Journal of the Faculty of Engineering and Architecture of Gazi University 40:4 (2025) 2191-2204

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

A classification study on predicting academic success using machine learning with oversampling and undersampling

Ayşe Alkan[1]*, Onur Sevli[2]
[1]Samsun Science and Art Center, 55070, İlkadım, Samsun, Türkiye
[2]Department of Computer Engineering, Faculty of Engineering, Burdur Mehmet Akif Ersoy University, 06680, Burdur, Türkiye

Highlights:

- Determining academic success in advance
- Effect of resampling techniques on classification performance
- Comparing the performance of various machine learning algorithms

Keywords:

- Educational data mining
- Machine learning
- Classification
- Prediction
- Resampling

Article Info:

Research Article
Received: 05.04.2024
Accepted: 19.05.2025

DOI:

10.17341/gazimmfd.1465283

Correspondence:

Author: Ayşe Alkan
e-mail: ayse.alkan55@gmail.com
phone: +90 362 231 0288

Graphical/Tabular Abstract

In this research, performance comparisons were made using various resampling techniques and different classifiers shown in Figure A in order to predict academic success.

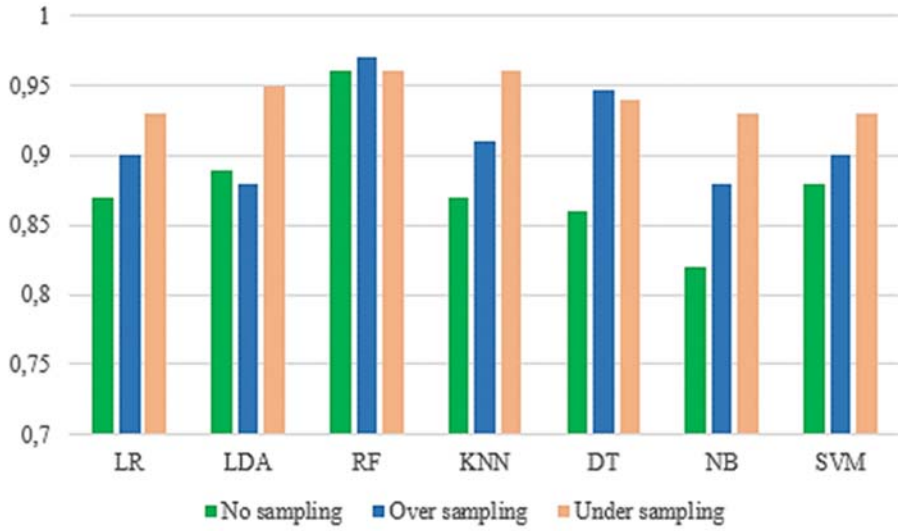


Figure A. Accuracy scores of classifiers obtained by resampling and non-sampling methods

Purpose: The aim of this study is to predict and predetermine students' academic success by using different machine learning algorithms on the data set. Accordingly, it includes classification studies for early detection of academic success with different machine learning algorithms and resampling techniques on the Open University Learning Analytics Dataset.

Theory and Methods: Open University Learning Analytics Dataset-OULAD" data set was used within the scope of the study. Logistic Regression (LR), Linear Discriminant Analysis (LDA), Rondem Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT, Support Vector Machine (SVM) and Naive Bayes (NB) methods were used in the classification processes and their performances were evaluated. Seven different methods were used. In addition to its performance without sampling, thirteen different resampling techniques were applied independently for each method used in this study. In this process, a total of ninety-one different classification processes were performed and the results are reported.

Results: Using the RandomOverSampler oversampling technique, 97% accuracy was obtained with the Random Forest classifier, and 96% accuracy was obtained with the K-Nearest Neighbors classifier using the AIIKNN undersampling method. The best results obtained with Random Forest in all classifications are 97% in accuracy, 96% in precision, 96% in recall and 96% in F1 Score. The best results obtained with K-Nearest Neighbors in all classifications are 96% in accuracy, 92% in precision, 92.6% in recall and 93% in F1 Score.

Conclusion: In the study, it was seen that RandomOverSampler, Kmeans SMOTE, which are oversampling techniques, and EditedNearestNeighbors and AIIKNN, which are undersampling techniques, provided a higher increase in success. The imbalance of classes in the data set is a factor that affects the success of the models. Balancing the existing data set with appropriate sampling techniques allows models to be more successful.

Journal of the Faculty of Engineering and Architecture of Gazi University 40:4 (2025) 2191-2204

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Fazla örnekleme ve az örnekleme ile makine öğrenmesi kullanarak akademik başarının tahmin edilmesine yönelik bir sınıflandırma çalışması

Ayşe Alkan[1]*, Onur Sevli[2]
[1]Samsun Bilim ve Sanat Merkezi, 55070, İlkadım, Samsun, Türkiye
[2]Burdur Mehmet Akif Ersoy Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 06680, Burdur, Türkiye

ÖNEÇIKANLAR

- Akademik başarının önceden tespit edilmesi
- Yeniden örnekleme tekniklerinin sınıflandırma performansına etkisi
- Çeşitli makine öğrenimi algoritmalarının performansının karşılaştırılması

Makale Bilgileri
Araştırma Makalesi
Geliş: 05.04.2024
Kabul: 19.05.2025

DOI:
10.17341/gazimmfd.1465283

Anahtar Kelimeler:
Eğitsel veri madenciliği, makine öğrenimi, sınıflandırma, tahmin, yeniden örnekleme

ÖZ

Makine öğrenimi, bir bilgisayarın verileri analiz etmeyi, modeller oluşturmayı ve karar vermeyi öğrenmesine olanak tanıyan bir yapay zekâ alanıdır. Makine öğrenimi, eğitimde önemli bir rol oynamaktadır. Bu araştırmada, Open University Learning Analytics veri seti üzerinde çeşitli makine öğrenimi teknikleri kullanılarak öğrencilerin akademik performanslarını tahmin etmek amacıyla sınıflandırmalar gerçekleştirilmiştir. Çeşitli yöntemlerin başarısı değerlendirilmiş, ayrıca sınıflandırıcıların tahmin başarısını arttırmak için literatürde sıkça görülen SMOTE, KMeansSMOTE, RandomOverSampler, ADASYN, BorderlineSMOTE ve SVMSMOTE fazla örnekleme teknikleri kullanılarak veri arttırma işlemi yapılmıştır. Ayrıca, EditedNearestNeighbours, AllKNN, NearMiss, NeighborhoodCleaningRule, OneSidedSelection, RandomUnderSampler ve TomekLinks az örnekleme teknikleri kullanılarak da veri azaltma işlemi yapılmıştır. Lojistik Regresyon, Linear Diskriminant Analizi, Rastgele Orman, K-en yakın komşu, Karar Ağaçları, Naive Bayes ve Destek Vektör Makineleri olarak yedi farklı makine öğrenme tekniği kullanılmıştır. Yeniden örnekleme olmadan sınıflandırmaya ayrıca altı farklı fazla örnekleme ve yedi farklı az örnekleme tekniği kullanılarak doksan bir sınıflandırma işlemi gerçekleştirilmiştir. Tüm sınıflandırma süreçleri dört farklı performans metriği ile raporlanmıştır. RandomOverSampler fazla örnekleme tekniği kullanılarak Rastgele Orman sınıflandırıcı ile %97, AIIKNN az örnekleme tekniği ile K-en yakın komşu sınıflandırıcıdan %96 doğruluk elde edilmiştir. Veri setinde sınıf dengesini sağlamak amacıyla fazla örnekleme ve az örnekleme yönteminin sınıflandırıcı başarısını arttırdığı görülmüştür.

A classification study on predicting academic success using machine learning with oversampling and undersampling

HIGHLIGHTS

- Determining academic success in advance
- Effect of resampling techniques on classification performance
- Comparing the performance of various machine learning algorithms

Article Info
Research Article
Received: 05.04.2024
Accepted: 19.05.2025

DOI:
10.17341/gazimmfd.1465283

Keywords:
Educational data mining, machine learning, classification, prediction, resampling

ABSTRACT

Machine learning represents a domain within artificial intelligence enabling computers to acquire the capacity for data analysis, model construction, and decision-making. Machine learning plays an important role in education. In this research, classifications were performed on the Open University Learning Analytics dataset using various machine learning techniques to predict students' academic performance. The success of various methods was evaluated, and to enhance the predictive accuracy of the classification algorithms data augmentation was performed using SMOTE, KMeansSMOTE, RandomOverSampler ADASYN, BorderlineSMOTE and SVMSMOTE oversampling techniques, which are frequently seen in the literature. Additionally, data reduction was performed using EditedNearestNeighbors, AllKNN, NearMiss, NeighborhoodCleaningRule, OneSidedSelection, RandomUnderSampler and TomekLinks undersampling techniques. Seven different machine learning techniques were used: Logistic Regression, Linear Discriminant Analysis, Random Forest, K-nearest neighbors, Decision Trees, Naive Bayes and Support Vector Machines. In addition to classification without resampling, ninety-one classification operations were performed using six different oversampling and seven different undersampling techniques. All classification processes are reported with four different performance metrics. Using the RandomOverSampler oversampling technique, 97% accuracy was achieved with the Random Forest classifier, and 96% accuracy was achieved with the K-nearest neighbors classifier using the AIIKNN undersampling technique. It has been noted that employing oversampling and undersampling techniques enhances the effectiveness of classifiers by addressing class imbalance within the dataset.

*Sorumlu Yazar / Yazarlar / *Corresponding Author/Authors: *ayse.alkan55@gmail.com, onursevli@mehmetakif.edu.tr / Tel: +90 362 231 02 88

1. Giriş (Introduction)

Bireylerde yaşantılar sonucu davranışlarda meydana gelen öğrenme sürecinde öğretmenler, öğrencilerin başarılarına ilişkin geribildirim alarak öğrenme ortamını, öğrenme programlarını, öğrencileri ve öğretmenin kendisini değerlendirerek sürece ilişkin gerekli düzenlemeleri yapabilirler. Eğitim alanında temel bir unsur olan eğitsel değerlendirme, öğrencinin anlama düzeyini belirlemek ve öğrenme sürecini geliştirmek amacıyla yapılan ilerlemeyi izleme sürecidir. Bu kritik aşama, eğitimin belirleyici güçlerinden biri olarak kabul edilir ve öğrenme stratejilerini önemli ölçüde etkileyebilir [1]. Klasik öğrenme ortamlarına kıyasla, teknolojik araçların eğitim süreçlerine hızla entegre edilmesi, öğrenci değerlendirmelerinin ve bilgilerinin dijital ortamlarda tutulmasını sağlar. Öğrenci verileri, büyük veri yığınlarını oluşturur. Bugünün gelişen teknolojileri, veri toplama, izleme, performans değerlendirme ve öğrencilere geri bildirim sağlama konularında olanaklar sunmaktadır. Eğitim sektöründe makine öğrenimi, yapay zekâ ve veri bilimi alanındaki hızlı ilerlemeler sayesinde öğrenme verileri toplanabilir ve öğrencilerin özel öğrenme ihtiyaçları belirlenebilir [2].

Bu çalışma, Ayşe Alkan'nın, Onur Sevli danışmanlığında yürüttüğü, Burdur Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı bünyesinde tamamlanan "Akademik başarının eğitsel veri madenciliği ve makine öğrenmesi teknikleri kullanılarak tahminlenmesi" başlıklı tez çalışmasına dayanmaktadır.

Karmaşık ilişkileri ve kalıpları öğrenebilen, makine öğrenmesi algoritmaları ile toplanan veriler arasında karar verebilmek, öğrencilerin performansını tahmin etmek ve öğrenme hedeflerine ulaşmada olası başarısızlığa karşı erken önlem almak mümkün olabilmektedir. Eğitim alanında sıklıkla kullanılan makine öğrenmesi algoritmaları akademik başarının tahmin edilmesinde de yaygın olarak kullanılan bir yöntem haline gelmiştir [3]. İnternetin eğitim alanında giderek yaygınlaşması, elektronik öğrenme, mobil öğrenme, çevrimiçi eğitim, web tabanlı eğitim ve kitlelerin katılabildiği açık çevrimiçi dersler gibi çeşitli platformların kullanımını artırmıştır. Bu popülerlik ve fırsatlar, dijital ortamda eğitim alan öğrenci sayısının artmasına paralel olarak veri miktarının da artmasını beraberinde getirmiştir. Online ortamlarda yapılan faaliyetler sonucunda öğrenciler dijital izlerini bırakmakta ve bu izler geniş veri havuzlarını oluşturmaktadır. Online eğitim ortamları, klasik sınıf ortamlarına göre öğrenci verilerini veri tabanlarına kaydetme imkânı sunmaktadır [4]. Veri tabanlarındaki veriler tek başına bir anlam taşımasa da çeşitli veri analiz yöntemleriyle işlenen veriler arasındaki bağlantılar anlamlı bilgiler sağlayabilmektedir. Öğrenme ortamlarının verimliliğini artırmak ve olası başarısızlık durumlarını tahmin etmek amacıyla öğrenci performansı, derse katılım, soru sorma sıklığı gibi verileri analiz etmek amacıyla araştırmacılar son dönemde veri madenciliği, eğitsel veri madenciliği gibi alanlara yönelmişlerdir. Veri madenciliği, 1960'larda verilerin bilgisayarda depolanmasıyla ortaya çıkmış ve büyük veri setlerindeki gizli ilişkileri inceleyerek anlamlı bilgiler elde etmeyi amaçlayan bir analiz yöntemidir [5, 6]. Veri setlerinin analizinde istatistik, bilgi teknolojileri, veri tabanı teknolojileri, yapay zekâ, makine öğrenmesi, örüntü tanımlama, veri görselleştirme gibi farklı yöntemler kullanılmaktadır [7]. Veri setlerinin analizinde kullanılan bir yöntem olan makine öğrenmesi, bilgisayarların insanlar gibi öğrenme yeteneğine sahip olmalarını ve mevcut verileri analiz ederek problemlere çözüm üretmelerini sağlayan bir disiplindir. Tarihi süreçte bilim insanları, "Makinalar düşünebilir mi?" sorusunun cevabını aramışlardır [8]. Makine öğrenmesi, yapay zekânın bir alt disiplini olup, bilgisayarların geçmiş veri deneyimlerinden yola çıkarak geleceğe yönelik tahminler yapmak için istatistiksel modellerin kullanıldığı bilimsel bir çalışma alanıdır [9, 10]. Makine öğrenmesi süreci, veri toplama, veri hazırlama, model eğitimi, model değerlendirmesi ve performansın artırılması olmak üzere beş adımdan oluşmaktadır [11]. Bu adımları takip ederek başarılı sonuçlar alan araştırmaların sayısı her geçen gün artmaktadır. Bu çalışmalardan bazıları bir sonraki bölümde verilmiştir.

2. İlgili Çalışmalar (Related Works)

Makine öğrenmesi algoritmaları ile mevcut verileri eğiterek ve en iyi modeli oluşturarak gelecekteki farklı durumlar için istikrarlı tahminler yapabilen pek çok araştırma mevcuttur. Örneğin, yükseköğretim programlarındaki öğrencilerin sonuçlarını tahmin etmek için kullanılan makine öğrenimi yaklaşımlarından oldukça doğru tahminler üretilmiştir [12, 13]. Üniversite sınav sonuçları ile öğrencilerin başarıları arasındaki ilişkiyi K-Means algoritmasıyla inceleyen araştırmada öğrencilerin bölümlerinin ve başarılarının beş farklı küme oluşturduğu belirtmiştir [14]. Üniversite öğrencilerinin başarısızlık ve devamsızlık tahminleri için Karar Ağaçları (KA), Yapay Sinir Ağları (YSA), Rastgele Orman (RO) ve Lineer Diskriminant Analizi (LDA) tekniklerini kullanan çalışmanın sonucunda en yüksek doğru sınıflandırma %57,35 ile LDA'dan elde edilmiştir [15]. KA, YSA ve LDA yöntemlerini kullanarak üniversite öğrencilerinin mezuniyet notlarının tahmin edildiği çalışmada kullanılan yöntemler %80 doğru sonuç vermiştir [16]. İstatistiksel sınıflandırıcılar, KA, Kural İndüksiyonu, Bulanık Kural Öğrenmesi ve YSA algoritmaları gibi yöntemlerle desteklenen bir Öğrenme Yönetim Sistemi (ÖYS) olan Moodle platformundan eğitim alan üniversite öğrencilerinin final notlarını tahmin etmeye yönelik yapılan bir araştırmada, Apriori algoritması ve Birliktelik Analizleri kullanılmıştır. Birliktelik analizi ile %95,5 oranında doğruluk elde edilmiştir [17]. Bir başka araştırmada KA yöntemlerinden J48 algoritması kullanılarak öğrencilerin derse devamları %80 başarı ile doğru olarak tahmin edilmiştir [18]. Expectation Maximization algoritması ile kümeleme ve KA kullanılarak sınıflandırma yapılan çalışmada öğrencilerin Öğrenci Seçme Sınavı (ÖSS) başarılarına etki eden faktörler tespit edilmeye çalışılmıştır [19]. KA, Destek Vektör Makineleri (DVM), YSA, Lojistik Regresyon (LR) modelleri kullanılan çalışmada üniversite birinci sınıf öğrencilerinin yıpranmasının ardındaki nedenler tahmin edilmeye çalışılmıştır ve çalışma sonucunda kullanılan yöntemlerin tamamının yaklaşık %80 oranında doğru sınıflandırma performansı gösterdiği belirtilmiştir [20]. WEKA veri madenciliği yazılımı ile yapılan çalışmada J48 ve RO algoritmaları kullanarak üniversite öğrencilerinin performanslarının tahmin edilmeye çalışıldığı araştırma sonucunda, J48 algoritmasıyla %88,37, RO ile ise %94,41 doğruluğa ulaşılmıştır [21]. Geliştirilmiş yapay zekâyı kullanan ve üniversite öğrencilerinin eğitimlerine devam edip etmediklerini tahmin eden bir model geliştirmeye çalışılan araştırma sonucunda KA algoritmasının %97,50 ile yüksek tahmin doğruluğu sağladığı belirtilmiştir [22]. Üniversite eğitimini bırakan öğrencilerin daha sonra eğitimlerine devam edip edemeyeceklerini bir kümeleme algoritması olan Ward yöntemiyle inceleyen araştırma sonucunda kümelere göre önerilerde bulunulmuştur [23]. YSA, DVM, LR ve KA yöntemlerini kullanarak yapılan çalışmada ortaokul yerleştirme sınavı başarısını en iyi tahmin eden değişkenin öğrencilerin iki yıl önce aldıkları seviye belirleme sınavı puanları olduğu belirtilmiştir [24]. Üniversite birinci sınıf öğrencilerinin öğrenimlerine devam durumunu tahmin etmek için KA, Naive Bayes (NB), YSA ve Rule Induction algoritmalarını kullanarak modeller geliştiren araştırmada, Rule Induction modelinin en yüksek genel doğruluk oranı olan %86,27'ye ulaştığı ifade edilmiştir [25]. Üniversite öğrencilerinin okulu bırakma eğilimlerini belirlemek amacıyla K-en Yakın Komşu (KNN) algoritması kullanarak küme analizi uygulanan araştırma sonucunda başarılı akademik geçmişe sahip öğrencilerin devam durumlarının yüksek olduğu belirtmiştir [26]. Lise öğrencilerinin okulu bırakma durumlarının tahmini üzerine

çalışılan araştırmada RO algoritmasını kullanan model %95 doğrulukla sonuç üretmiştir [27]. Yükseköğretim programlarında öğrencilerin sonuçlarını tahmin etmek için denetimsiz ve denetimli öğrenme tekniklerinin kullanıldığı iki aşamalı makine öğrenmesi yaklaşımında yüksek doğrulukta tahminler ürettiği belirtilmiştir [12]. Üniversitenin akademik itibarı, üniversitenin bulunduğu şehrin olanakları, üniversitenin sağladığı olanaklar ve kültürel olanaklar gibi faktörleri dikkate alarak öğrenci yerleştirme yüzdesinin tahmin edilmesine yönelik bir çalışma sonucunda Extreme Gradient Boosting (XGBoost) algoritmasının diğer makine öğrenmesi yaklaşımlarına kıyasla daha yüksek tahmin doğruluğu sergilediği bulgulanmıştır [13]. Farklı bir çalışmada, başarısızlık riski altındaki öğrencileri tahmin etmek için Öğrenim Yönetim Sistemi (ÖYS) veri kümesi ve internet kullanım günlük dosyaları makine öğrenmesi teknikleriyle analiz edilerek öğrenci demografik verileri birleştirilmiştir [28]. Çevrimiçi eğitim sisteminde etkileme stratejilerinin ikna ediciliği ile öğrencilerin derse yönelik davranışları arasındaki bağlantıların araştırıldığı çalışmada makine öğrenmesi algoritmaları başarılı olarak uygulanmıştır [29]. Başka bir çalışmada ise; başarılı bir tıklama tuzağı tespit performansının ve tıklama tuzağı cümlelerinin dil ve psikolojik yönlerinin ayrıntılı analizinde makine öğrenmesi algoritmaları uygulanmıştır [30]. Öğrenci özelliklerine göre kişilik tipini tahmin etmek için denetimli makine öğrenmesi ve sınıflandırma algoritmalarının kullanıldığı çalışmada ise; öğrencinin potansiyeline göre akademik rehberlik konusunda destek verilmiştir [31]. Denetimli makine öğrenmesi yaklaşımlarını kullanarak öğrencilerin notlarına ve not tahminlerine odaklanılan araştırma öğrenci performansını %96,64 tahmin etmede başarılı olmuştur [32]. Öğrencilerin eğitim düzeylerine, becerilerine ve önceki deneyimlerine göre kariyer yollarını keşfetmelerine yardımcı olmayı amaçlayan yapay zekâ tabanlı çalışmada da öğrencilerin becerilerini değerlendirmelerini ve ilişkili oldukları işleri bulmalarına yardımcı olunmaktadır [33].

3. Materyal ve Metot (Material and Method)

Bu araştırma, çeşitli sınıflandırıcılar kullanarak öğrencilerin başarısını tahmin etmeyi hedeflemektedir. Araştırmada kullanılan kamuya açık "Open University Learning Analytics Dataset-OULAD" veri setine kullanıcıların veri setlerini bulmasına ve yayınlamasına katkı sağlayan, Kaggle web sayfasından ulaşılmıştır [34]. Veri seti üzerinde deneysel çalışmalar ve gerekli analizler yapılmıştır. Sınıflandırma işlemlerinde LR, LDA, RO, KNN, KA, DVM ve NB yöntemleri kullanılmış ve performansları değerlendirilmiştir. Sınıflandırma çalışmalarında temel amaçlardan biri tahmin başarısını arttırmaktır. Bu nedenle yedi farklı yöntem ve her bir yöntem doğruluk, kesinlik, duyarlılık ve F1-Skor değerleri karşılaştırılarak incelenmiştir.

Veri manipülasyonu olmaksızın gerçekleştirilen sınıflandırma yanı sıra, altı çeşit fazla örnekleme ve yedi farklı az örnekleme tekniği kullanılarak toplamda doksan bir sınıflandırma işlemi gerçekleştirilmiştir. Tüm sınıflandırma süreçleri dört ayrı performans ölçütü ile detaylandırılmıştır.

3.1. Veri kümesi (Dataset)

Bu araştırmada, dünyanın en büyük uzaktan eğitim üniversitelerinden biri olan Açık Üniversite (OU)'nun "Open University Learning Analytics Dataset-OULAD" adlı veri seti kullanılmıştır. OULAD, OU öğrencilerinin 2013 ve 2014 yıllarına ait demografik verilerinin yanı sıra üniversitenin Sanal Öğrenme Ortamları (VLE) ile etkileşim verilerini içermektedir. 2013 ve 2014 yıllarına ait öğrenci verilerinin tablo halinde derlendiği OULAD, her bir tabloda açıklayıcı sütunlar ve diğer tablolardaki verilerle ilişkilendirilebilecek çeşitli bilgiler içermektedir. Veri seti yapısına göre öğrenci modülü, demografik bilgiler, kayıt bilgileri, değerlendirme bilgileri ve sanal öğrenme ortamları etkileşim bilgilerine bağlıdır.

OULAD veri seti, uzaktan eğitim sistemi için hazırlanmış, kamuya ücretsiz olarak sunulan bir kaynaktır. Uzaktan eğitimde öğrenci davranışlarını ve performansını analiz etmeye yönelik hazırlanmış olan bu veri seti, demografik bilgiler, modül kayıtları, değerlendirme sonuçları ve sanal öğrenme ortamı etkileşimlerini içermektedir. Veri tabanları arasında ilişkilerin bulunması, farklı tablolardaki verilerin birbiriyle ilişkilendirilmesini ve derinlemesine analiz yapılmasını sağlamaktadır. 32.593 öğrenci ve 22 modül sunumuna ait gerçek verilerden oluşması sebebi ile diğer küçük ölçekli ve sentetik veri setlerinden ayrılmaktadır. Bu özellikler OULAD'ı uzaktan eğitim ve dijital öğrenme süreçlerini analiz etmek için değerli bir kaynak haline getirmektedir.

Veri setindeki her bir öğe, bir veritabanı tablosuna denk gelmektedir ve bu tablolar arasında ilişkisel bir yapı bulunmaktadır. Veri kümesinde öğrenci bilgileri "StudentInfo" tablosunda, Öğrencilerin değerlendirilmesine ait bilgiler "student assessment" tablosunda, değerlendirme ile ilgili bilgiler "assessments" tablosunda, modül ile öğrenci etkileşimine ait bilgiler "studentVle" tablosunda, modüllerin kullanılmasının planlandığı haftalar ile bilgiler "Vle" tablosunda, modül içindeki kurslara ait bilgiler "courses" tablosunda, öğrencinin modüle kayıt olma bilgileri "studentRegistration" tablosunda bulunmaktadır. Bu tablolar, çalışma kapsamında kullanılan veri kaynaklarıdır.

Veri seti içerisindeki courses tablosunda bulunan alanlar Tablo 1'de, assessments tablosunda bulunan alanlar Tablo 2'de, Vle tablosunda bulunan alanlar Tablo 3'te, SudentInfo tablosunda bulunan alanlar Tablo 4'te, StudentRegistration tablosunda bulunan alanlar Tablo 5'te, StudentAssessment tablosunda bulunan alanlar Tablo 6'da, SudentVle tablosunda bulunan alanlar Tablo 7'de verilmiştir.

Tablo 1. Courses tablosunda bulunan öznitelikler (Features found in the Courses table)

Öznitelik Adı	Açıklama
code_module	Modülün belirleyici kodu
code_presentation	Şubat ayında başlayan dönem için 'B', Ekim ayında dönem için ise 'J' harflerinden oluşur.

Tablo 2. Assessments tablosunda bulunan öznitelikler (Features found in the Assessments table)

Öznitelik Adı	Açıklama
code_module	Değerlendirmenin ilgili modülün tanımlayıcı kodu
code_presentation	Değerlendirmenin ilgili sunumun tanımlayıcı kodu
id_assessment	Değerlendirmenin kimlik numarası
assessment_type	Değerlendirme türü. Üç farklı değerlendirme türü vardır: Öğretmen Tarafından İşaretlenen Değerlendirme (TMA), Bilgisayar Tarafından İşaretlenen Değerlendirme (CMA) ve Final Sınavı (Sınav).
date	Modül sunumunun başlangıcından itibaren geçen gün sayısı

Tablo 3. Vle tablosunda bulunan öznitelikler (Features found in the Vle table)

Öznitelik Adı	Açıklama
id_site	Materyalin kodu
code_module	Modülü tanımlamak için bir kod
code_presentation	Sunumu tanımlamak için bir kod
activity_type	Modül malzemesiyle ilişkilendirilmiş rol
week_from	Ders materyalinin başlangıç haftaları
week_to	Ders materyalinin bitiş haftaları

Tablo 4. StudentInfo tablosunda bulunan öznitelikler (Features found in StudentInfo table)

Öznitelik Adı	Açıklama
code_module	Öğrencinin kaydının bulunduğu modül için tanımlayıcı bir kod
code_presentation	Öğrencinin modüle kayıtlı olduğu sunumun kimlik kodu
id_student	Öğrencinin kimlik numarası
gender	Öğrencinin cinsiyeti
region	Öğrencinin sunumunu gerçekleştirdiği konum
highest_education	Modül sunumuna girişte öğrencinin eğitim seviyesi
imd_band	Modül sunumu esnasında öğrencinin yaşadığı bölgenin çoklu dezavantaj bandı endeksi.
age_band	Öğrencinin yaş aralığı
num_of_prev_attempts	Öğrencinin modülü deneme sayısı
studied_credits	Öğrencinin ait toplam kredi sayısı
Disability	Öğrencinin engelli olup olmadığını belirtir
final_result	Öğrencinin modül sunumundaki nihai sonucu

Tablo 5. StudentRegistration tablosunda bulunan öznitelikler (Features found in StudentRegistration table)

Öznitelik Adı	Açıklama
code_module	Modülün belirleyici kodu
code_presentation	Sunumun belirleyici kodu
id_student	Öğrencinin kimlik numarası
date_registration	Öğrencinin modül sunumuna kaydolma tarihi
date_unregistration	Öğrencinin modül sunumundan kaydının silindiği tarih

Tablo 6. StudentAssessment tablosunda bulunan öznitelikler (Features found in StudentAssessmen table)

Öznitelik Adı	Açıklama
id_assessment	Değerlendirmenin kimlik numarası
id_student	Öğrencinin numarası
date_submitted	Öğrencinin modül sunumunun başlangıcından itibaren geçen gün sayısıyla ölçülen teslim tarihi
is_banked	Değerlendirme sonucunun önceki bir sunumdan aktarıldığını gösteren işaretçi
score	Öğrencinin değerlendirmedeki notu

Tablo 7. StudentVle tablosunda bulunan öznitelikler

Öznitelik Adı	Açıklama
code_module	Modül için belirleyici bir kod
code_presentation	Modül sunumunun belirleyici kodu
id_student	Öğrencinin kimlik numarası
id_site	VLE (Sanal Öğrenme Ortamı) materyali için bir kimlik numarası
date	Öğrencinin materyale ulaştığı tarihler
sum_click	Öğrencinin materyalle etkileşim kurma sayısı

3.2. Veri Ön İşleme (Data Preprocessing)

Veri seti üzerinde makine öğrenmesi algoritmaları ile öğrenci başarısının tahmin edilebilmesi için başlangıçta belirli ön işlemelerin yapılması gerekmektedir. Analize geçmeden önce, veri ön işleme aşaması, gereksiz öznitelikleri ortadan kaldırarak, veri türlerini standartlaştırarak ve eksik veri örneklerini işleyerek ham veri kümesini rafine bir forma dönüştürür. Bu hazırlık aşaması, veri temizleme, normalleştirme ve öznitelik seçimi gibi önemli adımları kapsayan sonraki analiz için hayati öneme sahiptir.

3.2.1. Normalleştirme (Normalization)

Verilerin normalleştirilmesi, veri kümelerinin sınıflandırma veya tahmin algoritmaları için eğitilmesinden önce zorunlu bir ön işleme adımıdır. Özellikle rastgele dağıtılan veriler için önemlidir; tekdüze

Tablo 8. Tahminleme işlemi için kullanılacak öznitelikler (Features to use for estimation)

Öznitelikler	Açıklaması
num_of_prev_attempts	Öğrencinin modüle katılma sayısı
final_result	Öğrencinin dersi geçme durumu
weighted_grade	Öğrenciye ait ağırlıklı not ortalaması
pass_rate	Öğrencinin modülü geçme yüzdesi
exam_score	Final notu
date	Materyale erişim süresinin ortalama gün sayısı
sum_click	Materyale yapılan ortalama tıklama sayısı

dağılımın üstün sonuçlar vereceği varsayımıyla [0–1] veya [-1, 1] gibi daha dar aralıklarda ölçeklendirmeyi gerektirir. Bu çalışmada benimsenen yaklaşım Eş. 1'de belirtildiği gibi minimum-maksimum normalleştirme yöntemidir.

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

3.2.2. Eksik verileri silme (Deleting missing data)

Eksik veya mükerrer veri girişleri, veri kümelerinde sık karşılaşılan durumlardır ve genellikle insan hatasından veya veri aktarım sorunlarından kaynaklanır. Veri setinin doğruluğunu ve güvenilirliğini korumak için bu anormalliklerin sistematik olarak tespit edilmesi ve ortadan kaldırılması önemlidir. Bu çalışmada modülden kaydı silinen öğrenciler, öğrenci bilgilerinin yer aldığı StudentInformation tablosundan çıkarılmıştır.

3.2.3. Özniteliklerin seçimi (Features selection)

Öznitelik seçim süreci, hedef öğrenme problemlerini çözmek için veri kümesinden gereksiz ve ilgisiz öznitelikleri kaldırarak uygun bir öznitelik alt kümesini tanımlamayı amaçlamaktadır. Bu tür alakasız öznitelikler, öğrenme doğruluğunu ve model kalitesini olumsuz yönde etkileyebilir. Bu çalışmada öznitelikler önem düzeylerine göre seçilerek hedef öznitelikle karşılaştırılmıştır. Tahminde kullanılacak öznitelikler tek bir tabloda birleştirilmiştir. Bu tablodaki öznitelikler Tablo 8'de verilmiştir.

- İlk olarak, her değerlendirmenin puanı ve ağırlığıyla elde edilen son notlar hesaplanmıştır. Öğrencilerin 40 ve üstü not alması durumu geçer not olarak kabul edilmiş ve öğrencinin başarıyla tamamladığı değerlendirmelerin yüzdesi hesaplanarak ağırlıklı not ortalaması (weighted_grade) özniteliği oluşturularak yeni tabloya eklenmiştir.
- Öğrencilerin finaldeki başarı durumları (final_result) tahmin edilmeye çalışılmıştır.
- Her öğrencinin aldığı tüm derslerden ne kadarını geçtiğini gösteren başarı oranı hesaplanmış ve pass_rate adında öznitelik oluşturularak yeni tabloya eklenmiştir.
- StudentInfo tablosundan kayıtları silinen öğrenciler tespit edilerek, devam eden öğrenciler belirlenmiş ve bu öğrencilere ait exam_score özniteliği, yeni oluşturulan tabloya eklenmiştir.
- Vle ve StudentVle veri setleri, öğrencilerin verilen kaynaklarına etkileşimine dair bilgiler içerir. Bu verilerle öğrencilerin çalışma durumu, içeriğe tıklama sayıları, sistemde kalma süreleri gibi bilgilere ulaşılabilir. Öğrencinin materyali açtığı günler ve materyale erişim sayılarının ortalama değerleri hesaplanarak Date ve sum_click adlarında öznitelikler yeni oluşturulan tabloya eklenmiştir.

Oluşturulan yeni veri seti Tablo 8'de de gösterildiği üzere yedi öznitelikten oluşmaktadır. Veri setinde bulunan final_result dışındaki öznitelikler sayısal veri türüne sahipken, hedef değişken olan final_result özelliği kategorik bir yapıdadır. Final_result özniteliğine göre öğrencilerin başarı durumları incelendiğinde %69'unun "geçtiği" (pass), %18'inin "üstün başarı" (distinction) gösterdiği ve %13'ünün "başarısız" (fail) olduğu görülmektedir. Bu bulgu, genel olarak öğrencilerin büyük çoğunluğunun başarı kriterlerini karşıladığını, belirli bir kısmının ise yüksek düzeyde performans sergileyerek üstün başarı elde ettiğini ortaya koymaktadır. Bununla birlikte, başarısızlık oranı %13 olup, bu durumun eğitim sürecinde ek destek ve müdahale gerektiren bir öğrenci grubuna işaret ettiği söylenebilir. Veri setinde sınıf dağılımları analiz edildiğinde, geçme (pass) sınıfına ait örneklerin daha fazla olduğu, diğer iki sınıfın örnek sayılarının birbirine daha yakın olduğu gözlemlenmektedir.

3.2.4. Kullanılan Sınıflandırma Yöntemleri (Classification Methods)

K-En Yakın Komşu (KNN): Sınıflandırma ve regresyon problemlerine yönelik çözümlerde kullanılan temel bir makine öğrenimi algoritmasıdır. Veri noktalarını uzayda birbirlerine yakınlıklarına göre gruplandırmak ve tahminlerde bulunmak için kullanılır. Eş. 2'deki formül Öklid'e göre mesafenin belirlenmesinde kullanılır [35].

$$d_{(i,y)} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{yk}\right)^2} \quad (2)$$

Karar Ağacı (KA): Veri setindeki özelliklerin değerlerine göre alınan basit kararlarla verileri sınıflandırabilen bir makine öğrenme algoritmasıdır. Karar ağaçları hem sınıflandırma hem de regresyon problemlerinde kullanılabilir. Karar ağacı karar düğümleri, dallar ve yapraklardan oluşmaktadır.

Naive Bayes (NB): Makine öğrenmesi ve istatistiksel sınıflandırma için kullanılan olasılık tabanlı bir algoritmadır. Sınıflandırma sorunlarını çözmek için yaygın olarak kullanılır ve genellikle doğal dil işleme, spam filtreleme, duyarlılık analizi, tıbbi teşhis ve diğer sınıflandırma görevlerinde etkili sonuçlar verir. Bu algoritma, Bayes teoremini temel alır ve sınıf değişkeninin değerini dikkate alarak tüm değişkenlerin bağımsız olduğunu varsayar [36]. Bayes teoreminde kullanılan matematiksel formül Eş. 3'te açıklanmıştır.

$$(A|B) = \big(P(B|A) * P(A)\big)/P(B) \quad (3)$$

Rastgele Orman (RO): Makine öğrenmesi alanında yaygın olarak kullanılan ve sınıflandırma, regresyon, kümeleme gibi çeşitli görevler için etkili sonuçlar sağlayan bir topluluk öğrenme yöntemidir. Rastgele Orman, birçok karar ağacının birleştirilmesi ve bu ağaçların tahminlerinin ortalaması alınarak veya en çok oylanan sınıf veya değere göre sınıflandırma veya regresyon yapılarak oluşturulur. Rastgele Ormanlar, her bir karar ağacının tahminine dayalı varyansı işleyerek tahmin edilen ortalamanın esas gerçeğe (sınıflandırma) veya doğru değere (regresyon) yaklaşacak şekilde işleyebilir [37].

Destek Vektör Makinesi (DVM): Makine öğrenmesi alanında sınıflandırma ve regresyon problemlerinin çözümünde kullanılan

güçlü bir algoritmadır. DVM, uzaydaki veri noktalarını sınıflandıran bir hiper düzlem bulmaya odaklanır ve bu düzlem, sınıflar arasında maksimum marjı elde edecek şekilde optimize edilir. Sınıflar arasındaki marjın maksimuma çıkarılması, DVM'nin yüksek genelleme performansına ve iyi bir ayırt etme gücüne sahip olmasını sağlar.

Lojistik Regresyon (LR): İstatistik ve makine öğrenmesinde sınıflandırma problemlerini çözmek için kullanılan istatistiksel bir modeldir. Temel olarak, bir giriş verisi alır ve bu veriyi bir veya daha fazla kategoriye (sınıflara) atar. Bu nedenle sıklıkla ikili sınıflandırma problemlerinde kullanılır ancak çok sınıflı sınıflandırmaya da genişletilebilir.

Linear Diskriminant Analizi (LDA): İstatistik ve makine öğrenmesi alanlarındaki sınıflandırma problemlerinin çözümünde kullanılan bir örüntü tanıma ve boyut küçültme yöntemidir. LDA, çok boyutlu veri kümelerini daha düşük boyutlu bir alt uzaya dönüştürerek sınıflandırır ve sınıflar arasındaki farkları vurgular.

SMOTE (Synthetic Minority Oversampling Technique): SMOTE, azınlık sınıfının örneklerini artırmak için benzer örnekler arasında sentetik veri oluşturur. Bu yöntem, sınıf dengesizliklerini azaltmak için yaygın olarak kullanılır.

KMeansSMOTE: SMOTE tekniğinin bir geliştirmesidir. Veri kümelerini K-Means algoritması ile gruplara ayırır ve daha anlamlı sentetik veriler oluşturmak için bu grupları kullanır.

RandomOverSampler: Bu yöntem, azınlık sınıfından rastgele örnekler seçerek veri kümesini artırır. Yöntem basit ama aşırı öğrenmeye yatkın olabilir.

ADASYN (Adaptive Synthetic Sampling): Azınlık sınıfı örneklerinin yerel dağılımına göre sentetik örnekler oluşturur. Bu yöntem, öğrenmesi zor olan bölgelere daha fazla sentetik veri eklemeyi hedefler.

BorderlineSMOTE: Yalnızca sınıf sınırına yakın azınlık örnekleri için sentetik veri oluşturur. Bu sayede sınıf ayrımının güçlendirilmesi amaçlanır.

SVMSMOTE: SVM algoritmasını kullanarak azınlık sınıfının kritik sınırlarına yakın örneklerden sentetik veri oluşturur. Bu, daha karmaşık sınıf ayrımları için uygundur.

Edited Nearest Neighbours: Azınlık sınıfı örneklerini çevreleyen çoğunluk sınıfı örneklerini kaldırarak veri temizliği yapar. Bu yöntem sınıflar arası karışıklığı azaltmayı hedefler.

AllKNN: KNN algoritmasını kullanarak veri kümesindeki azınlık sınıfı örneklerine en yakın olan çoğunluk sınıfı örneklerini kaldırır ve gürültüyü azaltır.

NearMiss: Azınlık sınıfının çevresindeki çoğunluk örneklerini seçerek veri kümesini dengeler. Bu yöntem farklı versiyonlarla (NearMiss-1, NearMiss-2, vb.) uygulanabilir.

NeighborhoodCleaningRule: KNN tabanlı bir yaklaşımla sınıflar arası karışıklığı temizlemek için çoğunluk sınıfı örneklerini kaldırır. Daha dengeli bir veri kümesi oluşturur.

OneSidedSelection: Gürültülü ve gereksiz çoğunluk örneklerini kaldırarak veri kümesini küçültür. Bu yöntem, özellikle sınıf sınırını netleştirmek için etkilidir.

RandomUnderSampler: Bu yöntem, çoğunluk sınıfından rastgele örnekler kaldırarak veri kümesini dengeler. Basit bir yöntem olmakla birlikte önemli bilgi kayıplarına yol açabilir.

3.2.5. Performans Metrikleri (Performance Metrics)

Bir sınıflandırıcının başarısını değerlendirmek için Tablo 9'daki karmaşıklık matrisindeki değerler kullanılmaktadır.

Tablo 9. Karmaşıklık matrisi (Confusion matrix)

		Gerçek	
Tahmin		Pozitif	Negatif
	Pozitif	DP	YP
	Negatif	YN	DN

Karmaşıklık matrisinden farklı başarı ölçütleri oluşturulur. Bu çalışma kapsamında kullanılan metrikler Tablo 10'da verilmiştir.

Tablo 10. Performans metrikleri (Performance metrics)

Metrikler	Metriklerin matematiksel ifadesi
Doğruluk	(DP + DN) / (DP + YP + YN + DN)
Kesinlik	DP / (DP + YP)
Duyarlılık	DP / (DP + YN)
F1-Skor	2 * Kesinlik * Duyarlılık / (Kesinlik + Duyarlılık)

Doğruluk metriği bir modelin tahminlerinin ne kadar doğru olduğunu ölçerken, kesinlik metriği doğru sınıflandırılmış pozitif örneklerin toplam pozitif tahmin edilen örnekler arasındaki oranını ifade eder. Duyarlılık metriği, doğru sınıflandırılmış pozitif örneklerin toplam gerçek pozitif örnekler arasındaki oranını verirken, F1 skoru, kesinlik ve duyarlılık metriklerini birleştiren bir performans ölçüsüdür. Kullanılan parametre değerleri grid arama yöntemiyle elde edilmiştir.

4. Sonuçlar ve Tartışmalar (Results and Discussions)

Bu bölümde çalışma kapsamındaki deneysel sonuçlara yer verilmektedir. Veri setinde LR, LDA, RO, KNN, KA, NB, DVM sınıflandırıcıları kullanılmıştır. Yeniden örnekleme olmadan sınıflandırmaya ayrıca altı farklı fazla örnekleme ve yedi farklı az örnekleme tekniği her bir sınıflandırıcı için tek tek uygulanarak Tablo 10'da verilen performans ölçümleri yapılmıştır. Çalışmada kullanılan parametreler ve değerleri Tablo 11'de verilmiştir.

Tablo 11. Sınıflandırıcılar için kullanılan parametreler ve değerler (Parameters and values used for classifiers)

Sınıflandırıcı	Kullanılan parametreler ve değerler
LR	max_iter=250
LDA	Default
RO	n_estimators =100
KNN	n_neighbors=13
KA	Default
NB	Default
DVM	Default

LR modelinde max_iter=250 parametresi kullanılmıştır. Bu seçim, modelin yakınsaması için yeterli sayıda iterasyon yapılmasını sağlamak amacıyla tercih edilmiştir. LDA genellikle hiperparametre ayarı gerektirmeyen ve doğrusal sınırlarla çalışan bir algoritma olduğundan, varsayılan değerler çoğu durumda yeterlidir. Bu sebeple varsayılan ayarlar kullanılmıştır. RO algoritmasında n_estimators=100 parametresi iyi bir denge sağladığı için bu parametre seçilmiştir. KNN algoritmasında ise n_neighbors=13 olan

Tablo 12. LR için sınıflandırma sonuçları (Classification results for LR)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1-Skor
Örneklemsiz	-	0,87	0,84	0,81	0,82
Fazla örnekleme	SMOTE	0,87	0,87	0,86	0,87
	Kmeans SMOTE	*0,90*	*0,896*	*0,896*	*0,896*
	RandomOverSampler	0,85	0,853	0,853	0,85
	ADASYN	0,84	0,84	0,843	0,84
	BorderLineSMOTE	0,84	0,84	0,84	0,833
	SVMSMOTE	0,86	0,853	0,856	0,853
Az örnekleme	EditedNearestNeighbours	*0,93*	*0,93*	*0,90*	*0,913*
	AllKNN	0,90	0,91	0,843	0,87
	NearMiss	0,83	0,83	0,83	0,83
	NeighbourhoodCleaningRule	0,84	0,846	0,843	0,84
	OneSidedSelection	0,88	0,786	0,666	0,71
	RandomUnderSampler	0,86	0,856	0,86	0,856
	TomekLinks	0,89	0,866	0,816	0,84

Tablo 13. LDA için sınıflandırma sonuçları (Classification results for LDA)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1-Skor
Örneklemsiz	-	0,89	0,81	0,85	0,81
Fazla örnekleme	SMOTE	0,86	0,86	0,86	0,86
	Kmeans SMOTE	*0,88*	*0,89*	*0,883*	*0,876*
	RandomOverSampler	0,86	0,86	0,86	0,853
	ADASYN	0,83	0,84	0,833	0,82
	BorderLineSMOTE	0,83	0,846	0,833	0,82
	SVMSMOTE	0,85	0,86	0,846	0,836
Az örnekleme	EditedNearestNeighbours	*0,95*	*0,93*	*0,93*	*0,923*
	AllKNN	*0,94*	*0,923*	*0,916*	*0,916*
	NearMiss	0,89	0,893	0,89	0,893
	NeighbourhoodCleaningRule	0,84	0,85	0,836	0,826
	OneSidedSelection	0,89	0,806	0,726	0,76
	RandomUnderSampler	0,86	0,866	0,863	0,856
	TomekLinks	0,90	0,853	0,873	0,863

orta büyüklükte bir değer seçilmiştir. Bu parametre, karar verme sürecinde kaç komşunun dikkate alınacağını ifade etmektedir. KA ve NB genellikle başlangıç için hiperparametre optimizasyonu gerektirmediği için, DVM algoritmasında da varsayılan parametrelerin çoğu veri seti için iyi sonuç verdiği sebebi ile varsayılan ayarlar kullanılmıştır.

Veri setinde sınıf dağılımları analiz edildiğinde, geçme (pass) sınıfına ait örneklerin daha fazla olması ve üstün başarı (distinction) ile başarısız (fail) sınıflarının örnek sayılarının birbirine daha yakın olması veri setindeki dağılımlar arasında dengesizlik oluşturmaktadır. Bu durumun sınıflandırma modellerinin ayırt etme yeteneğinde sorunlara neden olması muhtemeldir.

Makine öğreniminde veri arttırma ve veri azaltma, dengesiz veri kümelerindeki sınıf dengesizliğini gidermek için kullanılan bir yöntemdir. Bu nedenle her sınıflandırıcı ile fazla örnekleme ve az örnekleme teknikleri uygulanmıştır. Bu çalışmada fazla örnekleme için; SMOTE [38], KMeansSMOTE [39], RandomOverSampler [40], ADASYN [41], BorderlineSMOTE [42] ve SVMSMOTE [43]; az örnekleme için: EditedNearestNeighbours [44], AllKNN [45], NearMiss [46], NeighborhoodCleaningRule [47], OneSidedSelection [48], RandomUnderSampler [49] teknikleri uygulanmıştır. LR, LDA, RO, KNN, KA, NB ve DVM olarak yedi farklı makine öğrenme tekniği kullanılmıştır. Yeniden örnekleme olmadan sınıflandırmaya ayrıca altı farklı fazla örnekleme ve yedi farklı az örnekleme tekniği kullanılarak doksan bir sınıflandırma işlemi yapılmıştır. Modellerin performansı doğruluk, kesinlik, duyarlılık ve F1-Skoru metrikleriyle ölçülerek kıyaslanmıştır.

Her sınıflandırma işleminin sonuçları Tablo 10'da verilen metriklerle ölçülmüştür. Sınıflandırma işlemlerinde 5 katlı çapraz doğrulama yöntemi uygulanmış ve sonuçların ortalamaları sunulmuştur. Elde edilen sonuçlar LR için Tablo 12'de, LDA için Tablo 13'te, RO için Tablo 14'te, KNN için Tablo 15'te, KA için Tablo 16'da, NB için Tablo 17'de, DVM için Tablo 18'dedir. En başarılı sonuçlar Tablo 19'dadır.

Tablo 12'de LR için yapılan sınıflandırma sonuçları değerlendirildiğinde bazı yeniden örnekleme tekniklerinin başarıyı arttırdığı görülmektedir. LR için örneklemsiz değerlendirmeyle elde edilen doğruluk %87, kesinlik %84, duyarlılık %81, F1-Skor %82'dir. Fazla örnekleme tekniklerinden Kmeans SMOTE'tan %90 doğruluk, %89,6 kesinlik, %89,6 duyarlılık, %89,6 F1-Skor ile yüksek değerler elde edilmiştir. Az örnekleme tekniklerinden AllKNN'den doğruluk %90, kesinlik %91, duyarlılık %84,3, F1-Skor %87 olarak yüksek değerler elde edilmiştir.

Tablo 14. RO için sınıflandırma sonuçları (Classification results for RO)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1-Skor
Örneklemsiz	-	0,96	0,87	0,87	0,85
Fazla örnekleme	SMOTE	0,94	0,936	0,94	0,936
	Kmeans SMOTE	0,95	0,95	0,95	0,95
	RandomOverSampler	*0,97*	*0,96*	*0,96*	*0,96*
	ADASYN	0,93	0,936	0,933	0,933
	BorderLineSMOTE	0,93	0,933	0,93	0,89
	SVMSMOTE	0,94	0,936	0,936	0,933
Az örnekleme	EditedNearestNeighbours	*0,96*	*0,933*	*0,933*	*0,953*
	AllKNN	*0,96*	*0,926*	*0,92*	*0,943*
	NearMiss	0,91	0,91	0,906	0,906
	NeighbourhoodCleaningRule	0,94	0,936	0,936	0,933
	OneSidedSelection	0,92	0,886	0,77	0,816
	RandomUnderSampler	0,89	0,883	0,883	0,886
	TomekLinks	0,93	0,93	0,876	0,90

Tablo 15. KNN için sınıflandırma sonuçları (Classification results for KNN)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1-Skor
Örneklemsiz	-	0,87	0,87	0,77	0,81
Fazla örnekleme	SMOTE	*0,91*	*0,91*	*0,91*	*0,91*
	Kmeans SMOTE	*0,91*	*0,93*	*0,933*	*0,93*
	RandomOverSampler	*0,88*	*0,876*	*0,876*	*0,876*
	ADASYN	0,90	0,903	0,896	0,89
	BorderLineSMOTE	0,90	0,91	0,906	0,903
	SVMSMOTE	*0,91*	*0,913*	*0,906*	*0,906*
Az örnekleme	EditedNearestNeighbours	0,94	0,956	0,906	0,93
	AllKNN	*0,96*	*0,92*	*0,926*	*0,93*
	NearMiss	0,87	0,87	0,866	0,866
	NeighbourhoodCleaningRule	0,91	0,913	0,906	0,906
	OneSidedSelection	0,90	0,873	0,77	0,813
	RandomUnderSampler	*0,87*	*0,863*	*0,87*	*0,866*
	TomekLinks	0,90	0,883	0,85	0,863

Tablo 13'te LDA'ya ilişkin sınıflandırma sonuçları değerlendirildiğinde bazı yeniden örnekleme tekniklerinin başarıyı artırdığı görülmektedir. LDA için örneklemsiz değerlendirmeyle elde edilen doğruluk %89, kesinlik %81, duyarlılık %85, F1-Skor %81'dir. Fazla örnekleme tekniklerinden, Kmeans SMOTE'tan %88 doğruluk, %89 kesinlik, %88,3 duyarlılık, %87,6 F1-Skor ile yüksek değerler elde edilmiştir. Az örnekleme tekniklerinden EditedNearestNeighbours'tan doğruluk %95, kesinlik %93, duyarlılık %93, F1-Skor %92,3 olarak yüksek değerler elde edilmiştir.

Tablo 14'te RO için yapılan sınıflandırma sonuçları değerlendirildiğinde bazı yeniden örnekleme tekniklerinin başarıyı arttırdığı görülmektedir. RO için örneklemsiz değerlendirmeyle elde edilen doğruluk %96, kesinlik %87, duyarlılık %87, F1-Skor %85'tir. Fazla örnekleme tekniklerinden yüksek değerler RandomOverSampler'dan %97 doğruluk, %96 kesinlik, %96 duyarlılık, %96 F1-Skor ile elde edilmiştir. Az örnekleme tekniklerinden EditedNearestNeighbours'tan doğruluk %96, kesinlik %93,3, duyarlılık %93,3, F1-Skor %95,3 olarak yüksek değerler elde edilmiştir.

Tablo 15'te KNN için yapılan sınıflandırma sonuçları değerlendirildiğinde bazı yeniden örnekleme tekniklerinin başarıyı arttırdığı görülmektedir. KNN için örneklemsiz değerlendirmeyle elde edilen doğruluk %87, kesinlik %87, duyarlılık %77, F1-Skor %81'dir. Fazla örnekleme tekniklerinden, Kmeans SMOTE'tan %91 doğruluk, %93 kesinlik, %93,3 duyarlılık, %93 F1-Skor ile yüksek değerler elde edilmiştir. Az örnekleme tekniklerinden AllKNN'den doğruluk %96, kesinlik %92, duyarlılık %92,6, F1-Skor %93 olarak yüksek değerler elde edilmiştir.

Tablo 16'da KA'ya ilişkin sınıflandırma sonuçları değerlendirildiğinde bazı yeniden örnekleme tekniklerinin başarıyı arttırdığı görülmektedir. KA için örneklemsiz değerlendirme ile elde edilen doğruluk %86, kesinlik %81, duyarlılık %83, F1-Skor %80'dir. Fazla örnekleme tekniklerinden yüksek değerler RandomOverSampler'dan %94,7 doğruluk, %95,6 kesinlik, %95,6 duyarlılık, %95,3 F1-Skor ile elde edilmiştir. Az örnekleme tekniklerinden EditedNearestNeighbours'tan doğruluk %94, kesinlik %93,6, duyarlılık %92,3, F1-Skor %93,3 olarak yüksek değerler elde edilmiştir.

Tablo 17'deki NB sınıflandırma sonuçları değerlendirildiğinde bazı yeniden örnekleme tekniklerinin başarıyı arttırdığı görülmektedir. NB için örneklemsiz değerlendirmeyle elde edilen doğruluk %82, kesinlik %75, duyarlılık %82, F1-Skor %78'dir. Fazla örnekleme tekniklerinden Kmeans SMOTE'tan %88 doğruluk, %87,6 kesinlik, %87,6 duyarlılık, %88 F1-Skor ile yüksek değerler elde edilmiştir. Az örnekleme tekniklerinden EditedNearestNeighbours'tan doğruluk %93, kesinlik %92, duyarlılık %92,3, F1-Skor %92,3 olarak yüksek değerler elde edilmiştir.

Tablo 16. KA için sınıflandırma sonuçları (Classification results for KA)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1-Skor
Örneklemsiz	-	0,86	0,81	0,83	0,80
Fazla örnekleme	SMOTE	0,91	0,906	0,91	0,913
	Kmeans SMOTE	0,92	0,923	0,923	0,923
	RandomOverSampler	0,947	0,956	0,956	0,953
	ADASYN	0,91	0,906	0,913	0,913
	BorderLineSMOTE	0,91	0,91	0,91	0,91
	SVMSMOTE	0,91	0,906	0,91	0,906
Az örnekleme	EditedNearestNeighbours	*0,94*	*0,936*	*0,923*	*0,933*
	AllKNN	0,93	0,933	0,92	0,93
	NearMiss	0,87	0,87	0,87	0,873
	NeighbourhoodCleaningRule	0,90	0,90	0,90	0,90
	OneSidedSelection	0,86	0,763	0,766	0,766
	RandomUnderSampler	*0,84*	*0,846*	*0,84*	*0,843*
	TomekLinks	0,89	0,856	0,833	0,843

Tablo 17. NB için sınıflandırma sonuçları (Classification results for NB)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1-Skor
Örneklemsiz	-	0,82	0,75	0,82	0,78
Fazla örnekleme	SMOTE	0,86	0,863	0,86	0,86
	Kmeans SMOTE	*0,88*	*0,876*	*0,876*	*0,88*
	RandomOverSampler	0,85	0,846	0,846	0,843
	ADASYN	0,82	0,816	0,816	0,806
	BorderLineSMOTE	0,83	0,84	0,836	0,823
	SVMSMOTE	0,86	0,86	0,856	0,85
Az örnekleme	EditedNearestNeighbours	*0,93*	*0,92*	*0,923*	*0,923*
	AllKNN	0,92	0,903	0,906	0,903
	NearMiss	0,86	0,866	0,856	0,856
	NeighbourhoodCleaningRule	0,83	0,836	0,833	0,823
	OneSidedSelection	0,86	0,736	0,713	0,723
	RandomUnderSampler	0,85	0,843	0,846	0,843
	TomekLinks	0,87	0,806	0,846	0,826

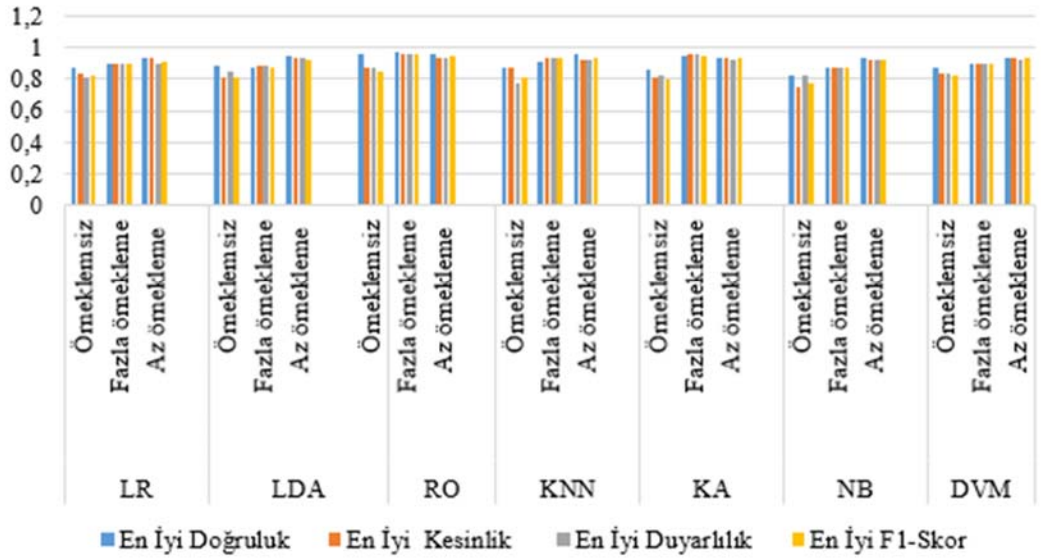
Tablo 18. DVM için sınıflandırma sonuçları (Classification results for DVM)

Örnekleme	Kullanılan Teknik	Doğruluk	Kesinlik	Duyarlılık	F1-Skor
Örneklemsiz	-	0,88	0,84	0,84	0,82
Fazla örnekleme	SMOTE	0,88	0,876	0,88	0,876
	Kmeans SMOTE	*0,90*	*0,896*	*0,896*	*0,90*
	RandomOverSampler	0,88	0,876	0,876	0,876
	ADASYN	0,84	0,843	0,843	0,833
	BorderLineSMOTE	0,85	0,856	0,853	0,843
	SVMSMOTE	0,88	0,88	0,88	0,876
Az örnekleme	EditedNearestNeighbours	0,93	0,93	0,92	0,936
	AllKNN	0,92	0,916	0,896	0,916
	NearMiss	0,89	0,89	0,886	0,886
	NeighbourhoodCleaningRule	0,86	0,86	0,86	0,85
	OneSidedSelection	0,90	0,90	0,713	0,766
	RandomUnderSampler	0,87	0,87	0,87	0,866
	TomekLinks	0,91	0,89	0,843	0,866

Tablo 18'deki DVM sınıflandırma sonuçları değerlendirildiğinde bazı yeniden örnekleme tekniklerinin başarıyı arttırdığı görülmektedir. DVM için örneklemsiz değerlendirme ile elde edilen doğruluk %88, kesinlik %84, duyarlılık %84, F1-Skor %82'dir. Fazla örnekleme tekniklerinden Kmeans SMOTE'tan %90 doğruluk, %89,6 kesinlik, %89,6 duyarlılık, %90 F1-Skor ile yüksek değerler elde edilmiştir. Az örnekleme tekniklerinden EditedNearestNeighbours'tan doğruluk %93, kesinlik %93, duyarlılık %92, F1-Skor %93,6 olarak yüksek değerler elde edilmiştir.

Tablo 19. En başarılı sonuçlar (The most successful results)

Örnekleme	Kullanılan Teknik	En İyi Doğruluk	En İyi Kesinlik	En İyi Duyarlılık	En İyi F1-Skor
LR	Örneklemsiz	0,87	0,84	0,81	0,82
	Fazla örnekleme	0,90	0,896	0,896	0,896
	Az örnekleme	0,93	0,93	0,90	0,913
LDA	Örneklemsiz	0,89	0,81	0,85	0,81
	Fazla örnekleme	0,88	0,89	0,883	0,876
	Az örnekleme	0,95	0,93	0,93	0,923
RO	Örneklemsiz	0,96	0,87	0,87	0,85
	Fazla örnekleme	0,97	0,96	0,96	0,96
	Az örnekleme	0,96	0,933	0,933	0,953
KNN	Örneklemsiz	0,87	0,87	0,77	0,81
	Fazla örnekleme	0,91	0,93	0,933	0,93
	Az örnekleme	0,96	0,92	0,926	0,93
KA	Örneklemsiz	0,86	0,81	0,83	0,80
	Fazla örnekleme	0,947	0,956	0,956	0,953
	Az örnekleme	0,94	0,936	0,923	0,933
NB	Örneklemsiz	0,82	0,75	0,82	0,78
	Fazla örnekleme	0,88	0,876	0,876	0,88
	Az örnekleme	0,93	0,92	0,923	0,923
DVM	Örneklemsiz	0,88	0,84	0,84	0,82
	Fazla örnekleme	0,90	0,896	0,896	0,90
	Az örnekleme	0,93	0,93	0,92	0,936



Şekil 1. Farklı Sınıflandırma Algoritmalarının Örnekleme Tekniklerine Göre Performans Karşılaştırması
(Performance Comparison of Different Classification Algorithms Based on Sampling Techniques)

Tablolardaki ölçümler değerlendirildiğinde bazı yeniden örnekleme tekniklerinin başarıyı arttırdığı görülmektedir. Yeniden örnekleme başarı oranları farklılık gösterse de bu çalışmada fazla örnekleme tekniklerinden RandomOverSampler, Kmeans SMOTE, az örnekleme tekniklerinden EditedNearestNeighbors ve AIIKNN tekniklerinin belirgin bir başarı artışının sağlandığı tespit edilmiştir. Kullanılan tüm sınıflandırıcılar ve örnekleme yöntemlerinden en başarılı olanları Tablo 19'da özetlenmiştir. Tablo 19'daki sonuçlar incelendiğinde RO yeniden örnekleme ile yüksek değerlere ulaşıldığı, tüm parametreler açısından en yüksek değerlerin elde edildiği görülmektedir. RO ile elde edilen en yüksek doğruluk değerine RandomOverSampler fazla örnekleme tekniği ile ulaşılmıştır. RO ile tüm sınıflandırmalarda elde edilen en iyi sonuçlar doğrulukta %97, kesinlikte %96, duyarlılık %96 ve F1 Skorunda %96'dır. RO sınıflandırıcıdan sonra en yüksek ölçümler KNN ile elde edilmiştir. Bu sınıflandırıcıyla elde edilen en üst düzey doğruluk değeri AIIKNN az örnekleme tekniği ile elde edilmiştir. KNN ile tüm sınıflandırmalarda elde edilen en iyi sonuçlar doğrulukta %96, kesinlikte %92, duyarlılıkta %92,6 ve F1 Skorunda %93'tür. Tablo 19'daki verilere ait farklı sınıflandırma algoritmalarının (LR, LDA, RO, KNN, KA, NB, DVM) örnekleme teknikleri (örneklemsiz, fazla örnekleme, az örnekleme) altında doğruluk, kesinlik, duyarlılık ve F1-Skor metrikleri bakımından performanslarına ait grafik Şekil 1'de verilmiştir.

Çalışma kapsamında LR, LDA, RO, KNN, KA, NB, DVM sınıflandırıcıları ile uygulanan yeniden örnekleme yöntemine sahip modellerden önerilen en iyi doğruluk değerleri Tablo 20'de verilmiştir.

Tablo 20. Bulguların literatürdeki çalışmalarla karşılaştırılması (Comparison of the findings with the studies in the literature)

Referans	Kullanılan yöntemler	En iyi doğruluk (%)
(Alkan, [50])	LR	87
	LDA	95
	RO	96
	KNN	87
	KA	86
	NB	82
	DVM	88
	Literatür ortalaması	88,71429
	Önerilen LR	93
	Önerilen LDA	95
	Önerilen RO	97
	Önerilen KNN	96
	Önerilen KA	94
	Önerilen NB	93
	Önerilen DVM	93
	Önerilen modellerin ortalaması	94,42857

Çalışmada kullanılan veri seti ile ilgili çalışmalar literatürde sınırlıdır. Ancak bu çalışmada kullanılan yeniden örnekleme ile daha yüksek başarı elde edilmiştir. Bir önceki çalışmada en yüksek doğruluk değerleri %96 iken bunu %95 ve %88 takip etmiştir. Ortalama başarı ise %88,71429'dur. Bu çalışmada; en yüksek doğruluk değerleri %97 olup, bunu %96 ve %95 takip etmektedir. Ortalama başarı da %94,42857'dir. Çalışma kapsamında RO en yüksek doğruluğa sahip olup onu KNN takip etmektedir. RO'da fazla örnekleme yöntemlerinden RandomOverSampler ile KNN'de az örnekleme yöntemlerinden AIIKNN ile yüksek başarı elde edilmiştir.

5. Sonuçlar (Conclusions)

Bu çalışmada, öğrenme ortamlarının daha verimli hale getirilmesi ve öğrencilerin olası başarısızlıklarının önceden tahmin edilmesi hedeflenmiştir. Bu amaçla, farklı sınıflandırıcılar kullanılarak öğrencilerin verileriyle dersten geçme durumları tahmin edilmeye çalışılmıştır. Çalışma kapsamında kamuya açık olan "Open University Learning Analytics Dataset -OULAD" kullanılmıştır. Veri seti üzerinde sınıflandırma işlemlerinde LR, LDA, RO, KNN, KA, DVM ve NB yöntemleri kullanılmış ve performansları değerlendirilmiştir. Çalışmada yedi farklı yöntem kullanılmış ve her bir yöntem doğruluk, kesinlik, duyarlılık ve F1-Skoru değerleri karşılaştırılarak incelenmiştir. Tahmin işleminde kullanılacak öznitelikler tek bir tabloda birleştirilmiştir. LR, LDA, RO, KNN, KA, DVM ve NB makine öğrenmesi modelleri için on üç farklı yeniden örnekleme ile elde edilen sınıflandırma başarıları karşılaştırılmıştır. Yeniden örnekleme yöntemlerinin doğruluk, kesinlik, duyarlılık ve F1-Skoru olmak üzere dört metrikle toplam doksan bir ayrı sınıflandırmanın başarısı değerlendirilmiştir. Bu çalışmada RO'da fazla örnekleme yöntemlerinden RandomOverSampler ile %97 doğruluk elde edilirken, KNN'de az örnekleme yöntemlerinden AIIKNN ile %96 doğruluk elde edilmiştir. Veri setinde sınıf dengesini sağlamak amacıyla yeniden örnekleme yapılmasının sınıflandırıcı başarısını arttırdığı görülmüştür. Bu çalışmada fazla örnekleme tekniklerinden RandomOverSampler ve Kmeans SMOTE ile az örnekleme tekniklerinden EditedNearestNeighbors ve AIIKNN ile yüksek başarı artışı sağlamıştır. Veri setindeki sınıfların dengesizliği modellerin başarısını etkileyen bir faktördür. Mevcut veri setinin uygun örnekleme teknikleriyle dengelenmesi modellerin daha başarılı olmasını sağlamaktadır, çalışmada bu durumu destekler niteliktedir. (Sevli, [51])'in araştırmasında da, en üstün performans %96,296 doğrulukla elde edilmiştir. Bu başarı, RO ile birlikte kullanılan InstanceHardnessThreshold fazla örnekleme tekniğiyle gerçekleştirilen sınıflandırma işleminden elde edilmiştir. (Börekçi ve Sevli [52])'in, SMOTE, KMeansSMOTE, ADASYN ve SVMSMOTE fazla örnekleme tekniklerini kullandıkları çalışmalarında ise, en yüksek doğruluk değerini %99,17 ile Light Gradient Boosting makine sınıflandırıcısının SMOTE fazla örnekleme tekniği ile birlikte kullanılması durumundan elde edilmiştir. Başka bir araştırmada LR, DVM, KA, RO ve Extra Tree Regressor (ETR) gibi farklı algoritmaların performansları karşılaştırılarak yeraltı suyunun akış tahmini yapılmaya çalışılmıştır [53]. Uygulanan modellerin her birinin başarı performansını artırmak amacıyla dalgacık dönüşümü kullanarak hibrit bir yöntem geliştirilmiş ve ETR en başarılı tahmin yöntemi olarak saptanmıştır. Bir diğer çalışmada ise, doğrusal regresyon, KA, RO, GBT regresyonu, izotonik regresyon gibi farklı algoritmalar kullanılmıştır [54]. Çalışma kapsamında, Türkiye'de ikinci el araçların büyük veri ve makine öğrenme teknikleriyle analizi ve fiyat tahmini yapılmış ve algoritmalar karşılaştırılmıştır. En yüksek başarı RO algoritmasından elde edilmiştir.

Bu çalışmada farklı makine öğrenmesi modellerinin başarıları karşılaştırılarak yeniden örnekleme teknikleriyle elde edilen sonuçlar modellerin performans durumu hakkında fikir oluşturmuştur. Bu unsurlar üzerinde çalışmanın ve öğrencilere yeni fırsatlar sunmanın başarıyı artıracağı söylenebilir. Araştırma sonuçlarına dayanarak aşağıdaki önerilerde bulunulabilir:

- Veri setlerinde dengesizlik durumlarına göre hem fazla hem de az örnekleme tekniklerinin birlikte kullanımı teşvik edilmelidir. Bu yaklaşım, model başarısını artırmada daha etkili sonuçlar sağlayabilir.
- Eğitim süreçlerinde ve öğrenme analitiklerinde veri dengesizliği sorununa odaklanılmalı ve uygun örnekleme yöntemleri uygulanmalıdır.
- Veri dengesizliğini gidermek için yeniden örnekleme tekniklerinin (RandomOverSampler, KMeansSMOTE, EditedNearestNeighbors, vb.) etkin bir şekilde kullanılması önerilir. Bu, sınıflandırma algoritmalarının performansını önemli ölçüde artırmaktadır.
- Modellerin doğruluk, kesinlik, duyarlılık ve F1-Skoru gibi performans metrikleri düzenli olarak analiz edilmeli ve daha yüksek performans sunan modeller tercih edilebilir.
- Eğitim yöneticilerine ve öğretmenlere, bu modellerin nasıl kullanılacağı ve tahmin sonuçlarının nasıl değerlendirileceği konusunda eğitim verilebilir.

- Öğretmenlerin, tahmin sonuçlarını ders planlarını uyarlamak ve öğrencilere daha etkili destek sağlamak için kullanmaları teşvik edilebilir.
- Öğrenci başarısını tahmin eden sistemler, erken uyarı sistemleri olarak kullanılmalı ve öğrencilere destek sunulabilir.
- Tahmin sonuçlarına dayanarak öğrencilere özel öğrenme materyalleri, rehberlik ve bireysel destek sağlanabilir.

Kaynaklar (References)

1. Crooks, T. J., The impact of classroom evaluation practices on students. Review of Educational Research, 58 (4), 438-481, 1988.
2. Luan, H., Tsai, C.-C., A review of using machine learning approaches for precision education. Educational Technology & Society, 24 (1), 250-266, 2021.
3. Agarwal, S., Data mining: data mining concepts and techniques, 2013 International Conference on Machine Intelligence and Research Advancement, 203-207.
4. Greller, W., Drachsler, H., Translating learning into numbers: a generic framework for learning analytics, Journal of Educational Technology & Society, 15 (3), 42-57, 2012.
5. Ledley, R. S., Using electronic computers in medical diagnosis, IRE Transactions on Medical Electronics, ME-7 (4), 274-280, 1960.
6. Kalıkov, A., Veri madenciliği ve bir e-ticaret uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2006.
7. Polat, A., Açık öğretim liseleri öğrencilerinin okul terki ve mezuniyet durumlarının eğitsel veri madenciliği ile incelenmesi, Doktora Tezi, Sakarya Üniversitesi, Eğitim Bilimleri Enstitüsü, Sakarya, 2021.
8. Turing, A., Computing machinery and intelligence. Mind, 49 (236), 433–460, 1950.
9. Alpaydin, E., Introduction to Machine Learning, fourth edition. MIT Press, 2020.
10. Koza, J. R., Bennett, F. H., Andre, D., Keane, M. A., Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming, J. S. Gero & F. Sudweeks (Ed.), Artificial Intelligence in Design 96, Springer Netherlands, 1996.
11. Lantz, B., Machine Learning with R: Expert techniques for predictive modeling, Packt Publishing Ltd., 2019.
12. Iatrellis, O., Savvas, I. K., Fitsilis, P., Gerogiannis, V. C., A two-phase machine learning approach for predicting student outcomes, Education and Information Technologies, 26, 69-88, 2021.
13. Çakıt, E., Dağdeviren, M., Predicting the percentage of student placement: A comparative study of machine learning algorithms, Education and Information Technologies, 27 (1), 997-1022, 2022.
14. Erdoğan, Ş., Timor, M., A data mining application in a student database, Havacılık ve Uzay Teknolojileri Dergisi, 2 (2), 53-57, 2005.
15. Vandamme, J.-P., Meskens, N., Superby, J.-F., Predicting academic performance by data mining methods, Education Economics, 15 (4), 405, 2007.
16. Ibrahim, Z., Rusli, D., Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression. 21st Annual SAS Malaysia Forum, 5th September, 2007.
17. Karabatak, M., Özellik Seçimi, Sınıflama ve Öngörü Uygulamalarına Yönelik Birliktelik Kuralı Çıkarımı ve Yazılım Geliştirilmesi, Doktora Tezi, Fırat Üniversitesi, Elazığ, 2008.
18. Dekker, G. W., Pechenizkiy, M., Vleeshouwers. J.M., Predicting students drop out: a case study. *EDM'09* - Educ. Data Min. 2009 2nd Int. Conf. Educ. Data Min. 41-50, 2009.
19. Bozkır, A. S., Sezer, E., Gök, B., Öğrenci Seçme Sınavında (ÖSS) öğrenci başarımını etkileyen faktörlerin veri madenciliği yöntemleriyle tespiti. 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), 13-15 Mayıs, Karabük Üniversitesi, Karabük, 37-43, 2009.
20. Delen, D., A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49 (4), 498-506, 2010.
21. Mishra, T., Kumar, D., Gupta, S., Mining students' data for prediction performance, 2014 Fourth International Conference on Advanced Computing & Communication Technologies, 255-262, 2014.
22. Sivakumar, S., Venkataraman, S., Selvaraj, R., Predictive modeling of student dropout indicators in educational data mining using improved decision tree, Indian Journal of Science and Technology, 9 (4), 1-5, 2016.
23. Schatzel, K., Callahan, T., Scott, C. J., Davis, T., Reaching the non-traditional stopout population: A segmentation approach, Journal of Marketing for Higher Education, 21 (1), 47-60, 2011.
24. Şen, B., Uçar, E., Delen, D. Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 39 (10), 9468-9476, 2012.
25. Djulovic, A., Li, D., Towards freshman retention prediction: A comparative study, International Journal of Information and Education Technology, 3 (5), 494-500, 2013.
26. Iam-On, N., Boongoen, T., Generating descriptive model for student dropout: A review of clustering approach. Human-centric Computing and Information Sciences, 7 (1), 1-24, 2017.
27. Chung, J. Y., Lee, S., Dropout early warning systems for high school students using machine learning, Children and Youth Services Review, 96, 346-353, 2019.
28. Trakunphutthirak, R., Lee, V. C., Application of educational data mining approach for student academic performance prediction using progressive temporal data, Journal of Educational Computing Research, 60 (3), 742-776, 2022.
29. Orji, F. A., Vassileva, J., Predicting the Persuasiveness of Influence Strategies From Student Online Learning Behaviour Using Machine Learning Methods, Journal of Educational Computing Research, 2023.
30. Genç, Ş., Surer, E., ClickbaitTR: Dataset for clickbait detection from Turkish news sites and social media with a comparative analysis via machine learning algorithms, Journal of Information Science, 49 (2), 480-499, 2023.
31. Al-Emran, M., Al-Nuaimi, M. N., Arpaci, I., Al-Sharafi, M. A., Anthony Jnr, B., Towards a wearable education: Understanding the determinants affecting students' adoption of wearable technologies using machine learning algorithms, Education and Information Technologies, 28 (3), 2727-2746, 2023.
32. Hussain, S., Khan, M. Q., Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning, Annals of data science, 10 (3), 637-655, 2023.
33. José-García, A., Sneyd, A., Melro, A., Ollagnier, A., Tarling, G., Zhang, H., ..., Arthur, R., C3-IoC: A career guidance system for assessing student skills using machine learning and network visualisation, International Journal of Artificial Intelligence in Education, 33 (4), 1092-1119, 2023.
34. Students Adaptability Level in Online Education. https://www.kaggle.com/datasets/rocki37/open-university-learning-analytics-dataset. Erişim tarihi Mart 01, 2023.
35. Kilinç, D., Borandağ, E., Yücalar, F., Tunali, V., Şimşek, M., Özçift, A., KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi, Marmara Fen Bilimleri Dergisi, 28 (3), Article 3, 2016.
36. Dimitoglou, G., Adams, J. A., Jim, C. M., Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability, arXiv preprint arXiv:1206.1121, 2012.
37. Liang, Y.C., Maimury, Y., Chen, A. H.L., Juarez, J. R. C., Machine learning-based prediction of air quality, Applied Sciences, 10 (24), Article 24, 2020.
38. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321-357, 2002.
39. Douzas, G., Bacao, F., Last, F., Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE, Information Sciences, 465, 1-20, 2018.
40. Drummond, C., Holte, R. C., C4. 5, Class imbalance, and cost sensitivity: Why under-sampling beats over-sampling, Workshop on learning from imbalanced datasets II, 11, 1-8, 2003.
41. He, H., Bai, Y., Garcia, E. A., Li, S., ADASYN: Adaptive synthetic sampling approach for imbalanced learning, 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), 1322-1328, 2008.
42. Han, H., Wang, W.-Y., Mao, B.-H., Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1, 878-887, 2005.
43. Nguyen, H. M., Cooper, E. W., Kamei, K., Borderline over-sampling for imbalanced data classification, International Journal of Knowledge Engineering and Soft Data Paradigms, 3 (1), 4-21, 2011.

44. Wilson, D. L., Asymptotic properties of nearest neighbor rules using edited data, IEEE Transactions on Systems, Man, and Cybernetics, 3, 408-421, 1972.
45. Tomek, I., An Experıment Wıth The Edıted Nearest-Nıeghbor Rule, 1976.
46. Mani, I., Zhang, I., KNN approach to unbalanced data distributions: A case study involving information extraction. Proceedings of workshop on learning from imbalanced datasets, 126, 1-7, 2003.
47. Laurikkala, J., Improving identification of difficult small classes by balancing class distribution, Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, Proceedings 8, 63-66, 2001.
48. Kubat, M., Matwin, S., Addressing the curse of imbalanced training sets: One-sided selection. Icml, 97 (1), 179, 1997.
49. Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., Napolitano, A., Using random undersampling to alleviate class imbalance on tweet sentiment data, 2015 IEEE international conference on information reuse and integration, 197-202, 2015.
50. Alkan, A., Akademik başarının eğitsel veri madenciliği ve makine öğrenmesi teknikleri kullanılarak tahminlenmesi, Yüksek Lisans Tezi, Mehmet Akif Ersoy Üniversitesi, Fen Bilimleri Enstitüsü, Burdur, 2023.
51. Sevli O., Diagnosis of diabetes mellitus using various classifiers, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (2), 989-1002, 2022.
52. Börekci A., Sevli O., A classification study for Turkish folk music makam recognition using machine learning with data augmentation techniques, Neural Computing and Applications, 36 (4), 1621-1639, 2024.
53. Akbulut U., Çifçi M.A., İşler B., Aslan Z., Comparison of different machine learning techniques in river flow prediction, Journal of the Faculty of Engineering and Architecture of Gazi University, 40 (1), 467-486, 2025
54. Gülmez B., Kulluk S., Analysis and price prediction of secondhand vehicles in Türkiye with big data and machine learning techniques, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (4), 2279-2290, 2023.