Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, Cilt 30, Sayı 3, 2025 ARAŞTIRMA

DOI: 10.17482/uumfd.1722270

DENGESİZ VERİ SETLERİ İÇİN İKİ AŞAMALI DENGELEME STRATEJİSİ: ADASYN İLE ÖRNEKLEM ARTIRMA, SVM TABANLI ÖRNEKLEM AZALTMA

Duygu YILMAZ EROĞLU *

Alınma: 18.06.2025 ; düzeltme: 05.09.2025 ; kabul: 04.11.2025

Öz: Bu çalışma, makine öğrenmesi alanında sıkça karşılaşılan dengesiz veri sorununu ele alarak, azınlık sınıf örneklerinin çoğunluk sınıf tarafından gölgede bırakıldığı durumlara odaklanmaktadır. Böyle bir dengesizlik, sağlık hizmetlerinden finansal sahtekârlık tespitine ve IoT tabanlı endüstriyel süreçlere kadar pek çok alanda model performansını ciddi biçimde zayıflatır. Sorunu gidermek için, azınlık sınıfını sentetik örneklerle zenginleştiren ADASYN yöntemi, SVM tabanlı uzaklık ölçümüyle belirlenen "en uzak" %10'luk çoğunluk örneklerinin çıkarılmasıyla birleştirilmiştir. Önerilen yaklaşım, SVM, RF, XGBoost ve KNN sınıflandırıcılarıyla on farklı veri seti üzerinde test edilmiştir. Bunlar arasında hem kalite kontrol verilerini hem de IoT sensör ölçümlerini içeren, gerçek üretim ortamından elde edilmiş 'Tekstil' veri seti de yer almaktadır. Özellikle iplik kopması gibi nadir ancak üretim açısından kritik olayları barındıran bu veri seti, yoğun dengesizlik nedeniyle standart yöntemlerde düşük başarı sergilemektedir. G-Ortalamalar metriğinde önemli iyileşmeler sunan yöntem, azınlık sınıfın daha başarılı tespitine katkıda bulunmuş ve on veri setinden beşinde en yüksek G-Ortalamalar değerini elde etmiştir.

Anahtar Kelimeler: Sınıflandırma, Dengesiz Veri, Örneklem Artırma, Örneklem Azaltma, Tekstil

A Two-Stage Balancing Strategy for Imbalanced Datasets: Oversampling with ADASYN and Undersampling Based on SVM

Abstract: This study tackles the frequently encountered problem of imbalanced datasets in machine learning, focusing on cases where minority-class examples are overshadowed by the majority class. Such imbalance significantly undermines model performance in diverse fields, including healthcare, fraud detection, and IoT-based industrial processes. To address this issue, we combine the ADASYN method—enriching the minority class with synthetic samples—with the removal of the "most distant" 10% of majority-class instances identified via an SVM-based distance measure. The proposed approach is tested with SVM, RF, XGBoost, and KNN classifiers on ten different datasets. Among these is the "Textile" dataset, which includes both quality control data and IoT sensor measurements and was collected from a real world production environment. Notably, this dataset includes rare yet critical events such as yarn breakage, which standard methods fail to detect effectively due to pronounced class imbalance. Our approach achieves considerable enhancements in the G-Mean metric, thereby improving the detection of minority cases and securing the highest G-Mean values on five out of ten datasets.

Keywords: Classification, Imbalanced Data, Oversampling, Undersampling, Textile

* Bursa Uludağ Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Görükle Kampüsü, 16059 Nilüfer/Bursa, Türkiye. E-posta: duygueroglu@uludag.edu.tr

1. GİRİŞ

Makine öğrenmesi ve veri madenciliği uygulamalarında dengesiz veri setleri, uzun yıllardır araştırmacıların odaklandığı temel sorunlardan biri olup, endüstriyel üretimden tıbbi teşhise uzanan geniş bir yelpazede sıklıkla karşımıza çıkmaktadır (López ve diğ., 2013). Bu tür veri dağılımlarında, azınlık sınıfına ait örneklerin baskın çoğunluk içinde kaybolması, standart sınıflandırma algoritmalarının performansını ciddi biçimde düşürebilmektedir (Abdelkhalek ve Mashaly, 2023; Weiss, 2004). Öte yandan, IoT tabanlı sensör ağlarının yaygınlaşması, gerek veri hacmini gerekse veri karmaşıklığını artırarak dengesizlik sorununu daha da görünür hâle getirmekte (Martikkala ve diğ., 2023; Wu ve diğ., 2023) ve gerçek zamanlı tahmin ile karar destek mekanizmalarında kritik gereksinimler doğurmaktadır. Bu bağlamda, azınlık sınıfını doğru şekilde yakalayabilmek adına örneklem artırma ve örneklem azaltma gibi dengeleme yöntemleri (He ve Garcia, 2009), büyük veri senaryoları dâhil olmak üzere (García-Gil ve diğ., 2024) geniş kapsamlı çözümler sunarak giderek daha fazla önem kazanmaktadır.

Bu çalışmada, TÜBİTAK TEYDEB destekli 3190654 numaralı araştırma projesi kapsamında elde edilen gerçek üretim verileri üzerinde dengesiz veri sorununa yönelik bir yaklaşım sunulmaktadır. Söz konusu Tekstil veri seti, iki ana kısımdan oluşmaktadır:

(i) Kalite kontrol parametreleri: Kopma Yükü, Mukavemet, Numara Denye, Uzama vb.

(ii) Üretim parametreleri: Final Bobin Sarım Tansiyonu, Final Bobin Tipi, Final Hız vb.

Kalite kontrol parametreleri giriş kontrol ile derlenirken, üretim parametreleri, dokuma üretim hattındaki sensörler ve kalite kontrol istasyonlarından otomatik olarak toplanmakta, zaman içinde büyüyen ancak çoğunlukla iplik kopması gibi istenmeyen olayların azınlıkta olduğu dengesiz bir yapı sergilemektedir. İpliğin kopma riskini tahmin edebilmek ve bu riski düşüren proses parametrelerini belirlemek, pratikte önemli bir gereksinim hâline gelmektedir.

Önerilen yaklaşımda, dengesiz veri setleri için sıklıkla başvurulan ADASYN (Adaptive Synthetic Sampling) yöntemi kullanılarak azınlık sınıfı sentetik örneklerle çoğaltılmakta, aynı zamanda SVM karar fonksiyonundan yararlanılarak marjin çizgisine en uzak %10 çoğunluk sınıfının çıkarılmasıyla bir tür örnek sayısı azaltma uygulanmaktadır. Böylece yüksek dengesizlikten kaynaklanan sorunlara, hem azınlık sınıfını yapay örneklerle zenginleştirme hem de baskın çoğunluk örneklerini kısmen azaltma yaklaşımıyla bütüncül bir çözüm sunulmaktadır. Ardından, SVM (Destek Vektör Makineleri), RF (Rassal Orman), XGBoost ve KNN (K-En Yakın Komşu) gibi farklı sınıflandırma algoritmalarıyla eğitim-test deneyleri gerçekleştirilmiştir.

Bu çerçevede, çalışmanın odağını belirginleştirmek üzere aşağıdaki araştırma soruları ortaya konulabilir.

S1: ADASYN ile yoğunluk temelli örneklem artırma ve SVM marjına en uzak %10 çoğunluk verisinin çıkarılması ile örneklem azaltma yaklaşımlarının birlikte kullanılmasının önerildiği yöntem, verinin dengeleme yapılmadan önceki ham haline göre ve sadece ADASYN ile örneklem artırma uygulanan yönteme göre G-Ortalamalar performans ölçütünde daha üstün sonuçlar üretir mi?

S2: Önerilen yaklaşım, karar sınırı çakışması ve çoğunluk baskısı etkilerini azaltarak hem literatürde kullanılan veri setlerinde hem de gerçek üretimden elde edilen Tekstil veri setinde G-Ortalamalar değerlendirme göstergesinde istikrarlı ve genellenebilir performans sağlar mı?

Çalışma sonunda, Tekstil veri seti dâhil olmak üzere toplam on veri seti üzerinde elde edilen deneysel sonuçlar paylaşılacak ve dengesiz veri setlerinin sınıflandırma sonuçlarının örneklem artırma ve örneklem azaltma yöntemleri ile nasıl iyileşebileceği gösterilecektir. Bulgular, özellikle aşırı dengesiz (Tekstil, Climate) ya da orta düzey dengesiz (Sentetik, Column_2c, vb.) veri setlerinde, azınlık sınıfını yakalama açısından kayda değer performans artışları sunduğunu göstermektedir. İlaveten, önerilen yöntem on veri setinin beş adedinde en iyi performansı göstermiştir.

2. LİTERATÜR TARAMASI

Bu bölümde, dengesiz veri setlerine yönelik literatür taramasını yaparken sınıflandırma algoritmaları bazında (SVM, RF, XGBoost, KNN vb.) sunulan yaklaşımlar ve ADASYN gibi yaygın kullanılan aşırı örnekleme yöntemi ile ilgili çalışmalar incelenmiştir. İlaveten, Tekstil sektöründeki iplik kopuşu problemi ve IoT tabanlı veri toplama yaklaşımları gibi spesifik örnekler de eklenerek, bu yöntemlerin pratik endüstriyel uygulamalarda nasıl değerlendirildiği vurgulanmıştır.

Dengesiz veri setlerinde SVM tabanlı yaklaşımlar, çoğunluk sınıfa doğru oluşan eğilim nedeniyle sıklıkla ek veri işleme aşamalarına gereksinim duyar. Bu kapsamda, Mu ve Zhao (2025) tarafından önerilen DCS-SOCP-SVM yöntemi, dengesiz veri setleri için özel tasarlanmış bir hiper-düzlem oluşturmayı hedeflemekte ve SV-SMOTE ile NSV-RUS gibi aşamalı örnekleme stratejilerini bütünleştirmektedir. Benzer biçimde, Wang ve Liu (2023) ise LSSASMOTE adlı bir aşırı örnekleme yaklaşımını, çekirdek fonksiyon parametreleri ile ceza parametrelerini optimize eden Levy Sparrow Arama Algortitması (LSSA) temelli SVM optimizasyonu ile birleştirerek, azınlık sınıfın daha etkin biçimde sınıflandırılmasını sağlamıştır. Öte yandan, Hussein ve diğ. (2023) çalışmalarında iyileştirilmiş Tavlama Benzetimi (SA) algoritmasını SVM ile bütünleştirerek, veri ön işleme aşamasında sentetik örnek üretimi ve gereksiz örneklerin elenmesini hedeflemiş, böylece dengesiz veri setlerinde ortalama %89,65 seviyesinde bir doğruluk elde etmişlerdir. Ayrıca, Guo ve diğ. (2024) tarafından ortaya konan SV-Borderline SMOTE-SVM yaklaşımı, çekirdek mesafesi tabanlı komşuluk analizini kullanarak yeni sentetik azınlık örnekleri üretmekte ve Gram matrisi üzerinden karar fonksiyonunu güncelleyerek G-Ortalamalar ve F-skor değerlerinde belirgin iyileşmeler sağlamaktadır. Son olarak, Yilmaz Eroglu ve Pir (2024) tarafından geliştirilen HOUM (Melez örneklem artırma ve örneklem azaltma yöntemi), Safe-Level SMOTE ve SVM'i hibrit bir çerçevede kullanarak hem örneklem artırmanın getirdiği aşırı uyum riskini hem de örneklem azaltmanın yol açtığı bilgi kaybını en aza indirgemeyi amaçlamıştır.

Karar ağaçlarında dengesizliğe duyarlılık, düğüm bölme ölçütlerinde çoğunluk sınıfın baskın hale gelmesiyle belirginleşir. Zhou ve diğ. (2024) bu soruna çözüm olarak SDDT ve WSD-CART algoritmalarını geliştirmiş, özellikle çok farklı oranlarda dengesizlik içeren UCI veri setlerinde yüksek F-skor ve AUC değerlerine ulaşmışlardır. Kredi skorlaması örneğini ele alan My ve Ta (2023) ise DTE (Decision Tree Ensemble) yönteminin, çoğunluk-azınlık dağılımını dengelemeye yönelik yeniden örnekleme teknikleriyle birleştirildiğinde, RF veya AdaBoost gibi popüler ağaç tabanlı yaklaşımlardan daha başarılı olduğunu göstermiştir. Büyük veri senaryolarında dengesizlik ve veri kalitesi sorunlarını eşzamanlı olarak ele almak üzere García-Gil ve diğ. (2024) tarafından öne sürülen SD_DeTE metodolojisi, farklı ön işleme (ör. rastgele ayrıklaştırma, PCA tabanlı dönüşümler) ve kümeleme tabanlı aşırı örnekleme aşamalarını birleştirerek karar ağaçlarının performansını artırmaktadır. Öte yandan Avizheh ve diğ. (2023), bir hayvancılık uygulaması olan buzağı doğum zorluğunu tahmin etme probleminde, azaltılmış örnekleme ve maliyet duyarlı yaklaşımların karar ağaçları üzerinde F-skorunda iyileştirme sağladığını, ancak veri eksikliği ve dengesizlik gibi sorunların da tam anlamıyla giderilemediğini belirtmektedir. RF (Random Forest), birden fazla karar ağacının oluşturduğu topluluk yapısıyla karmaşık veri dağılımlarını iyi öğrenebilir; ancak dengesiz veri setlerinde azınlık sınıfın temsil gücü sınırlı kalabilir. Bu noktada, Aymaz (2025) sağlık alanında dengesiz veri setleri için farklı aşırı örnekleme ve aşırı örnekleme tekniklerinin yanı sıra, denge oranı optimizasyonu üzerinde durarak RF ve SVM gibi algoritmaları karşılaştırmıştır. Wibbeke ve diğ. (2025) ise veri setindeki dengesizliği iyileştirmek için yeni bir yaklaşım sunmuş ve RF ile ağaç yapılarına dayalı modellerin, az temsil edilen örneklerde yüksek hata payına sahip olduğunu gözlemlemiştir. Ayrıca, IoT güvenliği gibi alanlarda da RF ile dengesiz veri sınıflandırma kombinasyonu dikkat çekicidir. Örneğin Almarshdi ve diğ. (2023), CNN+LSTM tabanlı bir derin öğrenme modelini RF ve Karar Ağacı gibi yöntemlerle kıyaslamış; SMOTE benzeri bir dengeleme tekniğinin

kullanımıyla %92,10 seviyesinde başarı elde etmişlerdir. Galán-Cuenca ve diğ. (2024) de COVID-19 göğüs röntgeni gibi kısıtlı ve dengesiz veri senaryolarında, Siyam ağları yardımıyla veri azlığı problemine çözüm getirirken, kıyaslama için RF gibi yöntemleri de değerlendirmiştir. Markwald ve Demidova (2024) ise REFUEL adı verilen kural çıkarım temelli bir yaklaşımla, azınlık sınıfın düğüm özelliklerini öğrenmede RF ve benzeri modellerin performansını kıyaslamış, az temsil edilen düğümlerde en az 4 puanlık doğruluk artışı sağlandığını bildirmiştir. Ercire ve Ünsal (2021), güç kalitesi bozulmalarının (GKB) RF ve dalgacık analiziyle sınıflandırılmasını ele almış; gürültülü ortamlarda dahi %99'un üzerinde doğruluk oranına ulaşarak yöntemin güvenilirliğini kanıtlamışlardır.

Gradient boosting (Gradyan Güçlendirme) temelli yöntemler de dengesiz verilerde yaygın biçimde kullanılmaktadır. Amirshahi ve Lahmiri (2024) iflas tahmininde XGBoost, LightGBM ve CatBoost yöntemlerini dengesiz veri setlerinde karşılaştırmış ve aşırı örnekleme'in bazı durumlarda ters etki yaratabildiğini gözlemlemiştir. Benzer biçimde Militino ve diğ. (2024), uydudan orman yangını tespiti gibi azınlık sınıfın son derece az olduğu alanlarda Lojistik Regresyon yaklaşımının dahi XGBoost'a yakın performans verebildiğini, ancak veri boyutu arttıkça XGBoost'un daha avantajlı hâle geldiğini vurgulamıştır. Karmaşık zemin koşullarının gerçek zamanlı tahmininde Wu ve diğ. (2023), XGBRF isimli bir karma model kullanarak Youden indisi ve AUC metriklerinde kayda değer bir iyileşme sağlamıştır. Dengesiz veri setini yeniden örneklemenin özellikle erken uyarı mekanizmasında büyük katkı sunduğu da belirtilmektedir. Dahası, Wang ve diğ. (2023b) isimli çalışmada ise anomali tespiti için XGBoost ve RF tabanlı topluluk modelleriyle çeşitli örnekleme stratejilerinin denenmesi sonucunda %98 üstü doğruluk ve %99'a yakın hassasiyet elde edildiği raporlanmıştır.

KNN algoritması, basitliği ve kolay yorumlanabilmesi nedeniyle tercih edilen bir yöntemdir; ancak dengesiz veri setlerinde çoğunluk sınıfın komşuluk yapısında baskın olma sorunu nedeniyle azınlık sınıfı çoğu zaman geri planda kalabilir. Wang ve diğ. (2023a) bu durumu iyileştirmek adına merkez değiştirme tabanlı bir yöntem geliştirerek, KNN'nin karar aşamasını daha dinamik ve homojen bir komşuluk yapısına oturtmuştur. Tarımsal uygulamalarda ise Pérez Tárraga ve diğ. (2025) don olaylarını tahmin etmek için SMOTE yöntemi ile veri dengesini artırdıktan sonra KNN kullanarak AUC değerini %13'lük bir artışla 0,9909 seviyesine yükseltmeyi başarmıştır. Diğer yandan, Morgan ve diğ. (2024) jeolojik haritalama senaryosunda, SVM, RF, XGBoost ve KNN gibi çeşitli modelleri kıyaslamış; KNN'nin kısıtlı veri koşullarında zaman zaman yetersiz kalabildiğini, ancak dağılımın dengelendiği durumlarda yine makul sonuçlar verdiğini göstermiştir. Altunkaynak ve diğ. (2020), dalgacık dönüşümüyle zenginleştirdikleri KNN yöntemini hava kirliliği tahmininde uygulayarak, klasik KNN modellerine kıyasla daha yüksek başarı sağladıklarını göstermiştir. Özdet ve İçer (2022), akciğer bilgisayarlı tomografi (BT) görüntülerinde tümör tespiti için farklı sınıflandırıcıları (Karar Ağaçları, SVM, KNN vb.) karşılaştırmış ve yüksek doğruluk oranlarının erken teşhiste umut verici olduğunu raporlamıştır.

Azınlık sınıfı örneklerini adaptif biçimde çoğaltan ADASYN (He ve diğ., 2008), dengesiz veri sorununda SMOTE'a benzer bir mantık izler ancak azınlık örneklerin dağılım yoğunluğuna göre sentetik noktalar ekleyerek sınıf sınırlarını daha iyi genişletir. Stanosheck ve diğ. (2024) tarafından yapılan deneylerde, E. coli tahmininde RF ve SVM modellerine ADASYN uygulanmasıyla %86,55 civarında doğruluk ve 0,924 AUC elde edilmiştir. Benzer biçimde Abdelkhalek ve Mashaly (2023), ağ saldırılarını (NIDS) tespit etmek için ADASYN ve Tomek Links tabanlı bir derin öğrenme modelini değerlendirmiş ve hem ikili hem de çok sınıflı senaryolarda önemli başarım raporlamıştır.

Tekstil sektöründe iplik kopuşu, üretim maliyetlerini ve kaliteyi doğrudan etkileyen kritik bir sorundur. Yao ve diğ. (2005), yapay sinir ağları kullanarak kopma yükü, mukavemet gibi kalite indeksleriyle atölye koşullarında elde edilen kopma oranları arasındaki ilişkiyi incelemiş; tek katmanlı ağ yapısı ile yüksek korelasyon tespit etmişlerdir. Roving gerilimini algılama üzerine çalışan Peng ve Yuan (2021) ise ring iplik eğirme sürecinde gerginlik değerlerini gerçek zamanlı izleyerek iplik kopuşlarını azaltmanın mümkün olduğunu aktarmaktadır. Dokuma sırasında

uygulanan gerginliğin kumaş mukavemeti ve kopma olayları üzerindeki etkisini inceleyen Mebrate ve diğ. (2022), hava jetli dokuma tezgâhlarında gerginlik arttıkça yırtılma ve çekme mukavemetinin azaldığını göstermiştir. Benzer şekilde, Lappage (2005) daha ince ve düzensiz ipliklerde ince noktalara bağlı kopuşların arttığını, bu nedenle üretim parametrelerinin iplik inceliği ve düzgünlüğüyle birlikte ayarlanması gerektiğini belirtmiştir.

IoT teknolojileri, tekstil alanında akıllı izleme ve süreç optimizasyonu için güçlü çözümler sunmaktadır. Ferreira ve diğ. (2025), polipirol ile geliştirilmiş kumaşların dokunma veya hareket enerjisini elektrik sinyaline dönüştürdüğünü ve böylece giyilebilir IoT sensörleri için potansiyel oluşturduğunu deneysel olarak göstermiştir. Üretim sürecinin sürdürülebilirlik boyutunda ise Tolentino-Zondervan ve DiVito (2024) dijitalleşme ve tedarik zinciri izlenebilirliğinin atık yönetimi ve çevresel etkilerin azalmasında önemli olduğunu vurgulamıştır. Atık toplama aşamasında IoT tabanlı sensör ve dinamik rota optimizasyonunu birleştiren Martikkala ve diğ. (2023), akıllı tekstil atık kutuları sayesinde toplama maliyetinde %7,4, zamandan %7,3 tasarruf ve %10,2 oranında CO_2 emisyonu azalması sağlandığını raporlamıştır. Bu sonuçlar, ileri dijitalleşme teknikleri ve sensör verilerinin gerçek zamanlı işlenmesiyle tekstil endüstrisinde sürdürülebilirlik ve verimliliğin artırılabileceğini göstermektedir.

Literatür çalışmaları gözden geçirildiğinde, dengesiz veri setlerine yönelik yaklaşımlar çoğunlukla yeniden örnekleme (SMOTE/ADASYN türevleri) ve/veya maliyet duyarlı öğrenmeye dayansa da, sınıf sınırında çakışma, sentetik gürültü riskleri, azınlık örneklerinin temsil yetersizliği sorunlarının bütünüyle giderilemediği görülmektedir. Nitekim karar ağaçlarında, düğüm bölme aşamalarında çoğunluk sınıfı baskın hâle gelebilmekte ve azınlık sınıfının temsil gücü sınırlı kalabilmektedir. KNN'de ise komşuluk yapısı gereği azınlık sınıfı sıklıkla geri planda kalmaktadır. Bu çalışmada, söz konusu eksikleri hafifletmek amacıyla, ADASYN ile, azınlık sınıfının yoğunluğunun düşük olduğu (az yoğun) bölgelerde sentetik örnek üretimi arttırılarak azınlık temsili güçlendirilmiş; ilaveten SVM karar sınırına en uzak %k çoğunluk örnekleri ayıklanarak öğrenmeye katkısı zayıf olduğu düşünülen örneklerin etkisi seyreltilmiştir. Çalışma kapsamında %k değeri için parametre seçimi yaparken, SVM karar sınırına en uzak %5, %10, %15, %25 bandında duyarlılık analizi yapılmış, %10 değeri seçilerek "En Uzak %10 Sil" kararında ilerlenmiştir. Ayrıca, literatürde sıklıkla kullanılan veri setlerine ilave olarak, TÜBİTAK TEYDEB destekli 3190654 numaralı araştırma projesi kapsamında üzerinde çalışılan "Tekstil" veri seti ele alınmış, gerçek üretim ortamından IoT ile toplanan üretim parametre verileri, hammadde kalite kontrol ölçümleriyle birlikte değerlendirilmiş; dokuma sırasında nadir görülen iplik kopuşlarının tahmini modellenmiş, böylece IoT tabanlı üretim verileri ile hammadde kalite kontrollerinin birlikte kullanımına ilişkin literatürdeki boşluğa uygulamalı bir katkı sağlanmıştır.

3. KULLANILAN YÖNTEMLER VE DEĞERLENDİRME ÖLÇÜTLERİ

Bu bölümde, dengesiz veri setlerini ele almak üzere kullanılan temel sınıflandırma algoritmaları (SVM, RF, XGBoost, KNN) ile azınlık sınıfını artırmaya yönelik örnekleme yöntemi ADASYN tanıtılmaktadır. Ek olarak, model başarısını adil bir biçimde değerlendirmesinde kullanılan değerlendirme ölçütleri irdelenecektir.

3.1. Destek Vektör Makineleri (Support Vector Machines (SVM))

Destek Vektör Makineleri (SVM), iki sınıfı, mümkün olan en geniş marjinle ayıran karar sınırını (hiper-düzlem) bulmayı amaçlayan denetimli bir öğrenme yöntemidir. Basitçe, veri üzerinde tanımlanan hiper-düzlem, sınıfları ayırırken marjini maksimize edecek şekilde optimize edilir.

3.1.1. Doğrusal SVM

Doğrusal SVM'de, veri noktalarını ayıran bir hiperdüzlem, Denklem (1) ile tanımlanır:

$$wx + b = 0 \tag{1}$$

Burada:

- w: Hiperdüzlemin ağırlık vektörüdür.
- b: Sapma (bias) terimidir.
- ‖w‖: w vektörünün Öklid normudur.

Margin, hiperdüzlem ile ona en yakın veri noktaları arasındaki mesafeyi temsil eder ve Denklem (2) ile ifade edilir.

$$margin = \frac{2}{\| \mathrm{w} \|} \tag{2}$$

Veri noktasının hiperdüzleme göre konumuna bağlı olarak hangi sınıfa ait olduğunu belirleyen karar fonksiyonu, Denklem (3) ile bulunur.

$$f(x_i) = \begin{cases} 1, if\ wx_i + b \geq 1 \\ -1, if\ wx_i + b \leq -1 \end{cases} \tag{3}$$

3.1.2. Doğrusal Olmayan SVM

Gerçek dünya problemlerinin çoğunda veriler tek bir hiperdüzlemle doğrusal olarak ayrılmaz. Bu sorunu çözmek için, veriler daha yüksek boyutlu bir uzaya dönüştürülür ve ardından bu uzayda bir hiperdüzlem tanımlanır. Radyal SVM'nin formülasyonlarında aşağıdaki semboller kullanılır.

- $f(x)$: Girdi x için sınıflandırma fonksiyonu.
- y_i : Sınıf etiketleri.
- a_i : SVM optimizasyonu sırasında elde edilen Lagrange çarpanları.
- $\varphi(x_j)\varphi(x_j)$: Daha yüksek boyutlu uzaydaki iç çarpımı temsil eder.
- $K(X_i, X_j)$: : Verileri açıkça yüksek boyuta dönüştürmeden iç çarpımı hesaplayan çekirdek (kernel) fonksiyonunu temsil eder.

Dönüştürme çözümü Denklem (4)'te gösterilen SVM formülüyle yapılır. Doğrusal olarak ayrılamayan bir SVM'de hesaplanması gereken $\varphi(x_i)\varphi(x_j)$ terimleri, önemli özelliklere sahip skaler çarpımlardır. Bu, çekirdek fonksiyonu (K) olarak adlandırılır. Çekirdek fonksiyonu kullanıldığında, SVM Denklem (5)'de gösterildiği gibi formüle edilir. Radyal Tabanlı (Radial Basis) ve Sigmoid çekirdekleri, çalışmalarda sıkça kullanılan fonksiyonlardır. Bu çekirdekler sırasıyla Denklem (6) ve Denklem (7) ile ifade edilir. Denklem (6)'daki σ, kullanılan Gauss fonksiyonunun yayılımını belirleyen bir parametredir. Denklem (7)'deki δ, Sigmoid çekirdek fonksiyonuna ait olup, çekirdek için kullanılan sigmoid fonksiyonunun şeklini etkiler.

$$f(x) = \text{sgn}\left(\sum_{i=1}^{i} y_i\, a_i \varphi(x_i)\varphi(x_j) + b\right) \tag{4}$$

$$f(x) = \text{sgn}\left(\sum_{i=1}^{i} y_i\, a_i K(x_i, x_j) + b\right) \tag{5}$$

$$Radyal\ Tabanlı: K(X_i, X_j) = \exp\left(-\frac{1}{2\sigma^2} \parallel X_i - X_j \parallel^2\right) \tag{6}$$

$$Sigmoid\ Tabanlı: K(X_i, X_j) = \tanh(kX_iX_j - \delta) \tag{7}$$

Bu çalışmada, Doğrusal Olmayan SVM kullanılmış olup, Radyal Tabanlı çekirdekler kullanılmıştır.

3.2. Karar Ağaçları ve Rassal Orman (Random Forest (RF))

Karar ağaçları, veriyi, yinelenen bölme adımlarıyla dallara ayırarak sınıflandırma veya regresyon yapan temel yöntemlerdendir. Her düğümde, bir özellik (ve varsa bir eşik) seçilip veri sol/sağ dal şeklinde ayrılır. Sınıflandırma için Gini veya Entropi gibi ölçütler kullanılarak safsızlık (impurity) değerlendirilir. Gini indeksi, sınıflandırma problemlerinde bir düğümdeki saf olmayan (impure) örneklerin oranını ölçen yaygın bir ölçüttür. Bu ölçüte göre, veri seti mümkün olduğunca saf alt kümelere ayrılmaya çalışılır; yani her bir alt düğümde tek bir sınıfa ait örneklerin oranı artırılmaya çalışılır. Gini indeksi değeri ne kadar düşükse, düğüm o kadar homojen kabul edilir. Bu nedenle algoritma, en düşük Gini indeksini sağlayacak şekilde bölünmeler yaparak karar ağacını büyütür. Bu yaklaşım, sınıflar arasındaki ayrımı daha etkili şekilde yakalayarak modelin genel sınıflandırma başarımını artırmayı hedefler.

RF algoritması, gözetimli sınıflandırma için birden çok karar ağacını bir araya getirir. Her ağaç, eğitim verisi kümesi S'den rastgele seçilmiş bir alt küme üzerinde eğitilir; nihai sınıf etiketi, tüm ağaçların oy çokluğuyla belirlenir.

Temel Kavramlar:

S: Tam eğitim kümesi

i: Ağaç indisi (1… k)

k: Ormandaki toplam ağaç sayısı

T_i: i. ağaç için karar ağacı modeli

S_i: T_i'nin öğrenmesi için rastgele örneklenmiş veri alt kümesi

Algoritma Akışı

1. Başlangıç: Tam eğitim kümesi S ile işe başlanır.
2. Ağaç Oluşturma: Her i=1,…,k için:
    1. S'den S_i alt kümesini örnekle.
    2. T_i'yi S_i üzerinde eğit.
3. Düğüm Bölümlendirme: Her düğümde, özellikler kümesinden rastgele seçilen F alt kümesindeki özellikler kullanılarak en uygun bölümlendirme bulunur.
4. Ağaç Büyütme: Ağaçlar, budanmadan tam derinliğe kadar büyütülür.
5. Tahmin: k ağacın çıktıları oy çokluğuyla birleştirilerek son sınıf etiketi atanır.

3.3. XGBoost (eXtreme Gradient Boost)

XGBoost, karar ağaçları üzerinden gradient boosting yöntemini hızlı ve optimize biçimde uygulayan bir yapıdır. Her yinelemede, mevcut modelin hataları (residuals) incelenerek yeni bir ağaç inşa edilir. Böylece model, adım adım iyileşir.

Temel Yapı

- Başlangıçta basit bir model (ör. sabit tahmin) kurulur.
- Her yinelemede bir ağaç eklenir, önceki modelin kalıntı hatalarını azaltmaya çalışır.
- Öğrenme oranı ve düzenlileştirme terimleri *(L1, L2)* aşırı uyumu önlemeye yardımcı olur.

Örnek Formül: Yeni eklenen ağaç $h_t(x)$ ile model, (8) numaralı formülde gösterildiği gibi güncellenir. Burada ν ($0 < \nu \leq 1$) öğrenme hızıdır. XGBoost, hatanın ikinci türevini de kullanarak daha hassas bölme kararları alır ve histogram tabanlı bölme özelliğiyle hızlıdır.

$$F_t(x) = F_{t-1}(x) + v.h_t(x) \tag{8}$$

3.4. K-En Yakın Komşu (K Nearest Neighbor (KNN))

K-En Yakın Komşu (KNN), tembel öğrenme sınıfında yer alan basit ama kullanışlı bir yöntemdir. Yeni bir örneğin etiketi, eğitim kümesindeki en yakın k komşusunun etiketlerinden çoğunluk yoluyla belirlenir. Temel mantık aşağıdaki gibidir:

1. K parametresini seçme: Dikkate alınacak en yakın komşu sayısını belirlenir.
2. Mesafe hesaplama: Sınıflandırılacak veri noktası ile veri kümesindeki diğer tüm noktalar arasındaki mesafeyi uygun bir mesafe metriği kullanarak hesaplayın.
3. Komşuları belirleme: Hesaplanan mesafeleri artan sıraya göre sıralayarak en yakın K komşuyu seçin.
4. Sınıf etiketi atama: Seçilen K komşu arasında en sık görülen sınıf etiketini hedef veri noktasına atayın.

Mesafe hesaplaması için bu çalışmada, Denklem (9)'da gösterilen Öklid Mesafesi (q=2) kullanılmıştır:

$$d(i,j) = \sqrt[q]{\left(\left|x_{i1} - x_{j1}\right|^q + \left|x_{i2} - x_{j2}\right|^q + \cdots + +\left|x_{ip} - x_{jp}\right|^q\right)} \tag{9}$$

3.5. ADASYN (Adaptive Synthetic Sampling)

ADASYN, dengesiz veri kümelerinde azınlık sınıfı sentetik örneklerle çoğaltarak (over-sampling) dengeyi sağlamayı amaçlayan bir yaklaşımdır (He ve diğ., 2008). Klasik SMOTE yönteminin (Synthetic Minority Over-sampling Technique) bir türevi olarak görülebilir, ancak ADASYN, daha seyrek bölgelerdeki azınlık örneklerine ağırlık vererek yapay örnekler üretir.

Temel İlke

- Azınlık sınıfın bulunduğu veri noktaları için, yerel yoğunluğu (komşu sayısı) az olan bölgeleri tespit eder.

- Bu "zayıf temsil" alanlarında daha fazla sayıda sentetik örnek üreterek, azınlık sınıfı çeşitlendirir.
- Böylelikle modelin bu kritik bölgeleri tanıması kolaylaşır

Örnek Formülasyon: Her bir azınlık örneği x_i için k komşusu belirlenir. Formülasyon (10) de gösterildiği gibi, eğer r_i, x_i'nin yeterince temsil edilmediği (seyrek) bölgede olma düzeyini gösteriyorsa,

$$r_i = \frac{\Delta_i}{\sum_{j=1}^{N_{azınlık}} \Delta_j} \tag{10}$$

Burada, Δi, x_i'nin çoğunluk/azınlık dağılımına göre eksik temsil düzeyidir. Her x_i için oluşturulacak sentetik örnek sayısı (11) numaralı formül ile belirlenebilir. $N_{azınlık}$, azınlık sınıfındaki toplam örnek sayısıdır. Sonra, G_i kadar yeni nokta, komşuları arasından rastgele seçilerek, lineer interpolasyon benzeri bir işlemle üretilir.

$$G_i = r_i \times G_{toplam} \tag{11}$$

Yapay Örnek Üretimi: Yeni sentetik örnek x_{yeni}, (12) formülüyle oluşturulur. Burada x_{nn}, x_i'nin komşularından biri, $\delta \in [0, 1]$ rastgele bir sayıdır. Böylece, azınlık sınıfın farklı bölgelerinde çeşitlilik sağlanmış olur.

$$x_{yeni} = x_i + \delta.(x_{nn} - x_i) \tag{12}$$

3.6. Performans Ölçütleri ve Değerlendirme

Makine öğrenmesinde model seçimi ve değerlendirme, bir sınıflandırıcının etkililiğini belirlemede kritik öneme sahiptir. Aşağıda, makine öğrenmesi çalışmalarında kullanılan temel ölçütler açıklanmış, son olarak da dengesiz veri setlerinin değerlendirilmesinde sıklıkla kullanılan ve bu çalışmada da karşılaştırmalarda dikkate alınacak olan G-Ortalamalar ölçütü açıklanmıştır.

- Doğruluk Oranı, TP (True Pozitif) ve TN (True Negatif) değerlerinin toplam örnek sayısına (TP+TN+FP+FN) oranıdır.
- F-skoru, kesinlik (PPV) ve duyarlılık (TPR) arasındaki harmonik ortalamadır.
- PPV (Kesinlik), pozitif olarak tahmin edilen örneklerin ne kadarının gerçekten pozitif olduğunu gösterir.
- TPR (Duyarlılık), gerçek pozitif örneklerin ne kadarının doğru tahmin edildiğini ifade eder.

Dengesiz veri kümelerinde bu metrikler yetersiz kalabildiğinden, G-Ortalamalar sıklıkla tercih edilir. Formülasyon (13) te de gösterildiği gibi G-Ortalamalar, azınlık (pozitif) ve çoğunluk (negatif) sınıfın başarı oranlarını (TPR ve TNR) ayrı ayrı değerlendirir ve geometrik ortalama üzerinden tek bir metrik sunar.

$$G_{ortalamalar} = \sqrt{TPR * TNR} \tag{13}$$

Bu yaklaşım, veri setindeki dengesizliğin neden olabileceği yanıltıcı sonuçları azaltarak Modelin her iki sınıftaki performansını da önemser.

4. ÖNERİLEN ALGORİTMA

Bu çalışmada, dengesiz veri kümeleri üzerinde sınıflandırma başarımını artırmak amacıyla ADASYN ile örneklem artırılmış, SVM ile oluşturulan marjine en uzak %10'luk çoğunluk örnekleri silinerek de örneklem azaltma sağlanmıştır. Bu bölüm, söz konusu iki tekniğin ayrıntılarını ve birbirleriyle nasıl entegre edildiklerini özetlemektedir.

4.1. ADASYN ile Örneklem Artırma

- Amaç: Azınlık sınıfa ait verileri çoğaltarak dengesizliği gidermek ve modelin azınlık sınıfını daha iyi tanımasını sağlamaktır.
- Çalışma Prensibi: Klasik SMOTE yaklaşımına benzer biçimde komşuluk ilişkilerinden yararlanarak yapay (sentetik) azınlık örnekleri oluşturur. Ancak seyrek bölgelerdeki azınlık örneklerine daha yoğun sentetik örnek ekler ve böylece sınıf dağılımında çeşitliliği artırır.
- Faydası: Azınlık sınıfın sınır bölgelerini daha iyi temsil ederken, çoğunluk sınıfın ezici etkisini azaltır ve sınıflar arası dengesizlikten kaynaklanan hataları minimuma indirmeye çalışır.

4.2. SVM ile Oluşturulan Marjine En Uzak %10 Çoğunluk Örneklerini Çıkarma

- Amaç: Karar sınırından çok uzakta kalan veya modelin öğrenme sürecine olumsuz etki edebilecek uç noktadaki çoğunluk örneklerini ayıklamaktır.
- Çalışma Prensibi: RBF çekirdekli bir Destek Vektör Sınıflandırıcısı modeliyle, her örneğin karar sınırına göre konumu belirlenir. Ardından, çoğunluk sınıf içinde bu değerin mutlak anlamda en yüksek olduğu (yani sınırdan en uzak) %10'luk kısım silinir.
- Faydası: Veri uzayındaki aşırı baskın veya atipik çoğunluk örneklerini temizleyerek, özellikle azınlık-çoğunluk sınırında daha dengeli bir veri dağılımı oluşturmayı hedefler.
- Parametre Seçimi: Sınırdan en uzak olup çıkarılacak olan örnek oranı %k, makalenin ilerleyen bölümlerinde sunulan duyarlılık analizi ($k \in \{5, 10, 15, 25\}$) ve Wilcoxon işaretli sıralar testi sonuçlarına dayalı olarak belirlenmiştir; bu kapsamda yöntem akışında varsayılan değer olarak "En Uzak %10 Sil" stratejisi benimsenmiştir.

Algoritmanın Özeti:

1. Etiket Haritalama: Veri kümesindeki sınıf etiketleri azınlık (1) ve çoğunluk (0) olacak şekilde yeniden düzenlenir.
2. K-Katlı Çapraz Doğrulama: Veri kümesi tabakalı 5-katlı çapraz doğrulama yöntemiyle beş katmana ayrılır. Her katmanda:
    - ADASYN: Eğitim seti üzerinde ADASYN uygulanarak azınlık sınıf örnek sayısı artırılır.
    - En Uzak %10 Sil: SVM ile bulunan marjine en uzak ve çoğunluk sınıfına ait en uzak %10'luk kısmı silinir.

    Ardından model bu veriyle eğitilir.
3. Performans Değerlendirme: Her katlamada G-Ortalamalar ölçütü hesaplanarak ortalama değerleri raporlanır.

Sonraki bölümlerde, elde edilen deneysel bulgular ve yöntemlerin karşılaştırmalı analizleri detaylı şekilde tartışılacaktır. Bu sayede önerilen yaklaşımın hem çoğunluk hem de azınlık sınıflarını daha "adil" bir biçimde temsil etme potansiyeli okuyucunun dikkatine sunulacaktır. Dengesiz verilerde azınlık ve çoğunluk sınıfları arasındaki örnek sayısı farkı,

makine öğrenmesi modellerinin performansını olumsuz etkileyebilmektedir. Söz konusu yaklaşım, bu soruna çift yönlü bir çözüm sunmayı hedeflemektedir:

5. DENEYSEL SONUÇLAR

5.1. Kullanılan Veri Setleri

Tablo 1 incelendiğinde, her veri setinin farklı sayıda öznitelik (değişken) ve örnek içerdiği görülmektedir. Ayrıca çoğunluk ve azınlık sınıflarının dağılımı da veri kümeleri arasında önemli ölçüde değişkenlik göstermektedir. Örneğin, Breast Cancer veri setinde çoğunluk/azınlık oranı 1,68 gibi düşük bir değerde kalırken, Tekstil veri setinde bu oran 25,28 gibi son derece yüksek bir seviyededir. Bu durum, modellerin dengesiz veriyle başa çıkma becerisinin önemini daha da vurgulamaktadır. Öte yandan, Sentetik olarak adlandırılan ve bir sonraki alt bölümde detaylandırılacak olan yapay veri seti, ağırlıklandırma (0,7; 0,3) yöntemiyle belirli bir orana (2,33) sahip olacak şekilde üretilmiş olup, diğer gerçek veri kümelerine benzer bir dengesizlik profili sergilemektedir. Tekstil veri seti de ilerleyen alt bölümlerde detaylandırılacaktır.

Tablo 1. Kullanılan veri setlerinin temel özellikleri

Veri seti	Öznitelik Sayısı	Örnek Sayısı	*Np*	*Nn*	*Np/Nn*
Climate	21	540	494	46	10,74
Column_2c	7	310	210	100	2,10
Haberman	4	305	224	81	2,77
Diabetes	9	768	500	268	1,87
Liver	11	583	416	167	2,49
Ionosphore	32	351	225	126	1,79
Transfusion	5	748	570	178	3,20
Breast Cancer	30	569	357	212	1,68
Tekstil	9	1708	1643	65	25,28
Sentetik	10	1000	700	300	2,33

5.1.1. Sentetik Veri Seti

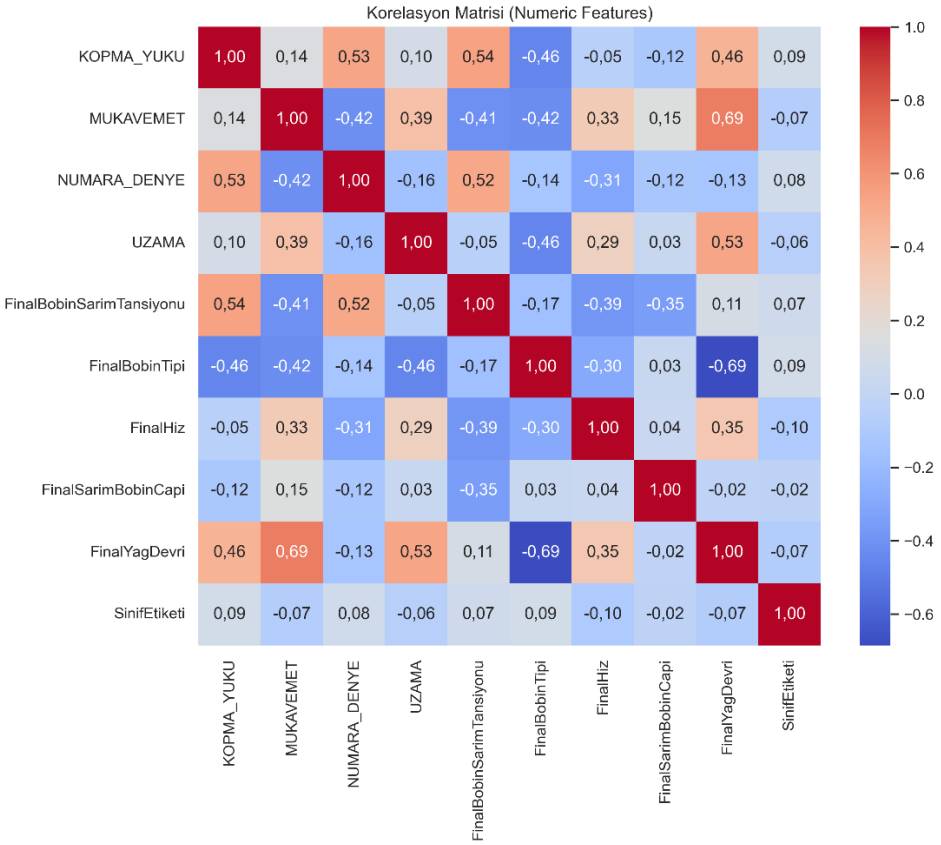
Gerçek veri setlerine ek olarak, modelin daha fazla veri seti üzerinde denenebilmesi ve farklı senaryoların simüle edilebilmesi amacıyla sentetik veri seti üretimi sıklıkla tercih edilmektedir. Bu yaklaşım sayesinde, veri kıtlığının yaşandığı durumlarda bile algoritmaların performansını incelemek ve deneyleri tekrarlanabilir kılmak mümkün olmaktadır. Aşağıda örnek olarak üretilen sentetik veri setinin temel özellikleri özetlenmiştir:

- Toplam 1.000 örnek ve 10 öznitelik içermektedir.
- Bu özelliklerden 6 tanesi bilgilendirici, 2 tanesi ise gereksiz tekrar içermektedir.
- Her sınıf için tek bir küme oluşturulmuştur.
- Sınıf oranları [0,7; 0,3] şeklinde ayarlanarak, iki sınıf arasında %70'e %30'luk bir dağılım sağlanmıştır.
- Rastgelelik için kullanılan "random_state"=42 ile tekrar üretilebilir sonuçlar elde edilebilir.

5.1.2. Tekstil Veri Seti

Bu çalışmada kullanılan "Tekstil" veri setinin bir bölümü üretim sürecinde otomatik olarak elde edilirken, diğer bölümü ise ilgili kumaşlarda kullanılan hammaddelerin kalite kontrolüne ait bilgiler sistemden çekilerek oluşturulmuştur. Proje kapsamında elde edilen veriler, dokuma

sırasında iplik kopma problemine etki eden temel parametreleri içermektedir. Bu verilerde hem üretime dair özellikler (ör. Final Bobin Sarım Tansiyonu, Final Hız, Final Yağ Devri) hem de kalite kontrol parametreleri (Kopma Yükü, Mukavemet, Denye, Uzama vb.) bulunmaktadır. Veri seti üzerinde sayısal değişkenler arasındaki ilişkinin ilk analizi, Şekil. 1'te sunulan korelasyon matrisi yardımıyla gerçekleştirilmiştir. Bu matriste, örneğin NUMARA_DENYE ile KOPMA_YUKU arasında orta düzeyde pozitif korelasyon gözlenirken, MUKAVEMET ve FinalYagDevri gibi parametreler arasında daha yüksek korelasyon değerleri dikkat çekmektedir. Çok yüksek korelasyon gözlemlenmediği için tüm parametreler çalışmada kullanılmıştır.



Şekil 1:
Sayısal değişkenler arasındaki korelasyon matrisi

Tablo 2 incelendiğinde, veri setine ait temel tanımlayıcı istatistiklerin oldukça farklı aralıklara sahip olduğu görülmektedir. Örneğin, KOPMA_YUKU ortalama değeri, 866,58 ve standart sapması 504,38 ile geniş bir yayılım sergilemektedir. Benzer biçimde, NUMARA_DENYE (ortalama 333,64, standart sapma 256,95) de üretim sürecindeki koşulların çeşitliliğini yansıtan yüksek bir varyasyon göstermektedir. MUKAVEMET ve UZAMA gibi değişkenler nispeten dar bir aralıkta kalmakla birlikte, örneğin MUKAVEMET'in 1'den 4'e kadar uzanması dikkat çekicidir. Ayrıca, FinalBobinSarimTansiyonu, FinalBobinTipi, FinalHiz, FinalSarimBobinCapi ve FinalYagDevri gibi diğer üretim parametreleri de ortalamaları ve maksimum–minimum değerleri dikkate alındığında belirgin biçimde farklılaşmaktadır (örneğin FinalHiz ortalama 597,67 olup 300–800 aralığında gözlenmektedir). Son olarak, SinifEtiketi değişkeni veri setini iki ayrı sınıfa ayırmakta olup, sınıflandırma modellemesinde bu etiket değerleri kullanılmaktadır. Sonraki bölümlerde, farklı sınıflandırma algoritmaları ve veri dengeleme yöntemleri ile elde edilen deneysel sonuçlar ayrıntılı şekilde ele alınacaktır.

Tablo 2. Tekstil veri setine ait tanımlayıcı istatistikler

Öznitelik	count	mean	std	min	25%	50%	75%	max
KOPMA_YUKU	1708	866,58	504,38	4	511	763,5	1112	2894
MUKAVEMET	1708	2,50	1,10	1	1	3	3	4
NUMARA_DENYE	1708	333,64	256,95	53	176	307	346	1810
UZAMA	1708	15,97	13,32	1	6	17	22	152
FinalBobinSarimTansiyonu	1708	26,94	7,83	2	28	28	30	40
FinalBobinTipi	1708	1,83	0,68	1	1	2	2	3
FinalHiz	1708	597,67	108,63	300	500	600	700	800
FinalSarimBobinCapi	1708	18,30	16,00	12	16	16	17	161
FinalYagDevri	1708	3,37	3,25	0	0	2	6	10
SinifEtiketi	1708	1,04	0,19	1	1	1	1	2

5.2. Sonuçların Karşılaştırılması

İlk aşamada, ADASYN sonrasında uygulanan SVM ile oluşturulan marjine en uzak %k çoğunluk verisini silme yaklaşımındaki k düzeyinin (En Uzak %5 Sil, En Uzak %10 Sil, En Uzak %15 Sil, En Uzak %25 Sil) sınıflandırma başarımı üzerindeki etkisi incelenmiştir; her veri kümesi × model kırılımı için dört ayar uygulanmış ve G-Ortalamalar performans değerleri Tablo 3'te raporlanmıştır. Karşılaştırmalar aynı örnekler üzerinde gerçekleştirilmiş olup, n=40 için iki yönlü Wilcoxon işaretli sıralar testi (α=0,05) temelinde yapılmıştır. Sonuçlara göre En Uzak %10'un En Uzak %15'e üstünlüğü istatistiksel olarak anlamlı bulunmuştur (W=224, p=0,0115; medyan $\Delta\approx$+0,0090). Ayrıca En Uzak %5'in En Uzak %25'e (W=230, p=0,0147; medyan $\Delta\approx$+0,0131) ve En Uzak %10'un En Uzak %25'e (W=253,5, p=0,0356; medyan $\Delta\approx$+0,0130) üstünlükleri anlamlıdır. En Uzak %5–En Uzak %15 karşılaştırması marjinal düzeydedir (W=263, p=0,0482), En Uzak %5–En Uzak %10 (W=305, p=0,342) ve En Uzak %15–En Uzak %25 (W=293, p=0,118) çiftlerinde ise anlamlı fark saptanmamıştır. Etkilerin büyüklüklerinin küçük olduğu gözlenmekle birlikte, En Uzak %25 düzeyi genel olarak zayıf, En Uzak %10 ise En Uzak %15'e kıyasla tercih edilebilir bir seçim olarak ortaya çıkmaktadır. Bu nedenle karşılaştırmalarda ADASYN + En Uzak %10 Sil kullanılmıştır.

Tablo 3. ADASYN + En Uzak %k Sil (k∈ {5, 10, 15, 25}) Uygulamasının Sınıflandırma Performansına Etkisi (G-Ortalamalar)

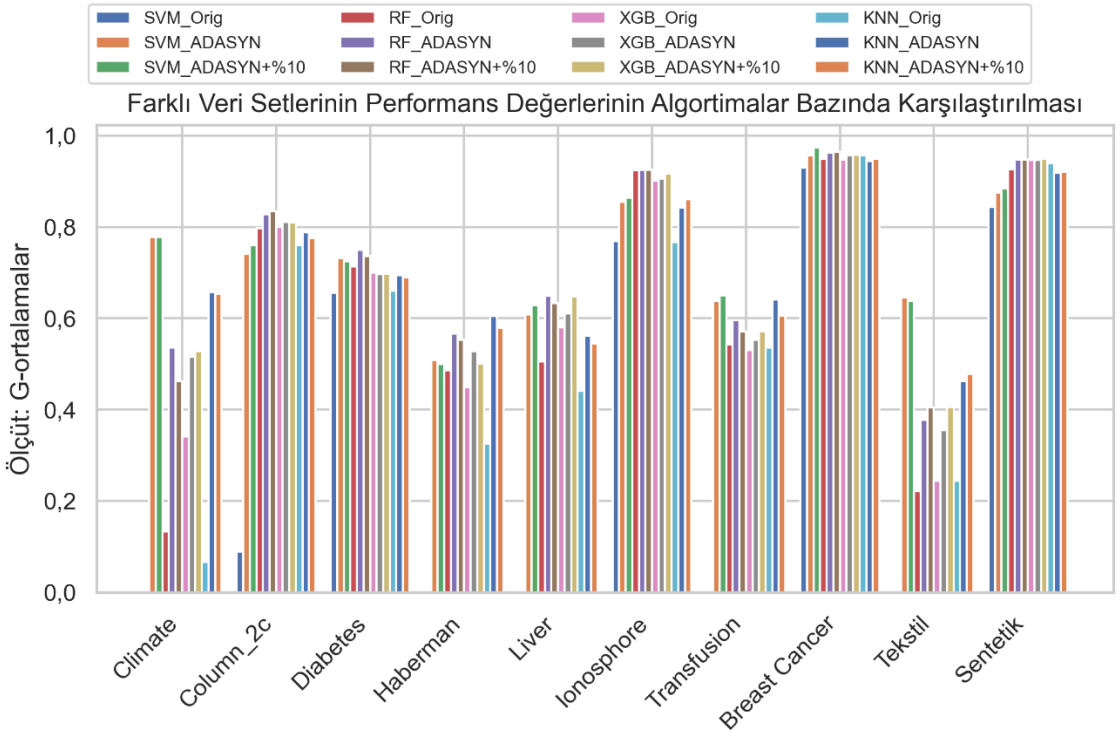
	SVM				RF				XGBoost				KNN			
Veri Setleri	ADASYN + En Uzak %5 Sil	ADASYN + En Uzak %10 Sil	ADASYN + En Uzak %15 Sil	ADASYN + En Uzak %25 Sil	ADASYN + En Uzak %5 Sil	ADASYN + En Uzak %10 Sil	ADASYN + En Uzak %15 Sil	ADASYN + En Uzak %25 Sil	ADASYN + En Uzak %5 Sil	ADASYN + En Uzak %10 Sil	ADASYN + En Uzak %15 Sil	ADASYN + En Uzak %25 Sil	ADASYN + En Uzak %5 Sil	ADASYN + En Uzak %10 Sil	ADASYN + En Uzak %15 Sil	ADASYN + En Uzak %25 Sil
Climate	0,7703	0,7777	0,7847	0,7845	0,5346	0,4628	0,5205	0,5559	0,5347	0,5279	0,5697	0,5451	0,6624	0,6534	0,6676	0,6396
Column_2c	0,7549	0,7595	0,7583	0,7540	0,8235	0,8346	0,8362	0,8109	0,8127	0,8095	0,7846	0,7825	0,7803	0,7762	0,7588	0,7394
Diabetes	0,7357	0,7252	0,7144	0,7015	0,7385	0,7358	0,7233	0,7247	0,7079	0,6980	0,6900	0,6859	0,6977	0,6899	0,6928	0,6731
Haberman	0,5092	0,4997	0,5170	0,5221	0,5783	0,5529	0,5426	0,5663	0,5035	0,5006	0,5174	0,5186	0,5765	0,5790	0,5681	0,5275
Liver	0,6389	0,6292	0,6108	0,5802	0,6316	0,6338	0,6094	0,6181	0,6103	0,6478	0,6316	0,6233	0,5513	0,5446	0,5322	0,5524
Ionosphore	0,8642	0,8642	0,8540	0,8491	0,9124	0,9254	0,9146	0,9091	0,9114	0,9170	0,9108	0,9123	0,8568	0,8612	0,8467	0,8262
Transfusion	0,6452	0,6506	0,6396	0,5747	0,6027	0,5713	0,5571	0,5367	0,5870	0,5709	0,5345	0,5293	0,6405	0,6061	0,5788	0,5941
Breast Cancer	0,9646	0,9744	0,9673	0,9507	0,9645	0,9649	0,9648	0,9612	0,9599	0,9582	0,9587	0,9631	0,9490	0,9490	0,9505	0,9535
Tekstil	0,6408	0,6375	0,6276	0,5929	0,3992	0,4050	0,4126	0,4999	0,3541	0,4062	0,3776	0,5005	0,4446	0,4785	0,4856	0,5315
Sentetik	0,8858	0,8848	0,8814	0,8803	0,9512	0,9484	0,9359	0,9151	0,9501	0,9486	0,9428	0,9177	0,9237	0,9210	0,9197	0,9217

Tablo 4'te, farklı veri setleri üzerinde, ADASYN ile işlenmiş veri, önerilen yöntem olan ADASYN + En Uzak %10 Sil ile işlenmiş veri ve işlenmemiş orijinal verinin, çeşitli sınıflandırıcılar (SVM, RF, XGBoost, KNN) kullanılarak ve 5-katlı çapraz doğrulama uygulanarak elde edilen G-Ortalamalar performans değerleri sunulmaktadır. G-Ortalamalar her veri setinin dengesiz doğasını dikkate alarak, azınlık sınıfın başarı oranını (recall) ve çoğunluk sınıfın doğruluğunu (specificity) birlikte yansıtan önemli bir metrik olarak tercih edilmiştir. Tablodaki her bir veri seti için ve her sınıflandırıcı yöntem kombinasyonu için elde

edilen değerler, Şekil 2'de de sütunlar halinde sunulmuştur. Her veri setine ait 12 sütun (SVM, RF, XGBoost, KNN × 3 strateji) gruplar halinde gösterilerek en yüksek skorun hangi kombinasyonla elde edildiği kolayca görülebilmektedir.

Tablo 4. Çapraz Doğrulama ile Sınıflandırıcıların G-Ortalamalar Performansı (kalın değerler satır bazında en yüksek performansı göstermektedir).

	SVM			RF			XGBoost			KNN		
Veri Setleri	Orijinal	ADASYN	ADASYN + En Uzak %10 Sil	Orijinal	ADASYN	ADASYN + En Uzak %10 Sil	Orijinal	ADASYN	ADASYN + En Uzak %10 Sil	Orijinal	ADASYN	ADASYN + En Uzak %10 Sil
Climate	0,0000	**0,7778**	0,7777	0,1330	0,5364	0,4628	0,3410	0,5151	0,5279	0,0667	0,6570	0,6534
Column_2c	0,0894	0,7410	0,7595	0,7969	0,8274	**0,8346**	0,8002	0,8106	0,8095	0,7601	0,7886	0,7762
Diabetes	0,6560	0,7313	0,7252	0,7140	**0,7493**	0,7358	0,7005	0,6968	0,6980	0,6607	0,6949	0,6899
Haberman	0,0000	0,5086	0,4997	0,4862	0,5661	0,5529	0,4487	0,5278	0,5006	0,3252	**0,6053**	0,5790
Liver	0,0000	0,6084	0,6292	0,5050	**0,6486**	0,6338	0,5806	0,6110	0,6478	0,4411	0,5621	0,5446
Ionosphore	0,7690	0,8555	0,8642	0,9237	0,9251	**0,9254**	0,9018	0,9058	0,9170	0,7668	0,8426	0,8612
Transfusion	0,0000	0,6382	**0,6506**	0,5424	0,5962	0,5713	0,5303	0,5533	0,5709	0,5359	0,6412	0,6061
Breast Cancer	0,9296	0,9568	**0,9744**	0,9493	0,9620	0,9649	0,9480	0,9565	0,9582	0,9566	0,9447	0,9490
Tekstil	0,0000	**0,6460**	0,6375	0,2211	0,3777	0,4050	0,2440	0,3550	0,4062	0,2439	0,4623	0,4785
Sentetik	0,8435	0,8758	0,8848	0,9267	0,9484	0,9484	0,9464	0,9465	**0,9486**	0,9398	0,9189	0,9210



Şekil 2:
Farklı veri setlerinin performans değerlerinin algoritmalar bazında karşılaştırılması

Her bir veri seti, dengesizlik oranı değerine göre aşağıda belirtilen üç grup içerisinde değerlendirilmiş ve yorumlanmıştır.

5.2.1. *Düşük Düzey Dengesizlik (Breast Cancer, Diabetes, Ionosphore)*

Bu grupta yer alan veri setleri (Breast Cancer, Diabetes, Ionosphore) görece düşük bir dengesizlik oranına sahiptir. Tablo 4 ve Şekil 2'den de görülebileceği gibi, genellikle Orijinal strateji dahi tatmin edici bir taban performans sağlamaktadır. Ancak ADASYN ve önerilen

yöntem olan ADASYN + En Uzak %10 Sil uygulamaları, özellikle SVM ve KNN gibi azınlık örneği azlığına karşı hassas yöntemlerde ilave iyileşmeler sunar. Breast Cancer veri setinde, orijinal SVM, 0,9296 gibi zaten yüksek bir değerle başlamaktadır. Önerilen yöntem ile 0,9744'e kadar çıkmaktadır. Benzer şekilde RF ve XGBoost da orijinalde yaklaşık 0,95 civarındayken aşırı örnekleme sonrasında az da olsa yükseliş göstermektedir. Diabetes veri setinde, Orijinal SVM 0,656 düzeyinde iken, ADASYN ile 0,7313'e çıkabilmekte, RF ise 0,714'ten 0,7493'e yükselmektedir. Bu da azınlık sınıfın ek örneklerle desteklenmesinin SVM ve RF üzerinde anlamlı bir katkı sağladığını gösterir. Ionosphore veri setinde orijinal SVM 0,769 düzeyinde olup, önerilen yöntem ile 0,8642'ye kadar çıkmaktadır. RF ise 0,9237 gibi zaten yüksek bir tabanla başlamakta ve ADASYN + En Uzak %10 Sil ile ek artışla 0,9254 seviyesine ulaşmaktadır. Düşük dengesizlik (1–2) seviyesinde orijinal strateji çoğu zaman makul olsa da ADASYN tabanlı yöntemler önemli ölçüde ek kazanç getirebilmektedir.

5.2.2. *Orta Düzey Dengesizlik (Column_2c, Haberman, Liver, Transfusion, Sentetik)*

Bu grupta bulunan veri setlerinde (Column_2c, Haberman, Liver, Transfusion ve Sentetik), Tablo 4 ve Şekil 2'den de gözlemlenebileceği gibi orijinal strateji kimi zaman makul sonuçlar vermekle birlikte, azınlık örneklerin arttırılması (ADASYN) veya ADASYN + En Uzak %10 Sil yaklaşımı çoğu modelde belirgin iyileşme sağlar. Column_2c veri setinde, Orijinal RF 0,7969'luk bir değere sahipken önerilen yaklaşım ile değer 0,8346'ya yükselmektedir. SVM ise Orijinal veri setinde elde edilen 0,0894'ten değerinden 0,741–0,759 aralığına kadar büyük sıçramalar gösterir. Haberman veri setinde orijinal KNN 0,3252'de kalırken, ADASYN ile 0,6053 düzeyine kadar çıkar; SVM de 0,00'dan 0,5086'ya yükselmektedir. Liver veri setinde orijinal SVM 0,00'dan ADASYN + En Uzak %10 Sil, önerilen yöntemi ile 0,6292'ye dek çıkabilirken, XGBoost sınıflandırmasında performans 0,5806'dan 0,6478'e yükselmektedir. RF de ADASYN ile 0,5050'den 0,6486 bandına çıkmaktadır. Transfusion veri setinde dengesizlik oranı biraz daha yüksekçe olsa da, orijinal SVM 0,00'dan ADASYN + En Uzak %10 Sil ile 0,6506'ya kadar iyileşme gösterir. KNN de 0,5359 seviyesinden ADASYN ile 0,6412 seviyesine çıkar. Sentetik veri setinde, orijinal SVM 0,8435 iken ADASYN + En Uzak %10 Sil ile 0,8848'e yükselir. RF ise 0,9267'den 0,9484 düzeyine uzanır; XGBoost da 0,9464 değerinden, önerilen yöntem olan ADASYN + En Uzak %10 Sil ile 0,9486'e yükselmektedir.

5.2.3. *Yüksek Düzey Dengesizlik (Climate, Tekstil)*

Bu gruptaki veri setleri, Climate (~ 10,74) ve Tekstil (~ 25,28) ile oldukça yüksek dengesizliğe sahiptir. Orijinal stratejide SVM/KNN gibi yöntemler 0,00 yakınında çok düşük değerlere inebilir; ancak ADASYN veya ADASYN + En Uzak %10 Sil uygulandığında performans son derece artmaktadır. Climate veri setinde orijinal SVM 0,0000 iken, ADASYN ile 0,7778'e kadar çıkabilmektedir; RF ise 0,1330'dan 0,5364 seviyesine ulaşır. Tekstil verisi, IoT destekli süreç parametrelerinden ve giriş kalite ölçümlerinden derlenmiş, dengesizlik oranı, 25,28 gibi oldukça yüksek bir orana sahiptir. Burada orijinal SVM 0,0000 olup, ADASYN ile 0,6460 değerine ve KNN de 0,2439'dan 0,4623–0,478 aralığına yükselmektedir. Bu, azınlık sınıfın yapay örneklerle güçlendirilmesinin çok kritik olduğunu göstermektedir. Büyük oranda dengesiz veri setlerinde ADASYN temelli yöntemler (gerekirse %10 çoğunluk örnek çıkarma eklemesiyle) modelin öğrenmesini hayli iyileştirmektedir. Özellikle SVM ve KNN gibi azınlık veriye duyarlı yöntemlerde bu sıçramalar belirgindir. RF ve XGBoost orijinalde kısmen daha dirençli olsa bile, ADASYN veya önerilen yöntem genelde ek kazanç sağlar.

Yapılan deneysel çalışmalar sonucunda, ADASYN veya ADASYN + En Uzak %10 Sil yaklaşımının, özellikle SVM ve KNN gibi azınlık veriye duyarlı sınıflandırıcılarda önemli performans artışları sağladığı gözlenmiştir. Orta veya yüksek dengesizlik oranına sahip veri setlerinde, orijinal stratejiyle elde edilen düşük sonuçlar, ADASYN temelli yöntemlerle anlamlı

derecede yükselmektedir. RF ve XGBoost ise çoğunlukla orijinal hâlinde dahi başarılı olmakla birlikte, ADASYN veya ADASYN + En Uzak %10 Sil yöntemleri ile performanslarını artırmışlardır. Sonuç olarak, önerilen ADASYN + En Uzak %10 Sil stratejisinin, incelenen on farklı veri setinin beşinde (Column_2c, Ionosphore, Transfusion, Breast Cancer ve Sentetik) en yüksek G-Ortalamalar değerini verdiği görülmektedir. Bu durum, ele alınan veri setlerinin yarısında "en iyi" performansı sunarak, özellikle dengesiz veri koşullarında azınlık sınıfını daha etkin temsil ettiğini ve sınıf ayrımını daha net bir biçimde ortaya koyduğunu göstermektedir.

6. BULGULAR VE DEĞERLENDİRME

Bu çalışmada, TÜBİTAK TEYDEB destekli 3190654 numaralı araştırma projesi kapsamında elde edilen gerçek üretim verilerinden oluşan Tekstil veri setini de bulunduran toplam on farklı veri seti üzerinde, dengesiz sınıflandırma problemini çözmeye yönelik çeşitli stratejiler incelenmiştir. Veri setleri arasında Sentetik (kurgusal olarak üretilmiş) bir örnek de yer almış; böylece dengesizlik oranları 1,68'den 25,28'e kadar uzanan ve farklı öznitelik sayıları barındıran geniş bir yelpaze oluşturulmuştur. Bu çalışma kapsamında önerilen yöntem, ADASYN ile örneklem artırma ve SVM ile oluşturulan marjin çizgisine en uzak %10 çoğunluk sınıfı örneklerini silmeyi öneren ADASYN + En Uzak %10 Sil stratejisidir. Önerilen yöntem, ADASYN ile örneklem artırma sonrası sınıflandırma performansı ile de karşılaştırılmıştır. Dengesiz veriye herhangi bir müdahale olmadan doğrudan sınıflandırmaya sokulduğu durum da Orijinal başlığı altında değerlendirilerek stratejilerin iyileştirme oranları ölçülmeye çalışılmıştır.

Dört farklı sınıflandırma algoritması (SVM, RF, XGBoost, KNN), 5-katlı çapraz doğrulama düzeniyle değerlendirilmiş ve performans ölçütü olarak G-Ortalamalar dikkate alınmıştır. Bulgular, dengesizliğin düşük seviyelerde olduğu (1–2 aralığı) veri setlerinde orijinal stratejinin dahi makul sonuçlar verebildiğini, ancak azınlık örneklerinin yetersiz kaldığı durumlarda ADASYN'in özellikle SVM ve KNN için ek fayda sağladığını göstermektedir. Orta düzey dengesizlik (2–3 aralığı) söz konusu olduğunda (Sentetik, Diabetes, Column_2c, Haberman, Liver vb.), orijinal verinin bir "eşik performans" sunduğu, ancak ADASYN ve önerilen yöntem olan ADASYN + En Uzak %10 Sil yönteminin pek çok modelde kayda değer ek kazanç getirdiği tespit edilmiştir. Yüksek dengesizlik oranlarında (3 ve üzeri), özellikle Tekstil (~ 25,28) ve Climate (~ 10,74) veri setlerinde, orijinal stratejinin çoğu zaman çok düşük kaldığı, buna karşın ADASYN veya ADASYN + En Uzak %10 Sil yöntemlerinin performansı çarpıcı biçimde yükselttiği gözlenmiştir.

ADASYN + En Uzak %10 Sil yönteminin, 10 veri setinin 5 adedinde (Column_2c, Ionosphore, Transfusion, Breast Cancer, Sentetik) ek iyileştirmeler sağladığı görülmüştür. Yine de, bazı durumlarda (örneğin Tekstil veya Climate veri setleri) yalnızca ADASYN yaklaşımıyla da yeterli düzeyde başarı elde edilebildiği için, "%10 çoğunluk örneği çıkarma" adımının her veri seti için zorunlu olmadığı anlaşılmaktadır. Özellikle SVM ve KNN gibi azınlık örneklerine duyarlı modellerde bu ek işlemin önemli katkılar sağlayabildiği, RF ve XGBoost 'un ise zaten daha yüksek başlangıç puanlarına rağmen veri işlemeden ek yarar gördüğü saptanmıştır.

Tekstil veri seti, kalite kontrol aşamasından geçmiş ipliğin, dokuma aşamasında kopup kopmadığını ikili şekilde etiketleyen, giriş kontrol ve IoT sensörlerinden (kırılma yükü, bobin sarım tansiyonu vb.) gelen üretim parametrelerini içeren bir yapıdadır. Söz konusu veri setinde dengesizlik oranının 25,28 gibi son derece yüksek olması, orijinal stratejide SVM veya KNN yöntemlerinin neredeyse başarısızlığa düştüğünü göstermiştir. Buna karşın, performans metriklerinin iyileştirilmesi için, örneklem artırma, azaltma yöntemlerinin kritik rol oynadığı ortaya koyulmaktadır. Öte yandan, düşük ve orta düzeyde dengesiz veri setlerinin yarısında ADASYN + En Uzak %10 Sil ile verinin orijinal haline göre ek iyileştirmeler gözlenmiş ve dengesizliğin daha etkin biçimde giderildiği anlaşılmıştır.

Gelecek çalışmalarda, yöntemin farklı veri tiplerine (zaman serisi IoT sinyalleri, metin/görüntüden türetilmiş öznitelikler) uygulanmasıyla elde edilecek sonuçlar

değerlendirilebilir. Ayrıca dengesiz veri bağlamında maliyet duyarlı veya ceza terimli sınıflandırıcılarla sistematik karşılaştırmalar yapılabilir. "En Uzak %k Sil" eşiğinin veri kümesine duyarlı olarak dinamik seçimi konusunda çalışılabilir. İlave olarak, verilerdeki dengesizlik problemini çözmeye yönelik, yeniden örnekleme (artırma/azaltma) stratejilerinin hibrit olarak ya da sınıflandırıcı modellerinin içine gömülerek çalıştırılması durumundaki performans değişimleri de araştırılabilir.

ÇIKAR ÇATIŞMASI

Yazar, bu çalışmayla ilgili olarak bilinen herhangi bir çıkar çatışması veya herhangi bir kurum/kuruluş ya da kişi ile ortak çıkar bulunmadığını onaylamaktadır.

YAZAR KATKISI

Bu çalışma tek yazar tarafından yürütülmüş ve tüm aşamaları yazar tarafından gerçekleştirilmiştir.

TEŞEKKÜR

Bu çalışmada kullanılan Tekstil gerçek veri seti, Güncel Yazılım Ar-Ge Merkezi ile Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK-TEYDEB-3190654) tarafından desteklenen bir proje kapsamında toplanmıştır.

KAYNAKLAR

1. Abdelkhalek, A. ve Mashaly, M., (2023) Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning, *The journal of* Supercomputing 79, 10611–10644. doi:10.1007/s11227-023-05073-x

2. Almarshdi, R., Nassef, L., Fadel, E. ve Alowidi, N. (2023) Hybrid deep learning based attack detection for imbalanced data classification, *Intelligent Automation & Soft Computing* 35. doi:10.32604/iasc.2023.026799

3. Altunkaynak, A., Başakın, E.E. ve Kartal, E. (2020) Dalgacik k-en yakin komşuluk yöntemi ile hava kirliliği tahmini, *Uludağ University Journal of The Faculty of Engineering,* 25, 1547–1556. doi:10.17482/uumfd.809938

4. Amirshahi, B. ve Lahmiri, S. (2024) Bankruptcy prediction using optimal ensemble models under balanced and imbalanced data, *Expert Systems*, 41, e13599. doi:10.1111/exsy.13599

5. Avizheh, M., Dadpasand, M., Dehnavi, E. ve Keshavarzi, H. (2023) Application of machine-learning algorithms to predict calving difficulty in Holstein dairy cattle, *Animal Production Science*, 63, 1095–1104. doi: 10.1071/AN22461

6. Aymaz, S. (2025) Unlocking the power of optimized data balancing ratios: a new frontier in tackling imbalanced datasets, *The Journal of Supercomputing*, 81, 443. doi:10.1007/s11227-025-06919-2

7. Ercire, M. ve Ünsal, A. (2021) Kısa Süreli Güç Kalitesi Bozulmalarının Dalgacık Analizi ve Rastgele Orman Yöntemi ile Sınıflandırılması, *Uludağ University Journal of The Faculty of Engineering,* 26, 903–920. doi:10.17482/uumfd.976342

8. Ferreira, G., Das, S., Coelho, G., Silva, R. R., Goswami, S., Pereira, R. N., Pereira, L., Fortunato, E., Martins, R. ve Nandy, S (2025) Energy harvesting and movement tracking by

polypyrrole functionalized textile for wearable IoT applications, *Journal of Energy Chemistry*, 102, 230–242. doi: 10.1016/j.jechem.2024.10.028

9. Galán-Cuenca, A., Gallego, A. J., Saval-Calvo, M. ve Pertusa, A. (2024) Few-shot learning for COVID-19 chest X-ray classification with imbalanced data: An inter vs. intra domain study. *Pattern Analysis and Applications*, 27, 69. doi: 10.1007/s10044-024-01285-w

10. García-Gil, D., García, S., Xiong, N. ve Herrera, F. (2024) Smart data driven decision trees ensemble methodology for imbalanced big data, *Cognitive Computation*, 16, 1572–1588. doi: 10.1007/s12559-024-10295-z

11. Guo, J., Wu, H., Chen, X. ve Lin, W. (2024) Adaptive SV-Borderline SMOTE-SVM algorithm for imbalanced data classification, *Applied Soft Computing*, 150, 110986. doi: 10.1016/j.asoc.2023.110986

12. He, H., Bai, Y., Garcia, E. A. ve Li, S. (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning, In *Proceedings of the IEEE International Joint Conference on Neural Networks*,1322–1328. doi: 10.1109/IJCNN.2008.4633969

13. He, H. ve Garcia, E. A. (2009) Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21, 1263–1284. doi: 10.1109/TKDE.2008.239

14. Hussein, H. I., Anwar, S. A. ve Ahmad, M. I. (2023) Imbalanced data classification using SVM based on improved simulated annealing featuring synthetic data generation and reduction, *Computers, Materials & Continua*, 75, 547–564. doi: 10.32604/cmc.2023.036025

15. Lappage, J. (2005) End breaks in the spinning and weaving of weavable singles yarns: Part 2: End breaks in weaving, *Textile Research Journal*, 75, 512–517. doi: 10.1177/0040517505053869

16. López, V., Fernández, A., García, S., Palade, V. ve Herrera, F. (2013) An insight into classification with imbalanced data, *Information Sciences*, 250, 113–141. doi: 10.1016/j.ins.2013.07.007

17. Markwald, M. ve Demidova, E. (2024) REFUEL: rule extraction for imbalanced neural node classification, *Machine Learning*, 113, 6227–6246. doi: 10.1007/s10994-024-06569-0

18. Martikkala, A., Mayanti, B., Helo, P., Lobov, A. ve Ituarte, I. F. (2023) Smart textile waste collection system – dynamic route optimization with IoT, *Journal of Environmental Management*, 335, 117548. doi: 10.1016/j.jenvman.2023.117548

19. Mebrate, M., Gessesse, N. ve Zinabu, N. (2022) Effect of loom tension on mechanical properties of plain woven cotton fabric, *Journal of Natural Fibers*, 19, 1443–1448. doi: 10.1080/15440478.2020.1776663

20. Militino, A. F., Goyena, H., Pérez-Goya, U. ve Ugarte, M. D. (2024) Logistic regression versus XGBoost for detecting burned areas using satellite images, *Environmental and Ecological Statistics*, 31, 57–77. doi: 10.1007/s10651-023-00590-7

21. Morgan, H., Elgendy, A., Said, A., Hashem, M., Li, W., Maharjan, S. ve El-Askary, H. (2024) Enhanced lithological mapping in arid crystalline regions using explainable AI and multi-spectral remote sensing data, *Computers & Geosciences*, 193, 105738. doi: 10.1016/j.cageo.2024.105738

22. Mu, X. ve Zhao, B. (2025) DCS-SOCP-SVM: A novel integrated sampling and classification algorithm for imbalanced datasets, *Computers, Materials & Continua*, 83, 2143–2159. doi: 10.32604/cmc.2025.060739

23. My, B. T. ve Ta, B. Q. (2023) An interpretable decision tree ensemble model for imbalanced credit scoring datasets, *Journal of Intelligent & Fuzzy Systems*, 45, 10853–10864. doi: 10.3233/JIFS-230825

24. Özdet, B. ve İçer, S. (2022) AKCİĞER BİLGİSAYARLI TOMOGRAFİ GÖRÜNTÜLERİNDE GÖRÜNTÜ İŞLEME UYGULAMALARI İLE TÜMÖRLERİNİN TESPİT EDİLMESİ, *Uludağ University Journal of The Faculty of Engineering,* 27, 135–150. doi:10.17482/uumfd.947619

25. Peng, C. ve Yuan, X. (2021) A study on detection of roving tension and fine control of yarn breakage, *Industria Textila,* 72, 250–255. doi:10.35530/IT.072.03.1757

26. Pérez Tárraga, J., Castillo-Cara, M., Arias-Antúnez, E. ve Dujovne, D. (2025) Frost forecasting through machine learning algorithms, *Earth Science Informatics,* 18, 183. doi:10.1007/s12145-025-01710-6

27. Stanosheck, J.A., Castell-Perez, M.E., Moreira, R.G., King, M.D. ve Castillo, A. (2024) Oversampling methods for machine learning model data training to improve model capabilities to predict the presence of Escherichia coli MG1655 in spinach wash water, *Journal of Food Science,* 89, 150–173. doi:10.1111/1750-3841.16850

28. Tolentino-Zondervan, F. ve DiVito, L. (2024) Sustainability performance of Dutch firms and the role of digitalization: The case of textile and apparel industry, *Journal of Cleaner Production*, 459, 142573. doi:10.1016/j.jclepro.2024.142573

29. Wang, A.X., Chukova, S.S. ve Nguyen, B.P. (2023a) Ensemble k-nearest neighbors based on centroid displacement, *Information Sciences*, 629, 313–323. doi:10.1016/j.ins.2023.02.004

30. Wang, X., Ren, H., Ren, J., Song, W., Qiao, Y., Ren, Z., Zhao, Y., Linghu, L., Cui, Y., Zhao, Z. ve diğ. (2023) Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data, *Computer Methods and Programs in Biomedicine*, 230, 107340. doi:10.1016/j.cmpb.2023.107340

31. Wang, Z. ve Liu, Q. (2023) Imbalanced data classification method based on LSSASMOTE, *IEEE Access*, 11, 32252–32260. doi:10.1109/ACCESS.2023.3262460

32. Weiss, G.M. (2004) Mining with rarity: a unifying framework, *ACM SIGKDD Explorations Newsletter*, 6, 7–19. doi: 10.1145/1007730.1007734

33. Wibbeke, J., Rohjans, S. ve Rauh, A. (2025) Quantification of data imbalance, *Expert Systems*, 42, e13840. doi: 10.1111/exsy.13840

34. Wu, L.j., Li, X., Yuan, J.d. ve Wang, S.j. (2023) Real-time prediction of tunnel face conditions using XGBoost random forest algorithm*, Frontiers of Structural and Civil Engineering*, 17, 1777–1795. doi: 10.1007/s11709-023-0044-4

35. Yao, G., Guo, J. ve Zhou, Y. (2005) Predicting the warp breakage rate in weaving by neural network techniques, *Textile Research Journal*, 75, 274–278. doi:10.1177/004051750507500

36. Yilmaz Eroglu, D. ve Pir, M.S. (2024) Hybrid oversampling and undersampling method (HOUM) via safe-level SMOTE and support vector machine, *Applied Sciences*, 14, 10438. doi: 10.3390/app142210438

37. Zhou, T., Gao, X., Sun, X. ve Han, L. (2024) Split difference weighting: An enhanced decision tree approach for imbalanced classification, *International Journal of Computers Communications & Control*, 19. doi: 10.15837/ijccc.2024.6.6702