2022

4(1), 110-127

Eğitsel Veri Madenciliği: Adayların Beden Eğitimi ve Spor Eğitimi Programına Yerleşme Durumlarının Tahmini

Mustafa Yağcı[1], Yusuf Ziya Olpak[*2], Kağan Gül[3], Sıdıka Seda Olpak[4]

Anahtar Sözcükler
Eğitsel veri madenciliği
Büyük veri
Özel yetenek sınavı

Makale Hakkında
Gönderim Tarihi
17 Mayıs 2022
Kabul Tarihi
13 Haziran 2022
Yayın Tarihi
29 Haziran 2022
Makale Türü
Araştırma Makalesi

Öz

Eğitsel veri madenciliğinin temel amacı, eğitimle ilgili konularda karar vermeyi desteklemek için eğitim verilerinden faydalı bilgiler çıkarmaktır. Eğitsel veri madenciliğinde en çok tercih edilen yöntemlerden biri de tahmindir. Mevcut çalışmanın birincil amacı, adayların Beden Eğitimi ve Spor Eğitimi programına kabul edilip edilmeyeceklerini farklı algoritmalar kullanarak tahmin etmektir. Bu araştırma kapsamında 2016-2020 yılları arasında Türkiye'de bir devlet üniversitesinin Beden Eğitimi ve Spor Eğitimi programına katılmak için başvuran 1.671 adaydan elde edilen verilerle çalışılmıştır. Random Forest, kNN, SVM, Logistic Regression ve Naïve Bayes algoritmalarının her biri, bir adayın ilgili programına kabul edilip edilmeyeceğini tahmin etmek için kullanılmıştır. Elde edilen bulgulara göre algoritmaların sınıflandırma doğruluğu en yüksekten en düşüğe doğru sırasıyla; Random Forest (.985), SVM (.845), kNN (.818), Naïve Bayes (.815) ve Logistic Regression (.701) şeklindedir. Başka bir deyişle, Random Forest algoritmasının, örnekleri neredeyse tam olarak doğru bir şekilde sınıflandırdığı bulunmuştur. Bu bağlamda, çalışmadan elde edilen diğer bulgular ayrıntılı olarak tartışılmış ve gelecek araştırmalar için önerilerde bulunulmuştur.

Educational Data Mining: Predicting Candidates' Placement Status in Physical Education and Sports Education Program

Keywords
Educational data mining
Big data
Special talent exam

Article Info
Received
May 17, 2022
Accepted
June 13, 2022
Published
June 29, 2022
Article Type
Research Paper

Abstract

Educational data mining's primary purpose being to extract useful information from educational data in order to support decision-making on educational issues. One of the most preferred methods in educational data mining is prediction. The primary purpose of the current study is to predict whether or not candidates will be admitted into the PESE program according to different algorithms. Within the scope of this research, data was obtained from 1,671 candidates who applied to join the PESE program of a state university in Turkey between 2016 and 2020 were studied. The Random Forest, kNN, SVM, Logistic Regression, and Naïve Bayes algorithms were each used to predict whether or not a candidate could admit to the PESE program. According to the findings, the algorithms' classification accuracy from highest to lowest is Random Forest (.985), SVM (.845), kNN (.818), Naïve Bayes (.815), and Logistic Regression (.701), respectively. In other words, the Random Forest algorithm is shown to have correctly classified the instances almost exactly. Other findings from the study are discussed in detail, and suggestions put forth for future research.

Atıf: Yağcı, M., Olpak, Y. Z., Gül, K., & Olpak, S. S. (2022). Eğitsel veri madenciliği: Adayların beden eğitimi ve spor eğitimi programına yerleşme durumlarının tahmini. *Bilgi ve İletişim Teknolojileri Dergisi, 4*(1), 110-127. https://doi.org/10.53694/bited.1118025

Cite: Yagci, M., Olpak, Y. Z., Gul, K., & Olpak, S. S. (2022). Educational data mining: Predicting candidates' placement status in physical education and sports education program. *Journal of Information and Communication Technologies, 4*(1), 110-127. https://doi.org/10.53694/bited.1118025

[*] **Sorumlu Yazar/Corresponding Author**: yusufziyaolpak@gmail.com
[1] Assoc. Prof. Dr., Kırşehir Ahi Evran University, Faculty of Engineering and Architecture, Kırşehir/Türkiye, mustafayagci06@gmail.com, http://orcid.org/0000-0003-2911-3909
[2] Assoc. Prof. Dr., Kırşehir Ahi Evran University, Faculty of Education, Kırşehir/Türkiye, yusufziyaolpak@gmail.com, http://orcid.org/0000-0001-5092-252X
[3] PhD Candidate, Kırşehir Ahi Evran University, Çiçekdağı Vocational School, Kırşehir/Türkiye, kagan.gul@ahievran.edu.tr, http://orcid.org/0000-0001-5959-8199
[4] M.Sc. Student, Kırşehir Ahi Evran University, Institute of Educational Sciences, Kırşehir/Türkiye, sedaolpak@gmail.com, https://orcid.org/0000-0001-7557-8227

Introduction

In line with the latest developments in technology, there is a continuous increase in the speed of data production as well as the volume of data generated on a daily basis. Meaningful information, patterns, and trends can be revealed from careful analysis of this data, which is then increasingly being used in many different areas of daily life (e.g., Agrawal & Prabakaran, 2020; Hallikainen et al., 2020; Line et al., 2020), and is referred to as big data (De Mauro et al., 2015). It may be said that one area considerably involved is education, as this data allows students, teachers, administrators, and other stakeholders to undertake new research and to explore new avenues previously unknown. Since each stakeholder group has its own mission and goals, educational information often needs to be analyzed from various perspectives (Hanna, 2004).

As an emerging discipline, educational data mining (EDM) was defined by Romero and Ventura (2010, p. 601) as a means to "exploits statistical, machine-learning, and data-mining algorithms over the different types of educational data" with its primary purpose being to extract useful information obtained from educational data in order to reinforce decision-making in educational studies (Calvet Liñán & Juan Pérez, 2015). There are various methods employed in the field of EDM to achieve this, and which have been classified in different ways by different researchers (Baker, 2011; Baker & Yacef, 2009; Bakhshinategh et al., 2018; Romero & Ventura, 2013). For instance, in a study by Bakhshinategh et al. (2018, p. 540), the most commonly used methods were specified as "classification and regression, clustering, association rule mining, discovery with models, outlier detection, social network analysis, text mining, sequential pattern mining, and visualization technique". For this reason, it should be considered that each method used for EDM has potentially different features that should be taken into consideration in its usage.

One of the most preferred methods in EDM is prediction (e.g., Aouifi et al., 2021; Sokkhey & Okazaki, 2020; Saa, 2016). Baker (2011), with research aimed at developing a model that can infer a single aspect of any data from a combination of various other aspects of the data using prediction methods such as classification (e.g., Karthikeyan et al., 2020; Kılıç Depren et al., 2017), regression (e.g., Karlos et al., 2020; Yulia, 2020), or density estimation (e.g., Gerber, 2014; Hernandez-Suarez et al., 2019). One of the important goals of today's educational systems is to maximize academic results, and to improve the quality of the educational experience by minimizing the failure rates of students. In this context, by employing prediction methods, learners' academic performance (Waheed et al., 2020) and their pass/fail status for a particular course (Abut et al., 2018) can be predicted, and early warning systems developed to qualify at-risk students (Akçapınar et al., 2019). These same methods can also be used to make early predictions about dropouts (Márquez-Vera et al., 2016).

Furthermore, by evaluating candidates according to different criteria, it may be predicted in advance whether or not students will be admitted to schools or other educational programs, or be accepted for employment, etc. (e.g., Acikkar & Akay, 2009). For instance, students can be admitted to certain higher education programs in Turkey based on special talent exams (STEs) organized directly by the higher education institutions themselves. As an example, numerous criteria (see Table 3) are taken into account in determining whether or not students may be enrolled to the Physical Education and Sports Education (PESE) program. The primary aim of the current research is to predict whether or not candidates will be admitted into the PESE program according to different algorithms. Thus, candidates who are or are considering applying to the PESE, or similar programs that accept students by

way of a STE, may be shown predictions of what they may encounter according to different scores and criteria used at the admission stage.

Moreover, a website could be developed where the candidates could enter their actual or anticipated scores based on the criteria used in the evaluation, and which could then be used to predict whether or not they would likely be admitted to the relevant program. Thus, by using the quota of the relevant program according to gender as a variable, candidates can gain a better idea about their likely admittance according to different scores and different criteria.

Methodology

In Turkey, students can be admitted to higher education institutions either by central placement or by STEs organized directly by higher education institutions themselves, taking into account the students' scores obtained from centralized national university entrance examinations conducted by the Measuring, Selection and Placement Center [Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM)]. The ÖSYM is a public institution that provides examination services to more than 10 million higher education candidates nationally each year in Turkey, and has both administrative and financial autonomy to conduct the processes of measuring, selection, and placement to international standards.

In order for students to be eligible to enroll to a higher education institution, they must first take the Higher Education Institutions Exam (HEIE), which is a centralized assessment that is administered once annually by the ÖSYM. The HEIE, in which thousands of candidates participate every year, is conducted in three different sessions/parts: Basic Proficiency Test (BPT), Field Proficiency Test (FPT), and Foreign Language Test (FLT). For example, according to 2020 data; 2.296.138 out of 2.424.718 candidates who applied to BPT, 1.672.376 out of 1.788.590 candidates who applied to FPT, and 105.579 out of 128.177 candidates who applied to FLT took the exam (ÖSYM, 2021).

Some departments in Turkey's higher education institutions accept students through a STE (e.g., coaching, sports management, physical education and sports education, etc.), and also in some academic units (e.g., conservatories, fine arts, schools of physical education, and sports, etc.). One example of this is the Physical Education and Sports Education (PESE) program, which is offered by some Sports Sciences Faculties or Physical Education and Sports Schools. Students who graduate from the PESE program can then apply to work as physical education teachers in either public or private schools in Turkey, and may also coach students at summer/winter sports camps during their spare time.

In this study, data was obtained from candidates who applied to join a PESE program, which is an example of programs in which students may be accepted through a STE at state universities in Turkey. In order for candidates to apply to the PESE program, they must first achieve a minimum of 180 points from the BPT or be in the top 800.000 candidates that year. Furthermore, candidates must then successfully complete the nine-stage STE (see Figure 1) as prepared by an authorized board of the relevant institution in order to determine their physical proficiency through activities of various sports branches. The scores of the candidates from the STE are determined according to the time taken to successfully complete the parkour (see Table 1).

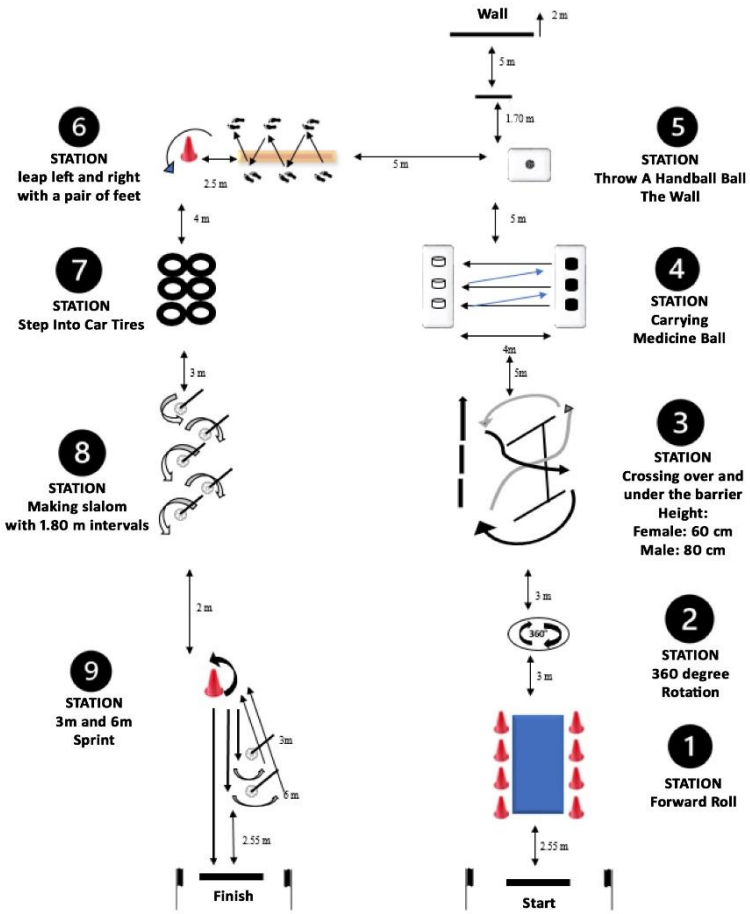

Figure 1. Special Talent Exam Parkour

Table 1. STE Scoring Table

Male		Female	
Time (seconds)	Points	Time (seconds)	Points
…-44.00	100	…-50.50	100
44.015-44.515	95	50.515-51.015	95
44.525-45.025	90	51.025-51.525	90
45.035-45.535	85	51.535-52.035	85
45.545-46.045	80	52.045-52.545	80
46.055-46.555	75	52.555-53.055	75
46.565-47.065	70	53.065-53.565	70
47.075-47.575	65	53.575-54.075	65
47.585-48.085	60	54.085-54.585	60
48.095-48.595	55	54.595-55.095	55
48.605-49.105	50	55.105-55.605	50
49.115-49.615	45	55.615-56.115	45
49.625-50.125	40	56.125-56.625	40
50.135-50.635	35	56.635-57.135	35
50.645-51.145	30	57.145-57.645	30
51.155-51.655	25	57.655-58.155	25
51.665-52.165	20	58.165-58.665	20

52.175-52.675	15	58.675-59.175	15
52.685-53.185	10	59.185-59.685	10
53.195-53.695	5	59.695-60.195	5
53.70-…	0	60.20-…	0

The scope of the STEs, their evaluation, and the student placement procedures are each self-regulated by the relevant higher education institutions themselves, and are thereby independent of the centralized ÖSYM. In other words, the STEs can be held by different higher education institutions using varying methods in order to determine a suitable form of examination that fits their institution's needs and infrastructure. While calculating the candidates' overall score for admittance to a PESE program, the BPT score, weighted high school grade point average (GPA) score, the placement status of any higher education program from the previous year, whether or not they graduated from a sports high school, and the score they achieved in their STE are all used. The data was obtained from 1.671 candidates who applied to join the PESE program of a state university in Turkey between 2016 and 2020, and that data was then subjected to analysis (see Table 2).

Table 2. Distributions of Candidates According to Application Year

Year	Gender	Average age (years)	Number of candidates	Quota
2016	Total	19	690	50
	Female	19	167	20
	Male	20	523	30
2017	Total	20	267	50
	Female	20	70	20
	Male	20	197	30
2018	Total	19	244	55
	Female	19	61	20
	Male	19	183	35
2019	Total	19	237	40
	Female	19	56	15
	Male	19	181	25
2020	Total	19	233	40
	Female	19	60	15
	Male	20	173	25
Total		19	1.671	235

Table 2 details the admittance quotas for both male and female candidates for the PESE program at the relevant higher education institutions between 2016 and 2020. Also, as indicated in Table 2, it can be seen that whilst there are decreasing candidate numbers each year, the numbers of male candidates consistently exceed those of female candidates for the PESE program.

Results

In the current study, Orange software was used to perform predictions according to various algorithms (Padmavaty et al., 2020). In this context, the Random Forest, Support Vector Machines (SVM), Logistic Regression, k-Nearest

Neighbors (kNN), and Naïve Bayes algorithms were each used to predict whether or not a candidate could admit to the PESE program. Figure 2 illustrates the details of the workflow of each of the developed models within the scope of this research.

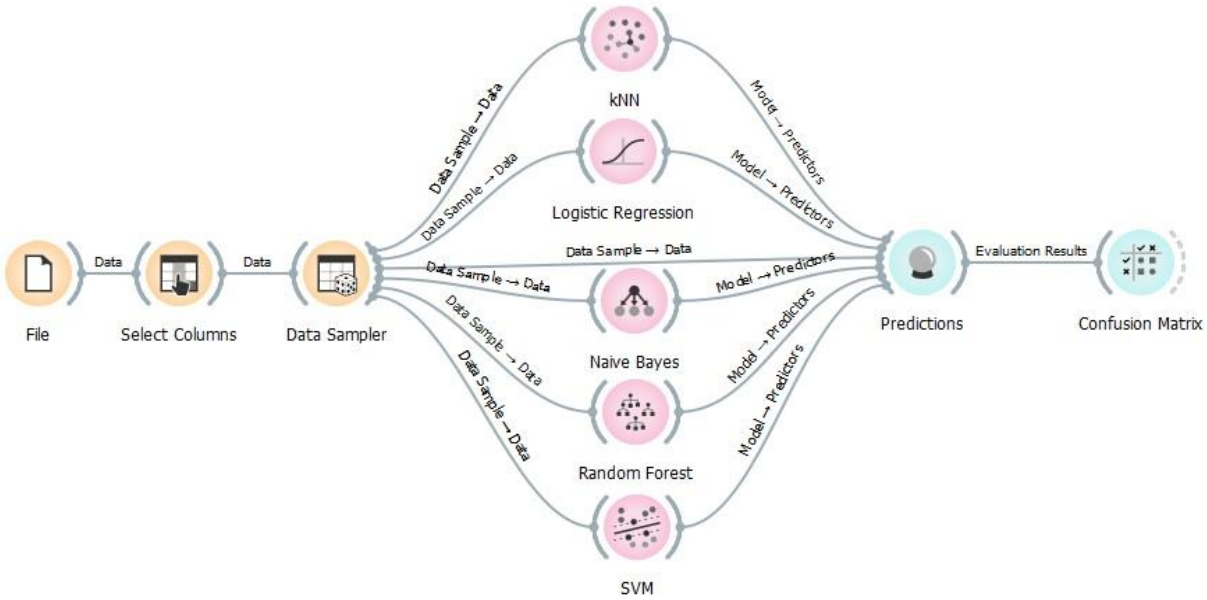

Figure 2. Workflow of Developed Models

In the models, to provide an assurance that the results obtained are both valid and may be generalized to make new predictions, the dataset (1.671 records) was partitioned into training (1.503 records) and test (168 records) sets via 10-fold cross-validation. Details regarding the variables for each of the models are presented in Table 3.

Table 3. Variables in the Models

Features	Target	Meta
- Application year (2016-2020) - Age (years) - Gender (Female; Male) - Placement status for previous year's higher education program (Placement; No placement) - High school specialization (Sports high school; Other high school) - BPT score (Centralized higher education entrance exam) - Weighted high school GPA score (Graduation grade calculated from an average of each annual GPA during high school education. Multiplying these by 5/100 gives the weighted high school GPA, which can be calculated as the lowest 250 and the highest 500) - STE score (From candidates' STE parkour)	Candidates' winning status - Winner - Reserve winner - Loser	CandidateID

Figure 3 provides an example of the analysis results used to predict the candidates' winning status. The values in the "Status" column are the actual values, whereas the values in the other columns are those pertaining to those predicted by each model (Random Forest, kNN, SVM, Logistic Regression, and Naïve Bayes).

	Random Forest	kNN	SVM	Logistic Regression	Naive Bayes	Status
1	1.00 : 0.00 : 0.00 → Loser	1.00 : 0.00 : 0.00 → Loser	0.62 : 0.34 : 0.04 → Loser	0.90 : 0.09 : 0.01 → Loser	0.99 : 0.01 : 0.00 → Loser	Loser
2	0.07 : 0.93 : 0.00 → Reserve Winner	0.20 : 0.80 : 0.00 → Reserve Winner	0.33 : 0.54 : 0.14 → Reserve Winner	0.90 : 0.10 : 0.00 → Loser	0.71 : 0.28 : 0.01 → Loser	Reserve Winner
3	1.00 : 0.00 : 0.00 → Loser	0.80 : 0.20 : 0.00 → Loser	0.34 : 0.31 : 0.35 → Loser	0.81 : 0.12 : 0.07 → Loser	0.97 : 0.01 : 0.02 → Loser	Loser
4	0.00 : 0.17 : 0.82 → Winner	0.40 : 0.20 : 0.40 → Loser	0.00 : 0.27 : 0.73 → Winner	0.49 : 0.34 : 0.17 → Loser	0.19 : 0.61 : 0.21 → Reserve Winner	Winner
5	0.00 : 0.00 : 1.00 → Winner	0.00 : 0.40 : 0.60 → Winner	0.00 : 0.14 : 0.86 → Winner	0.45 : 0.32 : 0.23 → Loser	0.08 : 0.45 : 0.47 → Winner	Winner
6	0.13 : 0.87 : 0.00 → Reserve Winner	0.80 : 0.20 : 0.00 → Loser	0.42 : 0.53 : 0.05 → Reserve Winner	0.80 : 0.15 : 0.05 → Loser	0.14 : 0.64 : 0.22 → Reserve Winner	Reserve Winner
7	1.00 : 0.00 : 0.00 → Loser	1.00 : 0.00 : 0.00 → Loser	0.65 : 0.32 : 0.03 → Loser	0.86 : 0.11 : 0.03 → Loser	0.99 : 0.01 : 0.01 → Loser	Loser
8	0.77 : 0.05 : 0.18 → Loser	1.00 : 0.00 : 0.00 → Loser	0.63 : 0.24 : 0.13 → Loser	0.83 : 0.03 : 0.14 → Loser	1.00 : 0.00 : 0.00 → Loser	Loser
9	1.00 : 0.00 : 0.00 → Loser	1.00 : 0.00 : 0.00 → Loser	0.67 : 0.31 : 0.02 → Loser	0.88 : 0.11 : 0.01 → Loser	0.99 : 0.01 : 0.00 → Loser	Loser
10	0.08 : 0.70 : 0.22 → Reserve Winner	0.00 : 0.20 : 0.80 → Winner	0.00 : 0.62 : 0.38 → Reserve Winner	0.63 : 0.30 : 0.07 → Loser	0.17 : 0.73 : 0.10 → Reserve Winner	Reserve Winner
11	0.00 : 0.15 : 0.85 → Winner	0.40 : 0.20 : 0.40 → Loser	0.01 : 0.16 : 0.83 → Winner	0.77 : 0.13 : 0.10 → Loser	0.44 : 0.32 : 0.24 → Loser	Winner
12	1.00 : 0.00 : 0.00 → Loser	1.00 : 0.00 : 0.00 → Loser	0.31 : 0.39 : 0.30 → Reserve Winner	0.69 : 0.17 : 0.15 → Loser	0.97 : 0.01 : 0.02 → Loser	Loser
13	0.06 : 0.92 : 0.02 → Reserve Winner	0.20 : 0.80 : 0.00 → Reserve Winner	0.11 : 0.69 : 0.20 → Reserve Winner	0.78 : 0.17 : 0.05 → Loser	0.19 : 0.67 : 0.14 → Reserve Winner	Reserve Winner
14	1.00 : 0.00 : 0.00 → Loser	1.00 : 0.00 : 0.00 → Loser	0.62 : 0.37 : 0.02 → Loser	0.80 : 0.16 : 0.04 → Loser	0.99 : 0.01 : 0.00 → Loser	Loser
15	0.00 : 0.95 : 0.05 → Reserve Winner	0.00 : 1.00 : 0.00 → Reserve Winner	0.00 : 0.80 : 0.20 → Reserve Winner	0.59 : 0.33 : 0.08 → Loser	0.17 : 0.62 : 0.21 → Reserve Winner	Reserve Winner
16	0.06 : 0.07 : 0.88 → Winner	0.00 : 0.20 : 0.80 → Winner	0.00 : 0.01 : 0.99 → Winner	0.29 : 0.14 : 0.57 → Winner	0.13 : 0.20 : 0.67 → Winner	Winner
17	0.00 : 1.00 : 0.00 → Reserve Winner	0.40 : 0.60 : 0.00 → Reserve Winner	0.01 : 0.87 : 0.12 → Reserve Winner	0.71 : 0.26 : 0.03 → Loser	0.26 : 0.69 : 0.06 → Reserve Winner	Reserve Winner
18	1.00 : 0.00 : 0.00 → Loser	1.00 : 0.00 : 0.00 → Loser	0.70 : 0.21 : 0.09 → Loser	0.86 : 0.08 : 0.06 → Loser	0.99 : 0.00 : 0.01 → Loser	Loser
19	0.01 : 0.97 : 0.02 → Reserve Winner	0.00 : 1.00 : 0.00 → Reserve Winner	0.01 : 0.82 : 0.18 → Reserve Winner	0.66 : 0.29 : 0.06 → Loser	0.17 : 0.62 : 0.21 → Reserve Winner	Reserve Winner
20	0.29 : 0.64 : 0.06 → Reserve Winner	0.60 : 0.40 : 0.00 → Loser	0.32 : 0.48 : 0.19 → Reserve Winner	0.66 : 0.16 : 0.17 → Loser	0.16 : 0.59 : 0.25 → Reserve Winner	Reserve Winner

Figure 3. Predictions

In order to evaluate the performance of each model based on the examples shown in Figure 3, a confusion matrix using information for the predicted and actual classifications can be used (e.g., Hasnain et al., 2020; Imamovic et al., 2020). In this context, the confusion matrices created for each algorithm used in the current study are presented in Figures 4a to 4e.

		Predicted			
		Loser	Reserve Winner	Winner	Σ
Actual	Loser	99.8 %	0.2 %	0.0 %	999
	Reserve Winner	1.7 %	97.3 %	1.0 %	293
	Winner	0.5 %	3.3 %	96.2 %	211
	Σ	1003	294	206	1503

Figure 4a. Confusion Matrix for Random Forest

		Predicted			
		Loser	Reserve Winner	Winner	Σ
Actual	Loser	93.4 %	4.9 %	1.7 %	999
	Reserve Winner	31.7 %	64.5 %	3.8 %	293
	Winner	38.9 %	12.3 %	48.8 %	211
	Σ	1108	264	131	1503

Figure 4b. Confusion Matrix for kNN

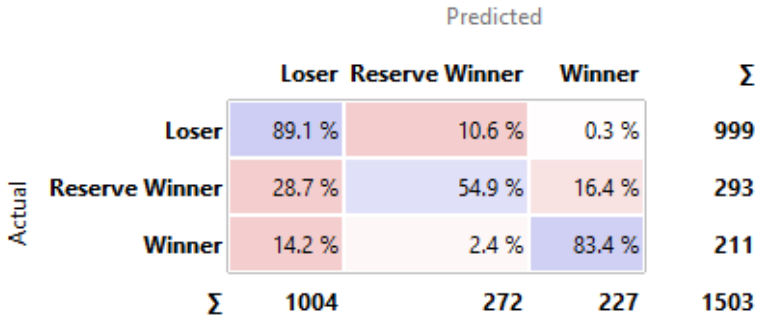

Figure 4c. Confusion Matrix for SVM

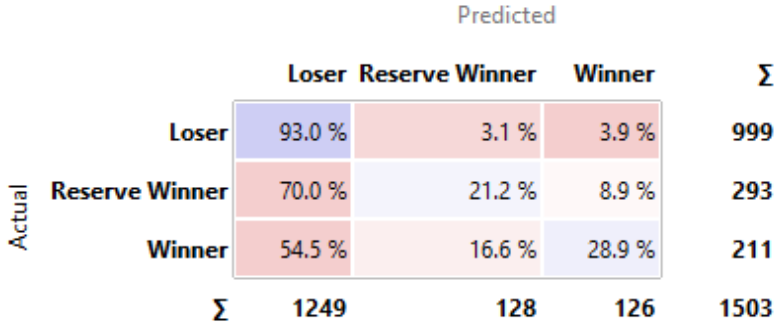

Figure 4d. Confusion Matrix for Logistic Regression

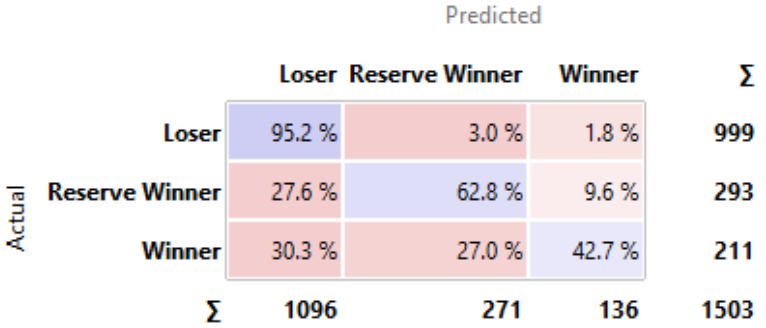

Figure 4e. Confusion Matrix for Naïve Bayes

The main diagonal of each 3x3 confusion matrix provides the percentage of correctly predicted samples, whilst the matrix elements outside of the main diagonal provide the percentage of predicted errors. Figure 4a shows that 99.8% of those not accepted into the PESE program, 97.3% of the reserve winners, and 96.2% of the winners were predicted correctly based on the Random Forest algorithm. The confusion matrices of the four other algorithms (kNN, SVM, Logistic Regression, and Naïve Bayes) are provided in Figures 4b, 4c, 4d, and 4e, respectively. Figure 4a showed that the algorithm that achieved the most accurate classification was Random Forest.

Other performance measures such as precision, recall, and specificity can also be used to evaluate the performances of models via a confusion matrix (Acikkar & Akay, 2009). For example, one measure is the Area under ROC Curve (AUC), which defines the area under the ROC curve, and is commonly used to evaluate the distinguishing

potential of prediction models (Janssens & Martens, 2020; Marzban, 2004). Therefore, in the current study, the basic concepts used in evaluating the performance of a model in terms of its prediction are as follows:

- Area under ROC Curve (AUC)
- Classification Accuracy (CA): Rate of correctly classified instances (e.g., Baldi et al., 2000; Pattiasina & Rosiyadi, 2020).
- Precision: Rate of true positives among instances classified as positive + false positives.
- Recall: Rate of true positives among all positive instances + false negatives.
- Specificity: Rate of true negatives among instances classified as negative + false positives.
- F1 (also known as balanced F-score): A weighted harmonic mean of values from Precision and Recall methods (Acikkar & Akay, 2009; Orange Data Mining, 2021).

In this context, the aforementioned six alternative performance measurements are indicated in Table 4 for each of the five models evaluated in the current study.

Table 4. Performance Measurements of the Models

Model	AUC	CA	Precision	Recall	Specificity	F1
Random Forest	1.000	.985	.985	.985	.989	.985
SVM	.940	.845	.855	.845	.899	.849
kNN	.932	.818	.811	.818	.757	.809
Logistic Regression	.805	.701	.662	.701	.573	.663
Naïve Bayes	.936	.815	.806	.815	.774	.803

According to Table 4, the algorithms' classification accuracy from highest to lowest were Random Forest (.985), SVM (.845), kNN (.818), Naïve Bayes (.815), and Logistic Regression (.701), respectively. In other words, the Random Forest algorithm was shown to have correctly classified the instances almost exactly.

Conclusion

In this study, a model is proposed to predict whether or not a candidate can be accepted into a PESE program by using EDM algorithms. Decision tree (Delen, 2011; Nandeshwar et al., 2011), SVM (Huang & Fang, 2013), Random Forest (Delen, 2011; Vandamme et al., 2007), and Artificial Neural Networks (Vandamme et al., 2007) are considered to be the more efficient models that provide more accurate results in estimating values based on more than one variable. In this context, the Random Forest, SVM, kNN, Logistic Regression, and Naïve Bayes algorithms were utilized in this study, which was conducted in order to predict the placement of candidate students in the PESE program. In addition, the performance indicators of each of these algorithms were then compared. Briefly, this research presents two main areas of focus. First, the study aims to predict whether or not candidates will be accepted into a PESE program; and second, compares the performance indicators of five algorithms frequently used for the purposes of prediction in EDM.

In this context, based on the analysis results of data obtained from 1.671 candidates between 2016 and 2020, it was estimated whether or not the candidates could be placed to the PESE program. The Random Forest algorithm

correctly predicted the placement status of the candidate with an excellent level of accuracy, whilst the other four algorithms were shown to have very high accuracy levels.

In the literature, there are but very few studies in which EDM algorithms are employed for a similar purpose. For example, Acikkar and Akay (2009) attempted to predict the placement status of PESE candidates to the program based on data from 2006 and 2007 belonging to 260 candidates, and using only the SVM algorithm. Their research results showed that SVM-based classification may be considered a promising tool in this area of application. In the current study, the performance indicators of five different algorithms were compared using data from 1.671 candidates for the 5-year period between 2016 and 2020, and it was seen that the Random Forest algorithm provided better results than SVM (Acikkar & Akay, 2009). Furthermore, the study carried out by Hussain et al. (2019), the performance indicators of different data mining algorithms for educational data were compared, and the findings indicated that the Neural Network was the best classifier.

In the literature, some studies compared the performance indicators of different algorithms in different disciplines (e.g., Deist et al., 2018; Uddin et al., 2019). Uddin et al. (2019) examined the performances of different algorithms in disease risk estimation in which 48 articles were examined, with the most frequently used algorithms found to be SVM (29 articles), and Naïve Bayes (23 articles). However, it was also stated that the algorithm with the highest degree of accuracy was SVM. Deist et al. (2018), on the other hand, compared the performances of classification algorithms using a radiotherapy-based dataset consisting of data from 3,496 patients. The researchers concluded that both the Random Forest and Elastic Net Logistic Regression (chemo) algorithms were found to be more discriminatory when compared to other classifiers in predicting toxicity accumulated from radiotherapy. Similarly, Asri et al. (2016) compared the performances of different algorithms (SVM, Decision Tree, Naïve Bayes, and kNN) in diagnosing breast cancer, and reported that SVM was shown to result in the highest accuracy. Belavagi and Muniyal (2016) compared the performance indicators of four algorithms (Random Forest, Gaussian Naïve Bayes, Logistic Regression, and SVM) in classifying data traffic to determine whether or not there was an attack on a network system. The study showed that Random Forest presented the highest performance level in determining whether or not there was an attack in the data traffic.

In summary, although it can be said that different algorithms can provide better results in the research conducted to date in different fields, the accuracy rate of the SVM and Random Forest algorithms can be said to be generally higher than other algorithms. For this reason, these algorithms should be included in the models used in future research. Furthermore, the current study was conducted with data obtained from 1,671 candidates who applied to the PESE program of a state university in Turkey. In future studies, more in-depth knowledge of the subject could be obtained by including data from candidates applying to PESE programs of different universities, as different types of STEs may be preferred by different universities (Kamuk, 2019; Kizir et al., 2014). Moreover, similar research studies could be conducted for other programs that also accept students via the STE route. In addition, similar studies may be carried out in different countries where similar exams are applied, and the situation between different cultures could then be compared. Additionally, it is recommended to investigate the reasons behind the decreases seen in the number of candidates applying to PESE programs in recent years. It would also be an interesting topic of research to investigate the reasons why males appear more interested in applying to PESE programs than females.

Based on these findings, it may be an interesting research subject to examine the relationship between the academic achievements of candidates who enrolled to PESE programs through until their graduation and the scores they initially received from their STEs. In other words, investigating whether or not those students who are more successful in the STE go on to graduate with a higher level of academic success from their respective PESE program may provide valuable information that could then be used to aid decision-making processes for faculties and administrators. Similarly, researching the relationship between students' achievement on the courses they take during their education and their STE scores may also provide important information in terms of supporting data-based decision-making processes.

Geniş Özet

Giriş

Teknolojideki son gelişmelere paralel olarak, günlük olarak üretilen veri hacminin yanı sıra veri üretim hızında da sürekli bir artış söz konusudur. Günlük yaşamın birçok farklı alanında giderek daha fazla kullanılmaya başlanan bu verilerin analizinden anlamlı bilgiler, modeller ve eğilimler ortaya çıkarılabilir (örn., Agrawal & Prabakaran, 2020; Hallikainen ve diğerleri, 2020; Line ve diğerleri, 2020) ve bu veriler büyük veri olarak tanımlanabilir (De Mauro ve diğerleri, 2015). Bu veriler eğitim alanında öğrencilerin, öğretmenlerin, idarecilerin ve diğer paydaşların yeni araştırmalar yapmalarına ve daha önce bilinmeyen yeni yolları keşfetmelerine olanak tanıyabilir. Her paydaş grubunun kendi misyon ve hedefleri olduğundan, eğitim bilgilerinin genellikle çeşitli açılardan analiz edilmesi gerekmektedir (Hanna, 2004).

Gelişmekte olan bir disiplin olarak eğitsel veri madenciliği, Romero ve Ventura (2010, s. 601) tarafından "farklı eğitim verileri üzerinde istatistiksel, makine öğrenimi ve veri madenciliği algoritmalarından yararlanma" aracı olarak tanımlanmıştır ve birincil amacı, eğitim çalışmalarında karar vermeyi güçlendirmek için eğitim verilerinden elde edilen faydalı bilgileri çıkarmaktır (Calvet Liñán & Juan Pérez, 2015). Bunu sağlamak için eğitsel veri madenciliği alanında kullanılan ve farklı araştırmacılar tarafından farklı şekillerde sınıflandırılan çeşitli yöntemler bulunmaktadır (Baker, 2011; Baker & Yacef, 2009; Bakhshinategh ve diğerleri, 2018; Romero & Ventura, 2013). Örneğin, Bakhshinategh ve diğerleri (2018, s. 540) çalışmasında en sık kullanılan yöntemleri; sınıflandırma ve regresyon, kümeleme, birliktelik kuralı madenciliği, modellerle keşif, aykırı değer tespiti, sosyal ağ analizi, metin madenciliği, sıralı örüntü madenciliği ve görselleştirme tekniği olarak belirtmiştir. Baker (2011) tarafından yapılan başka bir çalışmada ise bu yöntemler beş genel kategoriye ayırmıştır: ilişki madenciliği, kümeleme, tahmin, insan muhakemesi için verilerin damıtılması ve modellerle keşif. Bu nedenle, eğitsel veri madenciliği için kullanılan her yöntemin, kullanımında dikkate alınması gereken, potansiyel olarak farklı özelliklere sahip olabileceği göz önünde bulundurulmalıdır.

Eğitsel veri madenciliğinde en çok tercih edilen yöntemlerden biri ise tahmindir (Aouifi ve diğerleri, 2021; Sokkhey & Okazaki, 2020; Saa, 2016). Günümüz eğitim sistemlerinin önemli hedeflerinden biri ise, akademik sonuçları en üst düzeye çıkarmak ve öğrencilerin başarısızlık oranlarını en aza indirerek eğitim deneyiminin kalitesini artırmaktır. Bu bağlamda, tahmin yöntemleri kullanılarak öğrenicilerin akademik performansları (Waheed ve diğerleri, 2020) ve belirli bir ders için başarılı/başarısız olma olasılıkları (Abut ve diğerleri, 2018) tahmin edilebilir. Aynı yöntemler, okuldan ayrılmalar hakkında erken tahminlerde bulunmak için de kullanılabilir (Márquez-Vera ve diğerleri, 2016).

Ayrıca adayları farklı kriterlere göre değerlendirerek, okullara veya diğer eğitim programlarına kabul edilip edilmeyecekleri, işe kabul edilip edilmeyecekleri vb. önceden tahmin edilebilir (örn., Açıkkar & Akay, 2009). Örneğin, öğrenciler doğrudan yükseköğretim kurumlarının kendileri tarafından düzenlenen özel yetenek sınavlarına (STE) göre, Türkiye'deki belirli yükseköğretim programlarına kabul edilebilirler. Örnek olarak, öğrencilerin Beden Eğitimi ve Spor Eğitimi programına kayıt hakkı kazanabilmeleri için çok sayıda kriter (bkz. Tablo 3) dikkate alınmaktadır. Bu bağlamda bu araştırmanın birincil amacı da adayların Beden Eğitimi ve Spor Eğitimi programına kabul edilip edilmeyeceklerini farklı algoritmalar kullanarak tahmin etmektir.

Yöntem

Türkiye'de öğrencilerin bir yükseköğretim kurumuna kayıt yaptırabilmeleri için öncelikle, Ölçme, Seçme ve Yerleştirme Merkezi tarafından yapılan Yükseköğretim Kurumları Sınavına girmeleri gerekmektedir. Ardından, Ölçme, Seçme ve Yerleştirme Merkezi tarafından yapılan üniversite giriş sınavlarından aldıkları puanlar dikkate alınarak ve doğrudan yükseköğretim kurumlarının kendileri tarafından düzenlenen özel yetenek sınavlarından aldıkları puanlar dikkate alınarak çeşitli programlara kabul edilebilirler. Çünkü Türkiye'deki yükseköğretim kurumlarındaki bazı programlara (antrenörlük, spor yöneticiliği, beden eğitimi ve spor eğitimi vb.) özel yetenek sınavları ile öğrenci kabul edilmektedir. Örneğin, Spor Bilimleri Fakültelerinin Beden Eğitimi ve Spor Eğitimi programlarına özel yetenek sınavları ile öğrenci alınmaktadır. Bu bağlamda, bu araştırma kapsamında Türkiye'deki bir devlet üniversitesinin Beden Eğitimi ve Spor Eğitimi programına kayıt hakkı kazanabilmek için, 2016-2020 yılları arasında başvuruda bulunan 1.671 adaydan elde edilen verilerle çalışılmıştır.

Bulgular

Mevcut çalışmada, çeşitli algoritmalara göre tahminler gerçekleştirmek için Orange yazılımı kullanılmıştır (Padmavaty vd., 2020). Bu bağlamda, bir adayın Beden Eğitimi ve Spor Eğitimi programına kabul edilip edilemeyeceğini tahmin etmek için Random Forest, Support Vector Machines, Logistic Regression, k-Nearest Neighbors ve Naïve Bayes algoritmaları kullanılmıştır. Araştırma bulgularına göre, algoritmaların en yüksekten en düşüğe doğru sınıflandırma doğruluğu; Random Forest (.985), SVM (.845), kNN (.818), Naïve Bayes (.815) ve Logistic Regression (.701) şeklindedir. Başka bir ifadeyle, Random Forest algoritmasının, örnekleri neredeyse tam olarak doğru bir şekilde sınıflandırdığı görülmüştür.

Sonuç

Bu çalışmada, eğitsel veri madenciliği algoritmaları kullanılarak bir adayın Beden Eğitimi ve Spor Eğitimi programına kabul edilip edilemeyeceğini tahmin etmek için bir model önerilmiştir. Ayrıca, bu algoritmaların her birinin performans göstergeleri daha sonra karşılaştırılmıştır. Diğer bir ifadeyle, bu araştırmanın iki ana odak alanı bulunmaktadır. İlk olarak, adayların ilgili programa kabul edilip edilmeyeceklerini tahmin etmeyi amaçlamaktadır. İkincisi ise, eğitsel veri madenciliğinde tahmin amacıyla sıklıkla kullanılan beş farklı algoritmanın performans göstergelerini karşılaştırmaktır. Bu kapsamda Beden Eğitimi ve Spor Eğitimi programına kayıt hakkı kazanmak için, 2016-2020 yılları arasında başvuruda bulunan 1.671 adaydan elde edilen veriler analiz edilmiştir. Sonuçlar Random Forest algoritmasının, adayın yerleşme durumunu mükemmel bir doğruluk seviyesiyle tahmin edebildiğini göstermiştir.

Alanyazında eğitsel veri madenciği algoritmalarının benzer bir amaçla kullanıldığı çok az sayıda çalışma bulunmaktadır. Örneğin, Açıkkar ve Akay (2009) tarafından yapılan çalışmada, 2006 ve 2007 yıllarında, Beden Eğitimi ve Spor Eğitimi programına başvuruda bulanan 260 adayın ilgili programa yerleşip yerleşemeyeceği sadece SVM algoritması kullanılarak tahmin edilmeye çalışılmıştır. Bu çalışmada ise, 2016-2020 yılları arasındaki 5 yıllık dönem için 1.671 adaydan elde edilen veriler kullanılarak beş farklı algoritmanın performans göstergeleri karşılaştırılmış ve Random Forest algoritmasının SVM'den daha iyi sonuçlar verdiği görülmüştür.

Bu bulgulardan hareketle, Beden Eğitimi ve Spor Eğitimi programlarına kayıt yaptıran adayların mezuniyetlerine kadar olan akademik başarıları ile başlangıçta girdikleri özel yetenek sınavlarından aldıkları puanlar arasındaki

ilişkiyi incelemek ilginç bir araştırma konusu olabilir. Diğer bir ifadeyle, özel yetenek sınavından daha başarılı olan öğrencilerin ilgili programdan daha yüksek düzeyde akademik başarı ile mezun olup olmadıklarını araştırmak, yöneticiler için karar verme süreçlerine yardımcı olabilecek değerli bilgiler sağlayabilir. Benzer şekilde, öğrencilerin öğrenimleri süresince aldıkları derslerdeki başarıları ile özel yetenek sınavı puanları arasındaki ilişkinin araştırılması da veriye dayalı karar verme süreçlerini desteklemesi açısından önemli bilgiler sağlayabilir.

Yayın Etiği Bildirimi / Research Ethics

Araştırma ve yayın etiği konusunda bilimsel etik kaideler göz önünde bulundurulmuştur. / Scientific ethical principles have been taken into consideration in research and publication ethics.

Araştırmacıların Katkı Oranı / Contribution Rate of Researchers

Birinci araştırmacı makalenin genelinde ana sorumlu yazar olarak çalışmada yer alırken, ikinci yazar giriş, yöntem ve sonuç bölümlerinin yazımına katkı sağlamıştır. Üçüncü yazar verilerin toplanması ve analizinde görev almışken, dördüncü yazar sonuç bölümünün ve araştırmanın geniş özetinin yazılmasına katkı sağlamıştır. / While the first researcher took part in the study as the main responsible author throughout the article, the second author contributed to the writing of the introduction, method and conclusion sections. While the third author took part in the collection and analysis of the data and the writing of the results section, the fourth author contributed to the writing of the conclusion and the summary of the research.

Çıkar Çatışması / Conflict of Interest

Bu çalışmanın herhangi bir çıkar çatışması bulunmamaktadır. / This study has no conflict of interest.

Fon Bilgileri / Funding

Bu çalışma herhangi bir fon almamıştır. / This work has not received any funding.

Etik Kurul Onayı / The Ethical Commitee Approval

Bu araştırma makalesinin etik sorunu olmadığını beyan ederiz. / We hereby declare that this research article does not have an unethical problem.

Kaynakça/References

Abut, F., Yüksel, M. C., Akay, M. F., & Daneshvar, S. (2018). Predicting student's pass/fail status in an academic course using deep learning : A case study. *International Journal of Scientific Research in Information Systems and Engineering*, *4*(1), 87–91.

Acikkar, M., & Akay, M. F. (2009). Support vector machines for predicting the admission decision of a candidate to the School of Physical Education and Sports at Cukurova University. *Expert Systems with Applications*, *36*, 7228–7233. https://doi.org/10.1016/j.eswa.2008.09.007

Agrawal, R., & Prabakaran, S. (2020). Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity*, *124*(4), 525–534. https://doi.org/10.1038/s41437-020-0303-2

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

Aouifi, H. E., Hajji, M. E., Es-Saady, Y., & Douzi, H. (2021). Predicting learner's performance through video sequences viewing behavior analysis using educational data-mining. *Educational and Information Technologies*, *26*, 5799–5814. https://doi.org/10.1007/s10639-021-10512-4

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069. https://doi.org/10.1016/j.procs.2016.04.224

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., & 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

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, *16*(5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

Belavagi, M. C., & Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion detection. *Procedia Computer Science*, *89*, 117-123. https://doi.org/10.1016/j.procs.2016.06.016

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

Deist, T. M., Dankers, F. J. W. M, Valdes, G., Wijsman, R., Hsu, I.-C., Oberije, C., Lustberg, T., van Soest, J., Hoebers, F., Jochems, A., El Naqa, I., Wee, L., Morin, O., Raleigh, D. R., Bots, W., Kaanders, J. H., Belderbos, J., Kwint, M., Solberg, T.,…Lambin, P. (2018). Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers. *Medical Physics*, *45*(7), 3449-3459. https://doi.org/10.1002/mp.12967

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory and Practice*, *13*(1), 17–35. https://doi.org/10.2190/CS.13.1.b

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *American Institute of Physics*, *1644*, 97–104. https://doi.org/10.1063/1.4907823

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*(1), 115–125. https://doi.org/10.1016/j.dss.2014.02.003

Hallikainen, H., Savimäki, E., & Laukkanen, T. (2020). Fostering B2B sales with customer big data analytics. *Industrial Marketing Management*, *86*, 90–98. https://doi.org/10.1016/j.indmarman.2019.12.005

Hanna, M. (2004). Data mining in the e-learning domain. *Campus-Wide Information Systems*, *21*(1), 29–34. https://doi.org/10.1108/10650740410512301

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating trust prediction and confusion matrix measures for web services ranking. *IEEE Access*, *8*, 90847–90861. https://doi.org/10.1109/ACCESS.2020.2994222

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., & Villalba, L. J. G. (2019). Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. *Sensors*, *19*(7), Article 1746. https://doi.org/10.3390/s19071746

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*, *61*(1), 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

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. In R. Silhavy (Eds.), *Cybernetics and Algorithms in Intelligent Systems. CSOC2018. Advances in Intelligent Systems and Computing* (Vol. *765*, pp. 196–211). Springer https://doi.org/10.1007/978-3-319-91192-2_21

Imamovic, D., Babovic, E., & Bijedic, N. (2020). Prediction of mortality in patients with cardiovascular disease using data mining methods. In *Proceedings, 19th International Symposium INFOTEH-JAHORINA* (pp. 1-4). IEEE. https://doi.org/10.1109/INFOTEH48170.2020.9066297

Janssens, A. C. J. W., & Martens, F. K. (2020). Reflection on modern methods: Revisiting the area under the ROC Curve. *International Journal of Epidemiology*, *49*(4), 1397–1403. https://doi.org/10.1093/ije/dyz274

Kamuk, Y. (2019). Evaluation of the sports faculties' talent-based selection exams in the light of the new higher education examination system. *Spormetre the Journal of Physical Education and Sport Sciences*, *17*(3), 222–236. https://doi.org/10.33689/spormetre.510632

Karlos, S., Kostopoulos, G., & Kotsiantis, S. (2020). Predicting and interpreting students' grades in distance higher education through a semi-regression method. *Applied Sciences*, *10*(23), 1–19. https://doi.org/10.3390/app10238413

Karthikeyan, V. G., Thangaraj, P., & Karthik, S. (2020). Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. *Soft Computing*, *24*, 18477–18487. https://doi.org/10.1007/s00500-020-05075-4

Kizir, E., Temel, C., & Güllü, M. (2014). Examination of methods for student selection to the schools of physical education and sports in Turkey. *Spormetre the Journal of Physical Education and Sport Sciences*, *12*(2), 133–138. https://doi.org/10.1501/Sporm_0000000261

Kılıç Depren, S., Aşkın, Ö. E., & Öz, E. (2017). Identifying the classification performances of educational data mining methods: A case study for TIMSS. *Educational Sciences: Theory & Practice*, *17*(5), 1605–1623. https://doi.org/10.12738/estp.2017.5.0634

Line, N. D., Dogru, T., El-Manstrly, D., Buoye, A., Malthouse, E., & Kandampully, J. (2020). Control, use and ownership of big data: A reciprocal view of customer big data value in the hospitality and tourism industry. *Tourism Management*, *80*. https://doi.org/10.1016/j.tourman.2020.104106

Marzban, C. (2004). The ROC Curve and the area under it as performance measures. *Weather and Forecasting*, *19*(6), 1106–1114. https://doi.org/10.1175/825.1

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, *38*(12), 14984–14996. https://doi.org/10.1016/j.eswa.2011.05.048

Orange Data Mining. (2021). *Orange Widgets.* https://orangedatamining.com/widget-catalog/evaluate/testandscore/

Ölçme, Seçme ve Yerleştirme Merkezi. (2021). *2020 Yükseköğretim Kurumları Sınavı Sayısal Veriler* [*Numerical Data for the 2020 Higher Education Institutions Exam*]. https://dokuman.osym.gov.tr/pdfdokuman/2020/YKS/yks_sayisal_27072020.pdf

Padmavaty, V., Geetha, C., & Priya, N. (2020). Analysis of data mining tool Orange. *International Journal of Modern Agriculture*, *9*(4), 1146–1150. http://www.modern-journals.com/index.php/ijma/article/view/485

Pattiasina, T., & Rosiyadi, D. (2020). Comparison of data mining classification algorithm for predicting the performance of high school students. *Jurnal Techno Nusa Mandiri*, *17*(1), 23–30. https://doi.org/10.33480/techno.v17i1.1226

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*, 12–27. https://doi.org/10.1002/widm.1075

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5), 212–220. https://dx.doi.org/10.14569/IJACSA.2016.070531

Sokkhey, P., & Okazaki, T. (2020). Developing web-based support systems for predicting poor- performing students using educational data mining techniques. *International Journal of Advanced Computer Science and Applications*, *11*(7), 23–32. https://dx.doi.org/10.14569/IJACSA.2020.0110704

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics & Decision Making*, *19*, Article 281. https://doi.org/10.1186/s12911-019-1004-8

Vandamme, J.-P., Meskens, N., & Superby, J.-F. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics*, *15*(4), 405–419. https://doi.org/10.1080/09645290701409939

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*, Article 106189. https://doi.org/10.1016/j.chb.2019.106189

Yulia, L. W. S. (2020). Predicting student performance in higher education using multi-regression models. *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, *18*(3), 1354–1360. https://doi.org/10.12928/TELKOMNIKA.v18i3.14802