EDUCATIONAL DATA MINING IN DISTANCE EDUCATION: A SYSTEMATIC LITERATURE MAPPING STUDY

Assoc. Prof. Dr. Aslıhan TÜFEKCİ
asli@gazi.edu.tr
Gazi University, Turkey
Lec. Esra Ayça GÜZELDERELI YILMAZ
eguzeldereli@aku.edu.tr
Afyon Kocatepe University, Turkey

ABSTRACT

The education-training process and all activities related to this process have the power to direct the future of society. Nowadays, online distance education with rapidly spreading information and communication technology can overcome the various problems of traditional face to face education. It is a practical training system especially for women who do not work in our country. From this point of view, the process should be analyzed frequently concerning input, output, and other process elements. System requirements, changes in users' demand over time and open system design depend on changes in demands are essential factors to achieve maximum performance from the system at the design of web-based distance education systems. In web-based education, the restructuring of courses or contrary pages with taking into account individual differences provides a determination of individual needs and the best way for the adaptation of learners. Educational data mining is a multidisciplinary research area that develops methods and techniques for discovering data derived from various information systems used in education. This systematic mapping study aims to provide an overview of the current work of educational data mining practices in distance education. In this study, various search engines were used to search the academic literature. This systematic map and its results are based on 140 primary sources that involve articles published in conferences, articles published in the magazines, symposium articles, theses, and others.

Keywords: Distance education, Educational data mining, Systematic literature mapping

INTRODUCTION

Distance education has taken its place among the indispensables of education and training in the developed world as well as in most developed countries. This is due to changes in the needs and expectations of people in parallel with the developments in technology. The widespread use of the internet in our age has an essential place in accessing information. It also becomes an indispensable tool for information sharing and communication between communities and people living in the same society (URL-1, 2010). Distance education is an alternative to traditional educational problems, and it is a method of teaching by organizing educational activities and providing communication and interaction between practitioners and students through a specific center through specially prepared teaching units and various mediums (İçten, 2006).

Information and information technology from being just a tool, teachers, students and educational institutions-institutions that are effective in undertaking different tasks. At the same time, it is among the most important elements that have a say in the world economy. All societies aim to call today's economy as knowledge and technology economy; from education to healthcare using information technology to improve in every area of human resources and putting lifelong learning as the priority is to try to obtain a place for themselves in this area (Arı, 2010).

Successful distance education systems depend on students, faculty/institutions, assistants, technical staff and managers to work continuously and in teams. When designing distance education systems, design, interoperability, integration with other systems, scalability, performance levels, upgrade operations, support, security, and accessibility should be considered.

Technological developments that are now influential in educational environments have also enabled large volumes of data to be generated in the field of education with the integration of technology into education. The contribution of this data set to the quality of education is directly related to the uncovering of meaningful patterns within the data. For quality education, higher education institutions should be able to make the right decisions in an administrative and educational sense. Incorrect or incomplete academic planning, failing students, students who can leave school are the problems of higher education institutions. Solving these problems and taking precautions are very important for the quality of education.

The education-training process and all activities related to this process have the power to direct the future of society. From this point of view, it can be said that the process should be analyzed frequently regarding input, output, and other process elements. Although this analysis is carried out at micro and macro level achievement exams, the convergence of the achieved success to the desired one is controversial when the exam scores are considered as the only input. Therefore, it is important to predict the transition period of input to the desired output to establish the awareness of the situations in which interruptions in the process should be intervened. In this respect, educational data mining methods can be a powerful tool for academic interventions.

In the field of distance education, educational data mining studies are carried out widely, especially in the area of ethics, data protection, storage of data in appropriate systems, detailed data collection for in-depth analysis of learning processes, and development of reports that can be easily understood by teachers and learners.

Educational data mining concerns issues such as developing recommendations for students, providing feedback and support to trainers, modeling student data, grouping students, supporting planning activities, and identifying student analysis. Educational data mining is a relatively new field of study in scholarly research, and it increases its importance among educators day by day. It is an interdisciplinary field of study which is directly related to many areas such as computer science, statistics, mathematics, data visualization, etc. It aims to transform the data produced by the information and communication technologies used in education into meaningful information for actors involved in education by analyzing them with various methods. When the studies conducted in the field of education are examined, data mining seems to be used for classification, clustering, cohesion rules, methods and techniques for students' achievements, clustering according to their information, determination of their interests and trends, automatic presentation of learning contents and revealing misconceptions.

In this study, the educational data mining applications in distance education attempted to capture an overview of the systematic classification and mapping of current relevant studies. The contribution of the research is a systematic mapping of the primary sources that exist in the online resource repository on educational data mining distance education. The rest of this study is organized as follows; a summary of the educational data mining and its applications are given in Part 2. The research methodology, including the general systematic mapping process, the target followed in this study, and the

research questions are listed in Part 3. Part 4 discusses the source selection process. Section 5 analyzes the recurrent development of the map. The results of systematic mapping are presented in Part 6. Finally, Part 7 specifies the results of this mapping and future work to be done.

FIELD SUMMARY AND RELATED STUDIES

Distance education can be defined as methods and techniques that enable the learner who is far away from the teaching as time and space to reach the education program (Romero and Ventura, 2007). Many different methods have been used in distance education from past to present, such as learning by letter, audio and video cassette training, radio and TV broadcasting, teleconferencing and computer-aided education. Nowadays, in parallel with technological developments, these methods have been replaced by internet-based training which is easier to use and access. Internet-based education systems; course content preparation tools, simultaneous and asynchronous conference systems, questionnaires and quiz components, virtual work environments for resource sharing, whiteboard, note reporting system, diary book, homework publishing (Zaiane and Luo, 2001). However, the educational activities carried out by the students with these tools cannot be wholeheartedly followed up and evaluated by the educators. Although they present statistical reports on student activities, they do not have advanced tools to derive meaningful information for understanding student mobility. Therefore, data mining is used to reveal significant information, to define data patterns, to visualize and analyze data (Talavera and Gaudioso, 2004).

The technological developments that effect on educational environments have enabled the accumulation of large amounts of data in the field of education with the integration of technology in education. Data mining studies in education are driven by the fact that there is still unexplored knowledge available to students, teachers, administrative staff, and educational institutions in large volumes of data. Therefore, the information obtained from these data stacks will play an active role in the design of future educational environments.

Educational data mining concerns issues such as developing recommendations for students, providing feedback and support to trainers, modeling student data, grouping students, supporting planning activities, and identifying student analysis. Educational data mining is a relatively new field of study in educational research, and it increases its importance among educators day by day. It is an interdisciplinary field of study which is directly related to many areas such as computer science, statistics, mathematics, data visualization, etc. It aims to transform the data produced by the information and communication technologies used in education into meaningful information for actors involved in education by analyzing them with various methods. When the studies conducted in the field of education are examined, data mining seems to be used for classification, clustering, cohesion rules, methods and techniques for students' achievements, clustering according to their information, determination of their interests and trends, automatic presentation of learning contents and revealing misconceptions.

According to Calders and Pechenizkiy (2012), educational data mining is a multidisciplinary research area that develops methods and techniques for discovering data derived from various information systems used in education. Along with the increase in educational data, educational data mining has become a rich application area for data mining as well as learning knowledge. Educational data mining contributes to the understanding of the learning styles of learners and also enables data-driven decision-making to develop existing education practices and learning materials. Baker et al. (2010)

define educational data mining as a discipline that develops methods of discovering unique types of data coming from educational environments and using them to understand better students and how they learned. According to another definition, educational data mining is the application of data mining methods and techniques to specific sets of data coming from educational environments to find answers to educational questions (Romero & Ventura, 2010).

Many information is obtained such as the application of educational data mining techniques on the data in distance education systems, identifying frequently and rarely used paths, identifying pages that are not visited at all, and how learner-based groups use them. Obtaining this information for learners will play an active role in providing better learning in similar web-based applications. For example, in the previous form, the short paths of user activities as suggestions for learners or activity suggestions will play a useful role in the development of similar learning. In short, if the learning attitudes, interests and previous behaviors of the students can be adapted logically to the system content, this can be beneficial.

RESEARCH METHOD

An overview of the research method, the objectives and mapping questions are presented in this section.

The Purpose of the Study and the Mapping Questions

This work aims to identify the challenges and to find alternatives for future research from the perspectives of researchers and practitioners. To this end, the literature on educational data mining applications in distance education has been systematically mapped and reviewed to find the current approaches and trends in this area. Based on this reasoning, the following mapping questions (MQ) were created:

- **MQ 1- Mapping of studies by type of contribution to the field: The question of what type of contributions the scientific studies on educational data mining in distance education make to the field regarding the method, techniques, models, tools, processes, etc. According to Peterson's systematic mapping studies, the contribution type is a commonly used practice. Answering this question will help us to understand the tendency of the area that the current researches focus on in error determination.**
- **MQ 2- Mapping by type of research: It is aimed to answer the question of which research method was used to develop the studies. Peterson also introduced guidelines for classifying research approaches of the studies, and these principles were used to answer this question. To respond to MQ 2, primary studies were classified according to 6 different research methods. Each study was classified as including only one research method.**
- **MQ 3- Mapping according to educational data mining technique used: Mapping has been done according to data mining techniques used in the study of educational data mining. This step will help to understand the trend of existing data mining techniques and algorithms used.**
- **MQ 4- Mapping according to the purpose of use of data mining in distance education: In scientific studies covering data mining applications applied in distance education, the question of the purpose of using data mining techniques was sought. The answer to this question will help to understand how the scientific studies that are being developed in this area intensely serve for and what the general tendency is.**

An Overview of the Process

As mentioned before, this systematic mapping study is conducted based on the guidelines provided by Peterson et al. [1]. The process underlying this systematic map is summarized in Figure 1, which consists of three phases:

- Article selection (Part 4)
- Development of systematic mapping (Part 5)
- Results of systematic mapping (Part 6)

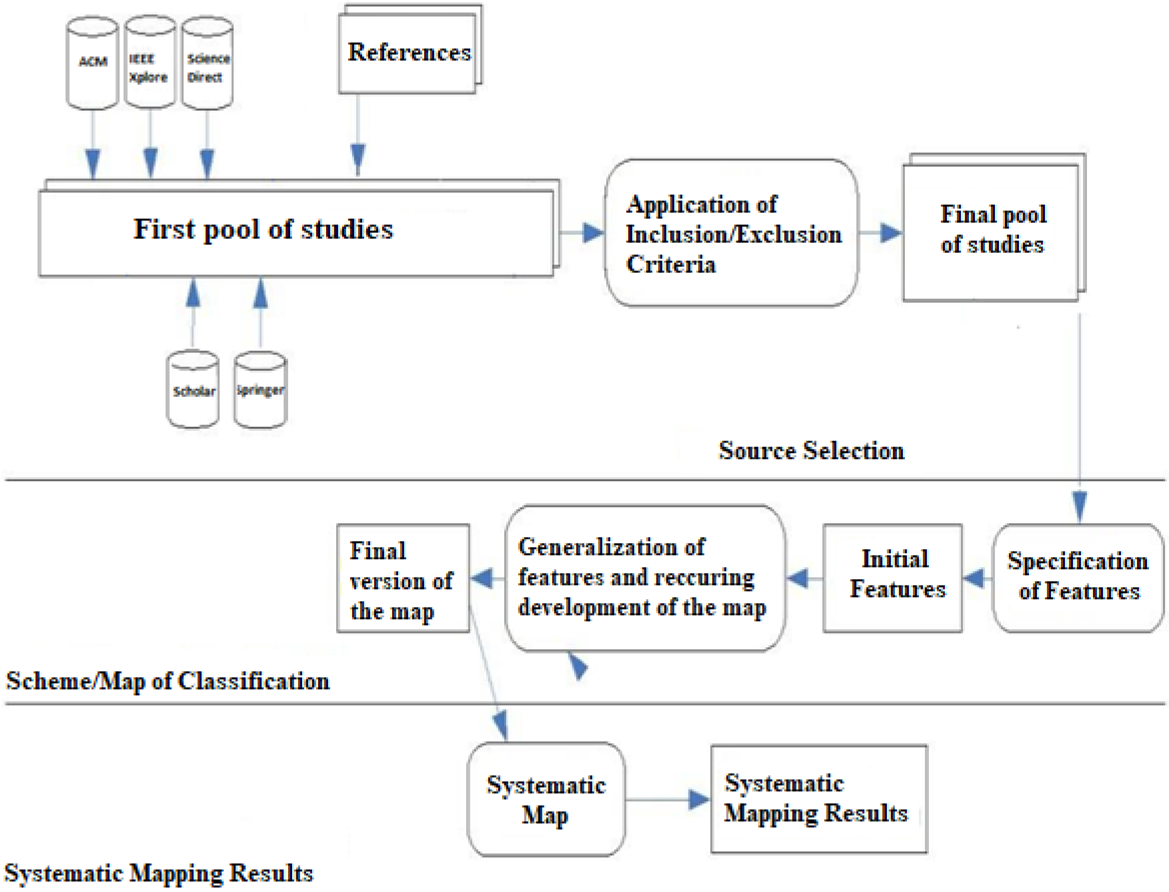
ACM
IEEE Xplore
Science Direct
References
First pool of studies
Application of Inclusion/Exclusion Criteria
Final pool of studies
Scholar
Springer
Source Selection
Final version of the map
Generalization of features and reccuring development of the map
Initial Features
Specification of Features
Scheme/Map of Classification
Systematic Map
Systematic Mapping Results
Systematic Mapping Results

Figure 1:
The protocol used for systematic mapping studies

Resource Selection

The first step in systematic mapping is the selection of resources. In this phase, the following steps were applied in order:

- Resource selection and search for key words (Part 4.1)
- Inclusion / exclusion criteria (Part 4.2)
- Completing the resource pool (Part 4.3)

Source Selection and Keywords to Search

The digital libraries used to find resources in the study are IEEE Xplore, (2) ACM Digital Library, (3) Science Direct, (4) Springer, and (5) Google Scholar. The search begins with the search keyword "educational data mining in distance education." This search resulted

in a total of 143 studies in the initial pool. By reviewing the summary and introduction parts of the studies obtained with this keyword, their relevance to the research field was evaluated concerning the reliability of the studies and the number of studies in the resource pool was reduced to 140 as a result of this evaluation. Also, some of the resources referenced by the studies were searched manually to minimize the risk of overlooking the related studies, and the studies that were not in the resource pool but were likely to be relevant were included in the study.

Inclusion/Exclusion Criteria

The inclusion criteria considered in this study were: (1) the relevance of each study to the context of educational data mining; (2) the level of coherence, evaluation and validity followed in the work. Only studies written in English and accessible only electronically were included. The studies that are related to the scope but do not have valid evidence were excluded. Articles related to the scope but not accessible for free were also excluded. To apply the inclusion/exclusion criteria in the first pool, each study was rated as "1" and "0" by evaluating the studies in the first pool. "1" indicates that the study might be included, and "0" indicates that the study might be excluded. The title, abstract, and keywords of the articles were reviewed to identify each study. If there was insufficient information available from these sources, a more in-depth assessment was made. As a result, the final pool was reduced from 143 to 140.

Recent Article Pool and Online Storage

The spreadsheet link (e-table) can be checked for the full reference list of 140 primary resources. The final pool of selected studies was published in an online repository using the Google Docs system. The classification of each publication chosen by the classification scheme described in Chapter 5 is also available in the online repository. The annual publication volume of educational data mining in distance education is shown in Figure 2. Regarding the beginning year of the publication period, educational data mining studies have begun to emerge in the year 2000, and there seems to be a growing concentration of work since 2013. Forty-three studies from 2017-2018 on educational data mining were included in the mapping.

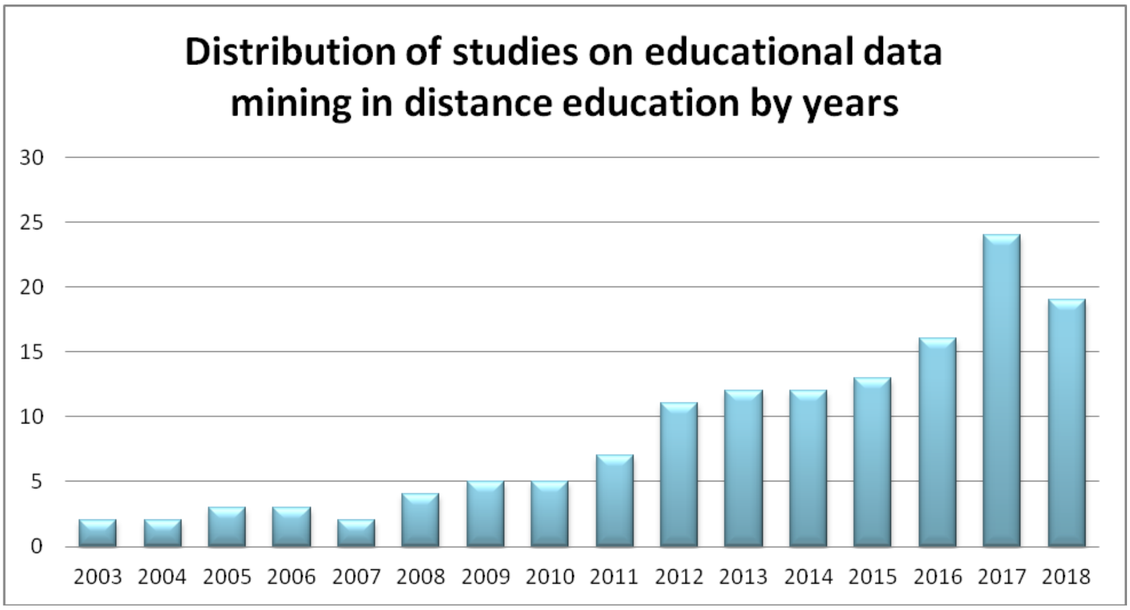

Figure 2:
Distribution of studies by years

THE DEVELOPMENT OF SYSTEMATIC MAP (CLASSIFICATION SCHEME)

Table 1 shows the final classification scheme developed after applying the processes described above. Column 1 in the table is the mapping question (MQ) list. Column 2 is the corresponding attribute/property. Column 3 is a set of all possible values for the property. Finally, column 4 specifies an attribute as to whether more than one selection can be applied.

Table 1:
The systematic map developed and used in the study

MQ	Attributes	Categories	(M)ultiple/ (S)ingle
1	Type of contribution	{Methods / Techniques, Models, Tools, Processes, Case studies, Others}	M
2	Type of research	{Basic research, Applied research, Experimental Development, Product development, Descriptive research, Others}	S
3	Data mining technique	{Association analysis, Clustering, Classification, Estimation by ANN, Others}	M
4	Intended use	{Identification, Predict, Knowledge discovery, Comparison, Others}	S

RESULTS

The results obtained in the context of the mapping questions asked in the systematic mapping study are presented in this section.

MQ1- Contribution Types of Studies to the Research Field

The first research question aims to find out how many studies have contributed to the literature through educational data mining methods/techniques, models, tools, processes, case studies, and others. Figure 3 shows the distribution of the contribution types of all 140 sources involved in the study.

Figure 3 shows that a large number of studies contribute to the field of the data mining process. It has also been determined that 49 studies have committed or developed an existing technique/methodology with new methods/techniques and models. Thirty-one studies that produced results by applying techniques on the case and 11 studies that contributed by the new tool were identified. According to contributions, an article can be included in more than one classification.

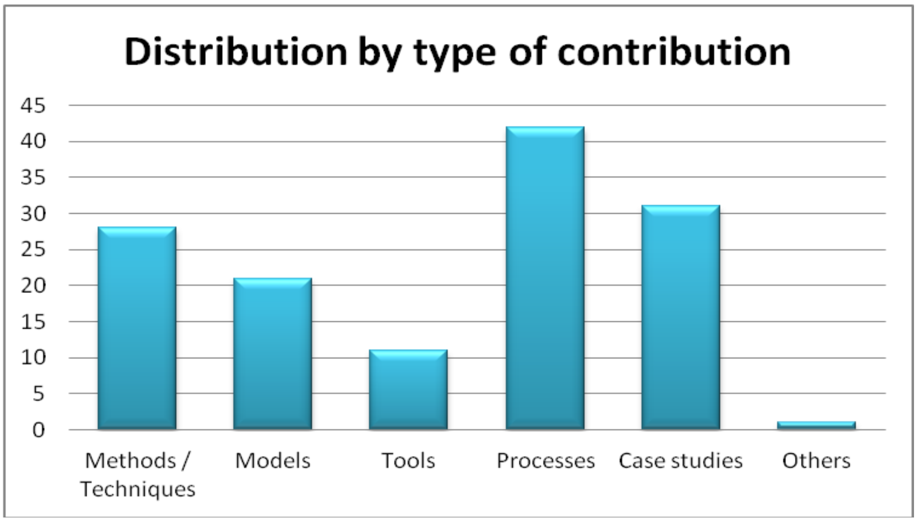

Figure 3:
Distribution by type of contribution

MQ2- Type of Research

This research question is aimed to determine what kind of research methods are used in the studies. Figure 4 shows the distribution of studies regarding research types.

In the pool of 140 studies included in the mapping study, Figure 4 shows that the most used research method is the necessary research. The results show that 54 articles use research questions or hypotheses. Following the primary analyses, there are also a lot of studies to apply data mining techniques on educational data in distance education systems. Nevertheless, the number of empirical or product-oriented studies that perform these functions is minimal. Especially in recent years, it has been determined that the emphasis is given to explanatory researches that present future proposals.

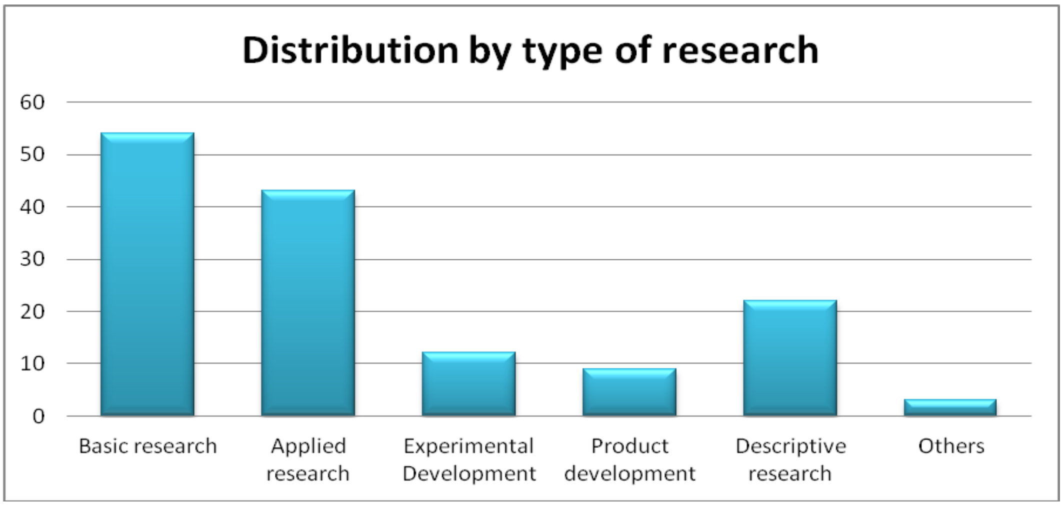

Figure 4:
Distribution by type of research

MQ3- Types of Data Mining Technique

To investigate which data mining techniques were used in educational data mining studies which are applied in distance education systems, the studies in the pool were

classified according to data mining techniques. Figure 5 shows the distribution of the methods used in all of the 140 studies examined.

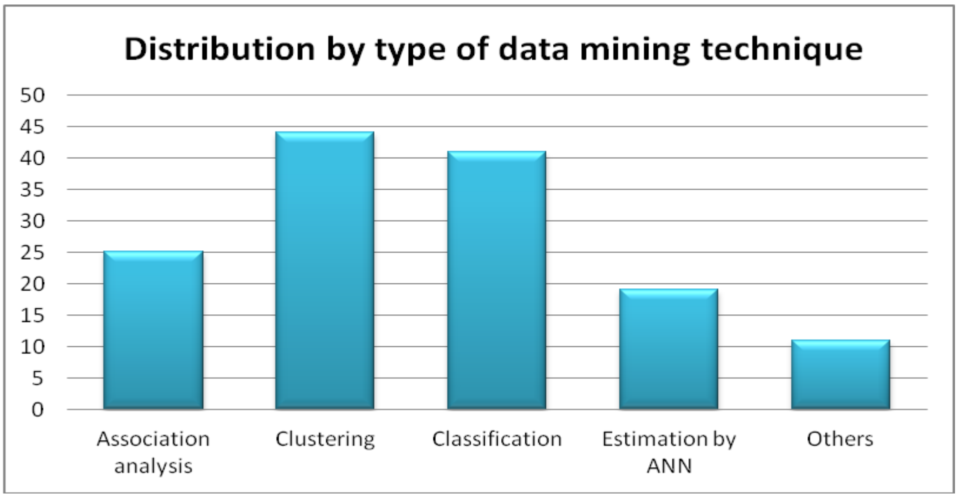

Figure 5:
Distribution by type of data mining technique

MQ4- Purpose of Educational Data Mining

To study how data mining techniques are being used for educational data mining studies in distance education systems, the data mining activities in the pool were classified by the purpose of use. Figure 5 shows the objects for which the techniques used in all of the 140 studies examined serve. Accordingly, it can be said that the reviews were generally developed towards the purpose of forecasting, identification, and discovery of information. The number of publications prepared for the comparison of techniques is relatively low.

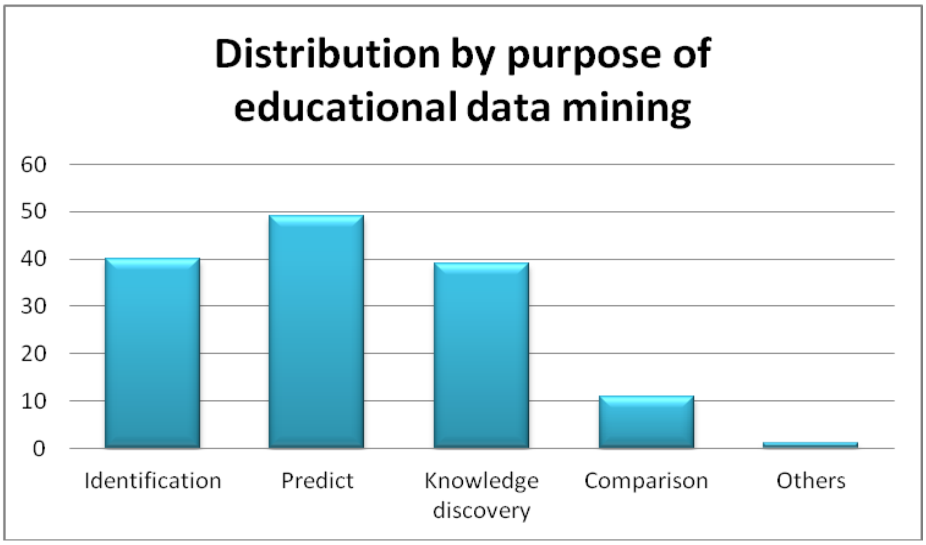

Figure 6:
Distribution by purpose of educational data mining

CONCLUSIONS

In this article, a systematic mapping study was conducted to characterize the educational data mining in the distance education area. A total of 143 primary studies were analyzed, and 140 sources remained in the final pool after filtering with inclusion and exclusion criteria. Subsequently, research questions were formed within the scope of the study. In response to the research questions, the charts related to the problems were analyzed. Accordingly, it is seen that studies which is educational data mining used in distance education systems began to emerge in the year 2000 and that since 2012 there has been a growing study intensity. It has been found that numerous studies have contributed to the data mining process. It has been determined that there is a density in the number of studies demonstrating new methods/techniques and models.

On the other hand, there are very few studies that put forth new tools. It has also been observed that studies in which classification and clustering techniques are predominantly studied, but studies that make predictions for the future are relatively small. As future work, based on this work, it is planned to carry out a Systematic Literature Review (SLR) study in the field of educational data mining by extending this work from different angles.

REFERENCES

Akcapinar, G., Altun, A., & Aşkar, P. (2015). Modeling Students' Academic Performance Based on Their Interactions with the Online Learning Environment, 14(2), 815–824. https://doi.org/10.17051/io.2015.03160

Akçapınar, G. (2014). Veri madenciliği algoritmalarını kullanarak öğrenci verilerinden birliktelik kurallarının çıkarılması (Tez No. 381422).

Akça, F. (2014). Veri madenciliği ile fen fakülteleri öğrenci profillerinin incelenmesi: Gazi Üniversitesi örneği (Tez No. 372909).

Aksoy, E. (2014) Matematik alanında üstün yetenekli ve zekalı öğrencilerin bazı değişkenler açısından veri madenciliği ile belirlenmesi (Tez No. 368273).

Alan, M. A. (2012). Veri madenciliği ve lisansüstü öğrenci verilerine üzerine bir uygulama. Dumlupınar Üniversitesi Sosyal Bilimler Dergisi (33), 165-174.

Alan, M. A. (2014). Karar Ağaçlarıyla öğrenci verilerinin sınıflandırılması. Atatürk Üniversitesi iktisadi ve idari bilimler dergisi, 28(4), 101-112.

Aydın, S. (2007). Veri madenciliği ve Anadolu Üniversitesi uzaktan eğitim sisteminde bir uygulama (Tez No. 220873).

Aydoğdu, Y. (2011). Evaluating e-learning environment by using data mining techniques/Elektronik öğrenme ortamlarının veri madenciliği teknikleri ile değerlendirilmesi (Tez No. 298383).

Ayesha, S., Mustafa, T., Sattar, A. R., and Khan, M. I. (2010). Data mining model for higher education system. Europen Journal of Scientific Research, 43 (1), 24-29.

Ayık, Y. Z., Özdemir, A., & Yavuz, U. (2007). Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği işe analizi. Atatürk Üniversitesi sosyal bilimler dergisi , 10 (2), 441-454.

Baker, R., & others. (2010). Data mining for education. International Encyclopedia of Education, 7(3), 112–118.

Baradwaj, B. K. and Pal, S. (2011). Mining educational data to analyze students' performance. International Journal of Advanced Computer Science and Applications, 2 (6), 63-69.

Baldi, P., & Brunak, S. (2001). Bioinformatics - The machine learning approach. Machine Learning.

Calders, T., & Pechenizkiy, M. (2012). Introduction to the Special Section on Educational Data Mining. SIGKDD Explor. Newsl., 13(2), 3–6. https://doi.org/10.1145/2207243.2207245

Fayyad, U. M., Weir, N., & Djorgovski, S. G. (1993). Automated Cataloging And Analysis Of Sky Survey Image Databases: The Skicat System. Cikm, 527–536. https://doi.org/10.1145/170088.170414

Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques" Morgan Kaufmann, Fourth Edition.

Gaafar, L. and Khamis, M. (2009). "Applications of Data Mining for Educational Decision Support", Proceedings of the 2009 Industrial Engineering Research Conference, 228-233.
İçten, T. (2006). Uzaktan eğitim öğrencileri için web tabanlı çevrimiçi sınav sistemi uygulaması geliştirilmesi (Yüksek lisans tezi), Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Kılınç, Ç. (2015). Üniversite öğrenci başarısı üzerine etki eden faktörlerin veri madenciliği yöntemleri ile incelenmesi (Tez No. 415460).

Kınay, E. (2016). Uzaktan eğitim sistemi analizi, Yaşar Üniversitesi , İstanbul (Tez No. 445016).

Krüger, A., Merceron, A., & Wolf, B. (2010). A data model to ease analysis and mining of educational data. In Educational Data Mining 2010.

Özbay, Ö. (2015). Öğretim yönetim sistemi üzerinde üniversite (lisans) düzeyindeki öğrenci hareketliliğinin very madenciliği yöntemleriyle analizi (Tez No. 414627).

Özbay, Ö. (2015). Veri madenciliği kavramı ve eğitimde veri madenciliği uygulamaları. Uluslararası eğitim bilimleri dergisi (5), 262-272.

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008) Systematic Mapping Studies in Software Engineering. EASE . 8, 68-77.

Romero, C. and Ventura, S. (2007) Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications, 33, 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured colla boration spaces. In Proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI 2004.

Taşdelen, A. (2014). Veri madenciliği yöntemleri ile mühendislik fakültesi uzaktan eğitim bölümlerinin analizi: Karabük Üniversitesi örneği (Tez No. 374695).

Thomas, E. H. and Galambos, N. (2004). What satisfies students? Mining studentopinion data with regression and decision tree analysis. Research in Higher Education, 45 (3), 251-269.

URL-1. (2010). Uzaktan eğitim: Türkiye' deki gelişmeler.
Erişim: http://www.scribd.com/doc/39002295/Uzaktan-Egitim-Turkiyede-Gelismeler.