



# Eğitim Sektöründe Öğrenci Terkinin Tahmini: Makine Öğrenmesi Tabanlı Karşılaştırmalı Bir Analiz

*Predicting Student Dropout in the Education Sector: A Comparative Machine Learning Analysis*

**Muhammed Akif Yenikaya**<sup>1</sup>

<sup>1</sup>Dr. Öğretim Üyesi, Kafkas Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü, m.akifykaya@mail.com, Orcid ID: 0000-0002-3624-722X

## MAKALE BİLGİSİ

### Anahtar Kelimeler

Öğrenci terki,  
Makine öğrenmesi,  
Tahmin modelleri,

### Makale Geçmişi:

Geliş Tarihi: 26 Ocak 2026  
Kabul Tarihi: 1 Mart 2026

## ARTICLE INFO

### Keywords

Student dropout,  
Machine learning,  
Predictive models

### Article History:

Received: 26 January 2026  
Accepted: 1 March 2026

## ÖZET

Yükseköğretimde öğrencilerin akademik başarısızlık yaşamaları ve okuldan ayrılmaları, bireysel eğitim süreçlerinin yanı sıra kurumların akademik etkinliği ve ekonomik sürdürülebilirliği üzerinde etkili olan küresel sorunlar arasında yer almaktadır. Bu çalışma, öğrencilerin akademik başarı durumları ile okul terk etme eğilimlerinin tahmin edilmesinde kullanılan farklı makine öğrenmesi algoritmalarının performanslarını karşılaştırmalı olarak incelemeyi amaçlamaktadır. Araştırmada Kaggle platformundan temin edilen açık kaynaklı bir veri seti kullanılmış olup veri seti 4.424 gözlem ve 36 değişkenden oluşmaktadır. Lojistik regresyon, karar ağacı, rastgele orman ve XGBoost algoritmaları kullanılarak sınıflandırma modelleri geliştirilmiştir. Modellerin performansı doğruluk, kesinlik, duyarlılık ve F1-skoru ölçütleri temelinde değerlendirilmiştir. Bulgular, lojistik regresyon modelinin %91 doğruluk ve %90 F1-skoru ile güçlü bir temel performans sunduğunu göstermektedir. Karar ağacı modeli %90,63 doğruluk ve %89,37 F1-skoru ile yorumlanabilir olmakla birlikte görece daha sınırlı bir başarı sergilemiştir. Buna karşılık, topluluk öğrenmesine dayalı rastgele orman ve XGBoost modelleri %92 doğruluk ve %92 düzeyinde ağırlıklı F1-skoru değerlerine ulaşarak daha dengeli sonuçlar ortaya koymuştur. Özellikle bu modellerin yüksek duyarlılık değerleri, akademik risk altındaki öğrencilerin erken aşamada belirlenmesine olanak sağlamaktadır. Çalışma sonuçları, makine öğrenmesine dayalı karşılaştırmalı yaklaşımların yükseköğretimde erken uyarı sistemleri ve veri temelli karar destek mekanizmalarının geliştirilmesine katkı sağlayabileceğini göstermektedir.

## ABSTRACT

In higher education, student academic failure and dropout are global challenges that affect individual educational trajectories as well as the academic effectiveness and economic sustainability of institutions. This study aims to comparatively examine the performance of different machine learning algorithms used to predict students' academic achievement and dropout tendencies. An open-access dataset obtained from the Kaggle platform was used, consisting of 4,424 observations and 36 variables. Classification models were developed using logistic regression, decision tree, random forest, and XGBoost algorithms. Model performance was evaluated based on accuracy, precision, recall, and F1-score metrics. The results indicate that the logistic regression model provides a strong baseline with 91% accuracy and a 90% F1-score. The decision tree model achieved 90.63% accuracy and an 89.37% F1-score, offering interpretability but relatively limited performance. In contrast, ensemble learning-based random forest and XGBoost models demonstrated superior and more balanced results, each achieving 92% accuracy and a weighted F1-score of 92%. Notably, the high recall values of these models provide a clear advantage in identifying students at academic risk early. Overall, the findings suggest that machine learning-based comparative approaches can contribute to the development of early warning systems and data-driven decision support mechanisms in higher education.

Yükseköğretimde öğrenci başarısızlığı ve okul terki, son yıllarda eğitim politikalarının ve akademik araştırmaların odağında yer alan yükseköğretim sistemlerinin sürdürülebilirliği açısından kritik bir sorun alanı olarak öne çıkmaktadır. Öğrencilerin eğitim süreçlerini tamamlayamadan sistem dışına çıkmaları, bireysel düzeyde akademik ve mesleki fırsatların azalmasına yol açarken, kurumsal düzeyde eğitim kalitesinin düşmesi ve kaynakların verimsiz kullanılması gibi sonuçlar doğurmaktadır. Nitekim uluslararası raporlar, birçok ülkede yükseköğretimde okul terk oranlarının hâlen kayda değer düzeylerde seyrettiğini ve bu durumun uzun vadeli sosyal ve ekonomik etkiler ürettiğini ortaya koymaktadır (OECD, 2022).

Öğrenci başarısızlığı ve okul terkine yol açan faktörlerin birden fazla bireysel ve kurumsal değişkenin eşzamanlı etkisiyle şekillendiği birçok çalışmada ortaya konulmuştur. Akademik performans göstergeleri, öğrencinin sosyo-ekonomik arka planı, aileye ilişkin değişkenler ve eğitim sürecine dair kurumsal faktörler bu karmaşık yapının temel bileşenleri arasında yer almaktadır (Aina vd., 2022). Özellikle ilk yıllarda yaşanan akademik uyum sorunları ve düşük başarı düzeyleri, öğrencilerin eğitimden kopma olasılığını artıran kritik risk unsurları olarak değerlendirilmektedir (Bargmann vd., 2022). Bu nedenle, risk altındaki öğrencilerin erken aşamada belirlenmesi hem önleyici müdahalelerin etkinliği hem de eğitim sistemlerinin sürdürülebilirliği açısından önemli bir gereklilik olarak öne çıkmaktadır.

Geleneksel istatistiksel yöntemler, öğrenci başarısızlığı ve okul terki gibi karmaşık olguları açıklamada belirli katkılar sunmakla birlikte, değişkenler arasındaki doğrusal olmayan ilişkileri yakalama konusunda sınırlılıklar barındırmaktadır. Bu durum, özellikle büyük ölçekli ve çok değişkenli eğitim verilerinin analizinde daha esnek ve güçlü yaklaşımlara duyulan ihtiyacı artırmıştır. Bu bağlamda, makine öğrenmesi yöntemleri, farklı veri türlerini aynı anda işleyebilme, karmaşık örüntüleri ortaya çıkarabilme ve özellikle erken risk belirleme süreçlerinde daha yüksek tahmin performansı sunabilmeleri nedeniyle eğitim alanında giderek daha yaygın biçimde kullanılmaktadır (Namoun & Alshançiti, 2020; Albreiki vd., 2021).

Son yıllarda yapılan çalışmalar, makine öğrenmesi tabanlı modellerin öğrenci başarısı ve okul terki tahmininde anlamlı sonuçlar üretebildiğini göstermektedir. Bununla birlikte, literatürdeki birçok araştırmanın tek bir model veya sınırlı sayıda algoritma üzerinden değerlendirme yaptığı, farklı yaklaşımların aynı deneysel koşullar altında karşılaştırılmasına yeterince odaklanmadığı görülmektedir (Rastrollo-Guerrero vd., 2020). Ayrıca, bazı çalışmalarda yüksek doğruluk değerlerine ulaşılmasına rağmen, eğitimsel karar verme açısından kritik öneme sahip olan duyarlılık ve hata türlerinin yeterince tartışılmadığı dikkat çekmektedir.

Bu çalışmanın amacı, yükseköğretimde öğrenci başarısı ve okul terk eğiliminin tahmin edilmesinde farklı makine öğrenmesi algoritmalarının karşılaştırmalı performanslarını incelemektir. Çalışmanın temel katkısı, farklı modelleme yaklaşımlarının aynı veri bölünmesi ve değerlendirme ölçütleri altında karşılaştırılarak, eğitim kurumları için erken uyarı ve veri temelli karar destek sistemlerine yönelik daha bütüncül bir bakış açısı sunmasıdır.

## 1. LİTERATÜR ÇALIŞMASI

Öğrenci başarısı ve okul terki olgusu, yükseköğretim sistemlerinin etkinliğini doğrudan etkileyen temel sorun alanlarından biri olarak ele alınmaktadır. Erken dönem çalışmalarda okul terki, ağırlıklı olarak öğrencilerin akademik entegrasyonu ve kurumsal uyum düzeyi çerçevesinde açıklanmıştır. Bu bağlamda Tinto'nun (1975; 2012) bütünlük kuramı, öğrencilerin akademik ve sosyal ortama uyum sağlayamamalarının okul terkinin tetikleyen temel unsurlar arasında yer aldığını ortaya koymuştur. Benzer şekilde Bean ve Metzner (1985), akademik performansın yanı sıra motivasyon, sosyal destek ve çevresel faktörlerin öğrencilerin eğitim sürecinde kalıcılığı üzerinde belirleyici olduğunu vurgulamıştır.

Son yıllarda yapılan çalışmalar, okul terkinin yalnızca bireysel akademik performansla açıklanamayacağını, öğrencilerin demografik özellikleri, öğrenme süreçlerine katılım düzeyleri, dijital öğrenme davranışları ve çevresel koşulların birlikte değerlendirilmesi gerektiğini ortaya koymaktadır. Bu çok boyutlu yapı, öğrenci başarısının öngörülmesinde daha kapsamlı ve bütüncül analitik yaklaşımlara duyulan ihtiyacı artırmıştır (Khan & Ghosh, 2021). Kurumsal yapı, öğretim kalitesi ve öğrenci destek mekanizmalarının da sürece dâhil edilmesi gerektiği, daha güncel çalışmalarda sıklıkla vurgulanmaktadır (Braxton vd., 2013).

Eğitim alanında üretilen veri miktarının ve çeşitliliğinin artmasıyla birlikte, öğrenci performansını ve okul terk riskini tahmin etmeye yönelik makine öğrenmesi temelli yöntemler giderek daha fazla benimsenmektedir. Sistemik derleme ve meta-analiz çalışmalarında, makine öğrenmesi algoritmalarının geleneksel istatistiksel yöntemlere kıyasla daha yüksek tahmin doğruluğu

sunduğu ve özellikle karmaşık, doğrusal olmayan veri yapılarında daha başarılı sonuçlar ürettiği belirtilmektedir (Andrade-Girón vd., 2023; Romero & Ventura, 2020). Bu çalışmalar, erken risk tespiti açısından makine öğrenmesi modellerinin önemli bir potansiyel taşıdığını göstermektedir.

Literatürde öne çıkan araştırmalar, öğrenci performansı ve okul terki tahmininde farklı algoritmaların değişen başarı düzeyleri sunduğunu ortaya koymaktadır. Pek vd. (2022), farklı makine öğrenmesi yaklaşımlarını karşılaştırdıkları çalışmalarında, topluluk öğrenme temelli modellerin tekil algoritmalara kıyasla daha istikrarlı ve genellenebilir performans sergilediğini belirtmiştir. Benzer şekilde, Huo vd. (2023), doğrusal ve geleneksel modellere kıyasla ağaç tabanlı ve artırmalı yöntemlerin, özellikle akademik risk altındaki öğrenci gruplarının belirlenmesinde daha etkili olduğunu rapor etmiştir. Rastgele orman ve gradyan artırma temelli modellerin, değişkenler arasındaki karmaşık ve doğrusal olmayan ilişkileri yakalama konusunda üstünlük sağladığı da literatürde sıklıkla vurgulanmaktadır (Breiman, 2001; Chen & Guestrin, 2016).

Bununla birlikte, literatürdeki çalışmaların önemli bir bölümünde metodolojik çeşitliliğin sınırlı olduğu dikkat çekmektedir. Bazı araştırmalar tek bir algoritma üzerinden değerlendirme yaparken, bazı çalışmalarda modellerin farklı veri bölünmeleri, dengesiz sınıf yapıları ve sınırlı performans ölçütleri altında test edildiği görülmektedir. Bu durum, model performanslarının doğrudan karşılaştırılmasını güçleştirmekte ve elde edilen bulguların genellenebilirliğini sınırlamaktadır (López-Zambrano vd., 2021). Ayrıca, yüksek doğruluk oranlarının raporlandığı çalışmalarda dahi, eğitimsel karar verme açısından kritik öneme sahip olan yanlış sınıflandırma türlerinin, özellikle de akademik risk altındaki öğrencilerin gözden kaçırılmasına yol açabilecek düşük duyarlılık değerlerinin yeterince tartışılmadığı görülmektedir. Bu bağlamda, çoklu algoritmaların ve farklı performans ölçütlerinin birlikte değerlendirildiği karşılaştırmalı çalışmaların, literatürde önemli bir boşluğu doldurma potansiyeline sahip olduğu söylenebilmektedir.

Mevcut literatür genel olarak değerlendirildiğinde, makine öğrenmesi temelli yaklaşımların öğrenci başarısı ve okul terki tahmininde güçlü bir analitik araç sunduğu açıkça görülmektedir. Ancak, farklı algoritmaların aynı deneysel koşullar altında karşılaştırıldığı, tutarlı veri bölünmeleri ve ortak değerlendirme ölçütleri kullanılan çalışmaların görece sınırlı olduğu anlaşılmaktadır. Bu durum, karşılaştırmalı ve bütüncül analizlere dayalı yeni çalışmalara duyulan ihtiyacı ortaya koymaktadır. Bu çalışma, söz konusu boşluğu doldurmayı hedefleyerek, farklı makine öğrenmesi algoritmalarının öğrenci başarısı ve okul terki tahminindeki görece performanslarını sistematik bir çerçevede incelemektedir.

## 2. MATERYAL VE METOT

### 2.1. Araştırma Tasarımı

Bu çalışma, yükseköğretimde öğrenci başarısı ve okul terk eğiliminin tahmin edilmesine yönelik olarak karşılaştırmalı bir makine öğrenmesi yaklaşımı benimsemektedir. Araştırma, sınıflandırma problemi çerçevesinde kurgulanmıştır. Farklı makine öğrenmesi algoritmalarının aynı veri seti ve aynı deneysel koşullar altında değerlendirilmesi amaçlanmıştır. Böylece modellerin görece performanslarının nesnel ve tutarlı biçimde karşılaştırılması hedeflenmiştir.

### 2.2. Veri Seti

Araştırmada, Kaggle platformu üzerinden erişilebilen açık kaynaklı bir veri seti kullanılmıştır (Realinho vd., 2021; Kaggle, 2025). Veri seti, yükseköğretim öğrencilerine ait akademik, demografik ve sosyo-ekonomik bilgileri içermekte olup toplam 4.424 gözlem ve 36 değişkenden oluşmaktadır. Hedef değişken, öğrencilerin akademik süreçlerini başarıyla tamamlamaları ya da okul terki yaşamaları durumunu yansıtan ikili bir sınıflandırma yapısına sahiptir. Veri setinin açık kaynaklı olması, araştırmanın şeffaflığı ve yeniden üretilebilirliği açısından önemli bir avantaj sağlamaktadır.

### 2.3. Veri Ön İşleme Süreci

Veri seti üzerinde yapılan ön incelemelerde eksik ya da tutarsız gözlemlere rastlanmamış olup, veri bütünlüğünün korunduğu belirlenmiştir. Modelleme sürecinde değişkenlerin farklı ölçeklerde değerlere sahip olması nedeniyle, tüm sayısal değişkenler için standartlaştırma işlemi uygulanmıştır. Bu kapsamda, ortalaması sıfır ve standart sapması bir olacak şekilde ölçeklendirme yapılmıştır. Ön işleme adımları, model performanslarının değişken ölçeklerinden etkilenmesini önlemek ve algoritmaların öğrenme sürecini daha dengeli hâle getirmek amacıyla gerçekleştirilmiştir.

## 2.4. Eğitim ve Test Veri Ayrımı

Makine öğrenmesi modellerinin performanslarının güvenilir biçimde değerlendirilebilmesi için veri setinin uygun oranlarda eğitim ve test alt kümelerine ayrılması kritik bir aşamadır. Bu çalışmada veri setinin %80'inin eğitim, %20'sinin ise test amacıyla kullanılması tercih edilmiştir. Bu oran, hem modelin yeterli miktarda veri üzerinden öğrenme gerçekleştirmesine olanak tanımakta hem de bağımsız test verisi üzerinden genellenebilirliğinin sağlıklı biçimde ölçülmesini mümkün kılmaktadır.

Literatürde öğrenci başarısı ve okul terki tahminine yönelik yapılan çalışmalarda, %80-20 oranındaki veri bölünmesinin dengeli ve etkili sonuçlar sunduğu görülmektedir. Nitekim, öğrenci başarısı ve okul terki tahminine odaklanan bir çalışmada, farklı eğitim-test oranları karşılaştırılmış ve %80 eğitim ve %20 test ayrımının doğruluk, kesinlik ve F1-skoru açısından en dengeli performansı sağladığı rapor edilmiştir. Söz konusu çalışmada, daha düşük eğitim oranlarında model öğrenme kapasitesinin sınırlı kaldığı; daha yüksek eğitim oranlarında ise test verisinin küçülmesine bağlı olarak performans değerlendirmesinin güvenilirliğinin azaldığı vurgulanmıştır (Ridwan vd., 2024).

Bu bulgular doğrultusunda, mevcut çalışmada da %80-20 veri ayrımı tercih edilerek, modellerin hem yeterli öğrenme sürecinden geçmesi hem de test aşamasında istikrarlı ve karşılaştırılabilir sonuçlar üretmesi amaçlanmıştır. Ayrıca, tüm makine öğrenmesi algoritmaları için aynı veri bölünmesinin kullanılması, deneysel tutarlılığın korunmasına ve modellerin görelî performanslarının nesnel biçimde karşılaştırılmasına olanak sağlamıştır. Bu yaklaşım, elde edilen sonuçların eğitimsel karar destek süreçlerinde daha güvenilir biçimde yorumlanmasına katkı sunmaktadır.

Veri bölünmesi sürecinde sınıflar arasındaki dağılımın korunması amacıyla stratified sampling yaklaşımı benimsenmiştir. Böylece hedef değişkenin eğitim ve test alt kümelerinde benzer oranlarda temsil edilmesi sağlanmıştır. Ayrıca modellere rastgelelikten kaynaklanabilecek farklılıkları kontrol altına almak ve sonuçların yeniden üretilebilirliğini sağlamak amacıyla sabit bir random seed değeri kullanılmıştır. Bu yaklaşım, elde edilen performans ölçütlerinin tutarlılığını ve deneysel şeffaflığı güçlendirmektedir.

## 2.5. Kullanılan Makine Öğrenmesi Modelleri

Bu çalışmada kullanılan ilk model lojistik regresyondur. Lojistik regresyon, ikili sonuç değişkenlerinin modellenmesinde uzun süredir kullanılan temel istatistiksel yaklaşımlardan biridir. Model, bağımsız değişkenler ile sonuç değişkeni arasındaki ilişkiyi olasılık temelli bir yapı üzerinden ele almakta ve katsayılar aracılığıyla yorumlanabilir sonuçlar sunmaktadır. Bu yönüyle lojistik regresyon, özellikle sosyal bilimlerde ve eğitim araştırmalarında yaygın biçimde temel karşılaştırma modeli olarak kullanılmakta ve daha karmaşık yöntemlerin performansını değerlendirmek için bir referans noktası oluşturmaktadır (Hosmer vd., 2013).

Çalışmada yer verilen karar ağacı modeli, verideki örüntüleri ardışık karar kuralları üzerinden açıklayabilmesi ve doğrusal olmayan ilişkileri yakalayabilmesi nedeniyle tercih edilmiştir. Karar ağaçlarının görsel olarak anlaşılabilir yapısı, özellikle eğitim bağlamında elde edilen sonuçların yorumlanmasını kolaylaştırmaktadır (Quinlan, 2014).

Rastgele orman yöntemi, birden fazla karar ağacının bir araya getirilmesiyle oluşturulan bir topluluk öğrenme yaklaşımıdır. Bu yapı, tekil karar ağaçlarının sınırlılıklarını azaltmakta ve daha kararlı tahminler elde edilmesini sağlamaktadır. Rastgele orman modelleri, özellikle yüksek boyutlu ve karmaşık veri setlerinde güçlü ve genellenebilir sonuçlar sunmaktadır (Breiman, 2001).

Çalışmada kullanılan ileri seviye yöntemlerden biri olan XGBoost ise gradyan artırma temelli öğrenme yaklaşımına dayanmaktadır. Model, önceki tahminlerde yapılan hataları kademeli olarak azaltmayı amaçlayan yapısı sayesinde yüksek tahmin başarısı sunmaktadır. Gradyan artırma yöntemlerinin, özellikle sınıflandırma problemlerinde güçlü performans sergilediği ve son yıllarda eğitim alanındaki tahmin çalışmalarında yaygın biçimde kullanıldığı literatürde vurgulanmaktadır (Friedman, 2001).

## 2.6. Model Hiperparametreleri ve Eğitim Süreci

Bu çalışmada kullanılan tüm modellerde hiperparametre seçimi eğitim verisi üzerinden yapılmış ve rastgelelik kontrolü sağlanarak analiz süreci tekrarlanabilir hâle getirilmiştir. Veri bölünmesinde sınıf dengesini korumak amacıyla tabakalı çapraz doğrulama tercih edilmiştir.

Lojistik regresyon modeli için hiperparametre araması Randomized Search yöntemi ile gerçekleştirilmiştir. Düzenleştirme katsayısını ifade eden C parametresi logaritmik ölçekte geniş bir aralıkta denenmiştir. Ayrıca penalty (l1, l2), solver (liblinear, saga) ve class\_weight (None, balanced) seçenekleri test edilmiştir. Maksimum iterasyon sayısı 2000 olarak belirlenmiş ve model seçimi F1-skoru temel alınarak yapılmıştır. Çapraz doğrulama 3 katlı ve karıştırmalı (shuffle=True) olarak uygulanmıştır.

Karar ağacı modeli için Grid Search kullanılmıştır. Splitter (best, random), max\_depth (1–100), min\_samples\_leaf (1–4) ve min\_samples\_split (2–4) parametreleri denenmiştir. Model seçimi doğruluk (accuracy) ölçütüne göre yapılmış ve en iyi parametre kombinasyonu 5 katlı çapraz doğrulama ile belirlenmiştir.

Rastgele orman modeli için yine 5 katlı çapraz doğrulama ile Grid Search uygulanmıştır. Bu modelde yalnızca ağaç sayısını ifade eden n\_estimators parametresi (100–400 aralığında) optimize edilmiştir. Diğer parametreler varsayılan değerlerinde bırakılmıştır.

XGBoost modeli için hiperparametre araması Randomized Search yöntemiyle gerçekleştirilmiştir. Min\_child\_weight, gamma, subsample, colsample\_bytree ve max\_depth parametreleri belirli aralıklarda test edilmiştir. Öğrenme oranı 0.02 ve ağaç sayısı 600 olarak sabit tutulmuştur. Model seçimi ROC-AUC ölçütüne göre yapılmış ve 3 katlı tabakalı çapraz doğrulama kullanılmıştır.

Tüm modellerde rastgelelik kontrolü için sabit tohum değeri kullanılmış ve nihai performans değerlendirmesi test verisi üzerinde gerçekleştirilmiştir.

## 2.7. Değerlendirme Ölçütleri

Bu çalışmada geliştirilen makine öğrenmesi modellerinin performansları, sınıflandırma problemleri için yaygın olarak kullanılan dört temel ölçüt üzerinden değerlendirilmiştir: doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1-skoru. Bu ölçütler, yalnızca genel sınıflandırma başarısını değil, aynı zamanda yanlış sınıflandırmaların eğitimsel karar süreçleri üzerindeki etkilerini de ortaya koymaktadır.

Bu ölçütlerin hesaplanmasında kullanılan temel kavramlar şu şekilde tanımlanmaktadır:

- TP (True Positive / Doğru Pozitif): Gerçek sınıfı pozitif olan bir gözlemin model tarafından doğru şekilde pozitif olarak sınıflandırılması.
- TN (True Negative / Doğru Negatif): Gerçek sınıfı negatif olan bir gözlemin model tarafından doğru şekilde negatif olarak sınıflandırılması.
- FP (False Positive / Yanlış Pozitif): Gerçek sınıfı negatif olan bir gözlemin model tarafından hatalı biçimde pozitif olarak sınıflandırılması.
- FN (False Negative / Yanlış Negatif): Gerçek sınıfı pozitif olan bir gözlemin model tarafından hatalı biçimde negatif olarak sınıflandırılması.

Eğitim sektörü bağlamında değerlendirildiğinde, bu hata türleri yalnızca istatistiksel birer gösterge değil, aynı zamanda doğrudan yönetsel ve pedagojik sonuçlar doğuran karar hatalarıdır. Özellikle FN (yanlış negatif) durumları, risk altındaki bir öğrencinin model tarafından tespit edilememesi anlamına geldiğinden, erken müdahale mekanizmalarının devreye girmesini engelleyerek eğitimsel açıdan daha yüksek maliyetler doğurabilmektedir. Buna karşılık, FP (yanlış pozitif) durumları ise risk taşımayan öğrencilerin gereksiz müdahale süreçlerine dâhil edilmesine yol açabilmektedir.

Doğruluk (accuracy), model tarafından doğru şekilde sınıflandırılan gözlemlerin toplam gözlem sayısına oranını ifade etmektedir:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Kesinlik (precision), modelin pozitif olarak sınıflandırdığı gözlemlerin ne kadarının gerçekten pozitif sınıfa ait olduğunu göstermektedir:

$$Kesinlik = \frac{TP}{TP + FP}$$

Duyarlılık (recall), modelin gerçek pozitif gözlemleri ne ölçüde doğru biçimde tespit edebildiğini ifade etmektedir:

$$Duyarlılık = \frac{TP}{TP + FN}$$

F1-skoru ise kesinlik ve duyarlılık ölçütlerinin harmonik ortalamasını temsil etmekte ve bu iki ölçüt arasındaki dengeyi yansıtmaktadır:

$$f1 - skoru = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık}$$

Bu ölçütler birlikte değerlendirildiğinde, modellerin yalnızca genel başarı düzeyleri değil, aynı zamanda eğitim kurumları açısından kritik öneme sahip olan hata türlerine karşı duyarlılıkları da kapsamlı biçimde analiz edilebilmektedir. Böylece geliştirilen modellerin, öğrenci başarısı ve okul terk riskinin tahmininde ne ölçüde güvenilir, dengeli ve karar destek süreçlerine uygun çıktılar ürettiği bütüncül olarak ortaya konulmaktadır.

### 3. BULGULAR

Bu çalışmada, öğrencilerin akademik başarı durumları ile okul terk riskinin tahmin edilmesine yönelik olarak çalışmada kullanılan makine öğrenmesi modellerinin performansları karşılaştırmalı biçimde değerlendirilmiştir.

Çalışmada kullanılan veri setinin orijinal yapısında hedef değişken üç sınıftan oluşmaktadır: Graduate (mezun), Enrolled (eğitime devam eden) ve Dropout (okul terk eden). Bu çalışmada, öğrencilerin okul terk riskinin tahmin edilmesine odaklanılması nedeniyle hedef değişken ikili sınıflandırma problemine dönüştürülmüştür. Bu kapsamda, Dropout (okul terk eden) kategorisi 1 olarak kodlanmış, Graduate (mezun) ve Enrolled (eğitime devam eden) kategorileri ise öğrencinin eğitim sürecini sürdürdüğünü veya tamamladığını ifade edecek şekilde 0 sınıfı altında birleştirilmiştir. Model performans ölçütleri, bu ikili sınıflandırma yapısı esas alınarak hesaplanmıştır.

Tüm modeller, veri setinin %80'inin eğitim ve %20'sinin test amacıyla ayrıldığı ortak bir deneysel tasarım çerçevesinde eğitilmiş ve test edilmiştir. Model performansları doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1-skoru ölçütleri üzerinden analiz edilmiştir.

#### 3.1. Lojistik Regresyon Modeli Bulguları

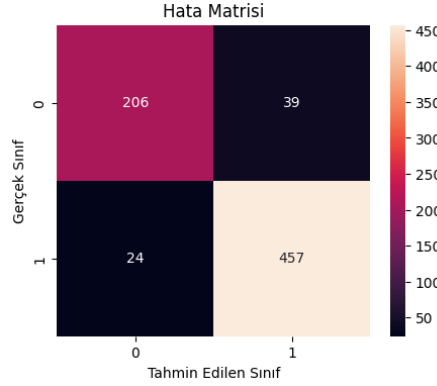
Bu bölümde, lojistik regresyon modeli kullanılarak elde edilen sınıflandırma performans sonuçları sunulmaktadır. Modele ait sınıf bazlı performans ölçütleri Tablo 1'de yer almaktadır.

**Tablo 1.** Lojistik Regresyon Modeline Ait Sınıf Bazlı Performans Ölçütleri

Sınıf	Kesinlik (Precision)	Duyarlılık (Recall)	F1-Skoru
0	0.90	0.84	0.87
1	0.92	0.95	0.94
<b>Doğruluk (Accuracy)</b>			<b>0.91</b>
<b>Makro Ortalama</b>	0.91	0.90	0.90
<b>Ağırlıklı Ortalama</b>	0.91	0.91	0.91

Lojistik regresyon modeli, test verisi üzerinde %91 doğruluk oranına ulaşmıştır. Sınıf bazlı sonuçlar incelendiğinde, birinci sınıf için yüksek kesinlik ve duyarlılık değerleri elde edildiği, ikinci sınıf için ise özellikle duyarlılık değerinin öne çıktığı görülmektedir. Makro ve ağırlıklı ortalama F1-skorlarının %90 düzeyinde gerçekleşmesi, modelin sınıflar arası dağılımdan sınırlı ölçüde etkilendiğini göstermektedir.

Lojistik regresyon modelinin sınıflandırma performansı, hata matrisi kullanılarak ayrıntılı biçimde incelenmiştir. Hata matrisi, modelin gerçek sınıf etiketleri ile tahmin edilen sınıflar arasındaki ilişkiyi göstermekte ve doğru ile hatalı sınıflandırmaların dağılımını ortaya koymaktadır. Şekil 1’de lojistik regresyon modeline ait hata matrisi sunulmaktadır.



Şekil 1. Lojistik Regresyon Modeline Ait Hata Matrisi

Bu matrise göre:

- 206 gözlem, gerçek sınıfı 0 olan öğrenciler arasından doğru şekilde 0 olarak sınıflandırılmıştır (TN – Doğru Negatif).
- 39 gözlem, gerçek sınıfı 0 olmasına rağmen 1 olarak sınıflandırılmıştır (FP – Yanlış Pozitif).
- 24 gözlem, gerçek sınıfı 1 olmasına rağmen 0 olarak sınıflandırılmıştır (FN – Yanlış Negatif).
- 457 gözlem, gerçek sınıfı 1 olan öğrenciler arasından doğru şekilde 1 olarak sınıflandırılmıştır (TP – Doğru Pozitif).

Bu sonuçlar, lojistik regresyon modelinin özellikle 1 sınıfına (Dropout – okul terk riski taşıyan öğrenciler) ait gözlemleri yüksek doğrulukla tanımladığını; buna karşılık sınırlı sayıda gözlemin yanlış negatif ve yanlış pozitif olarak sınıflandırıldığını göstermektedir. Hata matrisinde elde edilen bu dağılım, önceki bölümde sunulan doğruluk, kesinlik, duyarlılık ve F1-skoru değerleri ile tutarlılık göstermektedir.

### 3.2. Karar Ağacı Modeli Bulguları

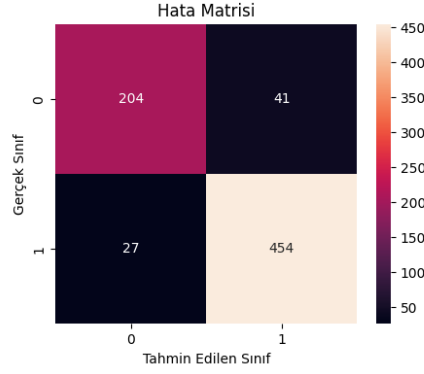
Bu bölümde, karar ağacı modeli kullanılarak elde edilen sınıflandırma performans sonuçları sunulmaktadır. Modele ait sınıf bazlı performans ölçütleri Tablo 2’de yer almaktadır.

Tablo 2. Karar Ağacı Modeline Ait Sınıf Bazlı Performans Ölçütleri

Sınıf	Kesinlik (Precision)	Duyarlılık (Recall)	F1-Skoru
0	0.8831	0.8327	0.8571
1	0.9172	0.9439	0.9303
<b>Doğruluk (Accuracy)</b>			<b>0.9063</b>
<b>Makro Ortalama</b>	0.9001	0.8883	0.8937
<b>Ağırlıklı Ortalama</b>	0.9057	0.9063	0.9056

Karar ağacı modeli, %90,63 doğruluk oranı ile lojistik regresyon modeline yakın bir performans sergilemiştir. Sınıf bazlı ölçütler değerlendirildiğinde, özellikle birinci sınıfa ilişkin duyarlılık değerinin yüksek olduğu, buna karşılık diğer sınıf için performansın görece daha düşük kaldığı görülmektedir. Makro ortalama F1-skorunun %89,37 olması, modelin sınıflar arasında tam anlamıyla dengeli bir performans sunmadığını ortaya koymaktadır.

Karar ağacı modelinin sınıflandırma performansı, hata matrisi kullanılarak ayrıntılı biçimde analiz edilmiştir. Hata matrisi, model tarafından yapılan doğru ve hatalı sınıflandırmaların dağılımını göstermekte olup, gerçek sınıf etiketleri ile tahmin edilen sınıflar arasındaki ilişkiyi ortaya koymaktadır. Karar ağacı modeline ait hata matrisi Şekil 2’de sunulmaktadır.



Şekil 2. Karar Ağacı Modeline Ait Hata Matrisi

Hata matrisi sonuçlarına göre:

- 204 gözlem, gerçek sınıfı 0 olan öğrenciler arasından doğru şekilde 0 olarak sınıflandırılmıştır (TN – Doğru Negatif).
- 41 gözlem, gerçek sınıfı 0 olmasına rağmen 1 olarak sınıflandırılmıştır (FP – Yanlış Pozitif).
- 27 gözlem, gerçek sınıfı 1 olmasına rağmen 0 olarak sınıflandırılmıştır (FN – Yanlış Negatif).
- 454 gözlem, gerçek sınıfı 1 olan öğrenciler arasından doğru şekilde 1 olarak sınıflandırılmıştır (TP – Doğru Pozitif).

Bu dağılım, karar ağacı modelinin özellikle 1 sınıfına (Dropout – okul terk riski taşıyan öğrenciler) ait gözlemleri yüksek oranda doğru sınıflandırdığını, buna karşılık sınırlı sayıda gözlemin yanlış pozitif ve yanlış negatif olarak sınıflandırıldığını göstermektedir. Hata matrisi üzerinden elde edilen bu sonuçlar, önceki tabloda sunulan doğruluk, kesinlik, duyarlılık ve F1-skoru değerleri ile tutarlılık göstermektedir.

### 3.3. Rastgele Orman Modeli Bulguları

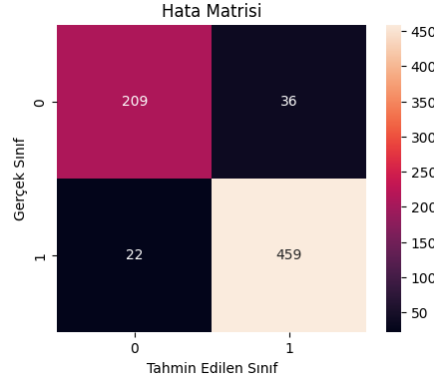
Bu bölümde, rastgele orman modeli kullanılarak elde edilen sınıflandırma performans sonuçları sunulmaktadır. Rastgele orman modeline ait sınıf bazlı performans ölçütleri ile genel başarı değerleri Tablo 3'te yer almaktadır.

Tablo 3. Rastgele Orman Modeline Ait Sınıf Bazlı Performans Ölçütleri

Sınıf	Kesinlik (Precision)	Duyarlılık (Recall)	F1-Skoru
0	0.90	0.85	0.88
1	0.93	0.95	0.94
<b>Doğruluk (Accuracy)</b>			<b>0.92</b>
<b>Makro Ortalama</b>	0.92	0.90	0.91
<b>Ağırlıklı Ortalama</b>	0.92	0.92	0.92

Rastgele orman modeli, %92 doğruluk oranı ile çalışmada kullanılan modeller arasında en yüksek genel başarı değerlerinden birine ulaşmıştır. Sınıf bazlı analizler, her iki sınıf için de yüksek kesinlik ve duyarlılık değerlerinin elde edildiğini göstermektedir. Ağırlıklı ortalama F1-skorunun %92 düzeyinde gerçekleşmesi, modelin hem genel sınıflandırma başarısı hem de sınıflar arası denge açısından istikrarlı sonuçlar ürettiğini ortaya koymaktadır.

Rastgele orman modelinin sınıflandırma performansı, hata matrisi kullanılarak ayrıntılı biçimde incelenmiştir. Hata matrisi, model tarafından gerçekleştirilen doğru ve hatalı sınıflandırmaların dağılımını göstermekte olup, gerçek sınıf etiketleri ile tahmin edilen sınıflar arasındaki ilişkiyi ortaya koymaktadır. Rastgele orman modeline ait hata matrisi Şekil 3'te sunulmaktadır.



Şekil 3. Rastgele Orman Modeline Ait Hata Matrisi

Hata matrisi sonuçlarına göre:

- 209 gözlem, gerçek sınıfı 0 olan öğrenciler arasından doğru şekilde 0 olarak sınıflandırılmıştır (TN – Doğru Negatif).
- 36 gözlem, gerçek sınıfı 0 olmasına rağmen 1 olarak sınıflandırılmıştır (FP – Yanlış Pozitif).
- 22 gözlem, gerçek sınıfı 1 olmasına rağmen 0 olarak sınıflandırılmıştır (FN – Yanlış Negatif).
- 459 gözlem, gerçek sınıfı 1 olan öğrenciler arasından doğru şekilde 1 olarak sınıflandırılmıştır (TP – Doğru Pozitif).

Bu dağılım, rastgele orman modelinin her iki sınıf için de yüksek düzeyde doğru sınıflandırma başarısı sergilediğini göstermektedir. Özellikle 1 sınıfına (Dropout – okul terk riski taşıyan öğrenciler) ait gözlemlerde doğru pozitif sayısının yüksek olması, modelin risk altındaki öğrencileri belirleme kapasitesinin güçlü olduğunu ortaya koymaktadır. Hata matrisi üzerinden elde edilen bu bulgular, performans tablosunda sunulan doğruluk, kesinlik, duyarlılık ve F1-skoru değerleri ile tutarlılık göstermektedir.

### 3.4. XGBoost Modeli Bulguları

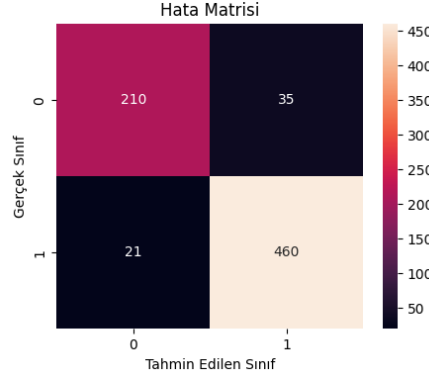
Bu bölümde, XGBoost modeli kullanılarak elde edilen sınıflandırma performans sonuçları sunulmaktadır. XGBoost modeline ait sınıf bazlı performans ölçütleri ile genel başarı değerleri Tablo 4'te yer almaktadır.

Tablo 4. XGBoost Modeline Ait Sınıf Bazlı Performans Ölçütleri

Sınıf	Kesinlik (Precision)	Duyarlılık (Recall)	F1-Skoru
0	0.91	0.86	0.88
1	0.93	0.96	0.94
<b>Doğruluk (Accuracy)</b>			<b>0.92</b>
<b>Makro Ortalama</b>	0.92	0.91	0.91
<b>Ağırlıklı Ortalama</b>	0.92	0.92	0.92

XGBoost, rastgele orman modeli ile benzer şekilde %92 doğruluk oranına ulaşmıştır. Sınıf bazlı sonuçlar incelendiğinde, özellikle pozitif sınıfa ilişkin duyarlılık değerinin yüksek olduğu ve yanlış negatif sınıflandırmaların sınırlı kaldığı görülmektedir. Makro ve ağırlıklı ortalama F1-skorumun %91-%92 aralığında gerçekleşmesi, modelin sınıflar arası performans dengesini koruduğunu göstermektedir.

XGBoost modelinin sınıflandırma performansı, hata matrisi kullanılarak ayrıntılı biçimde incelenmiştir. Hata matrisi, model tarafından gerçekleştirilen doğru ve hatalı sınıflandırmaların dağılımını göstermekte olup, gerçek sınıf etiketleri ile tahmin edilen sınıflar arasındaki ilişkiyi ortaya koymaktadır. XGBoost modeline ait hata matrisi Şekil 4'te sunulmaktadır.



Şekil 4. XGBoost Modeline Ait Hata Matrisi

Hata matrisi sonuçlarına göre:

- 210 gözlem, gerçek sınıfı 0 olan öğrenciler arasından doğru şekilde 0 olarak sınıflandırılmıştır (TN – Doğru Negatif).
- 35 gözlem, gerçek sınıfı 0 olmasına rağmen 1 olarak sınıflandırılmıştır (FP – Yanlış Pozitif).
- 21 gözlem, gerçek sınıfı 1 olmasına rağmen 0 olarak sınıflandırılmıştır (FN – Yanlış Negatif).
- 460 gözlem, gerçek sınıfı 1 olan öğrenciler arasından doğru şekilde 1 olarak sınıflandırılmıştır (TP – Doğru Pozitif).

Bu dağılım, XGBoost modelinin her iki sınıf için de yüksek düzeyde doğru sınıflandırma başarısı sunduğunu göstermektedir. Özellikle 1 sınıfına (Dropout – okul terk riski taşıyan öğrenciler) ait doğru pozitif sayısının yüksek olması, modelin risk altındaki öğrencileri belirleme kapasitesinin güçlü olduğunu ortaya koymaktadır. Hata matrisi üzerinden elde edilen bu sonuçlar, performans tablosunda sunulan doğruluk, kesinlik, duyarlılık ve F1-skoru değerleri ile tutarlılık göstermektedir.

#### 4. TARTIŞMA

Bu çalışmanın temel amacı, öğrenci terki tahmininde farklı makine öğrenmesi algoritmalarının karşılaştırmalı performanslarını incelemektir. Dolayısıyla analiz, değişkenlerin nedensel etkilerini test etmekten ziyade, çok boyutlu veri yapısı içerisinde hangi modelleme yaklaşımının daha yüksek ve dengeli tahmin başarısı sunduğunu ortaya koymaya odaklanmaktadır. Makine öğrenmesi modelleri özellikle yüksek boyutlu veri yapılarında değişkenler arasındaki karmaşık ve doğrusal olmayan ilişkileri bütüncül biçimde ele almakta; ancak bu yaklaşım doğrudan nedensel çıkarım üretmeyi amaçlamamaktadır. Hangi faktörlerin terki tetiklediğine yönelik ayrıntılı değişken etki analizleri, farklı modelleme teknikleri ve açıklanabilir yapay zekâ yöntemleri gerektiren ayrı bir araştırma tasarımı kapsamında ele alınabilir.

Çalışma kapsamında elde edilen bulgular, eğitim sektöründe öğrenci başarısı ve okul terk riskinin tahmin edilmesine yönelik olarak kullanılan farklı makine öğrenmesi yaklaşımlarının birbirinden ayrılan performans profilleri sunduğunu ortaya koymaktadır. Lojistik regresyon ve karar ağacı gibi tekil modelleme yaklaşımları, genel doğruluk ve F1-skoru açısından kabul edilebilir sonuçlar üretmiş olsa da topluluk öğrenme temelli modellerin (Rastgele Orman ve XGBoost) daha dengeli ve istikrarlı bir sınıflandırma performansı sergilediği görülmektedir. Bu durum, özellikle çok sayıda akademik, demografik ve sosyo-ekonomik değişken içeren eğitim verilerinde, tekil modeller yerine çoklu model birleşimlerine dayalı yaklaşımların daha güçlü genelleme kapasitesi sunduğunu göstermektedir.

Hata matrisleri üzerinden yapılan değerlendirmeler, modeller arasındaki farkların yalnızca genel doğruluk oranlarıyla sınırlı olmadığını ortaya koymaktadır. Özellikle yanlış negatif (FN) sınıflandırmalar açısından bakıldığında, topluluk öğrenme temelli modellerin daha düşük hata oranlarına sahip olduğu görülmektedir. Eğitim sektörü bağlamında yanlış negatif sınıflandırmalar, okul terk riski taşıyan bir öğrencinin model tarafından tespit edilememesi anlamına gelmekte ve erken müdahale fırsatlarının kaçırılmasına yol açabilmektedir. Bu nedenle, erken uyarı sistemlerinin geliştirilmesinde duyarlılık (recall) ve yanlış negatif oranlarının, doğruluk ölçütü kadar hatta bazı durumlarda daha fazla önem taşıdığı söylenebilir. Rastgele Orman ve XGBoost modellerinin bu açıdan sunduğu görece avantaj, bu modellerin eğitimsel karar destek sistemlerinde tercih edilebilirliğini artırmaktadır.

Karar ağacı modelinin sunduğu yorumlanabilir yapı, özellikle eğitim yöneticileri ve karar vericiler açısından önemli bir avantaj olarak değerlendirilebilir. Ancak elde edilen sonuçlar, tekil karar ağaçlarının performans açısından topluluk öğrenme temelli modellere kıyasla sınırlı kaldığını göstermektedir. Bu bulgu, eğitim verilerinin doğrusal olmayan ve karmaşık yapısının, tek bir karar ağacı tarafından tam anlamıyla temsil edilmesinin zor olduğunu düşündürmektedir. Buna karşılık, birden fazla karar ağacının bir araya getirilmesiyle oluşturulan Rastgele Orman yaklaşımı hem varyansı azaltmakta hem de daha kararlı tahminler üretmektedir.

Lojistik regresyon modeli ise güçlü bir temel karşılaştırma noktası sunmasına rağmen, özellikle değişkenler arasındaki karmaşık etkileşimlerin bulunduğu veri yapılarında sınırlı bir açıklama gücüne sahiptir. Bu durum, eğitim sektöründe veri odaklı karar destek sistemleri geliştirilirken, yalnızca basit ve yorumlanabilir modellerin değil, aynı zamanda daha gelişmiş ve esnek öğrenme mekanizmalarına sahip modellerin de dikkate alınması gerektiğini ortaya koymaktadır. Bununla birlikte, lojistik regresyonun sunduğu istikrarlı performans, bu modelin erken aşama analizler ve karşılaştırmalı çalışmalar için hâlen önemli bir referans noktası olduğunu göstermektedir.

Rastgele orman ve XGBoost modellerinin birbirine oldukça yakın performans değerleri sergilemesi, farklı öğrenme stratejilerine sahip olmalarına rağmen benzer düzeyde tahmin başarısı sunabildiklerini ortaya koymaktadır. Rastgele orman modeli, çoklu karar ağaçlarının rastgele örnekleme ve özellik seçimi yoluyla bir araya getirilmesine dayanırken, XGBoost modeli, artımlı hata düzeltme mekanizması sayesinde zayıf öğrenicileri ardışık biçimde güçlendirmektedir. Bu iki farklı yaklaşımın benzer sonuçlar üretmesi, eğitim verilerinin belirli bir örüntü yapısına sahip olduğunu ve bu yapının farklı topluluk öğrenme stratejileriyle etkili biçimde öğrenilebildiğini düşündürmektedir.

İşletme ve yönetim perspektifinden değerlendirildiğinde, bu çalışmanın bulguları eğitim kurumlarının yalnızca akademik faaliyet yürüten yapılar değil, aynı zamanda karmaşık hizmet organizasyonları olarak ele alınması gerektiğini ortaya koymaktadır. Öğrenci terk riski, yalnızca bireysel bir akademik başarısızlık göstergesi değil aynı zamanda kurumlar açısından kaynak kaybı, hizmet kalitesinde düşüş ve kurumsal sürdürülebilirlik sorunları doğuran stratejik bir yönetim problemidir. Bu bağlamda, makine öğrenmesi temelli tahmin modelleri, eğitim kurumları için operasyonel kararları destekleyen teknik araçların ötesinde, stratejik yönetim süreçlerine katkı sunan analitik bileşenler olarak değerlendirilebilir.

Bu çalışma karşılaştırmalı modelleme yaklaşımının önemini açık biçimde ortaya koymaktadır. Tek bir modelin sonuçlarına dayalı çıkarımlar yerine, farklı modelleme yaklaşımlarının aynı veri seti ve aynı deneysel koşullar altında değerlendirilmesi, eğitim sektörü için daha güvenilir ve genellenebilir karar destek sistemlerinin geliştirilmesine olanak tanımaktadır. Bu yaklaşım, eğitim kurumlarının reaktif müdahalelerden ziyade öngörüye dayalı, veri temelli ve proaktif bir yönetim anlayışına geçiş yapabilmesi açısından önemli bir katkı sunmaktadır.

Son olarak, Rastgele Orman ve XGBoost modellerinin benzer performans değerleri üretmesi, bu iki yöntem arasında belirgin bir üstünlük bulunmadığını göstermektedir. Bununla birlikte, modeller arasındaki performans farklarının istatistiksel anlamlılığına ilişkin ek bir test uygulanmamıştır. Bu durum çalışmanın yöntemsel sınırlılıklarından biri olarak değerlendirilebilir. Gelecek araştırmalarda, modeller arası performans farklarının istatistiksel olarak test edilmesi daha güçlü çıkarımlar yapılmasına katkı sağlayabilir.

## 5. SONUÇ VE ÖNERİLER

Bu çalışmada öğrenci başarısı ve okul terk riskinin tahmin edilmesine yönelik olarak lojistik regresyon, karar ağacı, rastgele orman ve XGBoost modelleri aynı deneysel koşullar altında karşılaştırmalı biçimde değerlendirilmiştir. Analizler, açık kaynaklı bir veri seti üzerinde, %80 eğitim ve %20 test ayrımı kullanılarak gerçekleştirilmiş; tüm modeller aynı performans ölçütleri ve hata matrisleri üzerinden değerlendirilmiştir. Bu yaklaşım, elde edilen performans değerlerinin doğrudan ve adil biçimde karşılaştırılmasına olanak sağlamıştır.

Yüzde değerler üzerinden yapılan değerlendirme, modeller arasında belirgin farklılıklar bulunduğunu göstermektedir. Lojistik regresyon modeli %91 doğruluk ve %91 ağırlıklı F1-skoru ile istikrarlı bir temel performans sunarken, karar ağacı modeli %90,63 doğruluk ve %90,56 ağırlıklı F1-skoru ile daha sınırlı bir başarı düzeyinde kalmıştır. Buna karşılık, topluluk öğrenme temelli modeller olan rastgele orman ve XGBoost, her iki ölçüt için de %92 düzeyine ulaşmış ve özellikle okul terk riski taşıyan öğrencileri temsil eden sınıf için yüksek duyarlılık değerleri (%95 ve %96) üretmiştir. Bu sonuçlar, risk sınıfına ilişkin yanlış negatif hataların azaltılması açısından söz konusu modellerin daha avantajlı bir yapı sunduğunu göstermektedir.

Bu çalışmada elde edilen performans değerleri, literatürde raporlanan bazı güncel ve erişilebilir örneklerle karşılaştırıldığında daha yüksek bir sınıflandırma başarısına işaret etmektedir. Örneğin Salloum ve arkadaşları (2024), öğrenci devamlılığı/terk tahmininde rastgele orman sınıflandırıcısı ile %76,72 doğruluk değerine ulaştıklarını raporlamıştır. Bu çalışmada ise rastgele orman ve XGBoost modelleri için %92 doğruluk ve %92 ağırlıklı F1-skoru elde edilmiş ve buna ek olarak risk sınıfına ilişkin duyarlılık değerleri %95–%96 aralığında gerçekleşmiştir. Ayrıca, çevrim içi eğitim bağlamında dropout tahminine odaklanan Yükseltürk ve arkadaşları (2014), çalışmasında en yüksek duyarlılık değerinin %87 düzeyinde raporlandığı, karar ağacı için ise duyarlılığın %79,7 olduğu belirtilmiştir. Bu bulgular birlikte değerlendirildiğinde, mevcut çalışmada elde edilen yüzde değerlerin (özellikle %92 doğruluk ve risk sınıfında %95–%96 duyarlılık) literatürde raporlanan örnek sonuçların üzerinde bir performans sunduğu görülmektedir. Bununla birlikte, veri seti, örneklem özellikleri ve doğrulama stratejileri farklı olabileceğinden, bu karşılaştırma “doğrudan eşdeğer” bir kıyas değil de elde edilen başarının literatürdeki tipik değerlerle konumlandırılması amacıyla yapılan bir performans karşılaştırması olarak değerlendirilmelidir.

Bu bağlamda, mevcut çalışmada rastgele orman ve XGBoost modelleri için elde edilen %92 doğruluk ve %92 ağırlıklı F1-skoru değerleri, literatürde raporlanan pek çok çalışmanın üzerinde bir performansa karşılık gelmektedir. Dahası, risk altındaki öğrencileri temsil eden sınıf için elde edilen %95 ve %96 düzeyindeki duyarlılık değerleri, yalnızca genel başarıyı değil, erken müdahale açısından kritik öneme sahip olan doğru risk tespitini de güçlendirmektedir. Bu durum, çalışmanın yalnızca istatistiksel açıdan değil, uygulamaya dönük sonuçlar bakımından da literatüre anlamlı bir katkı sunduğunu göstermektedir.

Çalışmadan elde edilen sonuçlar, öğrenci terk riskinin tahmin edilmesinde tek bir algoritmaya dayalı yaklaşımlar yerine, farklı modelleme tekniklerinin aynı koşullar altında karşılaştırılmasının önemini ortaya koymaktadır. Özellikle topluluk öğrenme temelli modellerin sunduğu daha yüksek ve dengeli performans, erken uyarı sistemlerinin geliştirilmesinde bu yaklaşımların öncelikli olarak değerlendirilmesi gerektiğini göstermektedir. Bununla birlikte, model seçiminin yalnızca performans değerlerine değil, yorumlanabilirlik, sürdürülebilirlik ve uygulama kolaylığı gibi boyutlara göre yapılması gerektiği de dikkate alınmalıdır.

Bu çalışmanın aynı zamanda bazı sınırlılıkları bulunmaktadır. Analizler tek bir veri seti üzerinden yürütülmüş olup, sonuçların farklı eğitim sistemleri veya farklı ülke bağlamları için genellenebilirliği sınırlı olabilir. Ayrıca veri seti, öğrencilerin zaman içindeki davranışlarını yansıtan dinamik değişkenleri içermemektedir. Gelecek çalışmalarda, farklı veri setlerinin kullanılması, zaman boyutunu içeren modellerin geliştirilmesi ve maliyet-duyarlı sınıflandırma yaklaşımlarının denenmesi, öğrenci terk riskinin daha derinlemesine analiz edilmesine katkı sağlayabilir.

Genel olarak değerlendirildiğinde, bu çalışma açık kaynak verilerle yürütülen karşılaştırmalı makine öğrenmesi analizlerinin öğrenci terk riskinin tahmininde yüksek ve literatürle karşılaştırıldığında üstün bir performans ortaya koyabildiğini göstermektedir. Yüzde değerler üzerinden yapılan ayrıntılı karşılaştırmalar ve hata türlerini merkeze alan değerlendirme yaklaşımı, gelecekte geliştirilecek erken uyarı sistemleri için sağlam bir metodolojik çerçeve sunmaktadır.

#### YAZAR BEYANI

**Araştırma ve Yayın Etiği Beyanı:** Bu çalışma bilimsel araştırma ve yayın etiği kurallarına uygun olarak hazırlanmıştır.

**Etik Kurul Onayı:** Bu araştırma etik kurul izni gerektiren analizleri kapsamadığından etik kurul onayı gerektirmemektedir.

**Yapay Zekâ Kullanım Beyanı:** Bu çalışmanın herhangi bir aşamasında yapay zekâ temelli araçlar kullanılmamıştır. Araştırmanın tasarımı, veri analizi, modelleme süreci, bulguların yorumlanması ve metnin yazımı tamamen yazar tarafından gerçekleştirilmiştir. Çalışmada kullanılan makine öğrenmesi algoritmaları, araştırma kapsamında yazar tarafından kodlanan modeller olup, herhangi bir yapay zekâ destekli içerik üretim aracıyla ilişkilendirilmemektedir.

**Yazar Katkıları:** Yazar çalışmanın tümünü tek başına gerçekleştirmiştir.

**Çıkar Çatışması:** Yazar açısından ya da üçüncü taraflar açısından çalışmadan kaynaklı çıkar çatışması bulunmamaktadır.

**KAYNAKÇA**

- Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, 79, 101102
- Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences*, 11(9), 552.
- Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodríguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., ... & Angeles-Morales, J. (2023). Predicting student dropout based on machine learning and deep learning: A systematic review. *EAI Endorsed Transactions on Scalable Information Systems*, 10(5).
- Bargmann, C., Thiele, L., & Kauffeld, S. (2022). Motivation matters: predicting students' career decidedness and intention to drop out after the first year in higher education. *Higher Education*, 83(4), 845-861.
- Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, 55(4), 485-540.
- Braxton, J. M., Doyle, W. R., Hartley III, H. V., Hirschy, A. S., Jones, W. A., & McLendon, M. K. (2013). *Rethinking college student retention*. John Wiley & Sons.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (s. 785-794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Huo, H., Cui, J., Hein, S., Padgett, Z., Ossolinski, M., Raim, R., & Zhang, J. (2023). Predicting dropout for nontraditional undergraduate students: A machine learning approach. *Journal of College Student Retention: Research, Theory & Practice*, 24(4), 1054-1077.
- Kaggle. (2025). Predict students' dropout and academic success. <https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data>
- Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, 26(1), 205-240.
- López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, 33(3), 456.
- Namoun, A., & Alshantiri, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), 237.
- OECD. (2022). *Education at a glance 2022: OECD indicators*. OECD Publishing. <https://doi.org/10.1787/3197152b-en>
- Pek, R. Z., Özyer, S. T., Elhage, T., Özyer, T., & Alhaji, R. (2022). The role of machine learning in identifying students at-risk and minimizing failure. *IEEE Access*, 11, 1224-1243.
- Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied sciences*, 10(3), 1042.
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2021). *Predict students' dropout and academic success* (Version 1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5777340>
- Ridwan, A., Priyatno, A. M., & Ningsih, L. (2024). Predict students' dropout and academic success with XGBoost. *Journal of Education and Computer Applications*, 1(2), 1-8.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3), e1355.
- Salloum, S. A., Basiouni, A., Alfaisal, R., Salloum, A., & Shaalan, K. (2024). Predicting student retention in higher education using machine learning. In *International Conference on Breaking Barriers with Generative Intelligence* (pp. 197-206). Cham: Springer Nature Switzerland.
- Quinlan, J. R. (2014). *C4.5: programs for machine learning*. Elsevier.
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1), 89-125.
- Tinto, V. (2012). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago press.

