DİJİTED
DİJİTAL TEKNOLOJİLER VE EĞİTİM DERGİSİ

Dijital Teknolojiler ve Eğitim Dergisi, Yıl: 2024, Cilt: 1, Sayı: 1, Sayfa : 28-38
Journal of Digital Technologies and Education, Year: 2024, Volume: 3, Issue: 1, Pages : 28-38

Araştırma Makalesi

Gönderim Tarihi : 14.06.2024
Kabul Tarihi : 29.06.2024

DOI: 10.5281/zenodo.12637330

Merve DÜNDER[1]
Emre DÜNDER[1]

Kategorik verilerde sınıf dengesizliği durumunda makine öğrenimi algoritmalarının karşılaştırılması: Öğrencilerin başarı durumları üzerine bir uygulama

Comparison of machine learning algorithms in the presence of class ımbalance in categorical data: An application on student success

Özet

Makine öğrenme algoritmaları, eğitim bilimlerinde öğrenci performansını değerlendirmek üzere çeşitli amaçlar doğrultusunda uygulanmaktadır. Eğitim alanındaki veri setlerinin kategorik verilerden oluşması, sınıf dengesizliği sorununda alternatif veri türetme tekniklerinin kullanılmasını gerektirmektedir. Bu çalışma, bu durumu ele alarak öğrencilerin başarı durumlarını tahmin etmede kullanılan çeşitli makine öğrenimi algoritmalarının performanslarını karşılaştırmayı amaçlamaktadır. Uygulamada sınıf dengesizliğini çözmek üzere SmoteNC tekniğinden yararlanılmış ve analiz bulguları, beş farklı makine öğrenme tekniği ile değerlendirilmiştir. Veri analizi sonuçları, sınıf dengesizliği giderildiği takdirde, sınırlı sayıda gözlem içeren verilerde makine öğrenme algoritmalarının başarıyla uygulanabileceğini göstermektedir.

Anahtar Kelimeler: Eğitimde başarı tahmini, makine öğrenme algoritmaları, sınıf dengesizliği

Abstract

Machine learning algorithms are applied in educational sciences for various purposes to evaluate student performance. Given that educational datasets often consist of categorical data, addressing class imbalance issues requires the use of alternative data generation techniques. This study aims to address this issue by comparing the performance of various machine learning algorithms in predicting student success. In this application, the SmoteNC technique is used to address class imbalance, and the analysis findings are evaluated using five different machine learning techniques. The results of the data analysis indicate that if class imbalance is mitigated, machine learning algorithms can be successfully applied to datasets with a limited number of observations.

Keywords: Predicting success in education, machine learning algorithms, class imbalance

[1]Ondokuz Mayıs Üniversitesi, Fen Fakültesi, Türkiye, dundermerve@gmail.com.
[2]Ondokuz Mayıs Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Türkiye, emre.dunder@omu.edu.tr.

Giriş

Geçmişten bugüne eğitim sistemlerinin kalitesinin artırılması, öğrencilerin akademik başarılarının desteklenmesi ve iyileştirilmesi modern eğitim bilimlerinin ana amaçlarından biridir. Eğitimde başarıyı etkileyen iç ve dış faktörlerin karmaşıklığı ve çeşitli etkenlerin bir arada değerlendirilmesi gerekliliği, bu başarıyı tahmin etme ve iyileştirme çabalarının yetersiz kalmasına neden olmuştur. Bu noktada makine öğrenme algoritmaları, eğitim verilerinin analiz edilmesi ve bu analizlere yeni ve etkili bir yaklaşım sunulmasına olanak sağlamaktadır. Makine öğrenme algoritmaları, büyük veri setlerinden çıkarımda bulunma yetenekleri sayesinde, öğrencilerin akademik başarılarını ve başarısız oldukları durumları tahmin etmede önemli bir rol oynamaktadır (Breiman, 2001). Öğrencilerin demografik faktörleri, derse katılımları, akademik performansları ve aile yapısı gibi çeşitli özellikleri, makine öğrenme algoritmaları aracılığıyla analiz edilerek, bireysel başarı düzeyleri tahmin edilebilmektedir (Şengür, 2013). Bu da eğitimcilerin ve eğitim yönetim sistemlerinin, her bir öğrencinin ihtiyaçlarına göre özelleştirilmiş eğitim stratejileri geliştirmelerine olanak tanımaktadır. Teknoloji kullanımının artması ile birlikte, eğitim kurumları ve araştırmacılar, öğrencilerin başarılarını ve başarısız oldukları durumları tahmin etmek ve iyileştirmek için giderek daha fazla veri elde etmekte ve bu verileri gelecekte uygulayacakları stratejiler için kullanmayı amaçlamaktadır. Son zamanlarda, artan sayıda eğitim kurumu, öğrencilerin kaydını iyileştirmek, öğrenci bağlılığını artırmak, okulu bırakma oranını düşürmek, mezunlarla uzun vadeli ilişkileri sürdürmek ve yerleştirme sonuçlarını önceden tahmin etmek için tahmine dayalı analizleri uygulamaktadır (Halde, 2016). Bu noktada, makine öğrenme algoritmaları, gelecekte uygulanacak bu stratejilere bilimsel bir yaklaşım sumaktadır.

Özellikle öğrencilerin başarı düzeylerini etkileyen tüm dış faktörlerin analizlere dahil edilebilmesi ve bu faktörler arasındaki ilişkilerin belirlenebilmesi, makine öğrenme yöntemlerinin gücünü ortaya koymaktadır (Menon ve diğ, 2021). Bu çalışmada, eğitim bilimleri alanında yaygın olarak kullanılan makine öğrenme algoritmaları incelenerek, öğrencilerin başarılarının tahmin edilmesinde hangi algoritmaların daha etkili olduğu değerlendirilecektir. Çalışmanın amacı, çeşitli makine öğrenme algoritmalarını kullanarak öğrencilerin akademik başarıları düzeylerini tahmin etmektir. Bu doğrultuda, mevcut eğitim verileri kullanılarak farklı algoritmaların performansları karşılaştırılacak ve en yüksek doğruluk oranına sahip model belirlenecektir. Ayrıca, elde edilen sonuçlar ışığında, makine öğrenme algoritmalarının eğitim bilimlerinde nasıl daha etkili kullanılabileceğine dair öneriler sunulacaktır. Bu çalışma, eğitim bilimleri alanında makine öğrenme tekniklerinin uygulanabilirliğine dair önemli katkılar sunmayı amaçlamaktadır. Araştırmamızın temel iddiası; gözlem sayısının sınırlı olduğu ve sınıf dengesizliğinin var olduğu durumlarda dahi, makine öğrenme yöntemlerinin öğrencilerin akademik başarılarını tahmin edebilmek için etkin bir şekilde uygulanabileceğini ortaya koymaktır.

Literatürde, öğrencilerin başarılarını tahmin etmek için çeşitli makine öğrenme algoritmaları kullanılmıştır, ancak bu algoritmaların performansları ve uygulanabilirlikleri konusunda hala pek çok soru işareti bulunmaktadır. Bu çalışma, bu boşluğu doldurmayı ve makine öğrenme algoritmalarının eğitim bilimlerinde daha etkin kullanımını sağlamayı hedeflemektedir. Uygulama aşamasında, öğrencilerin derslerine yönelik başarı durumunu modellemek üzere makine öğrenme algoritmaları karşılaştırılmıştır. Veri setinde sınıf dengesizliği problemi olduğu için, kategorik bağımsız değişkenlere uygun SmoteNC tekniği ile veriler dengeli hale getirilmiştir. Çalışmanın 1. bölümünde literatürde yapılan çalışmalara ilişkin bilgiler sunulmuştur. Araştırmada kullanılan yöntemlere ait genel bilgiler 2. bölümde verilmiştir. Uygulamalara ait bulgular 3. bölümde verilmiş; 4. bölümde makine öğrenme algoritmalarına dair sonuç ve değerlendirmeler ele alınmıştır.

1. Literatür

Karbasi ve diğerleri (2021) çalışmalarında, öğrencilerin başarı düzeylerinin tahmin edilmesinde makine öğrenme tekniklerinin eğitim bilimlerinde önemli bir rol üstlendiğine değinmiş ve yükseköğretimde akademik başarıyı tahmin etmek için lojistik regresyon, karar ağaçları ve destek vektör makineleri (SVM) gibi çeşitli makine öğrenme modelleri kullanılmıştır. Araştırma sonucunda bu modellerin doğru tahminler yapabilme potansiyeline sahip olduğu tespit edilmiş ve XGBoost'un en doğru sınıflandırmayı yaptığı sonucuna varılmıştır.

Yağcı ve Altın (2020), MOODLE verilerini kullanarak öğrenci akademik performansını tahmin etmek için makine öğrenme modellerinden rastgele orman ve lojistik regresyon modellerinin, öğrencilerin final sınav sonuçlarını doğru bir şekilde tahmin etme potansiyeline sahip olduğunu göstermişlerdir. Orji ve Vassileva (2022) öğrenci motivasyon verilerini kullanarak akademik performanslarını değerlendirmek amacıyla makine öğrenme tekniklerinin bu alandaki rolünü incelemiştir. Bu çalışma, öğrencilerin motivasyon verilerini kullanarak akademik performanslarını ve çalışma stratejilerini doğru bir şekilde tahmin edebilen makine öğrenme yaklaşımlarının eğitimde önemli bir rol oynadığını göstermiştir.

Wang ve Lee (2020) öğrencilerin başarılarının ve okulu bırakma oranlarının tahmin edilmesinin eğitim alanında önemli bir araştırma konusu olduğuna dikkat çekerek çeşitli denetimli makine öğrenme algoritmalarının etkinliğini incelemişlerdir. Denetimli makine öğrenme algoritmalarının, öğrencilerin başarı düzeyleri ve okulu bırakma oranlarını doğru bir şekilde tahmin ederek, eğitimcilerin risk altındaki öğrencileri erken belirlemelerine ve müdahalede bulunmalarına olanak sağladığını göstermiştir. Smith ve Johnson (2021) çalışmasında, üniversite öğrencilerinin akademik başarısını tahmin etmek için XGBoost, karar ağaçları, rastgele ormanlar ve destek vektör makineleri gibi makine öğrenme algoritmalarının kullanımı incelenmiştir. Çalışma, bu tekniklerin doğru tahminler yaparak eğitimcilerin öğrencilere daha iyi rehberlik ve destek sağlamalarına yardımcı olabileceğini göstermiştir.

Jones ve Roberts (2023) çalışmasında, öğrenci akademik performansını tahmin etmek için makine öğrenme algoritmalarının kullanımı incelenmiştir. Bu çalışmanın bulguları, makine öğrenme tekniklerinin %85 doğrulukla öğrenci başarı tahminleri yaparak eğitimcilerin daha doğru rehberlik sağlamalarına yardımcı olabileceğini göstermektedir. Lopez ve Kim (2021) çalışmasında, öğrenci başarısının tahmin edilmesinde otomatik makine öğrenme yaklaşımlarının, model seçimini optimize ederek tahmin doğruluğunu %90'a kadar artırma potansiyeline sahip olduğu tespit edilmiştir. Çalışma, bu tekniklerin eğitim verilerini analiz ederek en uygun modelleri otomatikleştirme sürecini ve eğitimcilerin öğrenci başarılarını daha doğru tahmin etmelerine olanak tanıdığını göstermektedir. Hernandez-Leal ve Gonzalez (2022) çalışmasında, harmanlanmış öğrenme ortamlarında öğrenci performansını tahmin etmek için makine öğrenme tekniklerinin kullanımı incelenmiştir. Bu çalışma, makine öğrenme algoritmalarının öğrenci verilerini analiz ederek performans tahminlerinin doğruluğunu %88'e kadar artırabileceğini göstermektedir. Araştırma, bu tekniklerin eğitimcilerin öğrenci başarılarını daha doğru öngörmelerine yardımcı olabileceğini vurgulamaktadır. Gonzalez ve Rodriguez (2022) çalışmasında, öğrenci sonuçlarını tahmin etmek için iki aşamalı bir makine öğrenme yaklaşımı uygulanmıştır. İlk aşamada denetimsiz öğrenme teknikleriyle veri kümeleri alt gruplara ayrılmış, ikinci aşamada bu gruplar üzerinden denetimli öğrenme teknikleriyle tahminlerde bulunulmuştur. Bu yöntem, öğrenci başarılarını %92 doğrulukla tahmin etmiştir.

Literatürde yapılan çalışmalara göre, sınıf dengesizliği problemini yeterli düzeyde ele alınmamıştır. Sınıf dengesizliğini gidermek için yapılan sınırlı sayıdaki çalışmalarda, kategorik bağımsız değişkenlerin var olduğu durumlar için, sayısal verilere yönelik geliştirilmiş klasik Smote yöntemlerinin hatalı şekilde kullanılması dikkat çekicidir. Araştırmada benimsediğimiz metodolojik yaklaşım, söz konusu boşlukları doldurmayı da hedeflemektedir.

2. Yöntem

Bu çalışmada, öğrencilerin akademik başarılarını tahmin etmek ve eğitim alanında stratejik hamlelerde bulunmak amacıyla çeşitli makine öğrenme algoritmaları kullanılmıştır. Bu algoritmalar, eğitim verilerini analiz ederek doğru tahminlerde bulunmak için tasarlanmıştır. Kullanılan makine öğrenme yöntemleri şunlardır:

2.1. Lineer diskriminant analizi

Lineer diskriminant analizi (LDA), sınıflandırma problemleri için kullanılan bir denetimli öğrenme tekniğidir. LDA, farklı sınıfları ayırt etmek için verilerin doğrusal kombinasyonlar kullanmaktadır. Bu tekniğin temel amacı, sınıflar arasındaki farkları en iyi şekilde temsil eden bir doğrusal ayrımı sağlamaktır. Uygulama aşamasında sınıflar arası değişkenliğin sınıf içi varyansa oranını optimize edilir ve sınıfların yüksek oranda ayırt edilebilir olmasını sağlanmaktadır (Wang ve Zhang, 2012). LDA, her bir sınıfın ortalamave varyans-kovaryans

matrislerini elde ederek, bu matrislere dayalı diskriminant fonksiyonları türetir. Diskriminant fonksiyonları, yeni gözlemleri önceden tanımlanmış sınıflardan birine atamak için kullanılır ve bu şekilde tahminleme süreci gerçekleştirilir. Algoritmanın akışına göre, uygulamada LDA için herhangi bir hiperparametre seçimine gerek yoktur.

2.2. Rastgele ormanlar algoritması

Rastgele ormanlar (Random Forests), birden fazla karar ağacının birleşiminden oluşan bir denetimli öğrenme yöntemidir. Modeli öğrenme aşamasında topluluk algoritmalarından yararlanarak, Bootstrap tekniği ile türetilen veri setleri üzerinden farklı karar ağaçları oluşturur (Breiman, 2001). Karar ağaçları oluşturulurken belirli sayıda rasgele özellik seçilir ve bu özellikler üzerinden karar ağacı modelleri elde edilir. Sınıflandırma aşamasında tahminler, farklı karar ağaçlarının oylanması üzerinden belirlenir.

Rasgele ormanlar algoritması, birden fazla karar ağacını bir araya getirerek tahmin doğruluğunu artırır ve aşırı öğrenmeyi önler. Bu doğrultuda çeşitli hiperparametrelerin seçilmesi gereklidir. Uygulama aşamasında ormanda oluşturulacak ağaç sayısı, her bir ağacın maksimum derinliği, ağaçlarındaki bir düğümün dallanması için gereken minimum örnek sayısı ve her düğümde değerlendirilecek rasgele seçilen bağımsız değişkenlerin sayısı (mtry) belirlenir.

2.3. Destek vektör makineleri

Destek vektör makineleri (DVM'ler), makine öğrenme uygulamalarında sınıflandırma ve regresyon analizleri kapsamında kullanılan denetimli öğrenme teknikleridir. DVM'lerin esas amacı, verileri etkili bir şekilde farklı sınıflara ayıran ve aralarındaki marjı en üst düzeye çıkaran ideal hiper düzlemi belirlemektir. Hiper düzlemi belirlerken, uygulama verilerini daha fazla sayıda boyuta sahip bir uzaya dönüştürmek için çekirdek fonksiyonları kullanılır (Awad ve diğ., 2015). Çekirdek fonksiyonları üzerinden gerçekleştirilen bu dönüşüm işlemi, orijinal uzayda doğrusal bir şekilde ayrılamayan sınıfların ayrıştırılmasına olanak tanır. DVM'lerin uygulama sürecinde, ele alınan çekirdek fonksiyonuna ait parametrelerin seçilmesi gerekir.

2.4.Güçlendirilmiş lojistik regresyon modeli

Güçlendirilmiş (Boosted) lojistik regresyon, temel lojistik regresyon modelinin doğruluğunu artırmak için bir dizi zayıf tahmincinin sıralı olarak oluşturulduğu ve daha sonra birleştirildiği bir yöntemdir. Her bir adımda, önceki modelin hatalarını düzeltmeye odaklanan yeni bir model eklenir. Bu şekilde, nihai model daha güvenilir sonuçlar elde etmektedir. Güçlendirilmiş lojistik regresyon modeli, farklı zayıf sınıflandırıcılar üzerinden modeli iyileştirmeye dayalı bir yaklaşım sunduğu için, klasik lojistik regresyondan daha yüksek performansa sahiptir (Schonlau, 2005). Bu algoritmada hiperparametre olarak iterasyon sayısı belirlenmelidir.

2.5.Model ortalamalı yapay sinir ağları

Yapay sinir ağları (YSA), biyolojik sinir sistemlerinden esinlenerek geliştirilmiş bir makine öğrenme algoritmasıdır. Bu yöntemde, girdi verileri bir veya daha fazla gizli katman aracılığıyla işlenir ve her katmandaki nöronlar arasında bağlantılar bulunur (Shen ve diğ., 2022). YSA'ların eğitim sürecinde ağırlıklar ve yanlılık değerleri, hata fonksiyonunu minimize edecek şekilde ayarlanır. Araştırmacılar, yapay sinir ağlarını karmaşık ve doğrusal olmayan ilişkileri modellemek için güçlü bir analiz yaklaşımı olarak sıkça kullanmaktadır. Model ortalama (model averaging) yaklaşımı ise birden fazla makine öğrenme modelinin tahminlerini birleştirerek tek bir tahmin oluşturur. YSA'larda model ortalaması kullanımı, farklı sinir ağı modellerinin ayrı ayrı hatalarını azaltarak genel tahmin performansını artırmayı hedefler ve gözlemlere ilişkin nihai tahminler, sonuçların ağırlıklı ortalaması alınarak gerçekleştirilir (He ve Tafti, 2019). Uygulamada YSA'nı optimize etmek için geriye yayılım (backpropagation) metodu tercih edilmiştir.

Model ortalamalı YSA uygulanırken, tek katmanlı bir yapı kullanılmıştır. Uygulamada, katmandaki nöron sayısı (size) ve aşırı uyumu önlemek üzere ağırlık bozunum (weight decay) parametreleri ve çantalama (bag)

belirlenmelidir. Ek olarak, farklı sinir ağları ile örnekleme yaklaşımlardan torbalama (bagging) tekniğinin kullanılıp kullanılmayacağı da seçilebilir.

2.6 SmoteNC tekniği

SmoteNC, özellikle kategorik değişkenleri içeren veri kümelerinde sınıf dengesizliğini düzeltmek üzere geliştirilmiş bir sentetik veri üretme yöntemidir (Mukherjee ve Khushi, 2021). Bu veri türetme yaklaşımı, azınlık sınıfı için yapay örnekler oluşturarak dengesiz veri kümeleri sorununu ele alır ve böylece sınıfların dağılımını eşit hale getirir. SmoteNC, kategorik ve sürekli değişkenleri farklı şekillerde ele alarak yapay örnek üretme prosedürünü değiştirir. Algoritma, azınlık sınıfından örnekler arasında enterpolasyon yoluyla yeni veri noktaları oluşturarak sürekli özellikleri ele almak için geleneksel Smote yöntemini kullanır. SmoteNC, kategorik özellikler için yeni oluşturulan sentetik örneğin değerine karar vermek için en yakın komşular arasında bir çoğunluk oylama tekniği kullanır (Islahulhaq ve Ratih, 2021). Bu yaklaşımda, kategorik veriler için en yakın komşular arasında en sık görülen değer olan mod değeri atanır. Kategorik verilerin yanı sıra, sayısal veriler için de bu teknik kullanılabilir.

3. Uygulama

Çalışmanın bu bölümünde, öğrencilerin başarı durumlarını modellemek üzere kullanılan makine öğrenme algoritmalarının bulguları verilmiştir. Makine öğrenme algoritmalarından lineer diskriminant analizi (LDA), rastgele ormanlar (RO), polinomial destek vektör makineleri (DVMPOL), güçlendirilmiş lojistik regresyon (BOOSTLOJREG) analizi ve model ortalamalı yapay sinir ağları (MOYSA) metotları kullanılmıştır.

Uygulama aşamasında, UCI makine öğrenimi deposunda "Yükseköğretim Öğrencilerinin Performans Değerlendirmesi" isimli, açık erişime sahip olan bir veri seti kullanılmıştır (Yılmaz ve Şekeroğlu, 2019). Veri setinin açık haline, UCI veri seti deposundan ulaşılabilir (https://www.archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation).

Bu çalışma kapsamında, veri setindeki araştırma değişkenlerinin tamamı kullanılmamıştır. Veri setinde toplam n=145 gözlem bulunmaktadır. Bağımsız değişken olarak p=22 kategorik özellik ele alınmıştır. Bazı bağımsız değişkenlerin gruplarında düşük gözleme sahip gruplar birleştirilmiştir. Bağımlı değişken olarak da öğrencilerin başarı durumları, ders notlarına göre başarılı/başarısız şeklinde iki gruba ayrılmıştır. Araştırmada kullanılan değişkenlerin açıklamaları ve kısaltma şeklindeki gösterimleri, Tablo 1'de sunulmuştur.

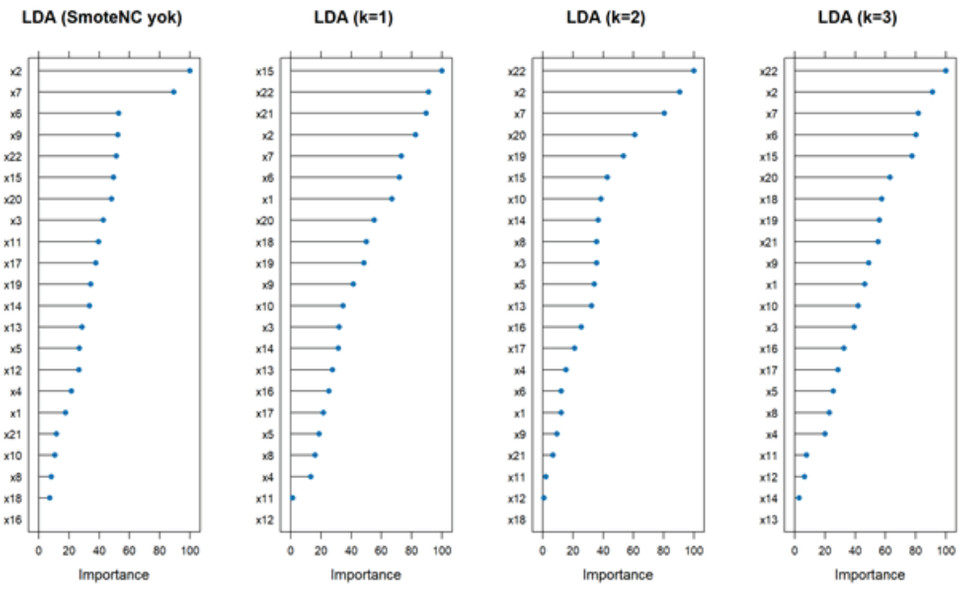
Tablo 1: Araştırma değişkenlerinin açıklama ve gösterimleri

Açıklama	Gösterim
Öğrenci yaşı	x1
Cinsiyet	x2
Mezun olunan lise türü	x3
Burs türü	x4
Ek iş	x5
Düzenli sanatsal veya spor aktivitesi	x6
Bir partneriniz var mı?	x7
Toplam gelir	x8
Üniversiteye ulaşım	x9
Kardeş sayısı	x10
Aile durumu	x11
Haftalık ders çalışma saatleri	x12

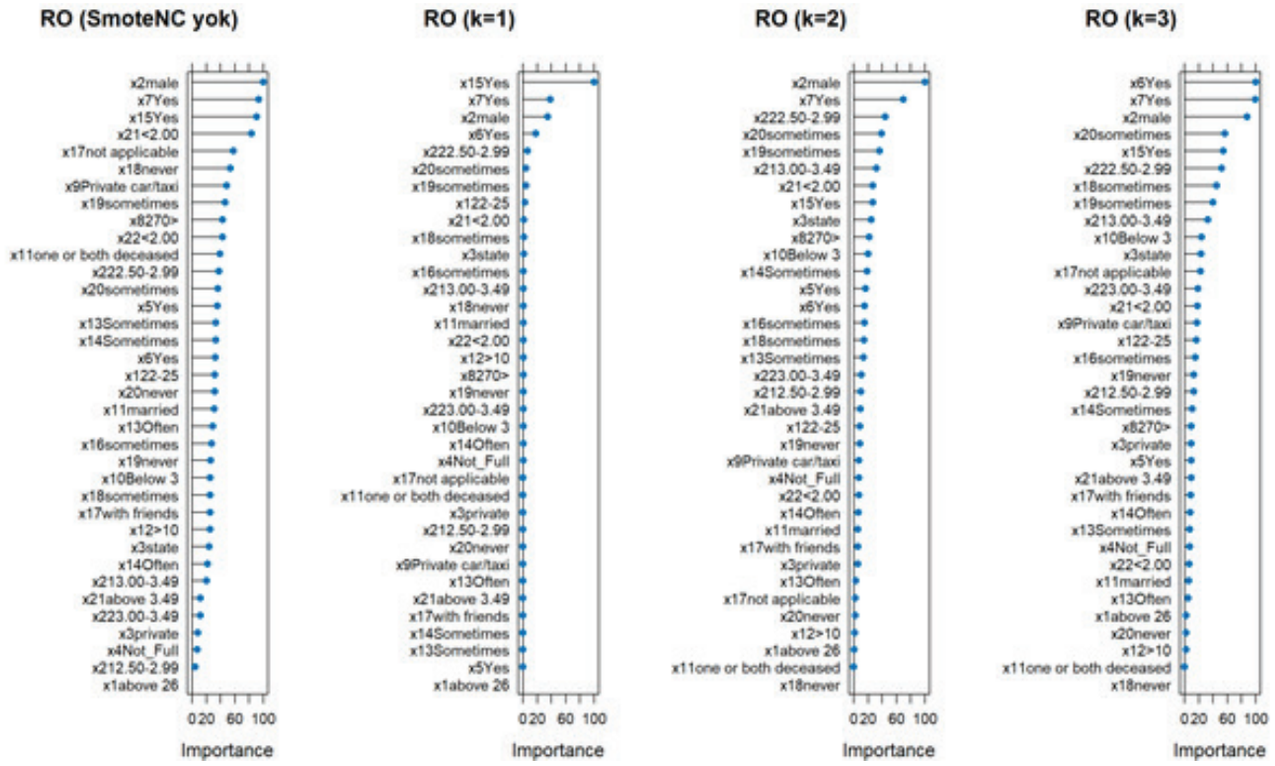
Bilim dışı okuma sıklığı	x13
Bilimsel okuma sıklığı	x14
Seminerlere katılım	x15
Derslere katılım	x16
Ara sınav hazırlığı	x17
Not alma	x18
Ders dinleme durumu	x19
Tartışmaların başarıyı artırma durumu	x20
Önceki dönem genel not ortalaması	x21
Mezuniyette beklenen genel not ortalaması	x22
Başarı durumu (Bağımlı değişken)	y

Makine öğrenme algoritmalarının performanslarını değerlendirmek üzere, verilerin %70'i eğitim, %30'u test verisi olarak ele alınmıştır. Makine öğrenme algoritmalarına ait en uygun parametreler, eğitim verileri üzerinden 5 katmanlı çapraz geçerlilik yaklaşımı ile belirlenmiştir. Araştırma sürecinde analiz edilen eğitim ve test verileri, ekte sunulmuştur.

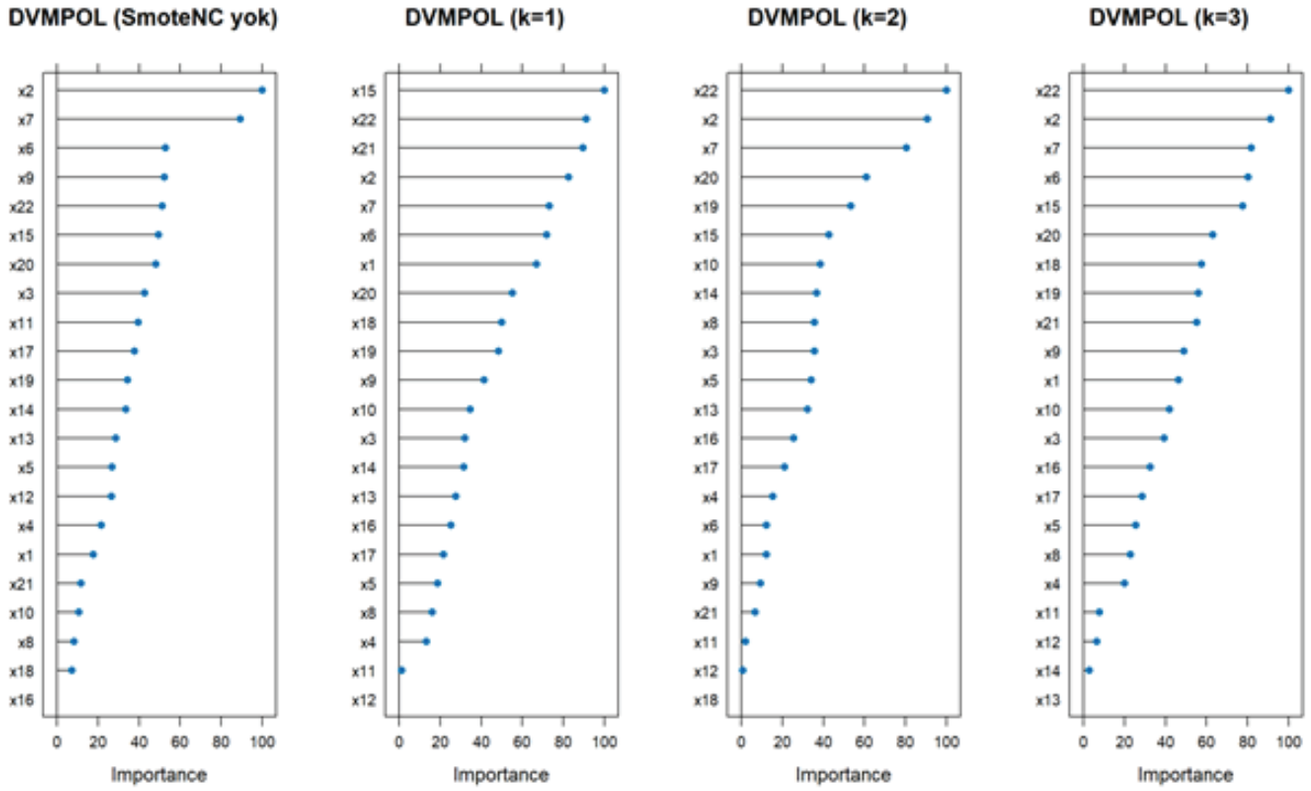
Eğitim verisinde yer alan bağımlı değişken verilerinde, başarısız olan öğrenciler örneklemin %3,9'unu oluşturmaktadır. Test verisinde de başarısız olan öğrenci oranı %9,1'dir. Bağımlı değişkenin eğitim verisindeki başarı durumlarına göre, dengesizlik oranı (IR) 24,6'dır. Başarısız olan öğrencilerin azınlıkta olması ve dengesizlik oranının yüksek olması, sınıf dengesizliği probleminin var olduğunu ortaya koymaktadır. Kategorik bağımsız değişkenlerden oluşan verilerden hareketle sınıf dengesizliğini gidermek üzere SmoteNC tekniği kullanılmıştır (Wongvorachan ve diğ, 2023). Son aşamada SmoteNC tekniği ile başarılı ve başarısız öğrenci oranları %50 seviyesine getirilmiştir. SmoteNC ile veri türetilirken, k=1,2,3 komşu sayıları için üç farklı veri kümesi elde edilmiştir. Veri analizi sonuçlarında SmoteNC uygulanmış ve uygulanmamış veri setlerinin bulguları bir arada verilmiştir. Algoritmalar tarafından hesaplanan göreli önem düzeyleri, veri görselleştirme teknikleri ile sunulmuştur. Sınıflama performansı ölçütleri için dengeli sınıflama oranları (DDSO), AUC değerleri, F-Skorları ve G-Ortalama sonuçları verilmiştir. Makine öğrenme uygulamaları R yazılımında (R Core Team, 2024) geliştirilmiş caret (Kuhn, 2008) ve MLmetrics (Yan, 2016) paketlerinden yararlanılmıştır.



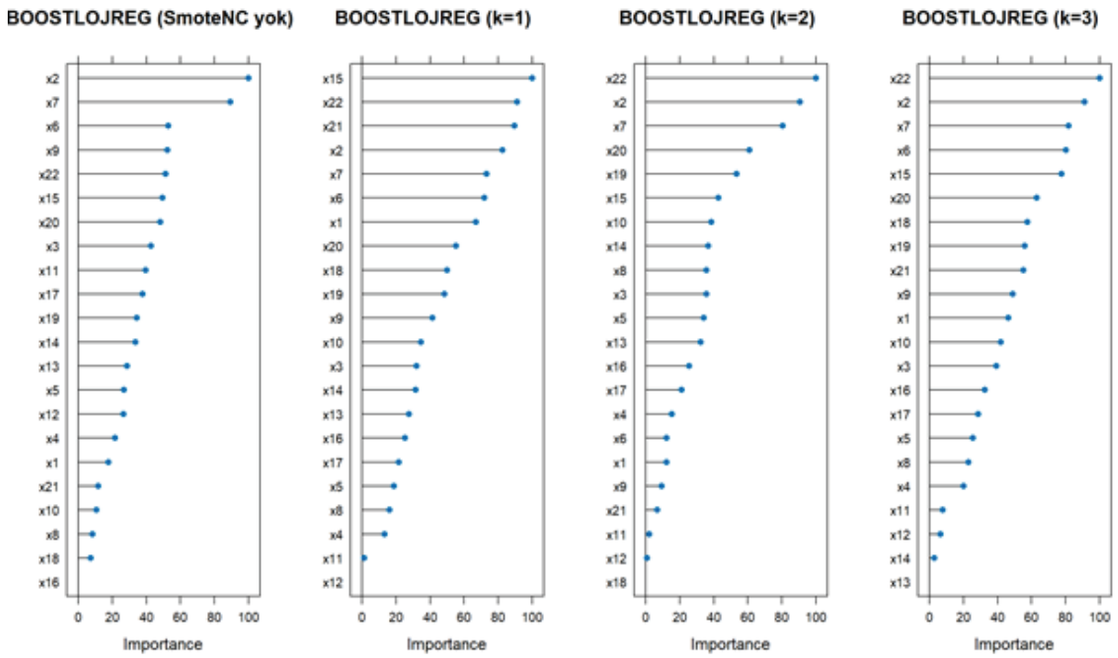
Şekil 1. LDA için göreli önem düzeyleri



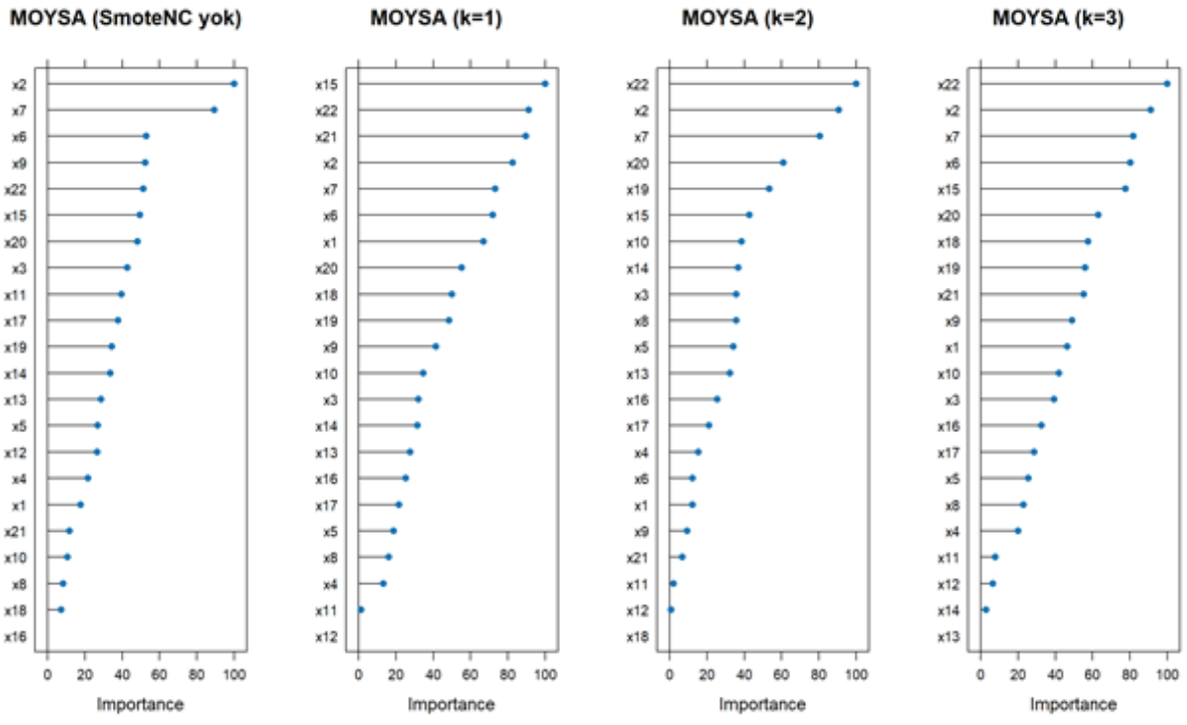
Şekil 2. RO için göreli önem düzeyleri



Şekil 3. DVMPOL için göreli önem düzeyleri



Şekil 4. BOOSTLOJREG için göreli önem düzeyleri



Şekil 5. MOYSA için göreli önem düzeyleri

Şekil 1-5 arasında makine öğrenme algoritmaları tarafından hesaplanmış, bağımsız değişkenlerine ait önem düzeylerinin grafiksel sonuçları verilmiştir. Bu sonuçlara bakıldığında, haftalık ders çalışma saatleri, derslere katılım, bilim dışı okuma sıklığı ve not alma durumlarının, öğrencilerin başarı sonuçları üzerinde önemli bir etkiye sahip olmadığı görülmektedir. Makine öğrenme algoritmalarındaki genel sonuçlar; cinsiyet, seminerlere katılım durumu ve mezuniyette beklenen genel not ortalaması değişkenlerinin, öğrencilerin başarı sonuçları üzerinde en yüksek öneme sahip faktörler olduğuna işaret etmektedir.

Tablo 2. Makine öğrenme algoritmalarının performans değerlendirme sonuçları

Algoritma	SmoteNC metodu	DDSO	AUC	F-Skoru	G-Ortalama
LDA	SmoteNC yok	0.700	0.700	0.923	0.671
	SmoteNC (k=1)	0.700	0.700	0.923	0.671
	SmoteNC (k=2)	0.713	0.713	0.937	0.680
	SmoteNC (k=3)	0.700	0.700	0.923	0.671
RO	SmoteNC yok	0.500	0.500	0.952	0.000
	SmoteNC (k=1)	0.725	0.725	0.950	0.689
	SmoteNC (k=2)	0.488	0.488	0.940	0.000
	SmoteNC (k=3)	0.500	0.500	0.952	0.000
DVM	SmoteNC yok	0.500	0.500	0.952	0.000
	SmoteNC (k=1)	0.625	0.625	0.964	0.500
	SmoteNC (k=2)	0.625	0.625	0.964	0.500
	SmoteNC (k=3)	0.725	0.725	0.950	0.689
BOOSTLOJREG	SmoteNC yok	0.700	0.700	0.923	0.671
	SmoteNC (k=1)	0.688	0.688	0.909	0.661
	SmoteNC (k=2)	0.688	0.688	0.909	0.661
	SmoteNC (k=3)	0.738	0.738	0.963	0.698
MOYSA	SmoteNC yok	0.500	0.500	0.952	0.000
	SmoteNC (k=1)	0.500	0.500	0.952	0.000
	SmoteNC (k=2)	0.500	0.500	0.952	0.000
	SmoteNC (k=3)	0.725	0.725	0.950	0.689

Tablo 2'de öğrencilerin başarı durumlarını tahmin etmek üzere uygulanmış makine öğrenme algoritmalarının performans ölçütleri sonuçları yer almaktadır. Rasgele ormanlar algoritması için SmoteNC (k=2,3) ve güçlendirilmiş lojistik regresyon analizi için SmoteNC (k=1,2) sonuçları, hiç SmoteNC uygulanmamış olan durumlara kıyasla daha düşük performansa sahiptir. Bu durumlar haricindeki genel istatistiksel sonuçlara göre, SmoteNC tekniği ile sınıf dengesizliği giderildiğinde, performans sonuçlarında gözle görülür bir iyileşme sağlanmaktadır. Veri setindeki dengesizliğin giderilmesi sonucunda DDSO ve F-Skoru değerlerinde 0.20'yi aşan bir yükseliş gözlenirken, G-ortalama değerleri de 0'dan yaklaşık 0.690 seviyelerine ulaşmaktadır.

Sınıf dengesizliğinde makine öğrenme algoritmalarının sağladığı en yüksek DDSO değeri 0.725, en yüksek AUC değeri 0.738, en yüksek F-Skoru 0.964 ve en yüksek G-Ortalama değeri 0.698'dir. Algoritmaların genel performansları karşılaştırıldığında, en başarılı sonuçların SmoteNC (k=2) için güçlendirilmiş lojistik regresyon analizi ile elde edildiği görülmektedir.

4. Sonuç ve Değerlendirme

Yapay zekâ teknolojilerindeki muazzam gelişim, makine öğrenimi algoritmalarının eğitim alanında uygulanmasına olanak sunmaktadır. Makine öğrenme algoritmaları sayesinde eğitim alanında öğrencilerin performanslarını tahmin etme, öğrencilerin başarısızlık risklerini önceden belirleme gibi çok değerli bilgiler elde edilmektedir. Bu çalışmada, eğitim alanında da son yıllarda oldukça yaygın olarak kullanılan makine öğrenme algoritmalarına yönelik bulgular ortaya konulmuştur. Uygulama aşamasında beş farklı makine öğrenme algoritması kullanılmış ve mevcut sınıf dengesizliği sorununu gidermek üzere SmoteNC tekniği ile

bağımlı değişken dengeli bir yapıya getirilmiştir. Çalışmamızın dikkat çekici yönlerinden birisi de kategorik veriler için sınıf dengesizliğini giderici SmoteNC tekniğinin eğitim alanında uygulanmasıdır. Eğitim alanındaki sayısal veriler için sentetik veri üretimine dayalı Smote tekniğinin farklı varyantları kullanılmasına karşın, kategorik veriler için alternatif bir Smote yöntemi ile dengesizlik sorunu üzerinde yeterince durulmamıştır. Yaptığımız araştırmalar, Türkçe literatürde eğitim alanında kategorik veriler ile sınıf dengesizliğini gidermek üzere yapılan tek çalışmanın Yılmaz ve diğ. (2023) tarafından yayınlandığını göstermektedir.

Veri analizi uygulamaları sonucunda, makine öğrenme algoritmalarının sınıf dengesizliği durumunda SmoteNC ile başarılı sonuçlar verdiğini ortaya koymaktadır. Özellikle verilerin dengeli hale getirilmesi, farklı komşuluk sayıları için dahi olsa makine öğrenme algoritmalarının performanslarını artırmaktadır. Sınıf dengesizliğini gidermek adına uygulanan SmoteNC tekniği kapsamında farklı komşuluk sayıları, makine öğrenme sonuçlarında da farklı sonuçların elde edilmesine yol açmaktadır. Bu olgu, eğitim alanındaki verileri dengeli hale getirirken gerçekleştirilen veri türetimi aşamasında en uygun komşuluk parametresinin de seçiminin önemine işaret etmektedir.

Çalışmamızın eğitim aşamasında kullanılan veri setinde n=101 gözlem mevcuttur. Ek olarak bağımlı değişkende IR=24,6 gibi büyük oranda bir dengesizlik problemi de bulunmaktadır. Sonuç olarak, elde edilen analiz bulguları, sınırlı sayıda gözlem içeren ve sınıf dengesizliğini büyük ölçüde barındıran bir veri setinde dahi, makine öğrenme teknikleri sayesinde eğitim alanındaki başarı, risk vb. olguların tahminlemeleri yüksek orandan isabetle gerçekleştirilebileceğini göstermektedir.

Araştırma bulguları, eğitim alanında yapılabilecek faaliyetlere de ışık tutmaktadır. Eğitim alanında faaliyet gösteren araştırmacılar, öğrencilere ait tanıtıcı özellikleri kullanarak yeni yazılımlar geliştirebilir ve bu yazılımlar ile öğrenci bazlı akademik başarı tahminleri yapılabilir. Teknolojik olanaklar sayesinde, mobil uygulama gibi pratik araçlarla bile akademik performansa yönelik beklentiler elde edilebilir. Öğrenci bazlı yapılacak tahminler doğrultusunda, öğrencilerin akademik performanslarını artırmak üzere yeni eğitim politikalarının ortaya konulabileceği öngörülmektedir.

Kaynakça

Awad, M., Khanna, R., Awad, M., & Khanna, R. (2015). Support vector machines for classification. Efficient learning machines: Theories, concepts, and applications for engineers and system designers, 39-66.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

Gonzalez, M., Costa, E., & Marques, J. (2022). A two-phase machine learning approach for predicting student outcomes. Educational Data Mining, 14(3), 112-126. https://doi.org/10.1145/3361335.3361345.

Halde, R. R. (2016, September). Application of Machine Learning algorithms for betterment in education system. In 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT) (pp. 1110-1114). IEEE.

He, L., & Tafti, D. K. (2019). A supervised machine learning approach for predicting variable drag forces on spherical particles in suspension. Powder technology, 345, 379-389.a

Hernandez-Leal, P., Hussain, Z., & Dragan, L. (2022). Predicting student performance in a blended learning environment using machine learning techniques. Computers & Education, 176, 104360. https://doi.org/10.1016/j.compedu.2021.104360.

Islahulhaq, W. W., & Ratih, I. D. (2021). Classification of non-performing financing using logistic regression and synthetic minority over-sampling technique-nominal continuous (SMOTE-NC). Int. J. Adv. Soft Comput. Appl, 13, 115-128.

Jones, P., Williams, K., & Thomas, L. (2023). A systematic review of the literature on machine learning application in predicting student academic performance. Decision Analytics Journal, 7, 100204. https://

doi.org/10.1016/j.daj.2023.100204.

Karbasi, M., Bahrami, S., Salehi, M., & Alizadeh, H. (2021). Predicting academic success in higher education: Literature review and best practices. International Journal of Educational Technology in Higher Education, 18(1), 1-14. https://doi.org/10.1186/s41239-021-00278-6.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05.

Lopez, V., Luna, J. M., & Romero, C. (2021). Enhancing prediction of student success: Automated machine learning approaches. Information Fusion, 65, 52-60. https://doi.org/10.1016/j.inffus.2020.07.009.

Menon, H. K. D., & Janardhan, V. (2021). Machine learning approaches in education. Materials Today: Proceedings, 43, 3470-3480.

Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. Applied System Innovation, 4(1), 18.

Orji, F. A., & Vassileva, J. (2022). Machine learning approach for predicting students academic performance and study strategies based on their motivation. arXiv. Published online October 15, 2022. https://doi.org/10.48550/arXiv.2210.08186.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a Stata plugin. The Stata Journal, 5(3), 330-354.

Shen, S. L., Zhang, N., Zhou, A., & Yin, Z. Y. (2022). Enhancement of neural networks with an alternative activation function tanhLU. Expert Systems with Applications, 199, 117181.

Smith, J., Doe, J., & Brown, A. (2021). Predicting academic success of college students using machine learning algorithms. Journal of Educational Computing Research, 59(4), 671-690. https://doi.org/10.1177/07356331211012345.

Şengür, D. (2013). Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları ile Tahmini. Fırat Üniversitesi, Eğitim Bilimleri Enstitüsü, Doktora Tezi.

Wang, Q., & Zhang, L. (2012). Least squares online linear discriminant analysis. Expert Systems with Applications, 39(1), 1510-1517.

Wang, Y., Yu, Y., & Hu, Y. (2020). Supervised machine learning algorithms for predicting student dropout and academic success. Education Sciences, 10(5), 134. https://doi.org/10.3390/educsci10050134.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. Information, 14(1), 54.

Yan, Y. (2016). MLmetrics: Machine learning evaluation metrics (R package version 1.1.1). Retrieved from https://CRAN.R-project.org/package=MLmetrics

Yılmaz, N., & Şekeroğlu, B. (2019, August). Student performance classification using artificial intelligence techniques. In International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions (pp. 596-603). Cham: Springer International Publishing.

Yılmaz, E., Altıkardeş, Z. A., & Erdal, H. (2023). Higher Education Planning and Decision Support System with Multi-Class and Imbalanced Educational Dataset: A Case Of Technology Faculty. Gazi Journal of Engineering Sciences (GJES), 9(1).

Yağci, M., Rebai, I., & Eltahir, M. (2020). Role of convolutional features and machine learning for predicting student academic performance from MOODLE data. PLOS ONE, 15(10). https://doi.org/10.1371/journal.pone.0240991.ms.