Öğrenme Yönetim Sistemi Log Kayıtlarının Akademik Başarı Tahmininde Kullanılması*

Araştırma Makalesi/Research Article

Mithat YAVUZARSLAN, Çiğdem Selçukcan EROL

Enformatik Bölümü, İstanbul Üniversitesi, İstanbul, Türkiye
mithat.yavuzarslan@ogr.iu.edu.tr, cigdems@istanbul.edu.tr

(Geliş/Received:08.12.2020; Kabul/Accepted:06.04.2022)
DOI: 10.17671/gazibtd.837884

Özet— Dünyada ve ülkemizde eğitim alanında dijitalleşme eğilimi arttıkça Öğrenme Yönetim Sistemleri (ÖYS) kullanımı da yaygınlaşmaktadır. Öğrenciler bu ortamlarla girdikleri etkileşimlerde kayda değer miktarda veri üretmekte ve bu veri üzerinde yapay zekâ algoritmaları kullanılarak öğrenme sürecini anlamaya dönük modeller geliştirilebilmektedir. Söz konusu modeller geliştirilirken eğitim ve öğrenme ortamına ait her türlü veri bu kapsama girebildiği gibi özellikle ÖYS'ler içerisindeki öğrenmeye harcanan zaman ve ders içeriğine erişim sıklığı gibi değişkenleri ölçmeye yarayan log (etkileşim) verisi öğrenme sürecinin anlaşılması bakımından büyük imkânlar barındırmaktadır. Bu çalışmada 2020 Bahar yarıyılı içerisinde açılan Temel Bilgisayar Uygulamaları dersine kayıtlı 93 öğrencinin 10 haftalık süre boyunca kullanmış oldukları Moodle tabanlı ÖYS içerisinden elde edilen log verisi üzerinde akademik başarı tahmini amacıyla KNN, Naive Bayes, SVM, CART ve C5.0 sınıflandırma algoritmaları uygulanmıştır. Elde edilen log dosyaları her bir öğrenci için ders ortamıyla olan etkileşimlerini ifade eden oturum açma sayısı, geçmiş konulara bakma sayısı, toplam ve ortalama görüntüleme sayısı, toplam ve ortalama oturum süresi, ödev materyalleri indirme sayısı, ödev deneme sayısı, ödeve harcanan zaman, sınav odaklı çalışma, eğitmene gönderilen mesaj sayısı, video sayfalarında geçirilen zaman ve yüklenen ödev sayısı özniteliklerine dönüştürülmüştür. Oluşturulan veri setinin dengesiz olmasından dolayı ayrıca yukarı örnekleme, SMOTE yöntemi ile sınıf örneklerini yakınlaştırma ve SMOTE yöntemi ile yukarı örnekleme yöntemleri kullanılarak 3 ayrı veri seti üzerinde de sınıflandırma algoritmaları uygulanmıştır. Çalışma sonucunda tüm veri setlerinde %80 üzeri sınıflandırma başarısına ulaşıldığı görülmüştür. En yüksek sınıflandırma başarıları SMOTE ile yukarı örnekleme uygulanan veri setinde negatif sınıfa ait örneklerin düşük miktarı ve benzer varyasyonların türetilmesi sebebiyle %100 başarı gösteren KNN algoritması göz ardı edildiğinde %97 başarı oranıyla CART ve SVM algoritmaları ile elde edilmiştir. Diğer yandan, Naive Bayes algoritmasının sonuçları daha güvenilir sayılabilecek olan rastgele alt örnekleme yöntemiyle en yüksek başarıyı gösterdiği görülmüştür. Sonuç olarak, ÖYS log kayıtlarının akademik başarı tahmininde kullanılabileceği görülmüş ve bulgular ilgili literatür ışığında tartışılmıştır.

Anahtar Kelimeler— veri madenciliği, sınıflandırma, öğrenme yönetim sistemleri, akademik başarı

Using Learning Management System Logs to Predict Undergraduate Students' Academic Performance

Abstract— Digitalization in education has fostered higher education institutes to use Learning Management Systems (LMS). Students are unintentionally generating large volumes of data known as logs while using LMSs, which can be utilized to build artificial intelligence models to predict educational variables. Unlike the ordinary web server logs, LMS log reports include information on students' specific interactions with the learning content/material, allowing variables to be created based on their interactions with the course. Five classification algorithms (KNN, Nave Bayes, SVM, CART, and C5.0) were used on the dataset created from the Moodle LMS log reports within a 10-week "Basic Computer Applications" course in which 93 undergraduate students registered. The log records for each student were transformed into a set of attributes that included the number of logins, material downloads, assignment attempts, uploaded assignments, messages sent to instructor, course page views, total time spent on the course page, average session time, total time spent on assignments and exams, and total time spent on video pages. Because the original dataset was imbalanced, over-sampling and SMOTE (Synthetic Minority Over-Sampling Technique) techniques were used to create three additional data sets besides the imbalanced dataset. The results showed that in each dataset, all classification performances were above 80%. If the KNN algorithm is ignored because of its extraordinarily high performance due to similar variations of the negative class generated by SMOTE technique, CART and SVM algorithms were found to be the most successful classifiers of students' academic performance with 97% accuracy. On the other hand, using random-sub-sampling technique which can be considered as more reliable, NB algorithm was found to be the most successful classifier. The findings of this study demonstrated that using classification algorithms, LMS logs can be utilized to predict academic performance of students.

Keywords— data mining, classification, learning management systems, academic performance

* Bu çalışma birinci yazar tarafından İstanbul Üniversitesi Fen Bilimleri Enstitüsü Enformatik programında hazırlanan doktora tezinden üretilmiştir ve özeti Future Learning 2020: 8th International Conference on Future Learning and Informatics: Data Revolution konferansında sunulmuştur.

1. GİRİŞ (INTRODUCTION)

Son yıllarda çevrimiçi ortamlarda kullanıcılar tarafından üretilen veri miktarının devasa boyutlara ulaştığı bilinmekte ve ham haliyle bir anlam ifade etmeyen verinin belirli bir amaç kapsamında ve uygun yöntemlerle işlenerek anlamlı bilgilere dönüştürülmesi ihtiyacı güncelliğini korumaktadır. Bu doğrultuda Veritabanlarında Bilgi Keşfi olarak adlandırılan ve verinin büyük oranda makine öğrenmesi teknikleri ile işlenerek karar verme süreçlerini destekleyici yeni bilgilere dönüştürülmesini ifade eden veri madenciliği uygulamaları sağlık, bankacılık, ticaret gibi alanlarda sıklıkla kullanılmaktadır [1]. Benzer şekilde Eğitsel Veri Madenciliği başlığı altında yürütülen çalışmalar da temel olarak öğrenme deneyim ve materyallerini geliştirmek ve öğretim tasarımı süreçlerinin düzenlenmesi ve planlanması konusundaki kararlara yol gösterici olmak amaçlarını taşımaktadırlar [2]. Eğitsel ortamlardan elde edilen her türlü veri üzerinde yürütülen veri madenciliği çalışmalarında kümeleme, sınıflandırma, uç değer tespiti, birliktelik kuralları analizi ve metin madenciliği başlıca kullanılan yöntemler olarak belirtilmekte ve yükseköğretim düzeyinde yapılan çalışmalarda en sık kullanılan yöntemin ise tahmine dayalı yöntemler oldukları görülmektedir [3, 4].

Hem e-öğrenme hem de yüz yüze öğrenme ortamları kapsamında yürütülen çalışmalarda tahmin edilen değişkenlerin başında akademik başarı gelmektedir [4]. Kimi çalışmalar öğrencilerin demografik bilgileri, eğitim geçmişleri, sosyo-ekonomik durum veya öğrenme ile ilgili bir ankete verilen cevaplar gibi çeşitli kaynaklardan elde edilen veriyi akademik başarı tahmininde kullanırken [5-9] özellikle Öğrenme Yönetim Sistemleri'nin (ÖYS) yaygınlaşmasıyla birlikte aynı amacı güden çalışmalarda bu sistemlere ait veritabanlarında depolanan log kaydı verisi de tahminleyici modellerin geliştirilmesi amacıyla sıklıkla kullanılmıştır [10]. Log kayıtları, kullanıcıların bir web sunucusunda yer alan sayfa ve kaynaklardan hangilerine ne zaman eriştiği bilgilerini gösteren dosyalardır [11] ve bu doğrultuda ÖYS log kayıtları öğrencinin derse ayırdığı zaman veya derse dönük ilgisi gibi e-öğrenme ortamları söz konusu olduğunda yüz yüze eğitime kıyasla takibi zor olan bilgilerin ortaya çıkarılması açısından büyük imkânlar barındırmaktadırlar.

ÖYS veya benzer çevrimiçi öğrenme ortamlarından elde edilen log kayıtları kullanılarak tahminleyici modellerin geliştirildiği çalışmalarda da akademik başarının en fazla tahmin edilen değişken olduğu görülmektedir[10, 12-19]. Ayrıca öğrenmeye dönük yaklaşım [20], öğrenme stilleri [11, 21] ve motivasyon [22, 23] gibi değişkenler de ÖYS log kayıtlarının temel alındığı çalışmalarda tahmin edilen değişkenler arasındadır.

Bu çalışmada bir ÖYS'den elde edilen log kayıtları kullanılarak üniversite öğrencilerinin akademik başarılarının tahmin edilmesi amaçlanmıştır. Dersin yürütüldüğü ÖYS, eğitsel içerik yayınlama ve yönetme amacıyla sıklıkla kullanılan ücretsiz bir hizmet olan Moodle ile yapılandırılmıştır. Moodle içerisinde yer alan log kayıtları doğrudan makine öğrenmesi algoritmalarının uygulanması için uygun olmadıklarından dolayı bu kayıtların sınıflandırma algoritmalarına uygun ve her bir öğrencinin dersin ÖYS'si ile olan etkileşimlerini ifade eden özniteliklere dönüştürülmeleri gerekmektedir [10]. Çalışma kapsamında derse ait ÖYS log kayıtlarından elde edilen özniteliklere yaş ve cinsiyet bilgilerinin de eklenmesiyle 13 adet özniteliğe sahip bir veri seti oluşturulmuştur. Akademik başarı değişkeninin tahmin edilmesi amacıyla ilgili veri seti üzerinde 5 ayrı sınıflandırma algoritması uygulanarak aşağıdaki sorulara cevap aranmıştır:

- Akademik başarının tahmin edilmesinde en yüksek başarıyı hangi sınıflandırma algoritması sergilemektedir?
- Akademik başarının tahmin edilmesi amacıyla geliştirilen modellerin performansı benzer çalışmalardaki modellerin performanslarıyla paralellik göstermekte midir?

2. YÖNTEM VE ARAÇLAR (METHODS AND INSTRUMENTS)

2.1. Veri Seti (Data Set)

Araştırmada kullanılan veri seti 2020 bahar döneminde temel bilgisayar kullanım becerilerini kazandırmayı amaçlayan Temel Bilgisayar Uygulamaları dersine kayıtlı 93 üniversite öğrencisinin dersin yürütüldüğü ÖYS içerisindeki hareket ve etkileşimlerini gösteren log kayıtları temel alınarak oluşturulmuştur. Derse kaydolan öğrencilerin yarıya yakını Mühendislik fakültesine kayıtlı öğrenciler olup geri kalanı ise Fen Edebiyat, İktisat, Ticari Bilimler, İletişim, Mimarlık ve Eğitim fakültelerine kayıtlıdır. Öğrenciler 10 haftalık süre içerisinde ÖYS aracılığıyla videolar, ödev sayfaları, ders anlatımı dosyaları, açıklayıcı görseller ve duyurular gibi içeriklere erişmişlerdir.

Moodle tabanlı ÖYS'ye ait ilişkisel verirtabanında depolanan log kayıtlarının filtrelenebilen bir sorgulamasına sistem yöneticisi arayüzünden erişilebilmekte ve öğrencilerin olay zamanı (Time), kullanıcı adı (User full name), etkilenen kullanıcı (Affected user), olay bağlamı (Event context), bileşen (Component), olay adı (Event Name), açıklama (Description), erişim kaynağı (Origin) ve IP adresi bilgileri görülebilmektedir (Şekil 1).

[31/05/20 03:12:30 PM, student1,- , Assignment: Upload Excel Assignment 3 Here, Assignment, Course module viewed, The user with id '87' viewed the 'assign' activity with course module id '131', web, 1xx.xx.1xx.1xx]

Şekil 1. Moodle log kaydı örneği (Example of a log record in Moodle)

Filtreleme sonucunda öğrencilerin dönem boyunca sistem içerisinde sergiledikleri hareketlerin depolandığı 31,573 kayıttan oluşan log dosyasına CSV (Comma Seperated Values) formatında erişilmiş ve ilk aşamada her bir öğrenci için Tablo 1'de açıklamaları verilen öznitelikler oluşturulmuştur.

Tablo 1. Veri seti öznitelikleri ve açıklamaları (Data set attributes and descriptions)

Öznitelik	Açıklama
OturumAcmaSayisi (*login_Count*)	Öğrencinin açtığı toplam oturum sayısı
GecmisKonularaBakmaSayisi (*pastSubjectsView_Count*)	Öğrencinin belirli bir haftada o hafta hariç geçmiş haftaların konularına bakma sayısı
ToplamGoruntulemeSayisi (*totalView_Count*)	Öğrencinin sistemdeki toplam sayfa görüntüleme sayısı
OrtalamaGoruntulemeSayisi (*averageView_Count*)	Öğrencinin toplam oturum açma sayısının toplam gezinme sayısına bölünmesi ile elde edilmiştir
ToplamOturumSuresi (*totalSessionTime*)	Öğrencinin sistemde geçirdiği toplam süre (saniye)
OrtalamaOturumSuresi (*averageSessionTime*)	Öğrencinin bir oturumda sistemde ortalama olarak geçirdiği süre toplam oturum süresinin oturum açma sayısına bölümünden elde edilmiştir (saniye)
OdevMateryalleriIndirmeSayisi (*assignmentMaterialsDown_Count*)	Paylaşılan ödev materyallerinin öğrenci tarafından indirilme sayısı
OdevDenemeSayisi (*assignmentAttempt_Count*)	Ödev yönerge, açıklama ve yükleme sayfalarının görüntülenme sayısı
OdeveHarcananZaman (*assignment_Time*)	Öğrencinin her bir ödevin ilgili haftası içerisinde ödev yönerge ve ayrıntılarını içeren sayfalarda geçirdiği süre (saniye)
SınavOdakliCalisma (*examOriented_Time*)	Öğrencinin derslerin son tarihinden sınav gününe kadar olan sınav haftası içerisinde sistemde geçirdiği süre (saniye)
VideoSuresi (*video_time*)	Öğrencinin videoların yayınlandığı sayfalarda geçirdiği süre (saniye)
EgitmeneGonderilenMesajSayisi (*messageToInstructor_Count*)	Öğrencinin dönem içerisinde dersin eğitmenine gönderdiği toplam mesaj sayısı
YuklenenOdevSayisi (*assignmentUpload_Count*)	Öğrencinin dönem içerisinde sisteme yüklediği ödev sayısı
Yas (*age*)	Öğrencinin yaşı
Cinsiyet (*gender*)	Öğrencinin cinsiyeti
Hedef Nitelik: SONNOT (*finalScore*)	Öğrencinin dönem sonu notu

Veri setindeki özniteliklerin oluşturulması ve değerlerinin hesaplanması amacıyla Microsoft Excel programının 2016 sürümü içerisindeki koşullu olarak toplam, ortalama veya miktar hesaplayan fonksiyonlar kullanılmıştır. Benzer çalışmalar incelendiğinde log kayıtlarından elde edilen özniteliklerin dersin içeriğine ve izlencesine göre değişmekle birlikte tamamlanan ödev sayısı, forumda harcanan süre, ödevler için harcanan süre, toplam ziyaret sayısı, yüklenen dosya sayısı gibi genel olarak öğrencilerin sistemle olan etkileşimlerini miktar ve süre cinsinden ifade eden değişkenler oldukları görülmektedir (örn. [15, 18]). Diğer yandan süre cinsinden hesaplanan değişkenlerle öğrencilerin belirli bir zamanda eriştikleri sayfadaki içerikle gerçek anlamda ilgilenip ilgilenmediklerini kesin olarak bilme olanağı bulunmamaktadır. Zhou ve diğerleri [24], log kayıtları aracılığıyla kullanıcıların bir sayfada harcadığı süreyi kesin olarak hesaplamanın kullanıcının ilgili sayfaya eriştiği süre içerisinde bilgisayarı terk edebilme ve başka bir program veya web sayfasına erişebilme ihtimallerinden dolayı kısıtlılıklara sahip olduğunu vurgulamakla birlikte eğer hesaplamalar oturum (session) süreleri göz önüne alınarak yapılırsa bu sürenin sayfalarda geçirilen gerçek süreyi varsayımsal olarak ortaya çıkarabileceğini belirtmişlerdir. Bu araştırmadaki veri setinde bir veya birden fazla sayfada geçirilen süre o sayfalara giriş ve çıkış kayıtlarını gösteren iki eylem arasındaki sürenin saniye cinsinden farkı alınarak hesaplanmış ve Moodle log dosyaları zaman aşımı kayıtlarını göstermediğinden dolayı uzun süreli eylemsizlik saptanan kullanıcı kayıtlarında derse ait ÖYS'nin maksimum oturum süresi olan 2 saatlik sürenin geçirildiği varsayılmıştır.

2.2. Veri Ön İşleme (Data Pre-processing)

Bu araştırmadaki veri ön işleme, modelleme ve performans değerlendirme süreçleri için R programlama dili ve paketleri kullanılmıştır. R dili; veri madenciliği ve makine öğrenmesi çalışmalarında sıklıkla kullanılmakta ve bu amaçla yürütülecek veri analizi, istatistiksel programlama, modelleme, veri görselleştirme vb. işlemler için gerekli paketlere açık kaynak ve ücretsiz olarak erişim imkânı sunmaktadır [25]. Veri ön işleme süreci aşağıdaki sırayla gerçekleştirilmiştir:

1- Veri setinde 11 öğrenciye ait yaş ve cinsiyet değerlerinin kayıp olduğu gözlenmiştir. Cinsiyet değişkenine ait kayıp değerler dersin öğrenci bilgi sistemi sayfasında yer alan bilgilerden yola çıkılarak tamamlanmış olup yaş değerleri ise kayıp değerler için bir yaklaşık değer atama yöntemi olan ve R kütüphanesine eklenen "mice" (Multivariate Imputation by Chained Equations) paketi kullanılarak doldurulmuştur [26].

2- Öznitelikler arasındaki korelasyonun incelenmesi amacıyla Pearson Korelasyon katsayısı hesaplanarak katsayı (p) değerleri .80'den yüksek olan ve birden fazla öznitelikle pozitif korelasyon gösteren *ToplamGoruntulemeSayisi*, *ToplamOturumSuresi* ve *VideoSuresi* öznitelikleri veri setinden çıkarılmıştır.

3- Geliştirilecek modelin hedef niteliği olan ve öğrencilerin dönem sonu notlarını içeren *SONNOT* vektörüne ait değerler dersin geçme notu olan 50 sınırı göz önüne alınarak "0: Kaldı, 1: Geçti" şeklinde ikili sınıfa indirgenmişlerdir:

$$SONNOT_i = \begin{cases} 1 \; eğer \; SONNOT_i \geq 50 \\ 0 \; eğer \; SONNOT_i < 50 \end{cases}$$

Eşitlik 1. SONNOT Vektörünün İkili Sınıfa İndirgenmesi (Reducing the Values in finalScore Vector to Binary Classes)

Eşitlik 1'de gösterildiği şekilde yapılan indirgeme işlemi sonucunda hedef niteliğe ait sınıfların dağılımları $N_1 = 82$ ve $N_0 = 11$ olarak saptanmıştır.

4- Özniteliklere ait numerik değerleri 1 ve 0 arasında ölçeklendirmek için min-max normalizasyonu yöntemi uygulanmıştır:

$$s' = \frac{s - min}{maks - min}$$

Eşitlik 2. Min-Max Normalizasyonu (Min-Max Normalization)

Eşitlik 2'de s' normalize edilmiş değerleri, s ise gerçek değerleri ifade etmektedir. min değeri bir öznitelik değerinin görülen en düşük değeri ve $maks$ ise o özniteliğe ait en yüksek değeri göstermektedir.

5- Hedef niteliğin ikili sınıfa indirgenmesi işlemi sonrası veri setinin dengesiz hale geldiği görülmüştür. Dengesiz veri seti, sınıflandırılacak kategorilerin birbirlerine yakın veya eşit sayıda dağılmadığı durumlarda ortaya çıkmaktadır [27]. Dengesiz dağılan veri setini dengeli hale getirmek için uygulanan başlıca yöntemler çoğunluk sınıfın sayılarının azınlık sınıfa yaklaştırıldığı aşağı örnekleme (undersampling), azınlık sınıfın sayılarının çoğunluk sınıfa yaklaştırıldığı yukarı örnekleme (oversampling) ve azınlık sınıfın örneklerinin sentetik olarak çoğaltıldığı Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) olarak sıralanmaktadır [28]. Kullanılan veri setinde azınlık sınıfa ait 11 örnek bulunduğu için aşağı örnekleme yöntemi ile elde edilecek toplam örnek sayısının bir tahmin modeli geliştirmek için yeterli olmayacağı görüldüğünden analizler dengesiz, rastgele alt örneklemeli, yukarı örneklemeli, SMOTE ile sınıf örnek sayılarının birbirlerine yaklaştırıldığı ve yine SMOTE ile azınlık sınıfın örnek sayılarının çoğunluk sınıf sayılarına eşitlendiği veri setleri üzerinde gerçekleştirilmiştir. Tekrarlı hold-out yöntemi olarak da bilinen rastgele alt örnekleme yönteminde kullanılan tekrar/iterasyon sayısı 200 olarak belirlenmiştir. Süreç sonunda tüm iterasyonlara ait değerlendirme ölçütü puanlarının aritmetik ortalaması hesaplanmıştır. Yukarı Örnekleme, SMOTE ile Yakınlaştırma ve SMOTE ile Yukarı Örnekleme yöntemleriyle elde edilen veri setleri uygulanan sınıflandırma algoritmalarına göre farklılaşmayıp bu algoritmalar ilgili kategori içindeki aynı veri setine uygulanmıştır. Farklı örnekleme yöntemleri sonucunda oluşan hedef nitelik sınıflarına ait frekansların dağılımı Tablo 2'de gösterilmiştir.

Tablo 2. Kullanılan örnekleme yöntemleri ve frekans dağılımları
(Sampling methods and class frequencies for data set)

Örnekleme Yöntemi	Frekans
Dengesiz Veri (N=93)	$N_1 = 82$, $N_0 = 11$
Rastgele Alt Örnekleme (N=93)	$N_1 = 82$, $N_0 = 11$
Yukarı Örnekleme (N=164)	$N_1 = 82$, $N_0 = 82$
SMOTE ile Yakınlaştırma (N=101)	$N_1 = 57$, $N_0 = 44$
SMOTE ile Yukarı Örnekleme (N=164)	$N_1 = 82$, $N_0 = 82$

2.3. Modelleme (Modelling)

Tahminleyici modelin geliştirilmesi aşamasında aynı probleme çözüm sunabilen birden fazla makine öğrenmesi algoritması bulunduğu için çeşitli algoritmalar aynı veriyle eğitilerek hedef niteliği tahmin edebilme başarıları karşılaştırılır [29]. Bu doğrultuda farklı yöntemlerle örneklenmiş olan veri seti üzerinde k-en yakın komşu (KNN), Naive Bayes (NB), Destek Vektör Makineleri (SVM), Sınıflandırma ve Regresyon Ağaçları (CART) ve C5.0 sınıflandırma algoritmaları uygulanmıştır.

1- k-en yakın komşu (KNN) algoritması Cover ve Hart [30] tarafından ortaya konmuş olup yeni bir örneğin sınıflandırmasında eğitim kümesinde yer alan örneklere olan benzerliği veya uzaklığını temel almaktadır. k-En Yakın Komşu algoritmasının adımları aşağıda verilmiştir:

- k parametresinin (sınıflandırmada temel alınacak en yakın komşu sayısının) belirlenmesi
- Her test kümesi örneği ile eğitim kümesi örnekleri arasındaki uzaklığın hesaplanması
- En yakın komşuların k parametresine göre belirlenmesi ve uzaklıklarının sıralanması
- En yakın komşulara ait kategorik çoğunluğun test kümesi örneğindeki değerin sınıflandırılması için kullanılması [31].

Bu çalışmada yürütülen k-en yakın komşu algoritması uygulamalarında k parametresi 3 olarak belirlenmiş ve uzaklık ölçütü olarak Öklid uzaklığı kullanılmıştır:

$$d(x,y) = \left(\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \right)$$

Eşitlik 3. Öklid Uzaklığı (Euclidian Distance)

Eşitlik 3'te x_i ve y_i bir uzaydaki iki örneğe ait olan noktaları göstermektedir.

2- Naive Bayes (NB) algoritması istatistik temelli bir sınıflandırma yöntemidir ve bir örneğin tanımlı sınıflardan hangisine ait olduğu her bir sınıf için olasılıksal olarak hesaplanır ve örnek yüksek olasılığa sahip sınıfta etiketlenir [32].

$$P(c|x) = \frac{P(x|c)\ P(c)}{P(x)}$$

Eşitlik 4. Naive Bayes Formülü (Naive Bayes Formula)

Eşitlik 4'te $P(c|x)$ bir x örneğinin c sınıfında olma olasılığını, $P(x|c)$ c sınıfında x örneklerinin görülme sıklığını, $P(c)$ c sınıfının genel olasılığını ve $P(x)$ ise x örneğinin eğitim verisi içerisindeki görülme sıklığını ifade etmektedir.

3- Destek Vektör Makineleri (SVM), Vapnik [33] tarafından ikili ve çoklu sınıflandırma problemlerinin

çözümü için önerilmiştir ve örneklerin ait olduğu sınıfları birbirinden ayıracak en uygun hiper-düzlemi elde etmeye çalışır. Doğrusal (lineer) olarak ayrılabilen sınıflandırma problemleri için karar fonksiyonu aşağıda verilmiştir [34]:

$$f(x) = sign\left(\sum_{i=1}^{\ell} y_i\ \lambda_i\,(\mathrm{x}\cdot \mathrm{x}_i) + b\right)$$

Eşitlik 5. DVM Karar Fonksiyonu (Decision Function of SVM)

Eşitlik 5'te x eğitim seti girdilerini, y sınıf etiketi çıktılarını ve b ise bias değerini ifade etmektedir.

4- CART algoritması sınıflandırma problemlerinin çözümü için ikili ağaç yapısını kullanan bir karar ağacı yöntemidir. Ağaç yapısının oluşumunda dalların hangi değişkenden ne şekilde ayrılacağının belirlenmesinde Gini indeksi kullanılmaktadır:

$$Gini_{sol} = 1 - \sum_{i=1}^{k}\left(\frac{L_i}{T_{sol}}\right)^2$$

$$Gini_{sağ} = 1 - \sum_{i=1}^{k}\left(\frac{R_i}{T_{sağ}}\right)^2$$

$$Gini_j = \left(\frac{1}{n}|T_{sol}|Gini_{sol} + |T_{sağ}|Gini_{sağ}\right)$$

Eşitlik 6. Gini İndeksi (Gini Index)

Eşitlik 6'da n örnek sayısını, k sınıf sayısını, T_{sol} ikili ağacın sol düğümündeki toplam örnek sayısını, $T_{sağ}$ ikili ağacın sağ düğümündeki toplam örnek sayısını, L_i sol düğümde i kategorisine ait olan örnek sayısını, R_i ise sağ düğümde i kategorisine ait olan örnek sayısını ifade etmektedir. İlk ayrım noktası olan kök düğüm belirlendikten sonra $Gini_j$ değeri en düşük olan değişken düğüm olarak seçilir ve ağacın yapılandırılması tüm yapraklara ulaşılana kadar kalan nitelikler için Gini indeksi hesaplanarak devam ettirilir [35].

5- C5.0 algoritması öncülü olan C4.5 algoritmasının kurallarını barındırmakla birlikte sınıflandırmada etkisiz olan özniteliklerin çıkarılması ve böylece daha küçük boyutlu ağaçlar ile daha anlaşılır kuralların yapılandırıldığı bir karar ağacı yöntemidir [36]. Ağacın oluşturulmasında ayırıcılığı en yüksek olan özniteliğin bulunmasında bilgi kazanımı (entropi) hesaplanır ve en yüksek kazanca sahip öznitelik ayrım için kullanılır.

$$E(S) = \sum_{i=1}^{m} -p_i\ \log_2(p_i)$$

Eşitlik 7. Bilgi Kazanımı (Entropi) Formülü (Information Gain Formula)

Eşitlik 7'de S eğitim setindeki örnekleri, m hedef niteliğe ait sınıf sayısını, p_i ise entropisi hesaplanan sınıfın eğitim setindeki örnek sayısının tüm örnek sayılarına oranını ifade etmektedir [37].

Bu aşamadan sonra tahminleyici modellerin seçilen algoritmalar kullanılarak eğitilmesi ve performanslarının değerlendirilmesi gerekmektedir. Bir performans değerlendirme yöntemi olan hold-out (dışarıda tutma), veri setindeki örneklerin bir defa seçilecek ve birbirleriyle örtüşmeyecek şekilde eğitim ve test kümelerine ayrılarak eğitim kümesinin modelin eğitilmesinde, test kümesinin ise sınıflandırma oranının kontrol edilmesinde kullanıldığı bir yöntemdir [38]. Bu çalışmada uygulanan sınıflandırma algoritmalarının performanslarının değerlendirilmesi için hold-out yöntemi seçilmiş olup veri setindeki örnekler %70 eğitim kümesi ve %30 test kümesi oranlarında bölünmüştür.

Sınıflandırma başarılarının ölçülmesinde ise aşağıda açıklamaları verilen doğruluk (accuracy), duyarlılık (sensitivity), özgüllük (specifity) ve F1-Skoru ölçütleri kullanılmıştır:

- Doğruluk değeri modelin doğru tahmin ettiği örnek sayısının veri setindeki tüm örneklere bölünmesiyle elde edilmektedir.
- Duyarlılık ya da geri çağırma (recall) ve özgüllük (specifity) değerleri ise sınıflandırma başarısını pozitif ve negatif sınıflar için ayrı olarak göstermektedir. Bu iki değer ilgili sınıfın doğru tahmin edilen örnek sayısının o sınıfa ait tüm gerçek örnek sayısına bölünmesiyle hesaplanmaktadır.
- F1-Skoru ise yalnızca pozitif veya negatif sınıfın başarısının değil tüm sınıfların duyarlılık ve kesinlik değerlerinin ağırlıklı ortalamaları hesaplanarak göz önüne alındığı bir performans değerlendirme ölçütüdür [39]. Kesinlik (precision) değeri pozitif sınıf tahminlerinin ne kadarının doğru tahminler olduğunu göstermekte ve doğru tahmin edilen pozitif sınıf örnekleri sayısının tüm pozitif sınıf tahminleri sayısına bölünmesiyle elde edilmektedir.

3. BULGULAR (FINDINGS)

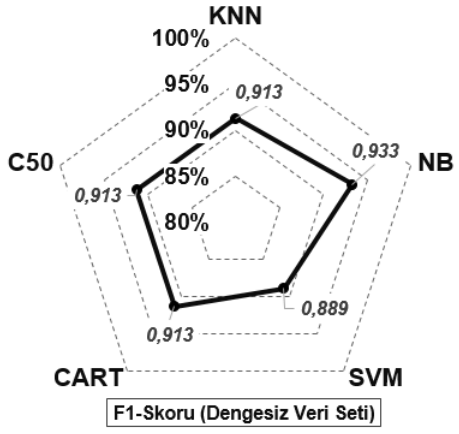
Tablo 3'te veri setinin farklı örneklemeleri üzerinde uygulanan algoritmaların öğrencilerin akademik başarılarını sınıflandırma performansları doğruluk, duyarlılık, özgüllük, kesinlik ve F-1 Skoru ölçütlerine ait değerleri gösterecek şekilde sunulmuştur.

Tablo 3 incelendiğinde algoritmaların sınıflandırma performanslarına ait Doğruluk ve F-1 Skoru değerlerinin tümünün %80'in üzerinde olduğu saptanmıştır. Aynı örnekleme yöntemleri birbirleriyle kıyaslandığında; Dengesiz veri setinde ve Rastgele Alt Örnekleme uygulanan veri setinde NB, Yukarı Örnekleme uygulanan veri setinde SVM, SMOTE ile Yakınlaştırma yönteminin kullanıldığı veri setinde aynı değerlere sahip olmak üzere NB dışındaki tüm algoritmalar ve SMOTE ile Yukarı Örnekleme yönteminin kullanıldığı veri setinde ise SVM

ve CART algoritmalarının en yüksek performansa sahip oldukları görülmüştür.

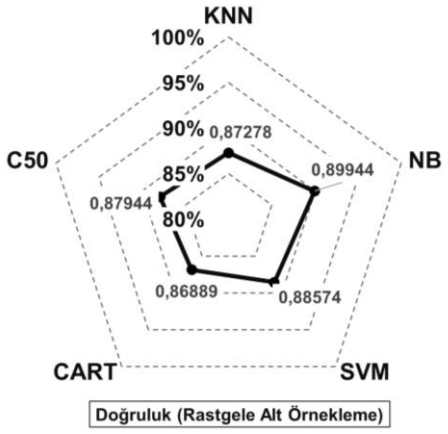
Tablo 3. Akademik başarı tahmini sonuçları (Results of the predictions for academic performance)

		Doğr.	Duy.	Özg.	Kes.	F1
Dengesiz Veri (N=93)	KNN	0,85	0,87	0,66	0,95	0,91
	NB	0,88	0,87	1	1	0,93
	SVM	0,81	0,83	0,66	0,95	0,88
	CART	0,85	0,87	0,66	0,95	0,91
	C5.0	0,85	0,87	0,66	0,95	0,91
Rastgele Alt Örnekleme (N=93)	KNN	0,87	0,94	0,31	0,92	0,93
	NB	0,89	0,91	0,78	0,97	0,94
	SVM	0,88	0,94	0,45	0,93	0,94
	CART	0,86	0,91	0,52	0,94	0,92
	C5.0	0,87	0,93	0,46	0,93	0,93
Yukarı Örnekleme (N=164)	KNN	0,89	0,91	0,87	0,88	0,89
	NB	0,87	0,87	0,87	0,87	0,87
	SVM	0,95	0,91	1	1	0,95
	CART	0,93	0,87	1	1	0,93
	C5.0	0,97	0,95	0,87	0,88	0,92
SMOTE ile Yakınlaştırma (N=101)	KNN	0,86	0,76	1	1	0,86
	NB	0,83	0,76	0,92	0,92	0,83
	SVM	0,86	0,76	1	1	0,86
	CART	0,86	0,76	1	1	0,86
	C5.0	0,86	0,76	1	1	0,86
SMOTE ile Yukarı Örnekleme (N=164)	KNN	1	1	1	1	1
	NB	0,93	0,95	0,91	0,92	0,93
	SVM	0,97	0,95	1	1	0,97
	CART	0,97	0,95	1	1	0,97
	C5.0	0,89	0,91	0,87	0,88	0,89



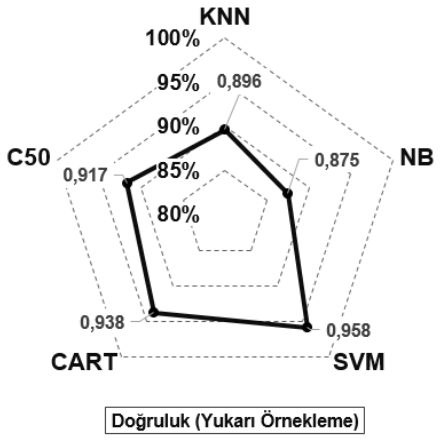
Şekil 2. Dengesiz Veri Seti'ne Ait F-1 Skoru Sonuçları (F-1 Scores of Imbalanced Data Set)

F1-Skoru ölçütünün özellikle dengesiz halde bulunan ve ikili sınıflandırma problemi içeren durumlarda başarının saptanmasına dönük işlevsel bir değerlendirme kriteri olarak kullanıldığı bilinmektedir [40]. Şekil 2'de gösterilen Dengesiz veri seti üzerinde uygulanan sınıflandırma algoritmalarının F-1 Skoru sonuçları karşılaştırıldığında en yüksek başarıyı gösteren algoritmanın NB olduğu görülmüştür.



Şekil 3. Rastgele Alt Örnekleme ile Oluşturulan Veri Seti'ne Ait Doğruluk Değerleri (Accuracy Scores of Random Subsampled Data Set)

Şekil 3'te Rastgele Alt Örnekleme yöntemiyle elde edilen veri setine ait doğruluk puanlarının birbirine yakın değerlere sahip olduğu ve bir defa örneklenen dengesiz veri setinde olduğu gibi NB algoritmasının en yüksek başarıyı sergilediği görülmektedir.

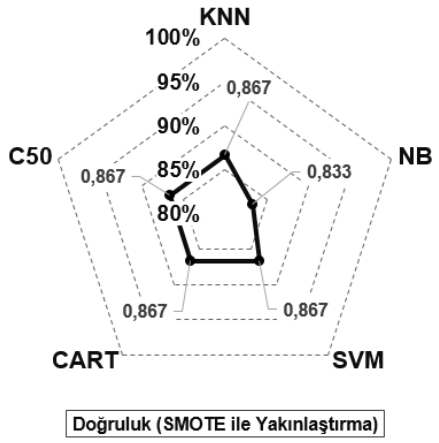


Şekil 4. Yukarı örnekleme ile oluşturulan veri seti'ne ait doğruluk değerleri (Accuracy scores of oversampled data set)

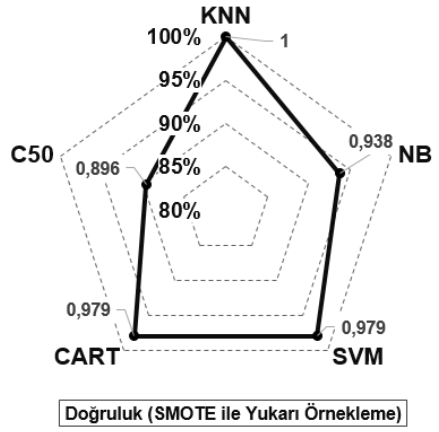
Şekil 3, 4, 5 ve 6'da bu çalışmada kullanılan algoritmaların doğruluk değerleri farklı örnekleme yöntemlerine göre kıyaslanmaktadır. Örnekleme yöntemleri göz önüne alındığında SMOTE ile Yukarı Örnekleme yönteminin kullanıldığı veri setine uygulanan KNN, NB, SVM ve CART algoritmalarının Doğruluk değerlerinin diğer örnekleme yöntemleriyle elde edilen değerlere göre daha yüksek oldukları görülmektedir.

4. TARTIŞMA VE SONUÇ (DISCUSSION AND CONCLUSION)

Bu çalışmada üniversite öğrencilerinin akademik başarıları Moodle tabanlı bir ÖYS ile olan etkileşim kayıtlarının tutulduğu log dosyaları kullanılarak tahmin edilmeye çalışılmıştır. Log dosyaları içindeki kayıtlar öğrencilerin akademik başarı hedefleriyle ilişkili olduğu düşünülen özniteliklere dönüştürülmüş ve örnekleme yöntemine göre farklılaşan veri setlerine uygulanan sınıflandırma algoritmalarının (KNN, NB, SVM, CART, C5.0) performansları karşılaştırılmıştır.



Şekil 5. SMOTE ile yakınlaştırma yöntemi ile oluşturulan veri seti'ne ait doğruluk değerleri
(Accuracy scores of data set whose class values were approximated with SMOTE)



Şekil 6. SMOTE ile yukarı örnekleme yöntemi ile oluşturulan veri seti'ne ait doğruluk değerleri
(Accuracy scores of data set oversampled with SMOTE)

Sonuçlar benzer çalışmalarda olduğu gibi ÖYS log kayıtlarının makine öğrenmesi teknikleriyle işlenerek akademik başarı tahmininde kullanılabileceğini göstermiştir [10, 12, 13, 15, 17, 18]. Moodle tabanlı ÖYS'ler dışında farklı sistem ve çevrimiçi öğrenme ortamlarından elde edilen log kayıtlarının da akademik başarı tahmininde paralel sonuçlara ulaştıkları görülmektedir [14, 19]. Log kayıtları ile akademik başarının tahmin edilmesini amaçlayan çalışmaların öğrencilerin dersle olan etkileşim düzeylerinin daha iyi anlaşılmasına olanak sağladığı gibi özellikle akademik başarısı düşük olması muhtemel öğrencilerin erkenden saptanarak gerekli önlemlerin alınması doğrultusunda da eğitmenlere yol gösterici olabileceği belirtilmektedir [13].

Bu çalışmada SMOTE ile yukarı örnekleme yöntemiyle elde edilen puanlar daha yüksek olsa da Rastgele örnekleme yöntemiyle gerçekleştirilen tekrarlı deneylerin doğruluk puanlarının daha güvenilir olduğu düşünülmektedir. Özellikle düşük miktarda örnekleme sahip ve sınıf etiketleri dengesiz dağılım gösteren veri setlerinde eğitim ve test süreçlerini rastgele ve farklı tekrar/iterasyonlar ile gerçekleştirmenin verinin etkili kullanımını sağladığı ve daha açıklayıcı sonuçlar elde edildiği görülmektedir [42].

Akçapınar [17] tarafından farklı bir çevrimiçi öğrenme ortamı kullanılarak yürütülen benzer bir çalışmada da en yüksek sınıflandırma başarısının KNN algoritması ile elde edildiği görülürken birden fazla çalışmada ise log kayıtları ile akademik başarının tahmin edilmesinde Karar Ağaçları temelli CART, C4.5 ve J48 algoritmaları öne çıkmaktadır [13, 14, 18, 42]. Romero ve diğerleri [15], çalışmalarının sonucuna dayanarak akademik başarı tahmininde en iyi sınıflandırma performansına sahip olan tek bir algoritma bulunmadığını vurgulamışlar fakat etkililiği ve karar vericiler tarafından anlaşılabilir olması açısından ise karar ağaçları başta olmak üzere kural çıkarımı ve bulanık algoritmaların kullanımını önermişlerdir. Diğer yandan bir e-öğrenme ortamının log kayıtları ile akademik başarı tahmini amacıyla çalışma yürüten Bravo ve Ortigosa [14] ise karar ağaçları ve benzeri yöntemlerde verinin dağılım ve bölünme oranlarına bağımlı olarak gerçekleşen küçük değişikliklerin birbirinden farklı sonuçlara sahip modeller ortaya çıkardığını vurgulayarak bu yöntemlerin zayıflıklarına dikkat çekmişlerdir. En yüksek sınıflandırma performanslarının gözlemlendiği SMOTE ile Yukarı Örnekleme uygulanan veri setinde KNN algoritması aşırı uyum göstermesi nedeniyle göz ardı edildiğinde aynı veri setinde CART algoritmasının elde ettiği yüksek performans literatürdeki bulgulara paralel olarak karar ağacı algoritmalarının log kayıtları ile akademik başarının tahmin edilmesini amaçlayan uygulamalardaki etkililiğini göstermiştir. C5.0 algoritmasının performans değerlerinin CART algoritmasıyla karşılaştırıldığında daha düşük olması ise CART algoritmasının ikili ağaç yapısının bu çalışmadaki ikili sınıflandırma problemiyle daha uyumlu olması ile açıklanabilir.

Çalışmanın sonuçlarıyla paralel olarak SMOTE yöntemi ile dengeli hale getirilen veri setlerinin daha yüksek sınıflandırma başarısına sahip tahminleyici modeller oluşturdukları görülmektedir [27]. Bu noktada azınlık sınıfa ait örneklerin doğrudan kopyalanarak çoğaltıldığı yukarı örnekleme yöntemine kıyasla SMOTE ile azınlık sınıfa ait sentetik örneklerin oluşturulduğu yukarı örnekleme sınıflandırmada yüksek puanların elde edilmesinde etkili olduğu söylenebilir. Ayrıca Romero ve diğerleri [15] de, yukarı örnekleme yöntemiyle 25

sınıflandırma algoritmasının 17'sinde dengesiz ve kategorik değerlere indirgenen veri setlerine göre daha iyi performans elde edildiğini belirtmişlerdir fakat KNN ve CART algoritmaları performansı düşüş gösteren algoritmalar arasındadır. Bu çalışmada ise yukarı örnekleme uygulanmış veri setinde KNN ve CART algoritmalarının Doğruluk skorları dengesiz veri setine göre sırasıyla %5 ve %10 oranında artmıştır. Kullanılan veri setinin dengesizbir veri seti olmasının yanında dersten kalan öğrencilere ait azınlık sınıf örneklerinin sayısının azlığı ($N_0 = 11$) da göz önüne alındığında küçük miktardaki yükseliş veya düşüşlerin dengesiz ve yukarı örneklenmiş veri setleri arasındaki farklı tam anlamıyla yorumlamaya olanak vermedikleri görülmektedir.

Benzer çalışmalarda log kayıtları kullanılarak oluşturulan veri setlerinin öğrencilerin ÖYS ile olan etkileşimlerini miktar veya yüzde cinsinden ifade eden görüntüleme sayısı, oturum açma sayısı, ödev tamamlama yüzdesi vb. öznitelikleri içerdiği görülmekle beraber [12, 14, 18, 19], bu çalışmada olduğu gibi hem süre hem de miktar ifade eden özniteliklerin kullanıldığı çalışmalar da bulunmaktadır [10, 13, 15, 17]. Bununla birlikte Hu, Lo ve Shih [43], ÖYS logları ile çevrimiçi öğrenme performansının tahmin edilmesinde zaman-bağımlı ve zaman-bağımsız özniteliklerle oluşturdukları veri setlerini kullanarak geliştirdikleri modellerin sınıflandırma performanslarını karşılaştırarak zaman-bağımlı özniteliklerden oluşan veri setinin kullanıldığı modelin daha yüksek başarı gösterdiğini belirtmişlerdir.

Çalışmanın yürütüldüğü dersin ÖYS'si içerisinde forum kullanılmaması bir kısıtlılık olarak değerlendirilebilir. Forumların ÖYS'lerde hem öğrenciler arasında hem de eğitmen ve öğrenci arasındaki etkileşimi arttırdığı ve öğrencilerin forumlar aracılığıyla soru sorma veya tartışmaya katılma sıklıklarının akademik başarılarıyla ilişkili olduğu bilinmektedir [44]. Pascual-Miguel ve diğerleri [45], Moodle tabanlı bir ÖYS içerisindeki etkileşimlerden forumda tartışma açma, cevap yazma ve güncelleme gibi eylemleri aktif etkileşim; forumla ilgili erişim ve görüntüleme eylemlerini ise pasif etkileşim başlıklarında kategorize etmişler fakat bu kategorilerin akademik başarının yordanması açısından farkları bulunamasa da etkileşimi zayıf fakat başarılı öğrencilerin saptanmasında kullanılabileceklerini öne sürmüşlerdir.

Sonuç olarak, gelecekte ÖYS log kayıtları kullanılarak öğrencilerin akademik başarılarının tahmin edilmesi amacıyla daha fazla örnek içeren dengeli veri setlerinin kullanıldığı, öğrencilerin ÖYS ile olan etkileşimlerini süre cinsinden ifade eden daha fazla zaman-bağımlı değişkenin veri setine dahil edildiği ve başta forumlar olmak üzere ÖYS içerisinde eğitmen ile ve öğrenciler arasındaki etkileşimi artıracak olanakların kullanıldığı çalışmaların yürütülmesi ve bu çalışmanın henüz tamamlanmamış bir dersin ÖYS log kayıtları ile test edilecek şekilde gerçekleştirilmesi önerilmektedir.

KAYNAKLAR (REFERENCES)

[1] U. Fayyad, G. Piatetsky-Shapiro & P. Smyth, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, 17(3), 37-54, 1996.

[2] C. Romero & S. Ventura, "Data Mining in Education", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27, 2013.

[3] C. Romero & S. Ventura, "Educational data mining: A survey from 1995 to 2005", *Expert Systems with Applications*, 33(1), 135-146, 2007.

[4] H. Aldowah, H. Al-Samarraie & W. M. Fauzy, "Educational Data Mining and Learning Analytics for 21st Century Higher Education: A Review and Synthesis", *Telematics and Informatics*, 37, 13-49, 2019.

[5] J. P. Vandamme, N. Meskens & J. F. Superby, "Predicting Academic Performance by Data Mining Methods", *Education Economics*, 15(4), 405, 2007.

[6] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang & H. Y. Seong, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques", **2009 Second International Conference on Computer and Electrical Engineering**, Dubai, 357-361, 2009.

[7] B. K. Bhardwaj & S. Pal "Data Mining: A Prediction for Performance Improvement Using Classification", *International Journal of Computer Science and Information Security*, 9(4), 136-140, 2011.

[8] F. Ahmad, N. H. Ismail & A. A. Aziz, "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques", *Applied Mathematical Sciences*, 9, 6415-6426, 2015.

[9] A. Mueen, B. Zafar & U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques", *International Journal of Modern Education and Computer Science*, 8(11), 36-42, 2016.

[10] C. Romero, S. Ventura & E. García, "Data Mining in Course Management Systems: Moodle Case Study and Tutorial", *Computers & Education*, 51(1), 368-384, 2008.

[11] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir & Y. El Allioui, "A Fuzzy Classification Approach for Learning Style Prediction Based on Web Mining Technique in E-Learning Environments", *Education and Information Technologies*, 24(3), 1943-1959, 2018.

[12] M. D. Calvo-Flores, E. G. Galindo, M. C. P. Jiménez & O. Pérez, "Predicting students' marks from Moodle logs using neural network models", *Current Developments in Technology-Assisted Education*, 1(2), 586-590, 2006.

[13] C. Romero, S. Ventura, P. G. Espejo & C. Hervás, "Data Mining Algorithms to Classify Students", **1st International Conference on Educational Data Mining**, Canada, 8-17, 2008.

[14] J. Bravo & A. Ortigosa, "Detecting Symptoms of Low Performance Using Production Rules", **2nd International Conference on Educational Data Mining**, Spain, 31-40, 2009.

[15] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero & S. Ventura, "Web Usage Mining for Predicting Final Marks of Students That Use Moodle Courses", *Computer Applications in Engineering Education*, 21(1), 135-146, 2010.

[16] Á. F. Agudo-Peregrina, S. Iglesias-Pradas, M. Á. Conde-González & Á. Hernández-García, "Can We Predict Success from Log Data in VLEs? Classification of Interactions for Learning Analytics and Their Relation with Performance in VLE-Supported F2F and Online Learning", *Computers in Human Behavior*, 31, 542-550, 2014.

[17] G. Akçapınar, **Çevrimiçi öğrenme ortamındaki etkileşim verilerine göre öğrencilerin akademik performanslarının veri madenciliği yaklaşımı ile modellenmesi**, Doktora Tezi, Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, 2014.

[18] N. Ademi, S. Loshkovska & S. Kalajdziski, "Prediction of Student Success Through Analysis of Moodle Logs: Case Study", **International Conference on ICT Innovations,** North Macedonia, 27-40, 2019.

[19] A. Y. Q. Huang, O. H. T. Lu, J. C. H. Huang, C. J. Yin, & S. J. H. Yang, "Predicting Students' Academic Performance by Using Educational Big Data and Learning Analytics: Evaluation of Classification Methods and Learning Logs", *Interactive Learning Environments*, 28(2), 206-230, 2020.

[20] G. Akçapınar, "Predicting students' approaches to learning based on Moodle logs", **In 8th International Conference on Education and New Learning Technologies**, Spain, 2347-2352, 2016.

[21] M. Abdullah, A. Alqahtani, J. Aljabri, R. Altowirgi & R. Fallatah, "Learning Style Classification Based on Student's Behavior in Moodle Learning Management System", *Transactions on Machine Learning and Artificial Intelligence*, 3(1), 13, 2015.

[22] M. Cocea & S. Weibelzahl, "Eliciting Motivation Knowledge from Log Files Towards Motivation Diagnosis for Adaptive Systems", **User Modeling 2007,** Cilt 4511, Editör: Conati C., McCoy K. & Paliouras G., Springer, Berlin, 197-206, 2007.

[23] A. Hershkovitz & R. Nachmias, "Learning about Online Learning Processes and Students' Motivation through Web Usage Mining", *Interdisciplinary Journal of E-Learning and Learning Objects*, 5(1), 197-214, 2009.

[24] Q. Zhou, W. Quan, Y. Zhong, W. Xiao, C. Mou & Y. Wang, "Predicting high-risk students using Internet access logs", *Knowledge and Information Systems*, 55(2), 393-413, 2017.

[25] R Core Team, **R: A language and environment for statistical computing**, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[26] S. V. Buuren, & K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R", *Journal of Statistical Software*, 45(3), 1-67, 2011.

[27] N.V. Chawla, "Data Mining for Imbalanced Datasets: An Overview" **Data Mining and Knowledge Discovery Handbook** Editör: Maimon O., Rokach L., Springer, Boston, 853-867, 2009.

[28] E. Kartal & Z. Özen, "Dengesiz Veri Setlerinde Sınıflandırma", **Mühendislikte Yapay Zeka ve Uygulamaları**, Editörler: Torkul O., Gülseçen S., Uyaroğlu Y., Çağıl G., Uçar M. K., Sakarya, Sakarya Üniversitesi Kütüphanesi Yayınevi, 109-131, 2017.

[29] R. Wirth & J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining" **4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining,** London, UK, 29-39, 2000.

[30] T. Cover & P. Hart, "Nearest Neighbor Pattern Classification", *IEEE Transactions on Information Theory*, 13(1), 21-27, 1967.

[31] M. Kantardzic, **Data Mining: Concepts, Models, Methods, and Algorithms,** John Wiley & Sons, New Jersey, ABD, 2011.

[32] D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval", **Machine Learning: ECML-98,** Cilt: 1398, Editörler: Nédellec C. & Rouveirol C., Springer, Berlin, 4-15., 1998.

[33] V. Vapnik, **The Nature of Statistical Learning Theory**. Springer, New York, 1995.

[34] E. E. Osuna, **Support Vector Machines: Training and Applications,** Doktora Tezi, Massachusetts Institute of Technology, 1998.

[35] Y. Özkan, **Veri Madenciliği Yöntemleri,** Papatya Yayıncılık, İstanbul, 2008.

[36] R. Pandya & J. Pandya, "C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning" *International Journal of Computer Applications*, 117(16), 18-21, 2015.

[37] S. Pang & J. Gong, "C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks", *Systems Engineering - Theory & Practice*, 29(12), 94-104, 2009.

[38] S. Yadav & S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification", **2016 IEEE 6th International Conference on Advanced Computing (IACC)**, India, 78-83, 2016.

[39] M. Sokolova, N. Japkowicz & S. Szpakowicz, "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation", **AI 2006: Advances in Artificial Intelligence,** 4304, Editörler: Sattar A. & Kang B., Springer, Berlin Heidelberg 1015-1021, 2006.

[40] G. Forman & M. Scholz, "Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement", *ACM SIGKDD Explorations Newsletter*, 12(1), 49–57, 2010.

[41] D. M. Hawkins, "The problem of overfitting", *Journal of chemical information and computer sciences*, 44(1), 1-12, 2004.

[42] V. Cerqueira, L. Torgo & I. Mozetič. "Evaluating time series forecasting models: an empirical studyon performance estimation methods", *Machine Learning*, 109:1997–2028, 2020.

[43] Y.H. Hu, C.-L. Lo & S. P. Shih, "Developing Early Warning Systems to Predict Students' Online Learning Performance", *Computers in Human Behavior*, 36, 469-478, 2014.

[44] C. Romero, P.González, S. Ventura, M. J. del Jesus & F. Herrera, "Evolutionary Algorithms for Subgroup Discovery in E-Learning: A Practical Application Using Moodle Data", *Expert Systems with Applications*, 36(2), 1632-1644, 2009.

[45] F. Pascual-Miguel, J. C. Pelaez, A. H. Garcia & S. I. Pradas, "A Characterisation of Passive and Active Interactions and Their Influence on Students' Achievement Using Moodle LMS Logs", *International Journal of Technology Enhanced Learning*, 3(4), 403, 2011.