Exploring Strategies Computer Science Educators need to use to Prepare Machine-Learning Datasets for Predicting MOOCs Dropout Rates

Houssen Himeda Nafed
h.nafed1@my.denver.coloradotech.edu

Mohamed A. Lotfy
mlotfy@regis.edu

Samuel Sambasivam
ssambasivam@coloradotech.edu

Department of Computer Science
Colorado Technical University
Aurora, Colorado 80014, USA

Abstract

Different strategies to prepare machine learning datasets for predicting student dropout rates in massive open online courses (MOOCs) 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 MOOCs. There is a critical need to examine the effectiveness and shortcomings of MOOCs and student retention in these courses. The current study investigated a sample of 25 participants from LinkedIn machine learning groups, who have experience in computer science, machine learning, and MOOCs. The data analysis resulted in the emergence of the following three major themes, the predictive models or algorithms used to predict dropout rates, the elements of the MOOCs experience, and the data elements needed in the machine-learning datasets.

Keywords: Machine learning, datasets, Massive Open Online Courses, Students dropout rates, computer science Educators.

1. INTRODUCTION

Computer science and information technology online education is the fastest growing form in the field of technology education since 1999 (Nicholson, 2007). Recently, there has been an evolution in e-learning that led to the emergence of a modern educational phenomenon the use of massive open online courses (MOOCs) (Jordan, 2014). MOOCs are a relatively new educational technology with a history dating back to 2008 (Anderson, 2013; Fini, 2009). The strategic objective of MOOCs is to open up education to the public (Zheng, Chen, & Burgos, 2018). MOOCs started in 2008 by George Siemens and David Cormier at the University of Manitoba, Canada (Cormier, 2010). 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).

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, White, Abdullah, & Davis, 2017). Research showed that by enrolling students from around the world, participation in MOOCs is much lower than actual classroom students (Maitland & Obeysekare, 2015).

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 student dropout rates (Hone & El Said, 2016). Student's skills and the quality of the MOOCs 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 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). 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).

There is concern regarding MOOCs student dropout rates where learners do not finish their courses (Zheng, Rosson, Shih, & Carroll, 2015). Researchers are 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 the assignment and giving feedback (Gütl, Rizzardini, Chang, & Morales, 2014).

MOOC platforms can generate different types of user data with can be given to machine-learning predictive models to help 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 student dropout rates in MOOCs as well as identifying and classifying students regarding withdrawals rates, continuation rates, and incompletions (Zheng et al., 2015). 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.

To fill the gap in the literature regarding student drop out in MOOCs, the current research sought answers to eleven related MOOCs questions from 25 participants. The participants have different experiences in computer science, machine learning, and MOOCs. Answers to these questions can help computer science educators to explore the strategies need to use to prepare machine-learning datasets for predicting MOOCs student dropout rates.

2. LITERATURE REVIEW

The literature review used current scholarly and practitioner to identify 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 MOOCs, analyzing big data to predict student dropouts in MOOCs, and the elements of data used by algorithms to predict drop-out in MOOC. The strategies MOOC computer science educators experienced under each of these eight categories formed the contextual framework for the literature review. Several components /categories of the strategies used to prepare machine learning datasets for predicting dropouts faced by computer science educators in MOOCs were investigated.

MOOCs has attracted researchers and opened a discussion on the impact of this educational phenomenon and the problems online education faces (Alumu & Thiagarajan, 2016). Guo and 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).

Recent studies indicated the need to study student motivation in MOOCs. Balakrishnan and Coetzee (2013) exploratory study focused on the behavior of students who dropped out of the university. 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 rates is lower than classroom-based courses when classroom interactions between students are infrequent (Sunar et al., 2017).

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). Predictive models play an effective role and have been used recently to support getting an appropriate decision for the future (Libbrecht & Noble, 2015). Data analysis plays an essential role in many companies where data can be analyzed using machine learning and big data algorithms on distributed computing platforms (Fisher, DeLine, Czerwinski, & Drucker, 2012).

Recent studies have increasingly used predictive models to understand the pattern and type of data in analyzing learning in MOOCs (Khalil, 2018). Hmedna et al., (2017) 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 have a broad impact on learning across educational platforms (Joseph, 2017). Understanding students' interaction across educational platforms helps to characterize student learning patterns that can help reduce dropout rates and require less teacher intervention (Ramesh et al., 2014).

To address MOOCs student dropout rates, there is a need to develop machine-learning algorithms that can predict student dropout using a thorough understanding of the types of data and algorithms that can accurately predict students who might dropout (Xing & Du, 2018). He, Bailey, Rubinstein, and Zahang (2015) proposed a machine learning framework using support vector machine (SVM) algorithm to predict student dropout rates in MOOCs from clickstream data. Kloft et al. (2014) indicated that the support vector machine algorithm helped diagnose the problem of MOOCs dropout rates. 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). Predictive analytics are interpreted by using the Pipelines Model as a design that helps researchers in provide the most accurate interpretation of the dropout rates (Nagrecha, Dillon, & Chawla, 2017). The Pipelines Model approach in MOOC dropout prediction helps to get a simple idea on past student behavior to predict future results (Nagrecha et al., 2017).

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, which must be analyzed and processed, introduces many challenges and opportunities for organizations to extract valuable information (Srinivasa & Bhatnagar, 2012). Analysis of 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). 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).

The different structure of classes on MOOC platforms presents different types of data, such as online learning behavior, discussion postings, and videos watching (Wang & Baker, 2015). These types of MOOC courses may lead to student dropout. The collected MOOC data contains information related to students' participation across the MOOC courses, which helps researchers explore students' performance (Abubakar & Ahmad, 2017). 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 machine-learning predictive models and algorithms to predict the student dropout potential (Márquez et al., 2013). Predicting student dropout in MOOCs by using a different dataset, which was used previously and classified whether the elements of data indicates drop out or not, can help 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 machine learning algorithms was mainly based on increasing the elements of data in order to get accurate predictions compared to other algorithms.

3. METHODOLOGY

A qualitative exploratory research was used for this study.

3.1 Study population

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 professionals were recruited from LinkedIn groups 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.. The size of the LinkedIn groups, which formed the population for this study, was about 94 thousand machine-learning professionals. The selection criteria for participants in the research study required that respondents had at least one year of experience in computer science, one year in machine-learning, and one year in MOOCs. Twenty-Five computer science professionals within machine-learning responded to the study questionnaire.

3.2 Research Procedure

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 them. Twenty-five professionals consented to participate in the study. All the participants that consented had to click on the questionnaire link, which took them to SurveyMonkey to participate in this study.

3.3 Instrumentation

The questionnaire instrument used (see Appendix A) explored the strategies computer science educators need to use to prepare machine-learning datasets for predicting student dropout rates in MOOCs (Dörnyei & Taguchi, 2009). The researcher also met three participants who met the selection criteria and revised the questionnaire questions before starting to collect data. 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.

3.4 Validity

Validity is one of the most important strengths of qualitative research. It ensures the accuracy of the investigation results from the perspective of the researcher (Leung, 2015). The questionnaire instrument was verified using a pilot study to identify any problems or flaws in the measuring instrument. 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 data collection. The pilot study results ensured that the questionnaire questions were applicable to ensure reliability and validity of the research results (Srinivasan & Lohith, 2017). Member checking was conducted 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 to ensure the most reliable and accurate account of what has transpired.

4. RESULTS

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. Appendix C provides the demographics of the respondents. 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.

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 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 MOOCs student dropout rates, (b) MOOCs experiences, and (c) machine-learning datasets for predicting the MOOCs student dropout rates. The MOOCs experience was further broken down into sub-themes, namely course design, course content, and instructor feedback to students in MOOCs.

The NVivo software was able to capture these words and describe them in a "word cloud." The

more frequent the participant used the word, the bolder, and bigger, the word appears in the word cloud. Figure 1 illustrates a word frequency cloud, which describes the frequency of words that appeared in the themes from the data analysis process.

Figure 1.Word frequency cloud

Appendix B displays the three principles and the themes supported by quotes from the participants as identified in the data analysis.

5. DISCUSSION

The problem addressed in this study was to identify the strategies computer science educators need to use to prepare machine-learning datasets for predicting MOOCs student dropout rates. The purpose of this qualitative exploratory study was to identify the predictive models or algorithms used to predict MOOCS student dropout rates, the MOOC experiences, and the machine-learning datasets for predicting MOOCs student dropout probability. 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 MOOCs student dropout rates (Xing et al., 2016).

The research study was subject to several limitations. 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 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.

Major Theme 1: Algorithms or predictive models for predicting MOOCs student dropout rates.

There are many algorithms and predictive models that help to reduce MOOCs student dropout. Umer, Susnjak, Mathrani, and Suriadi (2017) used different machine-learning algorithms in their experiments to identify which type of algorithms outperformed the others. 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 literature review 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, et al., (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 MOOCs student dropout rates. 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

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 MOOCs experience was based on several factors that help to improve the MOOC platforms 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 MOOCs (Kolowich, 2013).

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 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 MOOCs student dropout rates. 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 and 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 MOOCs Student dropout rates

Regarding the performance of the algorithms and predictive models depend on increasing the data elements in the datasets to get accurate results for predicting MOOCs student dropout rates. With the diversity of machine learning algorithms and techniques, such as deep learning and other methods, there is a need to identify the required datasets in MOOCs (Hernández, Herrera, Tomás, Tomás &, Navarro,2019).

The participants responses regarding this theme were 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 responses indicated the importance of these data elements in the datasets collected. In addition to these data elements, the participants suggested focusing on other types of data that would help to predict MOOCs student dropout rates 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. 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). Regarding the data elements that should be part of the datasets for predicting student dropout rates, 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 students behavior data 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 MOOCS student dropout rates, 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 courses and course registration records, 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 rates among students. This agrees with the 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 students dropping out. 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 MOOCs student dropout rates. 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 MOOCS student dropout rates.

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.

6. CONCLUSIONS

This qualitative exploratory study was designed to explore the strategies computer sciences educators need to use to prepare machine learning datasets for predicting MOOCs student dropout rates. This study explored the strategies of participants who have experience in machine learning, computer science, and MOOCs. Three principal themes were identified in this study: algorithms or predictive models to predict dropout rates, the data elements in the datasets, and the MOOC experience.

The findings of the study confirmed that the logistic regression algorithm is the best algorithm for predicting MOOCs student dropout rates. 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 MOOCs student dropout rates.

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.

MOOCs Computer science educators and scholars need to continue studying algorithms or predictive models for predicting MOOCs student dropout rates 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 MOOCs student dropout rates 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 dropouts in MOOCs. 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.

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.

7. 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).*

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

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

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.

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.

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

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

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.

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.

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

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. Routledge.

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

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 Learning, 10*(5).

Guo, P. J., & Reinecke, K. (2014). *Demographic differences in how students navigate through MOOCs.* Paper presented at the Proceedings of the first ACM conference on Learning@ scale conference.

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.

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*.

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*.

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.

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.

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

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*.

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.

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

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, Stiehler, Zheng, & (2014). 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).).

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

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

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

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. *Journal of family medicine and primary care*, *4*(3), 324.

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

Li, W., Gao, M., Li, H., Xiong, Q., Wen, J., & Wu, Z. (2016). *Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning.* Paper presented at the Neural Networks (IJCNN), 2016 International Joint Conference on.

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*.

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.

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.

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

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.

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

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.

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.*

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*.

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

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.

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.

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.

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

Srinivasa, S., & Bhatnagar, V. (2012). Big data analytics. Paper presented at *the Proceedings of the First International Con*

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

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., 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., 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].

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

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.

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, Rosson, M., Shih, P., & Carroll, J. (2015). *Understanding student motivation, behaviors and perceptions in MOOCs.* Paper presented at the Proceedings of the 18th ACM conference on computer supported cooperative work & social computing.

Appendix A. Survey Instrument

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:

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 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 B. Aggregated Themes

Algorithms or predictive models		**MOOCs experience**		**Datasets Content**	
Themes	Frequency of the algorithms throughout all Responses	Themes	Frequency of the MOOCs experience throughout all responses	Themes	Frequency of the MOOCs experience throughout all responses
Logistic regression	29	Course design	11	Online learning behavior	18
Decision Trees	21	Course content	9	Student behavior	17
Deep neural networks	18	Instructor Feedback to students	9	Assignment records	12
Support Vector Machine	11	Current problems	1	Age	12
K-means	10	Determine success or failure students	1	Graded activities within courses	11
Natural Language Processing Techniques	7	Extra references	1	Forum posts and discussions	11
Recurrent Neural Networks	6	Areas a student struggling with	1	The effective period of attending the course	10
Hidden Markov Models	6	MOOC synchronous	1	Gender	10
Bayesian Networks	5	Improve productive models	1	Stream server logs	7
Survival Analysis	3	Registered courses	1	Country	7
Temporal Difference	3	Current education system	1	Most viewed pages	7
Adversarial Networks	3	Challenging problem	1	Browsers	5
Random forests	2	Content delivery	1	Operating system	4
Supported Semi-supervised	2	Real-world problem	1	Other	2

Appendix C: 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