Review Article

Analysis of Mooc Data With Educational Data Mining: Systematic Literature Review

Rukiye Orman[1a], Nergiz Ercil Cağıltay[1b], Hasan Cakır[1c]

[1]Ankara Yıldırım Beyazıt University, Ankara, Turkiye
[2]Cankaya University, Ankara, Turkiye
[3]Gazi University, Ankara, Türkiye

rukiyeorman@aybu.edu.tr

DOI : 10.31202/ecjse.1581942
Received: 08.11.2024 Accepted: 07.04.2025
How to cite this article:
Rukiye Orman, Nergiz Ercil Cağıltay, Hasan Cakır, " Analysis of Mooc Data With Educational Data Mining: Systematic Literature Review",
El-Cezeri Journal of Science and Engineering, Vol: 12, Iss: 2, (2025), pp.(191-204).
ORCID: [a]00000-0003-1385-0939; [b]00000-0003-0875-9276; [c] 00000-0002-4499-9712.

Abstract : Participants' performance is one of the critical factors for the success of the platforms. There is a lot of data in Massive Open Online Course platforms that are free and open to everyone, and due to this large amount of educational data, it is difficult to make accurate predictions and inferences. The primary purpose of this research is to conduct a literature review to discover the existing Educational Data Mining methods and techniques used to analyze Massive Open Online Course data. For this purpose, the focus is on the source from which the data is collected, which Educational Data Mining methods and techniques are used, and which tools are used in the analysis to compare different approaches. A total of 32 articles published between 2013-2024 were included in the scope of the study. According to the findings, there are many algorithms used for Educational Data Mining methods and techniques in the analysis of Massive Open Online Course data. The most preferred algorithm in the studies is "K-Means", followed by "Support Vector Machines", "Decision Trees" and "Random Forest". Coursera and Edx are among the platforms used and preferred worldwide. It is anticipated that making the data available on these platforms public will contribute to further research and guide studies in the education field. Privacy and ethics also come to the fore within the scope of open data publication. In this context, developing some standards and new approaches to share data with researchers in a standard form that does not include privacy violations will significantly contribute to studies conducted in this field.

Keywords : Educational Data Mining (EDM), Machine Learning Algorithms, Massive Open Online Courses (MOOCs).

1 Introduction

Information technologies impact every field, especially education, health, and banking. While economic, political, and health-related change factors bring about new problems, digital technologies and scientific changes help develop new ways to solve these problems. At this point, performing and concretizing data-based operations is essential to better understand the issues experienced and develop solutions. Massive Open Online Courses (MOOCs), which open a new page in education and enable the revision of open and distance education, have opened the doors to taking online courses from world-famous universities. As in university education, opportunities such as badges and certificates are also offered to those who complete the course by accessing all kinds of course resources, participating in exams through the portal, asking questions, receiving answers, and submitting homework. All behaviors of participants registered in these environments on the system are recorded, and data about the students increases daily. Long-term daily data can be used for student and course evaluation [1]. Providing training online and recording detailed data about students' behavior during the training process by the systems significantly contributes to a better understanding of these learning processes. MOOCs and micro-qualifications, which contribute to the digital transformation of education, redefine society's perspective on learning and the roles of institutions/organizations that provide education. The transfer of learning from school desks to lifelong learning brings a new system change in education, as well as time and space independence. This change aims to keep students up to date with technological and economic developments. Micro certificates come to the fore in determining the framework and validity of courses, and valid certificates have an essential place in the sector [2]. While MOOCs were offered free of charge and open to everyone in the early years, later on, as MOOC platforms became independent education companies, there was a need for paid courses to finance the courses and ensure the sustainability of the platforms. For this reason, in addition to standard paid courses, the production of new content formats that provide financial resources, such as micro certificates and corporate training, has come to the fore. Universities prefer MOOCs because they provide instant data on students' participation in online courses and enable developers to stay up-to-date by developing their courses in line with this data [3]. At the same time, the data obtained from MOOC platforms contributes greatly to online learning research [4].

MOOCs, which contribute to lifelong learning, have attracted attention from many different segments, especially universities, due to the opportunities they offer, they also have disadvantages such as high dropout rates despite the high number of course enrollments, accreditation, high costs of preparation, technical problems, motivation of participants, and measurement and evaluation [5]–[7]. This situation causes the literature's perspective on MOOCs to be questioned again. Despite the high hardware, infrastructure, labor, and time costs, the number of participants who complete MOOCs and receive certificates is very low. Although thousands of participants enroll, the completion rate of most courses is below 13% [8]. In fact, the number of students who continue their courses after the first registration may be less than half [9], [10]. In fact, dropout rates have been a problem in online education even before the emergence of MOOC platforms. However, these two problems differ from each other. When a participant does not complete a course in online education, their self-confidence decreases, and they are discouraged from participating in different online courses. However, the dropout rate in MOOCs prepared with high costs means ineffective courses and cost loss for institutions. For this reason, research is being conducted to increase course completion rates and determine the reasons for dropout.

Studies conducted with data mining (DM) in educational environments are carried out using data collected from traditional classroom environments or online educational environments. [11]. It is more challenging to achieve learning outcomes in face-to-face education environments than in online education environments. For this reason, DM applications are carried out in a more limited way in face-to-face education environments. Educational Data Mining (EDM) applications, which have gained increasing momentum in recent years, mostly use data obtained from online education environments. Models are developed to understand the learning process by applying DM algorithms. These studies are essential in detecting students' interactions and mobility in online environments, modeling student profiles, and predicting their academic success [12]–[14]. EDM is a research area that focuses on applying DM, machine learning, and statistical methods to detect patterns in large-scale educational data. EDM utilizes e-learning platforms such as LMS, Intelligent Tutoring Systems (ITS), and, in recent years, Massive Open Online Courses (MOOCs) to obtain rich and versatile information from student learning interactions in educational environments [11], [15], [16]. These platforms record when and how often students access learning material, whether the answer to an exercise is correct, and how much time they spend reading a text or watching a video. With this recorded information, student performance can be determined, student profiles can be extracted, recommendations can be created, adaptive systems can be developed, and it can be analyzed to address different educational issues such as automatic grading of students' homework. Different EDM methods and techniques have been used to analyze this data. Thus, EDM has significantly influenced recent developments in education and has provided new opportunities for technologically developed learning systems according to the needs of students [17].

All participants' behaviors registered to MOOC are recorded on the system [18]. The instructor using this system can prepare and upload his content according to the system. When the data on these systems is analyzed using EDM methods and techniques, it guides educators and administrators in solving problems by creating participant profiles, personalizing the environment according to the participant, and improving the quality of the educational environment.

When the studies conducted with EDM in the literature are examined, it is seen that studies such as modeling participant/student performances and behaviors, examining participant/student academic success and attendance status, grouping according to participant/student characteristics, evaluation, feedback, pedagogical support, grouping according to participant/student characteristics are carried out [11], [12], [14], [15], [19]–[21]. Studies conducted in recent years focus on different sub-fields of EDMs. New ones were added to the eleven classifications made by Romero and Ventura in the field of EDM in 2010 [11]. With the added classifications, thirteen classification areas have emerged. Studies are being conducted on thirteen different sub-areas, including predicting student performance, detecting undesirable student behaviors, profiling and grouping students, social network analysis, providing reports, creating alerts for stakeholders, planning, and programming, creating course software, developing concept maps, creating recommendations, adaptive systems, evaluation, and scientific research [22].

In the literature, there are generally studies focusing on a single topic or area related to EDM in studies conducted with systematic literature reviews. These are studies using text mining techniques on student-MOOC interactions [23], using predictive video analytics [24], analyzing data obtained from environments such as Facebook and [25], identifying students who dropped out of school and students at risk [26], and measuring self-regulated learning strategies for students [27]. A systematic review was conducted. Studies in the literature on Educational Data Mining (EDM) often focus on a single subfield or a specific data mining method. For example, most studies focus on specific topics, such as student performance prediction or behavior detection, while a systematic review of a wide range of EDM applications is limited.

This study aims to fill an essential gap in the literature by providing a systematic perspective by focusing on the analysis of Massive Open Online Courses (MOOC) data using Educational Data Mining (EDM) methods and techniques. In particular, unlike the studies focusing on a single EDM field or method in existing research on the analysis of MOOC data, this study provides a broad review covering different subfields of EDM and the methods used. In addition, it offers practical information on the data sources, tools, and techniques used to analyze MOOC data to study large data sets on MOOC platforms. It provides a guide for future research in this field. This study's following research questions were determined to investigate the EDM methods and techniques applied to profiling and grouping students, predicting student performance, identifying student behaviors, and evaluating classifications.

1) What are the studies classified according to Educational Data Mining subfields?
2) From what sources was MOOC data collected?
3) What are the Educational Data Mining methods and techniques used in analyzing data?
4) What are the tools used for analysis?

2 Materials And Methods

It is a systematic literature review investigating current approaches to classifying EDM methods and techniques for analyzing MOOC platform data. The Systematic Literature Review approach is used to obtain comprehensive results for analyzing and discussing different published articles [28]. It is an open and repeatable approach based on a search strategy structured according to the publications' predetermined inclusion and exclusion criteria. The results obtained from the articles are analyzed within the framework of the determined research questions. [29], [30]. The search strategy, inclusion and exclusion criteria, and the analyses of the obtained data are explained in the subsections below.

2.1 Design of the Study

The most widely used Web of Science and Education Resources Information Center (EBSCO ERIC) databases were searched to conduct a systematic review. These databases include articles that meet the quality standards of journals of famous publishers such as Elsevier, Springer, IEEE, ACM, etc. Accreditation institutions also prefer them because they meet quality standards [31].

Figure 1 shows the search and selection process used for systematic literature review. Our study used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Boers, 2018; PRISMA, 2020) flow chart, which is the most preferred method for systematic reviews and meta-analyses. A flow chart was used. This flow chart contributed to clearly reporting the work steps and the derivation of meaningful syntheses and conclusions.

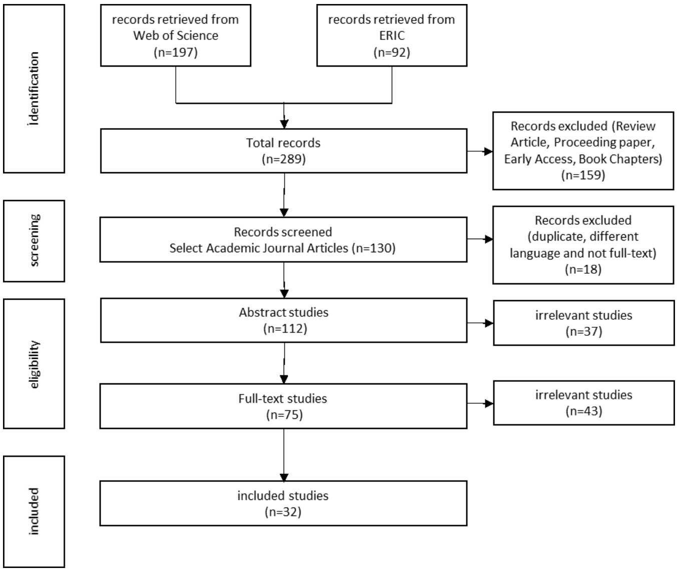
identification
screening
eligibility
included
records retrieved from Web of Science (n=197)
records retrieved from ERIC (n=92)
Total records (n=289)
Records excluded (Review Article, Proceeding paper, Early Access, Book Chapters) (n=159)
Records screened Select Academic Journal Articles (n=130)
Records excluded (duplicate, different language and not full-text) (n=18)
Abstract studies (n=112)
irrelevant studies (n=37)
Full-text studies (n=75)
irrelevant studies (n=43)
included studies (n=32)

Figure 1: Systematic Search and Selection Process

2.1.1 Search String

First, the Web of Science (WOS) and EBSCO Eric databases followed an automatic search strategy. The search string was built around predefined keywords containing wildcards. The words in the search string were scanned in the title, abstract, and keywords.

Search string: TS= MOOC* AND EDU* AND ("data mining" OR datamining OR "data-mining")

Indices: CPCI-S, SCI-EXPANDED, SSCI, CPCI-SSH, ESCI, A&.HCI, BKCI-SSH.

2.1.2 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria created to access relevant studies and prevent bias in selecting studies are in the table below. Studies were selected by considering general and specific criteria such as the type of study, year of publication, language, and whether it has a complete text and research method.

Table 1: Inclusion and Exclusion Criteria of the Study

Inclusion criteria:	Exclusion criteria:
1. Journal article	1. Literature review, book, book chapter, proceedings
2. Regarding the use of EDM methods in MOOCs	2. Studies outside the field of education
3. Published between 2013 and 2023	3. If the study is not relevant to the research questions
4. Published in English	
5. The full text of the article is available	

2.1.3 Search and Select

For the research to be suitable, it is essential to structure reliable, well-planned data sources and a search and selection strategy. Therefore, a comprehensive and systematic search was conducted to get a proper idea of its usage in the analysis of MOOCs. As a result of the database search, a total of 289 studies, 197 in WOS and 92 in ERIC were published between 2013-2024. A follow-up search was conducted in the databases to check if any new studies were published, and studies published until the first half of 2024 were included. After excluding the remaining articles, such as books, conference proceedings, systematic reviews, and duplicate studies, 130 articles were selected. Since WOS and ERIC index articles meet essential quality factors, this number of articles is anticipated to be suitable for a research pool. Due to the systematic review process, WOS and ERIC search results were included in this study's collection. At this stage, a direct abstract reading activity was performed to create the initial pool studies for filtering.

2.1.4 Summary Reading Activity

After the search and selection process, an abstract reading activity was completed for the remaining 130 articles. The following questions were prepared for the articles that would be excluded from the scope of the study. Exclusion of irrelevant articles is performed based on the following questions:

- Are there any areas of EDM implementation on MOOC platforms to be measured or reviewed?
- Are there any EDM methods used to measure or review?
- If there is a clear answer to either of these two questions, the article is kept; otherwise, the following question applies:
- Are there any EDM evaluation methods and techniques?
- As a final check, the article is excluded if there is no clear answer to this question.

After the summary reading activity, 37 articles that were not related to the field of education, did not contain MOOC data, did not conduct EDM analysis on MOOC data, and did not use any EDM methods and techniques were excluded. At the end of the summary reading activity, 75 articles were included in the study. Since EDM applications are a new application area in education, the number of articles is lower compared to other fields.

2.1.5 Full Reading and Filtering

Each step in this study was systematically addressed, and the first author recorded the process. The process was then comprehensively reviewed by the co-authors. The criteria were considered at each stage of the process. The selected articles (n=32) were read in full text, which required more intensive work than the other steps. In the full reading stage, the relevance of each study to the research questions and the EDM methods and techniques used in the studies was reviewed. Inclusion and exclusion criteria were also applied at this stage. All criteria ensured that the study was conducted according to its purpose.

2.1.6 Analysis of Studies

The studies were classified by focusing on the sub-areas of a) creating and grouping students' profiles, b) estimating student performance, c) determining student behaviors, and d) evaluating within the scope of applying EDM methods to MOOC data. The first research question was determined to identify the data sources in the reviewed articles. Then, the second research question was written to reveal the EDM methods and techniques used in each sub-area. Finally, analyses were conducted within the scope of the third research question in order to identify the tools used in the analyses.

3 Results and Discussion

Our study focused on four sub-areas of EVM in applying EVM methods and techniques to MOOC data. For this purpose, the findings obtained from the analyses were included, including articles published by year, a review of EVM sub-areas, a data source review, an evaluation of EVM methods and techniques, and tools used to compare different approaches. Four research questions were created for this purpose. The findings obtained according to each research question were discussed in this section, and the results were summarized for stakeholders who could benefit from the study. The analyzed articles are in the table below based on their published years. When the time trends of the publications are examined, it is seen that there is a regular increase from 2014 to 2022, and the publications increase and reach their peak in 2022. Although the number of studies is low in the first years, it increases in the following years. 2022 is the year in which the most studies were conducted, with 20 articles. The year most studies were conducted was 2019, with 17 articles. No articles were published on the subject in 2013. This is because MOOCs are new, and the platforms on which they are published have become popular in the years following. In addition, the fact that the EDM field is new is another effective factor. It is seen that the studies conducted using both EDM and MOOC started to increase after 2017.

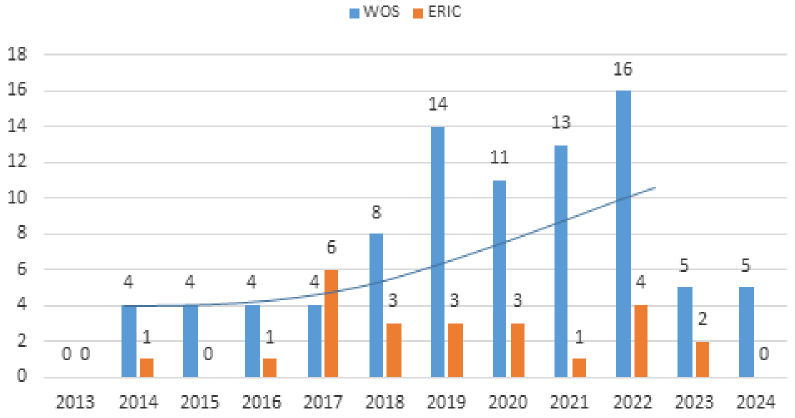

Figure 2: Distribution of publications by year between 2013-2024

When the first pool was created, many studies were conducted. However, when the studies conducted within the scope of education were focused on, the number of studies considered for analysis (n=32) decreased. However, reasons such as the pandemic affected the world in 2020 and later increased the interest in online education. Afterward, online environments gained popularity. Therefore, a large amount of data has been generated in environments such as MOOCs, and a need has arisen for new studies to be conducted in order to interpret this increasing data. It is seen that the studies have gained momentum with the increasing interest in areas such as data science, analytics, artificial intelligence, and machine learning. The number of studies is predicted to increase in the following years. Studies on the application of EVM methods and techniques to MOOC data provide important insights into various educational scenarios. However, the findings of these studies present both opportunities and important limitations in the digital transformation of education systems. Below, the findings from the literature are discussed in detail and in a controversial manner.

3.1 Review of Studies According to EDM Sub-Fields

The articles are examined within the scope of the four areas of EDM; 13 studies were conducted in predicting student performance, 9 in creating and grouping students' profiles, 9 in determining student behaviors, and 1 in evaluation. The table below includes studies conducted within the scope of the four areas of EDM.

Table 2: EDM Sub-domains and Publications

EDM Subfields	Ref	Number	%
Predicting student performance	(Ahmed, 2024; Alghamdi, 2024; Ani &. Khor, 2023; Lemay & Doleck, 2019, 2020; Liang et al., 2014; Monllaó Olivé et al., 2019; Onan, 2020; Pillutla et al., 2020; Swai & Mangowi, 2022; Tomkins & Getoor, 2019; Wan et al., 2019; Youssef et al., 2019)	13	41
Creating profiles and grouping students	(Cohen & Holstein, 2018; Dyulicheva, 2021; Lee, 2018; Li et al., 2022; Nilashi et al., 2022; Rizvi et al., 2019; Saqr et al., 2022; Tang et al., 2018; van den Beemt et al., 2018)	9	28
Determination of student behavior	(Assami et al., 2022; Benoit et al., 2024; Brinton et al., 2014; Geigle & Zhai, 2017; Gupta, 2019; Ruipérez-Valiente et al., 2021; Xu et al., 2022; Yang et al., 2016; Zhong et al., 2017)	9	28
Evaluation	(Nie et al., 2020)	1	3

Predicting student performance in this sub-field of EDM, the main goal is to predict the student's future performance based on their past activities. In the study conducted by Ani and Khor in 2023, machine learning models were used to predict student performance in MOOCs with the interaction of demographic information, academic background, and course materials. Data were tested in eight supervised machine-learning algorithms (Linear Regression, Logistic Regression, Random Forest, K-nearest neighbors, Support Vector Regression, Linear Discriminant Analysis, Deep Learning, and Decision Trees). It was seen that the data collected from MOOC platforms can be used to predict student performance with over 77% accuracy, and the Random Forest classifier model offers higher accuracy than others in predicting student performance [32]. Students' video interactions, exams, forums, navigation, etc., developed a prediction model to predict students who are at risk of dropping out of the course, those who are likely to fail, and successful students using the data. It was seen that the Random Forest (RF) model provided more accurate predictions than other models, with an average accuracy of 98.6% [33]. In support of this study, a study using the interaction data of school teachers [34] and a model were developed in studies using student, MOOC, and learning activity characteristics [35]. In these models, it was concluded that the RF algorithm produced 96.4% and 95% accurate predictions compared to other algorithms.

In the study conducted in 2024, the K- Means algorithm from the clustering method was used to predict the final result of a student based on gender, region, entryresult, previousattemptnumber, studiedcredits, and disability using various prediction models, and the best prediction model, SVM (96% accuracy), was selected [36]. This study used the clustering method and prediction models to help predict performance. In the other study, classification methods were used to evaluate students' preferences accurately by taking into account the comments on the platform, and the method with the best performance, Support Vector Regression (SVR), produced better results than other algorithms with 80% accuracy [37]. Some studies use the Decision Tree algorithm to find factors that may affect students' participation and performance [38]–[42]. Studies also classify student interactions to determine which students work together and the type of collaboration used [40], [43].

Consequently, predicting student performance involves allowing early prediction and identification of slow student progress and implementing teaching methodologies based on predicted performance. The results of predictive models are essential because they allow instructors to identify low-risk students early and intervene in at-risk students promptly. Timely intervention can enhance the student's learning experience and increase the effectiveness and engagement of the learning process. It can also include providing feedback to the student focused on their progress. Predictive methods can help make accurate recommendations and suggest learning solutions based on many factors. When the above studies are examined, it is seen that DM has a high potential to be a valuable tool for discovering how people learn, predicting learning, and understanding actual learning behavior.

Students' profiles are created and grouped based on daily course data, video viewing, and exam data. This information is used to group students for various purposes. In a study by Li, Du, and Sun in 2022, students' final grades were investigated by statistical analysis, lag sequence analysis, and DM methods to investigate their learning engagement, time organization, content visit sequences, and activity participation patterns. Data were collected from a finance course on the MOOC platform called XuetangX in 2018 (n=535 students). Three groups of students were identified. These were unsuccessful, satisfactory, and excellent [44]. In another study conducted in 2020, a FutureLearn The Educational Process Mining method was used to analyze MOOC (n = 2,086 students). As a result of the analysis, three groups of students were identified and categorized according to the clickstream data. These are markers, partial markers, and non-markers [45]. Another study conducted in 2022 examined the strategies used by teacher candidates in an in-service teacher training MOOC while teaching a programming course. Unsupervised clustering and process mining were applied to analyze MOOC daily data (n=8,547). As a result of the analysis, three groups with different strategies were identified. These are efficient clickers and intermediates [46]. In another study conducted in 2018, clickstream, forum, and exam score data of students (n=607) enrolled in a MOOC on the Canvas Network platform were analyzed. Three groups with different longitudinal participation trajectories were identified using the clustering method from EDM techniques. These are cluster A-infrequent participants, cluster B-gradual dropouts, and cluster C-constant participants [47].

In the above studies, three participant groups were determined using the clustering method of EDM. Data were collected from a single MOOC located on different MOOC platforms in the studies. Only one of the studies included teacher data instead of student data. When teachers were evaluated in terms of learning strategies, it was seen that they were similar to other student groups. There were differences between all three groups determined in the studies in terms of both learning participation and learning patterns. In two of the four studies [45], [47] clickstream data was used rather than daily data. It is seen that the stored daily data does not have any meaning on its own. This is because clicking on the data does not necessarily mean a behavioral interaction, leaving aside cognitive processing or learning. Therefore, it becomes clear that using other data and clickstream data will be important in deriving meaningful results.

In a 2019 study, Lee applied self-organizing map (SOM) and hierarchical clustering algorithms to the log files of MIT's summer 2014 Newtonian mechanics (8.MReVx) physics MOOC to investigate how students solved their weekly homework and exam problems in order to identify clusters of students who exhibited similar problem-solving patterns. Data were collected from the Edx platform (n= 4,337 students). As a result of the hierarchical clustering analysis performed with SOM, four student groups were determined. These are cluster1: those who did not receive a certificate, cluster2: those who received and did not receive a certificate, cluster3: those who struggle, and cluster4: early starters [48]. It was seen that the findings obtained from this study cannot be generalized to MOOCs that emphasize different types of learning activities and pedagogies (social studies, literature, etc.) where results cannot be obtained with the answer to a single question. As the number of clusters in SOM increases, the students assigned to each cluster become more homogeneous regarding problem-solving patterns. Hierarchical clustering without SOM could not determine the cluster of students who earned a course certificate. Combining SOM and hierarchical clustering algorithms allows a more leisurely exploration of complex, multi-dimensional diary data. In a study conducted in 2018, process mining techniques were applied to the video viewing behavior and exam submission process data of 16,224 students in a MOOC on Coursera. As a result of the analysis, four groups were determined. These are samplers, discriminators, initiators, and achievers. Process mining combined with traditional statistics applied from a personal constructivist perspective shows a fruitful approach to investigating learning behavior and learning processes in MOOCs. Data was collected from a MOOC in both studies, and four student groups were formed due to the analyses.

Cohen & Holstein, in 2018, 3,460 data from 5 different MOOCs from the CourseTalk website were analyzed using content analysis and DM and semantic analysis. The population of the analyzed courses included students from various and different countries. The aim was to reveal the features that contribute to the success of xMOOCs in science and management fields according to students' perceptions and to cluster students with similar preferences. In this study, instead of the commonly analyzed MOOC diary data, the focus was on student comments. As a result of the cluster analysis, five student groups were identified. These are social, no specific preference, cognitive, teaching-teacher, and teaching exam [49]. Although five student profiles were created in the study, the student profile is also related to various features such as previous knowledge, learning style, learning rhythm, etc. A study was conducted on positive student comments about successful courses, but it does not include the comments of all students who attended the courses. It seems that it would be essential to consider all student comments on successful and unsuccessful courses in order to provide a general perspective. In another study conducted in 2021, 38 math MOOCs on Udemy and 1,898 students' definitions of math anxiety were analyzed using text mining techniques (VADER sentiment analysis, k- means, BERT), and five groups were identified [50].

Short videos are the most potent learning objects in online courses and are highly preferred by all students. Badges and micro-credentials are also used in MOOCs. Using them together is anticipated to motivate the participants and encourage them to do the activities. Well-managed discussion forums can replace peer support in a physical classroom and give students a sense of participation. Although there is a significant and robust relationship between students' course completion status and their activity in the forums, it has also been observed that the group of students who did not complete the course actively participated in the forums.

In only one of the reviewed studies, six student groups were identified due to the analysis of MOOC data. The study conducted in 2022 classified different types of participant behavior in MOOCs into clusters using the DM methodology and based on video lectures, discussion forums, and assessment activities. Data were obtained from a MOOC on the Udemy platform. Se cluster analysis identified six participant groups [51]. Adaptive for Prediction Neuro-Fuzzy Inference Systems (ANFIS) were used.

As A result, nine studies were examined within the scope of the subfield of creating profiles and grouping students. Four studies determined three student groups, two that determined four student groups, two that determined five student groups, and 1 study that determined six student groups. The K-means method from cluster analysis was mainly used in the analyses. Determining student groups with the same characteristics can better adapt the course design to the needs and learning styles of the students. As seen in the studies examined, video lectures, discussion forums, and assessments are the primary learning resources in MOOCs. Therefore, analyzing a participant's activity in these components reflects their behavior in the course.

Detection of student behavior: Studies focusing on detecting student behaviors have faced three subtasks: predicting dropout on MOOC platforms, addressing the problem of students' engagement in their learning, and evaluating social functions. In the study by Geigle and Zhai in 2017, a two-layer hidden Markov model (2L-HMM) was used to discover student behavior patterns with click log data obtained from the MOOC platform. Although this study did not extract dropout behaviors, students' positive behaviors were extracted. It was observed that high-performing students tended to have longer [52]. In a different study similar to this study, which aimed to investigate and understand the learning processes and behaviors of students in MOOCs, EDM methodologies (K- means clustering, support vector machine, artificial neural networks) were used [53]. In support of this study, students' navigation traces were extracted to predict student motivation (Assami et al., 2022), and four supervised machine learning algorithms were used. In the study of dropout behavior in online courses, student behavior, perception [38], clickstream [54]and forum data [55]were widely used.

As a result, it can be concluded that most students can be divided into various groups according to their learning styles. Moreover, learning styles can be easily predicted according to the student's learning behaviors. It is seen that students with different learning styles behave differently in MOOCs. It also means that learning style can be a factor that determines the students' learning behaviors and even measures whether a student is suitable for learning through MOOCs. Understanding the behaviors and characteristics of the participants will allow the courses to be better adapted to the needs of different students and thus maximize the impact of MOOCs in providing lifelong learning on a large scale.

Evaluation: The proliferation of MOOCs emphasizes the need to develop accurate and valid evaluation methods to evaluate the quality and effectiveness of courses. A 2021 study by Nie, Luo & Sun proposed a MOOC evaluation (DME) method that combines the Analytical Hierarchy Process algorithm and text mining to integrate expert opinions, standardized rubrics, and student feedback data into the evaluation process [56]. 30 MOOCs were selected from the Coursera website and evaluated using the DME method, and the results were compared with expert evaluation and student rating scores. The result supports the suitability of the DME method as a low-cost, advanced, and accurate method for MOOC evaluation.

3.2 Data Sources (Platforms)

The studies conducted within the scope of the four areas of EDM, the first of which is the research questions, are examined in terms of data sources and are shown in the table below.

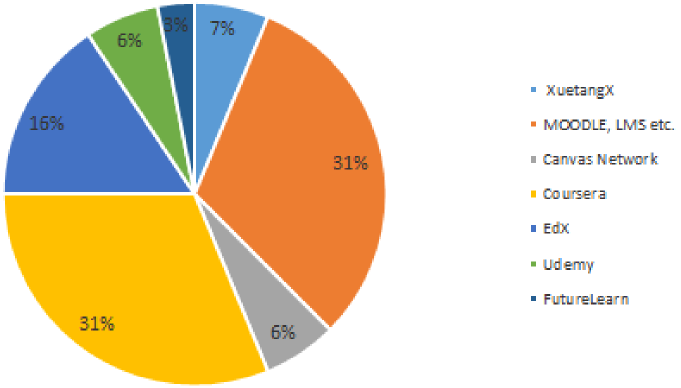

Figure 3: Data Sources

Determining the sources from which data is collected in the studies conducted will help researchers choose appropriate sources and platforms when collecting their data. Data sources in EDM can be online or offline. In the studies examined, data were generally collected from two sources: university courses and MOOC platforms. According to the findings, the most preferred platforms are LMSs such as Moodle (31%). The reason for this is that educational institutions can customize and use such platforms according to their own needs and that higher education institutions are preferred more in studies conducted for educational purposes. Coursera (31%) and Edx (16%) are the other most used platforms. In addition to Coursera and Edx, data was also collected from XuetangX (7%), Canvas Network (6%), Udemy (6%) and FutureLearn (3%) platforms. These platforms are among the platforms used and preferred worldwide. This is because these platforms are provided with MOOC support from universities worldwide, are accessible to a broad audience, and a large amount of data is stored on them. Another reason researchers prefer MOOCs is that they provide educators with a large amount of data that can be used to explain how students interact with the platform based on various factors.

As a result, it is anticipated that making the data available on these platforms public will contribute to further research and guide studies in the education field. Privacy and ethics come to the fore within the scope of open data publication. In this context, developing some standards and new approaches to sharing data with researchers in a standard form that does not include privacy violations will significantly contribute to the studies conducted in this field.

3.3 EDM Methods and Techniques Used

DM is divided into two predictive and descriptive methods. Descriptive methods are divided into classification and regression, predictive methods are divided into clustering and association rules. Algorithms used in the classification method: Decision trees, naive Bayes, k-nearest neighbors, artificial neural networks, support vector machines, time series analysis, and other methods. Linear and logistic regression algorithms are used in the regression method. Algorithms used in descriptive methods are clustering, association analysis, sequential sequence analysis, summarization, descriptive statistics, exception analysis, and other methods.

Massive Open Online Courses (MOOCs) generate a large amount of data that can be used based on various factors to predict and evaluate student performance. Among the machine learning tools used in the existing literature for similar purposes to predict student performance, discriminant analysis, support vector machines, naive Bayes, decision trees, K-nearest neighbors, random forest, linear and logistic regression, bayesian network and community methods [32]. Accuracy, precision, recall, and F-measure are the most commonly used evaluation metrics in MOOCs. Since the prediction models are based on a classifier, many evaluation metrics such as prediction accuracy measurements, confusion matrix, ROC curve, and Area Under the Curve (AUC) are used to measure the prediction quality to evaluate the classification model. Kappa, ROC, AUC, and F evaluation metrics were used in the studies. The primary purpose of presenting the evaluation metrics in the studies is to ensure that the results that may occur by chance are not reported. The table below shows the features, methods, and accuracy rates used in the studies.

Table 3: EDM Methods Used and Accuracy Results

Method	Ref and Year	Attributes	Accuracy
Support Vector Machine (SVM)	(Ahmed, 2024)	Gender, Region, Entrance Result, Number of Previous Attempts, Studied Credits, Disability, Final result	96%
	(Brinton et al., 2014)	Forum activity data	86%
	(Tomkins & Getoor, 2019)	Forum, assignments, quizzes, and exams data	76%
	(Pillutla et al., 2020)	Discussion board data	79%
	(Alghamdi, 2024)	Comments on the platform	80%
Random Forest (RF)	(Youssef et al., 2019)	Video Interaction, Transcript Interaction, Quiz Interaction, Effort, personal information, Performance, Prerequisites, Forum, Navigation, Weekly Final Test, Supplementary Resources	98%
	(Swai & Mangowi, 2022)	Interaction data: FPF (the frequency of participation in the forum), FDS (the frequency of discussion teaching strategies), KPI (the knowledge level related to PI strategy), KTPS (the knowledge level related to TPS strategy), TSK (the general teaching strategy knowledge level of teacher)	96%
	(Assami et al., 2022)	Learner features, MOOC features, and learning activity features)	95%
	(Ani & Khor, 2023)	studentInfo , studentAssessment and studentVle	77%
Decision Tree (DT)	(Gupta, 2019)	Student Behavior (enrolled, viewed, discovered, certified, gender), Student Perception (number of activities, certified or not, active days, videos played, number of sections), and Student records (course ID, user ID, year of birth, gender, class, and forum post)	98%
	(Liang et al., 2014)	activity completion Reported data (online group meeting, question discussion, reference reading, wiki editing, exam taking, assignment uploading, courseware downloading, and watching videos), learning records, and feedback collected from students ' surveys	80%
	(Wan et al., 2019)	learning learning behaviors (total_duration, total_video_duration, total_coursware_access etc.) habits (avg_start_submission_time, time_first_ visit,avg _time_between_problem_submission)	86%
	(Lemay & Doleck, 2020)	Video- Viewing Features	80%
	(Lemay & Doleck, 2019)	Video-Viewing Features (Videos Watched per Week, Avg Frac Spent watching, Total Number of Pauses, Avg Playback Rate, etc.)	70%
Artificial Neural Network (ANN)	(Zhong et al., 2017)	survey data and daily activity data	100%
	(Monllaó Olivé et al., 2019)	student registrations, users, courses	89%
	(Xu et al., 2022)	Clickstream data (Access, Discussion, navigate, page_close, problem, video, wiki)	69%
Logistic Regression (LR)	(Yang et al., 2016)	course content click behavior and course discussion forum data	80%
Naive Bayesian (NB)	(Onan, 2020)	Platform data	79%

When the features used in the studies are examined, the most preferred features are course and forum features, followed by student and exam features. Studies with activity and video diaries follow this. Assignments and word clouds are among the least preferred features in the studies. Analyses are usually performed using data sets belonging to more than one feature in the studies. In cases where high accuracy cannot be achieved using a single feature, analyses are supported using additional features. In addition to these, it is seen that data obtained from survey studies are also used in studies where statistical analysis is performed. There are a limited number of studies on assignments and word clouds. This situation reveals that not enough work has been done on these data sets and that it is essential to include studies in this area.

The clustering method, one of the descriptive models, has been used in studies to create and group students' profiles. Descriptive models allow the identification of patterns in existing data that will guide decision-making instead of prediction. The primary purpose is to find relationships, connections, and behaviors between the data in the data set. The two most commonly used algorithms in the clustering model are K-means and K-medoids.

The reviewed articles use various analysis techniques to analyze MOOC data, including algorithms, evaluations, tools, and statistical methods. Researchers decide which algorithm to use depending on the structure of the data set, the type of problem, and the requirements. Figure 3 shows the methods and techniques used in the studies. The most preferred algorithm in the studies is "K- Means ", followed by "Support Vector Machines", "Decision Trees" and "Random Forest". Then, "Artificial Neural Networks", "Logistic Regression" and " Naive Bayes " is coming. In addition to the Classification and Regression models from predictive models, the "K-means" algorithm used for clustering from descriptive models is included in the studies.

Table 4: EDM Methods Used and Num of Clusters

Method	Ref and Year	Attributes	Number of Clusters
Clustering	(Nilashi et al., 2022)	online reviews and ratings, survey responses	6
	(Ruipérez-Valiente et al., 2021)	Academic engagement (grades and submissions) and behavioral Engagement with the platform (general activity levels, interaction with videos, and Discussion forums).	3
	(Li et al., 2022)	learning engagement, time organization, content Visited sequences, and activity participation patterns	3
	(Cohen & Holstein, 2018)	teaching, social and cognitive, and technological data	5
	(Dyulicheva, 2021)	comments, reviews, social media profiles	5
	(Ahmed, 2024)	Gender, Region, Entrance Result, Number of Previous Attempts, Studied Credits, Disability, Final result	3
	(Saqr et al., 2022)	Daily data ((1) N lessons, (2) N successful lessons, (3) videos, (4) total views, (5) lesson evaluation, (6) interval, (7) duration, and (8) average lesson duration.)	3
	(Tang et al., 2018)	clickstream data, forum data, and Quiz scores	3
	(Benoit et al., 2024)	Activity data, namely, the Exercise ID, the start time and completion time of each exercise for each student, and subscription data, i.e., when the Subscription started and whether they canceled it.	2
	(Lee, 2018)	homework, problems, quizzes, midterm exams, final exam date	4
	(van den Beemt et al., 2018)	video viewing and exam data	4
Process Mining	(Rizvi et al., 2019)	Click behaviors	3

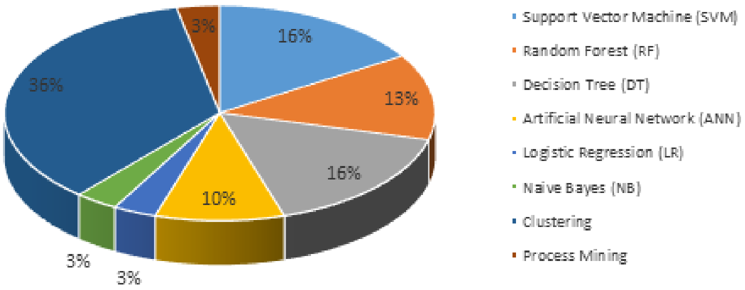

Figure 4: Distribution of Methods Used in Studies

Clustering algorithms have been used to support or direct predictive data. At the same time, statistical analyses used with DM methods are also found in the studies. These algorithms are basic algorithms used by researchers. K-means, Support Vector Machines, Decision Trees, and Random Forests were used the most in the examined studies. This is because more studies have been conducted in creating and grouping students' profiles and predicting student performance/motivation than in other areas. In addition, Support Vector Machines (SVM) and Decision Trees (DT) are widely used algorithms in the prediction model.

In fact, it is not easy to draw a clear conclusion about which algorithm is used the most in the EDM field because this may vary depending on the data sets used and the requirements of specific projects. However, some studies and application examples show that algorithms such as support vector machines and decision trees are used more. These algorithms offer techniques to extract useful information from training data and improve training processes [57]. However, since each algorithm has its advantages, disadvantages, and suitable usage scenarios, which algorithm to use depends on the characteristics and purpose of the data set. These methods and techniques continue to develop today, and their impact on training is expected to increase.

3.4 Tools Used for Analysis

In most DM applications, Weka, Rapid Miner, and Knime tools provide minimum qualifications for analysis. Keel, Orange, Rapid Miner, and Weka are used in text mining analysis to analyze large amounts of data in video analysis. At the same time, researchers prefer Apache Spark and Apache Hadoop [58]. Among these tools, Rapid Miner and Knime provide an interface for data visualization, and Weka offers a command line interface [59], [60]. The Orange tool, written in Python, allows script writing. Similarly, Knime and Rapid Miner also offer the ability to write snippets in Java and Python. The tools mentioned generally work after installation, but the Keel software can work without an installation requirement. [61]–[63]. Table 6 includes the analysis tools used in the studies.

The studies examined show that essential tools such as Python, Weka, Rstudio, and Discovery are used for analysis. Google Collabs has also been used in a few articles published in recent years. In addition, the name of the tool used for analysis is

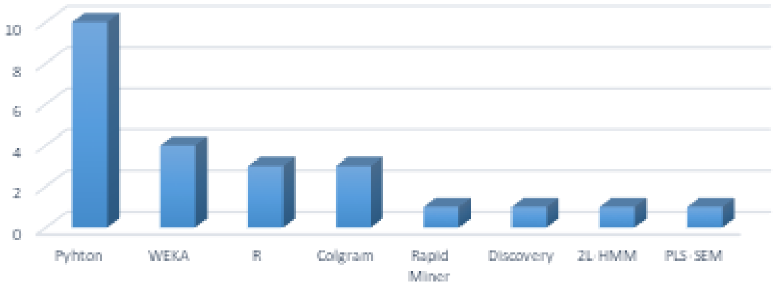

Figure 5: Analysis Tools

not specified in some of the articles examined. When selecting these tools, which have various features that support DM and machine learning processes, it is essential to consider factors such as the type of data to be used, the operating system, the budget, and user requirements.

The findings obtained from the studies can be used in the design of learning systems. These systems can be designed to produce detailed reports for participants, instructors, and administrators. It is envisaged that the reports can be used for participants to see their progress, for instructors to identify at-risk students, to take precautions and update their courses, and for administrators to develop data-based strategies.

4 Strengths and Limitations of the Study

Focusing on studies that apply EDM methods and techniques to MOOC data, the visualization of the findings obtained from data sources, tools used, methods, and techniques reveals the study's strength in providing an easy examination opportunity. Despite this, the study has some limitations. First, the study only examines publications in WOS and ERIC. Examining publications in different databases could have provided more data and awareness about this research. Second, the study only examines academic articles and assumes that they include reliable findings. However, examining various publications, such as announcements, books, book chapters, reports, etc., could have revealed different findings. Third, some sources were excluded because they did not contain sufficient information and did not meet the study's inclusion criteria. The findings of our study provide a general perspective to researchers who will apply EDM methods and techniques to MOOC data, as well as awareness of decision-making processes for educators and administrators.

5 Conclusions and Suggestions

Studies in the literature on Educational Data Mining (EDM) generally focus on a single subfield or a specific data mining method. While most studies focus on specific topics such as student performance prediction or behavior detection, a systematic review of a wide range of EDM applications is limited. Examining MOOC data in a broader context with multiple EDM techniques stands out as a topic that has not yet been fully addressed in the literature and is aimed at filling this gap. This study examines different subfields of EDM (student profiling, performance prediction, etc.) and various data mining methods used in these subfields together. In addition, the sources and tools used in the analysis of MOOC data were evaluated, and suggestions were presented regarding the missing aspects of the literature.

In the findings obtained, the majority of the studies in the literature focus on predicting student performance. For example, algorithms such as Random Forest (RF) and Support Vector Machines (SVM) have been found to be effective in predicting students' failure or success with high accuracy. Such prediction models allow instructors to intervene by identifying students at risk of low success early. In particular, the RF algorithm has attracted attention as one of the most effective prediction models in the literature with an accuracy rate of 98%. Early interventions on student motivation and learning strategies in education can reduce learning losses. Although these results seem positive, the accuracy of these models depends on the homogeneity of the data sets used and the complexity of the algorithms. The same success rate cannot be guaranteed for different demographic groups or students at different educational levels. In addition, incorrect predictions can reduce students' learning motivation and lead to ethical problems. EVM methods are widely used to understand and group students' learning behaviors. Methods such as K-means and Hierarchical Clustering classify student behaviors into three to six different categories. Studies have identified groups of students with different learning strategies. For example, groups such as "Achievers," "Demotivators," and "Insufficient" provide valuable information in understanding student profiles. Analyzing MOOC data in this way enables personalization of course content and increased participation. However, dividing students into specific groups may ignore individual differences and may not adequately reflect the complexity of the learning process. In addition, the behaviors identified are often based on a limited data set, and the generalizability of these findings is limited. High dropout rates in MOOCs (completion rates below 13%) have been frequently emphasized in the literature. Lack of motivation, content difficulties, and time management problems are prominent reasons for students to drop out of their learning processes. Studies analyzing dropout rates aim to develop motivation-enhancing strategies by predicting students' tendency to leave the platform early. For example, data obtained from forum activities have been correlated with students' participation status. However, strategies aimed at reducing dropout rates may not always yield the expected results. For example, motivation-enhancing elements such as awards or badges are not always effective in the long run. In addition, the methods used to understand the reasons for dropout rates do not take into account students' offline factors (e.g. personal life conditions). EVM studies have focused on classification, clustering and regression models. Decision Trees, Support Vector Machines and Random Forest stand out among the most commonly used methods. Algorithms such as Support Vector Machines and Decision Trees provide effective results in terms of both accuracy and speed. These algorithms are seen as valuable tools in dealing with the wide range of MOOC data. However, the effectiveness of these algorithms depends on the quality and size of the dataset used. In addition, in some cases, the features used (e.g. demographic information) are seen to be ethically sensitive. In the literature, studies comparing the results of different algorithms are limited, which makes it difficult to choose the ideal method. Tools such as Python, Weka and RapidMiner are widely used in EDM analyses. Tools such as Python offer analytical flexibility thanks to their extensive libraries. In addition, cloud-based solutions such as Google Colab make the process of working with large data sets easier. Despite such advantages, the experience level of researchers is an important factor in the use of analysis tools. This may limit the reproducibility of analyses for less experienced researchers. The findings suggest that EVM is a powerful tool for analyzing MOOC data. However, given the limitations of these studies, issues such as generalizability of algorithms, ethical concerns, and applicability of personalized learning strategies remain controversial. Future studies should aim to fill these gaps with larger datasets and more comprehensive methods.

In the studies examined, data was generally collected from two sources: university courses and MOOC platforms. Coursera and Edx are among the platforms used and preferred worldwide. It is anticipated that making the data on these platforms publicly available will contribute to more research and guide the studies to be conducted in the field of education. Privacy and ethics come to the fore within the scope of open data publication. In this context, developing some standards and new approaches for sharing data with researchers in a standard form without privacy violations will significantly contribute to the studies conducted in this field. Stanford University has taken a step on the subject. Datastage: Lagunita, the Stanford example of the NovoEd, Coursera, and OpenEdX platforms, maintains learning research data from courses offered on all three platforms. Access to this data is available upon request. Although access to some data (such as student certification) is limited, making MOOC data publicly available is essential. We hope that researchers can conduct their work on publicly available data in the future. EDM studies and students' privacy should be considered, and appropriate measures should be taken to protect data confidentiality. EDM studies may pose ethical problems. In particular, ethical problems may arise in data collection, data analysis, and interpretation of results. Therefore, it is anticipated that ethical issues will need to be considered in future EDM studies.

In many of the studies examined, the analysis tool is not specified. Specification of the analysis tools is essential to provide ideas for future studies, and it is recommended that researchers pay attention to the analysis tools they use in their future studies.

MOOCs were initially offered free of charge and open to everyone, but later on, as MOOC platforms became independent educational companies, there was a need for paid courses to finance the courses and ensure the sustainability of the platforms. Therefore, in addition to standard paid courses, producing new content formats that provide financial resources, such as micro-certifications and corporate training, has become a current issue. No study was found in the reviewed studies comparing paid and free MOOCs. It is anticipated that it will be essential to include research on whether paid and free MOOCs impact student success, attendance, or certification in future studies.

Authors' Contributions

Rukiye ORMAN: Conceptualization, methodology, formal analysis, research, data curation, visualization, writing-original draft preparation, writing-review and editing; Nergiz Ercil CAGILTAY; supervision, project management, writing – original draft preparation, writing – review and editing; Hasan CAKIR; supervision, project management, writing – original draft preparation, writing – review and editing. All authors have read and accepted the published version of the manuscript.

Competing Interests

The authors have no competing interests to declare.

References

[1] M. Liu and D. Yu, "Towards intelligent e-learning systems," *Education and Information Technologies*, vol. 28, no. 7, pp. 7845–7876, 2023.

[2] R. Orman, E. Şimşek, and M. Kozak Çakır, "Micro-credentials and reflections on higher education," *Higher Education Evaluation and Development*, vol. 17, no. 2, pp. 96–112, 2023.

[3] J. Goopio and C. Cheung, "The mooc dropout phenomenon and retention strategies," *Journal of Teaching in Travel & Tourism*, vol. 21, no. 2, pp. 177–197, 2020.

[4] P. Diver and I. Martinez, "Moocs as a massive research laboratory: Opportunities and challenges." *Distance Education*, vol. 36, no. 1, pp. 5–25, 2015.

[5] A. Bozkurt, "Bağlantıcı kitlesel açik çevrimiçi derslerde etkileşim örüntüleri ve öğreten-öğrenen rollerinin belirlenmesi," Ph.D. dissertation, Anadolu University (Turkey), 2015.

[6] T. Jadin and M. Gaisch, "Extending the moocversity a multi-layered and diversified lens for mooc research." *Proceedings of the European MOOC Stakeholder Summit*, pp. 73–78., 2014.

[7] K. Jordan, "Massive open online course completion rates revisited: Assessment, length and attrition," *The International Review of Research in Open and Distributed Learning*, vol. 16, no. 3, pp. 341–358., 2015.

[8] B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli, "A survey of machine learning approaches for student dropout prediction in online courses," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020.

[9] N. Çağıltay, K. Çağıltay, and B. Çelik, "An analysis of course characteristics, learner characteristics, and certification rates in mitx moocs." *The International Review of Research in Open and Distributed Learning*, vol. 21, no. 3, pp. 121–139, 2020.

[10] D. F. Onah, J. Sinclair, and R. Boyatt, "Dropout rates of massive open online courses: behavioural patterns." in *EDULEARN14 proceedings,*, 2014, Conference Proceedings, pp. 5825–5834.

[11] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[12] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions." *JEDM Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[13] L. N. M. Bezerra and M. T. Silva, "Educational data mining applied to a massive course," *International Journal Of Distance Education Technologies*, vol. 18, no. 4, pp. 17–30, 2020.

[14] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135–146, 2007.

[15] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works. expert systems with applications," *Expert systems with applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[16] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, pp. 135–146, 2017.

[17] K. Aulakh, R. K. Roul, and M. Kaushal, "E-learning enhancement through educational data mining with covid-19 outbreak period in backdrop: A review," *International journal of educational development*, vol. 101, p. 102814, 2023.

[18] J. M. Gallego-Romero, C. Alario-Hoyos, I. Estévez-Ayres, and C. D. Kloos, "Analyzing learners' engagement and behavior in moocs on programming with the codeboard ide," *Etr&D-Educational Technology Research and Development*, vol. 68, no. 5, pp. 2505–2528, 2020.

[19] N. Bousbia and I. Belamri, "Which contribution does edm provide to computer-based learning environments?" *Educational data mining: Applications and trends*, 2014.

[20] A. Hicham, A. Jeghal, A. Sabri, and H. Tairi, "A survey on educational data mining [2014-2019]," *IEEE*, 2020.

[21] R. A. Razak, M. Omar, and M. Ahmad, "A student performance prediction model using data mining technique." *International Journal of Engineering & Technology*, vol. 7, no. 2.15, pp. 61– 63, 2018.

[22] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years." *Education and Information Technologies*, vol. 23, pp. 537–553, 2018.

[23] S. Shatnawi, M. M. Gaber, and M. Cocea, "Text stream mining for massive open online courses: review and perspectives," *Systems Science & Control Engineering*, vol. 2, no. 1, pp. 664–676, 2014.

[24] O. R. Yürüm, T. Taskaya-Temizel, and S. Yildirim, "Predictive video analytics in online courses: A systematic literature review," *Technology Knowledge And Learning*, 2023.

[25] M. E. Buitrago-Ropero, M. S. Ramírez-Montoya, and A. C. Laverde, "Digital footprints (2005-2019): a systematic mapping of studies in education," *Interactive Learning Environments*, vol. 31, no. 2, pp. 876–889, 2023.

[26] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, 2021.

[27] E. Araka, E. Maina, R. Gitonga, and R. Oboko, "Research trends in measurement and intervention tools for self-regulated learning for e-learning environments-systematic review (2008-2018)," *Research And Practice In Technology Enhanced Learning*, vol. 15, no. 1, 2020.

[28] M. Bearman, C. D. Smith, A. Carbone, S. Slade, C. Baik, M. Hughes-Warrington, and D. L. Neumann, "Systematic review methodology in higher education," *Higher Education Research & Development*, vol. 31, no. 5, pp. 625–640, 2012.

[29] B. N. Green, C. D. Johnson, and A. Adams, "Writing narrative literature reviews for peer-reviewed journals: secrets of the trade," *J Chiropr Med*, vol. 5, no. 3, pp. 101–17, 2006.

[30] M. L. Pan, *Preparing literature reviews: Qualitative and quantitative approaches.* Taylor & Francis., 2016.

[31] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in moocs: A review and future research directions," *IEEE Transactions on Learning Technologies*, vol. 12, no. 3, pp. 384–401, 2018.

[32] A. Ani and E. T. Khor, "Development and evaluation of predictive models for predicting students performance in moocs," *Education and Information Technologies*, 2023.

[33] M. Youssef, S. Mohammed, E. Hamada, and B. Wafaa, "A predictive approach based on efficient feature selection and learning algorithms competition: Case of learners dropout in moocs," *Education and Information Technologies*, vol. 24, no. 6, pp. 3591–3618, 2019.

[34] C. T. Swai and S. E. Mangowi, "Mining school teachers' mooc training responses to infer their face-to-face teaching strategy preference," *The International Journal of Information and Learning Technology*, vol. 39, no. 1, pp. 82–94, 2022.

[35] S. Assami, N. Daoudi, and R. Ajhoun, "Implementation of a machine learning-based mooc recommender system using learner motivation prediction," *International Journal of Engineering Pedagogy (iJEP)*, vol. 12, no. 5, pp. 68–85, 2022.

[36] E. Ahmed, "Student performance prediction using machine learning algorithms." *Applied Computational Intelligence and Soft Computing*, vol. 1, p. 4067721, 2024.

[37] A. Alghamdi, "Evaluating factors influencing learner satisfaction in massive open online course selection: A data-driven approach using machine learning," *Arabian Journal for Science and Engineering*, vol. 1, no. 26, 2024.

[38] S. Gupta and A. Sabitha, "Deciphering the attributes of student retention in massive open online courses using data mining techniques," *Education and Information Technologies*, vol. 24, pp. 1973–1994, 2019.

[39] D. J. Lemay and T. Doleck, "Grade prediction of weekly assignments in moocs: mining video-viewing behavior," *Education and Information Technologies*, vol. 25, no. 2, pp. 1333–1342, 2019.

[40] ——, "Predicting completion of massive open online course (mooc) assignments from video viewing behavior," *Interactive Learning Environments*, vol. 30, no. 10, pp. 1782–1793, 2022.

[41] D. Liang, J. Y. Jia, X. M. Wu, J. M. Miao, and A. H. Wang, "Analysis of learners' behaviors and learning outcomes in a massive open online course," *Knowledge Management & E-Learning-An International Journal*, vol. 6, no. 3, pp. 281–298, 2014.

[42] H. Wan, K. Liu, Q. Yu, and X. Gao, "Pedagogical intervention practices: Improving learning engagement based on early prediction," *IEEE Transactions on Learning Technologies*, vol. 12, no. 2, pp. 278–289, 2019.

[43] V. S. Pillutla, A. A. Tawfik, and P. J. Giabbanelli, “Detecting the depth and progression of learning in massive open online courses by mining discussion data,” *Technology, Knowledge and Learning*, vol. 25, no. 4, pp. 881–898, 2020.

[44] S. Li, J. Du, and J. Sun, “Unfolding the learning behaviour patterns of mooc learners with different levels of achievement,” *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, 2022.

[45] S. Rizvi, B. Rienties, J. Rogaten, and R. F. Kizilcec, “Investigating variation in learning processes in a futurelearn mooc,” *Journal of Computing in Higher Education*, vol. 32, no. 1, pp. 162–181, 2019.

[46] M. Saqr, V. Tuominen, T. Valtonen, E. Sointu, S. Väisänen, and L. Hirsto, “Teachers learning profiles in learning programming: The big picture!” *Frontiers in Education*, vol. 7, 2022.

[47] H. Tang, W. Xing, and B. Pei, “Exploring the temporal dimension of forum participation in moocs,” *Distance Education*, vol. 39, no. 3, pp. 353–372, 2018.

[48] Y. Lee, “Using self-organizing map and clustering to investigate problem-solving patterns in the massive open online course: An exploratory study,” *Journal of Educational Computing Research*, vol. 57, no. 2, pp. 471–490, 2018.

[49] A. Cohen and S. Holstein, “Analysing successful massive open online courses using the community of inquiry model as perceived by students,” *Journal Of Computer Assisted Learning*, vol. 34, no. 5, pp. 544–556, 2018.

[50] Y. Dyulicheva, “Learning analytics in moocs as an instrument for measuring math anxiety,” *Voprosy Obrazovaniya / Educational Studies Moscow*, no. 4, pp. 243–265, 2021.

[51] M. Nilashi, R. A. Abumalloh, M. Zibarzani, S. Samad, W. A. Zogaan, M. Y. Ismail, S. Mohd, and N. A. M. Akib, “What factors influence students satisfaction in massive open online courses? findings from user-generated content using educational data mining,” *Education and Information Technologies*, vol. 27, no. 7, pp. 9401–9435, 2022.

[52] C. Geigle and C. Zhai, “Modeling mooc student behavior with two-layer hidden markov models,” *Journal of Educational Data Mining*, vol. 9, no. 1, pp. 1–24, 2017.

[53] S.-H. Zhong, Y. Li, Y. Liu, and Z. Wang, “A computational investigation of learning behaviors in moocs,” *Computer Applications in Engineering Education*, vol. 25, no. 5, pp. 693–705, 2017.

[54] C. Xu, G. Zhu, J. Ye, and J. Shu, “Educational data mining: Dropout prediction in xuetangx moocs,” *Neural Processing Letters*, vol. 54, no. 4, pp. 2885–2900, 2022.

[55] D. Yang, R. E. Kraut, and C. P. Rose, “Exploring the effect of student confusion in massive open online courses,” *Journal of Educational Data Mining*, vol. 8, no. 1, pp. 52–83, 2016.

[56] Y. Nie, H. Luo, and D. Sun, “Design and validation of a diagnostic mooc evaluation method combining ahp and text mining algorithms,” *Interactive Learning Environments*, vol. 29, no. 2, pp. 315–328, 2020.

[57] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, “Educational data mining to predict students’ academic performance: A survey study,” *Education and Information Technologies*, vol. 28, no. 1, pp. 905–971, 2023.

[58] M. M. George and P. S. Rasmi, “Performance comparison of apache hadoop and apache spark for covid-19 data sets.” pp. 1659–1665, 2022, January 2022.

[59] D. M. Dener, Murat and A. Orman, “Açık kaynak kodlu veri madenciliği programları: Weka’da örnek uygulama,” *Akademik Bilişim*, vol. 9, pp. 11–13, 2009.

[60] WEKA, 2024. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[61] A.-F. J., S. L., G. S., del Jesus M. J., V. S., G. J. M., O. J., R. C., B. J., R. V. M., F. J. C., and H. F., “Keel: A software tool to assess evolutionary algorithms to data min-ing problems,” *Soft Computing*, vol. 13, no. 3, pp. 307–318, 2009.

[62] ORANGE, 2024. [Online]. Available: http://orange.biolab.si/, (Erişim Tarihi: 2024).

[63] R, 2024. [Online]. Available: http://www.r-project.org/