T.C.

AYDIN ADNAN MENDERES ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK (TIP)

DOKTORA PROGRAMI

DENGESİZ VERİ SETLERİNDE FARKLI DENGELEME ALGORİTMALARININ OPTİMUM DENGE ORANLARININ SINIFLANDIRMA VE REGRESYON AĞAÇLARI YÖNTEMİ İLE İNCELENMESİ: SİMÜLASYON ÇALIŞMASI

HAKAN ÖZTÜRK

DOKTORA TEZİ

DANIŞMAN

Prof. Dr. Mevlüt TÜRE

AYDIN–2022

KABUL VE ONAY

T.C. Aydın Adnan Menderes Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı (Tıp) Doktora Programı çerçevesinde Hakan ÖZTÜRK tarafından hazırlanan "Dengesiz Veri Setlerinde Farklı Dengeleme Algoritmalarının Optimum Denge Oranlarının Sınıflandırma ve Regresyon Ağaçları Yöntemi ile İncelenmesi: Simülasyon Çalışması" başlıklı tez, aşağıdaki jüri tarafından Doktora Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 09/02/2022

Üye (T.D.)	: Prof. Dr. Mevlüt TÜRE	ADÜ	… (imza) …
Üye	: Prof. Dr. Kevser Setenay DİNÇER ÖNER	ESKİŞEHİR OSMANGAZİ ÜNİVERSİTESİ	… (imza) …
Üye	: Prof. Dr. Gökay BOZKURT	ADÜ	… (imza) …
Üye	: Prof. Dr. Canan BAYDEMİR	KOCAELİ ÜNİVERSİTESİ	… (imza) …
Üye	: Prof. Dr. İmran KURT ÖMÜRLÜ	ADÜ	… (imza) …

ONAY:

Bu tez Aydın Adnan Menderes Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun görülmüş ve Sağlık Bilimleri Enstitüsünün ……..……..… tarih ve …… sayılı oturumunda alınan …… nolu Yönetim Kurulu kararıyla kabul edilmiştir.

Prof. Dr. Süleyman AYPAK

Enstitü Müdürü V.

TEŞEKKÜR

Doktora öğrenimim süresince desteğini hiçbir zaman esirgemeyen, bilgi ve birikimiyle tez çalışmamın tüm aşamalarında yanımda olan ve bana yol gösteren tez danışmanım Prof. Dr. Mevlüt TÜRE'ye,

Tez çalışmam süresince desteğini esirgemeyen, bilgi ve birikimiyle tezime ve mesleki gelişimime katkıda bulunan Prof. Dr. İmran KURT ÖMÜRLÜ'ye,

Tez çalışmam süresince stresli sürecime katlanan, destek olan çalışma arkadaşım Arş. Gör. Fulden CANTAŞ TÜRKİŞ'e,

Desteklerini hiçbir zaman esirgemeyen varlıklarıyla bana güç veren annem Gülhan ÖZTÜRK ve babam Mustafa ÖZTÜRK'e teşekkür ederim.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR DİZİNİ

$\boldsymbol{x_i}$: Azınlık sınıfı birimi

$\boldsymbol{x_i'}$: Çoğunluk sınıfı birimi

$\boldsymbol{x_{ik}}$: x_i'nin azınlık sınıfı içerisindeki k en yakın komşusu

$\boldsymbol{x_{s_i}}$: Sentetik gözlem

$\boldsymbol{D}$: Eğitim veri seti

$\boldsymbol{Y}$: İki kategorili bağımlı değişken

$\boldsymbol{y_i}$: Bağımlı değişkenin kategorileri

$\boldsymbol{n_{az}}$: Azınlık sınıfı birim sayısı

$\boldsymbol{n_{\text{çoğ}}}$: Çoğunluk sınıfı birim sayısı

$\boldsymbol{d}$: Sınıf dengesizliğinin derecesi

$\boldsymbol{\beta}$: Yeniden örnekleme oranı

Δ_i : x_i'nin k en yakın komşuları içinde çoğunluk sınıfına ait olan birimlerin sayısı

$\boldsymbol{k_1}$: Gürültülü azınlık sınıfı birimlerini tahmin etmek için kullanılan komşu sayısı

$\boldsymbol{k_2}$: Bilgilendirici azınlık sınıfını oluşturmak için kullanılan çoğunluk sınıfı komşu sayısı

$\boldsymbol{k_3}$: Bilgilendirici azınlık sınıfını oluşturmak için kullanılan azınlık sınıfı komşu sayısı

$\boldsymbol{NN(x_i)}$: x_i'nin en yakın k_1 komşularının kümesi

$\boldsymbol{N_{\text{çoğ}}(x_i)}$: x_i'nin en yakın k_2 çoğunluk sınıfı komşularının kümesi

$\boldsymbol{N_{\text{çoğ}}(x_i)}$: x_i''nin en yakın k_3 azınlık sınıfı komşularının kümesi

$\boldsymbol{S_{az}}$: Azınlık sınıfı kümesi

$\boldsymbol{S_{\text{çoğ}}}$: Çoğunluk sınıfı kümesi

$\boldsymbol{S_{azf}}$: Filtrelenmiş azınlık kümesi

$\boldsymbol{S_{iaz}}$: Bilgilendirici azınlık kümesi

$\boldsymbol{S_{b\text{çoğ}}}$: Sınır çoğunluk kümesi

$\boldsymbol{S_{oaz}}$:Sentetik birimlerin kümesi

$\boldsymbol{z_i}$: Bilgilendirici azınlık kümesi birimi

$\boldsymbol{G}$: Üretilecek sentetik gözlemlerin sayısı

$\boldsymbol{h_t}$: Sınıflandırma modeli

$\boldsymbol{T}$: Tekrar sayısı

$\boldsymbol{t}$: 1 ile T arasındaki tekrar adımı

$\boldsymbol{w_t(i)}$: i'nci birimin t'inci tekrardaki ağırlığı

$\boldsymbol{hata_t}$: t'inci tekrardaki hata

$\boldsymbol{\alpha_t}$: Ağırlık güncelleme parametresi

$\boldsymbol{Z_t}$: $w_t(i)$ ağırlıklarının toplamı

X : Bağımsız değişkenler matrisi

AdaBoost : Adaptif Boosting

ADASYN : Adaptif Sentetik Örnekleme Yaklaşımı

AUC : Eğri Altında Kalan Alan

CART : Sınıflandırma ve Regresyon Ağaçları

MWMOTE : Çoğunluk Ağırlıklı Azınlık Aşırı Örnekleme Tekniği

SMOTE : Sentetik Azınlık Aşırı Örnekleme Tekniği

SMOTEBoost : Sentetik Azınlık Aşırı Örnekleme Tekniği Boosting

ROC : Alıcı işlem karakteristiği

ROS : Rastgele Aşırı Örnekleme

RUS : Rastgele Alt Örnekleme

RUSBoost : Rastgele Alt Boosting

UB : Alt Bagging

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

ÖZET

DENGESİZ VERİ SETLERİNDE FARKLI DENGELEME ALGORİTMALARININ OPTİMUM DENGE ORANLARININ SINIFLANDIRMA VE REGRESYON AĞAÇLARI YÖNTEMİ İLE İNCELENMESİ: SİMÜLASYON ÇALIŞMASI

Öztürk H. Aydın Adnan Menderes Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Programı, Doktora Tezi, Aydın, 2022.

Amaç: Bu çalışmada, orijinal simülasyon senaryoları ışığında, farklı korelasyon yapıları, değişken sayıları ve azınlık sınıfı prevalans oranları altında yedi farklı dengeleme algoritması için optimal azınlık-çoğunluk sınıfı dengeleme oranlarının sınıflandırma ve regresyon ağaçları (CART) ile incelenmesi amaçlandı.

Gereç ve Yöntem: Azınlık sınıfı prevalans oranları, korelasyon yapıları ve değişken sayıları dikkate alınarak toplum veri setlerinden örneklenen dengesiz veri setleri, rastgele aşırı örnekleme (ROS), sentetik azınlık aşırı örnekleme tekniği (SMOTE), çoğunluk ağırlıklı azınlık aşırı örnekleme tekniği (MWMOTE), adaptif sentetik örnekleme yaklaşımı (ADASYN), rastgele alt örnekleme (RUS), rastgele alt boosting (RUSBoost) ve alt bagging (UB) algoritmaları ile kademeli olarak dengelendi ve her kademede CART yöntemi ile toplum parametreleri tahmin edildi.

Bulgular: Tüm simülasyon senaryolarında, dengeleme algoritmalarının, sınıflandırma başarısını artırdığı gözlendi. Bu artışın, dengeleme oranının artmasıyla paralel olduğu ve tüm dengeleme algoritmalarının en yüksek alıcı işlem karakteristiği (ROC) eğrisi altında kalan alan (AUC) değerine genellikle tam denge (50:50) durumunda ulaştığı gözlendi. Ayrıca, yapılan sınıflandırmalarda, en yüksek AUC değerleri, RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde elde edildi. Türetilen toplum veri setlerinden hesaplanan AUC değerleri referans alınarak değerlendirilen optimal azınlık-çoğunluk sınıfı denge oranları, kullanılan dengeleme algoritmalarına bağlı olarak farklılık gösterdi. Bununla birlikte, değişkenler arasındaki korelasyon yapısı, bağımsız değişken sayısı ve azınlık sınıfı prevalans oranları da dengeleme algoritmaları için

optimal azınlık-çoğunluk sınıfı denge oranlarını etkiledi. Değişkenler arasındaki ilişki düzeyinin ve bağımsız değişken sayısının artışına paralel olarak dengeleme algoritmaları ile dengelenen veri setlerinin sınıflandırılmasından elde edilen AUC değerlerinin toplum veri setlerinden elde edilen AUC değerlerine yakınsama oranı arttı.

Sonuç: Sonuç olarak, RUSBoost ve UB algoritmalarının simülasyon senaryolarının çoğunda belirli denge oranlarından sonra parametre değerinden istatistiksel olarak yüksek sonuçlar ürettiği gözlendi. Hem ilişki düzeyindeki hem de bağımsız değişken sayısındaki artış RUSBoost ve UB algoritmalarının parametre değerinden yüksek sonuçlar üretme eğilimini artırdı. ROS, SMOTE, MWMOTE, ADASYN algoritmalarının, simülasyon senaryolarının çoğunda, RUS algoritmasının ise simülasyon senaryolarının hiçbirinde parametre değerinden istatistiksel olarak yüksek sonuçlar üretmediği gözlendi.

Anahtar kelimeler: Dengesiz veri, Topluluk öğrenme, Sınıflandırma ve regresyon ağaçları, Optimal sınıf dağılımı, Simülasyon

ABSTRACT

EXAMINING THE OPTIMUM BALANCE RATIOS OF DIFFERENT BALANCING ALGORITHMS IN IMBALANCED DATA SETS BY CLASSIFICATION AND REGRESSION TREES: SIMULATION STUDY

Öztürk H. Aydın Adnan Menderes University, Health Sciences Institute, Biostatistics Program, Doctorate Thesis, Aydın, 2022.

Objective: In this study, it was aimed to examine the optimal minority-majority class balancing ratios for seven different balancing algorithms by classification and regression trees (CART) under different correlation structures, variable numbers, and minority class prevalence rates in the light of original simulation scenarios.

Material and Methods: Imbalanced datasets were sampled from population datasets were derived by considering minority class prevalence rates, correlation structures, and variable numbers. Imbalanced datasets were gradually balanced with random oversampling (ROS), synthetic minority over-sampling technique (SMOTE), majority weighted minority oversampling technique (MWMOTE), adaptive synthetic sampling approach (ADASYN), random undersampling (RUS), random under boosting (RUSBoost), and under bagging (UB) algorithms and classified by CART method at each step.

Results: In all simulation scenarios, classification performance gradually increased in data sets that were gradually balanced with balancing algorithms. This increase is in parallel with the increase in the balancing ratio, and all balancing algorithms reached the highest area under the receiver operation characteristic (ROC) curve (AUC) value generally at fully balanced (50:50). In addition, the highest AUC values were obtained in the datasets balanced with the RUSBoost and UB algorithms. Optimal minority-majority class balance ratios, evaluated regarding the AUC values calculated from the derived population datasets, differed depending on the balancing algorithms used. However, the correlation structure between the variables, the number of

independent variables, and the minority class prevalence rates also affected the optimal minority-majority class balance ratios for the balancing algorithms. In parallel with the increase in the level of the relationship between the variables and the number of independent variables, the rate of convergence of the AUC values obtained from the classification of the data sets balanced with the balancing algorithms to the AUC values obtained from the population datasets increased.

Conclusion: In conclusion, statistically higher results than the population parameters were obtained when certain balancing ratios were exceeded in the datasets balanced with the RUBoost and UB algorithms in most of the simulation scenarios. The increase in both the level of correlation and the number of independent variables increased the tendency of RUSBoost and UB algorithms to produce results higher than the population parameters. ROS, SMOTE, MWMOTE, ADASYN algorithms produced statistically higher results than population parameters only for some scenarios with four and five independent variables where the correlation between variables was high. In none of the simulation scenarios, the RUS algorithm did not produce statistically higher results than the population parameters.

Keywords: Imbalanced data, Ensemble learning, Classification and regression trees, Optimal class distribution, Simulation

1. GİRİŞ

Sağlık alanı başta olmak üzere birçok alandaki gerçek veri setleri sınıf değişkeni bakımından genellikle dengesiz yapıdadır. Bir veri setinde farklı sınıfların birim sayılarının orantısız bir şekilde dağılması durumunda sınıf dengesizliği problemi ortaya çıkar. İkili sınıflandırmada, sınıflardan biri (azınlık sınıfı) diğerine (çoğunluk sınıfı) kıyasla yetersiz temsil ediliyorsa, bu veri setinin dengesiz olduğu söylenir. Dengesiz veri setleri ile oluşturulan sınıflandırma modelleri, genellikle çoğunluk sınıfının etkisi altında kalarak azınlık sınıfı gözlemlerini yanlış sınıflandırma eğiliminde olurlar. Özellikle sağlık alanında hastalık teşhisi, prognoz belirleme, hastalık izlemi veya sağlık hizmetlerinin kalite kontrolünü belirlemeye yönelik yapılan çalışmalarda sınıf dengesizliği problemi ile sıklıkla karşılaşılmaktadır (Acharya ve diğerleri, 2016; Bach ve diğerleri, 2017; Krawczyk ve diğerleri, 2015; Ren ve diğerleri, 2017; Saxena ve diğerleri, 2021; Wang ve diğerleri, 2020; Zięba ve diğerleri, 2014). Gerçek veri uygulamalarında veri setlerinin birçoğunun sınıf değişkeni bakımından dengesiz bir yapıda olması nedeniyle bu kadar basit bir tanıma sahip olan sınıf dengesizliği problemi, araştırmacıların ilgisini çekmektedir. Bu problemin üstesinden gelebilmek için etkinliği ispatlanmış, uygulanması basit ve sınıflandırıcıdan bağımsız çok sayıda *veri dengeleme algoritması* önerilmiştir (Barua ve diğerleri, 2012; Chawla ve diğerleri, 2002; Fernández ve diğerleri, 2018; He ve diğerleri, 2008; Santos ve diğerleri, 2018; Seiffert ve diğerleri, 2009; Sun ve diğerleri, 2015).

Veri dengeleme algoritmalarının amacı, model eğitim sürecinde eğitim veri setini yeniden örnekleyerek çarpık sınıf dağılımının etkisini azaltmaktır. Veri dengeleme algoritmalarında, aşırı örnekleme ve alt örnekleme olmak üzere iki temel yaklaşım söz konusudur. Aşırı örnekleme, azınlık sınıfının gözlemlerini artırma, alt örnekleme ise çoğunluk sınıfının gözlemlerini azaltma sürecini ifade eder. Literatürde çok sayıda aşırı örnekleme ve alt örnekleme algoritmaları ile bu algoritmaları bagging ve boosting gibi topluluk öğrenme teknikleriyle birleştiren çeşitli algoritmalar önerilmiştir (Barua ve diğerleri, 2012; Chawla ve diğerleri, 2002; He ve diğerleri, 2008; Seiffert ve diğerleri, 2009; Sun ve diğerleri, 2015). Yapılan birçok çalışmada, sınıf dengesizliği problemine çözüm olarak önerilen dengeleme algoritmalarının sınıflandırma başarısını kayda değer ölçüde iyileştirdiği ifade edilmiştir (Batista ve diğerleri, 2004; Galar ve

diğerleri, 2011; Hasanin ve Khoshgoftaar, 2018; Japkowicz, 2000; Kamei ve diğerleri, 2007; López ve diğerleri, 2013; Tyagi ve Mittal, 2020; Van Hulse ve diğerleri, 2007).

Çeşitli alanlarda gerçek veri setleri kullanılarak veri dengeleme algoritmalarının karşılaştırıldığı bazı çalışmalar yapılmıştır. Bu çalışmalarda dengeleme algoritmalarının performansları çeşitli sınıflandırma yöntemleri ve performans ölçütleri kullanılarak değerlendirilmiştir. Kamei ve diğerleri (2007) rastgele aşırı örnekleme (ROS), sentetik azınlık aşırı örnekleme tekniği (SMOTE) ve rastgele alt örnekleme (RUS) algoritmalarının performanslarını, iki gerçek veri seti kullanarak farklı sınıflandırma yöntemleri ile değerlendirmiş ve dengeleme algoritmalarının performanslarının hemen hemen eşit olduğunu ifade etmişlerdir. He ve diğerleri (2008) adaptif sentetik örnekleme yaklaşımı (ADASYN) ve SMOTE algoritmalarının performanslarını, beş gerçek veri seti kullanarak karar ağaçları yöntemi ile karşılaştırmış ve ADASYN algoritmasının en yüksek performansı gösterdiğini bildirmişlerdir. Bennin ve diğerleri (2016) RUS, ROS ve SMOTE algoritmalarının performanslarını 10 gerçek veri seti üzerinde 10 farklı sınıflandırma modeli kullanarak karşılaştırmış ve ROS algoritmasının en iyi performansı gösterdiğini ifade etmişlerdir. Rashu ve diğerleri (2014) RUS, ROS ve SMOTE algoritmalarının performanslarını gerçek bir veri seti üzerinde 3 farklı sınıflandırma modeli kullanarak karşılaştırmış ve SMOTE algoritmasının en yüksek performansı gösterdiğini bildirmişlerdir. Amin ve diğerleri (2016) çoğunluk ağırlıklı azınlık aşırı örnekleme tekniği (MWMOTE), SMOTE ve ADASYN dahil 6 farklı veri dengeleme algoritmasının performanslarını 4 gerçek veri seti üzerinde karşılaştırmış, SMOTE ve ADASYN'nin, MWMOTE algoritmasından daha iyi performans gösterdiklerini ifade etmişlerdir. Zhong ve diğerleri (2009) RUS, ROS ve SMOTE algoritmalarının performanslarını 15 gerçek veri seti üzerinde iki farklı sınıflandırıcı kullanarak karşılaştırmış ve ROS algoritmasının en iyi performansı gösterdiğini bildirmişlerdir. Chen ve diğerleri (2010) SMOTE ve ADASYN dahil 7 farklı dengeleme algoritmasını, 19 gerçek veri seti kullanarak çok katmanlı algılayıcı ile karşılaştırmış, SMOTE ve ADASYN algoritmalarının benzer performans gösterdiğini bildirmişlerdir. Bastia ve diğerleri (2004) ROS, RUS ve SMOTE dahil 10 farklı dengeleme algoritmasını, 13 farklı gerçek veri seti kullanarak karar ağaçları yöntemi ile karşılaştırmış ve ROS algoritmasının daha karmaşık algoritmalarla rekabet edebilecek sonuçlar sağladığını ifade etmişlerdir. Van Hulse ve diğerleri (2007) 35 gerçek veri seti ve 5 farklı sınıflandırma yöntemi kullanarak yedi farklı dengeleme algoritmasının performanslarını incelemiş ve hem RUS hem de SMOTE'un çok etkili dengeleme algoritmaları olduğunu bildirmişlerdir.

Seiffert ve diğerleri (2009) 15 gerçek veri seti üzerinde rastgele alt boosting (random under boosting, RUSBoost), sentetik azınlık aşırı örnekleme tekniği boosting (SMOTEBoost), adaptif boosting (AdaBoost), RUS ve SMOTE algoritmalarının performanslarını karşılaştırmış, RUSBoost yönteminin diğer yöntemlerden daha iyi performans gösterdiğini ifade etmişlerdir. Sun ve diğerleri (2015) RUS, ROS, SMOTE, Alt Bagging (under bagging, UB) ve RUSBoost dahil 13 farklı dengeleme algoritmasını 46 gerçek veri seti üzerinde 6 farklı sınıflandırıcı kullanarak karşılaştırmıştır. RUSBoost algoritmasının, oldukça hızlı, basit ve etkili olduğu ifade edilmiştir. Barua ve diğerleri (2012) MWMOTE, SMOTE, ADASYN dahil 4 farklı dengeleme algoritmasını, 20 gerçek ve 4 türetilmiş veri seti üzerinde farklı sınıflandırma yöntemleri ile karşılaştırmış ve gerçek veri setlerinin çoğunda, türetilmiş veri setlerinin ise tamamında sınıflandırma yöntemlerinden bağımsız olarak MWMOTE algoritmasının en iyi performansı gösterdiğini bildirmişlerdir. Bu çalışmaların çoğunda dengeleme algoritmalarının performansları ya gerçek veri setleri ya da gerçek veri setlerinden örneklenen dengesiz veri setleri kullanılarak değerlendirilmiştir. Sınıflandırma sonucu kullanılan performans ölçütüne göre en yüksek değere ulaşan dengeleme algoritmalarının en başarılı algoritmalar olduğu ifade edilmiştir. Bununla birlikte, yapılan çalışmalarda, kullanılan veri setlerine ve uygulanan prosedürlere göre en iyi performansı gösteren dengeleme algoritmaları farklılık göstermektedir.

Sınıflandırma analizlerinde bir diğer problem ise sınıf değişkenin alt kategorilerindeki birim sayıları bakımından kabul edilmiş optimal bir dağılımın olmamasıdır. Buna rağmen sınıf değişkeninin alt kategorilerindeki birim sayılarının eşit olduğu (dengeli (50:50)) bir dağılımın optimale yakın olduğu düşünülür (Weiss ve Provost, 2003). Ancak yapılan bazı çalışmalarda, sınıf değişkeni bakımından dengeli bir dağılımın optimal sınıf dağılımı olmadığı ifade edilmiştir. Khoshgoftaar ve diğerleri (2007) sınıflandırma performansının en yüksek olduğu optimal sınıf dağılımını belirlemek için farklı alanlardan toplam 10 gerçek veri seti ve 11 farklı sınıflandırıcı kullanarak yaptıkları uygulamalar sonucu optimal azınlık-çoğunluk sınıf dağılımının yaklaşık olarak 35:65 olduğunu belirtmişlerdir. Bir başka çalışmasında, Khoshgoftaar ve diğerleri (2010) dört gerçek veri seti ve dört sınıflandırıcı kullanarak, boosting ve bagging tabanlı dört dengeleme algoritmasının performanslarını karşılaştırmıştır. Dengeleme algoritmaları ile azınlık-çoğunluk sınıf dağılımını 35:65 ve 50:50 olacak şekilde dengelemiş ve dengelenen veri setlerini sınıflandırmışlardır. Sonuç olarak, azınlık-çoğunluk sınıf dağılımının 35:65 olacak şekilde dengelenmesi durumda çoğunlukla daha yüksek sınıflandırma performansı elde ettiklerini

bildirmişlerdir. Weiss ve Provost (2003) 26 gerçek veri seti üzerinde yaptıkları sınıflandırmalarda, performans ölçütü olarak doğruluk oranı kullanıldığında, doğal sınıf dağılımının; alıcı işlem karakteristiği (Receiver Operating Characteristic, ROC) eğrisi altında kalan alan (AUC) kullanıldığında ise tam dengeli (50:50) bir dağılımın en iyi performansı gösterme eğiliminde olduğunu ifade etmişlerdir. Ayrıca, optimal sınıf dağılımın bir veri setinden diğerine farklılık gösterdiğini söylemişlerdir. Albisua ve diğerleri (2013) ROS, SMOTE ve RUS dahil sekiz dengeleme algoritması için optimal dengeleme oranını belirlemeye çalışmıştır. Çalışmalarında, 29 gerçek veri seti ile iki farklı sınıflandırıcı kullanmış ve en yüksek AUC değerinin elde edildiği denge oranını optimal olarak değerlendirmişlerdir. Sonuç olarak, her veri setinin kendine ait bir optimal sınıf dağılımı olduğunu ve kullanılan dengeleme algoritmasına bağlı olarak optimal sınıf dağılımının genellikle dengeli bir sınıf dağılımı olmadığını ifade etmişlerdir. Literatürde yer alan bu sınırlı sayıdaki çalışmada, optimal azınlık-çoğunluk sınıfı denge oranı, gerçek veri setleri ya da RUS algoritması ile gerçek veri setlerinden örneklenen veri setleri üzerinde incelenmiştir. Birkaç çalışmada ise bazı dengeleme algoritmaları için optimal sınıf dağılımı sadece gerçek veri setleri üzerinde yapılan uygulamalarla incelenmiştir.

1.1. Tezin Amacı

Literatür taraması sonucunda dengeleme algoritmaları için optimal dengeleme oranlarının incelendiği bir simülasyon çalışmasına rastlanmamıştır. Literatürdeki bu eksikliği gidermek için bu çalışmada, orijinal simülasyon senaryoları ışığında, farklı korelasyon yapıları, değişken sayıları ve azınlık sınıfı prevalans oranları altında yedi farklı dengeleme algoritması (ROS, SMOTE, MWMOTE, ADASYN, RUS, RUSBoost ve UB) için optimal azınlık-çoğunluk sınıfı dengeleme oranlarının sınıflandırma ve regresyon ağaçları (CART) ile incelenmesi amaçlandı. Bu amaç doğrultusunda çalışmanın hipotezleri aşağıda verildi:

- Dengeleme algoritmalarının optimal azınlık-çoğunluk sınıfı denge oranları farklıdır.
- Dengeleme algoritmaları ile dengelenen veri setlerinin sınıflandırılmasından elde edilen AUC değerleri, toplum veri setlerinden hesaplanan AUC parametre değerlerinden anlamlı düzeyde yüksek olabilir.

2. GENEL BİLGİLER

2.1. Sınıf Dengesizliği Problemi

Bağımlı değişkenin kategorik olduğu veri setlerinde, bir sınıftaki birimlerin sayısının diğer sınıfın birim sayısından önemli ölçüde az olması sınıf dengesizliği problemini ortaya çıkarır. Bu durum, azınlık sınıfının veri setinde yetersiz temsil edilmesine ve eğitilen sınıflandırma modellerinin, genellikle çoğunluk sınıfının lehine yanlı tahminler üretmesine neden olur. Sınıf dengesizliği problemine çözüm olarak etkinliği ispatlanmış, uygulanması basit ve sınıflandırıcıdan bağımsız çok sayıda veri dengeleme algoritması geliştirilmiştir (Barua ve diğerleri, 2012; Chawla ve diğerleri, 2002; He ve diğerleri, 2008; Seiffert ve diğerleri, 2009; Sun ve diğerleri, 2015).

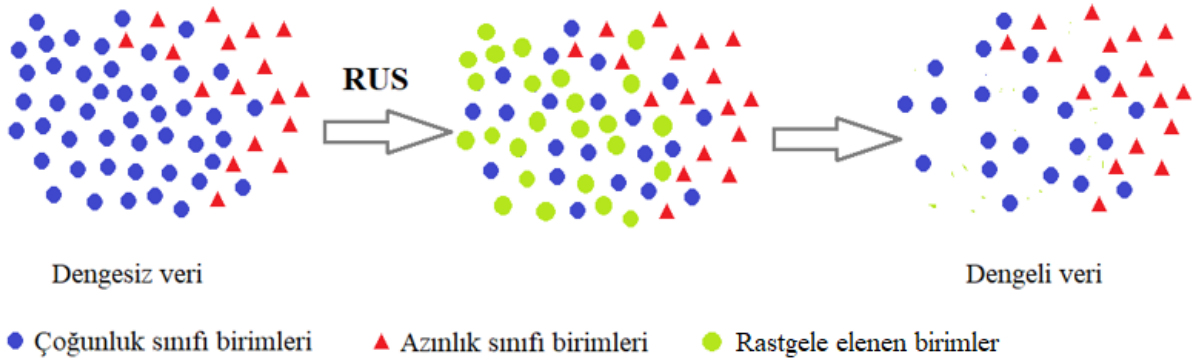
2.2. Veri Dengeleme Algoritmaları

Sınıf dengesizliği probleminin çözümüne yönelik genel bir yaklaşım veri dengeleme (ya da yeniden örnekleme) algoritmalarıdır. Veri dengeleme algoritmaları, alt örnekleme ve aşırı örnekleme denen iki temel yaklaşım üzerine inşa edilmiştir. Aşırı örnekleme yaklaşımında, azınlık sınıfının birim sayısı artırılarak, alt örnekleme yaklaşımında ise çoğunluk sınıfının birim sayısı azaltılarak sınıfların çarpık dağılımı dengelenmeye çalışılır. Literatürde çok sayıda aşırı örnekleme ve alt örnekleme algoritmaları ile bu algoritmaları bagging ve boosting gibi topluluk öğrenme teknikleriyle birleştiren çeşitli algoritmalar önerilmiştir (Barua ve diğerleri, 2012; Chawla ve diğerleri, 2002; He ve diğerleri, 2008; Seiffert ve diğerleri, 2009; Sun ve diğerleri, 2015).

Bu çalışmada, RUS, ROS, SMOTE, MWMOTE, ADASYN, RUSBoost ve UB olmak üzere yedi farklı dengeleme algoritması kullanıldı.

2.2.1. Rastgele Alt Örnekleme (RUS)

RUS algoritması, dengeli bir veri seti elde etmek için çoğunluk sınıfındaki birimleri rastgele eleyerek sınıf dağılımını dengelemeyi amaçlayan bir alt örnekleme yöntemidir (Şekil 1).

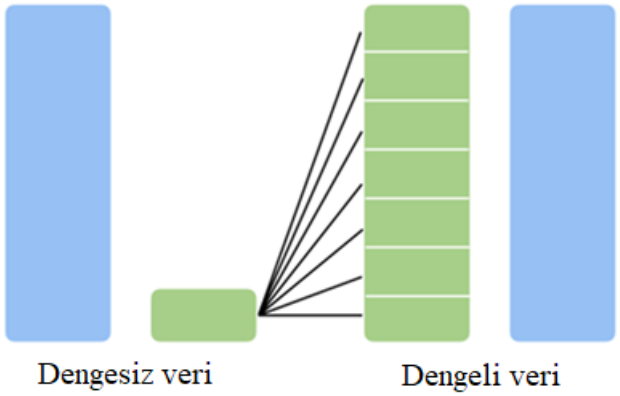


Şekil 1. RUS

RUS algoritması, birim sayısını azaltacağından sınıflandırma süresinin de azalmasını sağlar. RUS'un en büyük dezavantajı, sınıflandırma için önemli olabilecek çoğunluk sınıfı birimlerini veri setinden atma olasılığıdır (Fernández ve diğerleri, 2018).

2.2.2. Rastgele Aşırı Örnekleme (ROS)

ROS algoritması, azınlık sınıfı birimlerinin rastgele kopyalanarak çoğaltılması yoluyla sınıf dağılımını dengelemeyi amaçlayan bir aşırı örnekleme yöntemidir (Şekil 2).



Şekil 2. ROS

ROS algoritması, azınlık sınıfı birimlerinin birebir kopyalarını azınlık sınıfına eklediğinden, aşırı uyum olasılığını artırabilir (Fernández ve diğerleri, 2018). Ayrıca, birim sayısı artacağından sınıflandırma süresi de artar.

2.2.3. Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE)

SMOTE, azınlık sınıfı birimlerini rastgele kopyalamak yerine birimlerin k en yakın komşusunu temel alarak sentetik azınlık sınıfı birimleri üretmeye dayanan en popüler aşırı örnekleme algoritmalarından biridir. SMOTE algoritması, sentetik birimleri üretmek için interpolasyon tekniğini kullanır. Bunun için azınlık sınıfı birimi ile k en yakın komşuları arasındaki farklar alınır. Böylece orijinal azınlık sınıfı birimleri arasındaki benzerliklere dayanarak yeni sentetik veri noktaları oluşturulur.

SMOTE algoritmasının çalışma adımları aşağıdaki gibi özetlenebilir(Chawla ve diğerleri, 2002; Fernández ve diğerleri, 2018):

Adım 1. Eğitim setindeki azınlık sınıfından rastgele bir birim (x_i) seçilir,

Adım 2. x_i'nin azınlık sınıfı içerisindeki k en yakın komşusu (x_{ik}) bulunur,

Adım 3. x_i ile x_{ik} arasındaki fark (Öklid uzaklığı) hesaplanır,

Adım 4. Adım 3'te bulunan fark, 0 ile 1 arasında rastgele seçilen bir sayı (c) ile çarpılır.

Adım 5. Sentetik birim aşağıda verilen eşitlik yardımı ile elde edilir,

$$x_{s_i} = x_i + (x_i - x_{ik}) * c$$

Adım 6. İstenen sayıda sentetik birim üretmek için Adım 1-5 tekrarlanır.

Sentetik birimler, sınıflandırıcının daha küçük ve daha spesifik bölgeler yerine daha büyük ve daha az spesifik karar bölgeleri oluşturmasını sağlar. Böylece SMOTE, sınıflandırıcının daha iyi genelleştirilmesine yardımcı olur (Chawla ve diğerleri, 2002).

2.2.4. Adaptif Sentetik Örnekleme Yaklaşımı (ADASYN)

ADASYN, azınlık sınıfı birimlerinin dağılımlarına göre uyarlamalı olarak azınlık sınıfı birimleri oluşturma fikrine dayanan bir aşırı örnekleme algoritmasıdır. Öğrenilmesi daha kolay olan azınlık sınıfı birimlerine kıyasla öğrenilmesi daha zor olan azınlık sınıfı birimleri için daha fazla sentetik birim üretir. ADASYN, yalnızca dengesiz veri dağılımının getirdiği öğrenme yanlılığını azaltmakla kalmaz, aynı zamanda öğrenilmesi zor olan birimlere odaklanmak için karar sınırını uyarlamalı olarak değiştirir (Fernández ve diğerleri, 2018; He ve diğerleri, 2008). Bu da, her bir azınlık sınıfı birimi için eşit sayıda sentetik birimin üretildiği SMOTE algoritması ile ADASYN arasındaki temel farktır.

ADASYN algoritmasının çalışma adımları aşağıdaki gibi açıklanabilir (He ve diğerleri, 2008):

$D = \{x_i, y_i\}$, $i = 1, \dots, n$ olmak üzere, n birimli eğitim setini temsil etsin. Burada, x_i, p boyutlu **X** bağımsız değişkenler matrisinin bir birimini ve $y_i \in Y = \{1, -1\}$ bağımlı değişkenin kategorilerini temsil etmektedir. n_{az} ve $n_{\text{çoğ}}$ sırasıyla azınlık ve çoğunluk sınıfı birim sayıları olmak üzere $n_{az} \leq n_{\text{çoğ}}$ ve $n_{az} + n_{\text{çoğ}} = n$'dir.

Adım 1. Sınıf dengesizliğinin derecesi $d = n_{az}/n_{\text{çoğ}}$ eşitliği yardımı ile hesaplanır. Burada $d \in (0,1]$'dir.

Adım 2. d_{th}, sınıf dengesizliği oranının tolere edilen maksimum derecesi için önceden ayarlanmış bir eşik değer olmak üzere eğer $d < d_{th}$ ise:

(a) Azınlık sınıfı için üretilmesi gereken sentetik birimlerin sayısı, $G = (n_{\text{çoğ}} - n_{az}) \times \beta$ eşitliği kullanılarak hesaplanır. Burada $\beta \in [0,1]$, sentetik verilerin oluşturulmasından sonra istenen denge oranını belirtmek için kullanılan bir parametredir. $\beta = 1$ olması dengeleme işleminden sonra tamamen dengeli bir veri seti oluşturulduğu anlamına gelir.

(b) Her x_i azınlık sınıfı biriminin p boyutlu uzayda Öklid uzaklığına göre k en yakın komşusu bulunur ve r_i oranı $i = 1, \dots, n_{az}$ olmak üzere $r_i = \Delta_i / k$ eşitliği kullanılarak hesaplanır. Burada Δ_i, x_i'nin k en yakın komşuları içinde çoğunluk sınıfına ait olan birimlerin sayısıdır, bu nedenle $r_i \in [0,1]$'dir.

(c) $\hat{r}_i = r_i / \sum_{i=1}^{n_{az}} r_i$ eşitliği kullanılarak r_i normalize edilir. Burada $\sum_i \hat{r}_i = 1$ olduğundan $\hat{r}_i$ bir yoğunluk dağılımıdır.

(d) Her x_i azınlık sınıfı birimi için üretilmesi gereken sentetik birimlerin sayısı $g_i = \hat{r}_i \times G$ eşitliği kullanılarak hesaplanır. Burada G, adım 2(a)'da tanımlanan azınlık sınıfı için üretilmesi gereken sentetik birimlerin toplam sayısıdır.

(e) Her x_i azınlık sınıfı birimi için g_i tane sentetik birim, aşağıdaki adımlara göre oluşturulur:

(i) x_i'nin en yakın k komşusu arasından rastgele bir x_{zi} azınlık sınıfı birimi seçilir.

(ii) x_{s_i} sentetik birimi $x_{s_i} = x_i + (x_{zi} - x_i) \times \mathrm{c}$ eşitliği kullanılarak hesaplanır. Burada, $\mathrm{c} \in [0,1]$, rastgele bir sayıdır.

2.2.5. Çoğunluk Ağırlıklı Azınlık Aşırı Örnekleme Tekniği (MWMOTE)

MWMOTE, temel amacı hem birim seçimini hem de sentetik birim oluşturma prosedürünü iyileştirmek olan bir aşırı örnekleme algoritmasıdır. MWMOTE algoritması sentetik birimleri oluştururken, bir kümeleme yaklaşımı kullanır. Kümeleme kullanılmasının amacı, herhangi bir yanlış veya gürültülü sentetik birim oluşumunu önlemek için üretilen birimlerin azınlık sınıf alanı içinde kalmasını sağlamaktır. MWMOTE, öğrenmesi zor olan azınlık sınıfı birimlerini etkin bir şekilde seçmekle kalmaz, aynı zamanda bunlara uygun şekilde ağırlık verir. MWMOTE algoritması basitçe üç aşamada özetlenebilir: İlk aşamada, orijinal azınlık sınıfının en önemli ve

öğrenmesi zor azınlık sınıfı birimleri belirlenir ve belirlenen birimler için yeni bir set oluşturulur. İkinci aşamada, yeni setin her üyesine set içindeki önemine göre bir seçim ağırlığı verilir. Üçüncü aşamada, önceki ağırlıklar kullanılarak sentetik birimler üretilir ve orijinal sete eklenerek çıktı seti elde edilir (Barua ve diğerleri, 2012; Fernández ve diğerleri, 2018). MWMOTE algoritmasının çalışma adımları aşağıdaki gibi açıklanabilir (Barua ve diğerleri, 2012):

$S_{çoğ}$: Çoğunluk sınıfı kümesi,

S_{az}: Azınlık sınıfı kümesi,

G: Üretilecek sentetik gözlemlerin sayısı,

k_1: Gürültülü azınlık sınıfı birimlerini tahmin etmek için kullanılan komşu sayısı,

k_2: Bilgilendirici azınlık sınıfını oluşturmak için kullanılan çoğunluk sınıfı komşu sayısı,

k_3: Bilgilendirici azınlık sınıfını oluşturmak için kullanılan azınlık sınıfı komşu sayısı, olmak üzere:

Adım 1. Her azınlık sınıfı birimi $x_i \in S_{az}$ için Öklid uzaklığına göre x_i'nin en yakın k_1 komşularından oluşan $NN(x_i)$ kümesi hesaplanır.

Adım 2. Komşularında hiç azınlık birimi olmayan azınlık sınıfı birimlerini çıkararak filtrelenmiş azınlık kümesi S_{azf} oluşturulur:

$$S_{azf} = S_{az} - \{x_i \in S_{az}: NN(x_i) \text{ azınlık sınıfı birimi içermiyor}\}$$

Adım 3. Her $x_i \in S_{azf}$ için Öklid uzaklığına göre x_i'nin en yakın k_2 çoğunluk sınıfı komşularından oluşan $N_{çoğ}(x_i)$ en yakın çoğunluk kümesi hesaplanır.

Adım 4. Tüm $N_{çoğ}(x_i)$ kümelerinin birleşimi alınarak $S_{bçoğ} = \bigcup_{x_i \in S_{azf}} N_{çoğ}(x_i)$ sınır çoğunluk kümesi bulunur.

Adım 5. Her çoğunluk sınıfı birimi $x_i' \in S_{bçoğ}$ için Öklid uzaklığına göre x_i''nin en yakın k_3 azınlık sınıfı komşularından oluşan $N_{az}(x_i')$ en yakın azınlık kümesi hesaplanır.

Adım 6. Tüm $N_{az}(x_i')$ kümelerinin birleşimi alınarak $S_{iaz} = \bigcup_{x_i' \in S_{bçoğ}} N_{az}(x_i')$ bilgilendirici azınlık kümesi bulunur.

Adım 7. Her $x_i' \in S_{bçoğ}$ ve $x_i \in S_{iaz}$ için $I_w(x_i', x_i)$ bilgi ağırlıkları hesaplanır.

Adım 8. Her $x_i \in S_{iaz}$ için seçim ağırlıkları $S_w(x_i) = \sum_{x_i' \in S_{bçoğ}} I_w(x_i', x_i)$ hesaplanır.

Adım 9. Her $S_w(x_i)$ seçim ağırlıkları, $S_p(x_i) = S_w(x_i)/\sum_{z_i \in S_{iaz}} S_w(z_i)$ seçim olasılıklarına dönüştürülür.

Adım 10. $L_1, L_2, \ldots, L_M$ olmak üzere M kümeden oluşan S_{az}'ın kümeleri bulunur.

Adım 11. $S_{oaz} = S_{az}$ olarak tanımlanır ve aşağıdaki adımlar $j = 1 \ldots G$ için tekrarlanır.

(a) Bir x_i birimi, $S_p(x_i)$ olasılık dağılımına göre S_{iaz} kümesinden seçilir. Burada x_i, L_k kümesinin bir üyesidir.

(b) L_k kümesinden rastgele bir x_i' birimi seçilir.

(c) x_{s_i} sentetik birimi, $x_{s_i} = x_i + c \times (x_i' - x_i)$ eşitliği kullanılarak oluşturulur. Burada $c \in [0,1]$, rastgele bir sayıdır.

(d) Üretilen x_{s_i} sentetik birimi S_{oaz} kümesine eklenir.

2.2.6. Alt Bagging (UB)

Leo Breiman (1996) tarafından önerilen bagging yöntemi, model kararlılığını artırarak tahmin varyansını azaltmaya yönelik kullanılan bir topluluk öğrenme yöntemidir. Bagging yönteminde, eğitim setinden iadeli seçim yoluyla yeni eğitim setleri örneklenir. Yerine koyarak örnekleme yapıldığı için eğitim setindeki bir birim, yeni eğitim setlerinde birden fazla sayıda olabilir ya da hiç bulunmayabilir. Ayrıca eğitim setindeki her birimin seçilme şansı eşittir. Bu tür bir örneklem bootstrap örneklem olarak adlandırılır (Efron ve Tibshirani, 1994). Daha sonra her bootstrap örnekleminden bir sınıflandırma ya da regresyon modeli oluşturulur. Son olarak, modeller sınıflandırma için oylama yapılarak ya da regresyon için ortalama alınarak birleştirilir.

UB algoritması, RUS ve bagging yöntemini birleştiren bir dengeleme algoritmasıdır. UB algoritmasının çalışma adımları aşağıdaki gibi açıklanabilir (Barandela ve diğerleri, 2003):

D eğitim seti iki sınıflı bir veri seti olsun.

Adım 1. D eğitim setinden bootstrap yöntemi ile D_b eğitim setleri örneklenir. Burada b, oluşturulacak eğitim setlerinin sayısını ifade etmektedir.

Adım 2. Yeniden örnekleme oranı β belirlenir ve istenilen denge oranına sahip eğitim setleri RUS algoritması ile elde edilir.

Adım 3. Her bir eğitim seti ile bir sınıflandırma modeli eğitilir.

Adım 4. Sınıflandırma modelleri ile test seti tahmin edilir.

Adım 5. Tahminler oylama yöntemi ile birleştirilir.

2.2.7. Rastgele Alt Boosting (RUSBoost)

Boosting, ilk kez 1990 yılında Schapire tarafından tanıtılmıştır. Boosting, amacı zayıf öğrenicileri güçlü öğrenicilere dönüştürmek olan bir dizi algoritma olarak tanımlanabilir (Schapire, 1990). Literatürde çok sayıda boosting algoritması geliştirilmiş olup Freund ve Schapire (1996) tarafından önerilen AdaBoost algoritması, boosting ailesindeki en popüler algoritmalardan biridir. Bagging'den farklı olarak AdaBoost algoritmasında, iadeli seçim yoluyla birbirinden bağımsız olarak elde edilen eğitim setleri yerine her birime önemini veya sınıflandırılmasındaki zorluğu ifade eden bir ağırlık atanarak sıralı ve aşamalı olarak oluşturulan eğitim setleri kullanılır.

AdaBoost algoritmasında, başlangıçta tüm birimlere eşit ağırlık verilir. Daha sonra her sınıflandırıcıyı tekrarlı olarak eğitmek için tüm veri setini kullanır, ancak her tekrardan sonra her birime verilen ağırlıklar güncellenir. Geçerli tekrar sırasında yanlış sınıflandırılan birimlere, bir sonraki tekrarda doğru şekilde sınıflandırmak amacıyla daha fazla odaklanılır. Bu nedenle, bir sonraki sınıflandırıcının temel amacı, önceki tekrarlarda sınıflandırılması daha zor olan birimleri daha iyi öğrenmektir. Bu birimlere daha fazla odaklanmak için her tekrardan sonra yanlış sınıflandırılan birimlerin ağırlıkları artırılır. Böylece bir birimin hatası ne kadar yüksek ise sonraki tekrarda sınıflandırıcıyı eğitmek için seçilecek örnekleme girme olasılığı o kadar yüksek olur. Ayrıca, daha sonra ağırlıklı bir oylama gerçekleştirmek için test aşamasında kullanılan genel doğruluğuna bağlı olarak her bir sınıflandırıcıya başka bir ağırlık (güven indeksi) atanır. Son olarak, yeni bir birimi sınıflandırmak için her sınıflandırıcı, ağırlıklı bir oy (sınıflandırıcının doğruluğu ne kadar yüksek ise ağırlıklı oyu da o kadar fazladır) verir ve birimin hangi sınıfa atanacağı çoğunluk tarafından belirlenir (Fernández ve diğerleri, 2018).

RUSBoost, RUS ve AdaBoost'u birleştiren bir dengeleme algoritmasıdır. RUSBoost algoritmasının çalışma adımları aşağıdaki gibi açıklanabilir (Seiffert ve diğerleri, 2009):

D: $D = \{x_i, y_i\}$, $i = 1, \dots, n$ olmak üzere, n birimli eğitim seti,

y_i: $y_i \in Y = \{0,1\}$, bağımlı değişkenin kategorisi,

t: 1 ile T arasındaki tekrar adımı,

h_t: sınıflandırma modeli,

$w_t(i)$: i'nci birimin t tekrarındaki ağırlığı olmak üzere:

Adım 1. Başlangıçta tüm birimlerin ağırlıkları eşit olacak şekilde $w_1(i) = 1/n$ olarak ayarlanır,

Adım 2. $t = 1, \dots, T$ için (a) – (d) adımları tekrarlanır.

(a) Yeni eğitim seti S_t', w_t' ağırlıkları kullanılarak $\%a$'sı azınlık sınıfı olmak üzere RUS algoritması ile elde edilir.

(b) S_t' eğitim seti ile h_t sınıflandırıcısı eğitilir.

(c) Hata $hata_t$, orijinal eğitim seti D ve ağırlık dağılımı w_t'ye göre hesaplanır:

$$hata_t = \sum_{(i,y):y_i \neq y} w_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y))$$

(d) Ağırlık güncelleme parametresi $\alpha_t = hata_t/1 - hata_t$ şeklinde hesaplanır.

(e) Sonraki tekrar için $w_{t+1}(i)$ ağırlık dağılımı güncellenir:

$$w_{t+1}(i) = w_t(i)\alpha_t^{\frac{1}{2}(1+h_t(x_i,y_i)-h_t(x_i,y:y\neq y_i))}$$

(f) $w_{t+1}(i)$ ağırlıkları, $Z_t = \sum_i w_{t+1}(i)$ olmak üzere $w_{t+1}(i) = w_{t+1}(i)/Z_t$ şeklinde normalize edilir.

Adım 3. T tekrardan sonra h_t sınıflandırma modelleri ağırlıklı oylama ile birleştirilir:

$$H(x) = \lim_{y \in Y} \sum_{t=1}^{T} h_t(x, y) log \frac{1}{\alpha_t}$$

2.3. Sınıflandırma ve Regresyon Ağaçları (CART)

CART, bağımlı değişkenin kategorik veya sayısal olmasına bağlı olarak sınıflandırma veya regresyon ağaçları üreten parametrik olmayan bir karar ağacı tekniğidir (Breiman ve diğerleri, 2017). CART, hem geniş örneklem hacmine hem de çok sayıda tahmin değişkenine sahip veri setlerine uygulanabilir ve sapan değerlere karşı son derece dirençlidir.

CART algoritması bağımsız değişkenlere göre bağımlı değişkenin ikili bölünmeler şeklinde iki gruba ayrılması temeline dayanır. Eğitim setindeki birimler sonuç değişkeni için benzer değerlere sahip birimlerin bulunduğu alt dallara tekrarlı olarak bölünür ve her dallanma sonucunda iki alt düğüm oluşur. CART algoritması, her düğüm için tüm bağımsız değişkenleri ve tüm olası ikili bölünme değerlerini dikkate alarak en uygun dallanmayı seçip ağacı büyütür (Han ve diğerleri, 2011; Kantardzic, 2011; Larose ve Larose, 2014). CART algoritmasında bölünmenin başlayacağı değişken önemlidir. En iyi bölünmeyi sağlayan bağımsız değişken, çeşitli safsızlık veya çeşitlilik ölçüleri kullanılarak seçilir. Amaç, bağımlı değişkene göre mümkün olan en homojen grupları üretmektir (Omurlu ve diğerleri, 2014). Çalışmamızda, tekrarlı ikili bölümleme için en iyi bilinen kurallardan olan Gini safsızlık ölçüsü (Gini indeks) kullanıldı. Gini indeksi bağımlı değişken kategorik olduğunda, sınıflandırma ağacı oluşturmak için kullanılan bir seçim kriteridir.

D eğitim seti, y_i, $i = 1, \dots, k$ olmak üzere sınıf değişkeni, $|y_i|$ ve $|D|$, sırasıyla, i.sınıfın ve eğitim setinin birim sayısı olmak üzere Gini indeks aşağıdaki gibi ifade edilebilir (Breiman ve diğerleri, 2017; Han ve diğerleri, 2011):

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

Burada, p_i, eğitim setindeki bir birimin y_i sınıfına ait olma olasılığıdır ve $|y_i|/|D|$ şeklinde hesaplanır.

D eğitim seti bir $\mathbf{x_i}$ bağımsız değişkene göre D_1 ve D_2 şeklinde ikiye bölünüyor ise $\mathbf{x_i}$ değişkenine göre Gini indeksi, ortaya çıkan her bölümün ağırlıklı toplamı şeklinde aşağıdaki gibi hesaplanır (Breiman ve diğerleri, 2017; Han ve diğerleri, 2011):

$$Gini_{\mathbf{x_i}}(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

Her bağımsız değişken için olası ikili bölünmelerin her biri dikkate alınır. Bağımsız değişken kategorik ise o değişken için minimum Gini indeksini veren alt seti, bölme alt seti olarak seçilir. Sürekli bağımsız değişkenler için değişkenin her noktası kesim noktası olarak ele alınır ve her noktaya göre Gini indeksi hesaplanır. İkili bir bölünmeden sonra safsızlık ölçüsündeki azalma aşağıdaki gibi hesaplanır (Breiman ve diğerleri, 2017; Han ve diğerleri, 2011):

$$\Delta Gini(\mathbf{x_i}) = Gini(D) - Gini_{\mathbf{x_i}}(D)$$

Gini indeksi, bir düğümdeki tüm birimler tek bir kategoriye düştüğünde en küçük değerine, düğümdeki birimler her kategoriye eşit olarak dağıldığında en büyük değerine ulaşır (Sutton, 2005). Safsızlıktaki azalmayı en büyük yapan (veya eşdeğer olarak en küçük Gini indeksine sahip olan) bağımsız değişken, bölme değişkeni olarak seçilir (Han ve diğerleri, 2011).

2.4. Performans Değerlendirme Ölçütleri

Sınıflandırma yöntemleri ile eğitilen tahmin modellerinin performansını değerlendirmek için çeşitli ölçütler vardır. Bunlardan bazıları aşağıdaki gibidir:

- Duyarlılık
- Özgüllük
- Doğruluk
- Pozitif kestirim değeri
- Negatif kestirim değeri
- AUC
- F-Ölçüsü

- G-Ortalama

İkili bir sınıflandırma problemi için yukarıda yer alan performans ölçütlerinin hesaplanmasında 2x2'lik sınıflandırma tablosundan yararlanılır (Tablo 1).

Tablo 1. İki kategorili değişken için sınıflandırma tablosu.

		GERÇEK DURUM		**Toplam**
		Pozitif	**Negatif**	
TAHMİN	**Pozitif**	Doğru Pozitif (DP)	Yalancı Pozitif (YP)	DP+YP
	Negatif	Yalancı Negatif (YN)	Doğru Negatif (DN)	YN+DN
Toplam		DP+YN	YP+DN	N

Duyarlılık: Sınıflandırma modeli tarafından pozitif olarak atanan birimlerin, gerçekte pozitif olan birimler içerisindeki oranına karşılık gelir. Diğer bir ifadeyle, gerçekte pozitif olduğu bilinen bir gözlemin kestirim sonucunun pozitif çıkma olasılığıdır (Yerushalmy, 1947). Duyarlılık, 0 ile 1 arasında değer alır ve 1'e yaklaştıkça modelin performansı artar.

$$Duyarlılık = \frac{DP}{DP + YN}$$

Özgüllük: Sınıflandırma modeli tarafından negatif olarak atanan birimlerin, gerçekte negatif olan birimler içerisindeki oranına karşılık gelir. Diğer bir ifadeyle, gerçekte negatif olduğu bilinen bir gözlemin kestirim sonucunda da negatif çıkması olasılığıdır (Yerushalmy, 1947). Özgüllük, 0 ile 1 arasında değer alır ve 1'e yaklaştıkça modelin performansı artar.

$$Özgüllük = \frac{DN}{YP + DN}$$

Doğruluk: Sınıflandırma modelinin, gerçekte pozitif olan bir birimi pozitif, negatif olan bir birimi de negatif olarak tanımlama oranıdır. Doğru pozitif ve doğru negatif sonuçların toplamının çalışmada bulunan toplam birim sayısına bölünmesi ile bulunur (Metz, 1978). Doğruluk, 0 ile 1 arasında değer alır ve 1'e yaklaştıkça modelin performansı artar.

$$Doğruluk = \frac{DP + DN}{N}$$

Pozitif kestirim değeri: Kestirim değeri pozitif olan bir birimin gerçekte de pozitif olma olasılığıdır (Fletcher, 2019). Pozitif kestirim değeri, 0 ile 1 arasında değer alır ve 1'e yaklaştıkça modelin performansı artar.

$$Pozitif\ Kestirim\ Değeri = \frac{DP}{DP + YP}$$

Negatif kestirim değeri: Kestirim değeri negatif olan bir birimin gerçekte de negatif olma olasılığıdır (Fletcher, 2019). Negatif kestirim değeri, 0 ile 1 arasında değer alır ve 1'e yaklaştıkça modelin performansı artar.

$$Negatif\ Kestirim\ Değeri = \frac{DN}{YN + DN}$$

F-ölçüsü: Duyarlılık ile pozitif kestirim değerinin harmonik ortalamasıdır. F-ölçüsü, 0 ile 1 arasında değer alır ve 1'e yaklaştıkça modelin performansı artar (He ve Ma, 2013).

$$F - \text{ölçüsü} = 2 * \frac{Duyarlılık * Pozitif\ Kestirim\ Değeri}{Duyarlılık + Pozitif\ Kestirim\ Değeri}$$

G-ortalama: Bu ölçü, sınıflandırma modelinin hem pozitif hem de negatif sınıflar üzerindeki performansının göreli dengesini hesaba katar. Bunu yapabilmek için sınıflandırıcının hem duyarlılığının hem de özgüllüğünün geometrik ortalaması olarak tanımlanır (Kubat ve diğerleri, 1998). G-ortalama, 0 ile 1 arasında değer alır ve 1'e yaklaştıkça modelin performansı artar.

$$G - ortalama = \sqrt{Duyarlılık * \text{Özg}\ddot{\text{u}}\text{ll}\ddot{\text{u}}k}$$

AUC: Farklı kesim noktaları için hesaplanan duyarlılık ve 1-özgüllük değerleri ile elde edilen ROC eğrisi altında kalan alandır (Hanley ve McNeil, 1982). AUC, 0 ile 1 arasında değer alır ve 1'e yaklaştıkça modelin performansı artar.

$$AUC = \int_0^1 ROCcurve(t)dt$$

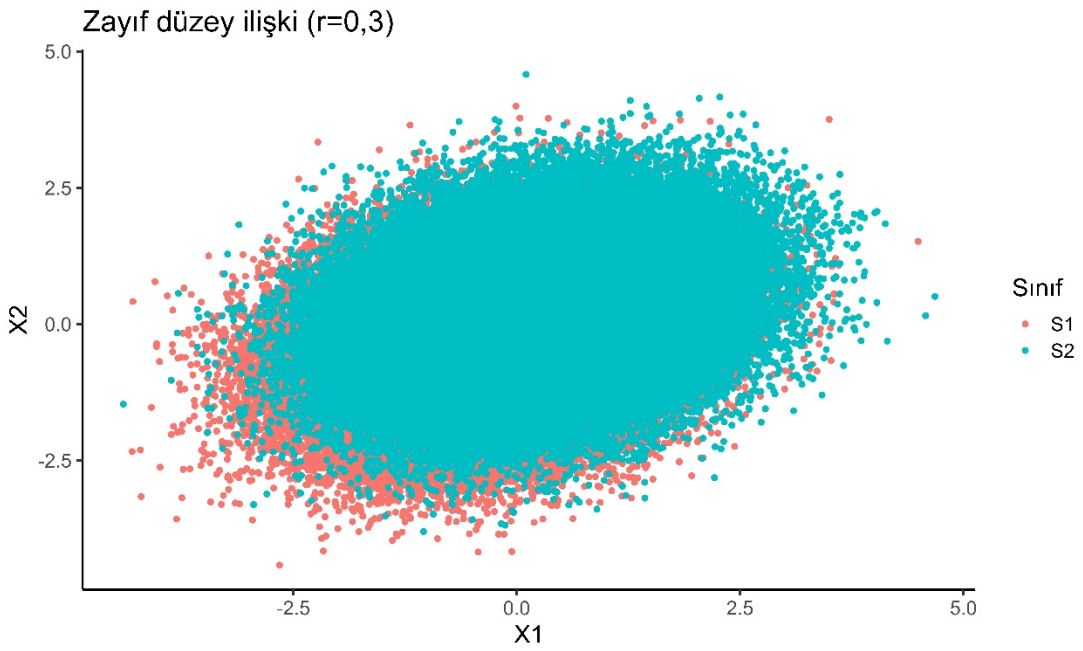
Dengeli bir veri seti için doğruluk oranı, sınıflandırma modelinin performansını değerlendirmede en sık kullanılan ölçüttür. Ancak, dengesiz veri setleri söz konusu olduğunda, sınıflandırma modelinin performansını değerlendirmek için doğruluk oranının kullanılması yanıltıcı sonuçlar verebilir. Örnek olarak, bir veri setinin %1'i azınlık sınıfı ve %99'u çoğunluk sınıfı birimlerinden oluşuyorsa, tüm birimleri çoğunluk sınıfı birimi olarak sınıflandıran bir modelin doğruluk oranı %99 olacaktır. Görünüşte, %99 gibi mükemmele yakın bir doğruluk oranına sahip olan sınıflandırma modelinin azınlık sınıfı birimlerini doğru sınıflandırma oranı %0'dır. Bu nedenle, özellikle dengesiz veri setleri ile elde edilen sınıflandırma modellerinin performansını değerlendirmede hem azınlık hem de çoğunluk sınıflarının doğruluğunu dikkate alan AUC, G-ortalama ve F-ölçüsü gibi performans ölçüleri kullanılmalıdır. Bu çalışmada, sınıflandırma modellerinin değerlendirilmesinde, performans ölçütü olarak AUC kullanıldı.

3. GEREÇ VE YÖNTEM

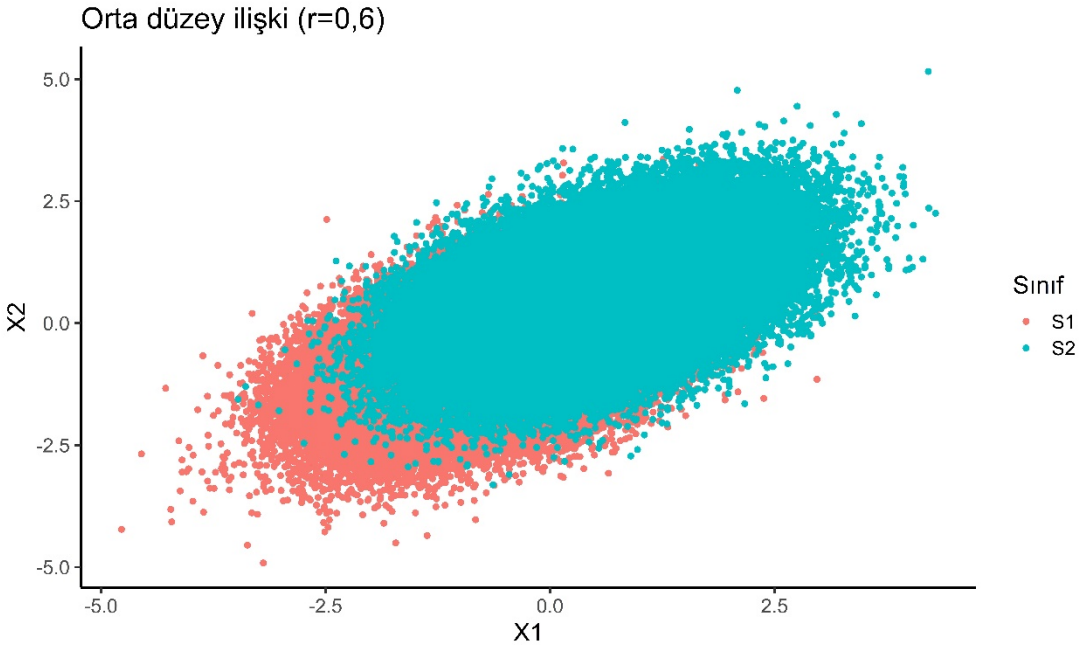
Bu çalışmada, 4 farklı aşırı örnekleme (ROS, SMOTE, MWMOTE ve ADASYN) ve 3 farklı alt örnekleme (RUS, RUSBoost ve UB) algoritmasının performansları ve optimal dengeleme oranları, orijinal bir simülasyon senaryosu ışığında, CART ile incelendi. Bu amaçla hazırlanan simülasyon senaryoları, dengeleme algoritmalarını etkileyebilecek olan korelasyon yapıları (zayıf, orta, yüksek), bağımsız değişken sayıları (2, 3, 4, 5) ve azınlık sınıfı prevalans oranları (0,0025, 0,005, 0,01, 0,05, 0,10, 0,15, 0,20, 0,25, 0,30, 0,35, 0,40) göz önünde bulundurularak oluşturuldu. Simülasyon uygulaması, R programında "stats", "caret", "imbalance", "DMwR", "ebmc", "smotefamily", "measures" ve "ggplot2" paketleri kullanılarak gerçekleştirildi. Simülasyon çalışmasına ilişkin uygulama adımları, sırasıyla, 3.1, 3.2, 3.3 ve 3.4 alt başlıkları altında verilmiştir.

3.1. Toplum Veri Setlerinin Türetimi ve Parametrelerin Hesaplanması

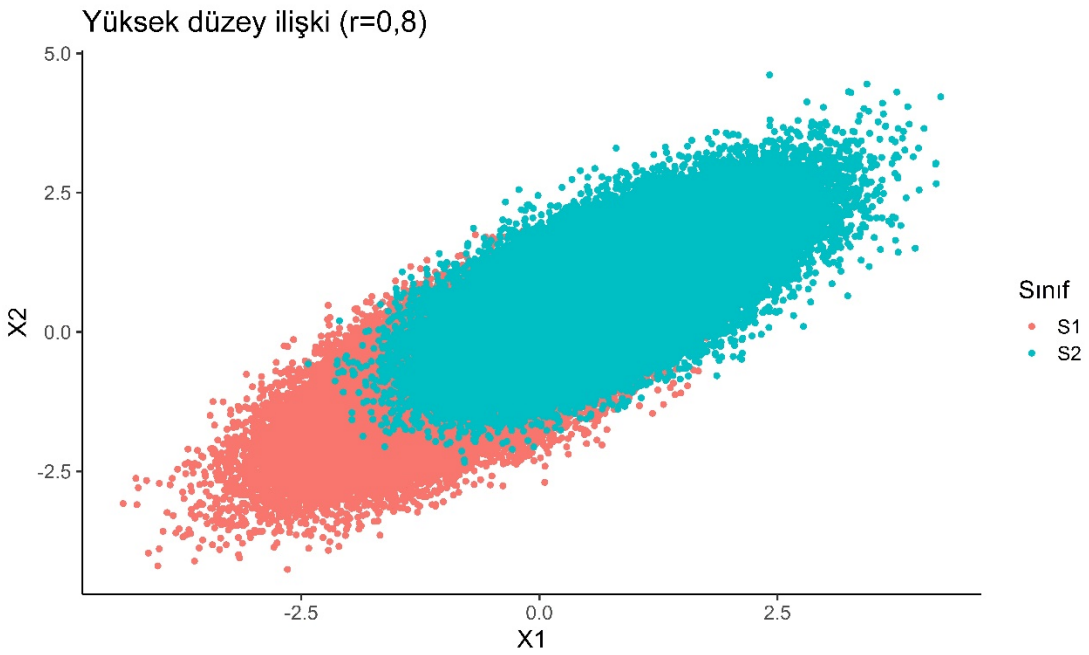
Toplumu temsil edecek olan 1000000 gözlemli ve değişkenler arası korelasyon katsayıları 0,3, 0,6 ve 0,8 olan çok değişkenli standart normal dağılıma sahip ($\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$) 3, 4, 5 ve 6 değişkenli toplam 12 tane veri setleri türetildi. İki sınıflı (S1, S2) ve sınıf değişkeni bakımından tam dengeli veri setleri (n_1=500000 ve n_2=500000) oluşturmak için tüm veri setlerinde, değişkenlerden bir tanesi 50.persantil değerine göre iki gruba ayrıldı. Böylece toplumu temsil edecek olan iki sınıflı, 2, 3, 4 ve 5 bağımsız değişkenli ve değişkenler arası zayıf, orta ve yüksek düzey korelasyona sahip toplum veri setleri elde edildi. Toplum veri setleri, CART algoritması kullanılarak 10 kat çapraz geçerlilik ile sınıflandırıldı ve bazı performans ölçütleri (duyarlılık, özgüllük, doğruluk, pozitif kestirim değeri, negatif kestirim değeri, F-ölçüsü, G-ortalama, AUC) hesaplandı. Sınıflandırma sonucu hesaplanan performans ölçütleri, toplum parametreleri olarak tanımlandı. Şekil 3, 4 ve 5'te sırasıyla, korelasyon yapısının zayıf, orta ve yüksek olduğu durumlara ilişkin türetilen 2 bağımsız değişkenli toplum veri setlerinin saçılım grafiği verildi.



Şekil 3. Zayıf düzey ilişki için iki bağımsız değişkenli toplum veri setinin saçılım grafiği.



Şekil 4. Orta düzey ilişki için iki bağımsız değişkenli toplum veri setinin saçılım grafiği.

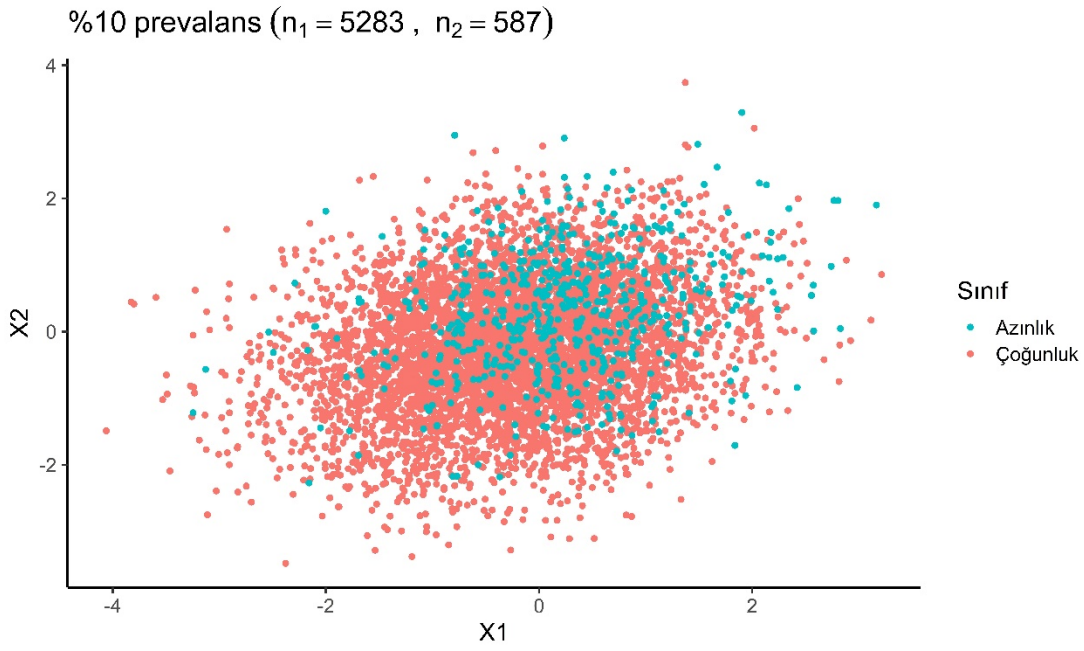


Şekil 5. Yüksek düzey ilişki için iki bağımsız değişkenli toplum veri setinin saçılım grafiği.

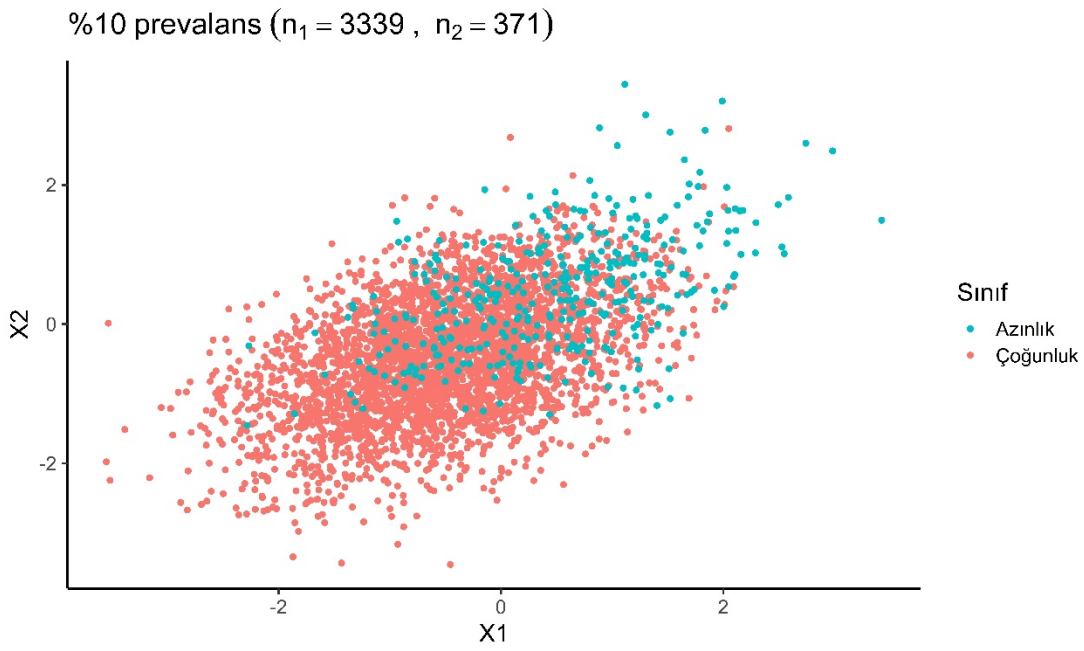
3.2. Dengesiz Veri Setlerinin Oluşturulması

Dengesiz veri setlerinin birim sayıları, prevalans, duyarlılık ve özgüllük oranları dikkate alınarak, %10 etki büyüklüğü, 0,80 güç ve 1.tip hata payı 0,05 olacak şekilde gerekli olan minimum birim sayıları hesaplanarak belirlendi (Bujang ve Adnan, 2016). Toplum veri setlerinin her birinden azınlık sınıfı prevalans oranları %0,25, %0,5, %1, %5, %10, %15, %20, %25, %30, %35 ve %40 olmak üzere 11 farklı dengesiz veri seti basit rastgele örnekleme ile oluşturuldu. Örneğin %10 prevalans oranına sahip dengesiz veri setini oluşturmak için dengeli olan toplum veri setlerinin ilk sınıfındaki birimler (n1=500000) arasından güç analizine göre belirlenen birim sayısının %90'nı diğer sınıftan (n2=500000) ise geriye kalan %10'u bağımsız olarak çekildi. Toplamda farklı korelasyon yapıları, değişken ve birim sayılarına sahip 132 tane dengesiz veri seti elde edildi.

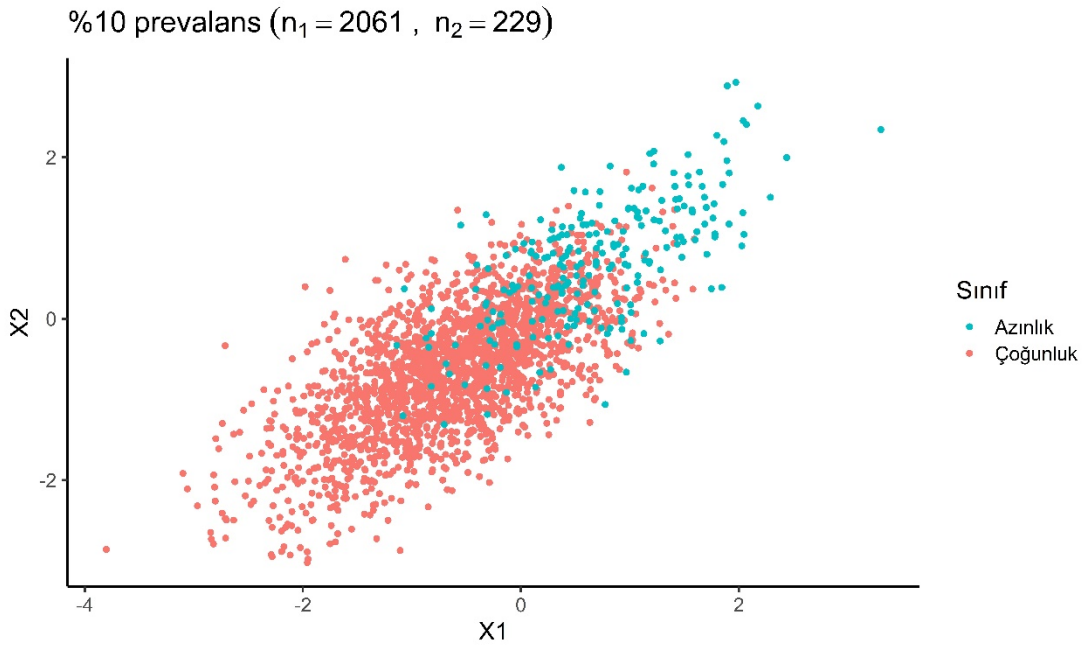
Şekil 6, 7 ve 8'de azınlık sınıfı prevalans oranının %10 ve değişkenler arasındaki ilişki yapısının, sırasıyla, zayıf, orta ve yüksek olduğu veri setlerine ilişkin saçılım grafikleri verildi.



Şekil 6. %10 prevalans ve zayıf ilişkili veri setinin saçılım grafiği.



Şekil 7. %10 prevalans ve orta düzey ilişkili veri setinin saçılım grafiği.



Şekil 8. %10 prevalans ve yüksek ilişkili veri setinin saçılım grafiği.

3.3. Dengesiz Veri Setlerinin Kademeli Olarak Dengelenmesi

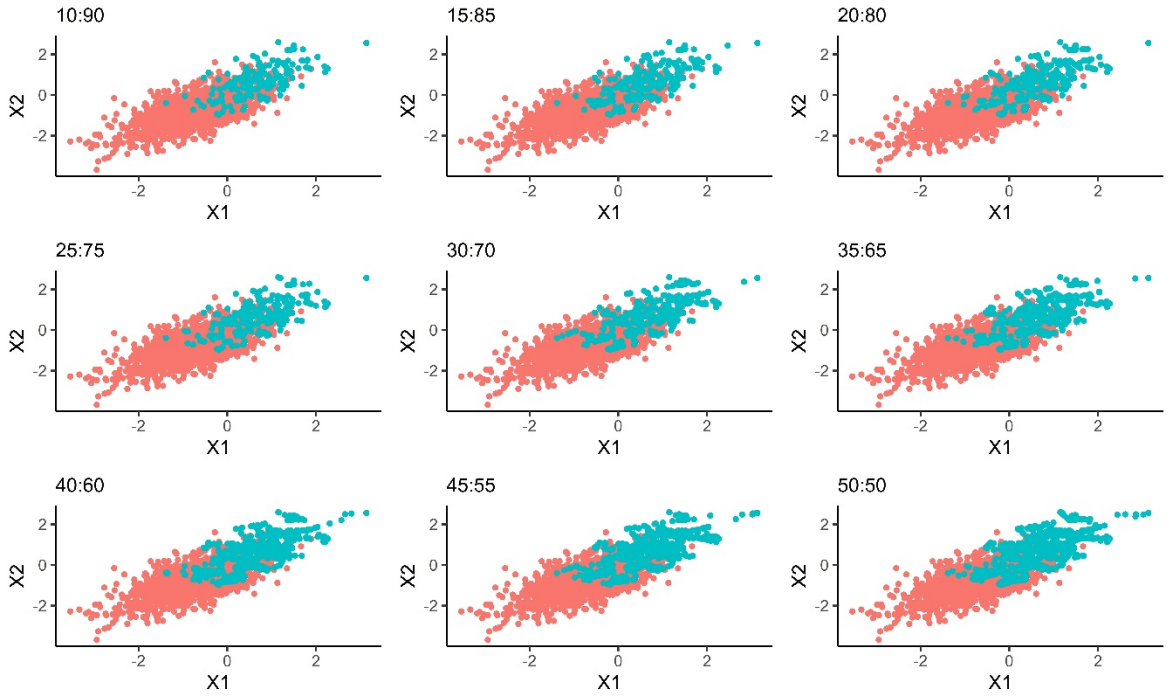
Dengeleme öncesinde, azınlık ve çoğunluk sınıfları, kendi içerisinde rastgele 10 parçaya bölündü. Daha sonra bu parçalar birleştirildi. Böylece herhangi bir parçaya azınlık sınıfından birim düşmeme olasılığı engellendi aynı zamanda her parçada orijinal veri setinin sınıf dağılımı korunmuş oldu. Parçalardan 9 tanesi eğitim seti ve kalan 1 tanesi ise test seti olarak belirlendi. Eğitim setleri, tam dengeye (50:50) ulaşana kadar Tablo 1'de gösterildiği gibi 7 farklı dengeleme algoritması ile kademeli olarak dengelendi. Her kademedeki azınlık ve çoğunluk sınıfı birim sayılarının, tüm dengeleme algoritmaları için birebir aynı olması sağlandı. Şekil 9 ve 10'da, sırasıyla, SMOTE algoritması için %10 prevalans oranına sahip olan veri setindeki azınlık gözlemlerinin kademeli artışı ve RUS algoritması için %10 prevalans oranına sahip olan veri setindeki çoğunluk gözlemlerinin kademeli azalışına ilişkin saçılım grafikleri verildi.

Tablo 2. Sınıf dağılımı ve dengeleme oranları.

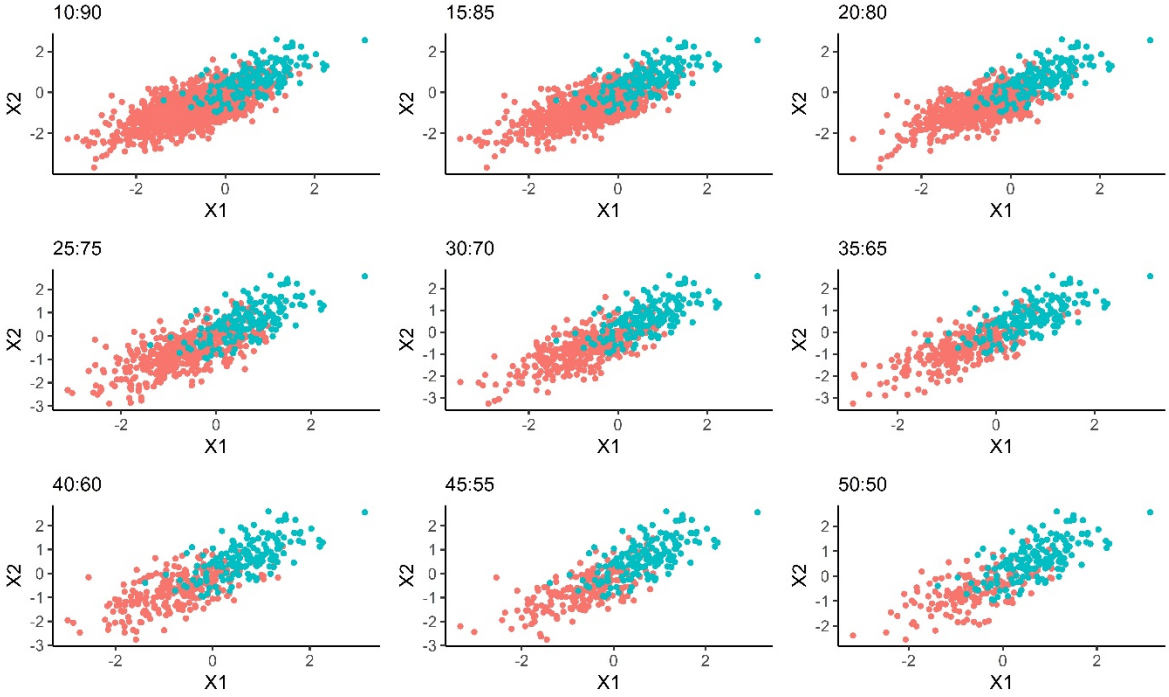
n_{az} (%)	n_{az}:$n_{çoğ}$ (%)						
0,25	10:90	20:80	30:70	35:65	40:60	45:55	50:50
0,5	10:90	20:80	30:70	35:65	40:60	45:55	50:50
1	10:90	20:80	30:70	35:65	40:60	45:55	50:50
5	10:90	20:80	30:70	35:65	40:60	45:55	50:50
10	-	20:80	30:70	35:65	40:60	45:55	50:50
15	-	-	30:70	35:65	40:60	45:55	50:50
20	-	-	30:70	35:65	40:60	45:55	50:50
25	-	-	-	35:65	40:60	45:55	50:50
30	-	-	-	35:65	40:60	45:55	50:50
35	-	-	-	-	40:60	45:55	50:50
40	-	-	-	-	-	45:55	50:50

n_{az}: azınlık sınıfı birim sayısı

$n_{çoğ}$: çoğunluk sınıfı birim sayısı



Şekil 9. SMOTE algoritması için azınlık gözlemlerinin kademeli artışı.



Şekil 10. RUS algoritması için çoğunluk gözlemlerinin kademeli azalışı.

3.4. Sınıflandırma

Bu çalışmada yapılan tüm sınıflandırmalar CART yöntemi ile gerçekleştirildi ve tamamında 10 kat çapraz geçerlilik kullanıldı. Çapraz geçerlilikte her veri seti, dokuzu modeli eğitmek, geri kalan parça ise modeli test etmek için toplam on eşit parçaya ayrıldı. Bu işlem on kez tekrarlandı, böylece her parça bir kez test seti olarak sınıflandırıldı. Bu çalışmada, 132 dengesiz veri seti için kademeli dengeleme aşamasında toplam 684 farklı eğitim seti oluşturuldu. Dengeleme yapılmayan 132 ve dengeleme yapılan 684 olmak üzere toplam 816 eğitim seti elde edildi. On kat çapraz geçerlilik aşamasında toplam 816x10=8160 farklı eğitim seti oluşturuldu. 8160 eğitim seti 7 farklı dengeleme algoritması ile dengelenerek toplam 8160x7=57120 eğitim seti elde edildi. Dengeleme algoritmalarının değerlendirilmesinde, örneklemden kaynaklanabilecek yanlılığı en aza indirmek için tüm dengeleme algoritmalarında aynı eğitim setleri kullanıldı. Bu işlemler 100 kez tekrar edildi. Tekrar sayısının 100 alınmasının nedeni simülasyon senaryolarının kompleks olması ve veri analiz sürecinin çok zaman almasıdır. Sonuç olarak, toplamda 57120x100=5712000 eğitim seti için CART yöntemi ile sınıflandırma modelleri eğitildi. Sınıflandırma modellerini değerlendirmek için her kademede 100 tekrarlı hesaplanan AUC değerleri, tek örneklem t testi kullanılarak toplum veri setinden hesaplanan AUC değeri ile karşılaştırıldı.

4. BULGULAR

Simülasyon çalışmasına ilişkin bulgular, korelasyon düzeylerine (düşük, orta, yüksek) göre üç farklı bölümde özetlendi. Her korelasyon düzeyi için 2, 3, 4 ve 5 bağımsız değişkenli veri setlerine ilişkin bulgular tablo ve grafiklerde verildi. Tablolarda dengeleme yapılmamış ve dengeleme algoritmaları ile kademeli olarak dengelenmiş veri setlerinin 10 kat çapraz geçerlilik ve 100 tekrarlı sınıflandırılmasından elde edilen AUC değerlerinin ortalamaları yer almaktadır. Tüm tablolarda, herhangi bir dengeleme algoritması kullanılmadan yapılan sınıflandırma sonucu elde edilen AUC değerleri "D-Yok" sütununda, toplum veri setinin sınıflandırılması sonucu elde edilen AUC değerleri "Parametre" sütununda ve azınlık:çoğunluk sınıfı birim sayıları yüzdeleri α sütununda yer almaktadır. Tablolarda, AUC ortalamalarının yer aldığı tüm hücreler, tek örneklem t testi kullanılarak toplum veri setlerinden hesaplanan AUC (parametre) değerleri ile karşılaştırılmıştır. Tablolardaki yeşil hücreler, parametre değerinden farklı olmayan, kırmızı hücreler, parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan AUC tahminlerini ifade etmektedir. Bulguların sunumundaki karmaşıklığı azaltmak için bazı azınlık sınıfı prevalans oranlarına (%0,5, %5, %15 ve %25) ilişkin sonuçlara tablo ve şekillerde yer verilmedi. Bulgular %0,25, %1, %10, %20, %30, %35 ve %40 olmak üzere yedi farklı azınlık sınıfı prevalans oranı için verildi. Ayrıca şekillerde görselliğin bozulmaması açısından bazı prevalans oranları için çizilen güven aralığı grafiklerinde "D-Yok" (herhangi bir dengeleme algoritmasının kullanılmadığı durum) durumuna ilişkin güven aralığı grafiği görünmemektedir. Bunun nedeni "D-Yok" durumuna ait AUC değerlerinin y ekseni alt sınır değerinden küçük olmasıdır.

4.1. Zayıf Düzey Korelasyona İlişkin Bulgular

Bu bölümde, değişkenler arası korelasyon katsayısının 0,3 olduğu 2, 3, 4 ve 5 bağımsız değişkenli türetilen toplum veri setlerinden örneklenen dengesiz veri setlerine ilişkin sonuçlar sırasıyla Tablo 3, 4, 5 ve 6'da yer almaktadır.

Tablo 3'te zayıf ilişkili ve iki bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 3'te görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. UB algoritması ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir.

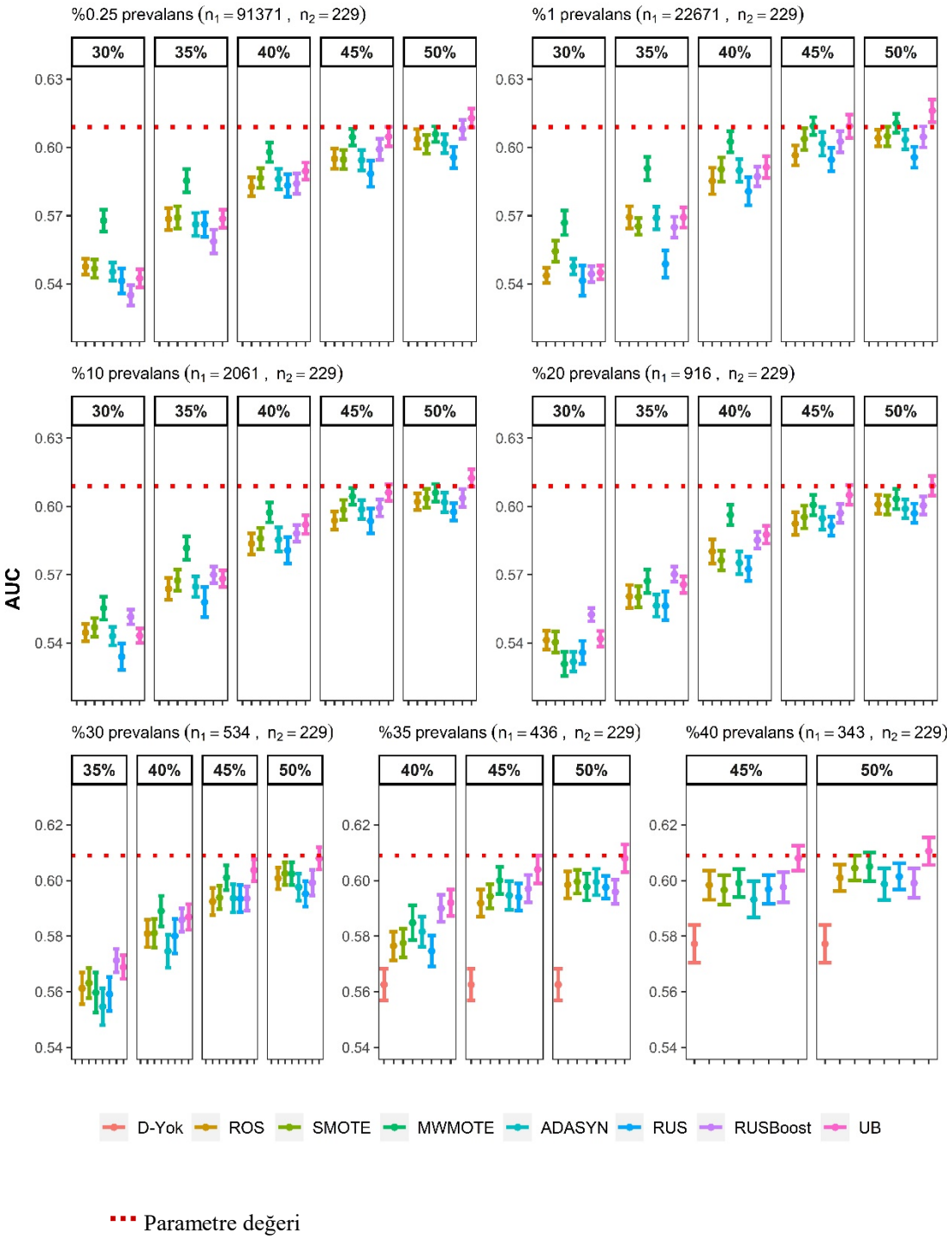
Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında, MWMOTE ve RUSBoost ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, UB ile yapılan dengelemede (45:55) denge oranında parametre değerinden farklı olmadığı, (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında, SMOTE ve RUSBoost ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında, MWMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 ve %35 olan veri setlerinde, UB ile yapılan dengelemede (50:50) denge oranlarında elde edilen AUC değerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Son olarak prevalans oranı %40 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında, SMOTE ve MWMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 3. Zayıf ilişkili ve iki bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5000	0,5041	0,5042	0,5040	0,5041	0,5001	0,5000	0,5000	0,6089
	20	0,5000	0,5144	0,5150	0,5254	0,5153	0,5085	0,5050	0,5015	0,6089
	30	0,5000	0,5476	0,5467	0,5679	0,5454	0,5414	05351	0,5425	0,6089
	35	0,5000	0,5686	0,5693	0,5854	0,5662	0,5661	0,5587	0,5686	0,6089
	40	0,5000	0,5866	0,5866	0,5979	0,5861	0,5832	0,5841	0,5896	0,6089
	45	0,5000	0,5951	0,5947	0,6045	0,5944	0,5884	0,5992	0,6048	0,6089
	50	0,5000	0,6037	0,6015	0,6058	0,6016	0,5956	0,6080	0,6129	0,6089
1:99	10	0,5000	0,5049	0,5039	0,5051	0,5050	0,5001	0,5002	0,5000	0,6089
	20	0,5000	0,5197	0,5199	0,5234	0,5197	0,5099	0,5103	0,5037	0,6089
	30	0,5000	0,5437	0,5544	0,5669	0,5477	0,5414	0,5444	0,5450	0,6089
	35	0,5000	0,5693	0,5653	0,5908	0,5690	0,5488	0,5649	0,5693	0,6089
	40	0,5000	0,5853	0,5904	0,6026	0,5899	0,5807	0,5872	0,5914	0,6089
	45	0,5000	0,5965	0,6038	0,6094	0,6016	0,5946	0,6026	0,6093	0,6089
	50	0,5000	0,6041	0,6049	0,6106	0,6034	0,5957	0,6046	0,6162	0,6089
10:90	20	0,5006	0,5172	0,5149	0,5071	0,5089	0,5059	0,5180	0,5033	0,6089
	30	0,5006	0,5445	0,5468	0,5553	0,5430	0,5340	0,5515	0,5432	0,6089
	35	0,5006	0,5638	0,5675	0,5817	0,5647	0,5579	0,5699	0,5682	0,6089
	40	0,5006	0,5836	0,5859	0,5974	0,5854	0,5807	0,5882	0,5920	0,6089
	45	0,5006	0,5938	0,5985	0,6044	0,5990	0,5936	0,5994	0,6061	0,6089
	50	0,5006	0,6021	0,6036	0,6060	0,6018	0,5976	0,6036	0,6124	0,6089
20:80	30	0,5072	0,5413	0,5403	0,5308	0,5318	0,5358	0,5525	0,5418	0,6089
	35	0,5072	0,5604	0,5603	0,5672	0,5564	0,5563	0,5702	0,5657	0,6089
	40	0,5072	0,5802	0,5763	0,5963	0,5752	0,5725	0,5852	0,5876	0,6089
	45	0,5072	0,5925	0,5983	0,6006	0,5947	0,5914	0,5971	0,6050	0,6089
	50	0,5072	0,6009	0,6006	0,6033	0,5990	0,5970	0,6004	0,6092	0,6089
30:70	35	0,5280	0,5612	0,5632	0,5597	0,5547	0,5592	0,5713	0,5689	0,6089
	40	0,5280	0,5809	0,5811	0,5890	0,5746	0,5801	0,5859	0,5869	0,6089
	45	0,5280	0,5924	0,5939	0,6011	0,5936	0,5935	0,5935	0,6037	0,6089
	50	0,5280	0,6008	0,6025	0,6024	0,5976	0,5952	0,5992	0,6080	0,6089
35:65	40	0,5626	0,5765	0,5775	0,5848	0,5816	0,5746	0,5900	0,5920	0,6089
	45	0,5626	0,5919	0,5943	0,6000	0,5947	0,5940	0,5971	0,6040	0,6089
	50	0,5626	0,5985	0,5996	0,5977	0,5995	0,5976	0,5959	0,6080	0,6089
40:60	45	0,5772	0,5984	0,5966	0,5991	0,5932	0,5969	0,5976	0,6081	0,6089
	50	0,5772	0,6010	0,6045	0,6050	0,5987	0,6014	0,5991	0,6106	0,6089

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 11'de verilmiştir.



Şekil 11. Zayıf ilişkili ve iki bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 4'te zayıf ilişkili ve üç bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 4'te görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, RUSBoost ve UB algoritmaları, bazı denge oranlarında parametre değerinden istatistiksel olarak anlamlı düzeyde yüksek sonuçlar üretmiştir. RUSBoost ve UB dışındaki algoritmalar ile dengelenen veri setleri için tam denge durumunda dahi parametre değerinden yüksek sonuçlar elde edilmemiş, genellikle toplum parametresine yakın fakat parametre değerinden düşük sonuçlar elde edilmiştir.

Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (45:55) denge oranında parametre değerinden farklı olmadığı, (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS ve MWMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). Prevalans oranı %10 olan veri setinde, UB ile yapılan dengelemede (45:55) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ve MWMOTE ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında, ROS ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ile yapılan dengelemede (45:55) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ile yapılan dengelemede (45:55)

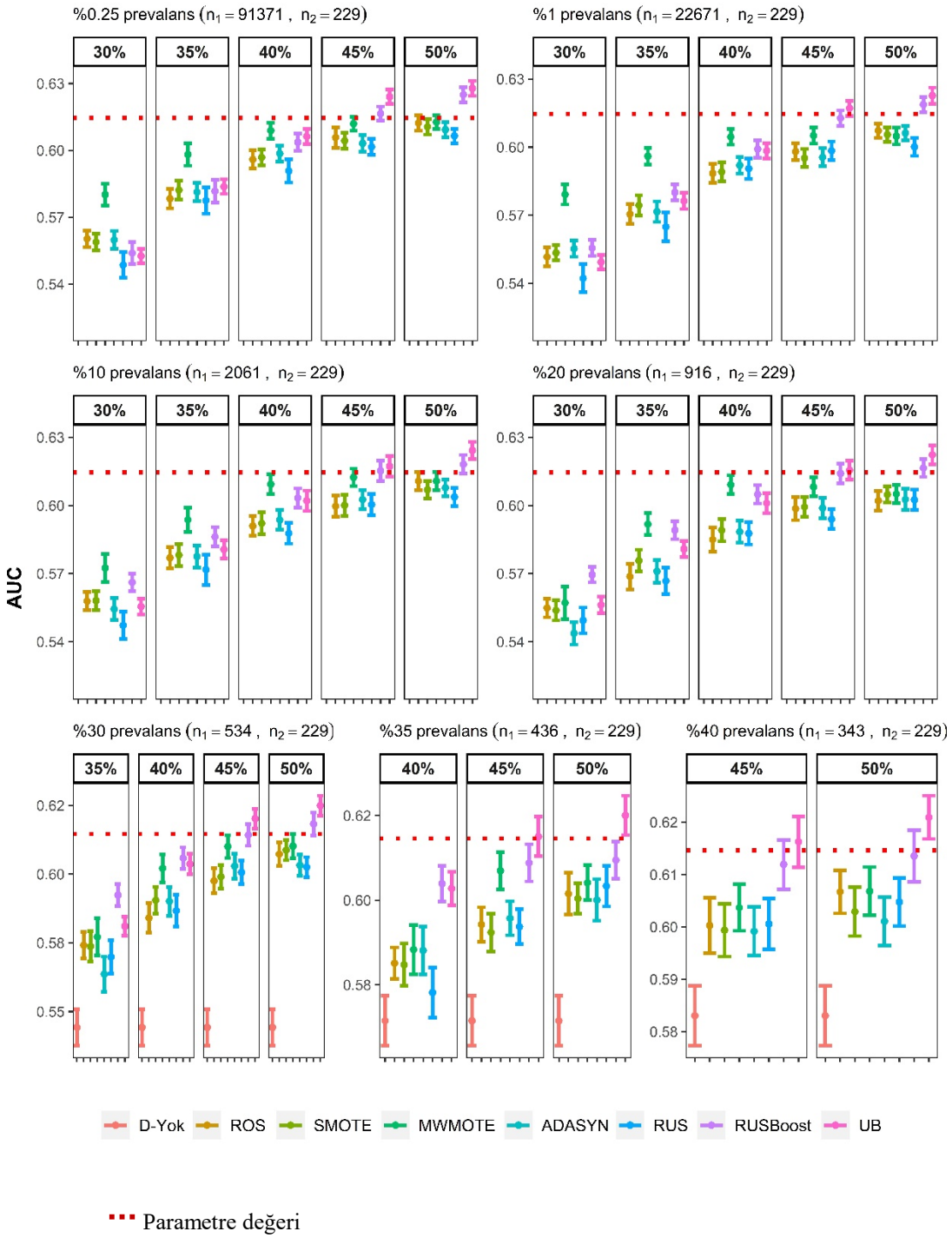
ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost dengelemede (45:55) ve (50:50) denge oranlarında, MWMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde, UB ile yapılan dengelemede (45:55) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). Son olarak prevalans oranı %40 olan veri setinde, UB ile yapılan dengelemede (45:55) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 4. Zayıf ilişkili ve üç bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5000	0,5043	0,5052	0,5044	0,5053	0,5007	0,5000	0,5000	0,6146
	20	0,5000	0,5203	0,5215	0,5310	0,5217	0,5101	0,5075	0,5023	0,6146
	30	0,5000	0,5604	0,5589	0,5801	0,5599	0,5486	0,5540	0,5526	0,6146
	35	0,5000	0,5784	0,5822	0,5982	0,5813	0,5776	0,5817	0,5838	0,6146
	40	0,5000	0,5959	0,5969	0,6089	0,5986	0,5908	0,6037	0,6062	0,6146
	45	0,5000	0,6058	0,6043	0,6120	0,6032	0,6016	0,6165	0,6240	0,6146
	50	0,5000	0,6123	0,6106	0,6127	0,6093	0,6065	0,6250	0,6278	0,6146
1:99	10	0,5001	0,5045	0,5047	0,5032	0,5047	0,5007	0,5005	0,5000	0,6146
	20	0,5001	0,5191	0,5203	0,5282	0,5207	0,5067	0,5117	0,5027	0,6146
	30	0,5001	0,5517	0,5535	0,5792	0,5553	0,5422	0,5556	0,5494	0,6146
	35	0,5001	0,5706	0,5743	0,5960	0,5716	0,5649	0,5800	0,5763	0,6146
	40	0,5001	0,5885	0,5891	0,6046	0,5920	0,5905	0,5991	0,5984	0,6146
	45	0,5001	0,5981	0,5953	0,6051	0,5956	0,5984	0,6128	0,6171	0,6146
	50	0,5001	0,6073	0,6055	0,6051	0,6062	0,6002	0,6187	0,6227	0,6146
10:90	20	0,5003	0,5225	0,5188	0,5092	0,5122	0,5086	0,5262	0,5046	0,6146
	30	0,5003	0,5579	0,5581	0,5724	0,5544	0,5472	0,5662	0,5555	0,6146
	35	0,5003	0,5770	0,5782	0,5937	0,5776	0,5718	0,5862	0,5806	0,6146
	40	0,5003	0,5910	0,5921	0,6094	0,5937	0,5877	0,6033	0,6021	0,6146
	45	0,5003	0,5997	0,6001	0,6124	0,6026	0,6004	0,6153	0,6172	0,6146
	50	0,5003	0,6107	0,6069	0,6107	0,6080	0,6038	0,6181	0,6242	0,6146
20:80	30	0,5115	0,5549	0,5538	0,5571	0,5437	0,5494	0,5695	0,5562	0,6146
	35	0,5115	0,5687	0,5757	0,5918	0,5710	0,5668	0,5891	0,5808	0,6146
	40	0,5115	0,5850	0,5891	0,6091	0,5885	0,5877	0,6049	0,6010	0,6146
	45	0,5115	0,5986	0,5993	0,6082	0,5989	0,5940	0,6141	0,6155	0,6146
	50	0,5115	0,6021	0,6048	0,6051	0,6027	0,6025	0,6165	0,6223	0,6146
30:70	35	0,5442	0,5741	0,5737	0,5771	0,5636	0,5699	0,5924	0,5810	0,6146
	40	0,5442	0,5841	0,5910	0,6021	0,5901	0,5867	0,6058	0,6037	0,6146
	45	0,5442	0,5976	0,5990	0,6100	0,6029	0,6007	0,6142	0,6202	0,6146
	50	0,5442	0,6073	0,6088	0,6102	0,6033	0,6025	0,6182	0,6249	0,6146
35:65	40	0,5715	0,5851	0,5847	0,5883	0,5881	0,5782	0,6039	0,6028	0,6146
	45	0,5715	0,5943	0,5924	0,6070	0,5957	0,5938	0,6088	0,6150	0,6146
	50	0,5715	0,6015	0,6004	0,6042	0,6001	0,6034	0,6095	0,6201	0,6146
40:60	45	0,5831	0,6003	0,5994	0,6037	0,5992	0,6006	0,6119	0,6163	0,6146
	50	0,5831	0,6067	0,6029	0,6068	0,6011	0,6048	0,6135	0,6209	0,6146

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 12'de verilmiştir.



Şekil 12. Zayıf ilişkili ve üç bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 5'te zayıf ilişkili ve dört bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 5'te görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, RUSBoost ve UB algoritmaları, bazı denge oranlarında parametre değerinden istatistiksel olarak anlamlı düzeyde yüksek sonuçlar üretmiştir. RUSBoost ve UB dışındaki algoritmalar ile dengelenen veri setleri için tam denge durumunda dahi parametre değerinden yüksek sonuçlar elde edilmemiş, genellikle toplum parametresine yakın fakat parametre değerinden düşük sonuçlar elde edilmiştir.

Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, UB ve RUSBoost ile yapılan dengelemelerde (40:60) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında, ROS ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) ve (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (40:60), ADASYN ve RUS ile yapılan dengelemelerde ise (50:50) denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (40:60), (45:55) ve (50:50), SMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan

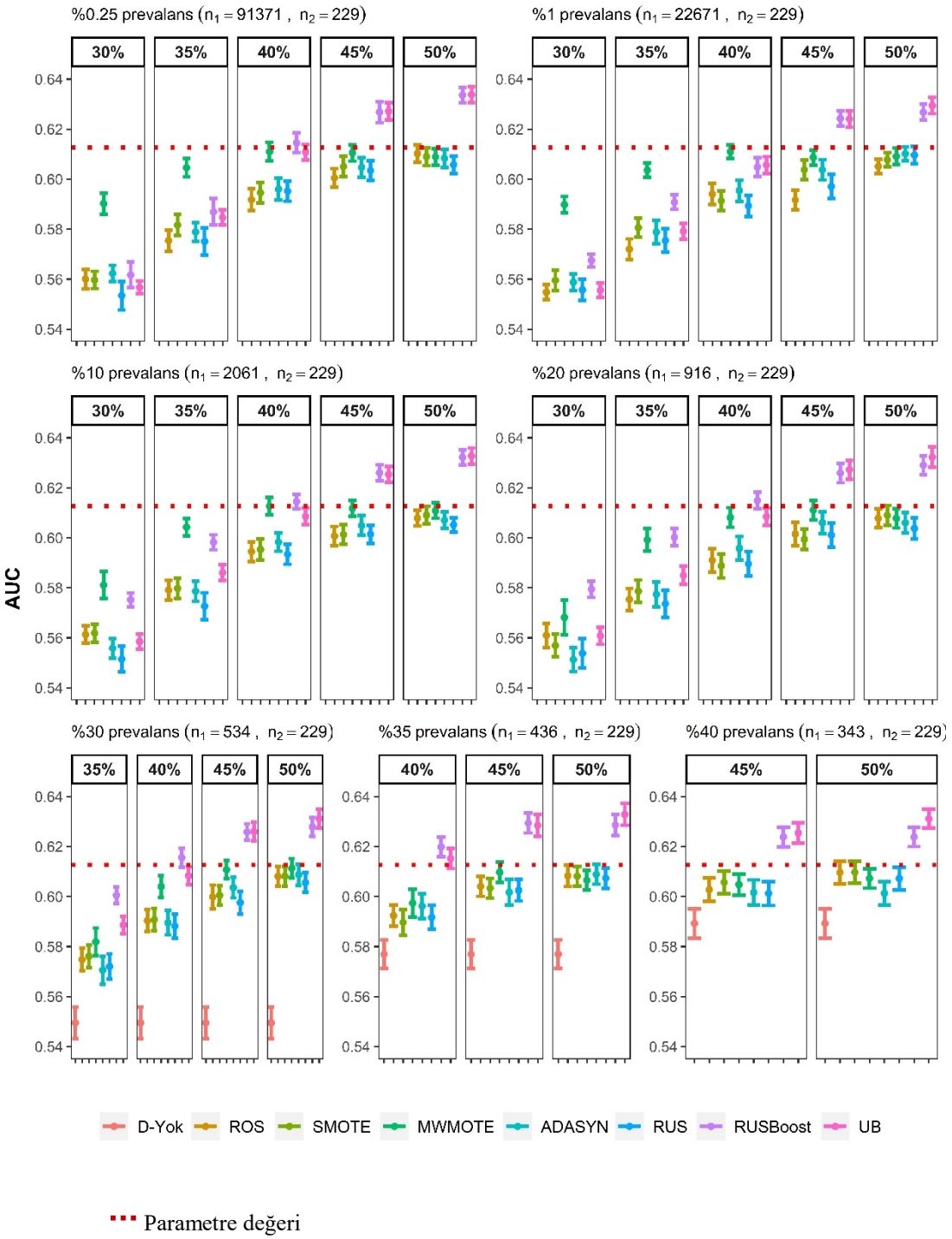
dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (45:55), SMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (45:55) ve (50:50), ADASYN ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde, RUSBoost ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). UB ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, MWMOTE ve ADASYN ile yapılan dengelemelerde, sırasıyla, (50:50), (45:55) ve (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Son olarak prevalans oranı %40 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) ve (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu ($p<0,05$), ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 5. Zayıf ilişkili ve dört bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5000	0,5053	0,5049	0,5058	0,5058	0,5009	0,5001	0,5000	0,6126
	20	0,5000	0,5247	0,5259	0,5411	0,5249	0,5148	0,5102	0,5028	0,6126
	30	0,5000	0,5600	0,5597	0,5902	0,5623	0,5535	0,5617	0,5567	0,6126
	35	0,5000	0,5754	0,5816	0,6046	0,5789	0,5751	0,5869	0,5847	0,6126
	40	0,5000	0,5918	0,5946	0,6110	0,5960	0,5952	0,6145	0,6108	0,6126
	45	0,5000	0,6005	0,6050	0,6105	0,6047	0,6034	0,6268	0,6271	0,6126
	50	0,5000	0,6102	0,6090	0,6089	0,6083	0,6058	0,6336	0,6338	0,6126
1:99	10	0,5000	0,5062	0,5071	0,5042	0,5077	0,5002	0,5006	0,5000	0,6126
	20	0,5000	0,5265	0,5260	0,5444	0,5259	0,5093	0,5156	0,5027	0,6126
	30	0,5000	0,5548	0,5595	0,5898	0,5588	0,5557	0,5675	0,5556	0,6126
	35	0,5000	0,5720	0,5806	0,6036	0,5788	0,5755	0,5908	0,5791	0,6126
	40	0,5000	0,5941	0,5914	0,6110	0,5954	0,5893	0,6049	0,6056	0,6126
	45	0,5000	0,5957	0,6037	0,6087	0,6038	0,5971	0,6242	0,6241	0,6126
	50	0,5000	0,6051	0,6077	0,6091	0,6102	0,6096	0,6268	0,6294	0,6126
10:90	20	0,5001	0,5270	0,5222	0,5124	0,5211	0,5111	0,5328	0,5058	0,6126
	30	0,5001	0,5613	0,5619	0,5811	0,5559	0,5515	0,5752	0,5585	0,6126
	35	0,5001	0,5791	0,5798	0,6043	0,5786	0,5726	0,5982	0,5861	0,6126
	40	0,5001	0,5945	0,5954	0,6127	0,5983	0,5934	0,6145	0,6086	0,6126
	45	0,5001	0,6007	0,6014	0,6117	0,6050	0,6014	0,6260	0,6254	0,6126
	50	0,5001	0,6080	0,6091	0,6109	0,6071	0,6052	0,6323	0,6328	0,6126
20:80	30	0,5095	0,5610	0,5570	0,5682	0,5514	0,5538	0,5795	0,5608	0,6126
	35	0,5095	0,5754	0,5787	0,5992	0,5774	0,5736	0,6002	0,5850	0,6126
	40	0,5095	0,5911	0,5888	0,6082	0,5958	0,5896	0,6149	0,6085	0,6126
	45	0,5095	0,6015	0,5995	0,6111	0,6061	0,6011	0,6259	0,6273	0,6126
	50	0,5095	0,6078	0,6090	0,6078	0,6061	0,6038	0,6291	0,6323	0,6126
30:70	35	0,5495	0,5748	0,5761	0,5819	0,5706	0,5720	0,6005	0,5886	0,6126
	40	0,5495	0,5903	0,5908	0,6040	0,5896	0,5881	0,6156	0,6083	0,6126
	45	0,5495	0,6000	0,6005	0,6107	0,6036	0,5976	0,6258	0,6260	0,6126
	50	0,5495	0,6080	0,6081	0,6113	0,6087	0,6056	0,6278	0,6312	0,6126
35:65	40	0,5770	0,5924	0,5896	0,5973	0,5962	0,5917	0,6198	0,6153	0,6126
	45	0,5770	0,6040	06034	0,6097	0,6017	0,6025	0,6294	0,6284	0,6126
	50	0,5770	0,6083	0,6082	0,6065	0,6089	0,6074	0,6286	0,6329	0,6126
40:60	45	0,5893	0,6027	0,6057	0,6048	0,6016	0,6013	0,6238	0,6254	0,6126
	50	0,5893	0,6096	0,6097	0,6073	0,6017	0,6072	0,6238	0,6311	0,6126

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 13'te verilmiştir.



Şekil 13. Zayıf ilişkili ve dört bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 6'da zayıf ilişkili ve beş bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 6'da görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, RUSBoost ve UB algoritmaları, bazı denge oranlarında parametre değerinden istatistiksel olarak anlamlı düzeyde yüksek sonuçlar üretmiştir. RUSBoost ve UB dışındaki algoritmalar ile dengelenen veri setleri için tam denge durumunda dahi parametre değerinden yüksek sonuçlar elde edilmemiş, genellikle toplum parametresine yakın fakat parametre değerinden düşük sonuçlar elde edilmiştir.

Prevalans oranı %0,25 olan oranlı veri seti için karşılaştırma sonuçları incelendiğinde, UB ve RUSBoost ile yapılan dengelemelerde (40:60) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (40:60) ve (50:50) denge oranlarında, RUS ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, RUSBoost ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). UB ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında, ROS, SMOTE ve ADASYN ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, RUSBoost ve UB ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında, ADASYN ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında, ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (40:60) denge

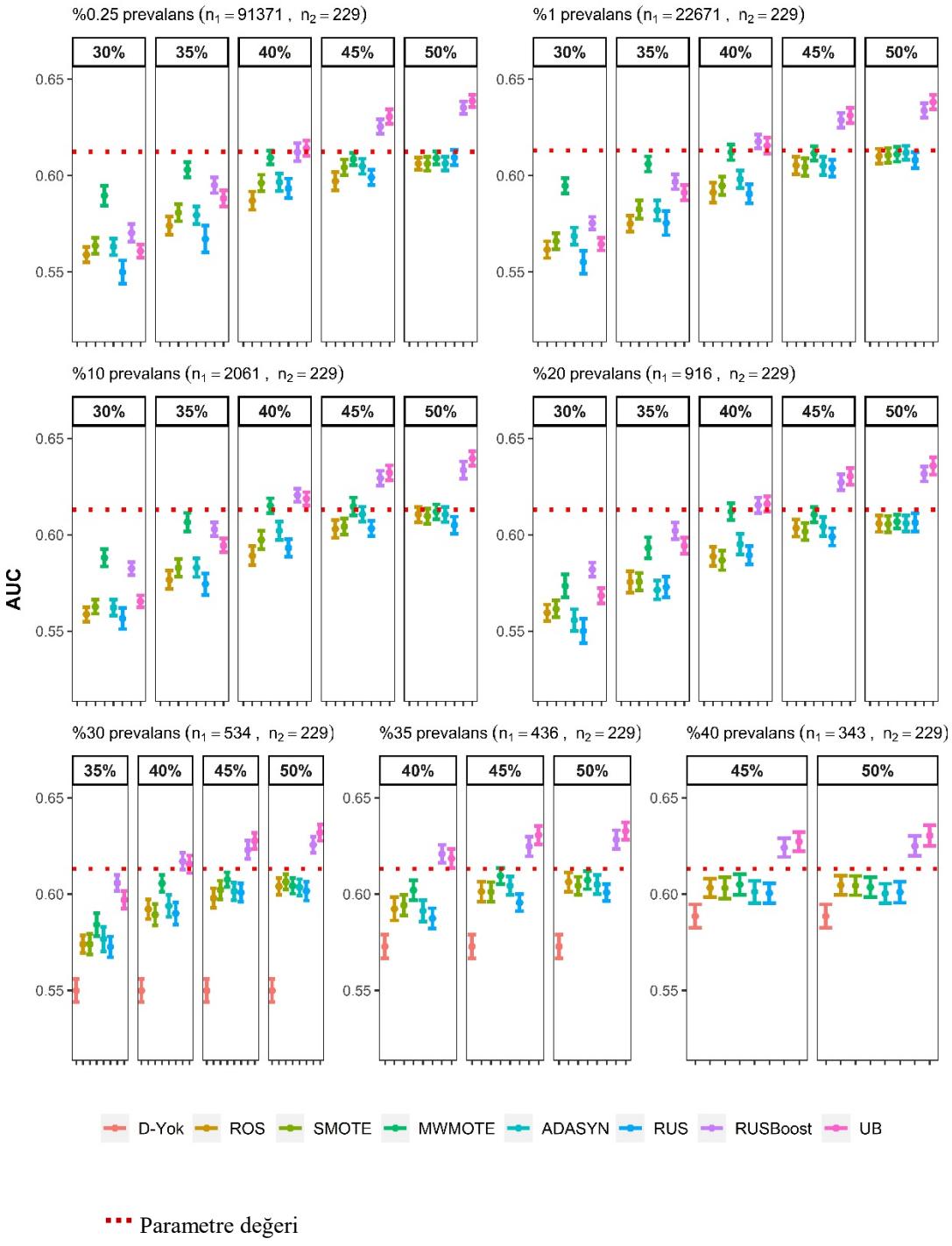
oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (40:60) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). Prevalans oranı %35 olan veri setinde, RUSBoost ve UB ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede ise (45:55) denge oranında elde edilen AUC değerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Son olarak prevalans oranı %40 olan veri setinde, RUSBoost ve UB ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 6. Zayıf ilişkili ve beş bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5000	0,5075	0,5077	0,5079	0,5062	0,5018	0,5011	0,5000	0,6130
	20	0,5000	0,5253	0,5271	0,5452	0,5271	0,5112	0,5217	0,5100	0,6130
	30	0,5000	0,5589	0,5636	0,5896	0,5630	0,5499	0,5703	0,5608	0,6130
	35	0,5000	0,5740	0,5807	0,6030	0,5794	0,5671	0,5950	0,5882	0,6130
	40	0,5000	0,5870	0,5962	0,6094	0,5966	0,5933	0,6122	0,6143	0,6130
	45	0,5000	0,5971	0,6042	0,6085	0,6048	0,5990	0,6254	0,6305	0,6130
	50	0,5000	0,6062	0,6061	0,6092	0,6063	0,6094	0,6352	0,6387	0,6130
1:99	10	0,5000	0,5086	0,5084	0,5072	0,5087	0,5011	0,5021	0,5000	0,6130
	20	0,5000	0,5284	0,5280	0,5443	0,5288	0,5124	0,5293	0,5106	0,6130
	30	0,5000	0,5615	0,5659	0,5947	0,5686	0,5551	0,5753	0,5644	0,6130
	35	0,5000	0,5750	0,5824	0,6060	0,5819	0,5754	0,5968	0,5912	0,6130
	40	0,5000	0,5912	0,5947	0,6121	0,5981	0,5905	0,6177	0,6156	0,6130
	45	0,5000	0,6053	0,6044	0,6115	0,6051	0,6038	0,6287	0,6312	0,6130
	50	0,5000	0,6101	0,6105	0,6111	0,6119	0,6080	0,6337	0,6381	0,6130
10:90	20	0,5012	0,5312	0,5286	0,5232	0,5244	0,5165	0,5385	0,5193	0,6130
	30	0,5012	0,5587	0,5627	0,5881	0,5622	0,5566	0,5825	0,5645	0,6130
	35	0,5012	0,5766	0,5830	0,6065	0,5830	0,5744	0,6029	0,5945	0,6130
	40	0,5012	0,5891	0,5974	0,6151	0,6021	0,5932	0,6205	0,6187	0,6130
	45	0,5012	0,6030	0,6043	0,6147	0,6107	0,6032	0,6294	0,6322	0,6130
	50	0,5012	0,6106	0,6097	0,6119	0,6105	0,6051	0,6336	0,6396	0,6130
20:80	30	0,5124	0,5596	0,5615	0,5735	0,5557	0,5501	0,5819	0,5684	0,6130
	35	0,5124	0,5755	0,5757	0,5933	0,5714	0,5728	0,6021	0,5943	0,6130
	40	0,5124	0,5888	0,5868	0,6121	0,5951	0,5894	0,6153	0,6160	0,6130
	45	0,5124	0,6035	0,6017	0,6105	0,6043	0,5989	0,6272	0,6304	0,6130
	50	0,5124	0,6058	0,6055	0,6068	0,6059	0,6064	0,6316	0,6358	0,6130
30:70	35	0,5499	0,5740	0,5740	0,5841	0,5766	0,5726	0,6058	0,5970	0,6130
	40	0,5499	0,5922	0,5893	0,6056	0,5939	0,5899	0,6170	0,6155	0,6130
	45	0,5499	0,5979	0,6021	0,6075	0,6015	0,6008	0,6229	0,6277	0,6130
	50	0,5499	0,6040	0,6063	0,6043	0,6035	0,6016	0,6256	0,6318	0,6130
35:65	40	0,5728	0,5924	0,5943	0,6020	0,5913	0,5874	0,6208	0,6185	0,6130
	45	0,5728	0,6013	0,6011	0,6094	0,6043	0,5956	0,6247	0,6306	0,6130
	50	0,5728	0,6063	0,6043	0,6073	0,6049	0,6008	0,6282	0,6327	0,6130
40:60	45	0,5885	0,6033	0,6032	0,6050	0,6011	0,6005	0,6241	0,6272	0,6130
	50	0,5885	0,6046	0,6044	0,6036	0,6003	0,6010	0,6249	0,6304	0,6130

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 14'te verilmiştir.



Şekil 14. Zayıf ilişkili ve beş bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

4.2. Orta Düzey Korelasyona İlişkin Bulgular

Bu bölümde, değişkenler arası korelasyon katsayısının 0,6 olduğu 2, 3, 4 ve 5 bağımsız değişkenli türetilen toplum veri setlerinden örneklenen dengesiz veri setlerine ilişkin sonuçlar sırasıyla Tablo 7, 8, 9 ve 10'da yer almaktadır.

Tablo 7'de orta düzey ilişkili ve iki bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 7'de görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. UB ve RUSBoost algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, UB algoritması, bazı denge oranlarında parametre değerinden istatistiksel olarak anlamlı düzeyde yüksek sonuçlar üretmiştir.

Prevalans oranı %0,25 olan veri setinde, UB ile yapılan dengelemede (45:55) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50), RUSBoost ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ve MWMOTE ile yapılan dengelemelerde ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında, RUSBoost ve MWMOTE ile yapılan dengelemelerde ise (45:55) ve (50:50)

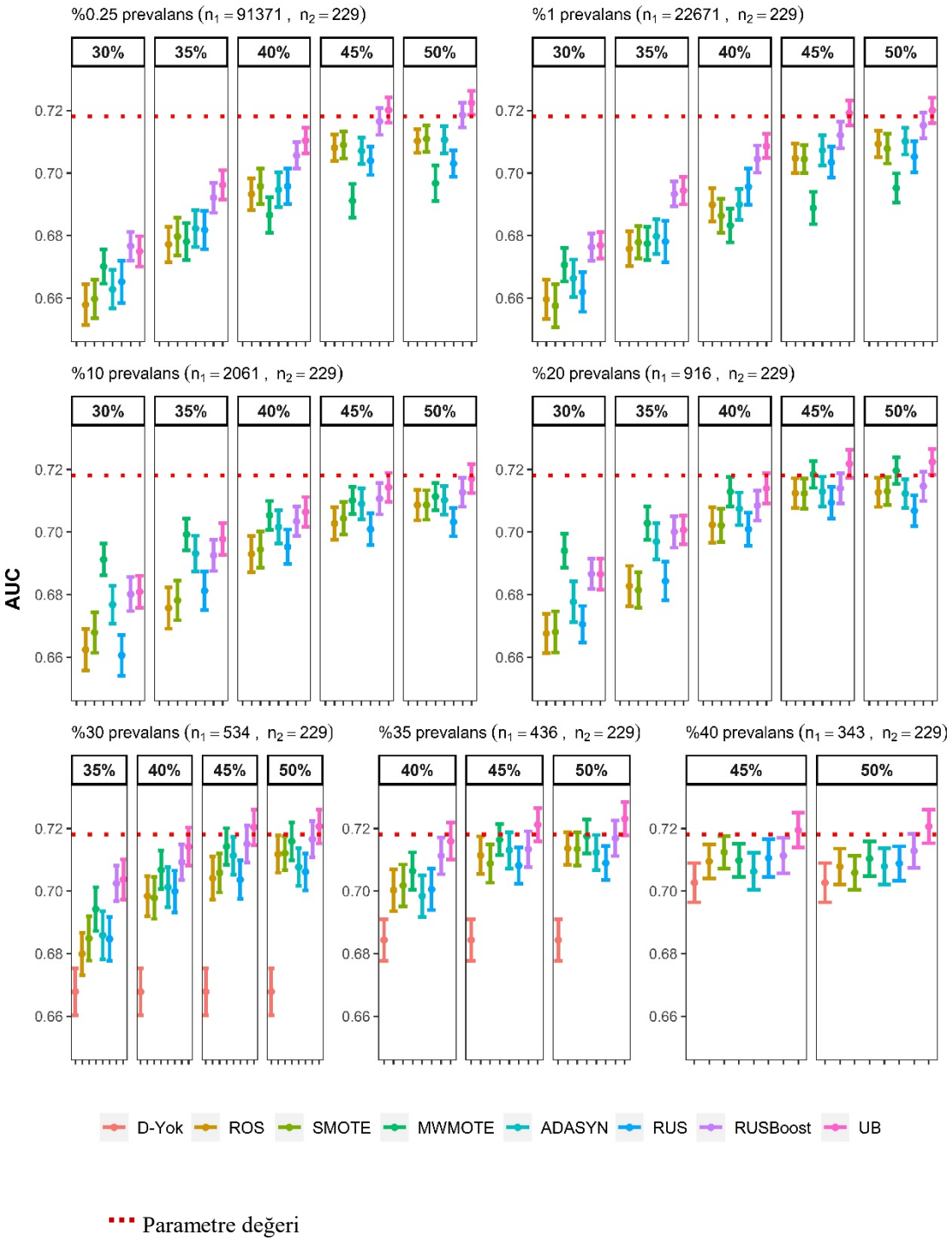
denge oranlarında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde, UB ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında, RUSBoost ve MWMOTE ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında, ADASYN ile yapılan dengelemede (45:55) denge oranında, ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Son olarak prevalans oranı %40 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50), RUSBoost ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 7. Orta düzey ilişkili ve iki bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5001	0,5637	0,5647	0,5895	0,5645	0,5650	0,5667	0,5680	0,7181
	20	0,5001	0,6195	0,6235	0,6438	0,6237	0,6189	0,6329	0,6292	0,7181
	30	0,5001	0,6579	0,6598	0,6701	0,6628	0,6653	0,6767	0,6750	0,7181
	35	0,5001	0,6772	0,6797	0,6781	0,6823	0,6818	0,6921	0,6962	0,7181
	40	0,5001	0,6933	0,6958	0,6866	0,6947	0,6958	0,7056	0,7104	0,7181
	45	0,5001	0,7081	0,7090	0,6911	0,7070	0,7039	0,7165	0,7202	0,7181
	50	0,5001	0,7103	0,7110	0,6967	0,7106	0,7030	0,7186	0,7225	0,7181
1:99	10	0,5011	0,5584	0,5572	0,5898	0,5621	0,5611	0,5697	0,5615	0,7181
	20	0,5011	0,6163	0,6224	0,6460	0,6218	0,6208	0,6339	0,6277	0,7181
	30	0,5011	0,6596	0,6576	0,6707	0,6664	0,6620	0,6764	0,6769	0,7181
	35	0,5011	0,6758	0,6779	0,6775	0,6798	0,6781	0,6933	0,6944	0,7181
	40	0,5011	0,6899	0,6863	0,6833	0,6899	0,6956	0,7045	0,7086	0,7181
	45	0,5011	0,7047	0,7045	0,6889	0,7072	0,7035	0,7121	0,7192	0,7181
	50	0,5011	0,7093	0,7078	0,6952	0,7102	0,7052	0,7152	0,7200	0,7181
10:90	20	0,5513	0,6175	0,6175	0,6464	0,6216	0,6191	0,6375	0,6301	0,7181
	30	0,5513	0,6624	0,6679	0,6912	0,6768	0,6606	0,6802	0,6809	0,7181
	35	0,5513	0,6757	0,6782	0,6992	0,6932	0,6812	0,6925	0,6978	0,7181
	40	0,5513	0,6930	0,6944	0,7053	0,7016	0,6953	0,7034	0,7064	0,7181
	45	0,5513	0,7027	0,7044	0,7101	0,7090	0,7009	0,7107	0,7143	0,7181
	50	0,5513	0,7086	0,7087	0,7114	0,7103	0,7032	0,7127	0,7171	0,7181
20:80	30	0,6297	0,6676	0,6681	0,6940	0,6777	0,6705	0,6866	0,6866	0,7181
	35	0,6297	0,6828	0,6815	0,7028	0,6970	0,6844	0,7001	0,7007	0,7181
	40	0,6297	0,7023	0,7021	0,7129	0,7075	0,7009	0,7085	0,7140	0,7181
	45	0,6297	0,7124	0,7123	0,7184	0,7130	0,7094	0,7140	0,7218	0,7181
	50	0,6297	0,7127	0,7130	0,7196	0,7123	0,7068	0,7147	0,7224	0,7181
30:70	35	0,6678	0,6799	0,6849	0,6942	0,6858	0,6847	0,7024	0,7037	0,7181
	40	0,6678	0,6984	0,6978	0,7068	0,7012	0,6999	0,7092	0,7142	0,7181
	45	0,6678	0,7041	0,7058	0,7142	0,7113	0,7036	0,7151	0,7204	0,7181
	50	0,6678	0,7118	0,7121	0,7159	0,7077	0,7062	0,7166	0,7207	0,7181
35:65	40	0,6843	0,7002	0,7018	0,7064	0,6984	0,7006	0,7113	0,7159	0,7181
	45	0,6843	0,7114	0,7088	0,7164	0,7131	0,7082	0,7134	0,7212	0,7181
	50	0,6843	0,7136	0,7135	0,7175	0,7123	0,7090	0,7169	0,7231	0,7181
40:60	45	0,7027	0,7079	0,7059	0,7098	0,7063	0,7088	0,7113	0,7195	0,7181
	50	0,7027	0,7095	0,7125	0,7104	0,7079	0,7105	0,7129	0,7207	0,7181

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 15'te verilmiştir.



Şekil 15. Orta düzey ilişkili ve iki bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 8'de orta düzey ilişkili ve üç bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 8'de görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, RUSBoost ve UB algoritmaları, bazı denge oranlarında parametre değerinden istatistiksel olarak anlamlı düzeyde yüksek sonuçlar üretmiştir. RUSBoost ve UB dışındaki algoritmalar ile dengelenen veri setleri için tam denge durumunda dahi parametre değerinden yüksek sonuçlar elde edilmemiş, genellikle toplum parametresine yakın fakat parametre değerinden düşük sonuçlar elde edilmiştir.

Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, UB ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, SMOTE ve ADASYN ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, UB ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p>0,05$). RUSBoost ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ve ADASYN ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, UB ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, SMOTE, MWMOTE ve ADASYN ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak

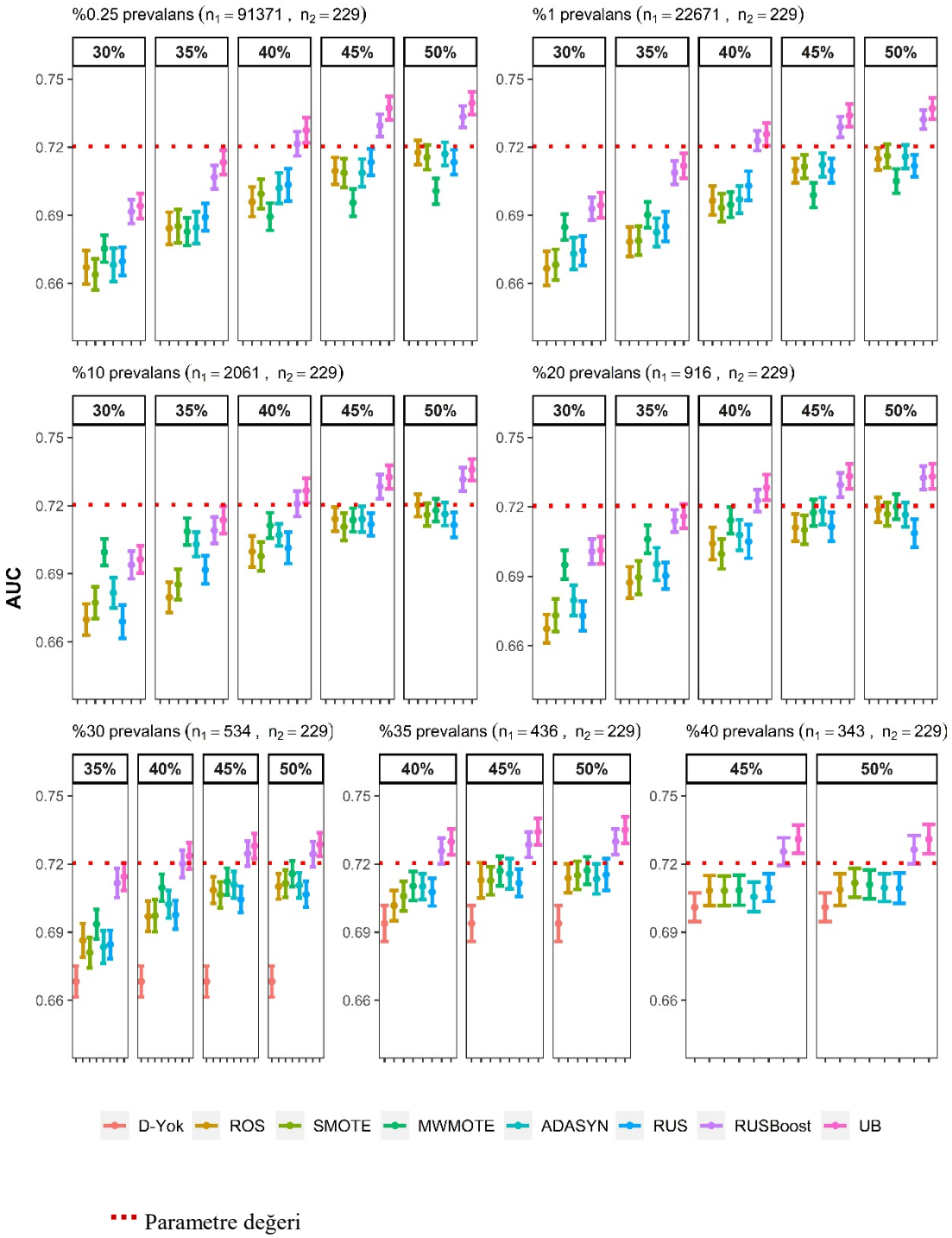
parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ile yapılan dengelemede (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ve ADASYN ile yapılan dengelemelerde (45:55) ve (50:50) denge oranında, ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROSBoost ile yapılan dengelemede (40:60), (45:55) ve (50:50), MWMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde, UB ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (45:55) ve (50:50), ADASYN ile yapılan dengelemede (45:55), ROS ve RUS ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Son olarak prevalans oranı %40 olan veri setinde, UB ile yapılan dengelemede (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu ($p<0,05$), RUSBoost ile yapılan dengelemede ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 8. Orta düzey ilişkili ve üç bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5005	0,5676	0,5689	0,6004	0,5704	0,5576	0,5791	0,5713	0,7204
	20	0,5005	0,6267	0,6289	0,6450	0,6309	0,6224	0,6465	0,6439	0,7204
	30	0,5005	0,6672	0,6639	0,6754	0,6682	0,6698	0,6916	0,6942	0,7204
	35	0,5005	0,6842	0,6852	0,6829	0,6846	0,6892	0,7069	0,7133	0,7204
	40	0,5005	0,6960	0,6995	0,6894	0,7020	0,7034	0,7215	0,7275	0,7204
	45	0,5005	0,7096	0,7087	0,6956	0,7087	0,7134	0,7296	0,7372	0,7204
	50	0,5005	0,7177	0,7156	0,7007	0,7170	0,7134	0,7334	0,7394	0,7204
1:99	10	0,5024	0,5699	0,5725	0,6076	0,5715	0,5658	0,5840	0,5753	0,7204
	20	0,5024	0,6272	0,6319	0,6540	0,6352	0,6270	0,6495	0,6457	0,7204
	30	0,5024	0,6666	0,6683	0,6848	0,6731	0,6745	0,6929	0,6945	0,7204
	35	0,5024	0,6784	0,6789	0,6902	0,6826	0,6851	0,7088	0,7118	0,7204
	40	0,5024	0,6965	0,6933	0,6947	0,6970	0,7031	0,7229	0,7257	0,7204
	45	0,5024	0,7098	0,7115	0,6989	0,7122	0,7097	0,7289	0,7340	0,7204
	50	0,5024	0,7148	0,7162	0,7051	0,7160	0,7117	0,7322	0,7371	0,7204
10:90	20	0,5657	0,6273	0,6299	0,6566	0,6331	0,6295	0,6552	0,6503	0,7204
	30	0,5657	0,6699	0,6772	0,6995	0,6816	0,6689	0,6939	0,6964	0,7204
	35	0,5657	0,6796	0,6853	0,7086	0,7030	0,6917	0,7091	0,7137	0,7204
	40	0,5657	0,6998	0,6977	0,7113	0,7071	0,7014	0,7208	0,7266	0,7204
	45	0,5657	0,7141	0,7107	0,7137	0,7141	0,7118	0,7284	0,7326	0,7204
	50	0,5657	0,7201	0,7160	0,7179	0,7163	0,7115	0,7316	0,7358	0,7204
20:80	30	0,6335	0,6673	0,6731	0,6949	0,6796	0,6729	0,7007	0,7013	0,7204
	35	0,6335	0,6873	0,6895	0,7060	06953	0,6903	0,7140	0,7160	0,7204
	40	0,6335	0,7042	0,6997	0,7142	0,7079	0,7050	0,7227	0,7284	0,7204
	45	0,6335	0,7111	0,7101	0,7175	0,7182	0,7114	0,7295	0,7333	0,7204
	50	0,6335	0,7188	0,7169	0,7201	0,7166	0,7086	0,7326	0,7332	0,7204
30:70	35	0,6683	0,6864	0,6811	0,6935	0,6836	0,6846	0,7117	0,7144	0,7204
	40	0,6683	0,6970	0,6973	0,7096	0,7023	0,6976	0,7201	0,7236	0,7204
	45	0,6683	0,7085	0,7065	0,7123	0,7109	0,7044	0,7246	0,7280	0,7204
	50	0,6683	0,7101	0,7114	0,7157	0,7107	0,7066	0,7246	0,7286	0,7204
35:65	40	0,6938	0,7017	0,7059	0,7103	0,7101	0,7077	0,7258	0,7298	0,7204
	45	0,6938	0,7128	0,7127	0,7169	0,7157	0,7117	0,7285	0,7342	0,7204
	50	0,6938	0,7138	0,7151	0,7172	0,7134	0,7153	0,7299	0,7350	0,7204
40:60	45	0,7010	0,7084	0,7083	0,7085	0,7056	0,7096	0,7255	0,7309	0,7204
	50	0,7010	0,7087	0,7117	0,7110	0,7096	0,7093	0,7263	0,7310	0,7204

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 16'da verilmiştir.



Şekil 16. Orta düzey ilişkili ve üç bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 9'da orta düzey ilişkili ve dört bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 9'da görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, RUSBoost ve UB algoritmaları, bazı denge oranlarında parametre değerinden istatistiksel olarak anlamlı düzeyde yüksek sonuçlar üretmiştir. RUSBoost ve UB dışındaki algoritmalar ile dengelenen veri setleri için tam denge durumunda dahi parametre değerinden yüksek sonuçlar elde edilmemiş, genellikle toplum parametresine yakın fakat parametre değerinden düşük sonuçlar elde edilmiştir.

Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, RUSBoost ve UB ile yapılan dengelemelerde (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, SMOTE ve ADASYN ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, RUSBoost ve UB ile yapılan dengelemelerde (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS ile yapılan dengelemede (45:55) ve (50:50), SMOTE, ADASYN ve RUS ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, RUSBoost ve UB ile yapılan dengelemelerde (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, SMOTE ve ADASYN ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ile yapılan dengelemede (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge

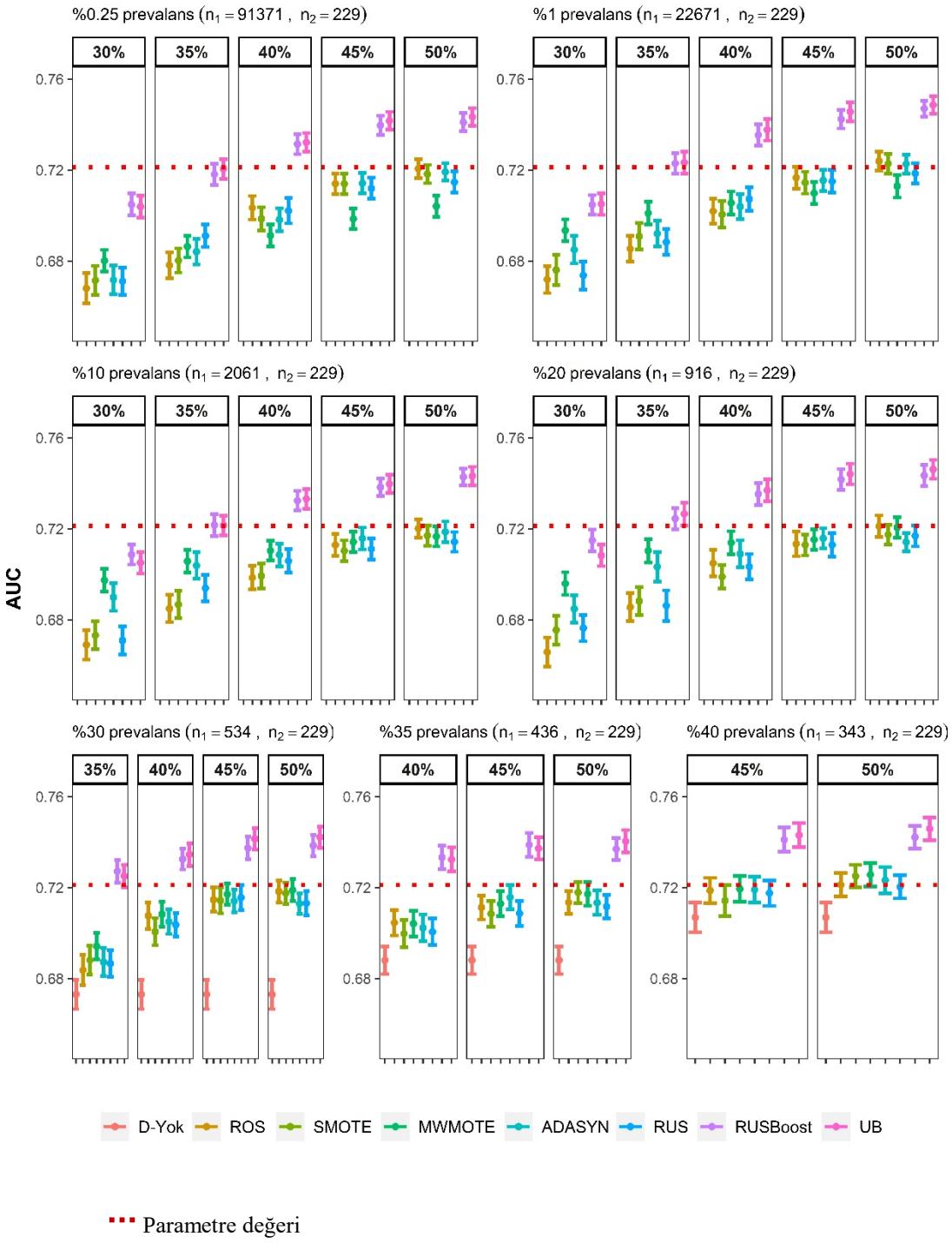
oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, SMOTE, MWMOTE ve RUS ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, RUSBoost ile yapılan dengelemede (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). UB ile yapılan dengelemede (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (45:55) ve (50:50), ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ve MWMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %40 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, MWMOTE, ADASYN ve RUS ile yapılan dengelemelerde (45:55) ve (50:50), SMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 9. Orta düzey ilişkili ve dört bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5004	0,5730	0,5734	0,6076	0,5765	0,5709	0,5936	0,5762	0,7212
	20	0,5004	0,6291	0,6320	0,6492	0,6323	0,6352	0,6587	0,6470	0,7212
	30	0,5004	0,6680	0,6715	0,6801	0,6717	0,6711	0,7049	0,7039	0,7212
	35	0,5004	0,6782	0,6803	0,6864	0,6841	0,6911	0,7181	0,7206	0,7212
	40	0,5004	0,7034	0,6986	0,6912	0,6982	0,7021	0,7314	0,7322	0,7212
	45	0,5004	0,7140	0,7140	0,6986	0,7142	0,7120	0,7397	0,7417	0,7212
	50	0,5004	0,7207	0,7183	0,7041	0,7192	0,7147	0,7410	0,7434	0,7212
1:99	10	0,5023	0,5775	0,5751	0,6166	0,5781	0,5744	0,6003	0,5792	0,7212
	20	0,5023	0,6321	0,6370	0,6623	0,6431	0,6432	0,6643	0,6528	0,7212
	30	0,5023	0,6719	0,6761	0,6935	0,6850	0,6737	0,7048	0,7051	0,7212
	35	0,5023	0,6854	0,6909	0,7011	0,6921	0,6884	0,7229	0,7233	0,7212
	40	0,5023	0,7019	0,7005	0,7056	0,7039	0,7073	0,7354	0,7378	0,7212
	45	0,5023	0,7166	0,7145	0,7099	0,7155	0,7151	0,7424	0,7457	0,7212
	50	0,5023	0,7239	0,7228	0,7129	0,7227	0,7185	0,7471	0,7486	0,7212
10:90	20	0,5756	0,6322	0,6336	0,6619	0,6421	0,6304	0,6691	0,6569	0,7212
	30	0,5756	0,6690	0,6732	0,6974	0,6900	0,6709	0,7087	0,7051	0,7212
	35	0,5756	0,6850	0,6867	0,7058	0,7039	0,6939	0,7216	0,7216	0,7212
	40	0,5756	0,6985	0,6993	0,7104	0,7085	0,7059	0,7324	0,7332	0,7212
	45	0,5756	0,7129	0,7104	0,7143	0,7158	0,7112	0,7383	0,7398	0,7212
	50	0,5756	0,7202	0,7170	0,7167	0,7187	0,7143	0,7428	0,7432	0,7212
20:80	30	0,6343	0,6658	0,6755	0,6959	0,6848	0,6764	0,7149	0,7083	0,7212
	35	0,6343	0,6856	0,6882	0,7104	0,7033	0,6862	0,7245	0,7266	0,7212
	40	0,6343	0,7049	0,6989	0,7139	0,7090	0,7034	0,7354	0,7370	0,7212
	45	0,6343	0,7134	0,7129	0,7153	0,7158	0,7129	0,7417	0,7441	0,7212
	50	0,6343	0,7212	0,7175	0,7207	0,7142	0,7169	0,7435	0,7462	0,7212
30:70	35	0,6731	0,6838	0,6881	0,6942	0,6873	0,6867	0,7272	0,7251	0,7212
	40	0,6731	0,7076	0,7006	0,7083	0,7050	0,7037	0,7325	0,7346	0,7212
	45	0,6731	0,7147	0,7144	0,7171	0,7141	0,7157	0,7374	0,7415	0,7212
	50	0,6731	0,7183	0,7175	0,7189	0,7129	0,7131	0,7385	0,7422	0,7212
35:65	40	0,6881	0,7046	0,6998	0,7042	0,7023	0,7006	0,7324	0,7324	0,7212
	45	0,6881	0,7113	0,7085	0,7129	0,7158	0,7087	0,7388	0,7373	0,7212
	50	0,6881	0,7135	0,7179	0,7172	0,7134	0,7117	0,7370	0,7404	0,7212
40:60	45	0,7069	0,7188	0,7142	0,7195	0,7191	0,7176	0,7411	0,7431	0,7212
	50	0,7069	0,7213	0,7251	0,7257	0,7233	0,7204	0,7422	0,7459	0,7212

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 17'de verilmiştir.



Şekil 17. Orta düzey ilişkili ve dört bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 10'da orta düzey ilişkili ve beş bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 10'da görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, RUSBoost ve UB algoritmaları, bazı denge oranlarında parametre değerinden istatistiksel olarak anlamlı düzeyde yüksek sonuçlar üretmiştir. RUSBoost ve UB dışındaki algoritmalar ile dengelenen veri setleri için tam denge durumunda dahi parametre değerinden yüksek sonuçlar elde edilmemiş, genellikle toplum parametresine yakın fakat parametre değerinden düşük sonuçlar elde edilmiştir.

Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, RUSBoost ve UB ile yapılan dengelemelerde (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, SMOTE ve ADASYN ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, RUSBoost ve UB ile yapılan dengelemelerde (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ve ADASYN ile yapılan dengelemelerde (45:55) ve (50:50), ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, RUSBoost ile yapılan dengelemede (30:70) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). UB ile yapılan dengelemede (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek

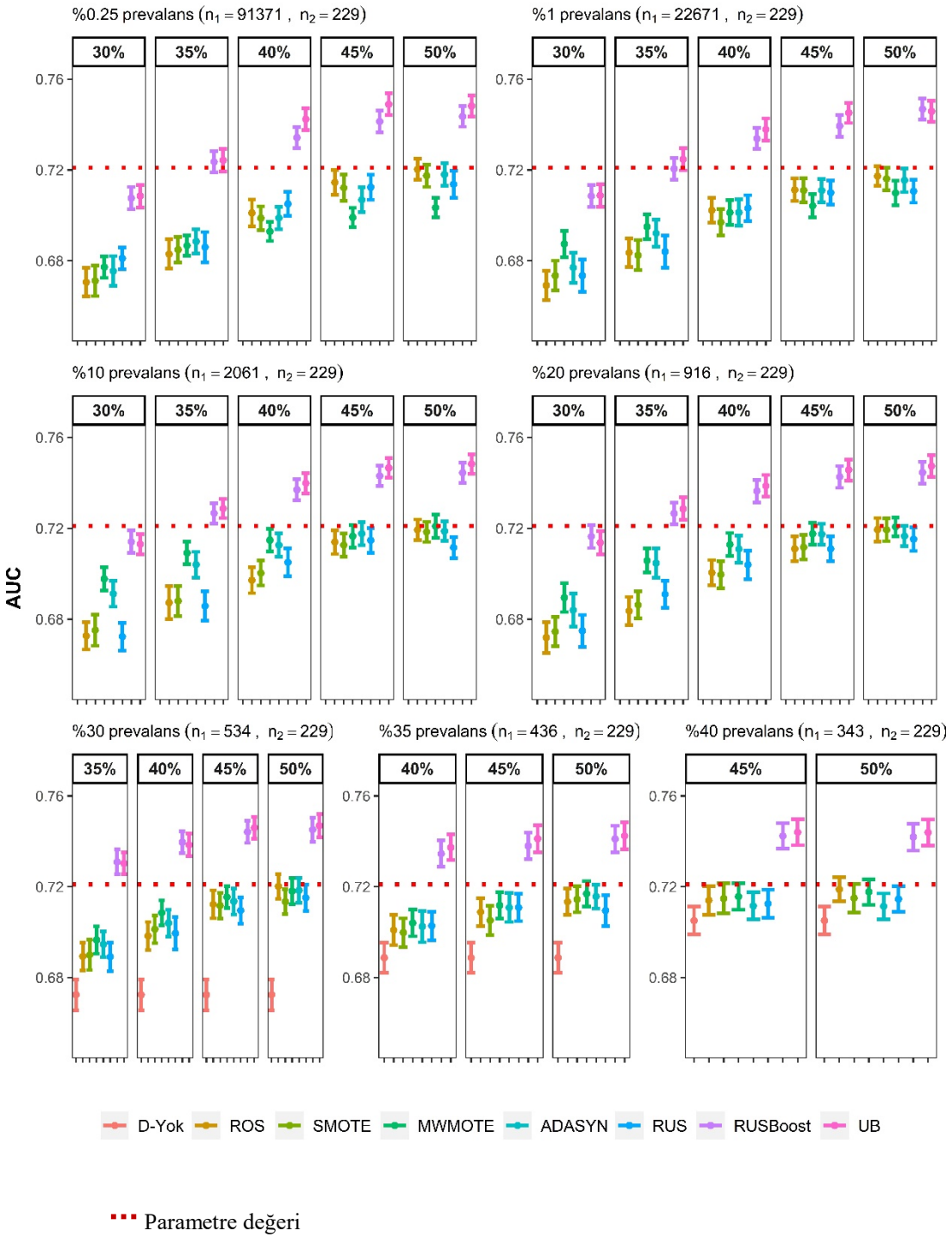
olduğu bulunmuştur ($p<0,05$). MWMOTE ve ADASYN ile yapılan dengelemelerde (45:55) ve (50:50), ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, MWMOTE, ADASYN ve RUS ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (50:50) denge oranında elde edilen AUC değerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Son olarak prevalans oranı %40 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ve MWMOTE ile yapılan dengelemelerde (45:55) ve (50:50), ROS ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerinin istatistiksel olarak parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 10. Orta düzey ilişkili ve beş bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5002	0,5801	0,5792	0,6062	0,5774	0,5676	0,5938	0,5823	0,7210
	20	0,5002	0,6293	0,6344	0,6433	0,6403	0,6253	0,6607	0,6536	0,7210
	30	0,5002	0,6706	0,6712	0,6772	0,6755	0,6811	0,7076	0,7085	0,7210
	35	0,5002	0,6830	0,6849	0,6867	0,6886	0,6860	0,7236	0,7243	0,7210
	40	0,5002	0,7011	0,6988	0,6929	0,6989	0,7050	0,7342	0,7424	0,7210
	45	0,5002	0,7145	0,7121	0,6991	0,7068	0,7124	0,7413	0,7489	0,7210
	50	0,5002	0,7203	0,7175	0,7034	0,7180	0,7137	0,7436	0,7490	0,7210
1:99	10	0,5028	0,5736	0,5770	0,6123	0,5809	0,5709	0,5992	0,5848	0,7210
	20	0,5028	0,6317	0,6383	0,6586	0,6362	0,6333	0,6622	0,6594	0,7210
	30	0,5028	0,6692	0,6735	0,6874	0,6770	0,6735	0,7085	0,7088	0,7210
	35	0,5028	0,6836	0,6825	0,6949	0,6921	0,6840	0,7205	0,7247	0,7210
	40	0,5028	0,7023	0,6969	0,7013	0,7013	0,7032	0,7338	0,7379	0,7210
	45	0,5028	0,7112	0,7110	0,7042	0,7110	0,7100	0,7394	0,7451	0,7210
	50	0,5028	0,7173	0,7161	0,7099	0,7155	0,7106	0,7468	0,7458	0,7210
10:90	20	0,5730	0,6329	0,6324	0,6608	0,6421	0,6372	0,6730	0,6626	0,7210
	30	0,5730	0,6727	0,6751	0,6978	0,6912	0,6724	0,7141	0,7131	0,7210
	35	0,5730	0,6873	0,6880	0,7092	0,7040	0,6858	0,7266	0,7287	0,7210
	40	0,5730	0,6971	0,7003	0,7148	0,7126	0,7051	0,7370	0,7399	0,7210
	45	0,5730	0,7140	0,7126	0,7165	0,7177	0,7147	0,7431	0,7466	0,7210
	50	0,5730	0,7194	0,7185	0,7209	0,7188	0,7115	0,7445	0,7484	0,7210
20:80	30	0,6321	0,6719	0,6745	0,6895	0,6840	0,6749	0,7164	0,7137	0,7210
	35	0,6321	0,6836	0,6863	0,7058	0,7047	0,6910	0,7265	0,7287	0,7210
	40	0,6321	0,7005	0,6996	0,7129	0,7109	0,7039	0,7364	0,7387	0,7210
	45	0,6321	0,7110	0,7117	0,7176	0,7174	0,7109	0,7427	0,7457	0,7210
	50	0,6321	0,7194	0,7193	0,7207	0,7166	0,7153	0,7446	0,7474	0,7210
30:70	35	0,6725	0,6894	0,6900	0,6966	0,6947	0,6891	0,7309	0,7303	0,7210
	40	0,6725	0,6983	0,7013	0,7085	0,7040	0,6995	0,7396	0,7384	0,7210
	45	0,6725	0,7123	0,7116	0,7152	0,7135	0,7095	0,7441	0,7460	0,7210
	50	0,6725	0,7202	0,7134	0,7181	0,7184	0,7155	0,7451	0,7469	0,7210
35:65	40	0,6887	0,7010	0,6998	0,7041	0,7024	0,7028	0,7345	0,7374	0,7210
	45	0,6887	0,7089	0,7052	0,7118	0,7108	0,7109	0,7380	0,7411	0,7210
	50	0,6887	0,7134	0,7145	0,7169	0,7153	0,7095	0,7410	0,7424	0,7210
40:60	45	0,7051	0,7140	0,7149	0,7157	0,7116	0,7126	0,7423	0,7440	0,7210
	50	0,7051	0,7189	0,7150	0,7177	0,7114	0,7145	0,7418	0,7439	0,7210

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 18'de verilmiştir.



Şekil 18. Orta düzey ilişkili ve beş bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

4.3. Yüksek Düzey Korelasyona İlişkin Bulgular

Bu bölümde, değişkenler arası korelasyon katsayısının 0,8 olduğu 2, 3, 4 ve 5 bağımsız değişkenli türetilen toplum veri setlerinden örneklenen dengesiz veri setlerine ilişkin sonuçlar sırasıyla Tablo 11, 12, 13 ve 14'te yer almaktadır.

Tablo 11'de yüksek ilişkili ve iki bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 11'de görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. UB ve RUSBoost algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, RUSBoost, UB ve ADASYN algoritmaları, bazı denge oranlarında parametre değerinden istatistiksel olarak anlamlı düzeyde yüksek sonuçlar üretmiştir.

Prevalans oranı %0,25 olan veri setinde, UB ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (40:60), (45:55) ve (50:50), ROS, SMOTE ve RUS için (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, UB ile yapılan dengelemede (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS ve SMOTE ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin

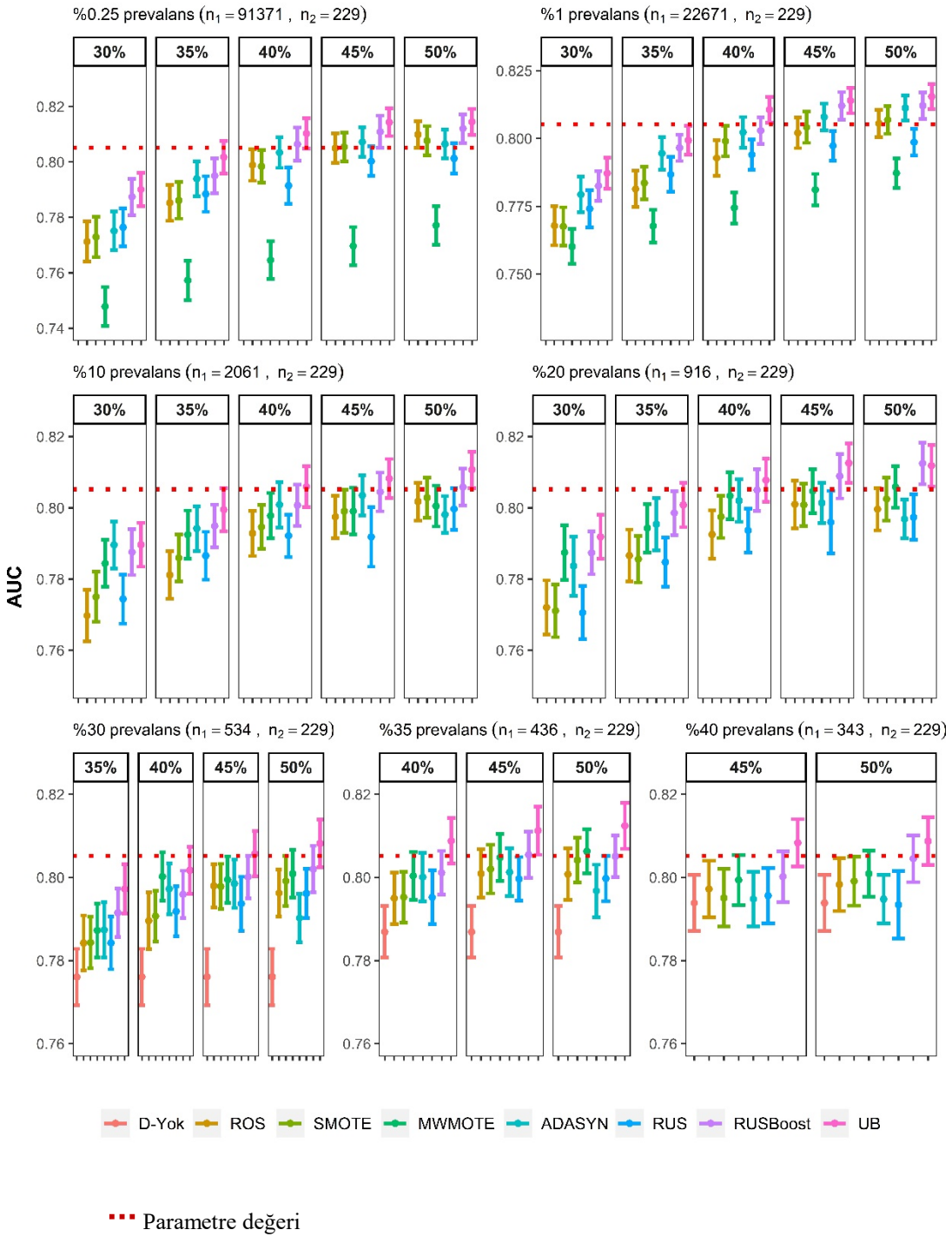
parametre değerinden farklı olmadığı bulunmuştur (p>0,05). Prevalans oranı %10 olan veri setinde, UB ile yapılan dengelemede (35:65), (40:60) ve (45:55) denge oranlarında elde edilen AUC değerinin parametre değerinden farklı olmadığı (p>0,05), (50:50) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur (p<0,05). RUSBoost ile yapılan dengelemede (40:60), (45:55) ve (50:50), ADASYN ile yapılan dengelemede (40:60) ve (45:55), SMOTE ve MWMOTE ile yapılan dengelemelerde (45:55) ve (50:50), ROS ve RUS ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur (p>0,05). Prevalans oranı %20 olan veri setinde, UB ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerinin parametre değerinden farklı olmadığı (p>0,05), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur (p<0,05). RUSBoost ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerinin parametre değerinden farklı olmadığı (p>0,05), (50:50) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur (p<0,05). ADASYN ile yapılan dengelemede (40:60) ve (45:55), MWMOTE ile yapılan dengelemede (40:60), (45:55) ve (50:50), ROS ve SMOTE ile yapılan dengelemelerde ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur (p>0,05). Prevalans oranı %30 olan veri setinde, UB ve MWMOTE ile yapılan dengelemelerde (40:60), (45:55) ve (50:50), RUSBoost ile yapılan dengelemede (45:55) ve (50:50), SMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur (p>0,05). Prevalans oranı %35 olan veri setinde, UB ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı (p>0,05), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur (p<0,05). RUSBoost ve MWMOTE ile yapılan dengelemelerde (40:60), (45:55) ve (50:50), ADASYN ile yapılan dengelemede (40:60) ve (45:55), ROS ve SMOTE ile yapılan dengelemelerde ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur (p>0,05). Prevalans oranı %40 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) ve (50:50), MWMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur (p>0,05). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur (p<0,05).

Tablo 11. Yüksek ilişkili ve iki bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5257	0,6692	0,6710	0,6958	0,6944	0,6815	0,6958	0,6886	0,8052
	20	0,5257	0,7295	0,7326	0,7275	0,7429	0,7397	0,7519	0,7534	0,8052
	30	0,5257	0,7713	0,7730	0,7479	0,7752	0,7764	0,7874	0,7901	0,8052
	35	0,5257	0,7853	0,7862	0,7573	0,7940	0,7885	0,7950	0,8018	0,8052
	40	0,5257	0,7989	0,7985	0,7646	0,8034	0,7915	0,8064	0,8103	0,8052
	45	0,5257	0,8050	0,8055	0,7697	0,8073	0,8003	0,8109	0,8144	0,8052
	50	0,5257	0,8099	0,8077	0,7772	0,8065	0,8013	0,8121	0,8145	0,8052
1:99	10	0,5585	0,6659	0,6691	0,7031	0,6885	0,6717	0,6924	0,6831	0,8052
	20	0,5585	0,7217	0,7193	0,7404	0,7445	0,7373	0,7479	0,7489	0,8052
	30	0,5585	0,7678	0,7676	0,7602	0,7794	0,7741	0,7825	0,7871	0,8052
	35	0,5585	0,7814	0,7836	0,7677	0,7945	0,7868	0,7966	0,7993	0,8052
	40	0,5585	0,7928	0,7991	0,7744	0,8023	0,7941	0,8029	0,8106	0,8052
	45	0,5585	0,8021	0,8042	0,7811	0,8080	0,7973	0,8121	0,8140	0,8052
	50	0,5585	0,8056	0,8069	0,7873	0,8113	0,7987	0,8121	0,8155	0,8052
10:90	20	0,6781	0,7291	0,7339	0,7708	0,7540	0,7391	0,7582	0,7558	0,8052
	30	0,6781	0,7698	0,7750	0,7844	0,7896	0,7744	0,7876	0,7897	0,8052
	35	0,6781	0,7812	0,7860	0,7925	0,7942	0,7866	0,7949	0,7995	0,8052
	40	0,6781	0,7929	0,7946	0,7978	0,8009	0,7922	0,8007	0,8059	0,8052
	45	0,6781	0,7975	0,7991	0,7991	0,8035	0,7919	0,8045	0,8083	0,8052
	50	0,6781	0,8018	0,8028	0,8005	0,7981	0,7997	0,8058	0,8107	0,8052
20:80	30	0,7074	0,7720	0,7711	0,7874	0,7837	0,7706	0,7874	0,7919	0,8052
	35	0,7074	0,7866	0,7856	0,7943	0,7954	0,7848	0,7986	0,8008	0,8052
	40	0,7074	0,7925	0,7975	0,8034	0,8020	0,7937	0,8050	0,8078	0,8052
	45	0,7074	0,8010	0,8008	0,8047	0,8014	0,7960	0,8089	0,8126	0,8052
	50	0,7074	0,7996	0,8025	0,8059	0,7969	0,7974	0,8125	0,8119	0,8052
30:70	35	0,7761	0,7842	0,7843	0,7873	0,7874	0,7842	0,7915	0,7972	0,8052
	40	0,7761	0,7896	0,7907	0,8002	0,7973	0,7919	0,7959	0,8017	0,8052
	45	0,7761	0,7980	0,7978	0,7995	0,7985	0,7937	0,8001	0,8057	0,8052
	50	0,7761	0,7962	0,7991	0,8009	0,7902	0,7962	0,8020	0,8082	0,8052
35:65	40	0,7869	0,7950	0,7953	0,8004	0,8001	0,7952	0,8012	0,8088	0,8052
	45	0,7869	0,8009	0,8020	0,8048	0,8013	0,7997	0,8055	0,8113	0,8052
	50	0,7869	0,8008	0,8042	0,8063	0,7968	0,7997	0,8051	0,8124	0,8052
40:60	45	0,7939	0,7972	0,7951	0,7994	0,7948	0,7956	0,8002	0,8083	0,8052
	50	0,7939	0,7983	0,7991	0,8009	0,7948	0,7935	0,8045	0,8087	0,8052

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 19'da verilmiştir.



Şekil 19. Yüksek ilişkili ve iki bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 12'de yüksek ilişkili ve üç bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 12'de görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, dengeleme algoritmaları, bazı denge oranlarında parametre değerinden anlamlı düzeyde yüksek sonuçlar üretmiştir.

Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, UB ile yapılan dengelemede (30:70) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (45:55) ve (50:50), ROS ve SMOTE ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (30:70) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ROS ve ADASYN ile yapılan dengelemelerde (45:55) ve (50:50), SMOTE ve RUS ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (30:70) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (35:65), (40:60), (45:55) ve (50:50), SMOTE, MWMOTE ve RUS ile yapılan dengelemelerde (45:55) ve (50:50), ROS ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (30:70) ve (35:65) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise

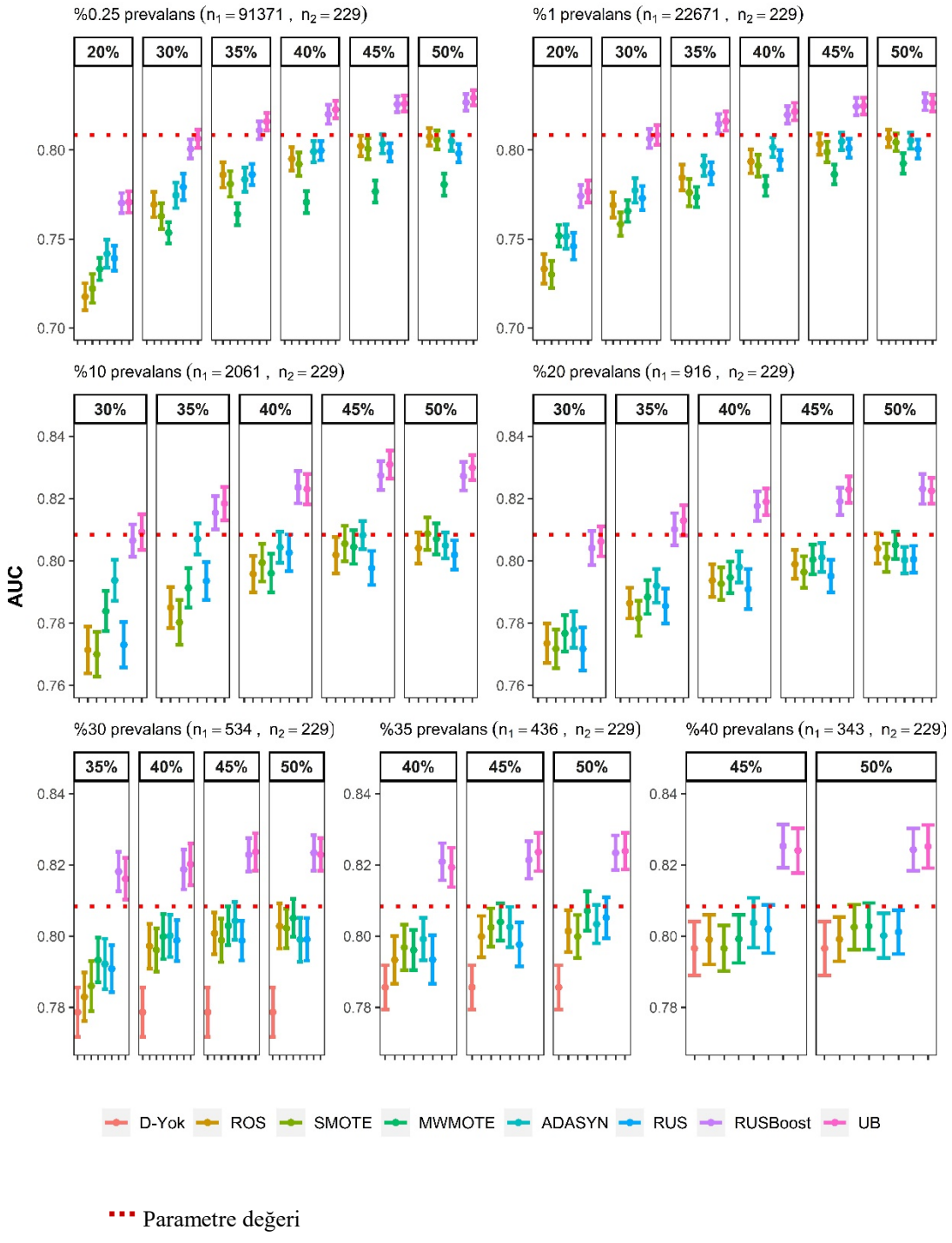
parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ROS ve MWMOTE ile yapılan dengelemelerde (50:50) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (45:55) ve (50:50), ADASYN ile yapılan dengelemede (45:55), ROS ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (45:55) ve (50:50), ADASYN ve RUS ile yapılan dengelemelerde ise (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %40 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). SMOTE, MWMOTE, ADASYN ve RUS ile yapılan dengelemelerde (50:50) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 12. Yüksek ilişkili ve üç bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5263	0,6711	0,6684	0,6918	0,6874	0,6755	0,7144	0,7085	0,8084
	20	0,5263	0,7176	0,7222	0,7332	0,7417	0,7392	0,7703	0,7708	0,8084
	30	0,5263	0,7694	0,7629	0,7536	0,7746	0,7792	0,8006	0,8063	0,8084
	35	0,5263	0,7860	0,7810	0,7640	0,7834	0,7862	0,8111	0,8161	0,8084
	40	0,5263	0,7951	0,7921	0,7707	0,7992	0,7997	0,8200	0,8227	0,8084
	45	0,5263	0,8023	0,8007	0,7767	0,8035	0,7988	0,8257	0,8261	0,8084
	50	0,5263	0,8074	0,8058	0,7807	0,8049	0,7982	0,8268	0,8294	0,8084
1:99	10	0,5639	0,6709	0,6760	0,7073	0,6978	0,6837	0,7174	0,7154	0,8084
	20	0,5639	0,7333	0,7301	0,7518	0,7514	0,7459	0,7742	0,7768	0,8084
	30	0,5639	0,7691	0,7584	0,7658	0,7773	0,7730	0,8064	0,8084	0,8084
	35	0,5639	0,7845	0,7761	0,7736	0,7911	0,7869	0,8147	0,8162	0,8084
	40	0,5639	0,7936	0,7913	0,7798	0,8015	0,7944	0,8197	0,8217	0,8084
	45	0,5639	0,8034	0,7990	0,7863	0,8048	0,8005	0,8245	0,8248	0,8084
	50	0,5639	0,8066	0,8043	0,7925	0,8052	0,8010	0,8271	0,8264	0,8084
10:90	20	0,6703	0,7312	0,7270	0,7520	0,7558	0,7321	0,7801	0,7765	0,8084
	30	0,6703	0,7714	0,7700	0,7839	0,7938	0,7731	0,8065	0,8093	0,8084
	35	0,6703	0,7850	0,7803	0,7914	0,8070	0,7936	0,8155	0,8184	0,8084
	40	0,6703	0,7958	0,7994	0,7961	0,8044	0,7977	0,8237	0,8230	0,8084
	45	0,6703	0,8019	0,8055	0,8045	0,8082	0,8026	0,8274	0,8309	0,8084
	50	0,6703	0,8041	0,8087	0,8070	0,8049	0,8019	0,8272	0,8299	0,8084
20:80	30	0,7362	0,7735	0,7718	0,7767	0,7779	0,7717	0,8041	0,8062	0,8084
	35	0,7362	0,7865	0,7816	0,7884	0,7920	0,7855	0,8101	0,8129	0,8084
	40	0,7362	0,7937	0,7927	0,7947	0,7980	0,7909	0,8176	0,8190	0,8084
	45	0,7362	0,7989	0,7964	0,8004	0,8011	0,7951	0,8191	0,8229	0,8084
	50	0,7362	0,8041	0,8010	0,8051	0,8002	0,8005	0,8231	0,8224	0,8084
30:70	35	0,7786	0,7829	0,7860	0,7933	0,7922	0,7908	0,8181	0,8161	0,8084
	40	0,7786	0,7972	0,7961	0,7999	0,8001	0,7988	0,8188	0,8202	0,8084
	45	0,7786	0,8008	0,7998	0,8030	0,8043	0,7987	0,8229	0,8237	0,8084
	50	0,7786	0,8028	0,8022	0,8051	0,7990	0,7991	0,8234	0,8229	0,8084
35:65	40	0,7856	0,7933	0,7968	0,7961	0,7992	0,7934	0,8209	0,8193	0,8084
	45	0,7856	0,7999	0,8024	0,8040	0,8026	0,7976	0,8214	0,8236	0,8084
	50	0,7856	0,8014	0,7999	0,8071	0,8034	0,8052	0,8234	0,8239	0,8084
40:60	45	0,7966	0,7990	0,7966	0,7992	0,8001	0,8012	0,8253	0,8241	0,8084
	50	0,7966	0,7991	0,8025	0,8028	0,8038	0,8020	0,8243	0,8252	0,8084

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 20'de verilmiştir.



Şekil 20. Yüksek ilişkili ve üç bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 13'te yüksek ilişkili ve dört bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 13'te görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, dengeleme algoritmaları, bazı denge oranlarında parametre değerinden anlamlı düzeyde yüksek sonuçlar üretmiştir.

Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, UB ile yapılan dengelemede (30:70) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (35:65) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ve ADASYN ile yapılan dengelemelerde (40:60), (45:55) ve (50:50), ROS ile yapılan dengelemede (45:55) ve (50:50), RUS ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %1 olan veri setinde, UB ile yapılan dengelemede (30:70), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUSBoost ile yapılan dengelemede (30:70) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS ile yapılan dengelemede (40:60) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUS ile yapılan

dengelemede (35:65), (40:60), (45:55) ve (50:50), MWMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (30:70) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (35:65), (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (35:65), (40:60), (45:55) ve (50:50), ROS ve RUS ile yapılan dengelemelerde ise (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (30:70) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (35:65), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (40:60) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). RUS ile yapılan dengelemede (40:60), (45:55) ve (50:50), SMOTE ile yapılan dengelemede ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (35:65), (40:60), (45:55) ve (50:50),

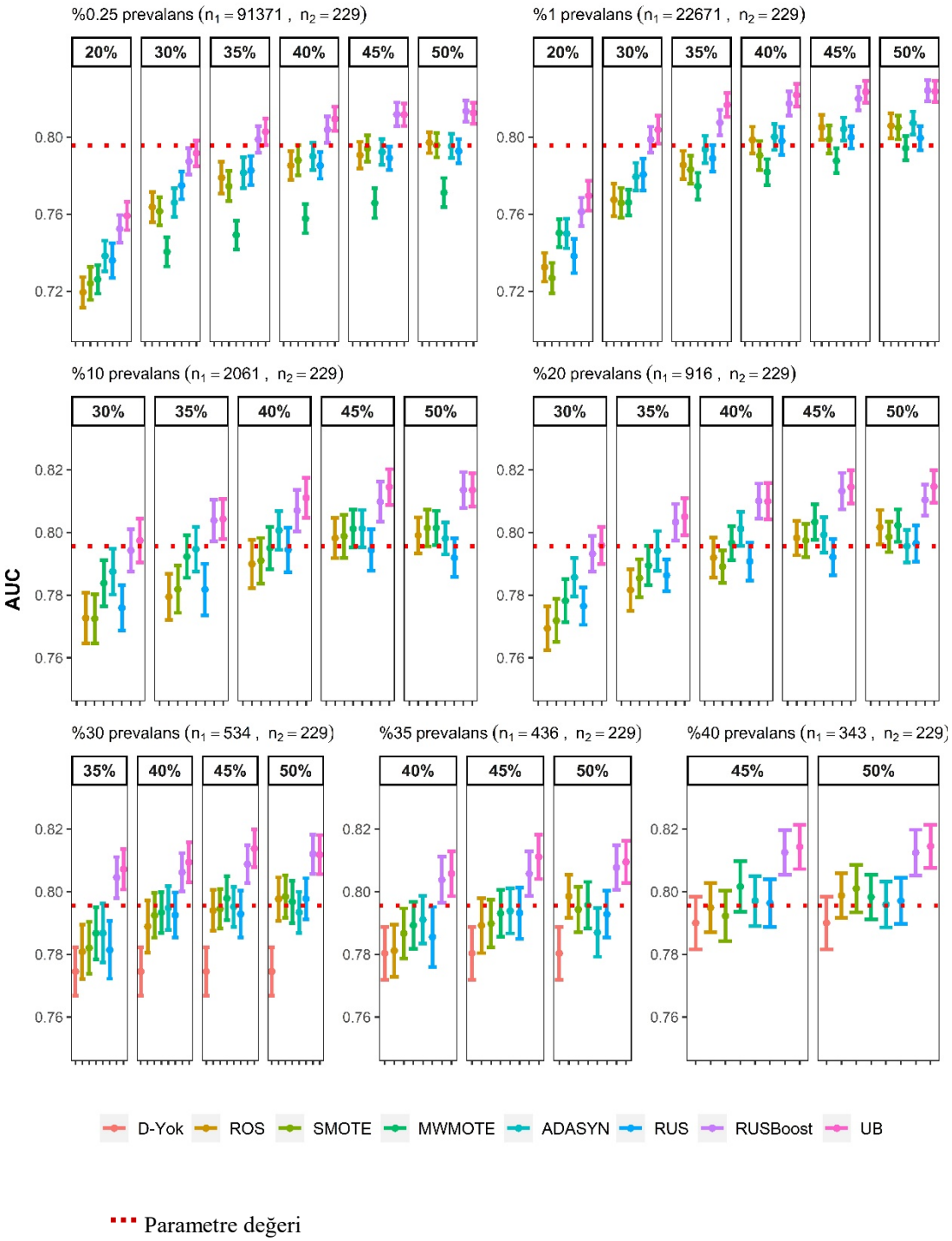
ROS, SMOTE, MWMOTE ve RUS ile yapılan dengelemelerde ise (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde UB ve RUSBoost ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (40:60), (45:55) ve (50:50), ADASYN ile yapılan dengelemede (40:60) ve (45:55), ROS, SMOTE ve RUS ile yapılan dengelemelerde ise (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Son olarak prevalans oranı %40 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). Dengelemenin yapılmadığı durumda ve diğer tüm algoritmalar ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 13. Yüksek ilişkili ve dört bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5264	0,6687	0,6681	0,6908	0,6854	0,6819	0,7030	0,7010	0,7956
	20	0,5264	0,7194	0,7241	0,7262	0,7383	0,7359	0,7523	0,7591	0,7956
	30	0,5264	0,7637	0,7615	0,7405	0,7660	0,7749	0,7873	0,7916	0,7956
	35	0,5264	0,7789	0,7746	0,7492	0,7815	0,7827	0,7987	0,8028	0,7956
	40	0,5264	0,7852	0,7880	0,7577	0,7900	0,7853	0,8039	0,8094	0,7956
	45	0,5264	0,7906	0,7941	0,7658	0,7922	0,7891	0,8118	0,8116	0,7956
	50	0,5264	0,7971	0,7958	0,7712	0,7954	0,7927	0,8134	0,8124	0,7956
1:99	10	0,5682	0,6773	0,6750	0,7055	0,6984	0,6835	0,7071	0,7031	0,7956
	20	0,5682	0,7325	0,7268	0,7502	0,7499	0,7382	0,7612	0,7695	0,7956
	30	0,5682	0,7674	0,7658	0,7660	0,7794	0,7806	0,7988	0,8038	0,7956
	35	0,5682	0,7855	0,7831	0,7745	0,7936	0,7889	0,8075	0,8167	0,7956
	40	0,5682	0,7985	0,7904	0,7818	0,8001	0,7980	0,8175	0,8218	0,7956
	45	0,5682	0,8051	0,7988	0,7877	0,8041	0,7999	0,8199	0,8235	0,7956
	50	0,5682	0,8058	0,8049	0,7942	0,8073	0,7995	0,8242	0,8237	0,7956
10:90	20	0,6784	0,7405	0,7298	0,7635	0,7548	0,7357	0,7701	0,7672	0,7956
	30	0,6784	0,7728	0,7725	0,7838	0,7875	0,7759	0,7942	0,7974	0,7956
	35	0,6784	0,7795	0,7819	0,7923	0,7947	0,7818	0,8038	0,8043	0,7956
	40	0,6784	0,7899	0,7910	0,7950	0,8007	0,7945	0,8070	0,8111	0,7956
	45	0,6784	0,7982	0,7988	0,8012	0,8013	0,7945	0,8098	0,8145	0,7956
	50	0,6784	0,7991	0,8015	0,8015	0,7981	0,7919	0,8136	0,8136	0,7956
20:80	30	0,7398	0,7694	0,7719	0,7782	0,7857	0,7765	0,7932	0,7959	0,7956
	35	0,7398	0,7816	0,7854	0,7894	0,7941	0,7863	0,8033	0,8050	0,7956
	40	0,7398	0,7919	0,7890	0,7966	0,8011	0,7907	0,8100	0,8099	0,7956
	45	0,7398	0,7982	0,7974	0,8033	0,7993	0,7922	0,8132	0,9145	0,7956
	50	0,7398	0,8017	0,7986	0,8022	0,7956	0,7965	0,8104	0,8147	0,7956
30:70	35	0,7745	0,7808	0,7821	0,7867	0,7868	0,7814	0,8045	0,8072	0,7956
	40	0,7745	0,7889	0,7925	0,7934	0,7948	0,7926	0,8062	0,8094	0,7956
	45	0,7745	0,7940	0,7945	0,7979	0,7952	0,7929	0,8088	0,8138	0,7956
	50	0,7745	0,7977	0,7984	0,7968	0,7934	0,7977	0,8120	0,8118	0,7956
35:65	40	0,7803	0,7813	0,7867	0,7893	0,7911	0,7855	0,8038	0,8058	0,7956
	45	0,7803	0,7892	0,7898	0,7931	0,7939	0,7932	0,8057	0,8112	0,7956
	50	0,7803	0,7985	0,7943	0,7957	0,7870	0,7928	0,8078	0,8095	0,7956
40:60	45	0,7900	0,7949	0,7923	0,8017	0,7971	0,7964	0,8126	0,8143	0,7956
	50	0,7900	0,7988	0,8010	0,7983	0,7960	0,7971	0,8125	0,8145	0,7956

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 21'de verilmiştir.



Şekil 21. Yüksek ilişkili ve dört bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

Tablo 14'te yüksek ilişkili ve beş bağımsız değişkenli toplum veri setine ilişkin bulgular verilmiştir. Tablo 14'te görüldüğü gibi dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde, sınıflandırma performansı kademeli olarak artmıştır. Bu artış, dengeleme oranının artmasıyla paraleldir ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaşmıştır. RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edilmiştir. Ayrıca, dengeleme algoritmaları, bazı denge oranlarında parametre değerinden anlamlı düzeyde yüksek sonuçlar üretmiştir.

Prevalans oranı %0,25 olan veri seti için karşılaştırma sonuçları incelendiğinde, UB ve RUSBoost ile yapılan dengelemelerde (30:70), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). RUS ile yapılan dengelemede (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). ROS ve SMOTE ile yapılan dengelemelerde (40:60) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). Prevalans oranı %1 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (30:70) denge oranında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ROS ve SMOTE ile yapılan dengelemelerde (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). RUS ile yapılan dengelemede (35:65), (40:60), (45:55) ve (50:50), MWMOTE ile yapılan dengelemede ise (50:50) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %10 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (30:70), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde

yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (30:70), (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ROS ve RUS ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %20 olan veri setinde, UB ile yapılan dengelemede (30:70), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROSBoost ile yapılan dengelemede (30:70) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (35:65), (40:60), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ve ADASYN ile yapılan dengelemelerde (35:65), (40:60), (45:55) ve (50:50), ROS ve RUS ile yapılan dengelemelerde ise (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %30 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). SMOTE ile yapılan dengelemede (35:65) ve (40:60) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) ve (50:50) denge oranlarında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (35:65), (40:60) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) denge oranında ise parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (35:65), (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden yüksek olduğu

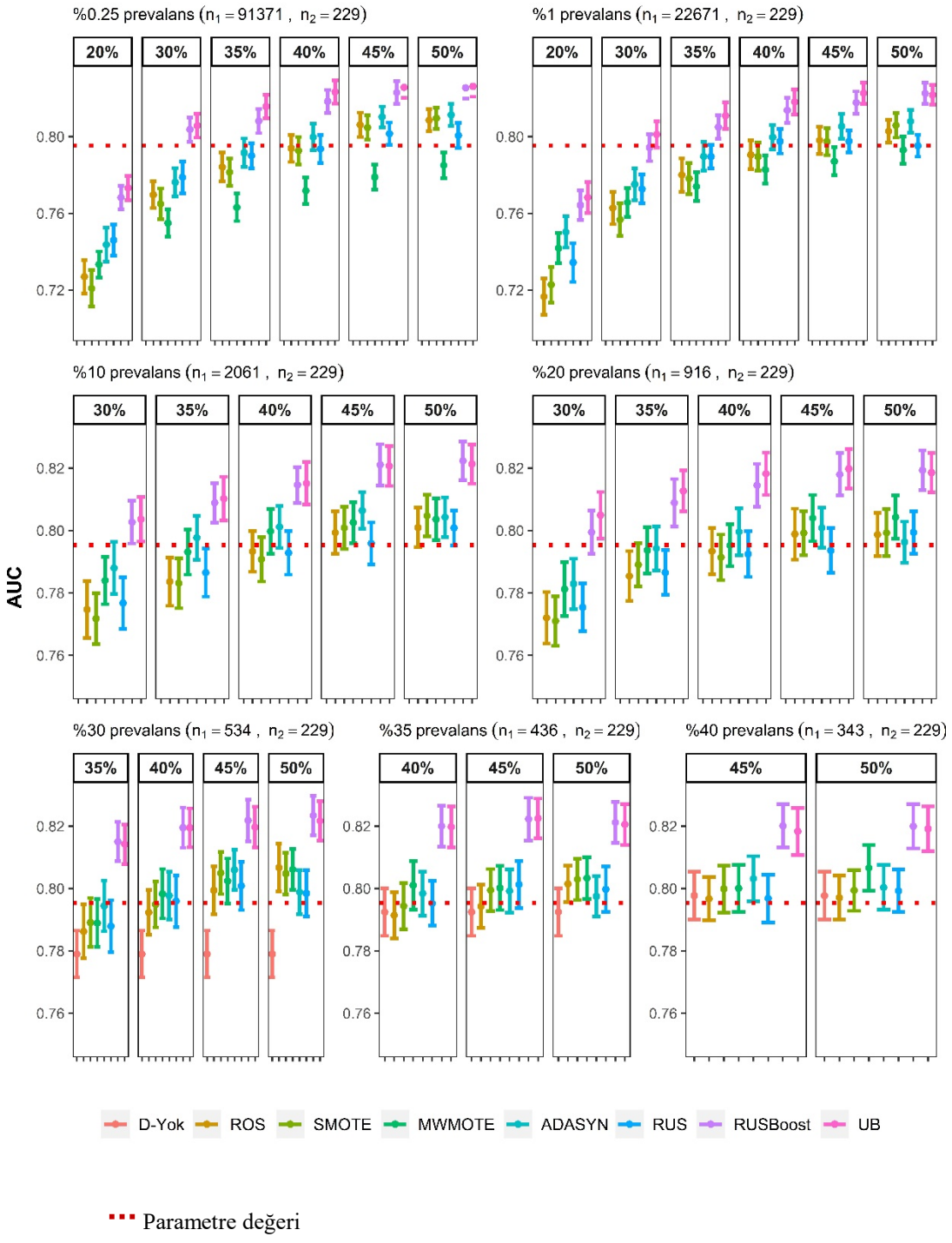
bulunmuştur ($p<0,05$). ROS ile yapılan dengelemede (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). RUS ile yapılan dengelemede ise (35:65), (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %35 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ROS, SMOTE ve MWMOTE ile yapılan dengelemelerde (40:60) ve (45:55) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ve RUS ile yapılan dengelemelerde (40:60), (45:55) ve (50:50) denge oranlarında ve dengeleme yapılmayan durumda elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Prevalans oranı %40 olan veri setinde, UB ve RUSBoost ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden anlamlı düzeyde yüksek olduğu bulunmuştur ($p<0,05$). ADASYN ile yapılan dengelemede (50:50) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (45:55) denge oranında ise parametre değerinden yüksek olduğu bulunmuştur ($p<0,05$). MWMOTE ile yapılan dengelemede (45:55) denge oranında elde edilen AUC değerinin parametre değerinden farklı olmadığı ($p>0,05$), (50:50) denge oranında ise parametre değerinden yüksek olduğu bulunmuştur ($p>0,05$). Dengelemenin yapılmadığı durumda ve diğer tüm algoritmalar ile yapılan dengelemelerde (45:55) ve (50:50) denge oranlarında elde edilen AUC değerlerinin parametre değerinden farklı olmadığı bulunmuştur ($p>0,05$). Diğer tüm durumlarda, tüm dengeleme algoritmaları için elde edilen AUC değerleri, parametre değerinden anlamlı düzeyde düşük bulunmuştur ($p<0,05$).

Tablo 14. Yüksek ilişkili ve beş bağımsız değişkenli toplum veri setine ilişkin gerçek ve tahmini AUC değerleri.

α	n_{az} (%)	D-Yok	ROS	SMOTE	MWMOTE	ADASYN	RUS	RUSBoost	UB	Parametre
0,25:99,75	10	0,5235	0,6733	0,6729	0,6952	0,6900	0,6811	0,7108	0,7104	0,7953
	20	0,5235	0,7270	0,7209	0,7334	0,7437	0,7460	0,7683	0,7732	0,7953
	30	0,5235	0,7697	0,7649	0,7550	0,7762	0,7788	0,8037	0,8057	0,7953
	35	0,5235	0,7842	0,7814	0,7631	0,7916	0,7900	0,8081	0,8157	0,7953
	40	0,5235	0,7938	0,7927	0,7718	0,7997	0,7935	0,8184	0,8232	0,7953
	45	0,5235	0,8061	0,8046	0,7789	0,8103	0,8006	0,8229	0,8257	0,7953
	50	0,5235	0,8085	0,8096	0,7850	0,8113	0,8015	0,8254	0,8262	0,7953
1:99	10	0,5617	0,6624	0,6667	0,7001	0,6921	0,6800	0,7083	0,7073	0,7953
	20	0,5617	0,7166	0,7228	0,7419	0,7503	0,7344	0,7643	0,7683	0,7953
	30	0,5617	0,7628	0,7567	0,7657	0,7751	0,7727	0,7943	0,8011	0,7953
	35	0,5617	0,7800	0,7781	0,7739	0,7897	0,7895	0,8049	0,8109	0,7953
	40	0,5617	0,7905	0,7896	0,7827	0,7997	0,7975	0,8137	0,8180	0,7953
	45	0,5617	0,7980	0,7974	0,7871	0,8053	0,7975	0,8177	0,8226	0,7953
	50	0,5617	0,8028	0,8059	0,7930	0,8080	0,7953	0,8225	0,8217	0,7953
10:90	20	0,6783	0,7378	0,7312	0,7609	0,7573	0,7423	0,7741	0,7790	0,7953
	30	0,6783	0,7747	0,7718	0,7840	0,7880	0,7768	0,8027	0,8036	0,7953
	35	0,6783	0,7836	0,7831	0,7931	0,7977	0,7865	0,8089	0,8102	0,7953
	40	0,6783	0,7933	0,7908	0,7997	0,8012	0,7929	0,8146	0,8151	0,7953
	45	0,6783	0,7993	0,8009	0,8026	0,8064	0,7959	0,8211	0,8207	0,7953
	50	0,6783	0,8010	0,8047	0,8036	0,8043	0,8008	0,8223	0,8213	0,7953
20:80	30	0,7374	0,7720	0,7710	0,7812	0,7829	0,7754	0,7995	0,8049	0,7953
	35	0,7374	0,7854	0,7891	0,7937	0,7943	0,7865	0,8089	0,8127	0,7953
	40	0,7374	0,7934	0,7914	0,7953	0,7996	0,7924	0,8145	0,8182	0,7953
	45	0,7374	0,7989	0,7992	0,8040	0,8009	0,7936	0,8180	0,8198	0,7953
	50	0,7374	0,7987	0,7993	0,8043	0,7963	0,7994	0,8193	0,8186	0,7953
30:70	35	0,7790	0,7863	0,7891	0,7889	0,7944	0,7879	0,8150	0,8142	0,7953
	40	0,7790	0,7923	0,7949	0,7983	0,7977	0,7960	0,8195	0,8195	0,7953
	45	0,7790	0,7994	0,8050	0,8024	0,8060	0,8008	0,8218	0,8197	0,7953
	50	0,7790	0,8067	0,8048	0,8061	0,7988	0,7985	0,8234	0,8217	0,7953
35:65	40	0,7925	0,7914	0,7944	0,8010	0,7984	0,7952	0,8200	0,8198	0,7953
	45	0,7925	0,7943	0,7995	0,8002	0,7992	0,8013	0,8223	0,8225	0,7953
	50	0,7925	0,8015	0,8029	0,8033	0,7975	0,7998	0,8212	0,8205	0,7953
40:60	45	0,7977	0,7967	0,7999	0,8001	0,8032	0,7968	0,8201	0,8183	0,7953
	50	0,7977	0,7971	0,7994	0,8066	0,8004	0,7992	0,8199	0,8192	0,7953

Tabloda yeşil hücreler parametre değerinden farklı olmayan, kırmızı hücreler parametre değerinden yüksek, diğer hücreler ise parametre değerinden küçük olan durumları ifade etmektedir.

Dengeleme algoritmalarının performanslarını görsel olarak inceleyebilmek için 100 tekrar sonucu elde edilen AUC değerlerinin güven aralığı grafikleri Şekil 22'de verilmiştir.



Şekil 22. Yüksek ilişkili ve beş bağımsız değişkenli veri setlerine ilişkin AUC değerlerinin güven aralığı grafiği.

5. TARTIŞMA

Son yıllarda dengesiz veri setlerinin analizi, hem teorik ve hem de pratik yönleriyle dikkate değer bir araştırma alanı haline gelmiştir. Sınıf dengesizliği problemine çözüm olarak çok sayıda veri dengeleme algoritması önerilmiş ve farklı dengeleme algoritmalarının performanslarının değerlendirildiği çalışmalar yapılmıştır. Çoğu çalışmada dengeleme algoritmalarının performansları sadece gerçek veri setleri üzerinden ya da gerçek veri setlerinden örneklenen dengesiz veri setleri üzerinden değerlendirilmiştir. Kullanılan performans ölçütüne göre en yüksek değere ulaşan yöntemlerin en başarılı yöntemler olduğu ifade edilmiştir. Önceki çalışmalara bakıldığında, en iyi performansı gösteren dengeleme algoritmalarının bir çalışmadan diğerine farklılık gösterdiği görülmektedir (Amin ve diğerleri, 2016; Barua ve diğerleri, 2012; Batista ve diğerleri, 2004; Bennin ve diğerleri, 2016; Chen ve diğerleri, 2010; He ve diğerleri, 2008; Rashu ve diğerleri, 2014; Van Hulse ve diğerleri, 2007). Bununla birlikte sınıf dengesizliği problemine çözüm olarak önerilen dengeleme algoritmalarının, sınıflandırma başarısını önemli ölçüde geliştirdiği yapılan birçok çalışmada ifade edilmiştir (Batista ve diğerleri, 2004; Galar ve diğerleri, 2011; Hasanin ve Khoshgoftaar, 2018; Japkowicz, 2000; Kamei ve diğerleri, 2007; López ve diğerleri, 2013; Tyagi ve Mittal, 2020; Van Hulse ve diğerleri, 2007). Çalışmamızda, önceki çalışmalara benzer şekilde, tüm simülasyon senaryolarında, dengeleme algoritmalarının, sınıflandırma başarısını artırdığı gözlendi. Bu artışın, dengeleme oranının artmasıyla paralel olduğu ve tüm dengeleme algoritmalarının en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaştığı gözlendi. Ayrıca, yapılan sınıflandırmalarda, en yüksek AUC değerleri, RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde elde edildi.

Araştırmacıların ilgisini çeken bir diğer konu ise sınıf değişkeni bakımından optimal bir denge oranının olup olmadığıdır. Genel olarak kabul edilmiş optimal bir sınıf dağılımı olmasa da, dengeli (50:50) bir dağılımın genellikle optimale yakın olduğu düşünülür (Weiss ve Provost, 2003). Ancak yapılan bazı çalışmalarda, dengeli bir dağılımın optimal azınlık-çoğunluk sınıf dağılımı olmadığı ifade edilmiştir. Khoshgoftaar ve diğerleri (2007) sınıflandırma performansının en yüksek olduğu optimal sınıf dağılımını belirlemeye çalışmıştır. Bu amaçla, 10 gerçek veri setindeki azınlık sınıflarından rastgele 5, 10, 20 ve 40 gözlem seçerek azınlık sınıflarını oluşturmuş daha sonra RUS

algoritması ile azınlık sınıfı gözlem sayısı oranları %1 ile %65 arasında değişen 14 dengesiz veri seti elde etmişlerdir. Dengesiz veri setlerini 11 farklı sınıflandırıcı kullanarak sınıflandırmışlar ve sonuç olarak, 56 dengesiz veri seti üzerinde yaptıkları uygulamalar sonucu optimal azınlık-çoğunluk sınıf dağılımının yaklaşık olarak 35:65 olduğunu ifade etmişlerdir. Weiss ve Provost (2003) C4.5 algoritmasını kullanarak sınıf dağılımı konusunu ele almış ve ideal dağılımın genellikle kullanılan performans ölçütüne bağlı olduğunu söylemiştir. Genel sınıflandırma doğruluğu göz önüne alındığında doğal sınıf dağılımının, AUC kullanıldığında ise dengeli bir dağılımın en iyi performansı gösterme eğiliminde olduğunu ifade etmişlerdir. Ayrıca, optimal sınıf dağılımın veri setinden veri setine farklılık gösterdiğini söylemişlerdir. Diğer bir çalışmada, Khoshgoftaar ve diğerleri (2010) dört gerçek veri seti ve dört sınıflandırıcı kullanarak, boosting ve bagging tabanlı dört dengeleme algoritmasının performanslarını karşılaştırmıştır. Dengeleme algoritmaları ile azınlık-çoğunluk sınıf dağılımını 35:65 ve 50:50 olacak şekilde dengelemiş ve dengelenen veri setlerini sınıflandırmışlardır. Sonuç olarak, azınlık-çoğunluk sınıf dağılımının 35:65 olacak şekilde dengelenmesi durumda çoğunlukla daha yüksek sınıflandırma performansı elde ettiklerini bildirmişlerdir. Albisua ve diğerleri (2013) ROS, SMOTE ve RUS dahil sekiz dengeleme algoritması için optimal dengeleme oranlarını, veri setleri tam dengeli durumda iken sınıflandırılmasından elde edilen AUC değerlerini referans alarak belirlemeye çalışmıştır. Çalışmalarında, 29 gerçek veri seti ile iki farklı sınıflandırıcı kullanmış ve en yüksek AUC değerinin elde edildiği denge oranını optimal olarak değerlendirmişlerdir. Sonuç olarak, her veri setinin kendine ait bir optimal sınıf dağılımı olduğunu ve kullanılan dengeleme algoritmasına bağlı olarak optimal sınıf dağılımının genellikle dengeli bir sınıf dağılımı olmadığını ifade etmişlerdir. Optimal denge oranının araştırıldığı önceki çalışmalarda, optimal azınlık-çoğunluk sınıfı denge oranı, gerçek veri setlerinden RUS algoritması ile örneklenen dengesiz veri setleri üzerinden belirlenmeye çalışılmıştır. Dengeleme algoritmaları için optimal azınlık-çoğunluk sınıfı denge oranlarının incelendiği kapsamlı bir simülasyon çalışması daha önce yapılmamıştır.

Örneklem, toplumun sadece küçük bir parçasıdır ve örneklemden hesaplanan istatistikler toplum parametrelerine ilişkin yapılan kestirimlerdir. Bu kestirimlerin belirli güven aralıkları içerisinde toplum parametresine yakın değerler olması beklenir. Kestirimlerin, sistematik olarak parametre değerinden küçük ya da büyük olması topluma ilişkin yapılan genellemelerin hatalı ve yanlı olmasına neden olur. Yansız istatistiksel kestirimler ve topluma ilişkin genellemeler yapılabilmesi, örnek veri setinin ait olduğu toplumu yeterli düzeyde temsil etmesine bağlıdır.

Gerçek veri setleri ya da bu veri setlerinden örneklenen dengesiz veri setlerinin kullanıldığı çalışmalarda, ilgili veri setinin hem ait olduğu toplumun parametreleri hem de toplumu yeterli sayı ve nitelikte temsil edip etmediği bilinmemektedir. Bu durumda yapılan kestirimlerin, toplum parametresine ne ölçüde yakınsadığı, istatistiksel olarak parametre değerinden farklı olup olmadıkları kontrol edilememektedir. Bu nedenle, yalnızca gerçek veri setleri üzerinden yapılan sınıflandırma sonuçları baz alınarak dengeleme algoritmalarının performanslarının değerlendirilmesi ve algoritmalara ilişkin genellemeler yapılması doğru bir yaklaşım değildir. Bu bağlamda dengeleme algoritmalarının performansları hakkında genellemeler yapabilmek kolay değildir. Çalışmamızda, farklı dengeleme algoritmaları için optimal denge oranları, kapsamlı bir simülasyon çalışması ışığında incelenmiştir

Önceki çalışmalardan farklı olarak çalışmamızda, örneklemden kaynaklanabilecek yanlılığı en aza indirmek için dengesiz veri setlerinin birim sayıları, prevalans, duyarlılık ve özgüllük oranları dikkate alınarak, %10 etki büyüklüğü, 0,80 güç ve 1.tip hata payı 0,05 olacak şekilde gerekli olan minimum birim sayıları hesaplanarak belirlendi. Ayrıca yapılan sınıflandırmaların tamamında 10 kat çapraz geçerlilik kullanıldı ve tüm senaryolar için 100 bağımsız tekrar gerçekleştirildi. Çalışmamızda, dengeleme algoritmalarının performanslarının değerlendirilmesi önceki çalışmalardan farklı şekilde ele alındı. Dengeleme algoritmalarına ilişkin performans değerlendirmeleri, türetilen toplum veri setlerinin sınıflandırılmasından elde edilen AUC değerleri (parametre) referans alınarak yapıldı. Dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinin sınıflandırılmasından elde edilen AUC değerleri, istatistiksel olarak, toplum veri setlerinden hesaplanan AUC değerleri ile karşılaştırıldı. Böylece, dengeleme algoritmaları ile dengelenen veri setleri üzerinden yapılan kestirimlerin hangi denge oranlarında istatistiksel olarak parametre değerlerine eşit, hangi denge oranlarında parametre değerlerinden küçük ya da büyük olduğu belirlendi. İstatistiksel olarak parametre değerlerinden farklı olmayan denge oranları, dengeleme algoritmaları için optimal denge oranları olarak değerlendirildi.

Çalışmamızın önceki çalışmalardan farkı, kurgulanan simülasyon senaryoları sayesinde; farklı ilişki yapıları ve değişken sayılarına göre türetilen toplum veri setlerinden parametreler hesaplanmış ve dengesiz veri setleri ait oldukları toplumu yeterli sayı ve nitelikte temsil edecek şekilde oluşturulmuştur. Ayrıca, dengeleme algoritmaları ile kademeli olarak dengelenen veri setleri üzerinden yapılan kestirimlerin parametre değerlerine nasıl ve ne ölçüde yakınsadığı, istatistiksel olarak parametre değerinden farklı olup olmadıkları belirlenebilmiştir. Böylece,

dengeleme algoritmaları için optimal azınlık-çoğunluk sınıfı denge oranları incelenebilmiştir. Çalışmamızda, türetilen toplum veri setlerinden hesaplanan AUC değerleri referans alınarak değerlendirilen optimal azınlık-çoğunluk sınıfı denge oranları, kullanılan dengeleme algoritmalarına bağlı olarak farklılık göstermiştir. Bununla birlikte, değişkenler arasındaki korelasyon yapısı, bağımsız değişken sayısı ve azınlık sınıfı prevalans oranları da dengeleme algoritmaları için optimal azınlık-çoğunluk sınıfı denge oranlarını etkilemiştir. Değişkenler arasındaki ilişki düzeyinin ve bağımsız değişken sayısının artışına paralel olarak dengeleme algoritmaları ile dengelenen veri setlerinin sınıflandırılmasından elde edilen AUC değerlerinin toplum veri setlerinden elde edilen AUC değerlerine yakınsama oranı artmıştır.

6. SONUÇ VE ÖNERİLER

Çalışmamızda, önceki çalışmalardan farklı olarak dengesizlik oranları, korelasyon yapıları ve bağımsız değişken sayıları göz önünde bulundurularak toplum veri setleri türetildi ve türetilen toplum veri setlerinden gerekli olan minimum örneklem hacmi belirlenerek dengesiz veri setleri örneklendi. Dengesiz veri setleri, dört farklı aşırı örnekleme (ROS, SMOTE, MWMOTE ve ADASYN) ve üç farklı alt örnekleme (RUS, RUSBoost ve UB) algoritması ile kademeli olarak dengelendi ve her bir kademede CART kullanılarak türetilen toplum veri setlerinden hesaplanan AUC değerleri tahmin edildi. Böylece, farklı senaryolar altında, dengeleme algoritmalarının hem toplum parametresini tahmin performansları hem de toplum parametreleri referans alınarak optimal azınlık-çoğunluk sınıfı denge oranları incelendi.

Sonuç olarak, tüm simülasyon senaryolarında dengeleme algoritmaları ile kademeli olarak dengelenen veri setlerinde sınıflandırma performansı kademeli olarak arttı. Bu artış, dengeleme oranının artmasıyla paraleldi ve tüm dengeleme algoritmaları en yüksek AUC değerine genellikle tam denge (50:50) durumda ulaştı. Sonuçlar ilişki yapısına göre değerlendirildiğinde; değişkenler arasındaki ilişki düzeyinin artmasına paralel olarak dengeleme algoritmaları ile dengelenen veri setlerinde sınıflandırma performansının arttığı gözlendi. Benzer şekilde, tüm ilişki yapılarında, bağımsız değişken sayısının artması sınıflandırma performansının artmasını sağladı. Topluluk öğrenme tabanlı alt örnekleme algoritmaları olan RUSBoost ve UB algoritmaları ile dengelenen veri setlerinde diğer yöntemlere kıyasla daha yüksek AUC değerleri elde edildi.

Dengeleme algoritmaları optimal dengeleme oranları bakımından incelendiğinde, RUSBoost ve UB algoritmalarının simülasyon senaryolarının çoğunda belirli denge oranlarından sonra parametre değerinden istatistiksel olarak yüksek sonuçlar ürettiği gözlendi. Hem ilişki düzeyindeki hem de bağımsız değişken sayısındaki artış RUSBoost ve UB algoritmalarının parametre değerinden yüksek sonuçlar üretme eğilimini artırdı. Zayıf ilişkili olan iki, üç, dört ve beş bağımsız değişkenli simülasyon senaryolarında, başlangıç azınlık:çoğunluk sınıf dağılımı dikkate alınarak, RUSBoost için optimal azınlık sınıfı gözlem oranları yaklaşık olarak, sırasıyla, %45-%50, %45-%50, %40-%45 ve %35-%40 aralığında, UB için sırasıyla, %45-%50, %40-%45, %40-%45 ve %35-%45 aralığında bulundu. Orta düzey ilişkili olan iki, üç, dört ve beş bağımsız değişkenli

simülasyon senaryolarında, başlangıç azınlık:çoğunluk sınıf dağılımı dikkate alınarak, RUSBoost için optimal azınlık sınıfı gözlem oranları yaklaşık olarak, sırasıyla, %45-%50, %40-%45, %35-%40 ve %30-%35 aralığında, UB için sırasıyla, %40-%50, %35-%40, %30-%35 ve %30-%35 aralığında bulundu. Yüksek ilişkili olan iki, üç, dört ve beş bağımsız değişkenli simülasyon senaryolarında, başlangıç azınlık:çoğunluk sınıf dağılımı dikkate alınarak, RUSBoost için optimal azınlık sınıfı gözlem oranları yaklaşık olarak, sırasıyla, %40-%50, %30-%35, %30-%35 ve %20-%30 aralığında, UB için sırasıyla, %35-%45, %30-%35, %30-%35 ve %20-%30 aralığında bulundu. ROS, SMOTE, MWMOTE, ADASYN ve RUS algoritmalarının, zayıf ve orta düzey ilişkili tüm simülasyon senaryoları ile yüksek ilişkili olan iki ve üç bağımsız değişkenli simülasyon senaryolarında, parametre değerinden istatistiksel olarak yüksek sonuçlar üretmediği ve optimal azınlık sınıfı gözlem oranlarının yaklaşık olarak %45-%50 aralığında olduğu gözlendi. Bununla birlikte, yüksek ilişkili olan dört ve beş bağımsız değişkenli bazı simülasyon senaryolarında ise RUS dışındaki algoritmaların bazı denge oranlarında parametre değerinden istatistiksel olarak yüksek sonuçlar ürettiği ve bu algoritmalar için optimal azınlık sınıfı gözlem oranının yaklaşık olarak %40 olduğu gözlendi.

Çalışmamızda elde edilen bulgular ışığında,

- Dengeleme algoritmaları ile dengelenen veri setlerinde sınıflandırma performansı önemli ölçüde artış göstermektedir.
- Dengeleme algoritmaları ile dengelenen veri setlerinde yapılan sınıflandırmalar sonucu elde edilen AUC değerleri, ilişki düzeyinin ve bağımsız değişken sayısının artışına paralel olarak artış göstermektedir.
- AUC ölçütü baz alındığında, RUSBoost ve UB algoritmalarının sınıflandırma performansına yaptıkları katkı ROS, SMOTE, MWMOTE, ADASYN ve RUS algoritmalarından fazladır.
- Değişkenler arasında yüksek ilişki ve bağımsız değişken sayısı ikiden fazla olan veri setlerinde azınlık sınıfı gözlem oranı %30 ya da %30'dan fazla ise RUSBoost ve UB algoritmaları ile dengelenen veri setlerinin sınıflandırılması sonucu elde edilen AUC değerleri sistematik olarak toplum AUC değerinden yüksek, yanlı sonuçlar elde edilmesine neden olur. Bu durum, orta düzey ilişkili ve bağımsız değişken sayısı üçten fazla olan veri setlerinde de geçerlidir.

- Bağımsız değişken sayısı ikiden fazla ve azınlık sınıfı gözlem oranı %30'dan az olduğu durumlarda, değişkenler arasındaki ilişki düzeyi yüksek ise azınlık sınıfı gözlem oranı en çok %30, ilişki düzeyi orta ise en çok %40 olacak şekilde dengeleme yapılmalıdır.
- Değişkenler arasında zayıf ve orta düzey ilişki bulunan veri setlerinde, ROS, SMOTE, MWMOTE, ADASYN ve RUS ile dengelenen veri setlerinin sınıflandırılması sonucu elde edilen AUC değerleri, toplum AUC değerlerinden istatistiksel olarak yüksek bulunmamıştır. Genellikle 45:55 ya da 50:50 denge oranlarında elde edilen AUC değerlerinin toplum AUC değerinden istatistiksel olarak farklı olmadığı gözlenmiştir. Bu nedenle ROS, SMOTE, MWMOTE, ADASYN ve RUS algoritmalarında denge oranının 50:50 alınmasında sakınca yoktur. Bu durum yüksek ilişkili olan iki ve üç bağımsız değişkenli veri setlerinde de geçerlidir.
- RUS algoritması, hiçbir simülasyon senaryosunda parametre değerinden istatistiksel olarak yüksek sonuçlar üretmemiştir. Bu nedenle, RUS algoritması ile yapılan dengelemelerin ilişki yapısı ve değişken sayısından bağımsız olarak (50:50) olmasında sakınca yoktur.

KAYNAKLAR

Acharya, U. R., Chowriappa, P., Fujita, H., Bhat, S., Dua, S., Koh, J. E., . . . Ng, K. H. (2016). Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images. *Knowledge-Based Systems, 107*, 235-245.

Albisua, I., Arbelaitz, O., Gurrutxaga, I., Lasarguren, A., Muguerza, J., Pérez, J. M. (2013). The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. *Progress in Artificial Intelligence, 2*(1), 45-63.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., . . . Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *Journal of IEEE Access, 4*, 7940-7957.

Bach, M., Werner, A., Żywiec, J., Pluskiewicz, W. (2017). The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences, 384*, 174-190.

Barandela, R., Valdovinos, R. M., Sánchez, J. S. (2003). New applications of ensembles of classifiers. *Pattern Analysis Applications, 6*(3), 245-256.

Barua, S., Islam, M. M., Yao, X., Murase, K. (2012). MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge data engineering, 26*(2), 405-425.

Batista, G. E., Prati, R. C., Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20-29.

Bennin, K. E., Keung, J., Monden, A., Kamei, Y., Ubayashi, N. (2016). *Investigating the effects of balanced training and testing datasets on effort-aware fault prediction models.* Paper presented at the 2016 IEEE 40th annual Computer software and applications conference (COMPSAC).

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123-140.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (2017). *Classification and regression trees*: Routledge.

Bujang, M. A. ve Adnan, T. H. (2016). Requirements for minimum sample size for sensitivity and specificity analysis. *Journal of clinical diagnostic research: JCDR, 10*(10), YE01-YE06.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chen, S., He, H., Garcia, E. A. (2010). RAMOBoost: Ranked minority oversampling in boosting. *IEEE Transactions on Neural Networks, 21*(10), 1624-1642.

Efron, B. ve Tibshirani, R. J. (1994). *An introduction to the bootstrap*: CRC press.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 11): Springer.

Fletcher, G. S. (2019). *Clinical epidemiology: the essentials*: Lippincott Williams & Wilkins.

Freund, Y. ve Schapire, R. E. (1996). *Experiments with a new boosting algorithm.* Paper presented at the iclm.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 42*(4), 463-484.

Han, J., Pei, J., Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Hanley, J. A. ve McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Hasanin, T. ve Khoshgoftaar, T. (2018). *The effects of random undersampling with simulated class imbalance for big data.* Paper presented at the 2018 IEEE International Conference on Information Reuse and Integration (IRI).

He, H., Bai, Y., Garcia, E. A., Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* Paper presented at the 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence).

He, H. ve Ma, Y. (2013). Imbalanced learning: foundations, algorithms, and applications.

Japkowicz, N. (2000). *The class imbalance problem: Significance and strategies.* Paper presented at the Proc. of the Int'l Conf. on Artificial Intelligence.

Kamei, Y., Monden, A., Matsumoto, S., Kakimoto, T., Matsumoto, K.-i. (2007). *The effects of over and under sampling on fault-prone module detection.* Paper presented at the First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007).

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*: John Wiley & Sons.

Khoshgoftaar, T. M., Seiffert, C., Van Hulse, J., Napolitano, A., Folleco, A. (2007). *Learning with limited minority class data.* Paper presented at the Sixth International Conference on Machine Learning and Applications (ICMLA 2007).

Khoshgoftaar, T. M., Van Hulse, J., Napolitano, A. (2010). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, Cybernetics-Part A: Systems Humans, 41*(3), 552-568.

Krawczyk, B., Schaefer, G., Woźniak, M. J. A. i. i. m. (2015). A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. *65*(3), 219-227.

Kubat, M., Holte, R. C., Matwin, S. J. M. l. (1998). Machine learning for the detection of oil spills in satellite radar images. *30*(2), 195-215.

Larose, D. T. ve Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining* (Vol. 4): John Wiley & Sons.

López, V., Fernández, A., García, S., Palade, V., Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113-141.

Metz, C. E. (1978). *Basic principles of ROC analysis.* Paper presented at the Seminars in nuclear medicine.

Omurlu, I. K., Ture, M., Unubol, M., Katranci, M., Guney, E. (2014). Comparing performances of logistic regression, classification & regression trees and artificial neural networks for predicting albuminuria in type 2 diabetes mellitus. *Int J Sci Basic Appl Res, 16*(1), 173-187.

Rashu, R. I., Haq, N., Rahman, R. M. (2014). *Data mining approaches to predict final grade by overcoming class imbalance problem.* Paper presented at the 2014 17th International Conference on Computer and Information Technology (ICCIT).

Ren, F., Cao, P., Li, W., Zhao, D., Zaiane, O. (2017). Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm. *Computerized Medical Imaging Graphics, 55*, 54-67.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. *IEEE Computational Intelligence Magazine, 13*(4), 59-76.

Saxena, S., Shukla, S., Gyanchandani, M. (2021). Breast cancer histopathology image classification using kernelized weighted extreme learning machine. *International Journal of Imaging Systems Technology, 31*(1), 168-179.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning, 5*(2), 197-227.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, Cybernetics-Part A: Systems Humans, 40*(1), 185-197.

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition, 48*(5), 1623-1637.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics, 24*, 303-329.

Tyagi, S. ve Mittal, S. (2020). Sampling approaches for imbalanced data classification problem in machine learning. In *Proceedings of ICRIC 2019* (pp. 209-221): Springer.

Van Hulse, J., Khoshgoftaar, T. M., Napolitano, A. (2007). *Experimental perspectives on learning from imbalanced data.* Paper presented at the Proceedings of the 24th international conference on Machine learning.

Wang, C., Deng, C., Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters, 136*, 190-197.

Weiss, G. M. ve Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of artificial intelligence research, 19*, 315-354.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports, 62*(2), 1432-1449.

Zhong, W., Raahemi, B., Liu, J. (2009). *Learning on class imbalanced data to classify peer-to-peer applications in ip traffic using resampling techniques.* Paper presented at the 2009 International Joint Conference on Neural Networks.

Zięba, M., Tomczak, J. M., Lubicz, M., Świątek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied soft computing, 14*, 99-108.