SAĞLIK ALANINDA HASTALIKLARIN ERKEN TAHMİNİNDE YAPAY ZEKA TEKNİKLERİNİN KULLANILMASI

Yasemin ÖZDEMİR

T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SAĞLIK ALANINDA HASTALIKLARIN ERKEN TAHMİNİNDE YAPAY ZEKA TEKNİKLERİNİN KULLANILMASI

Yasemin ÖZDEMİR
0009-0004-0692-5343

Prof. Dr. Nursel ÖZTÜRK
(Danışman)

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA – 2024
Her Hakkı Saklıdır

TEZ ONAYI

Yasemin ÖZDEMİR tarafından hazırlanan "SAĞLIK ALANINDA HASTALIKLARIN ERKEN TAHMİNİNDE YAPAY ZEKA TEKNİKLERİNİN KULLANILMASI" adlı tez çalışması aşağıdaki jüri tarafından oy birliği ile Bursa Uludağ Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman: Prof. Dr. Nursel ÖZTÜRK

Başkan	**:**	Prof. Dr. Nursel ÖZTÜRK 0000-0002-9835-0783 Bursa Uludağ Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Anabilim Dalı	İmza
Üye	**:**	Dr. Öğretim Üyesi İsmail Enes PARLAK 0000-0002-0694-9220 Bursa Teknik Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Endüstri Mühendisliği Anabilim Dalı	İmza
Üye	**:**	Dr. Öğretim Üyesi Seval ENE YALÇIN 0000-0001-8248-8924 Bursa Uludağ Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Anabilim Dalı	İmza

Yukarıdaki sonucu onaylarım

Prof. Dr. Ali KARA
Enstitü Müdürü
../../….

B.U.Ü. Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

24/06/2024

Yasemin ÖZDEMİR

TEZ YAYINLANMA
FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezin/raporun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma izni Bursa Uludağ Üniversitesi'ne aittir. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet hakları ile tezin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları tarafımıza ait olacaktır. Tezde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığını ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederiz.

Yükseköğretim Kurulu tarafından yayınlanan "**Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge**" kapsamında, yönerge tarafından belirtilen kısıtlamalar olmadığı takdirde tezin YÖK Ulusal Tez Merkezi / B.U.Ü. Kütüphanesi Açık Erişim Sistemi ve üye olunan diğer veri tabanlarının (Proquest veri tabanı gibi) erişimine açılması uygundur.

Danışman Adı-Soyadı Tarih Prof. Dr. Nursel ÖZTÜRK	Öğrencinin Adı-Soyadı Tarih Yasemin ÖZDEMİR
İmza Bu bölüme kişinin kendi el yazısı ile okudum anladım yazmalı ve imzalanmalıdır.	İmza Bu bölüme kişinin kendi el yazısı ile okudum anladım yazmalı ve imzalanmalıdır.

BURSA ULUDAĞ ÜNİVERSİTESİ
LİSANSÜSTÜ TEZ TANITIMI
ÖĞRENCİ VE DANIŞMAN FORMU

FR 3.4.6_27

ÖĞRENCİ
Adı SOYADI : Yasemin ÖZDEMİR
ÜNVANI : Endüstri Mühendisi
FEN BİLİMLERİ ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ABD
E-POSTA : ozdemir.ysminn@gmail.com
PROGRAMI: YÜKSEK LİSANS
ORCID : 0009-0004-0692-5343
TÜBİTAK ID :TBTK-0159-2132
WOS RESEARCHER ID :KUD-3847-2024
SCOPUS AUTHOR ID :
Google Scholar ID : dDSUEbsAAAAJ

DANIŞMAN
Adı SOYADI : Nursel ÖZTÜRK
ÜNVANI : Prof. Dr.
FEN BİLİMLERİ ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ABD
E-POSTA : nursel@uludag.edu.tr
YÖKSİS ARAŞTIRMACI ID : 156678
ORCID : 0000-0002-9835-0783
TÜBİTAK ID : TBTK-0014-0250
WOS RESEARCHER ID : AAG-9336-2021
SCOPUS AUTHOR ID : 7005688805
Google Scholar ID : 1zPAZlEAAAAJ

BURSA ULUDAĞ ÜNİVERSİTESİ YÜKSEK LİSANS EĞİTİMİ BOYUNCA BİLİMSEL ÇALIŞMALARI VE FAALİYETLERİ*

1. Özdemir, Y., Öztürk, N., An Artificial Intelligence Technique Application in the Diagnosis of Heart Disease Using Random Forest Algorithm, *8th International European Conference on Interdisciplinary Scientific Research*, July 13-15, 2023, Rome, Italy, pp.834-838.

BİRLEŞMİŞ MİLLETLER SÜRDÜRÜLEBİLİR KALKINMA HEDEFLERİ

☐

☐

☐

Diabetes

Heart Failure

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

ÖZET

Yüksek Lisans Tezi

SAĞLIK ALANINDA HASTALIKLARIN ERKEN TAHMİNİNDE YAPAY ZEKA TEKNİKLERİNİN KULLANILMASI

Yasemin ÖZDEMİR

Bursa Uludağ Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Nursel ÖZTÜRK

Kalp hastalığı ve diyabet hastalığı gibi hastalıklar son yıllarda giderek yaygınlaşan önemli hastalık gruplarındandır. Bu tez çalışmasının amacı, sağlık alanında yapay zeka tekniklerinin kullanılması ve yaygın görülen hastalıkların erken tahmini için literatüre katkı sağlamaktır. Çalışma kapsamında kalp hastalığı için UCI veri tabanından Cleveland veri seti ve diyabet hastalığı için Vanderbilt Üniversitesi Biyoistatistik Bölümünden alınan veri seti kullanılmıştır. Çalışmada, tahmin algoritmaları olarak Rassal Orman, Lojistik Regresyon ve XGBoost kullanılmıştır. Bu algoritmaların performansı, 5 katlı ve 10 katlı çapraz doğrulama kullanılarak değerlendirilmiştir. Performans ölçütleri olarak doğruluk, duyarlılık, kesinlik ve F1 skoru gibi kriterler kullanılmıştır. Diyabet hastalığı veri setinde dengesiz veri problemini çözmek için yapay veriler üretilerek veri seti daha dengeli hale getirilmiştir. Elde edilen sonuçlara göre, kalp hastalığı tahmininde Rastgele Orman algoritması ile %93,00, Lojistik Regresyon ile %94,00 ve XGBoost algoritması ile %95,00 başarı oranı elde edilmiştir. Diyabet hastalığı tahmininde ise Rastgele Orman algoritması %99,00, Lojistik Regresyon %96,00 ve XGBoost algoritması %98,00 başarı oranlarına ulaşmıştır. Çalışmanın devamında, kalp hastalığının tahmininde XGBoost tabanlı, diyabet hastalığı tahmininde ise Rastgele Orman tabanlı arayüz sistemi geliştirilmiştir. Çalışmanın sonucunda, yapay zeka tekniklerinin sağlık alanında hastalıkların tahmini için yüksek performans gösterdiği gözlemlenmiştir. Elde edilen sonuçlar, yapay zeka tekniklerinin hastalıkların erken tahmininde etkili olarak kullanılabileceğini göstermektedir.

Anahtar Kelimeler: Makine Öğrenmesi, kalp hastalığı tahmini, yapay zeka, diyabet hastalığı tahmini, rassal orman algoritması, XGBoost algoritması
2024, vii + 114 sayfa.

ABSTRACT

MSc Thesis

USE OF ARTIFICIAL INTELLIGENCE TECHNIQUES IN EARLY PREDICTION OF DISEASES IN THE FIELD OF HEALTH

Yasemin ÖZDEMİR

Bursa Uludağ University
Graduate School of Natural and Applied Sciences
Department of Industrial Engineering

Supervisor: Prof. Dr. Nursel ÖZTÜRK

Diseases such as heart disease and diabetes are among the important disease groups that have become increasingly common in recent years. The aim of this thesis is to contribute to the literature on the use of artificial intelligence techniques in the field of health and the early prediction of common diseases. Within the scope of the study, the Cleveland data set from the UCI database was used for heart disease and the data set from Vanderbilt University Biostatistics Department was used for diabetes. In the study, Random Forest, Logistic Regression and XGBoost were used as prediction algorithms. The performance of these algorithms was evaluated using 5-fold and 10-fold cross-validation. Criteria such as accuracy, sensitivity, precision and F1 score were used as performance measures. To solve the problem of unbalanced data in the diabetes data set, artificial data was produced and the data set was made more balanced. According to the results obtained, a success rate of 93,00% was achieved with the Random Forest algorithm, 94,00% with Logistic Regression and 95,00% with the XGBoost algorithm in predicting heart disease. In predicting diabetes, the Random Forest algorithm achieved success rates of 99,00%, Logistic Regression 96,00% and XGBoost algorithm 98,00%. In the continuation of the study, an XGBoost-based interface system was developed for predicting heart disease and a Random Forest-based interface system for predicting diabetes. As a result of the study, it was observed that artificial intelligence techniques showed high performance in predicting diseases in the field of health. The results obtained show that artificial intelligence techniques can be used effectively in the early prediction of diseases.

Keywords: Machine Learning, heart disease prediction, artificial intelligence, diabetes prediction, random forest algorithm, XGBoost algorithm
2024, vii + 114 pages.

ÖNSÖZ ve TEŞEKKÜR

Bu çalışmanın ana amacı yapay zeka teknikleri ile sağlık alanında hastalıkların erken tahminine yönelik katkı yapmaktır. Bu tez çalışmasında, kalp hastalığı ve diyabet hastalığının tahminine yönelik çalışmalar yapılarak kullanıcı arayüzü tasarlanmış ve bir sistem geliştirilmiştir.

Yüksek lisansa başladığım günden itibaren, tüm tez çalışması sürecim boyunca ve mesleki gelişimimde bana her zaman yol gösteren, destekleyen, cesaret veren ve ilham kaynağı olan değerli danışmanım Prof. Dr. Nursel Öztürk'e teşekkürlerimi sunarım.

Tüm eğitim hayatım boyunca bana desteklerini esirgemeyen annem Sevim Özdemir'e, babam Topçu Özdemir'e ve kardeşim Özgür Özdemir'e teşekkürlerimi sunarım.

Yasemin ÖZDEMİR
24/06/2024

İÇİNDEKİLER

Sayfa
ÖZET.....viii
ABSTRACT.....ix
ÖNSÖZ ve TEŞEKKÜR.....x
SİMGELER ve KISALTMALAR DİZİNİ xiiiii
ŞEKİLLER DİZİNİ.....xiv
ÇİZELGELER DİZİNİ xv
1. GİRİŞ…...……..1
2. KURAMSAL TEMELLER ve KAYNAK ARAŞTIRMASI 5
2.1. Kalp Hastalığı Tahminine Yönelik Yapılan Makine Öğrenmesi Çalışmaları 5
2.2. Diyabet Hastalığı Tahminine Yönelik Yapılan Makine Öğrenmesi Çalışmaları.....14
2.3. Çeşitli Hastalıkların Tahmininde Yapılan Makine Öğrenmesi Çalışmaları 20
3. MATERYAL ve YÖNTEM.....26
3.1. Makine Öğrenmesi 26
3.2. Rassal Orman Algoritması 28
3.3. Lojistik Regresyon Algoritması 32
3.4. Aşırı Gradyan Artırma Algoritması 36
3.5. Izgara Arama Algoritması.....40
3.6. Çapraz Doğrulama Yöntemi 42
4. BULGULAR ve TARTIŞMA.....44
4.1. Kalp Hastalığı ve Diyabet Hastalığı Veri Seti Temel Bilgiler.....44
4.1.1. Veri setlerinin incelenmesi ve dengesiz veri setinin dengeli hale getirilmesi…..46
4.1.2. Veri setlerinde istatistiksel bilgiler.....49
4.1.3. Korelasyon matrisinin incelenmesi 51
4.1.4. Isı haritasının incelenmesi.....54
4.1.5. Veri ön işleme 57
4.1.6. Verilerin normalizasyonu.....59
4.1.7. Öznitelik seçimi 61
4.1.8. Doğrulama.....62
4.2. Izgara Arama Algoritması.....63
4.3. Karışıklık Matrisi 64
4.4. Rassal Orman Algoritması Sonuçları.....66
4.4.1. Kalp hastalığı tahmininde rassal orman algoritması sonuçları 66
4.4.2. Diyabet hastalığı tahmininde rassal orman algoritması sonuçları.....69
4.5. Lojistik Regresyon Algoritması Sonuçları.....72
4.5.1. Kalp hastalığı tahmininde lojistik regresyon algoritması sonuçları.....72
4.5.2. Diyabet hastalığı tahmininde lojistik regresyon algoritması sonuçları.....74
4.6. XGBoost Algoritması Sonuçları 76
4.6.1. Kalp hastalığı tahmininde XGBoost algoritması sonuçları.....76
4.6.2. Diyabet hastalığı tahmininde XGBoost algoritması sonuçları.....78
4.6.3. Yapay zeka teknikleriyle kalp hastalığı tahmini literatür çalışmaları…… 80
4.6.4. Yapay zeka teknikleriyle diyabet hastalığı tahmini literatür çalışmaları.....81
4.6.5. Çapraz doğrulama yöntemi sonuçları.....82
4.7. Arayüz Sistemi 83
4.7.1. Kalp hastalığı tahmini için arayüz sistemi 84
4.7.2. Diyabet hastalığı tahmini için arayüz sistemi 86
5. SONUÇ …. 92

KAYNAKLAR 95
EKLER 103
EK1 Kalp hastalığı tahmini için rassal orman algoritması deney sonuçları 104
EK2 Diyabet hastalığı tahmini için rassal orman algoritması deney sonuçları 105
EK3 Kalp hastalığı tahmini için lojistik regresyon algoritması deney sonuçları 107
EK4 Diyabet hastalığı tahmini için lojistik regresyon algoritması deney sonuçları….. 109
EK5 Kalp hastalığı tahmini için XGBoost algoritması deney sonuçları 111
EK6 Diyabet hastalığı tahmini için XGBoost algoritması deney sonuçları 112
ÖZGEÇMİŞ 114

SİMGELER ve KISALTMALAR DİZİNİ

Simgeler	Açıklama
β_0	Regresyonun kesme noktası
β_n	Modelin eğim katsayısı
e	Euler Sayısı
FN	False Negative (Yanlış Negatif)
FP	False Positive (Yanlış Pozitif)
L1 ceza	L1 ceza katsayısı
L2 ceza	L2 ceza katsayısı
L oranı	L1 ve L2 arasındaki karışım oranı
N	Veri setindeki toplam veri sayısı
σ	Standart sapma değeri
TN	True Negative (Doğru Negatif)
TP	True Positive (Doğru Pozitif)
μ	Veri setinin ortalama değeri
x	Hesaplanacak verinin değeri
X_n	Bağımsız değişkenler
$\hat{y}_i$	Hedeflenen çıktı değeri
x_i	Veri setindeki tüm değerler
Z	Skor değeri

Kısaltmalar	Açıklama
CNN	Convolutional Neural Network (Evrişimsel Sinir Ağı)
IDF	International Diabetes Federation (Uluslararası Diyabet Federasyonu)
KA	Decision Tree (Karar Ağacı)
KNN	K-Nearest Neighbors (K- En Yakın Komşu)
LR	Logistic Regression (Lojistik Regresyon)
NB	Naive Bayes
RF	Random Forest (Rassal Orman)
SHDP	Smart Heart Disease Prediction (Akıllı Kalp Hastalığı Tahmini)
SMO	Sequential Minimal Optimization (Sıralı Minumum Optimizasyon)
SVM	Support Vector Machine (Destek Vektör Makinesi)
TÜİK	Türkiye İstatistik Kurumu
UCI	University of California Irvine
WHO	World Health Organization (Dünya Sağlık Örgütü)
XGBoost	Extreme Gradient Boosting (Aşırı Gradyan Artırma)
YSA	Artificial Neural Networks (Yapay Sinir Ağları)

ŞEKİLLER DİZİNİ

		Sayfa
Şekil 3.1.	Makine öğrenmesi akış diyagramı.....	27
Şekil 3.2.	Rassal orman algoritması karar ağaç yapısı.....	29
Şekil 3.3.	Lojistik regresyon algoritması olasılık dağılımı.....	33
Şekil 3.4.	XGBoost algoritması karar ağacı yapısı.....	38
Şekil 3.5.	Çapraz doğrulama yöntemi.....	42
Şekil 4.1.	Kalp hastalığında değişkenler arasındaki ilişkilerin ısı haritasında incelenmesi.....	56
Şekil 4.2.	Diyabet hastalığında değişkenler arasındaki ilişkilerin ısı haritasında incelenmesi.....	57
Şekil 4.3.	Kalp hastalığı tahmininde rassal orman algoritması karışıklık matris sonuçları.....	67
Şekil 4.4.	Diyabet hastalığı tahmininde dengesiz veri seti rassal orman algoritması karışıklık matrisi sonuçları.....	69
Şekil 4.5.	Diyabet hastalığı tahmininde rassal orman algoritması karışıklık matrisi sonuçları.....	70
Şekil 4.6.	Kalp hastalığı tahmininde lojistik regresyon algoritması karışıklık matrisi sonuçları.....	73
Şekil 4.7.	Diyabet hastalığı tahmininde lojistik regresyon algoritması karışıklık matrisi sonuçları.....	75
Şekil 4.8.	Kalp hastalığı tahmininde XGBoost algoritması karışıklık matrisi sonuçları.....	77
Şekil 4.9.	Diyabet hastalığı tahmininde XGBoost algoritması karışıklık matrisi sonuçları.....	79
Şekil 4.10.	Kalp hastalığı tahmini için kullanıcı arayüz sistemi.....	84
Şekil 4.11.	Kalp hastalığı tahmini arayüz sisteminde eksik bilgi uyarı ekranı.....	85
Şekil 4.12.	Kalp hastalığı tahmini arayüz sisteminde yanlış veri girişi uyarı sistemi.....	86
Şekil 4.13.	Diyabet hastalığı tahmini için kullanıcı arayüz sistemi.....	86
Şekil 4.14.	Diyabet hastalığı tahmini arayüz sistemi gerekli bilgilerin doldurulması.....	87
Şekil 4.15.	Diyabet hastalığı tahmini arayüz sistemi eksik bilgi uyarı ekranı.....	88
Şekil 4.16.	Diyabet hastalığı tahmini arayüz sistemi sonuç ekranı.....	89
Şekil 4.17.	Diyabet hastalığı tahmini arayüz sisteminde gerekli bilgilerin doldurulması.....	90
Şekil 4.18.	Diyabet hastalığı tahmini arayüz sistemi sonuç ekranı.....	91

ÇİZELGELER DİZİNİ

			Sayfa
Çizelge	4.1.	Kalp hastalığı veri seti öznitelik bilgisi…….…	45
Çizelge	4.2.	Diyabet hastalığı veri seti öznitelik bilgisi……	46
Çizelge	4.3.	Kalp hastalığı veri seti istatistiksel bilgiler……	50
Çizelge	4.4.	Diyabet hastalığı veri setindeki özniteliklerin kısaltmalarının açıklanması……	50
Çizelge	4.5.	Diyabet hastalığı veri seti için istatistiksel bilgiler……...	51
Çizelge	4.6.	Kalp hastalığı veri seti değişkenler arasındaki korelasyon matrisi.	53
Çizelge	4.7.	Diyabet hastalığı veri seti değişkenler arasındaki korelasyon matrisi……..……..	54
Çizelge	4.8.	Diyabet hastalığı veri setinde eksik verilerin tespit edilmesine ilişkin sonuçlar……...	58
Çizelge	4.9.	Kalp hastalığı veri setinde normalizasyon sonuçları…….	60
Çizelge	4.10.	Diyabet hastalığı veri setinde normalizasyon sonuçları……	61
Çizelge	4.11.	Karışıklık matrisi……...	65
Çizelge	4.12.	Kalp hastalığın tahmininde rassal orman algoritması literatür sonuçları……	68
Çizelge	4.13.	Diyabet hastalığı tahmininde rassal orman algoritması literatür sonuçları….……	71
Çizelge	4.14.	Kalp hastalığı tahmininde lojistik regresyon algoritması literatür sonuçları…….	74
Çizelge	4.15.	Diyabet hastalığı tahmininde lojistik regresyon algoritması literatür sonuçları…….……...……	76
Çizelge	4.16.	Kalp hastalığı tahmininde XGBoost algoritması literatür sonuçları…….……...	78
Çizelge	4.17.	Diyabet hastalığı tahmininde XGBoost algortiması literatür sonuçları…….……	80
Çizelge	4.18.	Kalp hastalığı tahmininde yapay zeka teknikleri kullanılarak elde edilen literatür sonuçları…….…….……..	81
Çizelge	4.19.	Diyabet hastalığı tahmininde yapay zeka teknikleri kullanılarak elde edilen literatür sonuçları…….……	82
Çizelge	4.20.	Kalp hastalığı ve diyabet hastalığı çapraz doğrulama sonuçları….	83

1. GİRİŞ

Günümüzdeki bilgisayar alanındaki teknolojik gelişmelerle birlikte birçok alanda olduğu gibi sağlık alanında da önemli gelişmeler olmuştur. Teknolojik gelişmeler sayesinde hastalıkların erken tahmininde önemli gelişmeler sağlanmıştır. Çalışmanın temel amacı, sağlık alanında yardımcı destek sistemlerinin kullanılmasıyla hastalıkların erken tahmininde yapay zeka tekniklerinin etkisini ve önemini vurgulamaktır. Bu bağlamda, gelecekteki sağlık uygulamalarında yapay zeka tekniklerinin yaygınlaşmasına katkı sağlamak amaçlanmıştır. Sağlık alanında yapay zeka tekniklerinin kullanımıyla, son dönemlerde giderek yaygınlaşan kalp hastalığı ve diyabet gibi hastalıkların erken tahminine yardımcı olmak amacıyla bir destek sistemi geliştirilmiş ve bu alana katkı sağlanması amaçlanmıştır.

Dünya Sağlık Örgütü (WHO), 2016 yılında yaşanan ölümlerin yaklaşık olarak 17,9 milyonunun, küresel toplam ölümlerin ise %31'inin kardiyovasküler hastalıklara bağlı olduğunu bildirmiştir (World Health Organization, 2018). Türkiye İstatistik Kurumu (TÜİK) verilerine göre, Türkiye'deki kalp hastalıkları sebebiyle birçok insan hayatını kaybetmektedir (TÜİK, 2021).

Kardiyovasküler hastalıklar, dünya genelinde ve ülkemizde en sık görülen hastalık ve ölüm nedenleri arasında yer almaktadır. Kardiyovasküler hastalıkların erken tahmin edilmesi, hastalığın takibi ve tedavisi açısından son derece önemlidir. Bu nedenle, hastalığın etkili bir şekilde tedavi edilmesi ve takip edilmesi kritik bir rol oynamaktadır.

Günümüzde obezite giderek artan bir sorun haline gelmektedir. Obezite, kalp hastalığı riskini artırmaktadır (Kalan ve Yeşil, 2010). Dünya genelinde insan davranışlarında ve yaşam tarzında meydana gelen değişiklikler nedeniyle diyabet hastalığına yakalanan kişi sayısı, 21. yüzyıldan itibaren hızla artmıştır (Zimmet vd., 2001). Dünya çapında hızla artan diyabet hastalığı, günümüzde ciddi bir sağlık sorunudur. 2035 yılına kadar, diyabet hasta sayısının 592 milyona ulaşması beklenmektedir (Coşansu, 2015).

Diyabet hastalığı, insülin ve tedavide kullanılan ilaçların yüksek maliyeti nedeniyle hastaların ve ailelerin sağlık harcamalarında önemli yük olmaktadır. Diyabet Atlas'ına göre, 2040 yılında dünya genelinde diyabet hastası kişi sayısının yaklaşık %35 artması ve tedavi maliyetinin ise %19 artması tahmin edilmektedir (Edition, 2015).

Bilgisayar teknolojilerinin ilerlemesiyle birlikte, insan beyninin çalışma şeklinin modellenmesi sonucunda oluşturulan yapay zeka uygulamaları yaygınlaşmaktadır. Yapay zeka çalışmalarının genel amacı, insan beyninin işleyişini analiz ederek, yeni yöntemler ve teknikler geliştirerek bilimin çeşitli alanlarına katkı sağlamaktır. Uzman sistemler, yapay zeka çalışmalarının odaklandığı önemli uygulama alanlarından biridir. Uzman sistemler, çeşitli sektörlerdeki karar vericilere yardımcı olmak için tasarlanmaktadır. Yapay zeka tekniklerinde sıkça kullanılan uzman sistemler, çeşitli hastalıkların tahmininde doktorlara yardımcı olmak için önemli bir destek sistemidir (Karaduman, 2019).

Teknolojik gelişmelerle birlikte dünyadaki trilyonlarca sağlık verisine erişilebilmektedir. Ancak, büyük miktarda verinin olması, verilerin karmaşık hale gelmesine ve analiz edilmesinin zorlaşmasına neden olmaktadır. Yapay zeka teknikleri, sağlık sektöründeki büyük miktardaki veriyi analiz ederek, veriler arasındaki karmaşık ilişkileri analiz etmektedir. Yapay zeka tekniklerinin sağlık alanında kullanılması, hastalık tahmininde doktorlara yardımcı olabilir (Bilgili, 2022).

Geçmiş yıllardan itibaren, sağlık alanında yapay zeka tekniklerinin kullanımı ve uygulama alanları araştırılmaktadır. Yapay zeka, cerrahi alanında ilk kez 1976 yılında Gunn tarafından akut karın ağrısının teşhis edilmesi üzerine yapılan araştırmada kullanılmıştır. Yapay zeka, sağlık çalışanlarının görevlerinde veri ve bilginin analiz edilmesinde yardımcı olmuştur (Ramesh vd., 2004).

Sağlık sektöründe yapay zekanın uygulanmasıyla birlikte yüksek maliyetli sağlık hizmetlerinin daha ekonomik hale getirilmesi, uzun süre gerektiren sağlık işlemlerinin daha kısa sürede tamamlanmasıyla zaman tasarrufu sağlanması, sağlık hizmetlerinin hızlı bir şekilde sunulması ve sağlık çalışanlarının iş yükünün azaltılması gibi potansiyel

faydaları bulunmaktadır. Ayrıca, sağlık alanında hızla artan karmaşık verileri kısa sürede analiz etmesi, yorumlayabilmesi ve yüksek başarı oranlarıyla tahmin yapabilme yeteneği, yapay zeka tekniklerinin sağlık hizmetlerinde kullanılmasının önemli avantajlarından olmaktadır (Güzel vd., 2022).

Sağlık alanında yapay zeka tekniklerinin kullanımı, hastalıkların tahmin edilmesinde ilgili doktorlara destek sağlayabilir. Ayrıca, yapay zeka tekniklerinin arayüz tasarımı ile birleştirilmesi, doktorların hastalık tahmininde daha etkili bir şekilde yapay zekayı kullanmalarına yardımcı olabilir. Bu çalışma, sağlık sektöründe yapay zeka tekniklerinin daha etkin bir şekilde kullanılmasına yönelik literatüre katkıda bulunmayı amaçlamaktadır.

Makine öğrenmesi, yapay zeka tekniklerinden en yaygın kullanılan ve başarılı performans veren yapay zekanın alt dallarından biridir. Makine öğrenmesi, veri analizi ve modelleme yaparken istatistiksel temellere dayanarak modeli öğrenip gelecekteki sonuçları tahmin etmeye çalışmaktadır (Davenport, 2018).

Teknolojinin gelişmesiyle birlikte erişilebilen veri miktarının artması, karmaşıklığın ve analiz sürecinin zorluğunun artmasına neden olmuştur. Bu durum, bilgisayar sistemlerine dayalı yapay zeka yöntemlerine olan ihtiyacı artırmıştır. Makine öğrenimi teknikleri, verileri kolayca analiz edebilir, eğitebilir ve öğrendiği bilgilere dayalı tahminler yapabilir. Ayrıca, belirsizlik altında kararlar verebilme yeteneği sayesinde makine öğrenimi teknikleri tercih edilmektedir. Makine öğrenimi, sınıflandırma ve regresyon problemlerini çözmek için yaygın olarak kullanıldığından, geniş bir kullanım alanına sahip ve farklı türde problemlere uyarlanabilme esnekliği nedeniyle tercih edilen yöntemlerden biridir (Murphy, 2012).

Makine öğrenmesi, farklı algoritmalar, teknikler ve yöntemler içeren geniş bir alandır. Makine öğrenmesi, finans, sağlık, otomotiv ve çeşitli diğer sektörlerde yaygın olarak kullanılmaktadır. Veri miktarının artması, geniş bir kullanım alanı, tahmin başarısı ve veriye dayalı sonuçlar elde etme yeteneği gibi nedenlerle, makine öğrenmesi sektörler tarafından tercih edilmektedir. Makine öğrenmesinde genellikle veri seti eğitim, test ve

doğrulama olarak üç farklı bölüme ayrılarak modelin performansı değerlendirilmektedir. Eğitim seti, modelin öğrenme sürecinde kullanılırken, test ve doğrulama setleri, modelin genelleme yeteneğini değerlendirmek amacıyla ayrılmaktadır. Eğitim setiyle eğitilen model, test seti üzerinde değerlendirilmektedir ve sonuçlar genelleme yeteneği ve modelin performansını vermektedir. Doğrulama seti ise eğitilen ve test edilen modelin performansını daha detaylı kontrol etmek için kullanılmaktadır (Géron, 2022).

Makine öğrenimi teknikleri, sağlık sektöründe önemli avantajları nedeniyle yaygın olarak kullanılmaktadır. Bu teknikler, tıbbi verileri analiz ederek hastalıkları hızlı ve yüksek doğruluk oranıyla tahmin edebilme yeteneğine sahiptir. Sağlık uzmanlarına hastalıkların tahmini ve tedavisinde yardımcı olmaktadırlar. Ayrıca, görüntü işleme yetenekleri sayesinde tıbbi görüntüleme alanında da yaygın olarak kullanılmaktadırlar. Bu teknikler, tıbbi görüntülerden önemli bilgiler elde etmek ve hastalıkların tahmini ve izlenmesinde büyük ölçüde etkili olmaktadır. Makine öğrenimi, büyük veri kümelerini kolayca öğrenebilir ve analiz edebilir, bu da yeni tedavi tekniklerinin geliştirilmesine katkıda bulunmaktadır. Ayrıca, makine öğrenimi hastalıkların tahmininde duyarlılıkları analiz etmek ve belirsizliği ölçmek için kullanılmaktadır (Peng vd., 2021).

2. KURAMSAL TEMELLER ve KAYNAK ARAŞTIRMASI

Teknolojinin ve bilgisayarların gelişmesiyle birlikte yapay zekanın, çeşitli alanlarda kullanımı yaygınlaşmıştır. Sağlık alanında yapay zeka tekniklerinin kullanımı giderek artmaktadır. Özellikle yapay zekanın alt dallarından olan makine öğrenmesi teknikleri, büyük veri analiziyle sağlık sektöründe çeşitli hastalıkların tahmininde başarılı sonuçlar elde etmesiyle ön plana çıkmaktadır. Literatürde benzer çalışmalara bakıldığında çeşitli hastalıkların tahmininde makine öğrenmesi tekniklerinin yüksek performans gösterdiği belirlenmiştir. Yapay zeka tekniklerinin çeşitli hastalıkların tahmininde ve tedavi aşamasında sağlık çalışanlarına yardımcı olabileceği gözükmektedir. Makine öğrenmesi tekniklerinin farklı hastalıkları kolayca anlayabilmesi, farklı veri setlerine uyum sağlayabilmesi, elektronik ortamda artan veri setlerinin çoğalması, zaman tasarrufu, tahmin sürecini hızlandırması ve başarılı sonuçlar vermesiyle kullanımının yaygınlaşacağı düşünülmektedir. Tez çalışması kapsamında, kalp hastalığı ve diyabet hastalığı için çeşitli makine öğrenmesi teknikleri kullanılarak tahmin çalışması gerçekleştirilmiştir. Ayrıca, çalışmada diyabet hastalığı veri seti dengesiz veri setiyken yapay veri kullanılarak dengeli hale getirilmiş ve bu sayede tahmin performans başarısı artırılmıştır. Tez çalışmasında, kalp hastalığı için XGBoost Algoritması'nın en yüksek doğruluk performansını sağladığı belirlenmiş, diyabet hastalığı için ise Rassal Orman Algoritması en yüksek doğruluk göstermiştir. Ayrıca, XGBoost Algoritması kullanılarak kalp hastalığı tahmini için bir arayüz tasarımı geliştirilmiş ve Rassal Orman Algoritması tabanlı diyabet hastalığı tahmini için ise başka bir arayüz tasarımı yapılmıştır. Tez çalışmasında yapay zeka tekniklerinin sağlık sektöründe kullanımının avantajlarından bahsedilmiş, doktorlara ve sağlık çalışanlarına yardımcı sistem olarak literatüre katkı sağlanması amaçlanmıştır.

2.1. Kalp Hastalığı Tahminine Yönelik Yapılan Makine Öğrenmesi Çalışmaları

Günümüzde yaygınlaşan kalp hastalığının erken tahmini, sağlık alanında büyük önem taşımaktadır. Bu bağlamda, makine öğrenmesi tekniklerinin kullanımı, farklı veri kaynaklarıyla esnek bir uyum sağlama kabiliyeti ve avantajları nedeniyle önemli bir

araştırma alanı olarak öne çıkmaktadır. Kalp hastalığının erken tahmininde makine öğrenimi yöntemlerinin kullanılmasıyla ilgili yapılan kaynak araştırmaları incelendiğinde önemli başarılar elde edildiği gözükmektedir. Farklı veri setleri ve çeşitli ülkelerde gerçekleştirilen araştırmalar, bu tekniklerin kalp hastalığı tahmininde oldukça başarılı sonuçlar verdiğini göstermektedir. Literatürdeki diğer çalışmalara bakıldığında, makine öğrenmesi modellerinin kullanılması ve arayüz tasarımının doktorlara ve sağlık çalışanlarına erken tahmin ve tedavi planlaması konusunda yardımcı olabileceği görülmektedir.

Palaniappan ve Awang (2008) tarafından gerçekleştirilen çalışmada, 909 hasta verisi ve 13 farklı öznitelik kullanılarak Akıllı Kalp Hastalığı Tahmin Sistemi geliştirilmiştir. Çalışmada, Naive Bayes, Yapay Sinir Ağları ve Karar Ağaçları gibi farklı Veri Madenciliği teknikleri kullanılarak tahmin modelleri oluşturulmuştur. Çalışmada, Naive Bayes yöntemi %88,12, Yapay Sinir Ağları %85,68 ve Karar Ağaçları yöntemi %80,40 doğruluk performansı göstermiştir.

Ani vd. (2016) çalışmalarında, koroner kalp hastalığını tahmin etmek için makine öğrenimi tekniklerini kullanarak bir karar destek sistemi geliştirmişlerdir. Çalışmada Naive Bayes, K-en yakın komşu, Karar ağacı-C4.5, ve Rassal orman algoritmaları kullanılarak 10 çapraz doğrulama yöntemi uygulanmıştır. Sonuçlara bakıldığında, Naive Bayes %84,00, K - En Yakın Komşu %77,00, Karar Ağacı-C4.5 %81,00 ve en yüksek doğruluk performansı Rassal orman algoritması kullanılarak %89,00 olarak elde edilmiştir.

Deepika vd. (2016) çalışmalarında, UCI sitesinden alınan diyabet ve kalp hastalığı veri setlerini kullanarak Naive Bayes, Karar Ağacı, Destek Vektör Makinesi ve Yapay Sinir Ağları makine öğrenmesi tekniklerini kullanarak tahmin çalışması gerçekleştirmişlerdir. Kullanılan sınıflandırıcılardan, kalp hastalığı tahmini için en yüksek doğruluk performansı %95,56 ile Destek Vektör Makinesi, diyabet hastalığı tahmini için ise en yüksek doğruluk oranı %73,59 ile Naive Bayes algoritmasıyla elde edilmiştir.

Weng vd. (2017) çalışmasında, İngiltere'deki aile hekimlerinden alınan 10 yıllık hasta kayıt verilerini kullanarak kardiyovasküler hastalıkların tahmin edilmesine yönelik makine öğrenimi algoritmalarıyla araştırma gerçekleştirmiştir. Makalede, lojistik regresyon, rastgele orman, gradyan artırma makineleri ve sinir ağları gibi farklı algoritmalar kullanıldığı ve bu algoritmaların kardiyovasküler hastalıkların tahmin edilmesinde geleneksel risk tahmin yöntemlerine göre daha başarılı sonuçlar verdiği görülmüştür. Bu sonuçlar, elektronik sağlık verilerinin kullanıldığı geniş bir veri setiyle desteklenmiştir. Bulgular, makine öğrenimi tekniklerinin kardiyovasküler hastalıkların tahmininde önemli bir araç olabileceğini ortaya koymaktadır.

Dwivedi (2018) tarafından gerçekleştirilen çalışmada, kalp hastalığını tahmin etmek için 270 hasta verisi ve 13 öznitelik kullanılarak çeşitli makine öğrenme tekniklerinin başarısı değerlendirilmiştir. Lojistik Regresyon kullanılarak %85,00 Naive Bayes ile %83,00 K-En Yakın Komşu kullanılarak %80,00, Destek Vektör Makinesi kullanılarak %82,00, Yapay Sinir Ağları yöntemi kullanılarak %84,00 doğruluk performans değerleri elde edilmiştir. Çalışmada Lojistik Regresyonun diğer tekniklere göre daha iyi performans gösterdiği belirlenmiştir.

Kohli ve Arora (2018) çalışmalarında, UCI kütüphanesinden elde edilen verilerle Kalp Hastalığı, Diyabet Hastalığı ve Meme Kanseri hastalıklarının tahmininde makine öğrenimi tekniklerini kullanmışlardır. Çalışmada, Cleveland veri seti üzerinde kalp hastalığını tahmin etmek için önem derecesi düşük olduğu belirlenen iki öznitelik çıkarılarak, kalan 11 öznitelik kullanılmıştır. Bu durumda, Lojistik Regresyon yöntemiyle elde edilen doğruluk performansı %87,10 ve Rassal Orman Algoritması kullanılarak doğruluk performansı %77,40 olarak elde etmişlerdir.

Özmen vd. (2018) Cleveland veri setinde 303 hasta verisi üzerinde ve 13 öznitelik kullanarak kalp hastalığı tahmini çalışması yapmıştır. Rassal Orman Algoritması kullanarak elde edilen sonuçlara göre, %82,50 oranında duyarlılık ve %84,21 doğruluk oranı elde ederken, Lojistik Regresyon yöntemi kullanarak %87,50 duyarlılık oranı ve %85,53 doğruluk performansı elde etmişlerdir.

Harjai ve Khatri (2019) çalışmalarında, Cleveland Kalp Hastalığı veri seti üzerinde 297 hasta verisi ve önem derecesi yüksek 9 öznitelik kullanarak Yapay Sinir Ağları yöntemiyle %89,20 doğruluk performansı elde etmişlerdir.

Krishnani vd. (2019) çalışmalarında, koroner kalp hastalıklarını tahmin etmek amacıyla Framingham veri setini kullanmışlardır. Çalışmada, Makine Öğrenimi teknikleri olarak Rassal Orman Algoritması, Karar Ağacı ve K-En Yakın Komşu gibi yöntemler kullanılmıştır. Modelin performansı K-katmanlı çapraz doğrulama yöntemi kullanılarak değerlendirilmiştir. Rassal Orman Algoritması kullanıldığında en yüksek doğruluk performansı elde edilmiş ve %96,80 olarak bulunmuştur. Karar Ağacı yöntemiyle elde edilen doğruluk %92,70, K-En Yakın Komşu yöntemiyle ise %92,89 olarak belirlenmiştir. Bu çalışma, koroner kalp hastalıklarının tahmininde Rassal Orman algoritmasının en etkili yöntem olduğunu ortaya koymaktadır.

Lakshmanarao vd. (2019) çalışmasında, kalp hastalığı tahmini için 4240 hasta verisi ve 15 öznitelik kullanarak makine öğrenmesi teknikleriyle tahmin çalışması yapmışlardır. Elde edilen sonuçlara göre, Rassal Orman Algoritması ve Ekstra Ağaç Sınıflandırıcı, diğer makine öğrenmesi yöntemlerine kıyasla daha başarılı bulunmuştur. Bu iki algoritma ile %90,30 doğruluk performansı elde edilmiştir ve bu algoritmaların tahminlemedeki yüksek başarısı gösterilmiştir.

Mohan vd. (2019) çalışmalarında, hibrit makine öğrenmesi tekniklerini Cleveland Kalp Hastalığı veri setindeki 297 hasta verisi üzerinde ve 13 öznitelik kullanarak gerçekleştirmişlerdir. Çalışmada, geleneksel makine öğrenimi algoritmalarının yanı sıra karma veya hibrit makine öğrenimi tekniklerini de kullanmışlardır. Çalışmalarında, rassal orman (RF) algoritması kullanılarak %90,10 oranında kesinlik, %92,8 oranında duyarlılık, %90,00 oranında F1 puanı, %88,40 doğruluk oranı değerleri elde edilmiştir. Ayrıca çalışmalarında lojistik regresyon (LR) algoritması kullanılarak %89,60 oranında kesinlik, %91,10 oranında duyarlılık, %90,20 duyarlılık ve %82,90 oranında doğruluk elde etmişlerdir.

Repaka vd. (2019) çalışmalarında, Akıllı Kalp Hastalığı Tahmini (SHDP: Smart Heart Disease Prediction) için UCI (University of California Irvine) Makine Öğrenimi Deposun'dan kalp hastalarının tıbbi verilerini kullanmışlardır. Çalışmada çeşitli teknikler kullanılmış ve en yüksek doğruluk performansı %89,77 ile Naive Bayes algoritmasıyla elde edilmiştir. Çalışmanın devamında, araştırmacılar kalp hastalığı tahmini için bir sistem tasarımı ve uygulaması geliştirmişlerdir.

Anbuselvan (2020) çalışmasında, Cleveland kalp hastalığı veri seti üzerinde kalp hastalığının tahmini için XGBoost algoritması tercih edilmiştir. Çalışmada, toplamda 303 hasta verisi ve bu hastaların 13 farklı özniteliği kullanılmıştır. Yapılan analiz sonucunda elde edilen doğruluk oranı %78,69 olarak belirlenmiştir.

Ekrem vd. (2020) çalışmalarında, yapay zeka teknikleri kullanarak kalp hastalığının tahminini incelemişlerdir. Çalışma, Cleveland Kalp Hastalığı veri setindeki 303 hasta üzerinde gerçekleştirilmiş olup, bu veri seti 13 öznitelik içermektedir. Hastalığın tahmininde önemi daha az olan öznitelikler çıkarılmıştır. RF kullanılarak altı ve dokuz öznitelikle yapılan analizler sonucunda %86,99 oranında doğruluk elde edilmiştir.

Farzana ve Veeraiah (2020) tarafından yapılan çalışmada, çoklu makine öğrenme teknikleri kullanılarak 304 hasta verisi ve 13 öznitelik üzerinden kalp hastalığının tahmini gerçekleştirilmiştir. Araştırma sonuçlarına göre, RF %88,52 doğruluk performansı sağlarken, XGBoost yöntemi %78,69 doğruluk oranı elde etmiştir. Bu bulgulara göre, RF, doğruluk performansı açısından XGBoost yönteminden daha yüksek bir başarıya ulaşmıştır.

Islam vd. (2020) çalışmalarında, 301 hasta kaydı ve 13 öznitelikten oluşan bir veri seti kullanmışlardır. Kalp hastalığı tahmininde, önem derecesi daha düşük olan bir öznitelik çıkarılmış ve makine öğrenimi teknikleriyle kardiyovasküler hastalık tahmini yapılmıştır. Lojistik Regresyon kullanılarak %89,00 kesinlik oranı, %86,00 duyarlılık oranı, %86,00 F1 puanı ve %82,90 doğruluk performans değerleri gözlemlenmiştir.

Li vd. (2020) çalışmalarında, Kalp Hastalığı'na ait 303 hasta kaydından oluşan ve 75 öznitelik içeren bir Cleveland veri seti kullanmışlardır. Çalışma kapsamında, 6 eksik veri kaydı çıkarılmış ve 13 öznitelik seçilerek kalp hastalığı tahmini yapılmıştır. Makine öğrenimi sınıflandırıcıları olarak Lojistik Regresyon, K-En Yakın Komşu, Yapay Sinir Ağları, Destek Vektör Makinesi, Naive Bayes ve Karar Ağacı algoritmaları kullanılmıştır. Ayrıca, çapraz doğrulama yöntemi de uygulanmıştır. Çalışmanın sonuçlarına göre, algoritmalar arasından en yüksek başarıyı Destek Vektör Makinesi göstermiş ve doğruluk performansı %92,37 olarak tespit edilmiştir.

Taşçı ve Şamlı (2020) çalışmalarında, Cleveland veri setini kullanarak kalp hastalığını tahmin etmek için WEKA programını kullanmışlardır. Çalışmada, 303 hasta üzerinde yapılan analizlerde 13 öznitelik kullanılmıştır. Rassal Orman Algoritması ile elde edilen sonuçlarda %83,80 doğruluk oranı belirlenirken, Lojistik Regresyon yöntemi kullanılarak %85,25 doğruluk oranı elde edilmiştir.

Asif vd. (2021) Cleveland veri setini kullanarak kalp hastalığı tahmini çalışmasında Lojistik Regresyon yöntemi ile %86,10 oranında kesinlik, %94,00 oranında duyarlılık, %89,80 oranında F1 puanı, %89,00 oranında doğruluk performansı elde ederken XGBoost kullanarak %85,20 kesinlik oranı, %88,00 duyarlılık oranı, %87,00 F1 puanı ve %85,20 doğruluk performansı elde etmiştir.

Bertsimas vd. (2021) çalışmalarında, sağlık sektöründe teknolojik gelişmelerle EKG'deki verilerin dijital ortamlarda bulunabilmesi, verilere kolayca erişilebilmesi ve makine öğrenimi tekniklerinin analizindeki başarının önemini vurgulayarak, kalp hastalığını tahmin etmede XGBoost algoritmasının yüksek performansından bahsetmektedir. Bununla birlikte, araştırma, kalp hastalığının tahmin edilmesinde öznitelik seçiminin ne kadar önemli olduğunu göstermiştir. Çalışmada, XGBoost algoritmasıyla eğitilen modellerin genellikle F1 Skorlarının %93,00 ile %99,00 arasında bulunduğu belirtilmektedir.

Coşar ve Deniz (2021) çalışmalarında, Cleveland veri seti kullanılarak 918 hasta verisi ve 11 öznitelik ile kalp hastalığı tahmini yapılmıştır. Çalışmalarında, makine öğrenmesi tekniklerinden Lojistik Regresyon ve Rassal Orman yöntemlerini kullanmışlardır. Rassal Orman ile %88,00 doğruluk elde edilirken, Lojistik Regresyon ile %85,00 doğruluk oranı elde edilmiştir.

Lin (2021) çalışmasında, makine öğrenmesi tekniklerinin karar verme ve tahmin yeteneği sayesinde çeşitli hastalıkların tahmininde önemli sonuçlar elde edilebileceği belirtilmektedir. Araştırmada, XGBoost algoritması kullanılarak kalp hastalığının tahmininde %80,20 F1 puanı ve %82,10 doğruluk oranına ulaşılmıştır.

Rahim vd. (2021) çalışmalarında, kardiyovasküler hastalıklar için Makine Öğrenimi tekniklerini kullanarak Cleveland veri seti dahil olmak üzere üç farklı veri setini incelemişlerdir. Çalışmada, üç farklı veri seti üzerinde öznitelik seçimi yapılan ve yapılmayan modeller oluşturulmuştur. Öznitelik seçimi yapılarak Framingham, Heart Disease ve Cleveland veri setlerinin performansları %99,10, %98,00 ve %95,50 doğruluk oranlarına yükseltilmiştir. Araştırma sonuçları, model başarısının veri ön işleme, eksik verilerin giderilmesi, öznitelik seçimi ve model eğitimi gibi faktörlerin önemine vurgu yapmaktadır. Özellikle, öznitelik seçiminin modelin performansı üzerinde oldukça etkili olduğu gözlemlenmiştir.

Vatansever vd. (2021) ABD Cleveland eyaletindeki hasta kayıt verilerinden kalp hastalığı tahmini için 303 hasta verisi ve 13 öznitelik kullanarak doğruluk tahmini için toplamda 28 deney gerçekleştirmişlerdir. Çalışmalarında Rassal Orman Algoritması yöntemiyle %83,60 doğruluk performansı elde etmişlerdir.

Budholiya vd. (2022) çalışmalarında Cleveland kalp hastalığı tahmini için veri setini XGBoost yöntemi kullanarak analiz etmişlerdir. Çalışmada, 303 hasta verisi ve 13 öznitelik kullanılarak elde edilen doğruluk performansı %91,80 olarak bulunmuştur.

Doki vd. (2022) tarafından yapılan çalışmada, kalp hastalığının tahmininde XGBoost algoritması tercih edilmiştir. Çalışmada, 303 hasta verisi üzerinde detaylı bir analiz gerçekleştirilmiş ve hastalığın tahmininde önem derecesi düşük olan 7 öznitelik belirlenerek bunlar çıkarılmıştır. Kalan 6 öznitelikle model oluşturulmuş ve elde edilen sonuçlar, %85,96 doğruluk oranıyla sonuçlanmıştır.

Gupta vd. (2022) kalp hastalığı tahmini için denetimli makine öğrenmesi yöntemini kullanmıştır. UCI'den aldığı veri setiyle Python programını kullanarak tahminleme çalışması yapmıştır. Çalışma, 303 hasta verisi üzerinde gerçekleştirilmiş olup, bu veri seti 13 farklı öznitelik içermektedir. Rassal Orman Algoritması yöntemi kullanılarak elde edilen sonuçlar; kesinlik oranı %88,00, duyarlılık oranı %86,27, F1 puanı %87,13 ve doğruluk oranı %85,71 olarak belirlenmiştir. Lojistik Regresyon yöntemi kullanılarak elde edilen sonuçlarda %90,74 kesinlik oranı, %96,08 duyarlılık oranı, F1 puanı %93,34 ve %92,30 doğruluk performansı elde edilmiştir.

Ramesh vd. (2022) çalışmalarında makine öğrenmesi teknikleri kullanılarak kalp hastalığının öngörülebilir analizi incelenmiştir. Cleveland veri setinde bulunan 303 hasta verisi ve 13 öznitelikten, korelasyon haritasına göre tahminlemede daha az öneme sahip olan öznitelikler çalışma kapsamından çıkarılmıştır. Bu nedenle, çalışmada toplam 7 öznitelik kullanılmıştır. Çalışma sonuçlarına göre, Rassal Orman Algoritması kullanılarak elde edilen duyarlılık oranı %83,20, F1 puanı %87,90 ve doğruluk oranı %88,35 olarak belirlenmiştir. Aynı çalışmada Lojistik Regresyon yöntemiyle elde edilen kesinlik oranı %88,10, duyarlılık oranı %86,10, F1 puanı %86,50 ve doğruluk oranı %84,22 olarak belirlenmiştir. Rassal Orman Algoritması'nın kalp hastalığını tahmin etme yeteneğinin, Lojistik Regresyon'a kıyasla daha yüksek olduğu gösterilmiştir.

Saboor vd. (2022), Cleveland veri setini kullanarak gerçekleştirdikleri çalışmada, 303 hasta verisi ve 13 öznitelik üzerinden makine öğrenmesi algoritmalarını kullanarak kalp hastalığı tahminini iyileştirmeye odaklanmışlardır. Araştırmanın sonuçlarına göre, XGBoost yöntemi kullanılarak %91,80 doğruluk oranına ulaşılmıştır.

Srinivas ve Katarya (2022), kardiyovasküler hastalıkların tahminine yönelik çalışmalarında XGBoost yöntemini uygulamışlardır. Çalışma sonuçlarına göre, %92,65 kesinlik, %95,34 duyarlılık, %92,98 F1 puanı ve %94,70 doğruluk performansı elde edildiği gözlemlenmiştir. Bu sonuçlar, XGBoost algoritmasının kardiyovasküler hastalıkların tahmininde başarılı olduğunu ortaya koymaktadır.

Vasanthi vd. (2022) çalışmalarında, Cleveland veri setini kullanarak kalp hastalığını Rassal Orman Algoritması yöntemiyle tahmin etmeye çalışmışlardır. Çalışmada, 303 hasta verisi üzerinde yapılan analizlerde 13 öznitelik kullanılmış ve %88,52 doğruluk performansı elde edilmiştir.

Baviskar vd. (2023) çalışmalarında iki farklı veri seti kullanarak hibrit derin öğrenme sınıflandırma modellerini kullanarak ve öznitelik seçiminin önemini vurgulayarak kalp hastalığı tahminini incelemektedir. Çalışma sonuçları, mevcut en iyi yöntemlere kıyasla daha yüksek performans ve doğruluk sağlandığını göstermektedir. UCI kütüphanesinden aldıkları Cleveland veri seti için %99,21 doğruluk elde edilirken, Klinik Kalp Yetmezliği veri seti için %93,56 doğruluk performansı elde edilmiştir.

Dissanayake ve Johar (2023) çalışmalarında, kalp hastalığının tahmininde makine öğrenmesi tekniklerini üç farklı veri setiyle değerlendirmiştir. Cleveland veri seti üzerinde XGBoost algoritması kullanıldığında %95,38 kesinlik, %85,19 duyarlılık, %90,20 F1 puanı ve %91,80 doğruluk oranı elde edilmiştir. Statlog veri setinde ise XGBoost algoritması kullanıldığında %82,58 doğruluk elde edilirken, VA Long Beach veri seti üzerinde ise XGBoost algoritması kullanılarak %75,50 doğruluk oranına ulaşılmıştır.

Uddin vd. (2023) çalışmalarında kardiyovasküler hastalıklara üç farklı veri setini birleştirerek makine öğrenmesi teknikleriyle incelemişlerdir. Çalışmada kullandıkları makine öğrenmesi tekniklerinden Karar Ağacı yöntemi %99,16 en yüksek performans değerini göstermiştir. Araştırma sonucuna göre, kardiyovasküler hastalıkların makine öğrenmesi teknikleriyle incelenmesinin önemi vurgulanmıştır.

Zeng (2023) çalışmasında, UCI veri tabanındaki birbirinden bağımsız Cleveland veri seti dahil olmak üzere beş farklı kalp hastalığı veri setini birleştirerek toplamda 918 veri kaydı ve 11 öznitelik kullanarak makine öğrenmesi teknikleriyle kalp hastalığı tahmini yapmıştır. Sonuçlara göre, XGBoost yöntemiyle elde edilen performans değerleri: Duyarlılık oranı %89,30, F1 skoru %90,40 ve doğruluk performans değeri %88,40; Destek Vektör Makinesi yöntemiyle elde edilen performans değerleri: Duyarlılık oranı %89,30, F1 skoru %90,70 ve doğruluk performans değeri %88,88, K-En Yakın Komşu yöntemiyle elde edilen performans değerleri: Duyarlılık oranı %87,60, F1 skoru %89,40 ve doğruluk performans değeri %87,30, Karar Ağacı yöntemiyle elde edilen performans değerleri: Duyarlılık oranı %83,40, F1 skoru %86,00 ve doğruluk performans değeri %83,33 olarak elde edilmiştir.

2.2. Diyabet Hastalığı Tahminine Yönelik Yapılan Makine Öğrenmesi Çalışmaları

Diyabet hastalığının erken tahmini, hastalığın yönetimi ve tedavisi açısından büyük önem taşımaktadır. Bu bağlamda, çeşitli veri kaynaklarından elde edilen klinik verilerin kullanılmasıyla geliştirilen makine öğrenmesi modelleri, diyabetin erken tahmini ve tedavisi konusunda önemli etkiye sahiptir. Literatürde yapılan araştırmalar, farklı veri setleri ve çeşitli hastane kayıtlarının analiz edilerek yapılan diyabet hastalığıyla ilgili çalışmalara odaklanmıştır. Bu çalışmalar, makine öğrenmesi modellerinin diyabet hastalığının erken tahmini ve tedavisi için etkin bir araç olarak kullanılabileceğini göstermektedir.

Nai-Arun ve Moungmai (2015) tarafından yapılan çalışmada, diyabet hastalığı riskinin tahmin edilmesinde makine öğrenimi sınıflandırıcıları karşılaştırılmaktadır. Veri seti Sawanpracharak Bölge Hastanesindeki kayıtlardan alınmıştır. Yaygın olarak kullanılan Karar Ağacı, Yapay Sinir Ağları, Lojistik Regresyon ve Naive Bayes sınıflandırıcıları incelenmiş ve modelin dayanıklılığını artırmak için Torbalama (Bagging) ve Güçlendirme (Boosting) teknikleri kullanılmıştır. Bu algoritmalardan en yüksek başarıyı %85,56 ile Rassal Orman Algoritması sağlamıştır. Rassal Orman Algoritması ile oluşturulan model, diyabet risk tahmini için web sitesi uygulamasında kullanılmıştır.

Kaur ve Kumari (2018) çalışmalarında, Pima Kızılderili Diyabet Hastalığı veri setini kullanarak beş farklı model geliştirmişlerdir. Ayrıca diyabet veri setinde öznitelik seçimi, Borusal Sarmal Algoritması kullanılarak yapılmış ve tüm modeller yüksek performans değerleri elde etmiştir. Sonuçlar, Destek Vektör Makinesi-Doğrusal modeli kullanıldığında %89,00 doğruluk ve %88,00 kesinlik performansı elde edilirken, KNN modeli kullanıldığında %90,00 geri çağırma ve %88,00 F1 skoru değeri elde edilmiştir. Destek Vektör Makinesi-Doğrusal modeli için AUC değeri %90,00, KNN modeli için ise %92,00 olarak belirlenmiştir. Araştırma sonuçlarına göre, Borusal sarmal algoritması kullanılarak öznitelik seçiminin performansı önemli ölçüde arttırdığı görülmüştür.

Kohli ve Arora (2018) çalışmalarında, Diyabet hastalığını tahmin etmek amacıyla 768 hasta verisi ve 8 öznitelikten oluşan Pima Kızılderili Diyabet Hastalığı veri setini kullanmışlardır. Önem derecesi daha az olduğu belirlenen 3 öznitelik çıkarılarak, kalan 5 öznitelik kullanılmıştır. Çalışma sonuçlarına göre, Lojistik Regresyon ile elde edilen doğruluk oranı %84,42, Karar Ağacı Algoritması ile %74,03, Rassal Orman Algoritması ile %81,82, Destek Vektör Makinesi ile %85,71 ve AdaBoost Algoritması ile %80,52 olarak belirlenmiştir.

Choi vd. (2019) çalışmasında, diyabet hastalığı tahmini için Kore Üniversitesi Guro Hastanesi'nden, diyabet dışı geçmişi olan 5 yıllık 8454 hasta kaydı verileri toplanarak makine öğrenimi algoritmaları kullanıldıktan sonra 10 kat çapraz doğrulama uygulanmıştır. Araştırma sonuçlarına göre, diyabet hastalığı tahmini yaparken Lojistik Regresyon yöntemi en yüksek doğruluk performansını göstermiştir.

Lukmanto vd. (2019) tarafından yürütülen çalışma kapsamında, Pima Indian Diabetes (PID) veri seti üzerinde gerçekleştirilen analizler sonucunda, Destek Vektör Makinesi (SVM) kullanılarak %89,02 gibi yüksek bir doğruluk oranı elde edilmiştir. Sonuçlara göre, öznitelik seçimi ve bulanık mantık kullanımının diyabet teşhisi ve sınıflandırmasında etkili bir yaklaşım olduğu görülmektedir.

Balasubramanian vd. (2020) çalışmalarında, diyabet hastalığının tahmininde 390 hasta kaydı ve 13 öznitelik kullanılarak Rassal Orman Algoritması, Lojistik Regresyon ve XGBoost algoritmalarının performansı değerlendirilmiştir. Rassal Orman Algoritması kullanıldığında, %67,00 kesinlik oranı, %50,00 duyarlılık oranı, 0,57 F1 puanı ve %88,00 doğruluk performansı elde edilmiştir. Lojistik Regresyon algoritması kullanıldığında, %80,00 kesinlik oranı, %33,00 duyarlılık oranı, 0,47 F1 puanı ve %88,00 doğruluk performansı belirlenmiştir. XGBoost algoritması kullanıldığında ise, %80,00 kesinlik oranı, %33,00 duyarlılık oranı, 0,47 F1 puanı ve %88,00 doğruluk performansı göstermiştir.

Nadesh vd. (2020) çalışmalarında, diyabet hastalığının tahmini için PIMA veri seti kullanılmış ve Derin Sinir Ağları (DSA) sınıflandırıcısı ile bir model geliştirilmiştir. Öznitelik seçimi için Ekstra Ağaçlar ve Rassal Orman Algoritması kullanılmıştır. Elde edilen sonuçlar, modelin %98,16 gibi yüksek bir doğruluk oranına sahip olduğunu göstermektedir. Araştırma, diyabet hastalığı tahmininde derin öğrenme tekniklerinin performansını ve öznitelik seçiminin önemini vurgulamaktadır.

Tigga ve Garg (2020) çalışmasında, makine öğrenmesi teknikleri kullanılarak 18 sorudan oluşan bir anketin sonuçları PIMA Kızılderili veri setine de uygulanarak diyabet hastalığının tahmini yapılmıştır. Diğer algoritmalarla karşılaştırıldığında, Rassal Orman Algoritması daha yüksek başarı elde etmiştir. Sonuçlara göre, bu analizlerde Rassal Orman Algoritması %94,10 doğruluk performansı sergilemiştir. İki veri setinde de yüksek başarı elde eden Rassal Orman Algoritması, diyabetin tahmininde etkili bir model olarak belirlenmiştir.

Veranyurt vd. (2020) çalışmalarında, sağlık sektöründe makine öğrenimi tekniklerinin kullanımını artırmak amacıyla Vanderbilt Üniversitesi Biyoistatistik Departmanı'ndan elde edilen 390 hasta verisi ve 15 öznitelikten oluşan diyabet veri setini kullanarak incelemiştir. Rassal Orman Algoritması ile yaptıkları analiz sonucunda %92,30 doğruluk oranı elde edilmiş ve yapay zeka tekniklerinin hastalık tahmini açısından önemini vurgulamıştır.

Akyol ve Kargacı (2021) çalışmalarında, diyabet hastalığının tahmin edilmesi için Bangladeş, Sylhet'teki Sylhet Diyabet Hastanesinden alınan 520 hasta kaydı ve 16 öznitelik kullanılarak çeşitli makine öğrenme algoritmalarının performansları değerlendirilmiştir. Çalışmalarında Gradyan Artırma, Rastgele Orman Algoritması, K-En Yakın Komşu, DSA kullanılmış olup en yüksek başarı Oylama Topluluğu Sınıflandırıcısı Modeli kullanılarak %100 sınıflandırma doğruluğu ile elde edilmiştir. Ayrıca 5 katlı çapraz doğrulama yöntemiyle ortalama %97,31 sınıflandırma doğruluğu sağlanmıştır.

Rajendra ve Latifi (2021) çalışmasında, diyabet hastalığı için PIMA Kızılderili veri seti ile Vanderbilt Üniversitesi'nden alınan hasta veri setleri kullanılmıştır. Vanderbilt veri setinde, öncelikle orijinal veri seti incelenmiş, daha sonra eksik veriler içeren bilgiler çıkarılarak 390 hasta verisi ve korelasyonda önemi yüksek 8 öznitelik seçilmiştir. Bu yaklaşım, Lojistik Regresyon kullanılarak %88,89 doğruluk performansı sergilemiştir. Benzer şekilde, PIMA Diyabet veri setinde Lojistik Regresyon kullanıldığında %75,32 doğruluk elde edilmiştir. Veri setindeki eksik özniteliklerin uygun şekilde doldurulması veya eksik verilerin çıkarılması gibi yaklaşımlar, makine öğrenimi tekniklerindeki başarının değerlendirilmesi açısından önemlidir.

Liza vd. (2021), Vanderbilt Üniversitesi'nden alınan 390 hasta kaydı üzerinde diyabetin erken tahmini için araştırma yapmışlardır. Çalışmada, Lojistik Regresyon yöntemi kullanılarak 6 önemli öznitelik incelenmiştir. Sonuçlar, %80,00 kesinlik oranı, %75,00 duyarlılık oranı ve %91,00 doğruluk performansıyla elde edilmiştir. Bu bulgular, diyabetin erken tahmini için Lojistik Regresyon yönteminin etkili bir yöntem olabileceğini göstermektedir.

Tiwari ve Singh (2021) çalışmasında, Rastgele Orman Algoritması ve önemli öznitelikleri kullanarak diyabet hastalığının tahmini üzerine bir araştırma gerçekleştirmiştir. Çalışmada, diyabet hastalığının giderek yaygınlaşan bir sağlık sorunu olduğu ve makine öğrenimi tekniklerinin sağlık kurumları tarafından kullanılmasının önemine vurgu yapılmıştır. Elde edilen sonuçlara göre, XGBoost algoritmasının %78,91

doğruluk performansı gösterdiği ve Yapay Sinir Ağlarına (YSA) göre daha yüksek bir başarı oranına sahip olduğu belirlenmiştir.

Madhav ve Goel (2022) çalışmalarında, diyabet hastalığı tahmininde 403 hasta kaydı ve 16 öznitelik kullanarak gerçekleştirdikleri araştırmada Rassal Orman Algoritması, Lojistik Regresyon ve XGBoost yöntemleri değerlendirilmiştir. Rassal Orman Algoritması ile %91,03 doğruluk performansı, Lojistik Regresyon ile %96,15 doğruluk performansı ve XGBoost yöntemi ile %92,32 doğruluk performansı elde edilmiştir. Sonuçlar, makine öğrenmesi tekniklerinin hastalıkların tahmininde kullanımının önemli olabileceğini göstermiştir.

Gill ve Pathwar (2022) çalışmalarında, Vanderbilt Üniversitesi açık erişim kütüphanesinden elde ettikleri verilerle öznitelik seçimi yaparak Tip-2 Diyabet hastalığının tahmin edilmesini amaçlamıştır. Rassal Orman Algoritması ile 518 hasta verisi ve 6 öznitelik kullanılarak %93,95 performans doğruluğu sağlanmıştır. Ayrıca, Lojistik Regresyon yöntemi ile 491 hasta verisi ve iki öznitelik kullanılarak %89,66 doğruluk oranı elde edilmiştir.

Mijwil ve Aljanabi (2023) tarafından yapılan çalışmada, Vanderbilt Üniversitesi'nden elde edilen veriler kullanılarak makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması gerçekleştirilmiştir. Çalışmada, Lojistik Regresyon kullanılarak 403 hasta kaydı ve 15 öznitelik incelenmiş ve %95,00 kesinlik oranı, %95,90 duyarlılık oranı, %95,40 F1 puanı ve %92,50 doğruluk performansı elde edilmiştir. Lojistik Regresyon'un hastalık tahminindeki yüksek performansı ve eksik verilerin uygun yöntemlerle kullanımıyla, veri setinin tamamının değerlendirilmesi durumunda diyabet hastalığı tahmininde yüksek başarı oranı elde edilmiştir.

Viswanatha vd. (2023), çalışmalarında, PIMA Kızılderili veri seti ve Vanderbilt diyabet veri setlerini kullanmışlardır. Eksik veriler çıkartılarak elde edilen 390 hasta veri kaydı ve en önemli 8 öznitelik kullanılarak lojistik regresyon yöntemiyle %88,89 doğruluk oranı elde edilirken, PIMA Kızılderili veri seti kullanılarak bu oran %74,03 olarak

belirlenmiştir. Ayrıca, çalışmalarında farklı yöntemler kullanarak performans sonuçları üzerinde yapılan değerlendirmelerle katkı sağlamıştır.

Ceh-Varela vd. (2024) çalışmalarında, makine öğrenmesi teknikleriyle diyabet hastalığı tahmininde Mendeley, PIMA, Diyabetin Erken Evresi ve Vanderbilt veri setlerinin doğruluğunu incelemiştir. Bu çalışma, özniteliklerin tahminlemedeki önemini belirlemek amacıyla tüm özniteliklerin yanı sıra önem derecesi yüksek iki ve üç en iyi özniteliklerin kullanıldığı modellerin performans analizini ve karşılaştırmasını içermektedir. Valderbilt veri seti 390 hasta kaydına ve 15 özniteliğe sahiptir. Korelasyon analizi sonucunda, diyabet hastalığının tahmininde belirleyici olan üç öznitelik belirlenmiş ve Rassal Orman Algoritması ile %91,81 kesinlik, %92,18 duyarlılık, %91,90 F1 puanı ve %91,82 doğruluk performansı elde edilmiştir. Lojistik Regresyon Algoritması ile %89,17 kesinlik, %81,94 duyarlılık, %85,23 F1 puanı ve %86.06 doğruluk performansı, XGBoost yöntemi ile ise %89,16 kesinlik, %91,45 duyarlılık, %90,17 F1 puanı ve %90,30 doğruluk performansı elde edilmiştir. Mendeley veri seti, 1000 hasta verisi ve 12 öznitelikten oluşmaktadır. Korelasyon analizi sonucunda, önem derecesi yüksek iki öznitelik kullanılarak Rassal Orman Algoritması ile %98,83, Lojistik Regresyon ile %97,16 ve XGBoost ile %98,77 doğruluk performansı elde edilmiştir. PIMA veri seti 768 hasta veri kaydı ve 8 öznitelikten, Diyabetin Erken Evresi veri seti ise 520 hasta veri kaydı ve 16 öznitelikten oluşmaktadır. Tüm özniteliklerin kullanılması ile iki ve üç önemli özniteliklerin kullanılmasının, algoritmanın en iyi performansının belirlenmesinde öznitelik seçiminin önemi üzerine bir çalışma sunmuşlardır.

Tez çalışması kapsamında Vandelbilt Üniversitesi Biyoistatistik Departmanı'ndan alınan 403 hasta kaydı verisi, Diyabet Hastalığının makine öğrenimi teknikleriyle tahmin edilmesi amacıyla incelenmiştir. Benzer çalışmalara ilişkin literatür incelendiğinde, veri setinde eksik veriler içeren 13 hasta kaydının çıkarıldığı gözlemlenmiştir. Böylece, 390 hasta kaydıyla yapılan çalışmanın sonuçlarına odaklanılmıştır. Eksik verilerin uygun yöntemlerle doldurulmadığı durumlarda, daha düşük doğruluk performansı elde edildiği gözlemlenmiştir. Diyabet hastalığı veri setinde hasta ve hasta olmayanların sayısının eşit dağılmadığı tespit edilmiştir. Bu nedenle, yapılan tez çalışmasında veri setinin dengelenmesi amacıyla yapay veriler üretilerek veri seti çoğaltılmıştır. Bu yöntemle, veri

seti daha dengeli hale getirilmiştir. Yapay veri kullanılarak oluşturulan dengelenmiş veri setinin doğruluk performans değerlerinin, 403 hasta kaydındaki eksik veriler çıkarılarak oluşturulan 390 hasta kaydına göre daha iyi sonuçlar verdiği belirlenmiştir. Tez çalışmasında veri setinin incelenmesi, eksik verilerin uygun yöntemle doldurulması, veri ön işleme aşamaları, öznitelik seçimi ve verinin dengeli hale getirilmesi için yapay veri üretilmesi gibi adımlar izlenmiştir. Ayrıca, arayüz tasarımıyla birlikte kalp hastalığı ve diyabet hastalığının tahmin edilmesini kolaylaştırıcı sistem geliştirilerek literatüre katkı sağlanması amaçlanmıştır.

2.3. Çeşitli Hastalıkların Tahmininde Yapılan Makine Öğrenmesi Çalışmaları

Hastalıkların erken tahmini ve tedavisi, sağlık alanında öncelikli hedefler arasında yer almaktadır. Literatürde yapılan araştırmalar, makine öğrenmesi tekniklerinin kullanılarak yapılan çalışmalara bakıldığında anemi, meme kanseri, alzheimer, kronik böbrek hastalıkları, göğüs röntgen görüntülerinin incelenmesi, Covid-19 gibi daha birçok önemli sağlık sorunlarının erken tahmininde oldukça başarılı sonuçlar verdiğini göstermektedir. Bu bağlamda, yapay zeka teknolojilerinin sağlık alanında kullanımının önemi vurgulanmaktadır. Makine öğrenmesi teknikleri, büyük veri setlerini kısa sürede analiz edebilir, bu da zaman kaybının önüne geçerek hastalıkların erken tahmini ve tedavisinde sağlık alanında önemli bir ilerleme ve yardımcı sistem olabilir.

Abd vd. (2017) tarafından gerçekleştirilen çalışmada, makine öğrenimi tekniklerinin kullanılmasıyla orak hücreli anemi hastalığı incelenmiştir. Çalışmada, hastaların akıllı telefonlarına yüklenen uygulama üzerinden hastanın evde bakımını ve izlenmesini sağlayan akıllı sistem önerilmiştir. Ayrıca, araştırmada uzman sistemlerin ve akıllı cihazların hastalıkların tedavisinin izlenmesinde ve sağlık hizmetlerinin gelişimindeki etkilerinden bahsedilmiştir. Modelin sınıflandırılması için Destek Vektör Makinesi (SVM), Sıralı Minumum Optimizasyon (Sequential Minimal Optimization: SMO), Naive Bayes (NB), ve Karar Ağacı (KA) algoritmalarıyla çalışılmış ve 10 katlı çapraz doğrulama yapılmıştır. Çalışmanın sonuçları, yüksek doğruluk oranlarının elde edildiğini göstermiş olup, sağlık alanında uzman sistemlerin önemini vurgulamaktadır.

Chen vd. (2017) çalışmalarında, makine öğrenimi algoritmalarının sağlık ve tıp alanındaki büyük veri kullanımıyla birlikte, tıbbi verilerin doğru analiziyle birçok alanda önemli bir rol oynadığını vurgulamışlardır. Makine öğrenimi tekniklerinin faydaları arasında hastalıkların erken tahmini, hastane hizmetlerinin iyileştirilmesi ve tedavi yöntemlerinin optimize edilmesi bulunmaktadır. Bu çalışmada, hastalıkların sık görüldüğü topluluklarda kronik hastalık salgınlarının (enfeksiyonel) makine öğrenmesi algoritmaları kullanılarak tahmin edilmesi amaçlanmıştır. Model, 2013-2015 yılları arasında Orta Çin'den alınan gerçek hastane verilerinin değiştirilmiş tahmin modellerinin deneysel olarak test edilmesinden oluşmaktadır. Hastane kayıtlarından alınan yapısal ve yapılandırılmamış veriler kullanılarak evrişimli sinir ağı (CNN: Convolutional Neural Network) hastalık riski tahmini algoritması kullanılmıştır. Algoritmalar arasından en yüksek doğruluk performansına %94,80 oranında ulaşılmıştır.

Kohli ve Arora (2018) Wisconsin Meme Kanseri Veri Seti üzerinde çeşitli makine öğrenimi tekniklerini incelemişlerdir. Veri seti 699 hasta verisi ve 9 öznitelikten oluşmaktadır. Lojistik Regresyon ile %95,71 doğruluk oranı, Karar Ağacı ile %94,29, Rassal Orman Algoritması ile %97,14, Destek Vektör Makinesi ile %97,14 ve AdaBoost ile %98,57 oranında yüksek performans değerleri elde edilmiştir. Bu çalışma, makine öğrenimi tekniklerinin sağlık alanında hastalık tahminindeki etkinliğini belirgin bir şekilde göstermektedir.

Londha vd. (2018) tarafından yürütülen bir çalışmada, alzheimer hastalığının beyin görüntülerinin analizi için makine öğrenimi teknikleri kullanılarak hastalık tahmini yapılmıştır. Araştırmada, Destek Vektör Makinesi, Rassal Orman Algoritması, Yapay Sinir Ağları, K-En Yakın Komşu ve Gradyan Artırma algoritmaları ile çalışılmıştır. Destek Vektör Makinesi %97,56, Gradyan Artırma %97,25, K-En Yakın Komşu ile %95,00 ve Rassal Orman Algoritması ile %97,86 doğruluk oranları elde edilmiştir. En yüksek başarı, yapay sinir ağları ile %98,36 olarak elde edilmiştir. Araştırma sonuçları, alzheimer hastalığının tahmininde yüksek başarı oranlarıyla yapay zeka tekniklerinin önemini vurgulamaktadır.

Tyagi vd. (2018) çalışmalarında, tiroid hastalığının erken tahmininde makine öğrenimi tekniklerini değerlendirmiştir. Araştırmada, Destek Vektör Makinesi, K-En Yakın Komşu, Karar Ağacı ve Yapay Sinir Ağları gibi farklı makine öğrenimi yöntemleri kullanılmıştır. Çalışmanın sonuçlarına göre, Destek Vektör Makinesi yöntemi %99,63 doğruluk oranıyla en yüksek başarıyı elde etmiştir. K-En Yakın Komşu %98,62 doğruluk oranıyla dikkate değer bir performans sergilerken, Yapay Sinir Ağları %97,50 doğruluk oranına sahip olmuştur. Ancak, Karar Ağacı yöntemi %75,76 doğruluk oranıyla diğer yöntemlere kıyasla daha düşük bir performans göstermiştir. Elde edilen sonuçlar, makine öğrenimi tekniklerinin tıbbi teşhislerdeki önemini ve potansiyelini vurgulamaktadır.

Almansour vd. (2019) çalışmalarında, kronik böbrek hastalığının tahmin edilmesi için yapay sinir ağları ve destek vektör makinesi gibi makine öğrenmesi tekniklerini uygulamışlardır. Çalışmada 400 hasta kaydı ve 24 öznitelik kullanılmıştır. Veri setindeki eksik veriler, ortalaması alınarak doldurulmuştur, bu şekilde veri setindeki eksik verilerin tamamlanması sağlanmıştır. Daha sonra, tahmin modeli için yapay sinir ağları ve destek vektör makinesi teknikleri kullanılmıştır. Destek Vektör Makinesi algoritması %97,75 doğruluk elde ederken, en yüksek doğruluk performansı %99,75 ile Yapay Sinir Ağları algoritmasıyla elde edilmiştir. Sonuçlara bakıldığında, her iki algoritmanın da yüksek doğruluk oranları verdiği ancak yapay sinir ağlarının daha yüksek doğruluk performansı gösterdiği gözlemlenmiştir.

Hung vd. (2019) çalışmalarında, diş kök çürüklerinin varlığını ve yokluğunu sınıflandırmak için makine öğrenimi yöntemlerini kullanarak model performansının değerlendirilmesini amaçlamıştır. Bu kapsamda, 2015-2016 Ulusal Sağlık ve Beslenme İnceleme Araştırması'ndan elde edilen 5135 hasta verisi ve 37 öznitelik kullanılmıştır. Çalışmada Destek Vektör Makinesi, XGBoost, Rassal Orman Algoritması, K-En Yakın Komşu ve Lojistik Regresyon gibi çeşitli makine öğrenimi teknikleri değerlendirilmiştir. Yapılan değerlendirmeler sonucunda, en iyi performansı gösteren modelin Destek Vektör Makinesi olduğu belirlenmiştir. Araştırma sonuçlarına göre, kök çürüklerinin belirlenmesinde Destek Vektör Makinesi %97,1 doğruluk, %95,1 hassasiyet, %99,6 duyarlılık, %94,3 özgüllük ve AUC eğrisi %99,70 değeriyle en yüksek performansı

göstermiştir. Araştırmada geliştirilen makine öğrenimi algoritmalarının yüksek performansı, diş hekimliği alanında olduğu kadar diş hekimliği dışındaki klinik uygulamalarda da kök çürüklerine erken müdahale ve tedavi sağlanması için makine öğrenimi algoritmalarının kullanımını teşvik etmektedir.

Xiao vd. (2019) çalışmalarında, makine öğrenimi tekniklerini kullanarak kronik böbrek hastalığının takibini kolaylaştırmayı ve hastalığın tahminini gerçekleştirmeyi amaçlamışlardır. Bu kapsamda, 551 hasta kaydı ve 18 öznitelik kullanılarak Lojistik Regresyon, Elastik Net, Lasso Regresyon, Ridge Regresyon, Destek Vektör Makinesi, Rastgele Orman, XGBoost, Sinir Ağı ve K-En Yakın Komşu gibi dokuz farklı tahmin modeli kurulmuş ve karşılaştırılmıştır. Yapılan değerlendirme sonuçlara göre, Lojistik Regresyon modelinin %82,00 doğruluk performansı ve %87,00 AUC değeri ile en iyi performansı gösterdiği belirlenmiştir.

Singh vd. (2020) çalışmalarında, Kaliforniya Üniversitesi Irvine veri tabanından elde edilen verileri kullanarak karaciğer hastalığını tahmin etme üzerine çalışma gerçekleştirmişlerdir. Çalışma kapsamında, sınıflandırma ve özellik seçimi tekniklerini kullanarak Akıllı Karaciğer Hastalığı Tahmin Yazılımı modeli geliştirmişlerdir. Akıllı Karaciğer Hastalığı Tahmin Yazılımı veri seti üzerinde Lojistik Regresyon, Sıralı Minimal Optimizasyon, Rassal Orman Algoritması, Naive Bayes, J48 ve K-En Yakın Komşu makine öğrenmesi teknikleri kullanılarak karaciğer hastalığı tahmini yapılmıştır. Veri setinin performansları iki aşamada ölçülmüştür. İlk aşamada veri setinde öznitelik seçimi yapılmadan algoritmaların doğruluk performansları ölçülmüştür. İkinci aşamada Greedy Stepwise yöntemiyle öznitelik seçimi yapılarak karaciğer hastalığının tahmininde önemli öznitelikler seçilmiş ve algoritmaların doğruluk performans değerleri tekrardan ölçülmüştür. Algoritmaların doğruluk performans değerleri 10 kat çapraz doğrulama kullanılarak değerlendirilmiştir. İlk aşamada öznitelik seçimi yapılmadan Rassal Orman Algoritması kullanılarak %72,53 doğruluk performansı, Lojistik Regresyon kullanılarak %72,50, Sıralı Minimal Optimizasyon kullanılarak %71,35, J48 kullanılarak %68,78, K-En Yakın Komşu ile %64,15 ve Naive Bayes %55,74 performans değerleri elde edilirken; ikinci aşamada öznitelik seçimi yapılarak Lojistik Regresyon kullanılarak %73,36, Rassal Orman Algoritması %71,87, Sıralı Minimal Optimizasyon

%71.36, J48 %70,67, K-En Yakın Komşu %67,41 ve Naive Bayes %55,90 oranında doğruluk belirlemişlerdir. Çalışma sonuçlarına göre, en iyi sonuç Lojistik Regresyon yöntemiyle öznitelik seçimi kullanılarak elde edilmiştir ve model üzerinde öznitelik seçimi uygulandıktan sonra işlem süresinin azaldığı gözlemlenmiştir.

Kukar vd. (2021) çalışmalarında, rutin kan testleri kullanarak COVID-19 teşhisi için makine öğrenimi modelleri geliştirmişlerdir. Toplamda 5493 hasta kaydı kullanılmıştır, bunların içinde 5333 hasta çeşitli bakteriyel ve viral enfeksiyonlar geçirirken, 160 hasta COVID-19 pozitif olarak saptanmıştır. Çalışmada toplam 35 öznitelik kullanılmıştır. Model oluşturulurken XGBoost yöntemi tercih edilmiş ve çapraz doğrulama yapılmıştır. XGBoost yöntemiyle belirlenen bir eşik değeri kullanılarak ROC eğrisi üzerinde belirli bir noktada %81,90 duyarlılık ve %97,90 özgüllük elde edilmiştir. Çapraz doğrulama sonucunda ise AUC değeri %97,00 olarak belirlenmiştir. Araştırma, COVID-19 tanısında makine öğrenimi teknikleri kullanılarak son derece başarılı sonuçlar elde edilebileceğini ve bu yöntemin kullanılabilir olduğunu göstermiştir.

Saha vd. (2021) tarafından yürütülen çalışmada, göğüs röntgen görüntülerinin incelenmesi yoluyla COVID-19 teşhisinde Evrişimsel Sinir Ağı algoritmaları ve makine öğrenimi tekniklerinin kullanıldığı görülmektedir. Çalışmada, Rassal Orman, Destek Vektör Makinesi, Karar Ağacı ve AdaBoost gibi çeşitli makine öğrenimi teknikleri değerlendirilmiştir. Çalışma sonuçlarına göre, Evrişimsel Sinir Ağı algoritması %96,52 doğruluk elde ederken, en yüksek doğruluk performansı %96,96 ile Destek Vektör Makinesi algoritmasıyla elde edilmiştir.

Olatunji vd. (2021) çalışmalarında, Suudi Arabistan'dan elde edilen bir veri kümesi kullanarak tiroid kanserinin erken aşamada tespiti için bir makine öğrenimi tabanlı sistem geliştirmişlerdir. Python programlama dili kullanılmış ve Rassal Orman, Yapay Sinir Ağları, Destek Vektör Makinesi ve Naive Bayes algoritmaları uygulanmıştır. Araştırmada, Rassal Orman algoritması tarafından elde edilen %90,91'lik doğruluk oranı en yüksek performansı sergilemiştir. Diğer algoritmalarla elde edilen doğruluk oranları sırasıyla Destek Vektör Makinesi %84,09, Yapay Sinir Ağları %88,64 ve Naive Bayes %81,82 olarak belirlenmiştir.

Ryzhikova vd. (2021) çalışmalarında, alzheimer hastalığının tahmininde beyin omurilik sıvısının yakın kızılötesi Raman spektroskopisi verilerinin makine öğrenimi teknikleriyle analizi gerçekleştirilmiştir. Çalışmada, 21 hasta ile 16 sağlıklı kontrol grubundan elde edilen veri seti kullanılmış ve yeni bir teknik geliştirilmiştir. Araştırmada %84 doğrulukla bir performans elde edilerek alzheimer hastalığının sınıflandırılması için modellerin başarısının önemi vurgulanmıştır.

Hassan vd. (2022) çalışmalarında, prostat kanseri tespiti üzerine ultrason ve manyetik rezonans görüntüleme verilerinden derin öğrenme tabanlı yapay zeka tekniklerini kullanarak bir çalışma gerçekleştirmişlerdir. Araştırmada, prostat kanserinin erken teşhisi ile sağkalım oranlarının önemli ölçüde artabileceği vurgulanmıştır. Araştırma sonuçlarına göre, en iyi model, test setinde ultrason görüntülerinde %97,00 ve manyetik rezonans görüntülerinde %80,00 maksimum doğruluk elde etmiştir. Modelin performansını artırmak için öznitelik çıkarımı yapılmış ve bu sayede modelin performansında önemli ölçüde artış sağlanmıştır. Yapay zeka tekniklerinin verimliliğinin arttığı ve bu nedenle hastanelerde kullanılabileceği belirtilmiştir.

Literatürdeki araştırmalar, farklı hastalıkların tahmininde çeşitli makine öğrenme tekniklerini kullanmaktadır ve algoritmaların modele uyum sağlayarak hastalık tahmininde yüksek doğruluk elde edildiğini göstermektedir. Bununla birlikte, sağlık çalışanlarına yardımcı olabilecek akıllı sistem tasarımlarının hastalık tahmini, tedavisi ve sürecin hızlandırılması gibi avantajlar sağladığı gözlemlenmiştir. Bu çalışmalar, sağlık sektöründe makine öğrenme tekniklerinin önemli avantajları olduğunu vurgulamaktadır.

3. MATERYAL ve YÖNTEM

Günümüzde yaygınlaşan hastalıkların erken tahmini, sağlık alanında büyük önem taşımaktadır. Yapay zeka, büyük veri setlerini analiz edebilme yeteneği, çeşitli sektörlerdeki kullanımı ve yüksek tahmin başarısıyla dikkat çekmektedir, bu da yapay zekanın geniş kapsamda kullanılmasını sağlamaktadır.

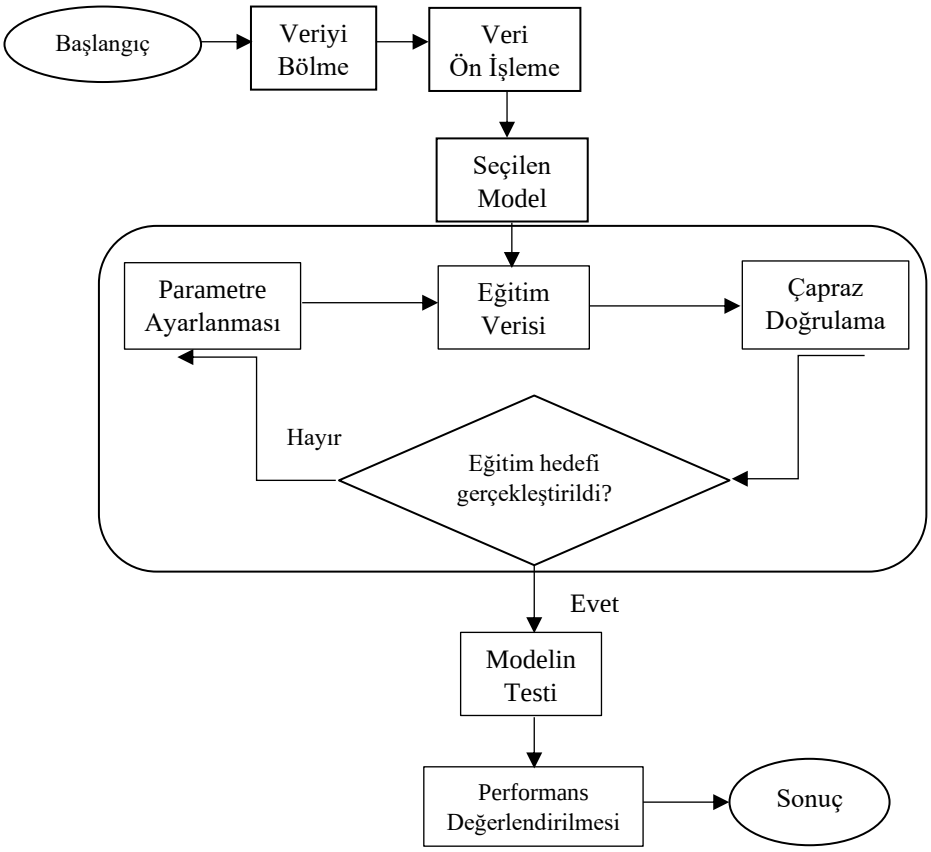
Yapay zekanın alt dalı olan makine öğrenmesi teknikleri, sağlık alanında büyük veri setlerinin analiz edilmesinde, hastalıkların erken teşhisinde ve tedavi süreçlerinin iyileştirilmesinde yardımcı olabilir. Yapılan tez çalışmasında, hastalıkların erken tahmininde farklı makine öğrenmesi teknikleri kullanılmıştır. Çalışmanın devamında en iyi performans gösteren algoritmalar ve en iyi hiperparametre değerleri kullanılarak arayüz tasarımı gerçekleştirilmiştir. Bu bölümde, kullanılan teknikler detaylı bir şekilde incelenmiş ve açıklanmıştır. Kullanılan tekniklerin nasıl uygulandığı, hangi veri setleriyle çalıştığı ve elde edilen sonuçlar üzerine detaylı bilgiler sunulmuştur.

3.1. Makine Öğrenmesi

Makine Öğrenmesi tarihte ilk defa bilgisayar alanında ve yapay zeka alanında uzman Arthur Samuel tarafından 1959 yılında ortaya atılmıştır (Samuel, 1959). Bilgisayarların ve teknolojinin gelişmesiyle birlikte farklı öğrenme türleri ortaya çıkmıştır. Makine öğrenimi, bilgisayar sistemlerinin belirli görevleri yerine getirmeleri için kullanılan algoritmaların ve istatistiksel modellerin bilimsel çalışmasıdır. Yapay Zekanın bir alt dalı olan Makine Öğrenmesi Algoritmaları uygulanan diğer yaklaşımlar arasından en iyi performans veren yöntemlerden biridir. Makine öğrenimi, verilerin daha iyi öğrenmesini sağlamaktadır ve veriden çıkacak sonuçların yorumlanmasına yardımcı olmaktadır. Farklı makine öğrenmesi algoritmalarından, çözülmek istenen problemin türüne ve değişken sayısına göre en uygun algoritma kullanılmaktadır (Mahesh, 2020).

Veri tabanlarına erişimin kolaylaşmasıyla birlikte sağlık alanında yapay zekanın hastalık teşhisindeki önemi ve kullanımı artmıştır. Kolay kullanımı ve çoğu veri tabanına

uygulanabilirliği nedeniyle makine öğrenmesi en çok tercih edilen yapay zeka yöntemlerinin başında yer almaktadır (Wuest vd., 2016). Şekil 3.1'de makine öğrenmesi akış diyagramı gösterilmiştir.



Şekil 3.1. Makine öğrenmesi akış diyagramı (Shams vd., 2024).

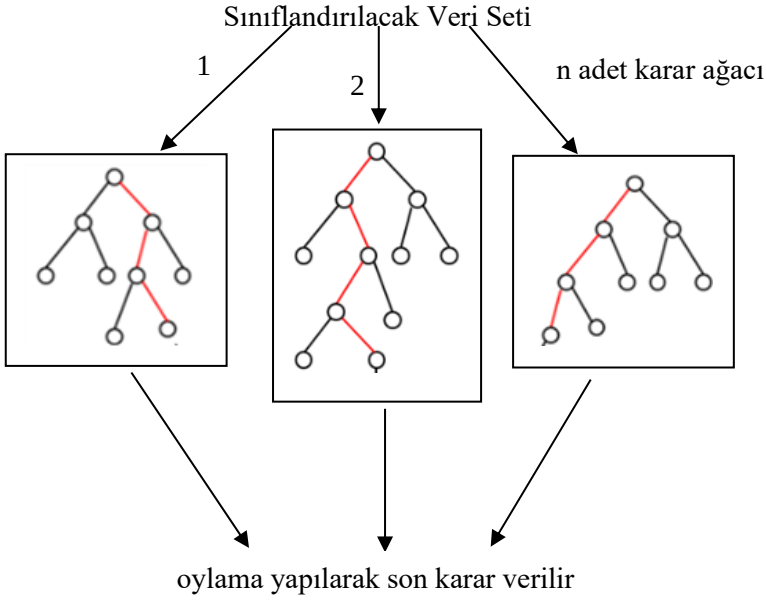
Makine öğrenmesi algoritmaları, sağlık, üretim, eğitim, perakende, bankacılık ve finansman sektörü gibi birçok farklı alanda kullanılmaktadır ve başarılı sonuçlar vermektedir. Özellikle sağlık alanında, makine öğreniminin kullanımı birçok avantaj sağlamaktadır. Makine öğrenimi, sağlık sektöründe veri analitiği, tıbbi görüntüleme analizi, hastalık tedavisi ve tedavi planlaması, büyük miktarda sağlık verisinin analiz edilebilmesi, hastalıkların erken tahmini, tedavisi, hastaların bekleme sürelerinin kısaltılması ve maliyetlerin azaltılması gibi alanlarda kullanılmaktadır. Makine öğrenmesi sağlık sektöründe önemli bir rol oynamasının yanı sıra, sağlık hizmetlerinin iyileştirilmesine de önemli katkıda bulunmaktadır (Shailaja, 2018).

3.2. Rassal Orman Algoritması

Rassal Orman Algoritması, 2001 yılında Leo Breiman tarafından geliştirilen bir makine öğrenmesi tekniğidir. Rassal Orman Algoritması, sınıflandırma ve regresyon problemlerinde kullanılabildiği için farklı problem tiplerine uyum sağlayabilme esnekliğine sahiptir. Bu özelliği, veri bilimi ve yapay zeka alanlarında çeşitli uygulamalarda yaygın olarak kullanılmasını sağlamaktadır (Breiman, 2001).

Rassal Orman Algoritması, veri setini rassal olarak parçalara ayırarak her bir parça için bir karar ağacı oluşturmaktadır. Bu algoritma, her bir rassal olarak seçilen veri örneği için birden fazla karar ağacı oluşturarak çalışmaktadır. Rassal Orman'ın çalışma yapısı, farklı karar ağaçlarının bir araya gelmesiyle oluşturulduğu için birçok algoritmadan daha güçlü bir model elde edilmesini ve yüksek tahmin performansı sağlamaktadır. (Ayyadevara, 2018).

Rassal Orman Algoritması, her bir ağacın rastgele alt örneklemeler üzerinde eğitilmesi sayesinde modelin ezberlemesinin önüne geçmektedir. Rassal Orman çalışma mantığı, her ağacın belirli bir alt kümeyle çalışmasını sağlamaktadır ve her ağaç birbirinden bağımsız olarak büyümektedir. Standart karar ağaçlarında her düğüm, tüm değişkenler arasından en iyi bölünmeyi seçerken, rassal ormanlarda her düğüm rastgele seçilen bir değişken alt kümesinden en iyi bölünmeyi seçmektedir. Rassal Orman modelin ezber ve aşırı uyumuna karşı dirençli olmaktadır. Bu nedenle, birçok tekniğe göre daha yüksek performans göstermektedir (Liaw ve Wiener, 2002). Şekil 3.2’de rassal orman algoritmasının karar ağacı yapısı gösterilmiştir.



Şekil 3.2. Rassal orman algoritması karar ağaç yapısı (Okumus ve Nuroglu, 2021)

Rassal Orman Algoritması, diğer geleneksel istatistiksel sınıflandırma yöntemlerine göre çeşitli avantajlara sahiptir. Rassal Orman, birçok sınıflandırma ağacını birleştirerek çalıştığı için daha güvenilir sonuçlar göstermektedir. Bu algoritmanın öne çıkan avantajları arasında yüksek sınıflandırma doğruluğu, karmaşık problemleri kolayca anlayıp modelleyebilme yeteneği, çeşitli problem türlerine uyum sağlama kabiliyeti ve eksik verilerin uygun değerlerle doldurulması yer almaktadır. Ayrıca Rassal Orman, her bir ağaç için rastgele seçilmiş değişkenleri kullanarak birçok örnek verisi oluşturmakta ve modelin aşırı uyumunun önüne geçmektedir (Cutler vd. 2007).

Rassal Orman Algoritması, eğitim verilerinin model tarafından ezberlenmesini engelleyerek aşırı uyumu önlemektedir. Rassal Orman, tahminlemede yüksek başarı oranı gösteren ve en çok tercih edilen makine öğrenmesi tekniklerindendir (Akşehir ve Kılıç, 2019).

- **Rassal orman algoritması hiperparametreleri**

Yapay zeka tekniklerinde, hiperparametrelerin doğru ayarlanması modelin performans başarısını önemli ölçüde etkilemektedir. Problemin yapısına göre hiperparametrelerin ayarlanması oldukça önemlidir. Hiperparametrelerin doğru ayarlanmasıyla yüksek

doğruluk elde edilebilirken, yanlış ayarlanmasıyla daha düşük doğruluk oranları elde edilir.

Hiperparametrelerin doğru ayarlanması, modelin eğitim ve tahmin sürelerini kısaltmaktadır. Bu aşama modelin daha hızlı eğitilmesini ve daha hızlı tahmin yapmasını sağlar.

Aşağıda Rassal Orman Algoritması hiperparametreleri açıklanmıştır:

- max_depth: Karar ağacının maksimum derinliğini belirlemektedir. Karar ağacının derinliği arttırıkça, modelin karmaşıklığı artar ve eğitim verisine daha fazla uyum sağlayabilir. Bu durum modelin aşırı öğrenme (overfitting) riskini artırır. max_depth hiperparametresi, karar ağacının derinliğini sınırlayarak modelin daha iyi genelleme yapmasını sağlamaktadır (Müller ve Guido, 2016; Owen, 2022).
- max_features: Karar ağacındaki her bir ağacın bölünürken maksimum özellik sayısını belirlemektedir. max_features parametresi, auto, sqrt, sayı ya da none değerler almaktadır. Hangi parametrenin kullanılacağı, veri setinin özelliklerine ve modelin performansına bağlı olarak en uygun olanı seçilmelidir.
  - auto: Genellikle, veri kümesindeki tüm özniteliklerin kullanılmasını sağlamaktadır.
  - sqrt: Öznitelik sayısının karekökünü alarak maksimum öznitelik sayısını belirlemektedir.
  - none: Öznitelik seçimi yapılmaz ve her bölünme işleminde tüm öznitelikleri kullanmaktadır.
- max_leaf_nodes: Karar ağacının derinliği ve aşırı öğrenmesinin engellenmesini kontrol etmektedir. Bu parametre, karar ağacının en fazla sahip olabileceği yaprak düğüm sayısını belirlemektedir. Karar ağacı, belirlenen bu maksimum yaprak düğüm sayısına ulaştığında, daha fazla dallanma gerçekleştirilmez ve ağaç büyümesi durdurulur. Böylece, aşırı karmaşıklığın önüne geçilerek modelin daha iyi genelleme yapması sağlanır.

- max_samples: Veri setindeki verilerin karar ağacına ne kadarının verileceğini belirlemektedir (Owen, 2022).
- min_impurity_decrease: Karar ağacındaki her bir düğümün ne zaman bölünmesi gerektiğini belirlemektedir (Owen, 2022).
- min_samples_leaf: Karar ağacındaki her bir düğümdeki minimum örnek sayısını belirlemektedir. Böylece, karar ağacının büyüklüğünü kontrol etmektedir ve modelin genelleme yeteneğini artırmaya yardımcı olmaktadır (Hastie vd., 2009; Owen, 2022).
- min_samples_split: Karar ağacındaki her bir düğümdeki bölmek için gereken minimum örnek sayısını belirlemektedir. Böylece, karar ağacının dallanma süreci kontrol edilmektedir ve modelin genelleme yeteneğini artırmaya yardımcı olmaktadır (Owen, 2022).
- Bootstrap: Karar ağacının eğitim verilerini oluştururken rasgele örneklem seçimi ile oluşturup oluşturulmayacağına karar verilmesini sağlamaktadır (Müller ve Guido, 2016; Owen, 2022).
- True: Karar ağacının eğitim verilerini rasgele örneklemi kullanarak seçer ve seçtiği örneklerle modelin eğitimini gerçekleştirmektedir.
- False: Karar ağacı tüm eğitim verilerini kullanarak eğitilmektedir. Böylece karar ağaçları arasında örneklem seçimi yapılmaz ve her karar ağacı tüm veri seti üzerinde eğitilmektedir.
- ccp_alpa: Modelin aşırı öğrenmesini önlemek için karar ağacındaki dalların budanmasını kontrol ederek karmaşıklığı önlemektedir (Owen, 2022).
- class_weight: Sınıf dengesizliklerinde modelin performansını arttırmak amacıyla kullanılmaktadır (Müller ve Guido, 2016; Owen, 2022).
- criterion: Karar ağaçlarındaki bölünmenin kalitesini sağlamak için, model en iyi bölünmeyi sağlamaktadır. Model genellikle Gini impurity veya entropy gibi kriterlerden birini seçmektedir.
- n_jobs: Çalıştırılacak işlerin sayısını ve işlemci çekirdeğini göstermektedir. Modelin işlem süresini kısaltarak eğitilmesinin hızlandırılmasını sağlamaktadır.
- oob_score: Modelin performansının tahmin edilmesini sağlamaktadır.

Rassal orman algoritmasında kullanılan diğer parametreler aşağıda açıklanmıştır:

- random_state: Random state parametresi, makine öğrenimi tekniklerinde rassallığı kontrol etmek için sıkça kullanılmaktadır. Modelde random state değeri belirlendiğinde, algoritma her çalıştırıldığında aynı rastgelelik seviyesinde işlem yaparak rastgeleliği kontrol etmektedir. Bu özellik, aynı veri seti üzerinde yapılan tekrarlı çalışmalarda sonuçların tutarlı olmasını sağlamaktadır. Random state parametresinin belirlenmesi modelin yapısal özelliklerine ve kullanılan algoritmaya göre doğru ayarlanması, modelin performansını ve en iyi doğruluk performansını elde etmek için önemlidir (Müller ve Guido, 2016; Owen, 2022).
- n_estimators: Modelin karmaşıklığını belirleyen ve karar ağaçlarının sayısını gösteren parametredir. Bu parametre, modelin aşırı uyum (overfitting) eğilimini kontrol eder ve modelin yüksek performans göstermesine yardımcı olur. Modele uygun bir n_estimators değerinin seçilmesiyle, modelin karmaşıklığı ayarlanabilir ve aşırı uyumu azaltılabilir.

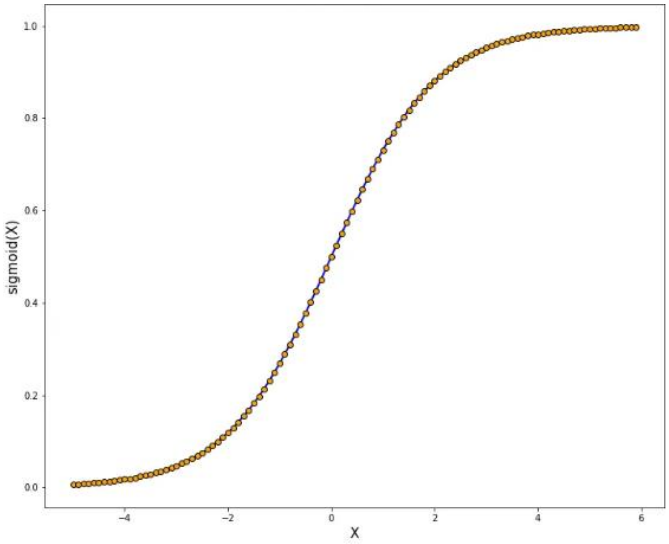
Rassal Orman Algoritmasının performansını arttırmak amacıyla ve en uygun hiperparametre değerlerinin bulunması için Izgara Arama Yöntemi kullanılmıştır.

3.3. Lojistik Regresyon Algoritması

Lojistik regresyon, David Cox'un 1958'de yayımladığı makalede, bağımlı değişkeni iki farklı sonuçtan oluşan sınıflandırma problemleri için nasıl kullanılabileceğini anlatmıştır. Bu çalışma, lojistik regresyonun gelişimine önemli katkılarda bulunmuş ve bu yöntemin kullanımının yaygınlaşmasını sağlamıştır (Cox, 1958)

Lojistik regresyon, makine öğrenimi algoritmalarına göre veri modellerine daha kolay uyarlanabilir ve analiz edilebilir olması, hızlı eğitim ve tahmin yeteneği ile büyük veri modellerinde sıklıkla tercih edilmektedir (Westreich vd., 2010).

Lojistik Regresyon, özellikle ikili ve kategorik sınıflandırma yapmak için sıkça kullanılan makine öğrenmesi yöntemidir. Bu yöntem, değişkenler arasındaki ilişkileri anlamak için özellikle ikili sınıflandırma problemlerinde kullanılan bir analiz yöntemidir. Bu analiz yöntemi, bağımsız değişkenlerin arasındaki ilişkiyi değerlendirerek bağımlı değişkenin kategoriye ait olma olasılığını tahmin etmekte kullanılır. Lojistik regresyon, sosyal bilimlerden pazarlamaya, finanstan tıbba kadar birçok alanda yaygın olarak kullanılan bir analiz yöntemidir. Özellikle, hastalık risk faktörlerinin belirlenmesi, tahmini ve tedavi sürecinde lojistik regresyon sıklıkla tercih edilmektedir. Sonuç olarak, lojistik regresyon, ikili sınıflandırma problemlerinde değişkenler arasındaki ilişkileri anlamak ve bağımlı değişkenlerin kategoriye ait olma olasılığını tahmin etmek için güçlü istatistiksel yöntemdir. Bu yöntem, bağımlı değişkenin kategorik olduğu durumlarda kullanılabilir ve çeşitli problem türlerinde değerlidir. Bağımsız değişkenlerin (girdi parametreleri) bağımlı değişkenler (sonuç değeri) üzerindeki etkilerinin belirlenmesi, analiz edilmesi ve modelin yorumlanmasını kolaylaştırmaktadır (Hosmer vd., 2013). Şekil 3.3'de lojistik regresyon algoritmasının olasılık dağılımı gösterilmektedir.



Şekil 3.3. Lojistik regresyon algoritması olasılık dağılımı (Thorn, 2020)

Lojistik regresyon, bağımlı değişken ile bağımsız değişkenler arasındaki doğrusal olmayan ilişkiyi sigmoid (lojistik) fonksiyonuyla göstermektedir. Bu fonksiyon, bağımsız değişkenlerin bir sınıfa ait olma olasılığını hesaplamak için kullanılmaktadır.

Lojistik regresyon genellikle bağımlı değişkenin kategorik olduğu durumlarda sıklıkla tercih edilmektedir. Ayrıca lojistik regresyon kategorik verilerde de kullanılmaktadır. Lojistik regresyon doğrusal regresyona benzemekte ve sonuçları tahmin etmek için Sigmoid fonksiyonu kullanılmaktadır. Sigmoid fonksiyonu, sonuçları 0 ile 1 arasında olasılığa dönüştürmektedir (Yıldız, 2014)

Doğrusal regresyon, genellikle sayısal değerleri analiz etmek için kullanılırken, lojistik regresyon sınıflandırma problemlerinde ve sınıflar arasındaki ilişkiyi tahmin etmek için kullanılmaktadır. Doğrusal regresyon, bağımlı değişkenlerin sürekli olduğu durumlarda tercih edilirken, lojistik regresyon genellikle bağımlı değişkenin kategorik olduğu durumlarda tercih edilmektedir (James vd., 2013) Lojistik regresyonun matematiksel formülü, Denklem 3.1'de gösterilmiştir (Stoltzfus, 2011).

$$\hat{y}_i = \frac{e^{\beta_0+\beta_2\cdot X_1+\beta_2\cdot X_2+\cdots+\beta_i\cdot X_i}}{1+e^{\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\cdots+\beta_i\cdot X_i}} \tag{3.1}$$

$\hat{y}_i$	Bağımlı değişkenin 1 olma olasılığını ve çıktı değerini ifade etmektedir.
e	Euler sayısını ifade etmektedir (e ≈ 2.71828).
β_0	Regresyonun gerçekleştiği kesme noktasıdır.
$\beta_1, \beta_2,\ldots,\beta_n$	Modelin katsayılarıdır. β katsayıları eğimi göstermektedir.
$X_1,X_2,\ldots,X_i$	Bağımsız değişkenlerdir.

$\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_i \cdot X_i$ Bağımsız değişken β katsayısına göre ağırlıklandırılır ve eğim artışı kadar bağımsız değişkende de artmaktadır.

- **Lojistik regresyon algoritması hiperparametreleri**

Hiperparametreler, makine öğrenimi modellerinin davranışlarını kontrol etmektedir ve veri setinin ihtiyacına göre en iyi performans değerlerine ulaşmasına yardımcı olmaktadır. Aşağıda, lojistik regresyon algoritmasının hiperparametreleri açıklanmıştır:

- Ceza (Penalty): Modelin karmaşıklığını kontrol ederek aşırı uyumu (overfitting) önlemektedir ve modelin daha iyi genelleme yapmasını sağlamaktadır. Ceza (Penalty), L1 ve L2 olmak üzere iki türe ayrılmaktadır (Hastie vd., 2009; Turgut vd., 2022).
- L1 (Lasso) Ceza: L1 cezası katsayılarını sıfıra yakınlaştırarak bazı öznitelikleri önemsiz hale getirmektedir. Modeldeki gereksiz özniteliklerin belirlenmesi ve kaldırılması için kullanılmaktadır.
- L2 (Ridge) Ceza: L2 cezası, ağırlık katsayılarının karelerinin toplamı kadar ceza uygulamaktadır. Katsayıları büyüklüklerine göre cezalandırmaktadır ve büyük katsayıların etkisini azaltmaktadır.
- fit_intercept: Modelin tahminlerindeki sapmayı belirleyen bir hiperparametredir. Bu parametre, modelin sabitini hesaba katıp katmayacağını kontrol etmektedir (Turgut vd., 2022).
- Çoklu Sınıf (Multi-class): Modelin bir veri örneğinin hangi sınıfa ait olduğunu belirleme yeteneğini ifade etmektedir. Veri örneğini çoklu sınıflara veya ikili sınıflara ayırma yeteneğini belirleyen bir kavramdır (Murphy, 2012).
- Dual: Çözüm algoritmasının belirlenmesini sağlamaktadır. Dual hiperparametresi, True ve False olmak üzere iki farklı değer almaktadır. Genellikle küçük veri setleri için False, büyük veri setleri için True seçilmektedir. Yüksek boyutlu büyük veri setleri için True olarak ayarlandığında, küçük veri setleri için False ayarlandığında çözüm algoritması daha etkili çalışmaktadır. Modelin yapısına, veri setine ve problem türüne göre uygun algoritma seçilmektedir (Murphy, 2012).
- C değeri: Aşırı uyumu kontrol etmek için kritik bir hiperparametredir. Büyük bir C değeri seçildiğinde, model aşırı uyuma daha yatkın hale gelir. Küçük bir C değeri ise modelin aşırı uyumu azaltır. Küçük seçilen C değeri modelin başarısını azaltırken, büyük seçilen C değeri modelin yüksek başarı göstermesine neden olabilir. C değerinin modele göre iyi ayarlanması tahminleme başarısında önemli olmaktadır (Müller ve Guido, 2016).
- Sınıf Ağırlığı (class_weight): Veri setlerindeki sınıflar arasında dengesizlik olduğunda, modelin performansında olumsuz sonuçlar gözlemlenebilmektedir.

Bu hiperparametre, sınıf dağılımlarının eşitlenmesini ve dengesizliklerin düzeltilmesini sağlamaktadır (Géron, 2022 ; Owen, 2022).

- Tolerans (Tol): Optimizasyon algoritmalarının durdurma kriterini belirlemektedir. Genelde çok küçük sayı seçilmektedir (Müller ve Guido, 2016 ; Géron, 2022).
- verbose: Eğitim sürecinin ne kadar ayrıntılı görüneceğini belirleyen parametredir. Eğitim sırasında her adımın ayrıntılarının ekran üzerinde ayrıntılı olarak gösterilip gösterilmeyeceğini ayarlamaktadır (Géron, 2022).
- warm_start: Modelin daha önceden kullanılan eğitim verisinden elde edilen çözümü kullanıp kullanılmayacağının belirlenmesini sağlar. Bu şekilde, model baştan eğitilmeden önceki çözümler kullanılarak eğitim süreci hızlandırılmaktadır (Hastie vd., 2009 ; Géron, 2022).
- Çözücü (Solver): Modelin eğitim sürecinde kullanılan bir optimizasyon algoritmasıdır. Bu algoritma, modelin en uygun parametrelerini bulmak ve veriye en uygun şekilde uymasını sağlamak için kullanılır (Müller ve Guido, 2016).
- intercept_scaling: Modelin sabit teriminin ölçeğini belirlemektedir. Genellikle lojistik regresyon gibi doğrusal modellerde kullanılmaktadır (Sormunen, 2022).
- L1_Oranı (L1_Ratio): L1 (Lasso) ve L2 (Ridge) düzenlemesi arasındaki karışım oranını belirlemektedir (Friedman vd., 2010).

3.4. Aşırı Gradyan Artırma Algoritması

Aşırı Gradyan Artırma (XGBoost: Extreme Gradient Boosting) algoritması, Tianqi Chen tarafından geliştirilen ve yaygın olarak kullanılan bir makine öğrenimi tekniğidir. Bu algoritma, Rassal Orman Algoritmasına benzer bir şekilde, karar ağaçları üzerine kurulmuş modeldir ve sınıflandırma ile regresyon modellerinde kullanılmaktadır. XGBoost'un temel prensibi, zayıf karar ağaçlarının bir araya getirilerek yüksek performanslı bir tahmin modeli elde edilmesidir. Her bir karar ağacı, önceki ağaçların tahmin hatalarını düzeltmek amacıyla eklenmektedir. Bu şekilde, artan sayıda karar ağacı modeli, genel tahmin performansını artırmaktadır. XGBoost, birçok farklı problem ve veri seti için etkili bir şekilde kullanılabilen güçlü bir algoritmadır (Owen, 2002).

XGBoost, aşırı gradyan artırma tekniğini kullanan ve karar ağaçlarına dayanan bir makine öğrenmesi algoritmasıdır. XGBoost, regresyon, sınıflandırma ve sıralama gibi çeşitli problemlerle ve farklı veri tiplerine uyum sağlamak amacıyla tasarlanmıştır. Farklı veri setlerinde etkili bir şekilde kullanılabilen bu algoritma, yüksek hız ve gelişmiş özelliklere sahip olduğundan geniş kullanım alanına sahiptir (Chen, 2015).

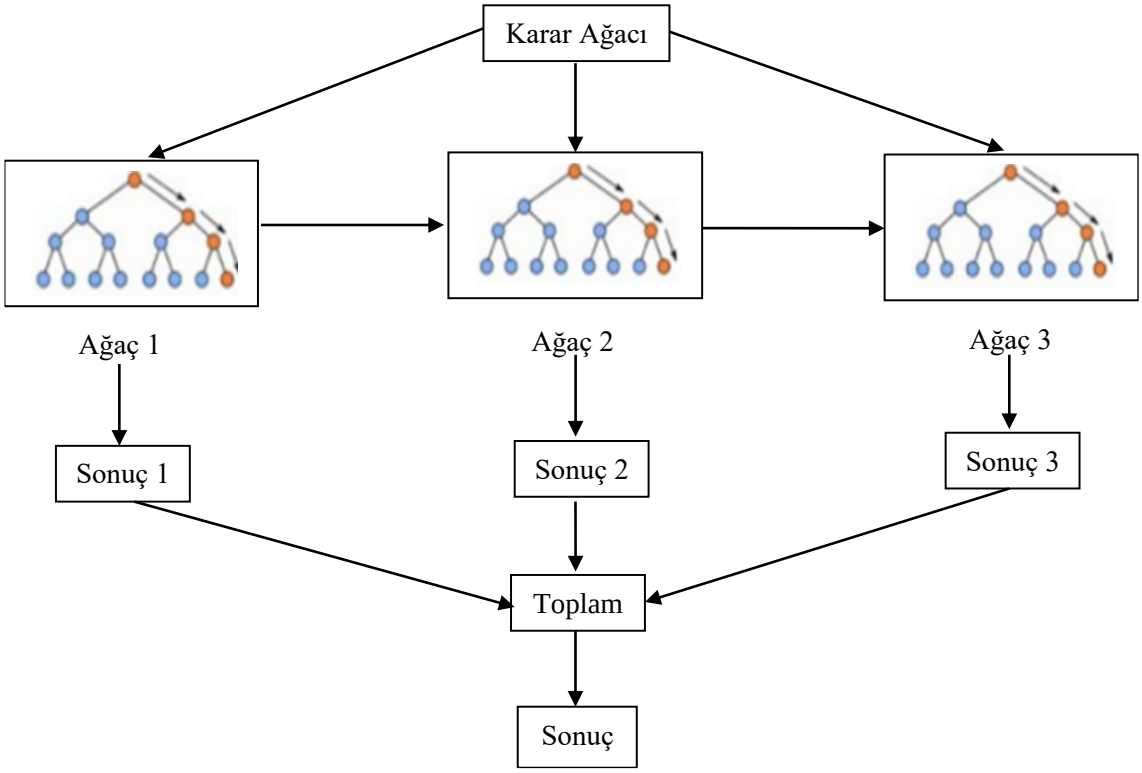
XGBoost, son derece başarılı performans gösteren ve ağaç artırma yöntemine dayanan bir makine öğrenimi tekniğidir. XGBoost'un teknolojik gelişmelerle birlikte sıkça kullanılmasının nedenlerinden biri, daha az kaynak kullanarak milyarlarca veriyi ölçekleyebilme yeteneğidir. Bu özelliği sayesinde XGBoost, model öğrenimini hızlandırarak oldukça yüksek performans göstermektedir. XGBoost'un diğer yapay zeka tekniklerine göre 10 kat daha hızlı çalışması da önemli bir avantajıdır. Ayrıca, eksik verilerle ve büyük veri kümeleriyle çalışabilme yeteneği, XGBoost'un yaygın bir şekilde kullanılmasını sağlamaktadır (Chen ve Guestrin, 2016).

XGBoost, Gradyan artırma (gradient boosting) temelli modelin hesaplama performansını artırmak amacıyla geliştirilmiştir. Bu algoritma, önceki modellerin hatalarını tahmin ederek yeni modeller oluşturmaktadır ve son tahmini yapmaktadır. Regresyon ve sınıflandırma problemleri gibi çeşitli alanlarda kullanılabilir ve özellikle karmaşık veri setlerinde yüksek performans göstermektedir (Ogunleye ve Wang, 2019).

XGBoost, genellikle hem regresyon hem de sınıflandırma tahmin modellemesi için kullanılan bir makine öğrenimi tekniğidir. XGBoost, çeşitli veri setlerinde uygulandığında birçok makine öğrenimi tekniğine kıyasla daha iyi performans göstermiştir (Friedman, 2001).

Deneysel olarak, XGBoost birçok makine öğrenme algoritmasından daha hızlı çalışmaktadır. XGBoost'un yüksek doğruluk performansıyla birlikte oldukça hızlı olması ve farklı veri setlerine uyum sağlaması, geniş çapta tanınmasını ve tercih edilen bir algoritma haline gelmesini sağlamıştır (Nielsen, 2016).

XGBoost, karar ağaçlarına dayanan ve gradyan artırma algoritmasıyla çalışan bir makine öğrenme tekniğidir. Bu algoritma, model performansını ve hesaplama hızını optimize etmek için özel olarak geliştirilmiştir. Zayıf modellerin birleştirilerek güçlü bir model oluşturulmasını sağlar ve aşırı uyumu azaltır. Özellikle, işlem hızını artırır ve bellek kullanımını optimize eder. Bu sayede modelin eğitim süresi azaltılır ve karmaşık veriler daha etkili şekilde analiz edebilir. XGBoost'un diğer özellikleri, derinlik öncelikli ağaç budama, paralel ağaç oluşturma gibi tekniklerin yanı sıra, eksik verilerin etkin şekilde işlenmesini sağlamaktadır ve yüksek doğruluk performansı göstermektedir. Ayrıca, XGBoost çapraz doğrulamada başarılı sonuçlar göstermektedir. Büyük veri kümelerine uyarlanabilirliği ve başarılı doğruluk performansları sayesinde geniş kullanım alanına sahiptir (Pathak ve Shrestha, 2020). Şekil 3.4'de XGBoost algoritmasının karar ağaç yapısı gösterilmiştir.



Şekil 3.4. XGBoost algoritması karar ağacı yapısı (Wang vd., 2020)

Makine öğrenmesi teknikleri arasında Rassal Orman ve XGBoost algoritmaları oldukça yaygın olarak kullanılan algoritmalardır. Rassal Orman ve XGBoost, her ikisi de karar

ağacı temelli algoritmalardır. Rassal Orman ve XGBoost, karar ağacı temelli makine öğrenimi algoritmaları olmalarının yanı sıra, farklı veri setleri ve problem alanları için genellikle etkili sonuçlar sağlayabilirler. İki teknik arasındaki çalışma farkları şöyledir: Rassal Orman Algoritması, eğitim sırasında birçok karar ağacı oluşturur ve her bir ağaçtan elde edilen tahminlerin ortalamasını alarak sonuç üretir. Her bir alt-ağaç modeli, eğitim verilerinden rastgele örnekler seçerek oluşturulur ve bu alt modeller bağımsız olarak çalışır. Bu nedenle, Rassal Orman her ağacı farklı alt veri kümeleriyle eğitir. XGBoost ise her bir ağaç modelini, önceki ağaç modelinden kalan hatayı azaltmak için bir sonraki ağaç modeli ile besler. Bu şekilde, her ağaç modeli, önceki ağaçlardan kalan hatayı minimize etmektedir. XGBoost'un gradyan artırma yöntemi kullanması, her bir ağaç modelinin önceki ağaçların hatalarını azaltmaya odaklanmasını sağlar ve genellikle daha iyi tahminler elde etmeye yardımcı olmaktadır (Wang vd., 2020).

- **XGBoost algoritması hiperparametreleri**

Hiperparametreler, modelin davranışını ve performansını kontrol etmektedir. Farklı problem tipleri ve modeller için uygun hiperparametrelerin seçimi, modelin başarısını belirleyen en önemli aşamalardandır.

XGBoost Algoritması Hiperparametreleri aşağıda açıklanmıştır:

- Öğrenme Oranı (learning_rate): Öğrenme oranı, makine öğrenimi modelinde her iterasyonda ağaç eklenirken önceki adımlardaki tahminlere verilen ağırlığı belirleyen bir hiperparametredir. Düşük bir öğrenme oranı seçildiğinde, daha fazla ağaç ve iterasyon gerekebilir, bu da modelin optimum sonuçlara ulaşma olasılığını artırabilir. Ancak, daha yüksek bir öğrenme oranı seçmek, eğitim sürecini hızlandırabilir; ancak bu durum modelin aşırı uyum riskini artırabilir. Bu nedenle, modelin yüksek performans elde edebilmesi için uygun bir öğrenme oranının seçilmesi kritik öneme sahiptir ve bu, modelin başarımını önemli ölçüde etkilemektedir (Müller ve Guido, 2016; Abualdenien ve Borrmann, 2022; Owen, 2022).

- Örnekleme Oranı (subsample): XGBoost'da her bir ağaç oluşturulurken kullanılacak eğitim örneklerinin oranını belirlemektedir. Bu hiperparametre her bir ağaç için alt örneklerin oranını göstermektedir. Örnekleme oranı, aşırı uyumu önlemek ve modelin performansını optimize etmek için önemlidir (Abualdenien ve Borrmann, 2022; Owen, 2022).
- Ağaç için öznitelik örnekleme oranı (colsample_bytree): Modelde her ağaç oluşturulurken kullanılacak özniteliklerin oranını ve hangi özniteliklerin kullanılacağını belirlemektedir. Bu hiperparametre, her ağaç oluşturulduğunda, belirlenen colsample_bytree oranı kadar özellik rastgele seçmektedir (Budholiya vd., 2022; Owen, 2022).
- Gamma: Modelde ağaçların büyüme hızını ve karmaşıklığını kontrol etmektedir. Yaprak düğümünde kayıp azaltılmasını önler ve aşırı uyumu kontrol etmeye yardımcı olmaktadır (Abualdenien ve Borrmann, 2022; Owen, 2022).
- lambda: Modelin karmaşıklığını kontrol eder ve aşırı uyumu önlemektedir (Owen, 2022; Pan, 2022).
- min_child_weight**:** Modelde ağaçlar oluşturulurken yapının bölünmesi için gereken minimum örnek ağırlık toplamını göstermektedir. Ağaçların büyüklüğünü kontrol etmekte ve aşırı uyumu kontrol etmektedir (Abualdenien ve Borrmann, 2022; Budholiya vd., 2022; Owen, 2022).

3.5. Izgara Arama Algoritması

Izgara Arama Algoritması (Grid Search), ismini hiperparametre uzayındaki tüm değerlerin test edilmesi ve bu işlemin ızgara oluşturmasından almaktadır. Bu algoritma, hiperparametre uzayında belirli bir aralıkta bulunan her bir parametrenin belirli bir değer kümesini inceleyerek, en iyi performansı sağlayan hiperparametre kombinasyonunu bulmayı amaçlamaktadır. Izgara oluşturması, bu işlemi bir ızgara gibi düzenli bir şekilde yaptığını ifade etmektedir. Izgara Arama algoritma yapısı hiperparametrelerin olası değerlerini bulmak için öncelikle liste oluşturur ve oluşturduğu bu listeleri iç içe geçmiş for döngüleri kullanarak tüm kombinasyonları deneyerek en iyi hiperparametre değerlerini bulmaya çalışır. Izgara Arama Algoritması avantajları arasında kullanımının

ve algoritmalara uygulanmasının oldukça kolay olması yer almaktadır. Ancak Izgara Arama Algoritması, tüm olası kombinasyonları test etmesinden dolayı algoritmanın sonuçlanması zaman almaktadır (Owen, 2022).

Izgara Arama Algoritması, farklı yapay zeka tekniklerinde modelin en uygun hiperparametrelerini belirlemek için yaygın olarak kullanılan tekniktir. Bu teknikte, her bir hiperparametre için istenen performansı sağlayacak değer kümesi belirlenmektedir ve bu değerlerin kombinasyonlarıyla ızgara oluşturmaktadır. Izgara Arama Algoritmasıyla belirlenen en iyi hiperparametreler seçildikten sonra, model bu parametrelerle eğitilir ve sonuçlar değerlendirilir. Izgara Arama Algoritması, modelin genelleme yeteneğinin değerlendirilmesine ve en uygun doğruluk performansına ulaşılmasına yardımcı olmaktadır. (Müller ve Guido, 2016).

Çalışmada, Kalp Hastalığı ve Diyabet Hastalığı veri setleri üzerinde Rassal Orman, Lojistik Regresyon ve XGBoost algoritmaları kullanılarak hastalıkların erken tahmini yapılmıştır. Modellerin değerlendirilmesi ve en uygun hiperparametrelerin belirlenmesi için Izgara Arama Algoritması kullanılmıştır.

- Rassal Orman Algoritması için Izgara Arama Yöntemi

Rassal Orman algoritmasıyla Kalp ve Diyabet hastalıklarının tahmini için önemli hiperparametrelerin en iyi performans değerlerini belirlemek amacıyla Izgara Arama Algoritması kullanılmıştır. Rassal Orman algoritmasının bazı parametreleri olan bootstrap, oob_score, max_depth, min_samples_leaf için en uygun değerlerin bulunmasında Izgara Arama yöntemi kullanılmıştır.

- Lojistik Regresyon Algoritması için Izgara Arama Yöntemi

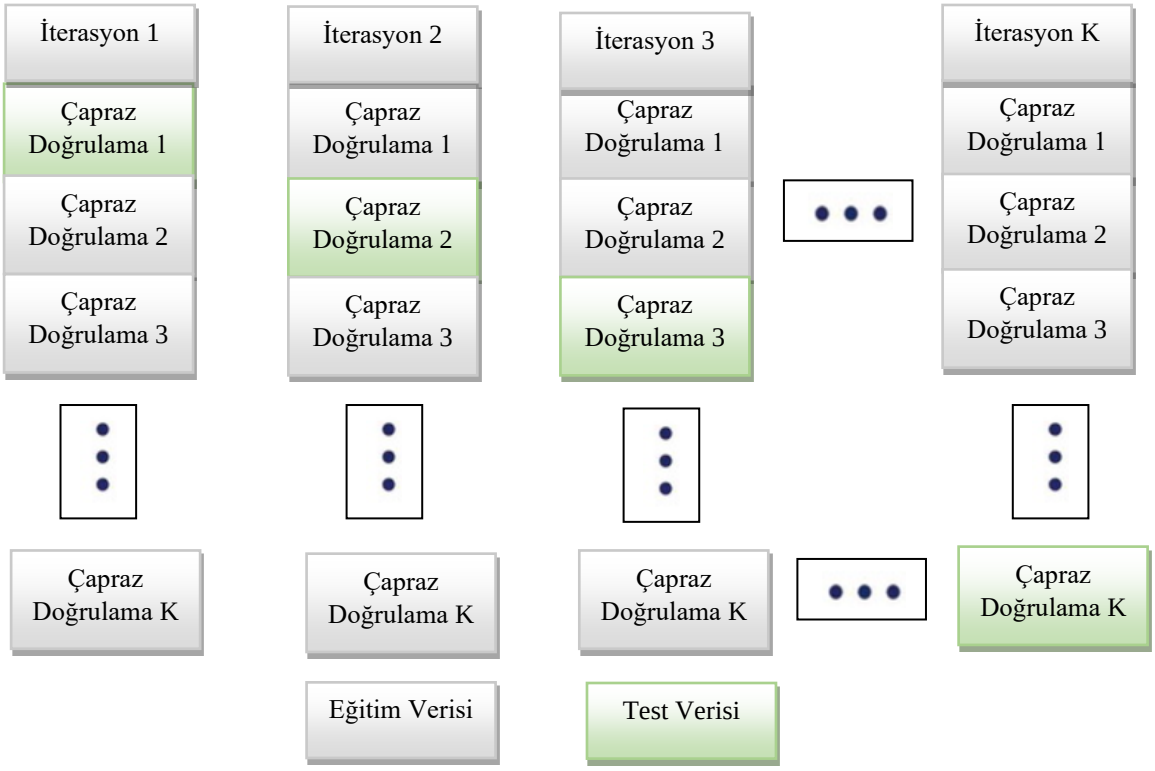
Lojistik regresyon algoritmasının bazı parametreleri penalty, C, solver, multi_class en uygun değerleri ızgara arama yöntemiyle bulunmuştur.

- XGBoost Algoritması için Izgara Arama Yöntemi

XGBoost algoritmasıyla Kalp ve Diyabet hastalıklarının tahmini için önemli hiperparametrelerin en iyi performans değerlerini belirlemek için Izgara Arama algoritması kullanılmıştır. XGBoost algoritmasının bazı parametreleri olan learning_rate, max_depth, min_child_weight, random_state için en uygun değerlere ulaşılmasında Izgara Arama yöntemi kullanılmıştır.

3.6. Çapraz Doğrulama Yöntemi

Çapraz doğrulama (cross-validation), modelin performansını değerlendirmek ve aşırı uyumu (overfitting) önlemek için kullanılan bir veri örnekleme yöntemidir. Bu teknik, veri setini birden çok parçaya ayırarak modelin performansını farklı veri kümeleri üzerinde test etmektedir. Böylece modelin yeni veya daha önce görülmemiş verilere nasıl uyum sağladığını göstermektedir. Çapraz doğrulama, modelin performansını değerlendirmede yaygın olarak kullanılan bir tekniktir (Berrar, 2019). Çapraz doğrulama yöntemi, Şekil 3.5'te gösterilmiştir.



Şekil 3.5. Çapraz doğrulama yöntemi (Ren vd., 2019)

Modelin en iyi hiperparametre ayarını belirlemede k-katlamalı çapraz doğrulama tekniği, veri setinin k eşit parçaya bölünmesini sağlamaktadır. Her bir parça hem eğitim hem de test verisinde kullanılmaktadır. Böylece modelin görmediği verileri kullanarak aşırı uyumu önlemektedir. Modelin başarısını değerlendirmede kritik öneme sahip olan bu yöntem, model performansı ve algoritma seçiminde oldukça önemlidir.

Çapraz doğrulama, makine öğrenimi tekniklerinde model performansını değerlendirmek için kullanılan bir yöntemdir. Veri seti k eşit alt kümeye bölünerek, her bir alt küme test verisi olarak seçilirken, kalan k -1 alt küme eğitim verisi olarak kullanılmaktadır. Çapraz doğrulama, seçilen her alt küme üzerinde aynı teknikle modelin tahmin performansını ölçmektedir. Bu yaklaşım regresyon ve sınıflandırma problemlerinde yaygın olarak kullanılmaktadır. Çapraz doğrulama, modelin performansını her bir test verisi üzerinde inceleyerek büyük veri setleri üzerinde daha güvenilir sonuçlar sunmaktadır ve modelin genelleme yeteneğini artırmaktadır (Ren vd., 2019; Raschka, 2018).

Tez çalışmasında, kalp hastalığı ve diyabet hastalığının tahmininde makine öğrenmesi tekniklerinden Rassal Orman Algoritması, Lojistik Regresyon ve XGBoost algoritmaları kullanılmıştır. Modellerin performansını değerlendirmek için iki veri setine çapraz doğrulama tekniği uygulanmıştır. Çapraz doğrulama kullanılarak modellerin genelleme yeteneği ve doğruluğu incelenmiştir.

4. BULGULAR ve TARTIŞMA

Çalışmada, sağlık alanında hastalıkların erken tahmininde yapay zeka tekniklerinin kullanılması incelenmiş ve yapay zeka teknikleri kullanılarak kalp hastalığı ve diyabet hastalıkları erken tahmini çalışması yapılmıştır. Kalp hastalığı tahmini için UCI Machine Learning Repository kütüphanesinden 297 veri ve 13 öznitelikten oluşan Clevaland Kalp Hastalığı veri seti kullanılmıştır. Diyabet Hastalığının tahmini için Vanderbilt Üniversitesi Biyoistatistik Departmanından alınan veri seti 403 kişiye ait olup 15 özellikten oluşmaktadır. Kalp hastalığı ve Diyabet hastalığı veri setlerinin incelenmesi ve değerlendirilmesinde yapay zeka tekniklerinden Rassal Orman Algoritması, Lojistik Regresyon ve XGBoost kullanılmıştır. Diyabet Hastalığı veri setinde 403 hasta kaydından modelin performansını test etmek amacıyla 7 hasta olmayan ve 2 hasta olan veriler ayrılmıştır. Ayrıca Boy, Ağırlık, Sistolik Kan Basıncı, Diastolik Kan Basıncı, Bel, Kalça, Kolestrol, HDL Kolestrol, Vücut Kitle İndeksi ve Bel/Kalça Oranı özniteliklerinde eksik veriler bulunmuştur. Eksik verilerde ortalama yöntemi kullanılarak eksikler tamamlanmıştır. En uygun hiperparametrelerin belirlenmesi için Izgara Arama Algoritması kullanılmıştır. Yapay zeka tekniklerinin performansı, Çapraz Doğrulama yöntemi kullanılarak değerlendirilmiştir. Çalışma, kalp hastalığı ve diyabet hastalığı tahmini için yapay zeka tekniklerinin etkin bir şekilde kullanılabilirliğini ortaya koymaktadır. Çalışmanın devamında Kalp hastalığı ve Diyabet hastalığının erken tahmininde makine öğrenmesi tekniklerinden en iyi performans gösteren algoritmalar kullanılarak arayüz geliştirilmiştir.

4.1. Kalp Hastalığı ve Diyabet Hastalığı Veri Seti Temel Bilgileri

Çalışmada kalp hastalığı ve diyabet hastalığı erken tahmininde kullanılan özniteliklerin özellikleri, veri tipleri ve açıklamalarıyla birlikte, kalp hastalığı ve diyabet hastalığına ait temel veri seti bilgileri yer almaktadır.

Nümerik ve kategorik veriler, veri analizlerinde kullanılan istatistiksel kavramlardır. Nümerik veriler, ölçülebilen ve matematiksel işlemlere uygun olan sayısal değerleri ifade

eder. Tamsayılar, ondalıklı sayılar ve kesirli sayılar nümerik verilere örnektir. Kategorik veriler ise, verilerin belirli kategoriler veya sınıflara ayrıldığı veri türleridir (Bruce vd., 2020).

- **Kalp hastalığı veri seti temel bilgileri**

Makine öğrenmesi teknikleri kullanılarak gerçekleştirilen kalp hastalığı tahmini çalışmasında, UCI (Machine Learning Repository) sitesinden erişilen 297 veriden oluşan veri seti ve 13 öznitelik kullanılmıştır. Çizelge 4.1'de Kalp hastalığı veri setindeki temel bilgiler gösterilmiştir.

Çizelge 4.1. Kalp hastalığı veri seti öznitelikleri bilgisi

Özellikler	Açıklama	Veri Tipi
Yaş	Hastanın Yaşı	Nümerik
İstirahat Kan Basıncı (Trestbps)	Dinlenme Durumundaki Kan Basıncı (mmHg)	Nümerik
Chol	Serum Kolestrol Düzeyi (mg/dl)	Nümerik
Thalach	Ulaşılan Maksimum Kalp Hızı	Nümerik
Oldpeak	Egzersizle Tektiklenen ST Segment Depresyonu	Nümerik
Ca	Floroskopide Boyanan Büyük Damar Sayısı (0-3)	Kategorik
Cinsiyet	Hastanın Cinsiyeti 1: Erkek , 0: Kadın	Kategorik
Cp	Göğüs Ağrısı Tipleri 1: Tipik Anjina 2: Atipik Anjina 3: Anjin Olmayan Ağrı 4: Asemptomatik	Kategorik
Fbs	Açlık Kan Şekeri >120 mg/dl 1: Var , 0:Yok	Kategorik
Restecg	İstirahat Kan Sonuçları 0: Normal 1: ST-T Dalga Anormallikleri 2: Sol Vertikül Hipertofisi	Kategorik
Exang	Egzersizle Tetiklenen Anjina 1:Var , 0:Yok	Kategorik
Slope	Pik Egzersiz ST Segmenti 1: Yukarı Eğimli 2: Düz Eğimli 3: Aşağı Eğimli	Kategorik
Thal	Thalyum Stres Testi 3: Normal 6: Fix Defekt 7: Reversible Defekt	Kategorik
Tahmin	Hastalık Tahmini 1:Hasta , 0: Hasta Değil	Kategorik

- **Diyabet hastalığı veri seti temel bilgileri**

Vanderbilt Üniversitesi Biyoistatistik Departmanı açık erişim sitesinden alınan diyabet hastalığı veri seti 403 kişiye ait hasta kaydı ve 15 öznitelikten oluşmaktadır. Vanderbilt diyabet hastalığı veri setindeki temel bilgileri Çizelge 4.2'de gösterilmiştir.

Çizelge 4.2. Diyabet hastalığı veri seti öznitelikleri bilgisi

Özellikler	Açıklama	Veri Tipi
Yaş	19-92	Nümerik
Boy	132-193 cm	Nümerik
Ağırlık	45-147 kg	Nümerik
Vücut Kitle İndeksi (BMI)	15,2-55,8 kg/m^2	Nümerik
Sistolik Kan Basıncı	92-250 mmHg	Nümerik
Diastolik Kan Basıncı	48-124 mmHg	Nümerik
Bel çevresi	66-142 cm	Nümerik
Bel/Kalça Oranı	0,68-1,14	Nümerik
Kalça çevresi	76-162 cm	Nümerik
Kolestrol	78-443 mg/dL	Nümerik
HDL Kolestrol	12-120 mg/dL	Nümerik
Kolestrol/HDL Kolestrol Oranı	1,5-19,3	Nümerik
Glukoz	48-385 mg/dL	Nümerik
Tahmin	Hastalık Tahmini 1:Hasta , 0: Hasta Değil	Kategorik

Veri setinde hasta olan ve hasta olmayan kişilerin sayısı eşit dağılmamaktadır. Hasta olmayan kişilerin sayısı hasta olan kişilerin sayısından daha fazladır ve veri seti dengesizdir. Veri setini dengeli hale getirmek için 147 yapay veri üretilmiştir. Böylece yapay veriler üretilerek veri seti daha dengeli hale getirilmiştir. Diyabet hastalığı veri seti 541 hasta kaydı ve 15 öznitelik içermektedir.

4.1.1. Veri setlerinin incelenmesi ve dengesiz veri setinin dengeli hale getirilmesi

Çalışmada kullanılan kalp hastalığı ve diyabet hastalığı veri setinin makine öğrenimi teknikleriyle hastalık tahminine başlamadan önce, veri setinin dengeli olup olmadığı incelenmiştir. Kalp hastalığı veri setinin dengeli olduğu, ancak diyabet hastalığı veri setinin dengesiz olduğu gözlemlenmiştir. Diyabet hastalığı veri setinde yapay veriler oluşturularak dengesiz veri seti daha dengeli hale getirilmiştir.

- **Kalp hastalığı veri seti**

Cleveland Kalp Hastalığı veri setinde 297 kişiye ait hasta kaydı incelendiğinde %53.87'sinin kalp hastası olmadığı, %46.13'ünün ise kalp hastası olduğu belirlenmiştir. Kalp hastalığı veri setinde kalp hastası olan ve kalp hastası olmayan kişi sayısının yaklaşık dengeli olduğu anlaşılmıştır.

- **Diyabet hastalığı veri seti**

Vanderbilt Üniversitesi Biyoistatistik Departmanı'nın açık erişim sitesinden alınan veri seti, 403 kişiye ait hasta kaydı ve 14 özellikten oluşmaktadır. Diyabet hastalığı veri setinde, doğruluk performansın test edilmesi için 7 hasta olmayan ve 2 hasta olan toplam 9 kişiye ait hasta kaydı verileri ayrılmıştır. Geriye kalan 394 kişiye ait veri seti incelendiğinde %85.28'inin diyabet hastası olmadığı, %14.72'sinin diyabet hastası olduğu belirlenmiştir.

Diyabet hastalığı veri seti incelendiğinde hasta olmayanların sayısı hasta olanların sayısına göre oldukça fazladır, ve bu durum veri setinin dengesiz olduğunu göstermektedir. Dengesiz veri setini daha dengeli hale getirmek için çeşitli yöntemler bulunmaktadır. Veri sayısı az olan sınıfın veri sayısı arttırılarak ve veri setini daha dengeli hale getirerek modelin bir sınıfı diğerine göre daha fazla ağırlıkta öğrenmesinin önüne geçilmiştir ve performans tutarlılığı arttırılmıştır.

Dengesiz veri setleri, bazı sınıfların diğerlerine göre çok daha fazla örneğe sahip olduğu durumlarda ortaya çıkmaktadır ve makine öğrenimi modellerinin performansını olumsuz etkilemektedir. Dengesiz veri setlerini dengeli hale getirmek için iki ana yaklaşım vardır (He ve Ma, 2013).

Aşırı Örnekleme (Oversampling): Veri setinde azınlıkta kalan sınıfların örnek sayısını artırarak veri setini dengelemek için kullanılan yöntemdir. Bu yöntemde, mevcut verileri

çoğaltarak veya yapay veri oluşturarak veri seti dengeli hale getirilmektedir (He ve Ma, 2013).

Düşük Örnekleme (Undersampling): Veri setinde dengesizliği azaltmak için çoğunluk sınıfındaki veri örneklerinin sayısını düşürmeyi amaçlayan yöntemdir. Bu yöntemde, örnek sayısının azaltılması, çoğunluk sınıfından rastgele örnekler çıkarmak veya veri setini küçültmek gibi yollarla yapılmaktadır (He ve Ma, 2013).

Her iki yöntem de dengesizliği azaltmak için kullanılmaktadır ve makine öğrenmesi modellerinin daha tutarlı ve doğru sonuçlar elde edilmesine yardımcı olmaktadır. Seçilecek yöntem, veri setinin büyüklüğüne, sınıf dengesizliği derecesine ve uygulama alanına göre değişmektedir (He ve Ma, 2013).

Diyabet hastalığı veri setinde toplam 403 hasta kaydı bulunmaktadır. Modelin performansını değerlendirmek amacıyla 7 sağlıklı ve 2 hasta olmak üzere toplam 9 kayıt ayrılmıştır. Kalan 394 kayıtta, hasta olmayan kişilerin sayısı hasta olan kişilerin sayısından daha fazladır. Bu dengesizlik nedeniyle makine öğrenmesi teknikleriyle veri seti incelendiğinde, modelin doğruluk oranı yüksek olmasına rağmen diğer performans metriklerinde bazı uyumsuzluklar gözlenmiştir. Model, sağlıklı verileri hasta verilere göre daha doğru tahmin etme eğilimindedir.

Modeldeki bu dengesizliği dengelemek amacıyla düşük örnekleme (undersampling) yöntemi kullanılmış, veri setinden rastgele seçilen sağlıklı kayıtlar kullanılarak 147 yapay veri üretilmiştir. Böylece veri seti toplamda 541 hasta kaydı haline getirilmiş ve model daha dengeli hale getirilmiştir. Makine öğrenmesi kullanılarak, modelin doğruluk oranı ve diğer performans metrikleri incelendiğinde, daha tutarlı ve dengeli sonuçlar elde edilmiştir. Yapay verilerin kullanımı ile modelin performansı artırılmıştır.

Birçok hastalığa ilişkin veri setlerinde genellikle, hasta ve hasta olmayan kayıtlar dengeli bir dağılıma sahip değildir. Bu çalışmada, veri setinin dengelenmesi, hastalık tahminlerinde daha tutarlı ve güvenilir sonuçlar elde edilmesine katkı sağlamıştır.

4.1.2. Veri setlerinde istatistiksel bilgiler

Bu bölümde kalp hastalığı ve diyabet hastalığına ait temel istatistiksel bilgiler yer almaktadır.

Count (Sayım): Veri noktalarının toplam sayısını ifade etmektedir. Bu terim, bir veri setinde bulunan toplam veri sayısını göstermektedir (Ross, 2017).

Mean (Ortalama): Veri setindeki değerlerin toplamının veri noktalarının sayısına bölünmesiyle veri setindeki ortalama değer bulunmaktadır (Cherry, 2010).

Std (Standart Sapma): Standart sapma, veri setindeki değerlerin ortalama değerden ne kadar uzaklaştığını göstermektedir. Düşük standart sapma, veri noktalarının ortalama değere yakın olduğunu ve veri setinin daha düzenli olduğunu gösterirken; yüksek standart sapma, veri noktalarının ortalama değerden daha uzak olduğunu ve veri setinin daha dağılmış olduğunu göstermektedir (Wan vd., 2014). Denklem 4.1.'de standart sapmanın matematiksel tanımı gösterilmiştir (Wachs, 2009)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{4.1}$$

σ Standart sapma değeri

x_i Veri setindeki tüm değerler

μ Veri setinin ortalama değeri

N Veri setindeki toplam veri sayısı

- **Kalp hastalığı veri seti için istatistiksel bilgiler**

Kalp hastalığı veri setini daha iyi anlamak amacıyla kalp hastalığı ile ilgili istatistiksel veriler incelenmiştir. Veri setinde yer alan yaş, cinsiyet, cp, trestbps, chol fbs, restecg, thalach, exang, oldpeak, slope, ca ve thal öznitelikleri Python programlama dili kullanılarak detaylı olarak analiz edilmiştir. Çizelge 4.3'de Kalp hastalığı veri setinin istatistiksel temel bilgileri verilmiştir.

Çizelge 4.3. Kalp hastalığı veri seti için istatistiksel bilgiler

	yaş	cinsiyet	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Tahmin
Sayım (count)	297	297	297	297	297	297	297	297	297	297	297	297	297	297
Ortalama (mean)	55	0,7	2,2	132	247	0,1	1	150	0,3	8,8	0,6	0,7	0,8	0,5
Standart Sapma (std)	9,1	0,5	1	18	52	0,4	1	23	0,5	12	0,6	0,9	1	0,5
Minimum Değer (min)	29	0	0	94	126	0	0	71	0	0	0	0	0	0
İlk Çeyrek 25%	48	0	2	120	211	0	0	133	0	0	0	0	0	0
Medyan 50%	56	1	2	130	243	0	1	153	0	4	1	0	0	0
Üçüncü Çeyrek 75%	61	1	3	140	276	0	2	166	1	14	1	1	2	1
Maksimum (max)	77	1	3	200	564	1	2	202	1	62	2	3	2	1

- **Diyabet hastalığı veri seti için istatistiksel bilgiler**

Diyabet hastalığının tahmini için istatistiksel veriler büyük önem taşımaktadır. Diyabet hastalığı veri seti özniteliklerinin kısaltmaları ve kısaltmalarının açıklaması Çizelge 4.4'de gösterilmiştir.

Çizelge 4.4. Diyabet hastalığı veri setindeki özniteliklerin kısaltmalarının açıklanması

Kısaltmalar	Açıklaması
A1	Cinsiyet
A2	Yaş
A3	Boy
A4	Ağırlık
A5	Vücut Kitle İndeksi
A6	Sistolik Kan Basıncı
A7	Diastolik Kan Basıncı
A8	Bel çevresi
A9	Kalça çevresi
A10	Bel / Kalça Oranı
A11	Kolestrol
A12	HDL Kolestrol
A13	Kolestrol / HDL Oranı
A14	Glukoz

Diyabet hastalığı veri setini daha iyi anlamak için Python programlama dili kullanılarak veri setinde yer alan Cinsiyet, Yaş, Boy, Ağırlık, Vücut Kitle İndeksi, Sistolik Kan Basıncı, Diastolik Kan Basıncı, Bel, Kalça, Bel/Kalça Oranı, Kolestrol, HDL Kolestrol, Kolestrol/HDL Oranı ve Glukoz öznitelikleri detaylı olarak analiz edilmiştir. Çizelge 4.5'de Diyabet hastalığı veri setinin istatistiksel temel bilgileri verilmiştir.

Çizelge 4.5. Diyabet hastalığı veri seti için istatistiksel bilgiler

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	Tahmin
Sayım (count)	541	541	541	541	541	541	541	541	541	541	541	541	541	541	541
Ortalama (mean)	0,6	50	168	82	29	140	84	98	111	0,9	212	49	4,8	131	0,4
Standart Sapma (std)	0,5	16	9,8	18	6,6	23	13	15	14	0,1	46	18	2	73	0,5
Minimum Değer (min)	0	19	132	45	15	90	48	66	76	0,7	78	12	1,5	48	0
İlk Çeyrek 25%	0	37	160	71	25	124	74	86	102	0,8	181	37	3,3	83	0
Medyan 50%	1	50	168	81	28	138	82	97	109	0,9	207	45	4,5	97	0
Üçüncü Çeyrek 75%	1	62	175	91	33	150	92	107	119	0,9	235	58	5,7	171	1
Maksimum (max)	1	92	193	147	56	200	124	142	142	1,1	443	120	19	385	1

Veri seti üzerinde yapılan istatistiksel incelemeler, verinin temel özelliklerini ve eğilimlerini göstermektedir.

4.1.3. Korelasyon matrisinin incelenmesi

Korelasyon, iki veya daha fazla değişkenler arasındaki ilişkiyi incelemek için makine öğrenmesi tekniklerinde sıklıkla tercih edilen istatistiksel yöntemdir. Korelasyon katsayısı, -1 ile +1 arasında değer almaktadır. Korelasyon katsayısı -1 'e yaklaştıkça negatif korelasyon yani incelenen özniteliğin önem derecesinin düşük olduğunu, +1'e yaklaştıkça da pozitif korelasyon yani önem derecesinin arttığını, 0 olduğunda ise değişkenler arasında bir ilişki olmadığını göstermektedir (Hastie, 2009).

Çalışmada korelasyon analizi kullanılarak değişkenler arasındaki ilişki incelenmiş ve modele en uygun özniteliklerin seçilmesinde kullanılmıştır. Bu analiz, önem derecesi düşük olan ve modele katkısı az olan gereksiz özniteliklerin belirlenmesinde yardımcı olmuştur. Böylece, kalp hastalığı ve diyabet hastalığının tahmininde önem derecesi yüksek ve düşük olan öznitelikler incelenerek, modelin performansını artırmak amacıyla diğer değişkenlere göre önem derecesi nispeten daha düşük olan öznitelikler çıkartılmıştır.

- **Kalp hastalığı veri seti korelasyon matrisi**

Korelasyon matrisi kullanılarak öznitelikler ve hastalık tahminine etkisi olan faktörler analiz edilmiştir. Korelasyon matrisine bakılarak kalp hastalığı ile en yüksek korelasyon ilişkisine sahip öznitelik Ca (Floroskopide Boyanan Büyük Damar Sayısı) olarak belirlenmiştir. Kalp hastalığı ile Ca özniteliği arasındaki korelasyon katsayısı 0.46 olarak bulunmuştur. İncelemeler sonucunda Ca değerinin kalp hastalık tahmininde diğer özniteliklere göre daha fazla etkili olduğu sonucuna ulaşılmıştır. Elde edilen korelasyon değeri kalp hastalığı tahmin modellerinde Ca'nın önemli olduğunu göstermektedir.

Çalışmanın devamında, kalp hastalığı tahmininde daha az etkili olduğu belirlenen öznitelikler korelasyon matrisi analizi sonucunda veri setinden çıkarılmıştır. Böylece, veri setinin gereksiz bilgilerden arındırılması ve modelin tahmin performansının artırılması hedeflenmiştir.

Korelasyon katsayısına bakıldığında önem derecesi 0,08 olarak belirlenen Fbs (Açlık Kan Şekeri) özniteliğinin veri setindeki diğer özniteliklere göre kalp hastalığı tahmininde etkisinin daha düşük olduğu anlamı çıkartılabilir. Korelasyon matrisi incelenerek öznitelikler ve hastalık tahminine etki eden öznitelikler incelenmiştir. Çizelge 4.6’da görüldüğü üzere değişkenler arasındaki ilişkileri incelemek için korelasyon matrisi kullanılmıştır.

Çizelge 4.6. Kalp hastalığı veri seti değişkenler arasındaki korelasyon matrisi

	yaş	Cinsiyet	Trestbps	chol	fbs	cp	resteeg	thalach	Exang	Oldpeak	slope	Ca	thal	Tahmin
yaş	1	-0,1	0,1	0,3	0,2	0,1	0,1	-0,4	0,1	0,2	0,2	0,4	0,1	0,2
cinsiyet	-0,1	1	0,01	-0,07	-0,2	0,04	0,03	-0,1	0,1	0,1	0,03	0,1	0,4	0,3
trestbps	0,1	0	1	0	0,1	-0,1	0,1	-0,3	0,4	0,1	0,2	0,2	0,3	0,41
chol	0,3	-0,1	-0,4	1	0,1	0,2	0,1	-0,05	0,01	0,2	0,1	0,1	0,1	0,15
fbs	0,2	-0.2	0,1	0,1	1	0	0,2	0	0,1	0,02	-0.01	0,12	0,02	0,08
cp	0,1	0,4	-0,1	0,2	0,1	1	-0,07	-0,01	0	-0,2	0,5	0,2	0,5	0
restecg	0,1	0	0,01	0,1	0,2	0,01	1	-0,01	0,01	0,1	0,1	0,1	0,01	0,17
thalach	-0,4	-0,1	-0,3	-0,4	0	-0,01	-0,1	1	-0,4	-0,3	-0,4	-0,3	-0,3	-0,42
exang	0,1	0,1	0,4	0,1	0,1	0	0,1	-0,4	1	0,2	0,3	0,1	0,3	0,42
oldpeak	0,2	0,1	0,1	0,2	0,02	-0,02	0,1	-0,3	0,2	1	0,5	0,2	0,3	0,34
slope	0,2	0	0,2	0,1	-0,01	0,05	0,1	-0,4	0,3	0,5	1	0,1	0,3	0,33
ca	0,4	0,1	0,2	0,1	0,1	0,2	0,1	-0,3	0,1	0,2	0,1	1	0,2	0,46
thal	0,1	0,4	0,3	0,1	0,02	0,1	0,01	-0,3	0,3	0,3	0,3	0,2	1	0,52
Tahmin	0,2	0,3	0,4	0,2	0,01	0	0,2	-0,4	1	0,3	0,3	0,5	0,5	1

- **Diyabet hastalığı veri seti korelasyon matrisi**

Diyabet hastalığı ile en yüksek korelasyona sahip özniteliğin Glukoz olduğu belirlenmiştir. Hastalık ile Glukoz (A14) özniteliği arasındaki korelasyon katsayısı 0.69 olarak bulunmuştur. Glukoz ’un korelasyon katsayısının yüksek olması, hastalık tahmini yapılırken diğer özniteliklere göre daha fazla etkili olacağı anlamına gelmektedir. Bu yüzden Glukoz özniteliğinin veri setinde olması gerekmektedir. Çizelge 4.7’de görüldüğü üzere değişkenler arasındaki ilişkileri incelemek için korelasyon matrisi kullanılmıştır.

Çizelge 4.7. Diyabet hastalığı veri seti değişkenler arasındaki korelasyon matrisi

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	Tahmin
A1	1	0	-0,7	-0,1	0,26	0,03	-0,1	0,06	0,29	-0,3	0,09	0,16	-0,1	-0,12	-0,02
A2	-0,04	1	-0,1	-0,05	0	0,48	0,09	-0,17	0,02	0,28	0,25	0,05	0,13	0,33	0,4
A3	-0,7	0	1	0.27	-0,3	-0,07	0,03	0,06	-0,1	0,26	-0,09	-0,17	0,12	0,16	0,03
A4	-0,1	0	0,27	1	0,86	0,11	0,15	0,84	0,8	0,29	0,07	-0,32	0,29	0,22	0,20
A5	0,27	0	-0,25	0,86	1	0,15	0,13	0,81	0,87	0,1	0,1	-0,22	0,22	0,13	0,18
A6	0,03	0,5	-0,1	0,11	0,15	1	0,6	0,26	0,17	0,21	0,29	0,04	0,15	0,2	0,29
A7	-0,1	0,1	0,03	0,2	0,1	0,6	1	0,17	0,1	0,15	0,2	0,07	0,08	0,01	0,06
A8	0,06	0,2	0,1	0,84	0,81	0,26	0,17	1	0,83	0,53	0,1	-0,3	0,28	0,23	0,27
A9	0,29	0,0	-0,1	0,81	0,87	0,17	0,1	0,83	1	-0,02	0,1	-0,2	0,19	0,12	0,18
A10	-0,3	0,3	0,3	0,3	0,1	0,21	0,15	0,53	0	1	0,1	-0,1	0,21	0,23	0,22
A11	0,1	0,3	-0,1	0,1	0,12	0,29	0,2	0,14	0,09	0,1	1	0,1	0,51	0,15	0,2
A12	0,16	0,1	-0,2	-0,32	-0,22	0,04	0,07	-0,3	-0,2	0,1	0,1	1	-0,7	-0,2	-0,17
A13	-0,1	0,1	0,1	0,29	0,22	0,15	0,08	0,28	0,19	0,21	0,51	-0,7	1	0,31	0,3
A14	-0,1	0,3	0,2	0,22	0,13	0,19	0	0,23	0,12	0,23	0,15	-0,2	0,3	1	0,69
Tahmin	0,0	0,4	0,03	0,2	0,18	0,28	0,07	0,28	0,18	0,22	0,2	-0,2	0,3	0,69	1

Korelasyon matrisi incelendiğinde hastalık ile en düşük korelasyon değerine (-0,02) sahip olan Cinsiyet (A1) özniteliğinin diğer özniteliklere göre daha az etkili olabileceği anlamı çıkarılabilir.

4.1.4. Isı haritasının incelenmesi

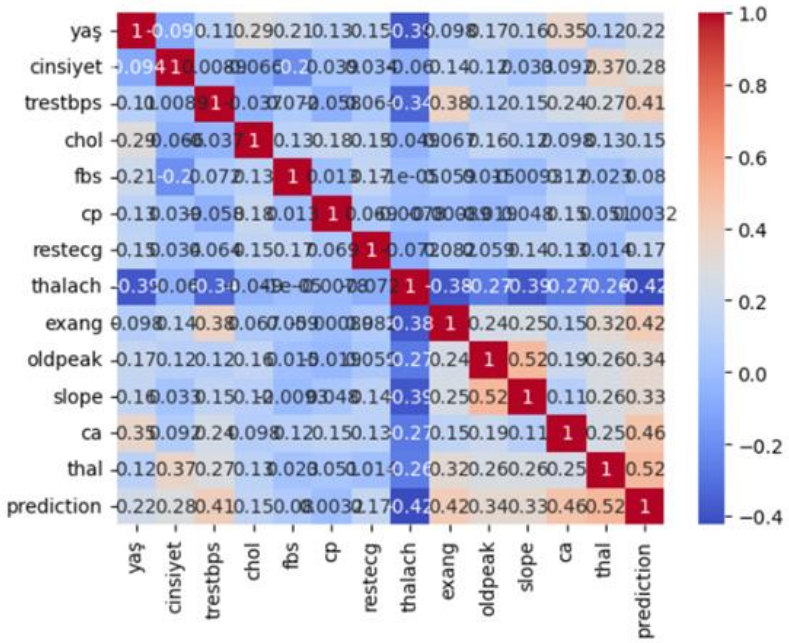
Isı Haritası (Heat Map): Korelasyon ısı haritası, modeldeki değişkenler arasındaki korelasyon katsayılarını görselleştirmek için kullanılan tekniktir. Isı haritasında, korelasyon katsayılarının değerlerini göstermek için renk tonları kullanılmaktadır. Genellikle koyu renkler yüksek korelasyonları, açık renkler ise düşük korelasyonları ifade etmektedir. Isı haritası, değişkenler arasındaki güçlü ilişkileri belirlemek, özniteliklerin modelin tahminine etkisini incelemek ve modelin tahmin performansını artırmak için kullanılmaktadır (Whitmire vd., 2021).

Çalışmada, kalp hastalığı ve diyabet hastalığı tahmininde öznitelikler arasındaki ilişkileri incelemek için ısı haritası kullanılmıştır. Isı haritası, özniteliklerin hastalık tahminine olan etkilerini ve birbirleriyle olan ilişkilerini analiz etmek için önemli araçtır. Isı haritası sayesinde, hangi özniteliklerin hastalık tahmininde daha yüksek öneme sahip olduğu belirlenmiştir. Isı haritası sayesinde veri seti daha iyi analiz edilmiş ve modelin anlaşılmasına yardımcı olmuştur.

- **Kalp hastalığı veri seti ısı haritası**

Isı haritaları, değişkenler arasındaki ilişkileri görsel olarak incelemek için kullanılmaktadır. Isı haritası, Python programlama dili kullanılarak pandas, seaborn ve matplotlib kütüphaneleri ile oluşturulmuştur. İlk olarak, veriler pandas kütüphanesi kullanılarak işlenip hazırlanmış, ardından seaborn kütüphanesi ile ısı haritası oluşturulmuş ve matplotlib kütüphanesi ile görselleştirilmiştir.

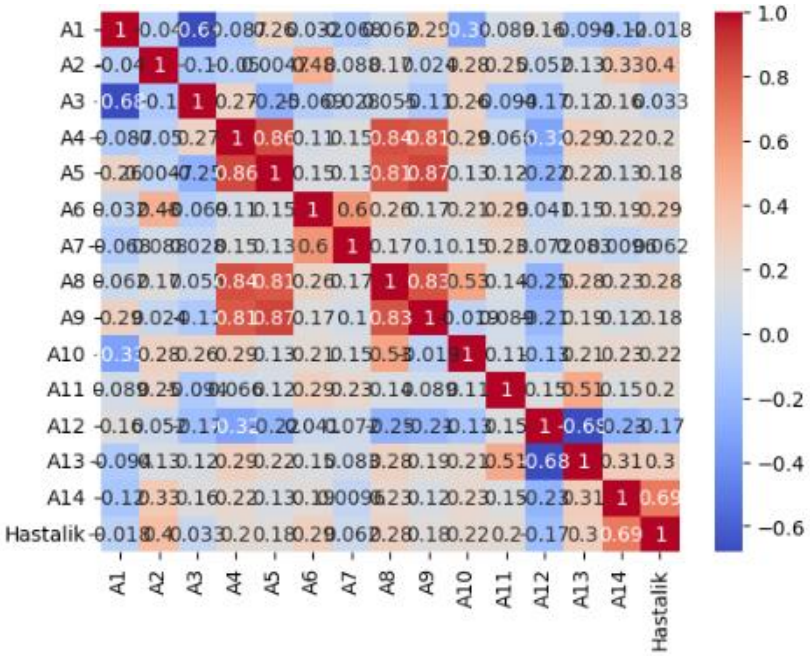
Isı haritasında mavi renkler değişkenler arasındaki düşük korelasyonu göstermektedir. Kırmızı renge yaklaştıkça değişkenler arasında yüksek korelasyon olduğunu göstermektedir. Kalp hastalığı veri setinin ısı haritasına bakıldığında öznitelikler arasında en yüksek öneme sahip değişkenler Ca (Floroskopide Boyanan Büyük Damar Sayısı), Thal (Ulaşılan Maksimum Kalp Hızı) ve Exang (Egzersizle Tetiklenen Anjina) olarak belirlenmiştir. Bunun yanı sıra, Slope (Pik Egzersiz ST Segmenti) değişkeninin ısı haritasında orta düzeyde bir öneme sahip olduğu gözlemlenmiştir. Şekil 4.1'de değişkenler arasındaki ilişkiler ısı haritasında gösterilmektedir.



Şekil 4.1. Kalp hastalığında değişkenler arasındaki ilişkilerin ısı haritası incelenmesi

- **Diyabet hastalığı veri seti ısı haritası**

Diyabet hastalığının tahmininde öznitelikler arasındaki ilişkiler, Python programlama dili ve pandas, seaborn, matplotlib kütüphaneleri kullanılarak oluşturulan ısı haritası ile incelenmiştir. Isı haritası incelendiğinde en yüksek korelasyona sahip öznitelik Glukoz olarak belirlenmiştir. Isı haritasına bakıldığında hastalık ile en düşük düşük korelasyon değerine sahip olan Cinsiyet özniteliğinin diğer özniteliklere göre daha az etkili olabileceği anlamı çıkarılmaktadır. Glukoz'un korelasyon katsayısının yüksek olması nedeniyle hastalık tahmini yapılırken diğer özniteliklere göre daha fazla etkili olacağı Cinsiyet'in ise daha az etkili olacağı belirlenmiştir. Ayrıca, diğer korelasyon katsayılarına göre daha düşük değerlere sahip olan boy ve diyastolik kan basıncı gibi özniteliklerin hastalık tahmininde nispeten daha az etkili olabileceği şeklinde yorum yapılabilir. Şekil 4.2'de değişkenler arasındaki ilişkiler ısı haritasında gösterilmektedir.



Şekil 4.2. Diyabet hastalığında değişkenler arasındaki ilişkilerin ısı haritası incelenmesi

4.1.5. Veri ön işleme

Veri ön işleme, makine öğrenmesi yöntemlerinin en önemli ve en çok zaman alan kısımlarından olmuştur. Başarı oranının en önemli etkenlerinden biri, veri ön işleme adımlarının iyi yapılmasından kaynaklanmaktadır. Veri ön işleme sayesinde veri setinde eksik, tutarsız, aynı boyutta olmayan verilerin tespitinin yapılması, verilerin normalizasyonu, nümerik, kategorik verilerin belirlenmesi sayesinde veri setinin analizinin iyi yapılması, veri setini tanımak ve performansını arttırmak için detaylı incelenmesi gerekmektedir. Kalp hastalığı ve diyabet hastalığı tahmininde makine öğrenmesi teknikleri kullanılmadan önce, veri ön işleme adımları gerçekleştirilmiştir.

- **Kalp hastalığı veri setinde eksik verilerin kontrol edilmesi**

Kalp hastalığı veri setinde eksik veri kontrolü yapılmış ve model tahminine başlamadan önce veri ön işleme adımları uygulanmıştır. Veri ön işleme yapılarak veri seti incelenmiş ve analiz için hazır hale getirilmiştir. Eksik verilerin belirlenmesi durumunda, modelin başarısını düşürmemek amacıyla eksik veriler, veri setinden çıkarılarak veya uygun

istatistiksel yöntemlerle doldurularak veri setinin bütünlüğü sağlanması amaçlanmaktadır. Eksik verilerin belirlenmesi, modelin performansını ve başarısını önemli ölçüde etkilemektedir. Eksik verilerin doğru yönetimi, modelin tahmin gücünü artırarak sonuçların doğruluğunu ve güvenilirliğini artırmaktadır. Kalp hastalığı veri setindeki 13 öznitelik için eksik veri kontrolü gerçekleştirilmiş ve eksik verinin olmadığı tespit edilmiştir.

- **Diyabet hastalığı veri setinde eksik verilerin kontrol edilmesi**

Diyabet hastalığını tahmin etmek için kullanılan veri setinde 14 özniteliğin eksik veriye sahip olup olmadığı kontrol edilmiştir. Çizelge 4.8'de, diyabet hastalığı veri setinde eksik verilerin tespit edilmesine ilişkin sonuçlar gösterilmiştir.

Çizelge 4.8. Diyabet hastalığı veri setinde eksik verilerin tespit edilmesine ilişkin sonuçlar

Öznitelik	Eksik veri sayısı
Cinsiyet	0
Yaş	0
Boy	5
Ağırlık	1
Vücut Kitle İndeksi	6
Sistolik Kan Basıncı	5
Diastolik Kan Basıncı	5
Bel çevresi	2
Kalça çevresi	2
Bel/Kalça Oranı	2
Kolestrol	1
HDL Kolestrol	1
Kolestrol/ HDL Kolestrol	1
Glukoz	0

Eksik veriler incelendiğinde, belirli özniteliklerde veri eksiklikleri olduğu gözlemlenmiştir. Boy, ağırlık, sistolik kan basıncı, diyastolik kan basıncı, bel çevresi, kalça çevresi, kolesterol ve HDL kolesterolde eksik veriler bulunmuştur. Ayrıca, vücut kitle indeksi (VKİ), bel/kalça oranı, kolestrol/HDL kolestrol oranında da, diğer değişkenlerdeki eksiklikler nedeniyle veri kayıpları tespit edilmiştir.

MCAR (Missing Completely At Random) yöntemi, veri setinde eksik verilerin tamamen rastgele oluştuğu durumlarda kullanılmaktadır. Eksik verileri doldurmak için eksik verilerin bulunduğu satırların ilgili sütununun ortalaması alınarak boşluklar doldurulmaktadır. Bu yöntem, eksik verilerin modelin performansı üzerindeki olumsuz etkisini azaltmaya yardımcı olmaktadır (Little ve Rubin, 2019).

Eksik verilerden, ilk gruptaki eksik değerler (Boy, ağırlık, sistolik kan basıncı, diastolik kan basıncı, bel çevresi, kalça çevresi, kolesterol ve HDL kolesterol) MCAR yöntemi kullanılarak ilgili özelliklerin ortalama değerleri ile doldurulmuştur. İkinci gruptaki eksik veriler ise (vücut kitle indeksi, bel/kalça oranı, kolestrol/HDL kolestrol oranı), ilk gruptaki eksiklikler giderildikten sonra tekrar hesaplanarak tamamlanmıştır. Bu işlemlerin ardından, veri setinde kontrol edildiğinde eksik veri kalmadığı belirlenmiştir.

4.1.6. Verilerin normalizasyonu

Normalizasyon, veri ön işleme adımlarının en önemli adımlarından biridir. Normalizasyon farklı değerdeki verileri belirli aralığa ölçeklendirmek amacıyla kullanılmaktadır. Veri setinde farklı değerlerdeki veriler algoritmaların performansını ve çalışmasını olumsuz yönde etkilemektedir. Bu nedenle normalizasyon kullanılarak genellikle veriler 0 ile 1 aralığına ya da -1 ile 1 aralığına getirilmektedir (Géron, 2022).

- **Z Skoru**

Normalizasyon tekniklerinden olan Z-Skoru normalizasyon belirli bir ölçüm değerinin ortalama değerden uzaklığını ve ölçüm değerinden ne kadar sapma olduğunu ölçmektedir (Chubb ve Simpson, 2012) Denklem 4.2.'de Z-Skoru'nun matematiksel tanımı gösterilmiştir.

$$Z = \frac{(x - \mu)}{\sigma} \tag{4.2}$$

Z-Skor yönteminde Z, skor değerini göstermektedir. x hesaplanmak istenilen verinin değeridir. μ veri örnekleminin ortalama değeridir. σ veri örnekleminin standart sapmasını açıklamaktadır (Curtis vd., 2016).

- **Kalp hastalığı veri seti normalizasyonu**

Kalp hastalığı tahmini için veri setini analiz etmek amacıyla veri türleri detaylı olarak incelenmiştir. Kalp hastalığı veri setinde nümerik veriler yaş, trestbps, chol, thalach, oldpeak Z Score yöntemi kullanılarak normalize edilirken kategorik veriler ca, cinsiyet, cp, fbs, restecg, exang, slope, thal için normalize işlemi yapılmamıştır. Veri setinde normalizasyon işlemi yapılarak analiz doğruluğu ve güvenilirliği artırılmıştır. Bu yöntemle, farklı özellikler arasındaki ölçek farklılıkları giderilmiş ve verilerin daha tutarlı hale gelmesi sağlanmıştır. Çizelge 4.9'da kalp hastalığı veri setinde Z-Score yöntemi kullanılarak elde edilen normalizasyon sonuçları gösterilmiştir.

Çizelge 4.9. Kalp hastalığı veri setinde normalizasyon sonuçları

	yaş	cinsiyet	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	1,6	1	0	1,6	-0,3	1	2	-0,8	0	-0,7	1	1	0
1	1,6	0	0	0,5	-0,2	0	0	-0,1	0	0,8	0	2	0
2	1,3	0	0	1	-0,4	0	0	-1,6	0	1,6	2	0	0
3	1,6	1	0	0,4	0,7	1	2	-1,1	0	0,5	1	1	0
4	1	1	0	-1,2	0,7	0	2	-0,2	1	0,8	1	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…
292	-1,6	1	3	1,1	-0,5	0	0	1	0	-0,8	0	0	2
293	-1,7	1	3	-0,8	-0,5	0	0	-0,4	0	0,3	1	0	2
294	-2,2	1	3	-0,7	-1	0	0	-0,3	1	0,6	1	0	2
295	-2,2	0	3	0,4	-1,2	0	0	-0,4	0	0,5	0	0	0
296	-2,2	1	3	-0,3	0,7	0	0	-0,3	1	-0,8	0	0	2

- **Diyabet hastalığı veri seti normalizasyonu**

Diyabet hastalığı veri seti incelendiğinde, tüm verilerin nümerik olduğu belirlenmiştir. Bu nedenle tüm veriler üzerinde Z-Skoru kullanılarak normalizasyon işlemi uygulanmıştır. Normalizasyon işlemi yapılarak makine öğrenmesi algoritmalarının performansının artırılması amaçlanmıştır. Çizelge 4.10'da, diyabet hastalığı veri setinde Z-Skoru yöntemi kullanılarak elde edilen normalizasyon sonuçları gösterilmiştir.

Çizelge 4.10. Diyabet hastalığı veri setinde normalizasyon sonuçları

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
0	0,8	-0,2	-1,1	-1,5	-1,1	-1,0	-1,8	-1,7	-1,0	-1,7	-0,2	0,4	-0,6	-0,7
1	0,8	0,7	-0,3	0,4	0,5	1,7	0,5	1,3	1,3	0,2	-0,3	-0,5	0	1,0
2	-1,2	0,9	1,3	1,1	0,3	-0,4	-1,3	0,9	0,6	0,7	-0,7	-1,4	1,4	0,6
3	0,8	-1,3	-0,5	0,9	1,2	-1,2	-1,2	1,3	0,8	1,0	-1,0	-1,4	1,1	-0,5
4	-1,2	0,3	2,3	0,1	-0,9	-0,8	0,7	-0,1	-1,0	1,5	-0,7	0,8	-1,0	1,5
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
536	0,8	0,1	-0,8	-0,7	-0,3	0	1,2	-1,1	-0,1	-2,0	1,6	0,3	0,2	-0,6
537	0,8	-1,3	0,7	-0,4	-0,7	-0,9	-1,0	-1,0	-0,6	-0,8	-0,4	-0,6	0,2	1,9
538	0,8	0,6	0,5	1,0	0,7	1,5	0,5	1,6	1,1	1,0	1,5	-0,5	1,1	1,9
539	0,8	-0,5	-0,8	0,4	0,9	-0,9	-0,4	0,4	0,8	-0,5	-0,3	0,2	-0,5	-0,8
540	0,8	1,1	-0,5	0,9	1,3	-1,8	-0,9	1,8	2,6	-0,6	-1,1	1,7	-1,4	-0,6

4.1.7. Öznitelik seçimi

Öznitelik seçimi, yapay zeka tekniklerinde yaygın olarak kullanılan veri ön işleme adımıdır. Bu adım, modelin performansını artırmak, aşırı uyumu önlemek ve öğrenme sürecinin verimliliğini artırmak amacıyla kullanılmaktadır. Öznitelik seçimi, modeli gereksiz bilgilerden arındırarak, hem daha iyi genelleme yapabilen hem de daha hızlı çalışan bir model elde edilmesini sağlamaktadır. Ayrıca, gereksiz özniteliklerin modelden

çıkarılması, hesaplama süresinin kısalmasına yardımcı olmaktadır. Öznitelik seçimi, yalnızca modelin performansını artırmakla kalmayıp aynı zamanda hesaplama maliyetlerini de azaltmaktadır (Li vd., 2017)

Çalışmada, kalp hastalığı ve diyabet hastalığı tahmininde, önem derecesi yüksek ve düşük öznitelikler belirlenerek modelin performansı değerlendirilmiştir. Öznitelik seçimi yapılarak modelin başarısı analiz edilmiştir.

- **Kalp hastalığı ve diyabet hastalığı veri seti öznitelik seçimi**

Çalışmada kalp hastalığı ve diyabet hastalığı tahmini yapılırken öznitelik seçimi korelasyon matrisi ve ısı haritaları incelenerek yapılmıştır. Bu nedenle veri setinde kullanılan özniteliklerin modelin performansına ve hastalık tahminine önemini anlamak için korelasyon ve ısı haritasına bakılarak sonuçlar analiz edilmiştir. Daha sonra makine öğrenmesi tekniklerinden rassal orman algoritması, lojistik regresyon ve XGBoost algoritmaları kullanılarak modelin yüksek performans göstermesi amacıyla en iyi hiperparametre değerleri bulunarak yüzlerce denemeler yapılmıştır ve sonuçlar incelenmiştir. Kalp hastalığını tahmin etmek amacıyla incelenen 13 öznitelikten fbs özniteliği çıkarıldığında model en iyi başarı performansını gösterirken diyabet hastalığı tahmini amacıyla 14 öznitelikten önem derecesi diğerlerinden düşük olan cinsiyet özniteliği çıkarıldığında model en uygun sonuca ulaşmıştır.

4.1.8. Doğrulama

Doğrulama (validation), makine öğrenimi modelini değerlendirmek için kullanılan bir doğrulama tekniğidir. Modelin performansını değerlendirmek ve aynı zamanda aşırı uyum (overfitting) sorununu tespit etmek amacıyla kullanılmaktadır (Balci, 1998).

Çalışmada, modelin performansını değerlendirmek amacıyla veri seti eğitim, test ve doğrulama (validation) setleri olarak ayrılmıştır. Kalp hastalığı ve diyabet hastalığı tahmini için makine öğrenmesi teknikleri kullanılarak yüzlerce deney yapılmış ve modeller en uygun hiperparametrelerle optimize edilmiştir.

Doğrulama seti, eğitim sırasında aşırı uyum riskini azaltmak için farklı hiperparametre ayarlarıyla modelin performansını değerlendirmek amacıyla kullanılmıştır. Bu sayede, modelin eğitim sırasında aşırı uyum riski en aza indirilmiş ve modelin daha iyi genelleme sağlaması amaçlanmıştır. Modelin eğitim aşaması tamamlandıktan sonra, test seti kullanılarak modelin gerçek dünya koşullarındaki performansı ve genelleme yeteneği ölçülmüştür.

4.2. Izgara Arama Algoritması

Izgara arama algoritması, modelin performansını optimize etmek amacıyla hiperparametrelerin tüm olası kombinasyonlarını çapraz doğrulama kullanarak deneyerek en iyi hiperparametre değerlerini bulmaktadır. Makine öğrenmesinde modelin performansını arttırmak için sıklıkla kullanılmaktadır (Géron, 2022).

Çalışmada kalp hastalığı ve diyabet hastalığı veri setleri kullanılarak makine öğrenmesi tekniklerinden rassal orman algoritması, lojistik regresyon algoritması ve XGBoost algoritmasının en iyi hiperparametre değerlerinin belirlenmesinde ızgara arama algoritması kullanılmıştır.

- **Kalp hastalığı tahmini için en uygun hiperparametre değerlerinin bulunması**

Kalp hastalığı tahmini için rassal orman, lojistik regresyon ve XGBoost algoritmalarının ızgara arama algoritması kullanılarak en iyi hiperparametre değerleri bulunmuştur.

- Rassal Orman Algoritması

Rassal Orman Algoritmasındaki en iyi hiperparametre değerleri max_depth= 150, min_samples_leaf= 3, oob_score= True alındığında modelin en iyi sonuçları elde edilmiştir.

- Lojistik Regresyon Algoritması

Lojistik Regresyon Algoritmasındaki en iyi hiperparametre değerleri penalty= L1, C= 3, solver= liblinear, multi_class= over alındığında modelin en iyi sonucuna ulaşılmıştır.

- XGBoost Algoritması

XGBoost Algoritmasındaki en iyi hiperparametre değerleri learning_rate: 0.3, max_depth: 1, min_child_weight: 1 , random_state: 19 alındığında modelin en iyi sonuçlarına ulaşılmıştır.

- **Diyabet hastalığı tahmini için en uygun hiperparametre değerlerinin bulunması**

Diyabet hastalığı tahmini için ızgara arama algoritması kullanılarak rassal orman, lojistik regresyon ve XGBoost algoritmalarının en iyi hiperparametre değerleri bulunmuştur.

- Rassal Orman Algoritması

Rassal Orman Algoritmasındaki en iyi hiperparametre değerleri max_depth= 150, min_samples_leaf= 3, oob_score= True alındığında modelin en iyi sonuçları elde edilmiştir.

- Lojistik Regresyon Algoritması

Lojistik Regresyon Algoritmasındaki en iyi hiperparametre değerleri penalty= L1, C= 1, solver= liblinear, multi_class= over alındığında modelin en iyi sonucu elde edilmiştir.

- XGBoost Algoritması

XGBoost Algoritmasındaki en iyi hiperparametre değerleri learning_rate: 0.3, max_depth: 1, min_child_weight: 1, random_state: 19 alındığında modelin en iyi sonuçlarına ulaşılmıştır.

4.3. Karışıklık Matrisi

Karışıklık matrisi, makine öğrenmesi tekniklerinde model performansını değerlendirmek amacıyla kullanılmaktadır. Karışıklık matrisinde yer alan önemli bilgiler aşağıda açıklanmıştır. Çizelge 4.11'de karışıklık matrisi tablosu gösterilmiştir.

Çizelge 4.11. Karışıklık matrisi

		Gerçek Değer	
		Pozitif	Negatif
Tahmin Edilen	Pozitif	Doğru Pozitif (TP)	Yanlış Pozitif (FP)
	Negatif	Yanlış Negatif (FN)	Doğru Negatif (TN)

TP (Doğru Pozitif): Modelin, gerçekten pozitif olan değeri doğru şekilde tahmin etmesidir (Raschka, ve Mirjalili, 2019; Kurt vd., 2020 ; Ustundag vd., 2022).
FN (Yanlış Negatif): Modelin, gerçekte pozitif olan değeri yanlış tahmin etmesidir (Raschka, ve Mirjalili, 2019; Kurt vd., 2020; Ustundag vd., 2022).
FP (Yanlış Pozitif): Modelin gerçekte yanlış olan değeri doğru tahmin etmesidir (Raschka, ve Mirjalili, 2019; Kurt vd., 2020; Ustundag vd., 2022).
TN (Doğru Negatif): Modelin gerçekte yanlış olan değeri yanlış tahmin etmesidir (Raschka, ve Mirjalili, 2019; Kurt vd., 2020; Ustundag vd., 2022).

Karışıklık matrisi (Confusion Matrix), sınıflandırma modellerinin performansını değerlendirmek için kullanılmaktadır. Matris, modelin tahmin ettiği değerlerle gerçek değerler arasındaki ilişkiyi göstermektedir. Böylece, modelin doğruluğu ve hata oranı analiz edilerek güçlü ve zayıf yönleri incelenmektedir (Hastie vd., 2009).

Doğruluk (Accuracy): Modelin yaptığı tahminlerin ne kadar doğru olduğunu değerlendirmek için kullanılmaktadır. Modelin, doğru olarak yapılan tahminlerinin tüm tahminlere oranını ifade etmektedir. Doğruluk oranı denklem (4.3)'deki gibi hesaplanmaktadır (Raschka, ve Mirjalili, 2019 ; Kurt vd., 2020 ;Ustundag vd., 2022).

$$Doğruluk = \frac{TP + TN}{TP + TN + FN + FP} \tag{4.3}$$

Duyarlılık (Recall) : Modelin gerçek pozitif durumları ne kadar başarılı şekilde tahmin ettiğini gösteren performans metriğidir. Yüksek duyarlılık, modelin pozitif olan örnekleri

doğru şekilde sınıflandırma konusundaki başarısını göstermektedir. (Raschka, ve Mirjalili, 2019 ; Kurt vd., 2020 ; Ustundag vd., 2022).

$$Duyarlılık = \frac{TP}{TP + FN} \tag{4.4}$$

Kesinlik (Precision): Modelin pozitif tahminlerinin ne kadarının gerçekten doğru olduğunu ölçer. Kesinlik, modelin yaptığı pozitif tahminlerin güvenilirliğini ve yanlış pozitiflerin olasılığını değerlendirmek için kullanılır (Raschka, ve Mirjalili, 2019 ; Kurt vd., 2020 ; Ustundag vd., 2022).

$$Kesinlik = \frac{TP}{TP + FP} \tag{4.5}$$

F1-Skoru: Kesinlik (precision) ve duyarlılık (recall) metriklerinin harmonik ortalamasını ifade eder. F1 skoru, kesinlik ve duyarlılığı birlikte değerlendirerek, iki ölçüt arasında denge sağlamaktadır (Raschka, ve Mirjalili, 2019 ; Kurt vd., 2020 ; Ustundag vd., 2022).

$$F1\ Skoru = \frac{2 * Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{4.6}$$

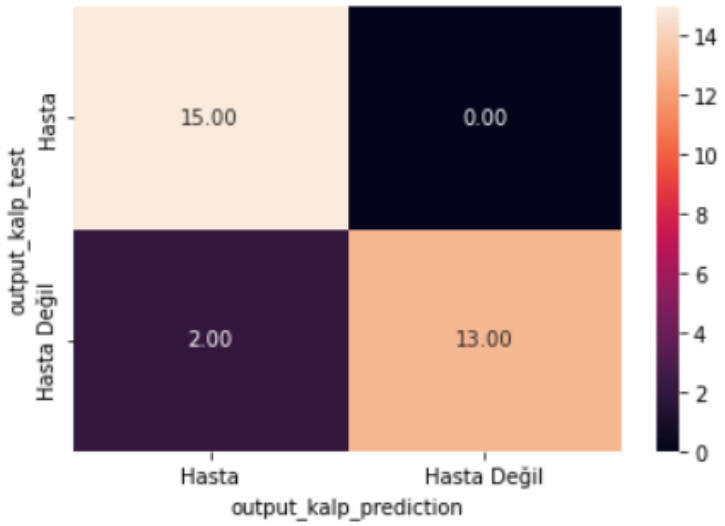
4.4. Rassal Orman Algoritması Sonuçları

Kalp hastalığı ve diyabet hastalığı veri setinde, rassal orman algoritması kullanılarak tahmin çalışması yapılmıştır. Veri setleri eğitim, test ve doğrulama olmak üzere üç kısma ayrılmış ve modelin yüksek performans göstermesi için en iyi hiperparametre değerlerini bulmak amacıyla çeşitli deneyler gerçekleştirilmiştir.

4.4.1. Kalp hastalığı tahmininde rassal orman algoritması sonuçları

Rassal orman algoritması kullanılarak kalp hastalığının tahminine yönelik çalışma gerçekleştirilmiştir. Bu çalışmada veri seti, eğitim, test ve doğrulama olarak üç bölüme ayrılmış ve çeşitli deneyler yapılmıştır. En yüksek doğruluk oranına sahip olan modelin sonuçları Çizelge 4.12.'de gösterilmiştir. Diğer deneylerin bir kısmı ise ekler bölümünde "Ek 1: Kalp hastalığı tahmini için rassal orman algoritması deney sonuçları" kısmında

yer almaktadır. Şekil 4.3'te, kalp hastalığı tahmini için rassal orman algoritmasıyla elde edilen karışıklık matrisi sonuçları gösterilmiştir.



Şekil 4.3. Kalp hastalığı tahmininde rassal orman algoritması karışıklık matrisi sonuçları

Rassal orman algoritması kullanılarak yapılan kalp hastalığı sınıflandırmasında doğruluk oranı %93,00 olarak bulunmuştur. Diğer performans metrik sonuçlarından %94,00 kesinlik, %93,00 duyarlılık ve %93,00 F1 elde edilmiştir.

297 hasta kaydı ve 13 öznitelikten oluşan kalp hastalığı veri setinde önem derecesi düşük fbs özniteliği çıkarılmıştır. Rassal orman algoritmasıyla en iyi sonuçları elde etmek için kullanılan hiperparametre değerleri: random state: 1905, estimator sayısı: 100, oob_score: False, bootstrap: True, criterion: entropy, max_features: sqrt, n_jobs: -1, min_samples_leaf: 10, ve min_samples_split: 2. Veri seti toplamda 297 hasta kaydından oluşmakta olup, bunun %80'i eğitim, %10'u test ve %10'u doğrulama için ayrılmıştır. 30 adet test verisinden sadece iki adet veride hasta değil sonucunun hasta olarak yanlış tespit edildiği belirlenmiştir. Modelin doğruluk performansı %93,00 olarak elde edilmiştir. Çizelge 4.12'de Kalp hastalığı rassal orman algoritması ile elde edilen literatür sonuçları bu tez çalışmasında elde edilen sonuçlarla karşılaştırılmıştır.

Çizelge 4.12. Kalp hastalığı tahmininde rassal orman algoritması literatür sonuçları

Kaynakça	Veri setindeki veri sayısı	Öznitelik Sayısı	Doğruluk Oranı	Kesinlik	Duyarlılık	F1 Puanı
Ani vd. (2016)	-	13	%89,00			
Ekrem vd. (2020)	303	6 ve 9	%86,99	-	-	-
Mohan vd. (2019)	297	13	%88,40	%90,10	%92,80	%90,00
Taşçı ve Şamlı (2020)	303	13	%83,60	-	-	-
Coşar ve Deniz (2021)	918	11	%88.00	-	-	-
Vasanthi vd. (2022)	303	12	%88,52	-	-	-
Gupta vd. (2022)	303	13	%85,71	%88,00	%86,27	%87,13
Özmen vd. (2018)	303	13	%84,21	-	%82,50	-
Vatansever vd. (2021)	303	13	%83,60	-	-	-
Farzana ve Veeraiah (2020)	304	13	%88,52	-	-	-
Ramesh vd. (2022)	303	7	%88.35	%88,70	%83,20	%87,90
Tez Çalışması	297	12	%93.00	%94,00	%93,00	%93,00

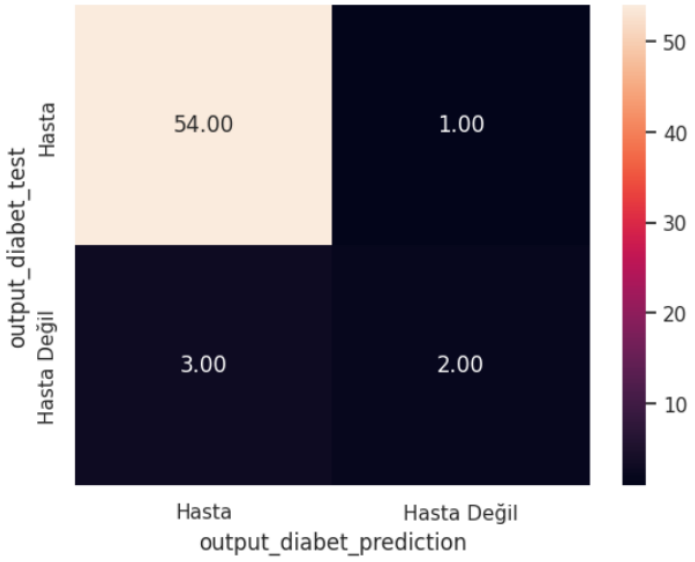
Literatürde Cleveland kalp hastalığı veri seti kullanılarak rassal orman algoritması ile elde edilen sonuçlar incelenmiştir. Bu çalışmalarda, farklı hasta kayıtları ve öznitelik seçimi yöntemleriyle ya da öznitelik seçimi olmadan elde edilen çeşitli sonuçlar ele alınmıştır. Tez çalışmasında, 297 hasta kaydından oluşan bir veri seti kullanılmış ve fbs (açlık kan şekeri) özniteliği çıkarılarak %93,00 doğruluk oranı elde edilmiştir. Diğer performans değerlerine bakıldığında %94,00 kesinlik, %93,00 duyarlılık ve F1 puanı %93,00 olarak bulunmuştur. Bu sonuçlar, tez çalışmasında elde edilen performans değerlerinin literatürdeki diğer çalışmalardan daha yüksek bir başarı oranına sahip olduğunu göstermektedir. Bu çalışma, veri setinin analizi, öznitelik seçimi işlemleri ve modele uygun en iyi hiperparametre değerlerinin bulunmasıyla literatüre katkı sağlamayı hedeflemiştir.

4.4.2. Diyabet hastalığı tahmininde rassal orman algoritması sonuçları

Diyabet hastalığının erken tahmininde rassal orman algoritması kullanılarak dengesiz veri seti dengeli hale getirilmiş ve veri seti sonuçları incelenmiştir.

- **Diyabet hastalığı dengesiz veri seti rassal orman algoritması sonuçları**

Diyabet hastalığı veri setinin 70'i eğitim, %15'i test ve %15'i doğrulama olarak ayrıldığında ve cinsiyet özniteliği çıkarıldığında %93,00 doğruluk oranı elde edilmiştir. Ancak diğer performans metriklerine bakıldığında doğruluk %93,00, duyarlılık %69,00, F1 Skoru %73,00 olarak bulunmuştur. Şekil 4.4'de, diyabet hastalığı dengesiz veri seti için rassal orman algoritmasıyla elde edilen karışıklık matris sonuçları gösterilmiştir.



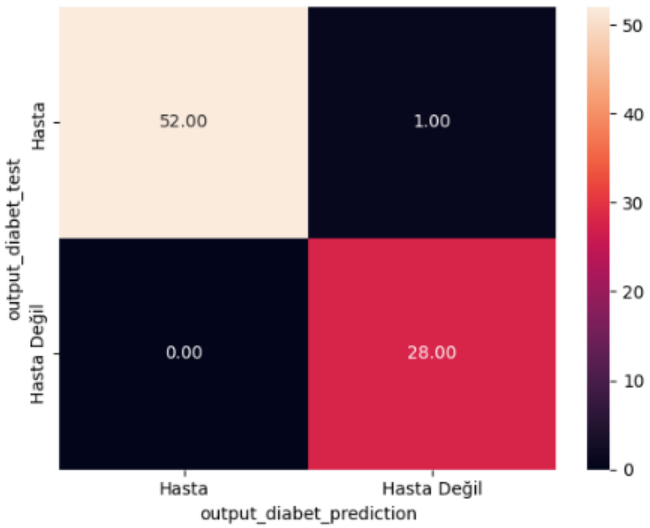
Şekil 4.4. Diyabet hastalığı tahmininde dengesiz veri seti rassal orman algoritması karışıklık matrisi sonuçları

Karışıklık matris sonuçlarına da bakıldığında başarı oranının yüksek olmasına rağmen, veri setinde hasta olmayan kişi sayısı hasta olanlara göre oldukça fazladır. Sonuçlara bakıldığında, veri setinin dengesiz olduğu belirlenmiştir. Yapay veri üretimi yoluyla veri seti daha dengeli hale getirilmiştir. Aynı hiperparametre değerleriyle rassal orman algoritması tekrar çalıştırılmıştır ve başarı oranı %99,00 elde edilmiştir. Bu sonuçlar, veri

setinin dengelenmesinin modelin genel performansını önemli ölçüde artırdığını göstermektedir.

- **Diyabet hastalığı tahmininde rassal orman algoritması sonuçları**

Diyabet hastalığı veri seti, eğitim, test ve doğrulama olarak üç bölüme ayrılmış ve en iyi hiperparametre değerleri kullanılarak modelin doğruluk performansını değerlendirmek amacıyla deneyler yapılmıştır. Deney sonuçları ekler bölümündeki "Ek 2: Diyabet hastalığı tahmini için rassal orman algoritması deney sonuçları" bölümünde yer almaktadır. Şekil 4.5'de, diyabet hastalığı veri seti için rassal orman algoritmasıyla elde edilen karışıklık matris sonuçları gösterilmiştir.



Şekil 4.5. Diyabet hastalığı tahmininde rassal orman algoritması karışıklık matrisi sonuçları

Karışıklık matrisine bakıldığında 81 adet test verisinden sadece bir adet veriyi yanlış sınıflandırmıştır. 52 veriyi hasta iken hasta olarak doğru ve 28 veriyi hasta değil iken hasta değil olarak doğru tahmin etmiştir. Sadece bir adet veride hasta sonucunu hasta değil olarak yanlış tespit ettiği gözükmektedir. Modelin doğruluk oranı %99,00 olarak bulunmuştur. Diğer performans değerlerine bakıldığında kesinlik, duyarlılık ve F1 puanı %99,00 olarak bulunmuştur. Yapılan deneylerin bir kısmı ise "Ek 2: Diyabet hastalığı tahmini için rassal orman algoritması deney sonuçları" bölümünde yer almaktadır.

Diyabet hastalığı tahmini için Rassal orman algoritmasında en iyi hiperparametre değerleri şu şekilde ayarlanmıştır; random state: 2, estimator: 100, oob_score: False, bootstrap: True, criterion: entropy, max_features: sqrt, max_depth: 150, n_jobs: -1, min_samples_leaf: 1, min_samples_split: 3. 541 hasta kaydından oluşan veri setinin %70'i eğitim, %15'i test ve %15'i doğrulama yapmak için ayrıldığında %99,00 başarı oranı elde edilmiştir. Çizelge 4.13'de Diyabet Hastalığı veri seti kullanılıp rassal orman algoritmasıyla elde edilen literatür sonuçları bu tez çalışmasında elde edilen sonuçlarla karşılaştırılmıştır.

Çizelge 4.13. Diyabet hastalığı tahmininde rassal orman algoritması literatür sonuçları

Kaynakça	Veri setindeki veri sayısı	Öznitelik Sayısı	Doğruluk Oranı	Kesinlik	Duyarlılık	F1 Puanı
Balasubramanian vd. (2020)	390	13	%88,00	%67,00	%50,00	%57,00
Veranyurt vd. (2020)	390	13	%92,30	-	-	-
Gill ve Pathwar (2022)	518	6	%93,95	-	-	-
Madhav ve Goel (2022)	403	16	%91,03	-	-	-
Ceh-Varela vd. (2024)	390	3	%91,82	%91,81	%92,18	%91,90
Tez Çalışması	541	13	%99,00	%99,00	%99,00	%99,00

Vanderbilt Diyabet hastalığı veri seti kullanılarak Rassal Orman algoritmasıyla yapılan çalışmalarda, eksik veriler içeren 13 hasta kaydı veri setinden çıkarıldıktan sonra veya çıkarılmadan, yapılan çeşitli literatür çalışmaları incelenmiştir. Tez çalışmasında, 541 hasta kaydından oluşan veri setinde cinsiyet özniteliği çıkarılarak %99,00 doğruluk oranı elde edilmiştir Diğer performans değerlerine bakıldığında, kesinlik, duyarlılık ve F1 puanı da %99,00 doğruluk oranını desteklemektedir.

Vanderbilt Diyabet hastalığı veri setinde hasta ve hasta olmayan bireylerin sayıları eşit dağılmadığı için, veri setini dengelemek amacıyla yapay veriler üretilmiştir. Bu sayede veri seti daha dengeli hale getirilerek modelin başarısı artırılmıştır. Aynı zamanda, Rassal

Orman Algoritması için hiperparametrelerin veri setine uygun şekilde ayarlanması ve hastalık tahmininde yüksek ve düşük öneme sahip özniteliklerin belirlenmesi, modelin performansını büyük ölçüde etkilemiştir. Bu çalışma, Rassal orman algoritmasının yapay zeka teknikleri arasında tahminleme gücünü gösterdiğini ortaya koymaktadır.

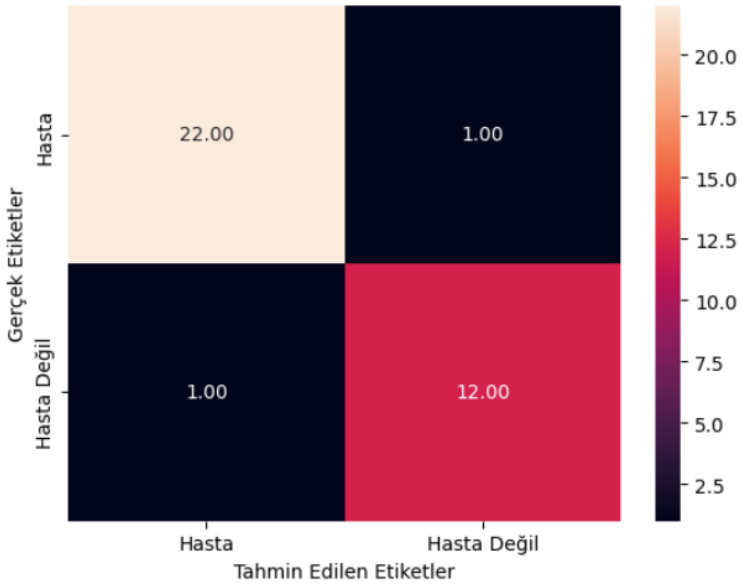
4.5. Lojistik Regresyon Algoritması Sonuçları

Kalp hastalığı ve diyabet hastalığı veri setinde, lojistik regresyon algoritması kullanılarak tahmin çalışması yapılmıştır. Bu çalışmada, veri setleri eğitim, test ve doğrulama olarak üçe ayrılmıştır. Modelin yüksek performans elde etmesini sağlamak amacıyla, en iyi hiperparametreleri belirlemek için çeşitli deneyler yapılmıştır.

4.5.1. Kalp hastalığı tahmininde lojistik regresyon algoritması sonuçları

Kalp hastalığını tahmini etmek için lojistik regresyon algoritması kullanılarak yapılan çalışmada, veri seti eğitim, test ve doğrulama olarak üç bölüme ayrılmıştır. Modelin performansını değerlendirmek için çeşitli deneyler yapılmıştır. Deney sonuçlarının bir kısmı ekler bölümündeki “Ek 3: Kalp hastalığı tahmini için lojistik regresyon algoritması deney sonuçları” bölümünde yer almaktadır.

Model, 36 adet test verisinden iki adet veriyi yanlış sınıflandırmıştır. 22 adet veriyi hasta olarak ve 12 adet veriyi hasta değil olarak doğru tahmin etmiştir. Bir adet veriyi hasta değil olarak ve bir adet veriyi hasta olarak yanlış tespit ettiği belirlenmiştir. Şekil 4.6’da, kalp hastalığı veri seti için lojistik regresyonla elde edilen karışıklık matrisi sonuçları gösterilmiştir.



Şekil 4.6. Kalp hastalığı tahmininde lojistik regresyon algoritması karışıklık matrisi sonuçları

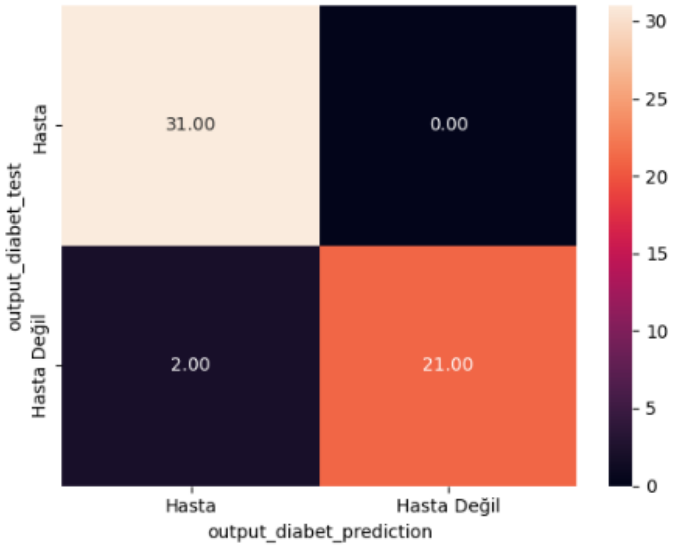
Kalp hastalığı tahmininde 297 hasta kaydı ve 13 öznitelik içeren veri setinden fbs (açlık kan şekeri) özniteliği çıkarılmıştır. Lojistik regresyon algoritması için en iyi hiperparametreler; random state: 19, max_iter: 100, penalty L2, C: 0.1, solver: liblinear, multi_class: None, class_weight: None, n_jobs: None olarak ayarlanmıştır. Veri setinin %80'i eğitim, %12'si test ve %8'i doğrulama için ayrılmıştır. Modelin doğruluk oranı %94,00 olarak bulunmuştur. Diğere performans değerlerine bakıldığında kesinlik, duyarlılık ve F1 puanı %94,00 olarak bulunmuştur. Bu sonuçlar, tez çalışmasındaki performans değerlerinin literatürdeki diğer çalışmalardan daha yüksek olduğunu göstermektedir. Çizelge 4.14'de Kalp Hastalığı veri seti kullanılıp lojistik regresyon algoritmasıyla elde edilen literatür sonuçları ve bu tez çalışmasında elde edilen sonuçlarla karşılaştırılmıştır.

Çizelge 4.14. Kalp hastalığı tahmininde lojistik regresyon algoritması literatür sonuçları

Kaynakça	Veri setindeki veri sayısı	Öznitelik Sayısı	Doğruluk Oranı	Kesinlik	Duyarlılık	F1 Puanı
Dwivedi (2018)	270	13	%85.00	%85.00	-	-
Kohli ve Arora (2018)	303	11	%87.10	-	-	-
Özmen vd. (2018)	303	13	%85.53	-	%87,50	-
Mohan vd. (2019)	297	13	%82,90	%89,60	%91,10	%90,20
Islam vd. (2020)	301	12	%82.90	%89,00	%86,00	%86,00
Taşçı ve Şamlı (2020)	303	13	%85.25	-	-	-
Asif vd. (2021)	303	13	%89,00	%86,10	%94,00	%89,80
Coşar ve Deniz (2021)	918	11	%85.00	-	-	-
Gupta vd. (2022)	303	13	%92.30	%90,74	%96,08	%93,34
Ramesh vd. (2022)	303	7	%84.22	%88,10	%86,10	%86,50
Tez Çalışması	297	12	%94.00	%94.00	%94.00	%94.00

4.5.2. Diyabet hastalığı tahmininde lojistik regresyon algoritması sonuçları

Diyabet hastalığı veri seti eğitim, test ve doğrulama olarak üç bölüme ayrılmıştır. Modelin performansını değerlendirmek amacıyla deneyler gerçekleştirilmiştir. Deney sonuçlarının bir kısmı ekler bölümündeki “Ek 4: Diyabet hastalığı tahmini için lojistik regresyon algoritması deney sonuçları" bölümünde yer almaktadır.” Şekil 4.7’de, diyabet hastalığı veri seti için lojistik regresyonla elde edilen karışıklık matrisi sonuçları gösterilmiştir.



Şekil 4.7. Diyabet hastalığı tahmininde lojistik regresyon algoritması karışıklık matrisi sonuçları

541 hasta kaydından oluşan diyabet hastalığı veri setinden cinsiyet özniteliği çıkartılarak veri seti eğitim, test ve doğrulama için üç bölüme ayrılmıştır. Veri setinin %80'i eğitim, %10'u test ve %10'u doğrulama amacıyla ayrılmıştır. Lojistik regresyon algoritması kullanılarak modelin doğruluk oranı % 96,00 olarak belirlenmiştir. 53 adet test verisinden sadece iki adet veriyi yanlış sınıflandırmıştır. Diğer performans değerlerine bakıldığında kesinlik %97,00, duyarlılık ve F1 puanı %96,00 olarak bulunmuştur. 31 veriyi hasta olarak doğru ve 21 veriyi hasta değil olarak doğru tahmin etmiştir. Sadece iki adet veriyi hasta olarak yanlış tahmin etmiştir.

Lojistik regresyon algoritması için en iyi hiperparametreler; random_state: 12, max_iter: 100, ceza türü: L1, C değeri: 1, çözümleyici: liblinear, multi_class: None, class_weight: None, n_jobs: None. Çizelge 4.15'de Diyabet Hastalığı veri seti kullanılıp lojistik regresyon algoritmasıyla elde edilen literatür sonuçları bu tez çalışmasında elde edilen sonuçlarla karşılaştırılmıştır.

Çizelge 4.15. Diyabet hastalığı tahmininde lojistik regresyon algoritması literatür sonuçları

Kaynakça	Veri setindeki veri sayısı	Öznitelik Sayısı	Doğruluk Oranı	Kesinlik	Duyarlılık	F1 Puanı
Balasubramanian vd. (2020)	390	13	%88,00	%88,00	%33,00	%47,00
Liza vd. (2021)	390	6	%91,00	%80,00	%75,00	-
Rajendra ve Latifi (2021)	390	8	%88,89	-	-	-
Gill ve Pathwar (2022)	491	2	%89,66	-	-	-
Madhav ve Goel (2022)	403	16	%96,15	-	-	-
Mijwil ve Aljanabi (2023)	403	15	%92,50	%95,00	%95,90	%95,40
Murthy vd. (2023)	390	8	%88,89	-	-	-
Ceh-Varela vd. (2024)	390	3	%86,06	%89,17	%81,94	%85,23
Tez Çalışması	541	13	%96,00	%97,00	%96,00	%96,00

Vanderbilt Diyabet hastalığı veri seti kullanılarak Lojistik regresyon algoritmasıyla %96,00 doğruluk oranı elde edilmiştir. Bu sonuçlar, tez çalışmasında elde edilen performans değerlerinin Lojistik regresyon algoritması kullanılarak elde edilen literatürdeki diğer çalışmalardan daha yüksek bir başarı oranına sahip olduğunu göstermektedir.

4.6. XGBoost Algoritması Sonuçları

Kalp hastalığı ve diyabet hastalığı veri setinde, XGBoost algoritması kullanılarak tahmin çalışması yapılmıştır. Veri setleri üç bölüme ayrılmıştır ve modelin performansını artırmak için çeşitli deneyler yapılarak en uygun hiperparametreler belirlenmiştir.

4.6.1. Kalp hastalığı tahmininde XGBoost algoritması sonuçları

XGBoost algoritması kullanılarak kalp hastalığı tahmini çalışması yapılmıştır. Şekil 4.8’de, kalp hastalığı veri setinde XGBoost algoritması ile elde edilen karışıklık matrisi sonuçları görülmektedir.

Şekil 4.8. Kalp hastalığı tahmininde XGBoost algoritması karışıklık matrisi sonuçları

297 hasta kaydından oluşan kalp hastalığı veri setinden fbs (açlık kan şekeri) özniteliği çıkarılmıştır. Veri setinin %80'i eğitim, %14'ü test ve %16'sı doğrulama olarak ayrılmıştır. Modelin doğruluk oranı %95,00 olarak belirlenmiştir. Diğer performans değerlerine bakıldığında kesinlik, duyarlılık, F1 puanı %95,00 bulunmuştur. 42 adet test verisinden sadece iki adet veriyi yanlış sınıflandırmıştır. 24 veriyi hasta olarak ve 16 veriyi hasta değil olarak doğru tahmin etmiştir. Bir adet veriyi hasta değil olarak ve bir adet veriyi hasta olarak yanlış tahmin etmiştir.

Kalp hastalığı veri seti kullanılarak XGBoost algoritması için hiperparametreler şu şekilde ayarlanmıştır: random state: 19, max_iter: 100, base_score: 0.5, booster: 'gbtree', colsample_bylevel: 1, colsample_bytree: 1, gamma: 0, learning_rate: 0.3, max_delta_step: 0, max_depth: 1, min_child_weight: 1, objective: 'binary:logistic', n_jobs: 0 ve n_estimators: 100. Modelin performansını değerlendirmek için deneyler yapılmıştır. Deney sonuçlarının bir kısmı “Ek: 5 Kalp hastalığı veri seti XGBoost algoritması deney sonuçları” bölümünde yer almaktadır. Çizelge 4.16’da Kalp Hastalığı veri seti kullanılıp XGBoost algoritmasıyla elde edilen literatür sonuçları bu tez çalışmasında elde edilen sonuçlarla karşılaştırılmıştır.

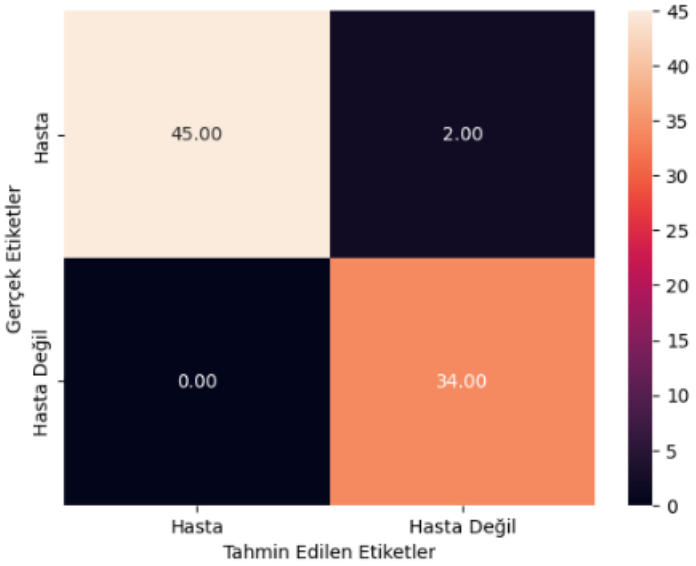
Çizelge 4.16. Kalp hastalığı tahmininde XGBoost algoritması literatür sonuçları

Kaynakça	Veri Setindeki Veri Sayısı	Öznitelik Sayısı	Doğruluk Oranı	Kesinlik	Duyarlılık	F1 Puanı
Anbuselvan (2020)	303	13	%78,69	-	-	-
Farzana ve Veeraiah (2020)	304	13	%78,69	-	-	-
Asif vd. (2021)	303	13	%85.20	%85,20	%88,00	%87,00
Lin (2021)	303	13	%82,10	-	-	%80,20
Budholiya vd. (2022)	303	13	%91,80	-	-	-
Doki vd. (2022)	303	6	%85,96	-	-	-
Saboor vd. (2022)	303	13	%91.80	-		-
Srinivas ve Katarya (2022)	303	13	%94,70	%92,65	%95,34	%93,98
Dissanayake ve Johar (2023)	303	13	%91,80	%95,83	%85,19	%90,20
Zeng (2023)	918	11	%88,84	-	%89,30	%90,40
Tez Çalışması	297	12	%95.00	%95,00	%95,00	%95,00

Tez çalışmasının, literatürdeki diğer çalışmalardan daha yüksek başarı oranı elde ettiği belirlenmiştir.

4.6.2. Diyabet hastalığı tahmininde XGBoost algoritması sonuçları

XGBoost algoritması kullanılarak diyabet hastalığı veri seti eğitim, test ve doğrulama olarak üçe ayrılarak modelin performansını belirlemek amacıyla deneyler yapılmıştır. Deney sonuçlarının bir kısmı “Ek 6: Diyabet hastalığı tahmini için XGBoost algoritması deney sonuçları” bölümünde yer almaktadır. Şekil 4.9’da, diyabet hastalığı veri setinde XGBoost algoritmasıyla elde edilen karışıklık matrisi sonucu gösterilmiştir.



Şekil 4.9. Diyabet hastalığı tahmininde XGBoost algoritması karışıklık matrisi sonuçları

Karışıklık matrisine bakıldığında 81 test verisinden iki adet veriyi yanlış sınıflandırılmıştır. 45 hasta verisini hasta olarak ve 34 hasta değil verisini hasta değil olarak doğru sınıflandırmıştır. Modelin doğruluk oranı %98,00 olarak belirlenmiştir.

Diyabet hastalığı veri setinde, XGBoost algoritmasında en iyi hiperparametre değerleri şu şekilde ayarlanmıştır: random state: 12, base_score: 0.5, booster: 'gbtree', colsample_bylevel: 1, colsample_bytree: 1, gamma: 0, learning_rate: 0.1, max_depth: 3, min_child_weight: 1, objective: 'binary:logistic', n_jobs: -1, n_estimators: 100.

XGBoost algoritması kullanılarak cinsiyet özniteliği çıkartılmış ve veri seti %70 eğitim, %15 test ve %15 doğrulama yapmak için ayrıldığında model %98,00 başarı oranı elde etmiştir. Diğer performans değerlerine bakıldığında F1 puanı %97,00, kesinlik, duyarlılık ve doğruluk %98,00 olarak bulunmuştur. Sonuçlara göre, veri setinin daha dengeli hale getirilmesinin modelin başarısında etkisinin önemli olduğu belirlenmiştir. Tez çalışmasından elde edilen başarı oranının literatürdeki diğer çalışmalara göre yüksek olduğu gözükmektedir. Çizelge 4.17’de Diyabet hastalığı veri seti kullanılıp XGBoost algoritmasıyla elde edilen literatür sonuçları bu tez çalışmasında elde edilen sonuçlarla karşılaştırılmıştır.

Çizelge 4.17. Diyabet hastalığı tahmininde XGBoost algoritması literatür sonuçları

Kaynakça	Veri setindeki veri sayısı	Öznitelik Sayısı	Doğruluk Oranı	Kesinlik	Duyarlılık	F1 Puanı
Balasubramain vd. (2020)	390	13	%88,00	%80,00	%33,00	%47,00
Madhav ve Goel (2022)	403	16	%92,32	-	-	-
Ceh-Varela vd. (2024)	390	3	%90,30	%89,16	%91,45	%90,17
Tez Çalışması	541	14	%98,00	%98,00	%98,00	%97,00

4.6.3. Yapay zeka teknikleriyle kalp hastalığı tahmini literatür çalışmaları

Literatürde, farklı kalp hastalığı veri setleri kullanılarak yapılan yapay zeka temelli çalışmalar incelendiğinde, kalp hastalığını tahmin etmede yüksek başarı gösterildiği görülmektedir. Yapay zeka tekniklerinin daha yaygın kullanılması, sağlık çalışanları için önemli bir destek aracı olabilir. Çizelge 4.18'de, kalp hastalığı ile ilgili yapay zeka teknikleri kullanılarak yapılan literatür araştırması sonuçları ile tez çalışması sonuçları gösterilmiştir.

Çizelge 4.18. Kalp hastalığı tahmininde yapay zeka teknikleri kullanılarak elde edilen literatür sonuçları

Kaynakça	Veri Seti	Kullanılan Yapay Zeka Tekniği	Doğruluk Oranı
Palaniappan ve Awang (2008)	Cleveland Veri Seti	Naive Bayes	%88,12
Deepika vd. (2016)	UCI	SVM	%95,56
Harjai ve Khatri (2019)	Cleveland Veri Seti	Yapay Sinir Ağları	%89,20
Krishnani vd. (2019)	Framingham	Rassal Orman	%96,80
Repaka vd. (2019)	UCI Kalp Hastalığı Veri Setleri	Naive Bayes	%89,77
Li vd. (2020)	Cleveland Veri Seti	SVM	%92,37
Rahim vd. (2021)	Framingham Heart Disease Cleveland	Ensemble (LR ve KNN) Ensemble Ensemble	%99,10 %98,00 %95.50
Baviskar vd. (2023)	Cleveland Klinik Kalp Yetmezliği	DPA-RNN+LSTM	%99,21 %93,56
Tez Çalışması	Cleveland Veri Seti	XGBoost	%95,00

4.6.4. Yapay zeka teknikleriyle diyabet hastalığı tahmini literatür çalışmaları

Literatürde, yapay zeka teknikleri kullanılarak farklı diyabet hastalığı veri setleri üzerinde yapılan çalışmalar incelendiğinde, yüksek başarı elde edildiği görülmektedir. Diyabet hastalığı üzerine yapay zeka teknikleriyle yapılan literatür araştırmasının sonuçları ile tez çalışması sonuçları Çizelge 4.19'da gösterilmiştir.

Çizelge 4.19. Diyabet hastalığı tahmininde yapay zeka teknikleri kullanılarak elde edilen literatür sonuçları

Kaynakça	Veri Seti	Kullanılan Yapay Zeka Tekniği	Doğruluk Oranı	AUC Değeri
Nai-Arun ve Moungmai (2015)	Sawanpracharak Bölge Hastanesi	Rassal Orman Algoritması	%85,56	
Deepika vd. (2016)	UCI	Naive Bayes	%73,59	
Kaur ve Kumari (2018)	PIMA veri seti	SVM-Linear modeli	%89,00	
Kohli ve Arora (2018)	PIMA Kızılderili	Destek Vektör Makinesi	%85.71	
Choi vd. (2019)	Kore Üniversitesi Guro Hastanesi	Lojistik Regresyon	-	%78,00
Lukmanto vd. (2019)	PIMA (PID) veri seti	SVM	%89,02	
Nadesh vd. (2020)	PIMA veri seti	Derin Sinir Ağları (DSA)	%98,16	
Tigga ve Garg (2020)	PIMA veri seti ve 18 soruluk anket	Rassal Orman Algoritması	%94,10	
Akyol ve Kargacı (2021)	Bangladeş, Sylhet Diyabet Hastanesi	Oylama Topluluğu Sınıflandırıcısı	%100,00	
Rajendra ve Latifi (2021)	PIMA veri seti	Lojistik Regresyon	%75,32	
Tiwari ve Singh (2021)	UCI diyabet veri setleri	XGBoost	%78,91	
Viswanatha vd. (2023)	PIMA veri seti Vanderbilt veri seti	Ensemble Ensemble	%78,00 %93,00	
Tez Çalışması	Vandelbilt Veri Seti	Rassal Orman Algoritması	%99,00	

Tez çalışmasında yapay zeka teknikleri kullanılarak kalp hastalığı ve diyabet hastalığı erken tahmini üzerinde çalışılmıştır. Aynı zamanda literatürde aynı ve farklı veri setleri ile yapılan çalışmalar da değerlendirilmiştir. Bu çalışmalar, yapay zekanın hastalık tahmininde yüksek başarı gösterdiğini ortaya koymaktadır. Yapay zeka tekniklerinin gelecekte sağlık alanında daha yaygın kullanılacağı düşünülmektedir.

4.6.5. Çapraz doğrulama yöntemi sonuçları

Kalp hastalığı ve diyabet hastalığı tahmininde kullanılan makine öğrenimi modellerinde, çapraz doğrulama yöntemi kullanılarak algoritmaların performansı değerlendirilmiştir.

- **Kalp hastalığı ve diyabet hastalığı veri setlerinin çapraz doğrulama yöntemi kullanılarak incelenmesi**

Kalp hastalığı ve diyabet hastalığı veri setleri için, model performansını değerlendirmek amacıyla çapraz doğrulama yöntemi kullanılmıştır. Rassal Orman, Lojistik Regresyon ve XGBoost algoritmalarına, $k = 5$ ve $k = 10$ katlı çapraz doğrulama yöntemleri uygulanmıştır. Çizelge 4.20'de Kalp hastalığı ve diyabet hastalığı çapraz doğrulama sonuçları gösterilmiştir.

Çizelge 4.20. Kalp hastalığı ve diyabet hastalığı çapraz doğrulama sonuçları

	Katman Sayısı	Rassal Orman	Lojistik Regresyon	XGBoost
Kalp Hastalığı veri seti	$k = 5$	%96,00	%95,00	%95,00
Kalp Hastalığı veri seti	$k = 10$	%97,00	%95,00	%95,00
Diyabet Hastalığı veri seti	$k = 5$	%96,00	%95,00	%96,00
Diyabet Hastalığı veri seti	$k = 10$	%97,00	%95,00	%96,00

Sonuçlar incelendiğinde algoritmaların hastalık tahmininde oldukça başarılı olduğu belirlenmiştir. Modelin performansı doğrulandıktan sonra, kalp hastalığı ve diyabet hastalığı tahmini için en iyi hiperparametre değerleri kullanılarak arayüz tasarımı çalışması yapılmıştır.

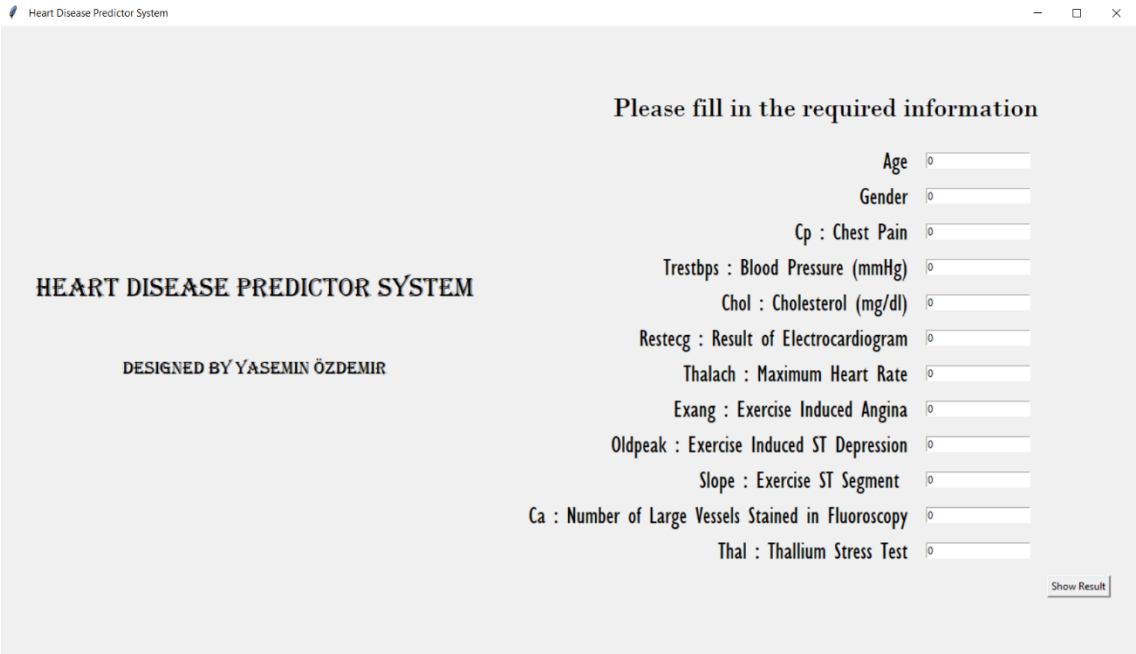
4.7. Arayüz Sistemi

Çalışmada, kalp hastalığı ve diyabet hastalığı erken tahmininde en yüksek performans değerleri veren yapay zeka teknikleri ve en uygun hiperparametreler kullanılarak arayüz tasarımı geliştirilmiştir. Kalp hastalığının tahmini için XGBoost algoritması ve Diyabet hastalığının tahmini için Rassal orman algoritması kullanılarak arayüz tasarımı geliştirilmiştir. Geliştirilen arayüz tasarımı, kalp hastalığı ve diyabet hastalığı ile ilgili önemli verilerin sisteme kolayca girilmesini ve sonuç ekranına hızlı erişim sağlamaktadır.

Arayüz tasarımının kalp hastalığı ve diyabet hastalıklarının tahmininde sağlık çalışanlarına yardımcı destek sistemi olması amaçlanmıştır.

4.7.1. Kalp hastalığı tahmini için arayüz sistemi

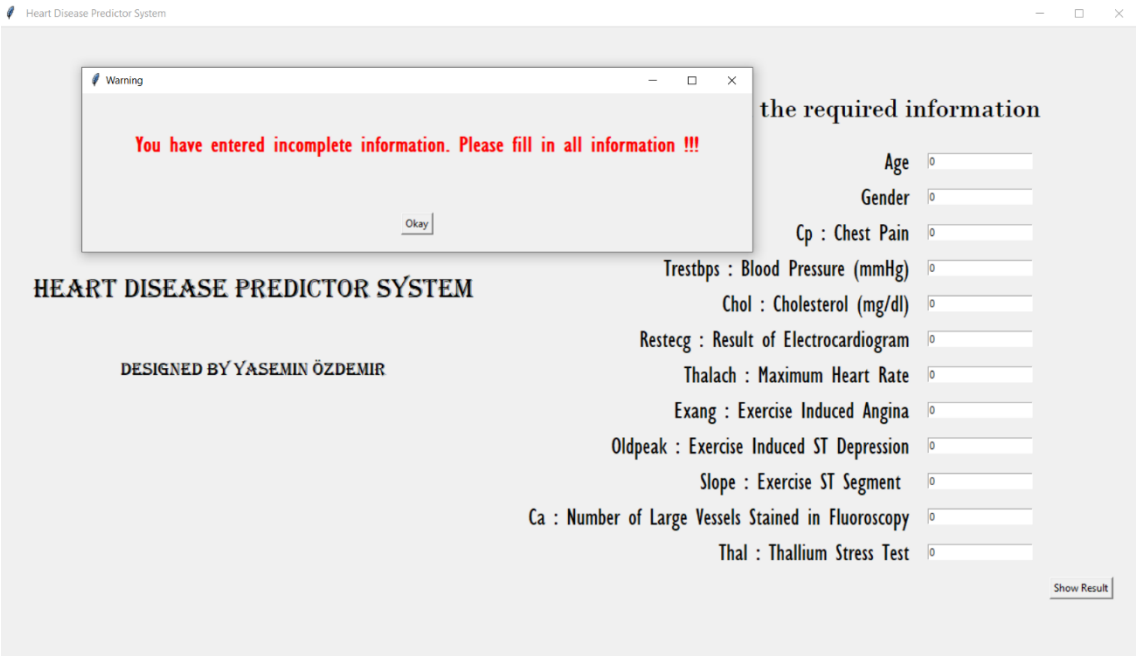
XGBoost'un kalp hastalığını tahmin etmek için etkili bir yöntem olduğu belirlendikten sonra geliştirilen arayüz tasarımının başarıyla çalıştığı doğrulanmıştır. Şekil 4.10'da, Kalp hastalığı tahmini için tasarımı yapılan kullanıcı arayüz sistemi ekran görüntüsü gösterilmiştir.



Şekil 4.10. Kalp hastalığı tahmini için kullanıcı arayüz sistemi

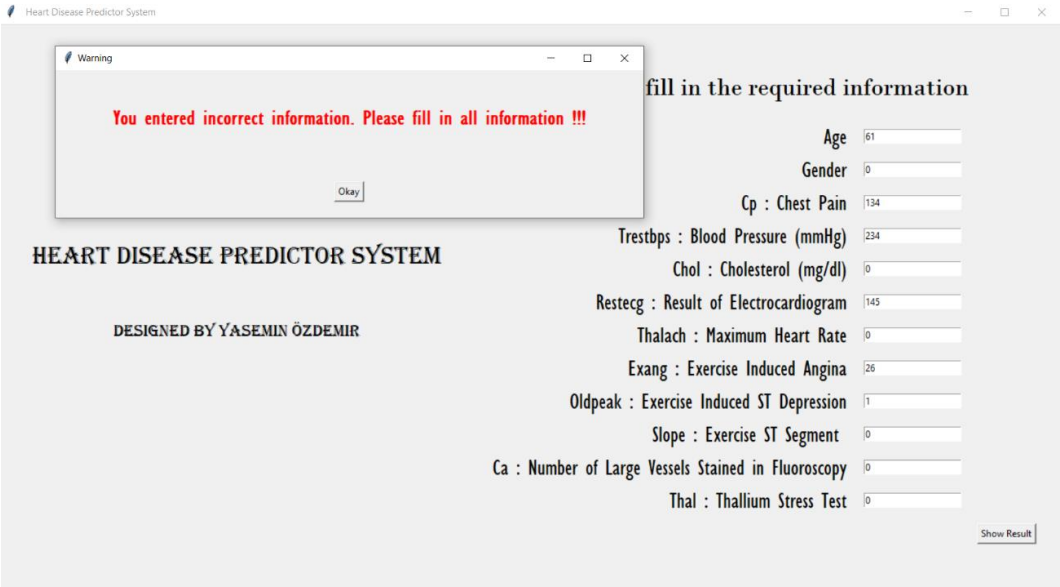
Kalp hastalığı tahmini için arayüz tasarımında tüm verilerin eksiksiz girilmesi sistemin doğru ve başarılı tahminler yapması açısından oldukça önemlidir. Bu nedenle, kullanıcılara tüm bilgileri doldurmaları konusunda uyarabilmek için geliştirilen arayüz tasarımında iki tane uyarı sistemi bulunmaktadır. İlk uyarı sisteminde kullanıcılar ekrandaki ilgili verileri boş girdiğinde "Sonucu Göster" butonuna bastığında, "Uyarı: Eksik bilgi girdiniz. Lütfen tüm bilgileri eksiksiz doldurun." şeklinde uyarı ekranı açılmaktadır. Kullanıcı gerekli tüm verileri ilgili bölümlere doldurduktan sonra "Sonucu

Göster” butonuna tıklayarak sonuç ekranında sonuç bilgisine ulaşmaktadır. Şekil 4.11’de, Kalp hastalığı tahmin sistemi arayüz sisteminde eksik bilgi uyarı ekranı gösterilmiştir.



Şekil 4.11. Kalp hastalığı tahmini arayüz sisteminde eksik bilgi uyarı ekranı

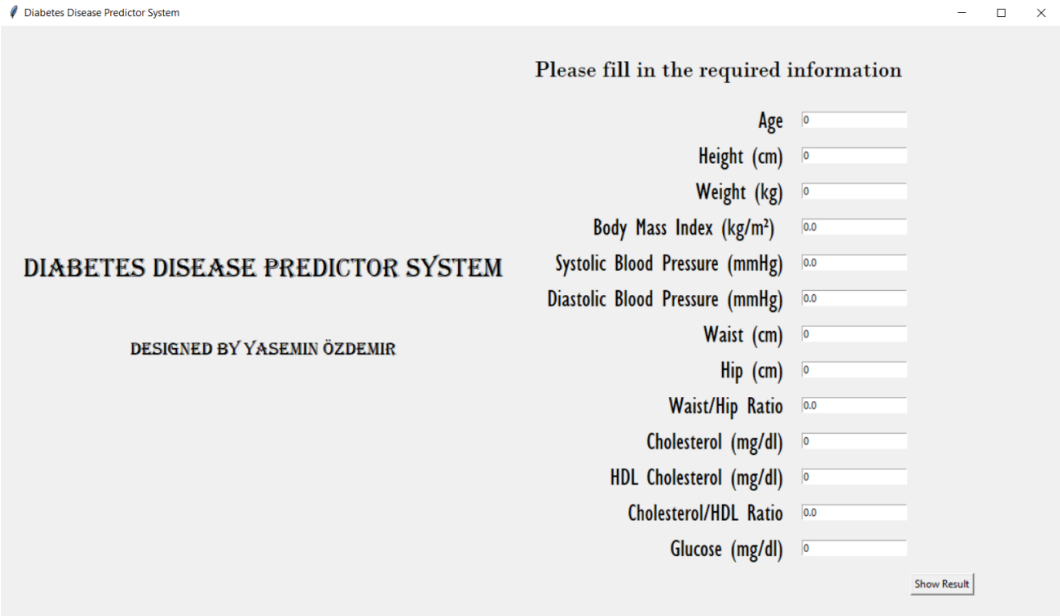
Arayüz tasarımında geliştirilen uyarı sistemi sayesinde kullanıcıların sisteme eksik veri girişleri önlenmiştir. Öznitelikler arasında 0 değerini alan kategorik değişkenler ve öznitelikler bulunmaktadır. Sisteme 0 alabilen değerlerin girişleri tanıtılarak kullanıcıların öznitelik verilerini doğru girmeleri amaçlanmıştır. Kullanıcı yanlış veri girişi yaptığında sistemde “Yanlış bilgi girdiniz. Lütfen tüm bilgileri doldurun.” şeklinde uyarı ekranı açılmaktadır. Geliştirilen arayüz tasarımı ile birlikte XGBoost algoritmasının kalp hastalığının tahmininde kullanılarak elde edilen sonuçlara bakıldığında oldukça başarılı sonuçlar elde ettiği belirlenmiştir. Şekil 4.12’de, Kalp hastalığı tahmin sistemi arayüz sistemi yanlış veri girişi uyarı sistemi gösterilmiştir.



Şekil 4.12. Kalp hastalığı tahmini arayüz sisteminde yanlış veri girişi uyarı sistemi

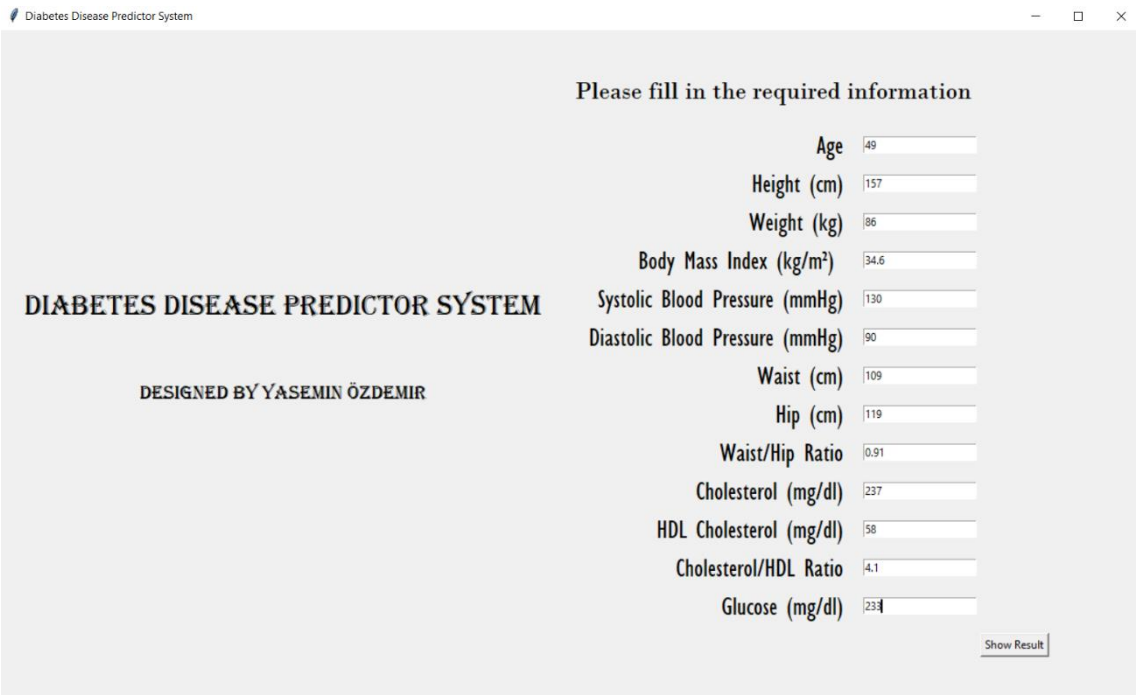
4.7.2. Diyabet hastalığı tahmini için arayüz sistemi

Diyabet hastalığı tahmininde rassal orman algoritması kullanılarak arayüz tasarımı geliştirilmiştir. Şekil 4.13'de, diyabet hastalığı tahmini için tasarımı yapılan arayüz sistemi ekranı gösterilmiştir.



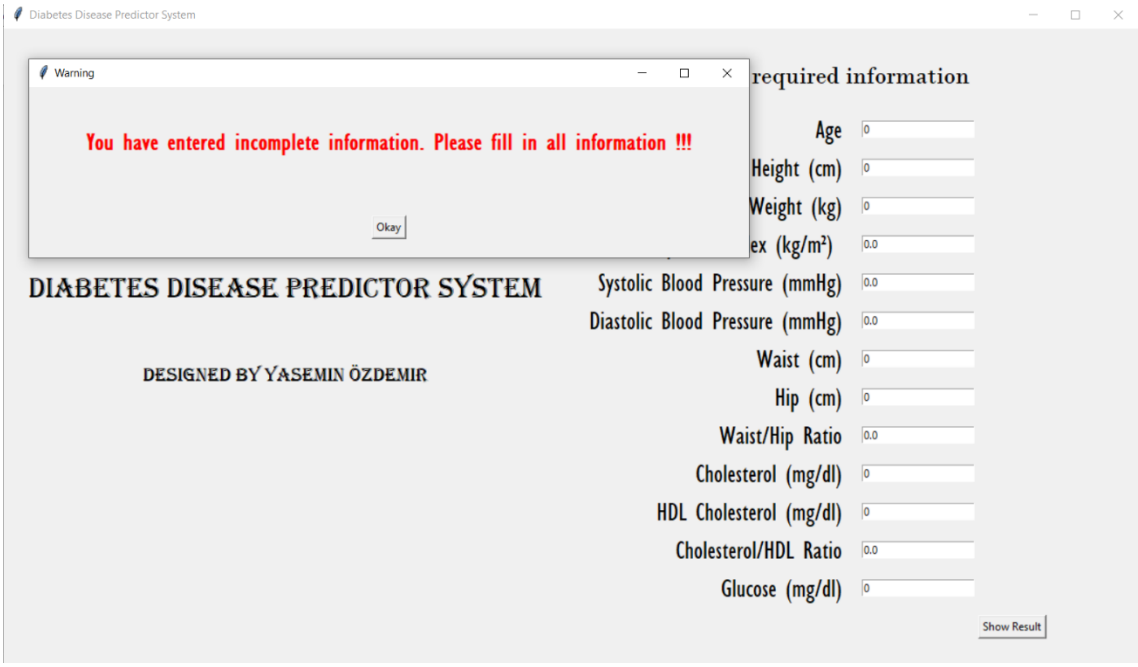
Şekil 4.13. Diyabet hastalığı tahmini için kullanıcı arayüzü sistemi

Rassal orman algoritması kullanılarak tasarlanan arayüz sisteminde sistemin performansını ölçmek amacıyla diyabet veri setinde, eğitim, test ve doğrulama işlemlerinden önce sistemin görmediği 7 hasta değil ve 2 hasta toplamda 9 veriden oluşan hasta kaydı verisi ayrılmıştır. Ayrılmış olan 9 veri sistemde tek tek denenmiştir ve geliştirilen ara yüz tasarımı 9 hasta kaydı verisi sonuçlarının hepsini doğru tahmin ederek %100,00 doğrulukla başarı göstermiştir. Şekil 4.14'de, diyabet hastalığı tahmini arayüz sisteminde gerekli bilgilerin doldurulmasına ilişkin ekran görüntüsü gösterilmiştir.



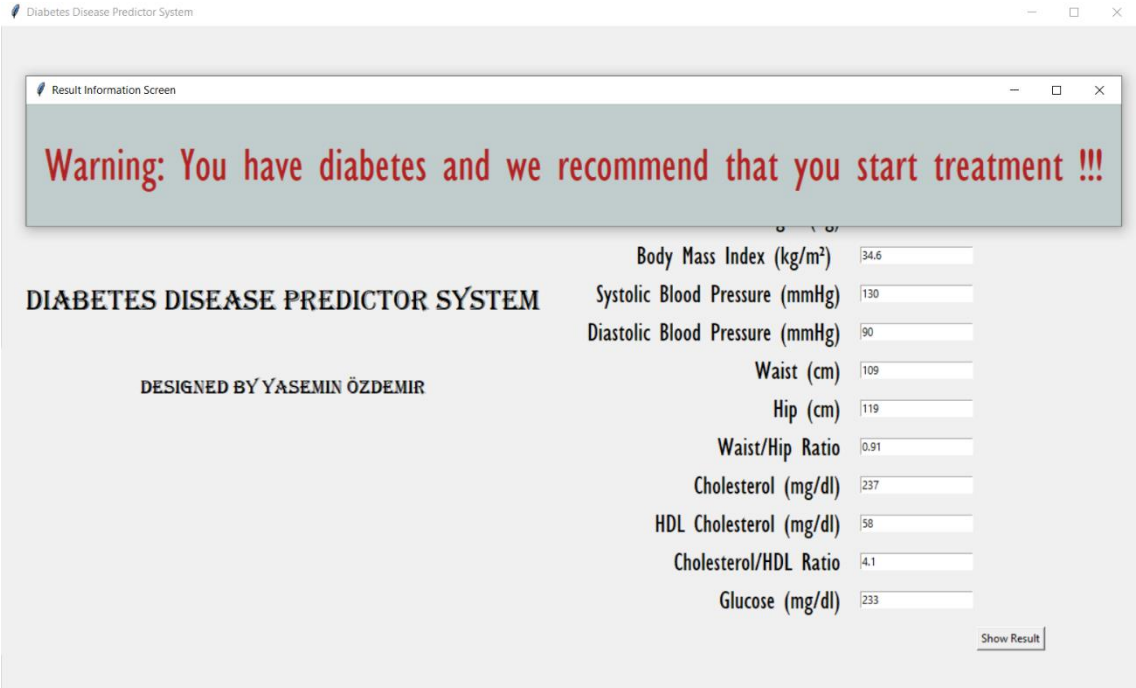
Şekil 4.14. Diyabet hastalığı tahmini arayüz sistemi gerekli bilgilerin doldurulması

Diyabet hastalığı tahmin sistemi arayüz sistemine daha önceden sistemin performansını test etmek amacıyla ayrılan veri kayıtları girilerek modelin başarısı test edilmiştir. Diyabet hastalığı tahmin sistemi arayüz sistemi uyarı ekranı Şekil 4.15'de gösterilmiştir.



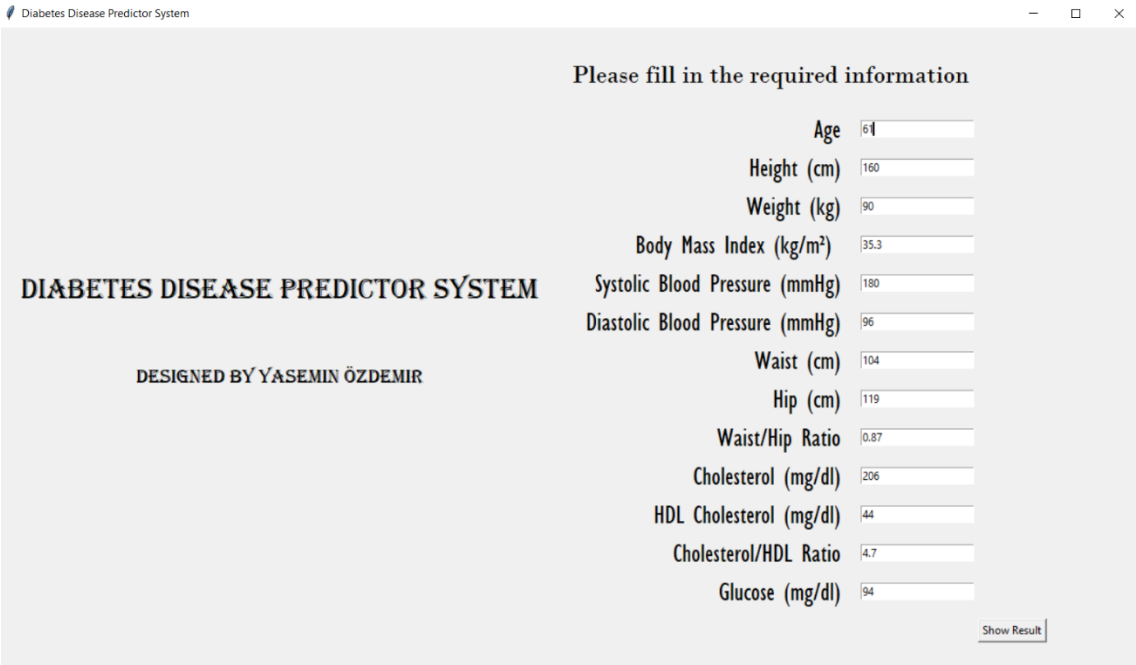
Şekil 4.15. Diyabet hastalığı tahmini arayüz sistemi eksik bilgi uyarı ekranı

Diyabet hastalığının tahmininde en önemli 13 öznitelikten oluşan arayüz sistemi Şekil 4.15'de gösterilmiştir. Hastalığın tahmininde kullanılan öznitelikler oldukça önemlidir. Bu nedenle arayüz tasarımında kullanıcıları gerekli tüm bilgileri eksiksiz doldurmaları konusunda uyarabilmek adına bir uyarı sistemi oluşturulmuştur. Kullanıcı, arayüz sistemine eksik bilgi girişi yaptığında ve sonucu göster butonuna bastığında "Uyarı: Eksik bilgi girdiniz. Lütfen tüm bilgileri doldurun." şeklinde bir uyarı almaktadır. Arayüz sistemi diyabet hastalığı sonuç bilgisi ekran görüntüsü Şekil 4.16'da, gösterilmiştir.



Şekil 4.16. Diyabet hastalığı tahmini arayüz sistemi sonuç ekranı

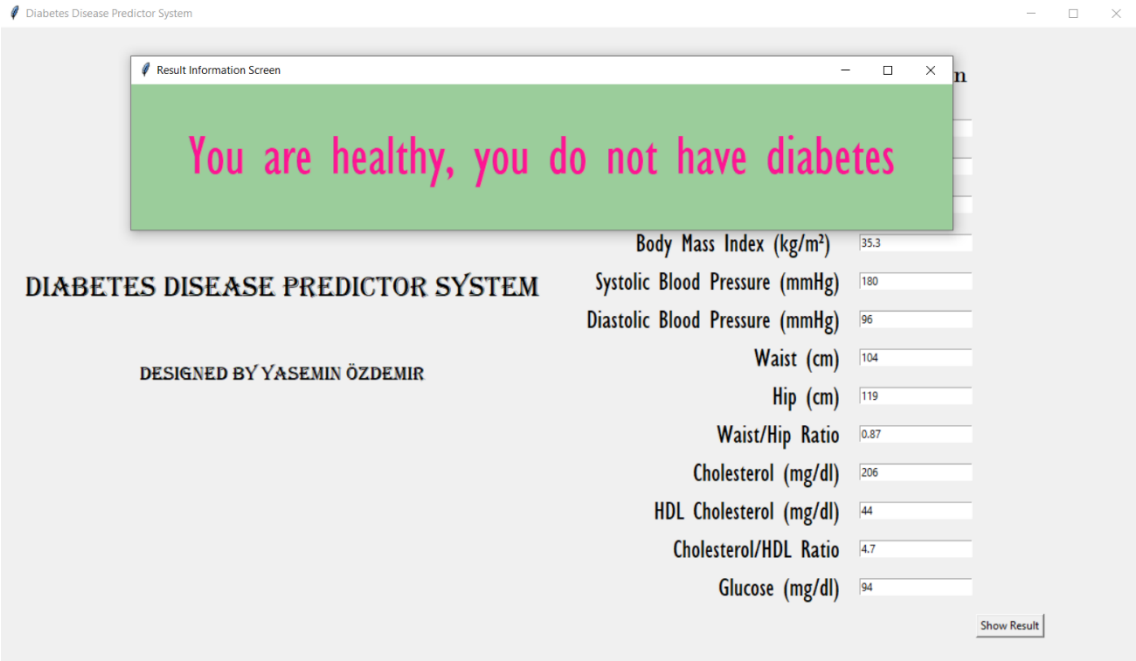
Sonuç bilgi ekranında, girişi yapılan verinin diyabet hastası olduğu belirlenmiştir. “Uyarı: Diyabet hastasısınız ve tedaviye başlamanızı öneririz.” yazısıyla sistem kullanıcıya uyarı vermektedir. Şekil 4.17’de, Diyabet hastalığı tahmini arayüz sisteminde gerekli bilgilerin doldurulması ekran görüntüsü gösterilmiştir.



Şekil 4.17. Diyabet hastalığı tahmini arayüz sisteminde gerekli bilgilerin doldurulması

Kullanıcı diyabet hastalık tahmin sistemine gerekli bilgileri girildikten sonra sonuç göster butonuna basıldığında sonuç bilgi sistemi ekranı açılmaktadır ve sonuç bilgi ekranına kullanıcı erişmektedir.

Diyabet Hastalığı Tahmininde Rassal orman algoritması hastalık tahmininde önemli bir başarı oranına sahiptir. Şekil 4.18'de, Diyabet hastalığı tahmini arayüz sistemi sonuç ekranı gösterilmiştir.



Şekil 4.18. Diyabet hastalığı tahmini arayüz sistemi sonuç ekranı

Sonuç bilgi ekranında, girişi yapılan verinin diyabet hastası olmadığı belirlenmiştir. “Sağlıklısın, diyabet hastası değilsin.” yazısıyla sistem kullanıcıya uyarı vermektedir.

Sonuçlara bakıldığında rassal orman algoritmasının kullanıcının sağlık durumu hakkında doğru tahminler yapma yeteneği görülmektedir. Yüksek doğruluk oranı, rassal orman algoritmasının arayüz tasarımı üzerindeki etkileyici performansını ve yapay zeka tekniklerindeki önemini göstermektedir.

5. SONUÇ

Çalışmanın amacı sağlık alanında yaygın görülen hastalıkların erken tahmini için literatüre katkı sağlamaktır. Kalp hastalığı ve diyabet hastalığının erken tahmininde yapay zeka tekniklerinin kullanımı incelenmiştir. Çalışma kapsamında, kalp hastalığı için UCI veri tabanından Cleveland veri seti ve diyabet hastalığı için Vanderbilt Üniversitesi Biyoistatistik Bölümünden alınan veri seti kullanılmıştır. Tüm analizler Python programlama dili kullanılarak gerçekleştirilmiştir.

Çalışmada tahmin algoritmaları olarak Rassal orman algoritması, Lojistik regresyon ve XGBoost kullanılmıştır. Kalp hastalığının erken tahmininde rassal orman algoritması kullanılarak modelin doğruluk performansı %93,00, Lojistik regresyon ile %94,00 ve XGBoost algoritması ile %95,00 başarı oranı elde edilmiştir.

Diyabet hastalığı veri setinde 13 adet eksik veri bulunmaktadır ve uygun yöntemlerle eksik veriler doldurulmuştur. Model performansını değerlendirmek amacıyla veri setinden hasta olmayan 7 veri ve hasta olan 2 veri olmak üzere toplam 9 hasta verisi ayrılmıştır. Veri setinde hasta olmayan kişilerin sayısı, hasta olan kişi sayısından oldukça fazladır. Diyabet hastalığı veri setinde dengesiz veri problemini çözmek amacıyla 147 yapay veri üretilerek veri seti daha dengeli hale getirilmiştir. Eksik verilerin uygun yöntemle doldurulması, en iyi hiperparametrelerin belirlenmesi ve dengesiz veri setinin daha dengeli hale getirilmesiyle veri ön işlemenin önemi belirlenmiştir.

Diyabet hastalığının erken tahmininde Rassal orman algoritması ile %99,00, Lojistik regresyon ile %96,00 ve XGBoost algoritması ile %98,00 başarı oranı elde edilmiştir. Rassal orman algoritması kullanılarak 1 veriyi yanlış tahmin ederek , %99,00 başarı oranı elde edilmiştir.

Çalışmanın devamında algoritmaların performansı 5 katlı ve 10 katlı çapraz doğrulama kullanılarak değerlendirilmiştir. Kalp hastalığının erken tahmininde 5 katlı çapraz doğrulama kullanılarak rassal orman algoritması ile %96,00, lojistik regresyon ile %95,00

ve XGBoost ile %95,00 doğruluk oranı elde etmiştir. 10 katlı çapraz doğrulamada ise Rassal orman algoritması ile %97,00, Lojistik regresyon ile %95,00 ve XGBoost ile %95,00 doğruluk oranlarına ulaşmıştır.

Diyabet hastalığının erken tahmininde 5 katlı çapraz doğrulama kullanılarak Rassal orman algoritması %96,00, Lojistik regresyon %95,00 ve XGBoost %96,00 doğruluk oranları elde etmiştir. 10 katlı çapraz doğrulama kullanılarak Rassal orman algoritması %97,00, Lojistik regresyon %95,00 ve XGBoost %96,00 doğruluk oranları elde edilmiştir.

Kalp hastalığı erken tahmini için aynı veri setiyle elde edilen literatür sonuçları tez çalışmasından elde edilen sonuçlarla karşılaştırılmıştır. Kalp hastalığının tahmininde Çizelge 4.12., Çizelge 4.14., Çizelge 4.16 sonuçları ve diyabet hastalığı için ise Çizelge 4.13., Çizelge 4.15. ve Çizelge 4.17. sonuçları incelendiğinde, tez çalışması literatürdeki diğer çalışmalardan daha yüksek başarı oranı göstermektedir. Çizelge 4.18, kalp hastalığı tahmininde yapay zeka teknikleri kullanılarak elde edilen literatür sonuçlarını içermektedir. Aynı şekilde, Çizelge 4.19'da Diyabet hastalığı tahmininde yapay zeka teknikleri kullanılarak elde edilen literatür sonuçları yer almaktadır

Çalışmanın devamında, kalp hastalığı tahmini için XGBoost algoritmasının en uygun hiperparametre değerleri Şekil 4.9 ve diyabet hastalığı tahmini için rastgele orman algoritmasının en uygun hiperparametre değerleri Şekil 4.12 kullanılarak bir kullanıcı arayüzü sistemi geliştirilmiştir. Diyabet hastalığı arayüz sisteminde, veri setinden çıkarılan 9 hasta kaydı ile modelin başarısı test edilmiştir. Şekil 15 ve Şekil 17'de gösterildiği gibi, 9 kişinin sağlık durumunu doğru bir şekilde tahmin edilmiştir. Kullanıcı arayüzünde, eksik verilerin ve yanlış bilgi girişlerinin önüne geçmek için uyarı sistemi oluşturulmuştur. Geliştirilen arayüz tasarımı hızlı sonuçlar vermesi ve kullanımının kolay olması bakımından hastalıkların erken tahmininde literatüre katkı sağlamaktadır.

Çalışmanın sonucunda yapay zeka tekniklerinin sağlık alanında hastalıkların erken tahmininde yüksek performans gösterdiği belirlenmiştir. Elde edilen sonuçlara göre

yapay zeka tekniklerinin sağlık alanındaki kullanımı ve öneminin gelecekte daha da artabileceği beklenmektedir.

KAYNAKLAR

Abd, D., Alwan, J. K., Ibrahim, M., & Naeem, M. B. (2017, March). The utilisation of machine learning approaches for medical data classification and personal care system management for sickle cell disease. 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT) (pp. 213-218). IEEE. https://doi.org/10.1109/NTICT.2017.7976147

Abualdenien, J., & Borrmann, A. (2022). Ensemble-learning approach for the classification of Levels of Geometry (LOG) of building elements. *Advanced Engineering Informatics*, *51*, 101497. https://doi.org/10.1016/j.aei.2021.101497

Akyol, K., & Karacı, A. (2021). Diyabet hastalığının erken aşamada tahmin edilmesi için makine öğrenme algoritmalarının performanslarının karşılaştırılması. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, *9*(6), 123-134.

Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., Alrashed, S., & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Computers in biology and medicine*, *109*, 101-111.

Anbuselvan, P. (2020). Heart disease prediction using machine learning techniques. *International Journal of Engineering Research & Technology*, *9*, 515-518.

Ani, R., Augustine, A., Akhil, N. C., & Deepa, O. S. (2016). Random forest ensemble classifier to predict the coronary heart disease using risk factors. In *Proceedings of the International Conference on Soft Computing Systems: ICSCS 2015, Volume 1* (pp. 701-710). Springer India.

Akşehir, Z. D., & Kılıç, E. (2019). Makine öğrenmesi teknikleri ile banka hisse senetlerinin fiyat tahmini. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *12*(2), 30-39.

Asif, S., Wenhui, Y., Tao, Y., Jinhai, S., & Jin, H. (2021, May). An ensemble machine learning method for the prediction of heart disease. Proceedings of the 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 98-103). IEEE. https://doi.org/10.1109/ICAIBD51990.2021.9459010

Ayyadevara, V. K. (2018). Pro machine learning algorithms. Apress: Berkeley, CA, USA.

Balasubramanian, S., Kashyap, R., CVN, S. T., & Anuradha, M. (2020, December). Hybrid prediction model for type-2 diabetes with class imbalance. *2020 IEEE international conference on machine learning and applied network technologies (ICMLANT)* (pp. 1-6). IEEE.

Balci, O. (1998). Verification, validation, and testing. *Handbook of simulation*, *10*(8), 335-393.

Baviskar, V., Verma, M., Chatterjee, P., & Singal, G. (2023). Efficient Heart Disease Prediction Using Hybrid Deep Learning Classification Models. *IRBM*, *44*(5), 100786.

Bertsimas, D., Mingardi, L., & Stellato, B. (2021). Machine learning for real-time heart disease prediction. *IEEE Journal of Biomedical and Health Informatics*, *25*(9), 3627-3637.

Berrar, D. (2019). Cross-Validation.

Bilgili, A. (2022). Sağlık İşletmelerinde Büyük Veri ve Makine Öğrenmesi Kullanımı. *Disiplinlerarası Boyutlarıyla Hizmet İşletmeleri*, 147.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media.

Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), 4514-4523.

Ceh-Varela, E., Maes, L., & Shakya, S. R. (2024). Machine Learning Analysis of Factors Contributing to Diabetes Development. *Cloud Computing and Data Science*, 157-182. https://doi.org/10.37256/ccds.5120243751

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*, 8869-8879.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., & Zhou, T. (2015). Xgboost: Extreme gradient boosting. *R Package Version 0.4-2*. (1)(4), 1-4.

Choi, B. G., Rha, S. W., Kim, S. W., Kang, J. H., Park, J. Y., & Noh, Y. K. (2019). Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. *Yonsei Medical Journal*, *60*(2), 191.

Cherry, K. (2010). *The everything psychology book: An introductory guide to the science of human behavior*. Simon and Schuster.

Chubb, H., & Simpson, J. M. (2012). The use of Z-scores in paediatric cardiology. *Annals of Pediatric Cardiology*, *5*(2), 179.

Coşansu, G. (2015). Diyabet: Küresel bir salgın hastalık. *Okmeydanı Tıp Dergisi*, *31*, 1-6.

Coşar, M., & Deniz, E. (2021). Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (28), 1112-1116.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *20*(2), 215-232.

Curtis, A. E., Smith, T. A., Ziganshin, B. A., & Elefteriades, J. A. (2016). The mystery of the Z-score. *Aorta*, *4*(04), 124-130.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

Davenport, T. H. (2018). *The AI advantage: How to put the artificial intelligence revolution to work*. MIT Press.

Deepika, K., & Seema, S. (2016, July). Predictive analytics to prevent and control chronic diseases. In *2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT)* (pp. 381-386). IEEE.

Dissanayake, K., & Johar, M. G. M. (2023). Two-level boosting classifiers ensemble based on feature selection for heart disease prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, *32*(1), 381-391.

Doki, S., Devella, S., Tallam, S., Gangannagari, S. S. R., Reddy, P. S., & Reddy, G. P. (2022, August). Heart disease prediction using XGBoost. *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)* (pp. 1317-1320). IEEE.

Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, *29*, 685-693.

Edition, S. (2015). IDF diabetes atlas. *Int Diabetes Fed*.

Ekrem, Ö., Salman, O. K. M., Aksoy, B., & İnan, S. A. (2020). Yapay Zekâ Yöntemleri Kullanılarak Kalp Hastalığının Tespiti. Mühendislik Bilimleri ve Tasarım Dergisi, 8(5), 241-254.

Farzana, S., & Veeraiah, D. (2020, October). Dynamic heart disease prediction using multi-machine learning techniques. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)* (pp. 1-5). IEEE.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Gill, S., & Pathwar, P. (2022). Prediction of Diabetes Using Various Feature Selection and Machine Learning Paradigms. In *Modern Approaches in Machine Learning & Cognitive Science: A Walkthrough* (pp. 133-146). Cham: Springer International Publishing.

Gupta, C., Saha, A., Reddy, N. S., & Acharya, U. D. (2022). Cardiac Disease Prediction using Supervised Machine Learning Techniques. In *Journal of Physics: Conference Series* (Vol. 2161, No. 1, p. 012013). IOP Publishing.

Güzel, Ş., Dömbekci, H. A., & Eren, F. (2022). Yapay zekânın sağlık alanında kullanımı: Nitel bir araştırma. *Celal Bayar Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi*, *9*(4), 509-519.

Harjai, S., & Khatri, S. K. (2019). An Intelligent Clinical Decision Support System Based on Artificial Neural Network for Early Diagnosis of Cardiovascular Diseases in Rural Areas. *In Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI)*, Dubai, United Arab Emirates, February 4–6, 2019 (pp. 729–736).https://doi.org/10.1109/AICAI.2019.8701237

Hassan, M. R., Islam, M. F., Uddin, M. Z., Ghoshal, G., Hassan, M. M., Huda, S., & Fortino, G. (2022). Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence. *Future Generation Computer Systems*, *127*, 462-472.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: Springer.

He, H., & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications.* New Jersey: The Institute of Electrical and Electronics Engineers.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression.* John Wiley & Sons.

Hung, M., Voss, M. W., Rosales, M. N., Li, W., Su, W., Xu, J., Bounsanga, J., Ruiz-Negrón, B., Lauren, E., & Licari, F. W. (2019). Application of machine learning for diagnostic prediction of root caries. *Gerodontology*, *36*(4), 395-404.

Islam, S., Jahan, N., & Khatun, M. E. (2020, March). Cardiovascular disease forecast using machine learning paradigms. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 487-490). IEEE.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: Springer.

Kalan, I., & Yeşil, Y. (2010). Obezite ile ilişkili kronik hastalıklar. *Diyabet ve obezite*, *78*.

Karaduman, T. (2019). *Yapay Zekâ Uygulama Alanları*. Gazi Universitesi, Bilisim Enstitusu, Adli Bilisim. ABD.

Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*, *18*(1/2), 90-100.

Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. *2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333). IEEE.

Kohli, P. S., & Arora, S. (2018, December). Application of machine learning in disease prediction. *2018 4th International conference on computing communication and automation (ICCCA)* (pp. 1-4). IEEE.

Kukar, M., Gunčar, G., Vovko, T., Podnar, S., Černelč, P., Brvar, M., Zalaznik, M., Notar, M., Moškon, S., & Notar, M. (2021). COVID-19 diagnosis by routine blood tests using machine learning. *Scientific Reports*, *11*(1), 10738.

Kurt, A., Buldu, B., & Cedimoğlu, İ. H. (2020). XGBoost ve rastgele orman algoritmalarının ağ tabanlı saldırı tespitine yönelik performanslarının karşılaştırılması. International Marmara Sciences Congress (pp. 730-736).

Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S. (2019, October). Prediction of coronary heart disease using supervised machine learning algorithms. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (pp. 367-372). IEEE.

Lakshmanarao, A., Swathi, Y., & Sundareswar, P. S. S. (2019). Machine learning techniques for heart disease prediction. *Forest*, *95*(99), 97.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1-45.

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access*, *8*, 107562-107582.

Lin, Y. (2021, April). Prediction and analysis of heart disease using machine learning. *2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI)* (pp. 53-58). IEEE.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Liza, F. R., Samsuzzaman, M., Azim, R., Mahmud, M. Z., Bepery, C., Masud, M. A., & Taha, B. (2021, December). An Ensemble Approach of Supervised Learning Algorithms and Artificial Neural Network for Early Prediction of Diabetes. In *2021*

3rd International Conference on Sustainable Technologies for Industry 4.0 (STI) (pp. 1-6). IEEE. https://doi.org/10.1109/STI53101.2021.9732413

Lodha, P., Talele, A., & Degaonkar, K. (2018, August). Diagnosis of Alzheimer's disease using machine learning. *2018 fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1-4). IEEE.

Lukmanto, R. B., Nugroho, A., & Akbar, H. (2019). Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. *Procedia Computer Science*, *157*, 46-54.

Madhav, G., & Goel, S. (2022, November). Type 2 Diabetes Prediction Using Machine Learning and Validation Using Weka Tool. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3* (pp. 297-310). Singapore: Springer Nature Singapore.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, *9*(1), 381-386.

Mijwil, M. M., & Aljanabi, M. (2023). A Comparative Analysis of Machine Learning Algorithms for Classification of Diabetes Utilizing Confusion Matrix Analysis. *Baghdad Science Journal*.

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

Vasanthi, R., M.E.,& Tamilselvi, J. (2022). Heart Disease Prediction Using Random Forest Algorithm. *Cardiometry*, (24), 982-988.

Viswanatha, V., Ramachandra, A. C., Murthy, D., & Thanishka, T. (2023). Diabetes prediction using machine learning approach. Strad Research, 10(8). https://doi.org/10.37896/sr10.8/008

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press.

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media, Inc.

Nadesh, R. K., & Arivuselvan, K. (2020). Type 2: diabetes mellitus prediction using deep neural networks classifier. *International Journal of Cognitive Computing in Engineering*, *1*, 55-61.

Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, *69*, 132-142.

Nielsen, D. (2016). *Tree boosting with XGBoost- Why Does XGBoost Win "every" machine learning competition?* (Master's thesis, NTNU).

Ogunleye, A., & Wang, Q. G. (2019). XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, *17*(6), 2131-2140.

Okumus, H., & Nuroglu, F. M. (2021). A random forest-based approach for fault location detection in distribution systems. *Electrical Engineering*, *103*(1), 257-264.

Olatunji, S. O., Alotaibi, S., Almutairi, E., Alrabae, Z., Almajid, Y., Altabee, R., Altassan, M., Basheer Ahmed, M. I. B., Farooqui, M., & Alhiyafi, J. (2021). Early diagnosis of thyroid cancer diseases using computational intelligence techniques: A case study of a Saudi Arabian dataset. Computers in Biology and Medicine, 131, 104267. https://doi.org/10.1016/j.compbiomed.2021.104267

Owen, L. (2022). *Hyperparameter Tuning with Python: Boost your machine learning model's performance via hyperparameter tuning*. Packt Publishing Ltd.

Özmen, Ö., Khdr, A., & Avcı, E. (2018). Sınıflandırıcıların kalp hastalığı verileri üzerine performans karşılaştırması. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 30(3), 153-159.

Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. *2008 IEEE/ACS international conference on computer systems and applications* (pp. 108-115). IEEE.

Pathak, S., & Shrestha, B. (2020). Financial Crime Detection: A Machine Learning Approach to Anomalous Financial Transaction Detection using SWIFT synthetic dataset.

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering*, *208*, 109520.

Peng, G. C. Y., Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W. W., Perdikaris, P., Petzold, L., & Kuhl, E. (2021). Multiscale modeling meets machine learning: What can we learn?. *Archives of Computational Methods in Engineering*, *28*, 1017-1037.

Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, A. W. (2021). An integrated machine learning framework for effective prediction of cardiovascular diseases. *IEEE Access*, *9*, 106575-106588.

Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, *1*, 100032.

Ramesh, A. N., Kambhampati, C., Monson, J. R. ve Drew, P. J. (2004). Artificial intelligence in medicine. Annals of the Royal College of Surgeons of England, 86(5), 334-338.

Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science*, 132-148.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Ren, Q., Li, M., & Han, S. (2019). Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. *Big Earth Data*, *3*(1), 8-25.

Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using Naives Bayesian. *2019 3rd International conference on trends in electronics and informatics (ICOEI)* (pp. 292-297). IEEE.

Ross, S. M. (2017). *Introductory statistics*. Academic Press.

Ryzhikova, E., Ralbovsky, N. M., Sikirzhytski, V., Kazakov, O., Halamkova, L., Quinn, J., Zimmerman, E.A., & Lednev, I. K. (2021). Raman spectroscopy and machine learning for biomedical applications: Alzheimer's disease diagnosis based on the analysis of cerebrospinal fluid. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *248*, 119188.

Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems*, *2022*.

Saha, P., Sadi, M. S., & Islam, M. M. (2021). EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers. *Informatics in medicine unlocked*, *22*, 100505.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018, March). Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 910-914). IEEE.

Shams, M. Y., Elshewey, A. M., El-Kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2024). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, *83*(12), 35307-35334.

Singh, J., Bagga, S., & Kaur, R. (2020). Software-based prediction of liver disease with feature selection and classification techniques. *Procedia Computer Science*, *167*, 1970-1980.

Sormunen, H. M. (2022). *Enhancing Customer Feedback Processing with Machine Learning in Microsoft Azure.* Microsoft Press.

Srinivas, P., & Katarya, R. (2022). hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. *Biomedical Signal Processing and Control*, *73*, 103456.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, *18*(10), 1099-1104.

Taşçı M. E., & Şamlı, R. (2020). Veri madenciliği ile kalp hastalığı teşhisi. *Avrupa Bilim ve Teknoloji Dergisi*, 88-95.

Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, *167*, 706-716.

Tiwari, P., & Singh, V. (2021). Diabetes disease prediction using significant attribute selection and classification approach. Journal of Physics: Conference Series, 1714(1), 012013. IOP Publishing.

Thorn, J. (2020). Logistic Regression Explained. Medium. URL: https://towardsdatascience.com/logisticregression-explained-9ee73cede081 (Date of access: 02.02.2023).

TÜİK. (2021). Ölüm ve Ölüm Nedeni İstatistikleri. https://data.tuik.gov.tr/Bulten/Index?p=Olum-ve-Olum-Nedeni-Istatistikleri-202145715 (Ziyaret tarihi: 16 Mart 2024).

Tyagi, A., Mehra, R., & Saxena, A. (2018, December). Interactive thyroid disease prediction system using machine learning technique. In *2018 Fifth international conference on parallel, distributed and grid computing (PDGC)* (pp. 689-693). IEEE.

UCI Irvine Machine Learning Repository. (Erişim tarihi: 10 Mart 2023). https://archive.ics.uci.edu/dataset/45/heart+disease

Uddin, K. M. M., Ripa, R., Yeasmin, N., Biswas, N., & Dey, S. K. (2023). Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset. *Intelligence-Based Medicine*, *7*, 100100.

Ustundag, A., Sivri, M. S., & Menguc, K. (2022). Feature Engineering. In *Business Analytics for Professionals* (pp. 153-169). Cham: Springer International Publishing.

Vanderbilt Department of Biostatistics. (Erişim tarihi: 10 Mart 2022). https://hbiostat.org/data/

Vatansever, B., Aydın, H., & Çetinkaya, A. (2021). Heart Disease Prediction with Machine Learning Algorithm Using Feature Selection by Genetic Algorithm. *Bilim, Teknoloji ve Mühendislik Araştırmaları Dergisi*, *2*(2), 67-80.

Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., Zhu, S., & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine*, *17*, 1-13.

Veranyurt, Ü., Deveci, A., Esen, M. F., & Veranyurt, O. (2020). Makine Öğrenmesi Teknikleriyle Hastalık Sınıflandırması: Random Forest, K-Nearest Neıghbour Ve Adaboost Algoritmaları Uygulaması. *Uluslararası Sağlık Yönetimi ve Stratejileri Araştırma Dergisi*, *6*(2), 275-286.

Wachs, S. (2009). What is a standard deviation and how do i compute it. *Integral Concepts, Inc*.

Wang, W., Chakraborty, G., & Chakraborty, B. (2020). Predicting the risk of chronic kidney disease (CKD) using machine learning algorithm. *Applied Sciences*, *11*(1), 202.

Wan, X., Wang, W., Liu, J., & Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology*, *14*, 1-13.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PloS one*, *12*(4), e0174944.

Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE 12*(4):e0174944

Whitmire, C. D., Vance, J. M., Rasheed, H. K., Missaoui, A., Rasheed, K. M., & Maier, F. W. (2021). Using machine learning and feature selection for alfalfa yield prediction. *AI*, *2*(1), 71-88.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, *63*(8), 826-833.

World Health Organization. (2018). *Noncommunicable diseases country profiles 2018.*

Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, *4*(1), 23–45. https://doi.org/10.1080/21693277.2016.1192517

Yıldız, A. (2014). Kurumsal Yönetim Endeksi ve Altman Z Skoruna Dayalı Lojistik Regresyon Yöntemiyle Şirketlerin Kredi Derecelendirilmesi. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *19*(3), 71-89.

Zeng, M. (2023). The Prediction of Heart Failure based on Four Machine Learning Algorithms. *Highlights in Science, Engineering and Technology*, *39*, 1377-1382.

Zimmet, P., Alberti, K. G. M. M., & Shaw, J. (2001). Global and societal implications of the diabetes epidemic. *Nature*, *414*(6865), 782-787.

EKLER

EK 1 Kalp hastalığı tahmini için rassal orman algoritması deney sonuçları
EK 2 Diyabet hastalığı tahmini için rassal orman algoritması deney sonuçları
EK 3 Kalp hastalığı tahmini için lojistik regresyon algoritması deney sonuçları
EK 4 Diyabet hastalığı tahmini için lojistik regresyon algoritması deney sonuçları
EK 5 Kalp hastalığı tahmini için XGBoost algoritması deney sonuçları
EK 6 Diyabet hastalığı tahmini için XGBoost algoritması deney sonuçları

EK Çizelge 1. Kalp hastalığı tahmini için rassal orman algoritması deney sonuçları

	Deney 1	Deney 2	Deney 3	Deney 4	Deney 5	Deney 6	Deney 7
Kullanılmayan Öznitelikler	chol	chol	chol,cp	Var	trestbps	cp	cp
Eğitim Oranı	%70	%70	%80	%80	%80	%70	%80
Test Oranı	%15	%15	%10	%10	%12	%15	%10
Doğrulama Oranı	%15	%15	%10	%10	%8	%15	%10
Random_state	19	10	10	1905	19	10	1905
Estimators	100	100	100	100	100	100	100
Oob_score	True	True	True	False	True	True	True
Bootstrap	True	True	True	True	True	True	True
Criterion	entropy	entropy	entropy	entropy	entropy	entropy	entropy
Max_features	sqrt	sqrt	sqrt	Sqrt	sqrt	sqrt	sqrt
Max_depth	none	none	150	None	150	150	150
N_jobs	-1	0	-1	-1	-1	-1	-1
Min_samples_leaf	2	3	6	10	9	2	7
Min_samples_split	2	3	2	2	5	2	2
Doğruluk Oranı	%89	%91	%87	%93	%92	%89	%90

	Deney 8	Deney 9	Deney 10	Deney 11	Deney 12	Deney 13	Deney 14
Kullanılmayan Öznitelikler	fbs	fbs	cp,fbs	chol,cp	chol,cp	chol,fbs	Chol, fbs, cp
Eğitim Oranı	%70	%80	%80	%80	%70	%80	%70
Test Oranı	%15	%10	%10	%10	%15	%10	%15
Doğrulama Oranı	%15	%10	%10	%10	%15	%10	%15
Random_state	19	1905	19	50	10	1905	50
Estimators	100	100	100	100	100	100	100
Oob_score	True	True	True	False	True	False	True
Bootstrap	True	True	True	True	True	True	True
Criterion	Entropy	entropy	entropy	entropy	entropy	entropy	entropy
Max_features	sqrt	Sqrt	Sqrt	Sqrt	Sqrt	Sqrt	Sqrt
Max_depth	150	150	150	150	150	150	150
N_jobs	-1	-1	-1	-1	-1	-1	-1
Min_samples_leaf	4	2	2	6	6	4	5
Min_samples_split	2	2	2	2	2	2	2
Doğruluk Oranı	%87	%90	%90	%87	%89	%90	%87

EK Çizelge 1. Kalp hastalığı tahmini için rassal orman algoritması deney sonuçları (devamı)

	Deney 15	Deney 16	Deney 17	Deney 18	Deney 19	Deney 20
Kullanılmayan Öznitelikler	chol, fbs, cp	chol, fbs, cp	Var	Var	Var	fbs, restecg
Eğitim Oranı	%80	%80	%70	%80	%80	%80
Test Oranı	%10	%10	%15	%10	%14	%10
Doğrulama Oranı	%10	%10	%15	%10	%6	%10
Random_state	1905	50	50	50	19	19
Estimators	100	100	100	100	100	100
Oob_score	True	True	True	True	False	True
Bootstrap	True	True	True	True	True	True
Criterion	Entropy	entropy	entropy	entropy	Gini	Entropy
Max_features	Sqrt	Sqrt	sqrt	Sqrt	Auto	Sqrt
Max_depth	150	150	none	None	150	None
N_jobs	-1	-2	-1	-1	-1	-1
Min_samples_leaf	2	8	3	3	6	8
Min_samples_split	2	2	2	2	2	2
Doğruluk Oranı	%87	%90	%87	%90	%90	%90

EK Çizelge 2. Diyabet hastalığı tahmini için rassal orman algoritması deney sonuçları

	Deney 1	Deney 2	Deney 3	Deney 4	Deney 5	Deney 6	Deney 7
Kullanılmayan Öznitelikler	Var	Var	Var	Var	Var	Var	Var
Eğitim Oranı	%80	%80	%70	%80	%70	%70	%70
Test Oranı	%10	%10	%15	%16	%21	%18	%15
Doğrulama Oranı	%10	%10	%15	%4	%9	%12	%15
Random_state	2	19	19	19	2	19	12
Estimators	100	150	150	100	100	100	100
Oob_score	False	False	False	False	False	False	False
Bootstrap	True	True	True	True	True	True	True
Criterion	entropy	entropy	entropy	entropy	entropy	entropy	gini
Max_features	sqrt	Sqrt	Sqrt	Sqrt	sqrt	Sqrt	auto
Max_depth	150	100	100	100	100	100	150
N_jobs	-1	-1	-1	-1	-1	-1	-1
Min_samples_leaf	1	1	4	1	1	4	2
Min_samples_split	7	4	3	3	2	10	2
Doğruluk Oranı	%98	%96	%96	%97	%96	%95	%98

EK Çizelge 2. Diyabet hastalığı tahmini için rassal orman algoritması deney sonuçları (devamı)

	Deney 8	Deney 9	Deney 10	Deney 11	Deney 12	Deney 13	Deney 14
Kullanılmayan Öznitelikler	Cinsiyet	Cinsiyet	Cinsiyet	Cinsiyet	Boy	Boy	Diastolik Kan Basıncı
Eğitim Oranı	%70	%80	%70	%70	%70	%80	%80
Test Oranı	%18	%10	%18	%15	%15	%12	%10
Doğrulama Oranı	%12	%10	%12	%15	%15	%8	%10
Random_state	2	2	2	2	42	2	19
Estimators	100	150	100	100	100	100	100
Oob_score	False	False	False	False	False	False	False
Bootstrap	True	True	True	True	True	True	True
Criterion	entropy	Gini	gini	entropy	entropy	entropy	entropy
Max_features	sqrt	sqrt	sqrt	sqrt	sqrt	sqrt	sqrt
Max_depth	150	150	150	150	150	100	150
N_jobs	-1	-1	-1	-1	-1	-1	-1
Min_samples_leaf	2	2	2	3	4	4	2
Min_samples_split	2	2	2	3	3	11	3
Doğruluk Oranı	%98	%98	%97	%99	%95	%95	%96

	Deney 15	Deney 16	Deney 17	Deney 18	Deney 19	Deney 20
Kullanılmayan Öznitelikler	Diastolik Kan Basıncı	Diastolik Kan Basıncı	Diastolik Kan Basıncı	Var	Boy, Cinsiyet, Diastolik Kan Basıncı	Boy, Cinsiyet, Diastolik Kan Basıncı
Eğitim Oranı	%70	%70	%70	%70	%80	%70
Test Oranı	%15	%15	%18	%24	%10	%15
Doğrulama Oranı	%15	%15	%12	%6	%10	%15
Random_state	12	2	2	2	2	2
Estimators	100	100	100	100	100	100
Oob_score	False	False	False	False	False	True
Bootstrap	True	True	True	True	True	True
Criterion	entropy	entropy	entropy	Gini	entropy	Entropy
Max_features	sqrt	Sqrt	sqrt	Auto	sqrt	Sqrt
Max_depth	150	150	150	100	150	None
N_jobs	-1	-1	-1	-1	-1	-1
Min_samples_leaf	3	3	1	3	1	1
Min_samples_split	3	5	5	2	3	4
Doğruluk Oranı	%95	%98	%98	%96	%98	%98

EK Çizelge 3. Kalp hastalığı tahmini için lojistik regresyon algoritması deney sonuçları

	Deney 1	Deney 2	Deney 3	Deney 4	Deney 5	Deney 6	Deney 7
Kullanılmayan Öznitelikler	fbs	fbs	fbs	fbs	restecg	chol	chol
Eğitim Oranı	%70	%80	%80	%80	%70	%80	%80
Test Oranı	%18	%12	%10	%10	%18	%10	%10
Doğrulama Oranı	%12	%8	%10	%10	%12	%10	%10
Random_state	19	19	1905	19	19	1905	19
Max_iter	100	100	100	100	100	100	100
Penalty	L1	L2	L2	L2	L1	L2	L2
C	1	0.1	0.1	0.1	1	0.1	0.1
Solver	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear
Multi_class	Over	Over	Over	Over	Over	Over	Over
Class_weight	None	None	None	None	None	None	None
N_jobs	None	None	None	None	None	None	None
Verbose	0	0	0	0	0	0	0
L1_ratio	None	None	None	None	None	None	None
fit_intercept	True	True	True	True	True	True	True
intercept_scaling	1	1	1	1	1	1	1
dual	False	False	False	False	False	False	False
tol	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
warm_start	False	False	False	False	False	False	False
Doğruluk Oranı	%91	%94	%90	%93	%92	%90	%93

	Deney 8	Deney 9	Deney 10	Deney 11	Deney 12	Deney 13	Deney 14
Kullanılmayan Öznitelikler	fbs, restecg, chol	restecg	cp, restecg	cp, fbs	cp, fbs	fbs, chol	fbs, chol
Eğitim Oranı	%80	%70	%80	%80	%70	%70	%70
Test Oranı	%10	%18	%12	%10	%15	%21	%15
Doğrulama Oranı	%10	%12	%8	%10	%15	%9	%15
Random_state	19	19	19	19	19	19	19
Max_iter	100	100	100	100	100	100	100
Penalty	L2	L1	L1	L2	L1	L1	L2
C	0.1	1	0.1	0.1	0.1	1	1
Solver	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear
Multi_class	Over	Over	Over	Over	Over	Over	Over
Class_weight	None	None	None	None	None	None	None
N_jobs	None	None	None	None	None	None	None
Verbose	0	0	0	0	0	0	0
L1_ratio	None	None	None	None	None	None	None
fit_intercept	True	True	True	True	True	True	True
intercept_scaling	1	1	1	1	1	1	1
dual	False	False	False	False	False	False	False
tol	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
warm_start	False	False	False	False	False	False	False
Doğruluk Oranı	%93	%91	%90	%87	%90	%92	%91

EK Çizelge 3. Kalp hastalığı tahmini için lojistik regresyon algoritması deney sonuçları (devamı)

	Deney 15	Deney 16	Deney 17	Deney 18	Deney 19	Deney 20
Kullanılmayan Öznitelikler	fbs, chol	fbs, restecg	Var	fbs, restecg, cp	Var	Var
Eğitim Oranı	%70	%70	%70	%80	%80	%70
Test Oranı	%21	%21	%15	%10	%10	%18
Doğrulama Oranı	%9	%9	%15	%10	%10	%12
Random_state	19	10	10	19	500	19
Max_iter	100	100	100	100	100	100
Penalty	L1	L2	L2	L1	L2	L1
C	1	0.1	1	0.01	1	1
Solver	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear
Multi_class	Over	Over	Over	Over	Over	Over
Class_weight	None	None	None	None	None	None
N_jobs	None	None	None	None	None	None
Verbose	0	0	0	0	0	0
L1_ratio	None	None	None	None	None	None
fit_intercept	True	True	True	True	True	True
intercept_scaling	1	1	1	1	1	1
dual	False	False	False	False	False	False
tol	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
warm_start	False	False	False	False	False	False
Doğruluk Oranı	%92	%92	%91	%90	%90	%91

EK Çizelge 4. Diyabet hastalığı tahmini için lojistik regresyon algoritması deney sonuçları

	Deney 1	Deney 2	Deney 3	Deney 4	Deney 5	Deney 6	Deney 7
Kullanılmayan Öznitelikler	Var	Var	Var	Var	Cinsiyet	Cinsiyet	Cinsiyet
Eğitim Oranı	%70	%70	%70	%80	%80	%70	%70
Test Oranı	%15	%15	%15	%10	%10	%18	%18
Doğrulama Oranı	%15	%15	%15	%10	%10	%12	%12
Random_state	2	2	12	12	12	12	12
Max_iter	100	150	200	100	100	100	100
Penalty	L1	L2	L2	L1	L1	L2	L2
C	0.01	0.01	0.01	0.001	0.01	0.001	0.1
Solver	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear
Multi_class	Over	Over	Over	Over	Over	Over	Over
Class_weight	None	None	None	None	None	None	None
N_jobs	None	None	None	None	None	None	None
Verbose	0	0	0	0	0	0	0
L1_ratio	None	None	None	None	None	None	None
fit_intercept	True	True	True	True	True	True	True
intercept_scaling	1	1	1	1	1	1	1
dual	False	False	False	False	False	False	False
tol	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
warm_start	False	False	False	False	False	False	False
Doğruluk Oranı	%89	%90	%94%	%93	%91	%92	%94

	Deney 8	Deney 9	Deney 10	Deney 11	Deney 12	Deney 13	Deney 14
Kullanılmayan Öznitelikler	Cinsiyet	Cinsiyet	Boy	Boy	Cinsiyet, Boy	Cinsiyet, Boy	Boy, Cinsiyet, Diastolik Kan Basıncı
Eğitim Oranı	%80	%80	%80	%70	%80	%70	%70
Test Oranı	%10	%10	%10	%15	%10	%15	%15
Doğrulama Oranı	%10	%10	%10	%15	%10	%15	%15
Random_state	12	12	12	12	12	2	2
Max_iter	100	100	100	100	100	100	100
Penalty	L2	L1	L2	L2	L2	L2	L2
C	0.001	1	0.1	0.1	0.001	0.001	0.001
Solver	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear
Multi_class	Over	Over	Over	Over	Over	Over	Over
Class_weight	None	None	None	None	None	None	None
N_jobs	None	None	None	None	None	None	None
Verbose	0	0	0	0	0	0	0
L1_ratio	None	None	None	None	None	None	None
fit_intercept	True	True	True	True	True	True	True
intercept_scaling	1	1	1	1	1	1	1
dual	False	False	False	False	False	False	False
tol	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
warm_start	False	False	False	False	False	False	False
Doğruluk Oranı	%92	%96	%91	%90	%86	%88	%90

EK Çizelge 4. Diyabet hastalığı tahmini için lojistik regresyon algoritması deney sonuçları (devam)

	Deney 15	Deney 16	Deney 17	Deney 18	Deney 19	Deney 20
Kullanılmayan Öznitelikler	Boy, Cinsiyet, Diastolik Kan Basıncı	Boy, Cinsiyet, Diastolik Kan Basıncı	Boy, Cinsiyet, Diastolik Kan Basıncı	Boy, Cinsiyet, Diastolik Kan Basıncı	Boy, Cinsiyet, Diastolik Kan Basıncı	Cinsiyet
Eğitim Oranı	%70	%80	%80	%70	%70	%70
Test Oranı	%15	%10	%10	%15	%15	%15
Doğrulama Oranı	%15	%10	%10	%15	%15	%15
Random_state	2	2	2	2	2	12
Max_iter	100	100	100	100	100	100
Penalty	L2	L2	L1	L1	L2	L2
C	0.1	0.1	1	1	1	1
Solver	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear	Liblinear
Multi_class	Over	Over	Over	Over	Over	Over
Class_weight	None	None	None	None	None	None
N_jobs	None	None	None	None	None	None
Verbose	0	0	0	0	0	0
L1_ratio	None	None	None	None	None	None
fit_intercept	True	True	True	True	True	True
intercept_scaling	1	1	1	1	1	1
dual	False	False	False	False	False	False
tol	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
warm_start	False	False	False	False	False	False
Doğruluk Oranı	%91	%91	%94	%93	%94	%94

EK Çizelge 5. Kalp hastalığı tahmini için XGBoost algoritması deney sonuçları

	Deney 1	Deney 2	Deney 3	Deney 4	Deney 5	Deney 6	Deney 7
Öznitelik Bilgisi	Var	Var	Var	Var	fbs, restecg	trestbps	fbs
Eğitim Oranı	%70	%80	%80	%80	%80	%80	%80
Test Oranı	%15	%12	%10	%14	%14	%10	%14
Doğrulama Oranı	%15	%8	%10	%6	%6	%10	%6
random_state	19	19	19	19	19	19	19
n_estimators	100	100	100	100	100	100	100
learning_rate	0.2	0.1	0.3	0.1	0.2	0.1	0.2
max_depth	1	1	1	1	1	1	1
max_depth_step	0	0	0	0	0	0	0
subsampe	1	1	1	1	1	1	1
min_child_weight	1	2	1	5	3	3	5
base_score	0.5	0.5	0.5	0.5	0.5	0.5	0.5
booster	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree
n_jobs	None	None	None	None	None	None	None
colsample_bytree	1	1	1	1	1	1	1
gamma	0	0	0	0	0	0	0
lambda	1	1	1	1	1	1	1
Doğruluk Oranı	%91	%92	%93	%93	%93	%90	%93

EK Çizelge 5. Kalp hastalığı tahmini için XGBoost algoritması deney sonuçları (devamı)

	Deney 8	Deney 9	Deney 10	Deney 11	Deney 12	Deney 13	Deney 14
Öznitelik Bilgisi	fbs	yaş	oldpeak	fbs,chol	cinsiyet	fbs	cp
Eğitim Oranı	%80	%80	%70	%80	%70	%80	%70
Test Oranı	%16	%10	%15	%10	%15	%14	%15
Doğrulama Oranı	%4	%10	%15	%10	%15	%6	%15
random_state	19	19	19	19	19	19	19
n_estimators	100	100	100	100	100	100	100
learning_rate	0.2	0.2	0.3	0.2	0.2	0.3	0.2
max_depth	3	1	2	1	1	1	1
max_depth_step	0	0	0	0	0	0	0
subsampe	1	1	1	1	1	1	1
min_child_weight	5	5	5	1	1	1	3
base_score	0.5	0.5	0.5	0.5	0.5	0.5	0.5
booster	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree
n_jobs	None	None	None	None	None	None	None
colsample_bytree	1	1	1	1	1	1	1
gamma	0	0	0	0	0	0	0
lambda	1	1	1	1	1	1	1
Doğruluk Oranı	%94	%90	%91	%90	%89	%95	%87

EK Çizelge 5. Kalp hastalığı tahmini için XGBoost algoritması deney sonuçları (devamı)

	Deney 15	Deney 16	Deney 17	Deney 18	Deney 19	Deney 20
Öznitelik Bilgisi	fbs, restecg, chol	cp, restecg	Var	fbs, restecg	chol	oldpeak
Eğitim Oranı	%80	%80	%80	%70	%80	%70
Test Oranı	%10	%14	%14	%21	%10	%24
Doğrulama Oranı	%10	%6	%6	%9	%10	%6
random_state	19	19	19	19	19	19
n_estimators	100	100	100	100	100	100
learning_rate	0.2	0.2	0.1	0.2	0.2	0.3
max_depth	1	1	1	1	1	2
max_depth_step	0	0	0	0	0	0
Subsampe	1	1	1	1	1	1
min_child_weight	5	3	5	5	1	5
base_score	0.5	0.5	0.5	0.5	0.5	0.5
Booster	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree
n_jobs	None	None	None	None	None	None
colsample_bytree	1	1	1	1	1	1
Gamma	0	0	0	0	0	0
Lambda	1	1	1	1	1	1
Doğruluk Oranı	%90	%90	%93	%92	%90	%91

EK Çizelge 6. Diyabet hastalığı tahmini için XGBoost algoritması deney sonuçları

	Deney 1	Deney 2	Deney 3	Deney 4	Deney 5	Deney 6	Deney 7
Öznitelik Bilgisi	Var	Var	Var	Var	Var	Cinsiyet	Cinsiyet
Eğitim Oranı	%70	%70	%70	%70	%70	%70	%70
Test Oranı	%15	%15	%18	%15	%18	%15	%21
Doğrulama Oranı	%15	%15	%12	%15	%12	%15	%9
random_state	19	2	2	12	12	2	2
n_estimators	100	100	100	100	100	100	100
learning_rate	0.1	0.1	0.2	0.01	0.1	0.1	0.1
max_depth	3	3	3	3	3	5	3
subsampe	1	1	1	1	1	1	1
min_child_weight	1	5	1	2	1	6	5
base_score	0.5	0.5	0.5	0.5	0.5	0.04	0.5
booster	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree
n_jobs	None	None	None	None	None	None	None
colsample_bytree	1	1	1	1	1	1	1
gamma	0	0	0	0	0	0	0
lambda	1	1	1	1	1	1	1
Doğruluk Oranı	%94	%96	%97	%95	%96	%95	%96

EK Çizelge 6. Diyabet hastalığı tahmini için XGBoost algoritması kullanılarak elde edilen deney sonuçları (devamı)

	Deney 8	Deney 9	Deney 10	Deney 11	Deney 12	Deney 13	Deney 14
Öznitelik Bilgisi	Cinsiyet	Var	Boy	Boy	Boy	Cinsiyet, Boy	Cinsiyet, Boy
Eğitim Oranı	%80	%70	%70	%80	%70%	%70	%70
Test Oranı	%12	%15	%15	%12	%18%	%15	%15
Doğrulama Oranı	%8	%15	%15	%8	%12%	%15	%15
random_state	2	12	2	2	12	2	2
n_estimators	100	100	100	100	100	100	100
learning_rate	0.1	0.1	0.01	0.2	0.3	0.02	0.01
max_depth	9	3	3	3	5	4	5
subsampe	1	1	1	1	1	1	1
min_child_weight	1	1	7	5	1	1	3
base_score	0.5	0.5	0.5	0.5	0.5	0.5	0.5
booster	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree
n_jobs	None	None	None	None	None	None	None
colsample_bytree	1	1	1	1	1	1	1
gamma	0	0	0	0	0	0	0
lambda	1	1	1	1	1	1	1
Doğruluk Oranı	%97	%98	%93	%94	%96	%95	%96

	Deney 15	Deney 16	Deney 17	Deney 18	Deney 19	Deney 20
Öznitelik Bilgisi	Cinsiyet, Boy	Cinsiyet, Boy	Cinsiyet, Boy, Diastolik Kan Basıncı	Cinsiyet, Boy, Diastolik Kan Basıncı	Cinsiyet, Boy, Diastolik Kan Basıncı	Var
Eğitim Oranı	%80	%70	%70	%80	%80	%70
Test Oranı	%10	%18	%15	%12	%10	%18
Doğrulama Oranı	%10	%12	%15	%8	%10	%12
random_state	2	2	2	2	2	2
n_estimators	100	100	100	100	100	100
learning_rate	0.01	0.3	0.01	0.2	0.2	0.1
max_depth	3	5	5	5	3	3
subsampe	1	1	1	1	1	1
min_child_weight	1	3	1	1	5	5
base_score	0.5	0.5	0.5	0.5	0.5	0.5
booster	Gbtree	Gbtree	gbtree	gbtree	gbtree	gbtree
n_jobs	None	None	None	None	None	None
colsample_bytree	1	1	1	1	1	1
gamma	0	0	0	0	0	0
lambda	1	1	1	1	1	1
Doğruluk Oranı	%96	%97	%95	%96	%96	%95

ÖZGEÇMİŞ

Adı Soyadı	: Yasemin Özdemir
Doğum Yeri ve Tarihi	: Pendik / İstanbul 18.02.1995
Yabancı Dil	: İngilizce

Eğitim Durumu

Lise	: Kurtköy Anadolu Lisesi
Lisans	: Atatürk Üniversitesi Endüstri Mühendisliği
Yüksek Lisans	: Bursa Uludağ Üniversitesi Endüstri Mühendisliği

Çalıştığı Kurum/Kurumlar	: Bosch Türkiye – Application & IT Manufacturing Support Smart Start (2022-2023)
İletişim (e-posta)	: ozdemir.ysminn@gmail.com

Yayınları :

Özdemir, Y., Öztürk, N., An Artificial Intelligence Technique Application in the Diagnosis of Heart Disease Using Random Forest Algorithm, *8th International European Conference on Interdisciplinary Scientific Research,* July 13-15, 2023, Rome, Italy, pp.834-838.