PAUEFD 61, 94-120 https://doi.org/10.9779/pauefd.1085483

Pamukkale Üniversitesi Eğitim Fakültesi Dergisi
ISSN: 2757-8127

Karar Ağacı ve Kural Tümevarımı ile Eğitsel Veri Madenciliği: SAÜ İLİTAM Örneği

Deniz DEMİRCİOĞLU DİREN[1], Mehmet Barış HORZUM[2]

Özet

Bu çalışma, karma bir lisans tamamlama programına (İLİTAM) kayıt yaptıran öğrencilerin profiline göre, öğrencinin başarılı olma ya da terk etme/başarısız olma durumlarını incelemeyi amaçlamaktadır. Ayrıca öğrenci verilerine ait değişkenlerin öznitelik ağırlıklarına göre öğrencinin başarılı olma ya da terk etme/başarısız olma durumları üzerindeki önem dereceleri de ele alınmıştır. Araştırma yöntemi olarak eğitsel veri madenciliği kapsamında kullanılan CRISP-DM süreç modelinden faydalanılmıştır. Öznitelik ağırlıkları ise bilgi kazanımı yöntemi ile tespit edilmiştir. Araştırmanın çalışma grubu Sakarya Üniversitesi (SAÜ) lisans tamamlama programına 2013-2016 yılları arasında programa giriş yapan öğrencilerden oluşmaktadır. Sistemsel kayıtlardan elde edilen veri seti öğrencinin üniversiteye giriş bilgilerini içermektedir ve buna karşılık hedef değer ise öğrencinin üniversiteden mezuniyet başarı durumları yani başarılı olma ya da terk etme/başarısız olma durumları ile oluşturmuştur. Sonuçlar hedef değere en çok etki eden parametrenin öğrencinin cinsiyeti olduğunu göstermektedir. Ayrıca en yakın komşu algoritması kullanılarak 91.30% tahmin doğruluğu oranıyla bir öğrencinin kayıt yaptırdığında sahip olduğu genel bilgilerine göre mezuniyet başarı durumlarının tahmini gerçekleştirilmiştir. Bu sayede öğrenciye yönelik planlama yapmak ve önerilerde bulunmak mümkün olacaktır. Araştırmada bulgulara yönelik sonuç ve öneriler geliştirilmiştir.

Makale Bilgileri
Araştırma Makalesi

Gönderim Tarihi
10/03/2022
Kabul Tarihi
24/01/2024
Yayın Tarihi
15/05/2024

Anahtar Kelimeler
Eğitsel Veri Madenciliği, Veri Ön İşleme, Karar Ağacı, Kural Tümevarım, ILITAM

1 Sakarya Üniversitesi, ORCID:0000-0003-3567-0779, ddemircioglu@sakarya.edu.tr
2 Sakarya Üniversitesi, ORCID:0000-0002-4280-0394, mhorzum@sakarya.edu.tr

Atıf:
Demircioğlu Diren, D. ve Horzum, M. B. (2024). Karar ağacı ve kural tümevarımı ile eğitsel veri madenciliği: SAÜ İLİTAM örneği. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi [PAUEFD], 61*, 94-120. https://doi.org/10.9779/pauefd. 1085483.

Giriş

Veri madenciliği ve makine öğrenme yöntemlerindeki gelişmeler ve bu yöntemlerin eğitim alanındaki verilerin analizi için kullanılması son zamanlarda oldukça yaygınlaşmıştır. Bu alanda; öznitelik seçimi, görselleştirme, sınıflandırma, tahmin, kümeleme, birliktelik kuralları, örüntü işleme ve metin madenciliği gibi çalışma konuları mevcuttur. Bu sayede verilerden bilgi ve sonuç çıkartarak eğitimcilere öğrenme süreçlerini geliştirmeleri için karar almalarına yardımcı olmak ve öğrencilerin de bu bilgileri kullanarak kendilerini geliştirmelerine yol göstermek hedeflenmektedir. Veri madenciliği, verileri anlama, analiz etme ve yararlı bilgi şekline dönüştürmeyi amaçlayan çok disiplinli bir yaklaşımdır. Diğer bir ifade ile büyük veri tabanlarından bilgi ve sonuç çıkarma sürecidir. Aranan örüntülere göre veri madenciliği ile özetleme, sınıflandırma, kümeleme, ilişkilendirme ve trend analizi gibi işlemler gerçekleştirilebilir. Sınıflandırma, bir nesnenin özelliklerine göre sınıfını belirlemeyi sağlamaktadır. Özetleme, verileri soyutlanmakta ve genelleştirilmektedir. Birliktelik kuralları, değişkenler arasındaki ilişkileri ortaya çıkartmaktadır. Kümelemede ise, sınıfı bilinmeyen bir dizi nesnenin tanımlanması amaçlanmaktadır (Fu, 2011).

Eğitsel Veri Madenciliği uluslararası eğitsel veri madenciliği topluluğu tarafından şu şekilde tanımlanır: "eğitim ortamlarından gelen benzersiz veri türlerini keşfetmek için yöntemler geliştirmek ve bu yöntemleri öğrencileri ve içinde öğrendikleri ortamları daha iyi anlamak için kullanmakla ilgilenen, gelişmekte olan bir disiplindir" (Educational Data Mining, 2021). Bu yaklaşım çerçevesinde, web tabanlı kurslar, öğrenme içerik yönetim sistemleri ve uyarlanabilir akıllı web tabanlı eğitim sistemlerinin her biri farklı veri kaynaklarına ve bilgi keşfi için hedeflere sahiptir (Romero ve Ventura, 2007). Son dönemlerde eğitsel veri madenciliği alanında çok farklı hedef ve amaçlara yönelik çok sayıda çalışmalar gerçekleştirilmektedir. Bu alandaki çalışma konularını sınıflandıran araştırmacılar bulunmaktadır (Bakhshinategh ve diğerleri, 2018; Dutt ve diğerleri, 2017; Rodrigues ve diğerleri, 2018). Bu çalışmalar incelendiğinde, en temel çalışma alanlarının veri analizi ve görselleştirme (Gonçalves ve diğerleri, 2017; Pascual-Cid ve diğerleri, 2010), öğrenci performansı tahmini (Aghalarova ve Keser, 2021; Miller ve diğerleri, 2015; Moradi ve diğerleri, 2014), eğitmenler için geri bildirim sağlama (Wang ve Lin, 2012) ve öğrenci modelleme (Kassim ve diğerleri, 2004; Kay, 2000; Moradi ve diğerleri, 2014) olarak belirlendiği görülmektedir.

Türkiye'deki eğitsel veri madenciliği ile ilgili çalışmalar incelendiğinde ise; verilerin analizi (Özçınar, 2006) öğrenci modelleme (Akçapınar, 2014; Ersöz 2017; Kismet, 2018) ve öğrencilerin performansını belirleme (Aydemir 2017; Polat, 2021; Özdemir, 2016) gibi konularda lisansüstü tezler ile karşılaşılmaktadır. Ayrıca performans tahmini (Baltaci, 2018;

Başer ve diğerleri, 2020; Bilen ve diğerleri, 2014; Şengür ve Tekin, 2014), verilerin analizi (Aydemir, 2019; Keskin ve diğerleri, 2019), destekleyici eğitimler için geri bildirim sağlama (Güre ve diğerleri, 2020) ile bu konularda genel anlatım ve literatür taraması (Aruğaslan ve Çivril, 2021; Özbay, 2015; Öztürk, 2018; Tekin ve Öztekin, 2018; Tosunoğlu ve diğerleri, 2021) gibi akademik çalışmalar mevcuttur.

Bu çalışmada ise verilerin analizi gerçekleştirilip, öğrenci performanslarının tahmini incelenerek destekleyici eğitmenler için geri bildirim sağlamak hedeflenmektedir. Bu alanda öğrenci performansı ile ilgili literatürde yapılmış sistematik alan yazın incelemeleri yol gösterici olmuştur (Abu Saa ve diğerleri, 2019; Hermaliani ve diğerleri, 2022; Namoun ve diğerleri, 2020). Öğrenci performansı tahmin etme konusu üzerine yüz yüze eğitim (Bliuc ve diğerleri, 2010; Morsy ve Karypis, 2017; Polyzou ve Karypis, 2016), uzaktan eğitim (Bliuc ve diğerleri, 2010; Gonçalves ve diğerleri, 2017; Howard ve diğerleri, 2016) ve karma eğitim (Sorour ve diğerleri, 2014; Zacharis, 2016) programlarında çalışmalar mevcuttur.

Bu verilerin işlenmesi, analiz edilmesi ve modellenmesi ile araştırmacılara yol gösterici sonuçlar elde edilebilmektedir. Geliştirilen modeller uzaktan eğitimin uygulama alanın geniş olmasından dolayı yaygın etkiye sahip olma potansiyeli taşımaktadır. Bu çalışmada Sakarya Üniversitesi (SAÜ) karma öğrenme programlarından biri olan İlahiyat lisans tamamlama (İLİTAM) öğrencilerinin verileri ele alınmaktadır. İLİTAM programı ilk olarak 2005 yılının Ekim ayında karma öğrenme yöntemi ile Ankara Üniversitesi'nde başlatılmış ve Türkiye'deki farklı üniversitelerde yürütülmeye başlanmıştır. Bu program, ön lisans programı mezunlarının lisans mezunu olabilmeleri için iki yıllık eğitim öğretim programını tamamlamalarını temel almaktadır. Alanyazında İLİTAM programı ile ilgili yapılan çalışmalar şu şekilde kategorilere ayrılabilmektedir;

1. Öğretim elemanlarının ve öğrencilerin karma ya da uzaktan eğitim programlarındaki süreçlere yönelik algılarını belirlemeyi hedefleyen çalışmalar (Arslan ve Korkmaz, 2019; Genç ve Ayhan, 2021; Gümrükçüoğlu ve Genç, 2020; Kablan, 2020; Karateke, 2020; Kaymakcan ve diğerleri, 2013)

2. Programın etkinliğini belirlemeyi hedefleyen çalışmalar (Imran ve diğerleri, 2019)

3. Programdaki müfredatta kullanılan materyal ve ders işleyiş tarzının uzaktan eğitim modeline uygunluğu, zorlukları ya da yaşanan sıkıntılara yönelik çalışmalar (Akaslan, 2020; Dağ, 2013)

İLİTAM ile ilgili yukarıda ifade edilen çalışmalarda genellikle nitel ve nicel yöntemler kullanılarak betimsel ve vaka çalışmalarının gerçekleştirildiği görülmüştür. Bu alanda yapılmış çalışmalarda daha

çok istatistiksel analizler kullanılmış olup veri madenciliği ve makine öğrenme uygulamalarına rastlanmamıştır. Bu yönüyle bu çalışmanın İLİTAM ve karma öğrenme öğrencileri verilerini ele alması bağlamında güncelliği bulunmaktadır. Ayrıca eğitsel veri madenciliği alanındaki literatür araştırmasında görülmüştür ki çalışmalar genellikle algoritmaların tahmin doğruluklarına ve kümelemeye odaklanmıştır. Ancak bu yapılırken veri boyutluluğu, sınıf dengesizliği, sınıflandırma hatası gibi problemler dikkate alınmamıştır (Imran ve diğerleri, 2019). Ayrıca kullanılan makine öğrenme algoritmalarının parametreleri için bazı çalışmalarda programın varsayılan değerleri kullanılırken bazı çalışmalarda ise bu değerler sezgisel olarak araştırmacı tarafından belirlenmektedir. Bu çalışmada diğer çalışmalardan farklı olarak sınıf dengesizliğini çözümlemek için değişkenler kategorileştirilmiş ve SMOTE yöntemi (Chawla ve diğerleri, 2002) kullanılmıştır. Ayrıca detaylı bir veri temizleme ve ön işleme adımı gerçekleştirilerek kullanılan algoritmaların parametre seçimi için hiper parametre (grid search) optimizasyonu kullanılmıştır. Bunun yanı sıra çalışmada algoritmaların parametrelerini optimize etmenin yanı sıra öğrenci mezuniyet başarı tahmin modelinde kullanılacak uygun makine öğrenimi algoritmalarının doğru şekilde seçilmesi ve başarıya etki eden değişkenlerin önem sırasının belirlenmesi hedeflenmektedir. Tahmin modeli başarılı olma ya da terk etme/başarısız olma durumları olmak üzere iki sınıftan oluşmaktadır. Başarılı sınıfı programdan başarı ile mezun olan öğrencileri, başarısız sınıfı ise programı başarısız olarak terk eden ve okuldan atılan öğrencileri temsil etmektedir. Bu yönüyle araştırma aşağıdaki sorulara cevap bulma amacıyla gerçekleştirilmiştir:

1. İLİTAM programında başarılı olma ya da terk etme/başarısız olma durumlarına etki eden değişkenlerin önem sıralaması nedir?

2. İLİTAM programına yeni kayıt yaptıran bir öğrencinin başarılı olma ya da terk etme/başarısız olma durumlarının eğitsel veri madenciliği teknikleri ile tahmin doğruluğu nedir?

3. Veri ön işleme adımlarının tahmin doğruluğuna katkısı nedir?

4. İLİTAM programından başarılı olma ya da terk etme/başarısız olma durumları ile bu durumlara etki eden değişkenler arasında nasıl ilişkiler ve kurallar vardır?

Yöntem

Araştırmada yöntem olarak eğitsel veri madenciliği kapsamında kullanılan (Chapman ve diğerleri, 2000) CRISP-DM süreç modelinden faydalanılmıştır. Çalışmada SAÜ İlahiyat Lisans Tamamlama karma öğrenme programındaki öğrencilerin üniversite genel bilgilerine göre başarılı olma ya da terk etme/başarısız olma durumlarının tahmin edilmesi ve bu durumu en çok etkileyen değişkenlerin tespit edilmesi hedeflenmiştir. Bunun yanında girdi bilgileri ile başarı durumu

arasındaki ilişkiler belirlenmiş ve programa yeni kayıt yaptıran öğrencinin olası başarı ya da başarısızlığını tahmini gerçekleştirilmiştir. Girdi değişkenlerinin sonucu ne şekilde etkilediğini tespit etmek için öznitelik ağırlıkları yöntemi, öğrencilerin başarı durumlarının tahmini için ise temel makine öğrenme algoritmaları kullanılmıştır. Öğrencilerin başarılı olma ya da terk etme/başarısız olma durumları ile değişkenler arasındaki ilişkilerin belirlenmesi için kural ve ağaç yapısı mantığı ile çalışan kural tümevarım (indüksiyon) yöntemi kullanılmıştır.

1999 yılında bir projede ortaya çıkartılmış olan CRISP-DM süreç modeli, veri madenciliği projelerini daha az maliyetli, daha güvenilir, daha tekrarlanabilir, daha yönetilebilir ve daha hızlı sonuç üretir hale getirmeyi amaçlamaktadır. Süreç, Tablo 1'de sunulan 6 adımdan oluşmaktadır (Wirth ve Hipp, 2000).

Tablo 1

CRISP-DM Modelinin Bileşenleri

Problemi Tanıma	Veriyi Anlama	Veri Ön İşleme ve Hazırlama	Modelleme	Değerlendirme	Dağıtım
Problemin hedeflerini belirlemek	Veri kalitesini doğrulamak	Veriyi seçmek	Modelleme tekniğini seçmek	Modeli değerlendirmek	Dağıtım planı
Problemi değerlendirmek	Verileri tanımlamak	Veriyi temizlemek	Test tasarımını geliştirmek	Süreci gözden geçirmek	Plan izleme ve bakım
Veri madenciliği hedeflerini belirlemek	Verileri keşfetmek	Veriyi oluşturmak	Model kurmak		Nihai rapor üretmek
Proje planı üretmek	Verileri toplamak	Verileri entegre etmek ve biçimlendirmek	Modeli değerlendirmek		Projeyi gözden geçirmek

Problemi Tanıma

Uzaktan eğitimde yüz yüze eğitimde olduğu gibi öğrencilerle temas halinde olup gözlemleme şansı biraz daha düşüktür. Bu nedenle uzaktan eğitim öğrencilerinin özellik ve davranışlarını inceleyerek değerlendirmek, öğrenme-öğretme süreçleri için büyük fayda sağlamaktadır. Uzaktan eğitimdeki öğrencilerin başarısını etkileyen en önemli değişkenlerin belirlenmesi ve öğrencilerin gelecekteki başarılı olma ya da terk etme/başarısız olma durumlarının belirlenmesi önemli bir problemdir. Çalışmada bu problemlere odaklanılmıştır.

Veriyi Anlama

Çalışmada kullanılan veri seti Sakarya Üniversitesi İLİTAM programında 2013-2016 yılları arasında giriş yapan öğrencilere ait veriler içermektedir. Veri seti iki veri tabanından toplanarak bir araya getirilmiştir. Verilerin bir kısmı öğrenci işleri veri tabanından (VT-1) elde edilen öğrenci profil bilgileri ve ÖSYM verileri, bir kısmı da SAÜ Bilgi Sisteminden (VT-2) elde

edilen başarı durumlarından oluşmaktadır. İki veri tabanındaki bilgiler öğrenci numaraları temel alınarak eşleştirilip bütünleştirilecek ve ardından analiz edilecektir. Çalışma ile ilgili etik onay, Sakarya Üniversitesi Etik Kurulu'nun 01/10/2021 tarihli 38 sayılı 57 nolu kararı ile alınan izin çerçevesinde yürütülmüştür. Öğrencilere ait değişkenler ile bu değişkenlerin açıklamaları Tablo 2'de sunulmaktadır.

Çalışmada sonuç değerini belirten hedef/çıktı değişkenin tahmin edilmesi amaçlanmaktadır. Bu değişken başarılı ve başarısız olarak kodlanmıştır. Başarılı kategorisi programdan başarı ile mezun olan öğrencileri, başarısız kategorisi ise programı başarısız olarak terk eden ve okuldan atılan öğrencileri içermektedir.

Tablo 2

Veri Önişleme Yapılmadan Oluşturulan İlk Veri Seti

Değişken Adı	Değişken Tanım	Veri Tabanı
Öğrenci Numarası	Öğrencinin Numarası	VT-1, VT-2
DoğumYılı	Öğrencinin doğum Yılı	VT-1
Cinsiyet	Öğrencinin cinsiyeti	VT-1
MezuniyetYılı	Liseden mezun olduğu sene	VT-1
YerleşmeKulPuan	ÖSYM puanı	VT-1
TercihSırası	Girdiği Bölüm Tercih Sırası	VT-1
ÖğretimYılı	Hangi öğretim yılında yerleştiği	VT-1
EnumOsymYerlesme	Yerleştirme türü	VT-1
Arşiv Bilgi	Başarı durumu	VT-2

Veri Önişleme ve Hazırlama

Veri seti ile çalışırken doğru analiz sonuçları, tahmin ve ilişkilere ulaşmak için ilk olarak problemi iyi tanımak ve incelenen veri setinin içerdiği verilerin düzgün ve anlaşılır olmasını sağlamak gerekmektedir. Bunun için veri önişleme süreci gerçekleştirilmektedir. Bu çalışmada gerçekleştirilen veri önişleme süreci, veri bütünleştirme, temizleme, dönüştürme ve azaltma işlemlerinden oluşmaktadır. Veri ön işleme ve hazırlama işlemleri Rapidminer 9.10.011 programı aracılığıyla gerçekleştirilmiştir.

Çalışmada ilk olarak iki ayrı veri tabanındaki veri setleri birleştirilerek bütünleştirilmiştir. Veri setinin girdi değerleri ile hedef değeri yani öğrencinin başarı durumu farklı veri tabanlarında bulunmaktadır. VT-1' de bulunan veriler, öğrencinin üniversiteye giriş ile ilgili genel özelliklerini içerirken VT-2' de bulunan veriler ise öğrencilerin mezuniyet durumlarını içermektedir. Giriş bilgileri için 3472 öğrenciye ait veri bulunmaktayken, başarılı olma ya da terk etme/başarısız olma durumları gibi arşiv bilgileri için 2189 öğrenciye ait veri bulunmaktadır. Veri setlerinin eşleştirme işlemi, arşiv bilgisi veri seti temel alınarak gerçekleştirilmiştir. Burada öğrenci numaraları iki veri seti için anahtar alan olarak kullanılmıştır. Eşleştirme sonucunda 2189 adet öğrenciye ait

bütünleşik yapıda bir veri seti elde edilmiştir. Veri ön işlemenin ikinci adımında veri setindeki eksik, hatalı, tekrarlı ve aykırı değerler incelenmiştir. Bu aşamada öğrenci numarası bilgisine göre 9 öğrenci verisinin tekrarlı olduğu tespit edilerek veri setinden çıkartılmıştır. 165 öğrencinin doğum yılı, cinsiyet, yerleşme puanı, girdiği bölüm tercih sırası, kontenjan ve öğretim yılı gibi bilgilerinde eksik kayıtlar olduğu için veri setinden çıkartılmıştır. Bunların yanında 782 öğrencinin kontenjan bilgisi hatalı olarak kaydedilmiş, 12 öğrencinin ise mezuniyet bilgisi hatalı olarak kaydedilmiş bunlar yorumlanamaz bir durum olduğu için veri setinden çıkartılmıştır. Arşiv durumlarından oluşan hedef değerde ise bir kişinin vefat ettiği için programı terk ettiği bilgisi bulunmaktadır. Bu nedenle bu öğrenci de veri setinden çıkartılmıştır. Tüm bu temizleme işlemleri sonucunda 1220 öğrenci verisini içeren bir veri seti elde edilmiştir. Ayrıca tüm değişkenler kutu grafiği yöntemi ile analiz edilmiştir ve veri setinde aykırı değer tespit edilmemiştir.

Veri setindeki bazı değişkenler mevcut hali ile değerlendirme ve yorumlama açısından uygun görülmemektedir. Doğum yılı değişkeninden üniversiteye giriş yaşını hesaplamak için öğretim yılı ile doğum yılı farkı alınmıştır. Bu şekilde "yaş" adında yeni bir nitelik oluşturulmuştur. Öğrencilerin liseden mezun olduktan kaç yıl sonra üniversiteye yerleştiğinin başarı durumu üzerindeki etkisini inceleyebilmek için üniversite öğretim yılından lise mezuniyet yılı farkı alınarak "üniversiteye başlama yılı" değişkeni elde edilmiştir. Ayrıca bölüme yerleşme puanını değerlendirirken her sene farklı taban ve tavan puanlar her sene farklı olduğu için puanlar farklı aralıklarda olabilmektedir. Bunun için giriş puanında z ve t dönüşümleri yapılarak puanlar ortak bir değere çekilmiştir. Üniversite öğrenci numarası kodlamasında son üç hane öğrencinin bölüme giriş sıralamasını temsil etmektedir. Bu nedenle "giriş sırası" değişkeni öğrenci numarası kullanılarak dönüştürülmüştür. Tercih sırası değişkeni daha anlamlı hale getirmek için kategorileştirilmiştir. Anlamlı değişkenler elde edebilmek için bu değişkenlerde Tablo 3' de sunulan dönüşüm işlemleri gerçekleştirilmiştir.

Tablo 3

Veri Dönüşüm İşlemleri

Eski değişken	Dönüşüm İşlemi	Yeni Değişken
Öğrenim Yılı-Doğum Yılı	Nitelik Oluşturma	Yaş
Öğrenim Yılı-Mezuniyet Yılı	Nitelik Oluşturma	Üniversiteye başlama yılı
YerleşmeKulPuan	Normalizasyon	Puan
Öğrenci numarası	Nitelik Oluşturma	Giriş Sırası
Tercih Sırası	Ayrıklaştırma	Sıra

Veri önişleme sonucunda 1220 adet öğrenci verisi için Tablo 4' deki değişkenler elde edilmiştir.

Tablo 4
Veri Seti-Özellikler

Özellik	Tanım	Türü	Değer
Yaş	Öğrencinin üniversiteye giriş yaşı	Sürekli	19-57
Cinsiyet	Öğrenci cinsiyeti	Nominal	Kız-Erkek
Üniversiteye başlama yılı	Lise sonrası programa kaç yılda yerleştiği	Sürekli	0-22
Puan	Yerleşme Puanı	Sürekli	34,73–107,93
TercihSıra	Yerleştiği program kaçıncı tercih	Nominal	1-3, 4-6, 7+
EnumOsymYerlesme	İlk yerleşme, ek yerleşme	Nominal	1,2
GirisSırası	Öğrencinin bölüme giriş sıralaması	Sürekli	1-599
BaşarıDurumu	Bölümden mezun olma durumu	Nominal	Başarılı, Başarısız

Verilerin analiz için hazırlanmasındaki bir diğer aşama ise, veri setindeki sınıfların denge durumunun incelenmesidir. Veri setindeki sınıfların yaklaşık olarak eşit sayıda veri içermesi tercih edilmelidir aksi durumda veri seti dengesiz olarak nitelendirilmektedir. Bu dengesizlik çoğunluk sınıfın yüksek doğruluklarla, azınlık sınıfının ise düşük doğruluklarla tahmin edilmesine neden olmaktadır. Sonuç olarak bu durumda genel doğruluk oranı ile yorum yapılması yanıltıcı olabilir. Sınıf dengesizliğini çözümlemek için uygunluğa göre, birbirine yakın sınıflar birleştirilebilir, yoğun olan sınıfa ceza puanı uygulanabilir ya da sentetik veri üretimi ile yeniden örnekleme yapılabilir (Chawla, 2005; Longadge ve Dongre, 2013; Romero ve diğerleri, 2008). Yeniden örnekleme için yaygın olarak kullanılan bir yöntem Sentetik Azınlık Aşırı Örnekleme (SMOTE) yöntemidir. SMOTE tekniği azınlık sınıfındaki veri sayısını çoğaltma prensibini temel alır ve iki benzer örneğin doğrusal kombinasyonlarını hesaplayarak işlem yapmaktadır (Chawla ve diğerleri, 2002). Veri dönüşümü süreci sonucunda elde edilen veri setinde temel etiket analizi yapılmıştır. Hedef değişkeninde başarılı kategorisinden 1138 adet örnek varken başarısız kategorisinden 82 örnek adet örnek bulunduğu görülmüştür. Veri setine SMOTE yöntemi uygulanarak başarısız kategorisindeki örnek sayısı başarılı kategorisindeki örnek sayısına eşitlenmiştir. Bunun sonucunda 2276 adet örnek elde edilmiştir.

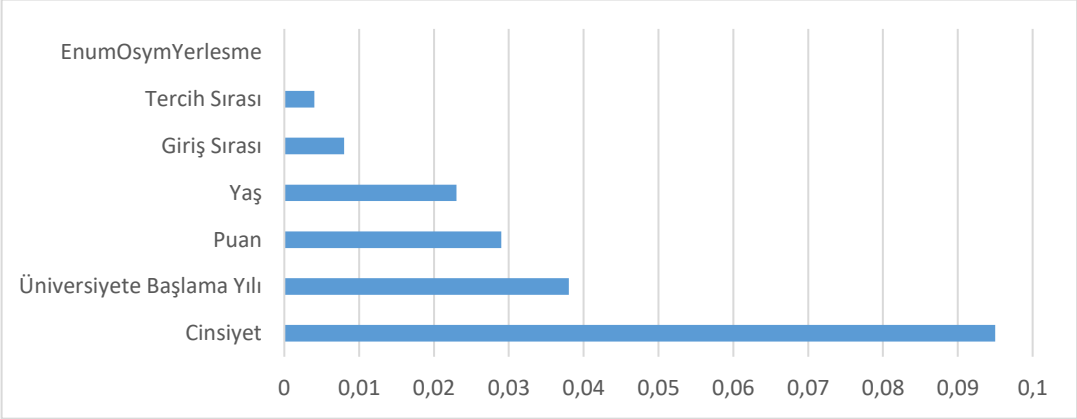
Öznitelik Ağırlıkları

Bilgi kazanımı, özniteliklerin yani değişkenlerin ağırlığını belirlemek için kullanılan popüler bir filtre modeli ve tekniğidir (Prasetiyowati ve

diğerleri, 2021; Sokkhey ve Okazaki, 2020). Her bir özniteliğin hedef değer üzerinde ne kadar etkiye sahip olduğunu görmeyi sağlayan bu yöntem eğitim alanında da kullanılmaktadır (Çifçi ve diğerleri 2018). Çalışmada öğrenci başarısına etki eden değişkenlerin ağırlıkları bilgi kazanımı yöntemi ile elde edilmiştir.

Çalışmada öznitelik ağırlıkları ve diğer tüm analizlerin uygulaması Rapidminer Studio 9.10.011 programı ile gerçekleştirilmiştir. Özniteliklerin ağırlıklarına göre incelendiğinde, sonuç değişkenine en çok etki eden ve en belirleyici olan değişken 0,095 değeri ile "Cinsiyet" değişkenidir. İkinci sırada ise 0,038 değeri ile "Üniversiteye başlama yılı" değişkeni yer almaktadır. Ardından 0,029 değeri ile "Puan", 0,023 değeri ile "Yas", 0,008 değeri ile "GirisSırası" ve 0,004 değeri ile "TercihSıra" değişkenleri gelmektedir. Sonuca en az etki eden değişken ise öğrencilerin asil ya da yedek olarak yerleştiğini belirten "EnumOsymYerlesme" değişkeni olarak elde edilmiştir. Özniteliklerin sıralaması ise Şekil 1'de gösterildiği gibidir.

Şekil 1
Öznitelik ağırlıkları



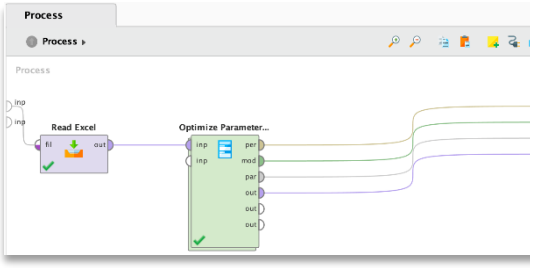
Modelleme

Çalışmada öğrenme yöntemi olarak k-kat çapraz doğrulama yöntemi kullanılmıştır. Bu yöntemde, verilerin bir parçası test yani doğrulama diğer parçaları ise eğitim için kullanılmaktadır. Yöntem ilk parçanın test diğer parçaların eğitim verisi olarak ele alınması ile başlar ve model çalıştırılır ardından ikinci parça test diğerleri eğitim verisi olarak ele alınır. Tüm parçalar test verisi olarak değerlendirildiğinde süreç tamamlanmış olmaktadır. Elde edilen tüm doğrulukların ortalamasıyla genel doğruluk elde edilmektedir (Refaeilzadeh ve diğerleri, 2009).

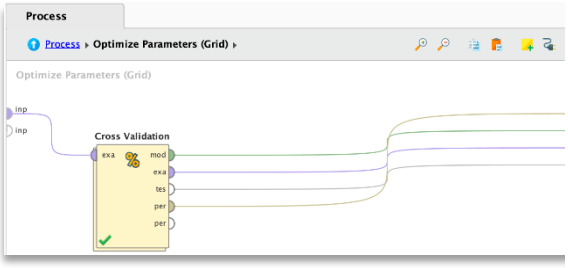
Çalışmada kullanılan algoritmaların parametreleri ise grid optimizasyon tekniği kullanılarak belirlenmişken geliştirilen tahmin modellerinin eğitim aşaması k-kat çapraz doğrulama ile gerçekleştirilmiştir.

Uygulanan optimizasyon ve eğitim modelleri Şekil 2 (a) ve Şekil 2 (b)'de görülmektedir.

Şekil 2 (a)
Optimizasyon modeli



Şekil 2 (b)
k-kat çapraz doğrulama



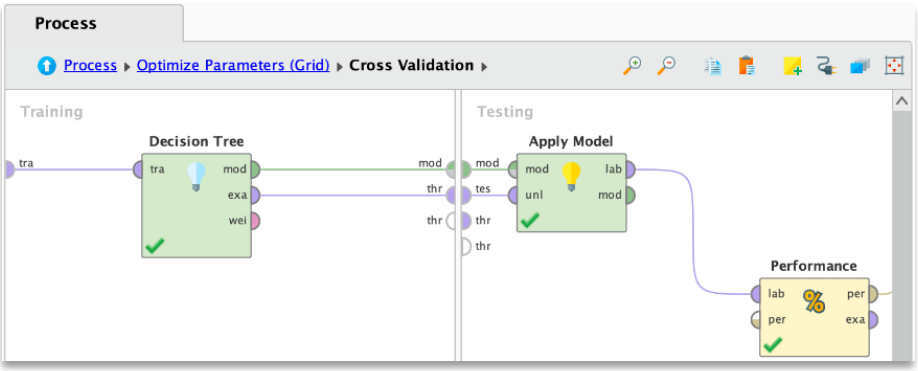
Çalışmada öğrenci performansının tahmin edilmesi için en temel makine öğrenme algoritmalarından; karar ağacı, k-en yakın komşu, naive bayes algoritmaları ve ilişki kurallarını çıkartmak için kural çıkarımı kullanılmıştır.

Karar Ağacı (KA)

KA, diğer algoritmalara göre daha yaygın kullanılan hem nominal hem de sayısal değişkenleri kabul eden bir yöntemdir. Algoritmanın uygulama adımları veri setini alt bölümlere ayırarak ağaç yapısına göre ilerler. Süreç kökten başlar ve ara düğümden devam ederek yaprak düğüme kadar dallanır, yapraklar ise sınıfları oluşturur. Yorumlanması ve anlaşılması kolay olan algoritmada EĞER-İSE kuralları oluşturulmaktadır (Bilgin, 2018; Mitchell, 1997). Kök düğümün hangi değişken olacağını belirlemek için bazı kriterler mevuttur. Bunlar, bilgi kazancı, gini indeksi ve kazanç oranıdır. Bu kriter algoritmanın çalışma performansına göre seçilebilir (Maimon ve Rokach, 2005).

Çalışmada Şekil 3' te görüldüğü gibi klasik karar ağacı kullanılmıştır. RapidMiner' da bulunan karar ağacı öğrenicisi, Quinlan'ın C4.5 veya CART'ına (Quinlan, 1986) benzer şekilde çalışmaktadır.

Şekil 3
Klasik Karar Ağacı Modeli



Grid optimizasyon kullanılarak elde edilen parametre sonuçları Tablo 5'de sunulmaktadır. Eğitim aşamasında kullanılan k-kat çapraz doğrulama yönteminin k değeri 1-10 arasında, örnekleme türü ise dört farklı örnekleme türü değişecek şekilde optimizasyon tekniği ile incelenmiştir. Bununla birlikte karar ağacının kök düğüm ve derinlik parametreleri de eşzamanlı olarak 144 iterasyon sonucunda optimizasyon tekniği ile belirlenmiştir.

Tablo 5
KA Parametreleri

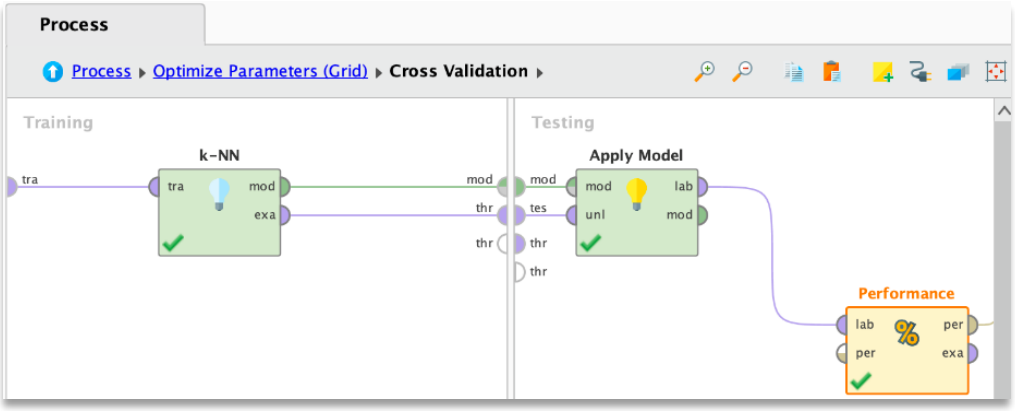
Parametre	Değer
k-kat çapraz doğrulama k sayısı	10
k-kat çapraz doğrulama örnekleme türü	Rastgele
Karar ağacı kök düğüm kriteri	Gini_index

K-en Yakın Komşu (K-NN)

En yakın komşu algoritmasının çalışma prensibi oldukça kolaydır. Kullanıcı tarafından belirlenen komşu sayısına göre seçim yapılarak o komşulara benzerliklere göre yeni örneğin ataması yapılır. Bu basit çalışma şekline rağmen algoritma oldukça etkilidir (Han ve diğerleri, 2011). Ancak sınıflandırma aşaması biraz yavaş ilerlemektedir ve eksik veri olduğunda ilave işlemler gerekmektedir (Lantz, 2019). En yakın komşuyu tespit ederken iki örnek arasındaki benzerlik ölçülmektedir. Bu mesafeyi hesaplamak için Öklid, Manhattan ve Minkowski uzaklığı kullanılabilmektedir (Bilgin, 2018).

Çalışmada K-NN temelli olarak uygulanan tahmin modeli Şekil 4' de sunulmaktadır.

Şekil 4
K-NN Modeli



K-NN modelinin eğitimi için kullanılan parametreler, en yakın komşu sayısını temsil eden k sayısı ve benzerlik ölçüleri için 484 iterasyon

sonucunda optimizasyon ile elde edilen parametre değerleri Tablo 6' da sunulmuştur.

Tablo 6
K-NN Parametreleri

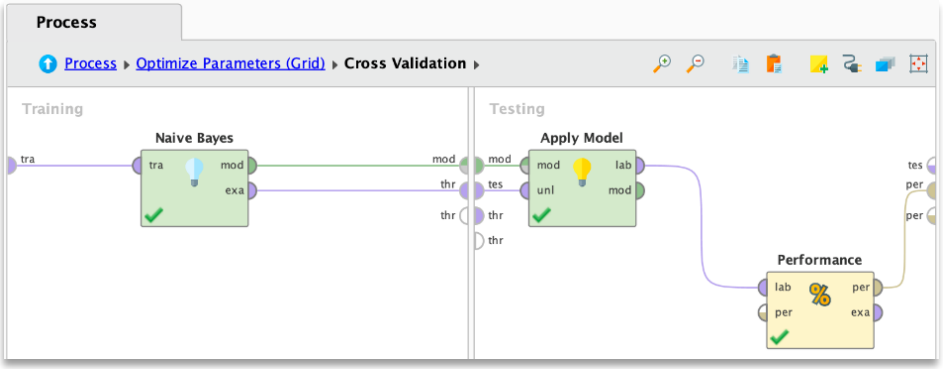
Parametre	Değer
k-kat çapraz doğrulama k sayısı	71
k-kat çapraz doğrulama örnekleme türü	Rastgele
K-NN k sayısı	1
Benzerlik türü	Karma Ölçüm
Benzerlik ölçüsü	Karma Oklid Uzaklığı

Naive Bayes (NB)

Öğrenme algoritmalarının arasındaki en pratik yöntemlerden biri olan Bayesci yaklaşım, önceki varsayım olasılığını, verilerin görülme olasılıklarını ve gözlemlenen verileri temel alarak bir hipotezin olasılığını hesaplamaktadır (Mitchell, 1997). Bayes yaklaşımını temel alan NB her çıktının görülme sıklığını ve bağımsız değişkenler ile bağımlı değişken kombinasyonunun kaç kere görüldüğünü incelemektedir (Bilgin, 2018). NB yönteminde bir varlığın belirli bir sınıfa ne kadar iyi uyduğu değerlendirilmektedir. Bunun için de üstünlük (odds) değeri temel alınmaktadır. Üstünlük değeri, bir varlığın bir hedef sınıfa ait olma olasılığının o sınıfa ait olmama olasılığına oranıdır (Byeon, 2022).

Çalışmada kullanılan NB temelli tahmin modeli Şekil 5' de sunulmaktadır.

Şekil 5
NB Modeli



NB algoritması için kullanılan tek parametre laplace korelasyonudur. Bunun dışında k-kat çapraz doğrulama ile ilgili parametre seçimi yapılmıştır. 88 iterasyon sonucunda optimizasyon yöntemi ile elde edilen parametre değerleri Tablo 7' de sunulmuştur.

Tablo 7
Örnekleme Türüne Göre NB Doğruluk Değerleri

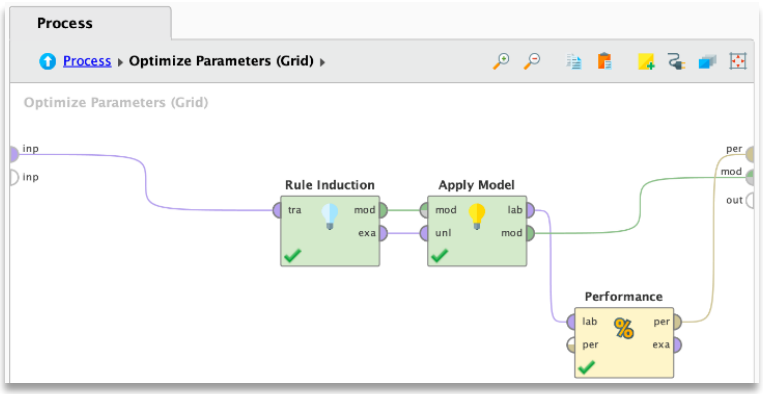
Parametre	Değer
k-kat çapraz doğrulama k sayısı	2
k-kat çapraz doğrulama örnekleme türü	Rastgele
Laplace korelasyon	Yok

Kural Tümevarımı

Kural tümevarım (rule induction) yönteminde daha az yaygın olan sınıflardan başlayarak, algoritma yinelemeli olarak büyümekte ve hiçbir pozitif örnek kalmayana veya hata oranı %50 'den fazla olana kadar kuralları budama işlemi devam etmektedir. Büyüme aşamasında, her kural için, her bir özelliğin olası her değerini deneyerek en yüksek bilgi kazancına sahip koşul seçilmektedir. Sonuçta elde edilen kurallar ile verideki değerlerin kolay anlaşılabilir bilgiye dönüşmesi sağlanmaktadır. Bilimsel olarak modeli temsil edebileceği gibi kısmen de açıklama yapabilir. Her ne kadar KA algoritmasına benzetilse de büyük eğitim verisinde kısmen kötü çalışma dezavantajı dışında karar kurallarının anlaşılması ve yorumlanması karar ağaçlarına göre daha kolaydır (Rapidminer, Rule Induction, 2020).

Çalışmada uygulanan kural tümevarımı Şekil 6' da sunulmaktadır.

Şekil 6
Kural Tümevarım



Kural tümevarımı için optimizasyon yöntemi ile elde edilen parametreler Tablo 8' de sunulmaktadır.

Tablo 8
Kural Çıkarımı Parametre Değerleri

Parametre	Değer
Kural çıkarımı kriteri	Doğruluk
Kural çıkarımı örnek oranı	1

Değerlendirme

Veri madenciliği problemlerinde modelin başarısı veri ön işleme, parametre seçimi ve test kümesinin belirlenmesi faaliyetlerine bağlıdır. Bu adımların ne kadar doğru şekilde gerçekleştiği ve uygulayıcı için kabul edilebilir bir düzeyde başarı elde edilip edilmediğinin belirlendiği kısım değerlendirme aşamasıdır. Sınıflandırma problemlerinde tahmin başarısının ölçülmesi için doğruluk kriteri ile ilgili ölçümler yapılmalıdır. Bunlar; doğruluk oranı, duyarlılık, kesinlik, F ölçütü ve kappa olarak sayılabilir. Bu kriterlerden bir ya da birkaçı ihtiyaca ve problemin gerekliliğine göre seçilebilmektedir.

Çalışmada algoritmalar doğruluk, duyarlılık, kesinlik ve kappa istatistik kritelerine göre değerlendirilmiştir. Doğrulukların elde edildiği hata matrisi Tablo 9' da yer almaktadır. Matriste sütunlar gerçek değerleri, satırlar ise tahmin değerlerini ifade etmektedir (Bilgin, 2018).

Tablo 9
Hata Matrisi

	Gerçek Başarılı Sınıf	Gerçek Başarısız Sınıf
Tahmin Başarılı Sınıf	G1	Y0
Tahmin Başarısız Sınıf	Y1	G0

Burada;

G1: test kümesindeki doğru sınıflandırılan başarılı örnek sayısı,
G0: test kümesindeki doğru sınıflandırılan başarısız örnek sayısı,
Y1: test kümesindeki yanlış sınıflandırılan başarılı örnek sayısı,
Y0: test kümesindeki yanlış sınıflandırılan başarısız örnek sayısı
olarak ifade edilmektedir.

Yapılan analizler sonucunda elde edilen sonuçlar Tablo 10'da sunulmuştur.

Tablo 10
Algoritma Sonuçları

Algoritma	Doğruluk	Duyarlılık	Kesinlik	Kappa
KA	%87,04	%87,04	%87,59	0,740
NB	%64,50	%64,49	%67,62	0,290
K-NN	%91,30	%91,46	%92,11	0,824

Başarı durumu tahmini için KA, NB ve K-NN olmak üzere üç algoritma seçilmiştir. Tablo 10 incelendiğinde K-NN algoritmasının %91,30 doğruluk, %91,46 duyarlılık, %92,11 kesinlik ve 0,824 kappa istatistiği değeri ile en başarılı tahmin değerlerine ulaştığı görülmektedir. Ardından %87,04 doğruluk, %87,04 duyarlılık, %87,59 kesinlik ve 0,740 kappa istatistiği değeri ile KA algoritması gelmektedir. Tahmin değerleri en düşük olan algoritma ise %64,50 doğruluk, %64,49 duyarlılık, %67,62 kesinlik ve 0,29 kappa istatistiği değeri ile NB

algoritması olduğu söylenebilmektedir. Bu nedenle çalışmada tahmin modeli için en uygun sınıflandırma algoritmasının K-NN algoritması olduğu söylenebilmektedir.

Dağıtım

Bu aşamada uygulama sonucunda elde edilen ilişki kurallarına göre politikalar geliştirilerek eğitmen ve yöneticilere yol gösterilmesi amaçlanmaktadır. Verilere göre geliştirilen politikaların eğitim-öğretim sistemleriyle bütünleştirilmesi, süreçlerin iyileştirilmesi ve daha başarılı toplumlara ulaşılmasını sağlayacaktır.

Bulgular

SAÜ İLİTAM öğrencilerinin genel bilgileri ve öğrencinin başarılı olma ya da terk etme/başarısız olma durumları ile ilgili gerçekleştirilen çalışma kapsamında dört farklı araştırma sorusu mevcuttur.

Bunlar;

a) Araştırma Sorusu 1: İLİTAM programında başarılı olma ya da terk etme/başarısız olma durumlarına etki eden değişkenlerin önem sıralaması nedir?

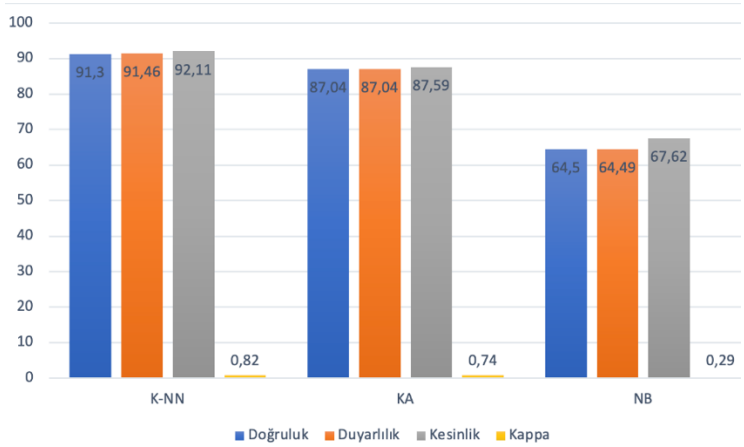
Özniteliklerin etki ağırlıkları için bilgi kazanımı algoritması kullanılmıştır. Öznitelik ağırlıklandırma sonucu olarak; en çok etki eden ve en belirleyici olan değişken “Cinsiyet” değişkeni olarak bulunmuştur. İkinci sırada etki eden değişken ise 0,038 değeri ile “Üniversiteye başlama yılı” değişkenidir. Ardından sırasıyla “Puan”, “Yas”, “GirisSırası” ve “TercihSıra” değişkenleri gelmektedir. Sonuca en az etki eden değişken ise öğrencilerin asil ya da yedek olarak yerleştiğini belirten “EnumOsymYerlesme” değişkeni olarak elde edilmiştir.

b) Araştırma Sorusu 2: İLİTAM programına yeni kayıt yaptıran bir öğrencinin başarılı olma ya da terk etme/başarısız olma durumlarının eğitsel veri madenciliği teknikleri ile tahmin edilme doğruluğu nedir?

Bir öğrenme algoritmasının maliyet ve çözüm süresinin yanındaki en önemli kriterlerinden biri de çözüm ve tahmin doğruluğudur. Tahmin doğrulukları algoritmaların performansına, verinin uygunluğuna ve kullanılan parametrelere göre değişmektedir. Ancak bu yüksek tahmin doğruluklarının göz ardı edilmesine sebep olabilmektedir. Çalışmada bu sorunun önüne geçebilmek için parametre optimizasyon tekniklerinden yararlanılmıştır. Ayrıca tahmin doğrulukları için en eski ve temel algoritmalardan olan ve çalışmalarda sıklıkla kullanılan (Akram ve diğerleri, 2019; Durairaj ve Vijitha, 2014; Rojanavasu, 2019) KA, K-NN ve NB algoritmaları kullanılmıştır. Algoritmaların tahmin başarısı ise, doğruluk, kesinlik, duyarlılık ve kappa ölçütüne göre değerlendirilmiştir.

Şekil 6'da sunulan sonuçlara göre; üç algoritma arasından en başarılı tahmin değerlerine ulaşan sınıflandırma algoritması K-NN algoritması ile elde edilmektedir. Ardından KA algoritması gelmektedir ve en düşük başarı oranına sahip algoritma ise NB algoritması olarak görülmektedir. Sonuçlara göre problem için uygun olan sınıflandırma modelinin K-NN algoritması ile geliştirilen model olduğu söylenebilmektedir.

Şekil 6
Tahmin Performans Sonuçları



c) Araştırma Sorusu 3: Veri ön işleme adımlarının tahmin doğruluğuna katkısı nedir?

Veri önişleme adımları uygulanmadan önce veri seti 1289 öğrenciye ait değer içermektedir. Veri ön temizleme adımları uygulandıktan sonra ise 1220 öğrenciye ait değer ile analizler yapılmıştır. Ardından veri dengesizliğini ortadan kaldırmak için uygulanan SMOTE yöntemi ile 2276 örnek elde edilmiştir. Elde edilen sonuçlarla ilgili bulgular aşağıda sıralanmıştır.

Algoritmaların, dört performans kriterine göre; veri ön işleme öncesi ve sonrasına ait tahmin doğrulukları Tablo 11'de sunulmaktadır. Değerler incelendiğinde, K-NN ve KA algoritmaları için tahmin performanslarının veri önişleme gerçekleştirilmeden önce düşük olduğu ve veri ön işleme adımlarının tahmin performanslarını olumlu yönde etkilediği görülmektedir. Bu iki algoritmanın aksine NB algoritmasının ise tahmin doğruluğunun veri ön işleme sonrasında azaldığı görülmektedir.

Tablo 11
Veri Önişleme öncesi ve sonrası tahmin performans değerleri

Algoritmalar	Doğruluk Oranları	
	Veri ön işleme öncesi	Veri ön işleme sonrası
KA	%78,49	%87,04
NB	%73,74	%64,50
K-NN	%76,91	%91,30

d) Araştırma Sorusu 4: İLİTAM programından başarılı olma ya da terk etme/başarısız olma durumları ile bu duruma etki eden değişkenler arasında nasıl ilişkiler ve kurallar vardır?

İLİTAM programında öğrenim gören öğrencilerin seçilen değişkenleri arasındaki ilişkilerin belirlenmesi için kural tümevarımı algoritmalarından yararlanılmıştır. Kural türeten bu algoritmalarda çoğu zaman, veri kümesi büyük olduğunda, fazla sayıda kural elde edilmektedir. Kurallardan bazıları sadece iyi bilinen alan bilgisine karşılık gelebilirken, bazıları değerlendirici açısından önemli olmayabilir. Bu nedenle önemli görülen kurallar seçilmelidir (McClean, 2003). Yani kurallar bilimsel olarak modeli temsil edebileceği gibi kısmen de açıklama yapabilmektedir. Bu çalışmada kural tümevarım algoritmalarından kural çıkarımı gerçekleştirilmiştir. Sonuç olarak 23 adet kural elde edilmiş ancak anlamlılık, kullanılabilirlik ve işe yararlılık açısından değerlendirilip uzman görüşü alınarak aşağıdaki 10 temel bulgu kullanılabilir olarak ele alınmıştır.

- Kız öğrenciler programa yerleşme için yaptıkları tercih sıralamasında bu programı 2. veya 3. sırada tercih ettilerse bölümden mezun olamamaktadırlar.
- Erkek öğrenciler programa yerleşme için yaptıkları tercih sıralamasında bu programı 3. sırada tercih ettilerse bölümden başarılı şekilde mezun olmaktadırlar.
- Giriş sırası bölüm sıralamasına göre sonlarda olsa bile yerleşme yılı yarım dönemden az olan öğrenciler başarılı şekilde mezun olmaktadır.
- Puan değeri ortalama ve düşük düzeyde olan erkek öğrenciler mezun olamamaktadır.
- Giriş sırası bölüm sıralamasına göre sonlarda olsa da yaşı 28’ den küçük olan tüm öğrenciler eğer puan değeri de ortalama düzeyde ise mezun olmaktadırlar.
- Giriş sırası bölüm sıralamasına göre sonlarda veya üst sıralarda olan kız öğrenciler programdan mezun olamamaktadır.

Tartışma ve Sonuç

Bu çalışma, İLİTAM karma öğrenme programlarında başarıya etki eden değişkenlerin belirlenmesi ve öğrencilere yönelik yol gösterecek eğitim politikalarının geliştirilmesi için gerçekleştirilmiştir. Bu amaçla, SAÜ İLİTAM programına 2013-2016 yılları arasında giriş yapan öğrencilerin verileri kullanılmıştır. Çalışmanın veri seti iki veri setinin eşleştirilerek birleştirilmesi ile oluşturulmuştur. Birinci veri seti öğrencinin üniversiteye giriş bilgilerine ait değişken değerlerini içeren girdi veri setidir. İkinci veri seti ise öğrencilerin üniversiteden mezun olma başarı durumlarını içeren çıktı veri setidir. İki veri setinde aynı öğrenciye ait değerler eşleştirilmiştir. Değişkenlerin belirlenebilmesi ve politikalara öngörü sağlayacak sürece ait tahminlerin yapılabilmesi için eğitsel veri

madenciliği kullanılmıştır. Bunun için ilk olarak veri, ön işleme adımları ile analiz için hazırlanmıştır. Tekrarlı, eksik değer temizleme, nitelik oluşturma ve veri çoğaltma işlemlerinden sonra 2276 öğrenciye ait değer içeren bir veri seti elde edilmiştir. Ayrıca veri dönüştürme işlemi ile değişkenlerde dönüşüm yapılarak daha anlamlı değişkenler elde edilmiştir.

İlk veri setinde 9 değişken mevcutken veri dönüşümünden sonra 8 değişken içeren, ilk veri setine göre daha dengeli bir veri seti elde edilmiştir. Elde edilen veri seti ile KA, K-NN ve NB modelleri eğitilmiştir. Sonuç olarak veri ön işleme süreci (Costa ve diğerleri, 2017; Romero ve diğerleri, 2008) çalışmaları ile tutarlı olarak modellerin tahmin doğruluklarının artmasını sağlamıştır. Doğruluk, kesinlik, duyarlılık ve kappa değerlerine göre incelendiğinde çalışmadaki sınıflandırma probleminde en uygun algoritmanın KNN olduğu sonucuna varılmıştır. Benzer şekilde Yükseltürk vd. (2014) tarafından okulu terk edecek öğrencilerin tahmin edilesi için gerçekleştirilen çalışmada da K-NN algoritması başarılı olarak belirtilmiştir. Bunlara karşın literatürde farklı algoritmaların da başarılı olduğu çalışmalar mevcuttur (Devasia ve diğerleri, 2016; Kabakchieva, 2013; Khasanah, 2017). Bunun nedeni bir makine öğrenmesi veya veri madenciliği analiz ve yorumunun kalitesinin, büyük ölçüde girdi verilerinin kalitesine bağlı olmasıdır (Lantz, 2015). Örneğin, doğru belirlenmiş değişkenler, algoritma ne kadar karmaşık olursa olsun diğer modellerden daha iyi performans gösteren modeller verme eğilimindedir. Genel olarak doğru değişken setine sahipseniz, basit bir model bile iyi performans göstererek istenen sonuçları vermektedir (Kamath ve Choppella, 2017).

Çalışmadan elde edilen diğer sonuç ise değişkenlerdeki önem sıralamasıdır. Sivakumar vd. (2016) öğrencilerin öğretim programını terk etme nedenlerini incelemişlerdir. Terk etme nedenleri ağırlıklı olarak öğrencilerin ailevi ve kişisel problemleri, ev ya da yurtta barınma gibi sosyal problemleri açısından ele alınmıştır. Programı terk etmelerinde en önemli etken ailevi nedenler olarak tespit edilmiştir. Bu çalışmada farklı olarak öğrencilerin terk etme durumları akademik açılardan ele alınmış ve bunlara yönelik eğitsel politikalar önermek amaçlanmıştır. Bu çalışmada sonuç üzerinde en etkili olan değişkenin öğrencilerin cinsiyeti olduğu görülmüştür. Erkek öğrencilerin bir kısmının farklı bir değişkene bağlı olmaksızın programı bırakma eğiliminde olduğu, kızların ise giriş puanı ve üniversiteye yerleşme sürelerine bağlı olarak programı bıraktıkları görülmektedir. Ayrıca kural çıkarımının diğer bulguları olan, kız öğrencilerin programa yerleşme için yaptıkları tercih sıralamasında bu programı 2 veya 3. sırada tercih etmeleri durumunda bölümden mezun olamamalarına karşın erkek öğrencilerin programa yerleşme için yaptıkları tercih sıralamasında bu programı 3. sırada tercih etseler dahi başarılı olmaları öznitelik önem sıralamasını destekler niteliktedir. Hakyemez (2015)'in çalışmasında da olduğu gibi cinsiyet

değişkenin akademik başarı durumu üzerinde önemli bir etkisi olduğu söylenebilmektedir. Öğrencilerin başarısız olma durumunu belirleyici olan ikinci değişkenin ise yerleşme yılı olduğu görülmektedir. İncelenen kurallara göre giriş sırası bölüm sıralamasına göre sonlarda olsa bile yerleşme yılı yarım dönemden az olan öğrenciler başarılı şekilde mezun olmaktadır. Bu durum öğrencinin çalışma motivasyonunun kesintisiz sürdürülmesi ile ilgili olabilir. Kural çıkarımından elde edilen bir diğer sonuç ise puan değeri ortalama ve düşük düzeyde olan erkek öğrencilerin mezun olamamasıdır. Bunun nedeni ders çalışma motivasyonu düşük olan erkek öğrencilerin eğitim hayatını bırakıp iş hayatına atılma istekleri olabilir. Bir diğer kural sonucu ise, öğrencinin giriş sırası bölüm sıralamasına göre sonlarda olsa da yaşı 28' den küçük olan tüm öğrenciler eğer puan değeri de ortalama düzeyde ise mezun olmasıdır. Ayrıca giriş sırası bölüm sıralamasına göre sonlarda veya sıralamanın üst sıralarında olan kız öğrenciler programdan mezun olamamaktadır. Bunun nedeni sıralamada aşağıda kalan öğrencilerin yaşadığı başarısız olma kaygısı, üst sıralardaki öğrencilerin ise başarılı olarak puanı daha yüksek bölümlere geçme isteği olabilir. Shahiri ve Husain (2015) tarafından öğrenci performansını belirlemeye yönelik gerçekleştirilen çalışmada genel not ortalaması değişkeninin yanında en çok etki eden değişkenler cinsiyet ve yaş olarak belirtilmiştir. Bu çalışmada da elde edilen sonuçlara benzer olarak yaş ve cinsiyet değişkenlerinin önemi ortaya konmaktadır.

Çalışmada elde edilen tahmin başarı oranları ve kurallar incelendiğinde bu alanda bu tekniklerin kullanılmasının verimli olacağı görülmektedir. Literatür incelemesi sonucunda karma eğitimde daha önce veri madenciliği teknikleri kullanılarak gerçekleştirilen çalışmalar olsa da İLİTAM programında bu alanda bir çalışmaya rastlanmamıştır. Bu nedenle çalışmanın alanda yol gösterici ve fikir verebilecek bir çalışma niteliğinde olduğu söylenebilir. Eğitim politikalarındaki yanlış stratejilerden dolayı, öğrenci başarılarında ve mezun kalitesinde sorunlar yaşanabilmektedir. Veriler arasındaki örüntü ve ilişkilerin analiz edilerek bunlara göre elde edilen sonuçlarla politika geliştirmek, öğrencilerin daha donanımlı şekilde mezun olmalarını ve okudukları bölümü terk etmemelerini önlemeye yardımcı olacaktır. İLİTAM bölüm koordinatörlerine ve ders veren öğretim üyelerine eğitim sürecinde öğrencilere yönelik ne gibi konulara dikkat etmesi, nasıl davranması ve ne gibi faaliyetler gerçekleştirmesinin faydaları olabileceği yönünde fikir verebilecek bir çalışma niteliğinde olduğu söylenebilir.

Çalışmadaki umut verici sonuçlara rağmen, çalışmanın bazı kısıtları bulunmaktadır. Türkiye'de 11 tane üniversitede İLİTAM programı bulunmasına karşın verilere ulaşma sorunu yaşanmaktadır. Bundan dolayı çalışmada veri toplama faaliyeti sadece SAÜ İLİTAM programında sistemsel olarak kayıtlara ulaşılabilmenin mümkün olduğu 2013-2016 yılları arasında programa kayıt yaptıran öğrenciler çerçevesinde

gerçekleştirilmiştir. Ayrıca çalışmada tahmin ve kural çıkarımı için temel algoritmalardan KA, K-NN ve NB algoritmaları kullanılmıştır. Kullanılabilecek çok fazla algoritma olduğu halde diğer algoritmaların kullanılmaması çalışmanın bir diğer sınırlılığı olarak belirtilebilir. Algoritma çeşitliliğinin yanında bu çalışma kapsamında toplanan veri setindeki nitelikler de kısıtlı olarak elde edilebilmiştir. Kullanılan niteliklerin yanına öğrencilerin ailevi ve sosyal yaşamlarını belirten niteliklerin katılması ile eğitilen modellerin performanslarının incelenmesinin yararlı olacağı düşünülmektedir. Ayrıca gelecek çalışmalarda literatürde eğitim alanı için önerilmiş olan veri madenciliği CRISP-EDM (Özdemir, 2016) yaklaşımının kullanılması planlanmaktadır.

Öneriler

Çalışmadan elde edilen sonuçlardan yola çıkarak bazı öneriler geliştirilebilir. İlk olarak programda görev alan eğitmenler derslerini alan öğrencilerin bölüme giriş tercih sıraları ile bölüme yerleşme sırasını belirten giriş sırası değişkenlerine dikkat ederek öğrencilerin derse katılım ve motivasyonunu sağlayamaya yönelik farklı materyal ya da uygulama kullanabilir ve farklı kaynak önerisi yapabilirler. Ayrıca programda sunulacak oryantasyon eğitiminde bu öğrencileri daha fazla motive etmek ve durumla ilgili farkındalık oluşturmak için etkinlikler yürütülebilir.

Etik Kurul İzin Bilgisi: *Bu araştırma, T.C. SAKARYA ÜNİVERSİTESİ REKTÖRLÜĞÜ Etik Kurulunun 01/10/2021 tarihli 38 sayılı 57 nolu kararı ile alınan izinle yürütülmüştür*

Yazar Çıkar Çatışması Bilgisi: *Yazarların beyan edeceği herhangi bir çıkar çatışması bulunmamaktadır.*

Yazar Katkısı: *Yazarlar bu makaleye eşit oranda katkıda bulunmuştur.*

Kaynakça

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24(4), 567-598. https://doi.org/10.1007/s10758-019-09408-7.

Aghalarova, S. ve Keser, S. B. (2021). Önerilen Yapay Sinir Ağı Algoritması ile Ortaokul Öğrencilerin Akademik Performansının Tahmini. *Veri Bilimi*, 4(2), 19-32.

Akaslan, Y. (2020). Mahiyet, nitelik ve müfredat açısından ilitam programlarında Kur'an-ı kerim dersleri (Ondokuz Mayıs Üniversitesi Örneği). *Ondokuz Mayıs Üniversitesi İlahiyat Fakültesi Dergisi*, 49, 9-37. https://doi.org/10.17120/omuifd.779343

Akçapınar, G. (2014). *Çevrimiçi öğrenme ortamındaki etkileşim verilerine göre öğrencilerin akademik performanslarının veri madenciliği yaklaşımı ile modellenmesi*. [Yayınlanmamış Doktora Tezi]. Hacettepe Üniversitesi.

Akgün, E. (2019). 2023 Eğitim vizyonunda eğitsel veri madenciliği. *Seta Perspektif*, 228,1-6.

Akram, A., Fu, C., Li, Y., Javed, M. Y., Lin, R., Jiang, Y., & Tang, Y. (2019). Predicting students' academic procrastination in blended learning course using homework submission data. *Ieee Access*, 7, 102487-102498. https://doi.org/10.1109/ACCESS.2019.2930867

Arslan, F., ve Korkmaz, Ö. (2019). İlahiyat lisans tamamlama uzaktan eğitim öğrencilerinin etkileşim kaygıları ve uzaktan eğitime dönük tutumları. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, 1(1), 12-25.

Aruğaslan, E. ve Çivril, H. (2021). Türkiye'de eğitim alanında yapılan veri madenciliği ve yapay zeka çalışmaları. *Uluslararası Teknolojik Bilimler Dergisi*, 13(2), 81-89.

Aydemir, B. (2017). *Veri madenciliği yöntemleri kullanarak meslek yüksek okulu öğrencilerinin akademik başarı tahmini* [Yüksek lisans tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü].

Aydemir, E. (2019). Ders Geçme Notlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (15), 70-76. https://doi.org/10.31590/ejosat.518899

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 23(1), 537-553. 10.1007/s10639-017-9616-z

Baltacı, A. (2018). The data mining: Measurement of academic achievement in faculty of divinity students by data mining. *Din ve Bilim –Muş Alparslan Üniversitesi İslami İlimler Fakültesi Dergisi*, 1(1), 1-23.

Başer, S. H., Hökelekli, O. ve Kemal, A. D. E. M. (2020). Ortaöğretimde Öğrenim Gören Öğrenci Performanslarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi. *Bilgisayar Bilimleri ve Teknolojileri Dergisi*, 1(1), 22-27.

Bilen, Ö., Hotaman, D., Aşkın, Ö. E. ve Büyüklü, A. H. (2014). LYS başarılarına göre okul performanslarının eğitsel veri madenciliği teknikleriyle incelenmesi: 2011 İstanbul örneği. *Eğitim ve Bilim*, 39(172), 78-94.

Bilgin, M. (2018). *Veri biliminde makine öğrenmesi makine öğrenmesi teorisi ve algoritmaları* (2. Baskı). Papatya Bilim.

Bliuc, A. M., Ellis, R., Goodyear, P., & Piggott, L. (2010). Learning through face-to-face and online discussions: Associations between students' conceptions,approaches and academic performance in political science. *British Journal of Educational Technology*, 41(3), 512-524. https://doi.org/10.1111/j.1467-8535.2009.00966.x

Byeon, H. (2022). Developing a predictive model for depressive disorders using stacking ensemble and naive Bayesian nomogram: using samples representing South Korea. *Frontiers in Psychiatry*, 12. https://doi.org/10.3389/fpsyt.2021.773290

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, 9 (13), 1-73.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357. https://doi.org/10.48550/arXiv.1106.1813

Chawla, Nitesh V. (2005). *Data mining for imbalanced datasets: an overview*. O. Maimon ve L. Rokach (Ed.), Data mining and knowledge discovery handbook. 853-867. Boston. Springer.

Çiftçi, F., Kaleli, C. ve Serkan, Ü. (2018). Öznitelik seçme ve makine öğrenmesi yöntemleriyle eğitmen performansının tahmin edilmesi. *Anadolu Journal of Educational Sciences International*, 8(2), 419-440. https://doi.org/10.18039/ajesi.454587

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in human behavior*, 73, 247-256. https://doi.org/10.1016/j.chb.2017.01.047

Dağ, M. (2013). İlahiyat lisans tamamlama (İLİTAM) programlarında Kur'an dersi-müfredat, materyal hazırlama ve karşılaşılan sorunlar. *Ekev Akademi Dergisi*, 17 (55), 37-54

Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). 91-95. IEEE. https://doi.org/10.14569/IJACSA.2016.070531

Durairaj, M., & Vijitha, C. (2014). Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies*, 5(4), 5987-5991. https://doi.org/10.1016/j.matpr.2021.05.646

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access*, 5, 15991-16005. https://doi.org/10.1109/ACCESS.2017.2654247

Educational Data Mining, (2022, Ocak 5). International Educational Data Mining Society. (2021). educationaldatamining.org. https://educationaldatamining.org/

Ersöz, A. R. (2017). *Eğitsel veri madenciliği ile öğrenci profillerinin belirlenmesi*. (Yayınlanmamış Doktora Tezi). Bursa Uludağ Üniversitesi.

Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164-181. https://doi.org/10.1016/j.engappai.2010.09.007

Genç, M. F. ve Ayhan, M. (2021). İLİTAM Bölümü Öğrencilerinin Hadis Dersine Yönelik Tutumları. *Mizanü'l-Hak: İslami İlimler Dergisi*, (12), 77-109. https://doi.org/10.47502/mizan.933285

Gonçalves, A. F. D., Maciel, A. M. A., & Rodrigues, R. L. (2017). *Development of a data mining education framework for data visualization in distance learning environments*. In International Conference on Software Engineering and Knowledge Engineering. https://doi.org/0.18293/SEKE2017-130

Gümrükçüoğlu, S. ve Genç, M. F. (2020). İLİTAM Bölümü Öğrencilerinin İlâhiyat Eğitimine Bakışı Kocaeli Üniversitesi İlâhiyat Fakültesi İLİTAM Örneği. *İHYA Uluslararası İslam Araştırmaları Dergisi*, 6(2), 640-656.

Güre, Ö. B., Kayri, M. ve Erdoğan, F. (2020). PISA 2015 matematik okuryazarlığını etkileyen faktörlerin eğitsel veri madenciliği ile çözümlenmesi. *Eğitim ve Bilim*, 45(202), 393-415. https://doi.org/10.15390/EB.2020.8477

Hakyemez, T. C. (2015). *İlk Yil Öğrencilerinin Akademik Performansina Etki Eden Faktörlerin Araştirilmasi ve Bu Faktörlere Bağli Olarak*

Başarılarının Tahminine Yönelik Bir Karar Destek Sistemi Tasarım. [Yayınlanmamış Yükseklisans Tezi]. Sakarya Üniversitesi.

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

Hermaliani, E. H., Fanani, A. Z., Santoso, H. A., Affandy, A., Purwanto, P., Muljono, M., Syukur, A., Setiadi, D., & Rafrastara, F. A. (2022). Systematic Review of Educational Data Mining for Student Performance Prediction using Bibliometric Network Analysis (SeBriNA). *In 2022 International Seminar on Application for Technology of Information and Communication* (iSemantic), 463-468. IEEE. https://doi.org/10.1109/iSemantic55962.2022.9920477.

Howard, S. K., Ma, J., & Yang, J. (2016). Student rules: Exploring patterns of students' computer-efficacy and engagement with digital technologies in learning. *Computers ve Education*, 101, 29-42. https://doi.org/10.1016/j.compedu.2016.05.008

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), 1-14. https://doi.org/10.3390/sym12020213

Imran, M., Latif, S., Mehmood, D. & Shah, M. S. (2019). Student Academic Performance Prediction using Supervised Learning Techniques. *International Journal of Emerging Technologies in Learning*, 14(14), 92-104. https://doi.org/ 10.3991/ijet.v14i14.10310

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1), https://doi.org/61-72. 10.2478/cait-2013-0006

Kablan, S. (2020). Koronavirüs (Covid-19) pandemi sürecinde çevrimiçi yapılan Kuran-ı Kerim dersi sınav değerlendirilmesi: İstanbul Üniversitesi İlahiyat Fakültesi ILITAM sınavları örneği. *Atlas international congress on social sciences 7.*

Kamath, U., & Choppella, K. (2017). *Mastering Java Machine Learning: A Java developer's guide to implementing machine learning and big data architectures*. Packt Publishing.

Karateke, T. (2020). İLİTAM öğrencilerinin bu programı seçme nedenleri ve karşılaştıkları sorunlar: Fırat Üniversitesi örneği. *Değerler Eğitimi Dergisi,* 18(39), 235-262. https://doi.org/10.34234/ded.634501

Kassim, A. A., Kazi, S. A., & Ranganath, S. (2004). A web-based intelligent learning environment for digital systems. *International Journal of Engineering Education*, 20(1), 13-23. https://doi.org/10.1108/02640470610689250

Kay, J. (2000). Stereotypes, student models and scrutability. *In International Conference on Intelligent Tutoring Systems*, Berlin, 19-30. https://doi.org/10.1007/3-540-45108-0_5

Kaymakcan, R., Meydan, H., Telli, A. ve Cevherli, K. (2013). Paydaşlarına göre ilahiyat lisans tamamlama (İLİTAM) programının değerlendirilmesi. *Değerler Eğitimi Dergisi*, 11(26), 71-110.

Keskin, S., Aydin, F. ve Yurdugül, H. (2019). Eğitsel veri madenciliği ve öğrenme analitikleri bağlaminda e-öğrenme verilerinde aykiri gözlemlerin belirlenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, 9(1), 292-309. https://doi.org/10.17943/etku.475149

Khasanah, A. U. (2017). A comparative study to predict student's performance using educational data mining techniques. *In IOP Conference Series: Materials Science and Engineering*, 215 (2017). https://doi.org/10.1088/1757-899X/215/1/012036

Kismet, E. (2018). *Eğitsel veri madenciliğinde kullanılmak üzere experience api (XAPI) temelli öğrenme deneyimi kayıtlarının işlenebilmesi için bir model geliştirilmesi.* [Yayınlanmamış yüksek lisans tezi. Kocaeli Üniversitesi]. Ulusal Tez Merkezi.

Lantz, B. (2015). *Machine Learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R* (2. Baskı). Packt Publishing.

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt Publishing.

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. *International Journal of Computer Science and Network (IJCSN)*, 2(1). https://doi.org/10.48550/arXiv.1305.1707.

Maimon, O., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook* (2. Baskı). Springer.

McClean, S. I. (2003). *Data mining and knowledge discovery. R. A. Meyers (Ed.), Encyclopedia of Physical Science and Technology* (3. Baskı). New York. Academic Press.

Miller, L. D., Soh, L. K., Samal, A., Kupzyk, K., & Nugent, G. (2015). A Comparison of Educational Statistics and Data Mining Approaches to Identify Characteristics That Impact Online Learning. *Journal of Educational Data Mining*, 7(3), 117-150.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill Science.

Moradi, H., Moradi, S. A., & Kashani, L. (2014). *Students' performance prediction using multi-channel decision fusion*. Educational Data Mining: Applications and Trends. 151-174.

Morsy, S., & Karypis, G. (2017). Cumulative knowledge-based regression models for next-term grade prediction. *In Proceedings of the 2017 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, 552-560.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), 237. https://doi.org/10.3390/app11010237

Özbay, Ö. (2015). Veri madenciliği kavramı ve eğitimde veri madenciliği uygulamaları. *The Journal of International Educational Sciences,* 2(5), 262-262. https://doi.org/10.16991/INESJOURNAL.162

Özçınar, H. (2006). *KPSS sonuçlarının veri madenciliği yöntemleriyle tahmin edilmesi* [Yüksek lisans tezi, Pamukkale Üniversitesi]. Ulusal Tez Merkezi

Özdemir, Ş. (2016). *Eğitimde veri madenciliği ve öğrenci akademik başarı öngörüsüne ilişkin bir uygulama*. [Doktora tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü]. Ulusal Tez Merkezi.

Öztürk, A. (2018). Açık ve uzaktan öğrenme ortamlarında eğitsel veri madenciliği. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, 4(2), 10-13.

Pascual-Cid, V., Vigentini, L., & Quixal, M. (2010). Visualising virtual learning environments: Case studies of the Website exploration tool. *In 2010 14th*

International Conference Information Visualisation, IEEE. 149-155. https://doi.org/10.1109/IV.2010.31

Polat, A. (2021). *Açık öğretim liseleri öğrencilerinin okul terki ve mezuniyet durumlarının eğitsel veri madenciliği ile incelenmesi*. [Doktora tezi, Sakarya Üniversitesi]. Ulusal Tez Merkezi.

Polyzou, A., & Karypis, G. (2016). Grade prediction with models specific to students and courses. *International Journal of Data Science and Analytics*, 2(3), 159-171. https://doi.org/10.48550/arXiv.1906.00792

Prasetiyowati, M. I., Maulidevi, N. U., & Surendro, K. (2021). Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *Journal of Big Data*, 8(1), 84. https://doi.org/10.21203/rs.3.rs-132775/v1

Quinlan, J.R. (1986). *Induction of decision trees*. Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251

Rapidminer (2020). Rapidminer Documentation Weight by rule (RapidMiner Studio Core;Version 9.9). "Rapidminer, Weight by Rule 2020". Erişim 17 Ocak 2022. http://docsupcoming.rapidminer.com/9.2/studio/operators/modeling/feature_weights/weight_by_rule.html.

Rapidminer (2020). Rapidminer Rule Induction (2020) (RapidMiner Studio Core; Version 9.9). "Rapidminer, rule induction 2020". Erişim 17 Ocak 2022. http://docs-upcoming.rapidminer.com/9.4/studio/operators/modeling/predictive/rules/rule_induction.html?upcoming-rapidminer%5Bpage%5D=9.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). *Cross-validation*. Encyclopedia of database systems, 532-538. https://doi.org/10.1007/978-0-387-399409_565

Rodrigues, M. W., Isotani, S., & Zarate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. *Telematics and Informatics*, 35(6), 1701-1717. https://doi.org/10.1016/j.tele.2018.04.015

Rojanavasu, P. (2019). Educational data analytics using association rule mining and classification. *In 2019 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering*. 142-145. IEEE. https://doi.org/ 10.1109/ECTI-NCON.2019.8692274

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

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008, June). Data mining algorithms to classify students. *In Educational data mining* 2008.

Şengür, D. ve Tekin, A. (2013). Öğrencilerin mezuniyet notlarının veri madenciliği metotları ile tahmini. *Bilişim Teknolojileri Dergisi*, 6(3), 7-16.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422. https://doi.org/ 10.1016/j.procs.2015.12.157

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, 9(4), 1-5. https://doi.org/10.17485/ijst/2016/v9i4/87032

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

Sorour, S. E., Mine, T., Goda, K., & Hirokawa, S. (2014). Predicting students' grades based on free style comments data by artificial neural network. *In 2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, 1-9. IEEE. https://doi.org/ 10.1109/FIE.2014.7044399

Tekin, A. ve Öztekin, Z. (2018). Eğitsel veri madenciliği ile ilgili 2006-2016 yılları arasında yapılan çalışmaların incelenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, 8(2), 108-124. https://doi.org/10.17943/etku.351473

Tosunoğlu, E., YILMAZ, R., Özeren, E. ve Sağlam, Z. (2021). Eğitimde makine öğrenmesi: Araştırmalardaki güncel eğilimler üzerine inceleme. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, 3(2), 178-199. https://doi.org/10.3815/akef.2021.16

Wang, C. S., & Lin, S. L. (2012). Combining fuzzy AHP and association rule to evaluate the activity processes of e-learning system. *In 2012 Sixth International Conference on Genetic and Evolutionary Computing* (s. 566-570). IEEE.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* 1. 29-40.

Yukselturk, E., Ozekes, S., & Turel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-learning*, 17(1), 118-133. https://doi.org/10.2478/eurodl-2014-0008.

Zacharis, N. Z. (2016). Predicting student academic performance in blended learning using artificial neural networks. *International Journal of Artificial Intelligence and Applications*, 7(5), 17-29. https://doi.org/10.5121/IJAIA.2016.7502

PUJE, 61, 94-120 https://doi.org/10.9779/pauefd. 1085483

Pamukkale University Journal of Education
ISSN: 2757-8127

Educational Data Mining with Decision Tree and Rule Induction: A Case Study of SAU ILITAM

Deniz DEMİRCİOĞLU DİREN [1], Mehmet Barış HORZUM [2]

Abstract

This study aims to examine whether a student is a successful graduate or a dropout/failure according to the profile of students enrolled in a blended bachelor's degree completion program (ILITAM). The weights of the attributes were also determined to examine the effect of the attributes on the output. In the study, the CRISP-DM process method used within the scope of educational data mining was employed. The information gain method was used to reveal the feature weights. The study group consisted of the students who entered the Sakarya University ILITAM program between 2013-2016. The data set consisted of systematic records, the student's university entrance information and the target value, that is, the student's graduation success status from the university. As a result, the attribute that had the most impact on the target value was obtained as the gender of the student. Also, when a new student enrolled, the prediction of his or her graduation success rate based on his or her general information was made using the nearest neighbor algorithm with an accuracy rate of 91.30%. Thus, plans for students can be made with improvement suggestions. The relevant conclusions and suggestions regarding the findings of the research are discussed accordingly.

Article Details
Research Article

Received
10/03/2022
Accepted
24/01/2024
Published
15/05/2024

Key words
Educational data mining, Data preprocessing, Decision tree, Rule induction, ILITAM

1 Sakarya University, ORCID:0000-0003-3567-0779, ddemircioglu@sakarya.edu.tr
2 Sakarya University, ORCID:0000-0002-4280-0394, mhorzum@sakarya.edu.tr

Suggested Citation:
Demircioğlu Diren, D., & Horzum, M. B. (2024). Educational data mining with decision tree and rule induction: A case study of SAU ILITAM. *Pamukkale University Journal of Education [PUJE], 61,* 94-120. https://doi.org/10.9779/pauefd. 1085483.

Introduction

Data mining is a multidisciplinary approach that aims to understand, analyze, and transform data into usable information. In other words, it can be defined as the process of extracting information and conclusions from large databases. According to the patterns sought, the processes in data mining can be classified as summarization, classification, clustering, association, and trend analysis. The classification method allows determining the class of an object according to its characteristics. With the summarization method, data is abstracted and generalized. Association rules enable revealing relationships between attributes. In the clustering method, the aim is to group and identify a series of objects of unknown class (Fu, 2011).

Educational data mining is defined by the international educational data mining community as "a discipline that deals with developing methods to make inferences from data in educational environments and understanding students and learning environments with the developed methods" (Educational Data Mining, 2021). Web-based courses, learning content management systems, and adaptive intelligent web-based education systems; each has different data sources and goals for knowledge discovery (Romero & Ventura, 2007). In recent years, there have been many studies on educational data mining with very different goals and objectives. Researchers classify the studies carried out in this field according to their subjects (Bakhshinategh et al., 2018; Dutt et al., 2017; Rodrigues et al., 2018). When the studies are examined, it is seen that the most basic areas of study are data analysis and visualization (Gonçalves et al., 2017; Pascual-Cid et al., 2010), student performance prediction (Aghalarova & Keser, 2021; Miller et al., 2015; 2015; Moradi et al., 2014), providing feedback for instructors (Wang & Lin, 2012) and student modeling (Kassim et al., Kay, 2000; 2004; Moradi et al., 2014).

In studies conducted on educational data mining in Turkey, the postgraduate thesis can be found on subjects such as data analysis (Özçınar, 2006), student modeling (Akçapınar, 2014; Ersöz 2017; Kismet, 2018), and determining student performance (Aydemir 2017; Özdemir, 2016; Polat, 2021). In addition, there are academic studies such as performance prediction (Baltaci, 2018; Başer et al., 2020; Bilen et al., 2014; Şengür & Tekin, 2014), data analysis (Aydemir, 2019; Keskin et al., 2019), providing feedback for supporting training (Güre et al., 2020), as well as general explanations and literature review on these subjects (Aruğaslan & Çivril, 2021; Özbay, 2015; Öztürk, 2018; Tekin & Özteki, 2018; Tosunoğlu et al., 2021).

This study aims to provide feedback for instructors based on student performance estimates obtained as a result of data analysis. In this sense, systematic literature reviews in the literature on student

performance have become directive (Abu Saa et al., 2019; Hermaliani et al., 2022; Namoun et al., 2020). There are a number of studies on the subject of predicting student performance in the fields of face-to-face learning (Bliuc et al., 2010; Morsy & Karypis, 2017; Polyzou & Karypis, 2016), distance learning (Bliuc et al., 2010; Gonçalves et al., 2017; Howard et al., 2016) and blended learning (Sorour et al., 2014; Zacharis, 2016). The amount of data obtained in distance education is considerably higher than in face-to-face education. By processing, analyzing, and modeling these data, results that will guide researchers can be achieved. The application area of distance education is quite wide, so it has the potential to have a widespread impact due to the models developed. In this research, the data of the Theology Undergraduate Completion Program (ILITAM) students, one of the blended learning programs of Sakarya University (SAU), were selected as the study group. The ILITAM program first started to be carried out at Ankara University as blended learning in October 2005 and was subsequently opened at different universities in Turkey. This program is based on associate degree program graduates completing a two-year education program in order to obtain a bachelor's degree. Research on the ILITAM program in the literature can be divided into categories as follows;

1. Studies to determine the perceptions of faculty members and students on blended or distance education programs (Arslan & Korkmaz, 2019; Genç & Ayhan, 2021; Gümrükçüoğlu & Genç, 2020; Kablan, 2020; Karateke, 2020; Kaymakcan et al., 2013)
2. Studies to determine the effectiveness of the education program (Imran et al., 2019)
3. Studies on the suitability of the materials used in the curriculum and the teaching style of the course with the distance education model, the difficulties or problems experienced (Akaslan, 2020; Dağ, 2013)

It has been observed that in the above-mentioned studies on ILITAM, descriptive and case studies were generally carried out using qualitative and quantitative methods. It has been observed that statistical analysis is mostly used in this regard, while data mining and machine learning applications have not been explored. This study is up-to-date in that it deals with data from ILITAM and blended learning students. In addition, in the relevant literature upon educational data mining, it was seen that studies generally focused on the prediction accuracy of algorithms and clustering. However, problems such as data dimensionality, class imbalance, and classification error were not taken into account during these processes (Imran et al., 2019). However, while in some studies, the default values of the program are used to determine the parameters of the machine learning algorithms in the models, in some studies, these values are determined heuristic by the

researcher. In this study, unlike other studies, attributes were categorized, and the SMOTE method (Chawla et al., 2002) was used to prevent class imbalance. Additionally, data cleaning and preprocessing steps were applied, and hyperparameter (grid search) optimization was used for parameter selection of the algorithms. In addition to parameter optimization, the study aims to correctly select the appropriate machine learning algorithms to be used in the student graduation success prediction model and to determine the order of importance of the attributes affecting success. The prediction model has two classes. The first is the success class, which represents students who successfully graduate from the program. The second is the dropout/failure class, which represents students who left the program as failures and dropped out from the program. In this respect, the research was conducted to find answers to the following questions:

1.What is the order of importance of the attributes that affect success or dropout/failure in the ILITAM program?

2. What is the accuracy of predicting the success or dropout/failure of a student newly enrolled in the ILITAM program using educational data mining techniques?

3. What is the contribution of data preprocessing steps to prediction accuracy?

4. What are the relationships and rules between success or dropout/failure in the ILITAM program and the attributes that affect this situation?

Method

In the research, the CRISP-DM process model, which is used within the scope of educational data mining (Chapman et al., 2000), was employed. The aim of the study is to predict the success or dropout/failure of students in the SAU ILITAM blended learning program according to their university entrance information and to identify the attributes that most affect this predicted situation. In addition, the success or dropout/failure status of a student newly enrolled in the ILITAM program was predicted, and the relationships between the result status and input attributes were identified. The student's success or failure status was determined by basic machine learning algorithms, and the relationships between input attributes and the outcome situation were determined by the feature weights method.

The CRISP-DM process model used in the study emerged in a project in 1999. The main purpose of the method is to make data mining projects less costly, more reliable, more repeatable, more manageable, and produce faster results. The six steps covered by the process are presented in Table 1 (Wirth & Hipp, 2000).

Table 1
Steps of the CRISP-DM Model

Description of the Problem	Data Understanding	Data Preprocessing and Preparation	Modelling	Evaluation	Deployment
Identifying the objectives of the problem	Validate data quality	Selecting data	Choosing the modeling technique	Evaluating the model	Deployment plan
Evaluate the problem	Identify data	Clearing data	Improving the test design	Review the process	Plan monitoring and maintenance
Determining data mining goals	Exploring data	Creating the data	Build a model	Determine Next Steps	Producing final report
Producing a project plan	Collect the data	Integrate and format data	Evaluate the model		Review the project

Description of the Problem

In distance education, there appears to be less chance to observe students compared to face-to-face education. Therefore, examining the characteristics and behaviors of distance education students is useful for learning-teaching processes. The problem of determining the most important attributes affecting the success of students in distance education and determining the success or dropout/failure of students in the future, which is the focus of the study, is an important problem.

Data Understanding

The data set of the study consists of data from students who entered the Sakarya University İLİTAM program between 2013 and 2016. The data set was obtained by combining data from two databases. Student profile information and OSYM data are from the student affairs database (DB-1). Success statuses are obtained from the SAU Information System (DB-2). The information in the two databases will be analyzed by matching and integrating according to student numbers. Ethical approval for the study was carried out within the framework of the permission obtained by the Sakarya University Ethics Committee, decision no. 38, no. 57, dated 01/10/2021. The input attributes of the students and the explanations of these attributes are presented in Table 2. The purpose of the study is to estimate the target/output attribute that expresses the outcome value. This target value is called pass and fail. The successful value includes information about students who successfully graduated from the program, and the dropout/failure value includes information about students who left the program failure and were dropout from the program.

Table 2
The first data set obtained without data preprocessing

Attributes Name	Attributes Definition	Database
Student number	Student's Number	DB-1, DB -2
Birth date	Student's Year of Birth	DB -1
Gender	Student's gender	DB -1
Graduation Year	The year he graduated from high school	DB -1
PlacementScore	OSYM score	DB -1
Order of Preference	Department Preference Order	DB -1
Education year	It is the academic year in which the student is placed.	DB -1
EnumOsymPlacement	Placement type	DB -1
Archive Information	Success status	DB -2

Data Preprocessing and Preparation

Before analyzing the data set, in order to reach correct results, predictions, and relationships, the problem must first be understood well, and the data must be prepared properly. For this reason, data preprocessing was applied. Data preprocessing includes data integration, cleaning, transformation, and reduction steps. In the study, the data preprocessing process was carried out through the Rapidminer 9.10.011 program.

First, the data sets in two separate databases were combined by applying the integration step. The input values of the data set and the student achievement status, that is, the target value, are in different databases. While the data in the DB-1 data set includes general information about students' entries, the data in the DB-2 data set includes their graduation status. The DB-1 data set contains data of 3472 students, and the DB-2 data set contains data of 2189 students. The merging of the data sets was carried out based on the DB-2 data set. Here, student numbers are used as the key field for two data sets. As a result of merging, an integrated data set of 2189 students was obtained. In the second step of data preprocessing, missing, erroneous, repetitive, and outlier values in the data set were examined. In this step, according to the student number information, nine student data were removed from the data set, which was determined to be repetitive. It was found that there were missing records in the information, such as birth year, gender, placement score, department preference order, quota, and academic year of 165 students, and these were also removed from the data set. Also, the quota information of 782 students and the graduation information of 12 students were recorded incorrectly, which cannot be interpreted, so they were removed from the data set. In the target value consisting of archive statuses, there is information that a

person left the program because he died. For this reason, the data of this student was also removed from the data set. The data set obtained at the end of the entire data cleaning process includes 1220 student data. In addition, the attributes were analyzed with the box plot method, and outliers were searched for in the data set, and no outliers were found.

Since some attributes in the data set were not deemed suitable for evaluation and interpretation, a new attribute creation process was applied through the transformation process. Within the scope of this process, the university entrance age was calculated by taking the difference between the academic year and the year of birth, and a new attribute called age was created. In order to examine the effect of how many years after students graduated from high school on their success, the difference between the university academic year and the high school graduation year was taken, and a new attribute called the placement year was obtained. Within the scope of attribute transformation, z, and t transformations were made on the entry score, and the scores were considered as a common value. The reason for this is that when evaluating the placement score for the department, the scores may be in different ranges since the base and top scores are different every year. Another transformation process was carried out on the student number. The last three digits in the number coding represent the student's entry ranking into the department. Therefore, the entry order attribute was transformed using the student number. The 'order of preference' attribute was categorized to make it more meaningful. The transformation processes performed are presented in Table 3.

Table 3

Data Transformation Processes

Current Attribute	Transformation Process	New Attribute
Academic year - Birth date	Attribute Generation	Age
Education year - Graduation Year	Attribute Generation	Year of Starting University
EnumOsymPlacement	Normalization	Score
Student number	Attribute Generation	Entry Order
Order of Preference	Discretization	Order

The attributes obtained for 1220 student data as a result of data preprocessing are presented in Table 4.

Table 4
Data Set-Features

Attributes	Description	Type	Value
Age	Student's university entry age	Continuous	19-57
Gender	Student Gender	Nominal	Female-Male
Year of Starting University	Year to enter the program after high school	Continuous	0-22
Score	Placement Score	Continuous	34,73–107,93
Order of Preference	Order of choosing the program you are placed in	Nominal	1-3, 4-6, 7+
EnumOsymPlacement	First placement additional placement	Nominal	1,2
Entry Order	Student's entry ranking to the department	Continuous	1-599
SuccessStatus	Graduation or non-graduation status	Nominal	Success, Non-Success

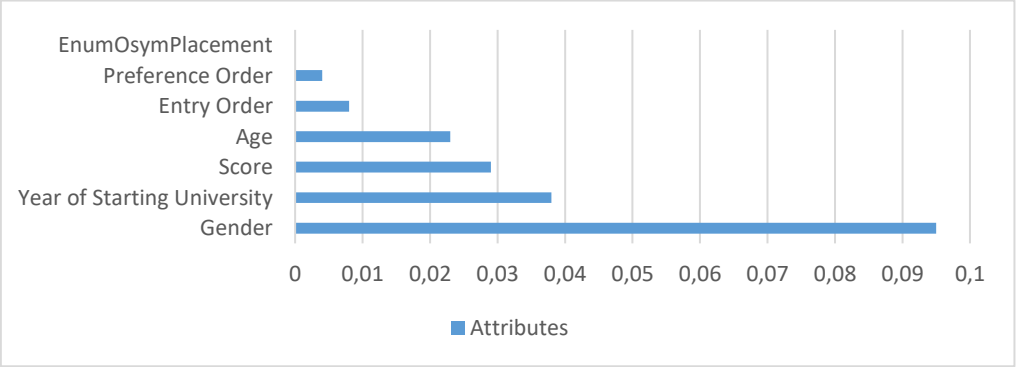
Another stage in the data preparation process is to examine the balance status of the classes in the data set. The classes in the data set must contain approximately equal numbers of data; otherwise, the data set is considered unbalanced. This imbalance causes the class with a large number of data to be predicted with high accuracy, while the minority class is predicted with low accuracy. In this case, commenting with general accuracy may be misleading. As a solution to class imbalance, classes that are close to each other can be combined depending on suitability, penalty points can be applied to the class with a large number of data, or resampling can be done with synthetic data generation (Chawla, 2005; Longadge & Dongre, 2013; Romero et al., 2008). A commonly used method for resampling is Synthetic Minority Oversampling (SMOTE). The SMOTE method is based on the principle of increasing the number of data in the minority class by calculating linear combinations of two similar samples (Chawla et al., 2002). By performing basic label analysis on the data set obtained as a result of the data transformation process, it was seen that there were 1138 examples from the successful category and 82 examples from the failure category in the target attribute. To resolve this imbalance, the SMOTE method was applied to the data set, and the number of samples in the failure category was equalized to the number of samples in the successful category. As a result, 2276 samples were obtained.

Attribute weights

Information retrieval is a popular filter model and technique used to determine the weight of features (Prasetiyowati et al., 2021; Sokkhey &

Okazaki, 2020). This method, which allows us to see how much influence each attribute has on the target value, is also used in the field of education (Çifçi et al., 2018). In the study, the weights of the attributes affecting student success were determined by the information acquisition method. In addition, the application of feature weights and all other analyses was carried out with the Rapidminer Studio 9.10.011 program. When the weights of the attributes are examined, it is seen that the attribute that has the most impact and is the most decisive on the result attribute is the "Gender," with a value of 0.095. Other attributes are listed as follows: "Year of starting university" with a value of 0.038, the "Score" with a value of 0.029, the "Age" with a value of 0.023, the "EntryOrder" with a value of 0.008 and the "PreferenceOrder" with a value of 0.004. It is seen that the attribute that has the least impact on the result is the "EnumOsymPlacement", which indicates whether the students are placed as primary or reserve students. The ranking of the attributes is presented in Figure 1.

Figure 1
Feature weights



Modelling

The k-fold cross-validation method was used as the learning method in the study. In this method, one part of the data is used for testing, and the other parts are used for training. The working strategy of the method starts by using the first part as testing and the other parts as training data, and then the model continue to work by using the second part as test and the others as training data. The process is completed when all parts are evaluated as test data. Accuracies are obtained at the end of each evaluation. The overall accuracy is obtained by averaging all these (Refaeilzadeh et al., 2009). The prediction models developed in the study were trained with k-fold cross-validation, and the parameters used in the algorithms were determined with the grid optimization technique. The optimization process and training models are seen in Figure 2 (a) and Figure 2 (b).

Figure 2 (a)
Optimization process

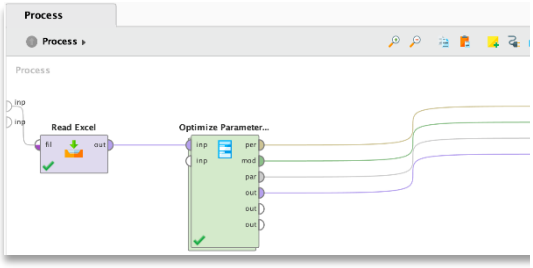
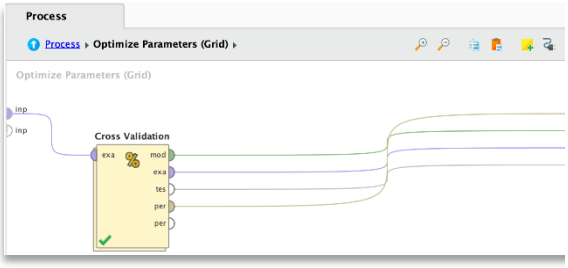


Figure 2(b)
k-fold cross validation

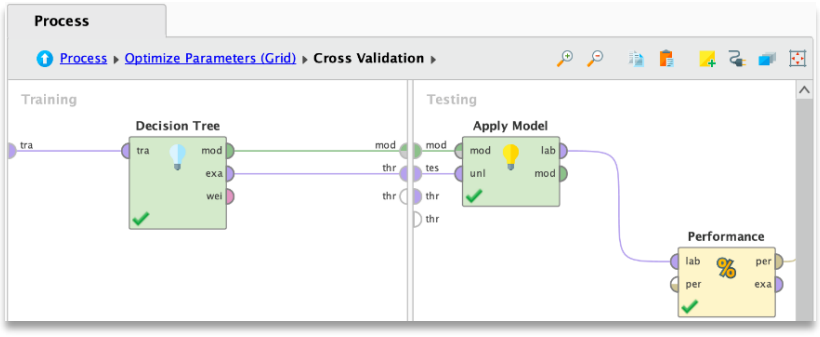


To predict student performance, basic machine learning algorithms such as decision tree, k-nearest neighbor, naive Bayes algorithms, and rule induction were used to extract relationship rules.

Decision Tree (DT)

DT can work with both nominal and numerical attributes and is a more widely used method than other algorithms. As a working principle, the algorithm proceeds according to the tree structure by dividing the data set into subsections. The process starts from the root and continues from the intermediate node and branches up to the leaf node, and the leaves form classes. The algorithm with IF-THEN rules is easy to interpret and understand (Bilgin, 2018; Mitchell, 1997). There are some criteria, such as information gain, gini index, and gain rate, to determine which attribute will be the root node. This criterion can be chosen according to the operating performance of the algorithm (Maimon & Rokach, 2005). In the study, a classical decision tree was used, as seen in Figure 3. This decision tree learner in RapidMiner works similarly to Quinlan's C4.5 or CART (Quinlan, 1986).

Figure 3
Classic decision tree model



Parameter results obtained using grid optimization are presented in Table 5. The k value of the k-fold cross-validation method was examined with the optimization technique, varying between 1-10, and the sampling type was changed into four different sampling types. The root node and depth parameters of the decision tree were determined

simultaneously with the optimization technique as a result of 144 iterations.

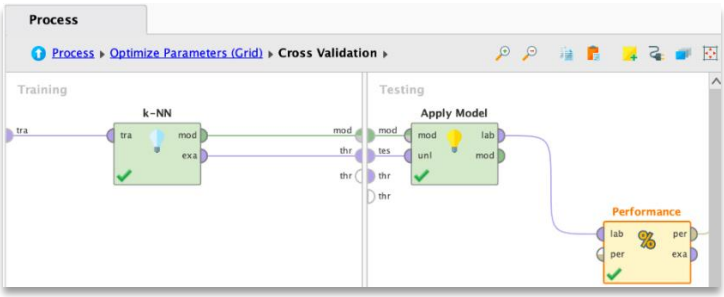
Table 5
DT parameters

Parameter	Value
k-fold cross validation	10
k-fold cross-validation sampling type	Random
Decision tree root node criterion	Gini_index

K-nearest Neighbor (K-NN)

In this algorithm, whose working principle is quite easy, a selection is made according to the number of neighbors determined by the user, and the new sample is assigned according to the similarities to those neighbors. Although it has an easy working principle, the algorithm is quite effective (Han et al., 2011). However, the classification stage progresses a little slowly, and additional processes are required when there is missing data (Lantz, 2019). The nearest neighbor is determined by measuring the similarity between two samples. This distance can be calculated using Euclidean, Manhattan, and Minkowski distances (Bilgin, 2018). The K-NN prediction model applied in the study is presented in Figure 4.

Figure 4
K-NN model



The parameter values used for training the K-NN model are presented in Table 6. The number k, which represents the number of nearest neighbors, and similarity measures were obtained by optimization as a result of 484 iterations.

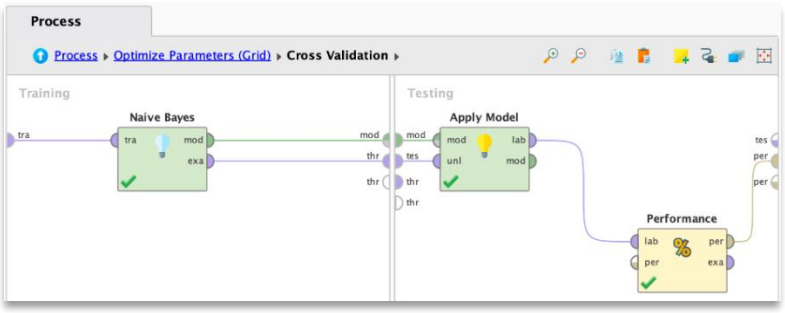
Table 6
K-NN parameters

Parameter	Value
k-fold cross validation k number	71
k-fold cross-validation sampling type	Random
K-NN k number	1
Measure type	Mixed Measurement
Mixed measure	Mixed Euclidean Distance

Naive Bayes (NB)

Bayesian approach is one of the most practical methods among learning algorithms. This method calculates the probability of a hypothesis based on the observed data with the probability of the previous assumption and the probability of the data appearing (Mitchell, 1997). Based on the Bayesian approach, NB examines the frequency of occurrence of each outcome and the number of times the combination of independent attributes and dependent attributes is seen (Bilgin, 2018). In this method, it is evaluated how well an asset fits into a particular class based on the odd value. Odds value is the ratio of the probability that an entity belongs to a target class to the probability that it does not belong (Byeon, 2022). The NB prediction model used in the study is presented in Figure 5.

Figure 5
NB model



For the NB algorithm, the parameter values obtained through optimization as a result of 88 iterations are presented in Table 7. The only parameter used in this algorithm is the Laplace correlation. Parameter selection related to k-fold cross-validation was made in the model.

Table 7
Accuracy values for NB by sampling type

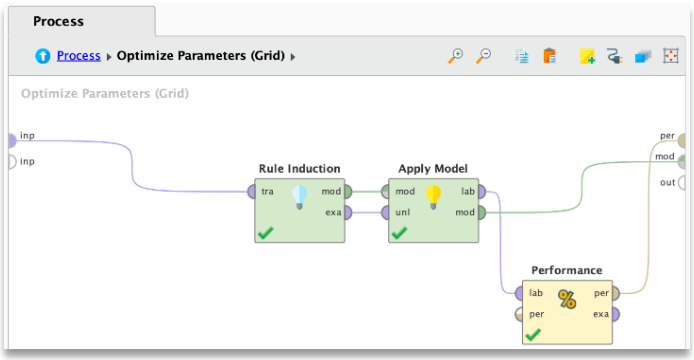
Parameter	Value
k-fold cross validation k number	2
k-fold cross-validation sampling type	Random
Laplace correlation	No

Rule Induction

In the rule induction method, the algorithm grows iteratively, and the process starts from less common classes. The rule pruning process continues until there are no positive examples left or the error rate is more than 50%. During the growth process, for each rule, the condition with the highest information gain is selected by trying the possible values of each feature. As a result, the values in the data are transformed into meaningful information with the rules obtained. It can represent the model scientifically as well as partially explain it. Although it is

similar to the DT algorithm, decision rules are easier to understand and interpret than decision trees. However, it can be stated as a disadvantage that it works partially poorly on large training data (Rapidminer, Rule Induction, 2020). The rule induction model in the study is presented in Figure 6.

Figure 6
Rule Induction



Parameters obtained by the optimization method for rule induction are presented in Table 8.

Table 8
Rule induction parameter values

Parameter	Value
Rule induction criterion	Accuracy
Sample rate	1

Evaluation

Data preprocessing, parameter selection, and test set determination are factors that affect the success of the model in data mining applications. Measurement of how accurate and acceptable these processes are carried out is made in the evaluation step. In order to measure prediction success in classification applications, evaluations regarding accuracy criteria should be made. These are accuracy rate, sensitivity, precision, F criterion, and kappa. One or more of these criteria can be used depending on the need and application requirements. In this study, algorithms were evaluated according to accuracy, sensitivity, precision, and kappa statistics criteria. The confusion matrix from which the accuracies were obtained is given in Table 9. In the matrix, columns represent actual values, and rows represent predicted values. (Bilgin, 2018).

Table 9
Confusion matrix

	Actual Successful Class	Actual Failure Class
Prediction Success Class	G1	Y0
Prediction Failure Class	Y1	G0

Here;

G1: The number of successfully classified samples in the test set,
G0: number of correctly classified failed samples in the test set,
Y1: number of misclassified successful examples in the test set,
Y0: number of misclassified failed samples in the test set

The results obtained according to the analyses are presented in Table 10.

Table 10
Algorithm results

Algorithm	Accuracy	Recall	Precision	Kappa
DT	%87,04	%87,04	%87,59	0,740
NB	%64,50	%64,49	%67,62	0,290
K-NN	%91,30	%91,46	%92,11	0,824

Success status was estimated by DT, NB, and K-NN algorithms. According to Table 10, it can be seen that the most successful prediction values are achieved with the K-NN algorithm. Performance values of the K-NN algorithm It was obtained as 91.30% accuracy, 91.46% sensitivity, 92.11% precision, and 0.824 kappa statistics. The K-NN algorithm is followed by the DT algorithm with 87.04% accuracy, 87.04% sensitivity, 87.59% precision, and 0.740 kappa statistic value. It is seen that the algorithm with the lowest predictive values is the NB algorithm, with 64.50% accuracy, 64.49% sensitivity, 67.62% precision, and 0.29 kappa statistic values. As a result, it can be said that the most suitable classification algorithm for the prediction model in the study is K-NN.

Deployment

The aim of this stage is to guide instructors and administrators by developing policies according to the relationship rules obtained as a result of the application. In addition, it is expected that the policies developed based on the data will be integrated with the education and training systems, improving the processes and achieving more successful societies.

Findings

There are four different research questions within the scope of the study. These;

a) Research question 1: What is the order of importance of the attributes that affect success or failure/dropout in the ILITAM program?

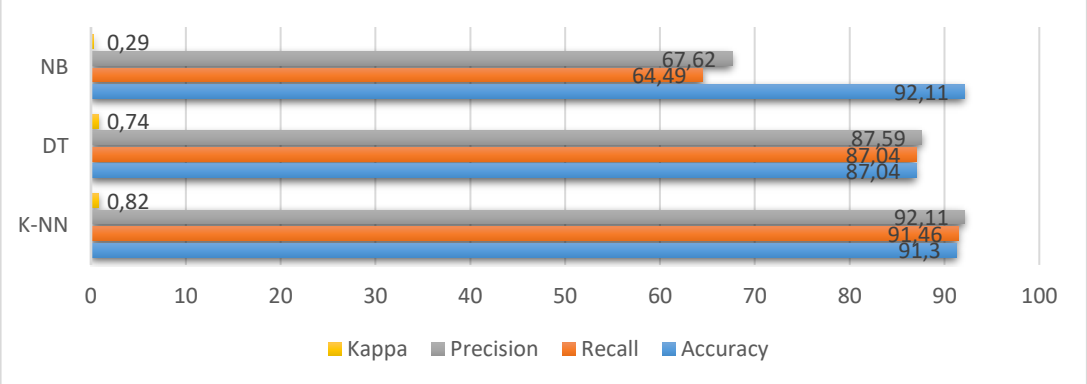
The impact weights of the features were calculated with the information gain algorithm. As a result, the most effective and decisive attribute was obtained as the "Gender". The second attribute is the " YearofStartingUniversity " with a value of 0.038. Then, the affecting attributes were obtained as "Score", "Age", "EntryOrder" and "OrderofPreference". The attribute that has the least impact on the result is the "EnumOsymPlacement", which indicates whether the students are placed as primary or reserve students.

b) Research question 2: What is the accuracy of predicting the success or dropout/failure of a student newly enrolled in the ILITAM program with educational data mining techniques?

One of the most important criteria of machine learning algorithms, besides cost and solution time, which are the success criteria, is solution and prediction accuracy. The prediction accuracy criterion depends on the performance of the algorithms, the suitability of the data, and the parameters used. The most decisive among these factors is the selection of data and parameters. Parameter selection can be done intuitively. However, this may cause possible high prediction accuracies to be ignored. In order to prevent this problem, parameter optimization techniques were used in the study. In addition, DT, K-NN, and NB algorithms, which are among the most basic algorithms for prediction accuracy and are frequently used in studies (Akram et al., 2019; Durairaj & Vijitha, 2014; Rojanavasu, 2019), were used. The prediction success of the algorithms was evaluated according to accuracy, precision, sensitivity, and kappa performance criteria.

According to the results presented in Figure 6, the classification algorithm that achieves the most successful prediction values among the three algorithms is the K-NN algorithm. Then comes the DT algorithm, and the algorithm with the lowest success rate is seen as the NB algorithm. It can be said that the classification model suitable for the problem is the model developed with the K-NN algorithm.

Figure 6
Prediction performance results



c) Research question 3: What is the contribution of data preprocessing steps to prediction accuracy?

The data set contained values of the information of 1289 students before the data preprocessing steps. After applying the data preprocessing steps, the data set contained values for 1220 students. Then, 2276 samples were obtained with the SMOTE method applied to eliminate data imbalance. Findings regarding the results obtained are listed below.

The prediction accuracies of the algorithms before and after data preprocessing are presented separately according to the performance criteria in Table 11. When the accuracy values are examined, it is seen that the prediction performances for K-NN and DT algorithms are low before data preprocessing. It can be said that data preprocessing steps positively affect prediction performances. Unlike these two algorithms, the prediction accuracy of the NB algorithm appears to decrease after data preprocessing.

Table 11

Prediction performance values before and after data preprocessing

	Accuracy Rate	
Algorithm	Before data preprocessing	After data preprocessing
DT	%78,49	%87,04
NB	%73,74	%64,50
K-NN	%76,91	%91,30

d) Research question 4: What are the relationships and rules between the success or dropout/failure of the ILITAM program and the attributes affecting this situation?

Inferences were made from rule induction algorithms in order to determine the relationships between the qualifications of the students studying in the ILITAM program. In these rule-generating algorithms, when the data set is large, a large number of rules are obtained. While some of the rules may simply correspond to well-known domain knowledge, some may not be important to the practitioner. For this reason, rules that are considered important and appropriate to the problem by practitioners and experts should be chosen (McClean, 2003). In other words, rules can scientifically represent the model or partially explain it. In this study, 23 rules were obtained and evaluated in terms of meaningfulness, usability, and usefulness. As a result of the evaluation, expert opinion was taken, and the following ten basic findings were accepted as usable.

- Female students cannot graduate from the department if their placement preference order is ranked second or third.

- Even if the entry order is at the end of the department, students whose placement year is less than half a semester graduate successfully.
- Male students with average or low placement score values cannot graduate.
- Even though the entry order is at the end of the department, all students under the age of 28 graduate if their score is at an average level.
- Female students whose entrance order is at the bottom or top of their department rankings cannot graduate from the program.

Discussion and Conclusion

This study was carried out to identify the attributes that affect success in ILITAM blended learning programs and to develop educational policies that will guide students. During the application phase, data from students who entered the SAU ILITAM program between 2013 and 2016 were used. The data set was created by combining two data sets. The first data set is the input data set containing students' information. The second data set is the output data set containing the success status of the students. Values belonging to the same student were matched in the two data sets. Educational data mining was used to determine attributes and make predictions about the process for policies. For this, first, the data set was prepared for analysis with preprocessing steps. After repeated missing value cleaning, attribute generation, and data augmentation processes, a data set containing values of 2276 students was obtained. In addition, more meaningful attributes were obtained by transforming the attributes with the data transformation process.

While there were nine attributes in the first data set, after data transformation, a more balanced data set containing eight attributes was obtained. DT, K-NN, and NB models were trained with the obtained data set. As a result, the prediction accuracy of the models increased, consistent with data preprocessing studies in the literature (Costa et al., 2017; Romero et al., 2008). When examined according to accuracy, precision, sensitivity, and kappa performance criteria, it was concluded that KNN was the most appropriate algorithm for the classification problem in the study. Similarly, in the study carried out by Yükseltürk et al. (2014) to predict the students who will drop out of school, the K-NN algorithm was stated to be successful. On the other hand, there are studies in the literature where different algorithms are successful (Devasia et al., 2016; Kabakchieva, 2013; Khasanah, 2017). This is because the quality of data mining analysis and interpretation largely depends on the quality of the input data (Lantz, 2015). For example, properly chosen attributes tend to outperform other models no matter how complex the algorithm. In general, when working with the right set of

attributes, even a simple model performs well and gives the desired results (Kamath & Choppella, 2017).

Another result obtained from the study is the importance ranking of the attributes. Sivakumar et al. (2016) examined the reasons for students leaving the program and discussed them mainly in terms of social problems such as students' family and personal problems. The most important factor in leaving the program was determined to be family reasons. In this study, students' dropout situations were discussed from an academic perspective, and it was aimed to suggest educational policies regarding them. It was observed that the most effective attribute of the result was the gender of the students. It is seen that some of the male students tend to drop out of the program regardless of any other attribute, while female students drop out of the program depending on their entrance score and university placement time. In addition, other findings of rule induction support the feature importance ranking. These findings indicate that while female students cannot graduate from the department if their preference order is second or third, male students will be successful even if their preference order for placement in the program is third. As in the study of Hakyemez (2015), it can be said that the 'Gender' has a significant effect on academic success. It seems that the second attribute that determines the failure of students is the year of placement. According to the rules examined, students whose placement year is less than half a semester graduate successfully, even if the entry order is at the end. This may be related to the uninterrupted maintenance of the student's motivation to study. Another result obtained is that male students with average or low score values cannot graduate. It can be said that the reason for this is that male students, who have low motivation to study, want to leave education and enter business life. In addition, according to the results, it is seen that all students under the age of 28 graduate if their score value is at the average level, even though the entry order is at the end. In addition, female students whose entry order are at the bottom or at the top cannot graduate from the program. The reason for this may be that students at the bottom of the rankings experience anxiety about failure, while students at the top want to be successful and move to departments with higher scores. Shahiri and Husain (2015) stated that while trying to determine student performance, the most influential attributes besides the general score attribute are 'Gender' and 'Age'. Similar to the results obtained in this study, the importance of age and gender attributes is revealed.

When the prediction success rates and rules obtained are examined, it is seen that the use of these techniques in this field will be efficient. As a result of the literature review, although there have been previous studies using data mining techniques in blended learning, no studies in this field were found in the ILITAM program. Thus, it can be said that

the study is a guide in the field. Due to wrong strategies in education policies, problems may occur in student success and graduate quality. Analyzing the patterns and relationships in the data and developing policies based on the results obtained will help students graduate better equipped and prevent them from leaving the department they study. It can be said that the study is a study that can give ideas to ILITAM department coordinators and faculty members on issues such as approaches to students and issues to be considered regarding course management during the education process.

Despite the successful and guiding results of the study, there are also some limitations. There are ILITAM programs in 11 universities in Turkey, but due to the problem of accessing the data, the data collection activity in the study was carried out only in the SAU ILITAM program. In addition, the data was collected within the framework of students enrolled between 2013 and 2016, when it was possible to access the records systematically. Although there are many algorithms that can be used, DT, K-NN and NB algorithms, which are among the basic algorithms, were used for prediction and rule extraction in the study. This can be stated as another limitation of the study. It can be said that another limitation is the characteristics of the collected data set. In addition to the qualities used, it is thought that it would be useful to examine the performances of the models trained by adding qualities that express the social lives of the students. In addition, it is planned to use the data mining CRISP-EDM (Özdemir, 2016) approach, which has been recommended for the field of education in the literature, in future studies.

Suggestions

Some suggestions can be offered according to the results obtained from the study. For example, by paying attention to the preference order and entry order attributes of the students in the courses they conduct, faculty members can use different materials or applications to increase students' participation and motivation in the course and suggest new resources. Additionally, orientation training can be planned to motivate students more, and activities can be carried out to raise awareness.

Ethics Committee Approval: *This research was conducted with the permission of the SAKARYA UNIVERSITY RECTORATE Ethics Committee, decision numbered 38, numbered 57, dated 01/10/2021.*

Conflict of Interest: *The authors have no conflict of interest to declare.*

Author Contribution: *The authors contributed equally to this article.*

References

Abu Saa, A., Al-Emran, M. & Shaalan, K. (2019). Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24(4), 567-598. https://doi.org/10.1007/s10758-019-09408-7

Aghalarova, S. & Keser, S. B. (2021). Önerilen Yapay Sinir Ağı Algoritması ile Ortaokul Öğrencilerin Akademik Performansının Tahmini. *Veri Bilimi*, 4(2), 19-32.

Akaslan, Y. (2020). Mahiyet, nitelik ve müfredat açısından ilitam programlarında Kur'an-ı kerim dersleri (Ondokuz Mayıs Üniversitesi Örneği). *Ondokuz Mayıs Üniversitesi İlahiyat Fakültesi Dergisi*, 49, 9-37. https://doi.org/10.17120/omuifd.779343

Akçapınar, G. (2014). *Çevrimiçi öğrenme ortamındaki etkileşim verilerine göre öğrencilerin akademik performanslarının veri madenciliği yaklaşımı ile modellenmesi*. [Yayınlanmamış Doktora Tezi. Hacettepe Üniversitesi]. Ulusal Tez Merkezi.

Akgün, E. (2019). *2023 Eğitim vizyonunda eğitsel veri madenciliği*. Seta Perspektif, 228,1-6.

Akram, A., Fu, C., Li, Y., Javed, M. Y., Lin, R., Jiang, Y. & Tang, Y. (2019). Predicting students' academic procrastination in blended learning course using homework submission data. *Ieee Access*, 7, 102487-102498. https://doi.org/10.1109/ACCESS.2019.2930867

Arslan, F., & Korkmaz, Ö. (2019). İlahiyat lisans tamamlama uzaktan eğitim öğrencilerinin etkileşim kaygıları ve uzaktan eğitime dönük tutumları. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, 1(1), 12-25.

Aruğaslan, E. & Çivril, H. (2021). Türkiye'de eğitim alanında yapılan veri madenciliği ve yapay zeka çalışmaları. *Uluslararası Teknolojik Bilimler Dergisi, 13(2)*, 81-89.

Aydemir, B. (2017). *Veri madenciliği yöntemleri kullanarak meslek yüksek okulu öğrencilerinin akademik başarı tahmini* [Yüksek lisans tezi, Pamukkale Üniversitesi]. Ulusal Tez Merkezi.

Aydemir, E. (2019). Ders Geçme Notlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi. *Avrupa Bilim ve Teknoloji Dergis*i, (15), 70-76. https://doi.org/10.31590/ejosat.518899

Bakhshinategh, B., Zaiane, O. R., ElAtia, S. & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 23(1), 537-553. 10.1007/s10639-017-9616-z

Baltacı, A. (2018). The Data Mining: Measurement of Academic Achievement in Faculty of Divinity Students by Data Mining. *Din ve Bilim –Muş Alparslan Üniversitesi İslami İlimler Fakültesi Dergisi*, 1(1), 1-23

Başer, S. H., Hökelekli, O. ve Kemal, A. D. E. M. (2020). Ortaöğretimde Öğrenim Gören Öğrenci Performanslarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi. *Bilgisayar Bilimleri ve Teknolojileri Dergisi*, 1(1), 22-27.

Bilen, Ö., Hotaman, D., Aşkın, Ö. E. & Büyüklü, A. H. (2014). LYS başarılarına göre okul performanslarının eğitsel veri madenciliği teknikleriyle incelenmesi: 2011 İstanbul örneği. *Eğitim ve Bilim*, 39(172), 78-94.

Bilgin, M. (2018). *Veri biliminde makine öğrenmesi makine öğrenmesi teorisi ve algoritmaları* (2. Baskı). Papatya Bilim.

Bliuc, A. M., Ellis, R., Goodyear, P. & Piggott, L. (2010). Learning through face-to-face and online discussions: Associations between students' conceptions,approaches and academic performance in political science. *British Journal of Educational Technology*, 41(3), 512-524. https://doi.org/10.1111/j.1467-8535.2009.00966.x

Byeon, H. (2022). Developing a predictive model for depressive disorders using stacking ensemble and naive Bayesian nomogram: using samples representing South Korea. *Frontiers in Psychiatry*, 12. https://doi.org/10.3389/fpsyt.2021.773290.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, 9 (13), 1-73.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357. https://doi.org/10.48550/arXiv.1106.1813

Chawla, Nitesh V. (2005). *Data mining for imbalanced datasets: an overview*. O. Maimon and L. Rokach (Ed.), Data mining and knowledge discovery handbook. 853-867. Boston. Springer.

Çiftçi, F., Kaleli, C. & Serkan, Ü. (2018). Öznitelik seçme ve makine öğrenmesi yöntemleriyle eğitmen performansının tahmin edilmesi. *Anadolu Journal of Educational Sciences International*, 8(2), 419-440. https://doi.org/10.18039/ajesi.454587

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F. & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in human behavior*, 73, 247-256. https://doi.org/10.1016/j.chb.2017.01.047

Dağ, M. (2013). İlahiyat lisans tamamlama (İLİTAM) programlarında Kur'an dersi-müfredat, materyal hazırlama ve karşılaşılan sorunlar. *Ekev Akademi Dergisi*, 17 (55), 37-54.

Devasia, T., Vinushree, T. P. & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. 91-95. IEEE. https://doi.org/10.14569/IJACSA.2016.070531

Durairaj, M. & Vijitha, C. (2014). Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies*, 5(4), 5987-5991. https://doi.org/10.1016/j.matpr.2021.05.646

Dutt, A., Ismail, M. A. & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access*, 5, 15991-16005. https://doi.org/10.1109/ACCESS.2017.2654247

Educational Data Mining, (2022, Ocak 5). International Educational Data Mining Society. (2021). educationaldatamining.org. https://educationaldatamining.org/.

Ersöz, A. R. (2017). *Eğitsel veri madenciliği ile öğrenci profillerinin belirlenmesi*. [Yayınlanmamış Doktora Tezi. Bursa Uludağ Üniversitesi].

Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164-181. https://doi.org/10.1016/j.engappai.2010.09.007

Genç, M. F. & Ayhan, M. (2021). İLİTAM Bölümü Öğrencilerinin Hadis Dersine Yönelik Tutumları. *Mizanü'l-Hak: İslami İlimler Dergisi*, (12), 77-109. https://doi.org/10.47502/mizan.933285

Gonçalves, A. F. D., Maciel, A. M. A. & Rodrigues, R. L. (2017). Development of a data mining education framework for data visualization in distance learning environments. *In International Conference on Software Engineering and Knowledge Engineering*. https://doi.org/0.18293/SEKE2017-130

Gümrükçüoğlu, S. & Genç, M. F. (2020). İLİTAM Bölümü Öğrencilerinin İlâhiyat *Eğitimine Bakışı Kocaeli Üniversitesi İlâhiyat Fakültesi İLİTAM Örneği*. İHYA Uluslararası İslam Araştırmaları Dergisi, 6(2), 640-656.

Güre, Ö. B., Kayri, M. & Erdoğan, F. (2020). PISA 2015 matematik okuryazarlığını etkileyen faktörlerin eğitsel veri madenciliği ile çözümlenmesi. *Eğitim ve Bilim*, 45(202), 393-415. https://doi.org/10.15390/EB.2020.8477

Hakyemez, T. C. (2015). *İlk Yil Öğrencilerinin Akademik Performansina Etki Eden Faktörlerin Araştirilmasi ve Bu Faktörlere Bağli Olarak Başarilarinin Tahminine Yönelik Bir Karar Destek Sistemi Tasarim*. [Yayınlanmamış Yükseklisans Tezi. Sakarya Üniversites]. Ulusal Tez Merkezi.

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

Hermaliani, E. H., Fanani, A. Z., Santoso, H. A., Affandy, A., Purwanto, P., Muljono, M., Syukur, A., Setiadi, D.R.I.M. & Rafrastara, F. A. (2022). Systematic Review of Educational Data Mining for Student Performance Prediction using Bibliometric Network Analysis (SeBriNA). *In 2022 International Seminar on Application for Technology of Information and Communication (iSemantic)* (s. 463-468). IEEE. https://doi.org/10.1109/iSemantic55962.2022.9920477

Howard, S. K., Ma, J. & Yang, J. (2016). Student rules: Exploring patterns of students' computer-efficacy and engagement with digital technologies in learning. *Computers & Education*, 101, 29-42. https://doi.org/10.1016/j.compedu.2016.05.008

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), 1-14. https://doi.org/10.3390/sym12020213.

Imran, M., Latif, S., Mehmood, D. & Shah, M. S. (2019). Student Academic Performance Prediction using Supervised Learning Techniques. *International Journal of Emerging Technologies in Learning*, 14(14), 92-104. https://doi.org/ 10.3991/ijet.v14i14.10310

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1), https://doi.org/61-72. 10.2478/cait-2013-0006

Kablan, S. (2020). Koronavirüs (Covid-19) pandemi sürecinde çevrimiçi yapılan Kuran-ı kerim dersi sınav değerlendirilmesi: İstanbul Üniversitesi ilahiyat fakültesi ilitam sınavları örneği. *Atlas international congress on social sciences* 7.

Kamath, U. & Choppella, K. (2017). *Mastering Java Machine Learning: A Java developer's guide to implementing machine learning and big data architectures*. Packt Publishing.

Karateke, T. (2020). İlitam öğrencilerinin bu programı seçme nedenleri ve karşılaştıkları sorunlar: Fırat Üniversitesi örneği. *Değerler Eğitimi Dergisi*, 18(39), 235-262. https://doi.org/10.34234/ded.634501

Kassim, A. A., Kazi, S. A. & Ranganath, S. (2004). A web-based intelligent learning environment for digital systems. *International Journal of Engineering Education*, 20(1), 13-23. https://doi.org/10.1108/02640470610689250

Kay, J. (2000). Stereotypes, student models and scrutability. *In International Conference on Intelligent Tutoring Systems* (s. 19-30). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45108-0_5

Kaymakcan, R., Meydan, H., Telli, A. & Cevherli, K. (2013). Paydaşlarına göre ilahiyat lisans tamamlama (İLİTAM) programının değerlendirilmesi. *Değerler Eğitimi Dergisi*, 11(26), 71-110.

Keskin, S., Aydin, F. & Yurdugül, H. (2019). Eğitsel veri madenciliği ve öğrenme analitikleri bağlaminda e-öğrenme verilerinde aykiri gözlemlerin belirlenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, 9(1), 292-309. https://doi.org/10.17943/etku.475149

Khasanah, A. U. (2017). A comparative study to predict student's performance using educational data mining techniques. *In IOP Conference Series: Materials Science and Engineering*, 215 (2017). https://doi.org/10.1088/1757-899X/215/1/012036

Kismet, E. (2018). *Eğitsel veri madenciliğinde kullanılmak üzere experience api (XAPI) temelli öğrenme deneyimi kayıtlarının işlenebilmesi için bir model geliştirilmesi*. [Yayınlanmamış yüksek lisans tezi. Kocaeli Üniversitesi].

Lantz, B. (2015). *Machine Learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R* (2. Baskı). Packt Publishing.

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt Publishing.

Longadge, R. & Dongre, S. (2013). Class imbalance problem in data mining review. *International Journal of Computer Science and Network (IJCSN)*, 2(1). https://doi.org/10.48550/arXiv.1305.1707

Maimon, O. & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook* (2. Baskı). Springer

McClean, S. I. (2003). Data mining and knowledge discovery. R. A. Meyers (Ed.), *Encyclopedia of Physical Science and Technology* (3. Baskı, s. 229-246). New York. Academic Press.

Miller, L. D., Soh, L. K., Samal, A., Kupzyk, K. & Nugent, G. (2015). A Comparison of Educational Statistics and Data Mining Approaches to Identify Characteristics That Impact Online Learning. *Journal of Educational Data Mining*, 7(3), 117-150.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill Science.

Moradi, H., Moradi, S. A. & Kashani, L. (2014). *Students' performance prediction using multi-channel decision fusion*. A. Peña-Ayala (Ed.), Educational Data Mining (s. 151-174). Springer.

Morsy, S. & Karypis, G. (2017). Cumulative knowledge-based regression models for next-term grade prediction. *In Proceedings of the 2017 SIAM International Conference on Data Mining* (s. 552-560). Society for Industrial and Applied Mathematics.

Namoun, A. & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), 237. https://doi.org/10.3390/app11010237

Özbay, Ö. (2015). Veri madenciliği kavramı ve eğitimde veri madenciliği uygulamaları. *The Journal of International Educational Sciences* 2(5), 262-262. https://doi.org/10.16991/INESJOURNAL.162

Özçınar, H. (2006). *KPSS sonuçlarının veri madenciliği yöntemleriyle tahmin edilmesi* [Yüksek lisans tezi, Pamukkale Üniversitesi]. Ulusal Tez Merkezi.

Özdemir, Ş. (2016). *Eğitimde veri madenciliği ve öğrenci akademik başarı öngörüsüne ilişkin bir uygulama*. [Doktora tezi, İstanbul Üniversitesi].

Öztürk, A. (2018). Açık ve uzaktan öğrenme ortamlarında eğitsel veri madenciliği. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, 4(2), 10-13.

Pascual-Cid, V., Vigentini, L. & Quixal, M. (2010). Visualising virtual learning environments: Case studies of the Website exploration tool. *In 2010 14th International Conference Information Visualisation* (s. 149-155). IEEE. https://doi.org/10.1109/IV.2010.31

Polat, A. (2021). *Açık öğretim liseleri öğrencilerinin okul terki ve mezuniyet durumlarının eğitsel veri madenciliği ile incelenmesi*. [Doktora tezi, Sakarya Üniversitesi]. Ulusal Tez Merkezi

Polyzou, A. & Karypis, G. (2016). Grade prediction with models specific to students and courses. *International Journal of Data Science and Analytics*, 2(3), 159-171. https://doi.org/10.48550/arXiv.1906.00792

Prasetiyowati, M. I., Maulidevi, N. U. & Surendro, K. (2021). Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *Journal of Big Data*, 8(1), 84. https://doi.org/10.21203/rs.3.rs-132775/v1

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. https://doi.org/10.1007/BF00116251

Rapidminer (2020). Rapidminer Documentation Weight by rule (RapidMiner Studio Core;Version 9.9). "Rapidminer, Weight by Rule 2020". Erişim 17 Ocak 2022. http://docsupcoming.rapidminer.com/9.2/studio/operators/modeling/feature_weights/weight_by_rule.html.

Rapidminer (2020). Rapidminer Rule Induction (2020) (RapidMiner Studio Core; Version 9.9). "Rapidminer, rule induction 2020". Erişim 17 Ocak 2022. http://docs-upcoming.rapidminer.com/9.4/studio/operators/modeling/predictive/rules/rule_induction.html?upcoming-rapidminer%5Bpage%5D=9.

Refaeilzadeh, P., Tang, L. & Liu, H. (2009). *Cross-validation*. L. Liu ve M. T. Özsu. (Ed.), Encyclopedia of Database Systems (s. 532-538). Boston. Springer. https://doi.org/10.1007/978-0-387-39940-9_565

Rodrigues, M. W., Isotani, S. & Zarate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. *Telematics and Informatics*, 35(6), 1701-1717. https://doi.org/10.1016/j.tele.2018.04.015

Rojanavasu, P. (2019). Educational data analytics using association rule mining and classification. *In 2019 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications*

engineering. 142-145. IEEE. https://doi.org/ 10.1109/ECTI-NCON.2019.8692274

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

Romero, C., Ventura, S., Espejo, P. G. & Hervás, C. (2008). Data mining algorithms to classify students. *In Educational data mining*.

Şengür, D. & Tekin, A. (2013). Öğrencilerin mezuniyet notlarının veri madenciliği metotları ile tahmini. *Bilişim Teknolojileri Dergisi*, 6(3), 7-16.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422. https://doi.org/ 10.1016/j.procs.2015.12.157

Sivakumar, S., Venkataraman, S. & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, 9(4), 1-5. https://doi.org/10.17485/ijst/2016/v9i4/87032

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

Sorour, S. E., Mine, T., Goda, K. & Hirokawa, S. (2014). Predicting students' grades based on free style comments data by artificial neural network. *In 2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (s. 1-9). IEEE. https://doi.org/ 10.1109/FIE.2014.7044399

Tekin, A. & Öztekin, Z. (2018). Eğitsel veri madenciliği ile ilgili 2006-2016 yılları arasında yapılan çalışmaların incelenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, 8(2), 108-124. https://doi.org/10.17943/etku.351473

Tosunoğlu, E., YILMAZ, R., Özeren, E. & Sağlam, Z. (2021). Eğitimde makine öğrenmesi: Araştırmalardaki güncel eğilimler üzerine inceleme. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi,* 3(2), 178-199. https://doi.org/10.3815/akef.2021.16

Wang, C. S. & Lin, S. L. (2012). Combining fuzzy AHP and association rule to evaluate the activity processes of e-learning system. *In 2012 Sixth International Conference on Genetic and Evolutionary Computing* (s. 566-570). IEEE.

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* 1. 29-40.

Yukselturk, E., Ozekes, S., & Turel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-learning*, 17(1), 118-133. https://doi.org/10.2478/eurodl-2014-0008

Zacharis, N. Z. (2016). Predicting student academic performance in blended learning using artificial neural networks. *International Journal of Artificial Intelligence and Applications*, 7(5), 17-29. https://doi.org/10.5121/IJAIA.2016.7502