T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

GÜRÜLTÜLÜ GÖZLEMLER DURUMUNDA DENGESİZ VERİDE ÖĞRENME İÇİN YENİ BİR YAKLAŞIM

FATİH SAĞLAM

İSTATİSTİK ANABİLİM DALI

SAMSUN
2020

Her hakkı saklıdır.

TEZ ONAYI

Fatih Sağlam tarafından hazırlanan "Gürültülü Gözlemler Durumunda Dengesiz Veride Öğrenme için Yeni Bir Yaklaşım" adlı tez çalışması …/…/2020 tarihinde aşağıdaki jüri tarafından Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda **Yüksek Lisans Tezi** olarak kabul edilmiştir.

Danışman	Prof. Dr. Mehmet Ali CENGİZ İstatistik Anabilim Dalı
İkinci Danışman	Dr. Öğr. Üyesi Emre Dünder Ondokuz Mayıs Üniversitesi İstatistik Anabilim Dalı

Jüri Üyeleri

Başkan	Prof. Dr. Mehmet Ali CENGİZ Ondokuz Mayıs Üniversitesi İstatistik Anabilim Dalı	……..
Üye	Prof. Dr. Yüksel TERZİ Ondokuz Mayıs Üniversitesi İstatistik Anabilim Dalı	……..
Üye	Dr. Öğr. Üyesi Faruk BULUT Rumeli Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı	……..

Yukarıdaki sonucu onaylarım. / / 2020

Prof. Dr. Bahtiyar ÖZTÜRK

Enstitü Müdürü

ETİK BEYAN

Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez içindeki bütün bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

…/…/2020

Fatih SAĞLAM

ÖZET

Yüksek Lisans Tezi

GÜRÜLTÜLÜ GÖZLEMLER DURUMUNDA DENGESİZ VERİDE ÖĞRENME İÇİN YENİ BİR YAKLAŞIM

Fatih Sağlam

Ondokuz Mayıs Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Prof. Dr. Mehmet Ali Cengiz

İkinci Danışman: Dr. Öğr. Üyesi Emre Dünder

Sınıflama çalışmalarında kullanılan sınıflama yöntemlerinin çoğunda sınıf gözlem sayılarının dengeli olduğu varsayımı vardır. Bu gibi durumlarda kurulan modeller çok gözleme sahip sınıfa ağırlık vererek tahminde bulunmaktadırlar. Böyle durumlarda sınıflayıcılar az gözlem sayısına sahip gözlemleri göz ardı ettikleri için çoğunluk sınıftan yana yanlı tahminde bulunmaktadırlar. Bu sınıf dengesizliği problemi bulunan veri setlerinde kullanılması önerilen performans ölçütleri olduğu gibi, bu problemi çözmek için önerilmiş yöntemler mevcuttur. Bu yöntemlerden en sık kullanılanlarından birisi yeniden örnekleme yöntemleridir.

Bu çalışmada yeniden örnekleme yöntemlerinden olan rastgele aşırı örnekleme (RAÖ) ve sentetik azınlık aşırı örnekleme (SMOTE) yöntemlerinin sorunları ele alınmış ve bu sorunları çözmeyi amaçlayan yeni bir yeniden örnekleme yöntemi önerilmiştir. Önerilen boosting ile SMOTE (B. SMOTE) yöntemi topluluk algoritmalarında kullanılan boosting prosedürünü kullanarak gürültü tespiti yapmakta ve bu gürültü bilgilerini kullanarak SMOTE algoritması içerisinde her bir gözlem için ayrı uygun komşu sayısı belirlemektedir.

Çalışmanın uygulama kısmında simülasyon verisi üzerinde yöntemler karşılaştırılmış ve görsel olarak RAÖ, ve SMOTE'un sorunları gösterildiği gibi B. SMOTE yönteminin bu sorunları aştığı ve daha iyi performans gösterdiği görülmüştür. Ayrıca 16 farklı veri seti ve 9 farklı sınıflayıcı üzerinden yapılan sınıflama modelerinin karşılaştırması sonucunda MKK ve F_1 performansları ve bu performansların sıra numaraları hesaplanmıştır. Sonuç olarak önerilen yöntemin her bir sınıflayıcıda ve tüm genel sonuçların ortalamasında diğer mevcut yeniden örnekleme yöntemlerinden daha iyi olduğu gösterilmiştir.

Ocak 2020, 78 sayfa
Anahtar Kelimeler: Sınıf dengesizliği, SMOTE, Boosting, Yeniden örnekleme

ABSTRACT

Master's Thesis

A NOVEL APROACH FOR LEARNING IN IMBALANCED DATA IN THE PRESENCE OF NOISE

Fatih Sağlam

Ondokuz Mayis University
Institute of Science
Department of Statistics

Supervisor: Prof. Dr. Mehmet Ali Cengiz

Co-Supervisor: Dr. Lecturer Emre Dünder

Most of the classification methods used in the classification studies have the asumption that the numbers of class observations are balanced. In such cases, models are predicted by giving biased weight to the the class with more observations. Therefore, the classifiers ignore the class with smaller number of observations and the majority class makes biased predictions. In data sets with class imbalance problem, there are suggested performance measures to be used as well as proposed methods to solve this problem. One of the most commonly used methods is resampling method.

In this study, the problems of random oversampling (ROS) and synthetic minority oversampling technique (SMOTE), which are some of the oversampling methods, are discussed and a new resampling method is proposed to solve these problems. The proposed SMOTE with boosting (B. SMOTE) method makes noise detection using the boosting procedure in ensemble algorithms and uses this information to determine the appropriate number of neighbors for each observation within SMOTE algorithm.

In the application section of the study, methods on both simulation data are compared and the problems of ROS and SMOTE are shown visually. Also, it is seen that B. SMOTE method overcame these problems and performed better. In addition, MCC and F_1 performances and ranks of these performances are calculated as a result of classification models made over 16 different data sets and 9 different classifiers. It is shown that the proposed method is better than the other resampling methods for each classifier and also in general.

January 2020, 78 pages

Keywords: Class imbalance, SMOTE, Boosting, Resampling

ÖNSÖZ VE TEŞEKKÜRLER

Akademik eğitim sürecimin bir üst noktası olan yüksek lisans tez boyunca yardım ve desteğini benden esirgemeyen danışman hocam Sayın Prof. Dr. Mehmet Ali Cengiz'e ve ikinci danışmanım Dr. Öğr. Üyesi Emre Dünder'e teşekkürü bir borç bilirim.

Bugünlere gelmemde en büyük paya sahip olan, gerek manevi gerekse maddi olarak her zaman destekleriyle yanımda olan, hayattaki en değerli varlıklarım başta annem ve babam olmak üzere tüm aileme ve arkadaşlarıma sonsuz teşekkür ederim.

Ocak 2020, Samsun

Fatih Sağlam

İÇİNDEKİLER

KISALTMALAR

AdaBoost	Adaptive Boosting (Uyarlamalı Boosting)
AÖ	Aşırı Örnekleme
B. SMOTE	Boosting ile SMOTE
CART	Classification and Regression Tree (Sınıflama ve Regresyon Ağacı)
DVM	Destek Vektör Makineleri
EAA (AUC)	Eğri Altındaki Alan (Area Under Curve)
SMOTE	Synthetic Minority Oversampling Technique (Sentetik Azınlık Aşırı Örnekleme Yöntemi)
EYK (KNN)	En Yakın K Komşuluk (K Nearest Neighbour)
MARS	Multivariate Adaptive Regression Splines (Çoklu Uyarlamalı Regresyon Splinleri)
MKK (MCC)	Matthew'in Korelasyon Katsayısı (Matthew's Correlation Coefficient)
RAÖ (ROS)	Rastgele Aşırı Örnekleme (Random Oversampling)
ROC	Receiver Operating Characteristic (Algılayıcı İşletim Eğrisi)
YÖ	Yeniden Örnekleme
YSA (ANN)	Yapay Sinir Ağları (Artificial Neural Networks)

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

ALGORİTMALAR DİZİNİ

1. GİRİŞ

Sınıf dengesizliği problemi, bazı sınıflara ait gözlemlerin diğer sınıflara göre daha fazla olması durumudur. Sınıf dengesizliği, veri setlerinde öğrenme sürecinde problemlere yol açar. Çünkü standart yöntemler çok gözleme sahip sınıflara ağırlık verir ve gözlem sayısı az olanları görmezden gelir. Çoğu sınıflayıcı algoritma, doğruluk oranını maksimize etmeyi veya hata oranını minimize etmeyi amaçlar. Gerçek hayatta, tıbbi teşhisler, sahtekarlık tespiti, yazı sınıflama ve yağ dökülme tespiti gibi alanlarda dengesiz veri setleri ile karşılaşılmaktadır. Azınlık sınıf tahmini, çoğunluk sınıf gözlemlerinden daha önemlidir. Bu nedenle sınıf dengesizliği git gide daha fazla ilgi çeken bir konu haline gelmektedir.

Sınıf dengesizliği problemi için bazı çözümler önerilmiştir. Bu çözümlerden birisi yeniden örnekleme yöntemidir. Bu yöntemde azınlık sınıf gözlemleri aşırı örneklenerek ve/veya çoğunluk sınıf gözlemleri az örneklenerek denge sağlanması amaçlanmaktadır. Yeniden örnekleme algoritmalarının kendi içinde türleri vardır. Rastgele yeniden örnekleme yöntemlerinde veri rastgele bir şekilde aşırı örneklenir ve/veya az örneklenir. Bazı yeniden örnekleme yöntemlerinde, rastgele veri türetmek yerine bir yönteme dayalı şekilde belirli bölgelere odaklanılır ve seçilmiş olan bu bölge üzerinden yeniden örnekleme yapılır. Diğer bir yeniden örneklem yönteminde ise yeni sentetik veri türetilir. Aşırı örnekleme ve az örneklemenin işe yaradığı birçok çalışmada gösterilmiş olsa da hala bazı dezavantajları vardır. Aşırı örnekleme yöntemi, kurulan modellerde aşırı uyum (overfitting) sorununa yol açabilmektedir. Az örnekleme yönteminde ise hali hazırdaki bilginin yok edilmesi söz konusudur.

Sınıflama algoritmalarında düzenleme yaparak da sınıf dengesizliği problemi ile çözülebilir. Sınıflara ait maliyetler (maliyet duyarlı öğrenme) değiştirilebilir. Tahmin kararını verecek olan eşik değeri değiştirilebilir. Elde edilecek olan olasılık tahminin elde ediliş şekli değiştirilebilir. Sınıfları ayrıştırmayı temel alan yöntemlerin yerine sadece bir sınıfa odaklı öğrenme yöntemleri kullanılabilir.

Topluluk sınıflayıcılarını kullanmak, zayıf sınıflayıcıların performansını artıran bir yöntemdir. Özellikle verideki gözlem sayısı az olduğunda veriden daha iyi bilgi

toplayabilmeyi sağlar. Torbalama (bagging) ve boosting gibi popüler topluluk (ensemble) algoritmalarının bazılarında her bir modele yeniden örnekleme yöntemleri uygulanabilmektedir.

Sınıf dengesizliği, veriyi öğrenmede zorluk yaratmanın yanı sıra elde edilen modelin performansını değerlendirmede de probleme neden olmaktadır. Sınıflayıcıları değerlendirirken en çok kullanılan doğruluk oranı veya hata oranı dengesiz sınıfların bulunduğu veri setinde doğru bir performans ölçüsü verememektedir. Bunun nedeni sınıflayıcıların çoğunluk sınıftan yana olması ve çoğunluk sınıfın gözlem sayısı fazla olduğu için modelin olduğundan daha iyi gözükmesidir. Bu durumlarda her iki sınıfı da eşit ağırlıkta dikkate alan performans ölçülerine yönelmek gerekir.

Bu çalışmada yeniden örnekleme yöntemleri ile sınıf dengesizliği probleminin çözümüne odaklanılmıştır. İlk bölümde sınıf dengesizliği, gürültü tespiti ve boosting ile ilgili literatür taraması yapılmıştır. İkinci bölümde sınıflamanın tanımı yapılmış ve sınıf dengesizliği problemi tanıtılmıştır. Burada sınıf dengesizliğinden bahsedilirken bu problemi çözmede kullanılan yaklaşımlardan ve dengesiz sınıf veri setleri için kullanılan performans ölçütlerinden bahsedilmiştir. Üçüncü bölümde topluluk algoritmalarından olan boosting algoritması anlatılmıştır. Dördüncü bölümde yeniden örnekleme yöntemlerinden bahsedilmiştir. İlk olarak rastgele aşırı örnekleme yöntemi tanımlanmıştır. Ardından literatürde en çok kullanılan SMOTE aşırı örnekleme yöntemi incelenmiş ve bu yöntemin eksik yanlarına vurgu yapılmıştır. Daha sonra SMOTE yönteminin eksik yanlarını çözmek için boosting ve gözlem komşuluklarının kullanıldığı yeni bir yöntem önerilmiş ve tanıtılmıştır. Beşinci bölümde dengesiz sınıflara sahip farklı veri setleri tanıtılmış ve her bir yeniden örnekleme yöntemi için bu veri setleri üzerinde farklı sınıflayıcılar ile modeller kurulmuştur. Ardından elde edilen performanslar karşılaştırılmıştır.

1.1. Sınıf Dengesizliği Problemi ile İlgili Çalışmalar

Han ve arkadaşları (2005), BL-SMOTE1 ve BLSMOTE2 isimli SMOTE yöntemini temel alan yeni iki aşırı örnekleme yöntemi önermişlerdir. Bu yöntemlerde sadece sınıflar arasında sınır bölgelerinde veri türetilmektedir. Deneysel çalışmalarda önerilen yöntemin SMOTE ve rastgele aşırıörnekleme yöntemine göre daha yüksek gerçek pozitif oranı ve F_1 değerine ulaştığı görülmüştür.

He ve arkadaşları (2008), dengesiz veri setleri için yeni bir sentetik örnekleme yaklaşımı olan ADASYN yöntemini önermiştir. Bu önerilen yaklaşımda temel fikir azınlık sınıf gözlemleri için öğrenme zorluklarına göre ağırlıklar kullanmaktır. Bu yöntemde öğrenmesi zor olan gözlemler için daha fazla veri türetilip, öğrenmesi kolay gözlemler için daha az veri türetilmektedir. Birçok veri setinde yapılan deneysel çalışmalar yöntemin beş farklı değerlendirme ölçütündeki etkinliğini göstermiştir.

Bunkhumpornpat ve arkadaşları (2009), SLSMOTE isimli SMOTE yönteminin geliştirilmiş bir yöntemini önermiştir. Bu yöntemde azınlık gözlemlere belirli ağırlıklar verilmekte ve güvenli seviyedeki gözlemler üzerinden sentetik veriler türetilmektedir. Yapılan deneylerde SLSMOTE yönteminin SMOTE ve B-SMOTE yönteminden daha iyi sonuçlar verdiği görülmüştür.

Bunkhumpornpat ve arkadaşları (2012), DBSMOTE isimli yeni bir aşırı örnekleme yöntemi önermiştir. Bu yöntem DBSCAN kümeleme yöntemi ile kabaca tespit edilmiş küme yoğunluklarını kullanmaktadır. Veri türetirken azınlık kümesinin merkezine doğru veri türetir. Sonuç olarak türetilen veriler küme merkezlerinde fazla ve merkezlerden uzakta azdır. Deneysel çalışmaların sonucunda DBSMOTE yönteminin SMOTE, B-SMOTE ve SLSMOTE yöntemlerinden daha yüksek F_1, AUC değerleri yakalayabildiği görülmüştür.

Zhang ve Li (2014), farklı sınıflara ait gözlemleri dengeli hale getirebilmek için veri dağılımını değiştirmeyen ve sınırları genişleten RWO isimli yeni bir yaklaşım önermişlerdir. Rastgele yürüme yöntemi ile gerçek veriden sentetik veri türeten bu yöntemde türetilen veri ile orijinal verinin ortalama ve standart sapmaları arasında anlamlı fark olmadığı gösterilmiştir. Deneyler sonucunda RWO yönteminin alternatif dengeleme yöntemlerinden birçok veri setinde daha iyi olduğu görülmüştür.

Qian ve arkadaşları (2014), dengesiz veri setleri için sınıflama problemlerine odaklanan bir yeniden örnekleme topluluğu algoritması geliştirmiştir. Bu yöntemde küçük sınıflar aşırı örneklenmiş ve büyük sınıflar az örneklenmiştir. Yeniden örnekleme ölçeği, minimum sınıf ile maksimum sınıf gözlem sayısı oranına göre belirlenmiştir. Birçok makine öğrenme yöntemi kullanılarak topluluk oluşturulmuştur. Azınlık gözlem sayısı oranı ve toplam değişken sayısının sonuçları yüksek derecede etkilediği tespit edilmiştir.

Charte ve arkadaşları (2015), REMEDIAL isimli çok sınıflı sınıflama yöntemlerinde kullanılmak üzere bir yeniden örnekleme yöntemi önermiştir. Bu yöntem aşırı örnekleme ve az örnekleme yapmak yerine kesişmiş dengesiz sınıf gözlemlerini birbirlerinden ayırmaya dayalı bir prosedürdür. Charte ve arkadaşları (2019), daha sonra bu yöntemi geliştirmişler ve REMEDIAL-Hwr isimli yeni bir prosedür önermişlerdir. Bu prosedürde ise REMEDIAL yöntemini bilinen en iyi yeniden örnekleme algoritmaları ile birleştirmişlerdir. Bu hibrit yöntemler incelenmiş ve kesişen etiketleri üzerindeki davranışları belirlenmiştir. Sonuçlar, önerilen yöntemin belirli sınıflayıcılarda aşırı dengesiz veri setlerinde performansı artırdığını göstermiştir.

Bulut (2016), sınıf dengesizliği problemi içeren veri setlerinde farklı sınıflayıcıların performanslarını incelemiştir. Çalışmada karar ağaçları, en yakın k komşular, Naive Bayes, destek vektör makinaları ve lojistik regresyon sınıflayıcıları kullanılmıştır. Kullanılan veri setlerinde lojistik regresyonun diğer sınıflayıcılardan daha iyi performans gösterdiği görülmüştür.

Loyola-Gonzàlez ve arkadaşları (2016), sınıf dengesizliği problemlerinde zıtlık düzeni temelli sınıflayıcıların performansını geliştirmede kullanılan yeniden örnekleme yöntemlerinin etkisini araştırmışlardır. Standart dengesiz veri setleri kullanıldığı zaman sonuçlarda yeniden örnekleme yapılmasından önce ve sonra arasında istatistiksel anlamlı farklılık olduğu görülmüştür. Bu çalışma, sınıf dengesi oranına göre yeniden örnekleme yönteminin seçimi hakkında da bir rehber olmuştur.

Yijing ve arkadaşları (2016), çok sınıflı dengesiz öğrenme problemi için AMCS isimli uyarlamalı çoklu sınıflama sistemi önermişlerdir. Bu yöntemde farklı dengesiz veri türleri arasında ayrım yapılmaktadır. Yöntem nitelik seçimi, yeniden örnekleme ve topluluk öğrenme tekniklerinin bir birleşimidir. İki farklı nitelik seçimi, üç farklı yeniden örnekleme ve beş farklı temel sınıflayıcı ile seçim havuzu oluşturulmuş ve deneysel çalışmalar ile AMCS yöntemi analiz edilmiştir. Sonuçlar AMCS yönteminin diğer yöntemlerden daha iyi veya rekabet edebilecek seviyede olduğunu göstermiştir.

Furundzic ve arkadaşları (2017), sınıflama problemlerinde dengesiz öğrenme problemine yeni bir metodolojik yaklaşım önermiştir. Önerilen yöntem örneklem entropisini maksimize etmeye dayalıdır. İdeal olarak dengelenmiş normal kafes

numunenin dağılım özelliklerinin tespitini ve bu özelliklerin temsilliğini artıran keyfi bir dengesiz numuneye kabul edilebilir transferini içermektedir. Yöntem yüksek olasılık yoğunluklu bölgelere az örnekleme, düşük yoğunluklu bölgelere aşırı örnekleme yapıldığını varsaymaktadır. Yöntem sonucunda örneklem entropisi artmış ve öğrenme algoritmasının bir sınıfa veya kümeye yönelik yanlılığı azalmıştır.

Ofek ve arkadaşları (2017), iki sınıflı sınıf dengesizliği problemini ele alan kümeleme tabanlı, hızlı ve yeni bir az örnekleme yöntemini tanıtmıştır. Bu yöntem, eğitim aşamasında azınlık sınıf gözlemlerini kümeleyen ve her bir kümeye yakın olan belirli miktarda çoğunluk sınıf gözlemini seçen bir yöntemdir. Her bir küme için ayrı bir sınıflayıcı kurulur ve hiçbir kümeye girmeyen örnek çoğunluk sınıf olarak etiketlenir. Kümelere ait sınıflayıcılardan elde edilen sonuçlar kümlerden uzaklıklarına göre ağırlıklandırılır. Birçok farklı veri setinde yapılan çalışmalar, önerilen yöntemin diğer yöntemlerden daha iyi olduğunu göstermiştir.

Charte ve arkadaşları (2017), dengesiz sınıflarda zor etiket problemini derinlemesine incelemiş ve çok sınıflı sınıflayıcılar üzerindeki etkisini ele alarak yeni bir yaklaşım önermiştir. Çok sınıflı veri setlerinde kullanılmak üzere iki farklı performans ölçütü ve çok sınıflı veri setlerine uygun yeni bir yeniden örnekleme yöntemi önerilmiştir.

Yu ve arkadaşları (2018), kredi kartı sınıflamada dengesiz veri problemini çözmek için derin inanç ağları temelli yeniden örnekleme destek vektör makinesi topluluk algoritması önermiştir. Bu yöntemde öncelikle dengeli alt kümeler oluşturabilmek için eğitim alt kümeleri oluştururken torbalama yöntemi kullanılmıştır. Daha sonra her bir alt küme için destek vektör makinesi modeli kullanılmıştır. Son olarak derin inanç ağları modeli bir topluluk yöntemi olarak girdi modellerinin sınıflama sonuçlarının toplamada uygulanmıştır. Ek olarak farklı sınıflara ait ağırlıklar, gelir duyarlı yöntem ile gelir matrisi kullanarak değiştirilmiştir. Deneysel sonuçlarda yeniden örnekleme ile kullanılan derin inanç ağları temelli topluluk stratejisinin özellikle sınıf dengesizliği probleminde performansı etkili derecede artırdığı görülmüştür.

García ve arkadaşları (2018), çalışmalarında çok sınıflı dengesizlik problemlerine odaklanmışlardır. Önerdikleri DES-MI isimli çok sınıflı dengesiz veri

setleri için dinamik topluluk seçimi yönteminde aday sınıflayıcıların yeterlilikleri komşuluklarındaki ağırlıklı örneklemlere göre belirlenmiştir. Bu yöntemde dengeli bir eğitim veri seti oluşturulmakta ve uygun sınıflayıcı seçilmektedir. Çok sınıflı veri setlerinde yapılan deneylerden elde edilen sonuçlar, DES-MI yönteminin sınıflama performansını artırdığını göstermiştir.

Sharififar ve arkadaşları (2019), toprak dağılımı tahmini yaparken sınıf dengesizliği problemini çözmede aşırı ve az örnekleme yöntemlerini kullanmışlardır. Çalışmada karar ağaçları, rastgele orman ve çok terimli lojistik regresyon modeli kurulmuştur. Sonuçlar, dengesiz sınıf dağılımı kullanarak yapılan tahminlerin azınlık sınıfa ait belirsiz haritaların komple kaybolmasına veya kısmen düşük doğrulukla tahmin edilmesine yol açtığını göstermiştir. Aşırı ve az örnekleme uyguladıktan sonra azınlık sınıfa ait gözlemler tüm modellerde önemli derecede performans artışı olduğu görülmüştür.

Ren ve arkadaşları (2019), SRE isimli karmaşık veri dağılımlı dengesiz veri akışlarından öğrenmek için topluluk temelli bir yöntem önermiştir. Bu yöntem kesişme problemini çözebilmek için periyodik olarak güncellenirken yeniden örneklem yapan bir yöntemdir. Güvenli olmayan gözlemlere ve eğilime odaklanan seçim temelli bir yeniden örnekleme mekanizması, en son bloğun sınıf dağılımını yeniden dengelemek için ele alınır. Daha sonra önceki topluluk üyeleri en son gözlemler kullanılarak periyodik olarak güncellenir. Çalışma sonuçlar SRE'nin durağan olmayan veri akışlarındaki etkisini göstermiştir.

Kim ve arkadaşları (2019), duyarlılık ve belirlilik üzerinde kullanıcı tarafından belirlenmiş kısıtlar koyan dengesizlik problemini çözecek yöntemler önermiştir. Öncelikle belirlenen kısıtlar altında maksimum hata oranını minimize edecek hedef oranı tespit edilmiştir. İkinci olarak orijinal örnekten etkilenmeyen bir tahmin edici ile sınırlanmış rastgele yeniden örnekleme kavramını öne sürmüşlerdir. Bu iki önerilen yöntem her sınıflama yönteminde kullanılabilir. Üçüncü olarak da rastgele yeniden örnekleme uygulanan seçilmiş sınıflayıcıların özellikleri tespit edilmiştir. Son olarak önerilen yöntem, resim tanıma içeriğinden erişimli sinir ağlarının son katmanından nitelikler çıkarılarak uygulanmıştır. Önerilen yöntem ile göz dibi veri seti üzerinde diyabetik retinopati sınıflaması yapılmıştır.

Ali-Gombe ve Elyan (2019), sınıf dengesizliği problemini çözebilmek için çekişmeli üretici ağlar kullanan bir veri artırma yaklaşımı önermişlerdir. Tek bir sahte sınıf kullanan diğer çekişmeli üretici ağ modellerinin aksine, önerilen yöntem birden fazla sahte sınıf kullanmış ve azınlık sınıfın sınıflanmasını ve veri üretiminin iyi olduğundan emin olmayı amaçlamıştır. Ayrıca önerilen yöntemde azınlık sınıf gözlemleri üretilerek veri setinin dengelenmesi sağlanmıştır. Yapılan deneyler sonucunda azınlık sınıfı aşırı az olduğunda bile veri üretilebildiğini göstermiş ve diğer veri artırma ve aşırı örnekleme yöntemleri ile kıyaslandığında sınıflama doğruluğu performansı daha yüksek bulunmuştur.

Zefrehi ve Altınçay (2019), heterojen toplulukların kullanarak sınıf dengesizliği problemini ele almışlardır. 66 farklı veri setinde yapılan incelemelerde toplulukların heterojenliği ile AUC ve F_1 skorları arasındaki ilişki incelenmiştir. Sonuç olarak sınıflamaya yöntemlerini sayısı arttıkça performans skorunun da arttığını göstermiştir. Homojen topluluklarla kıyaslandığında heterojen toplulukların çok daha yüksek skorlara ulaştığı görülmüştür. Ayrıca veri dengeleme yöntemlerinin homojen ve heterojen toplulukların performanslarını etkilediği fakat artışın anlamlı olmadığı görülmüştür.

Douzas ve Bacao (2019), Geometrik SMOTE isimli SMOTE algoritmasının geliştirimiş bir versiyonunu önermişlerdir. Geometrik smote yöntemi SMOTE yönteminin aksina seçilen azınlık gözleminin etrafında girdi uzayının geometrik bir bölgesinde sentetik veri üretmektedir. Geometrik SMOTE ile SMOTE arasında performans karşılaştırması yapılmıştır. Deneyler sonucunda Geometrik SMOTE yönteminin üretilen verinin kalitesinde anlamlı artış gösterdiği görülmüştür.

Kamalov (2019), sınıf dengesizliği probleminde KDE isimli, çekirdek yoğunluk tahmini yöntemini kullanarak azınlık yeniden örnekleme yöntemini incelemiştir. Çalışmada KDE'nin diğer yeniden örnekleme yöntemlerinden daha doğal şekilde veri türettiği iddia edilmektedir. Yapılan deneyler sonucunda KDE yönteminin diğer yeniden örnekleme yöntemlerini F_1 ve G ortalama ölçütlerinde geçtiği görülmüştür. Sonuçlar farklı sınıflama algoritmalarında da benzer şekildedir.

Maldonado ve arkadaşları (2019), çalışmalarında SMOTE algoritmasını çok boyutlu iki sınıflı durumlar için uyarlamışlardır. Her bir azınlık gözleme ait komşular

tespit edilirken kullanmak üzere yeni bir uzaklık ölçütü önermişlerdir. Düşük ve yüksek boyutlu veri setlerinde diğer yöntemlerle yapılan karşılaştırmalarda diğer SMOTE temelli yöntemlere göre daha iyi sonuç verdiği görülmüştür.

Wong ve arkadaşları (2020), çalışmalarında sınıf dengesizliği problemini aşmak için CSDNN ve CSDE isimli iki yeni maliyet duyarlı yöntem önermiştir. CSDNN, yığılmış gürültü yok eden otomatik kodlayıcıların maliyet duyarlı bir versiyonudur. CSDE ise CSDNN yönteminin topluluk versiyonudur. Yapılan deneylerde yöntemin performansı farklı iş alanlarına ait altı farklı büyük gerçek veri setinde ölçülmüştür. Elde edilen sonuçlar sınıf dengesizliği problemini aşmada önerilen yöntemin diğer yöntemlerden daha iyi olduğunu göstermiştir.

Vuttipittayamongkol ve Elyan (2020), iki sınıflı veri setlerinde sınıf dengesizliği problemi için, üst üste gelen veri noktalarını yok eden bir az örnekleme yöntemi önermişlerdir. Önerilen yöntem, üst üste gelen bölgelerdeki çoğunluk sınıf gözlemlerini yok etmeye dayalıdır. Bu gözlemlerin doğru tespiti ve elenmesi ile azınlık sınıf gözlemlerinin görünürlüğü artırılmış ve aşırı veri atımına engel olunmuştur. Bu çalışmada aynı felsefede dört farklı yöntem önerilmiştir. Simülasyon verilerinde ve gerçek veri setlerinde yapılan deneylerde farklı performans ölçülerinden elde edilen sonuçlarda hassasiyet ölçüsünde istatistiksel olarak anlamlı artış olduğu görülmüştür.

1.2. Gürültü Tespiti ile İlgili Çalışmalar

Garcia ve arkadaşları (2015a), gürültü tespit yöntemlerinin çok sınıf problemlerindeki perfromansını artıracak bir yaklaşım önermiştir. Bu yaklaşımda gürültü filtrelerken teketek ayrıştırma stratejisini kullanarak gürültü tespiti basitleştirilmiştir. İki sınıflı altproblemlerde elde edilen filtreleme sonuçlarını birleştirmek için sınıflayıcıların gürültü tahmin derecelerinin toplamına dayalı hafif oylama yaklaşımı kullanılmıştır. Deney sonuçları, teketek ayrıştırma stratejisinin genellikle gürültü filtresinin performansını artırdığını göstermiştir.

Garcia ve arkadaşları (2015b), çalışmalarında gürültünün sınıflama problemlerinin karmaşıklığını nasıl etkilediğini incelemişlerdir. Bunu yaparken farklı gürültü seviyelerinin varlığında farklı veri karmaşıklığı indekslerinin duyarlılığını incelemişlerdir. Sınıflama veri setinin karmaşıklığını karakterize ederken geometrik,

istatistiksel ve yapısal ölçümler hesaplanmıştır. Karmaşıklık ölçülerinden iki tanesini temel alan bir gürültü tespit yöntemi de önerilmiş ve gürültü varlığına daha hassas oldukları gösterilmiştir.

Xu ve arkadaşları (2015), yeni bir temsil temelli sınıflama yöntemi önermişlerdir. Bu yöntemde etkili bir şekilde test ve eğitim gözlemlerindeki gürültü azaltılmaktadır. Ayrıca önerilen yöntem hem orijinal hem de sanal eğitim gözlemlerindeki gürültüyü azaltıp test gözlemi hakkında etiket tahmini yapmaktadır. Orijinal yüz imaj verisinden oluşturulan sanal eğitim gözleminde yüzün farklı ölçüleri, pozları ve ifadeleri sınıflandırılmış ve sonuçlarda önerilen yöntemin çok iyi olduğu görülmüştür.

Bootkrajang (2016), rastgele etiket gürültüsü ve rastgele olmayan etiket gürültüsünün negatif etkisine karşı dayanıklı yeni bir genelleştirilmiş gürültü modeli öne sürmüştür. Sentetik ve gerçek veri setlerinde yapılan çalışmalarda önerilen yöntemin sınıflama doğruluk oranını mevcut gürültü model yaklaşımlarına kıyasla daha fazla artırdığı görülmüştür.

Luo ve arkadaşları (2016), sınıflama gürültüsü kavramı tanıtılmış ve CNSMO isimli ardışık minimal optimizasyon algoritmasına dayalı sınıflama gürültü tespiti yöntemi önerilmiştir. Bu yöntemde ayrıştırılamayan bir problem ayrıştırılabilir hale dönüştürülmüş ve böylece aşırı uyum sorununu dikkate almaya gerek kalmamıştır. Yapılan uygulamalarda önerilen algoritmanın çok daha hızlı bir eğitim sürecine tahmin doğruluğunu azaltmadan ulaştığı birçok veri setinde gösterilmiştir.

Rivera (2017), gürültü yoketmeye ve azınlık gruptan örneklem seçmeye odaklanan yeni bir aşırı örnekleme yöntemi önermiştir. Farklı veri setleri, sınıflayıcılarda ve yeniden örnekleme yöntemleri ile yapılan uygulamalarda duyarlılık ve G-ortalama ölçütlerinde diğer yöntemlere kıyasla daha iyi sonuçlar verdiği görülmüştür.

Rivera (2017), gürültü azaltmaya ve azınlık sınıftan örneklem seçmeye dayalı, azınlık sınıf tahmin doğruluk oranını artırmayı amaçlayan yeni bir aşırı örnekleme yöntemi önermiştir. Birçok veri seti, sınıflayıcı ve aşırı örnekleme yöntemi üzerinde

yapılan uygulamalar sonucunda, önerilen yöntemin duyarlılık ve G-ortalama ölçülerinde diğer yöntemlerden daha iyi olduğu görülmüştür.

Luengo ve arkadaşları (2018), yanlış etiketlenmiş gözlemleri doğru bir şekilde filtrelemek ve mümkün olduğunda etiketlerini doğru olana değiştirmek için gürültü filtreleyici topluluğunu teml alan yeni bir yöntem geliştirmiştir. Aynı zamanda filtreleme ve yeniden etiketleme işlemini desteklemek için bir gürültü skoru da hesaplanmaktadır. Yöndem CNC-NOS olarak adlandırılmış ve çalışmada diğer filtreler ve etiket düzenleyicileri ile karşılaştırılmıştır. Sonuçlara göre diğer yöntemlere kıyasla daha kaliteli eğitim veri seti ortaya çıkmıştır.

İnce ve arkadaşları (2019), matematiksel morfoloji temelli bir filtre yapısı kullanarak çoklu gürültü azaltma yöntemi olarak kenara korumalı doğrusal olmayan gürültü filtre yöntemi önermişlerdir. Yöntemin adı minimum saçılım indeksidir (minimum index of dispersion, MID). Bu yöntem MCV ve MLV filtreleme yöntemlerinin bir uzantısıdır. Çalışmada bu filtreleme yöntemi diğer yöntemlerle karşılaştırılmış ve hem kenar koruma hem gürültü yok etmede sağlamlığı açısından MCV ve MLV filtrelerinden daha iyi sonuç verdiği görülmüştür.

Maas ve arkadaşları (2019), rastgele orman sınıflayıcısını gürültü tespit yöntemiyle birleştiren bir yaklaşım önermişlerdir. Bu yaklaşım, bir eğitim gözleminin sadece bir sınıfa atanmasından ziyade tüm sınıflara belirli olasılıklarla atanması gerektiği fikrine dayalıdır. Önerilen rastgele orman yöntemi beş farklı test bölgesinde uygulanmış ve standart rastgele orman yönteminden daha yüksek doğruluğa sahip sonuçlar elde edilmiştir.

García-Gil ve arkadaşları (2019), büyük veri analizlerinde kullanılmak üzere iki gürültü yok etme yaklaşımı önermişlerdir. Bu yaklaşımlar homojen topluluk ve homojen olmayan topluluk filtrelerdir. Elde edilen sonuçlar bu önerilen yaklaşımların büyük veri sınıflama problemlerinde akıllı veri seti elde etmede daha etkili olduğunu göstermiştir.

Cano ve arkadaşları (2019), monoton sıralı sınıflama probleminde gürültü gözlemlerinin monotonluk şartlarını sağlamamasını dikkate alarak monoton veri seti oluşturmaya çalışmıştır. Önerdikleri gürültü filtreleme yöntemiyle hem modellerin

monotonluk indeksini hem de tahminlerin doğruluğunu artırmayı amaçlamışlardır. 12 farklı veri setinde kurulan sınıflama ve regresyon modellerinde önerilen monoton sınıflayıcıların orijinal sınıflayıcılardan daha iyi olduğu görülmüştür.

Chen ve arkadaşları (2019), kümeleme tabanlı, bir gözlemi sinyal veya gürültü olarak etiketleyen yeni bir dalga şekli sınıflama algoritması öne sürmüşlerdir. Bu dalga şekli sınıflama algoritması doğrudan gürültülü veriye uygulanarak seçilmiş sinyalleri tespit edebilmekte ve böylece geleneksel bir gürültü yoketme algoritmasının yerine kullanılabilmektedir. Chen ve arkadaşları (2019), ayrıca aynı çalışmada EWT isimli empirik dalga dönüşümü temelli yeni bir gürültü yoketme yöntemi de önermiştir. EWT yöntemi taslak olarak kullanılarak önerilen dalga şekli sınıflama yönteminin nasıl geleneksel gürültü tespit yöntemleri ile beraber kullanılarak daha iyi sinyal tespit edilebileceği gösterilmiştir.

1.3. Boosting ile İlgili Çalışmalar

Chawla ve arkadaşları (2003), SMOTEBoost isimli boosting prosedürü ile SMOTE algoritmasının bir birleşiminden oluşan bir yaklaşım önermiştir. Yanlış sınıflandırılan gözlemlere eşit ağırlıkların verildiği standart bostingin aksine SMOTEBoost yönteminde azınlık sınıftan sentetik gözlemler oluşturulur ve dolaylı biçimde çarpık dağılımlar için ağırlıkların güncellenişini değiştirir. Aşırı ve orta derecede dengesiz veri setlerinde uygulandığında azınlık sınıfına ait doğruluğu ve genel F değerlerini artırdığı görülmüştür.

Seiffert ve arkadaşları (2009), sınıf dengesizliği problemini çözmek için RUSBoost isimli yeniden örnekleme ile boosting algoritmasını birleştiren hibrit bir yöntem önermiştir. Yöntem Chawla ve arkadaşları'larının (2003) SMOTEBoost yöntemi ile benzer prosedüre sahiptir. Sadece SMOTE yerine rastgele az örnekleme uygular. 15 farklı veri setinde yapılan uygulamalar sonucunda RUSBoost ile SMOTEBoost yöntemlerinin diğer yeniden örnekleme yöntemlerinden daha iyi olduğu ve kendilerinin de birbirlerine yakın olduğu tespit edilmiştir.

Chen ve arkadaşları (2010), boostingde sıralı azınlık aşırıörnekleme yöntemini önermişlerdir. Bu yöntem her bir boosting iterasyonunda azınlık sınıfın örneklem olasılık dağılımına göre azınlık gözlemleri sıralar ve uyarlamalı biçimde karar sınırını

zor öğrenilen azınlıklara doğru çeker. Çalışma sonucunda 19 farklı gerçek veri setinde yöntemin performans sonuçları verilmiştir.

Wang ve Japkowicz (2010), destek vektör makinelerini zayıf sınıflayıcı olarak kullanarak çarpık dengesiz sınıflı veri setlerinde vektör uzay problemini çözmeye çalışmıştır. Daha sonra bu yaklaşımla gelen aşırı yanlılığı boosting algoritması ile karşılamışlardır. Bu destök vektör makinesi topluluğunun tahmin performansında hem azınlık hem de çoğunluk sınıf için etkileyici artış gösterdiği görülmüştür.

Soleymani ve arkadaşları (2018), ilerlemeli boosting (progressive boosting) isimli yeni bir topluluk algoritması önermiştir. Bu yöntem boosting prosedürüne ilişkisiz örnek gruplarını ilerlemeli bir şekilde ekleyip sınıflayıcılar gölü oluştururken bilgi kaybetmekten kaçınmayı amaçlamıştır. Sonuç olarak ilerlemeli boosting algoritmasının çarpıklık seviyesi değişken ve bilinmiyen veri setlerinde çalışırken daha güçlü olduğu görülmüştür. Ek olarak ilerlemeli bosting algoritmasının hesaplama karmaşıklığı, literatürde bulunan dengesiz sınıflarda az örnekleme yapan boosting topluluklarından daha düşüktür.

Wang ve arkadaşları (2018), yeni bir otomatik kaza sınıflama yöntemi önermiştir. Önerilen yöntemde trafik sinyalleri kayıtlı videolardan çıkartılmış ve spatiyal-temporal sinyallere dönüştürülmüştür. Daha sonra yerel bir ortalama filtresi kullanılarak gürültünün etkisi azaltılmış ve bu sinyallerden anlamlı nitelikler çıkartılmıştır. Çıkartılan nitelikler farklı tür trafik verileriyle beraber kullanılmıştır. Sınıflama aşamasında uyarlamalı boosting sınıflayıcısı verideki aykırı değerleri tespit etmek için kullanılmıştır. Destek vektör makine temelli bir yöntem ile aykırı değerlerin kategorileri belirlenmiş ve uyarlamalı boosting destek vektör makinesi yöntemi tanıtılmıştır. Deney sonuçlarında önerilen yöntemin %92'lik doğruluk oranına ulaştığı görülmüştür.

Sundar ve Punniyamoorthy (2019), Wang ve Japkowicz'in (2010) önerdiği yöntemde değişiklikler yaparak mevcut zaman karmaşıklığını artırmadan sınıflama performansını artırmışlardır. Değişiklik olarak ağırlık güncellerken uzaklık temelli bir kural kullanmışlar ve işaret temelli sınıflayıcı yok etme yöntemi kullanmışlardır. 43 farklı dengesiz sınıflı veri setinde yapılan uygulamalarda değişikliklerin her ikisininde boosting peformansında önemli artışa neden olduğu görülmüştür.

Fu ve arkadaşları (2019), sınıflama ağacının değişken seçimi ve değişken sıralamasındaki avantajlarını, boostingin tek bir dayanıklı ve güvenilir model geliştirme özelliğini temel alan, dayanıklı bilgilendirici değişken seçimi yapacak boosting sınıflama ağacı tasarlamıştır. Bu topluluk temelli değişken seçimi taslağı tekrar tekrar, orijinal veri setinin farklı ağırlıklarında karar ağcı sınıflandırıcısı kurmakta ve her bir tekrarda değişkenler ve önem sıralamaları hakkında bilgi toplamaktadır. Bu farklı bilgiler daha sonra yeni bir sınıflayıcıda birleştirilmektedir. Bu yaklaşımın performansını inceleyebilmek için akciğer hastalığına yakalanmış ve sağlık kontörlüne gelmiş hastalardan elde edilen veri seti kullanılmıştır. Uygulamada destek vektör makineleri, çok katmanlı perseptron ve önerilen yöntem uygulanmıştır. Sonuçlara göre önerilen yöntem dayanıklı ve güvenli bir şekilde diğer sınıflayıcılardan daha yüksek doğruluk oranına sahiptir.

Dev ve Eden (2019), yakın zamanda geliştirilmiş olan XGBoost, LightBM ve CatBoost yöntemlerini Daniudui ve Hangjinqi gaz istasynlarından alınan verileri birleştirerek uygulamışlardır. Daha sonra bu yöntemlerin performansları rastgele orman, AdaBoost ve GBM yöntemleri ile karşılaştırılmıştır. Analiz sonucunda en başarılı yöntemin LightGBM olduğu tespit edilmiştir.

Obregon ve arkadaşları (2019), RuleCOSI isimli, iki sınıflı karar ağaçları topluluğunu birleştiren bir yöntem önermişlerdir. Önerilen yöntem, bir kombinasyon matrisi kullanarak her bir ağacın ağırlıklarını ele almaktadır. Yöntemin perforansını ölçmek için üç farklı boosting algoritması ile birlikte birçok bilindik makine öğrenme veri setlerinde modeller kurulmuş ve performanslar ölçülmüştür. Sonuçlara bakıldığında birçok durumda kabul edilebilir performasnlar gösterdiği ve diğer boosting topluluk algoritmaları ile rekabet edebildiği görülmüştür.

Shi ve arkadaşları (2019), elektrokardiogram kalp atışlarındaki düzensizliği tespit edebilmek için XGBoost yöntemine dayalı hiyerarşik bir sınıflama yöntemi önermiştir. Önişlemden geçilirmiş kalp atışlarında 6 farklı nitelik çıkartılmıştır. Daha sonra önyinelemeli nitelik eleme yöntemi ile nitelikler seçilmiştir. Sınıflama aşamasında XGBoost sınıflayıcıları birleştirilerek hiyerarşik bir sınıflama modeli kurulmuştur. Sonuç olarak XGBoost yöntemi geliştirilmiş ve tek bir kalp atışı sınıflama modelinde ilk defa uygulanmıştır. Deneylerde yapılan karşılaştırmalarda yeni yöntemin etkinliği ortaya konulmuştur.

ALzubi ve arkadaşları (2019), büyük verideki akciğer kanseri hastalığı tespiti için maksimum olabilirlik boosting ile ağırlıkları optimize edilmiş sinir ağları topluluğu (WONN-MLB) kullanılmıştır. Önerilen yöntem nitelik seçimi ve topluluk sınıflaması olmak üzere iki aşamya ayırlmıştır. Deneysel sonuçlar, önerilen yaklaşımın diğer yöntemlere göre daha iyi yanlış pozitif oranı, doğruluk ve azaltılmış gecikme değerlerini yaklayabildiğini göstermiştir.

Ergul ve Bilgin (2019), çalışmalarında boosting temelli topluluk algoritması ile birleşik çekirdek fonksiyonlarını bir araya getirerek yüksek sınıflama performansı elde etmeyi amaçlamıştır. Ağırlıklı konveks kombinasyon yaklaşımını kullanarak spatial ve spektral birleşik çekirdek fonksiyonları oluşturulmuştur. Hesaplanması hızlı ve etkili olan aşırı öğrenme makine sınıflama algoritması kullanılmıştır. Hibrit birleşik çekirdekler Gauss, polinomiyal ve logaritmik kernel fonksiyonlarından oluşmaktadır. Elde edilen sonuçar mevcut yöntemlerle karşılaştırmalı şekilde sunulmuştur.

Junior ve Carmo Nicoletti (2019), yeni bir topluluk temelli sınıflama algoritması önermiştir. Bu önerilen yöntemde topluluğa bazı temel sınıflayıcılar ekleyerek boosting uygulamaya dayanılmaktadır. Güncelleme mekanizması modelin esnekliğini geliştirdiği için topluluk hızlı bir şekilde bilgi toplayarak yüksek hata oranlarından uzaklaşmakta ve tatmin edici sonuçlar elde etmektedir. Deney sonuçlarında önerilen topluluk algoritması diğer sekiz farklı topluluk algoritması ile karşılaştırmış ve önerilen yöntemin veri akış sınıflamasında diğer yöntemlerle rekabet edebildiği görülmüştür.

2. SINIFLAMA

Sınıflama yöntemleri, gözlem verisindeki düzeni tespit edip, bu düzeni temel alarak örneklemi kategorize etmek amacıyla geliştirilmiştir (Lavine, 2000). Sınıflama bir tür örüntü tanıma yöntemidir ve her bir girdi değerinin "sınıflar" vektöründeki bir değere atamasını yapar. Bireylere ait değişkenleri inceleyerek onların "hasta" veya "sağlıklı" olduğuna karar verilmesi buna örnek olarak verilebilir. Makine öğrenme terminolojisinde sınıflama, gözetimli öğrenme kategorisine girmektedir (Alpaydın, 2009). Bu öğrenme yaklaşımında doğru sınıflandırılmış, yani gerçek sınıfların bilindiği bir eğitim veri seti mevcuttur. Tahmin edilecek olan sınıf değişkeni genellikle istatistikte "bağımlı değişken", makine öğrenmede ise "sınıflar" olarak tarif edilir. Bu bağımlı değişkeni etkilediği düşünülen ve sınıfların tahmini için kullanılan diğer değişkenler ise istatistikte "bağımsız değişken", "açıklayıcı değişken" vb. olarak adlandırılırken makine öğrenmede genellikle "nitelik" olarak adlandırılmaktadır. İstatistik ve makine öğrenme dışındaki alanlarda, "kümeleme analizi" sınıflama olarak adlandırılabilmektedir. Fakat istatistik ve makine öğrenmede kümeleme ve sınıflama analizi birbirinden farklıdır.

Olasılıksal sınıflama yöntemleri, istatistiksel çıkarımlar yaparak gözleme uyan en iyi sınıfa karar verirler. Diğer algoritmalar doğrudan en iyi sınıfı tespit ederken, olasılıksal sınıflama yöntemlerinde her bir gözlemin aday sınıflara ait olma olasılıkları tespit edilir ve en yüksek olasılığa sahip sınıf tercih edilir. Bu tür yöntemlere örnek olarak lojistik regresyon ve bayes sınıflayıcısı verilebilir.

Sınıflayıcılar, lineer ve lineer olmayan olmak üzere ikiye ayrılmaktadırlar. Lineer sınıflayıcılar, her bir gözlem için olası sınıflara ait bir skor elde eden lineer fonksiyonları kullanmaktadır. Lineer sınıflayıcılarda en yüksek skora sahip olan sınıf tercih edilir. Bu skor fonksiyonlarına lineer tahmin edici fonksiyon denir ve aşağıdaki genel formda gösterilebilir;

$$G(\mathbf{X}_i, k) = \boldsymbol{\beta}_k \cdot \mathbf{X}_i. \qquad \textbf{(2.1)}$$

Burada $\mathbf{X}_i$, i-nci gözleme ait bağımsız değişkenler vektörü; $\boldsymbol{\beta}_k$, k sınıfına karşılık gelen ağırlıklar vektörü ve $G(\mathbf{X}_i, k)$, i-nci gözlemin k sınıfı ile olan ilişkisini belirten skordur. Bu temel yapıya sahip algoritmalara lineer sınıflayıcılar denir. Lineer sınıflayıcıların lineer olmayan sınıflayıcılardan farkı, optimal katsayıları belirleme yöntemi ve skorları yorumlama şeklidir. Lineer sınıflayıcılara örnek olarak lojistik regresyon, lineer diskriminant analizi, probit regresyonu ve çok katmanlı algılayıcılar verilebilir.

Sınıflama çalışmaları iki sınıflı (binary classification) ve ikiden fazla (multiclass classification) olmak üzere ikiye ayrılabilir. Bu çalışmada iki sınıflı sınıflama modelleri kurulmuştur.

2.1. Sınıf Dengesizliği

Gerçek verilerin neredeyse tümünde, sınıflara ait gözlem sayıları eşit olmayan dağılım göstermektedir yani sınıf dengesizliği vardır. Fakat sınıf dengesizliği tabiri, önemli dengesizlik oranın mevcut olduğu durumlar için kullanılmaktadır. Bu durum sınıflar-arası dengesizlik olarak adlandırılır ve sıklıkla karşılaşılan bir durumdur (He ve Garcia, 2009). Gözlemin sınıf gözlem sayıları arasındaki 100, 1000 ve 10000 gibi oranlar, aşırı dengesizlik durumuna işaret etmektedir ve sınıf dengesizliğinin varlığı kabul edilerek sınıflama yapılır. Bu durumlarda sınıflardan birisi, diğerine göre çok daha fazla temsil edilmiştir. Tüm sınıf dengesizliği problemleri iki sınıflı olmak zorunda değildir. Çok sınıflı dengesizlik problemleri de mevcuttur.

Bir diğer sınıf dengesizliği türü ise sınıflar-içi dengesizliktir. Bu dengesizlik türünde bir sınıfın kendi içerisinde ender gözlemler bulunmaktadır. Verilerin kısıtlı olmasından dolayı kaynaklanan bu problemde sınıflar-arası dengesizlik olsa da olmasa da sınıflayıcı zorlanacaktır. Sınıflar-içi dengesizlik problemi daha çok mevcut verinin dağılımı ile ilgili bir problemdir (He ve Garcia, 2009).

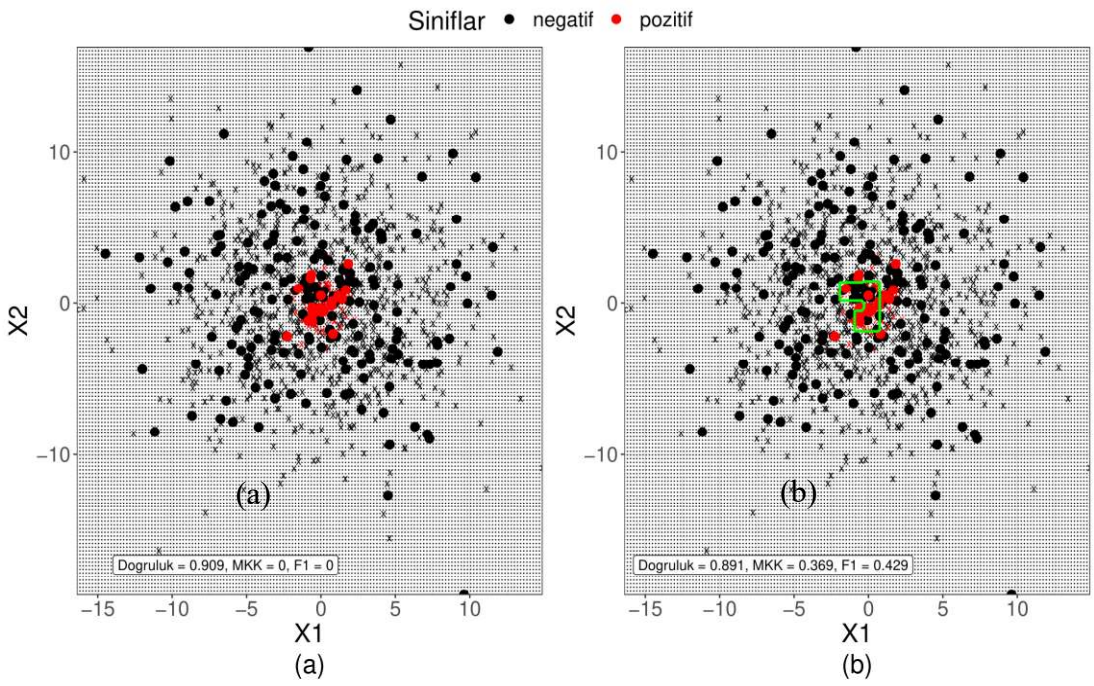
Sınıf dengesizliği durumunda sınıflayıcılar çoğunluk olan sınıftan yana yanlılık göstermektedir. Azınlık sınıf “pozitif”, çoğunluk sınıf ise “negatif” olarak ifade edilmektedir. Bu bölümden itibaren sınıflar bu şekilde adlandırılacaktır.

Sınıf dengesizliği olan Şekil 2.1’dekimodellere bakıldığında (a)’dasadece negatif sınıfı tercih eden bir model görülmektedir. Burada sınıf dengesizlik modeline

ait pozitif sınıf gözlemleri tahmin edilmeye çalışılmış, kötü bir model olmasına rağmen doğruluk oranını %90,9 olarak bulmuştur. Yani sadece %9,1'lik bir hata yapmıştır. Bu performans çok iyi gibi gözükse de hiçbir pozitif sınıfın doğru tahmin edilemediği ortadadır. Şekil 2.1(b)'de ise pozitif sınıfı tahmin etmeye çalışan daha iyi başka bir model görülmektedir. Bu modelde pozitif sınıftan doğru tahminler yapılmıştır fakat genel olarak %89,1 doğruluk oranı yakalanabilmiştir. Şekil 2.1(b)'deki model daha iyi olmasına rağmen, %10,9 hata oranı Şekil 2.1(a)'daki modelden daha yüksektir. Ayrıca ileride bahsedilecek olan MKK ve F_1 performans ölçütlerinin sınıf dengesizliğinde daha doğru ölçüler gösterdiği görülmektedir. Sınıf dengesizliği hakkında Şekil 2.1'den şu yorumlar yapılabilir:

- Sınıf dengesizliği durumunda, sınıflayıcılar negatif sınıftan yana ağırlık göstererek model kurmaktadır.

- Sınıf dengesizliği durumunda, doğruluk veya hata oranı iyi bir model performans ölçüsü değildir. Farklı ölçütler kullanılmalıdır.



Şekil 2.1 İki boyutlu iki sınıflı örnek bir veri setine ait farklı iki model için nokta grafikleri ve karar bölgeleri. Çarpı işareti ile belirtilen noktalar eğitim setine, içi dolu daire ile belirtilen noktalar ise test setine ait gözlemlerdir. Siyah bölge negatif sınıfın, kırmızı bölge pozitif sınıfın karar bölgesidir.

2.2. Sınıf Dengesizliği Sorununu Çözmede Kullanılan Yöntemler

Sınıf dengesizliği sorunu araştırmacıların uzun yıllardır ilgilendikleri bir problemdir. Bu problemi çözmede kullanılan yöntemler; yeniden örnekleme, maliyete duyarlı sınıflama ve topluluk algoritmaları başlıkları altında toplanabilir. Yeniden örnekleme yöntemlerinde veri dengelenene kadar, veri türetimi veya veri azaltma işlemi uygulanır ve böylece denge sağlanarak problem giderilmeye çalışılır (Das vd, 2014). Maliyete duyarlı sınıflama yönteminde, pozitif sınıfa, negatif sınıftan daha fazla ağırlık vererek dengesizliğin etkilerini ortadan kaldırılmak amaçlanır. Topluluk algoritmaları, tek bir hipotez kurarak model oluşturan geleneksel makine öğrenme yöntemlerinin aksine, birçok hipotez kurar ve bir fikir birliğinin oluştuğu bir model oluşturur. Böylece topluluk algoritmaları, içerisinde bulundurduğu modellerin hepsinden daha güçlü bir genelleme gücüne ulaşır ve bu sayede sınıf dengesizliğinden çok daha az etkilenir (Fernandes ve Carvalho, 2019). Bu çalışmada söz konusu yöntemlerden yeniden örnekleme yöntemleri üzerinde durulmuştur.

2.2.1. Yeniden örnekleme yöntemleri

Yeniden örnekleme yöntemleri, veri setindeki gözlemlere ait sınıflar daha dengeli hale gelecek şekilde veri setinin modifiye edildiği ön işlem yöntemleridir (Fernàndez vd, 2018). Yeniden örnekleme yöntemleriyle sınıf dengesini yakalayarak model kurma işleminin kullanışlı bir çözüm yöntemi olduğu daha önceki çalışmalarda gösterilmiştir (Chawla vd, 2008; Estabrooks ve Japkowicz, 2004; Garcia vd, 2012). Yeniden örnekleme yöntemleri üç ana başlıkta toplanabilir:

- Az örnekleme (Undersampling)
- Aşırı örnekleme (Oversampling)
- Hibrit yöntemler

Az örnekleme, negatif sınıfa ait gözlemler içerisinden alt küme seçerek pozitif sınıf ile denge sağlama işlemidir. Aşırı örneklemede, pozitif sınıfa ait gözlemler artırılır ve sınıf dengesi sağlanır. Hibrit yöntemler ise adından da anlaşılacağı gibi hem az örnekleme hem de aşırı örnekleme prosedürünü beraber kullanan yöntemlerdir. Yeniden örnekleme yöntemlerinin avantajı, kullanılacak olan sınıflayıcıdan bağımsız olmalarıdır (Fernàndez vd, 2018).

2.2.2. Performans ölçütleri

Sınıf dengesizliği ile olan bir veri setine, sınıflama modeli kurulurken bu modeli değerlendirecek olan performans ölçütünü seçmek çok önemlidir. Sınıflayıcıların performanslarını ölçerken en çok kullanılan ölçüt doğruluk veya hata oranıdır. Bu iki ölçüt birçok durumda işe yarayabilir. Fakat sınıf dengesizliği mevcut iken pozitif sınıfa ait doğruluk oranının da yüksek olması istenebilmektedir (Daskalaki vd, 2006). Doğruluk ve hata oranı, pozitif sınıfa ait doğruluk oranını göz ardı etmektedir. Dolayısıyla alternatif performans ölçütlerine bakmak gerekir. Performans ölçütleri hata matrisine bağlı ölçütler ve grafiksel performans ölçütleri olarak iki gruba ayrılabilir.

İki sınıflı bir sınıflayıcıya ait hata matrisi Çizelge 2.1'de verilmiştir. Sütunlar sınıflayıcı tahminlerini satırlar ise gerçek sınıf değerini temsil etmektedir. Çizelge 2.2.'de hata matrisi kullanılarak hesaplanan ve sıklıkla literatürde kullanılan performans ölçütleri verilmiştir. En sık kullanılan ölçüt doğruluk oranı ve hata oranıdır. İki ölçüt aslında birbirine denktir. Sınıf dengesizliği durumunda doğruluk oranı, pozitif sınıftan çok, negatif sınıfa ağırlık verdiği için yanıltıcı bir ölçüttür (Bekkar vd, 2013). Bu nedenle dengeli doğruluk oranı, G-ortalama, F_1 skoru ve Matthew'in korelasyon katsayısı (MKK) gibi ölçütler kullanılmaktadır. Doğruluk oranı, G-ortalama ve MKK her iki sınıfa eşit değer verirken F_1 skorunun yüksek olması modelin pozitif sınıfta daha iyi tahmin yaptığı anlamına gelmektedir (Bekkar vd, 2013).

Çizelge 2.1 Karışıklık matrisi (Confusion matrix)

		TAHMİN	
		Pozitif	Negatif
GERÇEK	Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
	Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Çizelge 2.2 Sınıflama modellerini değerlendirirken kullanılan performans ölçütleri

Ölçüt	Denklem	Aralık	Yorumu
Doğruluk Oranı	$\frac{DP+DN}{DP+DN+YP+YN}$	[0,1]	Doğru tahmin oranı
Hata Oranı	$1-\frac{DP+DN}{DP+DN+YP+YN}$	[0,1]	Yanlış tahmin oranı
Doğru Pozitif Oranı (Hassaslık) (Sensitivity)	$\frac{DP}{DP+DN}$	[0,1]	Pozitif gözlemlerin doğruluk oranı
Doğru Negatif Oranı (Belirginlik) (Specificity)	$\frac{DN}{YN+DN}$	[0,1]	Negatif gözlemlerin doğruluk oranı
Pozitif Tahmin Değeri (Kesinlik) (Precision)	$\frac{DP}{DP+YP}$	[0,1]	Pozitif tahminlerin doğruluk oranı
Negatif Tahmin Değeri	$\frac{DN}{DN+YN}$	[0,1]	Negatif tahminlerin doğruluk oranı
Dengeli Doğruluk Oranı	$\frac{1}{2}\left(\frac{DP}{DP+YP}+\frac{DN}{DN+YN}\right)$	[0,1]	Pozitif ve negatif tahmin değerlerinin aritmetik ortalaması
G-ortalama	$\sqrt{\frac{DP}{DP+DN}\times\frac{DN}{YN+DN}}$	[0,1]	Hassaslık ve belirginliğin geometrik ortalaması
F_1 skoru	$2\times\left(\frac{\frac{DP}{DP+YP}\times\frac{DP}{D}}{\frac{DP}{DP+YP}+\frac{DP}{D}}\right)$	[0,1]	Keskinlik ve hassaslığın harmonik ortalaması
Matthew'in Korelasyon Katsayısı (MKK)	$\frac{DP\times DN-YP\times YN}{\sqrt{(DP+YP)(DP+YN)(DN+YP)(DN+YN)}}$	[-1,1]	Tahminler ile gerçek sınıflar arasındaki ilişki seviyesi

Grafiksel performans ölçütü olarak en çok kullanılan yöntem ROC analizidir. Genellikle ROC eğrisi altında kalan alan (EAA) hesaplanarak bir değerlendirme yapılmaktadır. Hesaplanışı aşağıdaki gibi verilebilir (Haklı, 2018):

$$\widehat{EAA}=\int_0^1 \widehat{ROC}(t)dt. \qquad \textbf{(2.2)}$$

EAA ölçütü sınıf dengesizliği durumunda literatürde sıklıkla kullanılmaktadır. İstatistiksel olarak bir sınıflayıcının EAA'sı, o sınıflayıcının seçilen pozitif bir gözlemi negatif gözlemden daha güçlü seçme olasılığını verir (Bekkar vd, 2013).

Sınıf dengesizliği durumunda, literatürde en çok kullanılan performans ölçütleri F_1 skoru, G-ortalama ve EAA ölçütleridir. Bu ölçütler hangi sınıfın pozitif hangi sınıfın negatif sınıf olduğunu bilmek isteyen ölçütler olduğu için pozitif sınıfa yönelik değerlendirme yaparlar. Matthew'in korelasyon katsayısı (MKK) sınıflar arasında ayrım yapmayan ve pozitif sınıfa bağlı olmayan bir performans ölçütüdür (Chicco, 2017). Bu çalışmadaki modellerin performansları ölçülürken F_1 skoru ve MKK kullanılmıştır. MKK'da -1 en kötü, 1 en iyi performans, 0 ise rastgele sınıflayıcı anlamına gelmektedir. F_1 skoru pozitif ve negatif sınıf için ayrı ayrı hesaplanır. Çalışmada pozitif sınıf için hesaplanan F_1 skoru kullanılmıştır.

3. BOOSTING

Boosting son yirmi yıl içerisinde literatüre dahil edilmiş en güçlü öğrenme fikirlerinden birisidir. Başlangıçta sınıflama problemleri için tasarlanmış olsa da regresyon problemlerinde de kullanılmaktadır. Boosting prosedüründeki ana fikir, birçok "zayıf" sınıflandırıcıya ait çıktıları birleştirerek güçlü bir "topluluk" üretmektir (Hastie vd, 2005). Bu bakış açısıyla bakıldığında torbalama gibi diğer topluluk algoritmalarına benzerlik gösterir. Fakat temelde Boosting diğer torbalama algoritmalarından ayrılmaktadır.

Boosting prosedüründe mekanizmaya öncelikle y_i etiket değişkeni olacak şekilde $(x_1, y_1), \dots, (x_N, y_N)$ etiketlenmiş eğitim seti verilir. $t = 1, \dots, T$ iterasyonlarının her birinde mekanizma gözlemler üzerinde bir w_t dağılımı oluşturur ve bu w_t'ye bağlı olarak ε_t hatalı bir h_t hipotezi talep eder. Buradaki hata $\varepsilon_t = \Pr_{i\sim w_t}[h_t(x_i) \neq y_i]$ olarak verilebilir. Böylece w_t dağılımı her bir gözlemin (yani gözlemin) mevcut turdaki kısmi ağırlığını belirler. T iterasyondan sonra mekanizma zayıf hipotezleri birleştirerek tek bir tahmin kuralı oluşturur.

Popüler boosting algoritmalarından olan uyarlamalı boosting (AdaBoost), her bir modeli kurarken ilerlemeli bir şekilde gözlemlere ağırlık verir.

3.1. Uyarlamalı Boosting (AdaBoost)

Uyarlamalı boosting Freund ve Schapire tarafından 1996'da önerilmiştir. Genelleştirilebilmesi ve hızlı performansa ve düşük karmaşıklığa sahip olması nedeniyle en popüler ve etkili sınıflama araçlarından birisi olmuştur (Vezhnevets ve Vezhnevets, 2005).

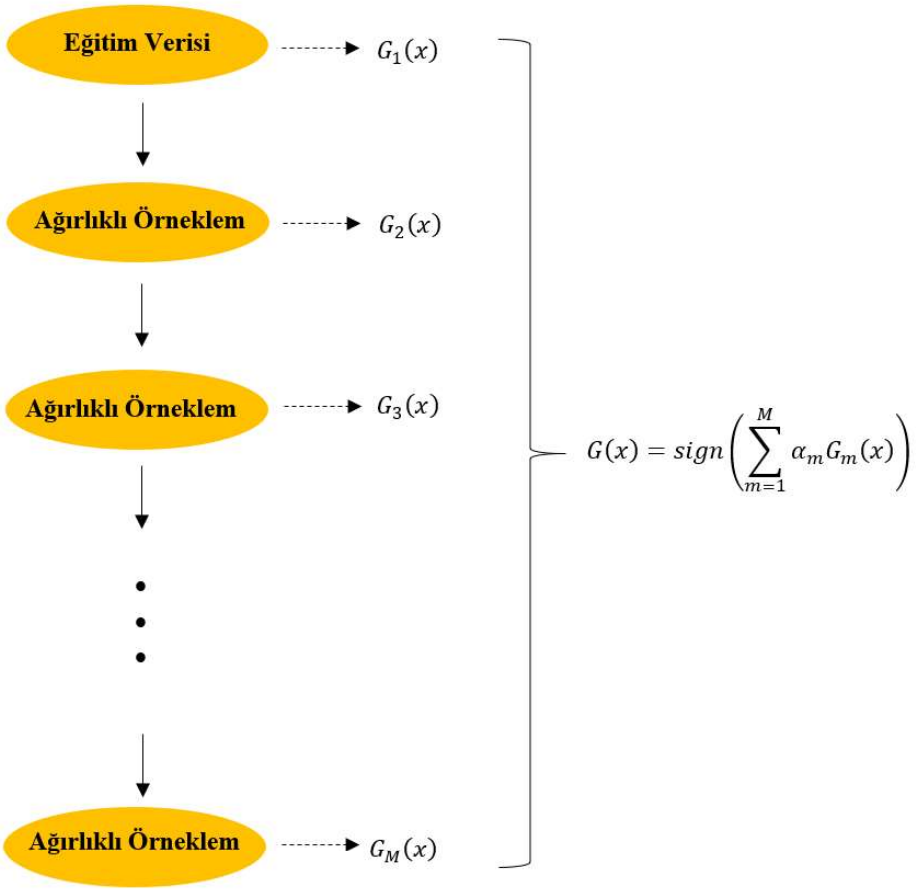
İki sınıflı bir sınıflama problemi ele alınsın. $Y \in \{-1,1\}$ olarak kodlanmış çıktı değişkeni olsun. X bağımsız değişkenleri verilsin. $G(X)$, $\{-1,1\}$ değerlerinden birini alan tahmin üreten bir fonksiyon olsun. Buna göre eğitim verisinin hata oranı;

$$\overline{hata} = \frac{1}{N}\sum_{i=1}^{N} I\big(y_i \neq G(x_i)\big) \qquad \textbf{(3.1)}$$

biçiminde hesaplanır. Gelecek tahminlerin beklenen hata oranı $E_{XY}I\big(Y \neq G(X)\big)$ olur. Zayıf sınıflayıcı, rastgele tahminden biraz daha iyi hata oranına sahip sınıflayıcıdır. Boostingin amacı, veri setinin tekrar tekrar değiştirilmiş versiyonlarına zayıf sınıflama algoritmalarını sırasıyla uygulamaktır. Bu nedenle $G_m(x), m = 1,2,\dots,M$ şeklinde zayıf sınıflayıcılar dizisi ortaya çıkar. Tüm bu sınıflayıcılardan elde edilen tahminler, ağırlıklı oylama yolu ile son tahmini aşağıdaki gibi oluşturur:

$$G(x) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right). \qquad \textbf{(3.2)}$$

Burada $\alpha_1, \alpha_2, \dots, \alpha_M$ boosting algoritması ile hesaplanan, modellere ait ağırlıklardır. Bu ağırlıklar kendilerine karşılık gelen modele ait $G_m(x)$'in katkısını ağırlıklandırır ve bunun etkisi ile model dizisindeki daha isabetli sınıflayıcılar daha yüksek etki kazanır. Uyarlamalı boosting'in çalışma şeması Şekil 3.1'de görülmektedir.



Şekil 3.1 Uyarlamalı boosting sınıflayıcısının çalışma şeması. Sınıflayıcılar ağırlıklandırılmış veri setleri üzerinde eğitilir ve daha sonra ağırlıklandırılmış bir şekilde son sınıflayıcı olarak birleştirilir.

Boosting adımlarındaki veri setleri, her bir (x_i, y_i) eğitim gözlemine ait $w_1, w_2, \dots, w_N$ ağırlıkları kullanılarak modifiye edilir. Başlangıçta tüm ağırlıklar eşit ve $w_i = 1/N$'dir. Bu nedenle ilk adım, sınıflayıcıyı normal bir şekilde eğitir. $m = 2,3, \dots, M$ iterasyon numarası olacak şekilde, her bir başarılı iterasyonda gözlem ağırlıkları bireysel olarak düzenlenir. Bu düzenleme örnekleme ve filtreleme şeklinde iki yolla yapılabilir (Freund ve Schapire, 1997). Bu çalışmada en sık kullanılan yöntem olan örnekleme yoluyla düzenleme ele alınacaktır. Eğitim veri seti içerisinden gözlemlere ait $w_1, w_2, \dots, w_N$ ağırlıkları iki şekilde kullanılabilir. Birincisinde bu ağırlıklar kullanılarak bootstrap ile "ağırlıklı örneklem" elde edilir ve bu örneklem üzerinden sınıflayıcı eğitilir. İkinci yöntemde sınıflayıcı, gözlem ağırlığı alabiliyorsa bu durumda $w_1, w_2, \dots, w_N$ ile sınıflayıcı eğitilir. m-nci adımda $G_m(x)$ sınıflayıcısına ait ağırlık olan α_m, modelin ağırlıklı hatası hata_m'ye göre aşağıdaki gibi hesaplanır:

$$\overline{\text{hata}}_m = \frac{\sum_{i=1}^{N} w_i I\left(y_i \neq G_m(x_i)\right)}{\sum_{i=1}^{N} w_i} \quad \textbf{(3.3)}$$

$$\alpha_m = \log\left(\left(1 - \overline{\text{hata}}_m\right)/\overline{\text{hata}}_m\right). \quad \textbf{(3.4)}$$

Bulunan model ağırlığına göre gözlemlere ait ağırlıklar aşağıdaki gibi güncellenir:

$$w_i \leftarrow w_i \cdot \exp\left[\alpha_m \cdot I\left(y_i \neq G_m(x_i)\right)\right], i = 1,2,\ldots,N. \quad \textbf{(3.5)}$$

Burada $G_m(x)$ tarafından yanlış sınıflandırılmış olan gözlemlerin ağırlıkları artırılır, aynı şekilde doğru tahmin edilen gözlemlerin ağırlıkları ise azaltılır. Bu işlemler M kez tekrar edilir.

Algoritma 3.1 Uyarlamalı Boosting

1. Girdi: $w_i = 1/N, i = 1,2,\ldots,N$ başlangıç gözlem ağırlıkları.
2. $m = 1,2,\ldots,M$
   a) Eğitim verisine w_i ağırlıklarını kullanarak $G_m(x)$ sınıflayıcısını uydur.
   b) Hesapla:
   $$\text{hata}_m = \frac{\sum_{i=1}^{N} w_i I\left(y_i \neq G_m(x_i)\right)}{\sum_{i=1}^{N} w_i}.$$
   c) Hesapla: $\alpha_m = \log\left((1 - \text{hata}_m)/\text{hata}_m\right)$.
   d) $w_i \leftarrow w_i \times \exp\left[\alpha_m \cdot I\left(y_i \neq G_m(x_i)\right)\right], i = 1,2,\ldots,N$
3. Çıktı: $G(x) = \text{sign}[\sum_{m=1}^{M} \alpha_m G_m(x)]$

Algoritma 3.1'de Uyarlamalı Boosting algoritması görülmektedir. 2a satırında $G_m(x)$ sınıflayıcısı üzerinde ağırlıklı gözlemler kullanılır. 2b satırında ağırlıklı hata oranı hesaplanır. 2c'de verilen $G_m(x)$ için son sınıflayıcı olan $G(x)$'de kullanılacak α_m ağırlığını hesaplar. 2d'de her bir gözleme ait ağırlıklar sıradaki iterasyon için güncellenir. $G_m(x)$ tarafından yanlış sınıflandırılan gözlemlerin ağırlıkları, $\exp(\alpha_m)$ değerine bağlı olarak sıradaki $G_{m+1}(x)$ sınıflayıcısında kısmi etkisini artıracak şekilde değiştirilir. Bu çalışmada uyarlamalı boosting yöntemindeki gözlem ağırlıklarından faydalanılmıştır.

4. YENİDEN ÖRNEKLEME

4.1. Rastgele Aşırı Örnekleme

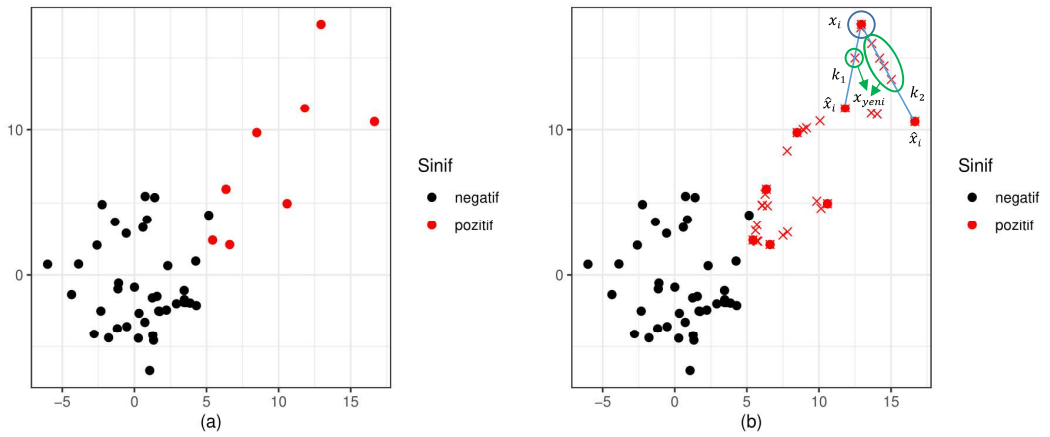
Rastgele aşırı örnekleme yöntemi pozitif sınıfa ait gözlemlerden örneklem çekerek veri setine ekleme işlemidir (He ve Garcia, 2009). Pozitif sınıfa ait gözlemler denge sağlanacak miktarda kopyalanır ve orijinal veri setine eklenir. Bu yöntem hem uygulaması hem de anlaması basit bir yöntemdir. Bu yöntemde\ aynı gözlemlerin tekrar etmesi sınıflayıcı kurallarının çok belirli olmasına yol açar. Bu nedenle sınıflayıcıların eğitim veri setine aşırı uyumu görülebilir. Bu da test veri setinde kötü performans görülebileceği anlamına gelmektedir.

4.2. Sentetik Azınlık Aşırı Örnekleme Yöntemi (SMOTE)

SMOTE, birçok çalışmada başarı göstermiş güçlü bir yöntemdir. Chawla ve arkadaşları (2002) tarafından önerilmiştir. SMOTE algoritması, mevcut pozitif gözlemler arasında nitelik uzayındaki boşluklara göre yapay veri üretir (He ve Garcia, 2009). $X_{poz} \in X$ alt kümesi için her biri $x_i \in X_{poz}$ olacak şekilde belirli bir gözlemlerin en yakın K adet komşusu ele alınsın. $X_{poz-\{i\}}$ ile x_i arasında$n_{poz} - 1$ adet Öklid uzaklığı hesaplanır. Bu $n_{poz} - 1$ uzaklık içerisinde, en küçük miktarı veren $X_{poz-\{i\}}$'nin K elemanı, en yakın K komşudur. Sentetik örnek üretmek için en yakın K komşudan rastgele birisi seçilir. Seçilen komşuya karşılık gelen vektör farkı, [0,1] aralığındaki rastgele bir sayı ile çarpılır. Sonuçta

$$x_{yeni} = x_i + (\hat{x}_i - x_i) \times \delta \qquad \textbf{(4.1)}$$

şeklinde sentetik veri türetilir. Burada $x_i \in X_{poz}$, ele alınan pozitif sınıf gözlemidir. $\hat{x}_i$, x_i için bulunan en yakın komşulardan birisidir. $\hat{x}_i \in X_{poz}$ ve $\delta \in [0,1]$'dir. Bu denklemle türetilen sentetik örnek (x_{yeni}), x_i ile rastgele seçilmiş $\hat{x}_i$ en yakın komşusu arasında çizgi üzerinde bir nokta olur. SMOTE algoritması Algoritma 4.1'de gösterilmiştir. Algoritma 4.1'te 2. adımda değişkenler [0,1] aralığına indirgenmiş ve 11. adımda tekrar normal haline getirilmiştir. Burada amaç, Öklid uzaklıkları hesaplanırken farklı değişkenlere ait ölçü birimlerinden etkilenmemektir.



Şekil 4.1 (a) Dengesiz sınıf dağılıma sahip iki boyutlu veri seti gözlemi. (b) Bu veri setine SMOTE (K = 2) uygulandıktan sonra (sentetik veriler "×" sembolü ile gösterilmiştir).

Algoritma 4.1 Uygulanan SMOTE kaba kodu (tam denge)

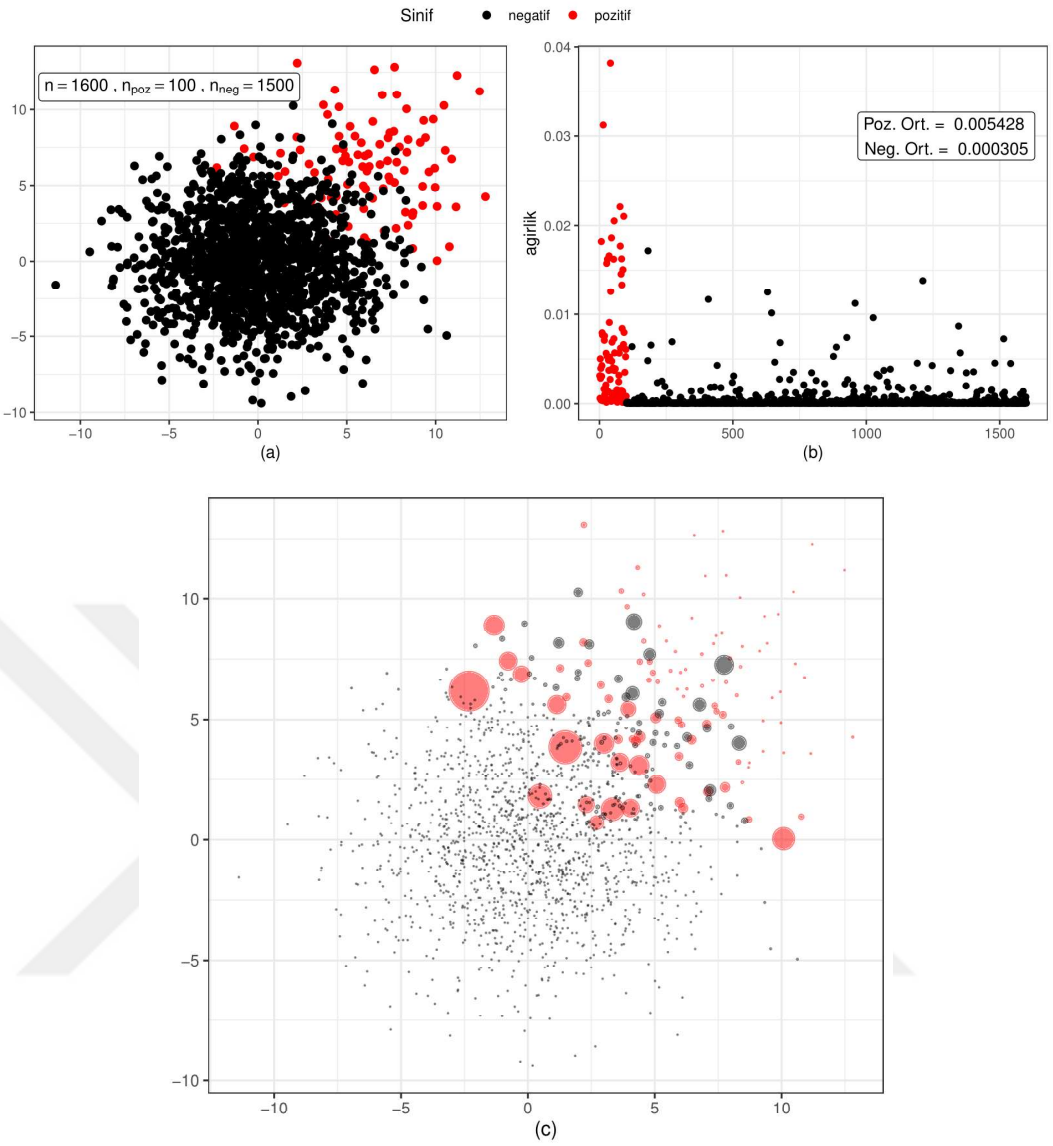
1. Girdi: $X_{n\times p}$, $Y_{n\times 1}$ ve K komşu sayısı.
2. $i = 1,2,\dots,p$ için X için
   $minler \leftarrow \min(x_{*i})$ leri içeren vektör
   $makslar \leftarrow \text{maks}(x_{*i})$ leri içeren vektör
   $$x_{*i} = \frac{x_{*i} - minler_i}{makslar_i - minler_i}$$
   hesaplanır ve tüm değişkenler için maksimum ve minimum değerler tutulur.
   NOT: Bu dönüşümü yapmanın amacı daha gerçekçi yakın komşular elde etmektir.
3. $N_{sentetik} \leftarrow N_{neg} - N_{poz}$
4. $C \leftarrow \left[\frac{N_{sentetik}}{N_{poz}}\right]$ değerlerinden oluşan N_{poz} uzunluğunda vektör.
5. C içerisinden rastgele $N_{sentetik} - toplam(C)$ gözlemi 1 artır.
   NOT: Tam denge sağlandı.
6. i = 1'den N_{poz}'a kadar
   $k_i \leftarrow X$ içerisinde X_i'ye kendisi dışındaki en yakın K komşunun indeksleri
   NOT: k, $N_{poz} \times K$'lık bir matris.
7. $X_{smote} \leftarrow N_{sentetik} \times p$'lik boş matris.
8. i = 1'den N_{poz}'a kadar
   a) $eyklar.id \leftarrow 1$ ile K arasında C_i adet sayı çek
   b) $eyklar \leftarrow k_i$ içerisinden $eyklar.id$ indeksli komşulara ait gözlemleri yerine koymalı şekilde çek
   c) $\delta \leftarrow 0$ ile 1 arasında C_i adet sayı türet
   d) $x_{yeni} = x_i + (eyklar - x_i) \times \delta$
   e) x_{yeni}'yi X_{smote}'a yerleştir.
9. $X_{yeni} \leftarrow X_{smote}$ ile X'i birleştir.
10. $N_{yeni} = N + N_{sentetik}$
11. $i = 1$'den p'ye kadar X_{yeni} için
    $$x_{*i} = x_{*i} \times (makslar_i - minler_i) + minler_i$$
    şeklinde 2. Adımda uygulanan dönüşüm geri alınır.
12. Çıktı: X_{yeni}

Şekil 4.1'de SMOTE uygulamasına ait bir örnek görülmektedir. Şekil 4.1(a)'da dengesiz sınıf dağılımına sahip iki boyutlu bir veri seti vardır. Kırmızı renkli noktalar pozitif, siyah renkli noktalar negatif sınıfa ait gözlemleri temsil etmektedir. Şekil 4.1(b)'de ise bu veri setine $K = 2$ komşu sayılı SMOTE uygulaması görülmektedir.

4.3. Boosting ile SMOTE (B. SMOTE)

SMOTE, bütün pozitif gözlem noktaları arasında ayırt etmeksizin veri türetir. Veri türetilirken dikkate alınan pozitif sınıf gözlemlerinin, diğer pozitif gözlemlerden uzak olması ve negatif gözlemlerin arasında bulunması gibi durumlarda gözlemler arasında veri türetmek, gürültü oluşmasına neden olabilmektedir. Bu durum normalde veri setinde olmayan problemlerin ortaya çıkmasına ve sınıflama modelinin yanılmasına neden olur.

Genel olarak SMOTE'un iki problemi vardır. Bunların birincisi, biraz önce bahsedildiği gibi, gürültü türetebilmesidir. Diğer sorunu ise sentetik veri üretimi için bağ kurulacak komşu sayısının doğru seçilmemesidir.



Şekil 4.2 (a) $X_{neg} \sim N(0,3)$ ve $X_{poz} \sim N(6,3)$ gözlemlere sahip iki sınıflı ve iki boyutlu örnek veri seti. (b) Veri seti 100 iterasyonluk uyarlamalı boosting sürecinden geçtikten sonra gözlemlere ait elde edilen ağırlıklar. (c) Bulunan ağırlıklara göre gözlem boyutları farklılık gösterdiğinde aynı veri seti.

Bu iki sorundan birincisini ortadan kaldırmak için pozitif gözlemler arasındaki gürültülerin tespit edilmesine ihtiyaç vardır. Uyarlamalı boosting yöntemi ile elde edilen gözlemlere ait güncellenen ağırlıklar, gözlemlerin bir nevi tahmin edilebilirlik seviyesini göstermektedir. Bu ağırlıklar dikkate alınarak hangi gözlemin iyi tahmin edilebilir, hangi gözlemin ise kötü tahmin edilebilir olduğu anlaşılmaktadır. Şekil 4.2'de örnek iki boyutlu bir veri seti ve bu veri setine uyaralamalı boosting uygulanarak elde edilen ağırlıklar görülmektedir. Pozitif gözlemlerin ağırlıklarının negatif gözlemlere göre daha yüksekte olduğu görülmektedir. Çünkü, dengesiz sınıflar

durumunda model negatif sınıftan yana yanlılık göstereceği için pozitif gözlemlerin zor tahmin edilmektedir. Pozitif gözlemlere ait ağırlıkların ortalaması 0.005428, negatif gözlemlerin ortalama ağırlıkları 0.000305'tir. Bu sonuca göre, pozitif gözlemlerin daha zor tahmin edildikleri açıkça görülmektedir. Bu ağırlıklar yardımı ile hem pozitif hem de negatif sınıflara ait olan gürültüler tespit edilebilir. Şekil 3.2(c)'de ağırlıklara göre boyutları farklılık gösteren gözlemler görülmektedir. Uyarlamalı boosting ile diğer sınıf gözlemlerinin bölgesine girmiş olan pozitif ve negatif gözlemler kolayca anlaşılmaktadır. Başlangıç gözlem ağırlığı olan $1/N$'in üzerine çıkan gözlemlerin zor tahmin edilen gözlemlerdir. $1/N$'in altında olan gözlemlerin ise kolay gözlemler olduğu söylenebilir. Fakat sınıf dengesizliğinin negatif sınıftan yanlılığı bilindiği için iki sınıf gözlemlerini aynı şekilde ele almak doğru olmaz. Mevcut ağırlıklara bağlı olan yeni bir eşik değeri tespit edip o eşik değerine göre gözlemler gürültü veya gürültü değil olarak etiketlenebilir. Bu durum için mutlak bir eşik değeri söylemek mümkün değildir. Fakat bu çalışmada pozitif ve negatif sınıflar için

$$Eşik_{poz} = \frac{1}{N}\left(\frac{2 \times N_{neg}}{N}\right) = \frac{2 \times N_{neg}}{N^2} \quad \textbf{(4.2)}$$

$$Eşik_{neg} = \frac{1}{N}\left(\frac{2 \times N_{poz}}{N}\right) = \frac{2 \times N_{poz}}{N^2} \quad \textbf{(4.3)}$$

eşik değerleri önerilmiştir. Burada $1/N$ başlangıç ağırlığıdır. Parantez içideki kısım ise karşı sınıfa ait gözlem fazlalığının oranıdır. Denklem 4.2 ve 4.3'te sınıflar tam dengeli olursa $\frac{(2\times N_{neg})}{N} = \frac{(2\times N_{poz})}{N} = 1$ olur ve eşik değerleri $1/N$ olur. Dengesiz sınıf dağılımı varken $1/N$ eşik değeri olarak alınırsa, pozitif gözlemler yüksek ağırlığa sahip olacağı için gerçekte olduğundan daha fazla pozitif gözlem, gürültü olarak ele alınır. Dengesizlikten ötürü ortaya çıkan yanlılığı ortadan kaldırmak için pozitif gözlem eşik değerinin dengesizlik oranına bağlı olarak artırılması ve negatif gözlem eşik değerinin aynı şekilde azaltılması mantıklıdır. Sınıf dengesizliği durumunda negatif gözlem sayısı $N/2$'den fazla olduğu için $2 \times N_{neg} > N$ olacağından $Eşik_{poz} > 1/N$ olacaktır. Bu eşik değeri ile pozitif sınıflarda daha gerçekçi gürültü tespiti yapılabilir. Aynı şekilde $Eşik_{neg} < 1/N$ olacak ve negatif sınıflarda daha gerçekçi gürültü tespiti yapılabilir. Sadece dengesizlik oranını dikkate aldığı ve dengesizlik oranına bağlı bir problem mevcut olduğu için eşik değerinde yapılan bu

değişiklik, makul olarak kabul edilmiştir. Bu gürültü belirleme prosedürü Algoritma 4.2'de verilmiştir.

Algoritma 4.2 Gürültü Belirleme

1. Girdi: M, zayıf sınıflayıcı sayısı.
2. Eğitim verisini Algoritma 3.1'den 3. aşama dışında zayıf sınıflayıcı sayısı M olacak şekilde geçir.
3. $w_{son} \leftarrow w_M$ son gözlem ağırlığını al.
4. w_{poz}, w_{neg} sınıflara göre ağırlıkları ayır.
5. Çıktı: Eğer $w_{poz_i} > \frac{2 \times N_{neg}}{N^2}$, $X_{poz_i} \leftarrow$ gürültü, değilse, $X_{poz_i} \leftarrow$ gürültü değil, $i = 1,2, \dots, N_{poz}$
6. Çıktı: Eğer $w_{neg_j} > \frac{2 \times N_{poz}}{N^2}$, $X_{neg_j} \leftarrow$ gürültü, değilse, $X_{neg_j} \leftarrow$ gürültü değil, $j = 1,2, \dots, N_{neg}$

B. SMOTE uygularken SMOTE'da olduğu gibi komşu sayısı, kullanıcı tarafından girilen bir parametre değildir. Önerilen yöntemde öncelikle maksimum komşu sayısı veri setindeki gözlem sayısına bağlı olarak denklem 4.4'teki gibi belirlenmektedir:

$$K_{maks} = \left\lfloor \frac{n_{neg}}{n_{poz}} \right\rfloor. \quad \textbf{(4.4)}$$

Bu oran, sınıf dengesinin sağlanması için ortalama gözlem başına düşen üretilecek sentetik örnek sayısının aşağı yuvarlanmış halidir. SMOTE'da denklem 4.4 düşük (dengesizlik oranı düşük) ve K komşu sayısı fazla olduğunda, az gözlem rastgele çok fazla komşuya atanır. Bu durumda komşularla oluşan bağların fazlalığı, yeni gözlemlerin anlamsız şekilde saçılmasına neden olur. Aynı denklem 4.4'ün yüksek olması (denesizlik oranı yüksek), bir pozitif gözlemden çok fazla veri türetileceği anlamına geldiği için komşu bağların azlığı, yeni verilerin küçük bir bölgede sıkışacağı ve modeli aşırı uyuma götüreceği anlamına gelir. Denklem 4.4'ten maksimum komşu bağı sayısı belirlenerek bu problemin üstesinden gelmek amaçlanmıştır.

K_{maks} belirlendikten sonra, her bir pozitif gözlem için aynı komşu sayısı atamak yerine her bir pozitif gözlem için farklı K_i komşu sayısı belirlenir. Bu prosedür Algoritma 4.3'te verilmiştir. Daha önce gürültü ve gürültü değil olarak belirlenen

pozitif gözlemler bu algoritma sonrasında iyi, kötü ve yalnız pozitif olarak üç gruba ayrılacaktır.

Pozitif gözlemlerin gürültü olmayan veri setindeki K_{maks} komşusu yakından uzağa doğru sıralandığında, en yakın ilk gürültü olmayan negatif gözlemin hangisi olduğu tespit edilir. Bu gözlemden daha uzaktaki pozitif gözlemlerle eşleşme yapıldığında gürültü olmayan negatif gözlemlerin bulunduğu bölgede veri üretilme olasılığı oluşacaktır. Bu nedenle tespit edilen en yakın negatif gözlemin komşu sıra sayısından bir çıkarılarak o gözlemin yakın komşu sayısı belirlenir:

$$K_i = en\ yakın\ negatif\ gözlem\ sırası - 1, i = 1,2, \dots, N_{poz}. \qquad \textbf{(4.5)}$$

Tüm pozitif gözlemler için bu prosedür gerçekleştirilir. Ardından $K_i > 0$ olan pozitif gözlemler, iyi pozitif olarak etiketlenmektedir. $K_i = 0$ fakat daha önce gürültü olarak tespit edilmemiş gözlemler, yalnız pozitif olarak etiketlenmektedir. $K_i = 0$ ve daha önce gürültü olarak tespit edilmiş pozitif gözlemler, kötü pozitif olarak etiketlenmektedir. Burada kötü pozitif olarak belirlenmiş negatif gözlemler üzerinden veri türetilmeyecektir. Daha önce gürültü olarak tespit edilmiş fakat güvenli bir bölgede olan gözlemler üzerinden veri türetilebilecektir.

B. SMOTE veri türetim işlemi Algoritma 4.4'te verilmiştir. B. SMOTE, SMOTE'a çok benzese de farklı komşu sayılarının gözlemden gözleme değişmesi ve yalnız gözlemler için gözlem kopyalayacak olması açısından farklıdır. Yalnız gözlemler, gürültü olmadıkları ve negatif gözlemlerin arasında kaldıkları için kendi çevresinde veri türetmez. Sadece gerekli miktarda kopyalanır. Böylece yeni gürültü oluşmasına engel olur. İyi gözlemler K_i komşu ile bağlarını kurarak standart SMOTE prosedürü ile veri türetir.

Algoritma 4.3 Komşu Sayısı Belirleme

1. Girdi: X_{iyi} gürültü olmayan pozitif ve negatif gözlemleri.
2. X_{poz}'e X_{iyi}'de en yakın

$$K_{maks} = \left\lfloor \frac{N_{neg}}{N_{poz}} \right\rfloor$$

komşuyu tespit et.
Not: Böylece hiçbir zaman pozitif gözlemden gürültüye doğru bağ kurulmamış olacaktır. Fakat pozitif gürültüden iyi pozitiflere doğru bağın kurulması mümkün bırakılmıştır.

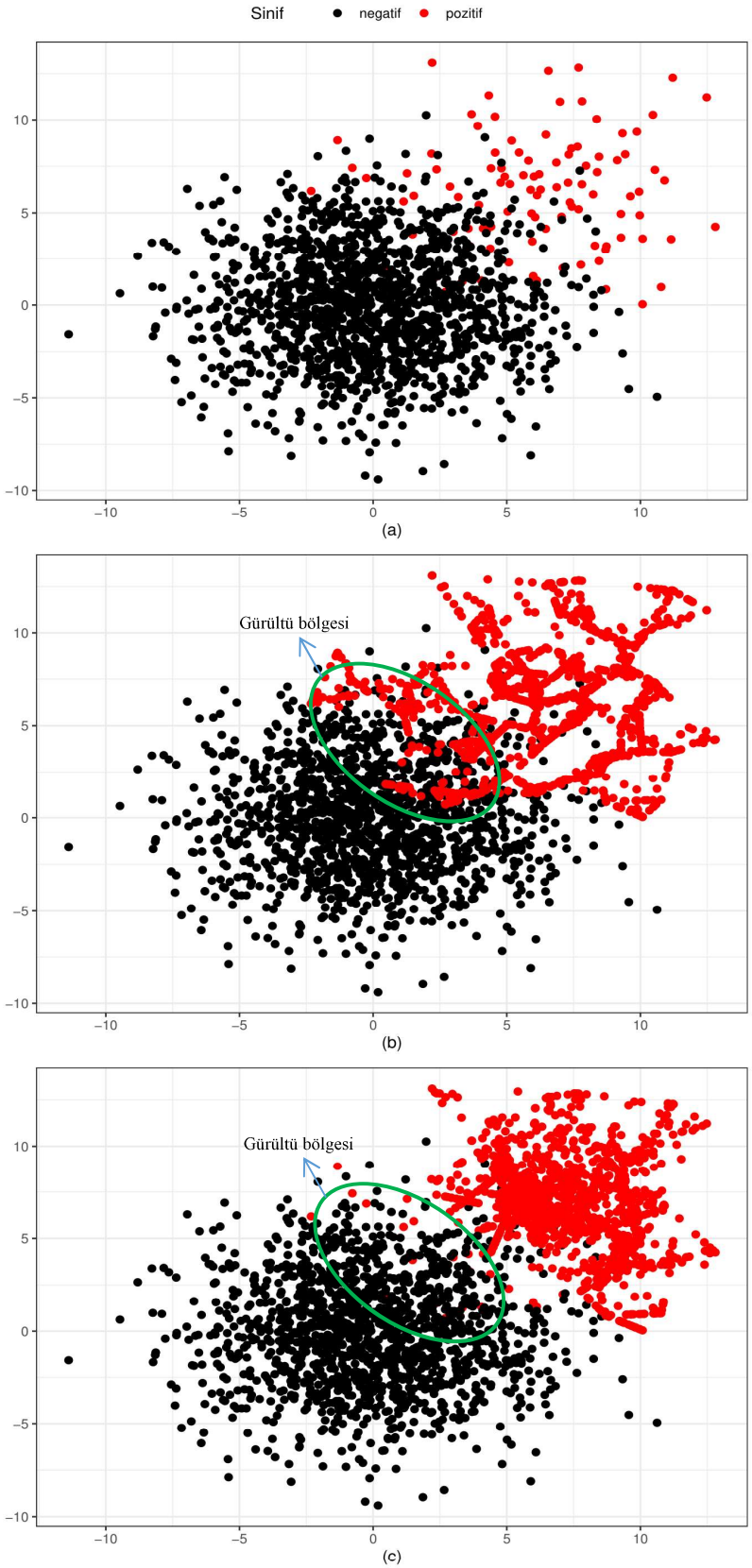
3. K_{maks} komşuyu yakından uzağa sırala ve ait oldukları sınıfları belirle.
4. $K_i \leftarrow$ Her bir X_{poz_i} için en K_{maks} komşudaki en yakın iyi negatif gözlemin sıra numarası -1
   Not: Bu şekilde gözlem uzayında bağ kurulacak en yakın komşulardan daha yakın negatif gözlem bulunmayacaktır. Böylece yeni üretilecek olan sentetik pozitif gözlemin herhangi bir iyi negatif gözlem üzerinden geçmesi mümkün olmayacaktır. Yakın komşulara bakarken negatif gürültüler dahil edilmediği için, negatif gürültü üzerinden geçmesi mümkün bırakılmıştır.
5. Çıktı: Eğer $K_i = 0$, ve X_{poz_i} bir pozitif gürültü ise $X_{poz_i} \leftarrow$ kötü pozitif
6. Çıktı: Eğer $K_i = 0$, fakat X_{poz_i} bir gürültü değil ise $X_{poz_i} \leftarrow$ yalnız pozitif
7. Çıktı: Eğer $K_i > 0$ ise $X_{poz_i} \leftarrow$ iyi pozitif

Algoritma 4.4 Boosting ile SMOTE veri türetimi (tam denge)

1. $X_{n\times p}$, $Y_{n\times 1}$ ve K komşu sayısı girilir.
2. $i = 1$'den p'ye kadar X için
   $minler \leftarrow \min(x_{*i})$ leri içeren vektör
   $makslar \leftarrow \text{maks}(x_{*i})$ leri içeren vektör
   $$x_{*i} = \frac{x_{*i} - minler_i}{makslar_i - minler_i}$$
   hesaplanır ve tüm değişkenler için maksimum ve minimum değerler tutulur.
   NOT: Bu dönüşümü yapmanın amacı daha gerçekçi yakın komşular elde etmektir.
3. Algoritma 4.2.
4. Algoritma 4.3.
5. $N_{sentetik} \leftarrow N_{neg} - N_{poz}$
6. $C \leftarrow$iyi pozitif ve yalnız pozitif indeksleri $\left\lceil \frac{N_{sentetik}}{N_{iyi\ poz}+N_{yalnız\ poz}} \right\rceil$, kötü pozitif indeksleri 0'dan oluşan N_{poz} uzunluğunda vektör.
7. C içerisinden iyi pozitif ve yalnız pozitif indeksleri arasında rastgele $N_{sentetik} - toplam(nvektör)$ gözlemi 1 artır.
   NOT: Tam denge sağlandı. Gürültü indekslerine karşılık değerler 0 oldu.
8. $X_{sentetik} \leftarrow N_{sentetik} \times p$'lik boş matris.

1) i = 1'den N_{poz}'a kadar
   a) Eğer x_i yalnız pozitif ise
      i. $x_{yeni} \leftarrow x_i$'yi C_i kez kopyala
      ii. x_{yeni}'yi $X_{sentetik}$'e yerleştir.
   b) Eğer x_i yalnız pozitif değil ise
      iii. $eyklar.id \leftarrow 1$ ile K arasında C_i adet sayı çek.
      iv. $eyklar \leftarrow k_i$ içerisinden $eyklar.id$ indeksli komşulara ait gözlemleri yerine koymalı şekilde çek.
      v. $\delta \leftarrow 0$ ile 1 arasında C_i adet sayı türet.
      vi. $x_{yeni} = x_i + (eyklar - x_i) \times \delta$
      vii. x_{yeni}'yi $X_{sentetik}$'e yerleştir.
9. $X_{yeni} \leftarrow X_{sentetik}$ ile X'i birleştir.
10. $i = 1$'den p'ye kadar X_{yeni} için
    $$x_{*i} = x_{*i} \times (makslar_i - minler_i) + minler_i$$
    şeklinde 2. adımda uygulanan dönüşüm geri alınır.
11. Çıktı: X_{yeni}

Şekil 4.3’te dengesiz sınıflı veri seti üzerinde SMOTE ve B. SMOTE uygulamasından sonra yeni veri setlerinin nokta grafikleri verilmiştir. Şekil 4.3(b)’de SMOTE uygulandığında ortaya çıkan gürültü gözlemleri görülmektedir. B. SMOTE uygulandığında gürültü bölgesinde veri türetilmez. Gözlemlerin kurulacak bağ sayısı güvenli komşu sayısı ile doğrudan ilişkili olduğu için güvenli bölgelerin daha iyi doldurulduğu görülmektedir.



Şekil 4.3 (a) Şekil 3.3'te verilen iki boyutlu dengesiz sınıf dağılımına sahip veri seti. (b) Bu veri setine SMOTE algoritması uygulandıktan sonra. (c) Boosting ile SMOTE algoritması uygulandıktan sonra.

5. UYGULAMA

5.1. Genel Bilgi

Öncelikle simülasyon çalışması 2 iki sınıflı veri seti 4 farklı sınıflayıcı ile kullanıldı. Simülasyon ve gerçek veri setleri ile ilgili bilgiler Kısım 5.2'de verildi. Ardından dengesiz sınıflara sahip 16 farklı iki sınıflı veri seti, 9 farklı sınıflayıcı kullanıldı. RAÖ, SMOTE ve B. SMOTE ve yeniden örnekleme yapılmadan kurulan modellerin performansları karşılaştırıldı. Performans ölçütü olarak F_1 skoru ve MKK kullanıldı. Bu ölçütlerin artması modelin iyi olduğu şeklinde yorumlandı. Kullanılan sınıflayıcılar, Çizelge 5.1'de görülmektedir.

Çizelge 5.1 Uygulamada kullanılan sınıflayıcılara ait bilgiler

Sınıflayıcı	R paketi	Parametreler
C4.5	RWeka	$C = 0.25, M = 2$
EYK	caret	$k = 5$
CART	rpart	$cp = 0.01$
Radyal DVM	kernlab	$sigma = 1, C = 0.25$
Lineer DVM	kernlab	$C = 0.25$
Yapay Sinir ağları	nnet	$size = 3, decay = 0.001$
Naive Bayes	klaR	$usekernel = TRUE, fL = 0, adjust = 0.5$
Lojistik regresyon	stats	Yok
MARS	earth	$nprune = 10, degree = 1$

Uygulamada gerçekleştirilen tüm prosedürler için R v3.6.1 programı (R Core Team, 2019), kullanıldı. R programında uygulamayı gerçekleştirebilmek ve gerekli grafik ve çizelge sonuçlarını alabilmek için RWeka (Hornik vd, 2009), rpart (Therneau vd, 2019), kernlab (Karatzoglou vd, 2004), nnet (Venables ve Ripley, 2002), klaR (Weihs vd, 2005), earth (Milborrow, 2019), mlbench (Leisch ve Dimitriadout, 2010), imbalance (Cordón vd, 2018), rattle.data (Williams, 2011), smotefamily (Siriseriwan, 2019), caret (Kuhn vd, 2019), ggplot2 (Wickham, 2016), xlsx (Dragulescu ve Arendt, 2018), GGally (Schloerke vd, 2018), data.table (Dowle ve Srinivasan, 2019), FNN (Beygelzimer vd, 2019), ada (Culp vd, 2016) ve lattice (Deepayan, 2008) paketlerinden faydalanıldı.

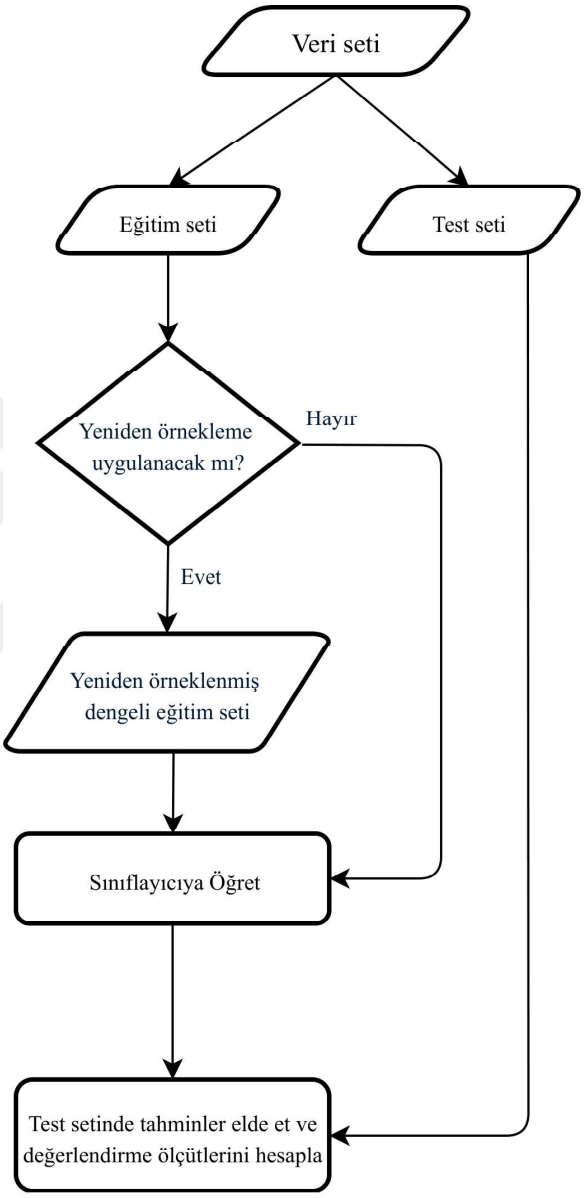
Uygulamada performans elde ederken kullanılan prosedür basitleştirilmiş şekilde Şekil 5.1'de görülmektedir. Bu işlem 10 katlı çapraz geçerlilik yapılırken her bir kat için tekrar edildi. Her bir durum için 10 farklı 10 katlı çapraz geçerlilik yapılarak 100'er sonuç ve 100'er farklı performans elde edildi. Bu 100 sonuç hem yeniden örnekleme yapılmadan hem de RAÖ, SMOTE ve B. SMOTE uygulanarak her bir durum için ayrı ayrı elde edildi. Performans belirlenirken elde edilen 100 sonucun aritmetik ortalaması alındı. Ayrıca her bir 100 durum için performanslarına göre yöntemlere sıra numaraları verildi ve bu sonuçlar da tutuldu. Bahsedilen sıra numaraları en iyi yönteme 1 ve en kötü yönteme 4 puan karşılık gelecek şekilde verildi. Aynı performansı gösterdikleri durumlarda ortalama sıra puanları verildi. Örneğin iki yaklaşım en iyi ve eşit performansa sahip ise her iki yönteme numara olarak $(1 + 2)/2 = 1.5$ verildi. Ardından gelen yönteme ise 3 değeri verildi. Her durumda bulunan 100 sonuç için elde edilen sıra numaralarının da ortalaması alınarak bulgular kısmında verildi.

Önerilen yöntem ile mevcut yöntemler arasında bulunan performans farkının anlamlı olup olmadığını tespit etmek için istatistiksel yöntemler kullanıldı. Bu yöntemlerin ne olacağına karar verme aşamasında şu hususlara dikkat edildi:

1. Elde edilen performanslarda veriye, sınıflayıcıya ve eğitim setine bağlı olarak yüksek değerler elde ettikçe bu performansı artırmak, düşük değerli sonuçlara göre kolay olmayacaktır. Performans değerleri arasındaki farkın boyutu az olsa da bu farklılık anlamlı olabilir.
2. Elde edilen MKK ve F_1 değerleri aynı koşullarda olduğu için yöntemlere ait değerler birbirine bağımlıdır. Gözlemin herhangi bir durum için bir yöntem çok iyi değerler elde ettiğinde, diğer değerlerin de çok iyi olması muhtemeldir. Çünkü bu durum verinin yapısına veya sınıflayıcıya bağlı olabilir.
3. Elde edilen sıra numaraları aynı koşullarda birbirine bağımlı değildir. Gözlemin herhangi bir durum için bir yöntem 1 sıra numarasına sahip olduğunda, diğer yöntemlerin de 1 sıra numarasına sahip olması beklenemez. Bu sıra numaraları kullanılan yöntemlere bağlıdır. Kendi aralarında bağımsızdır.

Bu hususlar dikkate alındığında MKK ve F_1 değerlerinin yöntemlere göre farklılıklarının istatistiksel anlamlılığına bakmak yanıltıcı olabilir. Fakat sıra

numaraları arasındaki fark incelenebilir. Bu nedenle ikiden fazla normal dağılıma sahip olmayan bağımsız gruplar arasındaki farklılığı incelemek için Kruskal Wallis H testi yapılmıştır. Anlamlı farklılık bulunulan durumlarda ikili karşılaştırmalar için Bonferroni düzeltmeli Dunn testi uygulanmıştır. Yapılan tüm testler için anlamlılık seviyesi 0.05 olarak alınmıştır.



Şekil 5.1 Veri seti üzerinde sınıflama modeli kurma ve performans hesaplama akış şeması

5.2. Veri Setleri

Uygulamada KEEL veri seti havuzundan (Alcalá-Fdez vd, 2011) Yeast4, Banana, Newthyroid1 ve Bupa veri setleri alınmıştır. UCI veri seti havuzundan (Dua ve Graff,

2019), Climatemodelsim, Ecoli, Cleveland, Statlog, Bloodtransfusion, Vehicle, Vertebral, Wholesalecustomers, Glass0, Seeds, PimaIndianDiabetes ve Ionosphere veri setleri alınmıştır. Tüm veri setleri içerisinden, kategorik bağımsız değişkenler çıkartılmış ve kayıp veriye sahip gözlemler atılmıştır. Vehicle veri seti için sınıf değişkeni 4 farklı sınıf içermektedir. Çalışmada bu sınıf değişkeni içerisindeki sınıflardan "van" ve "bus" sınıfları pozitif sınıf olarak alınarak iki farklı iki sınıflı veri seti oluşturulmuştur. Kullanılan veri setlerine ait bilgiler ve veri setlerindeki sınıflara ait gözlem sayıları ile dengesizlik oranları Çizelge 5.2'de verilmiştir. Çizelge 5.2'de veri setleri dengesizlik oranlarına göre sıralanmıştır. Bu duruma göre en yüksek dengesizlik oranı 28.10 ile Yeast4 veri setinde, en düşük dengesizlik oranı ise 1.38 ile Bupa veri setindedir. Bağımsız değişken sayısı en çok olan veri seti 32 ile Ionosphere, en az olan ise 2 ile Banana veri setidir.

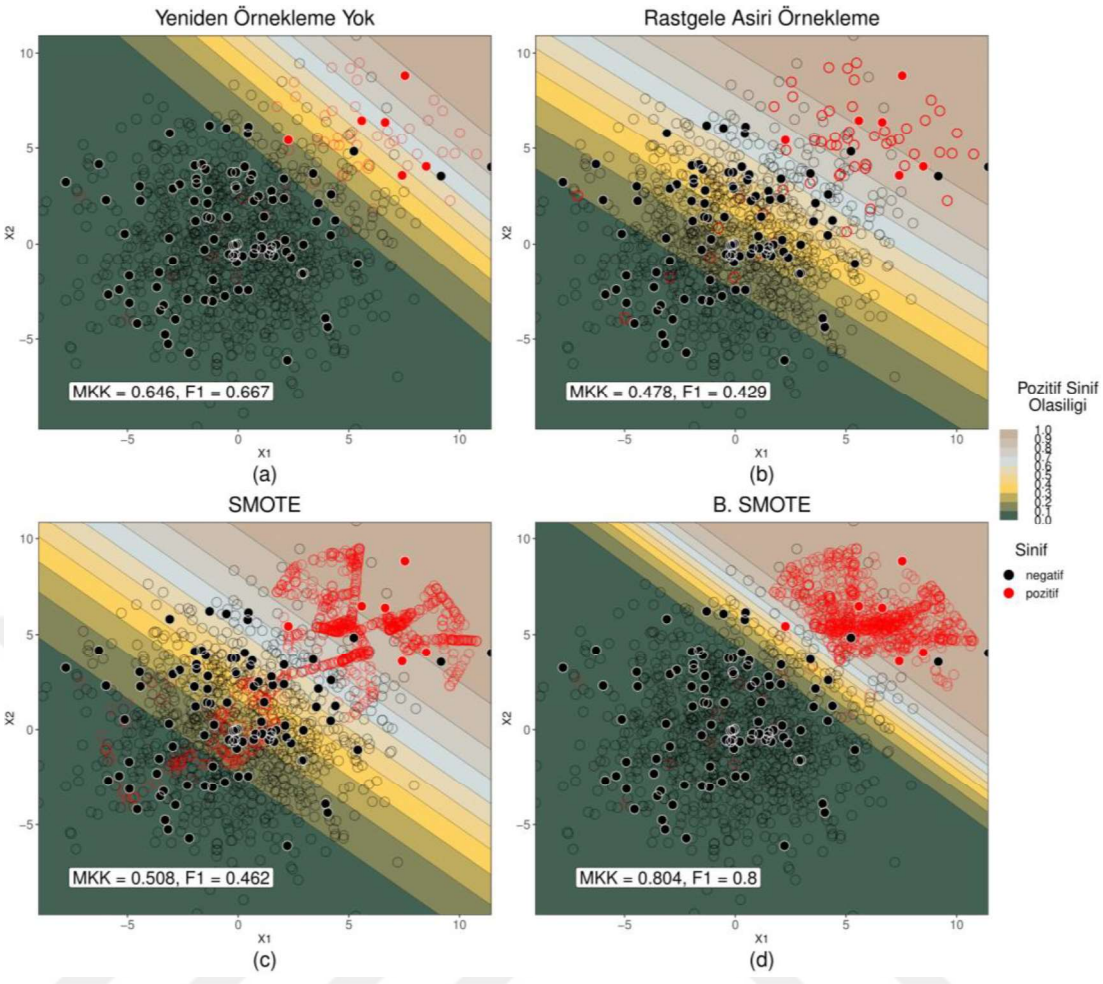
Çizelge 5.2 Çalışmada kullanılan veri setlerine ait bilgiler ve dengesizlik oranı

Veri Seti	P	N	N_{neg}	N_{poz}	Dengesizlik oranı (N_{neg}/N_{poz})
Yeast4	6	1484	1433	51	28,10
Climatemodelsim	18	540	494	46	10,74
Banana	2	2640	2376	264	9,00
Ecoli	5	335	301	34	8,85
Cleveland	5	296	262	34	7,71
Statlog	17	2310	1980	330	6,00
Newthyroid1	5	215	180	35	5,14
Vehicle.van	18	846	647	199	3,25
Vehicle.bus	18	846	628	218	2,88
Vertebral	6	310	210	100	2,10
Wholesalecustomers	7	440	298	142	2,10
Glass0	9	214	144	70	2,06
Seeds	7	199	133	66	2,02
PimaIndiansDiabetes	8	768	500	268	1,87
Ionosphere	32	351	225	126	1,79
Bupa	5	345	200	145	1,38

5.3. Bulgular

Yeniden örnekleme yöntemlerinin sınıflayıcılar üzerindeki etkisini ve önerilen yöntemin farkını görebilmek için simülasyon veri setleri oluşturuldu ve bu veri setlerinde uygulama yapıldı. Oluşturulan modellerin karar sınırlarını gösteren Şekil 5.2, 5.3, 5.4 ve 5.5 verildi. Sınıflayıcıların çalışma yapısı ele alınarak simülasyon verileri iki şekilde oluşturuldu. Lineer sınıflayıcılar üzerindeki etkisini görmek için lineer olarak ayrıştırılabilen bir veri seti oluşturuldu. Lineer olmayan sınıflayıcılar üzerindeki etkisini görmek için ise lineer olarak ayrıştırılamayan bir veri seti oluşturuldu.

Lineer olarak ayıştırılabilen veri seti için 1000 gözlemlik 2 değişkenli $X_{neg} \sim N(0, 3)$ ve 50 gözlemlik 2 değişkenli $X_{poz} \sim N(6, 2)$ gözlemler oluşturuldu. Ardından gürültü oluşturmak amacıyla negatif sınıfların içerisinden rastgele $50 \times 0.3 = 15$ adet gözlem seçildi ve sınıfı pozitif olarak değiştirildi. 1050 gözlemden oluşan veri seti %90 eğitim ve %10 test veri seti olarak ayrıldı. Eğitim veri seti üzerinde lojistik regresyon modeli kuruldu ve test veri seti üzerinde tahmin yapılarak MKK ve F_1 performansı hesaplandı. Bu performanslar ve pozitif sınıfı seçme olasılıklarına göre kontor grafikleri Şekil 5.2'de verildi. Halka şeklindeki noktalar eğitim veri setini, beyaz çizgili daireler ise test veri setini göstermektedir. Şekil 5.2(a)'da dengesiz veri seti üzerinde kurulmuş olan modelin performansı ve karar bölgeleri görülmektedir. Karar, olasılık miktarının 0.5 değerinin üzerine çıkılması veya altında kalınmasına göre verilir. Bir gözlemin pozitif sınıfa ait olması için pozitif olma olasılığının 0.5'in üzerine çıkması gerekir. Şekil 5.2(b) ve Şekil 5.2(c)'de görüldüğü gibi RAÖ ve SMOTE yöntemleri negatif gözlemlerin içerisinde bulunan gürültüler üzerinden de veri türetti. Bu nedenle karar sınırı negatif gözlemleri de kapsamaya başladı. Sınıf dengesizliği problemini çözmeye çalışırken dengesiz veri setinden bile daha kötü performansa sahip modeller elde edildi. Şekil 5.2(d)'de ise önerilen B. SMOTE yönteminin türettiği veriler ve bu veriler üzerinden kurulan modelin karar sınırları görülmtektedir. Gürültülerden hiçbir şekilde etkilenme olmadı ve bir yandan da pozitif ve negatif sınıfların çok iyi ayrıştığı bir veri seti elde edildi. Bu durumda kurulan lojistik regresyon modelinin MKK$= 0.804$ ve $F_1 = 0.8$ performanslarının da diğer lojistik regresyon modellerinden daha iyi olduğu görülmektedir.



Şekil 5.2 Lineer olarak ayıştırılabilen gürültülü gözlemlere sahip X_1 ve X_2 değişkenleri ile kurulan lojistik regresyon modeli pozitif sınıf olasılığı kontör grafikleri ve performansları.

Lineer olarak ayrıştırılamayan veri seti üretmek için 2 değişkenli $X \sim N(0,1)$ 1000 adet gözlemleri oluşturuldu. Daha sonra

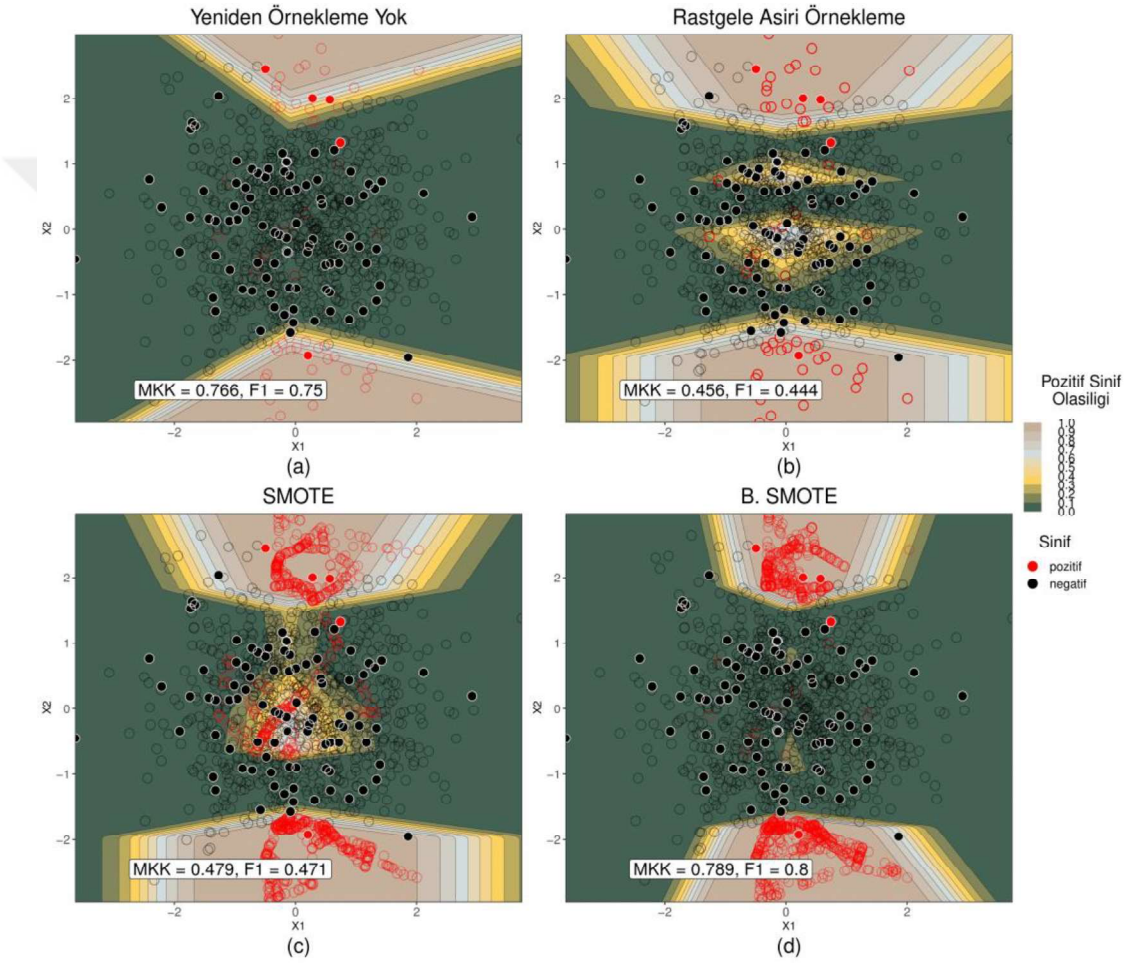
$$Y = \sin(X_1)^2 + \cos(X_2) + \sin(X_1)\cos(X_2) \qquad \textbf{(5.1)}$$

şeklinde Y bağımlı değişkeni oluşturuldu. Ardından Y değişkeni

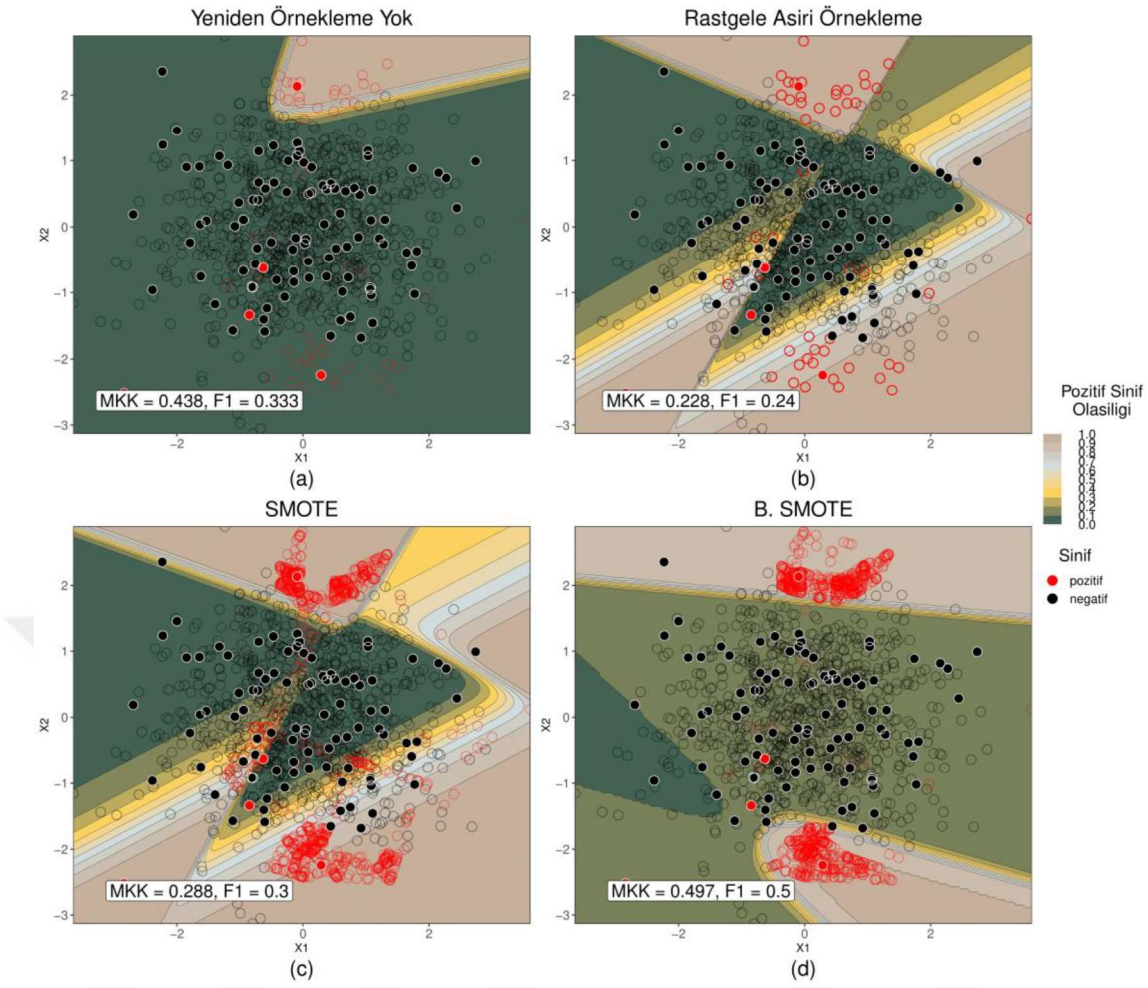
$$Y = \left\lfloor \frac{1}{1+e^{-Y}} + 0.5 \right\rfloor \qquad \textbf{(5.2)}$$

şeklinde önce $[0,1]$ aralığına indirgenip ardından yuvarlanarak 0 ve 1 şeklinde iki sınıf haline getirildi. Az gözlemi olan "pozitif" çok gözlemi olan "negatif" sınıf olarak adlandırıldı. Gürültü oluşturabilmek için negatif gözlemlerin içerisinden $\lfloor n_{pozitif} \times 0.3 + 0.5 \rfloor$ adet gözlem seçilerek sınıfı pozitif olarak değiştirildi. Böylece yüzde pozitif sınıfın %30'u kadar gürültü oluşturuldu. Elde edilen modeldeki dengesizlik miktarını artırmak için pozitif sınıflar içerisinden rastgele $\lfloor n_{pozitif} \times 0.2 + 0.5 \rfloor$ tanesi seçilip

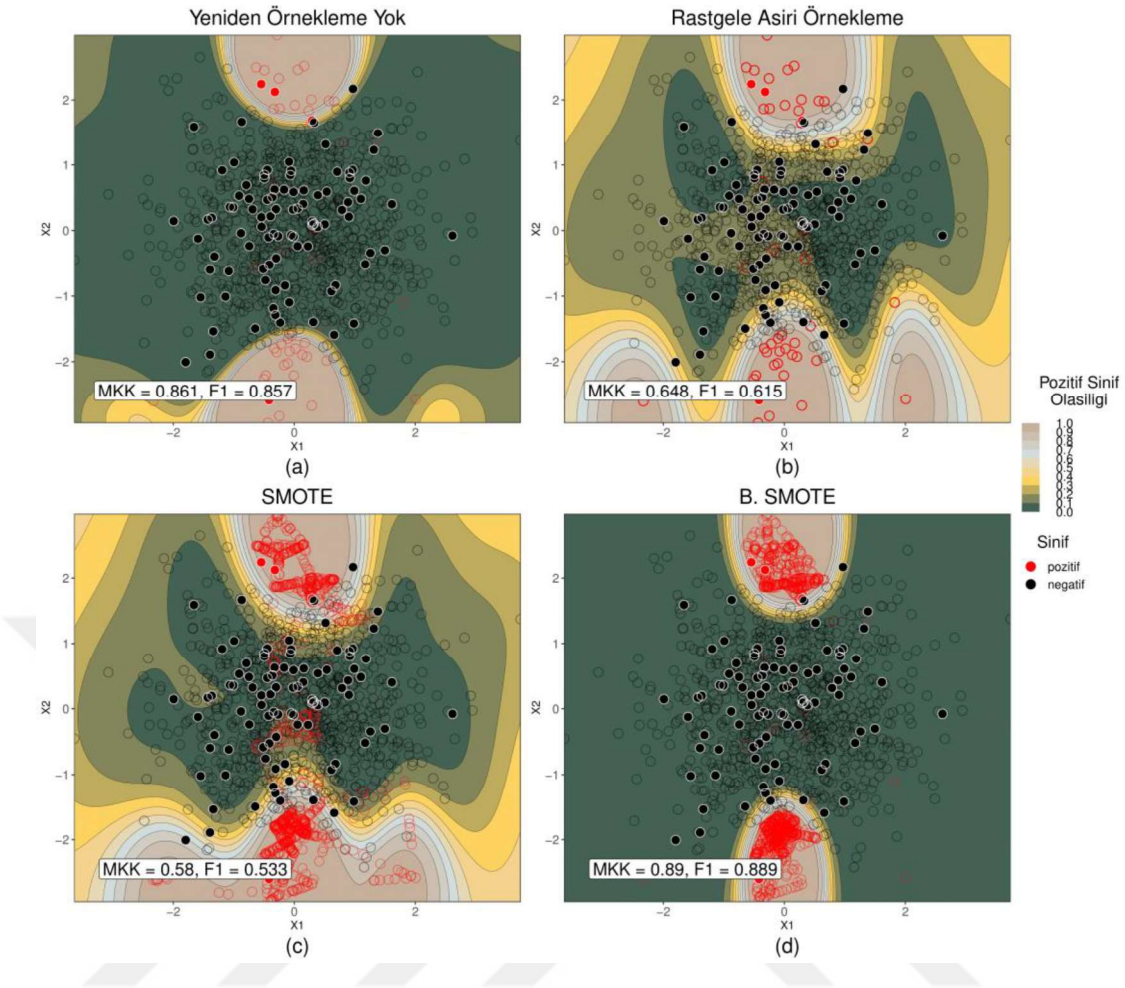
atıldı. Böylece yaklaşık 1'e 20 oranında negatif-pozitif sınıf oranı elde edildi. Elde edilen veri setleri üzerinde MARS, YSA ve Radyal DVM modelleri kuruldu. Şekil 5.3'te MARS sınıflayıcısına ait modeller görülmektedir. Şekil 5.3(b) ve Şekil 5.3(c)'de RAÖ ve SMOTE yöntemlerinin gürültüden etkilendikleri için model nasıl yanlış yönlendirdikleri görüldü. Aynı durum Şekil 5.4'ün (b) ve (c) ile Şekil 5.5'in (b) ve (c) şıklarında YSA ve Radyal DVM modellerinde de görülmektedir. B. SMOTE yönteminin, tüm durumlarda sınıf dengesizliği problemini çözerken gürültüye karşı dayanıklı olduğu ve performansları artırdığı görülmektedir.



Şekil 5.3 Lineer olarak ayrıştırılamayan gürültülü gözlemlere sahip X_1 ve X_2 değişkenleri ile kurulan MARS modeline ait pozitif olasılık kontor grafiği ve performansları.



Şekil 5.4 Lineer olarak ayrıştırılamayan gürültülü gözlemlere sahip X_1 ve X_2 değişkenleri ile kurulan YSA modeline ait pozitif olasılık kontor grafiği ve performansları.



Şekil 5.5 Lineer olarak ayrıştırılamayan gürültülü gözlemlere sahip X_1 ve X_2 değişkenleri ile kurulan Radyal DVM modeline ait pozitif olasılık kontor grafiği ve performansları.

C4.5 sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.3 ve 5.4'te ve bu ölçütlere ait sıra ortalamaları ise Çizelge 5.5 ve 5.6'da görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda 6, RAÖ'de 2, SMOTE'da 3, B. SMOTE'da 5 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında 8, RAÖ'de 3, SMOTE'da 1 ve B. SMOTE'da 4 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına B.SMOTE, en kötü sıra numarasına RAÖ ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına yeniden örnekleme yapılmadan modelleme, en kötü sıra numarsına ise SMOTE ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir. F_1 sıra ortalamalarına göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem SMOTE'dur.

Çizelge 5.3 C4.5 sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	0.251±0.151	**0.257±0.131**	0.241±0.163	0.238±0.164
cleveland	-0.002±0.068	0.064±0.215	0.095±0.214	**0.127±0.227**
ecoli	**0.583±0.244**	0.49±0.185	0.508±0.189	0.57±0.19
glass0	0.579±0.173	0.557±0.183	0.555±0.173	**0.581±0.174**
PimaIndiansDiabetes	0.439±0.104	0.433±0.084	**0.448±0.097**	0.436±0.089
banana	**0.744±0.08**	0.721±0.063	0.654±0.072	0.729±0.066
yeast4	**0.361±0.231**	0.326±0.184	0.338±0.144	0.376±0.173
newthyroid1	0.902±0.142	0.884±0.133	**0.911±0.119**	0.904±0.135
Ionosphere	**0.793±0.111**	0.783±0.104	0.765±0.117	0.782±0.114
Vehicle.bus	0.899±0.049	0.897±0.049	0.904±0.048	**0.906±0.053**
Vehicle.van	**0.827±0.062**	0.819±0.068	0.809±0.069	0.816±0.071
Climatemodelsim	0.499±0.232	0.423±0.219	0.463±0.196	**0.501±0.207**
Seeds	**0.844±0.136**	0.823±0.142	0.824±0.13	0.82±0.133
Statlog	0.971±0.023	**0.978±0.022**	0.977±0.021	0.974±0.022
Vertebral	0.573±0.158	0.533±0.151	**0.584±0.144**	0.576±0.141
Wholesalecustomers	0.786±0.086	0.763±0.094	0.79±0.083	**0.794±0.084**

Çizelge 5.4 C4.5 sınıflayıcısı için F_1 performans ölçütleri

Veri Seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.706±0.071**	0.678±0.066	0.665±0.089	0.662±0.087
cleveland	**0.934±0.016**	0.87±0.045	0.804±0.073	0.874±0.044
ecoli	**0.961±0.02**	0.939±0.024	0.931±0.033	0.946±0.027
glass0	0.848±0.066	0.834±0.074	0.831±0.072	**0.848±0.065**
PimaIndiansDiabetes	**0.811±0.036**	0.777±0.044	0.778±0.046	0.771±0.046
banana	**0.976±0.007**	0.969±0.008	0.954±0.013	0.97±0.008
yeast4	**0.983±0.005**	0.973±0.009	0.961±0.01	0.973±0.009
newthyroid1	0.984±0.022	0.98±0.022	**0.984±0.02**	0.984±0.022
Ionosphere	0.856±0.083	**0.856±0.071**	0.846±0.078	0.853±0.082
Vehicle.bus	0.974±0.013	0.973±0.013	0.974±0.013	**0.975±0.014**
Vehicle.van	0.958±0.015	**0.956±0.017**	0.953±0.019	0.955±0.018
Climatemodelsim	0.96±0.016	0.95±0.019	0.945±0.022	**0.956±0.018**
Seeds	**0.89±0.095**	0.878±0.098	0.878±0.089	0.875±0.092
Statlog	0.974±0.02	**0.981±0.019**	0.98±0.018	0.978±0.019
Vertebral	**0.867±0.055**	0.839±0.061	0.85±0.059	0.853±0.047
Wholesalecustomers	0.928±0.028	0.916±0.035	0.926±0.031	**0.931±0.028**

Çizelge 5.5 C4.5 sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.395	2.475	2.555	2.575
cleveland	2.775	2.620	2.410	2.195
ecoli	2.080	2.890	2.750	2.280
glass0	2.420	2.550	2.630	2.400
PimaIndiansDiabetes	2.585	2.660	2.230	2.525
banana	1.860	2.375	3.675	2.090
yeast4	2.295	2.720	2.750	2.235
newthyroid1	2.405	2.685	2.410	2.500
Ionosphere	2.380	2.495	2.690	2.435
Vehicle.bus	2.570	2.745	2.325	2.360
Vehicle.van	2.195	2.515	2.745	2.545
Climatemodelsim	2.270	2.870	2.610	2.250
Seeds	2.325	2.525	2.595	2.555

Statlog	2.870	2.240	2.360	2.530
Vertebral	2.500	2.795	2.335	2.370
Wholesalecustomers	2.455	2.900	2.345	2.300
Ortalama	2.398	2.628	2.588	2.384

Çizelge 5.6 C4.5 sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	1.970	2.600	2.665	2.765
cleveland	1.080	2.670	3.695	2.555
ecoli	1.760	2.810	3.090	2.340
glass0	2.240	2.595	2.820	2.345
PimaIndiansDiabetes	1.575	2.730	2.700	2.995
banana	1.340	2.495	3.885	2.280
yeast4	1.135	2.470	3.760	2.635
newthyroid1	2.535	2.705	2.380	2.380
Ionosphere	2.460	2.485	2.600	2.455
Vehicle.bus	2.410	2.715	2.525	2.350
Vehicle.van	2.215	2.505	2.745	2.535
Climatemodelsim	1.910	2.730	3.100	2.260
Seeds	2.345	2.505	2.595	2.555
Statlog	2.870	2.240	2.360	2.530
Vertebral	2.030	2.875	2.535	2.560
Wholesalecustomers	2.425	3.100	2.375	2.100
Ortalama	2.018	2.639	2.864	2.477

C4.5 sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.7'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0.05$, $p_{F_1} < 0.05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un yeniden örnekleme yapılmayan durumla arasında anlamlı fark bulunamamış fakat hem MKK hem de F_1 için B. SMOTE'un tüm yöntemlerle anlamlı farklılıkta olduğu görülmüştür.

Çizelge 5.7 C4.5 sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0001*	<0.0001*
YÖ yok – RAÖ (Dunn)	<0.0001*	<0.0001*
YÖ yok – SMOTE (Dunn)	<0.0001*	<0.0001*
RAÖ – SMOTE (Dunn)	0.1133	<0.0001*
YÖ yok – B. SMOTE (Dunn)	0.3778	<0.0001*
RAÖ – B. SMOTE (Dunn)	<0.0001*	<0.0001*
SMOTE – B. SMOTE (Dunn)	<0.0001*	<0.0001*

* p<0,05

EYK sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.8 ve 5.9'de ve bu ölçütlere ait sıra ortalamaları ise Çizelge 5.10 ve 5.11'da görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda

8, RAÖ'de 1, SMOTE'da 5, B. SMOTE'da 2 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında 13, RAÖ'de 0, SMOTE'da 2 ve B. SMOTE'da 1 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına B.SMOTE, en kötü sıra numarasına RAÖ ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına yeniden örnekleme yapılmadan modelleme, en kötü sıra numarsına ise RAÖ ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir. F_1 sıra ortalamalarıan göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en köyü yöntem RAÖ'dir.

Çizelge 5.8 EYK sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.273±0.159**	0.248±0.177	0.257±0.16	0.257±0.159
cleveland	-0.015±0.031	0.07±0.183	**0.11±0.181**	0.039±0.192
ecoli	**0.542±0.269**	0.513±0.141	0.541±0.161	0.526±0.173
glass0	0.528±0.201	**0.575±0.169**	0.574±0.166	0.564±0.172
PimaIndiansDiabetes	**0.352±0.101**	0.319±0.099	0.349±0.092	0.34±0.089
banana	**0.782±0.062**	0.605±0.055	0.64±0.063	0.715±0.061
yeast4	0.128±0.183	0.352±0.132	0.339±0.117	**0.372±0.132**
newthyroid1	0.845±0.166	0.851±0.141	0.898±0.132	**0.904±0.14**
Ionosphere	0.669±0.114	0.714±0.114	**0.741±0.124**	0.715±0.11
Vehicle.bus	**0.752±0.08**	0.722±0.065	0.75±0.064	0.742±0.063
Vehicle.van	**0.84±0.063**	0.767±0.057	0.816±0.062	0.813±0.06
Climatemodelsim	**0.352±0.271**	0.338±0.142	0.337±0.131	0.342±0.165
Seeds	0.779±0.148	0.791±0.135	**0.8±0.133**	0.795±0.127
Statlog	**0.959±0.027**	0.879±0.033	0.935±0.028	0.936±0.027
Vertebral	0.637±0.149	0.627±0.124	**0.64±0.135**	0.629±0.131
Wholesalecustomers	0.764±0.102	0.743±0.107	**0.777±0.085**	0.774±0.087

Çizelge 5.9 EYK sınıflayıcısı için F_1 performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.717±0.064**	0.662±0.087	0.673±0.078	0.669±0.078
cleveland	**0.934±0.013**	0.721±0.074	0.761±0.06	0.757±0.063
ecoli	**0.956±0.024**	0.9±0.042	0.919±0.038	0.922±0.037
glass0	**0.837±0.069**	0.822±0.081	0.816±0.082	0.809±0.085
PimaIndiansDiabetes	**0.786±0.034**	0.729±0.046	0.73±0.045	0.727±0.043
banana	**0.98±0.005**	0.939±0.011	0.951±0.01	0.967±0.008
yeast4	**0.98±0.004**	0.951±0.014	0.947±0.013	0.959±0.011
newthyroid1	0.978±0.021	0.965±0.037	0.98±0.024	**0.982±0.024**
Ionosphere	0.734±0.103	0.78±0.097	**0.806±0.101**	0.779±0.094
Vehicle.bus	**0.939±0.019**	0.902±0.03	0.922±0.022	0.919±0.023
Vehicle.van	**0.961±0.016**	0.923±0.025	0.944±0.022	0.943±0.022
Climatemodelsim	**0.962±0.012**	0.86±0.039	0.842±0.044	0.874±0.041
Seeds	0.84±0.108	0.854±0.092	**0.862±0.089**	0.858±0.086
Statlog	**0.965±0.023**	0.892±0.031	0.943±0.025	0.944±0.024
Vertebral	**0.879±0.051**	0.85±0.06	0.855±0.06	0.852±0.06
Wholesalecustomers	**0.922±0.033**	0.907±0.04	0.92±0.031	0.919±0.032

Çizelge 5.10 EYK sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.465	2.580	2.505	2.450
cleveland	3.050	2.430	2.025	2.495
ecoli	2.315	2.940	2.320	2.425
glass0	2.910	2.280	2.310	2.500
PimaIndiansDiabetes	2.200	3.035	2.285	2.480
banana	1.160	3.845	3.075	1.920
yeast4	3.530	2.270	2.520	1.680
newthyroid1	2.810	2.810	2.225	2.155
Ionosphere	3.250	2.430	1.940	2.380
Vehicle.bus	2.320	2.990	2.225	2.465
Vehicle.van	1.785	3.590	2.285	2.340
Climatemodelsim	2.200	2.640	2.825	2.335
Seeds	2.560	2.545	2.410	2.485
Statlog	1.505	3.905	2.330	2.260
Vertebral	2.415	2.575	2.430	2.580
Wholesalecustomers	2.555	2.970	2.180	2.295
Ortalama	2.439	2.864	2.368	2.327

Çizelge 5.11 EYK sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	1.555	2.930	2.775	2.740
cleveland	1.000	3.530	2.720	2.750
ecoli	1.170	3.635	2.720	2.475
glass0	2.010	2.440	2.700	2.850
PimaIndiansDiabetes	1.150	2.940	2.840	3.070
banana	1.000	3.915	3.085	2.000
yeast4	1.000	3.280	3.590	2.130
newthyroid1	2.380	2.970	2.365	2.285
Ionosphere	3.280	2.450	1.860	2.410
Vehicle.bus	1.650	3.530	2.355	2.465
Vehicle.van	1.295	3.840	2.435	2.430
Climatemodelsim	1.020	2.970	3.635	2.375
Seeds	2.670	2.535	2.360	2.435
Statlog	1.505	3.905	2.330	2.260
Vertebral	1.775	2.805	2.640	2.780
Wholesalecustomers	1.965	3.210	2.380	2.445
Ortalama	1.651	3.180	2.674	2.493

EYK sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.12'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0.05,\ p_{F_1} < 0.05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un SMOTE dışında diğer yöntemlerden anlamlı derecede farklı olduğu görülmüştür. F_1 için ise tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.12 EYK sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0001*	<0.0001*
YÖ yok – RAÖ (Dunn)	<0.0001*	<0.0001*
YÖ yok – SMOTE (Dunn)	0.0552	<0.0001*
RAÖ – SMOTE (Dunn)	<0.0001*	<0.0001*
YÖ yok – B. SMOTE (Dunn)	0.0017*	<0.0001*
RAÖ – B. SMOTE (Dunn)	<0.0001*	<0.0001*
SMOTE – B. SMOTE (Dunn)	0.0910	<0.0001*

* p<0.05

Naive Bayes sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.13 ve 5.14'de ve bu ölçütlere ait sıra ortalamaları ise Çizelge 5.15 ve 5.16'da görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda 6, RAÖ'de 3, SMOTE'da 5, B. SMOTE'da 2 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında 13, RAÖ'de 0, SMOTE'da 3 ve B. SMOTE'da 0 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına B.SMOTE, en kötü sıra numarasına RAÖ ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına yeniden örnekleme yapılmadan modelleme, en kötü sıra numarsına ise RAÖ ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir. F_1 sıra ortalamalarıan göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir.

Çizelge 5.13 Naive Bayes sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.202±0.167**	0.185±0.157	0.167±0.163	0.177±0.166
cleveland	**0.163±0.265**	0.064±0.199	0.034±0.173	0.046±0.196
ecoli	0.542±0.192	0.507±0.162	0.565±0.188	**0.562±0.187**
glass0	**0.57±0.146**	0.545±0.154	0.534±0.163	0.549±0.144
PimaIndiansDiabetes	0.433±0.083	**0.467±0.074**	0.465±0.08	0.463±0.075
banana	0.214±0.097	0.335±0.062	0.344±0.064	**0.35±0.063**
yeast4	0.237±0.172	0.231±0.1	**0.269±0.095**	0.256±0.093
newthyroid1	0.864±0.124	0.858±0.123	**0.871±0.115**	0.857±0.119
Ionosphere	0.813±0.091	0.805±0.093	**0.821±0.098**	0.802±0.095
Vehicle.bus	0.71±0.077	**0.724±0.077**	0.711±0.074	0.715±0.076
Vehicle.van	**0.58±0.055**	0.549±0.049	0.57±0.056	0.57±0.052
Climatemodelsim	0.179±0.246	**0.489±0.209**	0.45±0.217	0.467±0.19
Seeds	**0.79±0.132**	0.775±0.135	0.779±0.136	0.783±0.13
Statlog	0.944±0.029	0.947±0.028	**0.955±0.025**	0.954±0.028
Vertebral	**0.54±0.144**	0.531±0.137	0.534±0.14	0.539±0.144
Wholesalecustomers	0.755±0.092	0.743±0.088	**0.755±0.09**	0.752±0.089

Çizelge 5.14 Naive Bayes sınıflayıcısı için F_1 performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.675±0.075**	0.617±0.082	0.61±0.082	0.616±0.093
cleveland	**0.927±0.03**	0.798±0.061	0.785±0.057	0.811±0.058
ecoli	**0.937±0.035**	0.916±0.039	0.93±0.04	0.936±0.033
glass0	**0.805±0.083**	0.773±0.095	0.773±0.098	0.787±0.084
PimaIndiansDiabetes	**0.814±0.027**	0.796±0.029	0.795±0.032	0.795±0.031
banana	**0.948±0.004**	0.868±0.016	0.874±0.016	0.878±0.016
yeast4	**0.968±0.01**	0.9±0.021	0.904±0.02	0.903±0.022
newthyroid1	0.969±0.032	0.967±0.033	**0.97±0.032**	0.967±0.032
Ionosphere	0.875±0.061	0.87±0.061	**0.882±0.064**	0.868±0.063
Vehicle.bus	**0.92±0.022**	0.919±0.024	0.918±0.022	0.919±0.023
Vehicle.van	**0.821±0.036**	0.792±0.038	0.813±0.037	0.812±0.035
Climatemodelsim	**0.958±0.009**	0.949±0.022	0.949±0.019	0.953±0.019
Seeds	**0.855±0.091**	0.845±0.092	0.849±0.093	0.851±0.09
Statlog	0.952±0.025	0.954±0.025	**0.961±0.021**	0.96±0.024
Vertebral	**0.814±0.065**	0.802±0.065	0.807±0.067	0.811±0.067
Wholesalecustomers	**0.914±0.033**	0.907±0.033	0.911±0.035	0.912±0.033

Çizelge 5.15 Naive Bayes sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.125	2.510	2.745	2.620
cleveland	1.935	2.715	2.785	2.565
ecoli	2.315	3.015	2.380	2.290
glass0	2.100	2.670	2.720	2.510
PimaIndiansDiabetes	2.930	2.275	2.330	2.465
banana	3.790	2.655	1.915	1.640
yeast4	2.580	2.875	2.135	2.410
newthyroid1	2.425	2.580	2.400	2.595
Ionosphere	2.435	2.660	2.235	2.670
Vehicle.bus	2.660	2.105	2.695	2.540
Vehicle.van	1.875	3.445	2.345	2.335
Climatemodelsim	3.475	2.045	2.245	2.235
Seeds	2.355	2.605	2.555	2.485
Statlog	2.995	2.745	2.050	2.210
Vertebral	2.330	2.705	2.565	2.400
Wholesalecustomers	2.475	2.745	2.365	2.415
Ortalama	2.550	2.646	2.404	2.399

Çizelge 5.16 Naive Bayes sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	1.430	2.835	2.905	2.830
cleveland	1.000	3.015	3.340	2.645
ecoli	1.885	3.395	2.570	2.150
glass0	1.925	2.875	2.750	2.450
PimaIndiansDiabetes	1.745	2.640	2.805	2.810
banana	1.000	3.755	2.935	2.310
yeast4	1.000	3.225	2.755	3.020
newthyroid1	2.415	2.580	2.420	2.585
Ionosphere	2.495	2.660	2.145	2.700
Vehicle.bus	2.210	2.605	2.635	2.550
Vehicle.van	1.605	3.735	2.325	2.335
Climatemodelsim	1.925	2.875	2.745	2.455
Seeds	2.365	2.615	2.535	2.485
Statlog	2.995	2.745	2.050	2.210

Vertebral	2.180	2.835	2.565	2.420
Wholesalecustomers	2.205	2.895	2.485	2.415
Ortalama	1.898	2.955	2.622	2.523

Naive Bayes sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.17'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0{,}05$, $p_{F_1} < 0{,}05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un SMOTE hariç diğer tüm yöntemlerle arasında anlamlı derecede farklılık olduğu görülmüştür. F_1 için ise tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.17 Naive Bayes sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0001*	<0.0001*
YÖ yok – RAÖ (Dunn)	0.0012*	<0.0001*
YÖ yok – SMOTE (Dunn)	<0.0001*	<0.0001*
RAÖ – SMOTE (Dunn)	<0.0001*	<0.0001*
YÖ yok – B. SMOTE (Dunn)	<0.0001*	<0.0001*
RAÖ – B. SMOTE (Dunn)	<0.0001*	<0.0001*
SMOTE – B. SMOTE (Dunn)	0.4568	0.0011*

* p<0,05

YSA sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.18 ve 5.19'da ve bu ölçütlere ait sıra ortalamaları ise Çizelge 5.20 ve 5.21'de görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda 6, RAÖ'de 3, SMOTE'da 5, B. SMOTE'da 2 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında 13, RAÖ'de 0, SMOTE'da 3 ve B. SMOTE'da 0 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına B.SMOTE, en kötü sıra numarasına RAÖ ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına yeniden örnekleme yapılmadan modelleme, en kötü sıra numarsına ise RAÖ ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir. F_1 sıra ortalamalarıan göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir.

Çizelge 5.18 YSA sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.317±0.19**	0.317±0.18	0.277±0.197	0.257±0.204
cleveland	0.028±0.138	0.112±0.188	**0.131±0.171**	0.125±0.182
ecoli	0.169±0.273	0.38±0.184	**0.426±0.205**	0.424±0.19
glass0	0.423±0.215	0.456±0.193	**0.508±0.185**	0.477±0.185
PimaIndiansDiabetes	0.211±0.184	0.272±0.166	**0.318±0.166**	0.317±0.158
banana	**0.552±0.14**	0.419±0.106	0.436±0.116	0.444±0.104
yeast4	0.285±0.231	0.281±0.095	0.281±0.11	**0.294±0.114**
newthyroid1	**0.942±0.091**	0.897±0.225	0.905±0.205	0.871±0.273
Ionosphere	**0.747±0.124**	0.719±0.132	0.71±0.129	0.744±0.136
Vehicle.bus	0.476±0.378	0.605±0.296	**0.656±0.269**	0.615±0.297
Vehicle.van	0.673±0.323	**0.712±0.258**	0.709±0.26	0.705±0.254
Climatemodelsim	**0.606±0.212**	0.586±0.216	0.605±0.188	0.594±0.207
Seeds	0.888±0.106	**0.891±0.136**	0.884±0.16	0.883±0.169
Statlog	0.961±0.109	0.962±0.072	0.963±0.103	**0.967±0.101**
Vertebral	0.56±0.215	0.604±0.169	**0.614±0.134**	0.577±0.205
Wholesalecustomers	0.625±0.155	0.636±0.147	0.656±0.127	**0.658±0.114**

Çizelge 5.19 YSA sınıflayıcısı için F_1 performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.743±0.075**	0.704±0.128	0.675±0.148	0.648±0.201
cleveland	**0.934±0.018**	0.739±0.167	0.714±0.161	0.743±0.163
ecoli	**0.945±0.017**	0.867±0.095	0.894±0.054	0.893±0.069
glass0	**0.785±0.076**	0.757±0.091	0.768±0.098	0.761±0.099
PimaIndiansDiabetes	**0.775±0.046**	0.695±0.134	0.702±0.127	0.711±0.127
banana	**0.964±0.007**	0.892±0.068	0.904±0.058	0.9±0.068
yeast4	**0.983±0.004**	0.911±0.025	0.92±0.022	0.925±0.023
newthyroid1	**0.99±0.017**	0.99±0.018	0.988±0.019	0.988±0.021
Ionosphere	**0.818±0.097**	0.802±0.102	0.798±0.099	0.818±0.102
Vehicle.bus	**0.902±0.054**	0.845±0.169	0.853±0.181	0.846±0.178
Vehicle.van	**0.933±0.05**	0.914±0.084	0.916±0.082	0.916±0.076
Climatemodelsim	**0.966±0.017**	0.962±0.019	0.962±0.018	0.963±0.017
Seeds	0.921±0.074	**0.931±0.07**	0.923±0.092	0.927±0.088
Statlog	0.974±0.046	0.966±0.067	**0.976±0.031**	0.972±0.077
Vertebral	**0.841±0.082**	0.828±0.093	0.826±0.081	0.83±0.101
Wholesalecustomers	**0.867±0.049**	0.841±0.113	0.856±0.087	0.86±0.072

Çizelge 5.20 YSA sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.395	2.365	2.655	2.585
cleveland	3.060	2.495	2.180	2.265
ecoli	3.240	2.535	2.095	2.130
glass0	2.720	2.645	2.210	2.425
PimaIndiansDiabetes	2.935	2.605	2.270	2.190
banana	1.540	3.040	2.825	2.595
yeast4	2.380	2.615	2.595	2.410
newthyroid1	2.450	2.515	2.505	2.530
Ionosphere	2.360	2.690	2.595	2.355
Vehicle.bus	2.825	2.515	2.265	2.395

Vehicle.van	2.510	2.505	2.485	2.500
Climatemodelsim	2.365	2.520	2.450	2.665
Seeds	2.620	2.430	2.495	2.455
Statlog	2.485	2.675	2.465	2.375
Vertebral	2.675	2.400	2.420	2.505
Wholesalecustomers	2.655	2.565	2.425	2.355
Ortalama	2.575	2.569	2.433	2.420

Çizelge 5.21 YSA sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.035	2.480	2.660	2.825
cleveland	1.160	2.825	3.110	2.905
ecoli	1.435	3.120	2.730	2.715
glass0	2.245	2.825	2.475	2.455
PimaIndiansDiabetes	1.775	2.755	2.795	2.675
banana	1.040	3.140	2.915	2.905
yeast4	1.000	3.415	2.865	2.720
newthyroid1	2.450	2.520	2.500	2.530
Ionosphere	2.430	2.600	2.555	2.415
Vehicle.bus	2.445	2.585	2.450	2.520
Vehicle.van	2.355	2.485	2.535	2.625
Climatemodelsim	2.265	2.530	2.630	2.575
Seeds	2.630	2.430	2.495	2.445
Statlog	2.485	2.675	2.465	2.375
Vertebral	2.335	2.560	2.630	2.475
Wholesalecustomers	2.460	2.565	2.545	2.430
Ortalama	2.034	2.719	2.647	2.599

YSA sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.22'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0.05,\ p_{F_1} < 0.05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un diğer tüm yöntemlerle arasında anlamlı derecede farklılık olduğu görülmüştür. Aynı şekilde F_1 için de tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.22 YSA sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0001*	<0.0001*
YÖ yok – RAÖ (Dunn)	0.4826	<0.0001*
YÖ yok – SMOTE (Dunn)	0.0001*	<0.0001*
RAÖ – SMOTE (Dunn)	0.0001*	0.0292*
YÖ yok – B. SMOTE (Dunn)	<0.0001*	<0.0001*
RAÖ – B. SMOTE (Dunn)	<0.0001*	0.0007*
SMOTE – B. SMOTE (Dunn)	0.3540	0.0969

* p<0,05

CART sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.23 ve 5.24'te ve bu ölçütlere ait sıra ortalamaları ise Çizelge

5.25 ve 5.26'da görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda 4, RAÖ'de 6, SMOTE'da 2, B. SMOTE'da 4 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında 10, RAÖ'de 3, SMOTE'da 1 ve B. SMOTE'da 4 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına B.SMOTE, en kötü sıra numarasına yeniden örnekleme yapılmayan model ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına yeniden örnekleme yapılmadan modelleme, en kötü sıra numarsına ise RAÖ ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir. F_1 sıra ortalamalarıan göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir.

Çizelge 5.23 CART sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	0.273±0.165	0.289±0.151	0.288±0.158	**0.292±0.159**
cleveland	**0.15±0.258**	0.104±0.212	0.138±0.215	0.128±0.199
ecoli	0.519±0.26	**0.562±0.182**	0.497±0.163	0.517±0.179
glass0	0.511±0.208	**0.604±0.157**	0.574±0.182	0.586±0.198
PimaIndiansDiabetes	0.442±0.099	**0.504±0.091**	0.481±0.092	0.5±0.092
banana	**0.7±0.072**	0.574±0.074	0.577±0.061	0.603±0.074
yeast4	**0.343±0.218**	0.292±0.108	0.308±0.106	0.311±0.113
newthyroid1	0.782±0.187	0.844±0.146	**0.9±0.132**	0.892±0.163
Ionosphere	0.718±0.109	0.72±0.105	0.73±0.115	**0.755±0.108**
Vehicle.bus	**0.876±0.056**	0.838±0.066	0.868±0.06	0.866±0.063
Vehicle.van	0.795±0.08	**0.813±0.063**	0.771±0.067	0.769±0.071
Climatemodelsim	0.488±0.253	0.459±0.184	**0.529±0.176**	0.481±0.176
Seeds	0.731±0.18	**0.778±0.144**	0.738±0.146	0.739±0.162
Statlog	0.957±0.025	0.968±0.027	0.967±0.024	**0.968±0.024**
Vertebral	0.574±0.138	0.572±0.142	0.597±0.139	**0.604±0.13**
Wholesalecustomers	0.822±0.083	**0.828±0.078**	0.823±0.077	0.824±0.074

Çizelge 5.24 CART sınıflayıcısı için F_1 performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.704±0.073**	0.688±0.073	0.691±0.082	0.69±0.079
cleveland	**0.933±0.023**	0.817±0.063	0.816±0.066	0.83±0.056
ecoli	**0.957±0.023**	0.93±0.039	0.92±0.035	0.93±0.032
glass0	0.843±0.061	**0.854±0.061**	0.846±0.064	0.853±0.067
PimaIndiansDiabetes	**0.813±0.035**	0.792±0.049	0.791±0.044	0.793±0.044
banana	**0.971±0.007**	0.937±0.016	0.936±0.013	0.944±0.014
yeast4	**0.983±0.005**	0.924±0.02	0.923±0.018	0.931±0.017
newthyroid1	0.963±0.03	0.972±0.027	0.983±0.023	**0.983±0.022**
Ionosphere	0.803±0.083	0.816±0.071	0.825±0.075	**0.837±0.075**
Vehicle.bus	**0.968±0.014**	0.953±0.02	0.964±0.017	0.964±0.017
Vehicle.van	**0.952±0.018**	0.947±0.019	0.939±0.02	0.94±0.019
Climatemodelsim	**0.961±0.017**	0.939±0.024	0.946±0.021	0.945±0.022
Seeds	0.812±0.127	**0.848±0.099**	0.819±0.102	0.816±0.117
Statlog	0.963±0.022	0.972±0.024	**0.972±0.021**	**0.972±0.021**
Vertebral	**0.857±0.05**	0.842±0.059	0.852±0.054	**0.857±0.05**
Wholesalecustomers	0.941±0.028	**0.941±0.027**	0.939±0.027	0.939±0.026

Çizelge 5.25 CART sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.555	2.410	2.430	2.605
cleveland	2.530	2.610	2.385	2.475
ecoli	2.380	2.210	2.800	2.610
glass0	3.015	2.200	2.500	2.285
PimaIndiansDiabetes	3.250	2.035	2.555	2.160
banana	1.200	3.125	3.195	2.480
yeast4	2.180	2.830	2.530	2.460
newthyroid1	3.085	2.700	2.090	2.125
Ionosphere	2.665	2.775	2.465	2.095
Vehicle.bus	2.165	3.045	2.330	2.460
Vehicle.van	2.340	2.085	2.785	2.790
Climatemodelsim	2.395	2.780	2.255	2.570
Seeds	2.610	2.145	2.665	2.580
Statlog	3.085	2.235	2.370	2.310
Vertebral	2.690	2.620	2.460	2.230
Wholesalecustomers	2.540	2.390	2.535	2.535
Ortalama	2.542	2.512	2.521	2.423

Çizelge 5.26 CART sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.205	2.615	2.525	2.655
cleveland	1.030	3.100	3.085	2.785
ecoli	1.375	2.670	3.185	2.770
glass0	2.735	2.325	2.585	2.355
PimaIndiansDiabetes	1.890	2.625	2.785	2.700
banana	1.010	3.145	3.255	2.590
yeast4	1.000	3.250	3.120	2.630
newthyroid1	2.955	2.760	2.140	2.145
Ionosphere	2.785	2.665	2.405	2.145
Vehicle.bus	1.935	3.375	2.300	2.390
Vehicle.van	1.870	2.335	2.865	2.930
Climatemodelsim	1.595	3.150	2.665	2.590
Seeds	2.645	2.100	2.675	2.580
Statlog	3.085	2.235	2.370	2.310
Vertebral	2.340	2.830	2.550	2.280
Wholesalecustomers	2.450	2.340	2.665	2.545
Ortalama	2.056	2.720	2.698	2.525

CART sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.27'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0{,}05,\ p_{F_1} < 0{,}05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un diğer tüm yöntemlerle arasında anlamlı derecede farklılık olduğu görülmüştür. Aynı şekilde F_1 için de tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.27 CART sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	0.0070*	<0.0001*
YÖ yok – RAÖ (Dunn)	0.2444	<0.0001*
YÖ yok – SMOTE (Dunn)	0.3461	<0.0001*
RAÖ – SMOTE (Dunn)	0.3834	0.3180
YÖ yok – B. SMOTE (Dunn)	0.0008*	<0.0001*
RAÖ – B. SMOTE (Dunn)	0.0071*	<0.0001*
SMOTE – B. SMOTE (Dunn)	0.0030*	<0.0001*

* p<0,05

Lineer DVM sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.28 ve 5.29'da ve bu ölçütlere ait sıra ortalamaları ise Çizelge 5.30 ve 5.31'de görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda 4, RAÖ'de 1, SMOTE'da 3, B. SMOTE'da 9 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında 12, RAÖ'de 1, SMOTE'da 1 ve B. SMOTE'da 4 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına B. SMOTE, en kötü sıra numarasına yeniden örnekleme yapılmayan modelleme ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına yeniden örnekleme yapılmadan modelleme, en kötü sıra numarasına ise RAÖ ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir. F_1 sıra ortalamalarıan göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir.

Çizelge 5.28 Lineer DVM sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.335±0.172**	0.308±0.156	0.289±0.167	0.284±0.167
cleveland	0±0	0.22±0.184	0.235±0.19	**0.237±0.186**
ecoli	0.28±0.26	0.501±0.16	0.505±0.16	**0.533±0.155**
glass0	0.277±0.218	0.449±0.162	**0.458±0.151**	0.452±0.163
PimaIndiansDiabetes	0.475±0.1	0.479±0.092	**0.482±0.096**	0.473±0.096
banana	0±0	0.01±0.062	0.009±0.062	**0.031±0.063**
yeast4	0±0	0.316±0.099	0.309±0.1	**0.317±0.098**
newthyroid1	0.945±0.097	0.963±0.069	0.967±0.065	**0.967±0.062**
Ionosphere	**0.645±0.137**	0.628±0.143	0.63±0.144	0.639±0.14
Vehicle.bus	0.886±0.062	**0.893±0.055**	0.889±0.057	0.892±0.057
Vehicle.van	**0.904±0.045**	0.9±0.048	0.903±0.045	0.901±0.048
Climatemodelsim	**0.675±0.21**	0.573±0.142	0.655±0.162	0.635±0.139
Seeds	0.8±0.139	0.815±0.13	0.826±0.123	**0.829±0.126**
Statlog	0.988±0.013	0.99±0.012	**0.991±0.011**	**0.991±0.011**
Vertebral	0.663±0.129	0.662±0.115	0.666±0.115	**0.667±0.116**
Wholesalecustomers	0.805±0.079	0.798±0.078	0.803±0.076	**0.806±0.075**

Çizelge 5.29 Lineer DVM sınıflayıcısı için F_1 performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.756±0.061**	0.681±0.076	0.665±0.083	0.663±0.086
cleveland	**0.939±0.008**	0.808±0.063	0.804±0.06	0.809±0.063
ecoli	**0.948±0.015**	0.91±0.038	0.913±0.035	0.923±0.031
glass0	**0.796±0.059**	0.702±0.101	0.71±0.091	0.713±0.095
PimaIndiansDiabetes	**0.832±0.031**	0.806±0.038	0.806±0.039	0.803±0.039
banana	**0.947±0.001**	0.681±0.029	0.681±0.029	0.688±0.029
yeast4	**0.983±0.001**	0.924±0.015	0.928±0.013	0.926±0.013
newthyroid1	0.991±0.015	0.993±0.014	0.993±0.013	**0.994±0.012**
Ionosphere	0.729±0.121	0.744±0.106	0.744±0.106	**0.749±0.104**
Vehicle.bus	**0.971±0.016**	0.97±0.016	0.969±0.016	0.97±0.016
Vehicle.van	**0.977±0.011**	0.974±0.013	0.975±0.012	0.974±0.013
Climatemodelsim	**0.977±0.012**	0.939±0.022	0.959±0.02	0.958±0.018
Seeds	0.85±0.116	0.875±0.086	0.881±0.083	**0.883±0.085**
Statlog	0.99±0.011	**0.992±0.01**	**0.992±0.01**	**0.992±0.01**
Vertebral	**0.887±0.044**	0.858±0.06	0.861±0.059	0.863±0.057
Wholesalecustomers	**0.937±0.025**	0.93±0.028	0.932±0.027	0.934±0.027

Çizelge 5.30 Lineer DVM sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.260	2.390	2.675	2.675
cleveland	3.745	2.060	2.210	1.985
ecoli	3.150	2.490	2.400	1.960
glass0	3.295	2.280	2.180	2.245
PimaIndiansDiabetes	2.490	2.475	2.385	2.650
banana	2.735	2.805	2.840	1.620
yeast4	4.000	2.225	1.790	1.985
newthyroid1	2.590	2.515	2.460	2.435
Ionosphere	2.325	2.655	2.570	2.450
Vehicle.bus	2.795	2.335	2.510	2.360
Vehicle.van	2.565	2.590	2.360	2.485
Climatemodelsim	2.020	3.285	2.200	2.495
Seeds	2.830	2.525	2.330	2.315
Statlog	2.720	2.450	2.395	2.435
Vertebral	2.560	2.555	2.445	2.440
Wholesalecustomers	2.495	2.640	2.480	2.385
Ortalama	2.785	2.517	2.389	2.307

Çizelge 5.31 Lineer DVM sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	1.325	2.650	2.985	3.040
cleveland	1.000	2.950	3.110	2.940
ecoli	1.445	3.100	2.980	2.475
glass0	1.540	2.925	2.770	2.765
PimaIndiansDiabetes	1.560	2.695	2.745	3.000
banana	1.000	3.395	3.260	2.345
yeast4	1.000	3.295	2.580	3.125
newthyroid1	2.550	2.525	2.470	2.455
Ionosphere	2.720	2.525	2.420	2.335
Vehicle.bus	2.225	2.565	2.690	2.520
Vehicle.van	1.945	2.790	2.600	2.665
Climatemodelsim	1.340	3.665	2.440	2.555
Seeds	2.980	2.475	2.290	2.255
Statlog	2.720	2.450	2.395	2.435

Vertebral	1.920	2.775	2.705	2.600
Wholesalecustomers	1.975	2.840	2.640	2.545
Ortalama	1.827	2.851	2.692	2.628

Lineer DVM sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.32'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0{,}05$, $p_{F_1} < 0{,}05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un diğer tüm yöntemlerle arasında anlamlı derecede farklılık olduğu görülmüştür. Aynı şekilde F_1 için de tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.32 Lineer DVM sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0000*	<0.0000*
YÖ yok – RAÖ (Dunn)	<0.0000*	<0.0000*
YÖ yok – SMOTE (Dunn)	<0.0000*	<0.0000*
RAÖ – SMOTE (Dunn)	0.0001*	<0.0000*
YÖ yok – B. SMOTE (Dunn)	<0.0000*	<0.0000*
RAÖ – B. SMOTE (Dunn)	<0.0000*	<0.0000*
SMOTE – B. SMOTE (Dunn)	0.0038*	0.0144*

* p<0,05

Radyal DVM sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.38 ve 5.39'da ve bu ölçütlere ait sıra ortalamaları ise Çizelge 5.40 ve 5.41'de görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda 1, RAÖ'de 5, SMOTE'da 2, B. SMOTE'da 6 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında 9, RAÖ'de 3, SMOTE'da 4 ve B. SMOTE'da 5 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına B. SMOTE, en kötü sıra numarasına yeniden örnekleme yapılmayan modelleme ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına B. SMOTE, en kötü sıra numarasına ise yeniden örnekleme yapılmayan modelleme ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem SMOTE'tur. F_1 sıra ortalamalarına göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem SMOTE'tur.

Çizelge 5.33 Radyal DVM sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	0.305±0.148	0.303±0.145	0.317±0.151	**0.322±0.15**
cleveland	0±0	0.081±0.204	0.055±0.195	**0.083±0.207**
ecoli	0.017±0.105	0.521±0.207	**0.58±0.231**	0.542±0.229
glass0	0.498±0.16	**0.603±0.156**	0.597±0.158	0.61±0.161
PimaIndiansDiabetes	0±0	0.382±0.076	0.392±0.082	**0.392±0.077**
banana	**0.711±0.069**	0.69±0.057	0.698±0.055	0.712±0.056
yeast4	0±0	0.267±0.182	0.275±0.176	0.266±0.194
newthyroid1	0.302±0.295	0.929±0.095	**0.941±0.091**	0.939±0.086
Ionosphere	0.058±0.1	**0.409±0.077**	0.362±0.078	0.408±0.083
Vehicle.bus	0.443±0.1	0.696±0.07	0.667±0.069	**0.669±0.072**
Vehicle.van	0.182±0.112	**0.684±0.08**	0.609±0.085	0.611±0.085
Climatemodelsim	0±0	0±0	0±0	0±0
Seeds	0.751±0.144	**0.838±0.122**	0.824±0.124	0.819±0.124
Statlog	0.962±0.028	**0.983±0.016**	0.977±0.021	0.975±0.02
Vertebral	0.546±0.149	0.572±0.145	0.569±0.149	**0.579±0.139**
Wholesalecustomers	0.625±0.13	0.695±0.092	0.708±0.096	**0.717±0.098**

Çizelge 5.34 Radyal DVM sınıflayıcısı için F_1 performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.764±0.038**	0.746±0.05	0.751±0.055	0.754±0.05
cleveland	**0.939±0.008**	0.895±0.032	0.885±0.039	0.903±0.029
ecoli	0.944±0.01	0.93±0.036	**0.95±0.029**	0.946±0.028
glass0	0.861±0.036	0.863±0.055	0.863±0.055	**0.865±0.056**
PimaIndiansDiabetes	**0.789±0.003**	0.622±0.063	0.641±0.06	0.642±0.058
banana	**0.975±0.005**	0.959±0.009	0.962±0.008	0.965±0.008
yeast4	**0.983±0.001**	0.967±0.009	0.965±0.011	0.97±0.01
newthyroid1	0.928±0.021	0.985±0.022	0.988±0.019	**0.988±0.018**
Ionosphere	0.148±0.006	**0.638±0.039**	0.616±0.037	0.638±0.041
Vehicle.bus	0.886±0.014	**0.929±0.014**	0.924±0.014	0.924±0.014
Vehicle.van	0.874±0.006	**0.934±0.015**	0.921±0.014	0.921±0.014
Climatemodelsim	0.956±0.005	0.956±0.005	0.956±0.005	0.956±0.005
Seeds	0.787±0.145	**0.881±0.092**	0.867±0.101	0.864±0.098
Statlog	0.966±0.025	**0.985±0.014**	0.98±0.018	0.978±0.017
Vertebral	**0.872±0.036**	0.854±0.051	0.856±0.051	0.861±0.048
Wholesalecustomers	0.892±0.033	0.879±0.043	0.888±0.042	**0.893±0.043**

Çizelge 5.35 Radyal DVM sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.550	2.655	2.425	2.370
cleveland	2.565	2.465	2.750	2.220
ecoli	3.840	2.430	1.705	2.025
glass0	3.030	2.315	2.440	2.215
PimaIndiansDiabetes	4.000	2.345	1.870	1.785
banana	2.310	3.115	2.625	1.950
yeast4	3.630	2.200	2.170	2.000
newthyroid1	3.990	2.105	1.940	1.965
Ionosphere	3.990	1.640	2.700	1.670
Vehicle.bus	4.000	1.530	2.240	2.230
Vehicle.van	4.000	1.195	2.410	2.395
Climatemodelsim	2.500	2.500	2.500	2.500
Seeds	3.075	2.140	2.355	2.430
Statlog	3.365	1.810	2.330	2.495
Vertebral	2.620	2.485	2.540	2.355

Wholesalecustomers	3.240	2.665	2.150	1.945
Ortalama	3.294	2.224	2.321	2.159

Çizelge 5.36 Radyal DVM sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.270	2.715	2.545	2.470
cleveland	1.155	2.975	3.320	2.550
ecoli	2.400	3.220	2.035	2.345
glass0	2.560	2.515	2.555	2.370
PimaIndiansDiabetes	1.000	3.610	2.780	2.610
banana	1.260	3.625	2.965	2.150
yeast4	1.140	3.060	3.330	2.470
newthyroid1	3.880	2.155	1.970	1.995
Ionosphere	4.000	1.640	2.690	1.670
Vehicle.bus	4.000	1.530	2.240	2.230
Vehicle.van	4.000	1.205	2.400	2.395
Climatemodelsim	2.500	2.500	2.500	2.500
Seeds	3.305	2.070	2.285	2.340
Statlog	3.365	1.810	2.330	2.495
Vertebral	2.165	2.745	2.670	2.420
Wholesalecustomers	2.425	3.065	2.410	2.100
Ortalama	2.589	2.527	2.564	2.319

Radyal DVM sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.42'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0.05$, $p_{F_1} < 0.05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un diğer tüm yöntemlerle arasında anlamlı derecede farklılık olduğu görülmüştür. Aynı şekilde F_1 için de tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.37 Radyal DVM sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0001*	<0.0001*
YÖ yok – RAÖ (Dunn)	<0.0001*	0.0830
YÖ yok – SMOTE (Dunn)	<0.0001*	0.4773
RAÖ – SMOTE (Dunn)	0.0008*	0.0920
YÖ yok – B. SMOTE (Dunn)	<0.0001*	<0.0001*
RAÖ – B. SMOTE (Dunn)	0.0212*	<0.0001*
SMOTE – B. SMOTE (Dunn)	<0.0001*	<0.0001*

* p<0,05

Lojistik regresyon sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.43 ve 5.44'de ve bu ölçütlere ait sıra ortalamaları ise Çizelge 5.45 ve 5.46'da görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda 8, RAÖ'de 2, SMOTE'da 0, B. SMOTE'da 6 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında

13, RAÖ'de 2, SMOTE'da 1 ve B. SMOTE'da 1 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına B. SMOTE, en kötü sıra numarasına RAÖ ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına yeniden örnekleme yapılmayan modelleme, en kötü sıra numarasına ise RAÖ ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir. F_1 sıra ortalamalarına göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir.

Çizelge 5.38 Lojistik regresyon sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	0.313±0.172	0.312±0.154	0.315±0.156	**0.327±0.147**
cleveland	0.131±0.237	**0.263±0.179**	0.26±0.187	0.258±0.175
ecoli	0.513±0.245	0.475±0.15	0.503±0.172	**0.54±0.147**
glass0	**0.466±0.192**	0.46±0.186	0.46±0.187	0.452±0.183
PimaIndiansDiabetes	**0.488±0.092**	0.482±0.098	0.485±0.099	0.478±0.096
banana	0±0	0.07±0.067	0.072±0.067	**0.09±0.069**
yeast4	0.242±0.235	0.299±0.096	0.297±0.098	**0.305±0.096**
newthyroid1	0.916±0.127	**0.924±0.121**	0.92±0.121	0.922±0.122
Ionosphere	**0.64±0.141**	0.629±0.145	0.632±0.153	0.63±0.143
Vehicle.bus	**0.919±0.047**	0.902±0.049	0.911±0.047	0.911±0.047
Vehicle.van	**0.925±0.042**	0.911±0.046	0.916±0.045	0.915±0.046
Climatemodelsim	**0.699±0.201**	0.54±0.155	0.604±0.166	0.676±0.171
Seeds	0.906±0.112	0.898±0.111	0.905±0.112	**0.906±0.111**
Statlog	**0.967±0.023**	0.953±0.033	0.951±0.032	0.951±0.041
Vertebral	**0.665±0.135**	0.656±0.13	0.652±0.131	0.65±0.133
Wholesalecustomers	0.794±0.085	0.799±0.075	0.804±0.075	**0.811±0.074**

Çizelge 5.39 Lojistik regresyon sınıflayıcısı için F_1 performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.734±0.066**	0.684±0.075	0.684±0.072	0.686±0.076
cleveland	**0.941±0.014**	0.812±0.057	0.814±0.059	0.813±0.056
ecoli	**0.957±0.019**	0.901±0.037	0.915±0.035	0.925±0.031
glass0	**0.827±0.065**	0.774±0.078	0.777±0.079	0.775±0.078
PimaIndiansDiabetes	**0.836±0.028**	0.801±0.041	0.804±0.041	0.803±0.039
banana	**0.947±0.001**	0.678±0.028	0.679±0.027	0.697±0.027
yeast4	**0.983±0.004**	0.912±0.017	0.917±0.017	0.919±0.015
newthyroid1	0.986±0.022	**0.987±0.022**	0.986±0.021	0.986±0.022
Ionosphere	0.745±0.105	**0.75±0.102**	0.75±0.108	0.745±0.105
Vehicle.bus	**0.979±0.012**	0.973±0.014	0.976±0.013	0.976±0.013
Vehicle.van	**0.982±0.01**	0.978±0.012	0.979±0.011	0.979±0.012
Climatemodelsim	**0.977±0.013**	0.933±0.028	0.951±0.021	0.965±0.018
Seeds	**0.934±0.078**	0.93±0.076	**0.934±0.078**	0.934±0.076
Statlog	**0.971±0.02**	0.959±0.029	0.957±0.027	0.958±0.036
Vertebral	**0.888±0.046**	0.868±0.056	0.868±0.054	0.869±0.055
Wholesalecustomers	0.934±0.027	0.931±0.026	0.934±0.026	**0.937±0.025**

Çizelge 5.40 Lojistik regresyon sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.670	2.620	2.495	2.215
cleveland	3.205	2.240	2.295	2.260
ecoli	2.320	3.130	2.510	2.040
glass0	2.530	2.495	2.450	2.525
PimaIndiansDiabetes	2.410	2.580	2.420	2.590
banana	3.550	2.545	2.315	1.590
yeast4	2.590	2.750	2.445	2.215
newthyroid1	2.530	2.460	2.530	2.480
Ionosphere	2.305	2.610	2.490	2.595
Vehicle.bus	2.145	2.895	2.490	2.470
Vehicle.van	2.170	2.750	2.540	2.540
Climatemodelsim	1.790	3.500	2.695	2.015
Seeds	2.455	2.610	2.475	2.460
Statlog	2.125	2.570	2.755	2.550
Vertebral	2.425	2.475	2.555	2.545
Wholesalecustomers	2.745	2.625	2.450	2.180
Ortalama	2.497	2.678	2.494	2.329

Çizelge 5.41 Lojistik regresyon sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	1.635	2.845	2.815	2.705
cleveland	1.000	3.010	2.985	3.005
ecoli	1.210	3.530	2.830	2.430
glass0	1.630	2.830	2.725	2.815
PimaIndiansDiabetes	1.465	2.935	2.725	2.875
banana	1.000	3.585	3.385	2.030
yeast4	1.000	3.430	2.845	2.725
newthyroid1	2.520	2.480	2.520	2.480
Ionosphere	2.495	2.430	2.430	2.645
Vehicle.bus	1.955	3.035	2.540	2.470
Vehicle.van	1.970	2.850	2.590	2.590
Climatemodelsim	1.310	3.770	2.835	2.085
Seeds	2.455	2.610	2.475	2.460
Statlog	2.125	2.570	2.755	2.550
Vertebral	1.805	2.765	2.725	2.705
Wholesalecustomers	2.385	2.775	2.550	2.290
Ortalama	1.747	2.965	2.733	2.553

Lojistik regresyon sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.47'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0{,}05$, $p_{F_1} < 0{,}05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un diğer tüm yöntemlerle arasında anlamlı derecede farklılık olduğu görülmüştür. Aynı şekilde F_1 için de tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.42 Lojistik regresyon sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0001*	<0.0001*
YÖ yok – RAÖ (Dunn)	<0.0001*	<0.0001*
YÖ yok – SMOTE (Dunn)	0.3484	<0.0001*
RAÖ – SMOTE (Dunn)	<0.0001*	<0.0001*
YÖ yok – B. SMOTE (Dunn)	<0.0001*	<0.0001*
RAÖ – B. SMOTE (Dunn)	<0.0001*	<0.0001*
SMOTE – B. SMOTE (Dunn)	<0.0001*	<0.0001*

* p<0,05

MARS sınıflayıcısı tüm veri setlerine uygulandıktan sonra elde edilen MKK ve F_1 değerleri Çizelge 5.48 ve 5.49'de ve bu ölçütlere ait sıra ortalamaları ise Çizelge 5.50 ve 5.51'de görülmektedir. MKK ölçütüne göre yeniden örnekleme yapılmadığı durumda 8, RAÖ'de 2, SMOTE'da 2, B. SMOTE'da 4 kez en iyi performans elde edilmiştir. F_1 ölçütüne göre yeniden örnekleme yapılmadığında 13, RAÖ'de 1, SMOTE'da 1 ve B. SMOTE'da 2 kez en iyi performans elde edilmiştir. MKK sıra ortalamalarına bakıldığında en iyi sıra numarasına yeniden örnekleme yapılmayan modelleme, en kötü sıra numarasına RAÖ ulaşmıştır. F_1 sıra ortalamalarına bakıldığında ise en iyi sıra numarasına yeniden örnekleme yapılmayan modelleme, en kötü sıra numarasına ise RAÖ ulaşmıştır. MKK sıra ortalamalarına göre en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir. F_1 sıra ortalamalarına göre ise en iyi yeniden örnekleme yöntemi B. SMOTE iken en kötü yöntem RAÖ'dir.

Çizelge 5.43 MARS sınıflayıcısı için MKK performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	0.335±0.168	0.317±0.166	**0.347±0.175**	0.343±0.163
cleveland	0.067±0.198	0.13±0.204	0.074±0.179	**0.121±0.189**
ecoli	**0.579±0.263**	0.568±0.179	0.577±0.197	0.558±0.208
glass0	0.554±0.221	0.559±0.194	**0.592±0.169**	0.586±0.18
PimaIndiansDiabetes	0.483±0.092	**0.508±0.081**	0.488±0.088	0.503±0.082
banana	**0.426±0.091**	0.352±0.053	0.362±0.055	0.363±0.058
yeast4	0.292±0.229	0.294±0.11	0.309±0.113	**0.311±0.102**
newthyroid1	**0.9±0.125**	0.893±0.145	0.893±0.131	0.886±0.13
Ionosphere	**0.789±0.104**	0.781±0.11	0.737±0.12	0.777±0.104
Vehicle.bus	0.912±0.052	0.912±0.044	0.91±0.048	**0.918±0.041**
Vehicle.van	**0.905±0.057**	0.895±0.052	0.899±0.051	0.903±0.052
Climatemodelsim	**0.615±0.228**	0.543±0.165	0.523±0.191	0.583±0.175
Seeds	0.903±0.099	**0.912±0.098**	0.911±0.098	0.903±0.105
Statlog	**0.982±0.016**	0.968±0.024	0.969±0.026	0.974±0.023
Vertebral	**0.65±0.147**	0.637±0.136	0.639±0.128	0.628±0.139
Wholesalecustomers	0.794±0.09	0.792±0.08	0.804±0.082	**0.805±0.084**

Çizelge 5.44 MARS sınıflayıcısı için F_1 performans ölçütleri

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	**0.734±0.07**	0.696±0.082	0.711±0.085	0.708±0.081
cleveland	**0.935±0.017**	0.801±0.06	0.785±0.066	0.797±0.065
ecoli	**0.961±0.021**	0.937±0.031	0.943±0.03	0.946±0.025
glass0	**0.849±0.074**	0.824±0.082	0.841±0.066	0.84±0.072
PimaIndiansDiabetes	**0.831±0.029**	0.808±0.038	0.801±0.038	0.81±0.036
banana	**0.956±0.005**	0.86±0.016	0.866±0.017	0.871±0.017
yeast4	**0.982±0.005**	0.925±0.015	0.933±0.014	0.933±0.014
newthyroid1	**0.983±0.021**	0.983±0.022	**0.983±0.021**	0.981±0.022
Ionosphere	**0.855±0.075**	0.854±0.076	0.828±0.079	0.85±0.071
Vehicle.bus	0.977±0.014	0.976±0.012	0.976±0.013	**0.977±0.011**
Vehicle.van	**0.977±0.013**	0.973±0.014	0.975±0.013	0.975±0.014
Climatemodelsim	**0.972±0.014**	0.94±0.022	0.943±0.024	0.956±0.019
Seeds	0.932±0.069	**0.938±0.069**	0.938±0.07	0.932±0.074
Statlog	**0.985±0.014**	0.972±0.021	0.974±0.022	0.978±0.02
Vertebral	**0.881±0.051**	0.86±0.059	0.865±0.055	0.862±0.058
Wholesalecustomers	0.932±0.03	0.926±0.028	0.932±0.028	**0.933±0.029**

Çizelge 5.45 MARS sınıflayıcısı için MKK sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.555	2.670	2.295	2.480
cleveland	2.905	2.210	2.630	2.255
ecoli	2.300	2.635	2.480	2.585
glass0	2.545	2.755	2.375	2.325
PimaIndiansDiabetes	2.710	2.330	2.740	2.220
banana	1.760	3.200	2.605	2.435
yeast4	2.560	2.795	2.255	2.390
newthyroid1	2.470	2.520	2.475	2.535
Ionosphere	2.205	2.380	2.945	2.470
Vehicle.bus	2.500	2.550	2.550	2.400
Vehicle.van	2.535	2.585	2.510	2.370
Climatemodelsim	2.000	2.730	2.955	2.315
Seeds	2.570	2.420	2.465	2.545
Statlog	2.075	2.860	2.620	2.445
Vertebral	2.340	2.510	2.440	2.710
Wholesalecustomers	2.620	2.615	2.415	2.350
Ortalama	2.415	2.610	2.547	2.426

Çizelge 5.46 MARS sınıflayıcısı için F_1 sıra ortalamaları

Veri seti	YÖ yok	RAÖ	SMOTE	B. SMOTE
bupa	2.065	2.880	2.455	2.600
cleveland	1.000	2.845	3.165	2.990
ecoli	1.720	3.085	2.680	2.515
glass0	2.215	2.795	2.555	2.435
PimaIndiansDiabetes	1.740	2.730	3.000	2.530
banana	1.000	3.640	2.985	2.375
yeast4	1.000	3.475	2.795	2.730
newthyroid1	2.430	2.500	2.475	2.595
Ionosphere	2.295	2.330	2.905	2.470
Vehicle.bus	2.300	2.660	2.580	2.460
Vehicle.van	2.155	2.745	2.600	2.500
Climatemodelsim	1.300	3.390	3.025	2.285
Seeds	2.580	2.410	2.475	2.535
Statlog	2.075	2.860	2.620	2.445

Vertebral	1.910	2.870	2.510	2.710
Wholesalecustomers	2.360	2.835	2.465	2.340
Ortalama	1.884	2.878	2.705	2.532

MARS sınıflayıcısı için bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.52'de verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı farklılık bulunmuştur ($p_{MKK} < 0{,}05,\ p_{F_1} < 0{,}05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un diğer tüm yöntemlerle arasında anlamlı derecede farklılık olduğu görülmüştür. Aynı şekilde F_1 için de tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.47 MARS sınıflayıcısı için ortalama fark testi sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0001*	<0.0001*
YÖ yok – RAÖ (Dunn)	<0.0001*	<0.0001*
YÖ yok – SMOTE (Dunn)	0.0001*	<0.0001*
RAÖ – SMOTE (Dunn)	0.0394*	<0.0001*
YÖ yok – B. SMOTE (Dunn)	0.3033	<0.0001*
RAÖ – B. SMOTE (Dunn)	<0.0001*	<0.0001*
SMOTE – B. SMOTE (Dunn)	0.0006*	<0.0001*

* p<0,05

Genel bir sonuç elde edebilmek amacıyla tüm sıra durumlara ait sıra numaralarının ortalaması alınmış ve Çizelge 5.53'te verilmiştir. Görüldüğü gibi MKK ölçütüne göre en iyi sıra ortalamasına B. SMOTE ile ulaşılmış, en kötü sıra ortalamasına ise yeniden örnekleme yapılmadığında ulaşılmıştır. F_1 ölçütüne göre ise en iyi sıra ortalamasına yeniden örneklem yapılmadığında ulaşılmış, en kötü sıra ortalamasına ise RAÖ ile ulaşılmıştır. Yeniden örnekleme yöntemleri karşılaştırıldığında MKK için en başarılı yöntem B. SMOTE, en başarısız ise RAÖ olarak bulunmuştur. F_1 için ise en başarılı yöntem B. SMOTE, en başarısız yöntem yine RAÖ olarak bulunmuştur.

Çizelge 5.48 Tüm durumlar için yöntemlere ait sıra ortalamaları

Ölçüt	YÖ yok	RAÖ	SMOTE	B. SMOTE
MKK	2.727	2.541	2.414	2.317
F1	1.990	2.813	2.681	2.514

Tüm bulunan sıra numaraları arasında yeniden örnekleme yöntemlerine göre anlamlı farklılık olup olmadığına ilişkin testlerin sonuçları Çizelge 5.54'te verilmiştir. Bu sonuçlara göre yöntemler arasında hem MKK hem de F_1 ölçütleri için anlamlı

farklılık bulunmuştur ($p_{MKK} < 0.05,\ p_{F_1} < 0.05$). Çoklu karşılaştırma testleri sonucunda MKK için B. SMOTE'un diğer tüm yöntemlerle arasında anlamlı derecede farklılık olduğu görülmüştür. Aynı şekilde F_1 için de tüm yöntemlerle arasında anlamlı farklılık bulunmuştur.

Çizelge 5.49 Tüm sıra numaraları için anlamlı farklılık test sonuçları

	p_{MKK}	p_{F_1}
Kruskall-Wallis H	<0.0001*	<0.0001*
YÖ yok – RAÖ (Dunn)	<0.0001*	<0.0001*
YÖ yok – SMOTE (Dunn)	<0.0001*	<0.0001*
RAÖ – SMOTE (Dunn)	<0.0001*	<0.0001*
YÖ yok – B. SMOTE (Dunn)	<0.0001*	<0.0001*
RAÖ – B. SMOTE (Dunn)	<0.0001*	<0.0001*
SMOTE – B. SMOTE (Dunn)	<0.0001*	<0.0001*

* p<0,05

6. SONUÇLAR VE ÖNERİLER

Makine öğrenmesinde mevcut sınıflama algoritmalarının çoğu, sınıflara ait gözlem sayılarının birbirine dengeli olduğunu varsayarak hipotez üretmektedir. Fakat gerçek veri setlerinde bu durumla nadiren karşılaşılmaktadır. Sınıflara ait gözlemler arasındaki bu dengesizlik sınıf dengesizliği problemi olarak adlandırılmaktadır. Sınıf dengesizliği problemlerinde kullanılan model performans ölçütleri farklı olduğu için öncelikle mevcut ölçütler tanıtılmıştır. Ölçütler içerisinden kullanılmak adına seçilen MKK ve F_1 ölçütlerinin neden seçildiği belirtilmiştir.

Sınıf problemi çözmek için literatürde birçok yöntem önerilmiştir. En çok kullanılan yöntemlerden birisi yeniden örnekleme yöntemleridir. Yeniden örnekleme yöntemlerinde pozitif sınıf gözlemleri artırılarak veya negatif sınıf gözlemleri azaltılarak denge sağlamak amaçlanır. Bu çalışmada literatürde sıklıkla kullanılan yöntemlerden olan rastgele aşırı örnekleme ve SMOTE yöntemlerine ait problemler ele alınmış ve bu problemleri çözmeye yönelik B. SMOTE yöntemi önerilmiştir.

Çalışmada RAÖ ve SMOTE yönteminin veri setindeki gürültü gözlemlerine de önem vererek pozitif gözlemler türettiği gösterilmiştir. Ayrıca SMOTE yönteminin komşu sayısı parametresini sabit olarak almasının da bir problem olduğuna değinilmiştir.

Topluluk algoritmalarında kullanılan boosting prosedürü tanıtılmış ve bu prosedür kullanılarak gürültülerin tespit edilebileceği gösterilmiştir. Önerilen uygun komşu sayı seçme algoritması ile boosting ile gürültü gözlem belirleme algoritmasını bir arada barındıran Boosting ile SMOTE yöntemi tanıtılmıştır. Bu yöntem ile RAÖ ve SMOTE yönteminin türettikleri gözlemler grafikle karşılaştırılmıştır.

Yapılan simülasyon çalışmasında lineer olarak ayrıştırılabilen veri setinde gürültü bulunduğunda RAÖ ve SMOTE yöntemlerinin lojistik regresyon sınıflayıcısını nasıl yanılttıkları hem performans hem de grafikle gösterilmiştir. Lineer olarak ayrıştırılamayan simülasyon veri setinde ise lineer olmayan sınıflayıcılar kullanılarak bu yeniden örnekleme yöntemleri karşılaştırılmıştır. Burada da MARS, YSA ve Radyal DVM sınıflayıcılarında RAÖ ve SMOTE yöntemlerinin gürültüden etkilendikleri hem performanslarda hem de grafiklerde gösterilmiştir. Ayrıca önerilen

B. SMOTE yönteminin mevcut durumlarda gürültüden etkilenmeden veri türettiği ve model performansını artırdığı örneklendirilmiştir.

Literatürde sıklıkla kullanılan 16 hazır veri setinde 9 farklı sınıflayıcıda RAÖ, SMOTE ve B. SMOTE yöntemleri karşılaştırılmıştır. Tüm sınıflayıcılar için sıra numaraları verilerek ortalama sıra numaraları tespit edilmiştir. Bu sıra numaraları üzerinden 9 sınıflayıcının 9'unda da B. SMOTE yöntemi diğer yeniden örnekleme yöntemlerinden daha iyi sıra ortalaması elde etmiştir. Elde edilen sıra ortalamasının diğer yöntemlerin sıra ortalamalarından anlamlı derece farkı olup olmadığını tespit etmek için her sınıflayıcı sonuçlarında Kruskall-Wallis H testi yapılmış ve anlamlı sonuçlar elde edildiği anlaşılmıştır. Yöntemler arasında farklılık olup olmadığını görmek için de Bonferroni düzeltmeli Dunn testleri yapılmış ve sonuçları verilmiştir.

B. SMOTE yönteminin diğer yöntemlerle genel bir kıyaslamasını yapabilmek için tüm sınıflayıcıların sıra numaralarının genel bir ortalaması hesaplanmıştır. B. SMOTE yönteminin, diğer yöntemlere göre daha başarılı MKK ve F_1 skorlar sıra numaralarına sahip olduğu görülmüştür. Yine burada da anlamlı bir farklılık olup olmadığını görmek için Kruskall-Wallis H testi yapılmış ve bu farkın anlamlı olduğu tespit edilmiştir. Ardından yapılan Bonferroni düzeltmeli Dunn testi sonucunda ise B. SMOTE yönteminin diğer tüm yöntemlerle anlamlı farklılık gösterdiği görülmüştür.

Simülasyon ve hazır veri setleri üzerinde yapılan çalışmalardan elde edilen sonuçlar incelendiğinde gürültülü ve sınıf dengesizliği barındıran veri setlerinde önerilen B. SMOTE yönteminin RAÖ ve SMOTE yöntemlerinden daha başarılı olduğu gösterilmiştir. Ayrıca çalışma RAÖ ve SMOTE yöntemlerinin gürültüden nasıl etkilendiklerini hem performans hem de görsel yollarla göstermiştir. Önerilen yöntem sadece gürültülü veri setlerinde değil, gürültüsüz dengesiz sınıflı veri setlerinde de diğer yöntemlerle rekabet edecek seviyededir. Bu nedenle bu yöntemi sınıf dengesizliği problemi barındıran tüm veri setleri için önermekteyiz.

KAYNAKLAR

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F. 2011. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic ve Soft Computing*, 17.

Ali-Gombe, A., ve Elyan, E. 2019. MFC-GAN: Class-imbalanced dataset classification using Multiple Fake Class Generative Adversarial Network. *Neurocomputing*, 361, 212-221.

Alpaydin, E. 2009. *Introduction to machine learning*. MIT press.

ALzubi, J. A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., ve Thaventhiran, C. 2019. Boosted neural network ensemble classification for lung cancer disease diagnosis. *Applied Soft Computing*, 80, 579-591.

Bekkar, M., Djemaa, H. K., ve Alitouche, T. A. 2013. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3:10.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., ve Li, S. 2019. FNN: fast nearest neighbor search algorithms and applications. *R package version*, *1*(1).

Bootkrajang, J. 2016. A generalised label noise model for classification in the presence of annotation errors. *Neurocomputing*, 192, 61-71.

Bunkhumpornpat, C., Sinapiromsaran, K., ve Lursinsap, C. 2009. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining* 475-482. Springer, Berlin, Heidelberg.

Bunkhumpornpat, C., Sinapiromsaran, K., ve Lursinsap, C. 2012. DBSMOTE: density-based synthetic minority over-sampling technique. *Applied Intelligence*, 36:3, 664-684.

Bulut, F. 2016. Performance evaluations of supervised learners on imbalanced datasets. In *2016 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* 1-4. IEEE.

Cano, J. R., Luengo, J., ve García, S. 2019. Label noise filtering techniques to improve monotonic classification. *Neurocomputing*, 353, 83-95.

Charte, F., Rivera, A., del Jesus, M. J., ve Herrera, F. 2015. Resampling multilabel datasets by decoupling highly imbalanced labels. In *International Conference on Hybrid Artificial Intelligence Systems.* 489-501, Springer, Cham.

Charte, F., Rivera, A. J., del Jesus, M. J., ve Herrera, F. 2017. Dealing with difficult minority labels in imbalanced mutilabel data sets. *Neurocomputing*.

Charte, F., Rivera, A. J., del Jesus, M. J., ve Herrera, F. 2019. REMEDIAL-HwR: Tackling multilabel imbalance through label decoupling and data resampling hybridization. *Neurocomputing*, 326, 110-122.

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

Chawla, N. V., Lazarevic, A., Hall, L. O., ve Bowyer, K. W. 2003. SMOTEBoost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery* 107-119. Springer, Berlin, Heidelberg.

Chawla, N. V., Cieslak, D. A., Hall, L. O., ve Joshi, A. 2008. Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, 17:2, 225-252.

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

Chen, W., Bai, M., ve Song, H. 2019. Seismic noise attenuation based on waveform classification. *Journal of Applied Geophysics*, 167, 118-127.

Chicco, D. 2017. Ten quick tips for machine learning in computational biology. *BioData mining*, 10:1, 35.

Cordón, I., García, S., Fernández, A. ve Herrera, F. 2018. imbalance: Preprocessing Algorithms for Imbalanced Datasets. R package version 1.0.0. https://CRAN.R-project.org/package=imbalance

Culp, M., Johnson, K. and Michailidis, G. 2016. ada: The R Package Ada for Stochastic Boosting. R package version 2.0-5. https://CRAN.R-project.org/package=ada

Das, B., Krishnan, N. C., ve Cook, D. J. 2014. RACOG and wRACOG: Two probabilistic oversampling techniques. *IEEE transactions on knowledge and data engineering*, 27:1, 222-234.

Daskalaki, S., Kopanas, I., ve Avouris, N. 2006. Evaluation of classifiers for an uneven class distribution problem. *Applied artificial intelligence*, 20:5, 381-417.

Deepayan, S. 2008. Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

Dev, V. A., ve Eden, M. R. 2019. Formation Lithology Classification Using Scalable Gradient Boosted Decision Trees. *Computers ve Chemical Engineering*.

Douzas, G., ve Bacao, F. 2019. Geometric SMOTE A geometrically enhanced drop-in replacement for SMOTE. *Information Sciences*.

Dowle, M., ve Srinivasan, A. 2018. data. table: Extension ofdata. frame. R package version 1.11. 4.

Dragulescu, A. ve Arendt, C. 2012. *xlsx: Read, write, format Excel 2007 and Excel 97*. 2000/XP/2003 files.

Dua, D. and Graff, C. 2019. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Ergul, U., ve Bilgin, G. 2019. HCKBoost: Hybridized composite kernel boosting with extreme learning machines for hyperspectral image classification. *Neurocomputing*, 334, 100-113.

Estabrooks, A., Jo, T., ve Japkowicz, N. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20:1, 18-36.

Fernàndes, E. R., ve de Carvalho, A. C. 2019. Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning. *Information Sciences*, 494, 141-154.3

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., ve Herrera, F. 2018. *Learning from imbalanced data sets.* 1-377, Berlin: Springer.

Freund, Y., ve Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55:1, 119-139.

Fu, Y. J., Liu, H., Cui, Y. F., ve Zhou, Y. P. 2019. Boosting classification tree-radial basis function network: Application in metabonomics studies. *Chemometrics and Intelligent Laboratory Systems*, 193, 103829.

Furundzic, D., Stankovic, S., Jovicic, S., Punisic, S., ve Subotic, M. 2017. Distance based resampling of imbalanced classes: With an application example of speech quality assessment. *Engineering Applications of Artificial Intelligence*, 64, 440-461..

García, V., Sánchez, J. S., ve Mollineda, R. A. 2012. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25:1, 13-21.

Garcia, L. P., Sáez, J. A., Luengo, J., Lorena, A. C., de Carvalho, A. C., ve Herrera, F. 2015a. Using the One-vs-One decomposition to improve the performance of class noise filters via an aggregation strategy in multi-class classification problems. *Knowledge-Based Systems*, 90, 153-164.

Garcia, L. P., de Carvalho, A. C., ve Lorena, A. C. 2015b. Effect of label noise in the complexity of classification problems. *Neurocomputing*, 160, 108-119.

García, S., Zhang, Z. L., Altalhi, A., Alshomrani, S., ve Herrera, F. 2018. Dynamic ensemble selection for multi-class imbalanced datasets. *Information Sciences*, 445, 22-37.

García-Gil, D., Luengo, J., García, S., & Herrera, F. 2019. Enabling smart data: noise filtering in big data classification. *Information Sciences*, 479, 135-152.

Haklı, A. D. 2018. Sınıf Dengesizliği Sorununu Çözmek İçin Kullanılan Algoritmaların Farklı Sınıflandırma Yöntemlerinde Performanslarının Karşılaştırılması.

Han, H., Wang, W. Y., ve Mao, B. H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing* 878-887. Springer, Berlin, Heidelberg.

Hastie, T., Tibshirani, R., Friedman, J. ve Franklin, J. 2005. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27:2, 83-85.

He, H., Bai, Y., Garcia, E. A., ve Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* 1322-1328. IEEE.

He, H., ve Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21:9, 1263-1284.

Hornik K, Buchta C, Zeileis A 2009. “Open-Source Machine Learning: R Meets Weka.” _Computational Statistics_, 24:2, 225-232. doi: 10.1007/s00180-008-0119-7 (URL: https://doi.org/10.1007/s00180-008-0119-7).

Ince, I. F., Ince, O. F., ve Bulut, F. 2019. MID Filter: An Orientation-Based Nonlinear Filter For Reducing Multiplicative Noise. *Electronics*, 8:9, 936.

Junior, J. R. B., ve do Carmo Nicoletti, M. 2019. An iterative boosting-based ensemble for streaming data classification. *Information Fusion*, 45, 66-78.

Kamalov, F. (2019). Kernel density estimation based sampling for imbalanced class distribution. *Information Sciences*.Freund, Y. 1995. Boosting a weak learning algorithm by majority. *Information and computation*, 121:2, 256-285.

Karatzoglou, A., Smola, A., Hornik, K., ve Zeileis, A. 2004. kernlab-an S4 package for kernel methods in R. *Journal of statistical software*, 11:9, 1-20.

Kim, Y. G., Kwon, Y., ve Paik, M. C. 2019. Valid oversampling schemes to handle imbalance. *Pattern Recognition Letters*, 125, 661-667.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kengel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. Ve Hunt, T. 2015. caret: Classification and regression training. R package version 6.0–21. *CRAN, Vienna, Austria*.

Lavine, B. K. 2000. Clustering and classification of analytical data. Encyclopedia of analytical chemistry: instrumentation and applications.

Leisch, F., ve Dimitriadou, E. 2010. mlbench: Machine Learning Benchmark Problems. R package version 2.1-1.

Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., ve García-Borroto, M. 2016. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing*, 175, 935-947.

Luengo, J., Shim, S. O., Alshomrani, S., Altalhi, A., ve Herrera, F. 2018. CNC-NOS: Class noise cleaning by ensemble filtering and noise scoring. *Knowledge-Based Systems*, 140, 27-49.

Luo, Y., Xiong, Z., Xia, S., Tan, H., ve Gou, J. 2016. Classification noise detection based SMO algorithm. *Optik*, 127:17, 7021-7029.

Maas, A. E., Rottensteiner, F., ve Heipke, C. 2019. A label noise tolerant random forest for the classification of remote sensing data based on outdated maps for training. *Computer Vision and Image Understanding*, 188, 102782.

Maldonado, S., López, J., ve Vairetti, C. 2019. An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 76, 380-389.

Milborrow, S., 2019. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. earth: Multivariate Adaptive Regression Splines. R package version 5.1.2. https://CRAN.R-project.org/package=earth

Obregon, J., Kim, A., ve Jung, J. Y. 2019. RuleCOSI: Combination and simplification of production rules from boosted decision trees for imbalanced classification. *Expert Systems with Applications*, 126, 64-82.

Ofek, N., Rokach, L., Stern, R., ve Shabtai, A. 2017. Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem. *Neurocomputing*, 243, 88-102.

Qian, Y., Liang, Y., Li, M., Feng, G., ve Shi, X. 2014. A resampling ensemble algorithm for classification of imbalance problems. *Neurocomputing*, 143, 57-67.

R Core Team 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ren, S., Zhu, W., Liao, B., Li, Z., Wang, P., Li, K., Chen, M. ve Li, Z. 2019. Selection-based resampling ensemble algorithm for nonstationary imbalanced stream data learning. *Knowledge-Based Systems*, 163, 705-722.

Rivera, W. A. 2017. Noise reduction a priori synthetic over-sampling for class imbalanced data sets. *Information Sciences*, 408, 146-161.

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Larmarange, J. 2018. GGally: Extension to 'ggplot2'. R package version 1.4.0. https://CRAN.R-project.org/package=GGally

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

Sharififar, A., Sarmadian, F., Malone, B. P., ve Minasny, B. 2019. Addressing the issue of digital mapping of soil classes with imbalanced class observations. *Geoderma*, 350, 84-92

Shi, H., Wang, H., Huang, Y., Zhao, L., Qin, C., ve Liu, C. 2019. A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification. *Computer methods and programs in biomedicine*, 171, 1-10.

Siriseriwan, W. 2019. smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE. R package version 1.3.1. https://CRAN.R-project.org/package=smotefamily

Soleymani, R., Granger, E., ve Fumera, G. 2018. Progressive boosting for class imbalance and its application to face re-identification. *Expert Systems with Applications*, 101, 271-291.

Sundar, R., ve Punniyamoorthy, M. 2019. Performance enhanced Boosted SVM for Imbalanced datasets. *Applied Soft Computing*, 83, 105601.

Therneau, T., Atkinson, B., ve Ripley, B. 2015. rpart: Recursive partitioning and regression trees. *R package version*, 4, 1-9.

Venables, W. N. ve Ripley, B. D. 2002 Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Vezhnevets, A., ve Vezhnevets, V. 2005. Modest AdaBoost-teaching AdaBoost to generalize better. In *Graphicon* 12:5, 987-997.

Vuttipittayamongkol, P., ve Elyan, E. 2020. Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. *Information Sciences*, 509, 47-70.

Wang, B. X., ve Japkowicz, N. 2010. Boosting support vector machines for imbalanced data sets. *Knowledge and information systems*, 25:1, 1-20.

Wang, L. L., Ngan, H. Y., ve Yung, N. H. 2018. Automatic incident classification for large-scale traffic data by adaptive boosting SVM. *Information Sciences*, 467, 59-73.

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. 2005. klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335-343, Springer-Verlag, Berlin.

Wickham, H., ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Williams, G. J. (2011), *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!,* Springer.

Wong, M. L., Seng, K., ve Wong, P. K. 2020. Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain. *Expert Systems with Applications*, 141, 112918.

Xu, Y., Fang, X., You, J., Chen, Y., ve Liu, H. 2015. Noise-free representation based classification and face recognition experiments. *Neurocomputing*, 147, 307-314.

Yu, L., Zhou, R., Tang, L., ve Chen, R. 2018. A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data. *Applied Soft Computing*, 69, 192-202.

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., ve Jinling, L. 2016. Adapted ensemble classification algorithm based on multiple classifier system and feature

selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, 94, 88-104.

Zefrehi, H. G., ve Altınçay, H. 2019. Imbalance Learning Using Heterogeneous Ensembles. *Expert Systems with Applications*, 113005.

Zhang, H., ve Li, M. 2014. RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, 20, 99-116.

ÖZGEÇMİŞ

Adı ve Soyadı	:	Fatih Sağlam
Doğum Yeri	:	Zile/TOKAT
Doğum Tarihi	:	20.08.1989
Yabancı Dili	:	İngilizce
Eğitim Durumu		
Lise	:	Zile Anadolu Öğretmen Lisesi (2007)
Lisans	:	Ondokuz Mayıs Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü (2017)
Yayınlar	:	Pala, M., SAĞLAM, F., ve SÖZEN, Ç. (2019), Kapula Modelinin Belirlenmesinde AIC Değerinin Hatalı Seçimi. *The Journal of International Scientific Researches*, *4*(3), 335-343.
		Pala, M., ve SAĞLAM, F. (2019), PISA Fen, Matematik ve Okuma Puanları Arasındaki Bağımlılık Yapısının Kapula ile Modellenmesi. *Selçuk Üniversitesi Fen Fakültesi Fen Dergisi*, *45*(2), 149-162.
		Pala, M., Sağlam, F., ve SÖZEN, Ç. (2019), Investıgatıng The Dependency Structure Of Unemployment Rate Wıth Regard To Export And Import Usıng Copula Method.