Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(2) (2022) 92-105
DOI: 10.38016/jista.1069541

Farklı Sınıflandırıcılar ve Yeniden Örnekleme Teknikleri Kullanılarak Kalp Hastalığı Teşhisine Yönelik Karşılaştırmalı Bir Çalışma

Onur Sevli[1]*

[1]Burdur Mehmet Akif Ersoy Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Burdur, Türkiye

onursevli@mehmetakif.edu.tr

Öz

Kalp hastalıkları dünya genelinde yaygın olarak görülmekte ve küresel ölümlerin üçte birlik kısmına neden olmaktadır. Kalp hastalığının semptomlarını ayırt etmedeki zorluk ve çoğu kalp hastasının kriz anına kadar semptomların farkında olmaması, hastalığın tanısını zorlaştırmaktadır. Bir yapay zekâ disiplini olan makine öğrenmesi bilinen verilerden yola çıkarak, yeni vakaların teşhisi konusunda uzmanlar için başarılı karar destek çözümleri sunmaktadır. Bu çalışmada kalp hastalıklarının erken teşhisine yönelik çeşitli makine öğrenmesi teknikleri kullanarak sınıflamalar gerçekleştirilmiştir. Çalışma literatürde yaygın olarak kullanılan UCI kalp hastalığı veri seti üzerinde gerçekleştirilmiştir. Sınıflandırma başarısını arttırmak için, eldeki veri setinin sınıf dengesini sağlamaya yönelik olarak yeniden örnekleme teknikleri kullanılmıştır. Naive Bayes, Karar Ağaçları, Destek Vektör Makinesi, K En yakın Komşu, Lojistik Regresyon, Rastgele Orman, AdaBoost ve CatBoost olmak üzere 8 farklı makine öğrenmesi tekniğinin her biri için örneklemesiz sınıflama yanında fazla örnekleme ve az örnekleme tekniklerinden 8 farklı yöntem kullanılarak toplam 72 sınıflandırma işlemi gerçekleştirilmiştir. Her bir sınıflandırma işleminin sonucu doğruluk, kesinlik, duyarlılık, F1 skoru ve AUC olmak üzere 5 farklı parametre ile raporlanmıştır. En yüksek doğruluk değeri Rastgele Orman ve InstanceHardnessThreshold az örnekleme tekniğinin kullanıldığı sınıflamada %98.46 olarak elde edilmiştir. Elde edilen ölçümlerin literatürde son yıllarda yapılan benzer çalışmalarda ulaşılan sonuçlardan daha yüksek olduğu görülmüştür.

Anahtar kelimeler: Kalp hastalığı teşhisi, Makine öğrenmesi, Yeniden örnekleme

A Comparative Study of Heart Disease Diagnosis using Various Classifiers and Resampling Techniques

Abstract

Heart diseases are common worldwide and cause one-third of global deaths. The difficulty in distinguishing the symptoms of heart disease and the fact that most heart patients are not aware of the symptoms until the moment of crisis make the diagnosis of the disease difficult. Machine learning, an artificial intelligence discipline, provides experts with successful decision support solutions in diagnosing new cases based on known data. In this study, classifications were made using various machine learning techniques for the early diagnosis of heart diseases. The study was carried out on the UCI heart disease dataset, which is widely used in the literature. In order to increase the classification success, resampling techniques were used to ensure the class balance of the dataset. For each of 8 different machine learning techniques, namely Naive Bayes, Decision Trees, Support Vector Machine, K Nearest Neighbor, Logistic Regression, Random Forest, AdaBoost, and CatBoost, in addition to no-sampling classification, 8 different methods from oversampling and undersampling techniques were used to make a total of 72 classification processes were carried out. The result of each classification process is reported with 5 different parameters: accuracy, precision, recall, F1 score, and AUC. The highest accuracy value was obtained as 98.46% in the classification using Random Forest and InstanceHardnessThreshold undersampling technique. It was observed that the measurements obtained were higher than the results obtained in similar studies conducted in the literature in recent years.

Keywords: Heart disease diagnosis, Machine learning, Resampling

* Sorumlu yazar.
E-posta adresi: onursevli@mehmetakif.edu.tr

Alındı : 7 Şubat 2022
Revizyon : 2 Mart 2022
Kabul : 11 Mart 2022

1. Giriş (Introduction)

Kalp sağlığının korunması yaşam için son derece kritik bir öneme sahiptir. Kalp, yaşamın temeli olan kan tedarik sistemini yönetme, kan basıncını sağlama, tek yönlü kan akışını güvence altına alma ve oksijence zengin kanın dokulara, dokulardaki kirli kanın akciğerlere akışını sağlama gibi hayati fonksiyonları yerine getirir. Kaslı bir yapıya sahip olan kalp dakikada ortalama 70 kez kasılıp gevşeyerek, vücuda dakikada yaklaşık 5, saatte 300 ve günde 7200 litre kan pompalar.

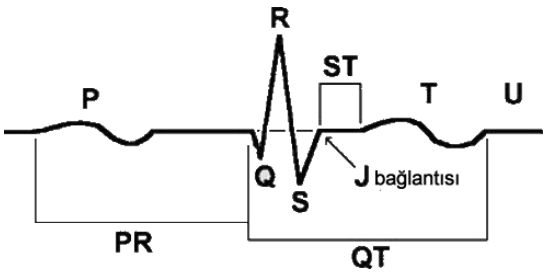
Kalp sağlığı üç temel sistemin birbiri ile uyum içinde çalışması ile mümkündür. Bunlar kardiyovasküler sistem, koroner arterler ve sinir ağıdır. Kardiyovasküler sistem, kanın tek yönde akışını sağlayan valflerle donatılmış kaslı bir pompalama sistemi ve vücudun en ince noktalarına kadar uzanan bir damar ağından oluşur. Kalp odacıkları sürekli kanla dolu olmasına rağmen kalp dokusu bu kanla beslenemez. Koroner arterler kalp yüzeyi boyunca yayılarak kalp kaslarına oksijence zengin kanın taşınmasını ve kalp dokusunun beslenerek düzenli şekilde çalışmasını sağlar. Kalbin kasılıp gevşemesi ise elektriksel sinyaller ve bu sinyalleri yöneten bir sinir ağı ile sağlanır.

Farklı sağlık problemleri bu sitemlerin çalışmasını olumsuz etkileyerek yaşamı tehdit eden kalp rahatsızlıkların oluşmasına sebep olur. Kalp sağlığını riske atan durumların ve belirtilerin erken tespiti hayati önem taşımaktadır. Kalp hastalıkları genetik etkenlere bağlı olabileceği gibi; sağlıksız yaşam, sigara kullanımı, diyabet, obezite, enfeksiyonlar, hipertansiyon gibi farklı sebeplerle kalp damar hastalıkları, kalp ritim bozuklukları, kalp yetmezliği ve kalp krizleri meydana gelebilir. Koroner arterlerdeki tıkanıklık veya daralma sonucu kalp kasını besleyen kan akışında kesilme meydana gelir. Bu durum kalp kasının yeterince oksijenlenmesine engel olarak kalp dokusunda hasar meydana getirebilir. Kalbi besleyen damarların duvarlarında biriken kolesterol gibi maddeler plaklar oluşturur. Zaman içinde çoğalan plaklar damarları tıkayarak kalp krizinin oluşumuna neden olur. Zamanında ve doğru müdahale ile tıkanan damarlar açılmazsa kalpte doku kaybı meydana gelir. Bu durum, kalbin pompalama gücünü azaltır ve kalp yetmezliğine sebep olur. Kalp dokusunun yeterince kanlanamadığı her an, kalıcı hasar oluşma riski artar.

Çağımızın en büyük sorunlarından biri olan stres, diğer tüm rahatsızlıklarda da olduğu gibi kalp hastalıkları ve krizlerinde tetikleyici bir role sahiptir. Ayrıca Kaliforniya merkezli Ailesel Hiperkolesterolemi Vakfı (FH Fountation) araştırmacıları tarafından yürütülen bir araştırmada, COVID-19'un genetik olarak yüksek kolesterol, kalp hastalığı veya her ikisi için risk taşıyanlarda kalp krizi oranını artırdığını ortaya konmuştur (Myers vd., 2021). Dünya Sağlık Örgütü'nün verilerine göre kalp hastalıklarının küresel ölçekte görülme oranı giderek artmakta olup, bulaşıcı olmayan hastalıklara bağlı ölümlerin %46'sı kalp damar hastalıklarından kaynaklanmaktadır. Kardiyovasküler hastalıklar son 15 yılın en yaygın sağlık sorunları olup, küresel ölümlerin tamamının %31'lik bölümü kalp ile ilgili hastalıklardan kaynaklanmaktadır (WHO, 2021). Türkiye İstatistik Kurumu verilerine göre ise son on yıldaki ölümlerin %40 oranı ile büyük çoğunluğunu kalp damar hastalıkları oluşturmaktadır (TUIK, 2021).

Kalp hastalıkları göğüste ağrı ve sıkışma, kısa aktiviteler sonrası nefes darlığı, baş dönmesi ve bayılmalar, kalbin çok sert ya da yavaş atması, kalp damarlarının sertleşmesine bağlı olarak bacak ve kollarda uyuşma, ani soğuk terlemeler, bulantı ve kusma gibi belirtiler verebilir. Bazı hastalarda belirgin şikâyetler görülmeden de kalp krizi oluşabilir. Kalp hastalıklarında erken müdahale son derece önemlidir. Erken müdahale can kaybı riskini ve kalp kasının zarar görme olasılığını azaltır.

Kalp hastalıklarının tanısı uzman hekimler tarafından yapılır. Tanı için fiziksel muayene, hastalığın durumuna bağlı olarak yapılacak tetkikler yanında elektrokardiyografi (EKG) ölçümleri kullanılır. EKG vücuda yapıştırılan elektrotlar ile kalbin elektriksel aktivitesini kaydeder. EKG kayıtları olan elektrokardiyogramlarda kalp atımları P,Q,R,S,T,U dalgalarından oluşan sinyaller şeklinde yorumlanır (Şekil 1). EKG dalgalarındaki değişimler, dalga düzeninin farklılaşması, dalgalar arasındaki sürelerde farklılıklar uzmanlara kalp hastalığı konusunda fikir verir (Kartal ve Köksal, 2020).



Şekil 1. EKG dalgası (ECG wave)

Kalp hastalığına sebep olan pek çok etken olmasından, belirtilerin bazen belirgin olmaması veya diğer hastalıklarla karıştırılabilmesinden dolayı teşhisi komplikedir. Pek çok kalp hastasında göğüs ağrısı ve yorgunluk gibi belirtiler görülürken, %50 gibi büyük bir bölümü kalp krizi geçirene kadar belirtilerin farkına varamamaktadır (Das vd., 2009). Bu durumda hastalığın erken tespiti uzmanlar için de zorlaşmaktadır. Çok sayıda değişken verinin yorumlanarak hastalık teşhisin zamanında ve doğru şekilde yapılabilmesi hayati öneme sahiptir. Bu konuda uzmanların doğru kararlar vermelerine yardımcı olacak destek sistemlere ihtiyaç duyulmaktadır.

Son yıllarda yapay zekâ teknolojisindeki gelişmeler sağlık alanında da alternatif çözümler sunmaktadır. Mevcut verilerden öğrenerek yeni vakalar hakkında sağlıklı tahminler üretmeyi sağlayan makine öğrenmesi, son zamanlarda popülerliği hızla artan bir teknolojidir. Makine öğrenmesinde belirli bir sonuca etki eden

parametreler üzerinden, daha önceki mevcut verilerden yola çıkılarak, yeni vakalar hakkında istikrarlı tahminler üretilebilmektedir.

Literatürde, kalp hastalıklarının teşhisine yönelik farklı makine öğrenmesi teknikleri kullanılarak gerçekleştirilen çalışmalar mevcuttur. Bu çalışmalarda Naive Bayes(NB), Destek Vektör Makinesi (DVM), Karar Ağaçları (KA), K En yakın Komşuluk (KNN), Rastgele Orman (RO) ve Yapay Sinir Ağları (YSA) gibi algoritmalar yaygın olarak kullanılmaktadır. Literatürde son yıllardaki benzer çalışmalar kronolojik sırada aşağıda özetlenmiştir.

Miranda vd. (2016) kalp hastalığı riskini tahminleme için açık uçlu sorulardan oluşan görüşmeler yapmışlar ve elde ettikleri 60589 kayıt ve 38 özellikten oluşan veri seti üzerinde gerçekleştirdikleri üç seviyeli risk sınıflamasında NB yöntemi ile en yüksek %87.98 doğruluk değerine ulaşmışlardır. Wiharto vd. (2016), 14 özellik ve 303 örnekten oluşan UCI veri seti üzerinde SMOTE aşırı örnekleme ve C4.5 tekniği ile gerçekleştirdikleri kalp hastalığı düzeyi sınıflamasında %84.2 AUC değeri elde etmişlerdir. Jabbar vd. (2016) ise RO yöntemi kullanarak gerçekleştirdikleri kalp hastalığı tahminleme çalışmasında %83.70 doğruluğa ulaşmış ve KA yöntemine göre daha yüksek bir başarı sağlandığını ortaya koymuşlardır.

Kim ve Kang (2017), Koreli 4146 adet bireyden elde edilen veriler üzerinde önce birbiriyle ilişkili nitelikleri tespit etmişler ve ardından YSA kullanarak gerçekleştirdikleri kalp hastalığı riski tahminleme çalışmasında %74.9 AUC değerine ulaşmışlardır. Arabasadi vd. (2017) sinir ağları kullanarak 303 örnek ve 54 özellikten oluşan Z-Alizadeh Sani kalp hastalığı veri seti üzerinde gerçekleştirdikleri sınıflama çalışmasında, ağ parametrelerinin genetik algoritmalar kullanılarak optimize edilmesi ile model başarısının arttığını ortaya koymuşlardır. Liu vd. (2017) 202 örnek ve 2 sınıftan oluşan Statlog kalp hastalığı veri seti üzerinde ReliefF and Rough Set (RFRS) özellik seçim yöntemi ve C4.5 sınıflayıcı kullanarak gerçekleştirdikleri çalışmada en yüksek %92.59 doğruluğa ulaşmışlardır.

David ve Belcy (2018), RO, KA ve NB sınıflayıcıları kullanarak UCI veri seti üzerinde gerçekleştirdikleri sınıflamada en yüksek doğruluğu %81 ile RO kullanımında elde etmişlerdir. Haq vd. (2018), aynı veri seti üzerinde farklı özellik seçim teknikleri ve yedi farklı makine öğrenmesi algoritması kullanarak gerçekleştirdikleri sınıflama çalışmasında en yüksek doğruluğu %86 ile DVM kullanarak elde etmişlerdir. Malav ve Kadam (2018), yine aynı veri seti üzerinde ilk olarak K-means uygulayarak elde ettikleri kümeleme sonucunu YSA'ya girdi olarak verip gerçekleştirdikleri sınıflamada %93.52 doğruluğa ulaşmışlardır. Poornima ve Gladis (2018), boyut indirgeme ve YSA kullanarak gerçekleştirdikleri sınıflama çalışmasında %94 doğruluk sağlamışlardır.

Ali vd. (2019b), UCI veri seti üzerinde istatistiksel özellik seçimi ve optimize edilmiş YSA kullanarak gerçekleştirdikleri sınıflamada %93.33 doğruluk elde etmişlerdir. Mohan vd. (2019), aynı veri seti üzerinde lineer model ile hibrit şekilde RO yöntemi kullanarak gerçekleştirdikleri sınıflandırma çalışmasında %88.7 doğruluğa ulaşmışlardır. Aynı veri seti üzerinde Ali vd. (2019a), bir yığın halinde iki ayrı DVM modelini kullanarak, lineer çekirdekli ilk model ile özellik seçimi gerçekleştirmiş, RBF çekirdekli model ile sınıflama gerçekleştirmişler ve %91.11 elde etmişledir. Latha ve Jeeva (2019) UCI veri seti üzerinde kolektif öğrenme algoritmaları ile gerçekleştirdikleri sınıflama çalışmasında zayıf sınıflayıcılara oranla maksimum %7 doğruluk artışı sağlandığını raporlamışlardır.

Mienye vd. (2020); Framingham, Massachusetts sakinlerinden toplanan 4238 adet örnekten oluşan kalp hastalığı veri seti üzerinde YSA kullanarak gerçekleştirdikleri sınıflama çalışmasında %90 doğruluğa ulaşmışlardır. Terrada vd. (2020), Z-Alizadeh Sani kalp hastalığı veri seti üzerinde YSA kullanarak gerçekleştirdikleri sınıflama çalışmasında %94 doğruluk elde etmişlerdir. Tama vd. (2020), UCI veri seti üzerinde kolektif öğrenme yöntemi ile gerçekleştirdikleri sınıflama çalışmasında %85.71 doğruluk elde etmişlerdir. Akalın vd. (2020), aynı veri seti üzerinde KNN, Gaussian Bayes ve RO olmak üzere üç farklı yöntem ile gerçekleştirdikleri sınıflamada %80, %80 ve %82 doğruluğa ulaşmışlardır.

Elhoseny vd. (2021), 13 adet özellik ve 270 örnekten oluşan kalp hastalığı veri seti üzerinde farklı makine öğrenmesi teknikleri kullanarak gerçekleştirdikleri sınıflama çalışmasında en yüksek doğruluk değerlerini %82.5, %81.5 ve %80.8 ile AdaBoost, LogitBoost ve NB yöntemleri ile elde etmişlerdir. Rani vd. (2021), yine aynı veri seti üzerinde SMOTE aşırı örnekleme yöntemi ve NB tekniği ile gerçekleştirdikleri sınıflamada %85.07 doğruluk değerine ulaşmışlardır. Aynı teknik ile aşırı örnekleme kullanmadan elde ettikleri doğruluk ise %84.79'dur. Katarya ve Meena (2021), farklı makine öğrenmesi teknikleri kullanarak UCI veri seti üzerinde gerçekleştirdikleri sınıflama çalışmasında RO yöntemi ile %95.60 doğruluğa ulaşmışlardır. Bharti vd. (2021) ise aynı veri seti üzerinde derin öğrenme yöntemi kullanarak %94.2 sınıflama doğruluğu sağlamışladır. Kavitha vd. (2021), KA ve RO yöntemlerini kapsayan hibrit bir model kullanarak UCI veri seti üzerinde gerçekleştirdikleri sınıflamada en yüksek %88.7 doğruluk elde etmişlerdir. Rajendran ve Vincent (2021), KA, DVM, KNN gibi farklı algoritmaların, RO meta sınıflayıcısı ile bir araya getirilmesi yoluyla gerçekleştirdikleri sınıflamada %87.64 doğruluğa ulaşmışlardır. Asif vd. (2021), UCI veri seti üzerinde kolektif öğrenme yöntemi kullanarak gerçekleştirdikleri sınıflamada en yüksek %92 doğruluk elde etmişlerdir. Maini vd. (2021), Güney Hindistan'da bir hastaneden elde edilen, 14 özellik 501 adet örnekten oluşan veri seti üzerinde RO yöntemi ile gerçekleştirdikleri kalp hastalığı tahminleme çalışmasında %93.8 doğruluğa ulaşmışlardır.

Bu çalışmada literatürde yaygın olarak kullanılan UCI veri seti üzerinde, yeniden örnekleme olmaksızın ve 8 farklı yeniden örnekleme tekniği ile kullanılarak, 8 farklı makine öğrenmesi yöntemi ile bireyin temel özellikleri ve klinik ölçümlere dayanarak, kalp hastalığı durumunun teşhisine yönelik sınıflandırma çalışmaları gerçekleştirilmiştir. Sınıflandırma işlemlerinin sonuçları doğruluk, kesinlik, duyarlılık, F1 skoru ve AUC olmak üzere 5 farklı parametre ile raporlanmış ve elde edilen 360 adet ölçümün sonuçları yorumlanmıştır.

2. Materyal ve Metot (Material and Method)

Bu çalışma, bireylerin kalp hastalığı riskini kişisel özellikler ve klinik ölçümlere bağlı olarak değerlendirmeyi amaçlamakta ve bu doğrultuda UCI kalp hastalığı veri seti üzerinde farklı makine öğrenmesi algoritmaları ile kalp hastalığının erken teşhisine yönelik sınıflandırma çalışmalarını içermektedir. Sınıflandırma için Naive Bayes, Karar Ağaçları, Destek Vektör Makinesi, K En yakın Komşuluk, Lojistik Regresyon, Rastgele Orman, AdaBoost ve CatBoost olmak üzere 8 farklı yöntem kullanılmıştır. Sınıflama çalışmalarındaki ana amaçlardan biri tahmin başarısını yükseltmektir. Bu amaçla, çalışmada kullanılan sekiz farklı sınıflandırıcının varsayılan durumundaki başarıları yanında, her biri için 8 ayrı yeniden örnekleme yöntemi bağımsız olarak uygulanmış ve ölçümleri raporlanmıştır.

2.1. Veri seti (Dataset)

Kalp hastalığının erken teşhisine yönelik bir çözüm sunmayı hedefleyen bu çalışmada, literatürde yaygın olarak kullanılan UCI kalp hastalığı veri seti ile gerçekleştirilmiştir. Kamuya açık olarak paylaşılan veri seti University of California Irvine çevrimiçi veri deposundan elde edilmiştir ("Heart Disease Data Set, UCI Machine Learning Repository," 1988). Macar ve İsveç bilim insanları tarafından derlenen veri seti 13 farklı tanı parametresi ile birlikte bireyin kalp hastalığı durumuna ilişkin bir adet tanı olmak üzere 14 adet özellik içermekte ve 303 örnekten oluşmaktadır. Veri seti içerisinde yer alan özellikler ve açıklamaları Tablo 1'de özet olarak verilmiştir. Veri setinde yer alan sayısal özelliklerin tanımlayıcı istatistikleri ise Tablo 2'de verilmiştir.

Tablo 1. Veri setine ait özellikler (Features of the dataset)

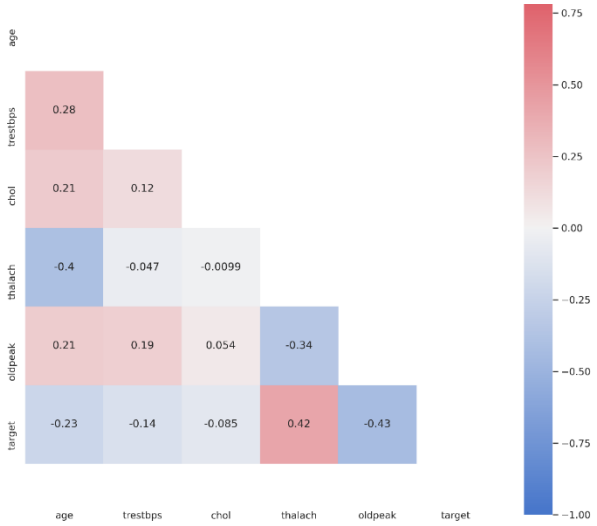
Özellik adı	Türü	Açıklama
age	Sayısal - sürekli	Hastanın yıl olarak yaşı
sex	Kategorik	Hastanın cinsiyeti (0 = kadın, 1 = erkek)
cp (chest pain)	Kategorik	Hastanın yaşadığı göğüs ağrısı türü (0 = tipik anjina, 1 = atipik anjina, 2 = anjinal olmayan ağrı, 3 = asemptomatik)
trestbps (resting blood pressure)	Sayısal - sürekli	Dinlenme durumundaki kan basıncı (mm Hg)
chol (cholestrol)	Sayısal - sürekli	Serumdaki kolesterol değeri (mg/dl)
fbs (fasting blood sugar)	Kategorik	Açlık kan şekeri fbs>120 ml/dl ise 1 (true) değilse 0 (false)
restecg	Kategorik	Dinlenme durumundaki elektrokardiyografik ölçüm (0 = normal, 1 = ST dalga anormalliği, 2 = olası sol ventrikül hipertrofisi)
thalach	Sayısal - sürekli	Ulaşılan maksimum kalp atış sayısı
exang	Kategorik	Egzersiz kaynaklı göğüs ağrısı (0 = yok, 1 = var)
oldpeak	Sayısal - sürekli	Egzersize bağlı dinlenmenin neden olduğu ST depresyonu (EKG'deki ST aralığı için)
slope	Kategorik	Egzersizin tepe noktasında ST segmentinin eğimi (0 = yukarı eğimli, 1 = düz, 2 = aşağı eğimli)
ca	Kategorik	Florosopi ile renklendirilen ana damarların sayısı (0-3)
thal	Kategorik	Talasemi durumu (1 = normal, 2 = sabit kusur, 3 = tersinir kusur)
target	Kategorik	Kalp hastalığı durumu (0 = sağlıklı, 1 = hasta)

Tablo 2. Veri setinin istatistiki karakteristiği (Statistical characteristic of the dataset)

Özellik	Minimum	Maksimum	Ortalama	Standart sapma
age	29	77	54.36	9.08
trestbps	94	200	131.62	17.53
chol	126	564	246.26	51.83
thalach	71	202	149.64	22.90
oldpeak	0	6.2	1.03	1.16

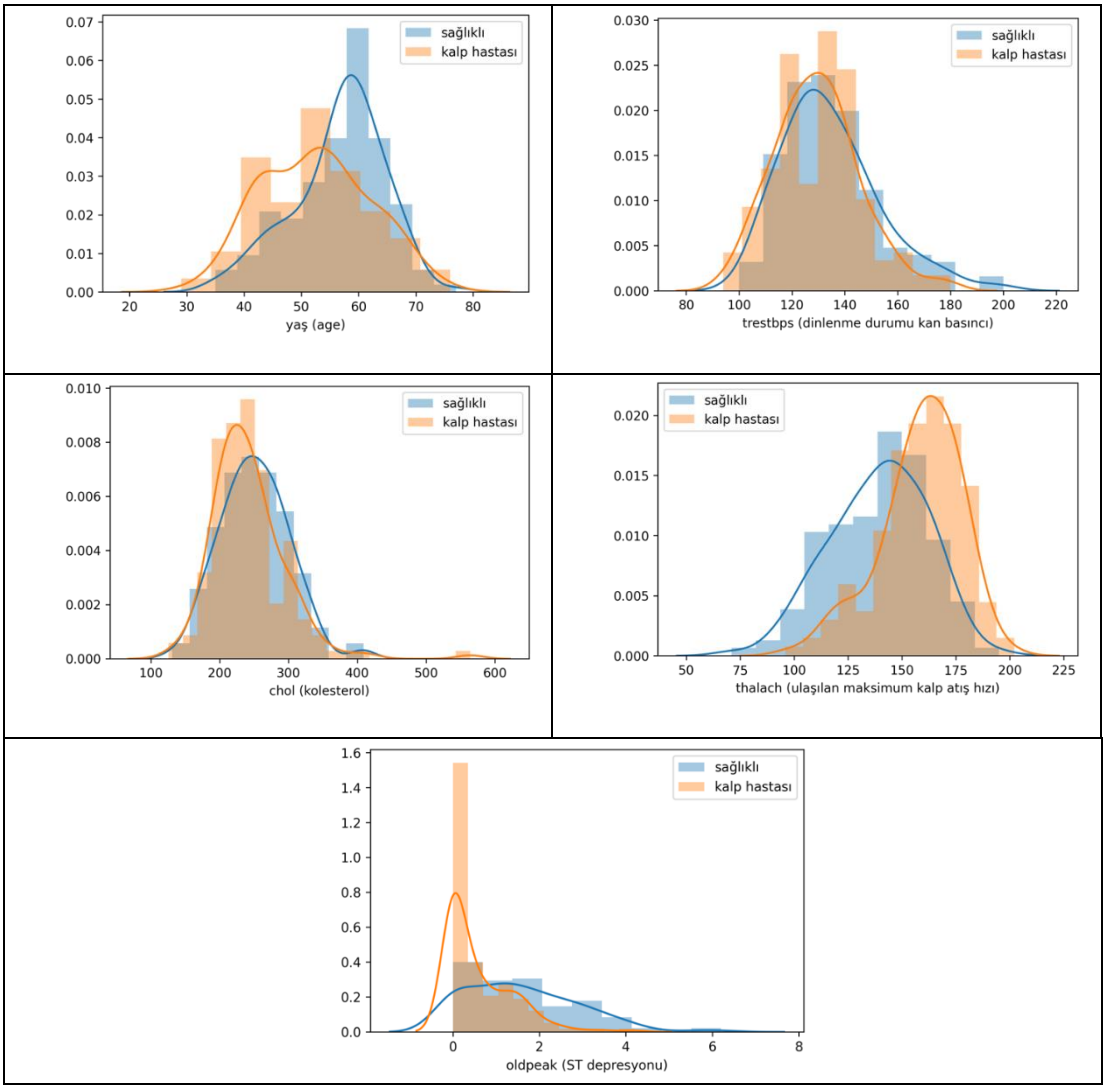
Veri setini oluşturan sayısal özellikler ve hedef değişken arasındaki korelasyonu gösteren matris Şekil 2'de verilmiştir.

Hedef özellik ile arasında 0.5 üzerinden korelasyon olan herhangi bir özellik bulunmadığından, tek başına hedefi tahminleme konusunda baskın değer yoktur. Hedef ile en yüksek korelasyona sahip özellik, ölçülen maksimum kalp atış hızı (thalac) değeridir. Matristeki her bir giriş özelliğinin, hedef değişkene göre dağılımını gösteren grafikler Tablo 3'te yer almaktadır.



Şekil 2. Korelasyon matrisi (Correlation matrix)

Tablo 3. Veri setindeki sayısal özelliklerinin hedef değişkene göre dağılımları (Distribution of numerical features in the dataset according to the target variable)



2.2. Kullanılan sınıflandırma yöntemleri (Classification methods used)

Bu çalışmada literatürde sıklıkla tercih edilen Naive Bayes, Karar Ağaçları (Decision Trees), Destek Vektör Makinesi (Support Vector Machine), K En yakın Komşuluk (K-Nearest Neighbor), Lojistik Regresyon (Logistic Regression), Rastgele Orman (Random Forest), AdaBoost (Adaptive Boosting) ve CatBoost (Categorical Boosting) sınıflandırma algoritmaları kullanılmıştır.

Naive Bayes (NB) algoritması, adını matematikçi Thomas Bayes'den alan bir sınıflandırma algoritmasıdır. NB sınıflandırıcı, olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile sınıfı bilinmeyen verilerin sınıfını tespit etmeyi amaçlar. Algoritma her durumun olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır.

Karar ağaçları (KA), sınıflama ve regresyon problemlerinden sıklıkla kullanılan ve karmaşık veri setleri ile çalışabilen, ağaç tabanlı bir algoritmadır. KA, özellikler ve hedefe bağlı olarak karar düğümleri ve yaprak düğümlerden oluşan bir ağaç yapısı formüle ederek bir sınıflandırma modeli oluşturur.

Destek vektör makinesi (DVM), Vapnik vd. (1997) tarafından literatüre kazandırılan, istatistiksel öğrenme teorisine dayalı, sınıflandırma ve regresyon analizi için kullanılan gözetimli bir makine öğrenmesi tekniğidir. DVM, her biri iki kategoriden birine ait olarak işaretlenmiş eğitim veri setinden öğrenerek, yeni örnekleri bu iki sınıftan birine olasılıklı olmayacak şekilde atayan bir model oluşturur. Veri örneklerinin yer aldığı düzlemde, sınıfları birbirinden ayırmak için, iki sınıfın üyelerinden en uzak mesafede olacak şekilde bir karar sınırının çizilmesi sağlanır. DVM'nin, aşırı uydurma problemi karşısındaki hassasiyetinin düşük olması ve yüksek doğruluk sağlaması kullanım yaygınlığını arttırmaktadır.

K En Yakın Komşuluk (KNN) algoritması, literatüre Fix ve Hodges (1952) tarafından kazandırılan, sınıflama ve regresyonda yaygın olarak kullanılan, parametrik olmayan bir yöntemidir. En temel makine öğrenmesi tekniklerinden biri olan KNN algoritmasındaki K değeri en yakın özellikleri seçmek amacıyla kullanılır (Bilgin, 2021). Sınıfı belirlenmek istenen nokta, K adet en yakın komşusuna bakılarak en yaygın olan sınıfa atanır. Sınıfı tahmin edilecek her bir örnek için, veri setindeki tüm örnekler arasında en yakın komşuluğun aranması nedeniyle veri setinin büyümesi durumunda işlem yükü artar. KNN algoritması mesafeye dayalı olduğundan, eğitim verilerinin normalizasyonu sınıflandırıcının doğruluğunu önemli ölçüde yükseltebilir.

Lojistik Regresyon (LR) bağımlı değişkenin süreksiz olduğu ikili sınıflama problemlerinde kullanılan istatistiksel bir modeldir. Bilgisayar bilimi, pek çok uygulamalı bilim ve gerçek dünya problemlerinde yaygın olarak kullanılmaktadır. Lojistik regresyon ikili bağımlı değişken ile bir dizi bağımsız değişken arasındaki ilişkiyi açıklamaya yönelik tahminleyici bir analizdir. Bu işi yerine getirmek için bir lojistik fonksiyon (logit fonksiyonu) kullanır. Bir olayın gerçekleşme olasılığı Eşitlik 1'deki formül ile ifade edilir. Olayın gerçekleşmeme olasılığı 1-p olmak üzere logit fonksiyonu ise Eşitlik 2'ye göre hesaplanır. Logistik regresyon logit dönüşümünü tahminlemek için bir formülün katsayılarını üretir.

$$p = \frac{e^{a+bx}}{1+e^{a+bx}} \qquad (1)$$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \qquad (2)$$

Rastgele orman (RO), eğitim esnasında çok sayıda karar ağacı oluşturarak, her bir ağacın ürettiği sonuçların modu veya ortalamasını alarak çıktı sınıfı belirleyen bir kolektif öğrenme algoritmasıdır. Ho (1995) tarafından oluşturulan yönteme dayanan RO, daha sonra Breiman (2001) tarafından geliştirilerek literatüre kazandırılmıştır. RO, geleneksel karar ağaçlarında yaygın olan problemlerden biri olan aşırı uydurma (overfitting) sorununa hem veri seti, hem öznitelikleri çok sayıda parçaya bölüp birden çok ağaç üzerinde işleyerek çözüm getirir.

AdaBoost(AB), Freund ve Schapire (1996) tarafından formüle edilen adaptif bir meta algoritmadır. Bireysel öğrenicilerin ve kararlarının birleştirilmesi mantığına dayanan kolektif bir öğrenme yöntemidir. Eğitim sürecinde bireysel öğrenicilerin durumları ağırlıklandırılarak, yapılan güncellemelerle nihai modelin güçlü bir öğrenmeye yakınsaması sağlanır. AB, kaynak tüketiminin etkin ve tahmin hızının yüksek olması nedeni ile kolektif modeller içerisinde yaygın olarak tercih edilir.

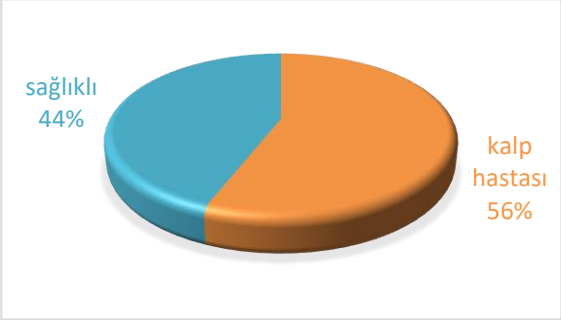
CatBoost(CB), Yandex mühendisleri Dorogush vd. (2018) tarafından formüle edilen açık kaynak kodlu bir algoritmadır. CB, klasik yönteme kıyasla kategorik özellikleri başarılı bir şekilde ele almak için permutasyona dayalı alternatif bir çözüm sunan yeni bir gradyan arttırma algoritmasıdır. Algoritmanın bir diğer avantajı, ağaç yapısını seçerken yaprak değerlerini hesaplamak için aşırı uydurmayı azaltmaya yardımcı olan yeni bir şema kullanmasıdır.

3. Kalp Hastalığı Teşhisine Yönelik Sınıflandırma (Classification for Diagnosis of Heart Disease)

Çalışmada kullanılan UCI kalp hastalığı veri seti 13 özellik ve 1 sınıf değeri ile tanımlanan, 303 örnekten oluşmaktadır. 13 adet girdi özelliğin 5 adedi sayısal, geriye kalan 8 adedi ise kategoriktir. Hedef sınıf değeri ise sağlıklı (0) ve kalp hastası (1) olmak üzere kategorik türdedir. Veri setinde içerisinde 165 adet hasta, 138 hasta olmayan birey kaydı yer almaktadır.

Sınıflandırma işlemi öncesinde veri seti üzerinde tahmin başarısını arttırmaya yönelik önişlemler uygulanmıştır. Kayıp veriler ilgili sütunun ortalama değeri ile doldurulmuştur. Ardından özelliklerin aykırı değerleri (outlier) tespit edilmiştir. Verilerin ilk ve

üçüncü çeyreği arasındaki farkın 1.5 katının alt sınırdan çıkarılması ve üst sınıra eklenmesi ile elde edilen aralık dışındaki değerler aykırı değer olarak kabul edilmiş ve bu değerleri içeren örnekler veri setinden temizlenmiştir. Daha sonra sayısal girdi özelliklerin tümü 0-1 aralığına ölçeklenmiştir. Son durumda veri setinde 159 adet hasta ve 125 adet sağlıklı birey verisi olmak üzere toplam 284 adet örnek yer almaktadır. Son durumda veri setindeki kalp hastası ve sağlıklı bireylerin dağılımlarını gösterir grafik Şekil 3'te yer almaktadır.



Şekil 3. Veri setinin sınıf dağılımı (Class distribution of the dataset)

Veri setindeki hasta ve sağlıklı bireylerin dağılımında kısmi bir dengesizlik söz konusudur. Dengesizlik durumunda, kullanılan sınıflama modeli baskın olan veriyi öğrenmeye yatkınlaşır. Veri setini daha dengeli hale getirerek bu soruna çözüm üretmek için yeniden örnekleme (resampling) tekniklerinden yararlanılır. Yeniden örnekleme, istatistiki yöntemlere dayanarak baskın olan sınıf verilerinin azaltılması ya da verinin istatistiki karakteristiğine uygun şekilde azınlık sınıf verilerinin çoğaltılması ile gerçekleştirilir.

Yeniden örnekleme, az örnekleme (undersampling) ve fazla örnekleme (oversampling) olmak üzere iki farklı yolla uygulanır. Baskın sınıfa ait olan verilerin azaltılması yoluyla dengeyi sağlamaya yönelik yaklaşım az örnekleme olarak adlandırılır. Fazla örnekleme yönteminde ise azınlık sınıfa ait veriler çeşitli tekniklerle arttırılır. Bu iş için, az olan veriyi rastgele olarak seçip kopyalama veya interpolasyon yöntemleri ile sentetik veri üretme yaklaşımları uygulanmaktadır. Bu çalışmada kullanılan veri setindeki dengesizlik büyük oranda olmasa da model başarısı üzerindeki etkisini ortaya koymak için varsayılan sınıflamaya ek olarak 8 farklı yeniden örnekleme tekniği kullanılmıştır. Çalışmada kullanılan fazla örnekleme ve az örnekleme teknikleri Tablo 4'te verilmiştir.

Tablo 4. Kullanılan yeniden örnekleme teknikleri (Resampling techniques used)

Fazla örnekleme	Az örnekleme
SMOTE	AllKNN
KMeansSMOTE	InstanceHardnessThreshold
ADASYN	NeighbourhoodCleaningRule
SVMSMOTE	OneSidedSelection

Tablo 4'te yer alan fazla örnekleme tekniklerinden SMOTE, Chawla vd. tarafından geliştirilen azınlık verilerden sentetik üretim yapan bir yaklaşımdır (Chawla vd., 2002). KMeansSMOTE, SMOTE tekniği üzerine, gürültü oluşumunu azaltmak için K-means kümeleme uygulayan bir tekniktir (Last vd., 2017). Benzer şekilde SVMSMOTE tekniği, SMOTE ile Destek Vektör Makinesini (SVM) birleştirir (Nguyen vd., 2011). ADASYN tekniği SMOTE tekniğine benzeyen ancak verideki sınıf dağılımlarına bağlı olarak değişen sayıda örnek üreten adaptif, sentetik bir fazla örnekleme tekniğidir (He vd., 2008).

Az örnekleme tekniklerinden AllKNN, en yakın komşuluk yöntemine göre karar sınırına yakın örnekleri kaldırmak suretiyle baskın sınıfı azaltma işlemini, komşuları çeşitlendirecek şekilde gerçekleştirir (Tomek, 1976a). InstanceHardnessThreshold veri setindeki örneklerin sınıflandırılma zorluğunun eşik değerlerine bağlı olarak gerçekleştirilen bir az örnekleme tekniğidir (Smith vd., 2014). NeighbourhoodCleaningRule, KNN ve genişletilmiş en yakın komşuluk (ENN) yöntemlerini kullanarak veri setindeki gürültüleri ortadan kaldırır (Laurikkala, 2001). Bir özellik uzayında, farklı sınıflara ait örneklerden birbirine en yakın Öklid mesafesine sahip olanlar, bu prosedürü ortaya koyan Ivan Tomek'e atfen Tomek's Link olarak adlandırılır. TomekLinks az örnekleme tekniği, veri seti içerisindeki Tomek's linkleri kaldırma üzerine kuruludur (Tomek, 1976b). OneSidedSelection tekniği, TomekLinks ve Yoğun En Yakın Komşuluk (CNN) kurallarını birleştiren bir az örnekleme tekniğidir (Kubat vd., 1997).

Veri seti üzerinde farklı sınıflandırıcıların her biri için yukarıda bahsi geçen yeniden örnekleme yöntemleri ayrı ayrı uygulanarak performans ölçümleri gerçekleştirilmiştir.

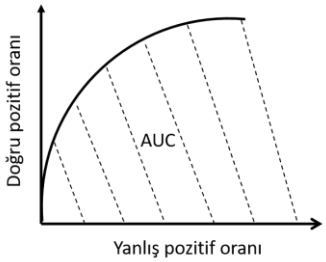
3.1. Performans Metrikleri (Performance Metrics)

Bir sınıflandırıcının başarısını ölçmek için karmaşıklık matrislerinden değerlerden yararlanılır. Karmaşıklık matrisleri sınıflandırma işlemi sonunda hangi sınıfların birbirinden ne düzeyde ayırt edilebildiği ile ilgili detaylı bilgiler sağlar. Gerçekte kalp hastası olan bir kayıt, sınıflandırıcı tarafından da hasta olarak doğru sınıflanmışsa Doğru Pozitif (DP), hastalığı yok şeklinde yanlış sınıflandırılmışsa Yanlış Negatif (YN) olarak adlandırılır. Benzer şekilde gerçekte kalp hastası olmayan kayıt, hastalığı yok şeklinde doğru sınıflandırılırsa Doğru Negatif (DN), hastalığı var şeklinde yanlış sınıflanırsa ise Yanlış Pozitif (YP) olarak adlandırılır. Karmaşıklık matrisinden elde edilen değerler kullanılarak farklı performans metrikleri üretilir. Bu çalışmada kullanılan metrikler Tablo 5'te verilmiştir.

Doğruluk metriği modelin genel başarısını ifade etmek için kullanılır. Doğruluk değeri, doğru sınıflanmış örnek sayısının, tüm örnek sayısına bölümü ile elde edilir. Kesinlik, pozitif olarak tahmin edilen değerlerin gerçekten ne kadarının pozitif olduğunu

gösteren metriktir. Aynı zamanda kesinlik metriği sınıflandırıcının yanlış pozitifleri eleme kabiliyetinin de göstergesidir. Duyarlılık, pozitif olarak tahmin edilmesi gereken değerlerin ne kadarının pozitif olarak tahmin edildiğini gösteren bir metriktir. Duyarlılık, sınıflandırıcının doğru pozitifleri tahmin etmedeki kabiliyetinin ölçütüdür. Kesinlik ve duyarlılık arasındaki dengeyi ifade etmek için F1 skoru kullanılır. F1 skoru, hesaplanan kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. Harmonik ortalama kullanılmasının nedeni uç durumların da göz ardı edilmemesinin gerekliliğidir. F1 skoru [0, 1] aralığında değer alır. Alıcı işlem karakteristiği (ROC) eğrileri, farklı sınıflar için bir olasılık eğrisidir. X ekseninde yanlış pozitif oranı, Y ekseninde ise doğru pozitif oranın yer aldığı bu eğri, kullanılan sınıflandırıcının tahminde ne kadar iyi olduğunu açıklar. Eğrinin altında kalan alan (AUC) [0,1] aralığında değer alır ve model performansının bir özeti kabul edilir. AUC değerinin 1'e yaklaşması veri setindeki sınıfların daha başarılı şekilde ayırt edilebildiğini gösterir.

Tablo 5. Performans metrikleri (Performance metrics)

Metrik	Matematiksel ifadesi
Doğruluk	(DP + DN) / (DP + YP + YN + DN)
Kesinlik	DP / (DP + YP)
Duyarlılık	DP / (DP + YN)
F1 Skoru	2 * kesinlik * duyarlılık / (kesinlik + duyarlılık)
ROC eğrisi ve AUC	

4. Bulgular ve Tartışma (Findings and Discussion)

Çalışmada kullanılan veri seti üzerinde NB, KA, DVM, KNN, LR, RO, AB ve CB sınıflandırıcıları ve her bir sınıflandırıcı için Tablo 4'te yer alan yeniden örnekleme teknikleri uygulanarak performans ölçümleri gerçekleştirilmiştir. Sınıflandırıcının hiperparametreleri ızgara arama yöntemi ile belirlenmiştir. Her bir yöntem için kullanılan parametreler Tablo 6'da verilmiştir.

Tablo 6. Sınıflandırıcıların parametreleri (Parameters of the classifiers)

Sınıflandırıcı	Kullanılan parametre ve değeri
NB	var_smoothing=1e-09
KA	min_samples_split=2, min_samples_leaf=1,
DVM	kernel='linear', C=3
KNN	n_neighbors=3
LR	C=0.1, penalty="l2", max_iter=250
RO	n_estimators =200
AB	n_estimators=100, learning_rate=0.01
CB	verbose=0, n_estimators=100, learning_rate=0.001

Yeniden örnekleme yapılarak ve yapılmadan gerçekleştirilen her bir sınıflama işleminin sonucu, Tablo 5'te verilen metrikler ile raporlanmıştır. Yeniden örnekleme oranı, azınlık ve baskın sınıf örnek sayılarını birbirine yaklaştıracak şekilde fazla örneklemede %30, az örneklemede %20 oranında uygulanmıştır. Fazla örnekleme sonucu veri seti örneklem büyüklüğü 321, az örnekleme sonucu 252'dir. Sınıflama işlemlerinde 10 kat çapraz doğrulama uygulanmıştır. Her bir aşamada %90 eğitim, %10 test seti olmak üzere, çapraz doğrulama işlemlerinden elde edilen sonuçların ortalamaları raporlanmıştır. Örnekleme tekniklerinin sınıflandırıcılar üzerindeki etkisini gösteren ölçümler NB için Tablo 7'de, KA için Tablo 8'de, DVM için Tablo 9'da, KNN için Tablo 10'da, LR için Tablo 11'de, RO için Tablo 12'de, AB için Tablo 13'te ve CB için Tablo 14'te verilmiştir.

Tablo 7. Naive Bayes (NB) sınıflandırıcıya ait ölçümler (Measurements of the Naive Bayes classifier)

Örnekleme	Kullanılan teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0.912281	0.935484	0.906250	0.920635	0.913125
Fazla örnekleme	SMOTE	0.904762	0.931034	0.870968	0.900000	0.904234
	KMeansSMOTE	0.890625	0.875000	0.903226	0.888889	0.891007
	ADASYN	0.901639	0.933333	0.875000	0.903226	0.903017
	SVMSMOTE	0.904762	0.931034	0.870968	0.900000	0.904234
Az örnekleme	AllKNN	0.934783	0.950000	0.904762	0.926829	0.932381
	InstanceHardnessThreshold	**0.960000**	**0.958333**	0.920000	**0.938776**	**0.940000**
	NeighbourhoodCleaningRule	0.934783	0.950000	0.904762	0.926829	0.932381
	OneSidedSelection	0.927273	0.933333	**0.933333**	0.933333	0.926667

Tablo 8. Karar Ağacı (KA) sınıflandırıcıya ait ölçümler (Measurements of the Decision Tree classifier)

Örnekleme	Kullanılan teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0.842105	0.870968	0.843750	0.857143	0.841875
Fazla örnekleme	SMOTE	0.936508	0.965517	0.903226	**0.933333**	**0.935988**
	KMeansSMOTE	**0.937500**	**1.000**	0.870968	0.931034	0.935484

	ADASYN	0.868852	0.900000	0.843750	0.870968	0.870151
	SVMSMOTE	0.888889	0.928571	0.838710	0.881356	0.888105
Az örnekleme	AllKNN	0.891304	0.900000	0.857143	0.878049	0.888571
	InstanceHardnessThreshold	0.92	0.956522	0.88	0.916667	0.92
	NeighbourhoodCleaningRule	0.888889	0.894737	0.850000	0.871795	0.885000
	OneSidedSelection	0.925926	0.931034	**0.931034**	0.931034	0.925517

Tablo 9. Destek Vektör Makinesi (DVM) sınıflandırıcıya ile ait ölçümler (Measurements of the Support Vector Machine classifier)

Örnekleme	Kullanılan teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0.929825	0.937500	0.937500	0.937500	0.928750
Fazla örnekleme	SMOTE	0.936508	**0.965517**	0.903226	0.933333	0.935988
	KMeansSMOTE	0.920635	0.933333	0.903226	0.918033	0.920363
	ADASYN	0.885246	0.878788	0.906250	0.892308	0.884159
	SVMSMOTE	0.906250	0.882353	0.937500	0.909091	0.906250
Az örnekleme	AllKNN	0.913043	0.904762	0.904762	0.904762	0.912381
	InstanceHardnessThreshold	**0.961460**	0.925926	**1.000**	**0.961538**	**0.961460**
	NeighbourhoodCleaningRule	0.934783	0.950000	0.904762	0.926829	0.932381
	OneSidedSelection	0.909091	0.931034	0.900000	0.915254	0.910000

Tablo 10. K En yakın Komşuluk (KNN) sınıflandırıcı ile elde edilen ölçümler (Measurements of the K Nearest Neighbor classifier)

Örnekleme	Kullanılan teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0.912281	0.935484	0.906250	0.920635	0.913125
Fazla örnekleme	SMOTE	0.888889	0.961538	0.806452	0.877193	0.887601
	KMeansSMOTE	0.904762	**1.000**	0.806452	0.892857	0.903226
	ADASYN	0.900000	0.906250	0.906250	0.906250	0.899554
	SVMSMOTE	0.890625	0.878788	0.906250	0.892308	0.890625
Az örnekleme	AllKNN	0.891304	0.833333	**0.952381**	0.888889	0.896190
	InstanceHardnessThreshold	**0.940000**	0.958333	0.920000	**0.938776**	**0.940000**
	NeighbourhoodCleaningRule	0.913043	0.947368	0.857143	0.900000	0.908571
	OneSidedSelection	0.909091	0.962963	0.866667	0.912281	0.913333

Tablo 11. Lojistik Regresyon (LR) sınıflandırıcı ile elde edilen ölçümler (Measurements of the Logistic Regression classifier)

Örnekleme	Kullanılan teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0.894737	0.861111	0.968750	0.911765	0.884375
Fazla örnekleme	SMOTE	0.890625	0.878788	0.906250	0.892308	0.890625
	KMeansSMOTE	0.890625	0.878788	0.906250	0.892308	0.890625
	ADASYN	0.885246	0.878788	0.906250	0.892308	0.884159
	SVMSMOTE	0.888889	**0.961538**	0.806452	0.877193	0.887601
Az örnekleme	AllKNN	0.891304	0.900000	0.857143	0.878049	0.888571
	InstanceHardnessThreshold	**0.961688**	0.925926	**1.000**	**0.961538**	**0.961688**
	NeighbourhoodCleaningRule	0.911111	0.863636	0.950000	0.904762	0.915000
	OneSidedSelection	0.942308	0.900000	**1.000**	0.947368	0.940000

Tablo 12. Rastgele Orman (RO) sınıflandırıcı ile elde edilen ölçümler (Measurements of the Random Forest classifier)

Örnekleme	Kullanılan teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0.894737	0.933333	0.875000	0.852941	0.813125
Fazla örnekleme	SMOTE	0.920635	**1.000**	0.83871	0.912281	0.919355
	KMeansSMOTE	0.921875	**1.000**	0.83871	0.912281	0.919355
	ADASYN	0.885246	0.931034	0.84375	0.885246	0.887392
	SVMSMOTE	0.904762	**1.000**	0.806452	0.892857	0.903226
Az örnekleme	AllKNN	0.956522	**1.000**	0.904762	0.950000	0.952381
	InstanceHardnessThreshold	**0.984600**	0.961538	**1.000**	**0.980392**	**0.984600**
	NeighbourhoodCleaningRule	0.933333	0.904762	0.950000	0.926829	0.935000
	OneSidedSelection	0.927273	0.933333	0.933333	0.933333	0.926667

Tablo 13. AdaBoost (AB) sınıflandırıcı ile elde edilen ölçümler (Measurements of the AdaBoost classifier)

Örnekleme	Kullanılan teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0.929825	**0.937500**	0.937500	0.937500	0.928750
Fazla örnekleme	SMOTE	0.921875	0.909091	0.93750	0.923077	0.921875
	KMeansSMOTE	0.875000	0.900000	0.84375	0.870968	0.87500
	ADASYN	0.918033	0.909091	0.93750	0.923077	0.917026
	SVMSMOTE	0.921875	0.909091	0.93750	0.923077	0.921875
Az örnekleme	AllKNN	0.891304	0.833333	0.952381	0.888889	0.896190
	InstanceHardnessThreshold	0.960784	0.928571	**1.000**	0.962963	**0.960000**

	NeighbourhoodCleaningRule	0.844444	0.809524	0.850000	0.829268	0.845000
	OneSidedSelection	**0.963636**	**0.937500**	**1.000**	**0.967742**	**0.960000**

Tablo 14. CatBoost (CB) sınıflandırıcı ile elde edilen ölçümler (Measurements of the CatBoost classifier)

Örnekleme	Kullanılan teknik	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	AUC
Örneklemesiz	-	0.929825	0.937500	0.937500	0.937500	0.928750
Fazla örnekleme	SMOTE	0.921875	0.909091	0.937500	0.923077	0.921875
	KMeansSMOTE	0.920635	0.933333	0.903226	0.918033	0.920363
	ADASYN	0.934426	0.937500	0.937500	0.937500	0.934267
	SVMSMOTE	0.920635	**0.964286**	0.870968	0.915254	0.919859
Az örnekleme	AllKNN	0.913043	0.869565	0.952381	0.909091	0.916190
	InstanceHardnessThreshold	**0.960784**	0.928571	**1.000**	**0.962963**	**0.960000**
	NeighbourhoodCleaningRule	0.913043	0.904762	0.904762	0.904762	0.912381
	OneSidedSelection	0.945455	0.935484	0.966667	0.950820	0.943333

Elde edilen ölçümler değerlendirildiğinde, yeniden örneklemenin bazı durumlarda belirli sınıflandırıcıların başarısını örnekleme olmayan duruma göre düşürdüğü görülse de genel anlamda yeniden örnekleme tekniklerinin sınıflama başarısını arttırdığı gözlemlenmiştir. Performans metrikleri açısından yeniden örnekleme teknikleri kullanılarak %10'a varan performans artışı sağlandığı görülmüştür. Yeniden örnekleme ile elde edilen bu artış literatürdeki benzer çalışmaları destekler nitelikte olup, bu çalışmadaki artışın emsallerinden yüksek olduğu sonucuna varılmıştır. Az ve fazla örnekleme tekniklerinin başarı durumları değişkenlik göstermekle birlikte, bu çalışmada az örneklemenin daha başarılı sonuç verdiği ve InstanceHardnessThreshold tekniğinin genel anlamda daha yüksek başarı artışı sağladığı görülmüştür. Kullanılan tüm sınıflandırıcılar ve örnekleme yöntemlerinin en iyi sonuçları özet olarak Tablo 15'te verilmiştir.

Tablo 15. Sınıflandırıcıların en iyi sonuçları (Best results of the classifiers)

Sınıflandırıcı	Örnekleme Yöntemi	En İyi Doğruluk	En İyi Kesinlik	En İyi Duyarlılık	En İyi F1 Skoru	En İyi AUC
Naive Bayes	Örnekleme olmadan	0.912281	0.935484	0.906250	0.920635	0.913125
	Fazla Örnekleme	0.904762	0.933333	0.903226	0.903226	0.904234
	Az Örnekleme	**0.960000**	**0.958333**	**0.933333**	**0.938776**	**0.940000**
Karar Ağaçları	Örnekleme olmadan	0.842105	0.870968	0.843750	0.857143	0.841875
	Fazla Örnekleme	**0.937500**	**1.000**	0.903226	**0.933333**	**0.935988**
	Az Örnekleme	0.925926	0.956522	**0.931034**	0.931034	0.925517
Destek Vektör Makinesi	Örnekleme olmadan	0.929825	0.937500	0.937500	0.937500	0.928750
	Fazla Örnekleme	0.936508	**0.965517**	0.937500	0.933333	0.935988
	Az Örnekleme	**0.961460**	0.950000	**1.000**	**0.961538**	**0.961460**
K En yakın Komşuluk	Örnekleme olmadan	0.912281	0.935484	0.906250	0.920635	0.913125
	Fazla Örnekleme	0.904762	**1.000**	0.906250	0.906250	0.903226
	Az Örnekleme	**0.940000**	0.962963	**0.952381**	**0.938776**	**0.940000**
Lojistik Regresyon	Örnekleme olmadan	0.894737	0.861111	0.968750	0.911765	0.884375
	Fazla Örnekleme	0.890625	**0.961538**	0.906250	0.892308	0.890625
	Az Örnekleme	**0.961688**	0.925926	**1.000**	**0.961538**	**0.961688**
Rastgele Orman	Örnekleme olmadan	0.894737	0.933333	0.875000	0.852941	0.813125
	Fazla Örnekleme	0.921875	**1.000**	0.83871	0.912281	0.919355
	Az Örnekleme	**0.984600**	**1.000**	**1.000**	**0.980392**	**0.984600**
AdaBoost	Örnekleme olmadan	0.929825	**0.937500**	0.937500	0.937500	0.928750
	Fazla Örnekleme	0.921875	0.909091	0.93750	0.923077	0.921875
	Az Örnekleme	**0.963636**	**0.937500**	**1.000**	**0.967742**	**0.960000**
CatBoost	Örnekleme olmadan	0.929825	0.937500	0.937500	0.937500	0.928750
	Fazla Örnekleme	0.934426	**0.964286**	0.937500	0.937500	0.934267
	Az Örnekleme	**0.960784**	0.935484	**1.000**	**0.962963**	**0.960000**

Tablo 15'teki sonuçlar incelendiğinde, tüm parametreler açısından en yüksek ölçümlerin Rastgele Orman sınıflandırıcı ile elde edildiği görülmektedir. Bu sınıflandırıcı ile ulaşılan en yüksek doğruluk değeri ise InstanceHardnessThreshold az örnekleme tekniğinin kullanıldığı durumda elde edilmiştir. Rastgele Orman ile tüm sınıflamalarda elde edilen en iyi sonuçlar doğruluk için %98.46, kesinlik için %100, duyarlılık için %100, F1 Skoru için %98.03 ve AUC için %98.46 olmuştur. Rastgele Orman modelinden sonra, en yüksek başarı AdaBoost ile elde edilmiş ve bu sınıflandırıcı ulaşılan en yüksek %96.36 doğruluk değeri yine bir az örnekleme tekniği olan OneSidedSelection ile sağlanmıştır. Yeniden örnekleme ile birlikte yapılan sınıflandırmalar içerisinde elde edilen en iyi doğruluk değerlerinin en düşüğü %93.75 ile Karar Ağaçları ile elde edilmiştir. Tüm sınıflandırıcılar için en iyi kesinlik değerleri %93.75 ile %100 arasında değişim göstermektedir. Elde edilen en iyi duyarlılık değerlerinin en düşüğü %93.10 olup geneli ise %100 seviyesindedir. Bu durum tüm sınıflandırıcıların doğru pozitifleri tespit ve yanlışları eleme kabiliyetlerinin yüksek olduğunu göstermektedir.

Bu çalışmada elde edilen en iyi sonuçlar ile literatürde son dönemdeki benzer çalışmaların sonuçları karşılaştırılmalı olarak Tablo 16'da verilmiştir.

Tablo 16. Elde edilen bulguların literatürdeki çalışmalar ile karşılaştırılması (Comparison of the obtained findings with the studies in the literatüre)

Referans	Veri seti	Kullanılan Yöntem	En iyi doğruluk (%)
Liu vd., 2017	Statlog	C4.5	92.59
David ve Belcy, 2018	UCI	RO	81
Haq vd., 2018	UCI	DVM	86
Malav ve Kadam, 2018	UCI	K-Means & YSA	93.52
Poornima ve Gladis, 2018	UCI	YSA	94
Ali vd., 2019b	UCI	YSA	93.33
Mohan vd., 2019	UCI	Hibrit RO	88.7
Ali vd., 2019a	UCI	Yığınlanmış DVM	91.11
Mienye vd., 2020	Framingham	YSA	90
Terrada vd., 2020	Z-Alizadeh Sani	YSA	94
Tama vd., 2020	UCI	Kolektif öğrenme	85.71
Akalın vd., 2020	UCI	RO	82
Elhoseny vd., 2021	UCI	AdaBoost	82.5
Rani vd., 2021	UCI	SMOTE & NB	85.07
Katarya ve Meena, 2021	UCI	RO	95.60
Bharti vd., 2021	UCI	Derin öğrenme	94.2
Kavitha vd., 2021	UCI	Hibrit model	88.7
Asif vd., 2021	UCI	Kolektif öğrenme	92
Maini vd., 2021	14 özellik, 501 örnek	RO	93.8
		Literatür ortalaması	*89.67*
Bu çalışma	UCI	Önerilen NB	96
		Önerilen KA	93.75
		Önerilen DVM	96.15
		Önerilen KNN	94
		Önerilen LR	96.17
		Önerilen RO	**98.46**
		Önerilen AB	96.36
		Önerilen CB	96.08
		Önerilen yöntemlerin ortalaması	*95.87*

Bu çalışmada, uygulanan yeniden örnekleme yöntemleri ile birlikte kullanılan sınıflandırıcıların literatürdeki benzer çalışmaların genelinden daha yüksek başarı gösterdiği görülmektedir. Çalışmada kullanılan Rastgele Orman modelinin az örnekleme tekniği ile birlikte, %98.46 doğrulukla Tablo 16'da yer alan son beş yıldaki benzer çalışmaların tümünden daha yüksek doğruluk sağladığı görülmektedir. Tabloda yer alan benzer 19 çalışmada elde edilen en yüksek doğruluk değeri %95.60, en düşük doğruluk ise %81'dir. Literatürdeki çalışmalarda ulaşılan ortalama doğruluk değeri ise %89.67'dir. Bu çalışmada elde edilen en düşük doğruluk değeri %93.75, önerilen sekiz modelin ortalama doğruluğu ise %95.87'dir. Bu çalışmada uygulanan modeller ile elde edilen doğruluk değerlerinin geneli, literatürdeki diğer çalışmalarda elde edilen doğruluk değerlerinden daha yüksektir.

Literatürdeki çalışmalarda en yüksek doğruluğun Rastgele Orman modeli ve benzer kolektif öğrenme yöntemleri ile elde edildiği görülmektedir. Bu çalışmada da en yüksek doğruluk değeri Rastgele Orman ve bunu takiben AdaBoost ve CatBoost kolektif öğrenme yöntemleri ile birlikte az örnekleme yaklaşımı kullanılarak sağlanmıştır. Az örnekleme tekniklerinden InstanceHardnessThreshold yönteminin diğer yeniden örnekleme yöntemlerine nazaran daha başarılı olduğu görülmektedir. Kolektif öğrenme algoritmaları aşırı uydurma (overfitting) problemlerine karşı daha az

hassastır. Yeniden örnekleme yoluyla veri setinin dengelenmesi de belirli bir sınıfı aşırı öğrenmeye olan yatkınlığı azaltmaktadır. Bu anlamda yeniden örneklemenin kolektif öğrenme ile birleşiminin aşırı uydurma probleminin çözümünde daha güçlü bir etki oluşturduğu ve bu yolla modelin genellenebilir başarısını arttırdığı elde edilen sonuçlardan yola çıkılarak söylenebilir.

Bunun yanında çalışmada kullanılan veri seti dengesiz yapıda olmasına rağmen sınıf dağılımları arasındaki farklılık yüksek oranda değildir. Normal seviyede dengesiz kabul edilebilecek bir veri seti ile gerçekleştirilen bu çalışmada elde edilen bulgular, aşırı dengesiz veri setlerinde farklılık gösterebilir. Çalışmada kullanılan veri setinin dengesizliğinin yüksek olmaması çalışmanın bir sınırlılığı olarak belirtilebilir.

5. Sonuçlar (Conclusions)

Bu çalışmada dünya genelinde yaygın görülen sağlık sorunlarından biri olan ve başlıca ölüm nedenleri içerisinde yer alan kalp hastalığının erken teşhisine yönelik farklı makine öğrenmesi algoritmaları kullanılarak alternatif bir çözüm önerilmiştir. Kalp hastalığı teşhisinin komplike olması ve hastaların pek çoğunun kriz anına kadar belirtilerin farkına varamaması, durumu uzmanlar için de zorlaştırabilmektedir. Bu anlamda hastaya ait temel bilgiler ve klinik ölçümlerden yola çıkarak 13 farklı tanı parametresi ile kalp hastalığı durumunu tahminlemeye yönelik 303 örnekten oluşan UCI veri seti üzerinde çalışılmıştır. 8 farklı makine öğrenmesi algoritması ve veri setini dengeleyerek tahmin doğruluğunu arttırmaya yönelik 8 farklı yeniden örnekleme yöntemi kullanılarak kalp hastalığı tahminlemesi için ölçümler gerçekleştirilmiştir. 72 ayrı sınıflama ve doğruluk, kesinlik, duyarlılık, F1 skoru ve AUC olmak üzere 5 ayrı parametre ile karakterize edilen 360 ölçümün sonuçlarına göre en yüksek doğruluk %98.46 olarak InstanceHardnessThreshold az örnekleme tekniği ile birlikte Rastgele Orman sınıflandırıcısının kullanıldığı durumda elde edilmiştir. Çalışmada önerilen yöntemlerin tümü literatürdeki, çoğunluğu bu çalışma ile aynı veri setini kullanan, son beş yıldaki benzer çalışmaların genelinden daha yüksek başarı göstermiştir. Bu çalışmada ulaşılan en yüksek doğruluk değeri ise sözü geçen çalışmaların tümünden daha yüksektir. Veriden öğrenen makine öğrenmesi algoritmalarının başarılarını belirleyen temel etken üzerinde çalışılan veri setidir. Veri setindeki sınıf dengesizlikleri modelin başarısını, sonuçların genellenebilirliğini olumsuz etkiler. Eldeki veri setinin uygun örnekleme teknikleri kullanılarak dengeli hale getirilmesi kullanılan modelin daha başarılı sınıflamalar yapmasına olanak tanımaktadır. Bu çalışmada elde edilen sonuçlar yeninden örnekleme tekniklerinin performansları hakkında bir görüş oluşturmuştur. Ayrıca farklı makine öğrenmesi modellerinin başarılarını karşılaştırma olanağı tanımıştır. Bu çalışma ile kullanılan kısmi dengesiz veri seti üzerinde yapılan sınıflamalarda yeniden örneklemenin %10'a varan başarı artışı sağladığı görülmüştür.

Kaynaklar (References)

Akalın, B., Veranyurt, Ü., Veranyurt, O., 2020. Classification of individuals at risk of heart disease using machine learning. Cumhuriyet Medical Journal 42, 283–289.

Ali, L., Niamat, A., Khan, J.A., Golilarz, N.A., Xingzhong, X., Noor, A., Nour, R., Bukhari, S.A.C., 2019a. An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access 7, 54007–54014.

Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., Khan, J.A., 2019b. An Automated Diagnostic System for Heart Disease Prediction Based on x^2 Statistical Model and Optimally Configured Deep Neural Network. IEEE Access 7, 34938–34945. https://doi.org/10.1109/ACCESS.2019.2904800

Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., Yarifard, A.A., 2017. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer Methods and Programs in Biomedicine 141, 19–26. https://doi.org/10.1016/j.cmpb.2017.01.004

Asif, S., Wenhui, Y., Tao, Y., Jinhai, S., Jin, H., 2021. An Ensemble Machine Learning Method for the Prediction of Heart Disease, in: 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE, pp. 98–103.

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., Singh, P., 2021. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Computational Intelligence and Neuroscience 2021, 8387680. https://doi.org/10.1155/2021/8387680

Bilgin, G., 2021. Makine öğrenmesi algoritmaları kullanarak erken dönemde diyabet hastalığı riskinin araştırılması. Journal of Intelligent Systems: Theory and Applications, 4(1), 55-64.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

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

Das, R., Turkoglu, I., Sengur, A., 2009. Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications 36, 7675–7680. https://doi.org/10.1016/j.eswa.2008.09.013

David, H., Belcy, S.A., 2018. HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES. ICTACT Journal on Soft Computing 9.

Dorogush, A.V., Ershov, V., Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. CoRR abs/1810.11363.

Elhoseny, M., Mohammed, M.A., Mostafa, S.A., Abdulkareem, K.H., Maashi, Mashael S., Garcia-Zapirain, B., Mutlag, A.A., Maashi, Marwah Suliman, 2021. A new multi-agent feature wrapper machine learning approach for heart disease diagnosis. Comput. Mater. Contin 67, 51–71.

Fix, E., Hodges Jr, J.L., 1952. Discriminatory analysis-nonparametric discrimination: Small sample performance. California Univ Berkeley.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm, in: Icml. Citeseer, pp. 148–156.

Haq, A.U., Li, J.P., Memon, M.H., Nazir, S., Sun, R., 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems 2018.

He, H., Bai, Y., Garcia, E., Li, S., 2008. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, in: Proceedings of the International Joint Conference on Neural Networks. pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Heart Disease Data Set, UCI Machine Learning Repository [WWW Document], 1988. URL https://archive.ics.uci.edu/ml/datasets/Heart+Disease (erişim tarihi: 4.8.21).

Ho, T.K., 1995. Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE, pp. 278–282.

Jabbar, M.A., Deekshatulu, B.L., Chandra, P., 2016. Prediction of Heart Disease Using Random Forest and Feature Subset Selection, in: Snášel, V., Abraham, A., Krömer, P., Pant, M., Muda, A.K. (Eds.), Innovations in Bio-Inspired Computing and Applications. Springer International Publishing, Cham, pp. 187–196.

Kartal, Mutlu, Köksal, Özlem, 2020. Akut Koroner Sendromlarda EKG.

Katarya, R., Meena, S.K., 2021. Machine learning techniques for heart disease prediction: a comparative study and analysis. Health and Technology 11, 87–97.

Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R., Suraj, R.S., 2021. Heart Disease Prediction using Hybrid machine Learning Model, in: 2021 6th International Conference on Inventive Computation Technologies (ICICT). pp. 1329–1333. https://doi.org/10.1109/ICICT50816.2021.9358597

Kim, J.K., Kang, S., 2017. Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis. Journal of Healthcare Engineering 2017, 2780501. https://doi.org/10.1155/2017/2780501

Kubat, M., Matwin, S., others, 1997. Addressing the curse of imbalanced training sets: one-sided selection, in: Icml. Citeseer, pp. 179–186.

Last, F., Douzas, G., Bacao, F., 2017. Oversampling for imbalanced learning based on k-means and smote. arXiv preprint arXiv:1711.00837.

Latha, C.B.C., Jeeva, S.C., 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked 16, 100203. https://doi.org/10.1016/j.imu.2019.100203

Laurikkala, J., 2001. Improving Identification of Difficult Small Classes by Balancing Class Distribution, in: Quaglini, S., Barahona, P., Andreassen, S. (Eds.), Artificial Intelligence in Medicine. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 63–66.

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Qiugen, Wang, Qian, 2017. A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Computational and Mathematical Methods in Medicine 2017, 8272091. https://doi.org/10.1155/2017/8272091

Maini, E., Venkateswarlu, B., Maini, B., Marwaha, D., 2021. Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India. Medical Journal Armed Forces India. https://doi.org/10.1016/j.mjafi.2020.10.013

Malav, A., Kadam, K., 2018. A hybrid approach for heart disease prediction using artificial neural network and K-means. International Journal of Pure and Applied Mathematics 118, 103–10.

Mienye, I.D., Sun, Y., Wang, Z., 2020. Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. Informatics in Medicine Unlocked 18, 100307.

Miranda, E., Irwansyah, E., Amelga, A.Y., Maribondang, M.M., Salim, M., 2016. Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. Healthcare informatics research 22, 196–205.

Mohan, S., Thirumalai, C., Srivastava, G., 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 7, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Myers, K.D., Wilemon, K., McGowan, M.P., Howard, W., Staszak, D., Rader, D.J., 2021. COVID-19 associated risks of myocardial infarction in persons with familial hypercholesterolemia with or without ASCVD. American Journal of Preventive Cardiology 7, 100197. https://doi.org/10.1016/j.ajpc.2021.100197

Nguyen, H., Cooper, E., Kamei, K., 2011. Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms 3, 4–21. https://doi.org/10.1504/IJKESDP.2011.039875

Poornima, V., Gladis, D., 2018. A novel approach for diagnosing heart disease with hybrid classifier. Biomed Res 29, 2274–2280.

Rajendran, N.A., Vincent, D.R., 2021. Heart Disease Prediction System using Ensemble of Machine Learning Algorithms. Recent Patents on Engineering 15, 130–139.

Rani, P., Kumar, R., Ahmed, N.M.S., Jain, A., 2021. A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments 1–13.

Smith, M.R., Martinez, T., Giraud-Carrier, C., 2014. An instance level analysis of data complexity. Machine Learning 95, 225–256. https://doi.org/10.1007/s10994-013-5422-z

Tama, B.A., Im, S., Lee, S., 2020. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. BioMed Research International 2020, 9816142. https://doi.org/10.1155/2020/9816142

Terrada, O., Hamida, S., Cherradi, B., Raihani, A., Bouattane, O., 2020. Supervised machine learning based medical diagnosis support system for prediction of patients with heart disease. Advances in Science, Technology and Engineering Systems Journal 5, 269–277.

Tomek, I., 1976a. An Experiment with the Edited Nearest-Neighbor Rule. IEEE Transactions on Systems, Man, and Cybernetics SMC-6, 448–452. https://doi.org/10.1109/TSMC.1976.4309523

Tomek, I., 1976b. Two modifications of CNN. IEEE Trans. Syst. Man Cybern. 6, 769–772.

TUIK (Türkiye İstatistik Kurumu), 2021.Ölüm Nedeni İstatistikleri. URL http://www.tuik.gov.tr/PreHaberBultenleri.do?id=27620 (erişim tarihi: 5.18.21).

Vapnik, V., Golowich, S.E., Smola, A., others, 1997. Support vector method for function approximation, regression estimation, and signal processing. Advances in neural information processing systems 281–287.

Wiharto, W., Kusnanto, H., Herianto, H., 2016. Interpretation of clinical data based on C4. 5 algorithm for the diagnosis of coronary heart disease. Healthcare informatics research 22, 186–195.

WHO (World Health Organization), 2021. Global status report on noncommunicable diseases. URL https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (erişim tarihi: 6.21.21).