T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KALP YETMEZLİĞİ HASTA VERİLERİNİN FARKLI SINIFLANDIRMA YÖNTEMLERİYLE ANALİZİ

Şevval Tuğçe BADİK

YÜKSEK LİSANS TEZİ
Matematik Anabilim Dalı
Matematik Programı

Danışman
Dr. Öğr. Üyesi Mutlu AKAR

Haziran, 2022

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KALP YETMEZLİĞİ HASTA VERİLERİNİN FARKLI SINIFLANDIRMA YÖNTEMLERİYLE ANALİZİ

Şevval Tuğçe BADİK tarafından hazırlanan tez çalışması 29.06.2022 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Anabilim Dalı, Matematik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Dr. Öğr. Üyesi Mutlu AKAR

Yıldız Teknik Üniversitesi

Danışman

Jüri Üyeleri

Dr. Öğr. Üyesi Mutlu AKAR, Danışman

Yıldız Teknik Üniversitesi _____

Doç. Dr. Özgür YILDIRIM, Üye

Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Fatih TEMİZ, Üye

Üsküdar Üniversitesi _____

Danışmanım Dr. Öğr. Üyesi Mutlu AKAR sorumluluğunda tarafımca hazırlanan Kalp Yetmezliği Hasta Verilerinin Farklı Sınıflandırma Yöntemleriyle Analizi başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Şevval Tuğçe BADİK

İmza

Bu çalışma, Yıldız Teknik Üniversitesi Bilimsel Araştırma Proje Koordinatörlüğü'nün FYL-2021-4416 numaralı projesi ile desteklenmiştir.

Aileme

TEŞEKKÜR

Yüksek lisans çalışmam süresince bilgi ve tecrübeleriyle destek olan tez danışmanım Dr. Öğr. Üyesi Mutlu Akar'a teşekkürlerimi sunarım. Lisans ve yüksek lisans eğitimimde katkı sağlayan tüm hocalarıma teşekkür ederim. Bu süreçte yanımda olan başta Halit Kandemir ve Sümeyye Çelebi olmak üzere tüm arkadaşlarıma ve eğitim hayatım boyunca maddi ve manevi her an yanımda olan beni hiçbir zaman yalnız bırakmayan sevgili anneme, babama ve biricik kardeşim Ecem'e teşekkür ederim.

Şevval Tuğçe BADİK

İÇİNDEKİLER

KISALTMA LİSTESİ ..... **vii**
ŞEKİL LİSTESİ ..... **viii**
TABLO LİSTESİ ..... **ix**
ÖZET ..... **x**
ABSTRACT ..... **xii**
1 GİRİŞ **1**
1.1 Literatür Özeti 1
1.2 Tezin Amacı 1
1.3 Hipotez 3
2 MAKİNE ÖĞRENMESİ **4**
2.1 Veri Ön İşleme 4
2.2 Öğrenme Türleri 5
2.2.1 Denetimli Öğrenme 5
2.2.2 Denetimsiz Öğrenme 5
2.2.3 Yarı Denetimli Öğrenme 5
2.3 Makine Öğrenmesi Modelleri 6
2.3.1 Regresyon Analizi 7
2.3.2 Sınıflandırma Yöntemleri 9
2.4 Başarı Ölçme Parametreleri 16
3 UYGULAMA VE ANALİZ **19**
3.1 Veri Seti 19
3.2 Sınıflandırma 21
3.3 Sonuçlar 25
4 SONUÇ VE ÖNERİLER **31**
KAYNAKÇA **32**
TEZDEN ÜRETİLMİŞ YAYINLAR **37**

KISALTMA LİSTESİ

ACC	Doğruluk Oranı
BP	Kan Basıncı
CPK	Kreatinin Fosfokinaz
FN	Yanlış Negatif Değer
FP	Yanlış Pozitif Değer
HFpEF	Korunmuş Ejeksiyon Fraksiyonu
HFrEF	Azaltılmış Ejeksiyon Fraksiyonu
KNN	K-En Yakın Komşuluk
KY	Kalp Yetmezliği
LVEF	Sol Ventrikül Ejeksiyon Fraksiyon
MCC	Matthews Korelasyon Katsayısı
NYHA	New York Heart Association
PCA	Temel Bileşen Analizi
PCC	Pearson Korelasyon Katsayısı
RF	Rastgele Orman
ROC AUC	ROC Eğrisi ve Eğrinin Altında Kalan Alan
PR AUC	PR Eğrisi ve Eğrinin Altında Kalan Alan
SE	Duyarlılık
SP	Özgüllük
SVM	Destek vektör makinesi
TN	Gerçek Negatif Değer
TP	Gerçek Pozitif Değer
YSA	Yapay Sinir Ağı

ŞEKİL LİSTESİ

Şekil 2.1 Veri ön işleme yöntemleri [12] 4
Şekil 2.2 Yanlılık-varyans değiş tokuşu [13] 7
Şekil 2.3 KNN ile sınıflandırma [14] 10
Şekil 2.4 SVM ile sınıflandırma [15] 11
Şekil 2.5 Kernel Trick ile boyut artırma [16] 12
Şekil 2.6 Yapay sinir hücresi [20] 14
Şekil 3.1 Standard Scaler sonrası sürekli değişkenlerin dağılımı……...22
Şekil 3.2 Bileşenlerin önem düzeyi 24
Şekil 3.3 Tüm değişkenlerin önem düzeyi 24
Şekil 3.4 Modellerin MCC değerleri 26
Şekil 3.5 Modellerin PCA uygulanmadan MCC değerleri 29

TABLO LİSTESİ

Tablo 2.1 Karışıklık matrisi..... 16
Tablo 3.1 Kategorik değişkenlerin hastalar üzerinde dağılımı..... 20
Tablo 3.2 Sürekli değişkenlerin ortalamasının hastalar üzerinde dağılımı 21
Tablo 3.3 Değişkenlerin ölüm değişkenine göre PCC hesaplaması 23
Tablo 3.4 Modele PCA uygulandıktan sonra performans ölçümü..... 25
Tablo 3.5 Modele PCA uygulanmadan performans ölçümü..... 28
Tablo 3.6 İki çalışma için en başarılı iki yöntemin karşılaştırılması..... 29

ÖZET

Kalp Yetmezliği Hasta Verilerinin Farklı Sınıflandırma Yöntemleriyle Analizi

Şevval Tuğçe BADİK

Matematik Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Dr. Öğr. Üyesi Mutlu AKAR

Kalp yetmezliği dünya genelinde ölüm oranında büyük ölçüde büyümeye sebep olan bir kardiyovasküler hastalıktır. Bu yazının amacı, çeşitli makine öğrenmesi yöntemlerini kullanarak hastaların çeşitli özelliklerine, hastalıklarına ve yaşam tarzlarına bakarak hayatta kalma olasılıklarını en başarılı şekilde tahmin etmektir.

Kullandığımız veri setinde 2015 yılında sol ventrikül sistolik disfonksiyonu olan 299 hasta NYHA sınıf III ve IV olarak sınıflandırılmıştır. Veri seti üzerinde çeşitli makine öğrenimi sınıflandırma yöntemlerini kullanarak hastaların hayatta kalma olasılığını tahmin etmeye çalışılmıştır. Bu çalışmada iki versiyon bulunmaktadır. Çalışmanın ilk versiyonunda, veri setinin boyutunu küçültmek için PCA kullanılmıştır. Makine öğrenimi algoritmalarının başarısı çeşitli performans ölçme metriklerine göre değerlendirilmiştir. İkinci versiyonda veri setine sadece makine öğrenmesi algoritmaları uygulanmış olup ardından performans ölçümü yapılmıştır.

İki sürüm karşılaştırıldığında, PCA uygulanmayan ikinci sürüm, genel olarak tüm makine öğrenmesi algoritmaları için daha başarılı olmuştur. İkinci versiyonda ise en başarılı sonucu CatBoost algoritması vermiştir.

Kalp yetmezliği olan hastaların ölüm durumları makine öğrenmesi teknikleri kullanılarak tahmin edilebilir. Bu makalenin amacı, hastanın ölüm durumunu değerlendirmek için çeşitli değişkenlere bakmaktır. Mortalite durumuna göre uygun tedavi seçilerek hastanın durumu iyileştirilebilir.

Anahtar Kelimeler: Kalp yetmezliği, sınıflandırma, PCA, CatBoost.

ABSTRACT

Analysis of Heart Failure Patient Data with Different Classification Methods

Şevval Tuğçe BADİK

Department of Mathematics

Master of Science Thesis

Supervisor: Asst. Prof. Mutlu AKAR

Heart failure is a cardiovascular disease that causes a large increase in the death rate worldwide. The purpose of this article is to most successfully predict the probability of survival by looking at the various characteristics, diseases and lifestyles of patients using various machine learning methods.

In the data set we used, 299 patients with left ventricular systolic dysfunction were classified as NYHA classes III and IV in 2015. An attempt was made to estimate the probability of survival of patients using various machine learning classification methods on the data set. There are two versions in this study. In the first version of the study, PCA was used to reduce the size of the dataset. The success of machine learning algorithms has been evaluated according to various performance measurement metrics. In the second version, only machine learning algorithms were applied to the data set and then performance measurement was performed.

Compared to the two versions, the second version, which did not implement PCA, was more successful for all machine learning algorithms in general. In the second version, the CatBoost algorithm gave the most successful result.

The death status of patients with heart failure can be estimated using machine learning techniques. The purpose of this article is to look at various variables to assess the patient's death status. The patient's condition can be improved by choosing the appropriate treatment according to the mortality situation.

Keywords: Heart failure, classification, PCA, CatBoost.

1
GİRİŞ

1.1 Literatür Özeti

Bu tezde kullanılan veri seti ilk olarak 2017 yılında yayınlanan Ahmad ve meslektaşlarının [1] yayınladığı makale ile ortaya çıkmıştır. Bu makalede 2015 yılındaki 299 kalp hastasının hayatta kalma durumu Cox Regresyon ve Kaplan Meier plot gibi istatistiksel yöntemler kullanılarak analiz edilmiştir. Daha sonra, 2019 yılında hastaların cinsiyet tabanlı hayatta kalma durum analizi Zahid ve meslektaşları [2] tarafından Cox Regresyon kullanılarak analiz edilmiştir. Son olarak 2020 yılında Chicco ve Jurman [3] makine öğrenmesi algoritmalarını kullanarak hastaların hayatta kalma durum analizini gerçekleştirmiştir. Chicco ve Jurman [3] sadece kreatinin ve ejeksiyon fraksiyon değişkenlerine bakılarak makine öğrenmesi algoritmalarıyla hastaların hayatta kalma analizinin yapılabileceğini öne sürmüşlerdir. Bu iki değişkene ek olarak zaman değişkeni eklenip başarı ölçüm parametreleri ile bakıldığında daha yüksek sonuçlar verdiği gözlenmiştir.

1.2 Tezin Amacı

Günümüzde veri hayatımızın vazgeçilmez bir parçası olmuştur. Gelişen teknolojilere bağlı olarak büyük veri setlerini elde etmek mümkün hale gelmiştir. Veri setleri üzerinde makine öğrenmesi, doğal dil işleme, sentiment analizi gibi ileri analitik teknikleri kullanılarak gerçek hayat problemleri üzerinde karar vermemiz mümkündür. Makine öğrenmesi teknikleri sağlık, finans, siber güvenlik, pazarlama gibi bir çok alanda kullanılabilir [4].

Dünya genelinde kardiyovasküler hastalıklar yetişkenlerde ölüm sebeplerinin başında gelmektedir [5]. Amerika Birleşik Devletleri'nde ortalama her 36.1 saniyede bir kişi kardiyovasküler hastalıklar sebebiyle hayatını kaybetmektedir. 2019 yılındaki veriler dikkate alındığında bir günde ortalama 2396 kişinin hayatını kaybettiği görülmüştür [6].

Kalp yetmezliği (KY) dünya genelinde ölüm oranında büyük ölçüde büyümeye sebep olan bir kardiyovasküler hastalıktır. Kalp yetmezliği Amerika Birleşik Devletleri'nde yaklaşık 6.5 milyon insanı, Avrupa'da yaklaşık 14 milyon insanı, dünya genelinde ise yaklaşık 26 milyon insanı etkilemekte ve sayı gün geçtikçe artmaktadır [7]. Kalp yetmezliğinin görülme oranı erkeklerde kadınlara oranla daha fazladır ve yaş geçtikçe görülme sıklığı artar [8].

Kalp yetmezliği, kalbin normal dolum basınçlarına rağmen veya sadece artan dolum basınçları nedeniyle dokuların metabolik ihtiyaçlarını karşılamak için yeterli oksijen sağlayamamasına neden olan kardiyak yapısal veya fonksiyonel bir bozukluktur. Kalp yetmezliği, kalpteki yapısal veya fonksiyonel bozukluğun neden olduğu, hastaların nefes darlığı, ayak bileği şişmesi ve güçsüzlük gibi tipik semptomlara ve yüksek juguler venöz basınç, akciğer hırıltısı ve yer değiştirmiş apeks atımı gibi belirtilere sahip olduğu klinik bir sendromdur [9].

Kalp yetmezliği genellikle sol ventrikül ejeksiyon fraksiyonuna (LVEF) göre sınıflandırılır, LVEF %50'den büyük olan KY, korunmuş ejeksiyon fraksiyonu (HFpEF) olarak adlandırılır ve LVEF %50'den az olan KY, azaltılmış ejeksiyon fraksiyonu (HFrEF) olarak adlandırılır. Son yıllarda, HFrEF iki bölüme ayrılmıştır. LVEF değeri, %40-%49 aralığında orta menzilli ejeksiyon fraksiyonu ve LVEF değeri %40'ın altında olduğunda azaltılmış ejeksiyon fraksiyonu olarak sınıflandırılır [10]. LVEF değeri %40'ın altında ise sol ventrikül sistolik disfonksiyonu olarak tanımlanır.

Semptomlara göre hastalar, New York Heart Association (NYHA) fonksiyonel sınıflandırması kullanılarak aşağıdaki gibi sınıflandırılabilir:

Sınıf I: KY, fiziksel aktivitede sınırlamalara neden olmaz; sıradan fiziksel aktivite semptomlara neden olmaz.

Sınıf II: KY, fiziksel aktivitede hafif sınırlamalara neden olur; hastalar istirahatte rahattır, ancak olağan fiziksel aktivite kalp yetersizliği semptomlarına neden olur.

Sınıf III: KY, fiziksel aktivitede belirgin sınırlamalara neden olur; hastalar istirahatte rahattır, ancak normal aktiviteden daha az aktivite KY semptomlarına neden olur.

Sınıf IV: KY hastaları, KY semptomları olmaksızın herhangi bir fiziksel aktiviteyi sürdüremezler veya istirahat halindeyken semptomları vardır [11].

Bu tezin amacı makine öğrenmesi tekniklerini uygulayarak çeşitli performans ölçme tekniklerine göre en başarılı sonucu elde edip kalp yetmezliğine sahip hastaların tedavi sürecinde yardımcı olabilecek modeller geliştirmektir.

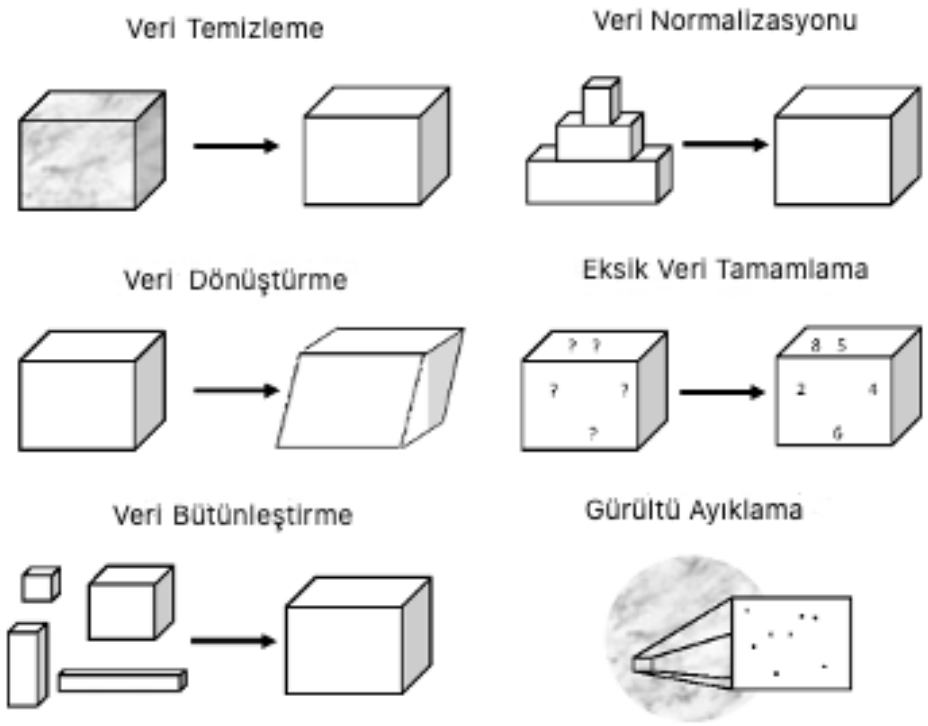
1.3 Hipotez

Makine öğrenmesi teknikleri sağlık alanında kullanılabilir ve sağlık çalışanlarına yardımcı olarak tedavi sürecinin daha hızlı belirlenip aksiyon alınmasında destek olabilir.

2
MAKİNE ÖĞRENMESİ

2.1 Veri Ön İşleme

Herhangi bir makine öğrenmesi tekniğini veri setine uygulamadan önce veri setini modele hazırlamak gerekir. Şekil 2.1 de gösterilen makine öğrenmesi modelleri kurulmadan önce veri seti üzerinde yapılan bir takım düzeltme, eksik veriyi tamamlama, dönüştürme, tekrarlanan verileri çıkarma, bütünleştirme, temizleme, normalize etme, boyut indirgeme işlemlerinin yapıldığı sürece veri ön işleme adı verilir [12].



Şekil 2.1 Veri ön işleme yöntemleri [12]

Veri ön işleme veriyi kaliteli hale getirmek için veri setine makine öğrenmesi teknikleri uygulanmadan önce yapılması gereken adımlardır. Veri ön işleme yöntemleri veri setine uygulanarak modelin performansını artırılıp daha yüksek doğruluk oranına ulaşılabilir.

Veri setine göre uygulanan veri ön işleme adımları değişkenlik gösterebilir. Örneğin bir veri setinde boy ve kilo değişkenleri mevcut olsun. Boy ve kilo değişkenlerinin aralıkları

birbirinden oldukça farklı olacağından veri seti normalize edilip boy ve kilo değişkenlerinin model tarafından anlaşılır hale getirilmesi gerekir, aksi halde model istenen performansı gösteremeyecektir.

Veri setini göz önüne alarak çok fazla bağımsız değişken mevcutsa bu modelin performansını düşürebileceğinden özellik seçimi yada özellik çıkarımı gibi veri ön işleme adımı uygulanabilir. Veri setinde eksik yada outlier değer bulunuyorsa eksik değer ortalama, alt veya üst sınır gibi değerlerle doldurulabilir yada veri setinden çıkarılabilir. Veri seti dengeli dağılmamış ise gözlemlerin az olduğu durumda gözlemler çoğaltılabilir fazla olduğu durumda ise gözlemler azaltılabilir. Veri ön işleme adımı veri setine göre uygulanır.

2.2 Öğrenme Türleri

Kullanılacak olan veri setinde bulunan verinin durumuna göre öğrenme türlerini üçe ayırabiliriz.

2.2.1 Denetimli Öğrenme

Denetimli öğrenmede kullanılan veri seti içinde eğitim seti etiketli verilerden oluşur. Denetimli öğrenme ile sonuçları bilinen verilerden yararlanılarak oluşturulan model temel alınır. Bu modele göre veri kümesinde etiket bilgisi bulunmayan verilerin sonuçları tahmin edilmeye çalışılır. Denetimli öğrenme yöntemlerine sınıflandırma ve regresyon algoritmalarını örnek verilebilir.

2.2.2 Denetimsiz Öğrenme

Denetimsiz öğrenme, veriyi etiketlemeden veri içerisindeki ilişkilerin ve yapıların öğrenilmesidir. Burada algoritmalar verilerden çıkarımlar yaparak verileri gruplara ayırır ve yeni veriyi en uygun gruba atamaya çalışır. Birbirine benzeyen nesneler aynı gruba aittir yaklaşımını uygular. Denetimsiz öğrenmeye örnek olarak temel bileşen analizi (PCA) ve kümeleme yöntemleri denetimsiz öğrenme yöntemleridir. Tavsiye sistemlerinde kullanımı örnek verilebilir.

2.2.3 Yarı Denetimli Öğrenme

Yarı denetimli öğrenme, veri setinde etiketli ve etiketsiz verilerin aynı anda bulunması durumunda uygulanabilir. Model, ödül ve ceza sistemi uygulanarak öğrenmeyi

gerçekleştirir. Örneğin bir robotun engele çarpması ceza olarak bildirilir ve bundan sonuç çıkararak engele çarpmamayı öğrenebilir. Oyunlar ve robotik alanlarında kullanımı örnek verilebilir.

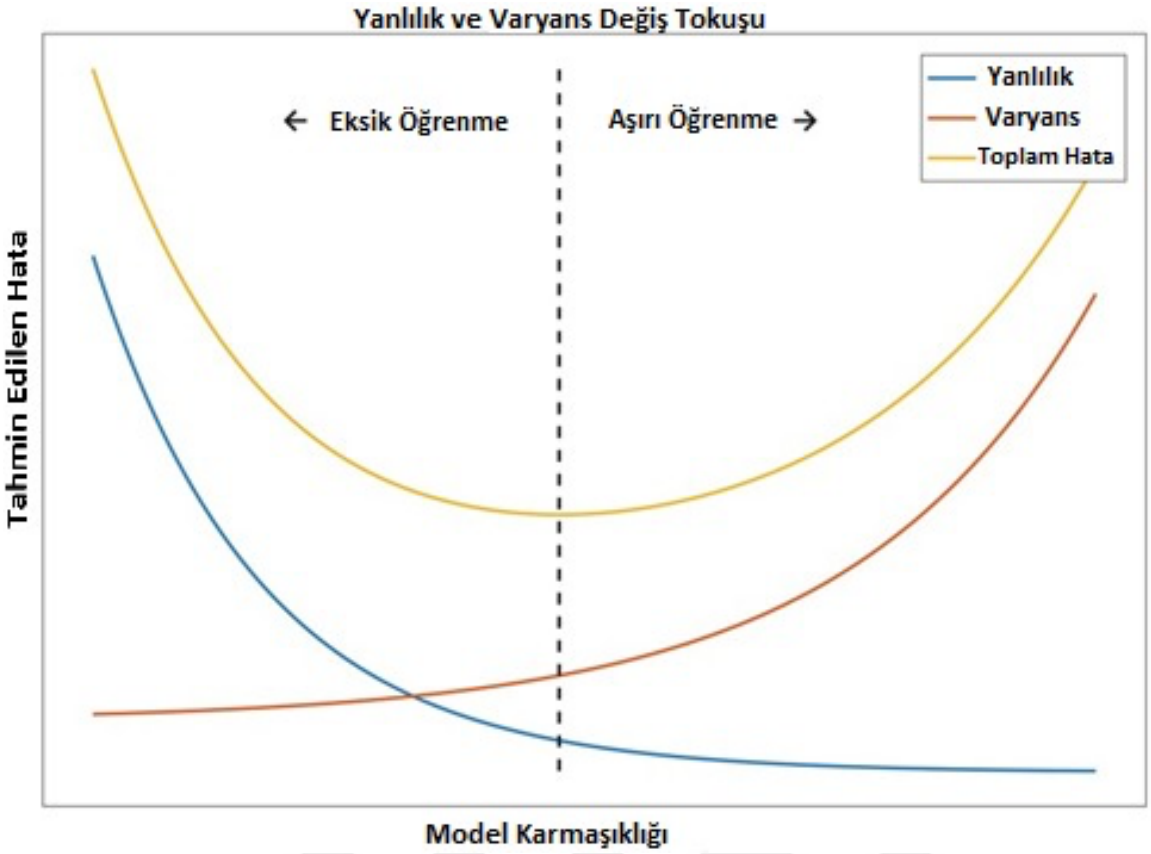
2.3 Makine Öğrenmesi Modelleri

Arthur Samuel'e göre Makine Öğrenimi, bilgisayarlara açıkça programlanmadan öğrenme yeteneği veren çalışma alanı olarak tanımlanır.

Veri setinde bulunan veriler X bağımsız değişkenleri girdi, Y bağımlı değişkeni ise çıktı olarak adlandırılabilir.

Makine öğrenmesi algoritmaları veri setine uygulandığında modelin öğrenme kabiliyeti yanlılık ve varyans kavramlarını ortaya çıkarır. Yanlılık, gerçek değerler ile tahmin edilen değerler arasındaki mesafeyi ifade eder. Yüksek yanlılığa sahip bir modelde eğitim verisi az öğrenilmiş olur ve modelin tahmin etme ihtimali zayıf kalır. Bu durum eksik öğrenmeye (underfitting) yol açar.

Varyans, bir algoritmanın esnekliğidir. Modelin tahmin etme kabiliyeti arttıkça varyansı artar, yanlılık azalır. Belli bir süre başarı artmaya devam eder ancak bir noktadan sonra model öğrenmeye değil ezberlemeye başlar. Bu durumda aşırı öğrenme (overfitting) ortaya çıkar. Model herhangi bir değişikliğe aşırı duyarlı hale gelir.



Şekil 2.2 Yanlılık-varyans değiş tokuşu [13]

Şekil 2.2'de görüleceği üzere optimum model düşük varyans ve düşük yanlılık olan modeldir.

2.3.1 Regresyon Analizi

Regresyon analizi, bağımlı bir değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkilerin tahmin edilmesi için kullanılan bir dizi istatistiksel yöntemdir. Regresyon analizi, verileri modelleyip bu verilere en uygun fonksiyonu bulmayı amaçlar. Basit lineer regresyon, çoklu lineer regresyon ve lojistik regresyon, regresyon analizinde en çok kullanılan yöntemlerdendir.

Basit lineer regresyonda bir bağımlı ve bir bağımsız değişken olmak üzere toplam 2 değişken üzerinde çalışılır. Basit lineer regresyonda amacımız bağımlı ve bağımsız değişken arasındaki ilişkiyi ifade eden lineer fonksiyonu bulmaktır.

Basit Lineer Regresyon Modeli Denklemi:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.1}$$

β_0, β_1: Regresyon katsayıları

ε: Hata terimi

X: Bağımsız değişken

Y: Bağımlı değişken

Örneklem Gerçek Değerler:

$$y_i = b_0 + b_i x_i + e_i \tag{2.2}$$

Tahmin Modeli:

$$\hat{y}_i = b_0 + b_i x_i \tag{2.3}$$

Buradaki b_i katsayılarını denklemde optimum olacak şekilde bulmayı amaçlıyoruz.

Hatalar:

$$e_i = y_i - \hat{y}_i \tag{2.4}$$

$$e_i = y_i - b_0 - b_i x_i \tag{2.5}$$

Birden fazla bağımsız değişkenin mevcut olduğu durumlarda çoklu lineer regresyon uygulanır. Çoklu lineer regresyon bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi bulmayı ve modeli kurmayı amaçlar.

Lineer regresyon modellerinde bağımsız değişkenler sürekli veya kategorik olabilirken bağımlı değişken sürekli olmak zorundadır. Lineer regresyonda bağımlı değişken $(-\infty, \infty)$ arasında değer alır.

Lojistik regresyonda ise bağımlı değişken kategorik olabilir. Lojistik regresyon modeli, sınıflandırma modeli oluşturarak bağımlı değişkeni en yüksek olasılık değerine sahip sınıfa atar. Lojistik regresyonda yapılacak tahmin 0 ile 1 arasında olur.

Lojistik fonksiyonda eşik değeri tanımlanır. Elde edilen bağımlı değişken eşik değeri dikkate alınarak sınıflandırılır.

Lojistik regresyon modelinin denklemi aşağıdaki gibi ifade edilir:

$$E(Y/x) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \qquad 0 \leq E(Y/x) \leq 1 \tag{2.6}$$

$E(Y/x)$: x gözleminin Y sınıfında olma olasılığı

2.3.2 Sınıflandırma Yöntemleri

Sınıflandırma veri setini sınıflara ayırma işlemidir. Bağımlı Y değişkeninin kategorik olduğu durumlarda sınıflandırma yöntemi kullanılarak model oluşturulur. Girdi olarak verilen X bağımsız değişkenlerinin sonucunda Y bağımlı değişkeninin hangi sınıfa ait olduğu tahmin edilir.

2.3.2.1 Karar Ağaçları

Karar ağacı algoritması ağaç yapısında bir sınıflandırma modelidir. If-then şeklinde işleyen kural tabanlı bir algoritmadır. Karar ağaçlarının ilk hücrelerine kök denir. Her bir gözlem kökteki özelliğe bakarak kurallara göre seçim yapılarak sınıflandırılır.

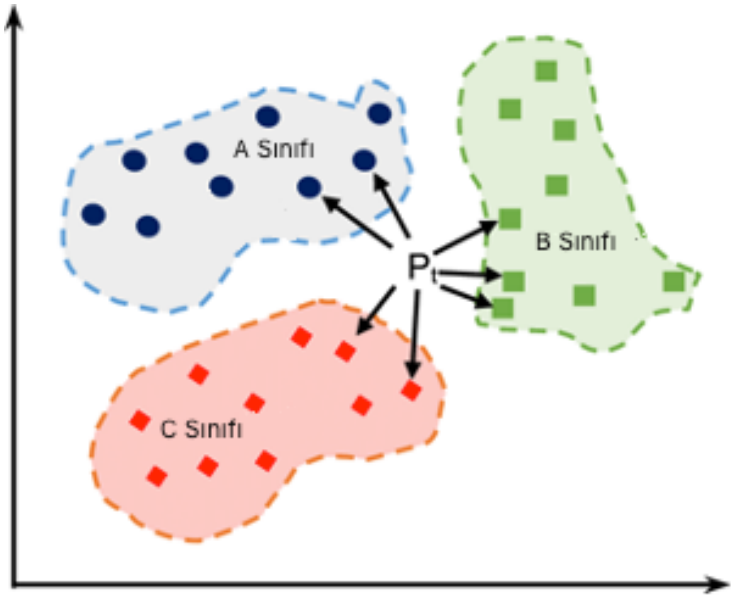
Kök hücrelerin altında düğümler bulunur. Her bir gözlem düğümler yardımıyla sınıflandırılır. Düğümler kökteki değişkenin durumuna göre verilen bir kararı gösterir. Bir sınıfa ulaşılıncaya kadar düğüm oluşmaya devam eder. Düğüm sayısı arttıkça modelin karmaşıklığı da artar.

Karar ağacının en altında bulunan yapılara yaprak adı verilir ve çıktı olan sınıfı gösterir.

Karar ağaçları başarılı ve hızlı sonuç verdiği için tercih edilebilir ancak veri setindeki değişken sayısı arttıkça model karmaşıklaşacağı için aşırı öğrenme söz konusu olabilir. Böyle durumlarda model, hiperparametre değerleri ile oynanarak en optimum parametre çeşitli yöntemlerle seçilebilir.

2.3.2.2 K-En Yakın Komşuluk

K-en yakın komşuluk (K-Nearest Neighbor=KNN) modelinde tahminler gözlem benzerliğine göre yapılır. Büyük veri setlerinde performans açısından çok iyi değildir. Öklid, Hamming veya Cosine gibi bir uzaklık hesabı ile her bir gözleme uzaklık hesaplanır. Sonrasında en yakın k adet gözlemin ortalaması alınır ve yeni değere atanarak sınıfı belirlenir.



Şekil 2.3 KNN ile sınıflandırma [14]

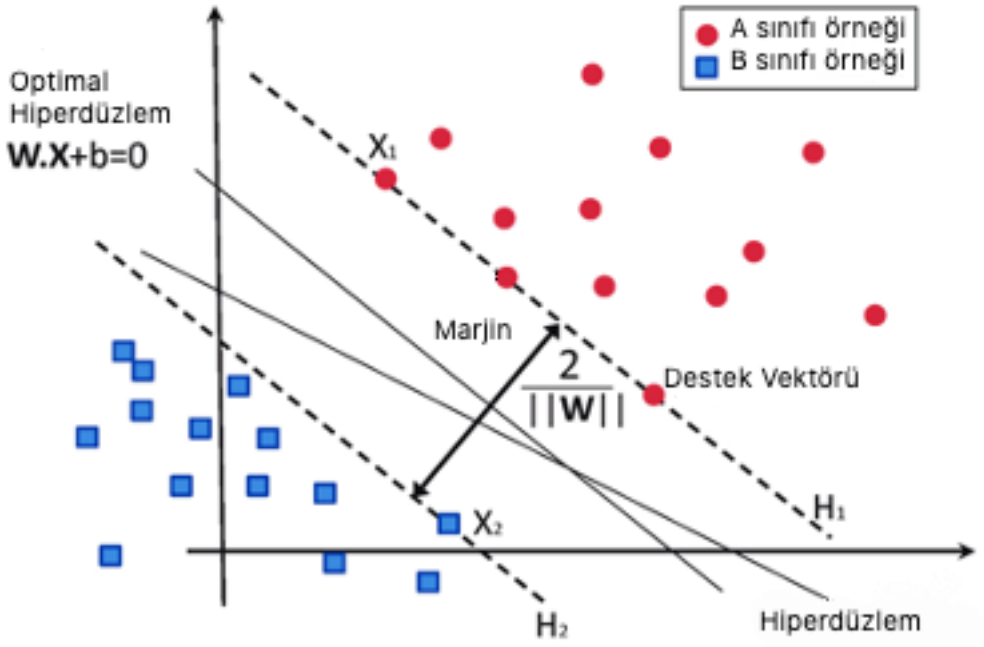
KNN Aşamaları:

1. k herhangi bir tam sayı olacak şekilde seçilir.
2. Belirlenen uzaklık ölçümü ile test verisi ile sınıfları belli olan eğitim verisi arasındaki uzaklık ölçülür.
3. Uzaklık ölçümüne göre sıralama yapılır.
4. Belirlenen k sayısına göre en yakın k gözlem seçilir.
5. Yeni gözlem en sık tekrar eden sınıfa atanır.

Şekil 2.3'de belirtilen modelde k=7 için en yakın 7 adet gözlem seçilip P_t gözleminin sınıfı belirlenmeye çalışılıyor.

2.3.2.3 Destek Vektör Makinesi

Destek vektör makinesi (support vector machine=SVM) modelinde amaç iki sınıf arasındaki ayrımın optimum olmasını sağlayacak hiperdüzlemi bulmaktır. Marjin, hiperdüzlem ile en yakın veri noktası arasındaki uzaklıktır. Bu veri noktalarına support vector denir. İki sınıf arasındaki marjinin maksimum olmasını sağlayacak şekilde hiperdüzlemi bulmayı hedefliyoruz. SVM, aykırı değerlere karşı dayanıklı olup aykırı değerleri göz ardı ederek maksimum marjin olacak şekilde hiperdüzlemi bulabilir.



Şekil 2.4 SVM ile sınıflandırma [15]

Şekil 2.4'de A ve B şeklinde iki sınıflı bir SVM modeli örneği bulunmaktadır. Veri setinde eğitim setinde m adet gözlem olduğunu varsayalım. $(x_i, y_i), i = 1,2,\ldots,m$ olacak şekilde

$$x_i^T = (x_{i1}, x_{i2}, \ldots, x_{id}) \in \mathbb{R}^d \tag{2.7}$$

i. gözlemin d boyutlu değişkeni şeklinde ifade edilir ve $y \in \{-1, +1\}$ şeklindedir.

x_i pozitif sınıfa aitse y_i $+1$, x_i negatif sınıfa aitse y_i -1 değerini alır. Eğitim seti aşağıdaki şekilde bir hiperdüzlemle ikiye ayrılabilir.

$$H(w, b) = \{x | w^t x + b = 0\} \tag{2.8}$$

w: Ağırlık Değerleri

b: Bias

x: Gözlem

H_1 ve H_2 marjinal hiperdüzlemler olarak adlandırılır ve aşağıdaki şekilde ifade edilir:

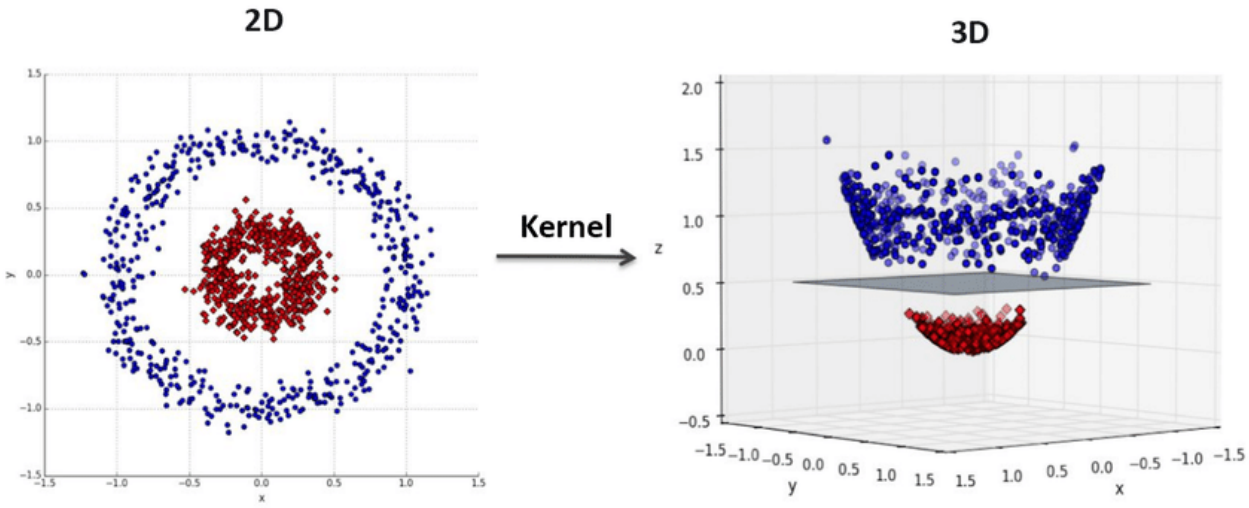
$$H_1 : (w^t x_i + b = 1) \tag{2.9}$$

$$H_2 : (w^t x_i + b = -1) \tag{2.10}$$

Buradan marjinal hiperdüzlemler arası uzaklık:

$$\text{Marjin} = \frac{2}{\|W\|} \tag{2.11}$$

şeklinde ifade edilir.



Şekil 2.5 Kernel Trick ile boyut artırma [16]

Sınıfların lineer olarak ayrılamadığı noktalarda kernel trick kullanılır. Kernel trick ile boyut artırılarak sınıf ayrımı yapılabilir. Örneğin iki boyutlu lineer olmayan iki sınıfı üç boyuta çevirerek ayrıştırma yapılabilir. Şekil 2.5'de lineer olarak ayrılamayan iki sınıfın kernel trick kullanılarak 2 boyuttan 3 boyuta geçirilerek ayrıştırılabildiği görülmektedir. Çeşitli yöntemlerle boyut artırmak mümkündür. Örneğin, Akar ve Sirakov [17] deri lezyonları veri seti üzerine, Clifford cebirini kullanarak verinin boyutunu artırarak sınıflandırma yöntemleri uygulamışlardır. Kernel kullanılarak boyut artırma dışında Akar, Sirakov ve Mete [18] Clifford cebirini kullanarak yeni bir kernel elde etmiş ve bu kerneller ile daha başarılı sonuçlar elde etmiştir.

2.3.2.4 Naive Bayes

Naive Bayes algoritması temeli Bayes Teoremine dayanan olasılık temelli bir modelleme yöntemidir. Amaç belirli bir örneğin her bir sınıfa ait olma olasılığının koşullu olasılık temelli hesaplanmasıdır.

Bayes teoremi rassal bir sürece bağlı olarak ortaya çıkan rastgele bir A olayı ile rastgele bir B olayı için koşullu olasılıklar ve marjinal olasılıklar arasındaki ilişkiyi tanımlar. Thomas Bayes tarafından ifade edilen Bayes Teoremi aşağıdaki gibidir:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2.12)$$

$P(A|B)$: B olayı gerçekleştiği durumda A olayının gerçekleşme olasılığı

$P(B|A)$: A olayı gerçekleştiği durumda B olayının gerçekleşme olasılığı

$P(A)$: A olayının gerçekleşme olasılığı

$P(B)$: B olayının gerçekleşme olasılığı

Naïve Bayes sınıflandırma algoritmasında her değişken birbirinden bağımsız kabul edilir.

$x_i = (x_1, x_2, \ldots, x_n)$ değeri sınıfı bilinmeyen bir gözlem olsun ve n adet değişkenden oluşsun.

K adet sınıf mevcut olsun. x_i değerinin C_i sınıfında olma olasılığı aşağıdaki şekilde ifade edilebilir:

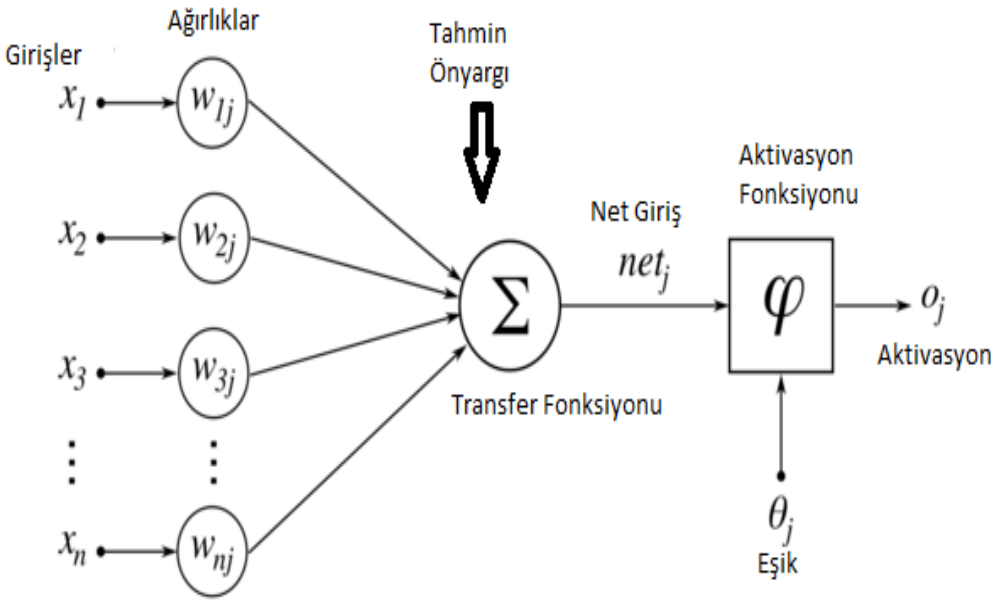
$$P(C_i|x_i) = \frac{P(x_i|C_i)P(C_i)}{P(x_i)} \quad (2.13)$$

2.3.2.5 Yapay Sinir Ağı

Yapay sinir ağı (YSA) algoritması insan beyninin bilgi işleme şeklini referans alarak bir model oluşturur.

Beyni oluşturan temel birimler sinir hücreleri (nöron) ve bunların uzantılarının diğer sinir hücreleri ile oluşturduğu değme noktaları olarak adlandırılan sinapslerdir. Nöronların oluşturduğu ağ örüntü sayısı ne kadar fazla olursa, bilgi işleme süreci o kadar güçlü olur. Her nöronun dentrit adı verilen çok sayıda kısa ve akson adı verilen bir tane uzun uzantısı vardır. Akson uçları ile başka nöronların dentritleri veya gövdeleri arasındaki bağlantıya sinaps adı verilir [19].

Dentritler sinyalleri çekirdeğe iletir. Çekirdek topladığı verileri akson aracılığı ile sinapslere gönderir. Sinapsler bu sinyaller üzerinde işlem yapıp hücrelere aktarır.



Şekil 2.6 Yapay sinir hücresi [20]

Yapay sinir ağları girdi katmanı, ara katman ve çıktı katmanından oluşur. Bu katmanlar bir araya gelerek bir ağ oluştururlar. Veriler öncelikle girdi katmanına gönderilir, (x_i) değişkenleri burada iletilir. Daha sonra ara katmanda veriler ağdaki ağırlık değerleri kullanılarak işlenir ve çıktı değerleri elde edilerek çıktı katmanına iletilir. Ağ içerisindeki girdilerden doğru çıktılar elde edebilmek için doğru ağırlıkların bulunması yani ağın eğitilmesi gerekmektedir [21].

Şekil 2.6'da yapay sinir ağı hücresi gösterilmiştir. Nöronlar aracılığıyla gelen sinyallerin ağırlıklandırılmış toplamları net girdiyi oluşturur. Nöron çıktı sinyalini belirlemek için gelen sinyallere bir aktivasyon fonksiyonu uygulanır [21].

Amaç hatayı minimum yapacak ağırlıkları bulmaktır. Yapay sinir ağında ağırlıklar kurulan modele göre belirlenir. Belirli iterasyon boyunca ağırlıklar optimum düzeye ulaşmaya çalışır. Bu süreçte makine öğrenme işlemi gerçekleşir.

2.3.2.6 Rastgele Orman

Rastgele orman (random forest=RF) algoritması, karar ağacı algoritmasını n defa çalıştırarak daha iyi tahmin yapabilmeyi sağlar. Karar ağaçları veri setinden rastgele seçilen alt kümelerdir. Seçilen karar ağaçları bir araya gelerek random forest algoritmasını oluşturur [22].

RF algoritması geleneksel sınıflandırma ve regresyon ağaçlarından farklı olarak çok sayıda karar ağacı oluşturmaktadır. Oluşturduğu karar ağaçlarının kombinasyonu üzerinden değerlendirme yapabilme imkanı sağlamaktadır [23]. Her bir karar ağacının dahil olduğu sınıf belirlenir ve en çok tekrar eden sınıf çıktı olarak kabul edilir.

2.3.2.7 Gradient Boosting

Gradient Boosting algoritması karar ağacı tabanlı bir algoritmadır. Gradient Boosting modelinde ilk olarak yaprak oluşturulur.

RF algoritmasındaki gibi veri setini çoklu sayıda alt veri setine bölme bu yaklaşımda uygulanmamaktadır. Veri setini kullanarak bir karar ağacı oluşturulur. Oluşturulan karar ağacının hatalarından yola çıkarak yeni bir karar ağacı daha oluşturulur. Bu şekilde ardışık şekilde karar ağaçları oluşturulur. Gradient Boosting ardışık ve bağımlı çok sayıda ağaç üretmektedir [24]. Bağımlı şekilde ağaç ürettiği için performans olarak daha yavaş ilerleyebilmektedir.

2.3.2.8 XGBoost

XGBoost, Gradient Boosting algoritmasının yüksek performanslı halidir. XGBoost algoritmasında ilk adım ilk tahmini yapmaktır. Bu tahmin, bundan sonraki adımlarda yapılacak işlemler ile yakınsayarak doğru sonuca ulaşılacağı için herhangi bir sayı olabilir. Bu sayı varsayılan olarak 0,5'tir. Daha sonra Gradient Boosting algoritmasında olduğu gibi hataları tahminleyen karar ağacı kurulur. Burada amaç hataları öğrenip doğru tahmine yaklaşmaktır. Hata oranı düşünceye kadar ya da belirlenen ağaç sayısına ulaşıncaya kadar ağaçlar üretilmeye devam eder [25].

2.3.2.9 CatBoost

Catboost, Yandex şirketi tarafından geliştirilmiş olan Gradient Boosting tabanlı bir makine öğrenmesi algoritmasıdır. Gradient Boosting algoritmasının performansını arttırmak amacıyla geliştirilen XGBoost algoritmasına alternatif olarak geliştirilmiştir. CatBoost, veri ön işleme evresini kısaltarak zaman açısından avantaj sağlar. Boş veriler ile başa çıkar ve kategorik verilere encoding yöntemini uygular. Catboost algoritması simetrik ağaçlar kurar. Bu sayede çok derin ağaçlar kurmadan yüksek tahmin oranı yakalar ve aşırı öğrenme sorununu aşar [26].

2.4 Başarı Ölçme Parametreleri

2.4.1 Karışıklık Matrisi (Confusion Matrix)

Sınıflandırma yöntemlerinde başarıyı ölçmek için birden fazla metrik kullanılabilir. Başarı ölçmede kullanılan parametrelerde karışıklık matrisi kullanılır. Karışıklık matrisi Tablo 2.1 deki gibi ifade edilir. 0: Negatif, 1: Pozitif değeri ifade etmektedir.

Tablo 2.1 Karışıklık matrisi

	Tahmin Edilen Değer:0	Tahmin Edilen Değer:1
Gerçek Değer:0	TN	FP
Gerçek Değer:1	FN	TP

Gerçek Pozitif Değer (True Positive=TP): Gerçek değerin 1, tahmin edilen değerin 1 olduğu durumlardır.

Gerçek Negatif Değer (True Negative=TN): Gerçek değerin 0 ve tahmin edilen değerin 0 olduğu durumlardır.

Yanlış Pozitif Değer (False Positive=FP): Gerçek değerin 0, tahmin edilen değerin 1 olduğu durumlardır.

Yanlış Negatif Değer (False Negative=FN): Gerçek değerin 1, tahmin edilen değerin 0 olduğu durumlardır.

2.4.2 Doğruluk (Accuracy=ACC)

Doğruluk oranı (ACC) gerçek değer ile tahmin edilen değerin aynı olduğu doğru tahminlerin, tüm tahminlere oranıdır.

$$ACC = \frac{TP+TN}{TP+TN+FN+FP} \tag{2.14}$$

2.4.3 Matthews Korelasyon Katsayısı (Matthews Correlation Coefficient=MCC)

Matthews korelasyon katsayısı (MCC) gerçek değerler ile tahmin edilen değerler arasındaki korelasyon ilişkisine bakarak değerlendirme yapar. MCC karışıklık matrisinde bulunan tüm değerleri baz alır [27]. MCC değeri aşağıdaki gibi ifade edilir:

$$\boldsymbol{MCC} = \frac{\boldsymbol{TP}\times\boldsymbol{TN}-\boldsymbol{FP}\times\boldsymbol{FN}}{\sqrt{(\boldsymbol{TP}+\boldsymbol{FP})(\boldsymbol{TP}+\boldsymbol{FN})(\boldsymbol{TN}+\boldsymbol{FP})(\boldsymbol{TN}+\boldsymbol{FN})}} \tag{2.15}$$

2.4.4 Duyarlılık (Sensitivity=SE)

Duyarlılık (SE) parametresi TP değerinin TP+FN değerine oranıdır. SE parametresinin gösterimi aşağıdaki gibidir:

$$SE = \frac{TP}{TP+FN} \tag{2.16}$$

2.4.5 Özgüllük (Specifity=SP)

Özgüllük (SP) parametresi TN değerinin TN+FP değerine oranıdır. SP parametresinin gösterimi aşağıdaki gibidir:

$$SP = \frac{TN}{TN+FP} \tag{2.17}$$

2.4.6 Hassasiyet

Hassasiyet parametresi TP değerinin TP+FP değerine oranıdır. Hassasiyet parametresinin gösterimi aşağıdaki gibidir:

$$Hassasiyet = \frac{TP}{TP+FP} \tag{2.18}$$

2.4.7 ROC AUC

ROC eğrisinin altında kalan alan (ROC AUC) yanlış pozitif oranına (FPR) ve gerçek pozitif oranına (TPR) bağlıdır. ROC eğrisinde Y ekseninde ise TPR değeri, X ekseninde ise FPR değeri bulunmaktadır [28]. TPR ve FPR aşağıdaki eşitlik ile sağlanır.

$$TPR=SE \tag{2.19}$$

$$FPR=1\text{-}SP \tag{2.20}$$

2.4.8 PR AUC

PR eğrisinin altında kalan alan (PR AUC) duyarlılık ve hassasiyet değerlerine bağlıdır. PR eğrisinde Y ekseninde ise hassasiyet değeri, X ekseninde ise duyarlılık değeri bulunmaktadır.

2.4.9 F_1 Score

F_1 Score hassasiyet ve duyarlılık parametrelerinin harmonik ortalamasıdır [29]. F_1 Score aşağıdaki gibi ifade edilir:

$$F_1 = 2 \times \frac{Hassasiyet \times Duyarlılık}{Hassasiyet + Duyarlılık} \tag{2.21}$$

3
UYGULAMA VE ANALİZ

3.1 Veri Seti

Veri seti Nisan-Aralık 2015 tarihleri arasında Faisalabad-Pakistan Kardiyoloji Enstitüsü ve Allied Hastanesine başvuran 299 hastadan oluşmaktadır. Veri seti 105 kadın hasta ve 194 erkek hastayı içermektedir. Veri setinde hastaların yaşı, sigara içme alışkanlığı olup olmadığı, çeşitli analiz sonuçları ve diyabet, kan basıncı ve anemi olup olmadığı verilmiştir.

Ahmad ve meslektaşlarının [1] yayınladığı çalışmada hastaların başka bir önemli hastalığı olup olmadığından bahsedilmemiştir. Hastaların yaşları 40 ile 95 arasında değişmekte olup, yaş ortalaması 60,83'tür. Hastaların kanında kreatinin fosfokinaz (CPK), trombosit, ejeksiyon fraksiyonu, kreatinin, sodyum değerlerine bakılmıştır. Bu değişkenler veri setinde sürekli değişkenler olarak verilmiştir. Sigara, diyabet, anemi, kan basıncı (BP) ve cinsiyet, veri setinde bulunan kategorik değişkenlerdir. Bu değişkenlere ek olarak, hastaların ölüm veya hayatta kalma süreleri, zaman değişkeni olarak tanımlanmış ve gün olarak verilmiştir. Hastaların tamamında sol ventrikül sistolik disfonksiyonu vardır ve NYHA sınıf III ve IV olarak sınıflandırılmıştır. Zaman değişkeni 4–285 gün arasında değişmekte olup, ortalama 130 gündür [1].

Veri seti eksik değer içermemektedir. Bu değişkenleri kullanarak hastaların ölüm ve hayatta kalma şanslarını mümkün olduğunca doğru bir şekilde sınıflandırmayı amaçlıyoruz. Veri setinde ölüm=1 vefat eden bir hastayı, ölüm=0 ise hayatta kalan bir hastayı ifade etmektedir.

CPK enzimi, kreatin ve adenosin trifosfatın (ATP) fosfokreatin ve adenosin difosfata (ADP) tepkimesini katalize eder. 20 ila 200 IU/L arasındaki CPK seviyeleri normal referans aralığında olarak kabul edilir. CPK seviyeleri, rabdomiyoliz, kalp hastalığı, böbrek hastalığı ve hatta bazı ilaçları içeren çeşitli faktörlerden etkilenebilir [30].

Trombositler, kan dolaşımında kan pıhtıları oluşturmaya yardımcı olan ve kalp krizi, felç ve periferik damar hastalığına karışan özel disk şeklindeki hücrelerdir [31]. HFrEF'li

hastalarda trombositler ile artmış mortalite arasında bir bağlantı vardır. Trombositler, HFrEF'li hastalarda kötü sonuçları öngörmenin bir yolu olabilir [32].

Kreatinin, veri setinde yer alan bir değişken olan serum kreatininini ifade eder. Daha yüksek kreatinin seviyeleri ve serum kreatininde daha fazla artış, hastanede daha uzun süre kalma, daha yüksek uzun vadeli mortalite ve daha yüksek yeniden hastaneye yatış oranları ile ilişkilendirilmiştir [33].

Ejeksiyon fraksiyonu, her kalp döngüsünde atılan kan hacminin yüzdesidir. Ejeksiyon fraksiyonu sol ventrikül sistolik performansını gösterir [34].

Sodyum, tüm dünyada diyetlerde kullanılan bir mineraldir. Sodyum, uygun kan hacmini ve kan basıncını korumak için gereklidir [35].

Tablo 3.1 Kategorik değişkenlerin hastalar üzerinde dağılımı

Kategorik Değişkenler	Ölen Hasta Sayısı	Hayatta Kalan Hasta Sayısı
Sigara Alışkanlığı (1: Var)	33	66
Sigara Alışkanlığı (0: Yok)	66	137
Diyabet (1: Var)	40	85
Diyabet (0: Yok)	56	118
Anemi (1: Var)	46	83
Anemi (0: Yok)	50	120
BP (1: Var)	39	66
BP (0: Yok)	57	137
Cinsiyet (1: Erkek)	62	132
Cinsiyet (0: Kadın)	34	71
Toplam Hasta Sayısı	96	203

Tablo 3.1'de görüldüğü üzere toplam ölen hasta sayısı 96, hayatta kalan hasta sayısı 203'tür. Dolayısıyla bu durumda veri setinin dengesiz olduğunu varsayabiliriz. Sürekli değişkenlerin ölen, hayatta kalan ve toplam hastalara göre ortalamaları Tablo 3.2'de gösterilmiştir.

Tablo 3.2 Sürekli değişkenlerin ortalamasının hastalar üzerinde dağılımı

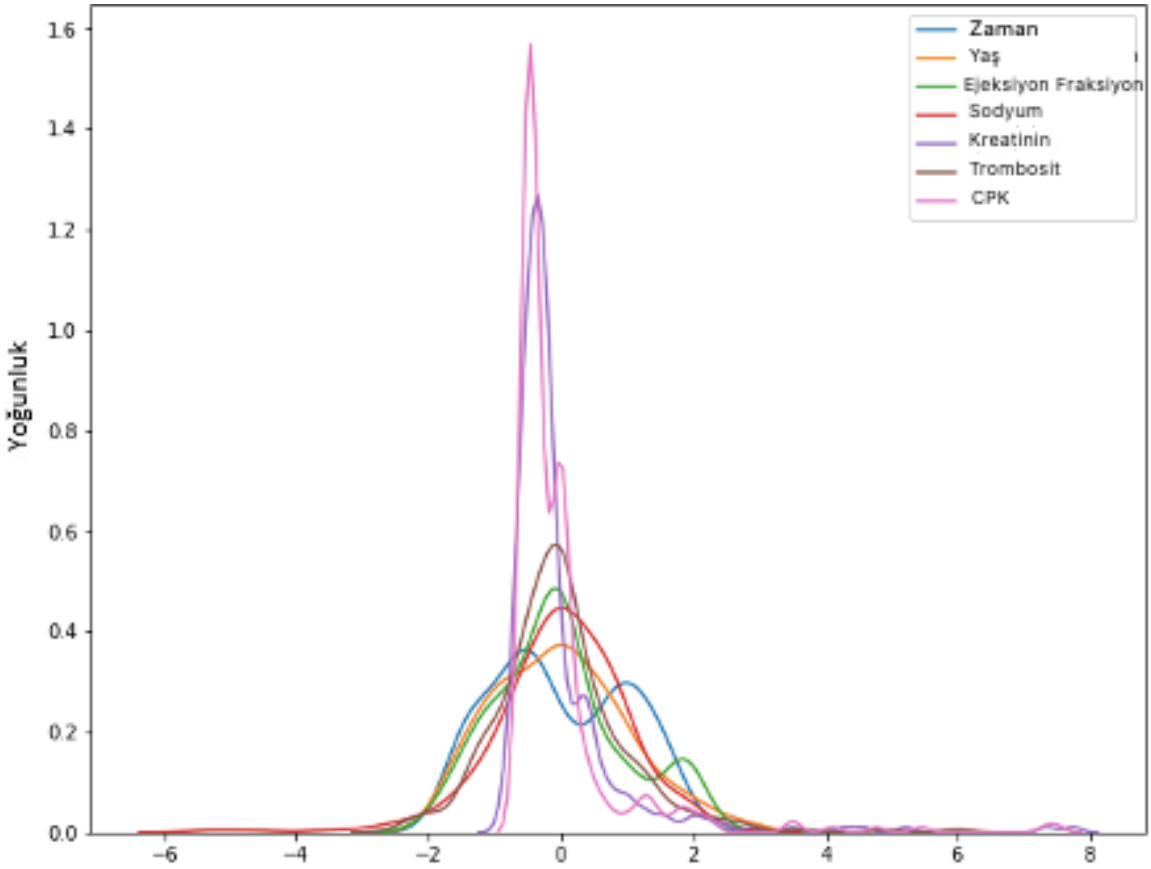
Sürekli Değişkenler	Ölen Hastaların Ortalaması	Hayatta Kalan Hastaların Ortalaması	Tüm Hastaların Ortalaması
Yaş	65.21	58.76	60.83
Ejeksiyon Fraksiyon	33.47	40.27	38.08
Sodyum	135.38	137.22	136.63
Kreatinin	1.83	1.18	1.39
Trombosit	256381.04	266657.49	263358.03
CPK	670.20	540.05	581.84
Zaman	70.89	158.34	130.26

3.2 Sınıflandırma

Veri seti, zaman değişkeni dahil olmak üzere bir bağımlı değişken yani ölüm değişkeninden ve toplam 12 bağımsız değişkenden oluşmaktadır. Bu çalışma, bu değişkenlere göre en başarılı sınıflandırmayı yapacak bir model oluşturmayı amaçlamaktadır. Çalışmanın tamamında JupyterLab programı kullanılmıştır.

İlk olarak, veri setinde "event" ifadesi "ölüm" ifadesine karşılık gelecek şekilde değiştirilmiştir. Ölüm=1 ise hasta vefat eder, ölüm=0 ise hasta hayatta kalır olarak kabul edilmiştir.

Daha sonra, kategorik değişkenleri veri setinden çıkararak veri setini ölçeklemek için Standard Scaler kullanılmıştır. Kategorik değişkenler daha sonra veri kümesine eklenmiştir.



Şekil 3.1 Standard Scaler sonrası sürekli değişkenlerin dağılımı

Veri setindeki sürekli değişkenleri ölçeklendirdikten sonra Şekil 3.1'de görüldüğü gibi sürekli değişkenler belirli bir aralıkta dağılım gösterirler. Değişkenleri ölçeklendirmek için Standard Scaler kullanmanın amacı, veri setindeki değişkenleri belirli bir aralığa koyarak değişkenlerin önemini daha net görebilmek ve modeli daha başarılı hale getirmektir.

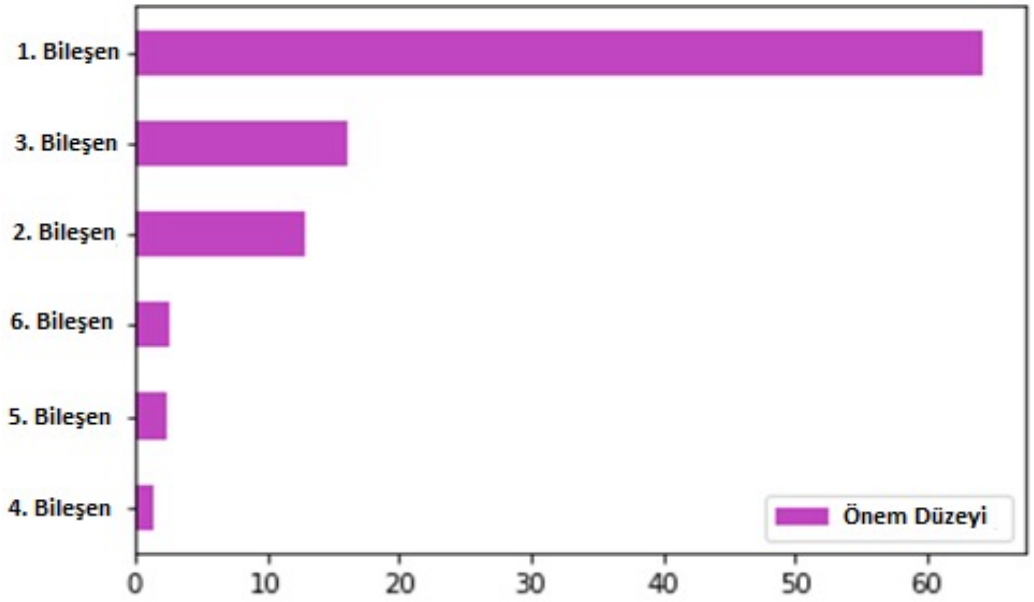
Pearson korelasyon katsayısı (pearson correlation coefficient=PCC) [36], veri seti ölçeklendirildikten sonra bağımlı değişken "ölüm" ile bağımsız değişkenler arasındaki ilişkiyi incelemek için kullanılır. Tablo 3.3'e göre, ölüm değişkeni için en önemli değişkenler, PCC ile gösterildiği gibi yaş, zaman, kreatinin ve ejeksiyon fraksiyonudur.

Tablo 3.3 Değişkenlerin ölüm değişkenine göre PCC hesaplaması

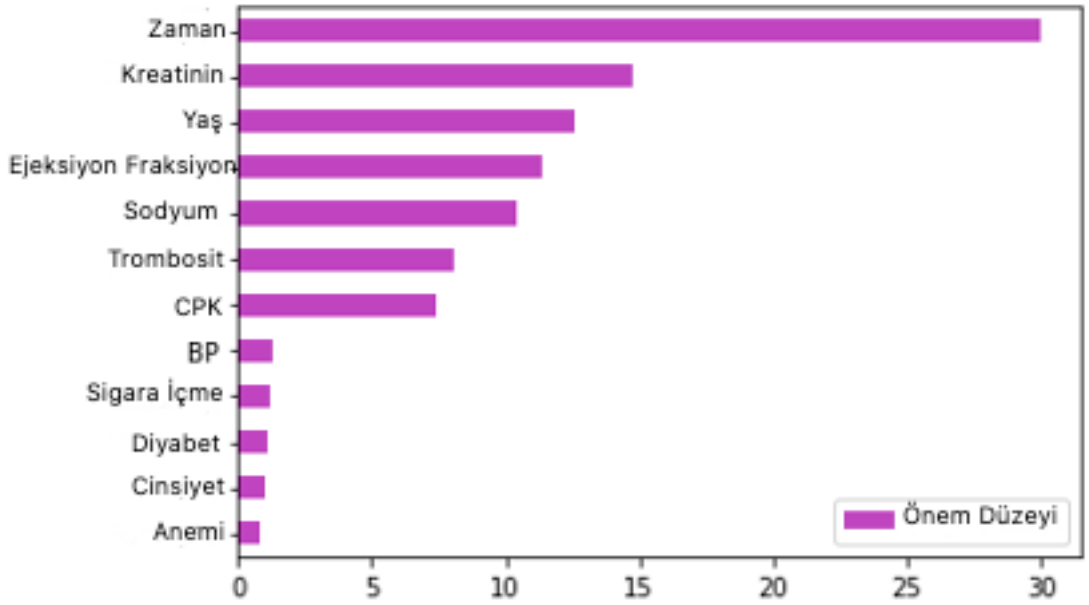
Değişken	PCC
Zaman	-0.5270
Yaş	0.2537
Ejeksiyon Fraksiyon	-0.2690
Sodyum	-0.1952
Kreatinin	0.2942
Trombosit	-0.0491
CPK	0.0627
Cinsiyet	-0.0043
Sigara Alışkanlığı	-0.0126
Diyabet	-0.0019
BP	0.0793
Anemi	0.0662

Yöntemler, veri setinden herhangi bir değişken çıkarılmadan uygulanmıştır. Öncelikle veri setinin %33'ü test seti olarak, %67'si ise eğitim seti olarak ayrılmıştır. Yani eğitim seti için 200, test seti için 99 hasta ayrılmıştır. Random state, modeli her çalıştırdığımızda farklı sonuç almayı önlemek için kullanılmıştır. Bu aşamalardan sonra veri setine iki farklı yöntem uygulanmıştır.

İlk versiyonda, modele PCA [37] uygulanarak 12 bağımsız değişken 6 bileşene indirgenmiştir. Ardından Lojistik Regresyon [38], Naive Bayes [39], KNN [40], SVM [41], YSA [42], Karar Ağacı [43], RF [44], Gradient Boosting [45], XGBoost [46] ve CatBoost [47] sınıflandırıcı algoritmaları modele uygulanmıştır. Modele PCA uygulandıktan sonra RF özellik seçimine göre 6 bileşenin yüzde olarak önem düzeyi Şekil 3.2'deki gibidir.



Şekil 3.2 Bileşenlerin önem düzeyi



Şekil 3.3 Tüm değişkenlerin önem düzeyi

İkinci versiyonda Lojistik Regresyon, Naïve Bayes, KNN, SVM, YSA, Karar Ağacı, RF, Gradient Boost, XGBoost gibi sınıflandırıcı algoritmalar ve CatBoost algoritması kullanılmıştır. Şekil 3.3, RF özellik seçimi kullanılarak tüm değişkenlerin önem düzeyini yüzde olarak göstermektedir.

3.3 Sonuçlar

Sınıflandırma algoritmalarına iki versiyonda da hiperparametre optimizasyonu uygulanmıştır. Model için en uygun parametreler GridSearch CV [48] kullanılarak seçilir ve ardından modelin performansı ölçülmüştür.

Tablo 3.4 Modele PCA uygulandıktan sonra performans ölçümü

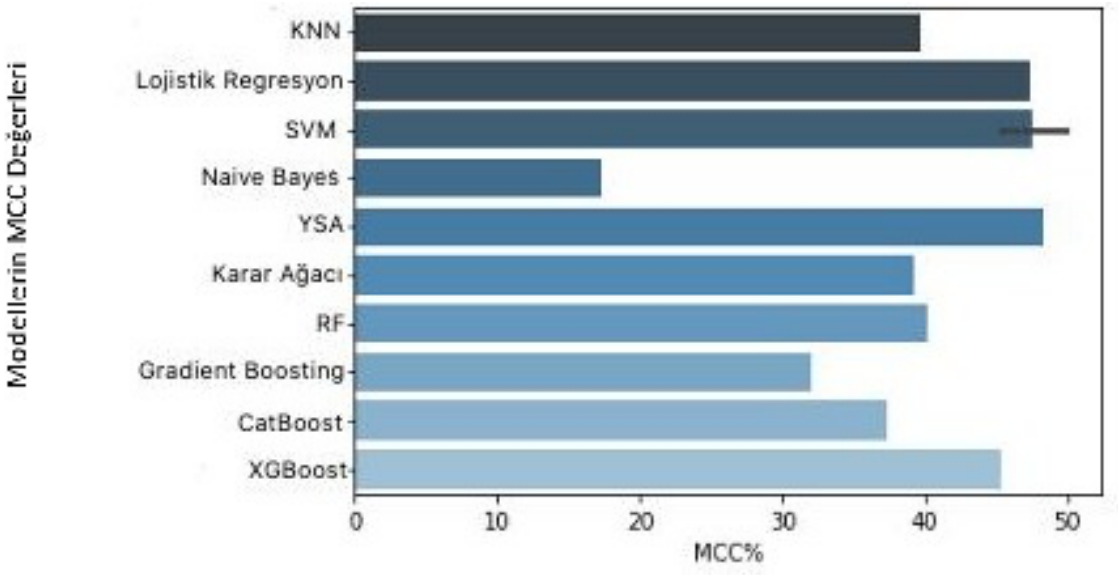
Yöntem	ACC	MCC	SE	SP	F_1 Score	ROC AUC	PR AUC
Lojistik Regresyon	0.7677	0.4739	0.5714	0.8750	0.6349	0.7232	0.5597
Naive Bayes	0.6465	0.1733	0.3429	0.8125	0.4068	0.5777	0.4038
KNN	0.7374	0.3969	0.4857	0.8750	0.5667	0.6804	0.5121
SVM Lineer	0.7777	0.4990	0.6000	0.8750	0.6563	0.7375	0.5759
SVM RBF	0.7576	0.4525	0.5714	0.8594	0.6250	0.7154	0.5456
YSA	0.7677	0.4827	0.6286	0.8437	0.6567	0.7361	0.5634
Karar Ağacı	0.7273	0.3924	0.5714	0.8125	0.5970	0.6920	0.5087
RF	0.7374	0.4011	0.5142	0.8594	0.5806	0.6868	0.5146
Gradient Boosting	0.6970	0.3208	0.5143	0.7969	0.5455	0.6556	0.4703
XGBoost	0.7576	0.4525	0.5714	0.8594	0.6250	0.7154	0.5456
CatBoost	0.7172	0.3735	0.5714	0.7969	0.5882	0.6842	0.4978

Başarıyı ölçmek için doğruluk oranı (acccuray=ACC), Matthews korelasyon katsayısı (Matthews Correlation Coefficient=MCC) [49], duyarlılık (sensitivity=SE), özgüllük (specifity=SP), F_1 score, ROC eğrisi ve eğrinin altında kalan alan (ROC AUC) and PR eğrisi ve eğrinin altında kalan alan (PR AUC) değerleri hesaplanmıştır.

Lojistik Regresyon algoritmasında solver olarak liblinear seçilmiştir. SVM'de kernel olarak lineer ve RBF seçilmiş ve iki farklı sonuç karşılaştırılmıştır. GridSearch CV, Naive Bayes ve Lojistik Regresyon dışındaki tüm modeller için en iyi parametreleri bulmak için kullanılmıştır.

Tablo 3.4, modele PCA uygulandıktan sonra yöntemlerin performans ölçümlerini göstermektedir.

GridSearch CV optimizasyonu yapıldıktan sonra, KNN modeli için n_neighbors= 3, SVM Lineer modeli için $C = 0.1$, SVM RBF modeli için $C = 50$ and gamma=0.01, YSA modeli için activation='relu', alpha=0.01, hidden_layer_sizes=(30, 30) ve solver='adam', Karar Ağacı modeli için max_depth=6, min_samples_split=5, Random Forest modeli için max_depth = 2, max_features=5, n_estimators=1000, min_samples_split=5, Gradient Boosting modeli için learning_rate=0.01, n_estimators=500, max_depth=5, min_samples_split=5, XGBoost modeli için learning_rate=0.01, n_estimators=100, max_depth=3, min_samples_split=2, subsample=0.6, CatBoost modeli için learning_rate=0.1, iterations=200, depth=5 değerleri seçilmiştir.



Şekil 3.4 Modellerin MCC değerleri

SVM Lineer modeli ACC, MCC, ROC AUC, PR AUC metriklerinde en başarılı sonucu veren yöntemdir. YSA modeli sensitivity, F_1 score metriklerinde en başarılı sonucu vermiştir. Tablo 3.4 de gösterildiği üzere Logistik Regresyon, KNN ve SVM Lineer

modellerinde specifity değeri aynı sonucu vermiştir. Tüm modeller için MCC değerleri Şekil 3.4'te gösterilmiştir. SVM Lineer modeli en yüksek MCC değerine sahiptir.

İkinci versiyonda yöntemler, modele PCA uygulanmadan 12 bağımsız değişken üzerinde uygulanmıştır. Bu versiyonda, XGBoost modeli için herhangi bir hiperparametre optimizasyonu yapılmamıştır. Hiperparametre optimizasyonunun olmaması, hiperparametre ayarı kullanılmadığında sonuçların XGBoost modeli için daha başarılı olmasından kaynaklanmaktadır.

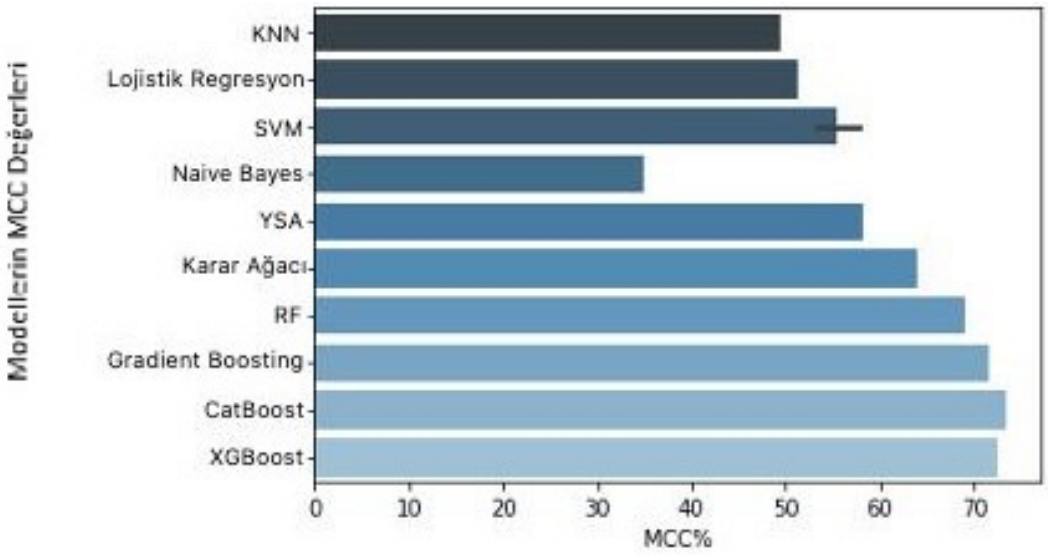
GridSearch CV optimizasyonu uygulandıktan sonra, KNN modeli için n_neighbors= 7, SVM Lineer modeli için $C = 6.0$, SVM RBF modeli için $C = 50$ and gamma=0.01, YSA modeli için activation='relu', alpha=0.005, hidden_layer_sizes=(50,) ve solver='adam', Karar Ağacı modeli için max_depth= 4, min_samples_split=16, RF modeli için max_depth=10, max_features=2, n_estimators=1000, min_samples_split=10, Gradient Boosting modeli için learning_rate=0.01, n_estimators=500, max_depth=3, min_samples_split=10, CatBoost modeli için learning_rate=0.01, learning_rate=0.1, iterations=200, depth=5 değerleri seçilmiştir.

Tablo 3.5'de, ikinci versiyon için performans ölçüm sonuçları gösterilmiştir. CatBoost sınıflandırma yöntemi ACC, MCC, F_1 Score, and PR AUC metriklerinde en başarılı sonuçları vermiştir. Öte yandan en başarılı sensitivity sonucu veren yöntem Gradient Boosting, en başarılı specifity sonucu veren yöntem KNN ve en başarılı ROC AUC sonucu veren yöntem XGBoost olmuştur.

Tablo 3.5 Modele PCA uygulanmadan performans ölçümü

Yöntem	ACC	MCC	SE	SP	F_1 Score	ROC AUC	PR AUC
Lojistik Regresyon	0.7778	0.5139	0.6857	0.8281	0.6857	0.7569	0.5813
Naive Bayes	0.7172	0.3483	0.4571	0.8594	0.5333	0.6583	0.4845
KNN	0.7778	0.4949	0.4857	0.9375	0.6071	0.7116	0.5750
SVM Lineer	0.7879	0.5330	0.6857	0.8438	0.6957	0.7647	0.5951
SVM RBF	0.8081	0.5775	0.7143	0.8594	0.7246	0.7868	0.6262
YSA	0.8081	0.5830	0.7429	0.8438	0.7324	0.7933	0.6274
Karar Ağacı	0.8182	0.6403	0.8857	0.7813	0.7750	0.8335	0.6506
RF	0.8586	0.6906	0.8000	0.8906	0.8000	0.8453	0.7107
Gradient Boosting	0.8586	0.7160	0.9143	0.8281	0.8205	0.8712	0.7107
XGBoost	0.8687	0.7259	0.8857	0.8593	0.8267	0.8725	0.7268
CatBoost	0.8788	0.7348	0.8286	0.9063	0.8286	0.8674	0.7471

Şekil 3.5, modelin MCC değerlerini göstermektedir. Buna göre en yüksek MCC değerinin CatBoost yöntemine ait olduğu, bunu XGBoost ve Gradient Boosting sınıflandırma yöntemlerinin takip ettiği görülmektedir.



Şekil 3.5 Modellerin PCA uygulanmadan MCC değerleri

İki versiyonu karşılaştırdığımızda, modele PCA uygulanan versiyonun tüm performans ölçüm metriklerinde önemli ölçüde daha düşük başarılı sonuçlar elde ettiğini görüyoruz.

İkinci versiyona derinlemesine baktığımızda ise RF, Gradient Boosting, XGBoost, CatBoost gibi ağaç tabanlı makine öğrenme yöntemlerinin diğer yöntemlere göre daha başarılı sonuçlar verdiğini görebiliyoruz. Bu yöntemler arasında CatBoost, performans ölçüm metriklerinin çoğu için en başarılı yöntemdir.

Tablo 3.6, Chicco ve Jurman'ın çalışmasında [3] kullanılan bu veri seti için en etkili makine öğrenimi yaklaşımı olan Lojistik Regresyon ile bu çalışmada en başarılı sonucu elde eden CatBoost yönteminin bir karşılaştırmasını göstermektedir.

Tablo 3.6 İki çalışma için en başarılı iki yöntemin karşılaştırılması

Yöntem	ACC	MCC	SE	SP	F_1 Score	ROC AUC	PR AUC
Logistik Regresyon	0.8380	0.6160	0.7850	0.8600	0.7190	0.8220	0.6170
CatBoost	0.8788	0.7348	0.8286	0.9063	0.8286	0.8674	0.7471

Lojistik Regresyon, Chicco ve Jurman'ın çalışmasında en iyi sonuçları vermiştir. Ejeksiyon fraksiyonu, serum kreatinin ve time değişkenlerini analiz etmek için Lojistik Regresyon kullanmışlardır ve 100 kez çalıştırdıktan sonra sonuçların ortalamasını almışlardır. Bu yöntemi kullanırken veri setini ölçeklendirmişlerdir ve veri setini %80 eğitim seti %20 test seti olarak bölmüşlerdir.

Bu çalışmada veri seti Standard Scaler kullanılarak ölçeklendirilmiş ve veri setinin %67'si eğitim seti, %33'ü test seti olacak şekilde bölünmüştür. Veri seti çok küçük olduğu için eğitim ve test seti bu şekilde bölünerek veri setinin dengesiz dağılımını önlemek istiyoruz. Bu çalışmada Chicco ve Jurman'ın çalışmasından farklı olarak tüm değişkenler üzerinde makine öğrenmesi algoritmaları kullanılmıştır.

İki sonuç karşılaştırıldığında, CatBoost yönteminin her bir performans metriği için daha iyi sonuçlar verdiği açıktır. Örneğin, Chicco ve Jurman'ın çalışmasında doğruluk 0.8380 iken bu çalışmada bu çalışma sonucu 0.8788 ve Chicco ve Jurman'ın çalışmasında MCC değeri 0.6160 iken bu çalışmada sonucu 0.7348'dir.

Sonuç olarak, Chicco ve Jurman'ın çalışmasının aksine, performans ölçüm metrikleri, veri setini ayırma, değişkenleri seçme ve CatBoost makine öğrenimi algoritmasını kullanma sürecimizin başarıyı arttırdığını açıkça göstermektedir.

4
SONUÇ VE ÖNERİLER

Kalp yetmezliği, dünya genelinde yaklaşık 26 milyon kişiyi etkileyen ciddi bir sağlık sorunudur. Diğer kronik rahatsızlıkları olan kişilerle ve genel nüfusla karşılaştırıldığında, kalp yetmezliği olan kişilerin yaşam kalitesi çok daha kötüdür [50]. KY tedavisindeki ilerlemelere rağmen morbidite ve mortalite oranı hala çok yüksektir. KY ile hastaneye yatış sonrası ölüm oranları 30 gün sonra %10 ve bir yıl sonra %22 olarak bildirilmektedir [51]. Kalp yetmezliği hastalığına yakalanma sayısı ve kalp yetmezliğinden ölüm oranı her geçen gün artmaktadır.

Makine öğrenimi, kalp yetmezliği olan hastaların ölüm durumunu tahmin edebilir. Bu çalışma, çeşitli değişkenlere bakarak hastanın ölüm durumunu belirlemeye yardımcı olmayı amaçlamaktadır. Mortalite durumuna göre uygun tedavi seçilerek hastanın durumu değiştirilebilir.

KAYNAKÇA

[1] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, ve M. A. Raza, "Survival analysis of heart failure patients: A case study," *PLOS ONE*, cilt 12, sayı 7, 2017.

[2] F. M. Zahid, S. Ramzan, S. Faisal, ve I. Hussain, "Gender based survival prediction models for heart failure patients: A case study in Pakistan," *PLOS ONE*, cilt 14, sayı 2, 2019.

[3] D. Chicco ve G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, cilt 20, sayı 1, 2020.

[4] I. H. Sarker, "Data Science and Analytics: An overview from data-driven smart computing, decision-making and applications perspective," *SN Computer Science*, cilt 2, sayı 5, 2021.

[5] B. Lu, D. Posner, J. L. Vassy, Y.-L. Ho, A. Galloway, S. Raghavan, J. Honerlaw, L. Tarko, J. Russo, S. Qazi, A. R. Orkaby, V. Tanukonda, L. Djousse, J. M. Gaziano, D. R. Gagnon, K. Cho, ve P. W. F. Wilson, "Prediction of cardiovascular and all-cause mortality after myocardial infarction in US veterans," *The American Journal of Cardiology*, cilt 169, ss. 10–17, 2022.

[6] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, M. S. V. Elkind, K. R. Evenson, C. Eze-Nliam, J. F. Ferguson, G. Generoso, J. E. Ho, R. Kalani, S. S. Khan, B. M. Kissela, K. L. Knutson, D. A. Levine, T. T. Lewis, J. Liu, M. S. Loop, J. Ma, M. E. Mussolino, S. D. Navaneethan, A. M. Perak, R. Poudel, M. Rezk-Hanna, G. A. Roth, E. B. Schroeder, S. H. Shah, E. L. Thacker, L. B. VanWagner, S. S. Virani, J. H. Voecks, N.-Y. Wang, K. Yaffe, and S. S. Martin, "Heart disease and stroke statistics—2022 update: A report from the American Heart Association," *Circulation*, cilt 145, sayı 8, 2022.

[7] J. H. Morris ve L. Chen, "Exercise training and heart failure: A review of the literature," *Cardiac Failure Review*, cilt 5, sayı 1, ss. 57–61, 2019.

[8] A. L. Bui, T. B. Horwich, ve G. C. Fonarow, "Epidemiology and risk profile of heart failure," *Nature Reviews Cardiology*, cilt 8, sayı 1, ss. 30–41, 2010.

[9] J. J. McMurray, S. Adamopoulos, S. D. Anker, A. Auricchio, M. Bohm, K. Dickstein, V. Falk, G. Filippatos, C. Fonseca, M. A. Gomez-Sanchez, T. Jaarsma, L. Kober, G. Y. Lip, A. P. Maggioni, A. Parkhomenko, B. M. Pieske, B. A. Popescu, P. K. Ronnevik, F. H. Rutten, J. Schwitter, P. Seferovic, J. Stepinska, P. T. Trindade, A. A. Voors, F. Zannad, A. Zeiher, J. J. Bax, H. Baumgartner, C. Ceconi, V. Dean, C. Deaton, R. Fagard, C. Funck-Brentano, D. Hasdai, A. Hoes, P. Kirchhof, J. Knuuti, P. Kolh, T. McDonagh, C. Moulin, B. A. Popescu, Z. Reiner, U. Sechtem, P. A. Sirnes, M. Tendera, A. Torbicki, A. Vahanian, S. Windecker, T. McDonagh, U. Sechtem, L. A. Bonet, P. Avraamides, H. A. Ben Lamin, M. Brignole, A. Coca, P. Cowburn, H. Dargie, P. Elliott, F. A. Flachskampf, G. F. Guida, S. Hardman, B. Iung, B. Merkely, C. Mueller, J. N. Nanas, O. W. Nielsen, S. Orn, J. T. Parissis, ve P. Ponikowski, "ESC guidelines for the diagnosis and treatment of acute and

chronic heart failure 2012: The Task Force for the diagnosis and treatment of acute and chronic heart failure 2012 of the European Society of Cardiology. developed in collaboration with the Heart Failure Association (HFA) of the ESC," *European Heart Journal*, cilt 33, sayı 14, ss. 1787–1847, 2012.

[10] F. F. Gong, M. V. Jelinek, J. M. Castro, J. M. Coller, M. McGrady, U. Boffa, L. Shiel, D. Liew, R. Wolfe, S. Stewart, A. J. Owen, H. Krum, C. M. Reid, D. L. Prior, ve D. J. Campbell, "Risk factors for incident heart failure with preserved or reduced ejection fraction, and valvular heart failure, in a community-based cohort," *Open Heart*, cilt 5, sayı 2, 2018.

[11] A. Inamdar ve A. Inamdar, "Heart failure: Diagnosis, management and Utilization," *Journal of Clinical Medicine*, cilt 5, sayı 7, p. 62, 2016.

[12] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, ve F. Herrera, "Big data preprocessing: Methods and prospects," *Big Data Analytics*, cilt 1, sayı 1, 2016.

[13] F.J.W.M. Dankers, A. Traverso, L. Wee, S.M.J. van Kuijk, "Prediction Modeling Methodology". P. Kubben, M. Dumontier, A. Dekker (eds) in *Fundamentals of Clinical Data Science.* Springer, Cham. 2019.

[14] D. M. Atallah, M. Badawy, A. El-Sayed, ve M. A. Ghoneim, "Predicting kidney transplantation outcome based on hybrid feature selection and Knn Classifier," *Multimedia Tools and Applications*, cilt 78, sayı 14, ss. 20383–20407, 2019.

[15] E. García-Gonzalo, Z. Fernández-Muñiz, P. García Nieto, A. Bernardo Sánchez, ve M. Menéndez Fernández, "Hard-rock stability analysis for span design in entry-type excavations with learning classifiers," *Materials*, cilt 9, sayı 7, p. 531, 2016.

[16] J. L. Rojo-Álvarez, M. Martínez-Ramón, J. Muñoz-Marí, ve G. Camps-Valls, "Support Vector Machine and Kernel Classification Algorithms," in *Digital signal processing with kernel methods*, New York, NY: John Wiley et Sons, 2018, ss. 433–502.

[17] M. Akar, N.M. Sirakov, M. Mete, "Clifford Algebra Multivectors and Kernels for Melanoma Classification", *Mathematical Methods in the Applied Sciences,* cilt 45, sayı 7, ss. 4056-4068, 2022.

[18] M. Akar, N.M. Sirakov, "Support Vector Machine Skin Lesion Classification in Clifford Algebra Subspaces", *Applications of Mathematics,* cilt 64, sayı 5, ss. 581 598, 2019.

[19] Ö. Korkmaz, A. Mahiroğlu. "Beyin, bellek ve öğrenme", *Kastamonu Eğitim Dergisi*, cilt 15, sayı 1, ss. 93-104, 2007.

[20] K. Öztürk, M. E. Şahin, "Yapay sinir ağları ve yapay zekâ'ya genel bir bakış", *Takvim-i Vekayi*, cilt 6, sayı 2, ss. 25-36, 2018.

[21] Ş. Can, M. Gerşil, "Manisa Pamuk Fiyatlarının Zaman Serisi Analizi ve Yapay Sinir Ağı teknikleri ̇ile Tahminlenmesi ve Tahmin Performanslarının

karşılaştırılması," *Yönetim ve Ekonomi: Celal Bayar Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, cilt 25, sayı 3, ss. 1017–1031, 2018.

[22] A. B. Varol, İ. İşeri, "Lenf Kanserine İlişkin Patoloji Görüntülerinin Makine Öğrenimi Yöntemleri ile Sınıflandırılması", *Avrupa Bilim ve Teknoloji Dergisi*, ss. 404-410, 2019.

[23] S. ÖZDEMİR, "Potential distribution modelling and mapping using random forest method: An example of Yukarigökdere District," *Turkish Journal of Forestry | Türkiye Ormancılık Dergisi*, cilt 19, sayı 1, ss. 51–56, 2018.

[24] Ş. İ. Serengil, "Gradient boosting vs Rastgele Orman," *Bilisim IO*, 09-Ocak-2022. [Online]. Available: https://bilisim.io/2022/01/08/gradient-boosting-vs-rastgele-orman/. [Erişim zamanı: 16-Jun-2022].

[25] E. R. Muratlar, "XGBoost Nasıl çalışır? neden˙iyi performans gösterir?," *Veri Bilimi Okulu*, 24-Mar-2020. [Online]. Available: https://www.veribilimiokulu.com/xgboost-nasil-calisir/. [Erişim zamanı: 16-Jun-2022].

[26] E. R. Muratlar, "Catboost Nedir? Diğer Boosting Algoritmalarından Farkı Nelerdir?," *Veri Bilimi Okulu*, 08-Ağustos-2020. [Online]. Available: https://www.veribilimiokulu.com/catboost-nedir-diger-boosting-algoritmalarindan-farki-nelerdir/. [Erişim zamanı: 16-Jun-2022].

[27] Y. Sener, "Veri Bilimi SINIFLANDIRMA model çıktılarını Değerlendiren metrikler (confusion matrix, accuracy...," *Medium*, 14-Ağustos-2020. [Online]. Available: https://yigitsener.medium.com/veri-bilimi-sınıflandırma-model-çıktılarını-değerlendiren-metrikler-confusion-matrix-accuracy-437f5633c82b. [Erişim zamanı: 26-Jun-2022].

[28] G. Ogundur, "Roc ve Auc," *Medium*, 12-Ocak-2020. [Online]. Available: https://medium.com/@gulcanogundur/roc-ve-auc-1fefcfc71a14. [Erişim zamanı: 26-Jun-2022].

[29] M. Noyan, "Regresyon ve Sınıflandırmada Hata metrikleri," *Medium*, 06-Ekim-2020. [Online]. Available: https://medium.com/deep-learning-turkiye/regresyon-ve-s%C4%B1n%C4%B1fland%C4%B1rmada-hata-metrikleri-143a40c6b656. [Erişim zamanı: 26-Jun-2022].

[30] R.S. Aujla, R. Patel, *Creatine Phosphokinase*. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2022.

[31] M. Metra, G. Cotter, M. Gheorghiade, L. Dei Cas, ve A. A. Voors, "The role of the kidney in heart failure," *European Heart Journal*, cilt 33, sayı 17, ss. 2135–2142, 2012.

[32] M. K. Mojadidi, J. N. Galeas, D. Goodman-Meza, P. Eshtehardi, P. Msaouel, I. Kelesidis, M. O. Zaman, J. S. Winoker, S. C. Roberts, P. Christia, ve R. Zolty, "Thrombocytopaenia as a prognostic indicator in heart failure with reduced ejection fraction," *Heart, Lung and Circulation*, cilt 25, sayı 6, ss. 568–575, 2016.

[33] M. Metra, G. Cotter, M. Gheorghiade, L. Dei Cas, ve A. A. Voors, "The role of the kidney in heart failure," *European Heart Journal*, cilt 33, sayı 17, ss. 2135–2142, 2012.

[34] D. Bamira ve M. H. Picard, "Imaging: Echocardiology—assessment of cardiac structure and function," *Encyclopedia of Cardiovascular Research and Medicine*, ss. 35–54, 2018.

[35] Y. Patel ve J. Joseph, "Sodium intake and heart failure," *International Journal of Molecular Sciences*, cilt 21, sayı 24, ss. 9474, 2020.

[36] J. Benesty, J. Chen, Y. Huang, I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*, cilt 2, Berlin: Springer, 2009, ss. 1–4.

[37] S. Wold, K. Esbensen, ve P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, cilt 2, sayı 1-3, ss. 37–52, 1987.

[38] RE. Fan, KW. Chang, CJ. Hsieh, XR. Wang, CJ. Lin, "LIBLINEAR: A Library for Large Linear Classification," *The Journal of Machine Learning Research*. cilt 9, ss. 1871–1874, 2008.

[39] I. Rish, "An empirical study of the naive Bayes classifier.", *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*. cilt 3, ss. 41-46, 2001.

[40] T. Cover ve P. Hart, "Nearest neighbor Pattern Classification," *IEEE Transactions on Information Theory*, cilt 13, sayı 1, ss. 21–27, 1967.

[41] S. Amari ve S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, cilt 12, sayı 6, ss. 783–789, 1999.

[42] Y. LeCun, Y. Bengio, ve G. Hinton, "Deep learning," *Nature*, cilt 521, sayı 7553, ss. 436–444, 2015.

[42] W. Y. Loh, "Classification and regression trees," *WIREs Data Mining and Knowledge Discovery*, cilt 1, sayı 1, ss. 14–23, 2011.

[44] L. Breiman, *Machine Learning*, cilt 45, sayı 1, ss. 5–32, 2001.

[45] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, cilt 29, sayı 5, 2001.

[46] T. Chen ve C. Guestrin, "XgBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ss. 785–794, 2016.

[47] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, "CatBoost: unbiased boosting with categorical features," *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. ss. 6639–6649,2018.

[48] J. Bergstra, Y. Bengio, "Random Search for Hyper-Parameter Optimization", *Journal of Machine Learning Research*, cilt 13, sayı 10, ss. 281–305, 2012.

[49] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta (BBA) - Protein Structure*, cilt 405, sayı 2, ss. 442–451, 1975.

[50] S. Heo, T. A. Lennie, C. Okoli, ve D. K. Moser, “Quality of life in patients with heart failure: Ask the patients,” *Heart & Lung*, cilt 38, sayı 2, ss. 100–108, 2009.

[51] Y. Cavusoglu, M. Zoghi, M. Eren, et al. “Post-discharge heart failure monitoring program in Turkey: Hit-point,” *The Anatolian Journal of Cardiology*, cilt 17, sayı 2, ss. 107–112, 2016.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. S. T. Badik, M. Akar, “Machine Learning Classification Model for Heart Failure Patients,” in *5th International Online Conference on Mathematics*, 2021.