Mühendislik Bilimleri ve Tasarım Dergisi
12(2), 443–454, 2024
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.1380197

Araştırma Makalesi

Research Article

AKADEMİK BAŞARININ VERİ MADENCİLİĞİ YÖNTEMLERİYLE TAHMİN EDİLMESİ

Mustafa YAĞCI*

Kırşehir Ahi Evran Üniversitesi, MMF, Bilgisayar Mühendisliği Bölümü, Kırşehir, Türkiye

Anahtar Kelimeler

Veri Madenciliği, Makine Öğrenmesi, Eğitsel Veri Madenciliği, Akademik Performansın Tahmin Edilmesi, Erken Uyarı Sistemleri.

Öz

Bu çalışmada, öğrencilerin bir önceki döneme ait dönem sonu not ortalamalarını veri madenciliği yöntemleri ile analiz ederek sonraki dönemlerde alabileceği dönem sonu not ortalamalarını giderek genişleyen 3 kategoride (Bölüm, Fakülte, Üniversite bazında) tahmin edecek yeni bir model önerilmiştir. Veri seti, Türkiye'de bir Devlet Üniversitesindeki tüm öğrenci kayıtlarının tutulduğu Öğrenci Bilgi Sisteminden (ÖBS) alınmıştır. Veriler, Sınıf öğretmenliği bölümünden 426, Eğitim fakültesinden 2.379 ve Üniversite genelinde eğitim gören 5.149 öğrencinin 2017-2018 Güz ve Bahar Yarıyılı dönem sonu not ortalamalarını içermektedir. Öğrencilerin dönem sonundaki genel not ortalamalarını tahmin etmek için veri madenciliği algoritmalarından rastgele orman, lineer regresyon, destek vektör makineleri ve k-en yakın komşular algoritmalarının başarımı hesaplanmış ve karşılaştırılmıştır. Uygulanan tüm algoritmalar örnekleri %92 ile %94 arasında değişen oranlarda doğru bir şekilde sınıflandırmıştır. Önerilen model, öğrencilerin dönem sonu not ortalamalarını tek bir değişken ile 4 üzerinden 0,28 puanlık ortalama sapma ile doğru tahmin etmiştir. Dönem sonu not ortalamalarının tahmin edilmesi sayesinde başarısız olma riski yüksek olan öğrenciler önceden belirlenebilir.

PREDICTING ACADEMIC ACHIEVEMENT USING DATA MINING METHODS

Keywords

Data Mining, Machine Learning, Predict Of Academic Performance, Educational Data Mining, Early Warning Systems.

Abstract

This study proposes a new model to analyze the grade point averages (GPAs) in the previous semester using data mining algorithms and to predict the final GPAs that students may receive in the following semesters in three gradually expanding categories (department, faculty, and university). The performances of the Random Forest, Linear Regression, Support Vector Machines, and k-Nearest Neighbors algorithms, which are among the data mining algorithms, were calculated and compared to estimate the GPAs of the students at the end of the semester. All algorithms applied correctly classified the samples at rates varying between 92% and 94%. The proposed model correctly estimated students' grade point averages at the end of the semester with an average deviation of 0.28 points over a 4 with a single variable. Students with a high risk of failure can be determined in advance by estimating their final grade point averages.

Alıntı / Cite

Yağcı, M., (2024). Akademik Başarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi, Mühendislik Bilimleri ve Tasarım Dergisi, 12(2), 443-454

Yazar Kimliği / Author ID (ORCID Number)

M. Yağcı, 0000-0003-2911-3909

Makale Süreci / Article Process

Başvuru Tarihi / Submission Date	23.10.2023
Revizyon Tarihi / Revision Date	12.06.2024
Kabul Tarihi / Accepted Date	12.06.2024
Yayım Tarihi / Published Date	30.06.2024

* İlgili yazar / Corresponding author: mustafayagci@ahievran.edu.tr, +90-386-280-6057

PREDICTING ACADEMIC ACHIEVEMENT USING DATA MINING METHODS

Mustafa YAĞCI†

Kırşehir Ahi Evran University, Faculty of Engineering and Architecture, Department of Computer Engineering, Kırşehir, Türkiye

Highlights

- Machine learning algorithms are quite successful in predicting students' academic achievement
- Exploring the hidden relationships in educational data and predicting students' academic achievements Sentence
- The processing of educational data causes improvements in many areas such as predicting student behaviour, analytical learning, and new approaches to education policies

Graphical Abstract

	Model	RMSE	MAE	R^2
Department	LR	0.453	0.290	0.455
	RF	0.434	0.280	0.499
	SVM	0.457	0.289	0.443
	kNN	0.480	0.312	0.387
Faculty	LR	0.393	0.296	0.543
	RF	0.398	0.300	0.530
	SVM	0.393	0.295	0.542
	kNN	0.396	0.299	0.534
University	LR	0.413	0.315	0.723
	RF	0.422	0.320	0.712
	SVM	0.414	0.315	0.723
	kNN	0.419	0.319	0.715

Figure. LR, RF, SVM, and kNN models' performance criteria

Purpose and Scope

The study concerns predicting students' academic achievement using GPS only, no demographic characteristics and no socio-economic data. In this context, this study aimed to develop a new model that can analyze the grade point averages (GPAs) of the previous semester with data mining methods and predict the final GPAs in the following semesters in three categories (department, faculty, and university).

Design/methodolgy

Data mining (DM) methods analyze the data measured and predict the results of samples in similar situations. Two types of these methods are regression and classification algorithms. While a regression algorithms continuously predict values, the classification algorithms predict categorical values.

Findings

The model proposed here accurately estimated the GPAs of the students at the end of the semester with an average deviation of seven points out of a hundred with a single variable. By estimating the final GPAs, students who are at risk of failure or who are at risk of drop out can be identified. So, education and training authorities can be given opportunities to implement corrective actions for these students. Modules that predict academic performance with DM methods can also be added to the LMS. It will thus be possible to make the most accurate predictions automatically and quickly. In short, teaching-learning processes can be managed more effectively and more efficiently thanks to the predictions for academic performance made by DM methods. Timely and targeted individual interventions can be ensured.

Originality

This study proposed a new model based on DM algorithms to identify students who have the potential to fail and who may be likely to drop out the university. This new model analyzes the students' GPAs from the previous semester with DM algorithms and predicts the GPAs they may receive in the following semesters in three categories (department, faculty, and university).

† Corresponding author: mustafayagci@ahievran.edu.tr, +90-386-280-6057

1. Giriş (Introduction)

Veri Madenciliği (VM), farklı sınıflandırma algoritmaları kullanarak büyük veriden yeni yönelimler ve yeni desenler çıkarma sürecidir (Baker ve Yacef, 2009). Başka bir deyişle büyük veri setlerinden faydalı bilgilerin keşfedilmesidir. Eğitsel veri madenciliği (EVM) ise "temel olarak öğrenciler ve eğitmenler tarafından oluşturulan eğitsel verileri incelemek için istatistiksel, makine öğrenimi ve veri madenciliği yöntemlerini geliştirir ve uyarlar" (Calvet Liñán ve Juan Pérez, 2015, s.100). EVM, eğitim ortamlarından elde edilen verilerdeki anlamlı bölümleri, özgün yapıları ve gizli kalıpları ortaya çıkarmak için yeni yöntemler geliştirir. EVM'nin temel amacı, eğitimle ilgili konularda karar almayı desteklemek için eğitim verilerinden bilgi çıkarmaktır (Calvet Liñán ve Juan Pérez, 2015). Akademik performansı değerlendirmek, gelecekteki performansı tahmin etmek ve mevcut sorunları belirlemek vb. için öğrenciler tarafından üretilen verilerin toplanmasını ve yorumlanmasını içerir. Kısaca EVM, büyük miktarda veriden önceden bilinmeyen, gizli, anlamlı ve faydalı kalıpların otomatik olarak çıkarılmasını içerir.

Akademik performans açısından başarısızlık riski taşıyan öğrencileri belirlemek için çeşitli VM algoritmaları başarıyla uygulanmıştır (Hu, Lo ve Shih, 2014). Öğrenme/öğretme süreçlerini iyileştirmeye hizmet eden bu algoritmaların kullanımıyla keşfedilen bilgilerden hem öğrenciler hem de öğretmenler faydalanmaktadır (Akçapınar, Altun ve Aşkar, 2019). Günümüz eğitim sistemlerinde, demografik verileri ve öğrencilerin akademik notlarını içeren büyük miktarda veri elektronik ortamlarda saklanmaktadır. Bu veriler çeşitli öğrenme yönetim sistemlerinden (ÖYS) ve öğrenci bilgi sistemlerinden (ÖBS) elde edilmektedir. Eğitim verilerindeki bu hızlı artış, öğrencilerin öğrenme çıktılarının iyileştirilmesine katkıda bulunabilir (Shorfuzzaman vd., 2019; Viberg vd., 2018).

EVM, ilgili paydaşlar arasındaki ilişkiyi tanımlamak ve eğitim ortamındaki gizli örüntüleri keşfederek öğrenme ortamını optimize etmek için istatistiksel bilgi sağlar (Fernandes vd., 2019). Daha sonra hem öğrencilerin hem de öğretmenlerin olumsuz sonuçlardan kaçınmasını sağlayacak ve böylece öğretme-öğrenme süreçlerinin iyileştirilmesine hizmet edecek pedagojik yaklaşımların geliştirilebileceği bir model oluşturur. Örneğin, EVM ile ilgili bazı çalışmalar e-öğrenme sistemlerini karşılaştırmış (Lara vd., 2014), bazıları eğitim verilerini sınıflandırmış (Chakraborty vd., 2016), diğerleri ise öğrenci performansını tahmin etmeye çalışmıştır (Fernandes vd., 2019). Böylece hem başarılı hem de risk altındaki öğrencilerin belirlenmesiyle düzeltici stratejiler ve pedagojik yöntemler geliştirilebilir (Casquero vd., 2016; Fidalgo-Blanco vd., 2015).

Ahmad ve Shahzadi (2018), akademik açıdan başarısız olma potansiyeli olan öğrencileri belirlemek için makine öğrenmesi yöntemleri ile bir model geliştirmiştir. Öğrencilerin öğrenme becerilerini, çalışma alışkanlıklarını ve akademik etkileşim özelliklerini bağımsız değişkenler olarak belirlemişlerdir. Modelin başarımı %85 olarak bulunmuştur. Cruz-Jesus ve ark. (2020), öğrencilerin demografik özelliklerini kullanarak akademik performanslarını tahmin etmeye çalışmıştır. K-en yakın komşular, lojistik regresyon, rastgele orman ve destek vektör makineleri algoritmaları, öğrencilerin %65'inin akademik performansını doğru bir şekilde tahmin etmiştir. Fernandes ve ark. (2019) öğrencilerin demografik özelliklerini ve dönem içi etkinliklere yönelik başarı puanlarını kullanarak bir model geliştirmiştir. Musso ve ark. (2020) ise, öğrencilerin sosyo-ekonomik özelliklerine ve akademik performanslarına dayalı bir makine öğrenmesi modeli geliştirmiştir.

Waheed ve ark., (2020), öğrencilerin ÖYS'deki etkileşimlerini farklı bir bakış açısıyla kullanarak yeni bir makine öğrenmesi modeli geliştirmiştir. Geliştirdikleri modelin %85 oranında doğru tahminler yaptığını belirten araştırmacılara göre, daha önce çevrimiçi dersleri gezinen öğrenciler daha başarılı olmuştur. Xu vd., (2019) üniversite öğrencilerinin internet kullanım özellikleri ile akademik performansları arasındaki ilişkiyi incelemiştir. Geliştirdikleri model, öğrencilerin performansını yüksek doğruluk oranıyla tahmin etmiştir. Burgos ve ark. (2018) da benzer şekilde öğrencilerin geçmiş yarıyıllardaki akademik performansları ile sonraki yarıyıllardaki akademik performansları arasındaki ilişkiyi incelemiştir.

Sonuç olarak EVM öğrencilerin okulu bırakması veya derse olan ilgisinin azalması gibi durumların erken tahmin edilmesini, performanslarını etkileyen içsel faktörlerin analiz edilmesini ve öğrencilerin performansını ölçmek için istatistiksel tekniklerin yapılmasını sağlar. Öğrencilerin akademik başarılarını tahmin etmek, yavaş öğrenenleri ve okulu bırakmaları belirlemek için çeşitli veri madenciliği teknikleri kullanılmaktadır (Hardman, Paucar-Caceres ve Fielding, 2013; Kaur, Singh ve Josan, 2015). Bu bağlamda erken tahmin, bu alanda uygun düzeltici strateji ve politikalar önererek öğrencilere destek olmak için değerlendirme yöntemlerini kapsayan nispeten yeni bir olgudur (Akçapınar vd., 2019; Waheed vd., 2020).

Akademik performansı ve kalıcılığı tahmin etmeyi amaçlayan araştırmacılar, sinir ağları, karar ağaçları, logit, probit ve regresyon dahil olmak üzere bir dizi teknik uygulamışlardır (Nandeshwar, Menzies ve Nelson, 2011). Bununla birlikte, en yeni çalışmalarda rastgele orman (Hung vd., 2020), genetik programlama (Pillay, 2020) ve Naïve Bayes (Sutoyo ve Almaarif, 2020) algoritmaları benimsenmiştir. Bu alandaki literatür incelendiğinde çalışmalarda çok çeşitli değişkenlerin kullanıldığı görülmüştür:

Bunlardan bazıları şunlardır; öğrencilerin internette bıraktığı gezinme, ders izlemede harcanan süre, devam yüzdesi gibi çeşitli dijital izler (Fernandes vd., 2019; Waheed vd., 2020; Xu vd., 2019), cinsiyet, yaş, ekonomik durum, katıldığı ders sayısı, internet erişimi gibi öğrencilerin demografik özellikleri (Aydemir, 2017; Bernacki vd., 2020; Cruz-Jesus vd., 2020; García-González & Skrita, 2019; Rebai, Yahia ve Essid, 2020; Rizvi, Rienties ve Ahmed, 2019), öğrenme becerileri, çalışma yaklaşımları, çalışma alışkanlıkları (Ahmad ve Shahzadi, 2018), öğrenme stratejileri, sosyal destek algısı, motivasyon, sağlık, akademik performans özellikleri (Costa-Mendes vd., 2020; Musso vd., 2020; Kılınç, 2015; Gök, 2017), ödevler, projeler, quizler (Kardaş & Güvenir, 2020).

Bu tür çalışmalarda geliştirilen modellerin hemen hemen hepsinde sınıflandırma doğruluk oranının %70 ile %95 arasında değiştiği görülmektedir. Ancak bu kadar çeşitli verilerin toplanması ve işlenmesi çok zaman almakta ve uzman bilgisi gerektirmektedir. Ayrıca Hoffait ve Schyns (2017) sosyo-ekonomik verilerin (ebeveynlerin eğitim düzeyi ve mesleği vb.) gereksiz olduğunu ve bu kadar fazla veri toplamanın zor olduğunu belirtmiştir. Ayrıca bu demografik veya sosyo-ekonomik veriler başarısızlığın nasıl önleneceği konusunda her zaman doğru fikirleri sağlamayabilir (Bernacki vd., 2020).

Türkiye'de yükseköğretim, Bologna Süreci terminolojisi açısından kısa, birinci, ikinci ve üçüncü aşamalardan oluşan tüm ortaöğretim sonrası yükseköğretim programlarını kapsamaktadır. Türk yükseköğretim programlarının yapısı, tek kademeli bir sisteme sahip olan diş hekimliği, eczacılık, tıp ve veterinerlik programları dışında, iki kademeli bir sisteme dayanmaktadır. Bu tek aşamalı programların süresi, altı yıl (360 AKTS) süren tıp hariç, beş yıldır (300 AKTS). Bu tek aşamalı programlardaki yeterlilikler, birinci aşama (lisans) artı ikinci aşama (yüksek lisans) derecesine eşdeğerdir. Lisans eğitim düzeyi, sırasıyla tam zamanlı iki yıllık (120 AKTS) ve dört yıllık (240 AKTS) eğitim programlarının başarıyla tamamlanmasının ardından verilen kısa dönem (önlisans) ve birinci aşama (lisans) derecelerinden oluşur.

Öğrencinin akademik başarısı, Öğrenci İşleri Daire Başkanlıkları tarafından yarıyıl sonu notu (ara sınav puanının %40'ı ile yarıyıl sonu sınav puanının %60'ının toplamı) ve genel not ortalaması (AGNO) dikkate alınarak hesaplanır ve bu not öğrencinin kişisel dosyasına işlenir. Önlisans ve Lisans veya Yüksek Lisans derecesi elde edebilmek için öğrencilerin genel not ortalamasının 4,00 üzerinden 2,00'den az olmaması ve programdaki tüm dersleri ve yaz uygulamalarını başarıyla tamamlamış olmaları gerekir.

Bu çalışma, öğrencilerin akademik başarısını yalnızca dönem sonu not ortalamasını kullanarak, demografik özellikler ve sosyo-ekonomik veriler olmadan tahmin etmeye yöneliktir. Bu bağlamda bu çalışmada, bir önceki yarıyılın not ortalamalarını veri madenciliği yöntemleriyle analiz ederek sonraki yarıyılların dönem sonu not ortalamalarını üç kategoride (bölüm, fakülte ve üniversite) tahmin edebilen yeni bir model geliştirilmesi amaçlanmıştır. Veri seti bu üç kategoriye ayrılmıştır. Böylece geliştirilen modelin performansı grup bazında değerlendirilebilecektir. Bu genel amaç doğrultusunda VM'den hangi algoritmaların en yüksek performansa sahip olduğu belirlenmiştir. Bu durum öğrencilerin akademik gelişimlerine katkı sağlayacak pedagojik müdahalelerin ve yeni politikaların geliştirilmesine katkı sağlayacaktır. Bu sayede her akademik dönem sonunda yapılacak değerlendirmelerle başarısız olma potansiyeli olan öğrenci sayısı azaltılabilir.

2. Yöntem (Method)

Bu bölümde veri setinin, ön işleme tekniklerinin ve kullanılan makine öğrenmesi yöntemlerinin ayrıntıları sunulmuştur.

2.1. Veri kümesi (Dataset)

Eğitim kurumları öğrencilere ilişkin verileri düzenli olarak elektronik ortamda saklar. Bu veriler öğrencilerin demografik özelliklerinden akademik performanslarına kadar çok fazla çeşit ve hacimde olabilmektedir. Bu çalışmada veriler Türkiye'de bir Devlet Üniversitesinde eğitim gören tüm öğrenci kayıtlarının tutulduğu Öğrenci Bilgi Sisteminden (ÖBS) alınmıştır. Bu kayıtlar arasından Bölüm kategorisinden sınıf öğretmenliği programına kayıtlı öğrenciler, Fakülte kategorisinden Eğitim Fakültesi

öğrencileri ve Üniversite kategorisinden 2017-2018 Bahar Yarıyılında Üniversiteye kayıtlı toplam 5.649 öğrencinin dönem sonu not ortalamaları veri seti olarak alınmıştır. Veri seti, geliştirilen modelin farklı gruplarda göstereceği performansın anlamlılığını ve tutarlılığını değerlendirebilmek için bu şekilde üç kategoriye ayrılmıştır. Bir başka ifade ile modelin performansını bölüm, fakülte ve üniversite geneli olmak üzere 3 kategoride belirlemek için veri seti gruplandırılmıştır. Öğrencilerin birimlere göre dağılımı Tablo 1.'de verilmiştir.

Tablo 1. Öğrencilerin Akademik Birimlere Göre Dağılımı (The Distribution of The Students by The Academic Unit)

Akademik birim		Öğrenci Sayısı
İktisadi ve İdari Bilimler Fakültesi		1.464
Eğitim Fakültesi	Fen Bilgisi Öğretmenliği	281
	Sosyal Bilgiler Öğretmenliği	289
	Türkçe Öğretmenliği	261
	Sınıf Öğretmenliği	426
	Bilgisayar ve Öğretim Teknolojileri Öğretmenliği	148
	İlköğretim Matematik Öğretmenliği	193
	Rehberlik ve Psikolojik Danışmanlık	573
	Okul Öncesi Öğretmenliği	208
Fen Edebiyat Fakültesi		1.487
Ziraat Fakültesi		319
Toplam		5.649

Bu çalışmada 2017-2018 Eğitim Öğretim yılı Güz yarıyılı dönem sonu not ortalaması bağımsız değişken olarak Bahar yarıyılı dönem sonu not ortalaması ise bağımlı değişken olarak belirlenmiştir. Dönem sonu not ortalaması 0 ile 100 aralığında değerler alabilmektedir. Öğrencilerin güz dönemi not ortalamasına bağlı olarak geliştirilen model, bahar dönemi not ortalamasını tahmin etmektedir. Bir başka ifade ile öğrencinin güz döneminde gösterdiği performansın bahar döneminde gösterebileceği performansı hangi düzeyde açıkladığı incelenmiştir. Yapılan dönem sonu not ortalaması tahmini ile başarısız olma potansiyeli olan öğrencilere düzeltici faaliyetler yapabilmek için yaklaşık 5 ay gibi bir süre bulunmaktadır.

2.2. Veri Hazırlama (Data Preparation)

Verinin makine öğrenmesi modeline uygun hale dönüştürülmesi aşamasıdır. Verilerin kullanıma hazır hale getirilmesi sürecidir. Ham verilerin işlenebilir verilere dönüştürülmesi gürültüden arındırılmasıdır. Bu amaçla toplam 6.729 kayıttan 1080 (%16) tanesi (örneğin güz döneminde derslere girip bahar döneminde katılmayan veya kaydını sildiren öğrencilerin kayıtları) silinmiştir. ÖBS'den alınan veri setinde öğrencilerin her bir dersten aldığı vize, mazeret, final ve bütünleme sınavlarına ait puanlar vardır. Bu puanların her biri bir satır olarak kayıtlıdır. Öncelikle her bir öğrenci için çok sayıda satırdan oluşan bu kayıtları dönem bazında gruplandırılmıştır ve ortalamaları alınmıştır. Daha sonra satırlardan oluşan vize, mazeret, final ve bütünleme sınav notları sütunlara dönüştürülerek öznitelik haline getirilmiştir.

2.3. Algoritmaların Uygulanması (Applying the Algorithms)

Veri tanımlama ve toplama aşamasından sonra modelin geliştirilmesi aşamasına geçilmiştir. Bunun için makine öğrenmesi algoritmaları uygulanmıştır. Başarı ve başarısızlığın nedenlerini incelemek için yapılan analizlerde lojistik regresyon ve zaman serisi gibi istatistiksel yöntemler kullanılabilir (Ortiz ve Dehon, 2008; Ortiz ve Dehon, 2013). Ancak bir ya da birden çok değişkenin sonuçlarına bağlı olarak başka bir değişkenin alabileceği değerlerin tahmin edilmesinde karar ağaçları (Nandeshwar vd., 2011; Delen, 2011), destek vektör makineleri (Huang ve Fang, 2013), rasgele ormanlar (Vandamme vd., 2007; Delen, 2010), ve yapay sinir ağları (Vandamme vd., 2007; Delen, 2010) daha verimlidir ve daha doğru sonuçlar verir. Makine öğrenmesi yöntemleri bilinen verilere dayalı olarak yeni verilerin sonuçlarını başarılı bir şekilde tahmin edebilen modeller oluşturmaktadır. Bu bağlamda öğrencilerin akademik performansını tahmin etmek için daha önce yapılan çalışmalara benzer şekilde lineer regresyon (LR), rastgele orman (RO), destek vektör makineleri DVM) ve k-en yakın komşular (kNN) uygulanmıştır (Akçapınar vd., 2019; Cruz-Jesus vd., 2020; Zabriskie vd., 2019). Böylece başarısız olma potansiyeli ve dersi/okulu bırakma olasılığı olan öğrenciler

belirlenebilecektir. Bölüm, fakülte ve üniversite kategorilerinde verilerin %70'i eğitim verisi, %30'u test verisi olarak dağıtılmıştır. Tablo 2 eğitim ve test verilerinin kategorilere göre dağılımını göstermektedir.

Tablo 2. Eğitim ve Test Verilerinin Kategorilere göre Dağılımı
(The Distribution of Training and Test Data According to The Categories)

Akademik birim	Eğitim verisi	Test verisi	Veri Seti
Sınıf Öğretmenliği Bölümü	299	127	426
Eğitim Fakültesi	1.666	713	2.379
Üniversite	3.955	1.694	5.649

2.3.1. Regresyon modelleri (Regression models)

Regresyon modelleri, makine öğrenmesinin önemli bir dalı olup bağımlı bir değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modellemeyi amaçlar. Regresyon analizinin temel amacı, bağımlı değişkenin (hedef değişken) değerlerini tahmin edebilmek için bağımsız değişkenlerin (özellikler) etkilerini anlamaktır. Regresyon modelleri, özellikle sürekli ve nicel verilerin tahmin edilmesinde yaygın olarak kullanılır. Sürekli veriler, belirli bir aralıkta herhangi bir değeri alabilen veri türleridir. Bu çalışmada kullanılan veriler sürekli ve nicel verileri içerdiği için regresyon modelleri kullanılmıştır. Bu çalışmada kullanılan regresyon modelleri şunlardır: Rastgele Orman (RO), Lineer Regresyon (LR), K-En Yakın Komşu (kNN) ve Destek Vektör Makinesi (DVM).

2.3.1.1. Rastgele orman (Random forest)

Karar Ağaçları hem regresyon hem de sınıflandırma problemlerini ele alabilen denetimli bir makine öğrenmesi tekniğidir (Breiman, 2001; aktaran Costa-Mendes vd., 2020). Karar ağacı, ağacın iç düğümlerinin bağımsız değişkenleri içerdiği ve yaprakların olası hedef sınıflara karşılık geldiği bir ağaç yapısı biçimine sahip bir sınıflandırma modeli oluşturur. Her dâhili düğümün, değişkenin alabileceği olası değerlere karşılık gelen birkaç dalı vardır. Bir karar ağacının oluşturulması, her yinelemede, verilerin homojenliğine dayalı olarak ağaca girmek için bir değişkenin seçildiği yinelemeli bir sürece dayanır.

Bir karar ağacı, bir dizi sorunun cevaplarına bağlı olarak bir değişkeni başarılı başarısız şeklinde sınıflandırmaya izin verir. En ilgili soruların sırasını belirlemek için özyinelemeli bir "böl ve fethet" işlemi yürütür. Bu ağaçlar karar süreci ve temel faktörler (soru sırası) hakkında kapsamlı bilgi sağlar. Bir ağaçtan elde edilen sonuçların okunması, diğer yöntemlerle elde edilen sonuçların okunmasından daha kolaydır.

Rastgele Orman (Breiman, 2001; aktaran Costa-Mendes vd., 2020), rastgele karar ağaçlarının birleşiminden oluşan bir makine öğrenmesi yöntemidir. Karar ağacı hedef değişkenler açısından homojen sınıflar oluşturmak için tahmin değişkenlerini sırayla bir dizi bölüm ve alt bölüme ayıran bir makine öğrenme algoritmasıdır. Rastgele bir karar ağacında her düğümde en iyi bölünme rastgele değişken seçimiyle gerçekleştirilir (Amit ve Geman, 1997; aktaran Costa-Mendes vd., 2020).

2.3.1.2. Destek vektör makineleri (Support vector machines)

DVM, sınıflandırma ve regresyon sorunlarını ele almak için popüler bir makine öğrenmesi yöntemidir (Cortes & Vapnik, 1995). DVM her bir gözlemin olası iki sınıftan birine ait olduğu bir sınıflandırma problemine odaklanır. Böylece her bir sınıfın örneklerini ikinci sınıfın örneklerinden ayıran en iyi alt küme belirlenir (Hastie vd., 2017; Akt. Cruz-Jesus vd., 2020).

2.3.1.3. Lineer regresyon (Linear regression)

Regresyon, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi en iyi açıklayan fonksiyonu elde etmek için uygulanan istatistiksel tekniklerdir (Kudyaba ve Hoptroff, 2001; aktaran Cihan vd., 2017). Sınıf olarak tanımlanan öznitelik bağımlı değişken, geri kalan öznitelikler ise bağımsız değişkenler olarak tanımlanır. Regresyon denetimli öğrenmedir. Regresyon analizi değişkenler arasında neden-sonuç ilişkisinin bulunmasına olanak sağlayan bir analiz yöntemidir.

2.3.1.4. K-en yakın komşular (kNN) (K-nearest neighbors)

kNN, hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılabilecek bir makine öğrenmesi tekniğidir (Cover ve Hart, 1967). Bu yöntem, belirli sınıflardan oluşan bir örneklem setindeki gözlem

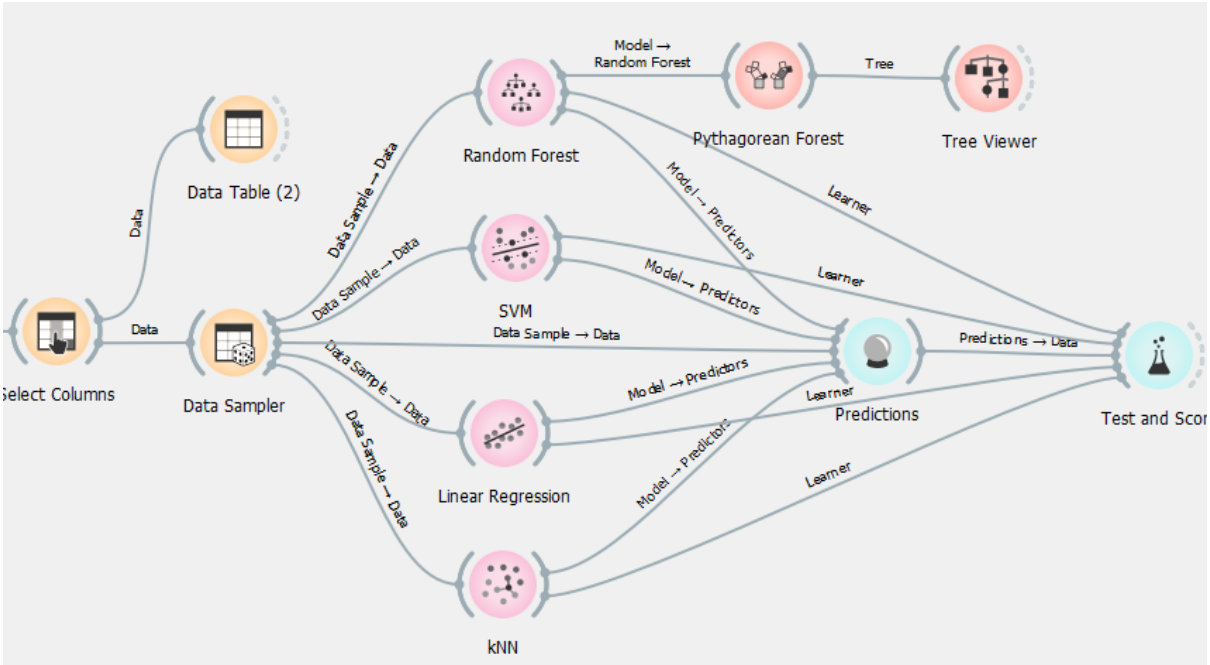
değerlerini kullanarak, örneğe dahil edilecek yeni bir gözlemin sınıfını belirlemek için kullanılır. Bir regresyon problemi durumunda algoritma, arama uzayındaki daha yakın (Öklid uzaklığına göre) veri noktalarının aynı hedef sınıfa ait olma olasılığının daha yüksek olması gerektiğini varsayar. Kullanılan formül Denklem 1'de verilmiştir.

$$d(x,y) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{ij})^2} \qquad (1)$$

kNN algoritması, eğitim setindeki her bir nokta çifti arasındaki mesafeyi hesaplar ve daha sonra, çoğunluk oyu kullanarak yeni bir veri noktası p'yi sınıflandırır. Eğitim setinden, p'ye daha yakın olan K noktalarını dikkate alarak p'yi K komşularının çoğunluğunun ait olduğu sınıfa atar.

3. Deneyler ve Sonuçlar (Experiments and Results)

Tüm deneysel aşama *Orange* makine öğrenmesi yazılımı ile gerçekleştirilmiştir (Ratra ve Gulia, 2020). Şekil 1'de bu çalışma için tasarlanan iş akışı şeması gösterilmektedir. Veriler, sınıf öğretmenliği bölümünde öğrenim gören 426, eğitim fakültesinde öğrenim gören 2.379 ve üniversitede öğrenim gören 5.649 öğrencinin 2017-2018 güz ve bahar dönemi genel not ortalamalarını içermektedir. Veri setindeki her bir gözlem yeterli sayıda eğitim verisi örneği ile temsil edilebildiği için ön işleme aşamasında herhangi bir veri seti dengesizliği oluşmamıştır. Modelin tasarımında güz dönemi genel not ortalamaları bağımsız değişken olarak kullanılmıştır. Açıklanması gereken değişken ise bahar dönemi genel not ortalamalarıydı. Tablo 3 model değişkenlerini göstermektedir.



Şekil 1. Geliştirilen Modelin İş Akış Şeması (Workflow of The Model Developed)

Tablo 3. Model Değişkenleri (Model Variables)

Öznitelikler	Hedef değişken	Meta
2017-2018 Güz dönemi	2017-2018 Bahar dönemi	stdID

Tablo 4'te 2017-2018 Bahar sütununda yer alan değerler gerçek değerlerdir. LR, RO, DVM ve kNN sütunlarındaki değerler ise ilgili modelin tahmin ettiği değerlerdir. Örneğin bölüm kategorisinde std1 numaralı öğrencinin bahar yarıyılı genel not ortalaması 3.05'dir. LR, RO, DVM ve kNN modellerinin tahmin edilen değerleri sırasıyla 3,16, 3,00, 3,17 ve 3,18'dir. İlk örnekte görüldüğü gibi modeller yaklaşık 0,10 puanlık sapma ile doğru tahmin yapmaktadır.

Tablo 4. Tahmin Modellerinin Olasılıkları ve Nihai Kararları (Probabilities and Final Decisions of Prediction Models)

	stdId	LR	RO	DVM	kNN	2017-2018 Bahar	2017-2018 Güz
Bölüm	std1	3,16	3,00	3,17	3,18	3,05	3,00
	std2	2,77	2,91	2,80	2,87	3,18	2,37
	std3	3,46	3,43	3,46	3,33	3,45	3,50
	std4	3,13	3,12	3,14	3,12	3,21	2,95
	std5	3,02	3,19	3,04	3,04	3,74	2,78
Fakülte	std1	2,98	3,01	3,00	3,07	3,27	2,86
	std2	2,44	2,40	2,48	2,41	1,85	2,22
	std3	2,74	2,81	2,77	2,85	2,88	2,58
	std4	3,87	3,65	3,86	3,68	4,00	3,92
	std5	3,40	3,35	3,40	3,40	3,44	3,36
Üniversite	std1	2,62	2,54	2,64	2,70	2,84	2,54
	std2	1,61	1,69	1,63	1,58	0,26	1,46
	std3	3,43	3,43	3,45	3,48	3,33	3,40
	std4	1,51	1,34	1,53	1,48	1,38	1,35
	std5	3,88	3,43	3,90	3,71	3,09	3,88

VM yöntemleri ölçülmüş verileri analiz eder ve benzer durumlardaki örneklerin sonuçlarını tahmin eder. Bu yöntemlerin iki türü regresyon ve sınıflandırma algoritmalarıdır. Regresyon algoritmaları sürekli olan değerleri tahmin ederken, sınıflandırma algoritmaları kategorik değerleri tahmin eder. Sonuç olarak temel fark, çıktı değişkeninin regresyon için sayısal (veya sürekli), sınıflandırma için ise kategorik (veya ayrık) olmasıdır. Yani bağımsız değişken sürekli bir değişkendir. Bu nedenle tahmin sonuçlarının doğruluğu regresyon metrikleri ile ölçülmüştür. Genel not ortalamaları için öngörülen değerler dört farklı metrik (Belirlilik Katsayısı-CoD, Ortalama Mutlak Hata-MAE, Ortalama Kare Hata-MSE ve Kök Ortalama Kare Hatası-RMSE) kullanılarak değerlendirilmiştir (Botchkarev, 2018; Botchkarev, 2019; Willmott & Matsuura, 2005). Bir VM modelinin doğruluk katsayısı (R^2) ne kadar yüksek olursa, tahmin edilen değerler gerçek değerlere o kadar yakın olur. MSE, RMSE ve MAE değerleri ise modelin hata ölçüsüdür. Düşük değerler modelin yüksek performans gösterdiği anlamına gelir. Bu çalışmada modellerin performansları RMSE, MAE ve R^2 metrikleri ile hesaplanmıştır. Korelasyon katsayısının (R^2) 1,00 olması bağımlı değişken ile bağımsız değişkenler arasında mükemmel bir pozitif ilişkiyi gösterir; -1,00 tamamen negatif bir ilişkidir; 0,00 ise herhangi bir ilişkinin olmadığını gösterir. Korelasyon katsayısının mutlak değeri 0,70-1,00 arasında ise yüksek düzeyde bir ilişki, 0,70-0,30 arasında ise orta düzeyde bir ilişki ve 0,30-0.00 arasında ise düşük düzeyde bir ilişki vardır (Büyüköztürk, 2008, sayfa 32). Tablo 5'te öğrencilerin dönem sonu genel not ortalamalarının tahminine ilişkin analiz sonuçları gösterilmektedir.

Tablo 5. LR, RO, DVM ve kNN Modellerinin Performans Ölçütleri (LR, RF, SVM, and kNN Models' Performance Criteria)

	Model	RMSE	MAE	R^2
Bölüm	LR	0,453	0,290	0,455
	RO	0,434	0,280	0,499
	DVM	0,457	0,289	0,443
	kNN	0,480	0,312	0,387
Fakülte	LR	0,393	0,296	0,543
	RO	0,398	0,300	0,530
	DVM	0,393	0,295	0,542
	kNN	0,396	0,299	0,534
Üniversite	LR	0,413	0,315	0,723
	RO	0,422	0,320	0,712
	DVM	0,414	0,315	0,723
	kNN	0,419	0,319	0,715

Bölüm kategorisinde en yüksek R^2 (0,499) değerini RO algoritması vermiştir. Bu bulguya göre bölüm kategorisinde tahmin edilen veriler ile gerçek veriler arasında orta düzeyde bir korelasyon vardır. Ayrıca MAE değerine (0,280) göre gerçek veriler 0,280 puan yukarı veya aşağı sapma ile doğru tahmin edilmiştir. Tahmin sonuçlarının gerçek sonuçlara ne kadar yakın olduğunu ölçmeye olanak tanıyan bu metrik hata değeri ne kadar düşük ise o kadar iyi sonuçlar üretilmektedir MAE için kullanılan formül Denklem 2'de verilmiştir. Sonuç olarak RO algoritması örnekleri %93,00 oranında doğru sınıflandırmıştır.

$$MAE = \frac{\sum_{i=1}^{n} |(Y_i - \lambda(X_i))|}{n} \qquad (2)$$

Fakülte kategorisinde en yüksek R^2 (0,543) değerini LR algoritması vermiştir. Bu bulguya göre fakülte kategorisinde tahmin edilen veriler ile gerçek veriler arasında orta düzeyde bir korelasyon vardır. Bununla birlikte MAE değerine (0,296) göre, gerçek veriler 0,296 puan yukarı veya aşağı sapma ile doğru tahmin edilmiştir. Sonuç olarak LR algoritması örnekleri %92,60 oranında doğru sınıflandırmıştır.

Üniversite kategorisinde ise en yüksek R^2 (0,723) değerini LR ve DVM algoritmaları vermiştir. Bu bulguya göre üniversite kategorisinde tahmin edilen veriler ile gerçek veriler arasında yüksek düzeyde bir korelasyon vardır. Ayrıca MAE değerine (0,315) göre, gerçek değeri 0,315 puan yukarı veya aşağı sapma ile doğru tahmin edilmiştir. Sonuç olarak LR ve DVM algoritmaları örnekleri %92,13 oranında doğru sınıflandırmıştır.

4. Tartışma ve Sonuç (Discussion and Conclusion)

Bu çalışmada başarısız olma potansiyeli ve ilerleyen dönemlerde dersi ya da okulu bırakma ihtimali yüksek olan öğrencileri belirlemek için makine öğrenmesi yöntemlerine dayalı yeni bir model önerilmiştir. Öğrencilerin bir önceki döneme ait dönem sonu not ortalamalarını veri madenciliği yöntemleri ile analiz ederek sonraki dönemlerde alabileceği dönem sonu not ortalamalarını 3 kategoride (Bölüm, Fakülte, Üniversite bazında) tahmin edecek yeni bir model önerilmiştir. Veri seti bölüm, fakülte ve üniversite geneli olacak şekilde kategorilere ayrılmıştır. Çünkü geliştirilen modelin başarımı kategori bazında da değerlendirilecektir. Ayrıca dört makine öğrenmesi yönteminin (LR, RO DVM ve kNN) performans göstergeleri karşılaştırılmıştır. Kısaca bu çalışmada üç parametreye odaklanılmıştır. Birincisi tek bir bağımsız değişken ile akademik başarı tahmini yapmak. İkincisi dört makine öğrenmesi yönteminin (LR, kNN, SVM ve RF) performans göstergelerini karşılaştırmak. Üçüncüsü ise 3 farklı kategoride (Bölüm, Fakülte, Üniversite) tahmin sonuçlarını karşılaştırmaktır.

Üniversite kategorisindeki LR, RO, DVM ve kNN algoritmalarında öğrencilerin bir önceki döneme ait genel not ortalamaları ile bir sonraki yarıyıl genel not ortalamaları arasında yüksek düzeyde bir korelasyon bulunmuştur. Algoritmaların yüksek performans göstergelerinin yanı sıra tahminlerin tek değişken kullanılarak yapılmış olması çalışmanın özgünlüğünü göstermektedir. Bulgular, öğrencilerin bir önceki yarıyıldaki genel not ortalamalarının bir sonraki yarıyılda alacakları genel not ortalamalarını yüksek düzeyde açıkladığını ifade etmeye olanak sağlamaktadır.

Bölüm ve fakülte kategorilerinde LR, RO, DVM ve kNN algoritmaları öğrencilerin bir önceki dönem genel not ortalamaları ile bir sonraki dönem genel not ortalamaları arasında orta düzeyde bir korelasyon olduğunu ortaya koymuştur.

Literatürde önceki dönem sonu not ortalamasına (tek bir değişken ile) dayalı olarak dönem sonu not ortalamalarının tahmin edildiği herhangi bir çalışmaya rastlanmamıştır. Bu yüzden araştırma sonuçları öğrencilerin akademik başarı puanlarını çeşitli demografik ve sosyo-ekonomik değişkenlere dayalı olarak tahmin etmeye çalışan araştırmalar ile karşılaştırılmıştır. Hoffait ve Schyns (2017) çeşitli demografik özelliklerine dayalı olarak başarısızlık riski yüksek olan öğrencileri belirlemek için veri madenciliği teknikleri ile yeni bir model geliştirmiştir. LR, yapay sinir ağları (YSA) ve RO yöntemlerinin performans göstergelerini karşılaştırmıştır. Başarısızlık riski yüksek olan öğrencileri %90 doğruluk ile tahmin edebilmiştir. Waheed ve ark. (2020) akademik açıdan düşük performans riski taşıyan ve dersi bırakma potansiyeli olan öğrencileri derin öğrenme modelleri ile belirlemiştir. Öğrencilerin demografik özellikleri ile birlikte LMS'deki toplam 54 öğrenci davranış özelliği ile bir model geliştirmiştir. Model ortalama %88 oranında doğru sınıflandırma yapmıştır. Elde ettiği sonuçların karar verme süreçlerine katkıda bulunacağını iddia etmektedir. Benzer şekilde Xu ve ark. (2019) de öğrencilerin internet kullanımı davranış özellikleri ile akademik performansları arasındaki ilişkiyi makine öğrenmesi yöntemleri ile incelemiştir. Benzer şekilde Bernacki ve ark. (2020) öğrencilerin LMS'de bıraktıkları dijital izlere dayalı olarak akademik başarı puanlarını tahmin etmeye çalışmıştır. Dersi tekrar etmesi gereken öğrencileri %75 başarı oranı ile doğru tahmin etmiştir. Ahmad ve Shahzadi (2018) ise öğrencilerin ders çalışma alışkanlıkları, öğrenme becerileri ve akademik etkileşim özellikleri ile akademik performansları arasındaki ilişkiyi makine öğrenmesi yöntemleri ile belirlemiştir. Önerdikleri modelin %85 oranında doğru tahmin yaptığı sonucunu bulmuştur. Sonuç olarak yüksek düzeyde bir ilişki bulmuştur ve makine öğrenmesi tekniklerinin eğitim-öğretim yönetiminin geliştirilmesine katkıda bulunacağını savunmaktadır. Makine öğrenmesi yöntemleri öğrencilerin demografik ve sosyo-ekonomik özellikleri ile akademik performansları arasındaki ilişkinin belirlenmesinde çok başarılı sonuçlar vermektedir (Cruz-Jesus vd., 2020; Costa-Mendes vd., 2020). Ancak dikkat edilirse bu çalışmaların hepsinde tahmin modelinin çok sayıda bağımsız değişken ile kurulduğu görülebilir.

Kısaca önerilen model öğrencilerin dönem sonu not ortalamalarını tek bir değişken ile yüz üzerinden ortalama 7 puan sapma ile doğru tahmin etmektedir. Dönem sonu not ortalamalarının tahmin edilmesi sayesinde başarısız olma riski olan ya da okulu bırakma riski olan öğrenciler önceden belirlenebilir. Eğitim-Öğretim otoritelerine bu öğrencilere düzeltici faaliyetler uygulayabilmeleri için fırsatlar verilebilir. LMS'lere makine öğrenmesi yöntemleri ile akademik performans tahmini yapan modüller eklenebilir. Böylece otomatik ve hızlı bir şekilde en doğru tahminler yapılabilir. Kısaca makine öğrenmesi yöntemleri ile yapılacak akademik başarı tahmini sayesinde öğrenme-öğretme süreçleri daha etkili ve daha verimli bir şekilde yönetilebilir. Zamanında ve hedef odaklı bireysel müdahaleler yapılması sağlanabilir.

Sonuç olarak; bu araştırma, öğrencinin akademik başarı düzeyini tespit etmede farklı değişkenler, farklı algoritmalar ve farklı bir yaklaşım kullansa da elde edilen sonuçlar önceki araştırmalarla uyumludur ve makine öğrenmesi yöntemlerinin öğrenci akademik motivasyonunu tahmin etmede etkili bir model oluşturabileceğini doğrulamaktadır. LR, RO, DVM ve kNN modellerinin başarım oranları çok yüksek düzeyde bulunmuştur. Ayrıca bu modellerin Bölüm, Fakülte ve Üniversite kategorileri için de uygulanabileceği görülmüştür. Bu tür veriye dayalı çalışmaların karar alma süreçlerine çok önemli katkılar sağlayabileceği söylenebilir. Ancak öğrencilerin desteklenmesi, karar alma süreçlerinin yönetilmesi ve öğrencilerin katılımının sağlanması için düzeltici stratejiler geliştirilmesi de gerekmektedir.

Bu çalışmada, Türkiye'deki bir devlet üniversitesindeki öğrencilerin verileri kullanılarak öğrencilerin önceki dönem genel not ortalamalarının veri madenciliği yöntemleri kullanılarak analiz edilmesi ve öğrencilerin sonraki dönemlerde alabilecekleri nihai genel not ortalamalarının tahmin edilmesi amaçlanmıştır. Bu nedenle gelecekteki araştırmalarda farklı eğitim düzeyindeki öğrenciler üzerinde çalışılabilir. Ayrıca öğrencilerin akademik performanslarını etkileyen çeşitli bireysel farklılıklar dikkate alınarak gelecek çalışmalar planlanabilir. Son olarak farklı ülkelerde de benzer çalışmalar yapılabilir. Böylece farklı kültürlerdeki durum karşılaştırılabilir.

Çıkar Çatışması (Conflict of Interest)

Yazar tarafından herhangi bir çıkar çatışması beyan edilmemiştir. No conflict of interest was declared by the author.

Kaynaklar (References)

Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. Bulletin of Education and Research, 40(3), 157–164.

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. https://doi.org/10.1186/s41239-019-0172-z

Aydemir, B. (2017). *Veri madenciliği yöntemleri kullanarak meslek yüksekokulu öğrencilerinin akademik başarı tahmini [Predicting academic success of vocational high school students using data mining methods]* [Master's Thesis]. Pamukkale University, Denizli, Turkey. http://hdl.handle.net/11499/2464

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

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education*, *158*. https://doi.org/10.1016/j.compedu.2020.103999

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. Retrieved from *http://www.gsrc.ca/metrics_typology2018.pdf* at 15 February 2021.

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge & Management*, *14*.

Burgos, C., Campanario, M. L., De, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance : A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, *66*(2018), 541–556. https://doi.org/10.1016/j.compeleceng.2017.03.005

Büyüköztürk, Ş. (2008). *Sosyal bilimler için veri analizi el kitabı. Ankara: PegemA Yayıncılık* (9th ed., p. 201). Ankara: PegemA.

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

Casquero, O., Ovelar, R., Romo, J., Benito, M., & Alberdi, M. (2016). Students' personal networks in virtual and personal learning environments: A case study in higher education using learning analytics approach. *Interactive Learning Environments*, *24*(1), 49–67. https://doi.org/10.1080/10494820.2013.817441

Chakraborty, B., Chakma, K., & Mukherjee, A. (2016). A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory. *Proceedings of 2nd IEEE International Conference on Engineering and Technology, ICETECH 2016*, *March*, 431–436. https://doi.org/10.1109/ICETECH.2016.7569290

Cihan, P., Gökçe, E., & Kalipsiz, O. (2017). Veteriner hekimlik alanında makine öğrenmesi uygulamaları üzerine bir derleme. *Kafkas Universitesi Veteriner Fakultesi Dergisi*, *23*(4), 673–680. https://doi.org/10.9775/kvfd.2016.17281

Cortes, C., & Vapnik, V. (1995). Supoort-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1109/64.163674

Costa-Mendes, R., Oliveira, T., Castelli, M., & Cruz-Jesus, F. (2020). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Education and Information Technologies*. https://doi.org/10.1007/s10639-020-10316-y

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1007/978-0-387-35973-1_862

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon*, *6*(6). https://doi.org/10.1016/j.heliyon.2020.e04081

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

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

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

Fidalgo-Blanco, Á., Sein-Echaluce, M. L., García-Peñalvo, F. J., & Conde, M. Á. (2015). Using learning analytics to improve teamwork assessment. *Computers in Human Behavior*, *47*, 149–156. https://doi.org/10.1016/j.chb.2014.11.050

García-González, J. D., & Skrita, A. (2019). Predicting academic performance based on students' family environment: Evidence for Colombia using classification trees. *Psychology, Society and Education*, *11*(3), 299–311. https://doi.org/10.25115/psye.v11i3.2056

Gök, M. (2017). Makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, *5*(3), 139–148.

Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm. *Systems Research and Behavioral Science*, *30*(2), 194–203. https://doi.org/10.1002/sres.2130

Hoffait, A., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*(2017), 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior, 36*, 469–478. https://doi.org/10.1016/j.chb.2014.04.002

Hung, H.-C., Liu, I.-F., Liang, C.-T., & Su, Y.-S. (2020). Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. *Symmetry, 12*(2). https://doi.org/10.3390/sym12020213

Kardaş, K., & Güvenir, A. (2020). Kısa sınavların , ödevlerin ve projelerin dönem sonu sınavına olan etkilerinin farklı makine öğrenmesi teknikleri ile araştırılması. *EMO Bilgisayar Dergisi*, *10*(1), 22–29.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, *57*, 500–508. https://doi.org/10.1016/j.procs.2015.07.372

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

Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European Higher Education Area - Application to student data from Open University of Madrid, UDIMA. *Computers and Education*, *72*, 23–36. https://doi.org/10.1016/j.compedu.2013.10.009

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. *Higher Education*, *80*(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7

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

Ortiz, E. A., & Dehon, C. (2008). What are the factors of success at university? A case study in Belgium. *CESifo Economic Studies*, *54*(2), 121–148. https://doi.org/10.1093/cesifo/ifn012

Ortiz, E. A., & Dehon, C. (2013). Roads to success in the Belgian French Community's Higher Education System: Predictors of dropout and degree completion at the Université Libre de Bruxelles. *Research in Higher Education*, *54*(6), 693–723. https://doi.org/10.1007/s11162-013-9290-y

Pillay, N. (2020). The impact of genetic programming in education. *Genetic Programming and Evolvable Machines, 21*, 87-97. https://doi.org/10.1007/s10710-019-09362-4

Ratra, R., & Gulia, P. (2020). Experimental evaluation of open source data mining tools (WEKA and Orange). *International Journal of Engineering Trends and Technology, 68*(8), 30-35. https://doi.org/10.14445/22315381/IJETT-V68I8P206S

Rebai, S., Yahia, F. B., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, *70*. https://doi.org/10.1016/j.seps.2019.06.009

Rizvi, S., Rienties, B., & Ahmed, S. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, *137*, 32–47. https://doi.org/10.1016/j.compedu.2019.04.001

Shorfuzzaman, M., Hossain, M. S., Nazir, A., Muhammad, G., & Alamri, A. (2019). Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. *Computers in Human Behavior*, *92*, 578–588. https://doi.org/10.1016/j.chb.2018.07.002

Sutoyo, E., & Almaarif, A. (2020). Educational data mining for predicting student graduation using the naïve bayes classifier algorithm. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4*(1), 95-101. https://doi.org/10.29207/resti.v4i1.1502

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

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110. https://doi.org/10.1016/j.chb.2018.07.027

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*. https://doi.org/10.1016/j.chb.2019.106189

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82.

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166–173. https://doi.org/10.1016/j.chb.2019.04.015

Zabriskie, C., Yang, J., DeVore, S., & Stewart, J. (2019). Using machine learning to predict physics course outcomes. *Physical Review Physics Education Research, 15*(2). https://doi.org/10.1103/PhysRevPhysEducRes.15.020120