Journal of Teacher Education and Lifelong Learning (TELL)

Volume: 4 Number: 2 Year: 2022 Research Article ISSN: 2687-5713

Using Big Data in Education: Curriculum Review with Educational Data Mining

Yusuf Ziya OLPAK[1] **Mustafa YAĞCI**[2]

[1] Kırşehir Ahi Evran University, Faculty of Education, Department of Computer Education and Instructional Technologies, Kırşehir, Türkiye
yusufziyaolpak@gmail.com

[2] Kırşehir Ahi Evran University, Faculty of Engineering and Architecture, Department of Computer Engineering, Kırşehir, Türkiye
mustafayagci06@gmail.com

Article Info

Article History
Received: 21/05/2022
Accepted: 11/11/2022
Published: 31/12/2022

Keywords:
Big data,
Educational data mining,
Association rules,
Apriori algorithm,
Relationship mining.

ABSTRACT

Today, most educational institutions have become more interested in big data. Because the importance of extracting useful information from educational data to support decision-making on educational issues has increased day by day. In this context, through educational data mining, this research study aims to reveal the association rules among compulsory courses in the Computer Education and Instructional Technology curriculum within the faculty of education of a state university in Turkey. In this context, the research was conducted with data obtained from 258 preservice teachers who had completed all of their compulsory courses (n = 42) for the Computer Education and Instructional Technology curriculum, having graduated from the Computer Education and Instructional Technology program between 2012 and 2020. According to the experimental results, the academic performance of preservice teachers in some courses could be used as a predictor of their academic performance in other courses. Other findings from the study are discussed in detail, and suggestions put forth for future research.

Citation: Olpak, Y. Z. & Yağcı, M. (2022). Using big data in education: Curriculum review with educational data mining. *Journal of Teacher Education and Lifelong Learning*, *4*(2), 181-195.

"This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)"

INTRODUCTION

Today, it can be said that the amount of data produced has increased in parallel with the development of information and communication technologies (ICT) in different areas of life. For example, since the 1800s, the number of scholarly and scientific journals has doubled approximately every 20 years (Mabe, 2003), making it now possible to access vast volumes of data. By processing this "big data" into meaningful information, patterns and trends can be revealed. Although no single common definition exists, big data can be defined as datasets whose size is beyond the capability of typical database software tools to capture, store, manage, and analyze (Yin & Kaynak, 2015). However, when the literature is examined, it can be seen that certain properties (known as the "6V's of Big Data") that make up big data have been defined in order to aid our full understanding (Alkatheri et al., 2019; Baaziz & Quoniam, 2013; Bozkurt, 2016; Daniel, 2015; Yin & Kaynak, 2015): 1) "Volume"—the quantity or size of the data; 2) "Variety"—the different types of data being generated (structured, unstructured, semi-structured); 3) "Velocity"—the speed of data in and out (batch, near time, real-time, streams); 4) "Value"—the purpose or the business outcome that the data brings in, to facilitate the decision-making process (useful meaning); 5) "Veracity"—the biases, noise, and abnormality in data; and, 6) "Verification"—the data verification and security. In addition to these properties, data must be collected, analyzed, and visualized in order to unlock the value of big data (Daniel, 2015).

Big data is used in many different areas such as agriculture (Lioutas & Charatsari, 2020), education (Fischer et al., 2020), healthcare (Shilo et al., 2020), and marketing (Jabbar et al., 2020), etc. For instance, in the study conducted by Yin and Kaynak (2015), it was stated that the main purpose behind the use of big data in industrial applications is to achieve desired performance levels, especially in terms of quality, whilst ensuring that the processes are conducted fault-free and with cost-efficiency. In this context, regarding the reflection of technology in education, it was reported that more than two-thirds of the global population live in districts covered by a mobile broadband network, and that mobile technology services have become more affordable than ever before, allowing technology to be readily integrated into education (International Telecommunication Union, 2016). Furthermore, most educational institutions have become more interested in the provision of online courses (Calvet Liñán & Juan Pérez, 2015). In the United States, almost 20 million students studied in postsecondary institutions in 2017, and 6.6 million of them took some form of distance education/online learning courses, including through mobile learning (Education Data, 2021). During the COVID-19 pandemic, many countries faced various forms of lockdown in order to prevent or slow the spread of the disease, and this resulted in widespread school closures affecting more than one billion learners worldwide (Tan et al., 2020). Therefore, alternative ways were employed for the continuation of education, with various tools used that provided both synchronous (e.g., chat, audio, and/or video conferencing) and asynchronous (e.g., e-mail, forums, blogs, websites) forms of communication. In educational environments, these interactive tools are mainly used in conjunction with learning management systems (LMS) such as Blackboard, Moodle, Sakai, Schoology, and TalentLMS, etc. Thus, detailed data on students' interactions with the LMS they use, their instructors, and other students has become available (e.g., information flows, navigation patterns, number of posts created, number of posts rated, number of clicks, number of sessions, reading files, social networks, etc.) (Akçapınar et al., 2019; Calvet Liñán & Juan Pérez, 2015; Pardo & Teasley, 2014; Siemens, 2013; Tempelaar et al., 2015). Calvet Liñán and Juan Pérez (2015) stated that educational data mining (EDM) and learning analytics (LA) are two different research areas devoted to the analysis of the aforementioned data. Additionally, in a study conducted by Elias (2011), it was stated that EDM and LA are closely related.

The main purpose of EDM and LA is to extract useful information from educational data in order to support decision-making on educational issues (Calvet Liñán & Juan Pérez, 2015). When the literature in this area is examined, it can be seen that the definitions of these two concepts also support this view. The most widely used definition of LA (Bozkurt, 2016) as "the measurement, collection, analysis and

reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (p. 3) was put forward by the organizers of the First International Conference on Learning Analytics and Knowledge (Long et al., 2011). On the other hand, EDM has been defined as "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in" by the International Educational Data Mining Society (Educational Data Mining Society, 2021, n.p.). In other words, EDM is an interdisciplinary field that applies statistical, machine-learning, and data-mining algorithms with different types of educational data in order to resolve issues related to educational research (Romero & Ventura, 2010).

Siemens and Baker (2012) stated five key differences between EDM and LA, which are "discovery," "reduction and holism," "origins," "adaptation and personalization," and "techniques and methods." In a study conducted by Calvet Liñán and Juan Pérez (2015), the differences and similarities between EDM and LA were addressed, and it was argued that while the focus of EDM is more on technique and methodology, LA is more concerned with applications. Bienkowski et al. (2012) reported there being no clear distinction between EDM and LA, and whilst LA focuses on applying known predictive models within instructional systems, EDM has its focus on looking for new patterns in data and developing new algorithms and/or new models.

The current research aims to utilize EDM to examine and reveal the association rules among the compulsory courses in the Computer Education and Instructional Technology (CEIT) curriculum within a faculty of education at a state university in Turkey. In research conducted by Altun and Ateş (2008), the problems that CEIT PSTs encountered during their undergraduate training and their professional concerns for the future were examined. The study's findings indicated that problems related to the CEIT curriculum were some of the most common (Altun & Ateş, 2008). For this reason, and due to the ease of accessibility for the researchers, the CEIT undergraduate program in Turkey was examined within the scope of the current study. Especially for CEIT PSTs, the courses they take mostly examine the different educational technologies on offer and how to employ them effectively within instructional activities.

The results of this study should help instructors make their educational activities more effective with data-driven decision support tools, resulting in increased grades and retention of preservice teachers (PSTs). Moreover, PSTs at risk of dropping out can be identified through the establishment of early-warning systems, and the opportunity taken to provide the necessary precautionary measures. Furthermore, the curriculum can be examined based on the analysis outcomes and updates applied to introduce improvements where needed and appropriate. Additionally, the study is aimed at helping educational administrators and policymakers recognize how EDM can best be applied for educational improvement. The current study, along with other similar studies, aims to develop a culture of utilizing big data in the process of educational decision-making.

THEORETICAL BACKGROUND

Data-rich educational systems can provide useful feedback (e.g., actionable, informative) to policymakers, administrators, instructors, and also to students. EDM processes that deal with this big data can be described as a system that "converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice" (Romero & Ventura, 2010, p. 601). In other words, EDM, which is gaining increased interest on a daily basis (Romero & Ventura, 2020), emphasizes the reduction of learning to components that support decision-making (Bienkowski et al., 2012). Although EDM is considered an emerging interdisciplinary research area (Romero & Ventura, 2013), many tools (e.g., Orange, Rapidminer, etc.) used in EDM have emerged over time, and even studies such as Slater et al. (2017) have been published that have reviewed these tools.

Although predicting academic performance is one of the most popular subject areas in EDM (Akçapınar et al., 2019; Fernandes et al., 2019), researchers also utilize EDM for purposes such as the

discovery or improvement of models regarding the knowledge structure of the domain, drop-out prediction, student profiling, and achieving a deeper understanding of educational phenomena (Baker, 2011; Calders & Pechenizkiy, 2012; Romero & Ventura, 2010). In order to achieve these and similar purposes, there exists a wide variety of methods that have proven popular within EDM, and that have been subjected to different classifications by different researchers (Baker, 2011; Baker & Yacef, 2009; Bakhshinategh et al., 2018; Romero & Ventura, 2007, 2020; Zaiane, 2002). For example, in a study conducted by Baker (2011), these methods were split between five general categories: "prediction" (Fernandes et al., 2019), "clustering" (Dutt et al., 2015), "relationship mining" (Hussain et al., 2019), "discovery with models" (Baker, 2007), and the "distillation of data for human judgment" (Baker & de Carvalho, 2008). Each method of using EDM works according to different features, and each should be taken into consideration when being employed. In this context, since the current study aims to determine the association rules among the Turkish CEIT curriculum's compulsory courses, relationship mining was conducted using the apriori algorithm.

Relationship mining, which aims to discover the relationships between variables in a dataset containing a large number of variables, has four types in general: "association rule mining," "correlation mining," "sequential pattern mining," and "causal data mining" (Baker, 2011). Association rule mining is an approach that supports future studies by analyzing past data and determining the associations that occur frequently together within a given dataset (Baker, 2011; Dunham, 2003). Furthermore, although there are different algorithms ("apriori," "predictive apriori," and "tertius") it is the apriori algorithm that has proven to be one of the most popular algorithms used in the literature when it comes to association rule mining (Dongre et al., 2014), as can be seen in many previous studies (Jhang et al., 2019; Natalia & Salvatore, 2020; Wu & Zeng, 2019), and was also preferred in the current research. Another reason for electing to employ the apriori algorithm was based on the research of Shweta and Garg (2013). In their study that compared the three association rule algorithms, apriori was shown to perform better than either predictive apriori or tertiary algorithms (Shweta & Garg, 2013).

Association rules are used frequently in many areas (Dunham, 2003) such as in marketing (Abinowi & Aminudin, 2020), advertising (Joshi & Sodhi, 2014), and product placement (Cil, 2012; Putra et al., 2018), and are also used in educational studies (Acharya & Madhu, 2012; Hung & Zhang, 2008; Hussain et al., 2019; Jha & Ragha, 2013; Kılınç, 2015; Kumar & Chadha, 2012; Moubayed et al., 2018; Ougiaroglou & Paschalis, 2012; Soimart & Mookdarsanit, 2016; Wu & Zeng, 2019). For example, Wu and Zeng (2019) revealed the association rules of 34 courses according to 100 students majoring from a Computer Science and Technology department using the apriori algorithm. The results of their study showed a strong correlation to exist between certain English-related courses. Furthermore, a strong correlation was found between hardware-related courses and between courses and their prerequisites. Moreover, Wu and Zeng's (2019) study also showed that the foundation of Discrete Mathematics is also important for the achievement of high success scores for certain basic professional courses (e.g., Programming Language and Introduction to Computer Science).

In a study conducted by Soimart and Mookdarsanit (2016), an apriori-based model was designed to recommend an appropriate discipline for high school students that was consistent with their skills. The model that was developed within the scope of Soimart and Mookdarsanit's (2016) research was related to dependent variables such as the students' GPA, their interests, skills, as well as their academic scores in English, Mathematics, Computer, Science, and Social Sciences, as well as independent variables such as management, mass communication, information system, and accounting. The researchers used 3,000 samples invalidating their model and found 11 meaningful association rules with a confidence ranging from 58% to 90%.

In a study conducted by Hung and Zhang (2008), a total of 17,934 server logs were analyzed in the examination of 98 higher education students' learning behaviors within an online course so as to construct knowledge on the typical patterns of the students' online learning behaviors. As a result of their study,

active and passive learners could be differentiated through the determination of the students' behavioral patterns and preferences within online learning processes. Moreover, the researchers also stated that these results showed how EDM could be employed so as to help improve educational activities with suggestions put forward for stakeholder groups such as online instructors and instructional designers.

The general purpose of a study conducted by Kumar and Chadha (2012) was to use the apriori algorithm for association rule mining in examining students' course performance. The generated association rules revealed that different variables can influence students who do not reach a sufficient level of performance at the post-graduate level. Furthermore, their study's results showed the association rules to be very helpful both for academic administrators and for curriculum planners.

Purpose of the Study

The cycle of applying data mining in educational systems prepared by Romero and Ventura (2007) provides detailed insight into how stakeholders such as students, teachers, and education researchers can benefit from big data. For instance, knowledge revealed from the processing of data collected within educational systems (e.g., students usage and interaction data, course information, academic data, etc.) using different data mining methods can be shown to educators and/or students in offering recommendations (Romero & Ventura, 2007). Furthermore, Yang and Hu (2011) stated that by analyzing data in educational systems using the apriori algorithm, useful rules for arranging courses, quality education, and educational models can be generated. In this context, the current study was conducted in order to reveal the association rules among compulsory courses of the CEIT curriculum in Turkey, which were examined using the apriori algorithm.

METHOD

Dataset

The CEIT program was examined within the scope of this research as one of 25 teacher education bachelor's degree programs in Turkey, which are generally applied as 4-year programs (Yükseköğretim Kurulu [Turkish Council of Higher Education], 2021b). However, some universities can apply for a 1-year compulsory English language preparatory program. Furthermore, specialized bachelor's degree programs may be longer than 4 years, such as medicine which is 6 years. In the 2006-2007 academic year, certain changes were introduced by the Turkish Council of Higher Education in the curriculum of education and educational science faculties; this update included changes introduced to the curriculum of the CEIT department, which produced its first graduates in 2002.

The Turkish Council of Higher Education is the governing body responsible for all higher education institutions throughout Turkey. The Council is an autonomous institution responsible for the planning, coordination, and governance of higher education in Turkey in accordance with the Turkish Constitution and the Higher Education Laws (Turkish Council of Higher Education, 2021).

In this context, a course related to computer hardware was added to the CEIT curriculum, the number of hours allocated to physics courses were reduced, and both biology and chemistry courses were removed altogether (Altun & Ateş, 2008). As a result of this update, the CEIT curriculum, to which students were enrolled between 2008 and 2018, consisted of 49 courses, of which seven were elective (Field Knowledge [FK]: one course; General Culture [GC]: two courses; and, Vocational Knowledge [VK]: four courses) delivered over five academic semesters, as well as 42 compulsory courses (FK: 22 courses; GC: nine courses; and, VK: 11 courses) delivered over eight academic semesters (Yükseköğretim Kurulu [Turkish Council of Higher Education], 2021a).

However, in 2018, the Turkish Council of Higher Education introduced wide-reaching changes across all teacher education degree program curricula (Yükseköğretim Kurulu [Turkish Council of Higher Education], 2021b). Since the PSTs who started their undergraduate education based on the curriculum updated in 2018 have yet to graduate, the current study uses data obtained from students who studied

under the previous curriculum (i.e., prior to the 2018 update). As such, the current study was conducted with data obtained from 258 PSTs (144 female and 114 male) who graduated between 2012 and 2020 from the CEIT program at a state university in Turkey. These 258 PSTs had therefore each completed all of the compulsory courses (n = 42) in the CEIT curriculum applicable at that time. The reason why elective courses in the curriculum were not added to the dataset used within the scope of this research is that PSTs were able to choose different alternatives in their elective courses. In this context, data from 15 of the elective courses taken by the participant PSTs were removed from the dataset prior to analysis. The final version of the dataset consisted of 10,836 (42x258) records, and includes information such as student number, course code, and the students' success scores for each course.

The general purpose of the research was to determine the association rules among the compulsory courses in the CEIT curriculum. Thus, the potential effect of students' success scores on one course on other courses was investigated. The data were therefore converted into a format suitable for the examination of association rules. For each student and for each course, the data in row format were converted into column format according to the basis of each course. The dataset obtained as a result of the data conversion process is shown in Table 1.

Table 1. *The dataset*

	FK Courses			GC Courses			VK Courses		
stdId	Feature1	…	Feature22	Feature1	…	Feature9	Feature1	…	Feature11
	411111	…	418121	411441	…	417413	411311	…	418328
std1	74		86	81		61	69		100
std2	63		79	67		86	58		100
std3	66		87	69		61	69		93
std4	67		89	65		86	62		100
std5	60		63	75		64	57		85
…	…		…	…		…	…		…
std257	70		89	73		82	78		100
std258	73		91	76		83	67		100

Creating a Model

Association rules, which were first addressed in the study of Agrawal et al. (1993), are one of the first methods used in data mining. Data mining methods that analyze the co-occurrence of events are called association rules (Agrawal & Srikant, 1994). Based on these association rules, it is decided which events will occur at the same time. The current study preferred the apriori algorithm, which is one of the most popular algorithms used in association rules mining in the literature (Parack et al., 2012).

The apriori algorithm has an iterative approach (Han et al., 2011) and is used to discover sets of items frequently mentioned in databases (Agrawal & Srikant, 1994). According to the algorithm, if the k-item set (item set with k elements) provides the minimum support value, its subsets will also provide the minimum support criteria (Agrawal & Srikant, 1994). Requirements of the apriori algorithm are as follows:

- The dataset to be used must have a tabular or transactional structure. Tabular data is structured as column-based, whereas transactional data is structured as row-based. Since the data in the current study was column-based, it was considered to be of tabular structure.
- The dataset should have a categorical structure. In the current study, PSTs' successes were scored according to the marking systems (0-100). These data were categorized according to the success scores of the PSTs on the basis of the course in question.
- The directions of the variables in the dataset should be defined as "in" (input), "out" (output), or "both," with each variable in the dataset expressed as only in, out, or both. The left side of the rule is

the antecedent, while the right is consequent. In the current study, the direction of the variables was determined as "both."

$$\left.\begin{matrix} X_{(in)} \Longrightarrow Y_{(out)} \\ Y_{(in)} \Longrightarrow X_{(out)} \end{matrix}\right\} \Longrightarrow X_{(both)} \Leftrightarrow Y_{(both)}$$

The apriori algorithm runs according to minimum support and minimum confidence parameters. The association rules among the elements are calculated with the support and confidence parameters. The greater the support and confidence parameters, the stronger the association rules. If the initial values of these parameters are too large, some rules may be overlooked; equally, if they are too small, it may lead to moving away from the desired hidden pattern.

The value of support indicates the rate of repetition of a relationship for all events, and is the percentage of combinations of the event items (Han et al., 2011). The value of support can be expressed according to the following formula:

$$Support(A \rightarrow B) = \frac{n(A \cup B)}{N}$$

A is "In" (Antecedent), B is "Out" (Consequent), and N: is the "Total number of events."

Confidence value determines the strength of relations among the events in association rules, and is the probability of Event B occurring if Event A occurs (Han et al., 2011). The value of confidence can be expressed according to the following formula:

$$Confidence(A \rightarrow B) = \frac{n(A \cup B)}{n(A)}$$

Finally, the lift value is a parameter examined where high confidence and support values exist, and is a coefficient obtained by dividing the confidence value by the support value (Han et al., 2011). The value of lift can be expressed according to following formula:

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(A \rightarrow B)}$$

EXPERIMENTS AND RESULTS

In the current study, Orange, an open-source data mining, machine learning, and data visualization software tool, was employed for the analysis (Demšar & Zupan, 2013). In the dataset used, the CEIT courses taken by the PSTs were determined as attributes. Each measurement includes data associated with a PST. Figure 1 illustrates the workflow of the model designed for the current study.

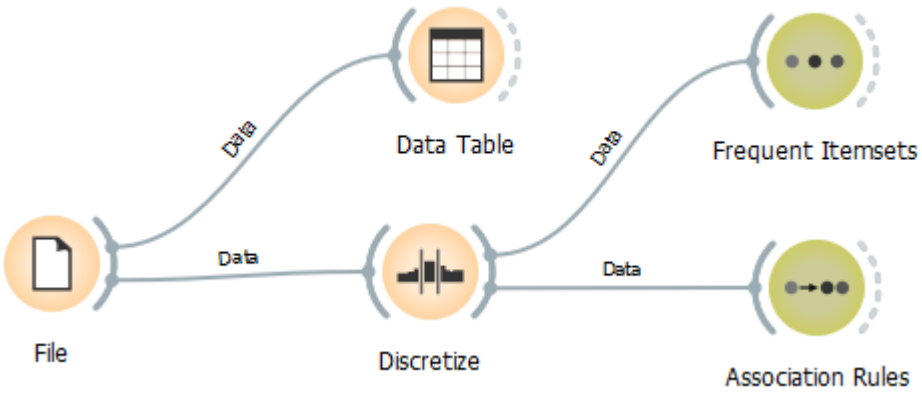

Figure 1. *Workflow of the developed model*

The basic parameters used in evaluating the model performance are support, confidence, and lift values. In the experimental process, different support and confidence values were tested, and the quality and number of association rules obtained were examined. Table 2 presents the numbers of rules obtained from different values of support and confidence parameters. These values are chosen in order to ensure that the rules are frequent and sufficiently meaningful to be taken into consideration.

Table 2. *Association rules*

Minimal support	0.10	Minimal confidence	Number of rules
		0.70	418
	FK	0.80	60
		0.85	11
		0.70	14
	GC	0.80	3
		0.85	1
		0.70	49
	VK	0.80	10
		0.85	6
Minimal confidence	0.65	Minimal support	Number of rules
		0.10	921
	FK	0.15	138
		0.20	17
		0.10	25
	GC	0.15	5
		0.20	0
		0.10	89
	VK	0.15	19
		0.20	7

According to Table 2, when the support parameter was kept constant as 0.10 and the confidence parameter values varied between 0.70 and 0.85, a minimum of 11 and a maximum of 418 rules were obtained in the FK courses category. When the confidence parameter value was kept constant at 0.65 and the support parameter values varied between 0.10 and 0.20, a minimum of 17 and a maximum of 921 rules were obtained in the FK courses category.

In the study, association rules with a minimum support value of 0.10 were determined to interpret the relations among the CEIT courses, as, for the association rules to be meaningful, at least 10% of the data must be similarly distributed. As a supplementary material, the association rules according to course categories are presented, with 60 association rules for the FK courses (s = 0.10, c = 0.80), 14 for the GC courses (s = 0.10, c = 0.70), and 49 association rules for the VK courses (s = 0.10, c = 0.70). Then, similar rules were filtered out where they were produced for the same course and with association rules in all three categories. Furthermore, association rules with similar characteristics were also removed. Example association rules with high support, confidence, and lift values are presented in Table 3a, 3b, and 3c.

Table 3a. *Association rules omitted where similar rules existed for same FK course/s (minimal support = 0.10, minimal confidence = 0.80)*

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.128	0.805	0.159	2.805	1.806	0.057	415112=50.00 - 66.67, 416122=50.00 - 66.67	→	414121=50.00 - 66.67
0.124	0.800	0.155	3.750	1.376	0.034	412122=50.00 - 66.67, 413111=50.00 - 66.67	→	413114=50.00 - 66.67
0.120	0.838	0.143	3.108	1.880	0.056	413111=50.00 - 66.67, 416123=49.00 - 65.33	→	414121=50.00 - 66.67

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.112	0.806	0.140	3.194	1.807	0.050	414124=50.00 - 66.67, 416123=49.00 - 65.33	→	414121=50.00 - 66.67
0.109	0.875	0.124	3.719	1.897	0.051	414121=≥ 83.33, 415111=≥ 83.33	→	417123=≥ 82.50
0.105	0.818	0.128	3.636	1.759	0.045	414121=50.00 - 66.67, 415111=66.67 - 83.33, 417123=66.00 - 82.50	→	415112=50.00 - 66.67
0.105	0.844	0.124	3.594	1.893	0.049	415112=50.00 - 66.67, 416123=49.00 - 65.33	→	414121=50.00 - 66.67
0.101	0.897	0.112	3.586	2.224	0.055	413112=63.33 - 79.17, 417123=66.00 - 82.50, 417111=86.67 - 93.33	→	418121=86.67 - 93.33

Table 3b. *Association rules omitted where similar rules existed for same GC course/s (minimal support = 0.10, minimal confidence = 0.70)*

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.112	0.725	0.155	3.650	1.281	0.025	412422=66.00 - 82.50, 412432=66.67 - 83.33, 416412=≥ 96.00	→	411441=62.00 - 77.50
0.112	0.725	0.155	3.300	1.417	0.033	412442=66.67 - 83.33, 412422=66.00 - 82.5	→	411421=66.00 - 82.50
0.109	0.700	0.155	3.650	1.237	0.021	411421=66.00 - 82.5, 412432=66.67 - 83.33, 416412=≥ 96.00	→	411441=62.00 - 77.50
0.101	0.812	0.124	4.562	1.436	0.031	411421=66.00 - 82.50, 412422=66.00 - 82.50, 416412=≥ 96.00	→	411441=62.00 - 77.50

Table 3c. *Association rules omitted where similar rules existed for same VK course/s (minimal support = 0.10, minimal confidence = 0.70)*

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.132	0.739	0.178	2.804	1.478	0.043	411310=65.33 - 81.67, 413312=≥ 81.67, 418328=≥ 97.00	→	418319=≥ 80.00
0.124	0.711	0.174	3.356	1.215	0.022	412311=63.33 - 79.17, 418319=64.00 – 80.00	→	414313=62.00 - 77.50
0.124	0.711	0.174	3.311	1.231	0.023	415314=81.67 - 89.33, 418319=≥ 80.00	→	418328=≥ 97.00
0.109	0.718	0.151	2.897	1.639	0.042	415314=74.00 - 81.67, 418319=64.00 – 80.00	→	417318=95.00 - 97.50
0.105	0.730	0.143	4.081	1.247	0.021	413312=65.33 - 81.67, 418319=≥ 80.00	→	414313=62.00 - 77.50
0.101	0.722	0.140	3.583	1.444	0.031	412311=47.50 - 63.33, 413312=≥ 81.67, 418328=≥ 97.00	→	418319=≥ 80.00

Since it is not possible to interpret all of these rules within the scope of the current study, one example has been interpreted as follows; however, other rules from the tables may also be similarly interpreted. According to Table 3a, 13% of the PSTs achieved scores ranging from 50.00 to 66.67 from the courses coded as both 415112 and 416122, and achieved scores ranging from 50.00 to 66.67 from the course coded as 414121. In addition, PSTs who achieved scores ranging from 50.00 to 66.67 from courses coded as both 415112 and 416122 showed an 80% probability of receiving scores ranging from 50.00 to 66.67 from the course coded as 414121. According to this finding, it may be stated that a PST's grade point average for courses coded as 415112 and 416122 can be used as a predictor of the grade point average for the course coded as 414121.

According to Table 3a, a strong relationship was found to exist between the software-related courses taught as part of the CEIT curriculum; for example, between "Programming Languages I" and "Programming Languages II," between "Internet-Based Programming" and "Programming Languages I/II," and between "Programming Languages I/II," and "Database Management Systems." Accordingly, it may be said that the PSTs' academic performance in the Programming Languages courses

was shown to be a predictor of their academic performance in the Database Management Systems course. However, a strong relationship was observed to exist between the Mathematics and Physics courses and the Programming Languages course. According to this finding, it may be stated that a PSTs's academic performance in courses requiring numerical intelligence such as Mathematics may affect their academic performance in software courses such as Programming Languages.

DISCUSSION AND CONCLUSION

Student information systems (SIS), in which the various forms of student information are recorded (e.g., ID, courses taken, grades, etc.) in universities and similar educational institutions, not only make things easier but also contribute to the generation of huge amounts of data. For example, Proliz Software is the most preferred SIS software in Turkey (used in more than 90 higher education institutions) and provides facilities to manage the data of more than four million students in total.

However, there are hidden patterns within this big data, hence it is necessary to attempt to obtain valuable/meaningful information from such large amounts of data in order to increase the quality of teaching and thereby to improve student learning. This outcome can be made possible through the application of various EDM methods. In this context, the current study applied the apriori algorithm to reveal the association rules among the compulsory courses of Turkey's CEIT curriculum. In other words, the association rules among the courses were investigated in order to reveal the factors that could cause high or low success scores.

According to the experimental results, a high level of positive correlation was found to exist among some courses in the current study. Also, according to the PSTs' performance in courses they had previously attended, their performance in some future courses may be estimated. In other words, the academic performance of PSTs in some courses can be used as a predictor of their academic performance in other courses. In this way, PSTs with a likelihood to fail can be identified in advance and appropriate supportive activities planned accordingly. The results also showed that the most association rules were found among the FK courses, VK courses, and GC courses, respectively.

In this study, the association rules between the courses taught in an undergraduate program were examined using the apriori algorithm. When the related literature is examined, a wide variety of studies using the apriori algorithm can be seen. For instance, Moubayed et al. (2018) examined the relationship between students' academic performances and their lesson participation with the apriori algorithm. Additionally, in a study conducted by Ko and Leu (2021), the relationship between students' self-efficacy beliefs and their academic performance was examined using the apriori algorithm, whilst Ougiaroglou and Paschalis (2012) examined the relationship among students' interest in lessons using association rules.

In the current study, a strong relationship was found between the "Mathematics II" course taught in the second semester and the "Programming Languages I" course taught in the third semester. Also, a strong relationship was found between the "Operating Systems and Applications" course in the fifth semester and the "Computer Networks and Communication" courses in the sixth semester. Similarly, Wu and Zeng (2019) found a strong relationship between a Mathematics and Programming Languages course and hardware-related courses and an Operating System and Compiling Principles course.

Furthermore, in the current research, it was observed that some courses had no association, and that the students' GPA from some courses were unable to be explained according to their GPA from other courses taken as part of the CEIT curriculum. In other words, some courses are not related to any of the other courses. This result shows that some courses taught in the curriculum were completely independent of each other. Similar results were also obtained in a limited number of studies in which the association rules between large numbers of courses were attempted to be determined (Soimart & Mookdarsanit, 2016; Wu & Zeng, 2019).

In a study conducted by Soimart and Mookdarsanit (2016), which had its focus on discipline recommendations for high-school students' future choices, 3,000 samples were studied, whilst the current study's dataset contained 10,836 undergraduate students' academic records. In other words, it can be said that the current study was conducted with a larger sample. Furthermore, in research conducted by Wu and Zeng (2019), their study attempted to determine the rules of association according to data from 34 courses taken by 100 students, whilst the current study examined data obtained from 42 courses taken by 258 PSTs. Therefore, it can be said that the dataset used in the current research was deemed to be quite large. In addition, in this study, the success points obtained in the 100-point system were categorized and analyzed. While some higher education institutions have a preference for marking-based systems, others may prefer grade-based systems (e.g., A, B, C, etc.); hence, the successes of students in marking-based systems were first converted into equivalent grading-based data.

Finally, with the association rules obtained in the current study, the academic performance of courses due to be taken forthcoming semesters can be estimated according to the academic performance in courses previous taken. On the other hand, Kumar and Chadha (2012) found beneficial relationships between graduate students' academic and their former success at the undergraduate level based on association rules. Moreover, using the current study's findings, as stated by Yang and Hu (2011), certain prerequisite courses can be determined. Furthermore, educational activities can be improved by rearranging relevant courses within the semesters. The methods and results of the current study are expected to become a point of reference, especially in curriculum improvement studies in higher education.

The current study's results do, however, present certain gaps which readers should consider. For instance, the data were obtained from 258 CEIT PSTs; therefore, larger samples or /datasets may help to reveal more effective and different association rules. Furthermore, in the current study, only the association rules for compulsory courses were attempted to be determined. Therefore, the association rules for elective courses taken by students who continued their education in different departments could also be determined in future studies.

Another point that should be mentioned here is that the courses in Turkey's CEIT curriculum were given by different instructors and within different academic semesters and years. Of course, different instructors may also have an effect on the students' grades, whilst students who started their educational life in different years will have entered university with different score levels. These and similar reasons may have had some effect on the association rules within the scope of the current research, which was studied based on data obtained for the 9-year period from 2012 to 2020.

REFERENCES

Abinowi, E., & Aminudin. (2020). Analysis of Instagram posting for marketing using apriori method. *Palarch's Journal of Archaeology of Egypt/Egyptology*, *17*(10), 3094-3101. https://archives.palarch.nl/index.php/jae/article/view/5445

Acharya, S., & Madhu, N. (2012). Discovery of students' academic patterns using data mining techniques. *International Journal on Computer Science and Engineering (IJCSE)*, *4*(06), 1054-1062.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD, 22*(2), 207-216. https://doi.org/10.1145/170035.170072

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487-499). Kaufmann.

Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, *16*, Article 40. https://doi.org/10.1186/s41239-019-0172-z

Alkatheri, S., Abbas, S. A., & Siddiqui, M. A. (2019). A comparative study of big data frameworks. *International Journal of Computer Science and Information Security (IJCSIS), 17*(1), 66-73. https://doi.org/10.5539/mas.v13n7p1

Altun, E., & Ateş, A. (2008). The problems and future concerns of computer and instructional technologies preservice teachers. *Elementary Education Online, 7*(3), 680-692. https://www.ilkogretim-online.org/fulltext/218-1596683177.pdf?1619079787

Baaziz, A., & Quoniam, L. (2013). How to use big data technologies to optimize operations in upstream petroleum industry. *International Journal of Innovation, 1*(1), 30-42. https://doi.org/10.5585/iji.v1i1.4

Baker, R. S. J. d. (2007). Is gaming the system state-or-trait? Educational data mining through the multi-contextual application of a validated behavioral model. In *Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007* (Vol. 2007, pp. 76-80). User Modeling. https://educationaldatamining.org/EDM_ORG/wp-content/uploads/2020/05/DM.UM07_proceedings_full.pdf

Baker, R. S. J. d. (2011). Data mining for education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International Encyclopedia of Education* (3rd ed., Vol. 7, pp. 112-118.). Elsevier.

Baker, R. S. J. d., & de Carvalho, A. M. J. A. (2008). Labeling student behavior faster and more precisely with text replays. *Proceedings of the First International Conference on Educational Data Mining*, (pp. 38-47).

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009 : A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-16. https://doi.org/10.5281/zenodo.3554657

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*, 537-553. https://doi.org/10.1007/s10639-017-9616-z

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. U.S. Department of Education, Office of Educational Technology. https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Bozkurt, A. (2016). Öğrenme analitiği: E-öğrenme, büyük veri ve bireyselleştirilmiş öğrenme [Learning analytics: E-learning, big data and personalized learning]. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *2*(4), 55-81. https://dergipark.org.tr/en/pub/auad/issue/34066/377071

Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. *ACM SIGKDD Explorations Newsletter*, *13*(2), 3-6. https://doi.org/10.1145/2207243.2207245

Calvet Liñán, L., & Juan Pérez, Á. A. (2015). Educational data mining and learning analytics: Differences, similarities, and time evolution. *RUSC. Universities and Knowledge Society Journal*, *12*(3), 98-112. https://doi.org/10.7238/rusc.v12i3.2515

Cil, I. (2012). Consumption universes based supermarket layout through association rule mining and multidimensional scaling. *Expert Systems with Applications*, *39*(10), 8611-8625. https://doi.org/10.1016/j.eswa.2012.01.192

Daniel, B. (2015). Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, *46*(5), 904-920. https://doi.org/10.1111/bjet.12230

Demšar, J., & Zupan, B. (2013). Orange: Data mining fruitful and fun - A historical perspective. *Informatica*, *37*(1), 55-60. http://www.informatica.si/ojs-2.4.3/index.php/informatica/article/viewFile/434/438

Dongre, J., Prajapati, G. L., & Tokekar, S. V. (2014). The role of apriori algorithm for finding the association rules in data mining. In A. Sharma, A. Ahlawat, A. Pandey, & V. Sharma (Eds.), *International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)* (pp. 657-660). IEEE. https://doi.org/10.1109/ICICICT.2014.6781357

Dunham, M. H. (2003). *Data mining introductory and advanced topics*. Pearson.

Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, *5*(2), 112-116. https://doi.org/10.7763/ijiee.2015.v5.513

Education Data. (2021). *Online education statistics*. https://educationdata.org/online-education-statistics.

Educational Data Mining Society. (2021). *Educational Data Mining.* https://educationaldatamining.org/

Elias, T. (2011). *Learning analytics: Definitions, processes and potential*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.7092&rep=rep1&type=pdf.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. V. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94*, 335-343. https://doi.org/10.1016/j.jbusres.2018.02.012

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130-160. https://doi.org/10.3102/0091732X20903304

Han, J., Kamber, M., & Pei, J. (2011). *Data mining : Concepts and techniques*. Kaufman.

Hung, J.-L, & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*, *4*(4), 426-436. https://jolt.merlot.org/vol4no4/hung_1208.pdf

Hussain, S., Atallah, R., Kamsin, A., & Hazarika, J. (2019). Classification, clustering and association rule mining in educational datasets using data mining tools: A case study. *Advances in Intelligent Systems and Computing*, *765*, 196-211. https://doi.org/10.1007/978-3-319-91192-2_21

International Telecommunication Union. (2016). *ICT facts and figures 2016*. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf

Jabbar, A., Akhtar, P., & Dani, S. (2020). Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Industrial Marketing Management*, *90*, 558-569. https://doi.org/10.1016/j.indmarman.2019.09.001

Jha, J., & Ragha, L. (2013). Educational data mining using improved apriori algorithm. *International Journal of Information and Computation Technology*, *3*(5), 411-418. https://www.ripublication.com/irph/ijict_spl/08_ijictv3n5spl.pdf

Jhang, K.-M., Chang, M.-C., Lo, T.-Y., Lin, C.-W., Wang, W.-F., & Wu, H.-H. (2019). Using the apriori algorithm to classify the care needs of patients with different types of dementia. *Patient Preference and Adherence*, *13*, 1899-1912. https://doi.org/10.2147/PPA.S223816

Joshi, A., & Sodhi, J. S. (2014). Target advertising via association rule mining. *International Journal of Advance Research in Computer Science and Management Studies*, *2*(5), 256-261. http://www.ijarcsms.com/docs/paper/volume2/issue5/V2I5-0066.pdf

Kılınç, Ç. (2015). *Üniversite öğrenci başarısı üzerine etki eden faktörlerin veri madenciliği yöntemleri ile incelenmesi [Examining the effects on university student success by data mining techniques]* [Master's Thesis]. Eskişehir Osmangazi University, Turkey. http://hdl.handle.net/11684/1256

Ko, C.-Y., & Leu, F.-Y. (2021). Examining successful attributes for undergraduate students by applying machine learning techniques. *IEEE Transactions on Education*, *64*(1), 50-57. https://doi.org/10.1109/TE.2020.3004596

Kumar, V., & Chadha, A. (2012). Mining association rules in students assessment data. *International Journal of Computer Science Issues*, *9*(5), 211-216. http://ijcsi.org/articles/Mining-association-rules-in-students-assessment-data.php

Lioutas, E. D., & Charatsari, C. (2020). Big data in agriculture: Does the new oil lead to sustainability? *Geoforum*, *109*, 1–3. https://doi.org/10.1016/j.geoforum.2019.12.019

Long, P., Siemens, G., Conole, G., & Gašević, D. (2011). *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. ACM. https://dl.acm.org/doi/proceedings/10.1145/2090116

Mabe, M. (2003). The growth and number of journals. *Serials*, *16*(2), 191-197. https://serials.uksg.org/articles/10.1629/16191/galley/729/download/

Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2018). Relationship between student engagement and performance in e-learning environment using association rules. In *IEEE World Engineering Education Conference (EDUNINE)*. IEEE. https://doi.org/10.1109/EDUNINE.2018.8451005

Natalia, D., & Salvatore, L. (2020). Apriori algorithm for association rules mining in aircraft runway excursions. *Civil Engineering and Architecture*, *8*(3), 206-217. https://doi.org/10.13189/cea.2020.080303

Ougiaroglou S., & Paschalis G. (2012). Association rules mining from the educational data of ESOG web-based application. In L. Iliadis, I. Maglogiannis, H. Papadopoulos, K. Karatzas, & S. Sioutas (Eds.), *Artificial Intelligence Applications and Innovations. AIAI 2012. IFIP Advances in Information and Communication Technology* (Vol 382, pp. 105-114). Springer. https://doi.org/10.1007/978-3-642-33412-2_11

Parack, S., Zahid, Z., & Merchant, F. (2012). Application of data mining in educational databases for predicting academic trends and patterns. In *IEEE International Conference on Technology Enhanced Education (ICTEE)* (paper 17). IEEE. https://doi.org/10.1109/ICTEE.2012.6208617

Pardo, A., & Teasley, S. (2014). Learning analytics research, theory and practice: Widening the discipline. *Journal of Learning Analytics*, *1*(3), 4-6. https://doi.org/10.18608/jla.2014.13.2

Putra, P. B. I. S. P., Suryani, N. P. S. M., & Aryani, S. (2018). Analysis of apriori algorithm on sales transactions to arrange placement of goods on minimarket. *International Journal of Engineering and Emerging Technology*, *3*(1), 13-17. https://ocs.unud.ac.id/index.php/ijeet/article/download/41250/25102

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics—Part C: Applications and Reviews*, *40*(6), 601-618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12-27. https://doi.org/10.1002/widm.1075

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), Article e1355. https://doi.org/10.1002/widm.1355

Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: Challenges and promises of big data in healthcare. *Nature Medicine*, *26*, 29-38. https://doi.org/10.1038/s41591-019-0727-5

Shweta, M., & Garg, K. (2013). Mining efficient association rules through apriori algorithm using attributes and comparative analysis of various association rule algorithms. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(6), 306-312. http://ijarcsse.com/Before_August_2017/docs/papers/Volume_3/6_June2013/V3I6-0192.pdf

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380-1400. https://doi.org/10.1177/0002764213498851

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In S. B. Shum, D. Gasevic, & R. Ferguson (Eds.), *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252-254). ACM. http://dx.doi.org/10.1145/2330601.2330661

Slater, S., Joksimovic, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85-106. https://doi.org/10.3102/1076998616666808

Soimart, L., & Mookdarsanit, P. (2016, September 22-23). *An admission recommendation of high-school students using apriori algorithm* [Conference presentation]. 6th International Conference on Sciences and Social Sciences, Mahasarakham, Thailand.

Tan, H. R., Chng, W. H., Chonardo, C., Ng, M. T. T., & Fung, F. M. (2020). How chemists achieve active learning online during the COVID-19 pandemic: Using the community of inquiry (CoI) framework to support remote teaching. *Journal of Chemical Education*, *97*(9), 2512-2518. https://doi.org/10.1021/acs.jchemed.0c00541

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, *47*, 157-167. https://doi.org/10.1016/j.chb.2014.05.038

Turkish Council of Higher Education. (2021). *Higher Education System in Turkey*. https://www.yok.gov.tr/en/institutional/higher-education-system

Wu, X., & Zeng, Y. (2019). Using apriori algorithm on students' performance data for association rules mining. *Advances in Social Science, Education and Humanities Research, 322*, 403-406. https://dx.doi.org/10.2991/iserss-19.2019.105

Yang, Q., & Hu, Y. (2011). Application of improved apriori algorithm on educational information. In J. Watada, P.-C. Chung, J.-M. Lin, C.-S. Shieh, & J.-S. Pan (Eds.), *5th International Conference on Genetic and Evolutionary Computing* (pp. 330-332). IEEE. https://doi.org/10.1109/ICGEC.2011.82

Yin, S., & Kaynak, O. (2015). Big data for modern industry: Challenges and trends. *Proceedings of the IEEE, 103*(2), 143-146. https://doi.org/10.1109/JPROC.2015.2388958

Yüksekoğretim Kurulu. (2021a). *Eğitim Fakültesi Öğretmen Yetiştirme Lisans Programlari [Faculty of Education Teacher Education Undergraduate Programs]*. https://www.yok.gov.tr/Documents/Yayinlar/Yayinlarimiz/egitim-fakultesi-ogretmen-yetistirme-lisans-programlari.pdf

Yüksekoğretim Kurulu. (2021b). *Programların Güncelleme Gerekçeleri, Getirdiği Yenilikler ve Uygulama Esasları [New Teacher Education Undergraduate Programs]*. https://www.yok.gov.tr/kurumsal/idari-birimler/egitim-ogretim-dairesi/yeni-ogretmen-yetistirme-lisans-programlari

Zaiane, O. R. (2002). Building a recommender agent for e-learning systems. In *Proceedings of the International Conference on Computers in Education* (pp. 55-59). IEEE. https://doi.org/10.1109/CIE.2002.1185862