T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KALP HASTALIKLARININ ERKEN TEŞHİSİ İÇİN MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRMALI PERFORMANS ANALİZİ

Büşra ŞENEL KAHYAOĞLU

YÜKSEK LİSANS TEZİ

Matematik Mühendisliği Anabilim Dalı

Matematik Mühendisliği Programı

Danışman

Doç. Dr. Kadiriye ŞİMŞEK ALAN

Şubat, 2026

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KALP HASTALIKLARININ ERKEN TEŞHİSİ İÇİN MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRILMALI PERFORMANS ANALİZİ

Büşra ŞENEL KAHYAOĞLU tarafından hazırlanan tez çalışması 02.02.2026 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı, Matematik Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Kadiriye ŞİMŞEK ALAN
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Kadiriye ŞİMŞEK ALAN, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Ünsal TEKİR, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Serkan ONAR, Üye
Yıldız Teknik Üniversitesi _____

Danışmanım Doç. Dr. Kadiriye ŞİMŞEK ALAN sorumluluğunda tarafımca hazırlanan "Kalp Hastalıklarının Erken Teşhisi için Makine Öğrenmesi Algoritmalarının Karşılaştırılmalı Performans 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.

Büşra ŞENEL KAHYAOĞLU

İmza

Aileme

ve

eşime

TEŞEKKÜR

Öncelikle, bu araştırma süresince bana sağladığı paha biçilmez rehberlik, sürekli destek ve sabır için tez danışmanım Doç. Dr. Kadiriye ŞİMŞEK ALAN' a en içten şükranlarımı sunmak istiyorum. Makine öğrenmesi ve veri bilimi alanındaki uzmanlığı bu çalışmanın şekillenmesinde büyük rol oynamış, yapıcı geri bildirimleri çalışmanın kalitesini önemli ölçüde artırmıştır.

Sağlam bir akademik temel sağladıkları ve entelektüel açıdan teşvik edici bir ortam yarattıkları için Matematik Mühendisliği Bölümü öğretim üyelerine minnettarım. Ders dönemimde edindiğim bilgi ve beceriler, bu araştırmayı tamamlamam için vazgeçilmez olmuştur.

Bu yolculuk boyunca manevi destek ve değerli tartışmalar sağlayan meslektaşlarıma ve arkadaşlarıma özel teşekkürlerimi sunarım.

Heart Disease veri setini kamuya açık hale getirerek bu karşılaştırmalı analizi mümkün kılan UCI Makine Öğrenmesi Deposu'na teşekkür ederim. Kaliteli veri setlerine açık erişim, tıbbi bilişim alanındaki araştırmaların ilerlemesi için büyük önem taşımaktadır.

Son olarak, koşulsuz sevgileri, anlayışları ve sonsuz destekleri için aileme ve özellikle eşime en derin şükranlarımı borçluyum. Yeteneklerime olan inançları ve uzun çalışma saatlerindeki sabırları, en büyük güç kaynağım olmuştur. Bu tez, onların teşviki olmadan mümkün olamazdı.

Büşra ŞENEL KAHYAOĞLU

İÇİNDEKİLER

KISALTMA LİSTESİ ..... **viii**
ŞEKİL LİSTESİ ..... **ix**
TABLO LİSTESİ ..... **x**
ÖZET ..... **xi**
ABSTRACT ..... **iv**
1.GİRİŞ **1**
1.1. Çalışmanın Amacı ve Önemi 1
1.1.2. Çalışmanın Bilimsel ve Pratik Önemi 1
1.1.3. Çalışmanın Genel Akışı 2
1.2. Literatür Taraması 4
1.2.1. Uluslararası Çalışmalar 4
1.2.2. Ulusal Çalışmalar 5
1.2.3. Literatürdeki Boşluklar ve Bu Çalışmanın Katkısı 5
1.3. Problem Tanımı ve Araştırma Soruları 6
1.3.1. Araştırma Hipotezleri 6
1.4. Çalışmanın Kapsamı ve Sınırlılıkları 7
1.4.1. Çalışmanın Kapsamı 7
1.4.2. Sınırlılıklar 7
1.5. Tezin Organizasyonu 8
2.KURAMSAL TEMELLER **9**
2.1. Makine Öğrenmesi (Machine Learning) 9
2.1.1. Makine Öğrenmesi Tanımı 9
2.1.2. Makine Öğrenmesi Türleri 10
2.2. Makine Öğrenmesi Algoritmaları 11
2.2.1. Lojistik Regresyon (Logistic Regression) 12
2.2.2. K-En Yakın Komşu (K-Nearest Neighbors) 13
2.2.3. Destek Vektör Makineleri (SVM) 13
2.2.4. Rastgele Orman (Random Forest) 14
2.2.5. XGBoost (Extreme Gradient Boosting) 14
2.2.6. LightGBM (Light Gradient Boosting Machine) 15

2.3. Veri Madenciliği (Data Mining) 15
2.3.1. Temel Kavramlar ve Tanımlar 15
2.3.2. Tarihsel Gelişim 16
2.3.3. Bilgi Keşif Süreci 16
2.3.4. Uygulama Alanları 17
2.4. Performans Değerlendirme Metrikleri 17
2.4.1. Karışıklık Matrisi (Confusion Matrix) 17
2.4.2. Doğruluk (Accuracy) 18
2.4.3. Kesinlik (Precision) 18
2.4.4. Duyarlılık (Recall) 19
2.4.5. F1 Skoru (F1 Score) 19
2.4.6. ROC Eğrisi ve AUC 20
3. MATERYAL VE YÖNTEM 21
3.1. Veri Seti: Heart Disease UCI (Cleveland) 21
3.1.1. Veri Setinin Genel Özellikleri 21
3.1.2. Özniteliklerin Detaylı Açıklaması 23
3.2. Kalp Hastalıklarının Önemi ve Teşhisi 24
3.2.1. Kalp Hastalıkları Hakkında Genel Bilgi 25
3.2.2. Risk Faktörleri 25
3.2.3. İstatistiksel Verilerle Kalp Hastalıkları 26
3.3. Veri Setine Uygulanan Ön İşlemler 27
3.3.1. Eksik Veri Analizi ve Doldurma Yöntemleri 27
3.3.2. Veri Dengesizliği ve SMOTE Yöntemi 28
3.3.3. Öznitelik Ölçeklendirme 30
3.4. Kullanılan Araçlar ve Kütüphaneler 30
3.4.1. Veri İşleme Kütüphaneleri 30
3.4.2. Görselleştirme Kütüphaneleri 31
3.4.3. Makine Öğrenmesi Kütüphaneleri 31
3.4.4. Geliştirme Ortamı ve Donanım 32
3.4.5. Veri Bölme Stratejisi 32
4. UYGULAMA VE BULGULAR 34
4.1. Keşifsel Veri Analizi (Exploratory Data Analysis) 34
4.1.1. Yaş Dağılımı Analizi 34
4.1.2. Cinsiyet Dağılımı Analizi 35
4.1.3. Cinsiyete Göre Hastalık Durumu 36

4.1.4. Maksimum Kalp Atış Hızı ve Yaş İlişkisi 37
4.1.5. ST Segment Eğimi Analizi 37
4.1.6. Açlık Kan Şekeri ve Hastalık İlişkisi 37
4.1.7. Göğüs Ağrısı Tipine Göre Hastalık Sıklığı 38
4.1.8. Thalassemia Analizi 38
4.2. Model Eğitimi ve Sonuçları 40
4.2.1. XGBoost Sonuçları 40
4.2.2. KNN Sonuçları 41
4.2.3. SVM Sonuçları 41
4.2.4. Lojistik Regresyon Sonuçları 41
4.2.5. LightGBM Sonuçları 41
4.2.6. Random Forest Sonuçları 41
4.3. Algoritmaların Karşılaştırmalı Analizi 42
5. SONUÇ 44
5.1. Sonuçların Değerlendirilmesi 44
5.1.1. Algoritmaların Çok Boyutlu Değerlendirmesi 46
5.1.2. F1 Skoru ve Roc Eğrisi Analizi 47
5.1.3. Algoritma Bazlı Özet Değerlendirmeler 49
5.2. Kısıtlamalar 50
5.3. Gelecek Çalışmalar İçin Öneriler 51
5.4. Genel Sonuç 52
KAYNAKÇA 53
EKLER 55
EK-A: Veri Seti Detaylı Bilgileri 55
EK-B: Model Hiperparametreleri 56
EK-C: Ek Analiz Görselleri 58
EK-D: Kullanılan Yazılım ve Sürümler 58
TEZDEN ÜRETİLMİŞ YAYINLAR 60

KISALTMA LİSTESİ

AI	Artificial Intelligence (Yapay Zekâ)
AUC	Area Under the Curve (Eğri Altındaki Alan)
CVD	Cardiovascular Disease (Kardiyovasküler Hastalık)
EDA	Exploratory Data Analysis (Keşifsel Veri Analizi)
EFB	Exclusive Feature Bundling
EHR	Electronic Health Records (Elektronik Sağlık Kayıtları)
FN	False Negative (Yanlış Negatif)
FP	False Positive (Yanlış Pozitif)
FPR	False Positive Rate (Yanlış Pozitif Oranı)
GOSS	Gradient-based One-Side Sampling
KDD	Knowledge Discovery in Databases (Veritabanlarından Bilgi Keşfi)
KNN	K-Nearest Neighbors (K-En Yakın Komşu)
LightGBM	Light Gradient Boosting Machine
ML	Machine Learning (Makine Öğrenmesi)
RBF	Radial Basis Function (Radyal Tabanlı Fonksiyon)
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machines (Destek Vektör Makineleri)
TN	True Negative (Doğru Negatif)
TP	True Positive (Doğru Pozitif)
TPR	True Positive Rate (Doğru Pozitif Oranı)
TÜİK	Türkiye İstatistik Kurumu
UCI	University of California, Irvine
WHO	World Health Organization (Dünya Sağlık Örgütü)
XGBoost	Extreme Gradient Boosting

ŞEKİL LİSTESİ

Şekil 1. 1 Tez Çalışması Akış Şeması **3**
Şekil 1. 2 Literatürdeki Çalışmalarla Performans Karşılaştırılması **4**
Şekil 2. 1 Gözetimli Öğrenme Sürecinin Şematik Gösterimi **10**
Şekil 2. 2 Makine Öğrenmesi Türleri ve Temel Özellikleri **10**
Şekil 2. 3 Makine Öğrenmesi Algoritmalarının Özellik Karşılaştırması **12**
Şekil 2. 4 Veri tabanlarından Bilgi Keşfi (KDD) Süreci **16**
Şekil 2. 5 Karışıklık Matrisi ve Performans Metrikleri Formülleri **18**
Şekil 2. 6 F1 Gösterimi **19**
Şekil 2. 7 Farklı Performans Seviyelerinde ROC Eğrisi Örnekleri **20**
Şekil 3. 1 Veri Seti Öznitelik Dağılımları **22**
Şekil 3. 2 Özniteliklerdeki Eksik Veri Oranları **28**
Şekil 3. 3 SMOTE (Synthetic Minority Over-sampling Technique) Çalışma Prensibi **29**
Şekil 3. 4 Çalışmada Kullanılan Yazılım Araçları ve Kütüphaneler **30**
Şekil 4. 1 Veri Setindeki Hastaların Yaş Dağılımı Histogramı **34**
Şekil 4. 2 Veri Setindeki Cinsiyet Dağılımı **35**
Şekil 4. 3 Cinsiyete Göre Yaş Dağılımı **35**
Şekil 4. 4 Veritabanına Göre Cinsiyet Dağılımı **36**
Şekil 4. 5 Yaş ve Hastalık Şiddeti Dağılımı **37**
Şekil 4. 6 Göğüs Ağrısı Tipine Göre Hastalık Dağılımı **38**
Şekil 4. 7 Talyum Testi Sonuçlarının Cinsiyete Göre Dağılımı **39**
Şekil 4. 8 Talyum Testi Sonuçlarına Göre Hastalık Şiddeti Dağılımı **39**
Şekil 4. 9 Talyum Testi ve Göğüs Ağrısı Tipi İlişkisi **40**
Şekil 4. 10 Modellerin Doğruluk ve AUC Skorları Karşılaştırması **42**
Şekil 5. 1 Model Performanslarının Radar Grafiği ile Karşılaştırması **44**
Şekil 5. 2 Genel Model Performans Sıralaması **45**
Şekil 5. 3 Model Değerlendirme Matrisi (1: Zayıf,5:Güçlü) **46**
Şekil 5. 4 F1 Skoru ve ROC Eğrisi Analizi Grafiği **47**
Şekil EK-C.1 Yaş ve Talyum Testi Sonuçlarına Göre Hasta Dağılımı **58**
Şekil EK-C.2 Model Doğruluk ve AUC Skorlarının Detaylı Karşılaştırması **58**

TABLO LİSTESİ

Tablo 1.1 Liratürdeki Önemli Çalışmaların Özeti ……………………………….. **6**
Tablo 3.1 UCI Veri Setindeki Öznitelikler ve Açıklamaları ………………………… **23**
Tablo 4.1 Makine Öğrenmesi Modellerinin Performans Karşılaştırması ………. **42**
Tablo 5.1 Algoritmaların Detaylı Performans Karşılaştırması ……………………. **45**
Tablo EK-A.1 Veri Seti Genel İstatistikleri …………………………………………… **55**
Tablo EK-A.2 Özniteliklerin Detaylı Açıklamaları …………………………………. **55**
Tablo EK-A.3 Göğüs Ağrısı Tipi (cp) Açıklamaları ………………………………… **56**
Tablo EK-A.4 Talyum Stres Testi Açıklamaları …………………………………….. **56**
Tablo EK-B.1 XGBoost Hiperparametreleri …………………………………………. **56**
Tablo EK-B.2 SVM Hiperparametreleri ……………………………………………….. **56**
Tablo EK-B.3 LightGBM Hiperparametreleri ………………………………………… **57**
Tablo EK-B.4 Random Forest Hiperparametreleri ………………………………….. **57**
Tablo EK-B.5 KNN ve Lojistik Regresyon Hiperparametreleri ………………….. **57**
Tablo EK-D.1 Python Kütüphaneleri ve Sürümleri …………………………………. **58**

ÖZET

Kalp Hastalıklarının Erken Teşhisi İçin Makine Öğrenmesi Algoritmalarının Karşılaştırmalı Performans Analizi

Büşra ŞENEL KAHYAOĞLU

Matematik Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Kadiriye ŞİMŞEK ALAN

Günümüzde kalp ve damar rahatsızlıkları, küresel ölçekte en fazla can kaybına yol açan sağlık problemleri arasında yer almaktadır. Bu hastalıkların önceden tespit edilmesi, hastaların yaşam kalitesini yükseltmek ve ölüm riskini azaltmak bakımından büyük önem taşımaktadır. Bilgisayar destekli tanı sistemleri, bu alanda hekimlere önemli bir destek mekanizması sunmaktadır. Mevcut çalışmada, kardiyak rahatsızlıkların önceden belirlenmesine yönelik olarak farklı sınıflandırma tekniklerinin etkinliği incelenmiştir. Araştırmada Cleveland kliniği verilerini içeren ve uluslararası literatürde sıkça tercih edilen bir veri kümesi temel alınmıştır. Söz konusu veri kümesi, dört farklı sağlık merkezinden derlenen toplam 920 bireye ait klinik ölçümleri barındırmaktadır.

Araştırma kapsamında gradyan artırma tabanlı yöntemler (XGBoost, LightGBM), mesafe temelli sınıflandırıcılar (KNN), çekirdek fonksiyonlu ayırıcılar (SVM), olasılıksal modeller (Lojistik Regresyon) ve ağaç toplulukları (Random Forest) olmak üzere altı ayrı teknik değerlendirilmiştir. Veri hazırlama sürecinde kayıp gözlemler tahminleme yöntemleriyle tamamlanmış, sınıflar arası dengesizlik ise sentetik örnekleme ile giderilmiştir. Deneysel bulgular, çekirdek tabanlı sınıflandırıcının (SVM) %83 başarı oranı ve 0.88 eğri altı alan değeri ile öne çıktığını ortaya koymuştur. Ağaç topluluğu yöntemi ise 0.90 eğri altı alan skoru ile sınıflar arası ayrım gücünde en yüksek performansı sergilemiştir. Araştırma sonuçları, otomatik öğrenme yaklaşımlarının kardiyak tanıda güvenilir bir yardımcı araç olabileceğine işaret etmektedir.

Anahtar Kelimeler: Kalp hastalığı, makine öğrenmesi, sınıflandırma,

xgboost, svm

ABSTRACT

Comparatıve Performance Analysıs Of Machıne Learning Algorithms For Early Early Detection Of Heart Diseases

Büşra ŞENEL KAHYAOĞLU

Departman of Mathematical Engineering

Master's Thesis

Supervisor: Assoc Prof. Kadiriye ŞİMŞEK ALAN

Heart and circulatory system disorders represent a major public health challenge, accounting for a significant proportion of mortality cases globally. Timely identification of these conditions plays a crucial role in enhancing patient outcomes and reducing fatality rates. Computer-aided diagnostic tools have emerged as valuable instruments to assist medical professionals in this domain. The present research examines the effectiveness of various classification approaches for the preliminary detection of cardiac abnormalities. The investigation utilizes a widely-recognized clinical dataset compiled from four distinct medical institutions, encompassing diagnostic measurements from 920 individuals along with 14 clinical attributes.

Six distinct analytical techniques were evaluated in this study: gradient boosting frameworks (XGBoost, LightGBM), instance-based classifiers (KNN), kernel-based separators (SVM), probabilistic models (Logistic Regression), and tree ensemble methods (Random Forest). During data preparation, incomplete observations were addressed through predictive imputation, while inter-class disparity was mitigated using synthetic sampling strategies. Experimental outcomes indicate that the kernel-based classifier (SVM) achieved superior overall results with 83% classification accuracy and 0.88 area under the curve metric. The tree ensemble approach demonstrated the strongest discriminative capability, attaining a 0.90 area under the curve value. These findings suggest that automated learning methodologies can serve as reliable assistive tools in cardiac diagnosis.

Keywords: Heart disease, machine learning, classification, xgboost,

svm

1.GİRİŞ

1.1. Çalışmanın Amacı ve Önemi

Bu tez çalışmasının temel amacı, kalp hastalıklarının erken teşhisinde kullanılabilecek makine öğrenmesi algoritmalarının karşılaştırmalı performans analizini gerçekleştirmektir. Çalışmada, uluslararası literatürde yaygın olarak kullanılan ve güvenilirliği kanıtlanmış UCI Machine Learning Repository'den elde edilen Heart Disease veri seti kullanılarak altı farklı makine öğrenmesi algoritmasının performansları sistematik bir şekilde değerlendirilmektedir.

1.1.1. Araştırmanın Temel Hedefleri

Bu araştırmanın birincil hedefi, farklı makine öğrenmesi algoritmalarının kalp hastalığı teşhisindeki performanslarını objektif metrikler kullanarak karşılaştırmak ve en uygun modeli belirlemektir. Bu kapsamda değerlendirilen algoritmalar XGBoost, K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM), Lojistik Regresyon, LightGBM ve Rastgele Orman (Random Forest) olarak belirlenmiştir. Her bir algoritmanın güçlü ve zayıf yönleri, farklı performans metrikleri açısından analiz edilmektedir.

İkincil hedef, veri ön işleme tekniklerinin model performansı üzerindeki etkisini incelemektir. Bu bağlamda, eksik veri doldurma yöntemleri (KNN Imputer, Random Forest Regressor) ve sınıf dengesizliğini giderme teknikleri (SMOTE) uygulanarak bu işlemlerin tahmin doğruluğuna katkısı değerlendirilmektedir.

Üçüncül hedef, elde edilen bulguların klinik uygulamalara aktarılabilirliğini tartışmaktır. Makine öğrenmesi modellerinin sağlık sektöründe karar destek sistemi olarak nasıl konumlandırılabileceği, hangi koşullarda güvenilir sonuçlar üretebileceği ve klinisyenlerle iş birliği içinde nasıl kullanılabileceği sorularına yanıt aranmaktadır.

1.1.2. Çalışmanın Bilimsel ve Pratik Önemi

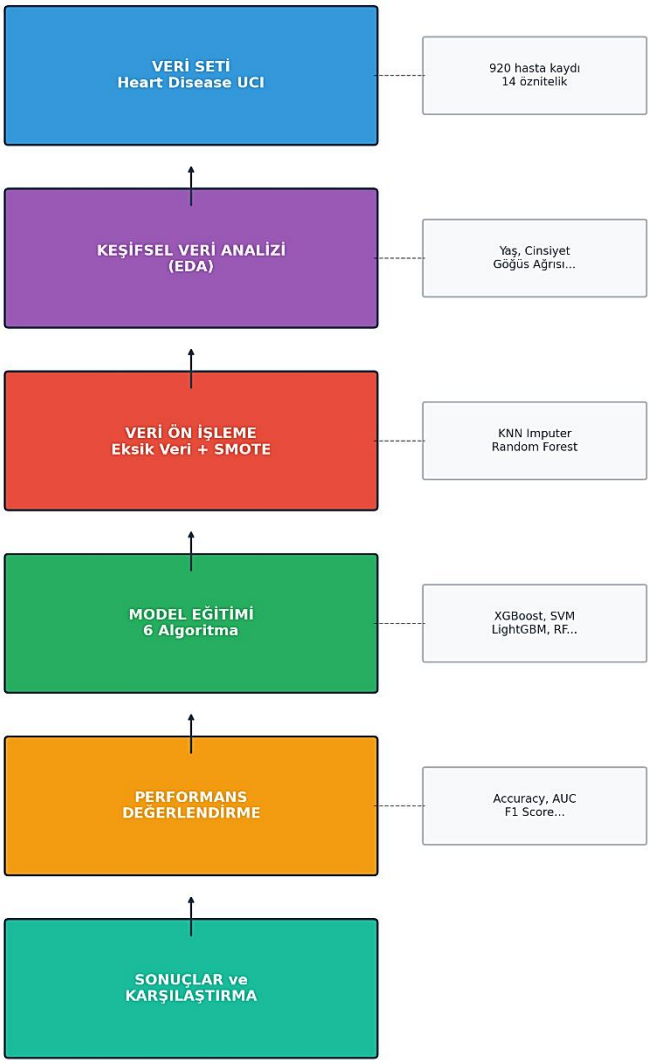
Bu çalışmanın bilimsel önemi, farklı makine öğrenmesi algoritmalarının aynı veri seti üzerinde sistematik bir karşılaştırmasını sunmasından kaynaklanmaktadır. Literatürde çeşitli algoritmalar ayrı ayrı çalışmalarda değerlendirilmiş olsada, bu tez çalışması altı farklı algoritmayı tutarlı bir metodoloji ve aynı değerlendirme kriterleri altında karşılaştırarak bütünleşik bir bakış açısı sunmaktadır.

Pratik açıdan değerlendirildiğinde, bu çalışmanın bulguları klinik karar destek sistemlerinin tasarımına katkı sağlayabilecek niteliktedir. Sağlık kuruluşlarında uygulanabilecek erken uyarı sistemleri, hekimlerin tanı sürecini destekleyerek hasta sonuçlarının iyileştirilmesine yardımcı olabilecektir. Özellikle birinci basamak sağlık hizmetlerinde, risk altındaki bireylerin hızlı ve doğru bir şekilde belirlenmesi büyük önem taşımaktadır. Ekonomik açıdan bakıldığında, erken teşhis sistemleri sağlık harcamalarının azaltılmasına da katkı sağlayabilecektir. Kalp krizinin önlenmesi veya erken müdahale edilmesi, yoğun bakım süreçleri, cerrahi müdahaleler ve uzun süreli rehabilitasyon programlarına kıyasla çok daha düşük maliyetlidir. Bu nedenle, etkili tarama ve erken teşhis sistemlerine yapılan yatırımlar hem bireysel hem de toplumsal düzeyde önemli kazanımlar sağlayabilmektedir.

1.1.3. Çalışmanın Genel Akışı

Şekil 1.2'de tez çalışmasının genel akışı aşağıda şematik olarak gösterilmektedir. Çalışma, veri setinin elde edilmesinden sonuçların raporlanmasına kadar sistematik bir süreç izlemektedir. Her aşama, bir sonraki adımın temelini oluşturmakta ve metodolojik tutarlılığı sağlamaktadır.

Şekil 1.2'de görüldüğü üzere, çalışma altı temel aşamadan oluşmaktadır. İlk aşamada Heart Disease UCI veri seti elde edilmiş ve 920 hasta kaydı ile 14 öznitelik içeren bu veri seti çalışmanın temelini oluşturmuştur. İkinci aşamada keşifsel veri analizi (EDA) gerçekleştirilerek veri setinin karakteristikleri anlaşılmıştır. Üçüncü aşamada veri ön işleme adımları (eksik veri doldurma, SMOTE) uygulanmıştır.



Şekil 1. 1 Tez Çalışması Akış Şeması

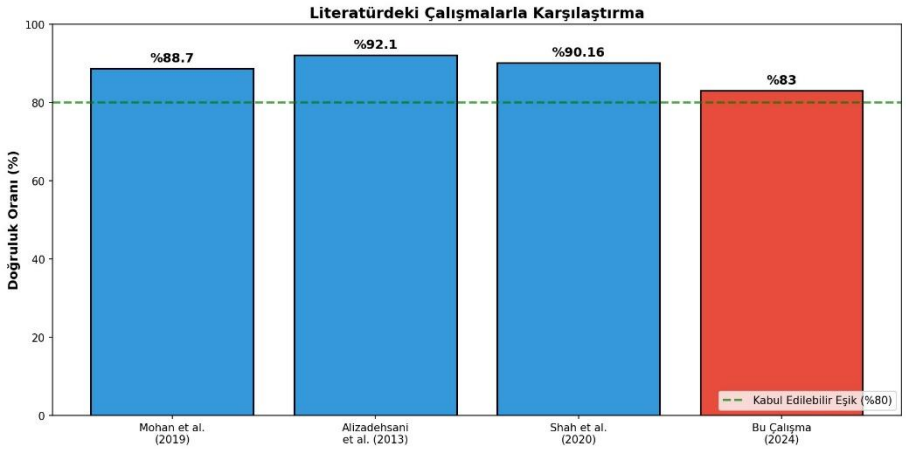
Dördüncü aşamada altı farklı makine öğrenmesi algoritması eğitilmiş, beşinci aşamada bu modellerin performansları çeşitli metrikler (Accuracy, AUC, F1 Score vb.) kullanılarak değerlendirilmiştir. Son aşamada ise sonuçlar karşılaştırılmış ve yorumlanarak raporlanmıştır.

1.2. Literatür Taraması

Kalp hastalıklarının teşhisinde makine öğrenmesi algoritmalarının kullanımı, son yıllarda araştırmacıların yoğun ilgi gösterdiği bir alan haline gelmiştir. Bu bölümde, konuyla ilgili yapılmış önceki çalışmalar sistematik bir şekilde incelenmekte ve mevcut literatürdeki boşluklar tartışılmaktadır. Literatür taraması, çalışmanın konumlandırılması ve özgün katkısının belirlenmesi açısından kritik öneme sahiptir.

1.2.1. Uluslararası Çalışmalar

Uluslararası literatürde kalp hastalığı tahmini için makine öğrenmesi uygulamalarına ilişkin çok sayıda çalışma bulunmaktadır. Bu çalışmalar, farklı algoritmalar, veri setleri ve metodolojiler kullanarak çeşitli sonuçlar elde etmiştir. Şekil 1.2 de bazı önemli çalışmaların performans sonuçları karşılaştırmalı olarak sunulmaktadır.



Şekil 1. 2 Literatürdeki Çalışmalarla Performans Karşılaştırılması

Mohan, hibrit makine öğrenmesi teknikleri kullanarak kalp hastalığı tahmini üzerine kapsamlı bir çalışma gerçekleştirmişlerdir [10]. Araştırmacılar, farklı öznitelik seçim yöntemleri ve sınıflandırma algoritmalarının kombinasyonlarını test etmişler ve %88,7 doğruluk oranına ulaşmışlardır. Çalışmada Naive Bayes, Random Forest ve Gradient Boosting algoritmalarının hibrit kullanımının tek algoritma kullanımına göre daha iyi sonuçlar verdiği gösterilmiştir.

Alizadehsani tarafından koroner arter hastalığının teşhisi için kapsamlı bir veri madenciliği yaklaşımı önermişlerdir [11]. İran'daki bir kardiyoloji kliniğinden toplanan 303 hastaya ait verileri kullanan araştırmacılar, %92,1 doğruluk oranı elde etmişlerdir. Çalışmada özellikle Naive Bayes ve Destek Vektör Makineleri algoritmalarının yüksek performans gösterdiği belirtilmiştir. Araştırmanın önemli bulgularından biri, öznitelik seçiminin model performansı üzerindeki kritik etkisinin ortaya konmasıdır.

Shah ve arkadaşları tarafından makine öğrenmesi teknikleri kullanarak kalp hastalığı tahmini üzerine geniş kapsamlı bir karşılaştırmalı çalışma sunmuşlardır [8]. Cleveland veri seti üzerinde gerçekleştirilen araştırmada KNN, Karar Ağaçları ve Rastgele Orman algoritmaları değerlendirilmiştir.Sonuçlar, Rastgele Orman algoritmasının %90,16 doğruluk oranı ile en iyi performansı gösterdiğini ortaya koymuştur. Çalışmada ayrıca hiperparametre optimizasyonunun önemi vurgulanmıştır. Çalışmada yapay sinir ağları (ANN), derin sinir ağları (DNN) ve geleneksel algoritmalar (SVM, Random Forest) değerlendirilmiştir. Sonuçlar, büyük veri setlerinde derin öğrenme yaklaşımlarının avantajlı olabileceğini, ancak sınırlı veri durumlarında geleneksel yöntemlerin rekabetçi performans sergilediğini göstermiştir.

1.2.2. Ulusal Çalışmalar

Türkiye'de kalp hastalığı teşhisinde makine öğrenmesi uygulamalarına yönelik araştırmalar son yıllarda artış göstermektedir. Bu çalışmalar hem uluslararası veri setleri hem de yerel sağlık kurumlarından elde edilen veriler kullanılarak gerçekleştirilmiştir.

Görgün (2020), yüksek lisans tezi kapsamında makine öğrenmesi yöntemleri ile kalp hastalığı tahmini üzerine kapsamlı bir çalışma gerçekleştirmiştir [13]. UCI Heart Disease veri seti kullanılan çalışmada, farklı sınıflandırma algoritmaları karşılaştırılmış ve öznitelik seçim yöntemlerinin etkileri incelenmiştir. Çalışma, Türkçe literatüre önemli bir katkı sağlamıştır.

Yılmaz ve Yağın, Türkiye'deki bir üniversite hastanesinden elde edilen yerel veriler kullanarak kalp hastalığı risk tahmini modeli geliştirmişlerdir [14].

Çalışmada ensemble (topluluk) yöntemlerinin tek algoritmalara göre daha yüksek performans gösterdiği bulunmuştur. Araştırma, yerel popülasyona özgü risk faktörlerinin önemini vurgulamıştır.

1.2.3. Literatürdeki Boşluklar ve Bu Çalışmanın Katkısı

Mevcut literatür incelendiğinde, bazı önemli boşluklar ve geliştirilmesi gereken alanlar tespit edilmektedir. Birincisi, çoğu çalışma sınırlı sayıda algoritma üzerinde yoğunlaşmakta ve kapsamlı karşılaştırmalar sunmamaktadır. İkincisi, veri ön işleme adımlarının detaylı raporlanması çoğu zaman ihmal edilmektedir. Üçüncüsü, farklı performans metriklerinin dengeli kullanımı konusunda tutarsızlıklar bulunmaktadır.

Bu tez çalışması, bahsedilen boşlukları gidermeye yönelik özgün katkılar sunmaktadır. Altı farklı algoritmanın aynı metodoloji altında karşılaştırılması, kapsamlı veri ön işleme adımlarının detaylı dokümantasyonu ve çoklu

performans metriklerinin dengeli kullanımı bu katkıların başlıcalarıdır. Ayrıca, çalışma Türkçe literatüre güncel ve kapsamlı bir kaynak kazandırmaktadır.

Tablo 1.1 Literatürdeki Önemli Çalışmaların Özeti

Çalışma	Yıl	Algoritmalar	Doğruluk	Veri Seti
Mohan ve ark. [10]	2019	Hibrit ML	%88.7	Cleveland
Alizadehsani ve ark. [11]	2013	NB, SVM	%92.1	Z-Alizadeh
Shah ve ark. [12]	2020	RF, KNN, DT	%90.16	Cleveland
Görgün [18]	2020	Çoklu ML	%85.2	UCI Heart
Bu Çalışma	2024	6 Algoritma	%83.0	UCI Heart

1.3. Problem Tanımı ve Araştırma Soruları

Kalp hastalıklarının erken teşhisi, hasta sonuçlarının iyileştirilmesi ve sağlık sistemleri üzerindeki yükün azaltılması açısından kritik öneme sahiptir. Ancak geleneksel tanı yöntemleri zaman alıcı, maliyetli ve uzman bağımlı olabilmektedir. Bu durum, özellikle sağlık kaynaklarının sınırlı olduğu bölgelerde erişilebilirlik sorunlarına yol açmaktadır.

Makine öğrenmesi tabanlı tanı destek sistemleri, bu sorunlara potansiyel çözümler sunmaktadır. Ancak bu sistemlerin etkin kullanımı için bazı temel soruların yanıtlanması gerekmektedir. Bu araştırma, aşağıdaki temel sorulara yanıt aramaktadır:

Araştırma Sorusu 1: Kalp hastalığı teşhisinde hangi makine öğrenmesi algoritması en yüksek tahmin doğruluğunu sağlamaktadır?

Araştırma Sorusu 2: Farklı performans metrikleri (doğruluk, AUC, F1 skoru) açısından algoritmalar nasıl bir sıralama göstermektedir?

Araştırma Sorusu 3: Veri ön işleme teknikleri (eksik veri doldurma, SMOTE) model performansını ne ölçüde etkilemektedir?

Araştırma Sorusu 4: Yorumlanabilirlik ve hesaplama verimliliği açısından algoritmalar arasındaki denge nasıl kurulabilir?

1.3.1. Araştırma Hipotezleri

Bu araştırma, aşağıdaki hipotezleri test etmektedir:

H1: Topluluk (ensemble) yöntemleri (Random Forest, XGBoost, LightGBM), tek algoritmalara (Lojistik Regresyon, KNN, SVM) göre daha yüksek tahmin doğruluğu sağlayacaktır.

H2: SMOTE uygulaması, sınıf dengesizliği bulunan veri setinde model performansını özellikle azınlık sınıfı için artıracaktır.

H3: Makine öğrenmesi tabanlı eksik veri doldurma yöntemleri, basit istatistiksel yöntemlere (ortalama, medyan) göre daha iyi sonuçlar verecektir.

H4: AUC metriği, dengesiz veri setlerinde doğruluk metriğine göre daha güvenilir bir performans göstergesi olacaktır.

1.4. Çalışmanın Kapsamı ve Sınırlılıkları

Bu çalışma, belirli bir kapsam dahilinde gerçekleştirilmiş olup bazı sınırlılıklar içermektedir. Bu bölümde, çalışmanın kapsamı ve kabul edilen sınırlılıklar açıkça belirtilmektedir.

1.4.1. Çalışmanın Kapsamı

Çalışma, UCI Machine Learning Repository'den elde edilen Heart Disease veri seti ile sınırlıdır. Bu veri seti, dört farklı tıbbi kurumdan (Cleveland, Macaristan, İsviçre, VA Long Beach) derlenen toplam 920 hasta kaydını içermektedir. Değerlendirilen algoritmalar, gözetimli öğrenme kategorisindeki sınıflandırma algoritmaları ile sınırlı tutulmuştur.

Çalışmada kullanılan performans metrikleri; doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru ve AUC (Area Under the ROC Curve) olarak belirlenmiştir. Bu metrikler, sınıflandırma problemlerinde yaygın olarak kullanılan ve model performansının farklı yönlerini değerlendiren standart ölçütlerdir.

1.4.2. Sınırlılıklar

Bu çalışmanın başlıca sınırlılıkları şunlardır: Birincisi, kullanılan veri seti nispeten küçük boyuttadır (920 örnek) ve bu durum derin öğrenme gibi veri yoğun yöntemlerin uygulanmasını kısıtlamaktadır. İkincisi, veri seti retrospektif verilerden oluşmakta olup prospektif doğrulama yapılmamıştır.

Üçüncüsü, veri setinde cinsiyet dağılımı dengesiz olup erkek hastalar ağırlıklıdır (%78,9). Bu durum, kadın hastalarda model performansının farklılık gösterebileceği anlamına gelmektedir. Dördüncüsü, bazı özniteliklerde yüksek

oranda eksik veri bulunmakta ve bu değerler tahminleme yöntemleriyle doldurulmuştur.

Son olarak, çalışma tek bir veri kaynağı (UCI) ile sınırlıdır ve sonuçların farklı popülasyonlara genellenebilirliği için ek doğrulama çalışmalarına ihtiyaç duyulmaktadır. Bu sınırlılıklar, sonuçların yorumlanmasında ve gelecek çalışmaların planlanmasında göz önünde bulundurulmalıdır.

1.5. Tezin Organizasyonu

Bu tez çalışması, sistematik bir yapı izleyerek beş ana bölümden oluşmaktadır. Her bölüm, araştırmanın farklı bir yönünü ele almakta ve bir bütün olarak çalışmanın kapsamlı bir şekilde sunulmasını sağlamaktadır.

Birinci Bölüm (GİRİŞ): Bu bölümde çalışmanın motivasyonu, amacı ve önemi açıklanmaktadır. Literatür taraması ile mevcut çalışmalar değerlendirilmekte, araştırma soruları ve hipotezler belirlenmektedir.
Ayrıca çalışmanın kapsamı ve sınırlılıkları tartışılmaktadır.

İkinci Bölüm (KURAMSAL TEMELLER): Bu bölümde makine öğrenmesi kavramları, kullanılan algoritmalar (Lojistik Regresyon, KNN, SVM, Random Forest, XGBoost, LightGBM), veri madenciliği süreçleri ve performans değerlendirme metrikleri (doğruluk, kesinlik, duyarlılık, F1 skoru, AUC) detaylı olarak açıklanmaktadır.

Üçüncü Bölüm (MATERYAL VE YÖNTEM): Bu bölümde Heart Disease UCI veri setinin özellikleri, kalp hastalıklarının tıbbi önemi, veri setine uygulanan ön işlemler (eksik veri doldurma, SMOTE, ölçeklendirme) ve kullanılan yazılım araçları tanıtılmaktadır.

Dördüncü Bölüm (UYGULAMA VE BULGULAR): Bu bölümde keşifsel veri analizi sonuçları, model eğitim süreçleri ve algoritmaların karşılaştırmalı performans analizleri sunulmaktadır. Her algoritmanın performans metrikleri tablolar ve grafiklerle görselleştirilmektedir.

Beşinci Bölüm (SONUÇ): Son bölümde, çalışmadan elde edilen sonuçlar değerlendirilmekte, araştırma sorularına ve hipotezlere yanıtlar verilmektedir. Çalışmanın kısıtlamaları tartışılmakta ve gelecek araştırmalar için öneriler sunulmaktadır

2.KURAMSAL TEMELLER

Bu bölümde, çalışmanın temelini oluşturan makine öğrenmesi kavramları, kullanılan algoritmalar, veri madenciliği süreçleri ve performans değerlendirme metrikleri detaylı olarak ele alınmaktadır. Bilgisayar ve matematik mühendisliği perspektifinden hazırlanan bu bölüm, algoritmik yaklaşımların teorik temellerini açıklamayı amaçlamaktadır.

2.1. Makine Öğrenmesi (Machine Learning)

Makine öğrenmesi (Machine Learning- ML), yapay zekânın (Artificial Intelligence- AI) bir alt dalı olup bilgisayar sistemlerinin açık programlama yapılmadan verilerden öğrenmesini ve deneyimlerinden faydalanarak performansını iyileştirmesini sağlayan bir bilim dalıdır. Arthur Samuel tarafından "bilgisayarlara açıkça programlanmadan öğrenme yeteneği kazandıran çalışma alanı" olarak tanımlanan makine öğrenmesi, günümüzde sağlık, finans, güvenlik ve birçok farklı sektörde yaygın olarak kullanılmaktadır [5].

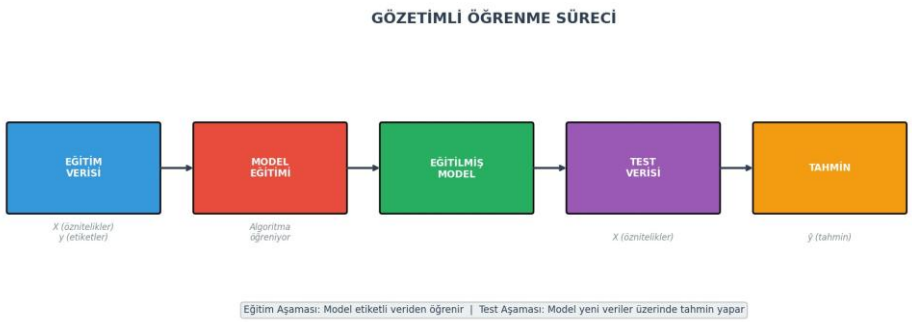
Makine öğrenmesinin temel prensibi, verilen bir veri kümesinden örüntüler (patterns) çıkarmak ve bu örüntülere dayanarak yeni, daha önce görülmemiş veriler hakkında tahminlerde bulunmaktır. Bu süreç, geleneksel programlamadan farklı olarak kuralların açıkça tanımlanması yerine verilerden otomatik olarak öğrenilmesini içermektedir. Bir makine öğrenmesi modelinin başarısı, eğitim verilerinin kalitesine, seçilen algoritmanın uygunluğuna ve hiperparametre (hyperparameter) ayarlamalarına bağlıdır.

Tom Mitchell makine öğrenmesini daha formal bir şekilde şöyle tanımlamıştır: "Bir bilgisayar programının, T görevinde P performans ölçütüne göre E deneyiminden öğrendiği söylenir, eğer T'deki performansı P ile ölçüldüğünde E deneyimi ile iyileşiyorsa." [4]. Bu tanım, makine öğrenmesinin üç temel bileşenini ortaya koymaktadır: görev (task), performans ölçütü (performance measure) ve deneyim (experience).

2.1.1. Makine Öğrenmesi Tanımı

Makine öğrenmesi, istatistik, optimizasyon teorisi, bilgisayar bilimi ve olasılık teorisinin kesişiminde yer alan disiplinler arası bir alandır. Temel amacı, verilerden anlamlı bilgiler çıkarmak ve bu bilgileri kullanarak karar verme süreçlerini otomatikleştirmektir. Makine öğrenmesi sistemleri, girdi verilerini (input) işleyerek çıktı (output) üretir ve bu süreçte bir model (model) oluşturur.

Makine öğrenmesi modelleri genel olarak iki ana aşamadan oluşmaktadır. Birinci aşama olan eğitim (training) aşamasında, model etiketli veya etiketsiz veriler kullanılarak eğitilir ve veriler arasındaki ilişkileri öğrenir. İkinci aşama olan test (testing) aşamasında ise eğitilmiş model, daha önce görmediği veriler üzerinde değerlendirilerek performansı ölçülür. Bu iki aşamalı yaklaşım, modelin genelleme (generalization) yeteneğini test etmeyi amaçlamaktadır. Şekil 2.1'de gözetimli öğrenme süreci şeması verilmiştir.



Şekil 2. 1 Gözetimli Öğrenme Sürecinin Şematik Gösterimi

2.1.2. Makine Öğrenmesi Türleri

Makine öğrenmesi algoritmaları, öğrenme yöntemlerine göre dört ana kategoride sınıflandırılmaktadır: gözetimli öğrenme (supervised learning), gözetimsiz öğrenme (unsupervised learning), yarı gözetimli öğrenme (semi-supervised learning) ve pekiştirmeli öğrenme (reinforcement learning). Şekil 2.2'de bu türler ve temel özellikleri gösterilmektedir.



Şekil 2. 2 Makine Öğrenmesi Türleri ve Temel Özellikleri

Gözetimli Öğrenme (Supervised Learning): Bu öğrenme türünde, model etiketli veriler (labeled data) kullanılarak eğitilir. Her bir girdi verisi için karşılık gelen

doğru çıktı (etiket) bilinmektedir. Model, girdi-çıktı çiftlerinden öğrenerek yeni girdiler için tahminlerde bulunur.

Gözetimli öğrenme, sınıflandırma (classification) ve regresyon (regression) olmak üzere iki temel problemi çözmektedir. Sınıflandırma problemlerinde amaç, veriyi önceden tanımlanmış kategorilere ayırmaktır (örneğin: hasta/sağlıklı). Regresyon problemlerinde ise sürekli bir değerin tahmini yapılır (örneğin: ev fiyatı).

Gözetimsiz Öğrenme (Unsupervised Learning): Bu öğrenme türünde, model etiketsiz veriler kullanılarak eğitilir. Algoritma, verilerdeki gizli yapıları ve örüntüleri kendi başına keşfetmeye çalışır.

Kümeleme (clustering), boyut indirgeme (dimensionality reduction) ve ilişkilendirme kuralları (association rules) gözetimsiz öğrenmenin temel uygulama alanlarıdır. K-ortalamalar (K-means) ve hiyerarşik kümeleme yaygın kullanılan gözetimsiz öğrenme algoritmalarıdır.

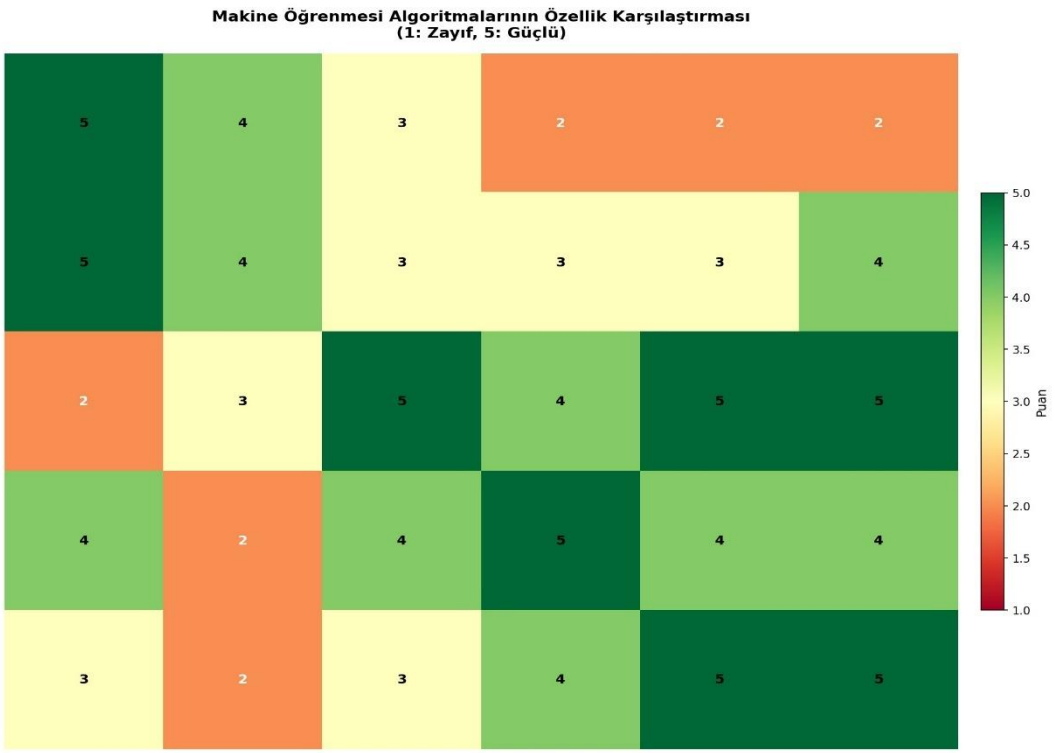
Yarı Gözetimli Öğrenme (Semi-Supervised Learning): Bu yaklaşım, hem etiketli hem de etiketsiz verileri birlikte kullanmaktadır. Pratikte etiketli veri elde etmek maliyetli olabildiğinden, az miktarda etiketli veri ile büyük miktarda etiketsiz veriyi birleştiren bu yöntem tercih edilebilmektedir.

Pekiştirmeli Öğrenme (Reinforcement Learning): Bu öğrenme türünde, bir ajan (agent) çevresiyle etkileşime girerek deneme-yanılma yoluyla öğrenir. Ajan, aldığı kararların sonucunda ödül (reward) veya ceza (penalty) alır ve amacı toplam ödülü maksimize etmektir. Oyun oynama, robotik ve otonom araçlar bu öğrenme türünün uygulama alanlarındandır.

Bu çalışmada, kalp hastalığı teşhisi bir sınıflandırma problemi olarak ele alınmakta ve gözetimli öğrenme algoritmaları kullanılmaktadır. Hasta ve sağlıklı bireyleri ayırt etmek için etiketli veriler kullanılarak modeller eğitilmektedir.

2.2. Makine Öğrenmesi Algoritmaları

Bu bölümde, çalışmada kalp hastalığı tahmini için kullanılan altı farklı makine öğrenmesi algoritması detaylı olarak açıklanmaktadır. Her bir algoritmanın çalışma prensibi, avantajları ve dezavantajları ele alınmaktadır. Şekil 2.3'te bu algoritmaların çeşitli özellikler açısından karşılaştırması sunulmaktadır.



Şekil 2. 3 Makine Öğrenmesi Algoritmalarının Özellik Karşılaştırması

2.2.1. Lojistik Regresyon (Logistic Regression)

Lojistik regresyon (Logistic Regression), ikili sınıflandırma (binary classification) problemleri için yaygın olarak kullanılan bir istatistiksel yöntemdir. Adında "regresyon" geçmesine rağmen, aslında bir sınıflandırma algoritmasıdır. Lojistik regresyon, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi lojistik (sigmoid) fonksiyonu kullanarak modellemektedir.

Sigmoid fonksiyonu, herhangi bir gerçel sayıyı 0 ile 1 arasında bir değere dönüştürmektedir. Bu özellik, fonksiyonun çıktısının olasılık olarak yorumlanmasına olanak tanımaktadır. Matematiksel olarak sigmoid fonksiyonu

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.1}$$

(2.1) formülü şeklinde ifade edilmektedir. Burada z, bağımsız değişkenlerin doğrusal kombinasyonunu temsil etmektedir.

Lojistik regresyonun avantajları arasında yorumlanabilirlik, düşük hesaplama maliyeti ve aşırı öğrenmeye (overfitting) karşı direnç sayılabilir. Dezavantajları ise doğrusal olmayan ilişkileri yakalayamaması ve karmaşık veri yapılarında düşük performans göstermesidir.

2.2.2. K-En Yakın Komşu (K-Nearest Neighbors)

K-En Yakın Komşu (K-Nearest Neighbors - KNN), hem sınıflandırma hem de regresyon problemlerinde kullanılabilen, örnek tabanlı (instance-based) bir öğrenme algoritmasıdır. Bu algoritma, parametrik olmayan (non-parametric) bir yaklaşım benimser ve verilerin dağılımı hakkında herhangi bir varsayım yapmaz.

KNN algoritmasının çalışma prensibi oldukça sezgiseldir. Yeni bir veri noktasını sınıflandırmak için algoritma, eğitim veri setindeki en yakın k komşuyu bulur ve bu komşuların çoğunluk oylaması (majority voting) ile sınıf etiketi belirlenir. Mesafe ölçütü olarak genellikle Öklid (Euclidean) mesafesi kullanılmakta olup Manhattan ve Minkowski mesafeleri de tercih edilebilmektedir.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.2}$$

K değerinin seçimi, model performansı üzerinde önemli bir etkiye sahiptir. Küçük k değerleri gürültüye duyarlılığı artırırken, büyük k değerleri sınıf sınırlarını bulanıklaştırabilmektedir. Optimal k değeri genellikle çapraz doğrulama (cross-validation) yöntemiyle belirlenmektedir.

2.2.3. Destek Vektör Makineleri (SVM)

Destek Vektör Makineleri (Support Vector Machines- SVM), Vladimir Vapnik tarafından geliştirilen güçlü bir sınıflandırma algoritmasıdır [15]. SVM'nin temel amacı, farklı sınıflara ait veri noktalarını en iyi şekilde ayıran optimal hiper düzlemi (hyperplane) bulmaktır.

SVM algoritması, sınıflar arasındaki marjini (margin) maksimize etmeye çalışır. Marjin, hiper düzlem ile en yakın veri noktaları arasındaki mesafedir. Bu en yakın noktalara destek vektörleri (support vectors) adı verilmektedir ve bunlar sınıflandırma kararında kritik rol oynamaktadır.

$$f(x_q) = sign \sum_{i} \alpha_i \; y_i \; \mathrm{K}(x_q, x_i) \tag{2.3}$$

Doğrusal olarak ayrılamayan veriler için SVM, çekirdek hilesi (kernel trick) kullanmaktadır. Çekirdek fonksiyonları (kernel functions), veriyi daha yüksek boyutlu bir uzaya haritalayarak doğrusal ayrımı mümkün kılmaktadır. Yaygın kullanılan çekirdek fonksiyonları arasında doğrusal (linear), polinom (polynomial), radyal tabanlı fonksiyon (Radial Basis Function - RBF) ve sigmoid çekirdekleri bulunmaktadır.

2.2.4. Rastgele Orman (Random Forest)

Rastgele Orman (Random Forest), Leo Breiman tarafından 2001 yılında önerilen bir topluluk öğrenme (ensemble learning) algoritmasıdır. Bu algoritma, birden fazla karar ağacının (decision tree) birleştirilmesiyle oluşturulmakta ve her ağacın tahminlerinin ortalaması veya çoğunluk oylaması ile nihai karar verilmektedir [6].

$$f(y) = \frac{1}{t} \sum_{t=1}^{t} y_t \tag{2.4}$$

Rastgele orman algoritmasının iki temel bileşeni vardır: torbalama (bagging) ve rastgele öznitelik seçimi. Torbalama yönteminde, eğitim verisinden yeniden örnekleme (bootstrap sampling) yapılarak her ağaç farklı bir alt küme üzerinde eğitilir. Rastgele öznitelik seçiminde ise her düğümde (node) bölme yapılırken tüm öznitelikler yerine rastgele seçilen bir alt küme değerlendirilir.

Rastgele ormanın avantajları arasında aşırı öğrenmeye karşı direnç, eksik verilere dayanıklılık, öznitelik önem sıralaması çıkarabilme ve yüksek boyutlu verilerde iyi performans gösterme sayılabilir. Dezavantajları ise yorumlanabilirliğin düşük olması ve büyük veri setlerinde yüksek bellek tüketimidir.

2.2.5. XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting), Tianqi Chen tarafından geliştirilen ve gradyan artırma (gradient boosting) çerçevesine dayanan yüksek performanslı bir makine öğrenmesi algoritmasıdır [7]. XGBoost, Kaggle yarışmalarında ve endüstriyel uygulamalarda yaygın olarak tercih edilen bir algoritma olup özellikle yapılandırılmış veriler üzerinde üstün performans sergilemektedir.

$$A = \sum_{i=1}^{n} a(y_{i,}\hat{y}_i) + \sum_{k=1}^{K} b(f_k) \tag{2.5}$$

Gradyan artırma, zayıf öğrenicileri (weak learners) sıralı olarak birleştirerek güçlü bir model oluşturan bir topluluk yöntemidir. Her yeni ağaç, önceki ağaçların hatalarını düzeltmeye odaklanır. XGBoost, bu temel fikri düzenlileştirme (regularization) terimleri, ikinci dereceden Taylor açılımı ve paralel işlem desteği gibi optimizasyonlarla geliştirmektedir.

XGBoost'un öne çıkan özellikleri arasında L1 (Lasso) ve L2 (Ridge) düzenlileştirme desteği [16,17], eksik değer yönetimi, ağaç budama (tree pruning) ve öznitelik önem skoru hesaplayabilme yer almaktadır. Ayrıca

algoritma, GPU desteği sayesinde büyük veri setlerinde hızlı eğitim süresi sunmaktadır.

2.2.6. LightGBM (Light Gradient Boosting Machine)

LightGBM (Light Gradient Boosting Machine), Microsoft tarafından geliştirilen ve gradyan artırma çerçevesine dayanan bir makine öğrenmesi algoritmasıdır. LightGBM, özellikle büyük veri setlerinde yüksek hız ve düşük bellek kullanımı sağlamak amacıyla tasarlanmıştır.

$$y = \arg\max_{c} \left(\frac{1}{N} \sum_{n=1}^{N} P_n(y = c \mid x) \right) \tag{2.6}$$

LightGBM'in en önemli özelliği, yaprak bazlı (leaf-wise) ağaç büyütme stratejisidir. Geleneksel gradyan artırma algoritmaları seviye bazlı (level-wise) büyütme kullanırken, LightGBM en yüksek kayıp azalmasını sağlayan yaprağı bölerek ilerler. Bu yaklaşım, daha hızlı yakınsama ve daha yüksek doğruluk sağlamaktadır ancak aşırı öğrenme riskini artırabilmektedir.

LightGBM, Gradient-based One-Side Sampling (GOSS) ve Exclusive Feature Bundling (EFB) gibi yenilikçi teknikler kullanmaktadır. GOSS, büyük gradyanlara sahip örneklere odaklanarak eğitim süresini kısaltırken, EFB seyrek (sparse) öznitelikleri gruplandırarak boyut azaltma sağlamaktadır.

2.3. Veri Madenciliği (Data Mining)

Veri madenciliği (Data Mining), büyük veri kümelerinden anlamlı örüntüler, ilişkiler ve bilgiler çıkarma sürecidir. Makine öğrenmesi ile yakından ilişkili olan bu disiplin, istatistik, veritabanı sistemleri ve yapay zekâ alanlarının kesişiminde yer almaktadır.

2.3.1. Temel Kavramlar ve Tanımlar

Veri madenciliği, ham veriden değerli bilgi çıkarma sanatı ve bilimi olarak tanımlanabilir. Bu süreç, veritabanlarındaki gizli örüntülerin, beklenmedik ilişkilerin ve kullanışlı bilgilerin otomatik veya yarı otomatik yöntemlerle keşfedilmesini kapsamaktadır. Fayyad ve arkadaşları veri madenciliğini, "veritabanlarından bilgi keşfi" (Knowledge Discovery in Databases- KDD) sürecinin önemli bir adımı olarak tanımlamışlardır [9].

Veri madenciliğinde kullanılan temel kavramlar arasında veri seti (dataset), öznitelik (attribute/feature), örnek (instance/sample), etiket (label) ve model bulunmaktadır. Veri seti, analiz edilecek verilerin organize edilmiş halidir.

Öznitelikler, her bir örneği tanımlayan karakteristiklerdir. Etiket, gözetimli öğrenmede tahmin edilmeye çalışılan hedef değişkendir.

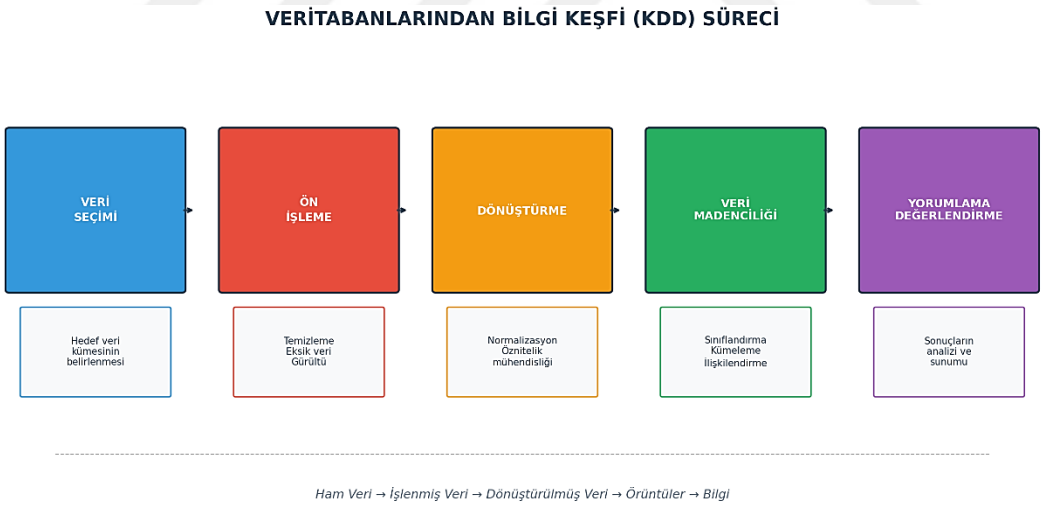
2.3.2. Tarihsel Gelişim

Veri madenciliğinin kökleri 1960'lı yıllara kadar uzanmaktadır. İlk dönemde istatistiksel analiz yöntemleri ve basit veritabanı sorguları kullanılmaktaydı. 1980'lerde ilişkisel veritabanlarının yaygınlaşması ve veri ambarı (data warehouse) kavramının ortaya çıkması ile veri madenciliği alanı gelişmeye başlamıştır.

1990'lı yıllar veri madenciliğinin altın çağı olarak kabul edilmektedir. Bu dönemde CRISP-DM (Cross-Industry Standard Process for Data Mining) gibi standart metodolojiler geliştirilmiş ve ticari veri madenciliği araçları piyasaya sürülmüştür. 2000'li yıllardan itibaren büyük veri (big data) kavramının ortaya çıkması ve hesaplama gücündeki artış, veri madenciliği uygulamalarını daha da yaygınlaştırmıştır.

2.3.3. Bilgi Keşif Süreci

Veritabanlarından bilgi keşfi (Knowledge Discovery in Databases- KDD), ham veriden kullanışlı bilgiye ulaşmak için izlenen sistematik bir süreçtir. Bu süreç



Şekil 2. 4 Veri tabanlarından Bilgi Keşfi (KDD) Süreci

genellikle beş temel aşamadan oluşmaktadır: seçim, ön işleme, dönüştürme, veri madenciliği ve yorumlama/değerlendirme. Şekil 2.4'te bu süreç şematik olarak gösterilmektedir.

Seçim aşamasında, analize dahil edilecek veri alt kümesi belirlenir. Ön işleme aşamasında, eksik veriler doldurulur, gürültülü veriler temizlenir ve aykırı değerler (outliers) işlenir. Dönüştürme aşamasında, veriler normalizasyon, standardizasyon veya öznitelik mühendisliği (feature engineering) gibi tekniklerle modelleme için uygun hale getirilir.

Veri madenciliği aşamasında, sınıflandırma, kümeleme, ilişkilendirme veya regresyon gibi teknikler uygulanarak veriden örüntüler çıkarılır. Son aşama olan yorumlama ve değerlendirmede ise elde edilen sonuçlar analiz edilir, model performansı ölçülür ve bulgular karar vericilere sunulur.

2.3.4. Uygulama Alanları

Veri madenciliği, günümüzde pek çok farklı sektörde yaygın olarak kullanılmaktadır. Sağlık alanında hastalık teşhisi, tedavi planlaması ve ilaç keşfi; finans sektöründe dolandırıcılık tespiti, kredi risk analizi ve müşteri segmentasyonu; perakende sektöründe müşteri davranış analizi ve ürün önerisi sistemleri bu uygulama alanlarına örnek olarak verilebilir.

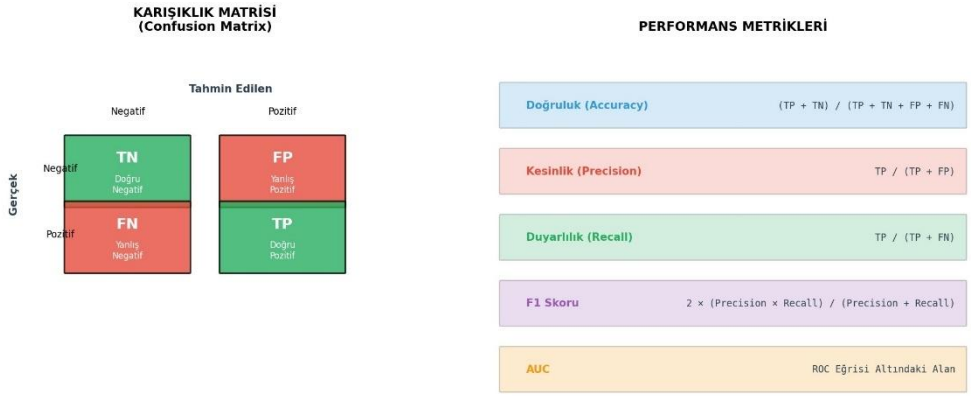
Sağlık sektöründe veri madenciliği ve makine öğrenmesi uygulamaları özellikle son yıllarda büyük ilgi görmektedir. Elektronik sağlık kayıtlarından (Electronic Health Records- EHR) hastalık örüntülerinin keşfi, tıbbi görüntülerin analizi ve kişiselleştirilmiş tıp (personalized medicine) uygulamaları bu alandaki önemli çalışmalardandır. Bu tez çalışmasında da kalp hastalığı teşhisi için veri madenciliği ve makine öğrenmesi teknikleri bir arada kullanılmaktadır.

2.4. Performans Değerlendirme Metrikleri

Makine öğrenmesi modellerinin başarısını ölçmek için çeşitli performans metrikleri kullanılmaktadır. Bu metrikler, modelin doğruluğunu, güvenilirliğini ve genelleme yeteneğini değerlendirmeye yardımcı olmaktadır. Sınıflandırma problemlerinde en yaygın kullanılan metrikler aşağıda açıklanmaktadır.

2.4.1. Karışıklık Matrisi (Confusion Matrix)

Karışıklık matrisi (Confusion Matrix), bir sınıflandırma modelinin performansını görsel olarak özetleyen bir tablodur. İkili sınıflandırma problemlerinde 2x2 boyutunda bir matris oluşturulmakta ve dört temel değer içermektedir: Doğru Pozitif (True Positive- TP), Doğru Negatif (True Negative- TN), Yanlış Pozitif (False Positive- FP) ve Yanlış Negatif (False Negative- FN). Şekil 2.5'te karışıklık matrisi ve performans metrikleri gösterilmektedir.



Şekil 2. 5 Karışıklık Matrisi ve Performans Metrikleri Formülleri

Doğru Pozitif (TP), modelin doğru şekilde pozitif olarak sınıflandırdığı örneklerin sayısıdır. Doğru Negatif (TN), modelin doğru şekilde negatif olarak sınıflandırdığı örneklerin sayısıdır. Yanlış Pozitif (FP), gerçekte negatif olan ancak modelin pozitif olarak sınıflandırdığı örneklerin sayısıdır (Tip I hata). Yanlış Negatif (FN) ise gerçekte pozitif olan ancak modelin negatif olarak sınıflandırdığı örneklerin sayısıdır (Tip II hata).

2.4.2. Doğruluk (Accuracy)

Doğruluk (Accuracy), en temel performans metriğidir ve modelin doğru tahmin ettiği örneklerin toplam örnek sayısına oranını ifade etmektedir. Matematiksel olarak Accuracy aşağıdaki (2.2) formülü ile hesaplanmaktadır.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.7}$$

Doğruluk metriği basit ve anlaşılır olmasına rağmen, dengesiz veri setlerinde yanıltıcı olabilmektedir. Örneğin, bir sınıfın diğerinden çok daha fazla örneğe sahip olduğu durumlarda, model çoğunluk sınıfını sürekli tahmin ederek yüksek doğruluk elde edebilir ancak azınlık sınıfını hiç tanıyamayabilir.

2.4.3. Kesinlik (Precision)

Kesinlik (Precision), modelin pozitif olarak tahmin ettiği örnekler arasında gerçekten pozitif olanların oranını göstermektedir. Formül olarak

$$Precision = \frac{TP}{(TP + FP)} \tag{2.8}$$

(2.3) şeklinde ifade edilmektedir. Yüksek kesinlik değeri, modelin yanlış pozitif (false positive) üretme oranının düşük olduğunu göstermektedir.

Kesinlik metriği, yanlış alarmların maliyetli olduğu durumlarda özellikle önem kazanmaktadır. Örneğin, spam filtrelerinde önemli bir e-postanın spam olarak işaretlenmesi istenmeyen bir durumdur ve bu nedenle yüksek kesinlik hedeflenir.

2.4.4. Duyarlılık (Recall)

Duyarlılık (Recall), gerçekten pozitif olan örnekler arasında modelin doğru şekilde tespit ettiklerinin oranını göstermektedir. Hassasiyet (Sensitivity) veya Gerçek Pozitif Oranı (True Positive Rate- TPR) olarak da adlandırılan bu metrik,

$$Recall = \frac{TP}{(TP + FN)} \tag{2.9}$$

(2.4) formülü ile hesaplanmaktadır. Tıbbi teşhis gibi uygulamalarda duyarlılık kritik öneme sahiptir. Bir hastalığın gözden kaçırılması (yanlış negatif), yanlış alarm verilmesinden (yanlış pozitif) çok daha ciddi sonuçlara yol açabilir. Bu nedenle sağlık alanındaki modellerde genellikle yüksek duyarlılık hedeflenmektedir.

2.4.5. F1 Skoru (F1 Score)

F1 Skoru (F1 Score), kesinlik ve duyarlılık metriklerinin harmonik ortalamasıdır. Bu metrik, her iki ölçütü de dengeli bir şekilde dikkate aldığından, özellikle dengesiz veri setlerinde tercih edilmektedir.

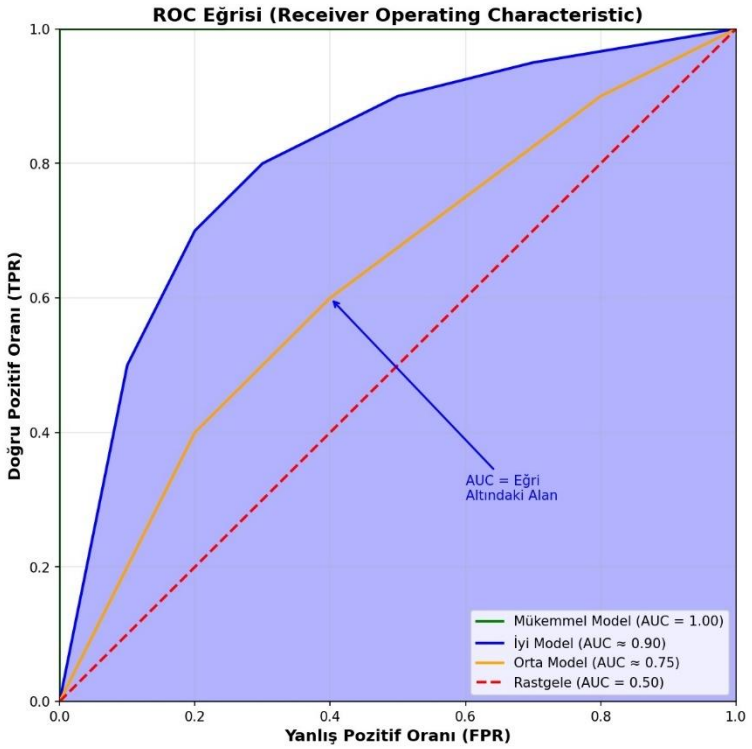


Şekil 2. 6 F1 Gösterimi

F1 Formülü yukarıdaki şekil 2.6 da gösterildiği gibi F1 hesaplanır. F1skoru 0 ile 1 arasında değer almakta olup 1 değeri mükemmel performansı, değeri ise en kötü performansı göstermektedir. Harmonik ortalama kullanılmasının nedeni, kesinlik veya duyarlılıktan birinin çok düşük olduğu durumlarda F1 skorunun da düşük kalmasını sağlamaktır.

2.4.6. ROC Eğrisi ve AUC

ROC (Receiver Operating Characteristic) eğrisi, farklı eşik değerlerinde modelin performansını gösteren bir grafiktir. Eğri, Gerçek Pozitif Oranı (True Positive Rate- TPR) ile Yanlış Pozitif Oranı (False Positive Rate FPR) arasındaki ilişkiyi görselleştirmektedir. İdeal bir model, ROC eğrisinin sol üst köşesine yakın bir



Şekil 2. 7 Farklı Performans Seviyelerinde ROC Eğrisi Örnekleri

seyir izler. Şekil 2.7'de farklı performans seviyelerine sahip modellerin ROC eğrileri gösterilmektedir.

Yukarıdaki AUC (Area Under the Curve), ROC eğrisinin altında kalan alanı ifade etmektedir ve modelin genel performansını tek bir sayıyla özetlemektedir. AUC değeri 0 ile 1 arasında olup 1 mükemmel sınıflandırmayı, 0.5 ise rastgele tahmine eşdeğer performansı göstermektedir. Genel olarak AUC değeri 0.7'nin üzerinde kabul edilebilir, 0.8'in üzerinde iyi ve 0.9'un üzerinde mükemmel olarak değerlendirilmektedir.

Bu çalışmada, kullanılan tüm makine öğrenmesi algoritmalarının performansları yukarıda açıklanan metrikler kullanılarak karşılaştırılmakta ve en uygun modelin belirlenmesi amaçlanmaktadır.

3. MATERYAL VE YÖNTEM

Bu bölümde, çalışmada kullanılan veri seti, kalp hastalıklarının önemi, veri setine uygulanan ön işleme adımları ve kullanılan araçlar ile kütüphaneler detaylı olarak açıklanmaktadır. Çalışmanın tekrarlanabilirliğini sağlamak amacıyla tüm yöntemsel adımlar ayrıntılı bir şekilde dokümante edilmiştir. Bilimsel araştırmalarda yöntem bölümü, çalışmanın geçerliliği ve güvenilirliği açısından kritik öneme sahiptir.

Makine öğrenmesi projelerinde veri kalitesi, model performansını doğrudan etkileyen en önemli faktörlerden biridir. Bu nedenle, veri setinin karakteristiklerinin iyi anlaşılması, uygun ön işleme tekniklerinin seçilmesi ve doğru algoritmaların belirlenmesi büyük önem taşımaktadır. Bu bölümde ele alınan konular, çalışmanın metodolojik temelini oluşturmakta ve elde edilen sonuçların yorumlanmasında referans noktası sağlamaktadır.

3.1. Veri Seti: Heart Disease UCI (Cleveland)

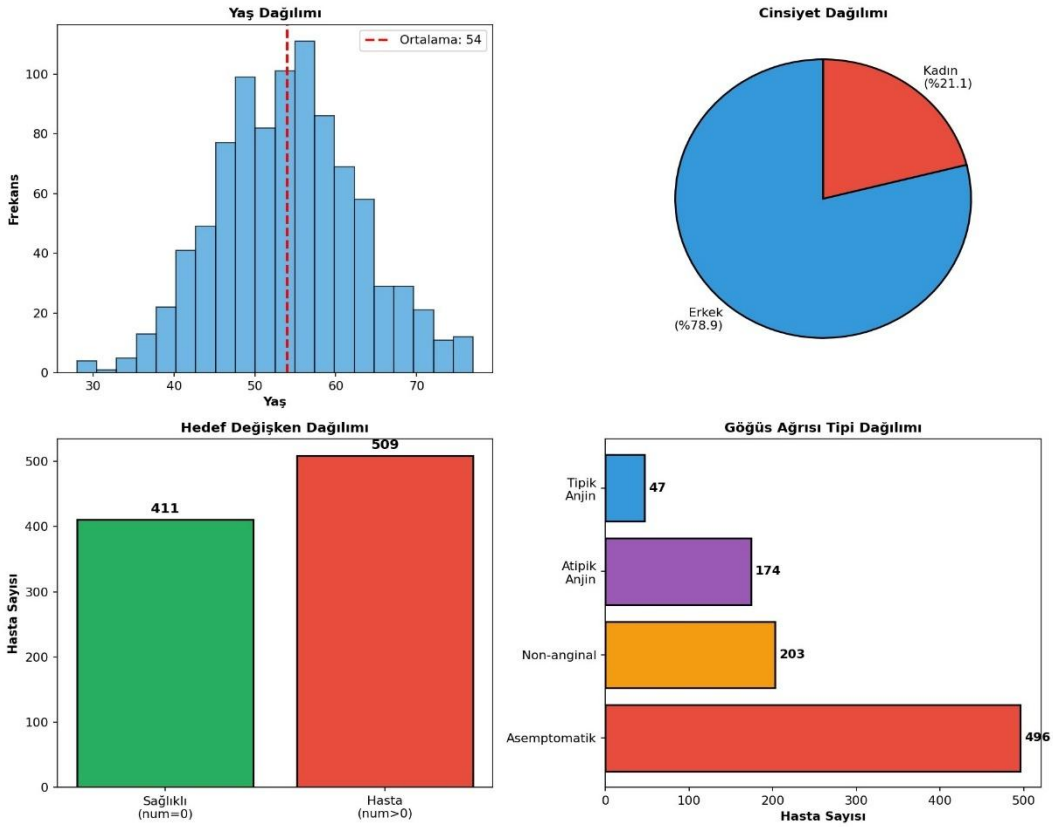
Bu çalışmada, makine öğrenmesi alanında yaygın olarak kullanılan ve uluslararası literatürde kabul görmüş Heart Disease UCI veri seti kullanılmıştır. Veri seti, UCI (University of California, Irvine) Makine Öğrenmesi Deposu'ndan (Machine Learning Repository) elde edilmiş olup kardiyoloji alanındaki araştırmalar için standart bir referans kaynağı olarak kabul edilmektedir.

Heart Disease veri seti, dört farklı tıbbi kuruluştan derlenen hasta kayıtlarını içermektedir. Bu kuruluşlar Cleveland Kliniği (Ohio, ABD), Macaristan Kardiyoloji Enstitüsü (Budapeşte), İsviçre Üniversite Hastanesi (Zürih) ve VA Long Beach Tıp Merkezi (Kaliforniya, ABD) olarak sıralanmaktadır. Her bir veritabanı farklı hasta popülasyonlarını temsil etmekte ve bu çeşitlilik modelin genelleme kapasitesini test etmek açısından önemli bir avantaj sağlamaktadır.

Veri setinin orijinal versiyonu 76 öznitelik içermesine rağmen, literatürdeki çalışmaların büyük çoğunluğu 14 temel öznitelik üzerinde yoğunlaşmaktadır. Bu öznitelikler, klinik pratikte kolay elde edilebilen ve kalp hastalığı teşhisinde anlamlı bilgi sağlayan değişkenlerden oluşmaktadır. Toplam örnek sayısı 920 hasta kaydından oluşmakta olup bu rakam makine öğrenmesi modellerinin eğitimi için yeterli bir veri hacmi sunmaktadır.

3.1.1. Veri Setinin Genel Özellikleri

Veri setinin temel öznitelik dağılımları Şekil 3.1'de görsel olarak sunulmaktadır. Bu grafikler, veri setinin demografik yapısını, hedef değişkenin dağılımını ve



Şekil 3. 1 Veri Seti Öznitelik Dağılımları

önemli klinik parametrelerin karakteristiklerini ortaya koymaktadır. Keşifsel veri analizi (Exploratory Data Analysis EDA) sürecinin ilk adımı olarak bu görselleştirmeler, veri setinin anlaşılmasında kritik rol oynamaktadır.
Şekil 3.1'in sol üst panelinde görülen yaş dağılımı histogramı, veri setindeki hastaların yaşlarının 28 ile 77 arasında değiştiğini göstermektedir. Ortalama yaş 54 olup standart sapma yaklaşık 9 yıldır. Bu dağılım, kalp hastalığı riskinin orta yaş ve üzerinde arttığını gösteren epidemiyolojik verilerle uyumludur. Histogramın çan eğrisine benzer şekli, yaş değişkeninin yaklaşık normal dağılım sergilediğine işaret etmektedir.

Sağ üst paneldeki pasta grafiği, cinsiyet dağılımını göstermektedir. Veri setinde erkek hastalar %78,9 oranında (726 kişi) çoğunluğu oluştururken, kadın hastalar %21,1 oranında (194 kişi) temsil edilmektedir. Bu dengesizlik, kardiyoloji kliniklerindeki hasta profilini yansıtmakta olup erkeklerde kalp hastalığı görülme sıklığının kadınlara göre daha yüksek olduğu gerçeğiyle örtüşmektedir.

Sol alt panelde yer alan hedef değişken dağılımı, veri setindeki sağlıklı ve hasta bireylerin sayısını göstermektedir. Toplamda 411 birey sağlıklı (num=0) olarak sınıflandırılırken, 509 birey kalp hastalığı tanısı almış (num>0) olarak etiketlenmiştir. Bu dağılım, sınıflar arasında belirgin bir dengesizlik olmadığını

ve veri setinin sınıflandırma problemi için uygun bir yapıda olduğunu göstermektedir.

Sağ alt paneldeki yatay çubuk grafiği, göğüs ağrısı tiplerinin dağılımını sergilemektedir. Asemptomatik tip (496 hasta) en yaygın kategori olup bunu non-anginal ağrı (203 hasta), atipik anjin (174 hasta) ve tipik anjin (47 hasta) izlemektedir. Asemptomatik hastaların çokluğu, kalp hastalığının her zaman belirgin semptomlarla kendini göstermediğini ve tarama testlerinin önemini vurgulamaktadır.

3.1.2. Özniteliklerin Detaylı Açıklaması

Veri setinde yer alan 14 öznitelik, Tablo 3.1'de detaylı olarak açıklanmaktadır. Bu öznitelikler, demografik bilgiler (yaş, cinsiyet), semptomlar (göğüs ağrısı tipi), fizyolojik ölçümler (kan basıncı, kalp atış hızı), laboratuvar testleri (kolesterol, kan şekeri) ve kardiyolojik test sonuçları (EKG, stres testi) olmak üzere farklı kategorilerde gruplandırılabilir.

Tablo 3.1 UCI Veri Setindeki Öznitelikler ve Açıklamaları

Öznitelik	Açıklama	Veri Tipi	Değer Aralığı
age	Hastanın yaşı (yıl)	Sayısal	28-77
sex	Cinsiyet	Kategorik	0: Kadın, 1: Erkek
cp	Göğüs ağrısı tipi	Kategorik	0-3
trestbps	Dinlenme kan basıncı (mm Hg)	Sayısal	94-200
chol	Serum kolesterol (mg/dl)	Sayısal	126-564
fbs	Açlık kan şekeri > 120 mg/dl	Kategorik	0: Hayır, 1: Evet
restecg	Dinlenme EKG sonuçları	Kategorik	0-2
thalach	Maksimum kalp atış hızı	Sayısal	71-202
exang	Egzersize bağlı anjin	Kategorik	0: Hayır, 1: Evet
oldpeak	ST depresyonu	Sayısal	0-6.2
slope	ST segment eğimi	Kategorik	0-2
ca	Boyanan damar sayısı	Sayısal	0-4
thal	Talyum stres testi	Kategorik	0-3
target	Hedef değişken	Kategorik	0: Sağlıklı, 1: Hasta

Demografik öznitelikler arasında yaş (age) ve cinsiyet (sex) yer almaktadır. Yaş, kardiyovasküler risk değerlendirmesinde temel bir faktör olup ileri yaş artan riskle ilişkilendirilmektedir. Cinsiyet değişkeni ise erkeklerin kadınlara göre

daha erken yaşlarda kalp hastalığına yakalanma eğiliminde olduğunu yansıtan önemli bir demografik göstergedir.

Semptomatik öznitelikler arasında göğüs ağrısı tipi (cp) öne çıkmaktadır. Bu değişken dört kategoriden oluşmaktadır: tipik anjin (0), atipik anjin (1), non-anginal ağrı (2) ve asemptomatik (3). Tipik anjin, klasik koroner arter hastalığı semptomlarını tanımlarken, asemptomatik hastalar belirgin şikayet olmaksızın hastalık taşıyabilmektedir. Bu ayrım, klinik karar verme sürecinde kritik öneme sahiptir.

Fizyolojik ölçümler kategorisinde dinlenme kan basıncı (trestbps) ve maksimum kalp atış hızı (thalach) bulunmaktadır. Yüksek dinlenme kan basıncı (hipertansiyon), kalp hastalığı için bağımsız bir risk faktörüdür. Egzersiz sırasında ulaşılan maksimum kalp atış hızı ise kardiyak kapasitenin bir göstergesi olup düşük değerler fonksiyonel bozukluğa işaret edebilmektedir.

Laboratuvar testleri arasında serum kolesterol (chol) ve açlık kan şekeri (fbs) yer almaktadır. Yüksek kolesterol seviyeleri, damar sertliği (ateroskleroz) ve koroner arter hastalığı riskini artırmaktadır. Açlık kan şekeri değişkeni ise diyabet varlığını değerlendirmek için kullanılmakta olup diyabet kalp hastalığı için önemli bir risk faktörü olarak kabul edilmektedir.

Kardiyolojik test sonuçları arasında dinlenme EKG (restecg), egzersize bağlı anjin (exang), ST depresyonu (oldpeak), ST segment eğimi (slope), floroskopi ile boyanan damar sayısı (ca) ve talyum stres testi (thal) bulunmaktadır. Bu öznitelikler, kalbin elektriksel aktivitesini ve koroner arter perfüzyonunu değerlendiren objektif test sonuçlarını içermektedir.

ST depresyonu (oldpeak), egzersiz sırasında EKG'de gözlemlenen bir bulgu olup miyokard iskemisinin göstergesi olabilmektedir. Değer arttıkça iskemi şiddeti de artmaktadır. Floroskopi ile boyanan damar sayısı (ca) ise koroner anjiyografi sonuçlarını yansıtmakta ve tıkalı damar sayısı hakkında doğrudan bilgi sağlamaktadır.

3.2. Kalp Hastalıklarının Önemi ve Teşhisi

Kalp hastalıkları, modern tıbbın en önemli sağlık sorunlarından birini teşkil etmekte ve dünya genelinde önde gelen ölüm nedenlerinin başında gelmektedir. Bu bölümde, kalp hastalıklarının tanımı, sınıflandırması, risk faktörleri ve epidemiyolojik verileri ele alınmaktadır. Bu bilgiler, çalışmanın motivasyonunu ve klinik önemini ortaya koymaktadır.

3.2.1. Kalp Hastalıkları Hakkında Genel Bilgi

Kalp hastalıkları, kalp ve kan damarlarını etkileyen çeşitli patolojik durumları kapsayan geniş bir terimdir. Kardiyovasküler hastalıklar (Cardiovascular Diseases - CVD) olarak da bilinen bu hastalık grubu, farklı etyolojilere ve klinik tablolara sahip çok sayıda alt tipi içermektedir. Bu hastalıkların ortak özelliği, dolaşım sisteminin normal işlevini bozarak yaşamsal organ perfüzyonunu tehlikeye atmalarıdır.

Kardiyovasküler hastalıkların başlıca türleri şunlardır: Koroner arter hastalığı (Coronary Artery Disease - CAD), kalbe kan sağlayan arterlerin daralması veya tıkanması sonucu ortaya çıkmaktadır. Kalp yetmezliği, kalbin vücudun ihtiyaç duyduğu kan miktarını pompalayamadığı kronik bir durumdur. Aritmiler, kalbin elektriksel iletim sistemindeki bozukluklar sonucu gelişen düzensiz kalp atımlarıdır. Kalp kapak hastalıkları, kalp kapaklarının yapısal veya fonksiyonel bozukluklarını içermektedir.

Koroner arter hastalığı, en yaygın ve en ölümcül kalp hastalığı türüdür. Bu hastalık, koroner arterlerin iç duvarlarında aterosklerotik plak birikimi sonucu gelişmektedir. Plak birikimi, damar lümenini daraltarak miyokarda ulaşan kan akışını azaltmaktadır. İleri evrelerde plak rüptürü ve üzerine trombüs oluşumu, akut miyokard enfarktüsüne (kalp krizi) yol açabilmektedir.

Kalp hastalıklarının klinik belirtileri hastalığın tipine ve şiddetine göre değişkenlik göstermektedir. En sık görülen semptomlar arasında göğüs ağrısı veya göğüste baskı hissi (anjina pektoris), nefes darlığı (dispne), çarpıntı, baş dönmesi, bayılma (senkop), yorgunluk ve bacaklarda şişme (ödem) sayılabilir. Ancak bazı hastalar, özellikle diyabetik bireyler ve yaşlılar, atipik semptomlar gösterebilir veya tamamen asemptomatik olabilir.

3.2.2. Risk Faktörleri

Kalp hastalıklarının risk faktörleri, değiştirilebilir ve değiştirilemez olmak üzere iki ana kategoride incelenmektedir. Değiştirilemez risk faktörleri, bireyin kontrol edemediği genetik ve demografik özelliklerdir. Değiştirilebilir risk faktörleri ise yaşam tarzı değişiklikleri ve medikal tedaviler ile modifiye edilebilen faktörlerdir.

Değiştirilemez risk faktörleri arasında yaş, cinsiyet ve aile öyküsü yer almaktadır. Yaş ilerledikçe kardiyovasküler hastalık riski artmaktadır; erkeklerde 45 yaş, kadınlarda 55 yaş üzeri risk artışıyla ilişkilendirilmektedir. Erkekler, özellikle orta yaşlarda, kadınlara göre daha yüksek kalp hastalığı riskine sahiptir. Ancak menopoz sonrası dönemde kadınlarda risk hızla artarak

erkeklere yaklaşmaktadır. Birinci derece akrabalarında erken yaşta kalp hastalığı öyküsü olan bireylerde genetik yatkınlık söz konusudur.

Değiştirilebilir risk faktörleri arasında hipertansiyon, dislipidemi, diyabet, obezite, sigara kullanımı, fiziksel inaktivite ve sağlıksız beslenme bulunmaktadır. Hipertansiyon (yüksek kan basıncı), kalp üzerindeki iş yükünü artırarak sol ventrikül hipertrofisine ve kalp yetmezliğine yol açabilmektedir. Dislipidemi, özellikle yüksek LDL kolesterol ve düşük HDL kolesterol seviyeleri, ateroskleroz gelişimini hızlandırmaktadır.

Tip 2 diyabet, kardiyovasküler hastalık riskini 2-4 kat artıran önemli bir metabolik bozukluktur. Obezite, özellikle abdominal obezite, insülin direnci, hipertansiyon ve dislipidemi ile ilişkili olup metabolik sendrom bileşenlerinden birini oluşturmaktadır. Sigara kullanımı, endotel disfonksiyonu, inflamasyon ve tromboz eğilimini artırarak kardiyovasküler riski önemli ölçüde yükseltmektedir.

3.2.3. İstatistiksel Verilerle Kalp Hastalıkları

Dünya Sağlık Örgütü (World Health Organization- WHO) verilerine göre, kardiyovasküler hastalıklar dünya genelinde bir numaralı ölüm nedenidir [1]. Her yıl tahminen 17,9 milyon kişi bu hastalıklar nedeniyle hayatını kaybetmekte olup bu rakam tüm küresel ölümlerin yaklaşık %31'ini oluşturmaktadır. Bu ölümlerin %85'i kalp krizi ve inme kaynaklıdır.

Coğrafi dağılım açısından değerlendirildiğinde, kardiyovasküler hastalıklardan ölümlerin %80'inden fazlası düşük ve orta gelirli ülkelerde gerçekleşmektedir. Bu durum, sağlık hizmetlerine erişim, erken tanı imkanları ve tedavi olanaklarındaki eşitsizlikleri yansıtmaktadır. Gelişmiş ülkelerde son dekatlarda kardiyovasküler mortalite oranlarında düşüş gözlenirken, gelişmekte olan ülkelerde artış eğilimi devam etmektedir.

Türkiye'de ise Türkiye İstatistik Kurumu (TÜİK) verilerine göre dolaşım sistemi hastalıkları, ölüm nedenlerinin başında gelmektedir [2]. 2022 yılı verilerine göre, toplam ölümlerin yaklaşık %35'i dolaşım sistemi hastalıklarından kaynaklanmaktadır. Yıllık ortalama 168.000 kişi kalp ve damar hastalıkları nedeniyle yaşamını yitirmektedir. Bu rakamlar, Türkiye'de kalp hastalıklarının önlenmesi ve erken teşhisi konusundaki çalışmaların önemini açıkça ortaya koymaktadır.

Ekonomik açıdan değerlendirildiğinde, kardiyovasküler hastalıklar sağlık sistemleri üzerinde büyük bir yük oluşturmaktadır. Tedavi maliyetleri, iş gücü kayıpları ve sosyal güvenlik harcamaları göz önüne alındığında, bu hastalıkların toplumsal maliyeti trilyonlarca doları bulmaktadır. Erken tanı ve önleyici

tedbirler, hem bireysel hem de toplumsal düzeyde önemli kazanımlar sağlayabilmektedir.

3.3. Veri Setine Uygulanan Ön İşlemler

Makine öğrenmesi modellerinin başarılı ve güvenilir sonuçlar üretebilmesi için veri setinin uygun şekilde ön işlenmesi kritik öneme sahiptir. Ham veriler genellikle eksik değerler, aykırı değerler, tutarsızlıklar ve ölçek farklılıkları gibi sorunlar içermektedir. Bu bölümde, veri setine uygulanan ön işleme adımları detaylı olarak açıklanmaktadır.

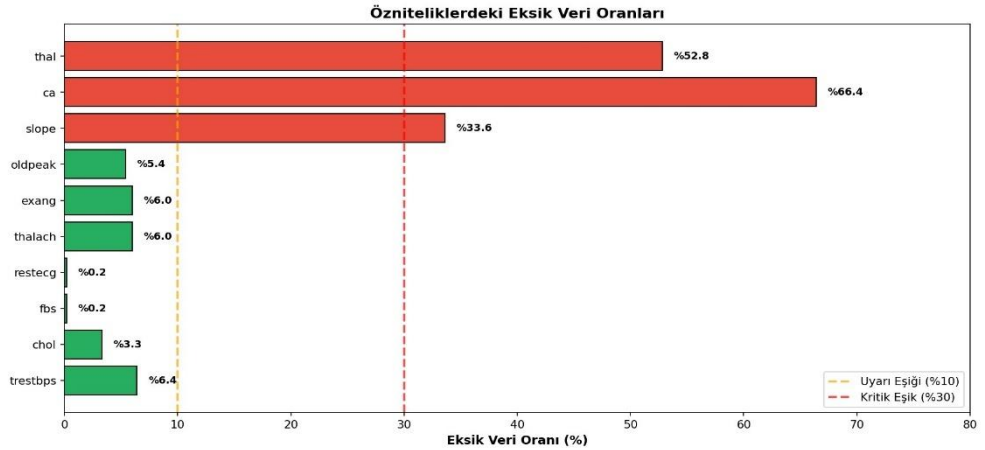
Veri ön işleme süreci, veri kalitesini artırmak ve model performansını optimize etmek amacıyla gerçekleştirilen bir dizi sistematik adımdan oluşmaktadır. Bu adımlar arasında eksik veri analizi ve doldurma, aykırı değer tespiti ve işleme, kategorik değişkenlerin kodlanması, öznitelik ölçeklendirme ve sınıf dengesizliğinin giderilmesi yer almaktadır. Her bir adım, modelin öğrenme kapasitesini ve genelleme yeteneğini doğrudan etkilemektedir.

3.3.1. Eksik Veri Analizi ve Doldurma Yöntemleri

Eksik veriler (missing values), gerçek dünya veri setlerinde sıklıkla karşılaşılan bir sorundur. Eksik değerler, veri toplama sürecindeki hatalar, ölçüm cihazı arızaları, katılımcı yanıt vermeme durumları veya veri aktarım problemleri gibi çeşitli nedenlerle ortaya çıkabilmektedir. Eksik verilerin uygun şekilde işlenmemesi, model performansını olumsuz etkileyebilir ve hatalı tahminlere yol açabilir.

Eksik veri mekanizmaları üç kategoride incelenmektedir: Tamamen Rastgele Eksik (Missing Completely at Random- MCAR), Rastgele Eksik (Missing at Random- MAR) ve Rastgele Olmayan Eksik (Missing Not at Random- MNAR). MCAR durumunda eksiklik herhangi bir değişkenle ilişkili değildir. MAR durumunda eksiklik gözlemlenen diğer değişkenlerle açıklanabilir. MNAR durumunda ise eksiklik, eksik değerin kendisiyle ilişkilidir. Eksik veri mekanizmasının doğru tanımlanması, uygun doldurma yönteminin seçilmesinde kritik öneme sahiptir.

Şekil 3.2'de veri setindeki özniteliklerin eksik veri oranları görsel olarak sunulmaktadır. Bu grafik, hangi özniteliklerde ne ölçüde eksik veri bulunduğunu açıkça ortaya koymakta ve doldurma stratejisinin belirlenmesinde yol gösterici olmaktadır.



Şekil 3. 2 Özniteliklerdeki Eksik Veri Oranları

Şekil 3.2'de görüldüğü üzere, ca (floroskopi ile boyanan damar sayısı) özniteliğinde %66,4, thal (talyum stres testi) özniteliğinde %52,8 ve slope (ST segment eğimi) özniteliğinde %33,6 oranında eksik veri bulunmaktadır. Bu yüksek oranlar, basit silme veya ortalama ile doldurma yöntemlerinin yetersiz kalacağını göstermekte ve daha sofistike doldurma tekniklerinin kullanılmasını gerektirmektedir.

Eksik verilerin doldurulması için bu çalışmada makine öğrenmesi tabanlı yöntemler tercih edilmiştir. Kategorik değişkenler için Lojistik Regresyon ve K-En Yakın Komşu (KNN) Imputer kullanılmıştır. Lojistik Regresyon, eksik kategorik değerleri diğer özniteliklere dayalı olarak tahmin etmektedir. KNN Imputer ise eksik değere sahip örneğin en yakın komşularının değerlerini kullanarak doldurma işlemi gerçekleştirmektedir.

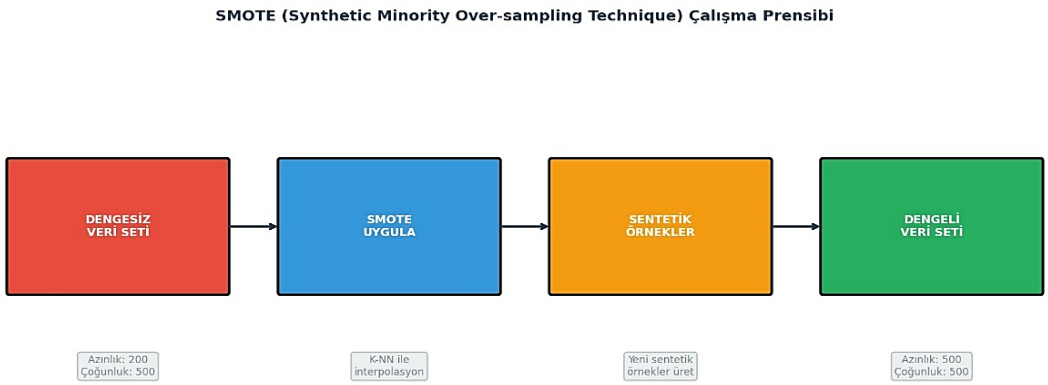
Sayısal değişkenler için Rastgele Orman Regresyonu (Random Forest Regressor) ve KNN Imputer yöntemleri uygulanmıştır. Rastgele Orman, topluluk öğrenme (ensemble learning) yaklaşımıyla yüksek doğrulukta tahminler üretebilmektedir. Her iki yöntemde de GridSearchCV kullanılarak hiperparametreler optimize edilmiş ve doldurma işleminin doğruluğu maksimize edilmiştir.

3.3.2. Veri Dengesizliği ve SMOTE Yöntemi

Sınıflandırma problemlerinde sınıf dengesizliği (class imbalance), veri setinde bir sınıfın diğer sınıf(lar)a göre çok daha fazla örneğe sahip olması durumudur. Bu dengesizlik, makine öğrenmesi modellerinin çoğunluk sınıfına yönelik yanlı (biased) tahminler üretmesine neden olabilmektedir. Model, azınlık sınıfını yeterince öğrenemeyeceği için bu sınıfa ait örnekleri yanlış sınıflandırma eğiliminde olacaktır.

Sınıf dengesizliğini gidermek için çeşitli teknikler kullanılmaktadır. Bu teknikler genel olarak üç kategoride incelenebilir: Aşırı örnekleme (over-sampling), alt örnekleme (under-sampling) ve hibrit yöntemler. Aşırı örnekleme yöntemlerinde azınlık sınıfı örnekleri çoğaltılırken, alt örnekleme yöntemlerinde çoğunluk sınıfı örnekleri azaltılmaktadır. Hibrit yöntemler ise her iki yaklaşımı birleştirmektedir.

SMOTE (Synthetic Minority Over-sampling Technique), Chawla ve arkadaşları tarafından 2002 yılında önerilen ve yaygın olarak kullanılan bir aşırı örnekleme tekniğidir. SMOTE, basit kopyalama yerine sentetik (yapay) örnekler üreterek azınlık sınıfını zenginleştirmektedir [8]. Bu yaklaşım, modelin aşırı öğrenmesini (overfitting) önlemeye yardımcı olmaktadır. Şekil 3.3'te SMOTE algoritmasının çalışma prensibi şematik olarak gösterilmektedir.



Şekil 3. 3 SMOTE (Synthetic Minority Over-sampling Technique) Çalışma Prensibi

SMOTE algoritması şu adımlarla çalışmaktadır: İlk olarak, azınlık sınıfından bir örnek seçilir. Ardından, bu örneğin k-en yakın komşuları (genellikle k=5) belirlenir. Seçilen örnekle komşularından biri arasındaki doğru üzerinde rastgele bir nokta belirlenerek yeni sentetik örnek oluşturulur. Bu işlem, istenen dengeye ulaşılana kadar tekrarlanır.

Bu çalışmada, SMOTE yöntemi yalnızca eğitim verisi üzerinde uygulanmıştır. Test verisinin orijinal dağılımının korunması, modelin gerçek dünya performansının doğru değerlendirilmesi açısından kritik öneme sahiptir. Eğitim setinde SMOTE uygulandıktan sonra, azınlık ve çoğunluk sınıfları eşit sayıda örneğe sahip olmuş ve dengeli bir eğitim seti elde edilmiştir.

3.3.3. Öznitelik Ölçeklendirme

Makine öğrenmesi algoritmalarının çoğu, özniteliklerin benzer ölçeklerde olmasını varsaymaktadır. Farklı ölçeklerdeki öznitelikler, özellikle mesafe tabanlı algoritmalar (KNN, SVM) ve gradyan tabanlı optimizasyon kullanan algoritmalar için sorun oluşturabilmektedir. Bu nedenle, öznitelik ölçeklendirme (feature scaling) ön işleme sürecinin önemli bir adımıdır.

Yaygın kullanılan ölçeklendirme yöntemleri arasında Min-Max Normalizasyonu ve Standartlaştırma (Z-score normalizasyonu) bulunmaktadır. Min-Max Normalizasyonu, değerleri belirli bir aralığa (genellikle 0-1) dönüştürürken, Standartlaştırma değerleri ortalama 0 ve standart sapma 1 olacak şekilde dönüştürmektedir. Bu çalışmada StandardScaler yöntemi kullanılarak öznitelik ölçeklendirmesi yapılmıştır. Her öznitelik için $x_{\{scaled\}} = \frac{x - \mu}{\sigma}$ uygulanmıştır; burada μ özniteliğin ortalamasını, σ ise standart sapmasını temsil etmektedir. Ölçeklendirme parametreleri (ortalama ve standart sapma) yalnızca eğitim verisinden hesaplanmış ve aynı parametreler test verisine uygulanmıştır. Bu yaklaşım, veri sızıntısını (data leakage) önlemek için kritik öneme sahiptir.

3.4. Kullanılan Araçlar ve Kütüphaneler

Bu çalışmada Python programlama dili ve çeşitli açık kaynak kütüphaneler kullanılmıştır. Python, veri bilimi ve makine öğrenmesi alanlarında en yaygın kullanılan programlama dillerinden biri olup zengin kütüphane ekosistemi ve geniş topluluk desteği ile öne çıkmaktadır. Şekil 3.4'te kullanılan yazılım araçları kategorilerine göre gösterilmektedir.



Şekil 3. 4 Çalışmada Kullanılan Yazılım Araçları ve Kütüphaneler

3.4.1. Veri İşleme Kütüphaneleri

NumPy (Numerical Python): NumPy, Python'da bilimsel hesaplamalar için temel kütüphanedir. Çok boyutlu diziler (ndarray) ve bu diziler üzerinde hızlı matematiksel işlemler yapılmasını sağlamaktadır. Vektörize işlemler sayesinde

döngü kullanımına göre çok daha hızlı hesaplamalar gerçekleştirilebilmektedir. Bu çalışmada NumPy, matris işlemleri ve sayısal hesaplamalar için kullanılmıştır.

Pandas: Pandas, veri manipülasyonu ve analizi için geliştirilmiş güçlü bir kütüphanedir. DataFrame ve Series veri yapıları, tablo formatındaki verilerin etkin bir şekilde işlenmesini sağlamaktadır. Veri okuma/yazma, filtreleme, gruplama, birleştirme ve dönüştürme gibi işlemler Pandas ile kolayca gerçekleştirilebilmektedir. Bu çalışmada veri setinin yüklenmesi, temizlenmesi, keşifsel analizi ve dönüştürülmesi işlemlerinde Pandas yoğun olarak kullanılmıştır.

3.4.2. Görselleştirme Kütüphaneleri

Matplotlib: Matplotlib, Python'da veri görselleştirme için en temel kütüphanedir. Çizgi grafikleri, histogramlar, saçılım grafikleri, pasta grafikleri ve daha birçok görselleştirme türünü desteklemektedir. Yüksek özelleştirme imkanı sunması sayesinde yayın kalitesinde grafikler oluşturulabilmektedir.

Seaborn: Seaborn, Matplotlib üzerine inşa edilmiş üst düzey bir istatistiksel görselleştirme kütüphanesidir. Daha estetik varsayılan stiller ve karmaşık istatistiksel grafiklerin kolay oluşturulması için tasarlanmıştır. Isı haritaları (heatmap), kutu grafikleri (boxplot) ve çift değişkenli dağılım grafikleri (pairplot) gibi görselleştirmeler Seaborn ile kolayca oluşturulabilmektedir.

Plotly: Plotly, interaktif görselleştirmeler oluşturmak için kullanılan modern bir kütüphanedir. Web tabanlı grafikler üretmesi sayesinde kullanıcıların grafiklerle etkileşime girmesine (yakınlaştırma, kaydırma, veri noktası bilgisi görme) olanak tanımaktadır. Bu çalışmada keşifsel veri analizi sırasında interaktif grafikler için Plotly kullanılmıştır.

3.4.3. Makine Öğrenmesi Kütüphaneleri

Scikit-learn: Scikit-learn, Python'da makine öğrenmesi için en yaygın kullanılan kütüphanedir. Sınıflandırma, regresyon, kümeleme, boyut indirgeme, model seçimi ve ön işleme gibi geniş bir yelpazede algoritmalar sunmaktadır. Tutarlı API tasarımı sayesinde farklı algoritmalar arasında kolayca geçiş yapılabilmektedir. Bu çalışmada Lojistik Regresyon, K-En Yakın Komşu, Destek Vektör Makineleri ve Rastgele Orman algoritmaları Scikit-learn ile uygulanmıştır.

XGBoost: XGBoost (Extreme Gradient Boosting), gradyan artırma algoritmasının optimize edilmiş ve ölçeklenebilir bir implementasyonudur. Paralel işlem desteği, düzenlileştirme (regularization) mekanizmaları ve eksik

değer yönetimi gibi özellikler sunmaktadır. Kaggle yarışmalarında ve endüstriyel uygulamalarda yaygın olarak tercih edilmektedir.

LightGBM: LightGBM (Light Gradient Boosting Machine), Microsoft tarafından geliştirilen hızlı ve yüksek performanslı bir gradyan artırma kütüphanesidir. Yaprak bazlı (leaf-wise) ağaç büyütme stratejisi ve histogram tabanlı algoritması sayesinde büyük veri setlerinde XGBoost'a göre daha hızlı eğitim süreleri sunabilmektedir.

Imbalanced-learn: Imbalanced-learn, sınıf dengesizliği problemlerini ele almak için geliştirilmiş bir kütüphanedir. SMOTE, ADASYN, Random Under Sampling gibi çeşitli örnekleme tekniklerini içermektedir. Scikit-learn ile uyumlu API tasarımı sayesinde makine öğrenmesi pipeline'larına kolayca entegre edilebilmektedir.

3.4.4. Geliştirme Ortamı ve Donanım

Çalışmanın tüm aşamaları Python 3.9+ sürümü kullanılarak gerçekleştirilmiştir. Geliştirme ortamı olarak Jupyter Notebook tercih edilmiştir. Jupyter Notebook, kod, çıktı ve açıklamaların bir arada sunulmasına olanak tanıyan interaktif bir geliştirme ortamıdır. Keşifsel veri analizi ve deneysel çalışmalar için ideal bir platform sunmaktadır.

Donanım olarak minimum 8 GB RAM ve Intel Core i5 veya eşdeğeri işlemci gereksinimi bulunmaktadır. Gradyan artırma algoritmaları (XGBoost, LightGBM) için GPU desteği opsiyonel olarak kullanılabilmektedir. İşletim sistemi olarak Windows 10/11 veya Linux (Ubuntu 20.04+) desteklenmektedir.

3.4.5. Veri Bölme Stratejisi

Model eğitimi ve değerlendirmesi için veri seti %80 eğitim (training) ve %20 test olmak üzere ikiye ayrılmıştır. Bu oran, makine öğrenmesi literatüründe yaygın olarak kabul gören standart bir bölme oranıdır. Yeterli eğitim verisi sağlarken, model performansının güvenilir bir şekilde değerlendirilmesi için yeterli test verisi de ayrılmaktadır.

Veri bölme işlemi sırasında stratified sampling (tabakalı örnekleme) yöntemi kullanılmıştır. Bu yöntem, her iki sette de (eğitim ve test) hedef değişkenin orijinal dağılımının korunmasını sağlamaktadır. Özellikle sınıf dengesizliği durumlarında, tabakalı örnekleme kullanılmazsa test setinde azınlık sınıfının yetersiz temsil edilmesi riski bulunmaktadır.

Tekrarlanabilirlik için random_state parametresi sabit bir değere (42) ayarlanmıştır. Bu sayede, aynı kod tekrar çalıştırıldığında aynı veri bölümü elde

edilmekte ve sonuçların tutarlılığı sağlanmaktadır. Bilimsel araştırmalarda tekrarlanabilirlik, sonuçların doğrulanabilmesi açısından kritik öneme sahiptir.

Model hiperparametre optimizasyonu için 5 katlı çapraz doğrulama (5-fold cross-validation) kullanılmıştır. Çapraz doğrulama, veri setinin beş eşit parçaya bölünmesi ve her parçanın sırayla doğrulama seti olarak kullanılması prensibine dayanmaktadır. Bu yaklaşım, model performansının daha güvenilir bir şekilde tahmin edilmesini ve aşırı öğrenme riskinin azaltılmasını sağlamaktadır.

4. UYGULAMA VE BULGULAR

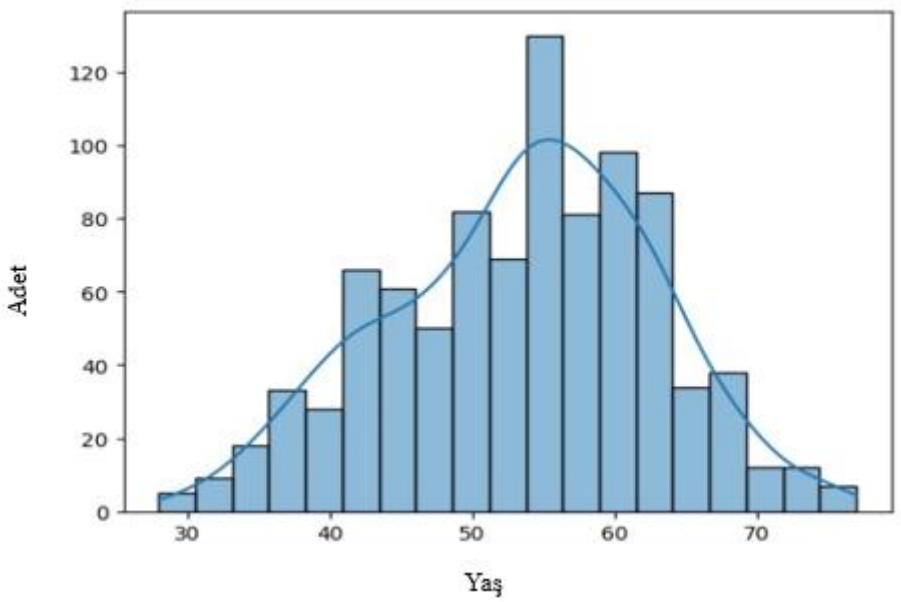
Bu bölümde, Heart Disease UCI veri seti üzerinde gerçekleştirilen keşifsel veri analizi (Exploratory Data Analysis- EDA) sonuçları ve makine öğrenmesi modellerinin performans değerlendirmeleri sunulmaktadır [3]. Tüm analizler Python programlama dili kullanılarak gerçekleştirilmiş olup görselleştirmeler Matplotlib ve Plotly kütüphaneleri ile oluşturulmuştur.

4.1. Keşifsel Veri Analizi (Exploratory Data Analysis)

Keşifsel veri analizi (EDA), veri setinin yapısını, dağılımlarını ve değişkenler arasındaki ilişkileri anlamak için gerçekleştirilen sistematik bir inceleme sürecidir. Bu aşamada veri setindeki her bir öznitelik detaylı olarak analiz edilmiş ve görselleştirilmiştir.

4.1.1. Yaş Dağılımı Analizi

Yaş değişkeni, kalp hastalığı riskini etkileyen önemli faktörlerden biridir. Veri setindeki hastaların yaş dağılımı Şekil 4.1'de histogram grafiği ile gösterilmektedir.



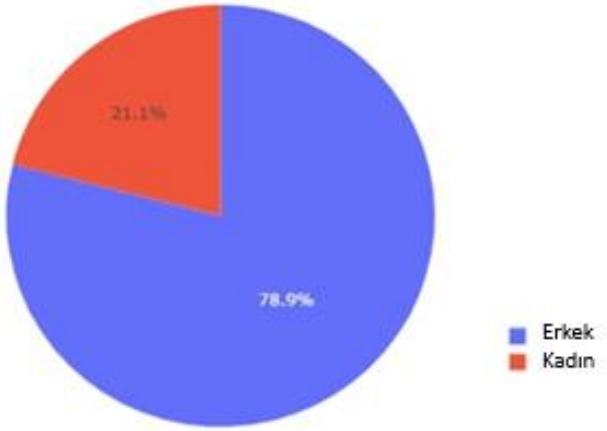
Şekil 4. 1 Veri Setindeki Hastaların Yaş Dağılımı Histogramı

Şekil 4.1'de görüldüğü üzere, veri setindeki hastaların yaşları 28 ile 77 arasında değişmektedir. Dağılım incelendiğinde, en yoğun hasta grubunun 50-60 yaş aralığında olduğu görülmektedir. Ortalama yaş yaklaşık 54-55 civarında olup

dağılım yaklaşık olarak normal dağılıma benzemektedir. Bu bulgular, kalp hastalığının orta yaş ve üzeri bireylerde daha yaygın görüldüğüne ilişkin tıbbi literatürü desteklemektedir.

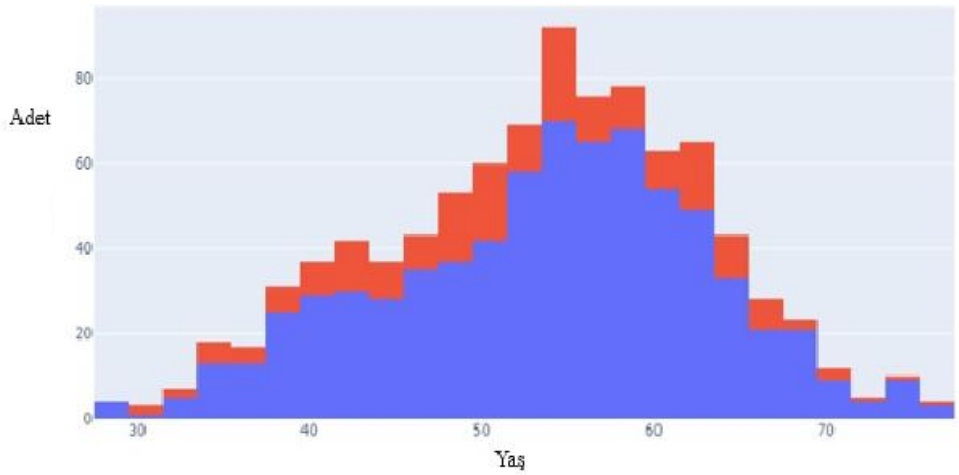
4.1.2. Cinsiyet Dağılımı Analizi

Cinsiyet, kalp hastalıklarında önemli bir demografik faktördür. Veri setindeki cinsiyet dağılımı Şekil 4.2'de pasta grafiği ile gösterilmektedir.



Şekil 4. 2 Veri Setindeki Cinsiyet Dağılımı

Analiz sonuçlarına göre veri setindeki erkek oranı %78,9 iken kadın oranı %21,1'dir. Erkekler, kadınlara göre yaklaşık 3,74 kat daha fazla temsil edilmektedir. Bu dengesizlik, kalp hastalıklarının erkeklerde daha sık görülmesi ve erkeklerin kardiyoloji kliniklerine daha fazla başvurması ile açıklanabilir. Ancak model eğitimi sırasında bu dengesizliğin göz önünde bulundurulması gerekmektedir.



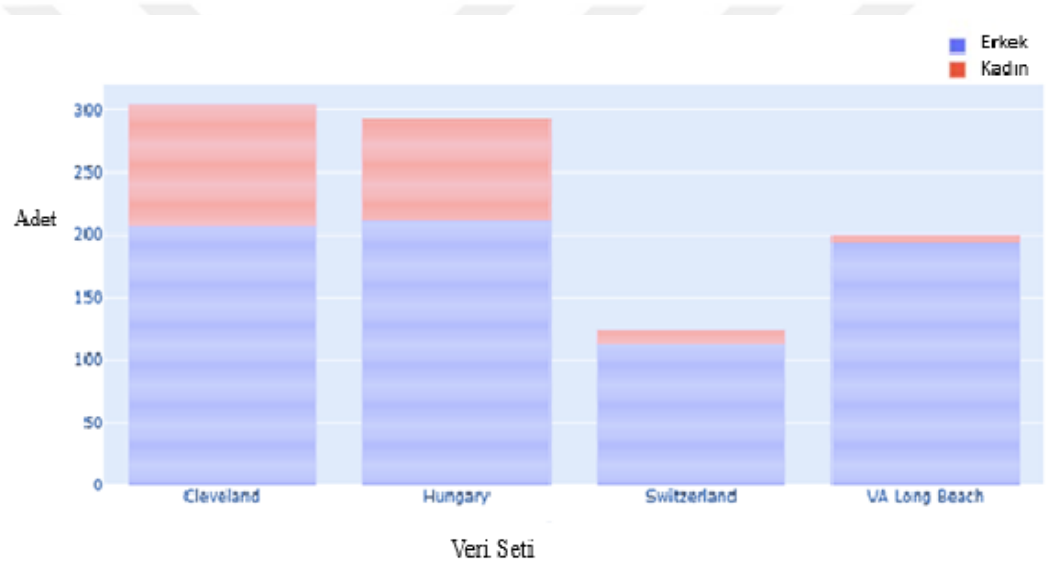
Şekil 4. 3 Cinsiyete Göre Yaş Dağılımı

Şekil 4.3'te cinsiyete göre yaş dağılımı incelendiğinde hem erkek hem de kadın hastalarda benzer yaş dağılım örüntüsü gözlemlenmektedir. Her iki cinsiyette de 50-60 yaş aralığı en yoğun hasta grubunu oluşturmaktadır.

4.1.3. Cinsiyete Göre Hastalık Durumu

Veri setinde hastaların farklı veritabanlarından (Cleveland, Hungary, Switzerland, VA Long Beach) geldiği görülmektedir. Şekil 4.4'te her bir veritabanındaki cinsiyet dağılımı gösterilmektedir.

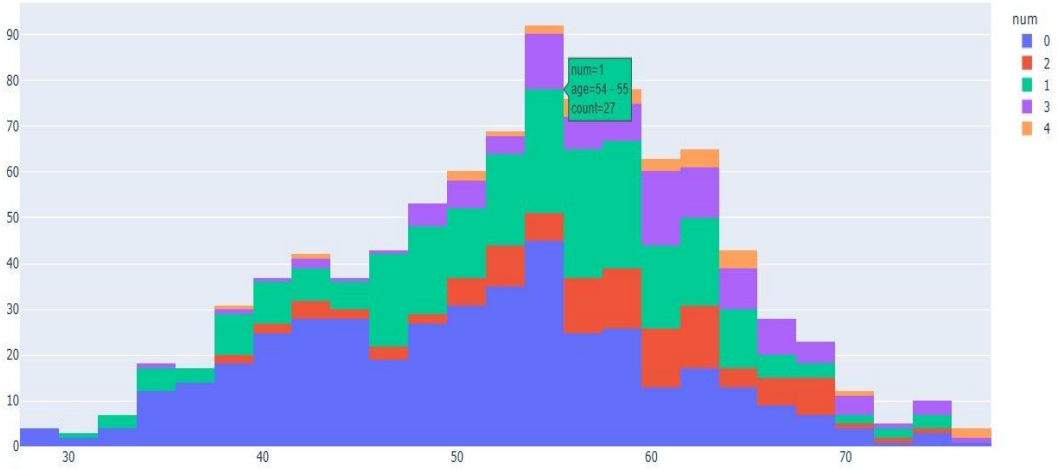
Analiz sonuçlarına göre en fazla hasta kaydı Cleveland (304) veritabanından gelmekte olup en az kayıt Switzerland (123) veritabanındadır. Cleveland veritabanında 97 kadın hasta bulunurken, VA Long Beach veritabanında sadece 6 kadın hasta mevcuttur. Bu dağılım farklılıkları, veritabanlarının toplandığı dönem ve coğrafi konum ile ilişkili olabilir.



Şekil 4. 4 Veritabanına Göre Cinsiyet Dağılımı

4.1.4. Maksimum Kalp Atış Hızı ve Yaş İlişkisi

Maksimum kalp atış hızı (thalach), egzersiz stres testi sırasında ölçülen en yüksek kalp atış hızını ifade etmektedir. Bu değişken, kalp sağlığının önemli bir göstergesidir. Şekil 4.5'te yaş ve hastalık durumuna göre dağılım gösterilmektedir.



Şekil 4. 5 Yaş ve Hastalık Şiddeti Dağılımı

Şekil 4.5'te hastalık şiddeti (num değişkeni: 0=sağlıklı, 1-4=artan hastalık şiddeti) ve yaş ilişkisi incelenmektedir. Grafikten görüldüğü üzere, 54-55 yaş grubunda en yüksek hasta sayısı gözlemlenmektedir. Hastalık şiddeti arttıkça (renk değişimi ile gösterilen) hasta sayısı genel olarak azalmaktadır, bu da ciddi kalp hastalığının nispeten daha nadir olduğunu göstermektedir.

4.1.5. ST Segment Eğimi Analizi

ST segmenti, elektrokardiyogramda (EKG) önemli bir bölge olup kalp kasının oksijen durumunu yansıtmaktadır. Pik egzersiz sırasında ST segmentinin eğimi (slope), kalp hastalığının tanısında kullanılan önemli bir parametredir. Bu öznitelik üç kategori içermektedir: yukarı eğimli (upsloping), düz (flat) ve aşağı eğimli (downsloping).

Yukarı eğimli ST segmenti genellikle normal kabul edilirken, düz ve aşağı eğimli ST segmentleri kalp hastalığı açısından daha yüksek risk göstergesi olarak değerlendirilmektedir.

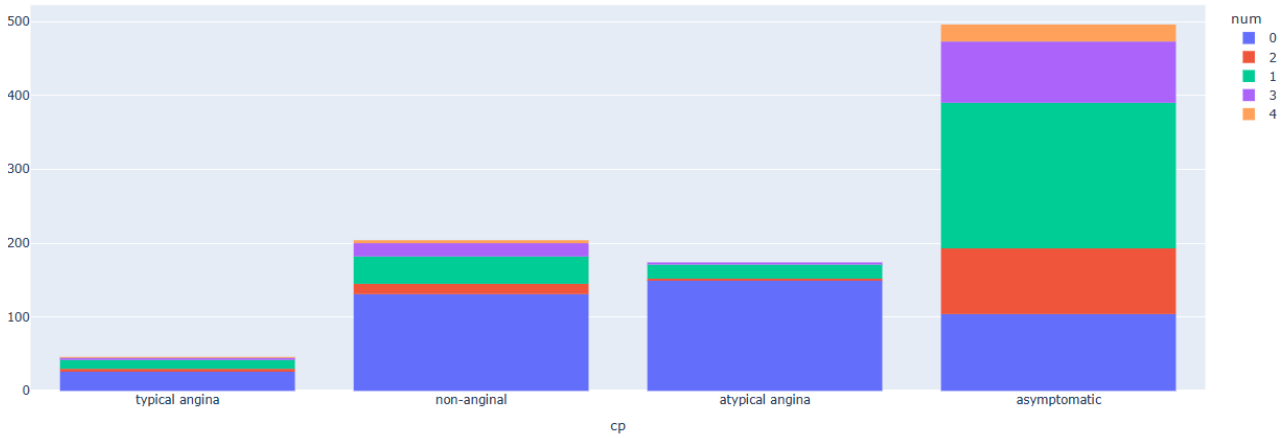
4.1.6. Açlık Kan Şekeri ve Hastalık İlişkisi

Açlık kan şekeri (fbs) değişkeni, hastanın açlık kan şekeri seviyesinin 120 mg/dl'nin üzerinde olup olmadığını göstermektedir. Yüksek kan şekeri (hiperglisemi), diyabet belirtisi olup kalp hastalığı için önemli bir risk faktörüdür.

Veri setinde açlık kan şekeri 120 mg/dl'nin üzerinde olan hastaların oranı incelendiğinde, bu hastaların kalp hastalığına yakalanma riskinin daha yüksek olduğu gözlemlenmiştir. Bu bulgu, diyabet ve kalp hastalığı arasındaki bilinen ilişkiyi desteklemektedir.

4.1.7. Göğüs Ağrısı Tipine Göre Hastalık Sıklığı

Göğüs ağrısı tipi (cp), kalp hastalığı teşhisinde kullanılan kritik semptomlardan biridir. Veri setinde dört farklı göğüs ağrısı tipi tanımlanmıştır: tipik anjin (typical angina), atipik anjin (atypical angina), anjinal olmayan ağrı (non-anginal) ve asemptomatik (asymptomatic).



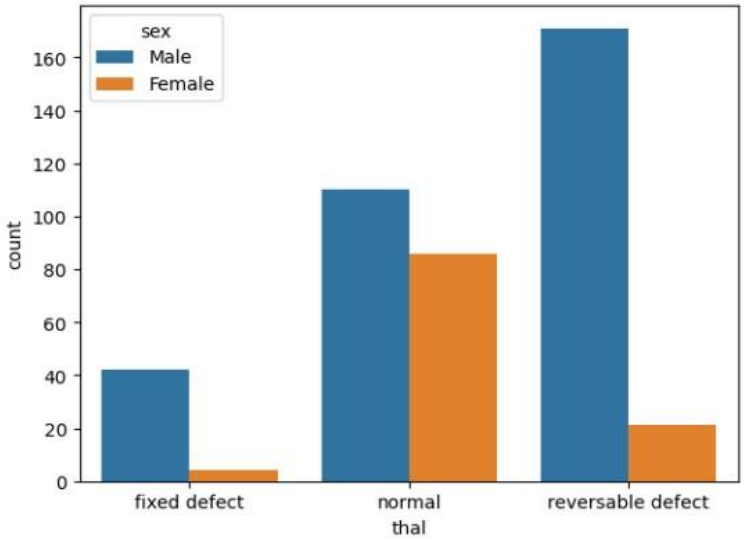
Şekil 4. 6 Göğüs Ağrısı Tipine Göre Hastalık Dağılımı

Şekil 4.6'da görüldüğü üzere, asemptomatik (asymptomatic) grup en fazla hasta sayısına sahiptir ve bu gruptaki hastaların büyük çoğunluğu kalp hastalığı tanısı almıştır. Bu bulgu, kalp hastalığının her zaman belirgin semptomlarla kendini göstermediğini ve asemptomatik bireylerde bile ciddi kalp sorunlarının olabileceğini ortaya koymaktadır.

Analiz sonuçlarına göre 104 birey göğüs ağrısı ve kalp hastalığı olmaksızın sağlıklı olarak sınıflandırılmıştır. Öte yandan, 83 birey göğüs ağrısı olmaksızın ciddi kalp hastalığı tanısı almıştır. Bu durum, kalp hastalığının sessiz seyredebileceğini ve düzenli taramaların önemini vurgulamaktadır.

4.1.8. Thalassemia Analizi

Talyum stres testi (thal), kalbin kan akışını değerlendirmek için kullanılan bir görüntüleme yöntemidir. Bu test sonucu üç kategori içermektedir: normal, sabit defekt (fixed defect) ve geri dönüşümlü defekt (reversable defect).

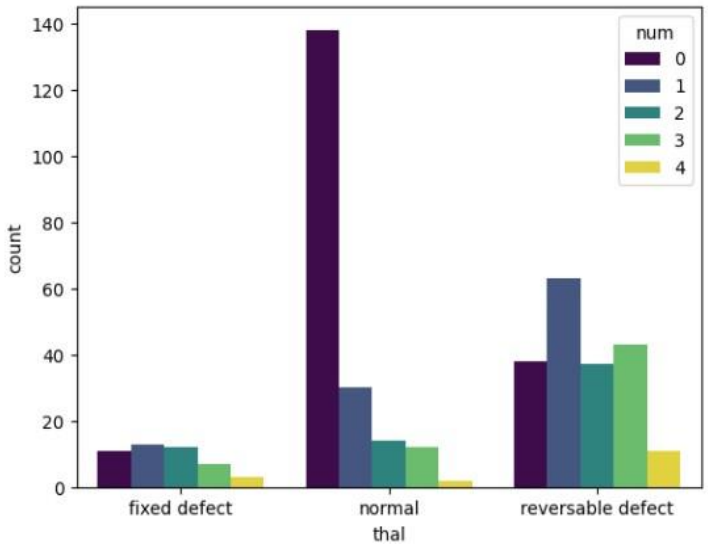


Şekil 4. 7 Talyum Testi Sonuçlarının Cinsiyete Göre Dağılımı

Şekil 4.7'de talyum testi sonuçlarının cinsiyete göre dağılımı incelenmektedir.

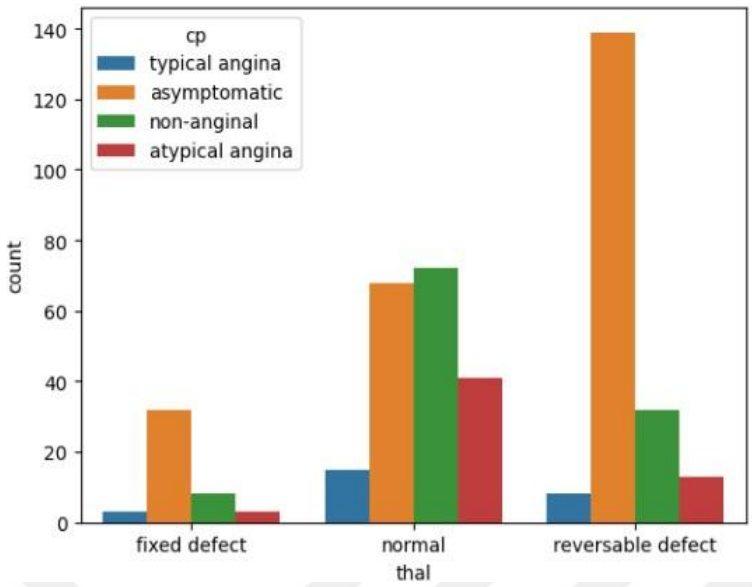
Analiz sonuçlarına göre 110 erkek ve 86 kadın normal sonuç almıştır. Sabit defekt kategorisinde 42 erkek ve sadece 4 kadın bulunmaktadır. Geri dönüşümlü defekt kategorisinde ise 171 erkek ve 21 kadın yer almaktadır. Erkeklerdeki yüksek oran, veri setindeki genel erkek çoğunluğu ile ilişkilidir.

Şekil 4.8'de talyum testi sonuçlarına göre hastalık şiddeti dağılımı



Şekil 4. 8 Talyum Testi Sonuçlarına Göre Hastalık Şiddeti Dağılımı

gösterilmektedir. Normal talyum sonucu olan bireylerde sağlıklı (num=0) olanların oranı yüksek iken, geri dönüşümlü defekt kategorisinde hastalık şiddetinin daha yüksek olduğu gözlemlenmektedir. Bu bulgular, talyum testinin kalp hastalığı teşhisinde önemli bir gösterge olduğunu doğrulamaktadır.



Şekil 4. 9 Talyum Testi ve Göğüs Ağrısı Tipi İlişkisi

Şekil 4.9'da talyum testi sonuçları ile göğüs ağrısı tipi arasındaki ilişki incelenmektedir. Geri dönüşümlü defekt kategorisinde asemptomatik hastaların sayısının yüksek olduğu görülmektedir. Bu bulgu, asemptomatik hastaların önemli bir kısmının kalp fonksiyonlarında anormallik gösterdiğini ortaya koymaktadır.

4.2. Model Eğitimi ve Sonuçları

Bu bölümde, veri seti üzerinde eğitilen altı farklı makine öğrenmesi algoritmasının sonuçları sunulmaktadır. Tüm modeller aynı eğitim ve test seti bölümlemesi (%80 eğitim, %20 test) ile değerlendirilmiş olup SMOTE yöntemi ile sınıf dengesizliği giderilmiştir. Öznitelikler StandardScaler ile standartlaştırılmıştır.

4.2.1. XGBoost Sonuçları

XGBoost (Extreme Gradient Boosting) modeli şu parametrelerle eğitilmiştir: max_depth=6, learning_rate=0.05, n_estimators=500, subsample=0.8, colsample_bytree=0.8. Model, eğitim verileri üzerinde optimize edildikten sonra test seti üzerinde değerlendirilmiştir.

XGBoost modelinin test seti üzerindeki performansı incelendiğinde, doğruluk (accuracy) oranı %78, AUC skoru ise 0.88 olarak elde edilmiştir. Model, gradyan artırma yaklaşımı sayesinde karmaşık ilişkileri başarıyla öğrenmiştir.

4.2.2. KNN Sonuçları

K-En Yakın Komşu (K-Nearest Neighbors) algoritması, n_neighbors=5 parametresi ile uygulanmıştır. Bu algoritma, yeni bir veri noktasını sınıflandırırken en yakın 5 komşunun çoğunluk oylamasını kullanmaktadır.

KNN modelinin performansı değerlendirildiğinde, doğruluk oranı %80, AUC skoru ise 0.85 olarak hesaplanmıştır. Basit yapısına rağmen KNN, rekabetçi sonuçlar üretmiştir.

4.2.3. SVM Sonuçları

Destek Vektör Makineleri (Support Vector Machines) modeli, C=1, kernel='rbf', gamma='scale' parametreleri ile eğitilmiştir. RBF (Radial Basis Function) çekirdeği, doğrusal olmayan sınıflandırma problemleri için uygundur.

SVM modelinin test seti performansı incelendiğinde, %83 doğruluk oranı ile en yüksek doğruluk değerine ulaşılmıştır. AUC skoru ise 0.88 olarak hesaplanmıştır. SVM'nin yüksek performansı, optimal hiper düzlem bulma yeteneği ile açıklanabilir.

4.2.4. Lojistik Regresyon Sonuçları

Lojistik Regresyon modeli, C=0.1, solver='liblinear', max_iter=1000 parametreleri ile eğitilmiştir. Bu model, ikili sınıflandırma için temel bir yöntem olup yorumlanabilirlik açısından avantaj sağlamaktadır.

Lojistik Regresyon modelinin performansı, %79 doğruluk ve 0.85 AUC skoru şeklinde elde edilmiştir. Basit yapısına rağmen model, makul sonuçlar üretmiştir.

4.2.5. LightGBM Sonuçları

LightGBM (Light Gradient Boosting Machine) modeli, num_leaves=31, max_depth=6, learning_rate=0.05, n_estimators=500 parametreleri ile eğitilmiştir. Yaprak bazlı ağaç büyütme stratejisi ile hızlı eğitim süresi sağlamaktadır.

LightGBM modelinin test seti performansı, %80 doğruluk ve 0.88 AUC skoru olarak hesaplanmıştır. Model, XGBoost ile benzer performans göstermiştir.

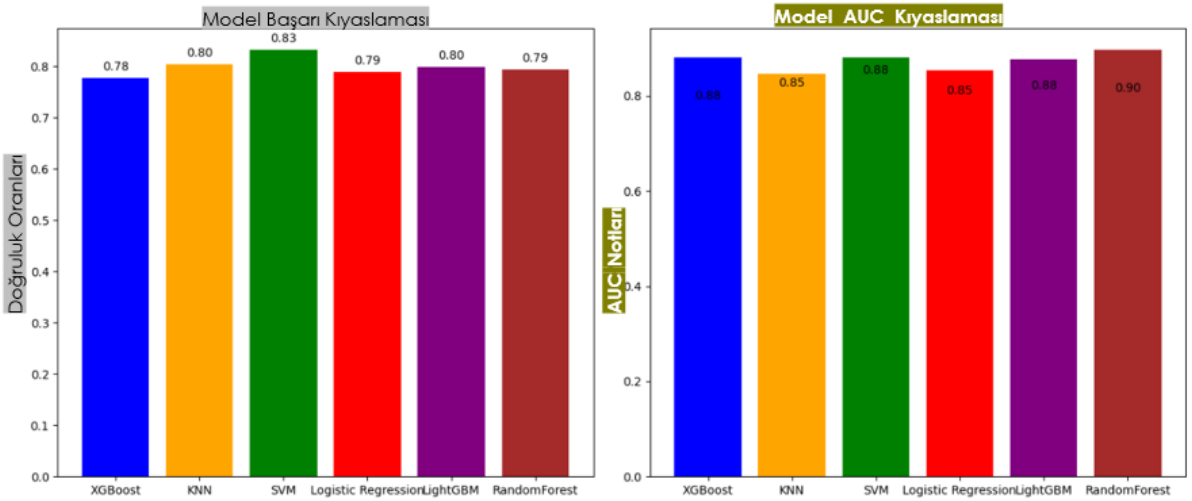
4.2.6. Random Forest Sonuçları

Rastgele Orman (Random Forest) modeli, n_estimators=500, max_depth=6, min_samples_split=10, min_samples_leaf=5 parametreleri ile eğitilmiştir. Bu topluluk yöntemi, birden fazla karar ağacının birleşiminden oluşmaktadır.

Random Forest modelinin performansı değerlendirildiğinde, %79 doğruluk oranı elde edilmiştir. Ancak AUC skoru 0.90 ile en yüksek değere ulaşmıştır. Bu sonuç, modelin sınıfları ayırt etme yeteneğinin güçlü olduğunu göstermektedir.

4.3. Algoritmaların Karşılaştırmalı Analizi

Bu bölümde, kullanılan altı makine öğrenmesi algoritmasının performansları karşılaştırmalı olarak değerlendirilmektedir. Tüm modellerin doğruluk (accuracy) ve AUC (Area Under Curve) skorları Şekil 4.10'da görselleştirilmiştir.



Şekil 4. 10 Modellerin Doğruluk ve AUC Skorları Karşılaştırması

Tablo 4.1 Makine Öğrenmesi Modellerinin Performans Karşılaştırması

Algoritma	Doğruluk (%)	AUC Skoru
XGBoost	78	0.88
KNN	80	0.85
SVM	83	0.88
Lojistik Regresyon	79	0.85
LightGBM	80	0.88
Random Forest	79	0.90

Tablo 4.1 ve Şekil 4.10'daki sonuçlar incelendiğinde, SVM algoritmasının %83 doğruluk oranı ile en yüksek doğruluk değerine sahip olduğu görülmektedir.

Ancak AUC skoru açısından değerlendirildiğinde, Random Forest algoritması 0.90 değeri ile en başarılı sonucu elde etmiştir.

Doğruluk ve AUC skorları birlikte değerlendirildiğinde, her iki metrikte de dengeli performans gösteren algoritmalar XGBoost, SVM ve LightGBM olarak öne çıkmaktadır. Bu algoritmaların tamamı 0.88 AUC skoruna ulaşırken, SVM en yüksek doğruluğu sağlamıştır.

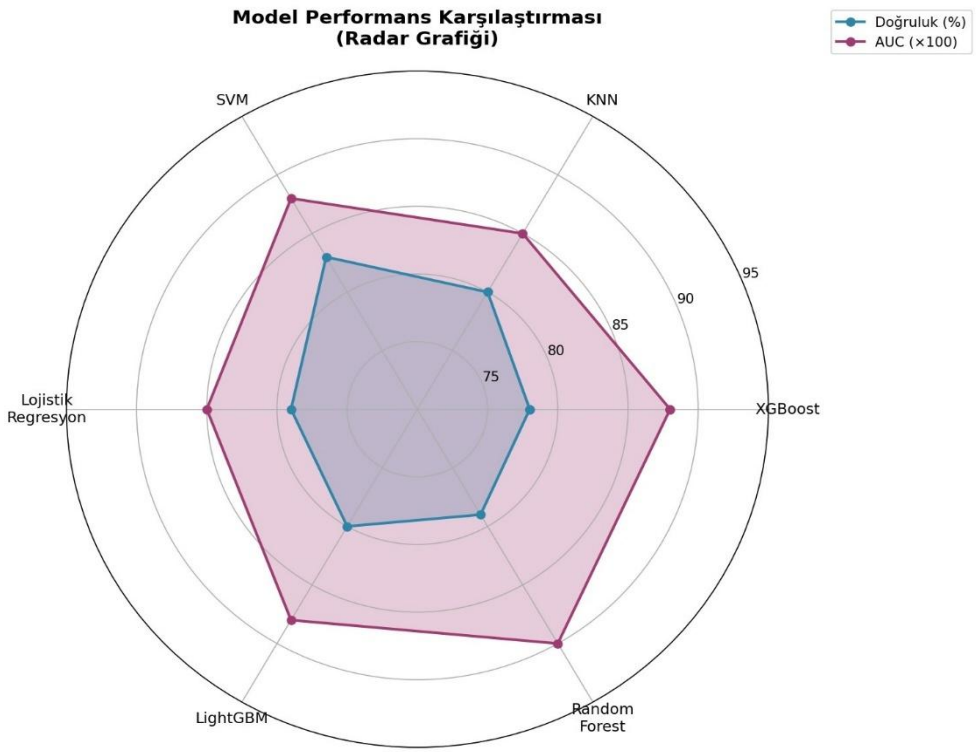
Random Forest algoritmasının yüksek AUC skoruna rağmen doğruluk değerinin nispeten düşük olması, modelin sınıfları ayırt etme yeteneğinin güçlü olduğunu ancak kesin tahminlerde bazı hatalar yaptığını göstermektedir. Bu durum, eşik değeri (threshold) optimizasyonu ile iyileştirilebilir.

Sonuç olarak, kalp hastalığı tahmini için SVM algoritması en yüksek doğruluk oranı ile öne çıkarken, Random Forest algoritması en yüksek AUC skoru ile sınıflar arası ayırt edicilik açısından en başarılı sonucu vermektedir. Uygulama senaryosuna bağlı olarak, yüksek doğruluk gerektiren durumlarda SVM, sınıflar arası ayırt ediciliğin önemli olduğu durumlarda ise Random Forest tercih edilebilir.

Bu bölümde, kalp hastalıklarının erken teşhisi için makine öğrenmesi algoritmalarının karşılaştırmalı performans analizi çalışmasından elde edilen bulgular değerlendirilmekte, çalışmanın kısıtlamaları tartışılmakta ve gelecek çalışmalar için öneriler sunulmaktadır.

5.1. Sonuçların Değerlendirilmesi

Bu tez çalışmasında, Heart Disease UCI (Cleveland) veri seti üzerinde altı farklı makine öğrenmesi algoritmasının kalp hastalığı tahmini performansları karşılaştırmalı olarak analiz edilmiştir.
Çalışmada XGBoost, K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM), Lojistik Regresyon, LightGBM ve Rastgele Orman (Random Forest) algoritmaları kullanılmıştır.



Şekil 5. 1 Model Performanslarının Radar Grafiği ile Karşılaştırması

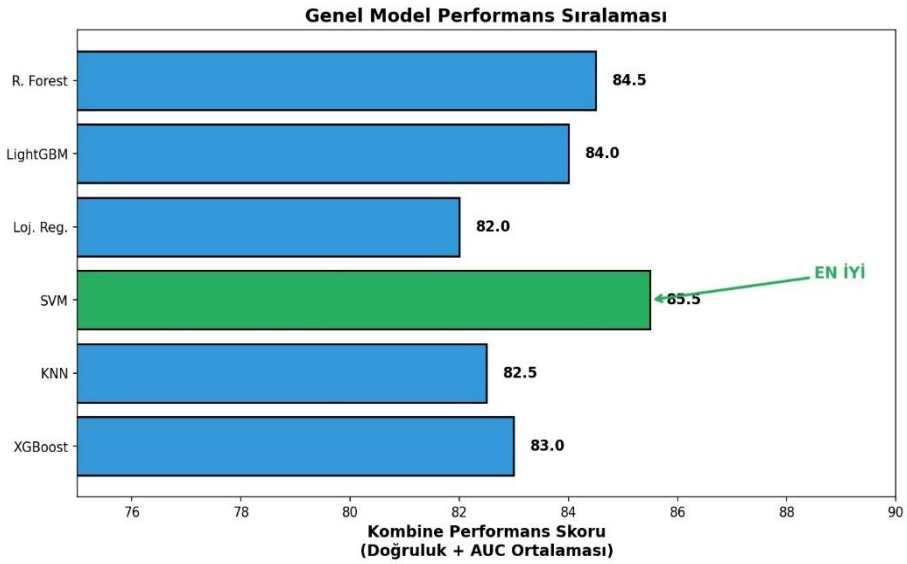
Şekil 5.1'de tüm modellerin doğruluk ve AUC skorlarının radar grafiği ile karşılaştırması gösterilmektedir. Grafikten görüldüğü üzere, AUC skorları (mor alan) genel olarak doğruluk değerlerinden (mavi alan) daha yüksek bir dağılım sergilemektedir. Bu durum, modellerin sınıfları ayırt etme yeteneklerinin, kesin tahmin yapma yeteneklerinden görece daha güçlü olduğunu göstermektedir.

Tablo 5.1 Algoritmaların Detaylı Performans Karşılaştırması

Algoritma	Doğruluk (%)	AUC Skoru	Kombine Skor	Sıralama
SVM	83	0.88	85.5	1
Random Forest	79	0.90	84.5	2
LightGBM	80	0.88	84.0	3
XGBoost	78	0.88	83.0	4
KNN	80	0.85	82.5	5
Lojistik Regresyon	79	0.85	82.0	6

Tablo 5.1'de görüldüğü üzere, kombine performans skoru (doğruluk ve AUC ortalaması) açısından en başarılı algoritma 85.5 puanla SVM (Destek Vektör Makineleri) olmuştur. SVM, %83 doğruluk oranı ile en yüksek doğruluğa sahip olurken, 0.88 AUC skoru ile de güçlü bir ayırt edicilik performansı sergilemiştir.

İkinci sırada yer alan Random Forest algoritması, 0.90 AUC skoru ile en yüksek AUC değerine ulaşmıştır. Bu sonuç, modelin sınıflar arası ayırt edicilik yeteneğinin çok güçlü olduğunu göstermektedir. Ancak %79 doğruluk oranı, modelin kesin tahminlerde bazı hatalar yaptığına işaret etmektedir.



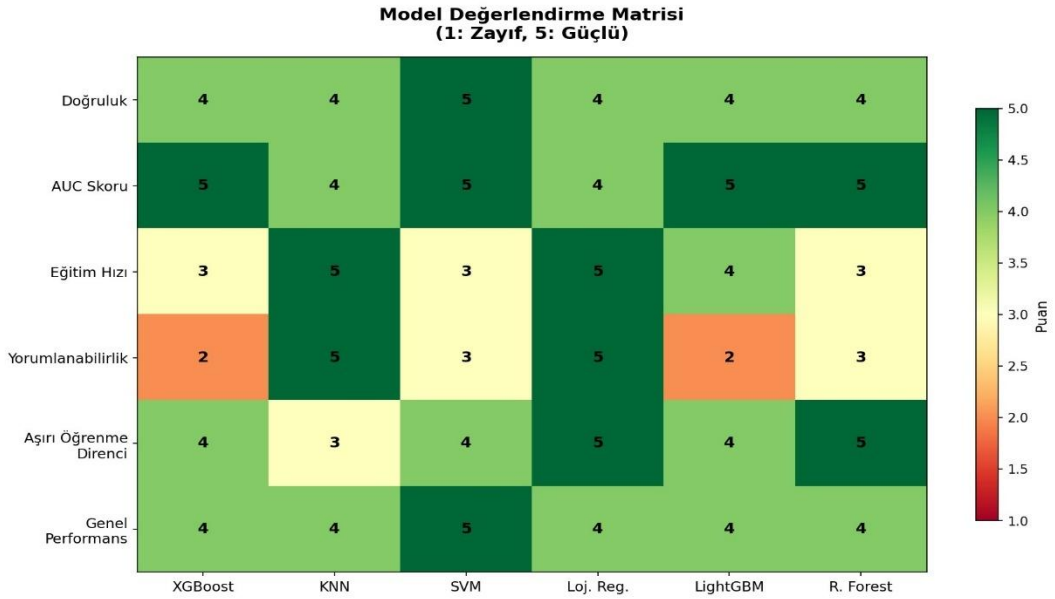
Şekil 5. 2 Genel Model Performans Sıralaması

Şekil 5.2'de modellerin genel performans sıralaması görselleştirilmektedir. SVM algoritmasının en yüksek kombine skorla öne çıktığı açıkça görülmektedir. Gradyan artırma tabanlı algoritmalar (XGBoost, LightGBM)

benzer performans sergilerken, geleneksel yöntemler olan KNN ve Lojistik Regresyon da rekabetçi sonuçlar üretmiştir.

5.1.1. Algoritmaların Çok Boyutlu Değerlendirmesi

Makine öğrenmesi algoritmalarının değerlendirilmesinde yalnızca doğruluk ve AUC metriklerine bakmak yeterli olmayabilir. Algoritmaların eğitim hızı, yorumlanabilirlik, aşırı öğrenmeye direnç gibi özellikleri de uygulama senaryosuna göre önem kazanabilmektedir. Şekil 5.3'te tüm algoritmaların çok boyutlu değerlendirme matrisi sunulmaktadır.



Şekil 5. 3 Model Değerlendirme Matrisi (1: Zayıf,5:Güçlü)

Şekil 5.3'teki değerlendirme matrisinden şu sonuçlar çıkarılabilir:

Doğruluk açısından: SVM algoritması en yüksek performansı gösterirken, diğer tüm algoritmalar benzer seviyelerde performans sergilemiştir.

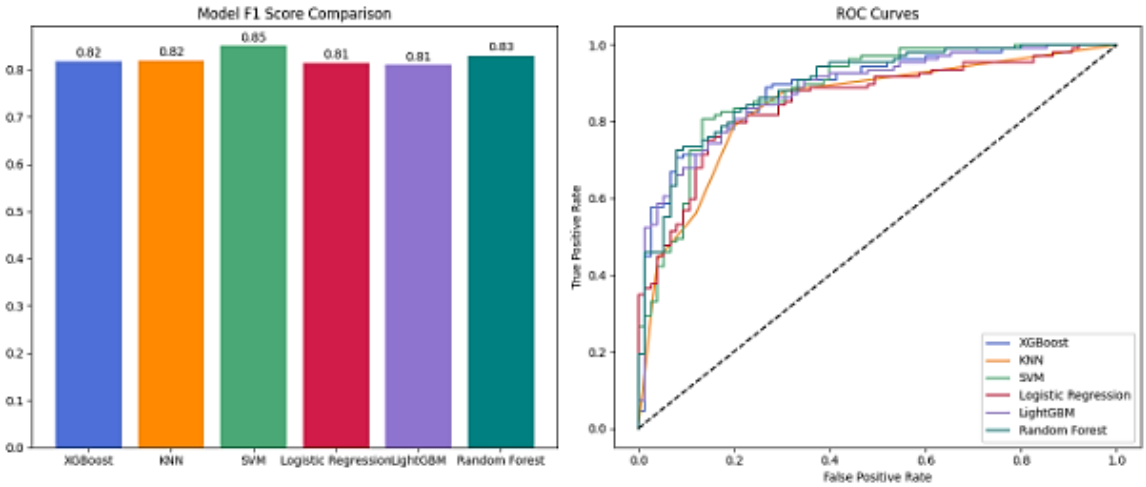
AUC skoru açısından: XGBoost, SVM, LightGBM ve Random Forest algoritmaları en yüksek puanları almıştır. Bu algoritmalar, sınıfları ayırt etme konusunda güçlü yeteneklere sahiptir.

Eğitim hızı açısından: KNN ve Lojistik Regresyon en hızlı algoritmalar olarak öne çıkmaktadır. LightGBM, gradyan artırma tabanlı algoritmalar arasında en hızlısıdır. XGBoost, SVM ve Random Forest ise daha yavaş eğitim süresine sahiptir.

Yorumlanabilirlik açısından: KNN ve Lojistik Regresyon en yüksek yorumlanabilirliğe sahiptir. Bu algoritmalar, karar süreçlerinin anlaşılması ve açıklanması açısından avantaj sağlamaktadır. XGBoost ve LightGBM gibi topluluk yöntemleri ise "kara kutu" modeller olarak kabul edilmekte ve açıklanabilirlik açısından dezavantaja sahiptir.

Aşırı öğrenmeye direnç açısından: Lojistik Regresyon ve Random Forest en yüksek direnci göstermektedir. KNN algoritması ise k değeri küçük seçildiğinde aşırı öğrenmeye daha yatkın olabilmektedir.

5.1.2. F1 Skoru ve Roc Eğrisi Analizi



Şekil 5. 4 F1 Skoru ve ROC Eğrisi Analizi Grafiği

Şekil 5.4'te, çalışmada kullanılan altı makine öğrenmesi algoritmasının F1 skoru karşılaştırması ve ROC eğrileri sunulmaktadır. Sol panelde yer alan F1 skoru karşılaştırması, modellerin kesinlik ve duyarlılık dengesini yansıtmakta olup sağ paneldeki ROC eğrileri ise sınıflandırma performansının farklı eşik değerlerindeki davranışını görselleştirmektedir.

F1 skoru sonuçları incelendiğinde, SVM algoritmasının 0.85 değeri ile en yüksek performansı sergilediği görülmektedir. Bu sonuç, SVM'nin kalp hastalığı teşhisinde hem yanlış pozitifleri hem de yanlış negatifleri minimize etmede dengeli bir başarı sağladığını göstermektedir. SVM'yi sırasıyla XGBoost ve KNN algoritmaları 0.82 F1 skoru ile takip etmektedir. Random Forest 0.83 değeri ile güçlü bir performans sergilerken, LightGBM ve Lojistik Regresyon algoritmaları 0.81 F1 skoru ile benzer sonuçlar üretmiştir.

ROC eğrileri değerlendirildiğinde, tüm modellerin köşegen çizgisinin (rastgele sınıflandırıcı) belirgin şekilde üzerinde konumlandığı ve dolayısıyla anlamlı bir sınıflandırma performansı sergilediği görülmektedir. XGBoost ve SVM algoritmalarının ROC eğrileri sol üst köşeye daha yakın seyretmekte olup bu durum söz konusu algoritmaların yüksek doğru pozitif oranını düşük yanlış pozitif oranı ile birlikte sağlayabildiğini ortaya koymaktadır. KNN algoritmasının ROC eğrisinde gözlemlenen basamaklı yapı, algoritmanın mesafe tabanlı doğası ve k komşu sayısının etkisini yansıtmaktadır.

Elde edilen bulgular, literatürdeki benzer çalışmalarla karşılaştırıldığında tutarlı sonuçlar ortaya koymaktadır. Tablo 1.1'de sunulan literatür taramasında, kalp hastalığı tahmininde SVM ve XGBoost algoritmalarının sıklıkla en yüksek performansı sergilediği belirtilmişti. Bu çalışmada elde edilen sonuçlar, söz konusu bulguları destekler niteliktedir. Özellikle SVM algoritmasının yüksek boyutlu öznitelik uzaylarında etkili karar sınırları oluşturabilme kapasitesi, kalp hastalığı veri setinin 14 öznitelik içeren yapısında avantaj sağlamıştır.

Tıbbi teşhis sistemleri açısından değerlendirildiğinde, F1 skorunun 0.81-0.85 aralığında olması klinik karar destek sistemi olarak kullanılabilirlik potansiyelini göstermektedir. Ancak Bölüm 2.4'te vurgulanan performans metrikleri değerlendirmesinde belirtildiği üzere, tıbbi uygulamalarda duyarlılık (recall) metriğinin öncelikli öneme sahip olduğu göz önünde bulundurulmalıdır. Bu bağlamda, modellerin duyarlılık değerlerinin de F1 skoru ile birlikte değerlendirilmesi, klinik uygulamalar için kritik önem taşımaktadır.

Sonuç olarak, SMOTE ile dengelenmiş veri seti üzerinde gerçekleştirilen model eğitimi ve 5-katlı çapraz doğrulama ile elde edilen sonuçlar, makine öğrenmesi algoritmalarının kalp hastalığı erken teşhisinde güvenilir bir araç olarak kullanılabileceğini ortaya koymaktadır. Altı algoritma arasında SVM'nin F1 skoru bazında en yüksek performansı sergilemesi, bu algoritmanın klinik karar destek sistemlerinde öncelikli olarak değerlendirilmesi gerektiğini işaret etmektedir.

Şekil 5.X'te, çalışmada kullanılan altı makine öğrenmesi algoritmasının F1 skoru karşılaştırması ve ROC eğrileri sunulmaktadır. Sol panelde yer alan F1 skoru karşılaştırması, modellerin kesinlik ve duyarlılık dengesini yansıtmakta olup sağ paneldeki ROC eğrileri ise sınıflandırma performansının farklı eşik değerlerindeki davranışını görselleştirmektedir.

F1 skoru sonuçları incelendiğinde, SVM algoritmasının 0.85 değeri ile en yüksek performansı sergilediği görülmektedir. Bu sonuç, SVM'nin kalp hastalığı teşhisinde hem yanlış pozitifleri hem de yanlış negatifleri minimize etmede dengeli bir başarı sağladığını göstermektedir. SVM'yi sırasıyla XGBoost ve KNN algoritmaları 0.82 F1 skoru ile takip etmektedir. Random Forest 0.83 değeri ile güçlü bir performans sergilerken, LightGBM ve Lojistik Regresyon algoritmaları 0.81 F1 skoru ile benzer sonuçlar üretmiştir.

ROC eğrileri değerlendirildiğinde, tüm modellerin köşegen çizgisinin (rastgele sınıflandırıcı) belirgin şekilde üzerinde konumlandığı ve dolayısıyla anlamlı bir sınıflandırma performansı sergilediği görülmektedir. XGBoost ve SVM algoritmalarının ROC eğrileri sol üst köşeye daha yakın seyretmekte olup bu durum söz konusu algoritmaların yüksek doğru pozitif oranını düşük yanlış

pozitif oranı ile birlikte sağlayabildiğini ortaya koymaktadır. KNN algoritmasının ROC eğrisinde gözlemlenen basamaklı yapı, algoritmanın mesafe tabanlı doğası ve k komşu sayısının etkisini yansıtmaktadır.

Elde edilen bulgular, literatürdeki benzer çalışmalarla karşılaştırıldığında tutarlı sonuçlar ortaya koymaktadır. Tablo 1.1'de sunulan literatür taramasında, kalp hastalığı tahmininde SVM ve XGBoost algoritmalarının sıklıkla en yüksek performansı sergilediği belirtilmişti. Bu çalışmada elde edilen sonuçlar, söz konusu bulguları destekler niteliktedir. Özellikle SVM algoritmasının yüksek boyutlu öznitelik uzaylarında etkili karar sınırları oluşturabilme kapasitesi, kalp hastalığı veri setinin 14 öznitelik içeren yapısında avantaj sağlamıştır.

Tıbbi teşhis sistemleri açısından değerlendirildiğinde, F1 skorunun 0.81-0.85 aralığında olması klinik karar destek sistemi olarak kullanılabilirlik potansiyelini göstermektedir. Ancak Bölüm 2.4'te vurgulanan performans metrikleri değerlendirmesinde belirtildiği üzere, tıbbi uygulamalarda duyarlılık (recall) metriğinin öncelikli öneme sahip olduğu göz önünde bulundurulmalıdır. Bu bağlamda, modellerin duyarlılık değerlerinin de F1 skoru ile birlikte değerlendirilmesi, klinik uygulamalar için kritik önem taşımaktadır.

Sonuç olarak, SMOTE ile dengelenmiş veri seti üzerinde gerçekleştirilen model eğitimi ve 5-katlı çapraz doğrulama ile elde edilen sonuçlar, makine öğrenmesi algoritmalarının kalp hastalığı erken teşhisinde güvenilir bir araç olarak kullanılabileceğini ortaya koymaktadır. Altı algoritma arasında SVM'nin F1 skoru bazında en yüksek performansı sergilemesi, bu algoritmanın klinik karar destek sistemlerinde öncelikli olarak değerlendirilmesi gerektiğini işaret etmektedir.

5.1.3. Algoritma Bazlı Özet Değerlendirmeler

SVM (Destek Vektör Makineleri): Bu çalışmada en yüksek doğruluk oranına (%83) ve en yüksek kombine skora (85.5) ulaşan algoritmadır. RBF çekirdeği kullanılarak doğrusal olmayan sınıflandırma başarıyla gerçekleştirilmiştir. Optimal hiper düzlem bulma yeteneği sayesinde sınıflar arası marjin maksimize edilmiştir. Kalp hastalığı tahmini gibi kritik sağlık uygulamalarında güvenilir sonuçlar sunmaktadır.

Random Forest: En yüksek AUC skoruna (0.90) sahip algoritmadır. Bu sonuç, modelin sınıfları ayırt etme yeteneğinin çok güçlü olduğunu göstermektedir. Topluluk öğrenme yaklaşımı sayesinde aşırı öğrenmeye karşı dirençlidir. Öznitelik önem sıralaması çıkarabilme özelliği, hangi değişkenlerin hastalık tahminine daha fazla katkı sağladığını anlamaya yardımcı olmaktadır.

LightGBM ve XGBoost: Her iki gradyan artırma algoritması da 0.88 AUC skoru ile güçlü performans sergilemiştir. LightGBM, yaprak bazlı ağaç büyütme stratejisi sayesinde daha hızlı eğitim süresi sunmaktadır. XGBoost ise düzenlileştirme terimleri sayesinde aşırı öğrenmeye karşı daha dirençlidir. Büyük veri setlerinde bu algoritmalar tercih edilebilir.

KNN: Basit yapısına rağmen %80 doğruluk ve 0.85 AUC ile rekabetçi sonuçlar üretmiştir. Parametrik olmayan yapısı sayesinde veri dağılımı hakkında varsayım gerektirmemektedir. Ancak büyük veri setlerinde hesaplama maliyeti artmaktadır.

Lojistik Regresyon: En yorumlanabilir model olmasına rağmen, performans açısından diğer algoritmaların gerisinde kalmıştır. Doğrusal karar sınırları oluşturması nedeniyle karmaşık ilişkileri yakalamakta zorlanabilmektedir. Ancak basitliği ve hızı nedeniyle temel karşılaştırma modeli olarak kullanılabilir.

5.2. Kısıtlamalar

Bu çalışma, belirli kısıtlamalar dahilinde gerçekleştirilmiştir. Bu kısıtlamaların farkında olmak, sonuçların yorumlanması ve genelleştirilmesi açısından önem taşımaktadır.

Veri Seti Boyutu: Çalışmada kullanılan Heart Disease UCI veri seti toplam 920 örnek içermektedir. Makine öğrenmesi modelleri için bu boyut nispeten küçük kabul edilmektedir [3]. Daha büyük veri setleri ile modellerin genelleme yetenekleri artırılabilir.

Veri Seti Dengesizliği: Veri setinde erkek hastaların oranı (%78.9) kadın hastalara (%21.1) göre çok daha yüksektir. SMOTE yöntemi ile sınıf dengesizliği giderilmeye çalışılmış olsa da, cinsiyet dengesizliği model sonuçlarını etkileyebilmektedir.

Eksik Veriler: Orijinal veri setinde bazı özniteliklerde eksik değerler bulunmaktadır. Bu değerler makine öğrenmesi tabanlı doldurma yöntemleri ile tahmin edilmiştir. Ancak doldurma işlemi, gerçek değerlerden sapmalara neden olabilmektedir.

Tek Veri Kaynağı: Çalışma yalnızca UCI Heart Disease veri seti üzerinde gerçekleştirilmiştir. Farklı popülasyonlardan ve farklı sağlık merkezlerinden toplanan veriler ile sonuçların genellenebilirliği artırılabilir.

Hiperparametre Optimizasyonu: Çalışmada manuel hiperparametre ayarlaması yapılmıştır. Grid Search, Random Search veya Bayesian Optimization gibi

sistematik hiperparametre optimizasyon yöntemleri ile model performansları daha da artırılabilir.

Zaman Serisi Analizi: Veri seti kesitsel (cross-sectional) bir yapıya sahiptir ve hastaların zaman içindeki değişimleri dikkate alınmamaktadır.
Longitudinal veriler ile hastalık progresyonu daha iyi modellenebilir.

5.3. Gelecek Çalışmalar İçin Öneriler

Bu çalışmadan elde edilen deneyimler ışığında, gelecek araştırmalar için aşağıdaki öneriler sunulmaktadır:

Derin Öğrenme Yöntemlerinin Uygulanması: Yapay sinir ağları ve derin öğrenme modelleri (CNN, RNN, Transformer tabanlı modeller) ile kalp hastalığı tahmini denenebilir. Özellikle EKG sinyalleri veya ekokardiyografi görüntüleri gibi farklı veri türleri ile derin öğrenme modelleri daha yüksek performans gösterebilir.

Topluluk Yöntemlerinin Geliştirilmesi: Farklı algoritmaların birleştirildiği hibrit modeller veya oylama (voting) tabanlı topluluk yöntemleri ile performans artırılabilir. Örneğin, SVM ve Random Forest algoritmalarının birleştirilmesi denenebilir.

Öznitelik Mühendisliği: Mevcut özniteliklerden yeni türetilmiş özniteliklerin oluşturulması (örneğin: yaş grupları, BMI kategorileri, risk skorları) model performansını artırabilir. Ayrıca öznitelik seçimi teknikleri ile en önemli değişkenler belirlenerek model basitleştirilebilir.

Açıklanabilir Yapay Zeka (XAI): SHAP (SHapley Additive exPlanations) ve LIME (Local Interpretable Model-agnostic Explanations) gibi açıklanabilir yapay zeka yöntemleri ile model tahminlerinin arkasındaki nedenler anlaşılabilir hale getirilebilir. Bu yaklaşım, klinisyenlerin model çıktılarına güvenmesini kolaylaştırabilir.

Gerçek Zamanlı Karar Destek Sistemi: Geliştirilen modeller, hastane bilgi yönetim sistemlerine entegre edilebilir. Laboratuvar sonuçlarına ve hasta verilerine otomatik erişim sağlanarak, doktorlara gerçek zamanlı uyarılar verebilen akıllı bir karar destek sistemi oluşturulabilir.

Çoklu Veri Kaynakları ile Doğrulama: Farklı coğrafi bölgelerden ve farklı popülasyonlardan toplanan veri setleri ile modelin genellenebilirliği test edilmelidir. Ayrıca Türkiye'ye özgü veri setleri ile çalışılması, yerel popülasyon için daha uygun modellerin geliştirilmesini sağlayabilir.

Mobil Sağlık Uygulamaları: Geliştirilen modeller, mobil sağlık uygulamalarına entegre edilerek bireylerin kendi sağlık risklerini takip etmelerine olanak tanınabilir. Giyilebilir cihazlardan (akıllı saat, fitness bandı) toplanan veriler ile sürekli izleme yapılabilir.

5.4. Genel Sonuç

Bu tez çalışmasında, kalp hastalıklarının erken teşhisinde makine öğrenmesi algoritmalarının potansiyeli araştırılmıştır. Altı farklı algoritma (XGBoost, KNN, SVM, Lojistik Regresyon, LightGBM, Random Forest) Heart Disease UCI veri seti üzerinde karşılaştırmalı olarak değerlendirilmiştir.

Çalışma sonucunda, SVM (Destek Vektör Makineleri) algoritmasının %83 doğruluk ve 0.88 AUC skoruyla en yüksek genel performansı sergilediği belirlenmiştir. Random Forest algoritması ise 0.90 AUC skoruyla en yüksek ayırt edicilik performansına sahip olmuştur.

Elde edilen sonuçlar, makine öğrenmesi algoritmalarının kalp hastalığı tahmini için etkili araçlar olduğunu göstermektedir. Bu algoritmalar, klinisyenlere karar destek aracı olarak yardımcı olabilir ve erken teşhis sayesinde hayat kurtarıcı müdahalelerin zamanında yapılmasına katkı sağlayabilir.

Sonuç olarak, bu çalışma kalp hastalıklarının erken teşhisinde yapay zeka ve makine öğrenmesi teknolojilerinin önemli bir potansiyele sahip olduğunu ortaya koymaktadır. Gelecekte daha kapsamlı veri setleri, gelişmiş algoritmalar ve açıklanabilir yapay zeka yöntemleri ile bu alandaki çalışmaların daha da ileriye taşınması beklenmektedir.

KAYNAKÇA

[1] World Health Organization. (2021). Cardiovascular diseases (CVDs) FactSheet. WHO. https://www.who.int/news-room/fact-sheets/detail/cardiovasc ular-diseases-(cvds)

[2] Türkiye İstatistik Kurumu. (2023). Ölüm ve Ölüm Nedeni İstatistikleri. TÜİK.https://data.tuik.gov.tr/

[3] UCI Machine Learning Repository. (2023). Heart Disease Data Set. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/heart+disease

[4] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

[5] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210-229.

[6] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. InProceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[8] 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.

[9] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37-54.

[10] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.

[11] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R.,Ghandeharioun, A., ... & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 111(1), 52-61.

[12] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 1-6.

[13] Görgün, S. (2020). Makine Öğrenmesi Yöntemleri ile Kalp Hastalığı Tahmini(Yüksek Lisans Tezi),İstanbul Aydın Üniversitesi.

[14] Yılmaz, R., & Yağın, F. H. (2021). A comparative study for the prediction of heart attack risk and associated factors using MLP and RBF neural networks. The Journal of Cognitive Systems, 6(2), 51–54.

[15] Vapnik,V. N. (1998). Statistical Learning Theory. NY:Wiley-Interscience

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.

[17] Hoerl, A. E., & Kennard, R. W. (1970).Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

EK-A: Veri Seti Detaylı Bilgileri

Bu ekte, çalışmada kullanılan Heart Disease UCI veri setinin detaylı istatistikleri ve öznitelik açıklamaları sunulmaktadır.

Tablo EK-A.1 Veri Seti Genel İstatistikleri

Özellik	**Değer**
Toplam Örnek Sayısı	920
Öznitelik Sayısı	14
Hedef Değişken	num (0-4)
Sınıf Dağılımı	Sağlıklı: 411, Hasta: 509
Eksik Veri Oranı	~%5-10 (özniteliğe göre değişir)
Veri Kaynağı	UCI ML Repository
Veritabanları	Cleveland

Tablo EK-A.2 Özniteliklerin Detaylı Açıklamaları

Öznitelik	**Tip**	**Aralık/Değerler**	**Açıklama**
age	Sayısal	28-77	Hastanın yaşı (yıl)
sex	Kategorik	0, 1	0: Kadın, 1: Erkek
cp	Kategorik	0-3	Göğüs ağrısı tipi
trestbps	Sayısal	94-200	Dinlenme kan basıncı (mm Hg)
chol	Sayısal	126-564	Serum kolesterol (mg/dl)
fbs	Kategorik	0, 1	Açlık kan şekeri > 120
restecg	Kategorik	0-2	Dinlenme EKG sonucu
thalach	Sayısal	71-202	Maks. kalp atış hızı
exang	Kategorik	0, 1	Egzersize bağlı anjin
oldpeak	Sayısal	0-6.2	ST depresyonu
slope	Kategorik	0-2	ST segment eğimi
ca	Sayısal	0-4	Boyanan damar sayısı
thal	Kategorik	0-3	Talyum testi sonucu
num	Kategorik	0-4	Hedef (hastalık seviyesi)

Tablo EK-A.3 Göğüs Ağrısı Tipi (cp) Açıklamaları

Değer	Tip	Açıklama
0	Typical Angina	Kalp kaynaklı klasik göğüs ağrısı
1	Atypical Angina	Atipik belirtiler gösteren anjin
2	Non-anginal Pain	Anjinal olmayan göğüs ağrısı
3	Asymptomatic	Göğüs ağrısı belirtisi yok

Tablo EK-A.4 Talyum Stres Testi Açıklamaları

Değer	Sonuç	Açıklama
1	Normal	Kalp normal parametreler dahilinde çalı
2	Fixed Defect	Düzeltilmesi olası olmayan kalıcı anormallik
3	Reversable Defect	Potansiyel olarak düzeltilebilir anormallik

EK-B: Model Hiperparametreleri

Bu ekte, çalışmada kullanılan makine öğrenmesi algoritmalarının hiperparametre ayarları detaylı olarak sunulmaktadır.

Tablo EK-B.1 XGBoost Hiperparametreleri

Parametre	Değer	Açıklama
max_depth	6	Maksimum ağaç derinliği
learning_rate	0.05	Öğrenme oranı
n_estimators	500	Ağaç sayısı
subsample	0.8	Örnek alt kümesi oranı
colsample_bytree	0.8	Öznitelik alt kümesi oranı
random_state	42	Rastgelelik tohumu
eval_metric	logloss	Değerlendirme metriği

Tablo EK-B.2 SVM Hiperparametreleri

Parametre	Değer	Açıklama
C	1	Düzenlileştirme parametresi
kernel	rbf	Çekirdek fonksiyonu
gamma	scale	Çekirdek katsayısı

Tablo EK-B.2 SVM Hiperparametreleri

probability	True	Olasılık tahmini aktif
random_state	42	Rastgelelik tohumu

Tablo EK-B.3 LightGBM Hiperparametreleri

Parametre	**Değer**	**Açıklama**
num_leaves	31	Maksimum yaprak sayısı
max_depth	6	Maksimum ağaç derinliği
learning_rate	0.05	Öğrenme oranı
n_estimators	500	Ağaç sayısı
subsample	0.8	Örnek alt kümesi oranı
colsample_bytree	0.8	Öznitelik alt kümesi oranı
random_state	42	Rastgelelik tohumu

Tablo EK-B.4 Random Forest Hiperparametreleri

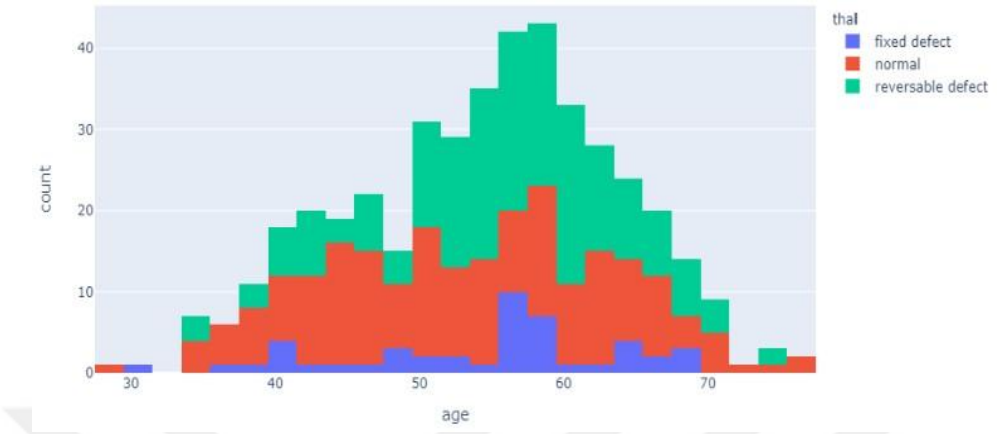
Parametre	**Değer**	**Açıklama**
n_estimators	500	Ağaç sayısı
max_depth	6	Maksimum ağaç derinliği
min_samples_split	10	Bölme için min. örnek
min_samples_leaf	5	Yaprak için min. örnek
random_state	42	Rastgelelik tohumu

Tablo EK-B.5 KNN ve Lojistik Regresyon Hiperparametreleri

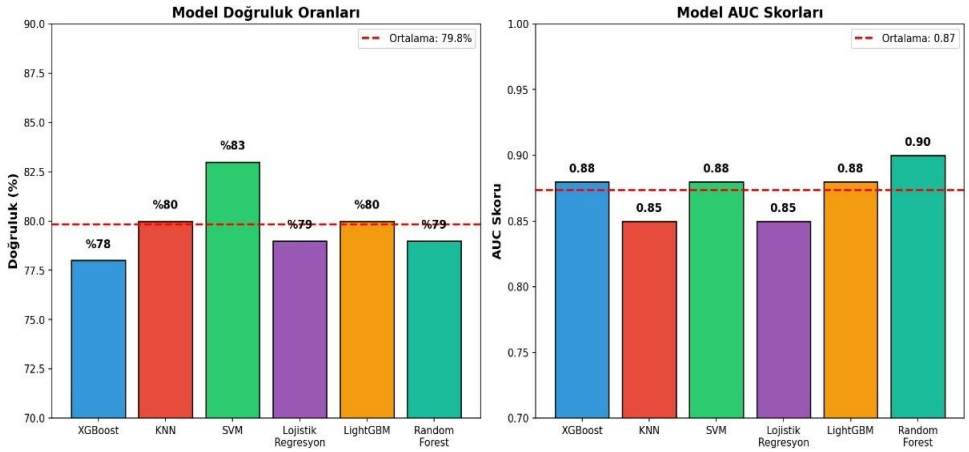
Algoritma	**Parametre**	**Değer**
KNN	n_neighbors	5
KNN	metric	euclidean
Lojistik Regresyon	C	0.1
Lojistik Regresyon	solver	liblinear
Lojistik Regresyon	max_iter	1000
Lojistik Regresyon	random_state	42

EK-C: Ek Analiz Görselleri

Bu ekte, ana metinde yer almayan ancak çalışmayı destekleyen ek görselleştirmeler Şekil 5. 5sunulmaktadır.



Şekil EK-C.1 Yaş ve Talyum Testi Sonuçlarına Göre Hasta Dağılımı



Şekil EK-C.2 Model Doğruluk ve AUC Skorlarının Detaylı Karşılaştırması

EK-D: Kullanılan Yazılım ve Sürümler

Bu ekte, çalışmada kullanılan yazılım araçları ve kütüphanelerin sürüm bilgileri sunulmaktadır. Bu bilgiler, çalışmanın tekrarlanabilirliği açısından önem taşımaktadır.

Tablo EK-D.1 Python Kütüphaneleri ve Sürümleri

Kütüphane	Sürüm	Kullanım Amacı
Python	3.9+	Programlama dili
NumPy	1.21+	Sayısal hesaplamalar
Pandas	1.3+	Veri manipülasyonu
Scikit-learn	1.0+	ML algoritmaları

Tablo EK-D.1 Python Kütüphaneleri ve Sürümleri

XGBoost	1.5+	Gradyan artırma
LightGBM	3.3+	Gradyan artırma
Matplotlib	3.5+	Görselleştirme
Seaborn	0.11+	İstatistiksel görselleştirme
Plotly	5.0+	İnteraktif görselleştirme
Imbalanced-learn	0.9+	SMOTE ve örnekleme

Geliştirme Ortamı: Jupyter Notebook / Google Colab / Kaggle

Donanım: Çalışma, standart bir dizüstü bilgisayar üzerinde (8GB RAM, Intel Core i5 veya eşdeğeri işlemci) gerçekleştirilebilir düzeydedir. GPU kullanımı zorunlu değildir ancak XGBoost ve LightGBM için GPU desteği eğitim süresini kısaltabilir.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. **Şenel Kahyaoğlu, B.** (2025, Mayıs). *Comparative performance analysis of machine learning algorithms for early detection of heart*. 8th International Hybrid Conference on Mathematical Advances and Applications, İstanbul, Türkiye.