Eğitim Teknolojisi

kuram ve uygulama

Kış 2019
Cilt 9
Sayı 1

Winter 2019
Volume 9
Issue 1

Educational Technology

theory and practice

ISSN: 2147-1908

EĞİTİM TEKNOLOJİSİ KURAM VE UYGULAMA / EDUCATIONAL TECHNOLOGY THEORY AND PRACTICE

Cilt 9, Sayı 1, Kış 2019
Volume 9, Issue 1, Winter 2019

Genel Yayın Editörü / Editor-in-Chief: **Dr. Halil İbrahim YALIN**
Editör / Editor: **Dr. Tolga GÜYER**

Basım Editörü / Publisher Editor: **Dr. Tolga GÜYER**
Redaksiyon / Redaction: **Mertcan ÜNAL, Dr. Burcu BERİKAN, Figen DEMİREL UZUN, Akça Okan YÜKSEL**
Dizgi / Typographic: **Dr. Tolga GÜYER**
Kapak ve Sayfa Tasarımı / Cover and Page Design: **Dr. Bilal ATASOY**
İletişim / Contact Person: **Dr. Tolga GÜYER**

Dizinlenmektedir / Indexed in: **ULAKBİM Sosyal ve Beşerî Bilimler Veritabanı (TR-Dizin), Türk Eğitim İndeksi, Sosyal Bilimler Atıf Dizini**

ETKU Dergisi **2011 yılından itibaren yılda iki defa** düzenli olarak yayınlanmaktadır.
Educational Technology Theory and Practice Journal is published regularly **twice a year since 2011.**

Editör Kurulu / Editorial Board*

* Liste isme göre alfabetik olarak oluşturulmuştur. / List is created in alphabetical order

Hakem Kurulu / Reviewers*

* Liste isme göre alfabetik olarak oluşturulmuştur. / List is created in alphabetical order.

İletişim Bilgileri / Contact Information

İnternet Adresi / Web: http://dergipark.gov.tr/etku
E-Posta / E-Mail: tguyer@gmail.com
Telefon / Phone: +90 (312) 202 17 38
Adres / Adress: Gazi Üniversitesi, Gazi Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü, 06500 Teknikokullar - Ankara / Türkiye

EĞİTİM TEKNOLOJİSİ Kuram ve Uygulama
Cilt:9 Sayı:1 Yıl:2019

Makale Geçmişi / Article History
Alındı/Received: 01.06.2018
Düzeltme Alındı/Received in revised form: 10.12.2018
Kabul edildi/Accepted: 11.12.2018

ÖĞRENCİLERİN STEM KARİYER TERCİHLERİNİN VERİ MADENCİLİĞİ YAKLAŞIMI İLE TAHMİN EDİLMESİ

Gökhan AKÇAPINAR[1], Erdal COŞGUN[2]

Öz

Bu çalışmada ortaokul öğrencilerinin, ASSISTments isimli zeki öğretim sistemindeki etkileşim verileri kullanılarak, eğitim ve mesleki kariyerlerine STEM ile ilgili bir alanda devam edip etmeyeceklerini tahmin edecek bir model oluşturulması amaçlanmıştır. Analizler 2017 yılında aynı amaçla düzenlenen ASSISTments Veri Madenciliği Yarışması'nda (ASSISTments Data Mining Competition 2017) katılımcılara sunulan veri seti ile gerçekleştirilmiştir. Veri seti, 2004-2007 yılları arasında sistemi kullanan 1709 öğrenciye ilişkin yaklaşık 1 milyon satırlık tıklama verisini içermektedir. Veriler, öğrencileri tanımlayan bilgiler silinerek katılımcılara sunulmuştur. Veri setinde 514 öğrencinin STEM kariyerine devam edip etmedikleri bilgisini içeren bir eğitim veri seti yer almaktadır. Tahmin modeli oluşturmak amacıyla Random Forest (RF), kNN, SVM (Support Vector Machine) ve GMB (Generalized Regression Models Boosted) algoritmaları kullanılmıştır. Veri setinde STEM tercih eden ve etmeyen öğrenciler arasında dengesiz dağılım bulunmaktadır. Bu nedenle farklı veri dengeleme yöntemlerinin modellerin tahmin performansına etkisi de test edilmiştir. Sonuçların değerlendirilmesi için 10-katlı çapraz geçerlilik yöntemi kullanılmıştır. Yapılan analizler sonucunda en iyi sınıflama performansına SVM algoritması ile yukarı örnekleme yönteminin birlikte kullanıldığı durumda ulaşılmıştır. Bu durumda oluşturulan tahmin modeli, STEM kariyeri tercih eden öğrencilerin %66'sını doğru olarak tahmin etmiştir. Aynı zamanda öğrencilerin STEM kariyer tercihlerini belirlemede önemli olan değişkenler de analiz edilmiştir.

Anahtar Kelimeler: Eğitsel veri madenciliği; tahmin; sınıflama; makine öğrenmesi; STEM eğitimi; ASSISTments

[1] Dr., Hacettepe Üniversitesi/Eğitim Fakültesi/BÖTE Bölümü, gokhana@hacettepe.edu.tr, orcid.org/0000-0002-0742-1612
[2] Dr., Microsoft Genomics, AI & Research, Microsoft Research, Redmond, Washington, ercosgun@microsoft.com, orcid.org/0000-0002-0742-1612

PREDICTING STUDENTS' STEM CAREER INTERESTS BY USING DATA MINING APPROACH

Abstract

In this study, it is aimed at creating a model that will predict whether secondary school students will continue their education and professional careers in an area related to STEM or not. Interaction dataset made available to the participants in ASSISTments Data Mining Competition 2017 is analyzed. This anonymized dataset consists of approximately 1 million click-stream records collected from 1709 students who used the intelligent tutoring system between 2004-2007. The dataset also contained a training dataset that includes information about whether 514 students in the dataset continued their STEM careers or not. For prediction, the performance of the Random Forest (RF), kNN, SVM (Support Vector Machine) and GMB (Generalized Regression Models Boosted) algorithms are compared. There was a class imbalance problem in training dataset, therefore, we compared various data balancing algorithms' effect on the prediction algorithms. A 10-fold cross-validation was used to evaluate the performance of prediction models. As a result, the best performance was obtained when SVM algorithm and oversampling method were used together. In this case, the prediction model predicted over the students who prefer STEM careers with an accuracy of 66%. Features that are important while predicting STEM career preferences of students were also analyzed.

Keywords: Educational data mining; prediction; classification; machine learning; STEM education; ASSISTments

Summary

This study aimed at predicting secondary school students' STEM career interests by using click-stream data collected from an intelligent tutoring system called ASSISTments. Data from the ASSISTments Data Mining Competition held in 2017 was used for prediction. The dataset consists of 12 files in total. Ten of these files are click-stream data regarding students' ASSISTments platform usage. The click-stream data include approximately 1 million lines of record from 1709 students who used the system between 2004-2007. Another one file contained data for model training including STEM career choices of 514 students. In another file (one file), the test dataset with the User ID of 172 students was provided. In the test dataset, information on whether the students continued their STEM careers was not mentioned. Within the scope of this study, permission was obtained to use the dataset for research purposes. The performance of the prediction models was tested on the training dataset since we did not have labels for the test data. We used 10-fold cross-validation to evaluate the performance of prediction models.

During preprocessing, first the files containing the raw data of the students' click-streams were imported into the database. It was found that there were duplicate records in data and these records were eliminated from the analysis (n = 47). Each click of the students

on the system is a one-line record on the data set, and there are 78 columns in each row. While some of these data are associated with the steps of problem solving process (e.g. the time spent in that step, the status of getting help, etc.), some of them are independent of the current action (e.g. type of school, average knowledge, etc.). From these data, initially 27 features were extracted to represent students' performance and affective states. Highly correlated features were also determined and removed from the data set. A correlation cut-off point is equal to 0.75 was taken into account. After this step, the number of features in the dataset was reduced to 12. These 12 features are: *SchoolId*, *MCAS*, *SYASSISTmentsUsage*, *AveCarelessness*, *AveResEngcon*, *AveResConf*, *AveResFrust*, *AveResOfftask*, *AveAttemptCount*, *TotalIsOriginal*, *TotalIsScaffold*, and *AveFrTimeOnSkill*. The data set also includes a class feature that specifies whether student preferred a career related to STEM or not (*isSTEM*).

In the dataset, there was an unbalanced distribution between students who do not prefer STEM (n = 350) and those who prefer (n = 117). Therefore, data balancing methods were applied to balance the dataset. In order to balance the dataset, oversampling, undersampling, SMOTE and ROSE methods were used. SMOTE and ROSE methods are used for balancing data with an equal number of items in each class. The performance of the prediction algorithms on the data sets generated by different balancing methods and original data were compared.

Random Forest (RF), kNN, SVM (Support Vector Machine), and GMB (Generalized Regression Models Boosted) algorithms were used to create prediction models. These four algorithms were tested in five different datasets created during the preprocess. This made it possible to compare the effects of sampling methods and different algorithms on prediction models. Analysis is conducted by R with the help of caret package. To compare the performance of the models ROC, Sensitivity and Specificity metrics were used. Average results obtained by 10-fold cross-validation were reported. In the calculation of metrics, students who prefer STEM are considered as a positive class. In selecting the best classification model, the overall performance of the model and the degree of prediction of students who prefer STEM were taken into consideration.

Within the scope of the first research problem, an answer was sought to question "Can students' STEM career preferences be predicted by using click-stream data in the intelligent learning system?". Performance of the prediction models in terms of ROC, Sensitivity and Specificity metrics reported in Table 5 and in Figure 1. Based on these results, we yield this conclusion that there are models which produce usable results.

Within the scope of the second research problem, the question "What is the effect of different algorithms and data balancing techniques on the performance of predicting students' STEM career preferences?" was sought. When the results given in Table 5 and Figure 1 are examined, it is noticed that, Specificity values of all algorithms are close to zero when the original data set is used. In other words, data balancing has increased the performance of algorithms, especially for predicting students who prefer STEM careers.

The best performance results are achieved in the case where SVM algorithm and oversampling method are used together. The mean ROC value obtained as a result of cross validation indicates that the model has a 63% chance of predicting students who prefer and do not prefer STEM. This model predicted average 61% of the students who did not prefer STEM and average 66% of the students who prefer STEM. The confusion matrix obtained by 10-fold cross validation for the best prediction model is shown in Table 6. When the confusion

matrix was examined, it was found that the prediction model correctly classified 291 of 467 students (62%) in total, 214 out of the 350 students who did not prefer STEM (61%) and 77 of 117 students who prefer STEM (66%). When the model's error rates are examined, it was found that the best performing model incorrectly classified 39% of the students who do not prefer STEM (n = 136), and 34% of the students who prefer STEM (n = 40).

Within the scope of the third research problem, the answers to the question "What are the important features in predicting the STEM career preferences of the students?" was sought. For this purpose, the feature importance of the above mentioned the best classification model was examined. In Figure 2, features are given in order based on their importance. The first five features that are important in predicting students' STEM preferences are: *AveCarelessness*, *AveAttemptcount*, *MCAS*, *TotalIsOriginal*, and *AveResFrust*.

Giriş

Veri madenciliği yöntemleri kullanılarak verideki gizli örüntülerin keşfedilmesi ya da veriyi kullanarak geleceğe yönelik kestirimler yapılması mümkün olmaktadır ancak araştırmacıların veri toplamak için araçlar geliştirmesi ve istedikleri nitelikte veriye ulaşması her zaman mümkün olmamaktadır. Son yıllarda yaygınlaşan kamusal veri setleri bu noktada araştırmacılara önemli kaynak sağlamaktadır. Bu sayede araştırmacılar yöntemlerini test etmek için hazır verilere kolayca ulaşırken, veriyi sağlayan kurumlar da elde edilen sonuçları kendi sistemlerinin iyileştirilmesi için kullanma olanağı bulmaktadır. Eğitsel alanda da çevrimiçi öğrenme ortamlarından elde edilen eğitsel büyük verilerin araştırma amaçlı sunulduğu platformalar oluşturulmaya başlanmıştır (Koedinger vd., 2010; Stamper vd., 2010). Bu veriler kullanılarak zaman zaman yarışmalar düzenlenmekte ve bu sayede çok sayıda araştırmacının belirli bir problem üzerine odaklanması olanaklı hale gelmektedir. Eğitim alanındaki en kapsamlı yarışma, Veri Madenciliği, Veri Bilimi ve Analitik Topluluğu (KDD) tarafından 2010 yılında düzenlenmiştir[3]. Bu yarışmada katılımcılardan zeki öğrenme sistemindeki etkileşim verilerini kullanarak öğrencilerin gelecekteki performanslarını tahmin etmeleri istenmiştir. Yarışmada Yu vd. (2010) en düşük hata oranına sahip modeli geliştirerek birinci olmuşlardır.

Eğitsel alanda düzenlenen bir diğer yarışma ise 2017 yılında düzenlenen ASSISTments Veri Madenciliği Yarışması'dır (ASSISTments Data Mining Competition 2017). Düzenlenen yarışmada amaç ortaokul öğrencilerinin ASSISTments isimli sistemi kullanırken geride bıraktıkları tıklama verilerinden eğitim ve mesleki kariyerine STEM alanında devam edip etmeyeceklerinin tahmin edilmesidir. Sunulan çalışmada da bu veri seti aynı amaçla kullanılmış ve analiz süreci ile birlikte elde edilen sonuçlar raporlanmıştır.

Araştırmanın amacı, ortaokul öğrencilerinin ASSISTments isimli zeki öğretim sistemindeki etkileşim verilerinden ileriki yaşamlarında (lise, üniversite ve iş hayatlarında) STEM ile ilgili bir kariyer tercih edip etmeyeceklerini tahmin edecek bir model oluşturulmasıdır. Bu amaçla aşağıdaki araştırma sorularına cevap aranmıştır;

1. Öğrencilerin STEM kariyer tercihleri zeki öğrenme sistemindeki tıklama verileri kullanılarak tahmin edilebilir mi?
2. Farklı algoritma ve veri dengeleme tekniklerinin, öğrencilerin STEM kariyer tercihlerinin tahmin edilmesine etkisi nasıldır?
3. Öğrencilerin STEM kariyer tercihlerini tahmin etmede önemli olan değişkenler nelerdir?

İlgili Çalışmalar

ASSISTments platformu Matematikten İngilizceye birçok konuda öğrencilerin öğrenme süreçlerine yardımcı olmak amacıyla geliştirilmiş bir çevrimiçi öğrenme ortamıdır (Feng, Heffernan, & Koedinger, 2009). Aynı zamanda öğretmenlere de öğrencilerinin ilerlemelerini izleme imkânı sunmaktadır. Bu platformda araştırmacılar ve öğretmenler kendi içeriklerini üretebilmekte ve paylaşabilmektedir. Öğretmenler, sistem üzerinden öğrencilere sınıf içi uygulamalar yaptırabilmekte ya da ev ödevi verebilmektedir. Platform tüm dünyada 50.000'den fazla öğrenci tarafından kullanılmakta ve sistem üzerinden her yıl 12.5 milyondan fazla problem çözülmektedir (Heffernan & Heffernan, 2014).

[3] http://www.kdd.org/kdd-cup/view/kdd-cup-2010-student-performance-evaluation

ASSISTments platformundan elde edilen tıklama verileri bugüne kadar eğitsel veri madenciliği çalışmalarında farklı amaçlarla kullanılmıştır. Bu çalışmalardan bir tanesinde Pedro, Ocumpaugh, Baker, ve Heffernan (2014) 363 öğrencinin ASSISTments sistemindeki kullanım verilerini analiz ederek öğrencilerin üniversitede STEM ile ilgili bir bölüme gidip gitmediklerini tahmin etmeye çalışmışlardır. Araştırmacılar oluşturdukları tahmin modelinde öğrencilerin sistem ile etkileşimlerini, bilgi düzeylerini, duygu durumlarını ve davranışlarını yansıtan 10 adet değişken kullanmışlardır. Oluşturdukları son modelde bu değişkenlerden iki tanesi ile (bilgi düzeyi ve sistemle oyun oynama) öğrencilerin STEM eğitimini tercih edip etmeyeceklerini %66 oranında doğru tahmin ettiklerini raporlamışlardır.

Pedro, Baker, Bowers, ve Heffernan (2013) tarafından yapılan bir diğer çalışmada araştırmacılar, öğrencilerin ASSISTments sistemindeki kullanım verileri ile üniversiteye devam edip etmeyeceklerini tahmin etmeye çalışmışlardır. Ortaokul döneminde sistemi kullanan 3747 öğrenciye ait verinin analiz edildiği bu çalışmada araştırmacılar öğrencilerin bilgi düzeylerini, duygu durumlarını ve davranışlarını yansıtan 9 adet değişkeni kullanarak üniversiteye devam eden ve etmeyen öğrencileri %69 oranında doğru olarak tahmin eden bir model oluşturmuşlardır.

Pardos, Baker, San Pedro, Gowda, ve Gowda (2014) tarafından yapılan bir diğer çalışmada ise araştırmacılar, öğrencilerin okul yıllarında ASSISTments sistemindeki kullanım verilerinden duygu durumlarını (sıkılma, çaresizlik, kafa karışıklığı vb.) ve davranışsal bağlılıklarını (görev dışı davranışlar sergileme, oyun oynama vb.) yansıtan değişkenler ile yılsonu matematik sınavından aldıkları sonuçlar arasındaki ilişkiyi incelemişlerdir. Kullandıkları veri seti, sistemi düzenli olarak kullanan (1 yıl süresince, haftada 2 gün ve 2'şer saat) 1393 öğrencinin verisini içermektedir. Araştırma sonucunda en yüksek pozitif yönlü ilişki bağlı konsantrasyon (engaged concentration) ile matematik puanları arasında bulunmuştur (r = 0,45). En yüksek negatif yönlü ilişki ise sistem ile oyun oynama (gaming the system) puanları ile matematik puanları arasında bulunmuştur (r = -0,43). Bunun dışında diğer değişkenler ile öğrencilerin matematik puanları arasında farklı düzeylerde pozitif ve negatif yönlü ilişkiler bulunmuştur. Araştırmacılar aynı zamanda bu değişkenlerin öğrencilerin dönem sonu performanslarını tahmin etmede kullanılabileceğini gösteren bir tahmin modeli oluşturmuşlardır.

Yukarıda bahsedilen çalışmalardan da görüleceği üzere öğrencilerin ASSISTments sistemindeki etkileşim verileri duygu durumlarını, bilgi düzeylerini, istenmeyen davranışlarını ve akademik başarılarını tahmin etmek amacıyla kullanılmıştır. Bu çalışma, analiz edilen veri seti, değişkenler ve kullanılan yöntemler açısından yukarıda bahsedilen çalışmalardan farklıdır. Elde edilen sonuçlar öğrencilerin ortaokul sonrası dönemde (lise, üniversite ve iş hayatında) STEM tercihlerini belirlemede önemli olan yeni değişkenlerin ve yöntemlerin belirlenmesi açısından alan yazına katkı sağlayacaktır.

Yöntem

Veri Seti

Çalışmada 2017 yılında düzenlenen ASSISTments Veri Madenciliği Yarışması'nda (ASSISTments Data Mining Competition 2017) katılımcılara sunulan veri seti kullanılmıştır. Veri seti toplam 12 adet dosyadan oluşmaktadır. Bu dosyalardan 10 tanesi öğrencilerin ASSISTments platformundaki tıklama verileridir. Tıklama verileri 2004-2007 yılları arasında sistemi kullanan 1709 tekil öğrenciye ilişkin yaklaşık 1 milyon satırlık kayıt içermektedir. Her

bir satır ise öğrencilerin kullanımına ilişkin 78 sütunluk veri içermektedir. Diğer bir dosyada katılımcıların tahmin modeli oluşturmak amacıyla kullanması için 514 öğrencinin STEM kariyerini tercih edip etmediği bilgisini içeren eğitim veri seti (training set) verilmiştir. Bir diğerinde ise 172 öğrencinin ID'sinin yer aldığı test veri seti verilmiştir. Test veri setinde öğrencilerin STEM kariyerine devam edip etmediği bilgisi katılımcılara verilmemiştir. Orijinal veri setleri ve değişkenler ile ilgili detaylı bilgiler yarışma web sayfasından incelenebilir[4]. Bu çalışma kapsamında yarışmada kullanılan veri setlerinin araştırma amacıyla kullanılabilmesi için gerekli izinler alınmıştır ancak 172 öğrencinin bilgisini içeren test veri setinde öğrencilerin STEM tercihleri katılımcılarla paylaşılmadığı için oluşturulan tahmin modellerinin performansı 10-katlı çapraz geçerlilik yöntemi kullanılarak 514 öğrenciye ilişkin eğitim veri seti üzerinde test edilmiştir.

Veri Ön İşleme

Veri madenciliği çalışmalarında ön işleme süreci son derece önemlidir ve düşük hata oranına sahip modellerin geliştirilmesi ile doğrudan ilişkilidir. Veri ön işleme amacıyla ilk olarak öğrencilerin tıklama verilerine ilişkin ham verileri içeren dosyalar birleştirilerek bir veri tabanına aktarılmıştır. Bu sayede verilerin işlenmesi ve analizde kullanılacak değişkenlerin oluşturulması olanaklı hale gelmiştir. Ön işleme sürecinde veride tekrarlı kayıtlar olduğu tespit edilmiş ve bu kayıtlar analizden çıkartılmıştır (n = 47). Veri setinde öğrencilerin sistemde problem çözmek amacıyla yaptıkları her bir tıklama bir satır olarak yer almaktadır ve her bir satırda 78 sütunluk veri bulunmaktadır. Bu verilerden bazıları her bir tıklama adımı ile ilişkili iken (o adımda harcanan süre, yardım alma durumu vb.) bazıları ise öğrencinin genel durumuna ilişkin ve her satırda aynı olan verilerdir (okul türü, ortalama bilgi düzeyi vb.). Bu veriler kullanılarak ve daha önce yapılan çalışmalardan da yararlanılarak analizlerde kullanılmak üzere 27 adet değişken oluşturulmuştur.

Öğrencilerin liseye geçiş sınavında aldıkları puanı gösteren MCAS değişkeninde kayıp veriler olduğu için ön işleme sürecinde bu kayıp veriler de ele alınmıştır. Bu amaçla model bazlı çalışan kNN algoritmasından (Kowarik & Templ, 2016) yararlanılarak, verilerin rastgele ya da ortalama ile doldurulması yerine benzer öğrencilerin verileri dikkate alınarak doldurulması sağlanmıştır. Daha sonra yüksek korelasyona sahip değişkenler belirlenmiş ve bu değişkenler veri setinden çıkartılmıştır. Korelasyon kesme noktası olarak 0,75 değeri alınmıştır. Diğer bir ifade ile aralarında 0,75 ve üzeri ilişki olan değişkenlerden bir tanesi analizde tutularak diğerleri analizden çıkartılmıştır. Bu aşamadan sonra veri setindeki değişken sayısı 27'den 12'ye düşmüştür. Bu değişkenler ve açıklamaları Tablo 1'de sunulmuştur. Değişken adları veri setinde olduğu gibi bırakılmıştır.

[4] https://sites.google.com/view/assistmentsdatamining/data-mining-competition-2017

Tablo 1. Seçilen değişkenler ve açıklamaları

Değişken	Açıklama
SchoolId	Verinin toplandığı zaman öğrencinin kayıtlı olduğu okulun kodu
MCAS	Öğrencilerin MCAS isimli matematik sınavından aldıkları puan
SYASSISTmentsUsage	Öğrencinin sistemi kullanıldığı akademik yıl
AveCarelessness	Öğrencinin ortalama dikkatsizlik (carelessness*)* puanı (M. O. C. Z. San Pedro, Baker, & Rodrigo, 2011)
AveResEngcon	Öğrencinin ortalama bağlı konsantrasyon (engaged concentration) puanı (M. O. C. Z. San Pedro vd., 2011)
AveResConf	Öğrencinin ortalama kafa karışıklığı (confusion) puanı (M. O. C. Z. San Pedro vd., 2011)
AveResFrust	Öğrencinin ortalama çaresizlik (frustration) puanı (M. O. C. Z. San Pedro vd., 2011)
AveResOfftask	Öğrencinin ortalama görev dışı (off-task) davranış puanı (Baker, 2007; M. O. C. Z. San Pedro vd., 2011)
AveAttemptCount	Öğrencinin bir problemi çözmek için ortalama deneme sayısı
TotalIsOriginal	Öğrencinin karşılaştığı toplam orijinal problem sayısı
TotalIsScaffold	Öğrencinin karşılaştığı toplam rehberli yardım (scaffolding) türündeki problem sayısı
AveFrTimeOnSkill	Öğrencinin problemlerin çözümü için gereken becerilerle ilk karşılaştığında ortalama cevap verme süresi

Veri setinde ayrıca öğrencilerin STEM ile ilgili bir alan tercih edip etmeme durumlarını belirten sınıf değişkeni yer almaktadır (isSTEM).

Veri setinde STEM tercih etmeyen öğrenciler (n = 350) ile tercih eden öğrenciler (n = 117) arasında 1'e 3 oranında bir dengesiz dağılım söz konusudur. Sınıf değişkeninin orantısal olarak dengesiz olması sınıflama algoritmalarının azınlık sınıfın elemanlarını tahmin etme performansını olumsuz olarak etkilemektedir (Chawla, 2005). Bu sorunun önüne geçmek için dört farklı veri dengeleme yöntemi kullanılarak veri setinin dengelenmesi yoluna gidilmiştir. Aynı zamanda tahmin algoritmalarının farklı dengeleme yöntemleri ile oluşturulan veri setlerindeki performansları karşılaştırılmıştır. Veri setinin dengelenmesi amacıyla yukarı örnekleme (oversampling), aşağı örnekleme (undersampling), SMOTE ve ROSE yöntemleri kullanılmıştır. SMOTE ve ROSE yöntemleri dengeli örneklem oluşturmak amacıyla kullanılmıştır.

Yukarı örnekleme yöntemi, az olan sınıftan (STEM tercih eden öğrenciler) benzer veriler türetilmesi, aşağı örnekleme yöntemi, fazla olan sınıftan (STEM tercih etmeyen öğrenciler) rastgele veri silinmesi, dengeli örnekleme yöntemi ise çok olan sınıftan veri silinmesi, az olan sınıftan ise yeni veri türetilmesi yoluyla örneklemi dengelemektedir. Oluşturulan sınıflama

modellerinde aşırı uyum sorunu oluşmaması için veri dengeleme işlemi sadece eğitim veri setine çapraz geçerlilik aşamasında uygulanmıştır.

Ön işleme sonucu elde edilen veri setinde her bir öğrenci için bir satır ve seçilen değişkenlere ilişkin bilgileri içeren 12 adet sütun yer almaktadır. Bu verilere ilişkin tanımlayıcı istatistikler STEM eğitimine devam (1) eden ve etmeyen (0) öğrenciler için Tablo 4'te sunulmuştur. *SchoolId* ve *SYASSISTmentsUsage* değişkenleri kategorik olduğu için tabloya eklenmemiştir.

Tablo 4. Sürekli değişkenlere ilişkin tanımlayıcı istatistikler

	STEM Tercih Edenler (n = 117)		**STEM Tercih Etmeyenler (n = 350)**	
	Ortalama	**Standart Sapma**	**Ortalama**	**Standart Sapma**
MCAS	35,61	12,21	30,78	12,45
AveCarelessness	0,15	0,08	0,12	0,07
AveResEngcon	0,65	0,03	0,65	0,03
AveResConf	0,11	0,04	0,11	0,04
AveResFrust	0,12	0,05	0,13	0,05
AveResOfftask	0,22	0,08	0,22	0,09
AveAttemptCount	1,99	0,54	2,29	0,71
TotalIsOriginal	120,5	74,12	100,8	62,31
TotalIsScaffold	124,95	76,24	123,99	82,07
AveFrTimeOnSkill	394,47	328,93	406,32	321,14

Tanımlayıcı istatistikler incelendiğinde ortalama değerler açısından STEM tercih eden öğrencilerin problemleri daha az denemede cevapladığı (*AveAttemptCount*), yeni bir problem ile karşılaştıklarında daha az zaman harcadıkları (*AveFrTimeOnSkill*) ve genel sınav puanlarının da (*MCAS*) daha yüksek olduğu görülmektedir. Duygusal durumlarını yansıtan değişkenler incelendiğinde ise STEM tercih eden öğrencilerin daha dikkatsiz (*AveCarelessness*) davrandığı görülmektedir. Çaresizlik (*AveResFrust*) ve kafa karışıklığı (*AveResConf*) puanlarının ise birbirine yakın olduğu görülmektedir. Tanımlayıcı istatistikler, aynı araç ile elde edilmiş farklı veri setleri ile uyumludur (Pedro vd., 2014; M. O. Z. San Pedro, Baker, Gowda, & Heffernan, 2013).

Tahmin Analizleri

Veri madenciliği analizlerinde elde edilecek sonuçlar veriye uygulanan ön işleme yöntemlerine, seçilen algoritmalara ve değerlendirme yöntemlerine göre değişiklik göstermektedir (Olmo, Romero, Gibaja, & Ventura, 2015). Bu nedenle eğitsel veri madenciliği çalışmalarında sıklıkla kullanılan dört farklı algoritma seçilmiş ve bu algoritmaların tahmin performansları karşılaştırılmıştır. Tahmin analizleri Random Forest (RF), kNN, SVM (Support

Vector Machine) ve GBM (Generalized Boosted Regression Models) algoritmaları ile R veri madenciliği yazılımı (R Core Team, 2017) kullanılarak gerçekleştirilmiştir. Aynı zamanda bu dört farklı algoritma, oluşturulan beş farklı veri setinde test edilmiştir. Bu sayede yapılan yirmi farklı analiz sonucunda (4 x 5) örnekleme yöntemlerinin ve farklı algoritmaların tahmin performanslarının karşılaştırılması olanaklı hale gelmiştir. Tahmin analizleri R yazılımında yer alan caret paketi (Kuhn, 2008) ile gerçekleştirilmiştir. Caret paketi parametre optimizasyonu yöntemi ile her bir modelin en iyi performans gösterdiği parametreleri otomatik olarak bulmakta ve sonuçları bu parametreler ile raporlamaktadır.

Model Performansının Değerlendirilmesi

Oluşturulan modellerin performansı eğitim veri seti üzerinde çapraz geçerlilik yöntemi kullanılarak test edilmiştir. Bu amaçla 10-katlı çapraz geçerlilik yöntemi uygulanmıştır. Çapraz geçerlilik yönteminde veri seti k tane eşit parçaya ayrılmakta (burada 10) bu parçalardan k-1 tanesi model oluşturmak için kullanılmakta ve oluşturulan model geriye kalan (k) veri seti üzerinde test edilmektedir. Bu işlem, her bir parça test seti olarak kullanılana kadar devam etmektedir (Refaeilzadeh, Tang, & Liu, 2016).

Model performanslarının karşılaştırılmasında ROC, Duyarlılık (Sensitivity) ve Seçicilik (Specificity) metrikleri kullanılmıştır. ROC metriği ROC eğrisi altında kalan alanı göstermektedir ve 0 ile 1 arasında değer almaktadır. Sıfır değeri modelin pozitif ve negatif sınıfı ayırt edemediğini gösterirken 1 olması tamamen ayırt edebildiğini göstermektedir. Duyarlılık, modelin negatif sınıfı ayırt etme derecesini, Seçicilik ise modelin pozitif sınıfı ayırt etme derecesini göstermektedir ve ikisi de 0 - 1 aralığında değerler almaktadır. Test veri setindeki dengesiz dağılımdan dolayı doğru sınıflama oranı yerine bu metrikler tercih edilmiştir. Metriklerin hesaplanmasında pozitif sınıf olarak STEM tercih eden öğrenciler alınmıştır. En iyi sınıflama modelinin seçilmesinde de modelin genel performansı ve STEM tercih eden öğrencileri tahmin etme derecesi dikkate alınmıştır. Aynı zamanda en iyi modelin çapraz geçerlilik sonucu elde edilen hata matrisi (confusion matrix) verilerek modelin STEM eğitimi tercih eden ve etmeyen öğrencileri sınıflama oranları karşılaştırılmıştır.

Bulgular

Birinci araştırma problemi kapsamında "Öğrencilerin STEM kariyer tercihleri zeki öğrenme sistemindeki tıklama verileri kullanılarak tahmin edilebilir mi?" sorusuna cevap aranmıştır. Yapılan tahmin analizleri sonucu elde edilen ROC, Duyarlılık ve Seçicilik değerleri Tablo 5'te ve Şekil 1'de sunulmuştur. Analiz sonuçları incelendiğinde tahmin amacıyla kullanılabilecek modellerin oluştuğu görülmektedir. İkinci araştırma problemi kapsamında "Farklı algoritma ve veri dengeleme tekniklerinin öğrencilerin STEM kariyer tercihlerinin tahmin edilmesine etkisi nasıldır?" sorusuna cevap aranmıştır. Tabloda 5 ve Şekil 1'de verilen sonuçlar incelendiğinde, orijinal veri setinin kullanıldığı durumda tüm algoritmaların STEM tercih eden öğrencileri ayırt etme performansının sıfıra yakın olduğu görülmektedir. Diğer bir ifade ile veri dengeleme işlemi algoritmaların performansını genel olarak artırmıştır.

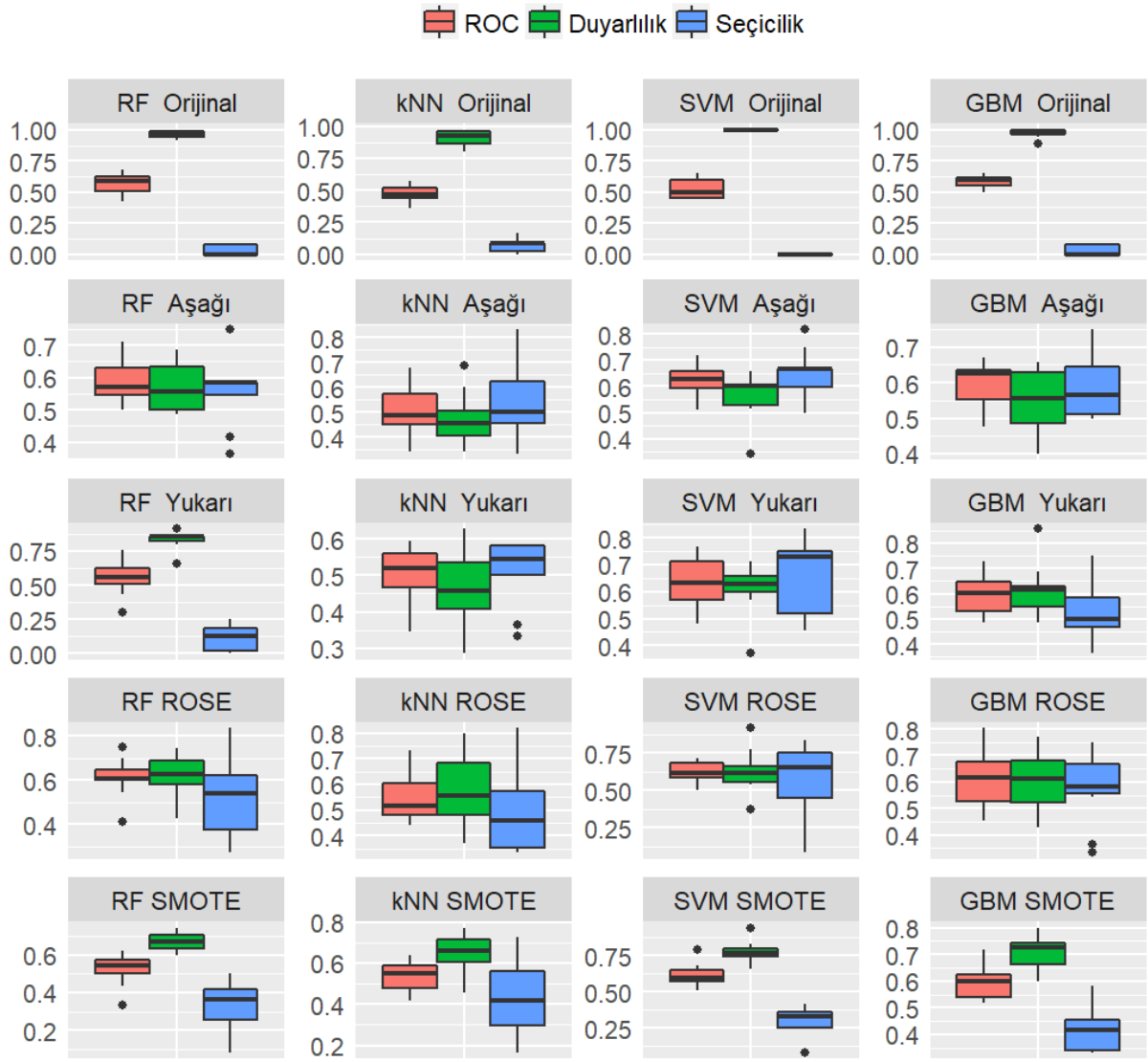
En iyi performans sonuçlarına SVM algoritması ve yukarı örnekleme yönteminin kullanıldığı durumda ulaşılmıştır. Çapraz geçerlilik sonucu elde edilen ortalama ROC değeri modelin STEM tercih eden ve etmeyen öğrencileri tahmin etmede %63 şansı olduğunu göstermektedir. Bunun dışında STEM tercih etmeyen öğrencilerin ortalama %61'ini, STEM tercih eden öğrencilerin de ortalama %66'ını doğru olarak tahmin etmiştir.

Tablo 5. Tahmin analizi sonuçları

	ROC			**Duyarlılık**			**Seçicilik**		
	Min	**Ort**	**Maks**	**Min**	**Ort**	**Maks**	**Min**	**Ort**	**Maks**
RF Orijinal	0,43	0,57	0,68	0,91	0,97	1,00	0,00	0,03	0,09
kNN Orijinal	0,36	0,47	0,57	0,80	0,91	0,97	0,00	0,07	0,17
SVM Orijinal	0,45	0,53	0,65	1,00	1,00	1,00	0,00	0,00	0,00
GBM Orijinal	0,50	0,58	0,65	0,89	0,97	1,00	0,00	0,03	0,09
RF Aşağı	0,50	0,58	0,71	0,49	0,57	0,69	0,36	0,55	0,75
kNN Aşağı	0,34	0,51	0,68	0,34	0,47	0,69	0,33	0,53	0,83
SVM Aşağı	0,51	0,62	0,72	0,34	0,57	0,66	0,50	0,65	0,82
GBM Aşağı	0,48	0,59	0,67	0,40	0,55	0,66	0,50	0,59	0,75
RF Yukarı	0,30	0,56	0,76	0,66	0,84	0,91	0,00	0,12	0,25
kNN Yukarı	0,35	0,51	0,59	0,29	0,47	0,63	0,33	0,51	0,58
SVM Yukarı	**0,48**	**0,63**	**0,77**	**0,37**	**0,61**	**0,71**	**0,45**	**0,66**	**0,83**
GBM Yukarı	0,49	0,60	0,73	0,49	0,61	0,86	0,36	0,54	0,75
RF ROSE	0,41	0,61	0,75	0,43	0,63	0,74	0,27	0,53	0,83
kNN ROSE	0,44	0,54	0,73	0,37	0,58	0,80	0,33	0,49	0,82
SVM ROSE	0,50	0,62	0,72	0,37	0,61	0,91	0,08	0,59	0,83
GBM ROSE	0,45	0,62	0,80	0,43	0,60	0,77	0,33	0,58	0,75
RF SMOTE	0,34	0,52	0,62	0,60	0,67	0,74	0,08	0,33	0,50
kNN SMOTE	0,42	0,53	0,64	0,46	0,65	0,77	0,17	0,43	0,73
SVM SMOTE	0,51	0,61	0,79	0,66	0,78	0,94	0,08	0,29	0,42
GBM SMOTE	0,52	0,59	0,72	0,60	0,71	0,80	0,33	0,42	0,58

Tablo 6'da ise en iyi sınıflama performansı sergileyen SVM algoritmasının yukarı örnekleme yöntemindeki çapraz geçerlilik sonucu elde edilen hata matrisine yer verilmiştir. Hata matrisi incelendiğinde, tahmin modelinin toplamda 467 öğrenciden 291'ini doğru sınıfladığı (%62), bunun yanında STEM tercih etmeyen 350 öğrenciden 214'ünü doğru sınıfladığı (%61), STEM tercih eden 117 öğrenciden ise 77'sini doğru olarak sınıfladığı (%66) görülmektedir. Hata oranları incelendiğinde ise en iyi performans gösteren modelin, gerçekte

STEM tercih etmeyen öğrencileri %39 oranında hatalı tahmin ettiği (n = 136), gerçekte STEM tercih eden öğrencileri ise %34 oranında hatalı tahmin ettiği (n = 40) görülmektedir.



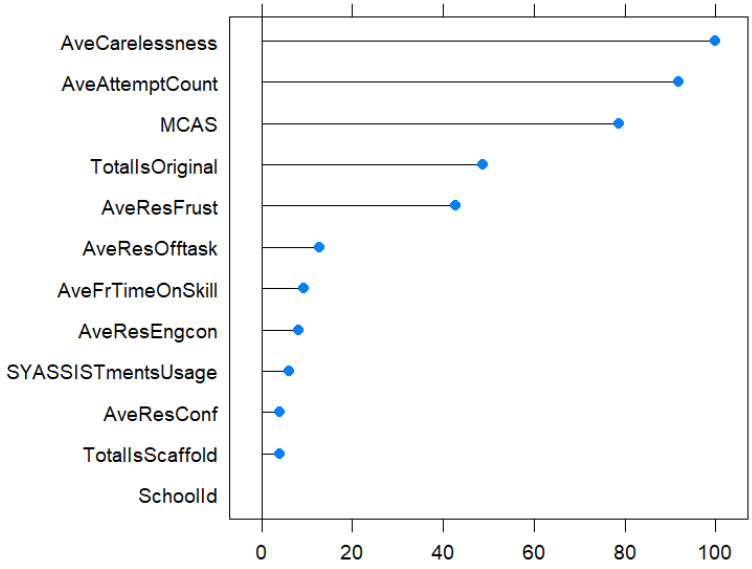
Şekil 1. Analiz sonucu elde edilen ROC, Duyarlılık (Sensitivity) ve Seçicilik (Specificity) metriklerinin görsel karşılaştırması

Tablo 6. Hata matrisi

Tahmin Sonucu	Gerçek Durum		
	STEM Tercih Etmeyen	STEM Tercih Eden	Toplam
STEM Tercih Etmeyen	214	40	254
STEM Tercih Eden	136	77	213
Toplam	350	117	467

Üçüncü araştırma problemi kapsamında "Öğrencilerin STEM kariyer tercihlerini tahmin etmede önemli olan değişkenler nelerdir?" sorusuna cevap aranmıştır. Bu amaçla yukarıda

belirtilen en iyi sınıflama modeline ilişkin değişken önemlilikleri incelenmiştir. Şekil 2'de değişkenler, önem sırasına göre verilmiştir. Buna göre öğrencilerin STEM tercihlerini tahmin etmede önemli olan ilk beş değişken; *AveCarelessness*, *AveAttemptcount*, *MCAS*, *TotalIsOriginal* ve *AveResFrust* değişkenleri olarak bulunmuştur.



Şekil 2. Öğrencilerin STEM tercihlerini tahmin etmede önemli olan değişkenler

Sonuç ve Tartışma

Bu çalışmada, ortaokul öğrencilerinin ASSISTments isimli zeki öğretim sistemindeki kullanım verilerinden, ileriki kariyerlerinde (lise, üniversite ve iş hayatında) STEM ile ilgili bir alan seçip seçmeyeceklerini tahmin edecek bir model oluşturulması amaçlanmıştır. Veri ön işleme aşamasında, veri birleştirme, tekrarlı veri temizleme, kayıp veri doldurma, veri dönüştürme, yüksek ilişkili değişkenlerin silinmesi gibi işlemler uygulanmıştır. Veri setinde STEM tercih eden ve etmeyen öğrenciler arasında dengesiz bir dağılım söz konusu oluğu için örneklem dengeleme yöntemleri kullanılmıştır. Analiz aşamasında ön işleme sonucu seçilen değişkenlerin, öğrencilerin STEM kariyerine devam edip etmeme durumlarını tahmin etme performansı test edilmiştir. Bu amaçla dört farklı tahmin algoritmasının, veri dengeleme yöntemleri kullanılarak oluşturulan beş farklı veri setindeki tahmin performansları karşılaştırmalı olarak test edilmiştir. Yapılan analizler sonucunda en iyi sınıflama yapan algoritma ve kullanılan örnekleme yöntemi belirlenmiştir. Araştırmada kullanılan tüm analizler 10-katlı çapraz geçerlilik yöntemi ile genelleştirilmiş, bu sayede makine öğrenmesi yöntemlerinde önemli bir sorun olan aşırı uyum sorunun önüne geçilmesi amaçlanmıştır.

Analiz sonucunda performans metrikleri açısından en iyi performansa SVM algoritması ile yukarı örnekleme yöntemi kullanılarak oluşturulan modelin kullanıldığı durumda ulaşılmıştır. Oluşturulan model STEM tercih eden öğrencileri ise ortalama %66 oranında doğru olarak sınıflamıştır. Pedro vd. (2014) aynı yazılımdan elde edilen farklı bir veri setinde, oluşturdukları lojistik regresyon modeliyle öğrencilerin STEM tercihlerini %66 oranında doğru olarak tahmin ettiklerini raporlamışlardır. Öğrenci başarısının ya da öğrenme çıktılarının tahmin edilmesi eğitsel veri madenciliği ve öğrenme analitiği alanlarında çalışılan en popüler araştırma alanıdır (Peña-Ayala, 2014). Bu konuda daha yüksek doğruluğa sahip modellerin geliştirilmesi, risk altındaki öğrencilerin zamanında belirlenmesi ve müdahale mekanizmalarının işe koşulması noktasında eğitimcilere yardımcı olacaktır.

Çalışmanın bir diğer bulgusu ise öğrencilerin STEM kariyeri tercihlerini tahmin etmede önemli olan değişkenlerin belirlenmesidir. Bu değişkenler *AveCarelessness*, *AveAttemptCount*, *MCAS*, *TotalIsOriginal* ve *AveResFrust* değişkenleridir. Bunlardan üç tanesi (*AveAttemptCount*, *MCAS, TotalIsOriginal*) öğrencilerin performansı ile ilgili iken diğer iki tanesi (*AveCarelessness*, *AveResFrust*) öğrencilerin aracı kullanırken sergiledikleri duygu durumları ile ilişkilidir. Başka bir ifade ile öğrencilerin duygusal durumları ile ilgili değişkenlerin, başarılarını yansıtan değişkenler kadar önemli olduğu görülmektedir. Desmarais ve Baker (2012) da yaptıkları tarama çalışmasında bu tür değişkenlerin en az öğrenci başarısı kadar önemli olduğunu belirtmişlerdir. Bu değişkenler aynı zamanda benzer veri setlerinde öğrenci performansını tahmin etmek amacıyla oluşturulan modellerde önemli değişkenler olarak belirlenen değişkenlerle de tutarlılık göstermektedir (Pardos vd., 2014; Pedro vd., 2013; Pedro vd., 2014). İleriki çalışmalarda bu değişkenlerin modellere katılması öğrenci performansını tahmin etmede başarıyı da artıracaktır.

Sınıf değişkeninin dengeli olarak dağılmadığı durumlar sınıflama algoritmalarının performansını olumsuz yönde etkilemektedir. Bu durumun önüne geçmek için dört farklı örneklem dengeleme yöntemi kullanılmış ve orijinal veri seti ile birlikte bu yöntemlerin algoritmaların tahmin performanslarına etkisi araştırılmıştır. Tablo 5 ve Şekil 6 incelendiğinde algoritmaların en iyi performansa yukarı örnekleme yöntemi kullanılan durumda ulaştığı görülmektedir. Diğer taraftan orijinal veri setinde STEM tercih eden öğrencileri tahmin etme performansının oldukça düşük olduğu görülmektedir. Bu sonuçlar veri dengeleme yöntemlerinin sınıflama algoritmalarının performansına etkisini göstermesi açısından önemlidir. Aşırı uyum sorunu yaşamamak için veri dengeleme algoritmalarının sadece eğitim veri setine uygulanmasına dikkat edilmelidir.

Eğitsel yazılımlardan elde edilen etkileşim verileri ile oluşturulan tahmin modelleri, öğrencilerin üniversite eğitimine devam etmelerine ya da STEM ile ilgili bir bölüm seçmelerinde yardımcı olabilecek kararların alınmasında bilinen tahmin modellerine göre daha önemli bilgiler sağlayacaktır (Pedro vd., 2013; Pedro vd., 2014). Bu amaçla, farklı yöntem ve tekniklerin eğitsel verilerde test edilerek daha yüksek tahmin oranına sahip yöntemlerin belirlenmesi önemlidir. Sunulan çalışma, kullanılan değişkenler ve algoritmalar ile sınırlıdır. İleriki araştırmalarda aynı veri seti üzerinde farklı ön işleme teknikleri, özellik belirleme yöntemleri ve algoritmalar test edilebilir.

Bilgilendirme

Prof. Neil T. Heffernan'a veri setinin araştırma amaçlı kullanımına izin verdiği için teşekkür ederiz. Çalışmada kullanılan veri setine https://goo.gl/forms/seAyF0aHUOxevhfF3 adresinden ulaşılabilir.

Kaynakça

Baker, R. S. J. d. (2007). *Modeling and understanding students' off-task behavior in intelligent tutoring systems*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, California, USA.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 853-867). Boston, MA: Springer US.

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1-2), 9-38. doi: 10.1007/s11257-011-9106-8

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction, 19*(3), 243-266. doi: 10.1007/s11257-009-9063-7

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470-497. doi: 10.1007/s40593-014-0024-x

Koedinger, K., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining, 43*. doi: citeulike-article-id:13242329

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *2016, 74*(7), 16. doi: 10.18637/jss.v074.i07

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *2008, 28*(5), 26. doi: 10.18637/jss.v028.i05

Olmo, J. L., Romero, C., Gibaja, E., & Ventura, S. (2015). Improving Meta-learning for Algorithm Selection by Using Multi-label Classification: A Case of Study with Educational Data Sets. *International Journal of Computational Intelligence Systems, 8*(6), 1144-1164. doi: 10.1080/18756891.2015.1113748

Pardos, Z. A., Baker, R. S. J. D., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *2014, 1*(1), 22. doi: 10.18608/jla.2014.11.6

Pedro, M. O., Baker, R., Bowers, A., & Heffernan, N. (2013). *Predicting college enrollment from student interaction with an intelligent tutoring system in middle school.* Paper presented at the Educational Data Mining 2013.

Pedro, M. O., Ocumpaugh, J., Baker, R., & Heffernan, N. (2014). *Predicting STEM and non-STEM college major enrollment from middle school interaction with mathematics educational software.* Paper presented at the Educational Data Mining 2014.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4, Part 1), 1432-1462. doi: doi.org/10.1016/j.eswa.2013.08.042

R Core Team. (2017). R: A language and environment for statistical computing: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-Validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 1-7). New York, NY: Springer New York.

San Pedro, M. O. C. Z., Baker, R. S. J. d., & Rodrigo, M. M. T. (2011). *Detecting Carelessness through Contextual Estimation of Slip Probabilities among Students Using an Intelligent Tutor for Mathematics*, Berlin, Heidelberg.

San Pedro, M. O. Z., Baker, R. S. J. d., Gowda, S. M., & Heffernan, N. T. (2013). Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. In H. C. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings* (pp. 41-50). Berlin, Heidelberg: Springer Berlin Heidelberg.

Stamper, J., Koedinger, K., Baker, R. S. J. d., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010). *PSLC DataShop: A Data Analysis Service for the Learning Science Community*, Berlin, Heidelberg.

Yu, H.-F., Lo, H.-Y., Hsieh, H.-P., Lou, J.-K., McKenzie, T. G., Chou, J.-W., . . . Lin, C.-J. (2010). *Feature Engineering and Classifier Ensemble for KDD Cup 2010*.