ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

Uluslararası Sürdürülebilir Mühendislik ve Teknoloji Dergisi
International Journal of Sustainable Engineering and Technology
ISSN: / Cilt: 9, Sayı: 1, (2025), 41 – 50
Araştırma Makalesi – Research Article

KONTROLLÜ DENGESİZLİK SENARYOLARINDA TOPLULUK ÖĞRENME MODELLERİN SİSTEMATİK KARŞILAŞTIRMASI

*Muhammed Abdulhamid KARABIYIK[1]

[1]Niğde Ömer Halisdemir Üniversitesi, Bor MYO, Bilgisayar Teknolojileri Bölümü, Niğde

(Geliş/Received: 19.05.2025, Kabul/Accepted: 06.06.2025, Yayınlanma/Published: 30.06.2025)

ÖZ

Bu çalışma, sınıf dengesizliğinin topluluk öğrenme algoritmaları üzerindeki etkisini kontrollü bir deneysel tasarım ile incelemeyi amaçlamaktadır. Çalışma kapsamında, Iris ve Wine veri setleri üzerinde dört farklı sınıf dağılımı senaryosu (orijinal, hafif, orta ve şiddetli dengesizlik) uygulanmış ve her senaryoda Random Forest, Gradient Boosting ve Bagging algoritmaları test edilmiştir. Değerlendirmelerde yalnızca doğruluk değil, aynı zamanda Macro-F1, Balanced Accuracy, G-Mean ve Cohen Kappa gibi çoklu performans metrikleri kullanılmıştır. Elde edilen bulgular, Gradient Boosting modelinin yüksek dengesizlik düzeylerinde ciddi performans kayıpları yaşadığını; buna karşılık Random Forest algoritmasının tüm senaryolarda kararlı ve güvenilir sonuçlar sunduğunu ortaya koymuştur. Bu yönüyle çalışma, sınıf dengesizliğine karşı dayanıklı model seçiminin ve çok boyutlu metriklerle yapılan değerlendirmelerin önemini vurgulamaktadır.

Anahtar kelimeler: Sınıf dengesizliği, topluluk öğrenme, Random Forest, performans metrikleri, dengesiz veri setleri

A SYSTEMATIC COMPARISON OF ENSEMBLE MODELS UNDER CONTROLLED CLASS IMBALANCE SCENARIOS

ABSTRACT

This study aims to investigate the impact of class imbalance on ensemble learning algorithms through a controlled experimental design. Four different class distribution scenarios (original, mild, moderate, and severe imbalance) were applied to the Iris and Wine datasets, and three ensemble models Random Forest, Gradient Boosting, and Bagging were evaluated in each scenario. Model performance was assessed using not only accuracy but also multiple metrics such as Macro-F1, Balanced Accuracy, G-Mean, and Cohen's Kappa. The findings reveal that Gradient Boosting suffered significant performance degradation under high imbalance conditions, while Random Forest consistently delivered stable and reliable results across all scenarios. These results highlight the importance of selecting imbalance-resilient models and conducting evaluations using multiple performance indicators in imbalanced classification tasks.

Keywords: Class imbalance, ensemble learning, Random Forest, performance metrics, imbalanced datasets

***Sorumlu Yazar/Corresponding Author:** abdulhamidkarabiyik@ohu.edu.tr

ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

1. Giriş (Introduction)

Makine öğrenmesi, günümüzde çok çeşitli alanlarda karar verme süreçlerini otomatikleştirmek ve optimize etmek amacıyla yaygın olarak kullanılmaktadır [1]. Bu uygulamaların önemli bir kısmı, sınıflandırma problemleri etrafında şekillenmektedir [2]. Ancak sınıflandırma modellerinin başarımı, yalnızca kullanılan algoritmanın gücüne değil, aynı zamanda verinin yapısal özelliklerine de doğrudan bağlıdır. Gerçek dünya veri setlerinin önemli bir kısmı, sınıflar arasında ciddi örnek sayısı dengesizlikleri içermektedir [3]. Bu durum, özellikle azınlık sınıfların yeterince temsil edilemediği durumlarda, makine öğrenmesi modellerinin bu sınıfları doğru tahmin edememesine neden olur [4]. Özellikle doğruluk (accuracy) gibi yüzeysel metrikler, dengesiz veri setlerinde yanıltıcı sonuçlar doğurabilir. Bu tür durumlarda, topluluk öğrenme yöntemleri hem genel başarıyı artırmak hem de model kararlılığını iyileştirmek amacıyla sıklıkla tercih edilmektedir [5]. Ancak, bu modellerin sınıf dengesizliğine karşı gösterdiği performansın ne ölçüde tutarlı ve güvenilir olduğu halen araştırılmaya değer bir konudur.

Topluluk öğrenme yöntemlerinin sınıflandırma performansını artırmadaki başarısı, literatürde birçok çalışma tarafından gösterilmiştir [6–8]. Bu modellerin dengesiz veri setlerinde kullanımı da farklı araştırmalarda ele alınmış; çoğu zaman örnekleme stratejileri veya ağırlıklandırma teknikleriyle birlikte değerlendirilmiştir. Ancak bu çalışmaların önemli bir bölümü, sınıf dengesizliğini ikili sınıflandırma problemleriyle sınırlı bir çerçevede ele almakta ve analizlerini çoğunlukla yalnızca doğruluk, AUC veya ROC eğrisi gibi sınırlı metrikler üzerinden gerçekleştirmektedir [9–12]. Bununla birlikte, çok sınıflı dengesiz veri setlerinde farklı dengesizlik seviyeleri altında topluluk modellerinin ne ölçüde kararlı ve başarılı sonuçlar verdiğine dair sistematik ve çok metrikli karşılaştırmalara oldukça sınırlı sayıda çalışmada yer verilmektedir. Ayrıca, küçük ölçekli ve çok sınıflı veri setleriyle yapılan kontrollü deneylerin azlığı, bu alandaki genel geçer çıkarımların geçerliliğini de sorgulanabilir hale getirmektedir. Bu çalışmada, çok sınıflı sınıflandırma problemlerinde sınıf dengesizliğinin etkisini sistematik biçimde incelemek amacıyla, farklı seviyelerde kontrollü dengesizlik senaryoları oluşturulmuştur. İki küçük ölçekli veri seti (Iris ve Wine) üzerinde, üç popüler topluluk öğrenme modeli olan Karar Ağaçları Topluluğu (Random Forest), Artırmalı Öğrenme (Gradient Boosting) ve Torbalama (Bagging) yöntemleri değerlendirilmiştir. Her veri seti için eğitim ve test verileri sabit tutulmuş; yalnızca eğitim verisine uygulanan %60, %80 ve %90 oranlarındaki örnek dengesizlikleriyle model kararlılığı gözlemlenmiştir. Böylece modellerin aynı test seti üzerinde, farklı eğitim senaryolarında gösterdiği performans farklılıkları adil bir biçimde karşılaştırılabilmiştir. Değerlendirme sürecinde yalnızca doğruluk değil, aynı zamanda Macro-F1 skoru, Dengeleştirilmiş Doğruluk (Balanced Accuracy), G-Mean, Cohen Kappa, Hassasiyet (Precision) ve Duyarlılık (Recall) gibi çoklu metrikler dikkate alınarak, her modelin güçlü ve zayıf yönleri çok boyutlu bir yaklaşımla analiz edilmiştir.

Sunulan yaklaşımın temel katkısı, topluluk öğrenme yöntemlerinin çok sınıflı sınıflandırma problemlerinde sınıf dengesizliği karşısındaki dayanıklılığını, çok metrikli ve sistematik bir yaklaşımla değerlendirmesidir. Özellikle sabit test seti kullanılarak yalnızca eğitim verisi üzerinde uygulanan farklı dengesizlik seviyeleri sayesinde, modellerin öğrenme davranışları ve genelleme kapasiteleri doğrudan gözlemlenebilmiştir. Ayrıca, çok sayıda değerlendirme metriğinin birlikte kullanılması, yalnızca doğruluk odaklı değil, bütüncül bir performans analizi yapılmasını sağlamıştır. Elde edilen bulgular, hem akademik çalışmalar için referans niteliğinde karşılaştırmalı sonuçlar sunmakta hem de dengesiz veri setleriyle çalışan uygulayıcılar için model seçimi konusunda pratik çıkarımlar üretmektedir. Makalenin devamında, ikinci bölümde araştırmanın yöntemi, kullanılan veri setleri ve uygulama adımları detaylandırılmış; üçüncü bölümde deneysel bulgular sunulmuş dördüncü bölümde sonuçlar tartışılmış son bölümde ise genel değerlendirmelere ve gelecekteki çalışma önerilerine yer verilmiştir.

2. Materyal ve Metot (Material and Method)

Bu bölümde, çalışmada kullanılan veri setleri, uygulanan sınıf dengesizliği senaryoları, tercih edilen topluluk öğrenme modelleri, performans ölçütleri ve deneysel düzenek ayrıntılı olarak açıklanmıştır.

2.1. Veri seti (Dataset)

Sınıf dengesizliğinin etkilerini inceleyebilmek amacıyla, çok sınıflı yapıya sahip iki küçük ölçekli veri seti kullanılmıştır. Bu veri setleri Iris ve Wine veriş setleridir [13,14]. Her iki veri setinin temel

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr

ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

istatistiksel özellikleri Tablo 1'de özetlenmiştir. Dengeli başlangıç yapıları sayesinde her iki veri seti de farklı dengesizlik düzeylerinin kontrollü olarak uygulanmasına olanak sağlamaktadır.

Tablo 1. Kullanılan veri setlerinin temel özellikleri

Veri Seti	Örnek Sayısı	Özellik Sayısı	Sınıf Sayısı	Sınıf DağlımıF1-Score
Iris	150	4	3	50/50/50
Wine	178	13	3	59/71/48

Tablo 1 incelendiğinde, her iki veri setinin de çok sınıflı yapıya sahip olduğu ve sınıflar arasında başlangıçta Iris veri seti tam ve Wine veri seti hafif dengeli bir dağılım gösterdiği görülmektedir. Iris veri seti tamamen dengeli bir yapıya sahipken, Wine veri setindeki küçük dengesizlik bile bazı sınıflarda temsil oranlarını anlamlı biçimde etkileyebilmektedir. Özellik sayısı bakımından Iris veri seti daha sade ve düşük boyutlu bir yapı sergilerken, Wine veri seti daha fazla özelliğe sahip olması nedeniyle modelleme açısından daha karmaşık bir örüntü sunmaktadır. Bu farklılıklar, topluluk öğrenme modellerinin hem sınıf dengesizliği hem de veri karmaşıklığı karşısındaki dayanıklılığını test etme açısından değerli bir karşılaştırma ortamı sağlamaktadır.

Her veri seti eğitim ve test olmak üzere ikiye ayrılmıştır. Dengesizlik senaryoları yalnızca eğitim verisi üzerinde uygulanmış; test seti tüm senaryolarda sabit tutularak modellerin farklı öğrenme koşulları altındaki performanslarının nesnel biçimde kıyaslanması sağlanmıştır.

2.2. Sınıf Dengesizliği Senaryoları (Class Imbalance Scenarios)

Gerçek dünya veri setlerinde sıklıkla karşılaşılan sınıf dengesizliği problemi, modellerin özellikle azınlık sınıfları doğru tahmin etme kapasitesini önemli ölçüde sınırlayabilmektedir. Bu durumu kontrollü bir biçimde analiz edebilmek amacıyla, eğitim verileri üzerinde dört farklı sınıf dağılımı senaryosu oluşturulmuştur: orijinal (tam dengeli) dağılım ve üç yapay dengesizlik düzeyi (hafif, orta ve şiddetli). Bu senaryoların temsil ettiği sınıf oranları Tablo 2'de özetlenmiştir.

Tablo 2. Uygulanan sınıf dengesizlik senaryoları

Senaryo	Sınıf 1 (%)	Sınıf 2 (%)	Sınıf 3 (%)	Açıklama
Orjinal	33.3	33.3	33.3	Tam dengeli eğitim verisi
Hafif	60.0	20.0	20.0	Hafif dengesizlik
Orta	80.0	10.0	10.0	Orta düzey dengesizlik
Şiddetli	90.0	5.0	5.0	Şiddetli dengesizlik

Tablo 2'de verilen oranlar, her bir veri setinin eğitim verisi üzerindeki örnek sayısına göre yeniden örnekleme yoluyla uygulanmıştır. Senaryolar, azınlık sınıfların giderek daha az temsil edildiği kurgusal durumları temsil etmektedir. Özellikle şiddetli dengesizlik senaryosunda, azınlık sınıflar toplamın yalnızca %5'ini oluşturmaktadır. Böylece topluluk öğrenme modellerinin bu zorlu koşullar altındaki kararlılığı ve duyarlılığı sistematik olarak test edilebilmiştir.

Dengesizlik yalnızca eğitim verisine uygulanmış; test verileri ise tüm senaryolarda sabit tutulmuştur. Bu yaklaşım sayesinde, modellerin farklı öğrenme senaryoları karşısında eşit koşullarda değerlendirilmesi sağlanmış ve performans farklarının yalnızca öğrenme sürecinden kaynaklanıp kaynaklanmadığı nesnel biçimde analiz edilebilmiştir.

2.3. Topluluk Öğrenme Modelleri (Ensemble Learning Models)

Bu üç modelin tercih edilmesinin temel nedeni, her birinin farklı bir topluluk stratejisini temsil etmesidir. Random Forest, eğitim sırasında kullanılan veri alt kümeleri ve rastgele seçilen özellikler aracılığıyla çeşitlilik sağlar; bu yönüyle aşırı öğrenmeye (overfitting) karşı dirençli bir yapı sunar. Gradient Boosting, modelleri art arda eğiterek önceki hataları düzeltmeye odaklanan artımlı bir yaklaşım benimser. Bu yapı, daha yüksek doğruluk potansiyeli sunsa da, özellikle azınlık sınıfların yanlış

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr

ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

öğrenilmesine karşı daha hassas bir davranış sergileyebilir. Bagging ise bootstrap örnekleme yöntemiyle oluşturulan veri alt kümeleri üzerinde öğreniciler eğiterek model varyansını azaltır ve öğrenme sürecini daha dengeli hale getirir. Bu çeşitlilik, her bir modelin güçlü ve zayıf yönlerinin farklı dengesizlik düzeylerinde nasıl ortaya çıktığını gözlemleme açısından değerli bir karşılaştırma zemini sunmaktadır.

2.3.1. Random Forest

Random Forest, birden fazla karar ağacının bağımsız olarak eğitilmesi ve sonuçların çoğunluk oyu ile birleştirilmesi esasına dayanan bir topluluk yöntemidir. Ağaçlar, bootstrap yöntemiyle oluşturulan farklı veri alt kümeleri üzerinde eğitilirken, her düğümde rastgele seçilen özellikler üzerinden en iyi bölünme noktası belirlenir. Bu yaklaşım, öğrenici modeller arasında çeşitlilik sağlayarak genelleme başarısını artırır [15,16]. Random Forest, sınıf dengesizliği karşısında göreli olarak kararlı sonuçlar verebilmesi ve sınıf etiketlerinden bağımsız olarak bölünme kararları alabilmesi nedeniyle bu çalışmada değerlendirmeye alınmıştır.

2.3.2. Gradient Boosting

Gradient Boosting, zayıf öğrenicilerin (çoğunlukla karar ağaçlarının) art arda eğitildiği ve her yeni modelin önceki modellerin hatalarını düzeltmeye çalıştığı artımlı bir öğrenme yaklaşımıdır. Modelin her adımda eksik kaldığı örüntüleri telafi etmesi, karmaşık karar sınırlarını öğrenebilmesini sağlar. Ancak bu hassasiyet, sınıf dengesizliğine karşı kırılganlığa da neden olabilir [17,18]. Bu modelin çalışmaya dahil edilme nedeni, doğruluk potansiyelinin yüksek olması ve dengesiz örnek dağılımlarında nasıl davrandığının sistematik biçimde gözlemlenmesinin istenmesidir.

2.3.3. Bagging Classifier

Bagging (Bootstrap Aggregating), farklı veri alt örnekleriyle eğitilen zayıf sınıflandırıcıların bir araya getirilmesiyle oluşturulan bir topluluk modelidir. Her bir öğrenici model, bootstrap yöntemiyle rastgele seçilmiş veri alt kümeleriyle eğitilir; daha sonra tahminler birleştirilerek nihai karar üretilir. Bu yaklaşım model varyansını azaltır ve öğrenme sürecini daha dengeli hale getirir [19,20]. Sınıf dengesizliği durumlarında sınıfları doğrudan ağırlıklandırmadan çalışabilmesi, Bagging yönteminin bu çalışmada kullanılmasını motive eden başlıca nedendir.

2.4. Değerlendirme Metrikleri (Evaluation Metrics)

Sınıf dengesizliği içeren çok sınıflı veri setlerinde, yalnızca doğruluk (accuracy) gibi temel metrikler kullanılarak yapılan değerlendirmeler yanıltıcı olabilir. Bu nedenle, model performansının daha adil ve bütüncül bir şekilde ölçülebilmesi için çok sayıda değerlendirme metriği birlikte kullanılmıştır. Kullanılan metrikler hem genel başarıyı hem de azınlık sınıflara yönelik duyarlılığı yansıtan özellikler taşımaktadır. Tablo 3, çalışmada kullanılan başlıca metrikleri ve her birinin kısa açıklamasını özetlemektedir.

Tablo 3. Çalışmada kullanılan değerlendirme metrikleri ve kullanım gerekçeleri

Metrik	Kullanım Gerekçesi
Accuracy (Doğruluk)	Tüm sınıflar dikkate alındığında doğru tahmin edilen örneklerin toplam içindeki oranıdır. Dengesiz veri setlerinde yüksek görünebilir ama azınlık sınıfları ihmal edebilir [21].
Macro-F1 Skoru	Her sınıf için ayrı ayrı hesaplanan F1 skorlarının ortalamasıdır. Sınıf dağılımından bağımsız olduğu için dengesiz veri setlerinde daha adil bir performans ölçütüdür [22].
Balanced Accuracy	Her sınıf için doğru sınıflandırma oranlarının ortalamasıdır. Özellikle sınıf sayıları dengesiz olduğunda daha güvenilir bir genel başarı ölçüsüdür [23].
Precision (Hassasiyet)	Pozitif tahminlerin ne kadarının gerçekten doğru olduğunu gösterir. Yanlış pozitifleri azaltmak açısından önemlidir [24].
Recall (Duyarlılık)	Gerçek pozitiflerin tahmin başarına odaklanan bir metirktir. Azınlık sınıfların tespiti için kritiktir [24].
G-Mean	Her sınıf için duyarlılık skorlarının geometrik ortalamasıdır. Sınıflar arası dengeyi yansıtır, özellikle dengesiz veri setleri için anlamlıdır [25].

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr

ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

Metrik	Kullanım Gerekçesi
Accuracy (Doğruluk)	Tüm sınıflar dikkate alındığında doğru tahmin edilen örneklerin toplam içindeki oranıdır. Dengesiz veri setlerinde yüksek görünebilir ama azınlık sınıfları ihmal edebilir [21].
Cohen's Kappa	Model başarımını rastgele sınıflandırmayla karşılaştırarak düzeltir. Yüksek değerler, modelin tahminlerinin istatistiksel olarak anlamlı olduğunu gösterir [26].
Log Loss	Sınıflara ait tahmin olasılıklarının doğruluğunu ölçer. Modelin ne kadar güvenle tahmin yaptığına dair bilgi verir [27].

Bu metriklerin birlikte kullanılması, yalnızca genel başarıyı değil, aynı zamanda azınlık sınıfların doğru tanınma düzeyini ve modelin kararlılığını da detaylı biçimde analiz etmeye olanak tanımaktadır. Böylece topluluk öğrenme algoritmalarının sınıf dengesizliği altındaki davranışları daha gerçekçi ve çok boyutlu şekilde değerlendirilmiştir.

2.5. Deneysel Altyapı (Experimental Framework)

Tüm deneysel süreçte veri setleri, sınıf dağılımları korunacak şekilde %80 eğitim – %20 test seti olarak ikiye ayrılmıştır. Bu ayrım sırasında katmanlı örnekleme (stratified sampling) yöntemi kullanılmış ve deneylerin tekrarlanabilirliğini sağlamak amacıyla sabit bir rastgelelik tohumu (random seed) tercih edilmiştir [28]. Değerlendirme sürecinde yalnızca eğitim verileri üzerinde dengesizlik senaryoları uygulanmış; test seti tüm senaryolarda sabit tutularak modellerin farklı öğrenme koşulları altında nesnel biçimde karşılaştırılması sağlanmıştır.

Dengesizlik senaryoları, sınıf oranları önceden belirlenmiş dört düzeye göre yapılandırılmıştır: orijinal, hafif dengesizlik, orta düzey dengesizlik ve şiddetli dengesizlik. Her senaryoda, sınıf dağılımları hedeflenen oranlara ulaşacak şekilde yeniden örneklenmiş; baskın sınıflardan rastgele örnek çıkarılarak rastgele alt örnekleme (random undersampling) uygulanmıştır. Bu yöntem, model eğitiminde sınıf dengesizliğini kontrollü biçimde simüle etmeye olanak tanımıştır.

Bu çalışmada, kullanılan üç topluluk öğrenme algoritması (Random Forest, Gradient Boosting, Bagging) varsayılan hiperparametre değerleri ile değerlendirilmiştir [29–31]. Hiperparametre optimizasyonu bilinçli olarak yapılmamıştır. Bunun temel nedeni, modellerin sınıf dengesizliğine karşı gösterdikleri tepkiyi yalnızca yapısal farklılıkları ve öğrenme stratejileri üzerinden karşılaştırmaktır. Hiperparametre ayarları, özellikle Boosting tabanlı yöntemlerde modelin duyarlılığını önemli ölçüde etkileyebilmekte ve sonuçları değiştirebilmektedir. Ancak bu tür optimizasyonlar, deneysel kontrolü azaltmakta ve model performanslarındaki farkların yalnızca algoritma kaynaklı mı yoksa parametre ayarından mı kaynaklandığını belirsizleştirebilmektedir. Bu çalışmanın amacı, algoritmaların sınıf dengesizliği karşısındaki doğal dayanıklılık düzeylerini doğrudan gözlemlemektir. Bu nedenle tüm modeller aynı koşullarda, ön ayarlarla test edilerek parametrik etkilerden arındırılmış daha nesnel bir karşılaştırma yapılması hedeflenmiştir.

Deneyler Python programlama dili ile gerçekleştirilmiş; modelleme sürecinde Scikit-learn kütüphanesi, veri setlerinin temininde ise Scikit-learn.datasets modülü kullanılmıştır [32,33]. Veri setlerinde yalnızca etiketlerin sayısal forma dönüştürülmesi ve temel biçimsel dönüşümler yapılmış; özellik standardizasyonu gerekli görülmemiştir.

3. Araştırma Bulguları (Research Findings)

Bu bölümde, elde edilen bulgular sistematik olarak sunulmuş ve hem model hem de dengesizlik düzeyine göre karşılaştırmalı analizler yapılmıştır. Ayrıca, modellerin hangi koşullarda daha kararlı ve etkili sonuçlar verdiği detaylı grafiklerle desteklenerek ortaya konulmuştur. Bulgular; dengesizlik düzeyinin artışına karşı metrik bazlı performans değişimleri, modeller arası farklar ve veri seti kaynaklı etkiler çerçevesinde çok boyutlu biçimde tartışılmıştır.

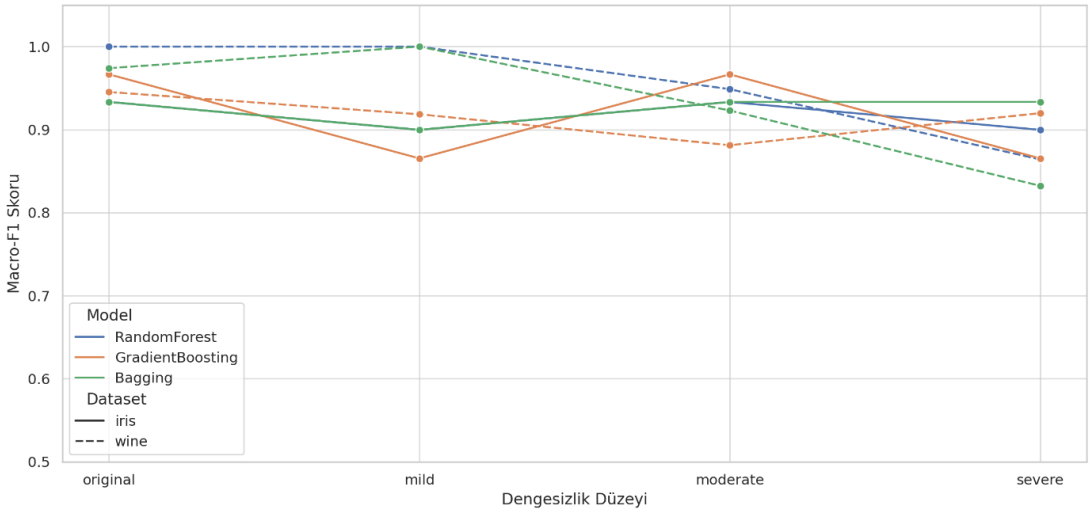
3.1. Modellerin Dengesizlik Düzeylerine Göre Genel Performans (Overall Performance of the Models Based on Imbalance Levels)

Şekil 1'de, dört farklı dengesizlik düzeyi altında her modelin Macro-F1 skorlarındaki değişim çizgisel olarak gösterilmiştir. Grafik incelendiğinde, dengesizliğin artmasıyla birlikte tüm modellerde belirgin

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr

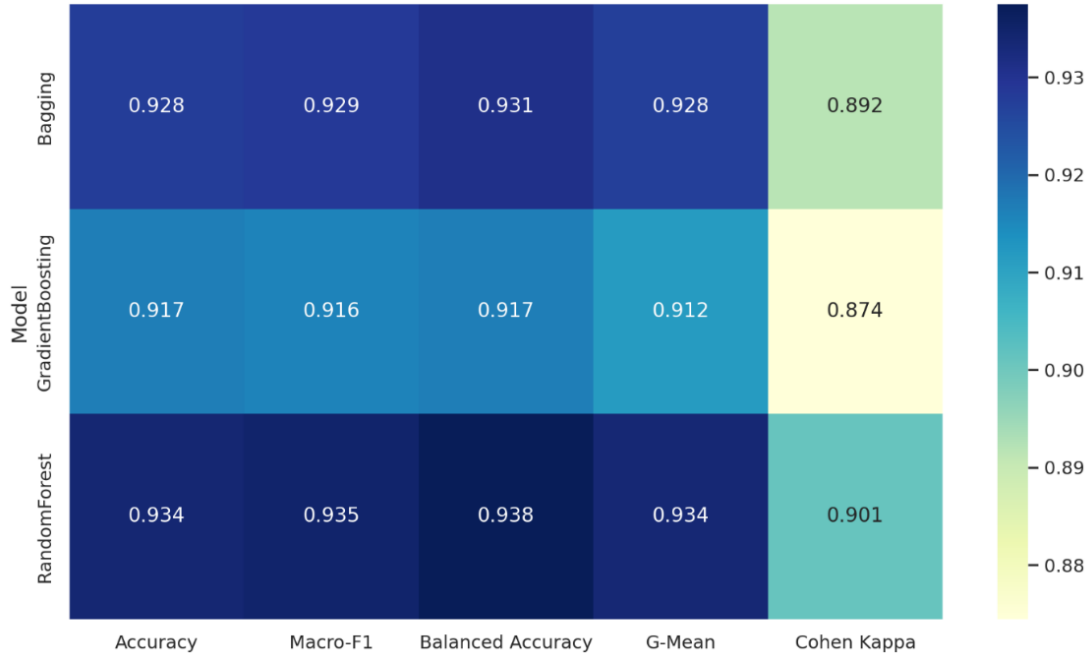
ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

bir performans düşüşü gözlemlenmektedir. Bu düşüş, özellikle Gradient Boosting modelinde daha keskin olup, azınlık sınıfların sayıca az olduğu senaryolarda bu modelin duyarlılık kaybına açık olduğunu göstermektedir. Buna karşın, Random Forest ve Bagging algoritmaları daha stabil bir düşüş sergileyerek yüksek dengesizlik oranlarında dahi daha başarılı sonuçlar vermiştir. Iris veri setinde gözlenen düşüş oranları, Wine veri setine kıyasla daha az belirgin olup, bu durum veri yapısındaki özellik sayısının sonuçlar üzerinde etkisini göstermektedir.



Şekil 1. Dengesizlik düzeyine göre Macro-F1 skorlarının değişimi

Modellerin genel başarı düzeylerini özetleyen Şekil 2 ise, beş farklı metriğe göre tüm senaryoların ortalama skorlarını sunmaktadır. Bu ısı haritası, özellikle Random Forest algoritmasının tüm metriklerde en yüksek ortalama değerleri sağladığını açıkça ortaya koymaktadır. Gradient Boosting ise, özellikle Cohen Kappa ve G-Mean gibi dengesizliğe duyarlı metriklerde daha düşük performans göstermiştir. Bagging modeli ise çoğu durumda Random Forest'a yakın sonuçlar üretmiş ancak özellikle Macro-F1 ve Balanced Accuracy gibi metriklerde zaman zaman daha değişken sonuçlar vermiştir.



Şekil 2. Her modelin genel ortalama performans metrikleri

Bu iki görselle birlikte değerlendirildiğinde, sınıf dengesizliği altında topluluk öğrenme modellerinin performans profilinin yalnızca doğruluk metriğine değil, daha duyarlı metriklere göre de farklılaştığı

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr

ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

açıkça görülmektedir. Bu da model seçiminin yalnızca genel doğruluğa değil, değerlendirme bağlamına göre yapılandırılması gerektiğini ortaya koymaktadır.

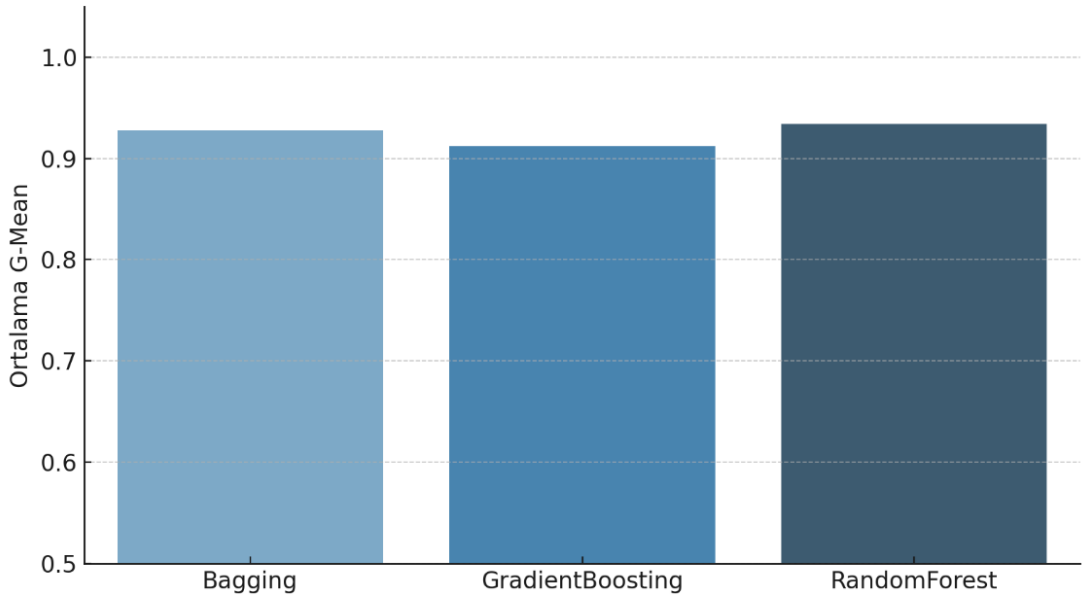
3.2.Model Bazlı Karşılaştırma Analizi (Model-Based Comparative Analysis)

Dört farklı dengesizlik düzeyi altında gerçekleştirilen değerlendirmelerde, topluluk öğrenme modelleri arasında performans açısından anlamlı farklar gözlenmiştir. Özellikle sınıf dağılımının bozulduğu senaryolarda, modellerin dengesizlik karşısındaki dirençleri ve genel kararlılık düzeyleri belirginleşmiştir.

Analiz sonuçları, Random Forest algoritmasının tüm dengesizlik senaryoları boyunca en istikrarlı performansı gösterdiğini ortaya koymaktadır. Bu model, özellikle azınlık sınıfların sayıca zayıf olduğu durumlarda dahi yüksek skorlar üretmiş ve metrikler arası dalgalanma düzeyi düşük kalmıştır. Bagging modeli de genel olarak Random Forest'a yakın sonuçlar vermiştir; ancak bazı senaryolarda (özellikle şiddetli dengesizlik altında) daha fazla performans dalgalanması gözlemlenmiştir.

Buna karşılık, Gradient Boosting algoritması, özellikle orta ve şiddetli dengesizlik senaryolarında belirgin şekilde düşük skorlar üretmiştir. Boosting yapısının, ardışıl öğrenme mekanizması nedeniyle azınlık sınıfları yeterince öğrenememesi bu düşüşün temel nedeni olabilir. Bu durum, özellikle Macro-F1, Balanced Accuracy ve G-Mean gibi dengesizliğe duyarlı metriklerde kendini açık biçimde göstermiştir.

Model karşılaştırmasını daha somut biçimde ortaya koyan Şekil 3, her modelin dört senaryo boyunca elde ettiği ortalama G-Mean skorlarını sunmaktadır. Grafik, Random Forest'ın sınıf dengesini en iyi şekilde koruyan model olduğunu açıkça göstermektedir. Gradient Boosting ise ortalama G-Mean skorlarında diğer modellere göre daha düşük seviyede kalmıştır. Bu da modelin dengesiz veri karşısında daha hassas olduğunu ve özellikle azınlık sınıflarda sınıflandırma başarısının düştüğünü göstermektedir.



Şekil 3. Modellerin Ortalama G-Mean Skorları (Tüm Senaryolar)

Sonuç olarak, sınıf dengesizliği bulunan veri setlerinde yalnızca doğruluk değil, duyarlılığı yüksek metriklere dayalı karşılaştırmalar da yapılmalı; model seçimi bu çok boyutlu değerlendirmeye göre belirlenmelidir. Bu bağlamda Random Forest, kararlılık ve dengesizlik toleransı açısından en güçlü aday olarak öne çıkmaktadır.

3.3. Veri Setine Özgü Bulgular (Findings Specific to the Dataset)

Deneysel analizlerin veri seti düzeyinde ayrıştırılması, sınıf dengesizliğinin etkisinin yalnızca model kaynaklı değil, aynı zamanda veri yapısına bağlı olarak da değiştiğini göstermektedir. Iris veri seti, başlangıçta tam dengeli bir sınıf dağılımına ve yalnızca dört özellik içeren düşük boyutlu bir yapıya

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr

ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

sahiptir. Buna karşılık, Wine veri seti, başlangıçta hafif dengesiz bir yapıya ve 13 özellikten oluşan daha karmaşık bir yapıya sahiptir. Bu yapısal farklılıklar, dengesizlik karşısındaki tepkiyi doğrudan etkilemiştir.

Özellikle Wine veri setinde, dengesizlik düzeyinin artmasıyla birlikte tüm modellerde metrik bazlı performans düşüşleri daha belirgin şekilde gözlemlenmiştir. Bu durum, hem veri setinin zaten başlangıçta tam dengeli olmaması hem de yüksek boyutluluğun modelin karar sınırlarını öğrenmesini zorlaştırmasıyla ilişkilendirilebilir. Gradient Boosting, Wine veri setinde dengesizliğe karşı en zayıf performansı sergileyen model olurken; Random Forest ve Bagging bu veri setinde daha kararlı bir başarı profili çizmiştir.

Iris veri setinde ise dengesizlik düzeylerinin etkisi daha sınırlı kalmış; özellikle Random Forest ve Bagging modelleri senaryolar arası daha az dalgalanma göstermiştir. Bu durum, veri setinin sade yapısı ve sınıflar arası ayrımın daha belirgin olması sayesinde modellerin daha güvenilir karar sınırları oluşturabilmesiyle açıklanabilir. Bununla birlikte, şiddetli dengesizlik senaryosunda Gradient Boosting modelinin performansının ciddi oranda düştüğü gözlemlenmiştir. Bu düşüş, modelin azınlık sınıf temsil sayısındaki azalmaya karşı oldukça hassas olduğunu bir kez daha ortaya koymuştur.

Genel olarak değerlendirildiğinde, veri setine özgü yapısal faktörler (özellik sayısı, sınıf ayrımı belirginliği, başlangıç dengesi) sınıf dengesizliği altında model davranışlarını anlamada kritik rol oynamaktadır. Bu nedenle model değerlendirmeleri yapılırken yalnızca metrik sonuçlara değil, veri setlerinin yapısal özelliklerine de dikkat edilmesi gerektiği ortaya konulmuştur.

4. Tartışma ve Sonuç (Results and Discussion)

Bu çalışmada, sınıf dengesizliğinin topluluk öğrenme algoritmaları üzerindeki etkisi kontrollü bir deneysel düzenekle analiz edilmiştir. Dört farklı dengesizlik düzeyinde, iki çok sınıflı veri seti ve üç yaygın topluluk modeli test edilmiş; doğruluk, duyarlılık ve kararlılığı yansıtan çok sayıda metrik üzerinden değerlendirme yapılmıştır. Elde edilen bulgular, sınıf dağılımı bozuldukça Gradient Boosting modelinin performans kaybına açık olduğunu, buna karşın Random Forest algoritmasının hem yüksek doğruluk hem de metrik bazlı istikrar sağladığını ortaya koymuştur. Bagging ise genel olarak dengeli bir performans göstermiş ancak bazı senaryolarda sınıflar arası dengesizliklere duyarlı olmuştur.

Sonuçlar, model seçiminde yalnızca doğruluk metriğine değil, Macro-F1, G-Mean ve Balanced Accuracy gibi sınıf dengesini daha iyi yansıtan ölçütlere de odaklanılması gerektiğini göstermektedir. Ayrıca, veri setinin yapısı ve sınıf ayrımının belirginliği, dengesizlik karşısında model başarısını doğrudan etkilemektedir. Bu bağlamda, dengesiz veri setleriyle çalışırken hem algoritma hem de metrik seçiminde çok boyutlu ve veriye özgü bir yaklaşım benimsenmesi önerilmektedir.

Gelecek çalışmalarda, daha büyük ve daha karmaşık yapıya sahip çok sınıflı veri setleriyle benzer karşılaştırmaların yapılması önerilmektedir. Ayrıca, örnekleme tabanlı dengeleme stratejilerinin (örneğin SMOTE, Borderline-SMOTE) entegre edildiği senaryolarda modellerin dayanıklılığı daha ayrıntılı şekilde analiz edilebilir. Bunun yanında, hiperparametre optimizasyonunun etkisini izole eden ek deneyler, modeller arası farkların daha adil bir zeminde değerlendirilmesini sağlayabilir. Farklı topluluk algoritmalarının (örneğin XGBoost, LightGBM, AdaBoost) değerlendirme kapsamına dahil edilmesi de çalışmanın kapsamını genişletmeye yönelik önemli bir adım olacaktır.

5. Kaynaklar (References)

[1] Ö. ÇELİK, A Research on Machine Learning Methods and Its Applications, Journal of Educational Technology and Online Learning 1 (2018) 25–40. https://doi.org/10.31681/jetol.457046.

[2] S.B. Kotsiantis, I.D. Zaharakis, P.E. Pintelas, Machine learning: a review of classification and combining techniques, Artif Intell Rev 26 (2006) 159–190. https://doi.org/10.1007/s10462-007-9052-3.

[3] S. Yadav, G.P. Bhole, Handling Imbalanced Dataset Classification in Machine Learning, in: 2020 IEEE Pune Section International Conference (PuneCon), IEEE, 2020: pp. 38–43. https://doi.org/10.1109/PuneCon50868.2020.9362471.

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr

ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

[4] H. Kaur, H.S. Pannu, A.K. Malhi, A Systematic Review on Imbalanced Data Challenges in Machine Learning, ACM Comput Surv 52 (2020) 1–36. https://doi.org/10.1145/3343440.

[5] Z.-H. Zhou, Ensemble Learning, in: Mach Learn, Springer Singapore, Singapore, 2021: pp. 181–210. https://doi.org/10.1007/978-981-15-1967-3_8.

[6] J. Beemer, K. Spoon, L. He, J. Fan, R.A. Levine, Ensemble Learning for Estimating Individualized Treatment Effects in Student Success Studies, Int J Artif Intell Educ 28 (2018) 315–335. https://doi.org/10.1007/s40593-017-0148-x.

[7] A. Mohammed, R. Kora, A comprehensive review on ensemble deep learning: Opportunities and challenges, Journal of King Saud University - Computer and Information Sciences 35 (2023) 757–774. https://doi.org/10.1016/j.jksuci.2023.01.014.

[8] O. Saidani, M. Umer, A. Alshardan, N. Alturki, M. Nappi, I. Ashraf, Student academic success prediction in multimedia-supported virtual learning system using ensemble learning approach, Multimed Tools Appl 83 (2024) 87553–87578. https://doi.org/10.1007/s11042-024-18669-z.

[9] M.R. Khalilpour Darzi, S.T.A. Niaki, M. Khedmati, Binary classification of imbalanced datasets: The case of CoIL challenge 2000, Expert Syst Appl 128 (2019) 169–186. https://doi.org/10.1016/j.eswa.2019.03.024.

[10] W. Lee, K. Seo, Downsampling for Binary Classification with a Highly Imbalanced Dataset Using Active Learning, Big Data Research 28 (2022) 100314. https://doi.org/10.1016/j.bdr.2022.100314.

[11] V. Kumar, G.S. Lalotra, P. Sasikala, D.S. Rajput, R. Kaluri, K. Lakshmanna, M. Shorfuzzaman, A. Alsufyani, M. Uddin, Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques, Healthcare 10 (2022) 1293. https://doi.org/10.3390/healthcare10071293.

[12] S. Cateni, V. Colla, M. Vannucci, A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing 135 (2014) 32–41. https://doi.org/10.1016/j.neucom.2013.05.059.

[13] P. Manikanta, D. Jayaprakash, D. Sharma, R. Bathla, Wine Quality Prediction Using Machine Learning Techniques, in: 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), IEEE, 2024: pp. 1015–1018. https://doi.org/10.1109/ICAICCIT64383.2024.10912091.

[14] M. De Marsico, M. Nappi, D. Riccio, H. Wechsler, Mobile Iris Challenge Evaluation (MICHE)-I, biometric iris dataset and protocols, Pattern Recognit Lett 57 (2015) 17–23. https://doi.org/10.1016/j.patrec.2015.02.009.

[15] G. Biau, E. Scornet, A random forest guided tour, TEST 25 (2016) 197–227. https://doi.org/10.1007/s11749-016-0481-7.

[16] A. Parmar, R. Katariya, V. Patel, A Review on Random Forest: An Ensemble Classifier, in: 2019: pp. 758–763. https://doi.org/10.1007/978-3-030-03146-6_86.

[17] Y. Zhang, A. Haghani, A gradient boosting method to improve travel time prediction, Transp Res Part C Emerg Technol 58 (2015) 308–324. https://doi.org/10.1016/j.trc.2015.02.019.

[18] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Artif Intell Rev 54 (2021) 1937–1967. https://doi.org/10.1007/s10462-020-09896-5.

[19] J.G. Dias, J.K. Vermunt, A bootstrap-based aggregate classifier for model-based clustering, Comput Stat 23 (2008) 643–659. https://doi.org/10.1007/s00180-007-0103-7.

[20] T. Hothorn, B. Lausen, Double-bagging: combining classifiers by bootstrap aggregation, Pattern Recognit 36 (2003) 1303–1309. https://doi.org/10.1016/S0031-3203(02)00169-3.

[21] C.W. Fisher, E.J.M. Lauria, C.C. Matheus, An Accuracy Metric, Journal of Data and Information Quality 1 (2009) 1–21. https://doi.org/10.1145/1659225.1659229.

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr

ISSN: 2618-6055 / 1, (9), 41 – 50 , 2025
DOI: 10.62301/usmtd.1701938

[22] M.C. Hinojosa Lee, J. Braet, J. Springael, Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores, Applied Sciences 14 (2024) 9863. https://doi.org/10.3390/app14219863.

[23] V. García, R.A. Mollineda, J.S. Sánchez, Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions, in: 2009: pp. 441–448. https://doi.org/10.1007/978-3-642-02172-5_57.

[24] H.R. Sofaer, J.A. Hoeting, C.S. Jarnevich, The area under the precision-recall curve as a performance metric for rare binary events, Methods Ecol Evol 10 (2019) 565–577. https://doi.org/10.1111/2041-210X.13140.

[25] H. Guo, H. Liu, C. Wu, W. Zhi, Y. Xiao, W. She, Logistic discrimination based on G-mean and F-measure for imbalanced problem, Journal of Intelligent & Fuzzy Systems 31 (2016) 1155–1166. https://doi.org/10.3233/IFS-162150.

[26] A. Figueroa, S. Ghosh, C. Aragon, Generalized Cohen's Kappa: A Novel Inter-rater Reliability Metric for Non-mutually Exclusive Categories, in: 2023: pp. 19–34. https://doi.org/10.1007/978-3-031-35132-7_2.

[27] A. Aggarwal, Z. Xu, O. Feyisetan, N. Teissier, On Log-Loss Scores and (No) Privacy, in: Proceedings of the Second Workshop on Privacy in NLP, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020: pp. 1–6. https://doi.org/10.18653/v1/2020.privatenlp-1.1.

[28] R. Singh, N.S. Mangat, Stratified Sampling, in: 1996: pp. 102–144. https://doi.org/10.1007/978-94-017-1404-4_5.

[29] M. Moeini, Hyperparameter tuning of supervised bagging ensemble machine learning model using Bayesian optimization for estimating stormwater quality, Sustain Water Resour Manag 10 (2024) 83. https://doi.org/10.1007/s40899-024-01064-9.

[30] R.I. Alkanhel, E.-S.M. El-Kenawy, M.M. Eid, L. Abualigah, M.A. Saeed, Optimizing IoT-driven smart grid stability prediction with dipper throated optimization algorithm for gradient boosting hyperparameters, Energy Reports 12 (2024) 305–320. https://doi.org/10.1016/j.egyr.2024.06.034.

[31] Y. Rimal, N. Sharma, A. Alsadoon, The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms, Multimed Tools Appl 83 (2024) 74349–74364. https://doi.org/10.1007/s11042-024-18426-2.

[32] O. Kramer, Scikit-Learn, in: 2016: pp. 45–53. https://doi.org/10.1007/978-3-319-33383-0_5.

[33] H. Bin Mehare, J.P. Anilkumar, N.A. Usmani, The Python Programming Language, in: A Guide to Applied Machine Learning for Biologists, Springer International Publishing, Cham, 2023: pp. 27–60. https://doi.org/10.1007/978-3-031-22206-1_2.

*Sorumlu Yazar/Corresponding Author: abdulhamidkarabiyik@ohu.edu.tr