 %

Libya	1	4 %
Turkey	1	4 %
Spain	1	4 %
Australia	1	4 %

Twenty-five participants responded to the survey, including two females and twenty-three males. Table 6 presents a summary of the participant's gender.

Table 6

The summary of the participant's gender

Gender	Number of Participants	Percentage of Participants
Male	23	92 %
Female	2	8 %

Presentation of the Data

Eleven questions were asked of each of the 25 participants. The survey questions for this study were

1. Which predictive models or algorithms do you think to use to predict dropout rate of students in MOOC? Check all that apply. If none of these apply, please describe the model or algorithm you use in the other textbox.
    a. Logistic Regression
    b. Deep Neural Network
    c. Support Vector Machine
    d. Hidden Markov Models
    e. Recurrent Neural Network
    f. Natural Language Processing Technique
    g. Decision Trees
    h. Survival Analysis
    i. Bayesian Network

Other (describe what other model or algorithm you use): _____

2. Based on your experience with machine-learning, what type of predictive models or algorithms that can be used to get better performance? Please elaborate

3. One type of machine learning is Supervised learning, Do you prefer to use logistic regression, Support Vector Machines (SVM), and Decision Trees when performing supervised learning? If this is not applicable to your work experience, put (N/A) in the box below.

4. Another type of machine learning is Unsupervised Learning. Do you prefer to use k-means clustering, and Association Rules when performing Unsupervised Learning? If this is not applicable to your work experience, put (N/A) in the box below.

5. The third type of machine learning is Semi-supervised, which is a mix of supervised learning and unsupervised learning. What algorithms do you prefer to use when you utilize this type of method? If this is not applicable to your work experience, put (N/A) in the box below.

6. The fourth type of machine learning is Reinforcement Learning. Do you prefer to use Adversarial Networks, and/or Temporal Difference (TD) Reinforcement Learning? If this is not applicable to your work experience, put (N/A) in the box below.

7. Based on your experiences MOOC, what improvements computer scientists educators need to make to increase interaction within the MOOC platforms?

8. Based on your experiences with MOOC, how does course content design impact student interaction in MOOC?

9. How does instructor involvement with the students help improve interaction within the MOOC platforms?. If this is not applicable to you, then type in the box below N/A.

10. How do you determine appropriate machine-learning datasets for predicting the dropout rates of students in MOOCs?
11. What type of data will help determine computer science, educators, to use to prepare a machine-learning dataset in the MOOC? Check all that apply. If you use other datasets, please describe in q. (other).

    a. Online learning behavior
    b. Student behavior
    c. Postings
    d. Demographics
    e. Clickstreams
    f. Stream server logs
    g. Graded activities within courses
    h. Forum posts and discussion
    i. Assignment records
    j. The effective period of attending the course
    k. Country
    l. Age
    m. Gender
    n. Most viewed pages
    o. Operating system
    p. Browser
    q. Other _____

After the data collection process, I selected the participants who have backgrounds in computer science and machine learning from various groups of machine learning in LinkedIn. The participants have various levels of experiences in computer science and machine learning in the LinkedIn community. Twenty-Five participants were emailed to participate in the Survey Monkey questionnaire. I collected all responses and imported them from Survey Monkey into Excel and PDF files. Once all responses were checked and scrubbed. I imported the responses into NVivo12 software to find the similarity themes among all responses. Member checking was used to ensure respondent validation to enhance study credibility, accuracy, and transferability. The following parts of this section present the results of the twenty--five participants' responses are described below.

Survey Question 1

The question was, "Which predictive models or algorithms do you think to use to predict dropout rate of students in MOOC? Check all that apply. If none of these apply, please describe the model or algorithm you use in the other textbox.

a. Logistic Regression
b. Deep Neural Network
c. Support Vector Machine
d. Hidden Markov Models
e. Recurrent Neural Network
f. Natural Language Processing Technique
g. Decision Trees
h. Survival Analysis
i. Bayesian Network

Other (describe what other model or algorithm you use): _____”

Aggregated data for survey question 1 yielded nine themes: (a) logistic regression, (b) decision tree, (c) deep neural network , (d) natural language Processing technique, (e) support vector machine, (f) survival Analysis, (g) recurrent neural network, (h) hidden Markova models, and (i) Bayesian network. The number of responses pertaining to each of the themes for survey question 1 is shown in Table 7.

Table 7

Themes for Survey Question 1

Themes	N	Percentage of Participants
Logistic regression	16	66.67%
Decision Trees	10	41.67%
Deep neural networks	9	37.50%
Support Vector Machine	7	29.17%
Natural Language Processing Techniques	7	29.17%
Recurrent Neural Networks	6	25.00%
Hidden Markov Models	5	20.83%
Bayesian Networks	5	20.83%
Survival Analysis	3	2.50%

Survey Question 2

The question was, “Based on your experience with machine-learning, what type of predictive models or algorithms that can be used to get better performance? Please elaborate”. Aggregated data for survey question 2 yielded six themes: (a) deep neural networks, (b) logistic regression, (c) decision trees, (d) random forests, (e) k-means

clustering, (f) and Hidden Markov Models. The number of responses pertaining to each of the themes for survey question 2 is shown in Table 8.

Table 8

Themes for Survey Question 2

Themes	N	Percentage of Participants
Deep neural networks	8	34.78 %
Logistic regression	6	26.09 %
Decision trees	3	13.04 %
Random forests	2	8.69 %
K-means clustering	1	4.34 %
Hidden Markov Models	1	4.34 %

Table 9 contains representative themes responses from survey question 2.

Table 9

Survey Question 2, Responses

	Responses
P20	I use an artificial neural network for the regression problem. It is easy to implement, and it will produce good prediction rates.
p19	Deep learning is one of the most powerful tools to improve the accuracy of MOOCs.
P16	learning-to-rank is accurate for ranking problems and relevant to a lot of software engineering problems
P15	A support vector machine would be a good choice for separating data into two camps. They tend to be highly efficient for these sorts of tasks and can intake high numbers of variables using the kernel trick.
P13	In the studies I have carried out, the random forest seems to be the most consistent, and it can achieve accurate results. recently, I have started exploring the world of neural networks with just the net package of r and results are also good despite not better than random forest

P12	Predictive models that have custom features designed by domain experts tend to do well; additionally, with large volumes of data available, DNNs are a good approach, especially if the architecture can be pretrained on a larger corpus
P9	Decision trees are a simple but powerful form of multiple variable analysis. They are produced by algorithms that identify various ways of splitting data into branch-like segments. Regression (linear and logistic) regression is one of the most popular methods in statistics. Regression analysis estimates relationships among variables, finding key patterns in large and diverse data sets, and how they relate to each other.
P6	You may study a collaborative filtering technique. So, you can find the similarities between students and may predict courses they are interested in, but they did not see yet.

Survey Question 3

The question was, "One type of machine learning is Supervised learning, Do you prefer to use logistic regression, Support Vector Machines (SVM), and Decision Trees when performing supervised learning? If this is not applicable to your work experience, put (N/A) in the box below. Aggregated data for survey question 3 yielded four themes:(a) decision trees, (b) logistic regression, (c) support victor machine, and (d) deep neural networks. The number of responses pertaining to each of the themes for survey question 3 is shown in Table 10

Table 10

Themes for Survey Question 3

Themes	N	Percentage of Participants
Decision trees	7	29.16 %
Logistic regression	6	25 %
Support victor machine	3	12.5 %
Deep neural networks	1	4.1 %

Table 11 contains representative themes responses from survey question 3.

Table 11

Survey Question 3, Responses

	Responses
P12	Yes. Although I think the definition of dropout might be a bit subjective, and it can vary depending on the study, I always define dropout, obtain the labels from data, and apply supervised learning.
p14	Using mostly deep neural networks for best performance classification and Random Forests / boosted trees to get good performance on smaller datasets
P25	Though there is some context to be considered here, SVM would be my first consideration

Survey Question 4

The question was, "Another type of machine learning is Unsupervised Learning, Do you prefer to use k-means clustering, and/or Association Rules when performing Unsupervised Learning? If this is not applicable to your work experience, put (N/A) in the box below. Aggregated data for survey question 4 yielded two themes:(a) k-means clustering, (b) and association rules. The number of responses pertaining to each of the themes for survey question 4 is shown in Table 12.

Table 12

Themes for Survey Question 4

Themes	N	Percentage of Participants
k-means clustering	9	37.5 %
Association Rules	1	4.16 %

Table 13 contains representative themes responses from survey question 4.

Table 13

Survey Question 4, Responses

	Responses
P4	Yes, I have used the K-means algorithm for predicting the class of an unknown sample.
P6	I prefer to use the Association Rule to find the relationship between the items
P7	It uncovers previously unknown patterns in data it better when you don't have data on the desired outcome
P14	Using language model pretraining (next word prediction etc.) tasks for unsupervised training on language
P15	k-means for prototypes and a quick study
P16	We can use k-means as it is more accurate in the predictions
P25	Though this is a bit restricted in choice, of those provided, clustering would be my first consideration

Survey Question 5

The question was, "The third type of machine learning is Semi-supervised, which is a mix between supervised learning and unsupervised learning. What algorithms do you prefer to use when you utilize this type of method? If this is not applicable to your work experience, put (N/A) in the box below."

Aggregated data for survey question 5 yielded nine themes: (a) graph-based algorithm, (b) logistic regression, (c) decision tree, (d) k-nearest neighbor algorithm, (e) supported Semi-supervised, (f) not supported Semi-supervised, (g) multi-task learners for language models, (h) genetic algorithm, and (i) support vector machine. The number of responses pertaining to each of the themes for survey question 5 is shown in Table 14.

Table 14

Themes for survey Question 5

Themes	N	Percentage of Participants
Supported Semi-supervised	2	8.33 %
Graph-based algorithms	1	4.1 %

Logistic Regression	1	4.1 %
Decision trees	1	4.1 %
K-nearest neighbor algorithms.	1	4.1 %
Not supported Semi-supervised	1	4.1 %
Multi-task learners for language models	1	4.1 %
Genetic algorithms	1	4.1 %
Support vector machine	1	4.1 %

Table 15 contains representative themes responses from survey question 5.

Table 15

Survey Question 5, Responses

	Responses
P1	I'm not a proponent of this technique, as it can introduce not needed "noise" and complicate reduction or clustering.
P24	This is some cool approaches that suggest labeling in a way to consider it weak supervision. I want to look into this. Open source snorkel.
P10	I strongly prefer this method

Survey Question 6

The question was, “The fourth type of machine learning is Reinforcement Learning. Do you prefer to use Adversarial Networks, and/or Temporal Difference (TD), Reinforcement Learning? If this is not applicable to your work experience, put (N/A) in the box below.”

Aggregated data for survey question 6 yielded five themes: (a) temporal difference, (b) adversarial networks, (c) q-learning, (d) reinforcement learning (e), and (f) game bots.The number of responses pertaining to each of the themes for survey question 6 is shown in Table 16.

Table 16

Themes for Survey Question 6

Themes	N	Percentage of Participants
Temporal Difference	3	12.5 %
Adversarial Networks	3	12.5 %
Q-learning	1	4.16 %
Reinforcement Learning	1	4.16 %
Game bots	1	4.16 %

Table 17 contains representative themes responses from survey question 6.

Table 17

Survey Question 6, Responses

	Responses
P16	I have used them very rarely for the creation of intelligent agents, such as game bots.
P9	Yes, I prefer to use Adversarial Networks, and/or Temporal Difference (TD) Reinforcement Learning
P1	This would depend on the context of the topic, as there is a goodly selection of techniques to use. Of this set, my preference would be Adversarial.

Survey Question 7

The question was, "Based on your experiences MOOC, what improvements computer scientists need to make to increase interaction within the MOOC platforms?"

Aggregated data for survey question 7 yielded sixteen themes: (a) course content , (b) course design , (c) determine success or failure for students, (d) extra references, (e) areas a student is struggling with, (f) MOOC synchronous, (g) improve productive models, (h) registered courses, (i) real courses, (j) challenging problems, (k) content delivery, (l) real word problem, (m) coursework, (n) current problems, (o) dropout systems, (p) and current education system . The number of responses pertaining to each of the themes for survey question 7 is shown in Table 18.

Table 18

Themes for Survey Question 7

Themes	N	Percentage of Participants
Course content	5	23.80 %
Course design	5	23.80 %
Determine success or failure for students	1	4.16 %
Extra references	1	4.16 %
Areas a student is struggling with	1	4.16 %
MOOC synchronous	1	4.16 %
Improve productive models	1	4.16 %
Registered courses	1	4.16 %
Real courses	1	4.16 %
Challenging problems	1	4.16 %
Content delivery	1	4.16 %
Real word problem	1	4.16 %
Coursework	1	4.16 %
Current problems	1	4.16 %
Dropout systems	1	4.16 %
Current education system	1	4.16 %

Table 19 contains representative themes responses from survey question 7.

Table 19

Survey Question 7, Responses

	Responses
P3	One limitation of MOOC education is that it is asynchronous, meaning that students and instructors are rarely in the same place at the same time, preventing direct communication. However, Internet-based platforms offer new opportunities that a traditional classroom does not. For example, screen-sharing technology can help

	students to get an insider's look at an instructor's creative process. Such platforms also create online forums where students from all over the world can exchange ideas. Computer Scientists are working to replace the current education system with MOOC.
P4	I think that the recommendation system is important. Recommendation system will provide trainers valuable course that meets their interests based on their registered courses and their evaluation.
P5	Extraction interesting patterns of users' behavior and actions.
P6	The computer scientists should focus on the futures of every algorithm in machine learning, also over courses that attract students related to the market to finish their courses effectively.
P8	More diverse MOOCs and more public knowledge sharing
P10	Machine Learning can continue to be used to increase the personalization of coursework or quizzes, as well as suggested resources depending on which areas a student is struggling with.
P12	I think, one of the current problems of dropout systems is that they are often made for research and despite they can show how powerful they can be, not many times they are used in real courses. Most of the research is posthoc analysis. I believe that predictive models can be improved but it would be important to start using them in real scenarios. That would raise the importance of them and would make an impact on learners.
P14	More interactive exercises, more challenging problems
P16	The courses need to provide tangible experience in the market
P17	High level of efficiency, speed up the user interaction, and daily feedback
P20	Add extra references related 100% that allow students to focus on these references instead of wasting time on another material. Also, focus on a real-world problem that can be solved and teach the students step by step.
P25	Broad subject. First, I'm not sure that a CS person is the one to make the changes unless they are the course content designer. - Understand the types of people who use them and the reasons why. - Determine why there is success or failure for students in the MOOC environment. - Gather a better understanding of human psychology and apply to course content/navigation, as opposed to primarily content delivery that makes sense from a functional perspective.

Survey Question 8

The question was, "Based on your experiences with MOOC, how does course content design impact student interaction in MOOC? ".Aggregated data for survey question 8 yielded twenty-two themes: (a)course design,(b) course content , (c) course content design ,(d) design options,(e) student disengages, (f) content design, (g) the quality of a course, (h) certain design

choices, (i) attractive design, (j) style preferences, (k) identifying prerequisites, (l) different styles, (m) different students, (n) practical courses, (o) educational design practices, (p) related courses, (q) course materials, (r) high- quality videos of courses, (s) users' satisfaction, (t) appropriate courses, (u) related courses, (v) learning course. The number of responses pertaining to each of the themes for survey question 8 is shown in Table 20.

Table 20

Themes for Survey Question 8

Themes	N	Percentage of Participants
Course design	6	24 %
Course conent	4	16 %
Course content design	4	16 %
Design options	4	16 %
Student disengages	4	16 %
The quality of a course	1	4.16 %
Certain design choices	1	4.16 %
Attractive design	1	4.16 %
Style preferences	1	4.16 %
Identifying prerequisites	1	4.16 %
Different styles	1	4.16 %
Different students	1	4.16 %
Practical courses	1	4.16 %
Educational design practices	1	4.16 %
Related courses	1	4.16 %
Course materials	1	4.16 %
High- quality videos of courses	1	4.16 %

Users' satisfaction	1	4.16 %
Appropriate courses	1	4.16 %
Related courses	1	4.16 %
Learning course	1	4.16 %

Table 21 contains representative themes responses from survey question 8.

Table 21

Survey Question 8, Responses

	Responses
P3	Course materials are very important and have a negative impact on students. People who are responsible for creating classes online know that very well, and they always choose appropriate courses for those students who study online.
P5	By offering to students the related courses they need based on their taught courses.
P8	Diversity allows student exposure to a wider variety of concepts
P10	Different students prefer different styles of courses, though a relation between style preferences and subject matter has not been made apparent to me.
P11	Depend, for me. I think attractive design should be considered as a priority after the content to deliver a fruitful ML course.
P12	Course content design is very important, and it can considerably affect predictive models. Currently, it is difficult to find a one-size-fits-all solution for the dropout problem because if you change your course design, your predictive models can be useless. Learners' interactions can change depending on the methodology, and that affects the models. More research is needed to try to find a solution to make predictive models generalizable or at least adaptable for large-scale scenarios (for example, in a university where there are hundreds of courses).
P15	Gives a good and deep introduction for people trying to migrate from traditional CV to ML
P17	The quality of a course be it a MOOC (Massive Open Online Course) or any other (open online) course, highly depends on the quality of the course design. Quality is a highly discussed topic in the literature as well as I don't really want to go into it in-depth, however, a very simple way of "measuring" the quality of a course design is by listening to the students the course is made for and the teachers who are teaching and maybe even did design the course. Users' (dis)satisfaction regarding the design can tell the user a lot about the impact certain design choices have on the learning experience. Designing a course takes a lot of time, effort, and expertise, and it remains a difficult task since the design options often seem to be endless. Sometimes the design options are rather limited by the technical possibilities of the platform or the environment. To get insight into educational design practices in MOOCs and to identify scalable best practices we developed an instrument called 'Educational

	Scalability Analysis Instrument'. The instrument is developed to analyze the course design from a qualitative perspective, and it can be used by teachers, designers but also the students themselves.
P20	It is a very important question. The videos must be in high quality with clear sound, and the design has to be friendly and easy to access.
P21	Course contents are very important to acquire the new knowledge, skills, and abilities in order to change their understandings and perspective
P25	Intrinsically connected. If the content sucks, the course sucks, and the student disengages. If the flow is inconsistent or differs from the material being taught, the student disengages. If the content is uninteresting, disengagement. If the aesthetics are difficult, disengagement.

Survey Question 9

The question was, "How does instructor involvement with the students help improve interaction within the MOOC platforms?. If this is not applicable to you, then type in the box below N/A".

Aggregated data for survey question 9 yielded nine themes: (a) feedback of in structure,(b)zero instructor interaction, (c) instructor visibility, (d) feedback to students' questions, (e) forms of interaction, (f) research classes online, (g) online classes, (h) posted a message, (i) explaining some issues. The number of responses pertaining to each of the themes for survey question 9 is shown in Table 22.

Table 22

Themes for Survey Question 9

Themes	N	Percentage of Participants
Feedback of Instructure	9	36 %
Zero instructor interaction	2	8.33 %
Instructor visibility	1	4.1 %
feedback to students' questions	1	4.1 %

Forms of interaction	1	4.1 %
Research classes online	1	4.1 %
Online classes	1	4.1 %
Posted a message	1	4.1 %
Explaining some issues	1	4.1 %

Table 23 contains representative themes responses from survey question 9.

Table 23

Survey Question 9, Responses

	Responses
P1	I like when the instructor or someone is at least looking at some of the stuff. Ideally not needed. For well put together courses. I don't mind zero instructor/ interaction.
P3	Even if you have online classes, having an instructor will help a lot in explaining some issues for students. For example, having research classes online will never help the students to write their dissertation.
P8	Choice of the right typical real-world problems and scenarios.
P11	One of the key point from my point of view is the course work and training, which might not be applied for all student because course work requires good programing skill in Python form example.
P12	So far, I have not explored this variable in the predictive models, but in other contributions, I found that there was usually an action-reaction in MOOCs. Whenever the instructor posted a message in the MOOC forum, the activity rose. This may imply that the involvement of the instructor can be an important factor to improve interaction within the MOOC.
P14	Message boards / piazza / forums interaction can help a lot with being stuck
P16	The instructor needs to give feedback to students' questions
P17	Embedded MOOC forums; and (2) rationales for social media use from both instructors' and students
P25	No instructor availability/involvement, then students with a high curiosity quotient, or those that are struggling with understanding to content will disengage.

Survey Question 10

The question was, “How do you determine appropriate machine-learning datasets for predicting the dropout rates of students in MOOCs?”

Aggregated data for survey question 10 yielded twenty-one: (a) exploring data analysis, (b) collected datasets from different courses, (c) datasets related to dropout students, (d) datasets related to dropout students, (e) depends on the performance of the algorithm, (f) the quality of data, (g) missing data, (h) create data sets, (i) accurate data, (j) limited data, (k) learning datasets, (l) appropriate machine learning datasets, (m) proprietary data of the MOOCs, (n) prior courses, (o) different courses, (p) computer vision models, (q) computer vision models, (r) predictive models, (s) much interaction, (t) exercises interaction, and (u) neural network . The number of responses pertaining to each of the themes for survey question 10 is shown in Table 24.

Table 24

Themes for Survey Question 10

Themes	N	Percentage of Participants
Exploring data analysis	2	8.33 %
Collecting datasets from different platforms	1	4.1 %
Collecting datasets from different courses	1	4.1 %
Datasets related to dropout students	1	4.1 %
Depends on the performance of the algorithm	1	4.1 %
The quality of data	1	4.1 %
Missing data	1	4.1 %
Create data sets	1	4.1 %
Accurate data	1	4.1 %
Limited data	1	4.1 %
Learning datasets	1	4.1 %
Appropriate machine learning datasets	1	4.1 %

proprietary data of the MOOCs	1	4.1 %
Prior courses	1	4.1 %
Different courses	1	4.1 %
Computer vision models	1	4.1 %
Predictive models	1	4.1 %
Much interaction	1	4.1 %
Exercises interaction	1	4.1 %

Table 25 contains representative themes responses from survey question 10.

Table 25

Survey Question 10, Responses

	Responses
P4	Datasets should not have missing data. Also, all dataset should be contained in discrete values.
P5	Based on its important features and number of records.
P6	Determining the appropriate machine learning datasets depends on the performance of the algorithm or model and understand the features for every algorithm by getting more dataset from MOOC platforms.
P7	Create data sets for computer vision models
P10	Data relating to the students in question, labeled with the dropout status, and with as many relevant variables as possible. Variables can later be cut down. Highly accurate data is extremely helpful, but it is more important to know exactly how accurate or inaccurate your data is.
P12	I think it is important in a dataset to have a representative sample of the data. This is important in MOOCs because many learners who enroll in the MOOC do not have any interaction. So, it is important to remove them. Moreover, as much interactions are provided (videos) exercises, etc.), the dataset would be more useful. However, despite this fact, it is possible to develop predictive models with limited data. For example, in a previous analysis, I could achieve powerful results with exercises interactions because videos were not available. Exercises, etc.), the dataset would be more useful. However, despite this fact, it is possible to develop predictive models with limited data. For example, in a previous analysis, I could achieve powerful results with exercises interactions because videos were not available.

P21	The appropriate machine-learning datasets are collected from different platforms and different courses.

Survey Question 11

The question was, “What type of data will help determine computer science, educators, to use to prepare a machine-learning dataset in the MOOC? Check all that apply. If you use other datasets, please describe in q. (other).

a. Online learning behavior

b. Student behavior

c. Postings

d. Demographics

e. Clickstreams

f. Stream server logs

g. Graded activities within courses

h. Forum posts and discussion

i. Assignment records

j. The effective period of attending the course

k. Country

l. Age

m. Gender

n. Most viewed pages

o. Operating system

p. Browser

q. Other _____?”

Aggregated data for survey question 10 yielded fourteen themes: (a) online learning behavior, (b) student behavior, (c) assignment records, (d) age, (e) graded activities within courses, (f) forum posts and discussion, (g) the effective period of attending course, (h)gender, (i) stream server logs, (j) country, (k) most viewed pages, (l) browser, (m) operating system, (n) and other.

The number of responses pertaining to each of the themes for survey question 11 is shown in Table 26.

Table 26

Themes for Survey Question 11

Themes	N	Percentage of Participants
Online learning behavior	18	72 %
Student behavior	17	68 %
Assignment records	12	48 %
Age	12	48 %
Graded activities within courses	11	44 %
Forum posts and discussions	11	44 %
The effective period of attending the course	10	40 %
Gender	10	40 %
Stream server logs	7	28 %
Country	7	28 %
Most viewed pages	7	28 %
Browsers	5	20 %
Operating system	4	16 %
Other	2	8 %

Table 27 contains representative themes responses from survey question 11.

Table 27

Survey Question 11, Responses

	Responses
P1	Activity and prior activity in the course and prior courses.
P20	History of attended courses

In the study, the participants used some words frequently to explore the strategies computer science educators need for predicting dropout rates in MOOCs. The strategies were based on three principals: the predictive models or algorithms used to predict dropout rate of students in MOOCs, MOOCs experiences, and machine-learning datasets for predicting the dropout rates of students in MOOCs. The NVivo software was able to capture these words and describe them in what is known as a "word cloud." The more frequent the participant used the word, the bolder, and bigger, the word appears in the word cloud. Figure 2 illustrates a word frequency cloud, which describes the frequency of words that appeared in the themes from the data analysis process.

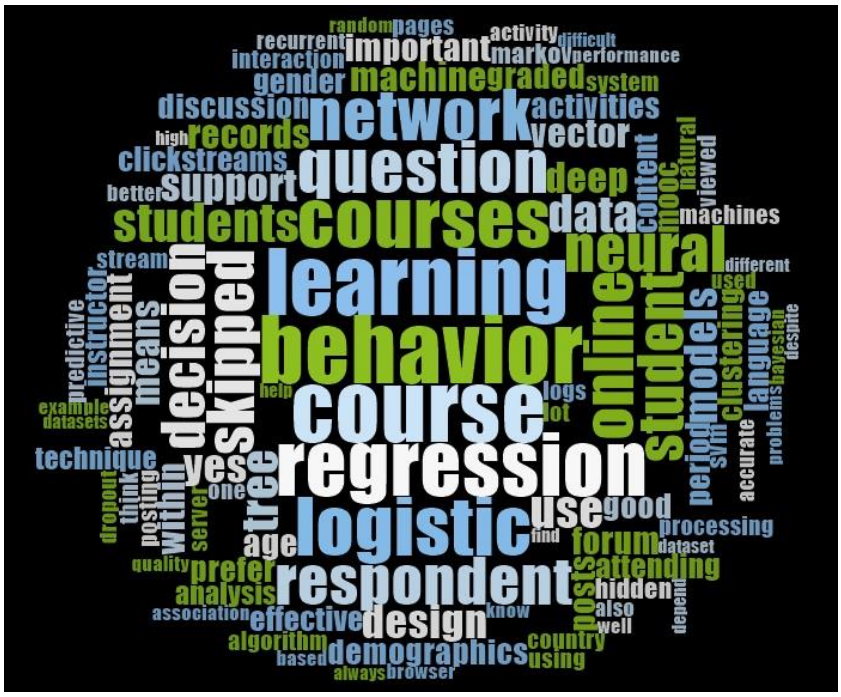

Figure 2. Word frequency cloud.

Presentation and Discussion of Findings

After the data was collected, imported, and uploaded into NVivo, the coding process was conducted. This process involved going through each survey question looking for major themes and supporting quotes. Aggregated themes data from the eleven questions that had been collected from the twenty-five participants included three principals that answered the research question. The findings in this study were based on three principals (a) the predictive models or algorithms used to predict dropout rate of students in MOOCs, (b) MOOCs experiences, and (c) machine-learning datasets for predicting the dropout rates of students in MOOCs. The following tables show the three principles and the themes supported by quotes from the participants as identified in the data analysis.

Finding One: Algorithms or Predictive Models for Predicting Dropout Rates in MOOCs

Table 28 is the cumulative total of the major themes found in this study related to algorithms or predictive models from question one to question six**.** An essential finding was that logistic regression, decision trees, deep neural networks, support vector machine, and K-means were the most used algorithms or predictive models. Most of the participants supported the use of these algorithms to predict the dropout rates in MOOCs. Table 28 shows the aggregated themes related to predictive models or algorithms used to predict dropout rate of students in MOOCs.

Table 28

Algorithms or predictive models themes

Themes	Frequency of the algorithms throughout all Responses
Logistic regression	29
Decision Trees	21
Deep neural networks	18
Support Vector Machine	11
K-means	10
Natural Language Processing Techniques	7
Recurrent Neural Networks	6
Hidden Markov Models	6
Bayesian Networks	5
Survival Analysis	3
Temporal Difference	3
Adversarial Networks	3

Random forests	2
Supported Semi-supervised	2
Association Rules	1
K-nearest neighbor algorithms.	1
Reinforcement Learning	1
Game bots	1
Q-learning	1
Graph based algorithms	1
Not supported Semi-supervised	1
Multi-task learners for language models	1
Genetic algorithms	1

The participants indicated that logistic regression is the best algorithm. One of the participants justified the use of this algorithm because “regression is one of the most popular methods in statistics. Regression analysis estimates relationships among variables, finding key patterns in large and diverse data sets and how they relate to each other”. Another participant indicted, “yes, I use logistic regression for accurate results.”

Also, the participants shared their thoughts on the decision tree algorithm. Some of the participants stated and supported the use of the decision trees algorithm. They indicated that “decision trees are a simple but powerful form of multiple variable analysis. They are produced by algorithms that identify various ways of splitting data into branch-like segments”. Another participant indicated, "using mostly boosted trees to get good performance on smaller datasets."

Some of the participants supported the use of deep neural networks. Participants stated that they used deep neural networks and justified their use. One justification was “deep learning

is one of the most powerful tools to improve the accuracy of MOOC." Another justification was that "using mostly deep neural networks for best performance."

There was varying support for the use of the remaining algorithms. One participant indicated the use of a support vector machine for prediction and justified that by his answer. "Support vector machine would be a good choice for separating data into two camps. They tend to be highly efficient for these sorts of tasks and can intake high numbers of variables using the kernel trick". Another participant supported the use of the K-means algorithm. He justified his answer "we can use k-means as it is more accurate in the predictions."

One of the participants supported the use of the random forest algorithm because as indicated in the answer that "in the studies, I have carried out, the random forest seems to be the most consistent, and it can achieve accurate results." Another participant stated that "I prefer to use the association rule to find the relationship between the items." The remaining participants had varying responses regarding the use of different models or algorithms for predicting dropout rates of students in MOOCs. These participants did not have clear justifications for their responses.

Finding Two: MOOCs Experience

Table 29 is the cumulative total of the major themes found related to MOOCs experience. The MOOCs experience mainly focused on three principles: improvements computer science educators need to make to increase interaction within MOOCs, how course content and design impact student interaction in MOOCs, and how does instructor involvement with the students help improve interaction within the MOOC platforms.

Table 29

MOOCs experience themes

Themes	Frequency of the MOOCs *experience* throughout all responses
Course design	11
Course content	9
Feedback of instructor to students	9
Current problems	1
Determine success or failure students	1
Extra references	1
Areas a student struggling with	1
MOOC synchronous	1
Improve productive models	1
Registered courses	1
Current education system	1
Challenging problem	1
Content delivery	1
Real-world problem	1
The attractive design of courses	1
Style preferences of courses	1
Identifying prerequisites	1
Different styles	1
Diffrrent students	1
Practical courses	1
Educational design practices	1

Course meterial	1
High- quality videos of courses	1
User satisfaction	1
Instructor visibility	1
Forums of interaction	1
Research classes online	1
Posted messages	1
Explaining some issues	1

The essential findings for MOOCs experience were (a) course design, (b) course content, and (c) instructor feedback to students. Most of the participants supported course design such as practical, styles, and quality or value of courses. One of the participants stressed the importance of designing the course in the MOOCs and focused on the quality of the course design and aspects of the design of the course from a qualitative perspective that can be used from designers, students, and instructors. The participant indicated in his response that

> The quality of a course is it a MOOC (Massive Open Online Course) or any other (open online) a course highly depends on the quality of the course design. Quality is a highly discussed topic in the literature as well as I don't really want to go into it in-depth, however, a very simple way of "measuring" the quality of a course design is by listening to the students the course is made for and the teachers who are teaching and maybe even did design the course. Users' (dis)satisfaction regarding the design can tell the user a lot about the impact certain design choices have on the learning experience. Designing a course takes a lot of time, effort, and expertise, and it remains a difficult task since the design options often seem to be

endless. Sometimes the design options are rather limited by the technical possibilities of the platform or the environment.

Another participant indicated the role of course content designer and its impact on student failure and success,

Broad subject. First, I'm not sure that a CS person is the one to make the changes unless they are the course content designer. Understand the types of people who use them and the reasons why. Determine why there is success or failure for students in the MOOC environment. Gather a better understanding of human psychology and apply to course content/navigation, as opposed to primarily content delivery that makes sense from a functional perspective.

As another participant indicated the importance of the course design, and how "machine learning can continue to be used to increase the personalization of coursework or quizzes, as well as suggested resources depending on which areas a student is struggling with."

Another participant indicated the importance of the course design. The response was, "I think attractive design should be considered as a priority after the content to deliver a fruitful course."

Also, regarding course content design and the styles of courses, one of the participants' response was "different students prefer different styles of courses, though a relation between style preferences and subject matter has not been made apparent to me."

One of the participants also indicated the importance of the quality of the videos in the course content, "the videos must be in high quality with clear sound and the design has to be friendly and easy to access."

Another participant stressed the importance of the content of the course and stressed that the content of the course needs to meet the requirements of the labor market, "the courses need to

provide tangible experience in the market." Another participant indicated that “it is very important to build on the right foundation, so the course content that does better about identifying prerequisites tends to do better.” As another participant pointed out, “different students prefer different styles of courses, though a relation between style preferences and subject matter has not been made apparent to me.” Another participant discussed the value of the courses in his response, “the courses should be in a practical way as hands-on, to gain more skills in their areas.” Another participant said, “course materials are very important and have a negative impact on students. People who are responsible for creating classes online know that very well, and they always choose appropriate courses for those students who study online.”

In addition to the theme of the course content, one participant stressed the importance of references in courses.

> Add extra references related 100% that allow students to focus on these references instead of wasting time on another material. Also, focus on a real-world problem that can be solved and teach the students to step feedback of instructor to students by step

With the importance of courses and their relationship to students in terms of scientific references and focus on real problems, another participant pointed out that there should be a real desire to register at the appropriate course and indicated in his response that course content design impact student interaction in a MOOC, “by offering to students the related courses they need based on their taught courses.”

The other MOOC experience theme was the instructor feedback to students. Some participants indicated the importance of the instructor feedback to increase interaction within MOOCs. One participant stressed that “the instructor needs to give feedback to students'

questions." Another participant indicated that instructors need to provide "daily feedback." The participant also stressed that the instructor feedback should be at a "high level of efficiency," in order to "speed up the user interaction."

Another participant stressed the importance of feedback between the instructor and the students,

> One limitation of MOOC education is that it is asynchronous, meaning that students and instructors are rarely in the same place at the same time, preventing direct communication. However, Internet-based platforms offer new opportunities that a traditional classroom does not. For example, screen-sharing technology can help students to get an insider's look at an instructor's creative process. Such platforms also create online forums where students from all over the world can exchange ideas. Computer Scientists are working to replace the current education system with MOOC.

The third theme was the feedback between the student and the instructor, This theme focused on interaction between the instructor and the student to increase the interaction in MOOCs, as one participant pointed out in the response that was "no instructor variability/involvement, then students with a high curiosity quotient, or those that are struggling with understanding of content, will disengage". Another participant indicated the importance of the role of forums and social media to increase interaction between students and instructors in MOOCs, "embedded MOOC forums; and (2) rationales for social media use from both instructors and students". Another participant indicated the importance of the feedback between the student and the instructor thus helping the students understand the many issues that need to be clarified, "even if you have online classes, having an instructor will help a lot in explaining

some issues for students. For example, having research classes online will never help the students to write their dissertation".

There were a variety of participants' responses to other themes related to MOOCs experience. One participant pointed out that one of the improvements that should be stressed is the predictive models used to address the real-time dropout problem while teaching courses,

> I think, one of the current problems of dropout systems is that they are often made for research and despite they can show how powerful they can be, not many times they are used in real courses. Most of the research is posthoc analysis. I believe that predictive models can be improved, but it would be important to start using them in real scenarios. That would raise the importance of them and would make an impact on learners.

Finding Three: Datasets for predicting dropout rates of students in MOOCs

The datasets for predicting dropout rates of students in MOOCs mainly focused on eight major themes. These themes were (a) online learning behavior, (b) student behavior, (c) assignment records, (d) age, (e) graded activities within courses, (f) graded activities within courses, (g) the effective period of attending the course, and (h) gender. Table 30 is the cumulative total of the major themes found in this study related to data sets for predicting dropout rates of students in MOOCs.

Table 30

Datasets Themes

Themes	Frequency of the MOOCs experience throughout all responses

Online learning behavior	18
Student behavior	17
Assignment records	12
Age	12
Graded activities within courses	11
Forum posts and discussions	11
The effective period of attending the course	10
Gender	10
Stream server logs	7
Country	7
Most viewed pages	7
Browsers	5
Operating system	4
Other	2
Exploring data analysis	2
Collecting datasets from different platforms	1
Collecting datasets from different courses	1
Datasets related to dropout students	1
Depends on the performance of the algorithm	1
The quality of data	1
Missing data	1
Create data sets	1
Accurate data	1
Limited data	1
Learning datasets	1
Appropriate machine learning datasets	1
proprietary data of the MOOCs	1

Prior courses	1
Different courses	1
Computer vision models	1
Predictive models	1
Much interaction	1
Exercises interaction	1

One participant stressed the importance of data representation by focusing on many interactions in MOOCs,

> I think it is important in a dataset to have a representative sample of the data. This is important in MOOCs because many learners who enroll in the MOOC do not have any interaction. So, it is important to remove them. Moreover, as much interactions are provided (videos, exercises, etc.), the dataset would be more useful. However, despite this fact, it is possible to develop predictive models with limited data. For example, in a previous analysis, I could achieve powerful results with exercises interactions because videos were not available.

The same participant indicated the importance of the variables and data that can be used to predict the dropout rates of students in MOOCs and stressed on the variables related to the exercises.

> I add some details for question 11. I have been researching the kind of variables. From previous findings, I consider that self-reported data is not useful because learners can say something and do another. Forum data may not be useful if there are not many posts, which is not unusual. Moreover, there are many people who complete the MOOC without using the forum so it may not be

representative. And from my experience, the best predictors are those related to exercises interactions.

Another participant supported the use of accurate data sets for predicting dropout rates of students in MOOCs,

> Data relating to the students in question, labeled with the dropout status, and with as many relevant variables as possible. Variables can later be cut down. Highly accurate data is extremely helpful, but it is more important to know exactly how accurate or inaccurate your data is.

Another participant indicated that the performance and features of the prediction algorithm depend on the appropriateness of the datasets, "determining the appropriate machine learning datasets depends on the performance of the algorithm or model, and understand the features for every algorithm by getting more dataset from MOOC platforms."

Another participant stressed the importance of data quality, "datasets should not have missing data. Also, all dataset should be contained discrete values". One of the participants supported the use of proprietary data available in the MOOC.

> You'd need the access to the proprietary data of the Coursera or someone like that. EDA will show if they capture needed metrics. Activity, prior activity, courses, prior courses. I would tend to think activity would strongly correlate with dropping out

Summary of Chapter Four

In this chapter, the data collection process, demographics of the participants, emerging themes, and a presentation and discussion of findings was provided. The purpose of this exploratory qualitative study was to explore strategies computer science educators need to use to

prepare machine learning dataset for predicting dropout rates of students in MOOCs. Survey Monkey was used to collect data from 25 professionals in machine learning from the LinkedIn community. Afterward, the combined data were coded and analyzed. The analysis of data revealed three principal themes, the predictive models or algorithms used to predict dropout rates of students in MOOCs, the MOOC experiences, and machine-learning datasets for predicting the dropout rates of students in MOOCs. Chapter five contains a discussion of these findings in more detail with a focus on implications for practice and recommendations for future studies.

CHAPTER FIVE

The high enrollment of students in massive open online courses (MOOCs) is misleading. Less than half of the learners enrolled in MOOCs actively engage in their courses, while the other learners either drop the course or do not participate, which contributes to the high dropout rate of students (Hone & El Said, 2016). As a nascent research design, there are no exploratory studies of a similar scope within big data analytics and machine-learning that identified the prediction strategies computer scientists educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in MOOCs (Xing et al., 2016). The problem addressed in this study was to identify the strategies computer science educators need to use to prepare machine-learning datasets for predicting dropout rates of students in massive open online courses.

The purpose of this qualitative exploratory study was to identify the predictive models or algorithms used to predict dropout rates of students in MOOCs, the MOOC experiences, and machine-learning datasets for predicting the dropout rates of students in MOOCs. A qualitative methodology was used for this exploratory research. Exploratory research is usually conducted to study an issue or a problem that is not clearly defined (Creswell, 2014). The qualitative methodology was appropriate for this research study because it enabled the use of open-ended survey questions.

The research study was subject to several limitations, which must be acknowledged. First, the research was limited to a sample of 25 participants from specific machine learning groups in the LinkedIn community. A second limitation is the inability to generalize the results of this study due to the small sample size and the different characteristics of the population, as well as the participants, were from different geographic regions. Thirdly, the answers of

participants were not optimal; there were variances in the responses due to the level of experience of the participants. Finally, the duration of the survey was open presented a limitation of time. The study questionnaire was available for a limited time.

Finding ways or strategies to reduce the dropout rates in MOOCs will enable the students to complete their courses successfully. Those who finish their MOOC courses successfully will contribute and use their knowledge to better their society. A delimitation in this study was that the participants were from LinkedIn, and they had to have at least one year of experience in machine learning.

In the following sections of this chapter, a summary of the findings and a detailed discussion of the conclusions as related to the research question and problem statement are presented. Limitations of the study, implications for the practice, and recommendations for the future researcher are also included in this chapter.

Findings and Conclusions

The study data was collected from 25 participants who had different experiences in computer science and machine learning from various groups of machine learning in LinkedIn. The collected responses were imported from Survey Monkey into Excel and PDF files. Once all responses were checked and scrubbed, the responses were imported into NVivo12 software to find the similarity themes among them. Member checking was used to ensure respondent validation to enhance study credibility, accuracy, and transferability.

The results of the analysis provided the following three themes, the predictive models or algorithms used, MOOCs experiences, and machine-learning datasets for predicting the dropout rates of students in MOOCs. The MOOCs experience was further broken down into sub-themes, namely course design, course content, and instructor feedback to students in MOOCs. In the

remaining parts of this section, an interpretation of the findings, as well as the significance of the results, is compared and contrasted with the literature.

Major Theme 1: Algorithms or Predictive Modules Used

This theme focused on the types of algorithms that can be used in the MOOC platforms for predicting dropout rates of students. The major algorithms that the study participants preferred to use were logistic regression, decision trees, deep neural networks, support vector machine, and K-means. The responses of the participants provided a clear vision for the understanding of this theme.

The literature review in this study indicated that logistic regression is the best algorithm for predicting dropout rates of students, which is similar to the findings of this study. Dalipi, Imran, and Kastrati (2018) showed that the logistic regression algorithm is the most widely used to address the problem of students dropping out from MOOCs. Umer, Susnjak, Mathrani, and Suriadi (2017) reported that the machine-learning algorithms used in a recent study included K nearest neighbor, logistic regression, random forest, and Naïve Bayes where the results showed that logistic regression outperformed other algorithms with the highest accuracy.

Due to the various experiences of the participants, there were diverse views to determine the appropriate algorithms other than logistic regression. Some of the participants provided justifications for using such algorithms like decision trees, deep neural networks, support vector machine, K-means, random forest algorithm, and association rule. There was a small difference in the sequence of use of these algorithms where the literature review indicated that support vector machine and decision trees are at the same level. Sequentially came other algorithms such as natural language processing, deep neural networks, hidden Markov moles, survival analysis, and Bayesian network algorithms for predicting dropout rates of students in MOOCs.

Major findings of the predictive models and algorithms in this study somewhat agreed with the findings in the literature. Dalipi, Imran, and Kastrati (2018) showed that support vector machine and decision trees algorithms came after logistic regression as the best algorithms for predicting the dropout rates of students in MOOCs. The order of the algorithms found in this study differs from the order found in the literature review. The findings of this study indicated that decision trees and deep neural networks instead of support vector machine came after logistic regression as best performing models for predicting dropout rates of students in MOOCs. A major finding of his study is the that some of the participants preferred the use of the K-means algorithm as an unsupervised machine learning model for predicting dropout rates of students which the literature review did not indicate. The K-means algorithm as an unsupervised machine learning model in this study came after algorithms such as logistic regression, decision trees, deep neural networks, and support vector machine.

Major Theme 2: MOOCs Experience

The second theme was the MOOCs experience. This theme consisted of three principal subthemes, improvements computer science educators need to make to increase interaction within MOOCs, how course content and design impact student interaction in MOOCs, and how does instructor involvement with the students help improve interaction within the MOOC platform. The literature review indicated the importance of the course design in MOOCs and its effective role in the retention of students. Also, the literature review stressed that the MOOC courses should be structured to reduce the dropout rates, which corresponds to the results of this study.

The study participants stressed that the course design, course content, delivery styles, course value, and the quality of the courses do increase the course interaction, thus reducing student dropout rates in MOOCs. Abeer and Miri (2014) reported that an ill-structured MOOC

would likely cause learners to fail to complete the course. This corresponds to the findings of this study. Schaffer et al. (2016) showed that the percentage of student dropouts differs according to the structure and type of course which agrees with the findings of this study.

As the literature review indicated, the role of instructors in MOOCs, by providing their recommendations to the learners based on their learning styles, improve the educational experience of MOOCs. The study findings indicated similarity with the literature review. The study participants stressed that the feedback of instructors to learners in MOOCs should be daily, discussing challenging problems, instructor visibility, how content is explained, forms of interaction, will increase the successful student compilation of the course. Alumu and Thiagarajan (2016) reported that instructors sought to improve the educational experience of MOOCs by providing recommendations to the learners based on their learning styles, which is similar to the finding of this study. As our findings showed, Alumu and Thiagarajan (2016) mentioned that the instructor feedback should be at a "high level of efficiency," in order to "speed up the user interaction."

Khalil (2018) showed that the content of the courses can increase the interaction in MOOCs. Khalil (2018) studied the expansion of the educational process through MOOCs and the interaction of students within these educational courses. He suggested that students who finish their courses successfully are likely to adopt MOOCs in the future. Khalil (2018) confirms the results of this study regarding the MOOC course content and its impact on course interaction.

Major Theme 3: Datasets for predicting dropout rates of students in MOOCs

The third theme was the data elements needed in datasets for predicting dropout rates of Students in MOOCs. The responses of the participants regarding this theme were essentially focused on online learning behavior, student behavior, assignment records, age, graded activities within courses, forum posts and discussions, the effective period of attending the course, and

gender. Participants indicated in their responses the importance of these data elements in the datasets collected. In addition to these data elements, the participants suggested focusing on another type of data that would help to predict dropout rates of students in MOOCs such as videos usage, exercise interactions, and the use of additional proprietary data available in the MOOC platform.

In addition to the above major data elements that should be included in the datasets, stream server logs, country, most viewed pages, and the browsers used should be in the datasets as well. The literature review indicated almost the same data elements that should be part of the datasets with some small differences than the findings of this study. (Li et al., 2016). The literature review indicated that data should be collected regarding online learning behavior, postings, frequency of watching included videos, behavior data, assignment grades, demographics, clickstreams, video data, and stream server logs (Wang & Baker, 2015)

In regard to the data elements that should be part of the datasets for predicting dropout rates of students, the literature review focused on online learning behavior and student behavior. Wang and Baker (2015) reported that the different structure of the MOOC classes presents different types of data, such as online learning behavior, postings, and the frequency of watching videos which is similar to the findings of this research. As reported by Dekker, Pechenizkiy, and Vleeshouwers (2009) studying the behavior data of the students in a certain period can help evaluate the educational process by focusing on the type of teaching and course presentation. Most of the participants in this study indicated the importance of using online learning behavior and student behavior in the datasets for predicting dropout rates of students in MOOCs, which correspond with the literature review. As Kizilcec, Piech, and Schneider (2013) reported, we

should focus on the behavioral data in the MOOC database, which can help in extracting patterns in the analyzed data that help predict the success of students in the MOOC courses.

.Wu and Zheng (2016) focused on the extraction of descriptive student information from training courses and course registration records in the edX platform, as well as user behavior while considering the privacy of data for students. Compared to the findings of this study, there was a similarity to use assignment records in the datasets for predicting dropout rates of students. Sinha (2014) reported that the diversity of the different data sources, assignment grades, demographics, and clickstreams play a decisive role in obtaining information on the student dropout phenomenon, which also corresponds to the findings of this study.

Nagrecha et al. (2017) indicated that we need to collect the use of clickstream and video data as well as student behavior, like video interaction, to determine the dropout rate among students. This agrees with this study findings that showed that clickstream or most viewed pages, and student behavior data should be in the collected datasets. In addition, Jiang, Williams, Schenke, Warschauer, and O'dowd (2014) reported that many researchers have analyzed stream server logs associated with MOOC platforms in regards to the video lectures viewing frequency, time spent by students studying the material, and the rate of completion of the various quizzes and homework-based assessments to predict dropout rates which confirm this study findings that these data elements should be in the datasets used to predict student dropout rates.

Researchers need to study student retention rates in MOOCs and analyze the causes of the dropout in MOOCs. This study findings showed that logistic regression is the best algorithm. The types of data elements in the collected datasets from the MOOCs will help faculty to determine the adequacy of these datasets and decide on which algorithms or predictive models should be run to determine the factors that will reduce the dropout rates of students in MOOCs.

This study findings will help MOOCs designers, researchers, faculties, and computer science educators to understand the strategies that need to be used for predicting student dropout rates in MOOCs.

The Limitations of the Study

This research study was subject to several limitations, which must be acknowledged. First, the research was limited to a sample of 25 participants from specific machine learning groups in the LinkedIn community. A second limitation is the inability to generalize the results of this study due to the small sample size and the different characteristics of the population, as well as the participants, were from different geographic regions. The answers of participants were not optimal; there were variances in the responses due to the level of experience of the participants. Finally, the short duration of the survey was open presented a limitation of time. The study questionnaire was available for a limited time. A delimitation in this study was that the participants were from LinkedIn, and they had to have at least one year of experience in machine learning. Also, the participants in this study had to answer three prequalifying questions to qualify before starting the questionnaire.

Implications for Practice

Hone and El Said (2016) showed that the high enrollment of students in massive open online courses (MOOCs) is misleading. Less than half of the learners enrolled in MOOCs actively engage in their courses, while the other learners either drop the course or do not participate, which contributes to the high dropout rate of students in these platforms. The results of this study will enable MOOCs designers to understand the MOOCs experience in regard to the content of the courses, design of courses, and the feedback of the instructors. This study results will help computer science educators and faculties to understand the types of algorithms or

predictive models and the associated datasets that will enable them to predict the dropout rates of students in MOOCs.

To implement the findings of this study, computer science educators and MOOCs designers may face some challenges such as the privacy of data in MOOCs while using the algorithms or predictive models in the real-time for predicting dropout rates of students in MOOCs. The following sections provide three recommendations for findings of this study, which are: MOOCs experience, algorithms or predictive, and datasets.

. Implications of Study and Recommendations for Future Studies

Study participants provided data that was explored, identified, and helped in understanding the strategies computer science educators need to use in preparing machine learning datasets to predict dropout rates in MOOCs. This qualitative study was designed to explore, identify, and understand three principal findings: the algorithms or predictive models, the datasets data elements, and the MOOCs experience for predicting the dropout rates of students in MOOCs.

Recommendation 1: Algorithms or predictive models

This recommendation is for computer science educators in MOOCs and scholars to continue further studying algorithms or predictive models for predicting dropout rates of students in MOOCs focusing on the specific algorithms or predictive models that are closely related with MOOCs platforms. While these algorithms have been discussed in this study, more can be done to find optimal algorithms to reduce the dropout rates of students in MOOCs and evaluate these algorithms using real-time processes such as those available in Apache Spark. The findings of this study can help computer science educators to choose the best algorithms or predictive models to reduce the dropout in MOOCs.

Recommendation 2: Datasets

Increased diversity of the data elements in the datasets helps to improve the performance of algorithms and predictive models for accurate results. Using the datasets in real-time scenarios will help computer science educators to get precise predictions. The findings of this study can help computer science educators to choose the best data elements that need to be in the datasets for predicting dropout rates of students in MOOCs.

Recommendation 3: MOOCs experience

MOOC designers need to increase student interaction in MOOCs to address the problem of dropouts. The findings of this study showed that course design, course content, and instructors' feedback are critical factors that can impact student success and can decrease student dropout in MOOCs. The designers of MOOCs can benefit from the findings of this study by providing ways to increase the interaction and efficiency of the MOOCs platforms. MOOC designers should focus on the course design as well as the value and quality of the course content and devise methods and techniques to increase the interaction between the instructors and students. There are many different MOOCs platforms around the world used by various universities and faculties. As this study found that the MOOCs experience is a critical factor that can address the dropout problem, MOOC designers need to find ways to increase the interactions within the MOOC platform and evaluate the interaction between the instructors and students. Future studies should provide a comprehensive analysis of the performance evaluation on different MOOC platforms and identify ways to increase course interaction. Also, future studies should investigate the impact of the course design, course content, and the instructor's feedback on student success in MOOCs.

Conclusion

This qualitative exploratory study was designed to explore the strategies that computer sciences educators need to use to prepare machine learning datasets for predicting student dropout rates MOOCs. This study explored the strategies of participants who have experience in machine learning, computer science, and MOOCs. A total of 25 participants were used to collect data for this qualitative study. A qualitative survey with 11 questioner's questions was used to assist in answering the research question.

Current scholarly and practitioner literature was used with a qualitative approach to establish the strategies that computer science educators need to use to prepare machine learning dataset for predicting dropout rates of students in MOOCs. The literature was organized into eight categories to assist in adding to the body of knowledge. Literature collection began with the history of Machine-Learning Datasets, the Influences of Machine-Learning for predicting, MOOCs, MOOC Student Dropout Rates, big data and analytics, the role of big data and analytics in MOOC, analyzing big data to Predict Student dropouts in MOOC, and the elements of data used by algorithms to predict drop-out in MOOC. The strategies computer science educators need in MOOCs experienced under each of these eight categories formed the contextual framework for the literature review. The literature review was compared and contrasted with the findings that were presented in Chapter 4.

Several components /categories of the strategies used to prepare machine learning datasets for predicting dropouts faced by computer science educators in MOOCs were discussed in Chapter 4. The key findings of this study were: algorithms or predictive models, the data elements needed in the datasets, and the MOOCs experience. The following conclusions presented contribute to the field of massive open online courses (MOOCs), machine learning,

and computer science by examining the literature alongside the study findings. As a result, these conclusions were a combination of the lived experiences of 25 participants at the LinkedIn community that they have a background in computer science, machine learning, and MOOCs.

Algorithms or predictive models for predicting dropout rates of students in MOOCs

There are many algorithms and predictive models that help to reduce student dropout in the MOOCs. Umer, Susnjak, Mathrani, and Suriadi (2017) used different machine-learning algorithms in their experiments to identify which type of algorithms outperformed the others. This study's goal was to explore the strategies computer science educators needed to use to prepare machine-learning datasets for predicting dropout rates of students in MOOCs. The findings of the study confirmed that the logistic regression algorithm is the best algorithm for predicting dropout rates of students in MOOCs

Datasets for predicting dropout rates of students in MOOCs

The performance of the algorithms and predictive models depend on increasing the data elements in the datasets to get accurate results for predicting dropout rates of students in MOOCs. With the diversity of many algorithms and techniques in the area of machine learning such as deep learning and other methods, there is a need to identify the required datasets in most fields such as educational technology and MOOCs (Hernández, Herrera, Tomás, Tomás &, Navarro,2019) The findings of the study confirmed that online learning behavior data, student behavior data, and assignment records were important date elements in the datasets used by algorithms or predictive models for predicting dropout rates of students in MOOCs.

MOOCs experience

The MOOCs experience was based on several factors that help to improve the platforms of the MOOCs and to increase interaction to address the problem of student dropouts. The MOOC experience components such as course content, course design, and the instructor's feedback are

critical to keeping students in the courses. Researchers are particularly interested in understanding why students are dropping out of MOOC (Kolowich, 2013). The findings of the study stressed the importance of the course content. MOOC courses should be of high quality and have value to the students. Also, course design should cater to the different learning styles of students to attract them to MOOCs platforms. Finally, instructors should give timely feedback and recommendations to the students, thus increasing the interaction in MOOCs.

Final Statements

Wang, Yu, and Miao (2017) mentioned that the strategies computer science educators need to use to prepare machine-learning datasets for predicting dropout rates of students in massive open online courses have not been established. Exploring the strategies needed by computer science educators has become important to increase interaction in the MOOCs and for predicting the dropout rates of students. Three principals themes that had discussed by the researcher in this study, which was: algorithms or predictive models to predict dropout rates, the data elements in the datasets, and the MOOC experience. The research question in this study was "what are the strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses." These findings generally align with previous research in areas such as computer science educators, machine learning, and MOOCs and that led to the conclusion and findings of this study that answered the research question such as algorithms or predictive models, datasets, and MOOCs experience for predicting dropout rates of students in MOOCs.

REFERENCES

Abeer, W., & Miri, B. (2014). Students' preferences and views about learning in a MOOC. Procedia-Social and Behavioral Sciences, 152, 318-323.

Abubakar, Y., & Ahmad, N. B. H. (2017). Prediction of students' performance in e-learning environment using random forest. *International Journal of Innovative Computing, 7(2).*

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

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.

Alumu, S., & Thiagarajan, P. (2016). Massive open online courses and E-learning in higher education. *Indian Journal of Science and Technology*, 9(6).

Anderson, G., & Arsenault, N. (2005). *Fundamentals of educational research*. Routledge.

Anderson, T. (2013). Promise and/or peril: MOOCs and open and distance education. *Commonwealth of learning*, *3*, 1-9.

Angluin, D. (1992, July). Computational learning theory: survey and selected bibliography. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing* (pp. 351-369). ACM.

Anney, V. N. (2014). Ensuring the quality of the findings of qualitative research: Looking at trustworthiness criteria. *Journal of Emerging Trends in Educational Research and Policy Studies (JETERAPS)*, *5*(2), 272-281.

Amnueypornsakul, B., Bhat, S., & Chinprutthiwong, P. (2014, October). Predicting attrition along the way: The UIUC model. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 55-59).

Alario-Hoyos, C., Estévez-Ayres, I., Pérez-Sanagustín, M., Kloos, C. D., & Fernández-Panadero, C. (2017). Understanding learners' motivation and learning strategies in MOOCs. *The International Review of Research in Open and Distributed Learning*, *18*(3).

Baker, R., Evans, B., Greenberg, E., & Dee, T. (2014). Understanding persistence in moocs (massive open online courses): Descriptive & experimental evidence. *EMOOCs*, 5-10.

Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, *7*(3), 112-118.

Bates, A. T. (2018). *Teaching in a digital age: Guidelines for designing teaching and learning*.

Baturay, M. H. (2015). An overview of the world of MOOCs. *Procedia-Social and Behavioral Sciences*, *174*, 427-433.

Baturay, M. H. (2015). An overview of the world of MOOCs. *Procedia-Social and Behavioral Sciences*, *174*, 427-433.

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

Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report*, *13*(4), 544-559.

Bell, E., Bryman, A., & Harley, B. (2018). *Business research methods*. Oxford university press.

Bernard, H. R., Wutich, A., & Ryan, G. W. (2016). *Analyzing qualitative data: Systematic approaches*. SAGE publications.

Berry, M. J., & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support*: John Wiley & Sons, Inc.

Bhadani, A. K., & Jothimani, D. (2016). Big data: challenges, opportunities, and realities. In *Effective Big Data management and opportunities for implementation* (pp. 1-24). IGI Global.

Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., & Dera, D. (2017). Machine learning in transportation data analytics. In *Data Analytics for Intelligent Transportation Systems* (pp. 283-307). Elsevier.

Bonk, C. J., Lee, M. M., Reeves, T. C., & Reynolds, T. H. (Eds.). (2015). *MOOCs and open education around the world*. Routledge.

Boynton, P. M., & Greenhalgh, T. (2004). Selecting, designing, and developing your questionnaire. *Bmj*, *328*(7451), 1312-1315.

Boyer, S., & Veeramachaneni, K. (2015, June). Transfer learning for predictive models in massive open online courses. In *International conference on artificial intelligence in education* (pp. 54-63). Springer, Cham.

Brace, I. (2018). *Questionnaire design: How to plan, structure and write survey material for effective market research*. Kogan Page Publishers.

Bradshaw, M., & Stratford, E. (2010). Qualitative research design and rigour.

Brinton, C. G., Buccapatnam, S., Chiang, M., & Poor, H. V. (2016). Mining MOOC clickstreams: Video-watching behavior vs. in-video quiz performance. *IEEE Transactions on Signal Processing*, *64*(14), 3677-3692.

Bromley, E., Mikesell, L., Jones, F., & Khodyakov, D. (2015). From subject to participant: Ethics and the evolving role of community in health research. *American Journal of Public Health*, *105*(5), 900-908.

Brooks, C., Thompson, C., & Teasley, S. (2015, March). A time series interaction analysis method for building predictive models of learners using log data. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 126-135). ACM.

Bryant, T. (2015). Bringing the social back to MOOCs. *Educause Review*.

Burgos, C., Campanario, M. L., de la Pena, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541-556.

Byrne, M. M. (2001). Evaluating the findings of qualitative research. *AORN journal*, *73*(3), 703-703.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE review*, 42(4), 40.

Carley, K. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological methodology*, 75-126.

Cassar, D. R., de Carvalho, A. C., & Zanotto, E. D. (2018). Predicting glass transition temperatures using neural networks. *Acta Materialia*, *159*, 249-256.

Castillo, N. M., Lee, J., Zahra, F. T., & Wagner, D. A. (2015). MOOCS for development: Trends, challenges, and opportunities. *International Technologies & International Development*, *11*(2), 35.

Chen, D., Feng, Y., Zhao, Z., Jiang, J., & Yu, J. (2014, December). Does MOOC really work effectively. In *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)* (pp. 272-277). IEEE.

Chen, Y., & Zhang, M. (2017, May). Mooc student dropout: Pattern and prevention. In *Proceedings of the ACM Turing 50th Celebration Conference-China* (p. 4). ACM.

Chaplot, D. S., Rhim, E., & Kim, J. (2015, June). Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks. In *AIED Workshops* (Vol. 53, pp. 54-57).

Charmaz, K., & Belgrave, L. L. (2007). Grounded theory. *The Blackwell encyclopedia of sociology*.

Chaturvedi, S., Goldwasser, D., & Daumé III, H. (2014, June). Predicting instructor's intervention in mooc forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1501-1511).

Chen, D., Feng, Y., Zhao, Z., Jiang, J., & Yu, J. (2014, December). Does MOOC really work effectively. In *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)* (pp. 272-277). IEEE.

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). Big data: related technologies, challenges and future prospects.

Chen, Y., Chen, Q., Zhao, M., Boyer, S., Veeramachaneni, K., & Qu, H. (2016, October). DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction. *In 2016 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 111-120). IEEE.

Chen, Y., Chen, Q., Zhao, M., Boyer, S., Veeramachaneni, K., & Qu, H. (2016, October). DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 111-120). IEEE.

Cheng, H., Tan, P. N., Gao, J., & Scripps, J. (2006, April). Multistep-ahead time series prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 765-774). Springer, Berlin, Heidelberg.

Clow, D. (2013, April). MOOCs and the funnel of participation. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 185-189). ACM.

Cohen, A., & Nachmias, R. (2006). A quantitative cost effectiveness model for Web-supported academic instruction. *The Internet and Higher Education*, *9*(2), 81-90.

Conijn, R., Van den Beemt, A., & Cuijpers, P. (2018). Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, *34*(5), 615-628.

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

Cook, M. (2016). *State of the MOOC 2016: A year of massive landscape change for massive open online courses*. Online Course Report.

Cormier, D. (2010). Through the open door. *EDUCAUSE review*, *45*(4), 30-39.

Corrin, L., De Barba, P. G., & Bakharia, A. (2017, March). Using learning analytics to explore help-seeking learner profiles in MOOCs. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 424-428). ACM.

Creswell. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*: Sage publications.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications.

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into practice*, *39*(3), 124-130.

Crossley, S., Dascalu, M., McNamara, D. S., Baker, R., & Trausan-Matu, S. (2017). Predicting success in massive open online courses (MOOCs) using cohesion network analysis. In: Philadelphia, PA: International Society of the Learning Sciences.

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016, April). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 6-14). ACM.

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018, April). MOOC dropout prediction using machine learning techniques: Review and research challenges. In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1007-1014). IEEE.

De Freitas, S. I., Morgan, J., & Gibson, D. (2015). Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. *British Journal of Educational Technology*, *46*(3), 455-471.

De Waard, I., Abajian, S., Gallagher, M. S., Hogue, R., Keskin, N., Koutropoulos, A., & Rodriguez, O. C. (2011). Using mLearning and MOOCs to understand chaos, emergence, and complexity in education. *The International Review of Research in Open and Distributed Learning*, *12*(7), 94-115.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*.

Dixon, P. G., Goodrich, G. B., & Cooke, W. H. (2008). Using teleconnections to predict wildfires in Mississippi. *Monthly Weather Review*, *136*(7), 2804-2811.

Dyumin, A. A., & Andrianova, S. V. (2016, November). Moocs and vendor trainings in academic curriculum: Yet another step towards global university. In *2016 International Conference on Engineering and Telecommunication (EnT)* (pp. 39-44). IEEE.

Elo, Satu, Maria Kääriäinen, Outi Kanste, Tarja Pölkki, Kati Utriainen, and Helvi Kyngäs. "Qualitative content analysis: A focus on trustworthiness." *SAGE open* 4, no. 1 (2014): 2158244014522633.

Engle, W. (2014). UBC MOOC Pilot: Design and Delivery. In: Vancouver BC: University of British Columbia.

Eriksson, P., & Kovalainen, A. (2015). *Qualitative methods in business research: A practical guide to social research*. Sage.

Ewais, A., & Samra, D. A. (2017, October). Adaptive MOOCs: A framework for adaptive course based on intended learning outcomes. In *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)* (pp. 204-209). IEEE.

Feng, S., Zhou, H., & Dong, H. (2019). Using deep neural network with small dataset to predict material defects. *Materials & Design*, *162*, 300-310.

Fini, A. (2009). The technological dimension of a massive open online course: The case of the CCK08 course tools. *The International Review of Research in Open and Distributed*

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with big data analytics. *interactions*, *19*(3), 50-59.

Fei, M., & Yeung, D. Y. (2015, November). Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE International Conference on Data Mining*

García, O. A., & Secades, V. A. (2013). Big Data & Learning Analytics: A Potential Way to Optimize eLearning Technological Tools. *International Association for Development of the Information Society*.

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, *28*(2), 127-203.

Gardner, J., Brooks, C., Andres, J. M., & Baker, R. (2018, June). Replicating MOOC predictive models at scale. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (p. 1). ACM.

Ghobrial, B. G. (2014). Invasion of the MOOCs: The Promises and Perils of Massive Open Online Courses. *Comunicar*, *22*(43), 214.

Ghoting, A., Krishnamurthy, R., Pednault, E., Reinwald, B., Sindhwani, V., Tatikonda, S., ... & Vaithyanathan, S. (2011, April). SystemML: Declarative machine learning on MapReduce. In *2011 IEEE 27th International Conference on Data Engineering* (pp. 231-242). IEEE.

Gitinabard, N., Khoshnevisan, F., Lynch, C. F., & Wang, E. Y. (2018). Your Actions or Your Associates? Predicting Certification and Dropout in MOOCs with Behavioral and Social Features. *arXiv preprint arXiv:1809.00052*.

Glaser, B. G., & Strauss, A. L. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Goggins, S. P., Xing, W., Chen, X., Chen, B., & Wadholm, B. (2015). Learning Analytics at" Small" Scale: Exploring a Complexity-Grounded Model for Assessment Automation. *J.*

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The qualitative report*, *8*(4), 597-606.

Goral, T. (2013). Make way for SPOCS: small, private online courses may provide what MOOCs can't. *University business*, *16*(7), 45.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Ectj*, *29*(2), 75.

Bowen, G. A. (2008). Naturalistic inquiry and the saturation concept: a research note. *Qualitative research*, *8*(1), 137-152.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field methods*, *18*(1), 59-82.

Guo, P. J., & Reinecke, K. (2014, March). Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 21-30). ACM.

Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014, September). Attrition in MOOC: Lessons learned from drop-out students. In *International workshop on learning technology for education in cloud* (pp. 37-48). Springer, Cham.

Hagen, C., Ciobo, M., Wall, D., Yadav, A., Khan, K., Miller, J., & Evans, H. (2013). Big data and the creative destruction of today's business models. *Retrieved January*, *5*, 2015.

Hammersley, M. (2012). *What is qualitative research?*. A&C Black.

Hanus, J. J., & Relyea, H. C. (1975). A policy assessment of the Privacy Act of 1974. *Am. UL Rev.*, *25*, 555.

Harju, M., Leppänen, T., & Virtanen, I. (2018). Interaction and Student Dropout in Massive Open Online Courses. *arXiv preprint arXiv:1810.08043*.

Harvey, W. S. (2011). Strategies for conducting elite interviews. *Qualitative research*, *11*(4), 431-441.

Hassan, Z. A., Schattner, P., & Mazza, D. (2006). Doing a pilot study: why is it essential?. *Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia*, *1*(2-3), 70.

Hathaway, R. S. (1995). Assumptions underlying quantitative and qualitative research: Implications for institutional research. Research in higher education, 36(5), 535-562.

Hathaway, R. S. (1995). Assumptions underlying quantitative and qualitative research: Implications for institutional research. *Research in higher education*, *36*(5), 535-562.

Haykin, S. S. (2009). *Neural networks and learning machines/Simon Haykin*. New York: Prentice Hall,.

He, J., Bailey, J., Rubinstein, B. I., & Zhang, R. (2015, February). Identifying at-risk students in massive open online courses. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Heale, R., & Forbes, D. (2013). Understanding triangulation in research. *Evidence-Based Nursing*, *16*(4), 98-98.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Complexity*, *2019*.

Hew, K. F. (2016). Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS. *British Journal of Educational Technology*, *47*(2), 320-341.

Hidalgo, J. J. (2018). *Exploring the Big Data and Machine Learning Framing Concepts for a Predictive Classification Model* (Doctoral dissertation, Colorado Technical University).

Hmedna, B., El Mezouary, A., Baz, O., & Mammass, D. (2017). Identifying and tracking learning styles in MOOCs: A neural networks approach. *International Journal of Innovation and Applied Studies*, *19*(2), 267.

Hmedna, B., El Mezouary, A., Baz, O., & Mammass, D. (2016). A machine learning approach to identify and track learning styles in MOOCs. In *2016 5th International Conference on Multimedia Computing and Systems (ICMCS)* (pp. 212-216). IEEE.

Holtzhausen, S. (2001). Triangulation as a powerful tool to strengthen the qualitative research design: The Resource-based Learning Career Preparation Programme (RBLCPP) as a case study.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?. *Brain Informatics*, *3*(2), 119-131.

Hone, K. S., & El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. *Computers & Education*, *98*, 157-168.

Hong, B., Wei, Z., & Yang, Y. (2017, August). Discovering learning behavior patterns to predict dropout in MOOC. In *2017 12th International Conference on Computer Science and Education (ICCSE)* (pp. 700-704). IEEE.

Irfan, G. (2018). *Exploring the pplication security measures in hive to secure data in column*. (Doctoral dissertation, Colorado Technical University).

Jackson, J. (2002). Data mining; a conceptual overview. *Communications of the Association for Information Systems*, *8*(1), 19.

Jamshed, S. (2014). Qualitative research method-interviewing and observation. *Journal of basic and clinical pharmacy*, *5*(4), 87.

Javadi, M., & Zarea, K. (2016). Understanding thematic analysis and its pitfall. *Journal of Client Care*, 1(1), 33-39.

Jiang, S., Williams, A., Schenke, K., Warschauer, M., & O'dowd, D. (2014, July). Predicting MOOC performance with week 1 behavior. In *Educational data mining 2014*.

Johnson, B., & Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches*: Sage.

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning, 15*(1).

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Joseph, A. M., & Nath, B. A. (2013). Integration of Massive Open Online Education (MOOC) System with in-Classroom Interaction and Assessment and Accreditation: An extensive report from a pilot study. In *Proceedings of the international conference on e-learning, e-business, enterprise information systems, and e-Government (EEE)* (p. 105). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Joseph , J. K. (2017). Reimagining the role of technology in education: 2017 National Education Technology Plan update Rerrived from https://tech.ed.gov/higherednetp/. In: Office of Educational Technology Washington, DC.

Kashyap, A., & Nayak, A. (2018, September). Different Machine Learning Models to predict dropouts in MOOCs. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 80-85). IEEE.

Kaur, K., & Kaur, K. (2015, September). Analyzing the effect of difficulty level of a course on students performance prediction using data mining. I*n 2015 1st International Conference on Next Generation Computing Technologies (NumerGCT) (pp. 756-761). IEEE.*

Kaur, K., & Kaur, K. (2015, September). Analyzing the effect of difficulty level of a course on students performance prediction using data mining. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)* (pp. 756-761). IEEE.

Kaveri, A., Gunasekar, S., Gupta, D., & Pratap, M. (2015, October). Decoding the Indian MOOC learner. In *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)* (pp. 182-187). IEEE.

Kaveri, A., Gunasekar, S., Gupta, D., & Pratap, M. (2016, December). Decoding Engagement in MOOCs: An Indian Learner Perspective. In *2016 IEEE Eighth International Conference on Technology for Education (T4E)* (pp. 100-105). IEEE.

Kennedy, G., Coffrin, C., De Barba, P., & Corrin, L. (2015, March). Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance. In *Proceedings*

of the fifth international conference on learning analytics and knowledge (pp. 136-140). ACM.

Khalifa, S., Elshater, Y., Sundaravarathan, K., Bhat, A., Martin, P., Imam, F., ... & Statchuk, C. (2016). The six pillars for building big data analytics ecosystems. *ACM Computing Surveys (CSUR)*, *49*(2), 33.

Khalil, M. (2018). Learning Analytics in Massive Open Online Courses. *arXiv preprint arXiv:1802.09344*.

Khalil, H., & Ebner, M. (2014, June). MOOCs completion rates and possible methods to improve retention-A literature review. In *EdMedia+ Innovate Learning* (pp. 1305-1313). Association for the Advancement of Computing in Education (AACE).

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014, March). Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 31-40). ACM.

Kivunja, C., & Kuyini, A. B. (2017). Understanding and Applying Research Paradigms in Educational Contexts. *International Journal of Higher Education,* 6(5), 26-41.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. Paper presented at *the Proceedings of the third international conference on learning analytics and knowledge.*

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014, October). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs* (pp. 60-65).

Knapp, T. R. (2015). The reliability of measuring instruments. Vancouver, BC: Edgeworth Laboratory for Quantitative Educational and Behavioral Science Series. Dostopno na URL: *http://www. educ. ubc. ca/faculty/zumbo/series/knapp/index. htm.*

Kohavi, R. (1998). Glossary of terms. *Special issue on applications of machine learning and the knowledge discovery process*, *30*(271), 127-132.

Kolowich, S. (2013). Coursera takes a nuanced view of MOOC dropout rates. *The chronicle of higher education*.

Kop, R., Fournier, H., & Mak, J. S. F. (2011). A pedagogy of abundance or a pedagogy to support human beings? Participant support on massive open online courses. *The International Review of Research in Open and Distributed Learning*, *12*(7), 74-93.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

Kotsiantis, S. (2009). Educational data mining: a case study for predicting dropout-prone students. *International Journal of Knowledge Engineering and Soft Data Paradigms*, *1*(2), 101-111.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, 8-17.

Labaree, R. (2013). Organizing Your Social Sciences Research Paper: Types of Research Designs. USC Libraries Research Guides.

Li, W., Gao, M., Li, H., Xiong, Q., Wen, J., & Wu, Z. (2016, July). Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. In *2016 international joint conference on neural networks (IJCNN)* (pp. 3130-3137). IEEE.

Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016, April). Big data application in education: dropout prediction in edx MOOCs. In *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)* (pp. 440-443). IEEE.

Laveti, R. N., Kuppili, S., Ch, J., Pal, S. N., & Babu, N. S. C. (2017, August). Implementation of learning analytics framework for MOOCs using state-of-the-art in-memory computing. *In 2017 5th National Conference on E-Learning & E-Learning Technologies (ELELTECH) (pp. 1-6). IEEE.*

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), 321.

Lincoln, Y. S., & Guba, E. G. (2012). Naturalistic inquiry (Vol. 75): *Sage.*

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distributed Learning*, 14(3), 202-227.

Lynda, H., & Dahmani, F. B. (2016, July). An assessment planner for MOOCs based ODALA approach. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (pp. 855-862). IEEE.

Mack, N. (2005). Qualitative research methods: A data collector's field guide.

Mack, N. (2014). Qualitative research methods: A data collector's field guide.

Mackness, J., Mak, S., & Williams, R. (2010). The ideals and reality of participating in a MOOC. Paper presented at *the Proceedings of the 7th international conference on networked learning 2010.*

Maitland, C., & Obeysekare, E. (2015). The creation of capital through an ICT-based learning program: a case study of MOOC camp. Paper presented at *the Proceedings of the Seventh International Conference on Information and Communication Technologies and Development.*

Makarova, A., Mikhailov, D., & Shako, V. (2014). Hydrodynamic Simulations of Dynamics of Near-Wellbore Zone Properties during Drilling and Cleanup Procedures for Open-Hole Well. *Oil & Gas Technologies*, 92(3).

Maltby, D. (2011). Big data analytics. Paper presented at the 74th *Annual Meeting of the Association for Information Science and Technology* (ASIST).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: *The next frontier for innovation, competition, and productivity*.

Marie, Mingyu, & Barbara. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Paper presented at *the Proceedings of conference on advanced technology for education.*

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, *38*(3), 315-330.

Marr, B. (2016). What is the Difference Between Artificial Intelligence and Machine Learning. *In: Forbes.*

Marshall, C., & Rossman, G. B. (2014). *Designing qualitative research*. Sage publications.

Martínez-Mesa, J., González-Chica, D. A., Bastos, J. L., Bonamigo, R. R., & Duquia, R. P. (2014). Sample size: how many participants do I need in my research?. *Anais brasileiros de dermatologia*, *89*(4), 609-615.

Mayer, I. (2015). Qualitative research with a focus on qualitative data analysis. *International Journal of Sales, Retailing & Marketing,* 4(9), 53-67.

Mayhew, L. A. (2017). *Perceived Right to Privacy by Information Systems Professionals in Big Data Collection* (Doctoral dissertation, Colorado Technical University).

McGuirk, P. M., & O'Neill, P. (2016). Using questionnaires in qualitative human geography.

Mduma, N., Kalegele, K., & Machuve, D. (2019). A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). *An artificial intelligence approach*. Springer Science & Business Media.

Miller, T., Birch, M., Mauthner, M., & Jessop, J. (2012). *Ethics in qualitative research*: Sage.

Misra, P. (2018). MOOCs for Teacher Professional Development: Reflections and Suggested Actions. *Open Praxis*, 10(1), 67-77.

Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, 18(3), 11.

Moin, K. I., & Ahmed, D. Q. B. (2012). Use of data mining in banking. *International Journal of Engineering Research and Applications*, 2(2), 738-742.

Morabito, V. (2015). Big data and analytics. *Strategic and organisational impacts*.

Morgan-Trimmer, S., & Wood, F. (2016). Ethnographic methods for process evaluations of complex health behaviour interventions. *Trials*, 17(1), 232.

Most, M. M., Craddick, S., Crawford, S., Redican, S., Rhodes, D., Rukenbrod, F., Group, D.-S. C. R. (2003). Dietary quality assurance processes of the DASH-Sodium controlled diet study. *Journal of the American Dietetic Association, 103(10), 1339-1346.*

Nagrecha, S., Dillon, J. Z., & Chawla, N. V. (2017). MOOC dropout prediction: lessons learned from making pipelines interpretable. Paper presented at *the Proceedings of the 26th International Conference on World Wide Web Companion.*

Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4), 049901.

Nawrot, I., & Doucet, A. (2014). Building engagement for MOOC students: introducing support for time management on online learning platforms. Paper presented at *the Proceedings of the 23rd International Conference on world wide web.*

Nicholson, P. (2007). A history of e-learning. *In Computers and education* (pp. 1-11): Springer.

Niemi, D., & Gitin, E. (2012). Using Big Data to Predict Student Dropouts: Technology Affordances for Research. *International Association for Development of the Information Society.*

Nordhagen, S., Calverley, D., Foulds, C., O'Keefe, L., & Wang, X. (2014). Climate change research and credibility: balancing tensions across professional, personal, and public domains. *Climatic change*, 125(2), 149-162.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., Estévez-Ayres, I., & Kloos, C. D. (2018, April). Sentiment Analysis in MOOCs: *A case study*. In 2018 *IEEE Global Engineering Education Conference (EDUCON) (pp. 1489-1496). IEEE.*

Northcraft, T. G. (2017). *A qualitative study of the relationship between a banking it troubled project and the executive project sponsor's project management maturity level* (Doctoral dissertation, Colorado Technical University).

O'Cathain, A., & Thomas, K. J. (2004). " Any other comments?" Open questions on questionnaires–a bane or a bonus to research? *BMC medical research methodology*, 4(1), 25.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13), 1216.

Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 proceedings*, 5825-5834.

Onwuegbuzie, A. J., Leech, N. L., & Collins, K. M. (2012). Qualitative analysis techniques for the review of the literature. *The qualitative report*, 17(28), 1-28.

Owonikoko, T. K. (2013). Upholding the principles of autonomy, beneficence, and justice in phase I clinical trials. *The oncologist*, 18(3), 242-244.

Padilla-Díaz, M. (2015). Phenomenology in educational qualitative research: Philosophy as science or philosophical science. *International Journal of Educational Excellence,* 1(2), 101-110.

Pappano, L. (2012). The Year of the MOOC-The New York Times. Retrrived from http://www. nytimes. com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace. html.

Penwarden, R. (2014). Exploratory Research: What is it? And 4 Ways to Implement it in Your Research! In.

Perrier, L., Buja, A., Mastrangelo, G., Baron, P. S., Ducimetière, F., Pauwels, P. J., . Favier, B. (2014). Transferability of health cost evaluation across locations in oncology: cluster and principal component analysis as an explorative tool. *BMC health services research*, 14(1), 537.

Pilli, O., & Admiraal, W. (2017). Students' Learning Outcomes in Massive Open Online Courses (MOOCs): Some Suggestions for Course Design. *Journal of Higher Education/Yükseköğretim Dergisi,* 7(1).

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

Rasmussen, C. E. (2004). Gaussian processes in machine learning. *In Advanced lectures on machine learning* (pp. 63-71): Springer.

Ramesh, A., Goldwasser, D., Huang, B., Daume III, H., & Getoor, L. (2014, March). Uncovering hidden engagement patterns for predicting learner performance in MOOCs. In Proceedings of *the first ACM conference on Learning@ scale conference (pp. 157-158). ACM.*

Reich, J. (2015). Rebooting MOOC research. *Science*, 347(6217), 34-35.

Ren, Z., Rangwala, H., & Johri, A. (2016). Predicting performance on MOOC assessments using multi-regression models. *arXiv preprint arXiv:1605.02269.*

Rivard, R. (2013). Measuring the MOOC dropout rate. *Inside Higher Ed*, 8, 2013.

Rodrigues, R. L., Ramos, J. L., Silva, J. C. S., Gomes, A. S., de Souza, F. D. F., & Maciel, A. M. A. (2016, July). Discovering level of participation in MOOCs through clusters analysis. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)* (pp. 232-233). IEEE.

Rodriguez, C. O. (2012). MOOCs and the AI-Stanford Like Courses: Two Successful and Distinct Course Formats for Massive Open Online Courses. *European Journal of Open, Distance and E-Learning.*

Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., & Sherer, J. (2014). Social factors that contribute to attrition in MOOCs. Paper presented at the Proceedings of *the first ACM conference on Learning @ scale conference - L@S '14.*

Sachdeva, A., Singh, P. K., & Sharma, A. (2015). MOOCs: A comprehensive study to highlight its strengths and weaknesses. Paper presented at *the MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference on.*

Sapp, C. (2017). Preparing and architecting for machine learning. *Gartner Technical Professional Advice*, 1-37.

Sargeant, J. (2012). Qualitative research part II: Participants, analysis, and quality assurance.

Schaffer, J., Huynh, B., O'Donovan, J., Höllerer, T., Xia, Y., & Lin, S. (2016, August). An analysis of student behavior in two massive open online courses. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 380-385). IEEE Press.

Scheider, S., Ostermann, F. O., & Adams, B. (2017). Why good data analysts need to be critical synthesists. Determining the role of semantics in data analysis. *Future Generation Computer Systems*, 72, 11-22.

Schulze, A. S., Leigh, D., Sparks, P., & Spinello, E. (2017). Massive Open Online Courses and Completion Rates: Are Self-Directed Adult Learners the Most Successful at MOOCs? *In Handbook of Research on Individualism and Identity in the Globalized Digital Age* (pp. 24-49): IGI Global.

Shafiq, H., Ashraf, Z., Mahajan, I. M., & Qadri, U. (2017). Courses beyond borders: A case study of MOOC platform Coursera. *Library Philosophy and Practice*, 1-15.

Shah, D. (2018). By the numbers: MOOCs in 2017. Retrieved from https://www.class-central.com/report/mooc-stats-2017/.

Sharkey, M., & Sanders, R. (2014). A process for predicting MOOC attrition. Paper presented at the Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs.

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for information*, 22(2), 63-75.

Sheth, B. D. (2013). A *learning approach to personalized information filtering. Massachusetts Institute of Technology*, (Doctoral dissertation, Massachusetts Institute of Technology).

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5), 30.

Silverman, D. (2016). *Qualitative research*: Sage.

Simon, M. K., & Goes, J. (2013). *Assumption, limitations, delimitations, and scope of the study*. In: Doctoral dissertation, Dissertation and scholarly Research: Recipes for success.

Sinha, T. (2014). " Your click decides your fate": Leveraging clickstream patterns from MOOC videos to infer students' information processing & attrition behavior. *arXiv preprint arXiv*:1407.7143.

Sinha, T., Li, N., Jermann, P., & Dillenbourg, P. (2014). Capturing" attrition intensifying" structural traits from didactic interaction sequences of MOOC learners. *arXiv preprint arXiv:*1409.5887.

Simon, M. K., & Goes, J. (2013). Assumptions, limitations, delimitations, and scope of the study. Retrieved from dissertationrecipes. com.

Skeels, M. M., & Grudin, J. (2009). When social networks cross boundaries: a case study of workplace use of facebook and linkedin. Paper presented at *the Proceedings of the ACM 2009 international conference on Supporting group work.*

Smith, J., & Noble, H. (2014). Bias in research. Evidence-based nursing, ebnurs-2014-101946.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. Paper presented at *the Advances in neural information processing systems.*

Sotiriadou, P., Brouwers, J., & Le, T.-A. (2014). Choosing a qualitative data analysis tool: A comparison of NVivo and Leximancer. *Annals of Leisure Research*, 17(2), 218-234.

Srilekshmi, M., Sindhumol, S., Chatterjee, S., & Bijlani, K. (2016). Learning Analytics to Identify Students At-risk in MOOCs. Paper presented at *the Technology for Education (T4E), 2016 IEEE Eighth International Conference on.*

Srinivasa, S., & Bhatnagar, V. (2012). Big data analytics. Paper presented at *the Proceedings of the First International Conference on Big Data Analytics BDA.*

Srinivasan, R., & Lohith, C. (2017). Pilot Study—Assessment of validity and reliability. *In Strategic Marketing and Innovation for Indian MSMEs* (pp. 43-49): Springer.

Stebbins, R. A. (2001). *Exploratory research in the social sciences* (Vol. 48): Sage.

Sunar, A. S., White, S., Abdullah, N. A., & Davis, H. C. (2017). How Learners' Interactions Sustain Engagement: A MOOC Case Study. *IEEE Transactions on Learning Technologies*, 10(4), 475-487. doi:10.1109/tlt.2016.2633268

SurveyMonkey and IRB Guidelines (n.d). Retrieved from https://help.surveymonkey.com/articles/en_US/kb/How-does-SurveyMonkey-adhere-to-IRB-guidelines.

Suter, W. N. (2012). Qualitative data, analysis, and design. Introduction to educational research: *A critical thinking approach*, 2, 342-386.

Taylor, C., Veeramachaneni, K., & O'Reilly, U. M. (2014). Likely to stop? predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*.

Teherani, A., Martimianakis, T., Stenfors-Hayes, T., Wadhwa, A., & Varpio, L. (2015). Choosing a qualitative research approach. *Journal of graduate medical education*, 7(4), 669-670.

Thaimoocs. (2017). Thaimooc Retrieved from https://www.thaimooc.org/.

Thaipisutikul, T., & Tuarob, S. (2017, August). MOOCs as an intelligent online learning platform in Thailand: Past, present, future challenges and opportunities. In *2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)* (pp. 1-5). IEEE.

Theuveny, B., Mikhailov, D., Spesivtsev, P., Starostin, A., Osiptsov, A., Sidorova, M., & Shako, V. (2013). Integrated approach to simulation of near-wellbore and wellbore cleanup. Paper presented at *the SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, S.(2017) Prediction of Students' Dropout in MOOC Environment.

Wang, W., Yu, H., & Miao, C. (2017, July). Deep model for dropout prediction in MOOCs. *In Proceedings of the 2nd International Conference on Crowd Science and Engineering*(pp. 26-32). ACM.

Wang , Y., & Baker, R. (2015). Content or platform: Why do students complete MOOCs. *MERLOT Journal of Online Learning and Teaching*, 11(1), 17-30.

Welsh, D. H., & Dragusin, M. (2013). The new generation of massive open online course (MOOCS) and entrepreneurship education. *Small Business Institute Journal*, 9(1), 51-65.

Wen, M., Yang, D., & Rose, C. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? Paper presented at *the Educational data mining 2014.*

Whitt, E. J., & Kuh, G. D. (1989). Qualitative Methods in Higher Education Research: A Team Approach to Multiple Site Investigation. ASHE Annual Meeting Paper.

Wildemuth, B. M. (2016). *Applications of social research methods to questions in information and library science:* ABC-CLIO.

Wisdom, J., & Creswell, J. W. (2013). *Mixed methods: integrating quantitative and qualitative data collection and analysis while studying patient-centered medical home models*. Rockville: Agency for Healthcare Research and Quality.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Wong, B. T.-m. (2016). Factors leading to effective teaching of MOOCs. *Asian Association of Open Universities Journal*, 11(1), 105-118.

Wu, B., & Chen, X. (2015). Research on MOOCs continuance. Paper presented at *the 3rd International Conference on Material, Mechanical and Manufacturing Engineering.*

Xiao, C., Qiu, H., & Cheng, S. M. (2019). Challenges and opportunities for effective assessments within a quality assurance framework for MOOCs. J*ournal of Hospitality, Leisure, Sport & Tourism Education*, 24, 1-16.

Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior,* 58, 119-129. doi:10.1016/j.chb.2015.12.007

Xing, W., & Du, D. (2018). Dropout Prediction in MOOCs: Using Deep Learning for Personalized *Intervention. Journal of Educational Computing* Research. doi:10.1177/0735633118757015

Xing, W., & Goggins, S. (2015). Learning analytics in outer space: a Hidden Naïve Bayes model for automatic student off-task behavior detection. Paper presented at the Proceedings of *the Fifth International Conference on Learning Analytics And Knowledge.*

Xing, W., Kim, S. M., & Goggins, S. (2015). Modeling performance in asynchronous CSCL: an exploration of social ability, collective efficacy and social interaction. In: International Society of the Learning Sciences, Inc.[ISLS].

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

Yaakob, R., Mustapha, N., Nuruddin, A. A. B., & Sitanggang, I. S. (2011). Modeling forest fires risk using spatial decision tree. Paper presented at *the Data Mining and Optimization (DMO), 2011 3rd Conference on.*

Ye, C., & Biswas, G. (2014). Early prediction of student dropout and performance in MOOCs using higher granularity temporal information. *Journal of Learning Analytics, 1*(3), 169-172.

Yilmaz, K. (2013). Comparison of quantitative and qualitative research traditions: Epistemological, theoretical, and methodological differences. *European Journal of Education*, 48(2), 311-325.

Yin, R. K. (2003). Case study research design and methods third edition. *Applied social research methods series, 5*.

Yin, R. K. (2015). *Qualitative research from start to finish*: Guilford Publications.

Chen, Y. (2014). Investigating MOOCs through blog mining. *The International Review of Research in Open and Distributed Learning*, 15(2).

Yu, X., & Wu, S. (2015, October). Typical applications of big data in education. In *2015 International Conference of Educational Innovation through Technology (EITT)* (pp. 103-106). IEEE.

Yuan, L., & Powell, S. J. (2013). MOOCs and open education: Implications for higher education.

Yuan, L., Powell, S. J., & Olivier, B. (2014). Beyond MOOCs: Sustainable online learning in institutions.

Zamawe, F. C. (2015). The implication of using NVivo software in qualitative data analysis: Evidence-based reflections. *Malawi Medical Journal*, 27(1), 13-15.

Zheng, Chen, L., & Burgos, D. (2018). The International Comparison and Trend Analysis of the Development of MOOCs in Higher Education. *In The Development of MOOCs in China* (pp. 1-9): Springer.

Zheng, S., Rosson, M. B., Shih, P. C., & Carroll, J. M. (2015, February). Understanding student motivation, behaviors and perceptions in MOOCs. *In Proceedings of the 18th ACM*

conference on computer supported cooperative work & social computing (pp. 1882-1895). ACM.

Zheng, Y., & Yin, B. (2015, October). Big Data Analytics in MOOCs. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing* (pp. 681-686). IEEE.

Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2013). *Business research methods*: Cengage Learning.

APPENDICIES
APPENDIX A: LETTER OF PERMISSION TO USE SURVEYMONKEY SITE"

SurveyMonkey Inc.
www. surveymonkey.com

For questions, visit our Help Center
help.survemonkey.com

Re: Permission to Conduct Research Using SurveyMonkey

To Whom It May Concern:

This letter is being produced in response to a request by a student at your institution who wishes to conduct a survey using SurveyMonkey in order to support their research. The student has indicated that they require a letter from SurveyMonkey granting them permission to do this. Please accept this letter as evidence of such permission. Students are permitted to conduct research via the SurveyMonkey platform provided that they abide by our Terms of Use at https://www.surveymonkey.com/mp/legal/terms-of-use/.

SurveyMonkey is a self-serve survey platform on which our users can, by themselves, create, deploy and analyze surveys through an online interface. We have users in many different industries who use surveys for many different purposes. One of our most common use cases is students and other types of researchers using our online tools to conduct academic research.

If you have any questions about this letter, please contact us through our Help Center at help.surveymonkey.com.

Sincerely,

SurveyMonkey Inc.

APPENDIX B: INFORMED CONSENT

Title of Study: EXPLORING THE STRATEGIES COMPUTER SCIENCE EDUCATORS NEED TO USE TO PREPARE MACHINE-LEARNING DATASETS FOR PREDICTING THE DROPOUT RATES OF STUDENTS IN MASSIVE OPEN ONLINE COURSES

Investigator: Houssen Nafed

Contact Number:

Purpose of the Study

You are invited to participate in a research study. The purpose of the research qualitative exploratory study is to explore the strategies Computer Scientists need to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses (MOOCs)

Participants

You are being asked to participate in the study because of your profile at LinkedIn website, and who has academic backgrounds related to machine learning in computer science. Your opinions, outlook, and insights with respect to your experiences are critical to explore the strategies Computer Scientists need to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses (MOOCs)

Procedures

If you volunteer to participate in this study, you will be asked to do the following: select Click on "Start Survey," read each of the questions carefully and select an answer for each question. Once you have answered all the questions, click on "Save."

Benefits of Participation

There may/may not be direct benefits to you as a participant in this study. However, we hope to learn different perspectives and experiences to determine the strategies Computer Scientists need to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses (MOOCs)

Risks of Participation

There are risks involved in all research studies. This study may include only minimal risks. You may not see the ultimate results of your input to this study. You may also feel somewhat uncomfortable in answering some of the questions since it may involve personal opinions, beliefs, or experiences in regarding the strategies Computer Scientists need to prepare machine-learning datasets for predicting the dropout rates of students in massive open online courses (MOOCs)

Cost/Compensation

There will not be financial cost to you to participate in this study. The study will take approximately 10 minutes of your time. You will not be compensated for your time. Colorado Technical University (CTU) may not provide compensation or free medical care for an unanticipated injury sustained as a result of participating in this research study.

Contact Information

If you have any questions or concerns about the study, you may contact Dr. Trish Elley, Committee Chair, at TElley@coloradotech.edu, or by mail at 4435 North Chestnut, Colorado Springs, Colorado 80901. For questions regarding the rights of research subjects, any complaints or comments regarding the manner in which the study is being conducted, you may contact the Colorado Technical University, Doctoral Programs, at 719.598.0200.

Voluntary Participation

Your participation in this study is voluntary. You may refuse to participate in this study or in any part of this study. You may withdraw at any time without prejudice to your relations with the university. You are encouraged to ask questions about this study at the beginning or any time during the research study.

Confidentiality

All information gathered in this study will be kept completely confidential. No reference will be made in written or oral materials that could link you to this study. Your name will be number coded and used for references purposes only. The researcher will be the only individual who will know your identity. All your answers will be stored in a password protected laptop that only the researcher has access and password knowledge. The laptop will be locked safe in Falcon, Colorado, where only the researcher has access for at least 5 years after completion of the study and dissertation publication. After the storage time has elapsed, all notes and answers gathered will be destroyed in accordance with CTU policy.

Participant Consent

I have read the above information and agree to participate in this study. I am at least 18 years of age. A copy of this form has been given to me.

By clicking on "Start Survey" I agree to all the above terms and conditions.
I Acknowledge the Informed Consent information
<<START SURVEY>>

_____ _____

Signature of Participant Date

_____

Participant Name (Please Print)

APPENDIX C: QUESTIONNAIRE

Questionnaire questions:

Pre-qualifying questions:

1. At least one year of experience in Computer Science (Yes/No)
2. At least one year of experience in MOOC (Yes/No)
3. At least one year of experience in machine learning (Yes/No)
4. Do you meet one of the criteria to participate in the survey such as computer science or machine learning or massive open online courses (MOOC)? If you choose "No" please stop to start the survey.

Questions:

2. Which predictive models or algorithms do you think to use to predict dropout rate of students in MOOC? Check all that apply. If none of these apply, please describe the model or algorithm you use in the other textbox.
    a. Logistic Regression
    b. Deep Neural Network
    c. Support Vector Machine
    d. Hidden Markov Models
    e. Recurrent Neural Network
    f. Natural Language Processing Technique
    g. Decision Trees
    h. Survival Analysis
    i. Bayesian Network

    Other (describe what other model or algorithm you use):_____

2. Based on your experience with machine-learning, what type of predictive models or algorithms that can be used to get better performance? Please elaborate

3.One type of machine learning is Supervised learning, Do you prefer to use logistic regression, Support Vector Machines (SVM), and Decision Trees when performing supervised learning?
If this is not applicaple to your work experince, put (N/A) in box below.

4.Another type of machine learning is Unsupervised Learning , Do you prefer to use k-means clustering, and/or Association Rules when performing Unsupervised Learning? If this is not applicaple to your work experince, put (N/A) in box below.

5.The third type of machine learning is Semi-supervised which is a mix between supervised learning and unsupervised learning. What algorithms do you prefer to use when you utilize

this type of method? If this is not applicaple to your work experince, put (N/A) in box below.

6. The fourth type of machine learning is Reinforcement Learning. Do you prefer to use Adversarial Networks, and/or Temporal Difference (TD) Reinforcement Learning? If this is not applicable to your work experience, put (N/A) in box below.

7. Based on your experience with MOOCs, what improvements computer scientists educators need to make to increase interaction within the MOOC platforms?

8. Based on your experiences with MOOC, how does course content design impact student interaction in MOOC?

9. How does instructor involvement with the students help improve interaction within the MOOC platforms?. If this is not applicable to you, then type in the box below N/A.

10. How do you determine appropriate machine-learning datasets for predicting the dropout rates of students in MOOCs?

11. What type of data will help determine computer scientists to prepare a machine-learning dataset in the MOOC ? Check all that apply. If you use other datasets, please describe in q. (other).

a. Online learning behavior
b. Student behavior
c. Postings
d. Demographics
e. Clickstreams
f. Stream server logs
g. Graded activities within courses
h. Forum posts and discussion
i. Assignment records
j. The effective period of attending course
k. Country
l. Age
m. Gender
n. Most viewed pages
o. Operating system
p. Browser
q. Other _____? ”

APPENDIX D: SURVEYMONKEY'S IRB GUIDELINES

SurveyMonkey and IRB Guidelines

Students are certainly permitted to conduct research via the SurveyMonkey platform, provided that they abide by our Terms of Use. Many students use SurveyMonkey to conduct research for their dissertations or graduate work.

This help article outlines the potential guidelines for using SurveyMonkey as a tool to survey research participants. These are criteria that most university IRB's recommend when using an online survey tool to collect data. It is important to engage your Institutional Review Board to approve.

Obtaining Written Permission to Conduct Research Using SurveyMonkey

We are happy to assist you with getting the approvals you need to perform your student research. Here is a letter on SurveyMonkey letterhead that you can provide to your IRB to evidence permission to use the SurveyMonkey platform to conduct your research: Permission to Conduct Research Using SurveyMonkey (PDF)

Secure Transmission

- It is important to enable SSL encryption. Sensitive data must be protected as it moves along communication pathways between the respondent's computer and SurveyMonkey servers.
- Disable IP address tracking to make the survey anonymous.

Informed Consent

- Include a consent form on the first page of your survey.
- SurveyMonkey records the respondent time stamp. This is important especially for respondents that consented to taking your survey.
- The survey should allow for "no response "or "prefer not to respond" as an option for every survey question. A survey where a respondent cannot proceed without answering the question is in violation of the respondent's right to withhold information.
- At the end of the survey, the respondent should be given an option to withdraw from survey.

Database and Server Security

- SurveyMonkey Privacy Policy
- Security Statement

HIPAA Compliance

If you're a covered entity regulated by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and want to collect protected health information in your health and well-being surveys, please see HIPAA Compliance at SurveyMonkey for more details (" SurveyMonkey and IRB Guidelines (n.d)

")

APPENDIX E: INVITATION TO PARTICIPATE IN SURVEY EMAIL

Hello mate, my name is Houssen Nafed. I am working on my Doctorate in Computer Science with a concentration in big data analytics through Colorado Technical University. I am now in the final stage of completing my dissertation! As part of my dissertation, I am conducting a research study in an area that is important to education technology and Machine learning. In recent years, massive open online courses (MOOCs) has become the norm for higher education institutions to offer online education. It is in some ways replacing the traditional classroom environment. As more educational institutions move towards the MOOC environment, there is a concern among the educational community with dropout rates in this environment. The purpose of my dissertation research is to explore the strategies computer science educators need to use to prepare machine-learning datasets for predicting the dropout rates of students in MOOCs. I am inviting you to participate in a brief survey as part of my dissertation. If you can participate, you may click on the link below to take the online survey. The survey includes two established personality attribute instruments. Participation in the survey is voluntary and your answers will be kept confidential according to the terms in the informed consent which must be acknowledged before taking the survey. Participating in the survey will take approximately 10-15 minutes. The survey responses and identities will be coded to protect your privacy. The survey results will be published in my dissertation, If you are interested in volunteering or have any questions, please contact Houssen Nafed.

Cell phone:

Click the button below to start the survey. Thank you for your participation!

https://www.surveymonkey.com/r/ZX6SZXS

APPENDIX F: PARTICIPANT DEMOGRAPHICS

Participant	Gender	Location	Education	Years of Experience in Computer Science	MOOCs Experience	Machine Learning Experience
1	M	USA	Master	12	Yes	Yes
2	M	USA	Master	10	Yes	Yes
3	M	USA	Master	7	Yes	Yes
4	M	USA	Bachelor	6	Yes	Yes
5	M	USA	Master	5	Yes	Yes
6	F	Turkey	PHD	12	Yes	Yes
7	M	USA	PHD	10	Yes	Yes
8	M	USA	Master	12	Yes	Yes
9	M	USA	PHD	10	Yes	Yes
10	M	USA	PHD	8	Yes	Yes
11	M	USA	Master	15	Yes	Yes
12	M	USA	PHD	9	Yes	Yes
13	M	Libya	PHD	14	Yes	Yes
14	M	USA	Master	8	Yes	Yes
15	M	USA	Master	15	Yes	Yes
16	M	Malaysia	PHD	7	Yes	Yes
17	M	USA	Master	5	Yes	Yes
18	M	USA	PHD	7	Yes	Yes
19	M	Malaysia	Master	9	NO	Yes
20	M	Australia	Master	12	NO	Yes
21	M	Spain	Bachelor	5	NO	Yes
22	M	USA	Bachelor	13	NO	Yes
23	F	USA	Master	6	NO	Yes
24	M	USA	Master	12	NO	NO
25	M	Serbia	Master	6	NO	NO