DOKUZ EYLÜL ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DENGESİZ VERİ SETLERİ İÇİN AŞIRI ÖRNEKLEME YÖNTEMLERİNİN PERFORMANS KARŞILAŞTIRMASI

Eren HATİPOĞLU

Mart, 2025

İZMİR

DENGESİZ VERİ SETLERİ İÇİN AŞIRI ÖRNEKLEME YÖNTEMLERİNİN PERFORMANS KARŞILAŞTIRMASI

Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü
Yüksek Lisans Tezi
İstatistik Anabilim Dalı, Veri Bilimi Tezsiz Yüksek Lisans Programı

Eren HATİPOĞLU

Mart, 2025
İZMİR

TEZSİZ YÜKSEK LİSANS DÖNEM PROJESİ SONUÇ FORMU

EREN HATİPOĞLU tarafından **PROF. DR. NESLİHAN DEMİREL** yönetiminde hazırlanan **"DENGESİZ VERİ SETLERİ İÇİN AŞIRI ÖRNEKLEME YÖNTEMLERİNİN PERFORMANS KARŞILAŞTIRMASI"** başlıklı Dönem Projesi tarafımdan okunmuş, kapsamı ve niteliği açısından bir Tezsiz Yüksek Lisans Dönem Projesi olarak kabul edilmiştir.

Kabul edilen Tezsiz Yüksek Lisans Dönem Projesi

- ☐ Kapsamlı bir derleme
- ☐ Eleştirel bir rapor
- ■ Uygulamaya dönük bir proje
- ☐ Deneysel bir çalışma dır.

……

Prof. Dr. Neslihan Demirel

Danışman

TEŞEKKÜR

Proje konumu seçmemde yardımcı olan, proje süresi boyunca gösterdiği ilgiyle projemi en iyi şekilde yapmamı sağlayan ve bilgi birikimiyle bana çok şey katmış danışman hocam Prof. Dr. Neslihan Demirel'e sonsuz teşekkürlerimi sunarım.

Dokuz Eylül Üniversitesi İstatistik Bölümünün tüm akademisyenlerine teşekkür ediyorum. Her birinden hayatım boyunca kullanacağım çok şey öğrendim.

Hem bu yüksek lisans sürecinde hem de her daim beni desteklemiş aileme sonsuz teşekkürlerimi iletiyorum. Onların özverileri olmadan bu proje olamazdı.

Eren HATİPOĞLU

DENGESİZ VERİ SETLERİ İÇİN AŞIRI ÖRNEKLEME YÖNTEMLERİNİN PERFORMANS KARŞILAŞTIRMASI

ÖZ

İkili sınıflandırma problemlerinde iki sınıfa ait gözlem sayısının birbirinden oldukça farklı olması durumu dengesiz veri olarak tanımlanmaktadır. Bu durum, makine öğrenmesi algoritmaları kullanılarak veri analiz edilirken problemlere yol açar. Dengesiz veri sorununu aşmak için sıklıkla aşırı örnekleme yöntemleri kullanılır. Bu çalışmada, 14 değişkene sahip olan ve bir kişinin kalp hastası olarak sınıflandırılıp sınıflandırılamayacağını belirlemek amacıyla kullanılan veri seti öncelikle yapay olarak, azınlık sınıfının çoğunluk sınıfına oranı %5, %15, %25 olacak şekilde dengesiz hale getirilmiştir. Bu veri setlerini dengelemek için aşırı örnekleme yöntemleri olarak Uyarlanabilir Sentetik Örnekleme (ADASYN), Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE), Borderline-SMOTE ve Rassal Aşırı Örnekleme Örnekleri (ROSE) yöntemleri kullanılmıştır. Makine öğrenmesi algoritmalarından lojistik regresyon, Naive Bayes, rassal orman, destek vektör makinesi, yapay sinir ağları ve aşırı gradyan artırma (XGBoost) ile sınıf tahminlemesi yapılmıştır. Dengesiz haldeki veri setleri üzerinde de aynı makine öğrenmesi algoritmaları uygulanmış ve kapsamlı bir karşılaştırma yapılmıştır. Tüm analizler R programlama dili ve gerekli paketler kullanılarak yapılmıştır. Sonuç olarak tüm veri setleri, algoritmalar ve metrikler göz önünde bulundurulduğunda aşırı örnekleme yöntemlerinden ADASYN, Borderline-SMOTE ve SMOTE, makine öğrenmesi algoritmalarından sırasıyla Rassal Orman ve XGBoost algoritmalarının en iyi performansı gösterdiği sonucuna varılmıştır.

Anahtar kelimeler: Aşırı örnekleme, dengesiz veri, makine öğrenmesi, sınıflandırma algoritmaları.

PERFORMANCE COMPARISON OF OVERSAMPLING METHODS FOR IMBALANCED DATASETS

ABSTRACT

In binary classification problems, when the number of observations belonging to two classes is quaite different from each other, it is defined as imbalanced data. This situation causes problems when data is analyzed with machine learning algorithms. Oversampling methods are often used for the problem of imbalanced data. In this study, the dataset, which has 14 variables and is used to determine whether a person can be classified as a heart disease patient, was first artificially unbalanced so that the ratio of the minority class to the majority class was 5%, 15%, and 25%. Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Oversampling Technique (SMOTE), Borderline-SMOTE and Random Oversampling Examples (ROSE) methods were used as oversampling methods to balance these datasets. Classification was performed with machine learning algorithms such as logistic regression, Naive Bayes, random forest, support vector machine, artificial neural networks, and XGBoost. The same machine learning algorithms were applied on imbalanced datasets, and a comprehensive comparison was made. All analyses were performed with the use of the R programming language and the necessary packages. As a result, considering all datasets, algorithms, and metrics, it was concluded that ADASYN, Borderline-SMOTE, and SMOTE from oversampling methods and random forest and XGBoost algorithms from machine learning algorithms showed the best performance, respectively.

Keywords: Oversampling, imbalanced datasets, machine learning, classification algorithms.

İÇİNDEKİLER

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

TABLOLAR LİSTESİ

BÖLÜM BİR
GİRİŞ

Dengesiz veriler, veri bilimi ve istatistik başta olmak üzere birçok alanda karşılaşılan yaygın bir problemdir. Dengesiz veri, hedef değişkendeki sınıflara ait gözlem değerlerinin birbirine eşit olmaması veya eşit olmaya yakın olmaması anlamına gelir. Oluşturulan makine öğrenmesi modeli çoğunluk olan sınıfa odaklanmaya yatkındır. Bu durumda azınlık olan sınıfın ihmal edilmesi, çoğunluk sınıfında aşırı öğrenme gibi çeşitli problemler oluşur ve genellikle çoğunluk olan sınıf daha az önemli olan sınıftır (Kotsiantis vd., 2006). Örneğin azınlık olan sınıf bir hastalık (örn. kanser olma veya olmama durumu) ise bu kritik bir konu olabilir ve böyle durumlarda modelin yanlış yönlendirici olması istenmez. Azınlık sınıfının, çoğunluk sınıfına oranının %1'den az olması durumunda dengesizlik çok, %1-20 arasındayken orta, %20-40 arasındayken ise dengesizlik az varsayılmaktadır (Azlim Khan ve Ahamed Hassain Malim, 2023).

Çağımızda, teknolojinin hızla gelişmesi sayesinde verilere erişilebilirlik çok fazla artmıştır ve bu verileri kullanarak çeşitli öngörüler elde etmek de daha kritik hale gelmiştir. Dengesiz veri problemi akademi ve endüstride göreceli olarak yeni bir problemdir ve bu dengesiz verileri nasıl kullanabileceğimiz önemli bir konudur (Haibo ve Garcia, 2009). Verilerin sınıflandırılması konusunda en iyi performans gösteren algoritmalar bile verinin dengeli olduğunu varsayar ve veri dengesiz olduğunda bu algoritmalar dezavantajlı konuma düşer (Sun vd., 2009).

Dengesiz veri problemini aşmak için birçok yöntem vardır ve bu yöntemler veri seviyesi ile algoritma seviyesi şeklinde iki farklı yaklaşımla uygulanmaktadır. Aşırı örnekleme yöntemleri veri seviyesi yaklaşım kapsamına girer ve sıklıkla kullanılırlar. Farklı aşırı örnekleme yöntemlerinin kullanılması çoğaltılmış verilerin farklı özellikler göstermesine sebep olur (Santoso vd., 2017).

Bütün bu bilgiler ışığında, çalışmanın ikinci bölümünde aşırı örnekleme yöntemleri ve makine öğrenmesi algoritmaları açıklanmıştır. Üçüncü bölüm uygulama kısmında makine öğrenmesi algoritmaları hem dengelenmiş hem de dengesiz veri setleri

üzerinde uygulanmış ve performansları karşılaştırılmıştır. Son olarak ise dördüncü bölümde çalışmanın sonuçları belirtilmiştir.

BÖLÜM İKİ
AŞIRI ÖRNEKLEME YÖNTEMLERİ VE SINIFLANDIRMA ALGORİTMALARI

Bu bölümde aşırı örnekleme yöntemleri olan SMOTE, Borderline-SMOTE, ADASYN ve ROSE yöntemleri ve denetimli makine öğreniminde sınıflandırma problemlerinde kullanılan lojistik regresyon, Naive Bayes, rassal orman, XGBoost, yapay sinir ağları ve destek vektör makinesi algoritmaları açıklanmıştır. Ayrıca uygulama bölümünde yöntemlerin performanslarını karşılaştırmak üzere performans metrikleri tanımlanmıştır.

2.1 Rassal Alt Örnekleme Yöntemi

Bu çalışmanın uygulama bölümünde, aşırı örnekleme yöntemleri ile karşılaştırılmak üzere rassal alt örnekleme yöntemi olan Random Undersampling (RUS) kullanılmıştır. Çoğunluk sınıfından gözlemleri rastgele bir şekilde silip azınlık sınıfıyla eşitleyerek denge sağlar fakat ciddi miktarda veri kaybına sebep olur ve performans düşüşü yaratır (Newaz vd., 2024). Verileri dengelemek için en basit yöntemlerden biri rassal alt örneklemedir (Sağlam, 2021).

2.2 Aşırı Örnekleme Yöntemleri

Bu bölümde SMOTE, Borderline-SMOTE, ADASYN, ROSE aşırı örnekleme yöntemleri açıklanmıştır.

2.2.1 SMOTE

SMOTE algoritması dengesiz verilerde azınlık sınıfı için yeni, "sentetik" veriler oluşturan bir yöntemdir. Bu yöntemde halihazırda var olan sınıfı rassal olarak kopyalayarak çoğaltmayıp farklı bir yaklaşım gerçekleştirilir. Azınlık sınıfından bir örneklem seçilir, aynı sınıf içerisindeki *k*-en yakın komşusu belirlenir ve bu noktaları birbirine bağlayan hat boyunca sentetik gözlemler üretilir.

Sentetik veriyi oluşturmak için SMOTE algoritmasında seçili bir alt küme ile en yakın komşularından biri arasındaki fark hesaplanır. Bu fark 0 ile 1 arasındaki rassal bir faktörle ölçeklenir, ardından orijinal örnekleme eklenerek sentetik veri, ikisini

birleştiren doğru parçası üzerinde bir yere yerleştirilir. Bu sayede azınlık sınıfı çoğaltılır ve kullanılan sınıflandırıcı algoritmaların belirli noktalara aşırı uyum sağlaması problemi çözülmüş olur. SMOTE yaklaşımı, verileri basit ve rassal bir şekilde çoğaltmaya oranla daha verimli sonuçlar vermektedir. (Chawla vd., 2002).

2.2.2 Borderline-SMOTE

Borderline-SMOTE, SMOTE algoritmasının dengesiz verilerde özellikle sınıflandırma doğruluğunu arttırmak için geliştirilmiş versiyonudur. Sınıflandırma problemlerinde, sınırlara yakın olan değerler yanlış sınıflandırılmaya açıktır ve Borderline-SMOTE bu sorunu gidermek üzere geliştirilmiştir. Bu algoritma, SMOTE algoritmasından ve rassal aşırı örneklemeden daha iyi performans göstermiştir.

Algoritmadaki süreç sınır değerleri tespit etmekle başlar. Her bir azınlık sınıfı için *k*-en yakın komşular bulunur. Bir azınlık sınıfı, çoğunluk sınıflarıyla çevriliyse sınır değer olarak tanımlanır ve yanlış sınıflandırmaya yatkın olduğu tespit edilir. Azınlık değerlerle çevrili azınlık değerler ise "güvenli" sayılır ve sınır değer olarak tanımlanmaz. Sınır değerler tespit edildiğinde bu değerler için özel olarak sentetik veriler üretilir (Han vd., 2005).

2.2.3 ADASYN

ADASYN ileri düzey bir aşırı örnekleme yöntemidir ve asıl odaklandığı kısım sınıflandırması en zor olan azınlık verileridir. Aşırı örnekleme sürecine veri setindeki dengesizlik derecesi hesaplanarak başlanır. Her bir azınlık sınıfı gözlemi için o gözlemin *k*-en yakın komşuları bulunur ve o komşuların etrafındaki çoğunluk sınıfı gözlemlerinin oranı hesaba katılır. Bu işlemin sonucuna göre her bir azınlık sınıfı gözlemine bir ağırlık atanır. Ağırlığı yüksek yani çoğunluk sınıfı gözlemleri tarafından etrafı sarılmış gözlemler için özellikle gözlemler üretilir. Basit bir aşırı örneklemeden farkı hangi sınıfa ait olduğu anlaşılması zor olan gözlemlere odaklanılmasıdır (Haibo vd., 2008).

2.2.4 ROSE

ROSE algoritması ile azınlık ve çoğunluk sınıfları hesaba katarak veri çoğaltma işlemi gerçekleştirilir ve bunu yaparken her bir sınıfın koşullu yoğunluğu çekirdek temelli yaklaşım gerçekleştirilerek bulunur. Bu çekirdek temelli yaklaşım ile sentetik noktalar gerçek verilerin etrafına koyulur ve verilerin dağılımı basit rassal aşırı örneklemeden daha iyi yakalanır. Algoritmadaki parametreler değiştirilerek yeni veri setinin denge durumu kontrol edilebilir. Ayrıca ROSE hem kategorik hem sürekli değişkenler üzerinde çalışabildiği için ikili sınıflandırma problemlerinde oldukça yaygın bir şekilde kullanılabilmektedir (Lunardon vd., 2021).

2.3 Sınıflandırma Algoritmaları

Sınıflandırma algoritmalarından lojistik regresyon, Naive Bayes, rassal orman, XGBoost, yapay sinir ağları ve destek vektör makinesi algoritmalarının çalışma prensibi kısaca açıklanmıştır.

2.3.1 Lojistik Regresyon

Lojistik regresyon ikili sınıflandırmalar için kullanılan popüler ve temel bir denetimli makine öğrenmesi algoritmasıdır. Bu yöntemde, girdinin belli bir sınıfa ait olması durumunun olasılığı sigmoid fonksiyonu kullanılarak bulunur. Sigmoid fonksiyonu sonucunda çıktı 0 veya 1 olan bir değer olarak verilir ve böylece sınıfın etiketine karar verilmiş olur.

Bir sınıflandırma yapmak için öncelikle girdi özelliklerinin ağırlıklı ortalaması alınarak başlanır, bir hata (bias) eklenir ve olasılığı elde edebilmek için sigmoid fonksiyonu uygulanır. Sonrasında, gözlem bir sınıfa veya diğer sınıfa dahil edilir. Çoklu sınıflandırma söz konusu olduğunda softmax fonksiyonu kullanılır ve birden fazla sınıf için olasılıklar aynı anda hesaplanır.

Etiketli bir veri üzerinden optimum ağırlıklar ve hatayı öğrenirken bir kayıp fonksiyonu kullanılır ve bu genellikle cross-entropy kayıp fonksiyonu olur. Cross-entropy tahmin edilen ve gerçek değerler arasındaki farkı ölçmenin bir yöntemidir. Optimizasyon gradyan inişi gibi teknikler kullanılarak sağlanır ve gradyan iniş yönteminde de kaybı azaltmak için ağırlıklar yinelemeli olarak tekrar ayarlanır.

Lojistik regresyonda aşırı uyumu engelleme yöntemlerinden biri olan regülarizasyon desteklenir, büyük ağırlıklar penalize edilip aşırı uyum engellenir ve model, eğitilmediği veriler üzerinde daha iyi performans gösterir (Jurafsky ve Martin, 2019).

2.3.2 Naive Bayes

Naive Bayes, Bayes Teoremi üzerine kurulmuş olasılıksal bir sınıflandırma algoritmasıdır. Basitliği ve etkili olması sebebiyle, özellikle yüksek boyutlu verilerde sıklıkla kullanılır. Naive Bayes koşullu bağımsızlık varsayımı üzerinde hareket eder, bu da tüm değişkenlerin birbirinden bağımsız olduğunu varsaymak anlamına gelir. Bu bağımsızlık varsayımı gerçekçilikten uzak olsa da, arka plandaki hesaplamaları anlamlı derecede basitleştirir böylece Naive Bayes bilgisayarlarda hızlı çalışabilir ve de etkili olur. Sınıflandırıcı algoritmada, bir gözlemde her bir sınıf için olasılıklar öncül (prior) olasılıklar ile kombine ederek hesaplanır. Her bir sınıf için, sınıfın öncül olasılığı o sınıf altındaki gözlemlenen değişkenlerin olasılığıyla çarpılır. Sınıfı tahmin etmek için ise Bayes kuralından geldiği şekilde en yüksek sonraki (posterior) olasılığa sahip sınıf seçilir.

Bağımsızlık varsayımına karşın Naive Bayes sıklıkla iyi sınıflandırma sonuçları verir, özellikle doğru sınıf sadece en muhtemel sınıf olduğu durumlarda. Yüksek performansı sayesinde gerçek problemlere uygulanması konusunda tercih edilen bir algoritmadır. (Vikramkumar ve Trilochan, 2014).

2.3.3 Rassal Orman

Rassal orman birden fazla karar ağacı kullanarak sınıflandırma ve regresyon yapılabilen bir topluluk öğrenmesi yöntemidir. Bu teknikte, bootstrap aggregating (bagging) yöntemi kullanılarak varyans azaltılır ve aşırı uyum engellenir. Rassal ormanda, her bir ağaç veriden rassal olarak çekilmiş bir örneklem içine konulur. Bu rassallık ağaçlar arasındaki korelasyonu önler, modeli daha sağlam yaparak gürültüye dayanıklı olmasını sağlar ve üzerinde eğitilmediği verilerde daha iyi performans göstermesini sağlar.

Buradaki süreç birden çok karar ağacı yaratarak başlar ve bunların hepsi veriden alınan farklı kesitlerin ortasına inşa edilir. Ağaçların eğitimi sırasında, her bir nodülde

sadece belli sayıda değişken baz alınır böylece ağaçlardaki çeşitlilik sağlanır. Bütün ağaçlar eğitildiğinde, sınıflandırma problemleri için tahmin edilen sınıfı çoğunluk oyuna göre belirler, eğer konu regresyon problemiyse tahminlerin ortalaması alınır. Rassal orman değişken önemini bulmak ve hata oranları gibi ekstra avantajlar sağlayabilir böylece model performansı daha iyi anlaşılabilir. (Breiman, 2001).

2.3.4 XGBoost

XGBoost algoritması, gradyan-yükseltilmiş karar ağaçlarının yüksek performans gösteren bir uygulamasıdır. XGBoost, yapısal (structured) veriler üzerinde regresyon ve sınıflandırma konularında oldukça başarılıdır ve çeşitli makine öğrenmesi yarışmalarında en iyi sonuçları veren algoritma seçilmiştir.

XGBoost, başarı oranını yinelemeli olarak karar ağaçları oluşturup her bir yinelemede önceki ağacın hatalarını düzelterek arttırır. Standart yükseltme yöntemlerinden farklı olarak düzenlenmiş objektif fonksiyon (regularized objective function) ve ağırlıklı kantil taslağı (weighted quantile sketch) kullanılır ve bu şekilde büyük veri setleri üzerinde çalışılabilir ve kayıp veriler de bir seyreklik algoritması (sparsity-aware algorithm) kullanarak doldurulur.

XGBoost'un önemli bir özelliği paralel ve dağıtılmış hesaplamayla veya hafıza optimizasyonuyla ölçeklenebilir olmasıdır. Algoritma, sütun alt örnekleme ve çekirdek dışı hesaplama gibi yöntemlerle milyarlarca veri üzerinde limitli kaynaklar olsa dahi çalışabilir. Efektif ve esnek bir algoritma olması makine öğrenmesi dünyasında popüler olmasını sağlamıştır ve diğer topluluk öğrenmesi yöntemlerinden birden fazla yönüyle daha başarılı olduğu ortaya çıkmıştır (Chen ve Guestrin, 2016).

2.3.5 Yapay Sinir Ağları

Yapay sinir ağları insan beyninden esinlenerek üretilmiş bir modeldir ve birçok basit, nöron adı verilen birimlerden oluşur. Bu basit nöronlar kolektif olarak çalıştığında, geleneksel makine öğrenmesinin zorlandığı veya başaramadığı örüntüleri tanıma ve karar verme gibi konularda başarı gösterir.

Yapay sinir ağlarında bir nöron tarafından girdi alınır, bu girdiye ağırlıklar eklenir, toplanır ve bir aktivasyon fonksiyonundan geçirilir ve sonucunda çıktı değeri elde

edilir. Yapay sinir ağlarının 3 katmanı vardır (giriş katmanı, gizli katman ve çıkış katmanı) ve veri bu katmanlardan eğitim yöntemiyle geçer ardından verideki örüntü tespit edilir. Eğitim sürecinde tahmin edilen ve gerçek değerler arasındaki farka göre ağırlıklar ayarlanır. Bu yinelemeli süreç sayesinde yapay sinir ağlarında öğrenme işlemi gerçekleşir ve ağırlıklar yeni veride kullanılmak üzere ayarlanır. Görüntü işleme, doğal dil işleme ve tıbbi tanı koyma gibi birçok alanda yapay sinir ağları yöntemi kullanılmaktadır. Dağıtılmış yapısı sayesinde gürültülü veya eksik verilerde bile etkili performans gösterir ve bu da onu modern makine öğrenmesi alanında güçlü bir araç yapar (Uhrig, 1995).

2.3.6 Destek Vektör Makinesi

Destek Vektör Makinesi bir denetimli öğrenme algoritmasıdır ve genellikle sınıflandırma problemleri için kullanılır. En temelde, 2 boyutlu düzlem üzerinde iki farklı sınıfı maksimum hiper düzlem ile ayırmak suretiyle çalışır. DVM'de, en yakın iki farklı veri sınıfı arasındaki boşluk (veya marjin) maksimize etmeye çalışılır. Bu marjin maksimizasyonu daha sağlam ve aşırı öğrenmeye dirençli bir model oluşturur.

DVM'nin çekirdek fonksiyonu değiştirilerek lineer olmayan veriler üzerinde de çalışması sağlanabilir. Çekirdek fonksiyonu değiştirilerek DVM'nin veriyi daha yüksek bir boyuta taşıması sağlanır böylece lineer ayırma gerçekleşebilir. Sıklıkla kullanılan bazı çekirdek fonksiyonları polinom fonksiyonu, radyal bazlı fonksiyon ve sigmoid fonksiyonudur.

DVM ayrıca spesifik veri eğitim noktalarına dayanır, buna "destek vektörleri" denir. Destek vektörleri karar sınırını oluştururken kritik rol oynar. DVM metin ve görüntü sınıflandırma ve biyoinformatik alanında sıklıkla kullanılır ve hem lineer hem de lineer olmayan problemlerin üstesinden gelebilir (Hearst vd., 1998).

2.4 Performans Metrikleri

Çalışmada performans metrikleri olarak karmaşıklık matrisi üzerinden hesaplanan doğruluk, dengelenmiş doğruluk, duyarlılık, özgüllük ve F1-Skoru kullanılmıştır.

Tablo 2.1 Karmaşıklık matrisi

		TAHMİN	
		Pozitif (1)	Negatif (0)
GERÇEK	Pozitif (1)	Gerçek Pozitif (GP)	Yanlış Negatif (YN)
	Negatif (0)	Yanlış Pozitif (YP)	Gerçek Negatif (GN)

2.4.1 Doğruluk

Doğruluk (accuracy) pozitif veya negatif olması fark etmeksizin doğru sınıflandırılmış değerlerin oranını gösterir ve genel bir performans ölçütü olarak kullanılır. Dengesiz verilerde yanıltıcı olabilir.

$$Doğruluk = \frac{\mathrm{GP} + \mathrm{GN}}{\mathrm{GP} + \mathrm{GN} + \mathrm{YP} + \mathrm{YN}}$$

2.4.2 Duyarlılık

Duyarlılık (sensitivity), sadece doğru olarak sınıflandırılan gerçek pozitiflerin oranını verir.

$$Duyarlılık = \frac{\mathrm{GP}}{\mathrm{GP} + \mathrm{YN}}$$

2.4.3 Özgüllük

Özgüllük (specificity), sadece doğru olarak sınıflandırılan gerçek negatiflerin oranını verir.

$$\text{Özgüllük} = \frac{\mathrm{GN}}{\mathrm{GN} + \mathrm{YP}}$$

2.4.4 Dengelenmiş Doğruluk

Dengelenmiş doğruluk (balanced accuracy) duyarlılık ve özgüllük metriklerinin ortalamasıdır ve dengesiz veriler üzerinde çalışırken etkilidir.

$$Dengelenmiş\ Doğruluk = \frac{\text{Duyarlılık} + \text{Özgüllük}}{2}$$

2.4.5 F1-Skoru

F1-Skoru, hassasiyet (precision) ve duyarlılığın harmonik ortalamasıdır. Bu şekilde bu iki metriği dengeleyerek tek bir metrik elde etmeyi sağlar. Dengesiz verilerle çalışılırken faydalıdır.

$$F_1 = 2 \cdot \frac{\text{Hassasiyet} \cdot \text{Duyarlılık}}{\text{Hassasiyet} + \text{Duyarlılık}}$$

F1-Skoru hesabında kullanılan hassasiyet değeri aşağıdaki eşitlikten hesaplanır.

$$Hassasiyet = \frac{\text{GP}}{\text{GP} + \text{YP}}$$

BÖLÜM ÜÇ
UYGULAMA

3.1 Veri Setinin Tanıtılması

Kalp hastalığı veri seti UC Irvine Machine Learning Repository'ye 1988 yılında bağışlanmıştır. Bu çalışmada Kaggle üzerindeki versiyonu kullanılmıştır (Lapp, 2019). Toplam 1025 gözlem ve 14 değişkenden oluşmaktadır, değişken isimleri ve açıklamaları Tablo 3.1'de verilmiştir.

Tablo 3.1 Değişken isimleri ve açıklamaları

Değişken	Açıklama
Age	Yaş
Trestbps	Dinlenme kan basıncı (hastaneye kabul anında mm Hg cinsinden)
Chol	Serum kolesterolü mg/dl cinsinden
Thalach	Maksimum kalp atış hızı
Oldpeak	Egzersizle oluşan ST baskılanması dinlenmeye göre
Ca	Majör damar sayısı (0-3)
Target	Yanıt değişkeni (1 = hasta; 0 = hasta değil)
Sex	(1 = erkek; 0 = kadın)
Cp	Göğüs ağrısı tipi (0-3)
Fbs	Açlık kan şekeri > 120 mg/dl (1 = doğru; 0 = yanlış)

Değişken	Açıklama
restecg	Dinlenme elektrokardiyografi sonuçları
exang	Egzersiz kaynaklı angina (1 = evet; 0 = hayır)
Slope	Zirve egzersiz ST segmentinin eğimi
thal	0 = normal; 1 = sabit kusur; 2 = geri döndürülebilir kusur

3.2 Veri Ön İşleme

Şekil 3.1'de gösterildiği üzere gerekli olan sütunlar ikili kodlanmıştır ve işleme böyle devam edilmiştir.

```
categorical_cols <- c("sex", "cp", "fbs", "restecg", "exang", "slope", "thal", "ca")
dummies <- dummyVars(~ ., data = heart[categorical_cols])
one_hot_data <- predict(dummies, newdata = heart[categorical_cols])
heart <- cbind(heart[ , !(names(heart) %in% categorical_cols)], one_hot_data)
```

Şekil 3.1 İkili Kodlama

3.3 Tanımlayıcı İstatistikler

Şekil 3.2'de verinin ilk 6 satırı gösterilmiştir.

```
> head(heart)
  age trestbps chol thalach oldpeak ca target sex cp fbs restecg exang slope thal
1  52      125  212     168     1.0  2      0   1  0   0       1     0     2    3
2  53      140  203     155     3.1  0      0   1  0   1       0     1     0    3
3  70      145  174     125     2.6  0      0   1  0   0       1     1     0    3
4  61      148  203     161     0.0  1      0   1  0   0       1     0     2    3
5  62      138  294     106     1.9  3      0   0  0   1       1     0     1    2
6  58      100  248     122     1.0  0      1   0  0   0       0     0     1    2
```

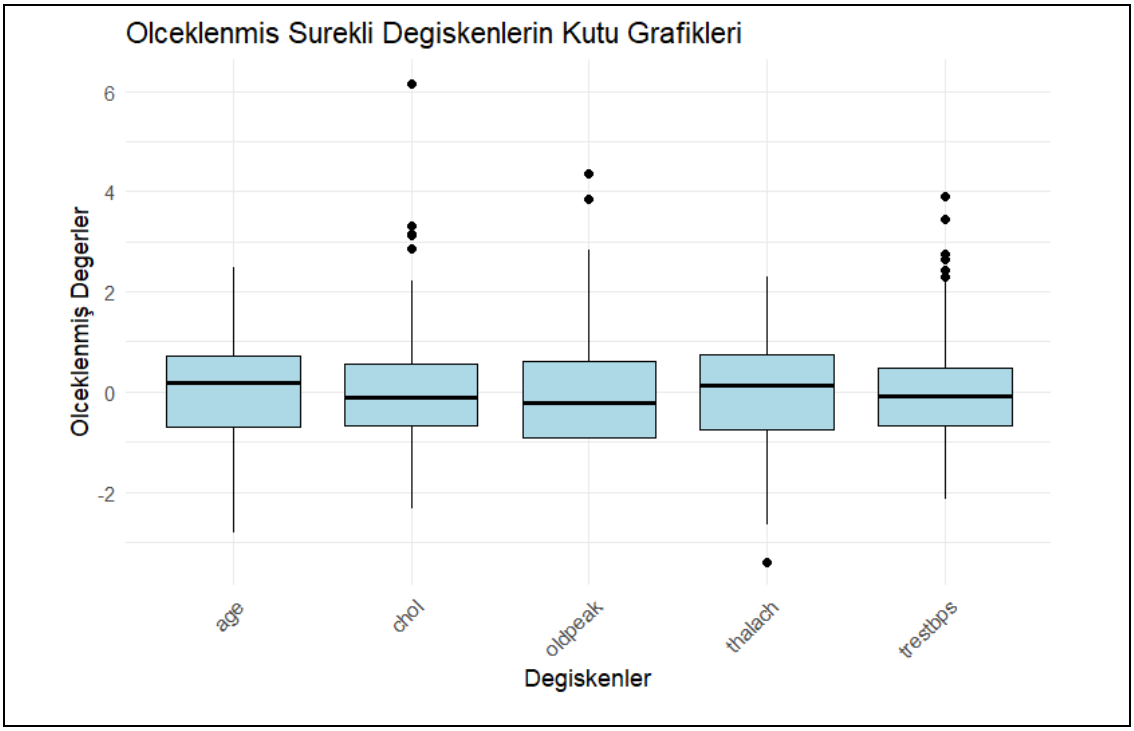
Şekil 3.2 Verinin ilk 6 satırı

Şekil 3.3'te veride kayıp değerler olmadığı, sürekli değişkenlerin basit tanımlayıcı istatistikleri ve kategorik değişkenlerin düzeylerine ait frekansları verilmiştir.

```
> summary(heart)
      age            trestbps          chol         thalach         oldpeak        target  sex     cp
 Min.   :29.00   Min.   : 94.0   Min.   :126   Min.   : 71.0   Min.   :0.000   1:526   0:312   0:497
 1st Qu.:48.00   1st Qu.:120.0   1st Qu.:211   1st Qu.:132.0   1st Qu.:0.000   0:499   1:713   1:167
 Median :56.00   Median :130.0   Median :240   Median :152.0   Median :0.800                   2:284
 Mean   :54.43   Mean   :131.6   Mean   :246   Mean   :149.1   Mean   :1.072                   3: 77
 3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:275   3rd Qu.:166.0   3rd Qu.:1.800
 Max.   :77.00   Max.   :200.0   Max.   :564   Max.   :202.0   Max.   :6.200
 fbs      restecg exang   slope  thal   ca
 0:872   0:497   0:680   0: 74   0:  7   0:578
 1:153   1:513   1:345   1:482   1: 64   1:226
         2: 15           2:469   2:544   2:134
                         3:410   3: 69
                                 4: 18
```

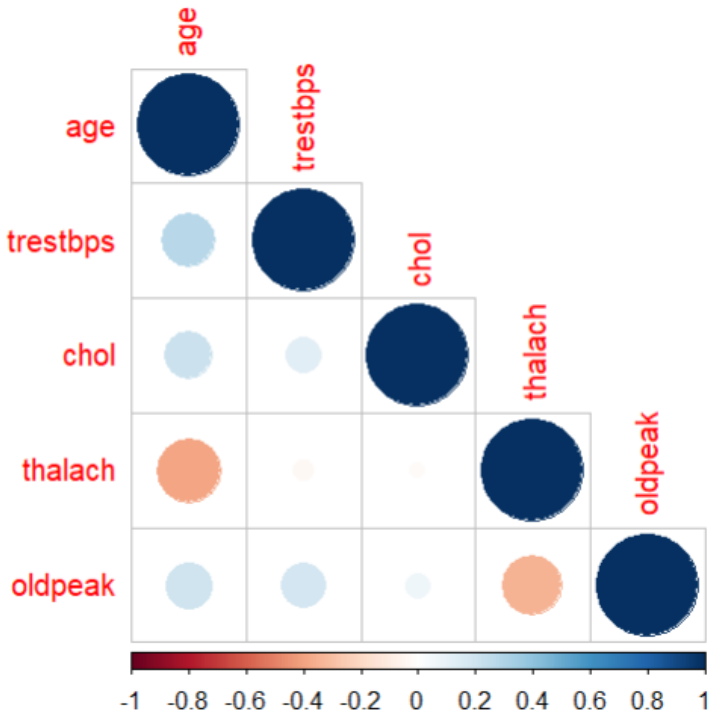
Şekil 3.3 Tanımlayıcı istatistikler

Şekil 3.4'te ölçeklenmiş haldeki sürekli değişkenlerin kutu grafikleri verilmiştir. "age" değişkeninde uç değer yoktur fakat diğer değişkenlerde uç değerler gözlemlenmektedir. "chol" değişkeni en ekstrem uç değere sahiptir, "oldpeak" değişkeninde 2 adet, "thalach" değişkeninde 1 adet uç değer gözlemlenirken "trestbps" değişkeninde birbirine yakın çokça uç değer vardır.



Şekil 3.4 Kutu grafikleri

Şekil 3.5'te "thalach" ve "age" ile "oldpeak" ve "thalach" değişkenleri arasında hafif seviyede bir negatif korelasyon görülmektedir. Bunların dışında değişkenler arasında ciddi bir korelasyon yoktur.



Şekil 3.5 Korelasyon matrisi

3.4 Verinin Dengesiz Hale Getirilmesi

Orijinalinde dengeli sınıflara sahip olan kalp hastalığı veri seti çalışmanın amacı doğrultusunda, azınlık sınıfının çoğunluk sınıfına oranı %5, %15, %25 olacak şekilde Şekil 3.6'daki kod kullanılarak dengesiz hale getirilmiştir.

```r
set.seed(123)
create_imbalanced <- function(majority, minority, ratio) {
  minority_sample <- minority %>% sample_frac(ratio)
  bind_rows(majority, minority_sample)
}

class_0 <- heart %>% filter(target == 0)
class_1 <- heart %>% filter(target == 1)

heart_5_imbalanced <- create_imbalanced(class_0, class_1, 0.05)
heart_15_imbalanced <- create_imbalanced(class_0, class_1, 0.15)
heart_25_imbalanced <- create_imbalanced(class_0, class_1, 0.25)
```

Şekil 3.6 Veriyi dengesiz hale getiren kod

Veriler dengesiz hale getirildiğinde oluşan yeni veri setlerinde hedef değişkendeki her iki gözlemin sayıları Tablo 3.2'de mevcuttur. "1" kalp hastası, "0" sağlıklı birey olarak kodlanmıştır.

Tablo 3.2 Dengesiz verilerde hedef değişkenin gözlem sayıları

Veri Seti	**1 / 0**
%5 Dengesiz	26 / 499
%15 Dengesiz	79 / 499
%25 Dengesiz	132 / 499
Orijinal veri (Dengeli)	499 / 526

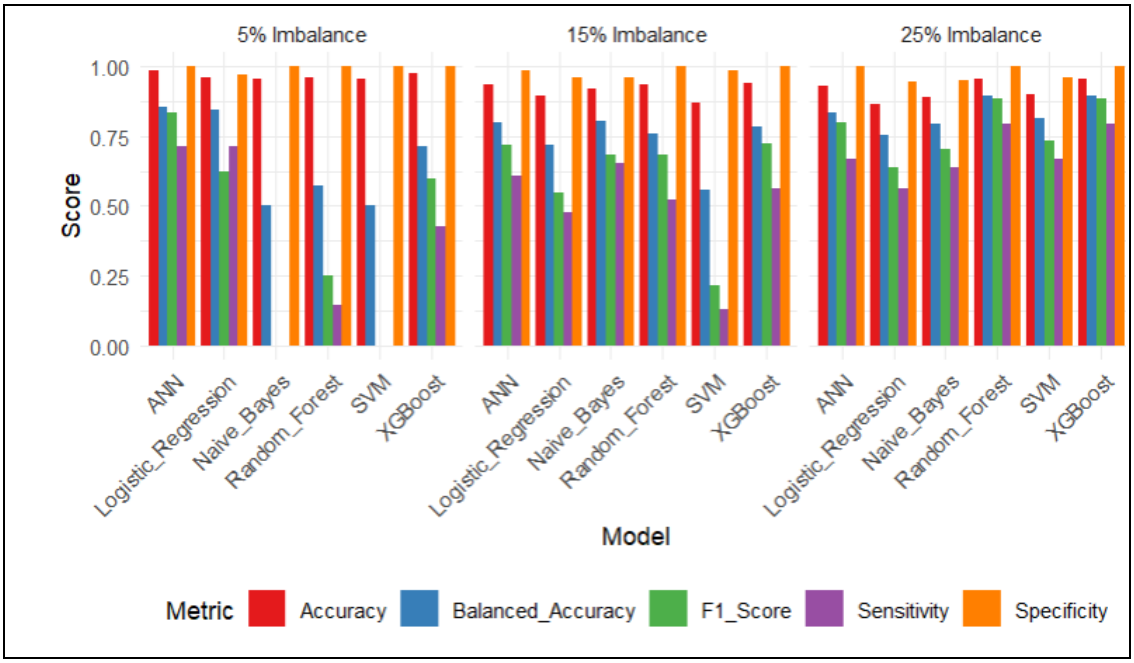
3.5 Analiz Sonuçları

Dengesiz haldeki veri setleri üzerindeki makine öğrenmesi algoritmalarının performans metrikleri Tablo 3.3'te verilmiştir. Şekil 3.7'de dengesizlik durumlarına göre sınıflandırma algoritmalarının performans değerleri bir arada değerlendirilmek üzere görselleştirilmiştir.

Tablo 3.3 Dengesiz veri setleri üzerindeki performans metrikleri

Veri Seti	Model	Doğruluk	Duyarlılık	Özgüllük	Dengelenmiş Doğruluk	F1-Skoru
%5 Dengesizlik	YSA	**0.987**	**0.714**	**1.000**	**0.857**	**0.833**
	Lojistik Reg.	0.961	**0.714**	0.973	**0.843**	0.625
	Naive Bayes	0.955	0.000	**1.000**	0.500	NA
	Rassal Orman	0.961	0.142	**1.000**	0.571	0.250
	DVM	0.955	0.000	**1.000**	0.500	NA
	XGBoost	0.974	0.428	**1.000**	0.714	0.600
%15 Dengesizlik	YSA	0.936	**0.986**	0.608	0.797	0.717
	Lojistik Reg.	0.895	0.478	0.959	0.718	0.550
	Naive Bayes	0.918	0.652	0.959	**0.805**	0.681
	Rassal Orman	0.936	0.521	**1.000**	0.760	0.685
	DVM	0.872	0.130	0.986	0.558	0.214
	XGBoost	**0.941**	0.565	**1.000**	0.782	**0.722**
%25 Dengesizlik	YSA	0.930	0.666	**1.000**	0.833	0.800
	Lojistik Reg.	0.867	0.564	0.946	0.755	0.637
	Naive Bayes	0.888	0.641	0.953	0.797	0.704
	Rassal Orman	**0.957**	**0.794**	**1.000**	**0.897**	**0.885**
	DVM	0.898	0.666	0.959	0.813	0.732
	XGBoost	**0.957**	**0.794**	**1.000**	**0.897**	**0.885**

Tablo 3.3 ve Şekil 3.7'ye göre, %5 dengesizlik durumunda en yüksek performans değerleri YSA ile elde edilmiştir. %15 dengesizlik durumunda sınıflandırma algoritmalarının performans metrikleri birbirlerine daha yakın sonuçlar vermiş, XGBoost yöntemi birçok metrikte daha yüksek değer almıştır. Ancak duyarlılık değeri en yüksek YSA ile elde edilmiştir. %25 dengesizlik durumunda en yüksek performans metrikleri rassal orman ve XGBoost algoritmalarından elde edilmiştir.

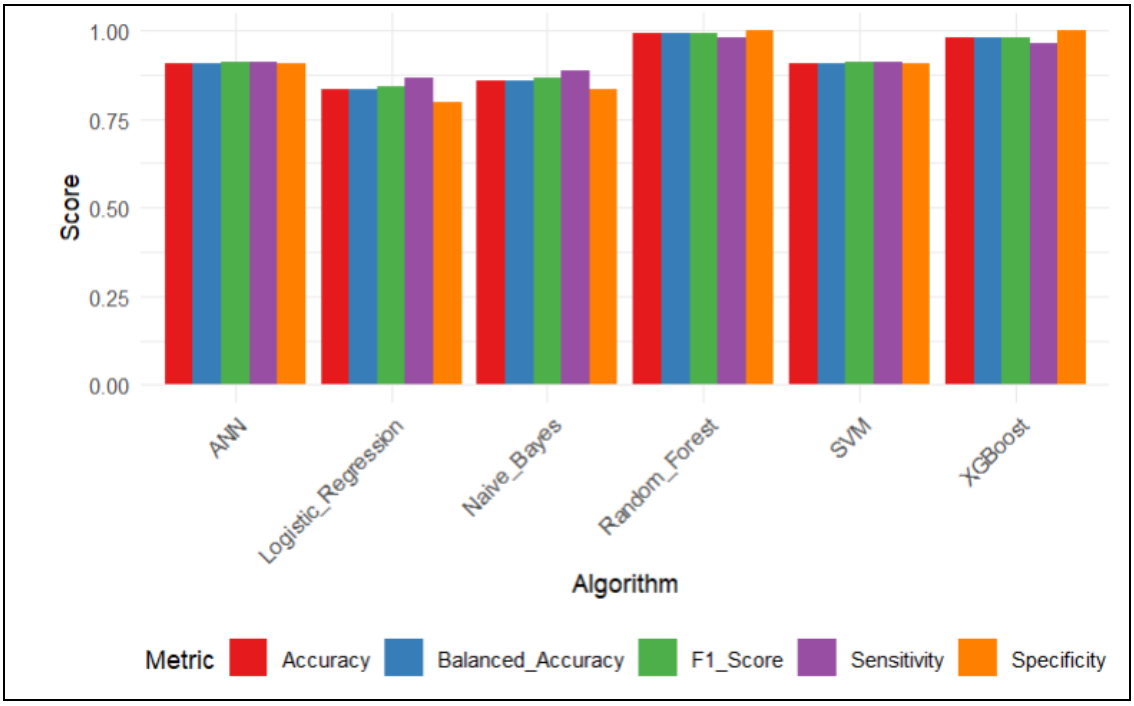


Şekil 3.7 Dengesiz hale getirilmiş veri setleri üzerindeki performanslar

Tablo 3.4'te dengesiz hale getirilmemiş, orijinal veri setinin sınıflandırma algoritmalarından elde edilen performans değerleri gösterilmiştir. Şekil 3.8'de karşılaştırma amaçlı görselleştirilmiştir. Rassal orman algoritmasının tüm metrikler dikkate alındığında en iyi performansı gösterdiği anlaşılmaktadır. Çok yakın bir farkla XGBoost algoritması yüksek performans gösterirken, lojistik regresyon diğer algoritmalara göre en düşük performansı göstermiştir.

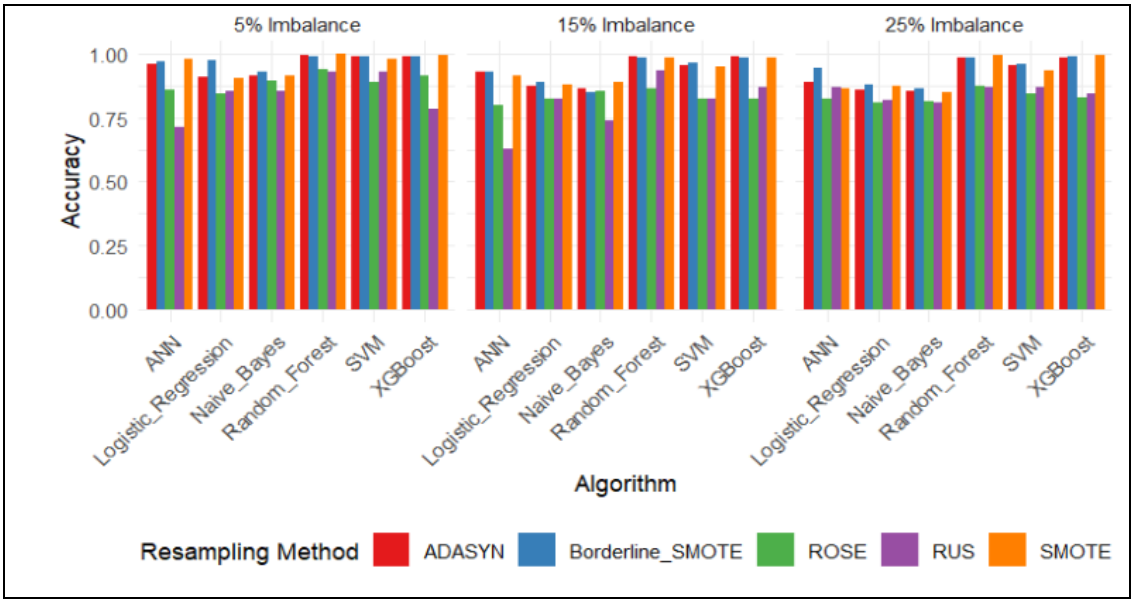
Tablo 3.4 Orijinal veri seti üzerindeki performans metrikleri

Veri Seti	Model	Doğruluk	Duyarlılık	Özgüllük	Dengelenmiş Doğruluk	F1-Skoru
Orijinal Veri	YSA	0.908	0.906	0.910	0.908	0.910
	Lojistik Reg.	0.833	0.798	0.866	0.832	0.842
	Naive Bayes	0.859	0.832	0.885	0.858	0.866
	Rassal Orman	**0.990**	**1.000**	**0.980**	**0.990**	**0.990**
	DVM	0.908	0.906	0.910	0.908	0.910
	XGBoost	0.980	**1.000**	0.961	0.980	0.980



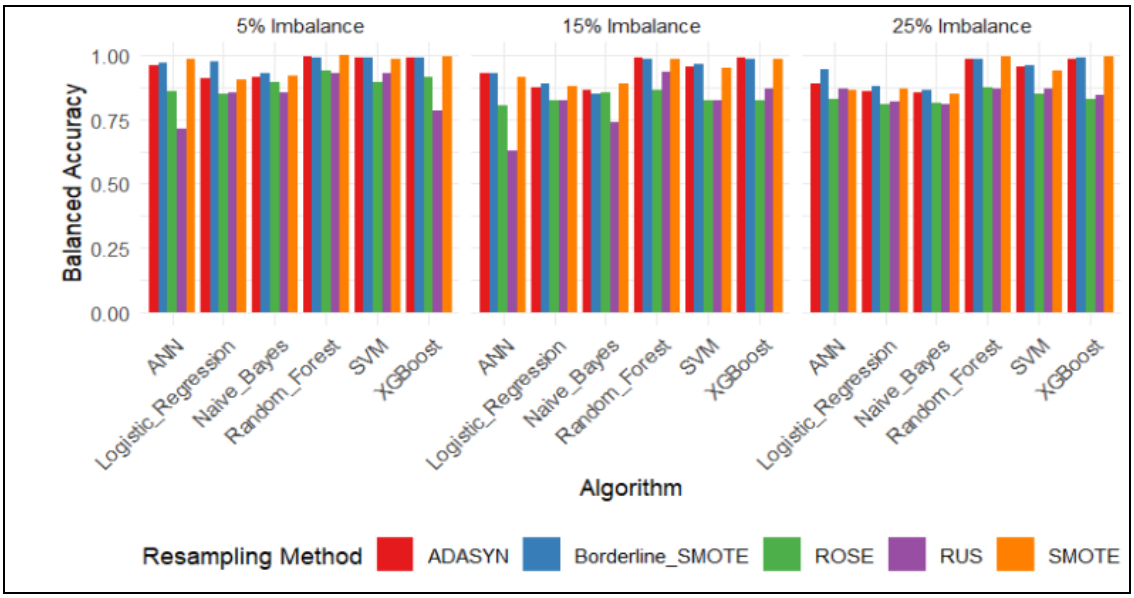
Şekil 3.8 Orijinal dengeli veri seti üzerindeki performanslar

%5, %15 ve %25 dengesizlikten dengelenme durumları için elde edilen tüm performans metrikleri Ekler kısmında tablo olarak verilmiştir. Şekil 3.9 aşırı örnekleme ile dengelenmiş veri setlerinde uygulanmış makine öğrenmesi algoritmalarının doğruluk performanslarını göstermektedir. ADASYN, Borderline-SMOTE ve SMOTE'un doğruluk oranları, RUS ve ROSE ile karşılaştırıldığında birçok sınıflandırma algoritmasında oldukça yüksek elde edilmiştir. Özellikle rassal orman algoritması farklı dengesizlik durumları ve aşırı örnekleme yöntemlerinde en yüksek duyarlılık değerine ulaşmıştır. RUS algoritmasıyla dengelenen veri setlerinde genel olarak düşük performans gözlemlenmiştir, özellikle %5 ve %15 dengesizlikten dengelenme durumlarında yapay sinir ağlarında bu durum belirgindir.



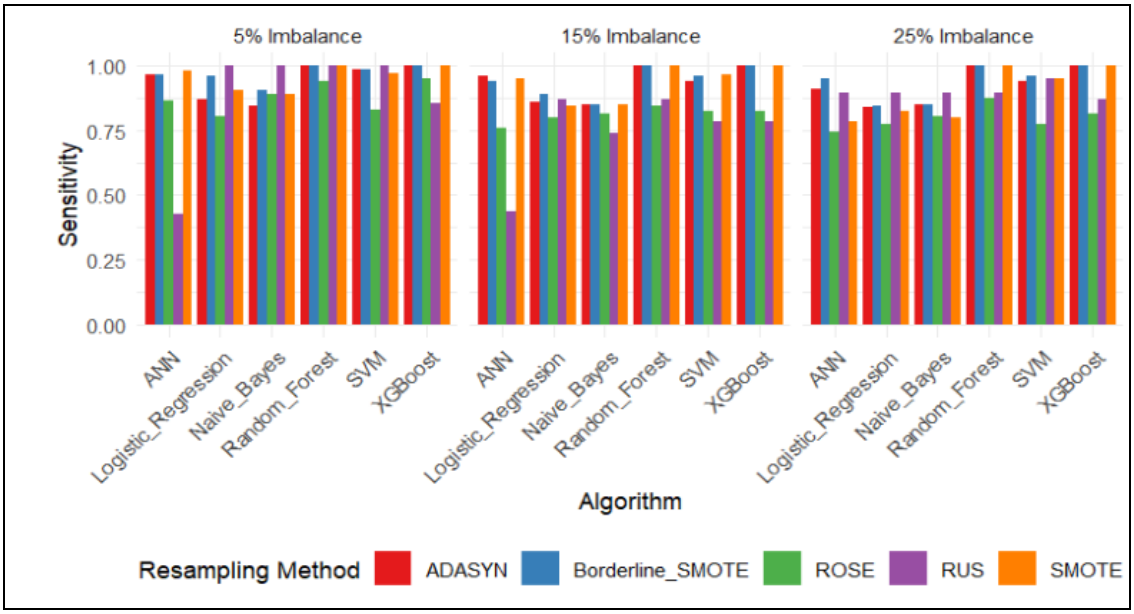
Şekil 3.9 Dengelenmiş veri setleri üzerindeki doğruluk performansları

Şekil 3.10'da görüldüğü üzere dengelenmiş doğruluk grafikleri, doğruluk grafikleriyle çok benzer çıkmıştır. ADASYN, Borderline-SMOTE ve SMOTE bu metrikte de yüksek performans göstermektedir. %5 dengesizlik durumunda özellikle RUS ile dengelenmiş veride performansta düşüşler görülmektedir. RUS ve ROSE yöntemleri ancak sınıflandırma algoritmalarından rassal orman algoritması ile birlikte kullanıldığında daha yüksek dengelenmiş doğruluk değeri elde edilebilmektedir. Rassal orman algoritması üç dengesizlik durumu için de dengelenmiş doğruluk değerinde en yüksek performansı göstermiştir.



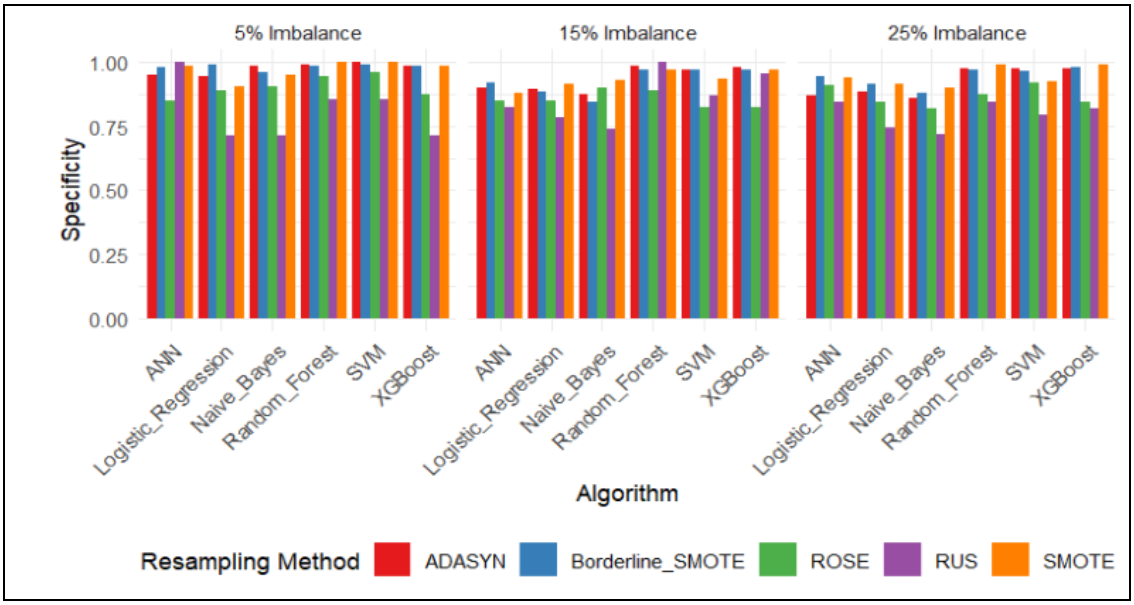
Şekil 3.10 Dengelenmiş veri setleri üzerindeki deng. doğruluk performansları

Şekil 3.11 dengelenmiş veri setleri için duyarlılık performansını göstermektedir. %5 ve %15 dengesizlik durumlarında, RUS yöntemi ile yapay sinir ağları ile en düşük duyarlılık değerleri elde edilmiştir. RUS ve ROSE yöntemleri haricinde, ADASYN, Borderline-SMOTE ve SMOTE ile dengelenmiş veri setlerinde genel olarak rassal orman ve XGBoost sınıflandırma algoritmaları ile en yüksek duyarlılık değerleri elde edilmiştir.



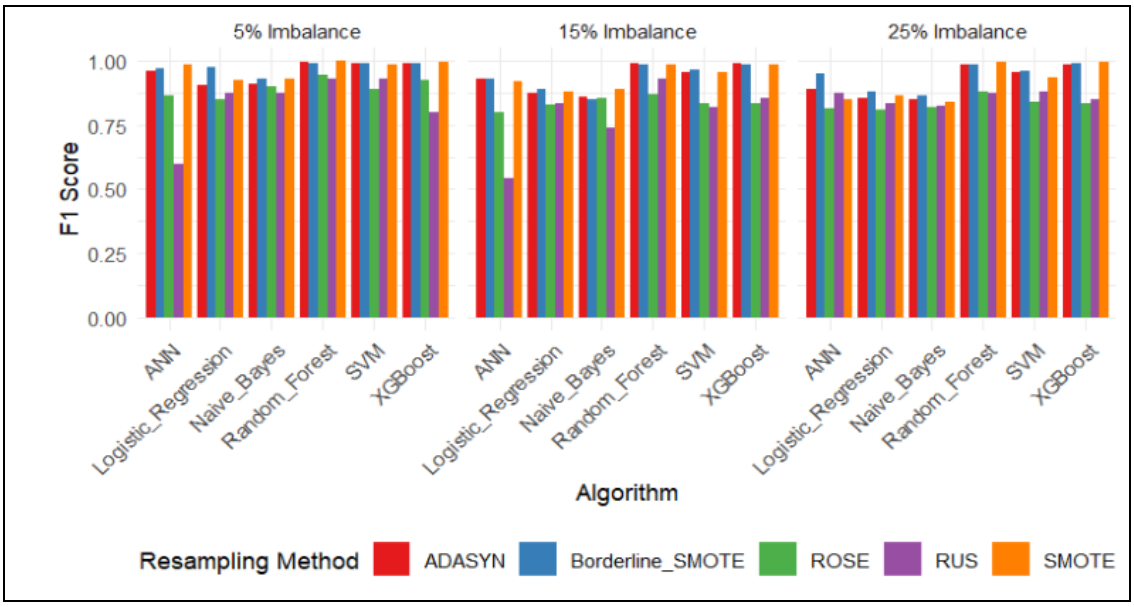
Şekil 3.11 Dengelenmiş veri setleri üzerindeki duyarlılık performansları

Şekil 3.12 dengelenmiş veri setleri için özgülllük performansını göstermektedir. Genel olarak ADASYN, Borderline SMOTE ve SMOTE yöntemleri ile kullanılan sınıflandırma algoritmalarında yüksek özgüllük değerleri elde edilmiştir. En düşük elde edilen özgüllük değerleri %5 dengesizlik durumu için RUS yöntemi kullanılarak lojistik regresyon ve Naive Bayes ile elde edilmiştir. Farklı dengesizlik durumlarına göre net bir sınıflandırma algoritması ön plana çıkmamıştır.



Şekil 3.12 Dengelenmiş veri setleri üzerindeki özgüllük performansları

Şekil 3.13'e bakıldığında, F1 skorları %5 dengeli durumunda RUS ve ROSE yöntemlerinin belli sınıflandırma algoritmalarıyla kullanımı haricinde genel olarak yüksek ve birbirine yakın değerler verdiği gözlenmiştir. RUS yönetimi %5 ve %15 dengesizlik durumlarında oldukça düşük performans gösterirken, %25 dengesizlikte performansını arttırmıştır. Genel olarak en iyi F1-skoru performansı ADASYN, Borderline-SMOTE ve SMOTE yöntemleri ile dengelenmiş verilerden ve sırasıyla rassal orman, XGBoost ve DVM sınıflandırma algoritmalarından elde edilmektedir.



Şekil 3.13 Dengelenmiş veri setleri üzerindeki F1-skoru performansları

BÖLÜM DÖRT
SONUÇ

Bu çalışmada hedef değişkendeki sınıflara ait gözlem değerlerinin birbirine eşit olmaması veya eşit olmaya yakın olmaması durumunda ortaya çıkan dengesiz veri sorununu gidermek üzere kullanılan aşırı örnekleme yöntemleri karşılaştırılmıştır. Veri setlerini dengelemek üzere aşırı örnekleme yöntemlerinden ADASYN, Borderline-SMOTE, ROSE ve SMOTE yöntemleri kullanılmıştır.

ADASYN, Borderline-SMOTE ve SMOTE yöntemleri, doğru bir sınıflandırma için önemli olan azınlık sınıfının iyi temsil edilmesini sağlayarak makine öğrenmesi algoritmalarının performansını anlamlı derecede arttırmıştır.

RUS yöntemi verileri azaltarak dengelediği için ciddi oranda bilgi kaybına sebep olmaktadır ve performans metriklerinde RUS'un sıklıkla oldukça düşük değerler verdiği gözlemlenebilir. Aynı zamanda ROSE algoritması ile dengelenen veri setlerinde de makine öğrenmesi modellerinin performansları ADASYN, Borderline-SMOTE ve SMOTE'a oranla arka planda kalmıştır.

Rassal orman algoritması bütün veri setleri arasında en iyi performansı sergileyen algoritma olmuştur. XGBoost algoritması ise genel olarak ikinci en iyi performansı gösteren sınıflandırma algoritmasıdır. Bu algoritmaların, dengesizlik derecesine bağlı olarak, dengesiz verilerde kullanılabileceğini ayrıca aşırı örnekleme yöntemleri ile dengelenmiş verilerle birlikte kullanıldıklarında daha iyi performans gösterdiği söylenebilir. Lojistik regresyon, yapay sinir ağları, destek vektör makinesi ve Naive Bayes sınıflandırma algoritmalarının performansları aşırı örneklenmiş veriler üzerinde çalışıldığında rassal orman ve XGBoost algoritmaları kadar iyi performans değerleri göstermemiştir.

Dengesiz verilerle çalışırken kullanılacak olan performans metriği oldukça önemlidir. Uygun aşırı örnekleme yönteminin ve uygun sınıflandırma algoritmasının belirlenmesinde bu çalışmada kullanılan performans metrikleri kritik bir rol oynamıştır. Sadece tek bir performans metriğine göre yorum yapılmasının yanıltıcı olabileceği bu çalışmayla tekrar gösterilmiştir.

KAYNAKLAR

Azlim Khan, A. K., ve Ahamed Hassain Malim, N. H. (2023). Comparative studies on resampling techniques in machine learning and deep learning models for drug-target interaction prediction. *Molecules*, *28*(4), 1663. https://doi.org/10.3390/molecules28041663

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

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. https://doi.org/10.1613/jair.953

Chen, T., ve Guestrin, C. (2016, June 10). *XGBoost: A scalable tree boosting system*. arXiv.org. https://doi.org/10.48550/arXiv.1603.02754

Haibo He, ve Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/tkde.2008.239

Haibo He, Yang Bai, Garcia, E. A., ve Shutao Li. (2008). Adasyn: Adaptive Synthetic Sampling Approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/ijcnn.2008.4633969

Han, H., Wang, W.-Y., ve Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. *Lecture Notes in Computer Science*, 878–887. https://doi.org/10.1007/11538059_91

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., ve Scholkopf, B. (1998). Support Vector Machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

Jurafsky, D., ve Martin, J. H. (2019). *Speech and language processing*. Stanford University.

Kotsiantis, S., Kanellopoulos, D., ve Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*(1). https://scholar.google.ca/citations?view_op=view_citation&hl=en&user=h6zwXYMAAAAJ&citation_for_view=h6zwXYMAAAAJ:9yKSN-GCB0IC

Lapp, D. (2019, June 6). *Heart disease dataset*. Kaggle. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data

Lunardon, N., Torelli, N., ve Menardi, G. (2021). *Rose: Random over-sampling examples*. Package 'ROSE.' https://cran.r-project.org/web/packages/ROSE/ROSE.pdf

Newaz, A., Mohosheu, Md. S., Noman, Md. A., ve Jabid, T. (2024). IBRF: Improved balanced random forest classifier. *2024 35th Conference of Open Innovations Association (FRUCT)*, 501–508. https://doi.org/10.23919/fruct61870.2024.10516372

Santoso, B., Wijayanto, H., Notodiputro, K. A., ve Sartono, B. (2017). Synthetic over sampling methods for handling class imbalanced problems : A Review. *IOP Conference Series: Earth and Environmental Science*, *58*, 012031. https://doi.org/10.1088/1755-1315/58/1/012031

Sağlam, F. (2021). Optimization based undersampling for imbalanced classes. *Adıyaman University Journal of Science*. https://doi.org/10.37094/adyujsci.884120

Sun, Y., Wong, A. K., ve Kamel, M. S. (2009). Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719. https://doi.org/10.1142/s0218001409007326

Uhrig, R. E. (1995). Introduction to artificial neural networks. *Proceedings of IECON ’95 - 21st Annual Conference on IEEE Industrial Electronics*, *1*, 33–37. https://doi.org/10.1109/iecon.1995.483329

Vikramkumar, B, V., ve Trilochan. (2014, April 3). *Bayes and naive bayes classifier*. arXiv.org. https://doi.org/10.48550/arXiv.1404.0933

EK 1. %5 Dengesizlik Durumundan Aşırı Örneklenmiş Veri Seti Üzerindeki Performans Metrikleri

Yöntem	Model	Doğruluk	Duyarlılık	Özgüllük	Dengelenmiş Doğruluk	F1-Skoru
ADASYN	YSA	0.959	0.966	0.952	0.959	0.960
	Lojistik Reg.	0.909	0.872	0.945	0.909	0.905
	Naive Bayes	0.915	0.845	0.986	0.916	0.909
	Rassal Orman	**0.996**	**1.000**	0.993	**0.996**	**0.996**
	DVM	0.993	0.986	**1.000**	0.993	0.993
	XGBoost	0.993	**1.000**	0.986	0.993	0.993
Borderline-SMOTE	YSA	0.972	0.966	0.979	0.972	0.972
	Lojistik Reg.	0.976	0.959	**0.993**	0.976	0.976
	Naive Bayes	0.932	0.906	0.958	0.932	0.931
	Rassal Orman	**0.993**	**1.000**	0.986	**0.993**	**0.993**
	DVM	0.989	0.986	**0.993**	0.989	0.989
	XGBoost	**0.993**	**1.000**	0.986	0.993	0.993
RUS	YSA	0.714	0.428	**1.000**	0.714	0.600
	Lojistik Reg.	0.857	**1.000**	0.714	0.857	0.875
	Naive Bayes	0.857	**1.000**	0.714	0.857	0.875
	Rassal Orman	**0.928**	**1.000**	0.857	**0.928**	**0.933**
	DVM	**0.928**	**1.000**	0.857	**0.928**	**0.933**
	XGBoost	0.785	0.857	0.714	0.785	0.800
ROSE	YSA	0.858	0.867	0.849	0.858	0.867
	Lojistik Reg.	0.846	0.807	0.890	0.848	0.848
	Naive Bayes	0.897	0.891	0.904	0.897	0.902
	Rassal Orman	**0.942**	0.939	0.945	**0.942**	**0.945**
	DVM	0.891	0.831	**0.958**	0.895	0.890
	XGBoost	0.916	**0.951**	0.876	0.914	0.923

Yöntem	Model	Doğruluk	Duyarlılık	Özgüllük	Dengelenmiş Doğruluk	F1-Skoru
SMOTE	YSA	0.982	0.979	0.988	0.984	0.986
	Lojistik Reg.	0.905	0.906	0.905	0.905	0.924
	Naive Bayes	0.914	0.892	0.952	0.922	0.930
	Rassal Orman	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	DVM	0.982	0.973	**1.000**	0.986	0.986
	XGBoost	0.995	**1.000**	0.988	0.994	0.996

EK 2. %15 Dengesizlik Durumundan Aşırı Örneklenmiş Veri Seti Üzerindeki Performans Metrikleri

Yöntem	Model	Doğruluk	Duyarlılık	Özgüllük	Dengelenmiş Doğruluk	F1-Skoru
ADASYN	YSA	0.930	0.959	0.901	0.930	0.931
	Lojistik Reg.	0.877	0.859	0.894	0.876	0.873
	Naive Bayes	0.863	0.852	0.875	0.863	0.861
	Rassal Orman	**0.993**	**1.000**	**0.986**	**0.993**	**0.993**
	DVM	0.956	0.939	0.973	0.956	0.955
	XGBoost	0.990	**1.000**	0.980	0.990	0.990
Borderline-SMOTE	YSA	0.931	0.939	0.922	0.931	0.933
	Lojistik Reg.	0.890	0.892	0.887	0.889	0.892
	Naive Bayes	0.848	0.852	0.845	0.848	0.852
	Rassal Orman	0.986	**1.000**	**0.971**	**0.985**	**0.986**
	DVM	**0.965**	0.959	**0.971**	0.965	0.966
	XGBoost	0.986	**1.000**	**0.971**	**0.985**	**0.986**
RUS	YSA	0.630	0.434	0.826	0.630	0.540
	Lojistik Reg.	0.826	**0.869**	0.782	0.826	0.833
	Naive Bayes	0.739	0.739	0.739	0.739	0.739
	Rassal Orman	**0.934**	**0.869**	**1.000**	**0.934**	**0.930**
	DVM	0.826	0.782	0.869	0.826	0.818
	XGBoost	0.869	0.782	0.956	0.869	0.857
ROSE	YSA	0.802	0.758	**0.851**	0.805	0.802
	Lojistik Reg.	0.825	0.802	**0.851**	0.827	0.829
	Naive Bayes	0.854	**0.901**	0.813	0.857	0.855
	Rassal Orman	**0.866**	0.888	0.846	**0.867**	**0.870**
	DVM	0.825	0.827	0.824	0.825	0.833
	XGBoost	0.825	0.827	0.824	0.825	0.833

Yöntem	Model	Doğruluk	Duyarlılık	Özgüllük	Dengelenmiş Doğruluk	F1-Skoru
SMOTE	YSA	0.917	0.880	0.953	0.917	0.922
	Lojistik Reg.	0.879	0.915	0.845	0.880	0.878
	Naive Bayes	0.890	0.929	0.852	0.890	0.890
	Rassal Orman	**0.986**	**0.971**	**1.000**	**0.985**	**0.986**
	DVM	0.951	0.936	0.966	0.951	0.953
	XGBoost	**0.986**	**0.971**	**1.000**	**0.985**	**0.986**

EK 3. %25 Dengesizlik Durumundan Aşırı Örneklenmiş Veri Seti Üzerindeki Performans Metrikleri

Yöntem	Model	Doğruluk	Duyarlılık	Özgüllük	Dengelenmiş Doğruluk	F1-Skoru
ADASYN	YSA	0.892	0.872	0.912	0.892	0.891
	Lojistik Reg.	0.862	0.885	0.838	0.862	0.856
	Naive Bayes	0.856	0.859	0.852	0.856	0.852
	Rassal Orman	**0.986**	**0.974**	**1.000**	**0.987**	**0.986**
	DVM	0.957	**0.974**	0.939	0.957	0.955
	XGBoost	**0.986**	**0.974**	**1.000**	**0.987**	**0.986**
Borderline-SMOTE	YSA	0.948	0.943	0.953	0.948	0.949
	Lojistik Reg.	0.879	0.915	0.845	0.880	0.878
	Naive Bayes	0.865	0.880	0.852	0.866	0.866
	Rassal Orman	**0.986**	0.971	**1.000**	0.985	0.986
	DVM	0.962	0.964	0.959	0.962	0.962
	XGBoost	**0.989**	**0.978**	**1.000**	**0.989**	**0.990**
RUS	YSA	**0.871**	**0.846**	0.897	**0.871**	0.875
	Lojistik Reg.	0.820	0.743	0.897	0.820	0.833
	Naive Bayes	0.807	0.717	0.897	0.807	0.823
	Rassal Orman	**0.871**	**0.846**	0.897	**0.871**	0.875
	DVM	**0.871**	0.794	**0.948**	**0.871**	**0.880**
	XGBoost	0.846	0.820	0.871	0.846	0.850
ROSE	YSA	0.824	0.911	0.744	0.828	0.815
	Lojistik Reg.	0.808	0.844	0.775	0.809	0.808
	Naive Bayes	0.813	0.822	0.806	0.814	0.818
	Rassal Orman	**0.877**	0.877	**0.877**	**0.877**	**0.882**
	DVM	0.845	**0.922**	0.775	0.848	0.839
	XGBoost	0.829	0.844	0.816	0.830	0.833

Yöntem	Model	Doğruluk	Duyarlılık	Özgüllük	Dengelenmiş Doğruluk	F1-Skoru
SMOTE	YSA	0.866	0.943	0.785	0.864	0.850
	Lojistik Reg.	0.872	0.917	0.825	0.871	0.863
	Naive Bayes	0.850	0.898	0.798	0.848	0.838
	Rassal Orman	**0.996**	**0.993**	**1.000**	**0.996**	**0.996**
	DVM	0.938	0.924	0.953	0.938	0.937
	XGBoost	**0.996**	**0.993**	**1.000**	**0.996**	**0.996**