T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DOKTORA TEZİ

SEZGİSEL HİBRİT ÖĞRENME YÖNTEMLERİ İLE SAĞLIK VERİLERİNİN ANALİZİ

Hatice NİZAM ÖZOĞUR

DANIŞMAN
Prof. Dr. Zeynep ORMAN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği, Doktora Programı

Mayıs, 2024

Eşim Gökhan ÖZOĞUR'a ithaf ediyorum…

BÜTÇE DESTEKLERİ

SEZGİSEL HİBRİT ÖĞRENME YÖNTEMLERİ İLE SAĞLIK VERİLERİNİN ANALİZİ

Bu tez çalışması için herhangi bir kurumdan bütçe desteği alınmamıştır.

TEŞEKKÜR

Doktora tez çalışmam süresince sağladığı değerli bilgi ve tecrübeleriyle bana yol gösteren değerli tez danışmanım Prof. Dr. Zeynep ORMAN'a,

Tez izleme komitemde yer alarak çalışmama fikirleri ve katkılarıyla destek olan Prof. Dr. Rüya ŞAMLI'ya ve Doç. Dr. Fatma PATLAR AKBULUT'a,

Hayatımın her aşamasında yanımda olan, maddi ve manevi olarak desteklerini hiçbir zaman esirgemeyen sevgili aileme,

Her koşulda olduğu gibi doktora çalışmalarım sırasında da her zaman bana inanıp, yanımda olan ve hedeflerime ulaşmam için beni cesaretlendiren ve güç veren yol arkadaşım sevgili eşim Gökhan ÖZOĞUR'a en içten teşekkürlerimi sunarım.

Mayıs 2024 Hatice NİZAM ÖZOĞUR

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar	Açıklama
LR	: Lojistik Regresyon
RF	: Rastgele Orman
SVM	: Destek Vektör Makinesi
DT	: Karar Ağacı
NB	: Naive Bayes
XGBoost	: Aşırı Gradyan Artırma
DNN	: Derin Sinir Ağı
MLP	: Çok Katmanlı Algılayıcı
ANN	: Yapay Sinir Ağı
FNN	: İleri Beslemeli Sinir Ağı
LightGBM	: Hafif Gradyan Arttırma Makinesi
KNN	: k-En Yakın Komşu
MICE	: Zincirlenmiş Denklemlerle Çok Değişkenli Atama
SMOTE	: Sentetik Azınlık Aşırı Örnekleme Tekniği
RUS	: Rastgele Eksik Örnekleme
ENN	: Düzenlenmiş En Yakın Komşu
ADASYN	: Uyarlanabilir Sentetik Örnekleme Yaklaşımı
PSO	: Parçacık Sürü Optimizasyonu
GA	: Genetik Algoritma
ALO	: Antlion Optimizasyonu
PID	: PIMA Indians Diyabet Veri Kümesi
SİT	: Serebral İnme Tahmini Veri Kümesi
KBH	: Kronik Böbrek Hastalığı Veri Kümesi
BMI	: Vücut Kitle İndeksi
MCC	: Matthews Korelasyon Katsayısı
AUC	: Eğri Altındaki Alan
ROC	: Alıcı İşletim Karakteristik Eğrisi
IQR	: Çeyrekler Arası Aralık

RMSE : Kök Ortalama Kare Hata

MAE : Ortalama Mutlak Hata

IQR : Çeyrekler Arası Aralık

PCA : Temel Bileşen Analizi

PCC : Pearson Korelasyon Katsayısı

k-means : k-ortalama

SGD : Stokastik Gradyan İniş

LOF : Yerel Aykırı Değer Faktörü

WHO : Dünya Sağlık Örgütü

ÖZET

DOKTORA TEZİ

SEZGİSEL HİBRİT ÖĞRENME YÖNTEMLERİ İLE SAĞLIK VERİLERİNİN ANALİZİ

Hatice NİZAM ÖZOĞUR

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği, Doktora Programı

Danışman: Prof. Dr. Zeynep ORMAN

Sağlık verilerinin analizi, hastalıkların teşhisi ve tahmini çalışmalarında kritik öneme sahiptir. Günümüzde artan veri miktarıyla birlikte araştırmacıların ve hekimlerin makine öğrenmesi yöntemleriyle tasarlanan doğru tanı sistemlerine olan talepleri açıktır. Makine öğrenmesi yöntemleri, dengeli veri kümeleri ve tam verilere dayanarak tasarlandığından genellikle dengesiz ve eksik veriler içeren sağlık veri kümelerinde hatalı sonuçlara neden olmaktadır. Bu tez çalışmasında, sınıf dengesizliği ve eksik değer problemlerini ele almak üzere hibrit bir ön işleme yöntemi geliştirilmiştir. Bu yöntem, eksik değerlerin tamamlanması için Zincirlenmiş Denklemlerle Çok Değişkenli Atama (MICE) yöntemiyle birlikte Genetik Algoritma (GA) sezgiseli kullanılarak geliştirilen GA-MICE yöntemini ve dengesiz dağılımlı sınıfların dengelemesi için Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) ve Düzenlenmiş En Yakın Komşu (ENN) eksik örnekleme yöntemini GA ve Parçacık Sürü Optimizasyon (PSO) sezgiselleriyle birleştirerek geliştirilen GASMOTEPSO_ENN yöntemini içermektedir. Önerilen yöntemin etkinliği, diyabet, inme ve böbrek hastalığı gibi önemli sağlık sorunlarının tespitinde, açık erişimli veri kümeleri üzerinde 6 farklı makine öğrenmesi sınıflandırma yöntemleriyle test edilmiştir. Elde edilen bulgulara göre, önerilen yöntem, üç veri kümesinde

%93 ile %100 arasında değişen doğruluk, kesinlik, duyarlılık, F1-skoru ve Eğri Altındaki Alan (AUC) değerleri elde etmiştir. Bu yöntem, sınıf dengesizliğini ve eksik değer sorunlarını ele almak için etkili bir şekilde çalışmış ve literatürdeki benzer yöntemlere kıyasla daha yüksek ve güvenilir sonuçlar vermiştir.

Mayıs 2024 , 101 sayfa.

Anahtar kelimeler: Dengesiz Veri Kümesi, Eksik Değer Problemi, Sezgisel Yöntemler, Hastalık Sınıflandırması, Makine Öğrenmesi

ABSTRACT

Ph.D. THESIS

ANALYSIS OF HEALTH DATA WITH HEURISTIC HYBRID LEARNING METHODS

Hatice NİZAM ÖZOĞUR

İstanbul University-Cerrahpaşa

Institute of Graduate Studies

Department of Computer Engineering

Computer Engineering Programme

Supervisor: Prof. Dr. Zeynep ORMAN

The analysis of health data holds critical importance in the diagnosis and prediction of diseases. With the increasing volume of data in today's world, there is a clear demand from researchers and healthcare professionals for accurately designed diagnostic systems using machine learning methods. However, machine learning methods, being designed based on balanced datasets and complete information, often lead to erroneous results in healthcare datasets due to their inherently imbalanced and incomplete nature. In this thesis, a hybrid preprocessing method has been developed to tackle class imbalance and missing value problems. This method includes the GA-MICE approach, which utilizes Genetic Algorithm (GA) heuristics and the Multiple Imputation by Chained Equations (MICE) method for completing missing values. Additionally, it incorporates the GASMOTEPSO_ENN method, which combines GA and Particle Swarm Optimization (PSO) heuristics with Synthetic Minority Over-Sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN) undersampling technique for balancing imbalanced class distributions. The effectiveness of the proposed method was tested on publicly available datasets for significant health issues such as diabetes, stroke, and kidney disease using six

different machine learning classification methods. The findings revealed that the proposed method achieved accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) values ranging from 93% to 100% across the three datasets. This method effectively addressed class imbalance and missing value issues, yielding higher and more reliable results compared to similar methods in the literature.

May 2024, 101 pages.

Keywords: Imbalanced Dataset, Missing Value Problem, Heuristic Methods, Disease Classification, Machine Learning

1. GİRİŞ

2020 yılında Dünya Sağlık Örgütü (WHO) tarafından yayınlanan rapora göre, 2019'da dünya çapında gerçekleşen 55,4 milyon ölümün %55'ini oluşturan ilk 10 ölüm nedeni belirlenmiştir [1]. Bu raporda, iskemik kalp hastalığının dünya genelinde toplam ölümlerin %16'sını oluşturarak ilk sırayı aldığı vurgulanmıştır. İnme ise ölümlerin yaklaşık %11'ini oluşturarak ölümlerin ikinci sırasında yer almaktadır. Böbrek hastalıkları, dünya genelinde önde gelen ölüm nedenleri arasında yer almakta olup, son yıllarda endişe verici bir artış göstererek 10. sıraya yükselmiştir. Ayrıca, diyabetin %70'lik önemli bir artıştan sonra en çok ölüme neden olan ilk 10 hastalık arasına girdiği belirtilmiştir. Bu bulgular, sağlık politikalarının yönlendirilmesinde önemli bir rehber olarak kullanılabilir. Ayrıca, makine öğrenmesi ve yapay zeka tekniklerinin sağlık alanındaki çalışmalarda uygulanmasıyla hastalıkların erken teşhisi ve tedavisi gibi alanlarda önemli gelişmeler sağlanabilir [2]. Bu da sağlık hizmetlerinin etkinliğini artırabilir ve genel halk sağlığını iyileştirebilir.

Sağlık veri kümelerinde yaygın olarak karşılaşılan dengesiz sınıf dağılımı ve eksik veriler, sağlık alanında yapılan araştırmalarda önemli zorlukları oluşturmaktadır. Eksik veri problemi, hastaların bazı özelliklerinin eksik veya yanlış kaydedilmiş olmasıyla ortaya çıkar, laboratuvar sonuçlarının eksik olması veya kayıt dışı randevular gibi durumlar bu sorunu oluşturabilir. Bu eksiklikler, veri analizi ve makine öğrenmesi modelleme süreçlerinde güvenilirliği azaltabilir ve sonuçların yanıltıcı ve yanlı olmasını tetikleyebilir. Diğer yandan, dengesiz dağılımlı sınıf problemi, özellikle nadir hastalıklar veya sağlık durumlarıyla ilgili veri kümelerinde belirgin hale gelmektedir. Örneğin; kanser gibi nadir görülen bir hastalık durumunda, hasta sayısı diğer sağlıklı bireylerin sayısına göre çok daha az olabilir. Bu durum, makine öğrenmesi algoritmalarının dengesiz sınıf dağılımı nedeniyle nadir sınıfları doğru bir şekilde tanımlama yeteneğini zayıflatabilir ve model değerlendirme metrik sonuçlarını olumsuz yönde etkileyebilir. Bu iki problem, sağlık veri kümelerinin analiz ve değerlendirilmesi süreçlerinde karşılaşılan karmaşıklığı artırır ve doğru tedavi planlaması veya hastalık risklerinin değerlendirilmesi gibi kritik karar alma süreçlerini etkileyebilir. Literatürde, sağlık alanında hastalık tespiti üzerine birçok çalışma [3-26] yapılmıştır. Bu çalışmalar, makine öğrenmesi modellerinin etkinliğinin artırılması, hastalıkların etkin bir şekilde sınıflandırılması, eksik veri

ve sınıf dengesizliği gibi önemli zorlukların üstesinden gelinmesi gibi çeşitli konulara odaklanmaktadır.

Sağlık verilerinin analizi ve tahmini çalışmaları, özel tetkikler yapılmadan hastalık belirtilerinin erken evrelerinde tespit edilmesinde, hastalıkların takibinde ve önlenmesinde önemli katkılar sağlamaktadır. Teknolojik gelişmelerin hızlı artışıyla birlikte, elektronik sağlık kayıtlarının toplanması ve saklanması giderek daha yaygın hale gelmektedir. Bu durum, sağlık alanındaki veri miktarında yaşanan hızlı artışla beraber araştırmacı ve hekimlerin makine öğrenmesi modelleriyle sağlık verilerinin analiz yöntem ve araçlarının geliştirilmesi konusundaki taleplerini artırmaktadır. Bu taleplerin temelinde, eksik değerlerin yönetimi ve dengesiz sınıf problemlerinin çözümü gibi zorlukların üstesinden gelme ihtiyacı bulunmaktadır. Bu tez çalışmasında, sağlık verilerinin analizi ve tahmini çalışmalarında karşılaşılan eksik değerlerin tamamlanması ve dengesiz dağılımlı sınıfların dengelenmesine yönelik bir ön işleme modelinin geliştirilmesi hedeflenmiştir. Tez çalışmasının kapsamı, WHO verilerine dayanarak belirlenen ölüm oranlarının başında gelen diyabet, inme ve böbrek hastalıklarının tespit edilmesine odaklanmıştır. Sağlık verilerindeki eksik değerlerin tamamlanması ve sınıfların dengelenmesi, makine öğrenmesi yöntemlerinin etkin bir şekilde uygulanabilmesi için kritik öneme sahiptir. Yapılan tez çalışması, bu zorlukları aşmak için özgün bir yaklaşım geliştirmeyi amaçlamaktadır.

Bu tez çalışmasında, sağlık veri kümelerinde eksik değerlerin doldurulması ve dengesiz dağılımlı sınıfların dengelenmesine yönelik bir hibrit ön işleme modeli geliştirilmiştir. Bu model, eksik değerlerin atanması için literatürde başarıyla kullanılan MICE yöntemini GA sezgisel yöntemi ile birleştirerek geliştirilen GA-MICE yöntemini içermektedir. Ayrıca, dengesiz sınıf dağılımıyla başa çıkmak için SMOTE aşırı örnekleme ve ENN eksik örnekleme yöntemlerinin GA ve PSO sezgisel yöntemleriyle entegre edilmesiyle geliştirilen GASMOTEPSO_ENN yöntemini de içermektedir. Tez çalışması kapsamında gerçekleştirilmiş çalışmaların temel katkıları aşağıda özetlenmiştir:

- Sağlık veri kümeleri için özel olarak tasarlanmış bir hibrit ön işleme modeli, eksik veri ve sınıf dengesizliği gibi zorluklarla başa çıkmak için literatüre sunulmuştur.

- Veri kümelerindeki eksik verilerin doldurulması için literatürde başarıyla kullanılan MICE atama yöntemi, GA sezgisel yöntemi ile geliştirilerek GA-MICE adında yeni bir atama yöntemi önerilmiştir.

- SMOTE aşırı örnekleme ve ENN eksik örnekleme yöntemlerini GA ve PSO sezgisel yöntemleriyle entegre ederek geliştirilen GASMOTEPSO_ENN adında yeni bir hibrit yeniden örnekleme yöntemi literatüre sunulmuştur.

- Hibrit ön işleme tekniği, dengesiz sınıfların düzenlenmesinde sentetik örneklerin kalitesini artırarak ve eksik değerlerin doldurulmasında veri tutarlılığını sağlayarak, veri kümelerindeki hasta ve sağlıklı bireyleri doğru bir şekilde ayırt etmek için makine öğrenmesi sınıflandırıcılarının performansını önemli ölçüde iyileştirmiştir.

- Önerilen hibrit ön işleme modelinin performansını analiz etmek için üç ayrı deney gerçekleştirilmiş ve elde edilen bulgular ayrıntılı olarak paylaşılmıştır.

- Önerilen hibrit ön işleme modeli, farklı özelliklere sahip üç sağlık veri kümesi üzerinde altı farklı makine öğrenmesi sınıflandırıcıları kullanılarak test edilmiş ve performans sonuçları, literatürdeki benzer çalışmalarla karşılaştırılmıştır.

Tezin geri kalan kısmı şu şekilde organize edilmiştir: Bölüm 2, makine öğrenmesi kavramı çerçevesinde veri analizi aşamalarını, makine öğrenmesi sınıflandırma yöntemlerini, eksik veri işleme ve sınıf dengeleme yöntemlerini, aykırı değer tespit yöntemlerini ve sezgisel yöntemleri detaylı bir şekilde ele almaktadır. Ayrıca, sağlık veri kümelerinde eksik değerlerin ve sınıf dengesizliğinin çözümü için geliştirilmiş yöntemlere yönelik literatürde yapılan çalışmaların incelemesini içermektedir. Bölüm 3'te sağlık veri kümelerinde eksik değer ve dengesiz dağılımlı sınıf sorunlarının çözümlenmesine yönelik tez kapsamında geliştirilen yeni hibrit ön işleme yöntemi açıklanmıştır. Bu bölümde, deneysel çalışmada kullanılan sağlık veri kümeleri, veri hazırlama süreci ve sınıflandırma modeli ayrıntılı olarak ele alınmıştır. Tez çalışması kapsamında gerçekleştirilen deneylerin sonuçları ve önerilen hibrit ön işleme modelinin makine öğrenmesi sınıflandırma yöntemlerinin performansına etkisi Bölüm 4'te sunulmuştur. Elde edilen bulguların sonuçları Bölüm 5'te detaylı bir şekilde tartışılmıştır. Son olarak, Bölüm 6'da tez çalışmasının sonuçları değerlendirilmiş ve gelecek çalışmalar için öneriler sunulmuştur.

2. KAVRAMSAL ÇERÇEVE

Bu bölümde, makine öğrenmesi, veri analizi, makine öğrenmesi sınıflandırma algoritmaları, veri ön işleme aşamasında kullanılan eksik veri işleme teknikleri ve sınıf dengesizliği sorununu ele alan yöntemler, sezgisel yaklaşımlar, aykırı değer tespit yöntemleri ve tez çalışmasına benzer konuları ve problemleri ele alan literatür incelemeleri detaylı olarak incelenmiştir.

2.1. MAKİNE ÖĞRENMESİ

Makine öğrenmesi, bilgisayar sistemlerinin girdi verilerini kullanarak belirli çıktı tahminlerini elde edebilmeyi veri analizi ile öğrenen ve bu öğrenilen bilgileri kullanarak belirli görevleri daha iyi yapabildiği bir yapay zeka alt dalıdır. Bir başka deyişle makine öğrenmesi, bir problemin çözümünde bilgisayar programlama dillerindeki gibi kuralları açık bir şekilde tanımlamak yerine, bilgisayarların örnek veriler veya önceden öğrendiği deneyimleri kullanarak başarımlarını artıracak biçimde programlanmasıdır [27]. Makine öğrenmesi genellikle büyük miktarda veriye dayalı karmaşık problemlerin çözümünde ve insan müdahalesini minimumda tutarak otomatik kararlar almak için kullanılır. Günümüzde teknolojinin hızlı bir şekilde ilerlemesiyle beraber verinin boyutunun artması sağlık hizmetleri, finans ve bankacılık, eğitim, ulaşım ve lojistik, perakende ve e-ticaret gibi çeşitli alanlarda karşımıza çıkan bir fenomen haline gelmiştir. Örneğin; internet kullanımının artması, sosyal medya platformlarının yaygınlaşması, nesnelerin interneti (IoT) cihazlarının artması ve dijitalleşmenin ilerlemesi gibi nedenlerle veriler hızla artmaktadır. Bu büyük veri akışı ile makine öğrenmesi yöntemleri kullanılarak kullanıcı davranışları analiz edilebilir ve anlamlı bilgiler çıkarılarak kişiselleştirilmiş içerik sunumu, reklam önerileri, çevresel izleme, enerji verimliliği gibi çeşitli işlemler için tahminlerde bulunulabilir.

Büyük miktarda veriye dayalı olarak makine öğrenmesi modellerinin geliştirilmesinde veri analizi oldukça önemlidir. Makine öğrenmesi projelerinde veri analizi süreci, verileri toplama, temizleme, dönüştürme ve görselleştirme aşamalarını içerir. Bu aşamalar, verinin anlaşılması ve işlenmesi için temel adımları oluşturur. Ardından, makine öğrenmesi aşaması gelir, burada veriler bu modellere beslenir ve modeller, verilerden öğrenme sürecini gerçekleştirir, desenleri tanır ve sonuçlar üretir. Makine öğrenmesi ve veri analizi yaklaşımları özellikle karmaşık

problemleri çözmek, tahminlerde bulunmak ve verilerden değerli tahminler elde etmek için birlikte kullanılır.

2.2. VERİ ANALİZİ

Veri, sayı, metin, sembol, görüntü, video, ses veya işaret gibi çeşitli biçimlerde olan bilgi parçacıklarıdır. Veriler tek başına bir anlam ifade etmezler. Verilerin bir anlam ifade ederek bilgiye dönüşmesi veri analizi ve verinin yorumlaması sonucu mümkün olmaktadır. Günümüzde, internet ve sosyal medya platformlarının yoğun bir şekilde kullanımı sonucu sürekli bir veri akışı olmaktadır. Verilerin yüksek hacimli olduğu durumlarda, gereksiz verilerin ayıklanarak yalnızca anlamlı verilerin toplanması önemlidir. Ham verilerin elde edilerek bunların yararlı bilgilere dönüştürülmesi sürecine veri analizi denilmektedir. Veri analizinde bilgisayar sistemlerinden yararlanılarak veriler alınabilir, işlenebilir, depolanabilir ve analiz edilebilir. Veri analizi sürecinin temel adımları aşağıda yer alan alt bölümlerde açıklanmaktadır.

2.2.1. Veri Toplama

Veri toplama süreci, bir problemin çözümünde gerekli olan verilerin toplandığı ve bunların depolanacağı kaynağın belirlendiği adımdır. Veri toplama aşamasında veriler çalışmanın kendi kaynaklarından, benzer başka çalışmaların kaynaklarından veya çalışmanın amacına uygun açık erişime sahip veri tabanlarından sağlanabilir.

2.2.2. Veri Temizleme

Veri temizleme süreci, veri kümesi içerisinden eksik, aykırı veya hatalı değerler içeren kayıtların tespit edildiği ve düzeltme işleminin gerçekleştirildiği adımdır. Temel olarak, eksik, hatalı veya aykırı olan verilerin belirlenmesi, ardından bu verilerin değiştirilmesi veya silinmesi işlemine dayanır. Makine öğrenmesi çalışmalarında veri temizliği özellikle sağlık çalışmalarında son derece önemlidir. Veri kümesinde yer alan hatalı veya aykırı değer içeren kayıtlar makine öğrenmesi algoritmalarının başarımını olumsuz yönde etkileyebilir ve dolayısıyla modelin yanlış sonuçlar üretmesine neden olabilir [28]. Sağlık alanında yapılan makine öğrenmesi çalışmalarında alınan yanlış kararlar ve hatalı sonuçlar ciddi sorunlara sebep olabilmektedir. Örnek olarak; eksik, hatalı veya aykırı veriler doğrultusunda yapılan model hataları hastaların uygun tedaviyi almamalarına veya gereksiz tedavilere maruz kalmalarına, sağlık sorunlarının kötüleşmesine neden olabilir. Veri kümesindeki eksik, hatalı veya aykırı

değerlerin ele alınması için birçok yöntem geliştirilmiştir. Bu yöntemler, veri kalitesini artırmak ve analizlerin güvenilirliğini sağlamak amacıyla kullanılmaktadır.

2.2.3. Veri Dönüştürme

Veri dönüştürme süreci, verilerin orijinal formatından veya dağılımından farklı bir forma veya yapıya dönüştürülmesi işleminin gerçekleştirildiği adımdır. Bu dönüşüm işlemleri, verilerin daha iyi anlaşılmasını, işlenmesini veya analiz edilmesini sağlamak için yapılmaktadır. Veri dönüşümü, verilerin makine öğrenmesi algoritmalarına uygun bir formata dönüştürülmesinde önemli bir aşama olarak kabul edilir. Makine öğrenmesi alanında, veri dönüştürme işlemleri genellikle kategorik verilerin sayısal formata dönüştürülmesi, verilerin ölçeklendirilmesi, verilerin birleştirilmesi veya ayrılması, boyut azaltma ve özellik mühendisliği gibi işlemleri içerebilir. Bu işlemlerden sonra veriler makine öğrenmesi algoritmalarına daha uygun bir şekilde sunulduğundan bu algoritmalar nihai bir hedefe yönelik tahmin veya sınıflandırma çalışmalarında daha iyi performans gösterebilmektedir.

2.2.4. Veri Görselleştirme

Veri görselleştirme süreci, verilerin kolay anlaşılır bilgilere dönüştürülebilmesi için grafikler veya görsel araçlar kullanarak temsil edilme işleminin gerçekleştirildiği adımdır. Böylece, veri kümesinin özellikleri, dağılımları ve ilişkileri daha açık gözlemlenerek ilgili çalışmalarda daha etkili analizler gerçekleştirilebilir. Makine öğrenmesi algoritmalarının kullanımı öncesinde veri kümesindeki sınıf dağılımlarına göre veri kümesinin dengeli veya dengesiz olduğu, veri kümesinde bulunan aykırı değerlerin veya eksik değerlerin dağılımı ile ilgili bilgiler elde edilebilir. Ardından, bu algoritmalarının performans sonuçlarının iyileştirilmesine yönelik eksik değerlerin doldurulması, aykırı değerlerin çıkarılması, dengesiz veri sınıflarının dengelenmesi gibi sorunların üstesinden gelinmesinde çeşitli yöntemler uygulanabilir.

2.2.5. Veri Bölme

Veri bölme süreci, makine öğrenmesi modelleri için veri kümesinin eğitim ve test veri kümesi olarak ayrılmasının gerçekleştirildiği adımdır. Bazı çalışmalarda veri kümesi eğitim, geçerleme ve test kümesi olmak üzere üç bölüme de ayrılabilir. Eğitim veri kümesi bir makine öğrenmesi algoritmasının parametrelerinin ayarlanması ve modelin temel bilgilerinin öğrenilmesi için kullanılmaktadır. Geçerleme ve test veri kümeleri ise modelin performansını değerlendirmek için kullanılmaktadır. Temel olarak, veri bölümlendirme yöntemi olarak sınama seti yöntemi (Hold-out method) yaygın olarak kullanılır. Bu yöntemde, veri kümesinin belirli bir oranı

eğitim kümesine atanırken geri kalan bölümü ise test kümesine atanır. Veri kümesinin yeterli büyüklükte olduğu durumlarda bu yöntem kullanışlıdır. Veri kümesinin yeterli büyüklükte olmadığı durumlarda ise genellikle k-katlamalı çapraz doğrulama yöntemi (k-fold cross validation method) kullanılmaktadır. Bu yöntemde, veri kümesi eşit boyuttaki k alt bölüme ayrılır. Bu bölümlerden biri test kümesi için kullanılırken, kalan bölümler eğitim kümesi olarak kullanılır. Bu işlem çapraz bir şekilde tüm alt bölümler için k kez tekrarlanır. Bu yöntemin genel başarısı veya genel hatası (error rate), elde edilen k sonucun ortalaması hesaplanarak elde edilir.

2.2.6. Model Seçimi

Model seçimi süreci, problemin amacınıza en uygun olan modelin belirlendiği adımdır. Model seçimi, veri analizi ve makine öğrenmesi çalışmalarının önemli bir aşamasıdır ve doğru modeli seçmek, analizin başarısını büyük ölçüde etkileyebilir. Model seçimi, genellikle deneme yanılma yoluyla belirlenir. Veri analizinin amacı, veri tipinin sayısal veya kategorik olması, veri dağılımı, veri boyutu, hiper-parametreler, veri kümesinin dengeli olup olmaması gibi faktörler model seçimini etkilemektedir. Modelleme aşamasında sınıflandırma, regresyon, kümeleme ve zaman serisine bağlı analiz gibi çeşitli problemlerin türüne uygun olarak farklı modeller kullanılmaktadır. Modelleme aşamasında analizin amacına göre denetimli öğrenme, denetimsiz öğrenme, yarı denetimli öğrenme ve pekiştirmeli öğrenme gibi uygun makine öğrenmesi yöntemlerinden modeller tasarlanmaktadır.

2.2.7. Model Eğitimi

Model eğitimi süreci, bir makine öğrenmesi yönteminin veri kümesi üzerinde öğrenme sürecini ifade eden adımdır. Bu süreçte, sınıflandırma ve regresyon, iki temel öğrenme türüdür. Regresyon, bir girdi verisiyle ilişkilendirilmiş sürekli bir çıktı tahmin etmeye odaklanan bir öğrenme türüdür. Örneğin, diyabet hastalarının geçmiş kan şekeri değerleri kullanılarak gelecekteki kan şekeri değerlerini tahmin etmek için bir regresyon modeli kullanılabilir. Sınıflandırma ise seçilen bir makine öğrenmesi modelindeki eğitim veri kümesindeki verilerle bu verilere karşılık gelen etiketleri veya çıktıları öğrenmeyi amaçlar. Örneğin, hastalık tahmin çalışmasında bir bireyin hasta veya sağlıklı olarak ayırt edilmesi için bir sınıflandırma modeli kullanılabilir. Bu süreç, modelin belirli bir görevde doğru kararlar verebilmesi için veri kümesindeki desenleri öğrenmesini içerir.

2.2.8. Model Ayarlaması

Model ayarlaması süreci, sınıflandırma veya regresyon gibi belirli bir görevin gerçekleştirilmesinde bir makine öğrenmesi modelinin en iyi performansı elde etmek için hiper-parametrelerin ayarladığı adımdır. Başlangıçta varsayılan hiper-parametre değerleriyle başlayan bu süreç, hiper-parametre değerlerini deneyerek, en iyi performansı elde edecek yapılandırmaları belirler. Bu adımlar, hiper-parametre aralıklarının belirlenmesi, çapraz geçerleme kullanılması, hiper-parametrelerin ayarlanması, sonuçların değerlendirilmesi, en iyi hiper-parametrelerin seçilmesi ve nihai modelin eğitimi ve değerlendirmesini içerir. Model ayarlaması, modelin aşırı uyum (overfittting) veya yetersiz uyum (underfitting) gibi sorunları önlemeye yardımcı olarak daha iyi tahminler elde etmeye olanak tanır.

2.2.9. Model Değerlendirmesi

Model değerlendirmesi süreci, bir makine öğrenmesi modelinin performansını ölçme ve analiz etme sürecini ifade eden adımdır. Bu süreç, modelin ne kadar iyi veya kötü performans gösterdiğini anlamamıza ve gerektiğinde iyileştirmeler yapmamıza olanak tanır. Bir sınıflandırma probleminde, eğitilen model test verileri üzerinde genellikle Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall veya Sensitivity), Özgüllük (Specificity), F1-skoru, Matthews Korelasyon Katsayısı (MCC) ve AUC gibi model değerlendirme ölçütlerine göre değerlendirilir. Model değerlendirme ölçütleri Tablo 2.1'de verilen değerlere göre hesaplanır. Bu matrise göre, dengesiz veri kümelerinde, azınlık sınıfı örnekleri pozitif olarak tasvir edilirken, çoğunluk sınıfı örneklerinin sınıf etiketi negatif olarak temsil edilir. Doğruluk oranı, model tarafından yapılan doğru tahminlerin tüm veri kümesindeki örneklerin sayısına oranıdır. Bu oran 0 ile 1 arasında bir değere sahiptir ve bu değer 1'e yaklaştıkça modelin performansının ne kadar güçlü olduğunu ifade eder. Doğruluk oranı Denklem 2.1 'de gösterildiği gibi hesaplanmaktadır. Kesinlik, gerçek pozitif sonuçların toplam pozitif tahminlere bölünmesiyle elde edilir. Bu metrik ile doğru yapılan tahminlerin sayısı belirlenir. Kesinlik Denklem 2.2'de gösterildiği gibi hesaplanmaktadır. Gerçek Pozitif Oran (TPR) olarak da bilinen duyarlılık, doğru pozitif tahminlerin, doğru pozitif tahminlerin sayısı ile doğru negatif tahminlerin sayısına bölünmesi ile elde edilir. Bu metrik, hedef sınıfın bir azınlık sınıfını temsil ettiği, veri kümesinin dengesiz bir sınıf dağılımı sergilediği senaryolarda önemli bir öneme sahiptir. Duyarlılık Denklem 2.3'te gösterildiği gibi hesaplanmaktadır. Yanlış Pozitif Oran (FPR) olarak da bilinen özgüllük, yanlış sınıflandırılmış negatiflerin veri kümesindeki toplam negatif örnek sayısına oranını ifade eder. Özgüllük Denklem 2.4'te gösterildiği gibi

hesaplanmaktadır. F1-skoru, kesinlik ve duyarlılık metriklerinin harmonik ortalamasının hesaplanmasıyla gerçekleştirilir. F1-skor Denklem 2.5'te gösterildiği gibi hesaplanmaktadır. MCC, ikili ve çok sınıflı sınıflandırma modellerinin değerlendirilmesi için yaygın olarak kullanılan bir performans ölçüsüdür. MCC, karışıklık matrisinin dört bileşeninin tümünü dikkate alır, böylece sınıflandırma kalitesine ilişkin kapsayıcı ve güvenilir bir gösterge sağlar. MCC Denklem 2.6'da gösterildiği gibi hesaplanmaktadır. AUC, bir Alıcı İşletim Karakteristik (ROC) eğrisinin altında kalan alanı ifade eder. ROC eğrisi, bir modelin performansını TPR ve FPR sınıflandırma ölçümleri açısından gösteren grafiksel bir temsildir. Sınıflandırma değeri 1'e ne kadar yakınsa sınıflandırıcının performansı o kadar iyidir. Bu metrikler, modelin belirli bir problemi ne kadar iyi çözebildiğini objektif bir şekilde değerlendirmemize olanak tanır. Modelin doğru ve güvenilir sonuçlar üretebilmesi için bu süreç titizlikle yürütülmelidir.

Tablo 2.1: Karışıklık matrisi.

		Gerçek (Actual)	
		Pozitif	**Negatif**
Tahmin (Predicted)	**Pozitif**	Doğru Pozitif (TP)	Yanlış Pozitif (FP)
	Negatif	Yanlış Negatif (FN)	Doğru Negatif (TN)

$$Doğruluk\ (Accuracy) = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$Kesinlik\ (Precision) = \frac{TP}{TP + FP} \tag{2.2}$$

$$Duyarlılık\ (Recall)(TPR) = \frac{TP}{TP + FN} \tag{2.3}$$

$$Özgüllük\ (Specificity)(FPR) = \frac{FP}{FP + TN} \tag{2.4}$$

$$F1 - skor\ (F1) = 2\ X \frac{Kesinlik\ X\ Duyarlılık}{Kesinlik + Duyarlılık} \tag{2.5}$$

$$MCC = \frac{TP + TN - FP\ X\ FN}{\sqrt{(TP + FP)x(TP + FN)X\ (TN + FP)X(TN + FN)}} \tag{2.6}$$

2.3. MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ

Makine öğrenmesi sınıflandırma yöntemleri, bir veri kümesindeki girdileri belirli sınıflara atayan modelleri ifade eder. Bu modeller, genellikle eğitim veri kümelerindeki desenleri öğrenir ve daha sonra test veri kümelerindeki veriler üzerinde sınıflandırma yapabilir. Bu amaçla geliştirilmiş çok sayıda makine öğrenmesi sınıflandırma yöntemi bulunmaktadır. Aşağıda tez çalışması kapsamında ele alınan sınıflandırma yöntemleri özetlenmiştir.

2.3.1. Lojistik Regresyon (LR)

Lojistik Regresyon (LR), istatistik ve makine öğrenmesi sınıflandırma problemlerinde yaygın olarak kullanılan istatistiksel bir analiz yöntemidir. Bu yöntem, genellikle bir veya daha fazla bağımsız değişken (tahmin edici) ile ikili bağımlı değişken (sınıf veya sonuç) arasındaki ilişkiyi tanımlamak için tercih edilir [29]. Örneğin; bir veri kümesindeki bireylerin hasta veya sağlıklı olarak kategorize edilmesinde kullanılabilir. Bağımlı bir değişkenin, iki sınıfın bir üyesi olma olasılığının tahmini, ilgili değişkenlerin belirli ağırlıklarla çarpılıp toplanması ile elde edilen doğrusal bir kombinasyonun, logit dönüşümü uygulanarak hesaplanmaktadır. Bu hesaplama Denklem 2.7 'de verilmiştir.

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}} \tag{2.7}$$

Burada, P(Y=1*)* bağımlı değişkenin 1 olma olasılığını temsil etmektedir. *e* Euler sayısıdır ve $b_0+b_1x_1+b_2x_2+\cdots+b_1x_k$ parametreleri modelin öğrenmesi gereken katsayılardır. x_0, x_1, x_2,...,x_k ise giriş değişkenleridir [30]. LR, modelin öğrenme sürecinde genellikle maksimum olabilirlik yöntemi veya gradyan inişi gibi optimizasyon tekniklerini kullanır. Böylece, veri kümesindeki örneklerden elde edilen bilgilerle model parametreleri güncellenerek sınıflandırma işlemi yapılır. Bu yöntem, özellikle iki sınıflı sınıflandırma problemlerinde yaygın olarak kullanılır ve etkili bir çözüm sunar.

2.3.2. Rastgele Orman (RF)

Breiman tarafından geliştirilen Rastgele Orman (RF) algoritması, sınıflandırma ve regresyon problemlerinde yaygın olarak kullanılan bir makine öğrenmesi yöntemidir [31]. Veri kümesindeki değişkenlerin kullanılmasıyla bir bireyin diyabet olup olmadığını tahmin etmek, bir sınıflandırma problemini temsil ederken; bir bireyin diyabet olma olasılığının hesaplanması ise regresyon problemlerine örnek olarak gösterilebilir. RF, veri kümesinden rastgele alt kümeler oluşturur ve her bir alt küme için bir karar ağacı eğitilir [31]. Her bir karar ağacı, veri

kümesindeki özelliklere dayalı olarak sınıflandırma veya regresyon yapar. Ardından, bu ağaçlardan elde edilen tahminler bir araya getirerek bir çıktı tahmin edilir. Sınıflandırma problemlerinde genellikle oylama kullanılırken, regresyon problemlerinde tahminlerin ortalaması alınır [32]. Bu yöntemin bir dizi avantajı bulunmaktadır. Bu avantajlar içerisinde, çoklu ağaçların çeşitliliğinin birleştirilmesiyle modelin aşırı öğrenme riskinin azaltılması ve daha güvenilir, etkili sonuçların elde edilmesi yer almaktadır. Veri kümesinin rastgele seçilmesiyle varyansın kontrol edilmesi ve önemli özelliklerin belirlenebilmesi de bu yöntemin diğer öne çıkan avantajlarıdır. Bu özellikler, bu yöntemi özellikle geniş veri kümeleri üzerinde etkili bir şekilde çalışan ve güçlü bir genelleme yeteneğine sahip bir model haline getirir.

2.3.3. Destek Vektör Makinesi (SVM)

Destek Vektör Makinesi (SVM), sınıflandırma ve regresyon problemlerinde yaygın olarak kullanılan bir makine öğrenmesi yöntemidir. Bu yöntem, veri kümesinde yer alan veri noktalarını optimal bir hiper düzlemle en iyi şekilde ayırmayı amaçlamaktadır. SVM'in dört temel konsepti vardır: ayırıcı hiper düzlem, destek vektörler, marj ve çekirdek fonksiyonu [33]. Ayırıcı hiper düzlem, veri kümesindeki farklı veri örneklerini birbirinden ayıran optimal bir çizgidir [33]. Destek vektörler, hiper düzlemle en yakın olan veri noktalarını ifade eder [33]. Bu noktalar, sınıflar arasındaki mesafeyi (marj) belirleyen kritik unsurlardır. Marj, destek vektörler arasındaki uzaklığı ifade eder. SVM, bu marjı maksimize etmeye çalışarak sınıflar arasındaki ayrımı en iyi şekilde sağlamak ve aynı zamanda modelin genelleme yeteneğini artırmak için kullanılır [33]. Çekirdek fonksiyonları, verileri düşük boyutlu bir uzaydan daha yüksek boyutlu bir uzaya yansıtarak SVM'in doğrusal olarak ayrılabilen olmayan veri kümelerinde de etkili olmasını sağlar [33]. SVM algoritmasının avantajları arasında çok boyutlu veri kümelerinde etkili performans, destek vektörler sayesinde aşırı öğrenmeye karşı dayanıklılık ve çeşitli çekirdek fonksiyonlarını kullanabilme esnekliği bulunmaktadır. SVM, özellikle karmaşık sınıflandırma problemleri ve regresyon analizleri için tercih edilir.

2.3.4. Karar Ağacı (DT)

Karar Ağacı (DT), sınıflandırma ve regresyon problemlerinde kullanılan bir makine öğrenmesi yöntemidir. Bu yöntem, veri kümesindeki öznitelik değerlerine dayanarak belirli bir dizi kurala göre kararlar alır. Bu kararlara göre, veri örneklerini sınıflandırma veya bir hedef değişkenin değerini tahmin etme işlemi gerçekleştirir [34]. DT yöntemi, kök düğüm, dallar, iç düğümler ve yaprak düğümlerden oluşan hiyerarşik bir ağaç yapısından oluşur [35]. Kök düğüm, tüm veri kümesini temsil eden ağacın başlangıç noktasıdır. Kök düğümden çıkan her bir öznitelik

değeri ve kararlar dallar aracılığıyla iç düğümlere bağlanır. İç düğümler, bir dalda bulunan ve bir öznitelik değerine göre karar almayı sağlayan yapılardır. Yaprak düğümler, sınıflandırma problemlerinde bir sınıfı veya regresyon problemlerinde bir tahmin değerine karşılık gelen sonuç değerlerini temsil eden yapılardır. DT, basit böl ve yönet yaklaşım kurallarını kullanma yeteneği ve çok yönlülüğü sayesinde birçok avantaja sahiptir. Bu özellikler, çeşitli alanlardaki karmaşık problemleri çözmek için DT yöntemini ideal kılarken, aynı zamanda anlaşılabilirlikleri modelin içerdiği kararları açıklamasını kolaylaştırır.

2.3.5. Naive Bayes (NB)

Bayes teorimine dayanan Naive Bayes (NB) algoritması, özellikle sınıflandırma problemlerinde yaygın olarak kullanılan bir makine öğrenmesi yöntemidir. Bir sınıflandırma probleminde, sınıflar arasındaki olasılık ilişkilerinin modellenmesinde Bayes teorimi kullanılır ve bu bilgiler ışığında verilen bir örnek ilgili sınıfa atanır [36]. Bayes sınıflandırma formülü Denklem 2.8'de verilmiştir.

$$P(Sınıf \mid Örnek) = \frac{P(Örnek \mid Sınıf).P(Sınıf)}{P(Örnek)} \tag{2.8}$$

Bu formülde:

- $P(Sınıf \mid Örnek)$: Örneğin bir sınıfa ait olma olasılığı.
- $P(Örnek \mid Sınıf)$: Sınıfa ait örnekteki özniteliklerin olasılıklarının çarpımı.
- $P(Sınıf)$: Sınıfın genel olasılığı.
- $P(Örnek)$: Örneğin genel olasılığı.

Bu formül, belirli bir örneğin verilen sınıfa ait olma olasılığını hesaplamak için kullanılır [36]. NB algoritması, bir örneğin hangi sınıfa ait olma olasılıkları hesaplamak için eğitim verisi üzerinde çalışır ve daha sonra test verisi üzerinde bu olasılıkları kullanarak sınıflandırma yapar.

2.3.6. Aşırı Gradyan Artırma (XGBoost)

Aşırı Gradyan Artırma (XGBoost) algoritması, ilk olarak 2011 yılında Tianqi Chen ve Carlos Guestrin tarafından önerilen sınıflandırma ve regresyon problemlerinde yaygın olarak bir

makine öğrenmesi algoritmasıdır [37]. Bu yöntem, güçlü bir öğrenici modelini üretmek için çoklu öğrenici modellerini birleştiren Boosting ağaç modellerine dayanan bir öğrenme yapısıdır. XGBoost yöntemi, öğrenme sürecini gerçekleştirirken önceki ağaçların hatalarını düzeltmeyi amaçlayan bir hedef fonksiyonu kullanır [38]. Bu süreçte, sınıflandırma problemlerinde, tahminler ile gözlemlenen örnekler arasındaki hataların farkları ölçülür [38]. Regresyon problemlerinde ise, her ağacın tahmin değeri, her bir ağacın ortalaması alınarak belirlenir [38]. Bu iteratif süreç, modelin performansını artırmaya yönelik olarak devam eder. XGBoost yöntemi, paralel işleme yetenekleri, ölçeklenebilirliği, düzenleme özellikleri ve güçlü tahmin gücü sayesinde büyük ölçekli veri kümelerinde hızlı ve verimli şekilde çalışır.

2.4. EKSİK VERİ İŞLEME YÖNTEMLERİ

Eksik değer, veri analizinin veri toplama sürecinde çeşitli nedenlerden dolayı eksik bilgilerin bulunma durumunu ifade eden bir problemdir. Bu eksik veriler genellikle katılımcıların verdiği eksik bilgiler, ölçüm cihazlarından alınan hatalı bilgiler veya kayıp verilerden kaynaklanabilir. Eksik değer problemi, tıbbi araştırmalar [39- 42], gaz analizi [43] ve ağ analizi [44] gibi gerçek dünya veri kümelerinde yaygın olarak görülmektedir. Örneğin; bir tıbbi teşhis çalışmasında görülen eksik değerler genellikle hasta katılımı ve/veya ilgili bazı soruları yanıtlama reddi, klinik kayıt hataları, sağlık personelinin verileri tam olarak girmemesi, tedavinin bırakılması ve laboratuvar testlerindeki teknik sorunlardan kaynaklanabilir. Veri kümesindeki eksik değerlerin varlığı, makine öğrenmesi algoritmalarının genelleme yeteneğini ve güvenilirliğini etkileyebilmektedir. Bu nedenle, eksik değer problemini yönetmek ve uygun eksik veri işleme yöntemlerini belirlemek araştırmacılar için önemli hale gelmiştir. Eksik veri işleme yöntemleri, veri analizi süreçlerinde karşılaşılan eksik değer problemlerini etkili bir şekilde yönetmek amacıyla kullanılır. Bu yöntemler, eksik değer içeren örnekleri silme yöntemleri ve eksik değerleri doldurma yöntemleri olmak üzere temelde iki ana strateji üzerine kurulmuştur.

2.4.1. Silme Yöntemleri

Eksik değerleri silme yöntemleri arasında satır bazlı silme, sütun bazlı silme ve eşlerin silinmesi (pairwise) gibi yöntemler bulunmaktadır. Satır bazlı silme yöntemi, en az bir eksik değer içeren örneklerin veri kümesinden çıkartıldığı bir eksik veri işleme stratejisidir. Satır bazlı silme yöntemi, basit ve hızlı bir çözüm sunmasına rağmen, eksik değer içeren örnekler analizin dışında bırakıldığından dolayı bilgi kaybına yol açabilir ve veri kümesinin genel yapısını değiştirebilir. Bu yöntem kullanılırken veri kümesinin yapısı, eksik verilerin dağılımı ve analiz

amacı dikkate alınmalıdır. Eksik değerlerin veri kümesinde rastgele dağıldığı ve küçük bir miktarı temsil ettiği durumlarda bu yöntem kullanılabilir. Ancak, bilgi kaybının kritik olduğu ve eksik değerlerin desenlere sahip olduğu durumlarda diğer eksik değer işleme stratejilerini kullanmak daha uygun olabilir.

Sütun bazlı silme, veri kümesindeki eksik değer içeren belirli bir değişkenin (sütun) veri kümesinden çıkartıldığı bir eksik veri işleme stratejisidir. Sütun bazlı silme yöntemi, genellikle belirli bir değişkende eksik verilerin yoğunlaştığı durumlar için tercih edilir. Bu yöntemin sağladığı avantajlar, eksik değer içeren belirli bir değişkenin tamamen veri kümesinden çıkarılması sonucunda analiz sürecini basitleştirmesi ve hızlandırmasıdır. Ayrıca, sadece eksik değer içeren değişkenlerin analizden çıkartılması diğer analizlerin etkilenmesini engelleyerek basit ve hızlı çözüm sunar. Ancak, bu yöntemin kullanılması bir dizi dezavantajı da beraberinde getirir. Eksik değer içeren değişkenin veri kümesinden çıkarılması, o değişkenin içerdiği diğer değerlere ait ilişkilerin kaybolmasına yol açarak bilgi kaybına neden olabilir. Ek olarak, eksik değerlerin belirli bir desen gösterdiği durumlarda bu yöntemin kullanılması veri kümesinin genel yapısını önemli ölçüde bozabilir. Sütun bazlı silme yöntemi, özellikle eksik değer içeren değişkenin analizle ilgili olmadığı ve bu değişkenin eksik değer içerme oranının düşük olduğu durumlarda tercih edilebilir.

Eşlerin silinmesi, veri kümesinde eksik değer içeren örneklerden oluşan çiftlerin analizden çıkartıldığı bir eksik veri işleme stratejisidir. Örneğin; bir analizde kullanılan bir çift değişken içerisinde en az bir eksik değer bulunması durumunda bu çiftler analizden çıkartılır. Bu strateji, eksik değer içeren değişkenlerin analizden tamamen çıkarılmasını engelleyerek, eksik verilerin sadece ilgili değişkenler üzerindeki etkilerini azaltmayı amaçlamaktadır. Böylece, diğer değişkenlerin etkileri minimum düzeyde tutularak bu yöntemin hızlı ve uygulanabilir olması sağladığı avantajlar içerisindedir. Ancak, veri kümesinden silinerek analiz dışında tutulan örnekler diğer değişkenlere ait ilişkileri ve desenleri temsil etme potansiyellerinden dolayı veri kümesinde bilgi kaybına neden olabilir. Eşlerin silinmesi yöntemi, eksik verilerin miktarının az olduğu ve belirli bir desen içermediği durumlarda tercih edilebilir.

2.4.2. Doldurma Yöntemleri

Eksik verileri doldurma yöntemleri, istatistiksel yöntemler ve makine öğrenmesi tabanlı yöntemler olarak iki kategoride gruplanabilir. İstatistiksel yöntemler kapsamında değerlendirilen yöntemlerin ilki basit doldurma yöntemleridir. Basit doldurma yöntemleri,

eksik veri sorununa yönelik temel stratejiler olarak kabul edilir ve veri kümesinin özelliklerine bağlı olarak tercih edilebilir. Veri kümesinde yer alan eksik değerleri doldurmak için iki temel yaklaşım bulunmaktadır. İlk yaklaşım, ortalama, mod ve medyan doldurma yöntemlerini içermektedir. Bu yaklaşım, veri kümesindeki eksik değerleri sırasıyla ilgili sütunun ortalaması, modu veya medyanı ile doldurur. Bu stratejide, verinin genel eğilimi korunsa da veri kümesindeki karmaşıklık tam olarak yansıtılamamaktadır. İkinci yaklaşımda ileri veya geri yayılma doldurma yöntemleri bulunmaktadır. İleri yayılma yönteminde, veri kümesi içerisinde yer alan bir sütundaki eksik değerler ilgili sütundaki bir önceki geçerli değer ile doldurulur. Geri yayılma yönteminde ise veri kümesi içerisinde yer alan bir sütundaki eksik değerler ilgili sütundaki bir sonraki geçerli değer ile doldurulur. Bu strateji, özellikle zamansal veya sıralı veri kümelerinde eksik değerleri etkili bir şekilde doldurmak amacıyla tercih edilmektedir. İstatistiksel yöntemler kapsamında değerlendirilen bir başka yaklaşım ise çoklu doldurma yöntemleridir. Bu yöntemler, veri kümesindeki eksik değerler ile başa çıkabilmek için birden fazla veri kümesi oluşturularak eksik verilerin tahmin edilmesini sağlar. Her bir veri kümesi, eksik değerleri farklı yöntemlerle tahmin etmek için kullanılır ve elde edilen her bir sonuç birleştirilerek eksik değerlerin tamamlandığı tek bir veri kümesi elde edilir. Bu yöntemler, genellikle bir dizi eksik değeri aynı anda ele almak için kullanılır ve böylece daha güvenilir sonuçlar sağlayabilir. Bu yöntemlerden biri literatürde başarıyla kullanılan bir yaklaşım olan MICE yöntemidir. MICE yöntemi, eksik değerler ile başa çıkmak için kullanılan istatistiksel bir yaklaşım içerir. Bu yöntemde, veri kümesindeki eksik değerlerin tahmin edilmesi için eksik değer içermeyen veriler ve diğer değişkenler kullanılır. Tahmin modeli olarak bir dizi regresyon yöntemi uygulanır. Bu işlem, belirli bir kriter sağlanana kadar veya belli bir iterasyon sayısı gerçekleşene kadar devam eder. Her iterasyonda, eksik değerlerin tahminleri güncellenir ve daha doğru tahminler elde edilir. MICE yöntemi, veri kümesindeki eksik değerlerin tahmin edilmesinde değişkenler arasındaki ilişkileri dikkate alarak esnek bir yaklaşım sunar. Bu nedenle, MICE yöntemi, karmaşık veri kümelerinde eksik değerlerin tahmin edilme sürecinde başarılı bir şekilde kullanılır.

Makine öğrenmesi tabanlı eksik değer atama yöntemleri, eksik verileri ele almak için tahmine dayalı öğrenme modellerinin kullanımına dayalı tekniklerdir. Bu yöntemler, verilerdeki eksik olmayan bilgileri kullanarak elde edilen bilgilere bağlı olarak eksik verileri tahmin eder. Bilimsel literatürde k-en yakın komşu (KNN), SVM, RF ve regresyon modelleri yaygın olarak kullanılan bazı makine öğrenmesi tabanlı atama yöntemleridir. KNN ataması, eksik değerleri

veri kümesindeki benzer veri noktalarının değerlerini kullanarak tahmin eden bir tekniktir. Bu yöntem, eksik değerleri verilerde mevcut olan ilişkileri ve kalıpları hesaba katarak doldurmayı amaçlar. KNN ataması, birden fazla değişkeni veya özelliği dikkate aldığı için çok değişkenli eksik değer atama yöntemi olarak bilinmektedir. SVM atama yöntemi, regresyon tabanlı eksik değerlerin tahmin edilmesinde etkili olan bir makine öğrenmesi algoritmasıdır. SVM ataması, çıktı değerleri ve karar değerlerini kullanır. İlk olarak, eksik olan durum nitelikleri belirlenir ve ardından bu eksik değerler bir SVM modeli ile tahmin edilir. Bu yöntem, büyük ölçekli veri kümelerinde etkili bir performans sergiler ve bellek kullanımını verimli bir şekilde yönetir. RF atama yöntemi, veri kümesinde bulunan eksik değerlerin tahmin edilmesi için kullanılan bir makine öğrenmesi algoritmasıdır. Bu yöntemde, eksik değerlerin tahmini birçok karar ağacının bir araya gelmesi ile elde edilir. Tahmin modelinde yer alan her bir karar ağacı, rastgele seçilen örnekler ve özellikler üzerinde eğitilir. Sonuç olarak, her bir eksik değer tüm karar ağaçlarından elde edilen tahminlerin birleşimi ile elde edilir. RF yöntemi, büyük veri kümelerinde etkili bir şekilde çalışabilir ve veri yapısını koruyabilir. Böylece, eksik değerlerin tahmini problemlerinde sıklıkla tercih edilir. Regresyon modelleri, eksik değerleri veri kümesindeki diğer değişkenler ile doğrusal ilişkilerini kullanarak tahmin eder. Bu yöntem, bir değişkenin değerinin diğer değişkenlerle doğrusal olarak değiştiğini varsayar. Veri kümesindeki eksik değerler doğrudan bir regresyon modeli aracılığıyla tahmin edilir ve böylece eksik değerlerin atanması sağlanır.

2.5. SINIF DENGELEME YÖNTEMLERİ

Sınıf dengesizliği, veri kümesindeki sınıflarda bulunan örneklerin dağılımının eşit olmadığı durumlarda ortaya çıkan bir problemdir. Sınıf dengesizliği genellikle gerçek dünya veri kümelerinde, özellikle tıbbi teşhis [45-49], dolandırıcılık tespiti [50, 51], siber güvenlik [52, 53] ve spam tespiti [54-56] gibi alanlarda yaygın olarak görülür. Örneğin; bir tıbbi teşhis ikili sınıflandırma probleminde, genellikle sağlıklı bireyleri temsil eden sınıf çoğunluk sınıfı oluştururken, hasta bireyleri temsil eden sınıf azınlık sınıfı oluşturur. Bu durumda, makine öğrenmesi sınıflandırma algoritmaları azınlık sınıfı örneklerini yeterince öğrenemediğinden dolayı yanlı sonuçlar üretebilir. Sınıf dengeleme yöntemleri, sınıflar arasındaki örnek dağılımını dengeleyerek bu tür problemlerin üstesinden gelmeyi amaçlar. Sınıf dengesizliğiyle başa çıkmak için eksik yeniden örnekleme yöntemleri, aşırı yeniden örnekleme yöntemleri, sentetik yeniden örnekleme yöntemleri ve hibrit yeniden örnekleme yöntemleri gibi çeşitli

yaklaşımlar kullanılabilir. Literatürde yaygın olarak kullanılan sınıf dengeleme yöntemlerinden bazıları aşağıda yer alan alt bölümlerde açıklanmaktadır.

2.5.1. Eksik Yeniden Örnekleme Yöntemleri

Eksik yeniden örnekleme yöntemleri, dengesiz sınıf dağılımlı veri kümelerinde çoğunluk sınıfına ait örneklerin sayısını azaltarak sınıf dengesizliğini ele alan bir veri örnekleme yöntemidir [57]. Bu yöntemlerden rastgele eksik yeniden örnekleme yöntemi (random under-sampling) en çok bilinendir [58]. Bu yöntem, çoğunluk sınıfından rastgele seçilen örneklerin veri kümesinden çıkarılmasıyla veri kümesini dengelemeyi amaçlar. Böylece, azınlık ve çoğunluk sınıfları arasında denge sağlanarak makine öğrenmesi modellerinin genelleme yeteneği iyileştirilir ve dengesizlik nedeniyle oluşabilecek aşırı uyum problemi minimize edilir. Ancak, bu yöntem çoğunluk sınıfında yararlı olabilecek bilgileri de silebileceğinden dolayı bilgi kaybına ve nadir durumların doğru bir şekilde temsil edilememesine yol açabilir.

2.5.2. Aşırı Yeniden Örnekleme Yöntemleri

Aşırı yeniden örnekleme yöntemi, dengesiz sınıf dağılımlı veri kümelerinde azınlık sınıfına ait örneklerin sayısını artırarak sınıf dengesizliğini ele alan bir veri örnekleme yöntemidir [58]. Rastgele aşırı yeniden örnekleme yöntemi (random over-sampling) sınıf dengesizliğini ele almak için en yaygın kullanılan aşırı yeniden örnekleme stratejisidir. Bu yöntem, azınlık sınıfına ait örnekleri kopyalayarak veya tekrar örnekleyerek veri kümesindeki azınlık sınıfının örnek sayısını artırır [59]. Böylece, azınlık ve çoğunluk sınıfları arasında denge sağlanarak makine öğrenmesi modellerinin azınlık sınıfını daha etkili bir şekilde öğrenmesi ve daha iyi bir performans sergilemesi sağlanır. Ancak, bu yöntem azınlık sınıfındaki örnekleri birçok kez kopyalayarak makine öğrenmesi modellerinde aşırı uyum sorunlarına yol açabilir. Ayrıca, bu yöntemin kullanılması ile veri kümesinin veri dağılım dengesi değişebilir ve modelin performansı olumsuz yönde etkilenebilir.

2.5.3. Sentetik Yeniden Örnekleme Yöntemleri

Sentetik yeniden örnekleme yöntemleri, veri kümelerinde sıklıkla görülen dengesiz sınıf dağılımları problemine çözüm sunan ve veri bilimi ile makine öğrenmesi alanlarında yaygın olarak kullanılan bir veri artırma tekniğidir. Yaygın olarak kullanılan sentetik örnekleme yöntemleri arasında, SMOTE ve Uyarlanabilir Sentetik Örnekleme Yaklaşımı (ADASYN) gibi yöntemler yer almaktadır. Bu yöntemler, veri kümesindeki örneklerinden yararlanarak yeni sentetik örnekleri oluşturur.

SMOTE yöntemi Chawla ve diğ. [60] tarafından 2002 yılında önerilen bir sentetik aşırı örnekleme algoritmasıdır. Bu yöntem, genellikle sınıflandırma problemlerinde dengesiz dağılımlı sınıfları dengelemek amacıyla yaygın olarak kullanılmaktadır. SMOTE yöntemi, azınlık sınıfında yeni sentetik örnekler oluşturmak için mevcut örneklerin k-en yakın komşuları arasında doğrusal enterpolasyon tekniğini kullanarak aşırı örnekleme yapar. Sınıfların dengelenmesinde kullanılacak k komşu sayısı, aşırı örnekleme miktarına bağlı olarak rastgele seçilerek belirlenir. Azınlık sınıfından (X_i ve X_j) iki örneğin doğrusal enterpolasyonuyla Denklem 2.9 'a göre yeni bir sentetik örnek oluşturulur.

$$X_{yeni} = X_i + (X_j - X_i) * \alpha \tag{2.9}$$

Burada X_{yeni} doğrusal enterpolasyon yoluyla oluşturulmuş yeni bir sentetik örnektir. Yeni bir sentetik örnek oluşturmak için öncelikle azınlık sınıfından bir veya birden fazla komşu arasından rastgele bir komşu (X_i) seçilir. Bu örnek, yeni sentetik örnekler oluşturmak için temel alınacak örneği belirler. Daha sonra, seçilen komşular (X_j ve X_i) arasındaki fark Öklid uzaklığına bağlı olarak hesaplanır. Bu, orijinal örneğin komşularına olan uzaklığını temsil eder. Bu fark, rastgele bir çarpan α değeri ile çarpılır. α değeri, 0 ile 1 arasında rastgele kayan bir sayıdır ve sentetik örneklerin orijinal örneğe ne kadar yakın veya uzak olacağını belirler. Son olarak, bu çarpılmış fark değeri orijinal örneğe eklenerek yeni bir sentetik örnek oluşturulur. Bu işlem, doğrusal interpolasyon yoluyla orijinal örneğin komşuları arasında bir nokta seçerek gerçekleşir. SMOTE yöntemi geleneksel aşırı örnekleme teknikleriyle karşılaştırıldığında üstün performans sergilemesine rağmen örtüşen, sınır ve gürültü örneklerinin oluşturulması gibi bazı sorunları da beraberinde getirebilir. Bu yöntem birçok çalışmada başarıyla kullanılmış ve literatürde birçok çalışmada SMOTE yönteminin geliştirilmesiyle elde edilen farklı uzantıları önerilmiştir.

ADASYN yöntemi, He ve diğ. [61] tarafından 2008 yılında önerilen bir sentetik aşırı örnekleme yöntemidir. Bu yöntem, dengesiz dağılımlı sınıfları dengelemek için sentetik örneklerin oluşturulmasında adaptif bir yaklaşım benimser. Bu yaklaşımda, ADASYN yöntemi azınlık sınıfı örneklerinin yoğunluk farklarına göre sentetik örnekler üretir. Buna göre, daha az temsil edilen azınlık sınıfı örnekleri için daha fazla sentetik örnekler üretilmesi sağlanır. Böylece, azınlık sınıfının daha iyi temsil edilmesi ve modelin genel performansının artması sağlanır. ADASYN, özellikle büyük ve yüksek boyutlu veri kümelerinde etkili bir şekilde çalışır ve sınıf dengesizliği sorununu azaltmada başarılı sonuçlar sağlar.

2.5.4. Hibrit Yeniden Örnekleme Yöntemleri

Hibrit yeniden örnekleme yöntemleri, sınıf dengesizliğini ele almak için hem aşırı örnekleme hem de eksik örnekleme gibi farklı yaklaşımları stratejik olarak birleştiren bir veri örnekleme yöntemidir. Bu yöntem, azınlık sınıflarının örnek sayısını artırırken, çoğunluk sınıflarının örnek sayısını azaltır. Bu yaklaşım, farklı örnekleme yöntemlerinin güçlü yönlerini kullanarak daha dengeli bir veri kümesi elde etmeyi hedefler. Örneğin, bir hibrit yöntem, azınlık sınıfındaki örneklerin sayısını artırmak için SMOTE gibi aşırı örnekleme yöntemlerini kullanırken, çoğunluk sınıfındaki örneklerin sayısını azaltmak için ENN veya Tomek gibi eksik örnekleme yöntemlerini kullanabilir. SMOTEENN yöntemi, rastgele seçilen homojen komşu örnekler üzerinde doğrusal enterpolasyon yoluyla sentetik örnekleri oluşturmak için SMOTE aşırı örnekleme yöntemini kullanır. Sentetik örnekler içerisindeki aykırı ve gürültülü örnekler ENN eksik yeniden örnekleme yöntemi ile veri kümesinden kaldırılır [62]. Böylece, SMOTEENN yöntemi, azınlık sınıfının temsilini artırırken aynı zamanda veri kümesindeki gürültüyü azaltarak makine öğrenmesi modelinin aşırı uyumunu önler. SMOTETomek yöntemi, azınlık sınıfında sentetik örnekler oluşturmak için SMOTE yöntemini kullanırken, sınıf sınırları Tomek bağlantıları kullanılarak netleştirilir [63]. Böylece, azınlık sınıfının temsili artırılırken sınıf sınırları iyileştirilerek sınıf dengesi için etkili bir strateji sağlanır. Hibrit yeniden örnekleme yöntemlerinin avantajları arasında, veri kümesinde dengeli sınıfların elde edilmesi, azınlık sınıfının daha etkili bir şekilde temsil edilmesi ve makine öğrenmesi model performansının iyileştirilmesi yer almaktadır. Ancak, hibrit yeniden örnekleme yöntemlerinin karmaşıklığı sebebiyle uygulanması karmaşıktır.

2.6. AYKIRI DEĞER TESPİT YÖNTEMLERİ

Aykırı değerler, istatistiksel analiz ve veri madenciliği alanlarında geniş bir şekilde ele alınan önemli bir konudur. Bu değerler genellikle bir veri kümesinde beklenen dağılımdan önemli ölçüde farklı olan nadir gözlemlerdir veya diğer verilere göre uygun olmayan değerlerdir. Aykırı değerlerin tespiti, veri analizinde ve makine öğrenmesi modellerinin geliştirilmesinde önemlidir ve bu değerler analiz sonuçlarını ve model performansını yanıltabilir. Bu bağlamda, Denklem 2.10'da verilen Çeyrekler Arası Aralık (IQR) gibi istatistiksel yöntemler, veri kümesindeki aykırı değerleri tanımlamak için yaygın olarak kullanılmaktadır. IQR yöntemi, verinin dörtte birliklerini hesaplayarak alt ve üst sınırları belirler ve bu sınırların dışındaki değerleri aykırı değerler olarak işaretler. Üst ve alt sınır aralıkları Denklem 2.11 ve Denklem 2.12’e göre hesaplanır.

$$IQR = (Q3 - Q1) \tag{2.10}$$

$$\text{Üst sınır} = (Q1 - 1{,}5) * IQR \tag{2.11}$$

$$Alt\ sınır = (Q3 + 1{,}5) * IQR \tag{2.12}$$

Aykırı değerlerin tespiti, veri kümesinin doğruluğunu artırabilir ve modelin genelleme yeteneğini iyileştirebilir. Ancak, aykırı değerlerin nasıl ele alındığı, belirli bir veri kümesinin karakteristiklerine ve analizin amaçlarına bağlı olarak değişebilir.

2.7. SEZGİSEL YÖNTEMLER

Sezgisel yöntemler, genellikle doğal sistemlerden esinlenerek tasarlanan karmaşık problemlerin çözümünde kullanılan optimizasyon algoritmalarıdır. Bu algoritmalar, popülasyona dayalı olarak çözüm alanını araştırır ve genellikle problem alanında optimize edilmek istenen bir hedef fonksiyonu ile optimal çözümü bulmaya çalışırlar. PSO ve GA gibi sezgisel yöntemler, bu yöntemlere örnek olarak verilebilir.

2.7.1. Parçacık Sürü Optimizasyonu (PSO)

PSO, kuşlar, balıklar ve böcekler gibi sürü halinde hareket eden bazı hayvanların davranışlarını modelleyerek optimize edilmesi gereken bir problemde çözüm alanını araştıran Kennedy ve Eberhart [64] tarafından geliştirilen popülasyona dayalı bir optimizasyon algoritmasıdır. Bu yöntem, sürü zekası kavramına dayanır ve aday çözümlerin bir parçacık sürüsü olarak temsil edildiği bir çözüm uzayında hareket etmelerini sağlar. Parçacıklar, kendi performanslarına ve komşularının en iyi performanslarına göre yönlendirilen izler oluşturur. PSO, diğer doğadan esinlenen algoritmalardan farklı olarak, iş birliği ve rekabetin nesiller boyunca bireyler arasında gerçekleştiği bir evrim sürecine dayanır. Bu bilgi akışı, lokal veya global olabilir ve algoritmanın temel bir özelliğidir [65]. PSO, sürünün bireysel üyeleri olan parçacıklardan oluşur. Başlangıçta, her bir parçacık rastgele bir şekilde başlatılır. Daha sonra, PSO geçmiş deneyimlerden öğrenerek her bir parçacığın sürü içindeki en iyi konumunu bulmak için bir optimizasyon süreci gerçekleştirir. Bu süreç, belirlenen bir durdurma kriteri sağlanana veya iterasyonun sonuna kadar devam eder [66]. Bir parçacığın hareketi, sırasıyla Denklem 2.13 ve Denklem 2.14'te verilen konum ve hız formülleri kullanılarak açıklanabilir.

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{2.13}$$

Bu denklemde $x_i(t)$, i numaralı parçacığın t zamanındaki konumunu temsil eder. $v_i(t+1)$ ise aynı parçacığın $t+1$ zamanındaki hızını ifade eder. Bu denklem, her parçacığın konumunu ve hızını bir sonraki zaman adımına güncellemek için kullanılır.

$$v_i(t+1) = wv_i(t) + c_1 r_1[p_{best} - x_i(t)] + c_2 r_2[g_{best} - x_i(t)] \quad (2.14)$$

Bu denklemde w ağırlık faktörünü, c_1 ve c_2 bilişsel ve sosyal ivmelenme katsayılarını, r_1 ve r_2 rastgele sayıları, p_{best} kişisel en iyi konumu, g_{best} ise genel en iyi konumu temsil eder. Bu denklem, her bir parçacığın hızını güncellemek için kullanılır ve parçacıkların kendi en iyi konumlarına ve hızlarının nasıl güncellendiğini açıklar. Bu güncellemeler, sürünün en iyi çözümüne doğru yakınsamayı sağlar.

2.7.2. Genetik Algoritma (GA)

GA yöntemi, Holland [67] tarafından biyolojik evrim süreçlerinden ilham alınarak geliştirilmiş popülasyon tabanlı bir optimizasyon algoritmasıdır. GA, doğal seçilim, çaprazlama ve mutasyon gibi biyolojiden esinlenilmiş ana operatörlere dayanmaktadır. Optimizasyon ve arama problemlerinin çözümlenmesinde bir popülasyon içindeki bireyleri iteratif olarak değiştirir ve adaptasyonlarına dayanarak çözüm alanında iyileşme sağlar. GA yönteminin sözde kodu Şekil 2.1'de sunulan Algoritma 1'de verilmiştir.

Algorithm 1 Genetik Algoritma

```
Başlangıç popülasyonunu oluştur
Uygunluk fonksiyonunu tanımla
while Durdurma koşulu sağlanana kadar do
    Ebeveynleri seç
    Çaprazlama ve mutasyon uygula
    Yeni popülasyonu oluştur
    Uygunluk tabanlı seçim yap
end while
En iyi çözümü raporla
```

Şekil 2.1: Genetik Algoritmasının sözde kodu.

Algoritma 1, GA'nın genel işleyişini açıklamaktadır. Başlangıçta, bir başlangıç popülasyonu oluşturulur ve bu popülasyonun her bir bireyi bir çözüm adayını temsil eder. Daha sonra, probleme özgü bir uygunluk fonksiyonu tanımlanır, bu fonksiyon her bir çözümün ne kadar iyi olduğunu değerlendirir. Algoritma, belirli bir durdurma koşulu sağlanana kadar bir döngü içinde çalışır. Bu durdurma koşulu, genellikle belirli bir uygunluk seviyesine ulaşılması veya belirli bir iterasyon sayısının tamamlanması gibi kriterlere dayanabilir. Her döngüde, mevcut

popülasyondan ebeveynler seçilir. Seçim genellikle uygunluk fonksiyonuna dayalı olarak yapılır; daha iyi uygunluk değerine sahip bireyler daha yüksek bir olasılıkla seçilir. Seçilen ebeveynler arasında çaprazlama (crossover) ve mutasyon işlemleri uygulanır. Çaprazlama işlemi, ebeveynlerin genetik materyalinin birleştirilmesini sağlar ve yeni çözümlerin oluşturulmasına katkıda bulunur. Mutasyon işlemi, rastgele seçilen bireylerin genetik materyalindeki değişiklikleri sağlar ve genetik çeşitliliği artırır. Yeni çözümlerden oluşan bir popülasyon oluşturulduktan sonra, bu popülasyonun daha iyi uygunluk değerlerine sahip bireylerinin seçilmesi için uygunluk tabanlı bir seçim yapılır. Bu adım, genellikle daha iyi çözümlerin bir sonraki nesillere aktarılmasını sağlar. Son olarak, GA en iyi çözümü raporlar. Bu, genellikle en yüksek uygunluk değerine sahip olan veya belirli bir kriteri karşılayan çözümü içerir.

2.8. LİTERATÜR İNCELEMESİ

Makine öğrenmesi yöntemleri, sağlık sektöründe hastalık teşhisi, tedavi planlaması, hasta izleme ve sağlık hizmetlerinin iyileştirilmesi gibi pek çok alanda önemli bir rol oynamaktadır. Ancak, sağlık veri kümeleri nadir hastalıkların teşhis edilmesi gibi durumlarda sıklıkla dengesiz sınıf dağılımı ve temel verilerin eksikliği gibi zorluklarla karşılaşmaktadır. Dengesiz sınıf dağılımı, makine öğrenmesi sınıflandırma modellerinin performansını olumsuz etkileyebilirken, eksik değerler ise modellerin doğruluğunu ve güvenilirliğini azaltabilir. Bu bağlamda, sağlık veri analizi alanında dengesiz sınıf dağılımı ve eksik veri sorunlarına etkili bir şekilde müdahale etmek, makine öğrenmesi modellerinin performansını artırmak ve modellerin karar verme süreçlerinin kalitesini iyileştirmek için kritik öneme sahiptir. Bu tez kapsamında, sağlık veri kümelerindeki dengesizlik ve eksiklik problemlerini ele almak için literatürdeki çalışmalar incelenmiştir.

Sailasya ve diğ. [3], çeşitli fizyolojik özelliklere dayalı inme oluşumunu tahmin etme amacıyla birden fazla makine öğrenmesi algoritmasının karşılaştırmalı bir analizini gerçekleştirmiştir. Veri kümesindeki eksik değerler, ilgili sütunun ortalama değeri kullanılarak doldurulmuştur, dengesiz sınıf dağılımı ise rastgele azaltma yöntemiyle dengelemiştir. Bu deneysel çalışmada, hastaları ve sağlıklı bireyleri sınıflandırmak için LR, DT, RF, KNN, SVM ve NB algoritmaları kullanılmıştır. Deneysel bulgulara dayanarak, yaklaşık %82 doğruluk oranıyla NB algoritması en iyi performansı göstermiştir. Ayrıca, elde edilen hassasiyet, duyarlılık ve F1-skorlarına göre, NB sınıflandırıcısının üstün performans sergilediği gözlemlenmiştir.

Al-Zubaid ve diğ. [4], dengesiz inme veri kümesini yeniden dengelemek için SMOTE tekniğini kullanırken, veri kümesindeki eksik değerlerin doldurulması için ortalama basit yöntemi kullanmıştır. Önerilen modeli test etmek için LR, RF, SVM, DT ve oy verme sınıflandırıcıları kullanılmıştır. Deneysel sonuçlara göre, RF algoritması diğer makine öğrenmesi modellerine göre %94,7 doğruluk oranı ile üstün bir performans sergilemiştir. Ayrıca, RF algoritması, hassasiyet, duyarlılık ve F1-skoru açısından üstün performans göstermiştir.

Liu, Fan ve Wu [5], eksik ve dengesiz bir tıbbi veri kümesinde inme oluşumunu tahmin etmeyi amaçlayan hibrit bir makine öğrenmesi yaklaşımı önermiştir. Bu yöntemde, dengesiz bir veri kümesi sorununu ele almak için otomatik hiper-parametre optimizasyonunu (AutoHPO) kullanılmıştır. Önerilen yaklaşım, iki temel aşamadan oluşmaktadır: eksik değerleri doldurma ve sınıflandırma. Veri kümesindeki eksik değerler, RF algoritması ile doldurulurken, Serebral İnme Tahmini (SİT) veri kümesinde inme tahmini için Derin Sinir Ağı (DNN) tekniğini kullanan AutoHPO önerilmiştir. Deneysel sonuçlara göre, model inme tahmininde dikkate değer doğruluk göstermiş ve yanlış negatif oranını azaltmıştır.

Wang ve Cheng [6], dengesiz tıbbi veri kümelerini ele almak için birden fazla yöntemi entegre eden hibrit bir yöntem önermiştir. Çalışmalarında, aşırı örnekleme için SMOTE, eksik örnekleme için Rastgele Eksik Örnekleme (RUS), özellik seçimi için PSO ve doğru sınıflandırma amacıyla MetaCost algoritmasını önerilmişlerdir. Deneysel çalışmada, önerilen yaklaşımın dokuz tıbbi veri kümesi üzerinde uygulanarak etkinliği değerlendirilmiştir. Deneysel bulgulara göre, önerilen yaklaşımın sınıf dengesizlik performansını önemli ölçüde artırma potansiyeline sahip olduğu belirlenmiştir.

Asniar, Maulidevi ve Surendro [7], dengesiz veri sorununu ele almak için SMOTE ve Yerel Aykırı Değer Faktörü (LOF) yöntemlerini entegre eden yeni bir yaklaşım olan SMOTE-LOF yöntemini önermiştir. Bu yöntemde, SMOTE yöntemi ile çoğunluk sınıflarında üretilen sentetik örnekler içerisinde gürültülü olanların tespit edilmesinde ve ortadan kaldırılmasında LOF yöntemi kullanılmıştır. Dengesiz veri kümeleri üzerindeki deneysel sonuçlara göre, SMOTE-LOF yöntemi, özellikle daha fazla sayıda örnek ve daha küçük dengesizlik oranına sahip veri kümelerinde doğruluk, F-skor ve AUC açısından SMOTE'dan daha iyi performans göstermiştir.

Pranto, Mehnaz, Momen ve Huq [8], diyabetin öngörülmesinde karşılaşılan sınıf dengesizliği sorununu ele almak için maliyet duyarlı öğrenme ve SMOTE yöntemini birleştiren bir yöntem önermiştir. Geliştirilen modeller PIMA Indians Diyabet (PID) veri kümesinde ve Bangladeş'in

Dhaka şehrinde bulunan Kurmitola Genel Hastanesi'nden elde edilen veri kümesinde test edilmiştir. Elde edilen bulgulara göre, önerilen yaklaşım diyabet tahmininin güvenilirliğini artırmıştır.

Chittora ve diğ. [9], Kronik Böbrek Hastalığı (KBH) veri kümesinde böbrek hastalığının tahmini için SMOTE ile desteklenmiş maliyet duyarlı öğrenme ve önemli özellik seçimi yöntemlerini içeren bir yaklaşım önermiştir. Yapılan deneylerde, SMOTE ile tam özelliklerin kullanılmasıyla lineer SVM en yüksek doğruluk oranını (%98,86) elde etmiştir. Ayrıca, LASSO regresyonu ile seçilen özelliklerin SMOTE ile birleştirilmesi, diğer sınıflandırıcılarla da iyi sonuçlar vermiştir. Derin sinir ağı modeli, en yüksek doğruluk oranını (%99,6) elde etmiştir.

Mienye ve Sun [10], tıbbi veri kümelerindeki sınıf dengesizliği sorununu ele almak için maliyet-duyarlı öğrenme yöntemlerini kullanarak sağlam sınıflandırıcılar geliştirmeyi önerdi. Dengesiz sınıf dağılımıyla başa çıkmak için, literatürde yaygın olarak kullanılan makine öğrenme yöntemlerinin amaç fonksiyonlarını değiştirerek orijinal yaklaşımlar geliştirildi. Bu yöntemin etkinliğini değerlendirmek için, PID, Haberman Meme Kanseri, Serviks Kanseri Risk Faktörleri ve KBH gibi tıbbi veri kümelerinde deneyler yapıldı. Deneysel sonuçlar, maliyet-duyarlı yöntemlerin standart algoritmalardan daha iyi performans gösterdiğini göstermektedir.

Gan ve diğ. [11], dengesiz tıbbi teşhis verileri için sınıflandırma performansını artırmak amacıyla maliyet-duyarlı bir sınıflandırma algoritması önermektedir. Deneysel çalışmalarında, modelin performansını değerlendirmek için Cleveland kalp, Hint karaciğer hastası, Dermatoloji ve Serviks kanseri risk faktörleri gibi veri kümeleri kullanılmıştır. Elde edilen deneysel bulgular, önerilen modelin diğer yöntemlere kıyasla daha yüksek etkinliğe sahip olduğunu ve klinik uzmanların daha etkili kararlar almasına yardımcı olabileceğini göstermektedir.

Bhattacharya ve diğ. [12], inme veri kümesindeki bilgi kalitesiyle ilgili zorlukların üstesinden gelmeyi amaçlayan bir ön işleme stratejisi önermektedir. Ön işleme prosedürü, işlenmemiş verilerin değiştirilmesi ve standartlaştırılması, gereksiz özelliklerin kaldırılması, herhangi eksik verilerin ele alınması ve dengesiz sınıf dağılımlarının düzeltilmesi gibi birçok adımı içermektedir. Ayrıca, DNN için optimal hiper-parametreleri etkili bir şekilde belirlemek için Antlion Optimization (ALO) algoritmasının kullanılması önerilmiştir. Önerilen modelin performansı diğer yaygın olarak kullanılan sınıflandırma modelleri ile karşılaştırılmıştır ve deneylerden elde edilen bulgulara göre, önerilen modelin eğitim süresi açısından daha etkili olduğu ve performans açısından diğer yöntemleri geride bıraktığı görülmüştür.

Xu, Shen, Nie ve Kou [13], tıbbi veri kümelerindeki sınıf dengesizliği problemine çözüm getirmek için SMOTE yönteminin örnek oluşturma ile ilgili kısıtlamalarını ele almayı amaçlayan yeni bir yaklaşım olan yanlış sınıflandırma sentetik azınlık aşırı örnekleme tekniğini önermişlerdir. Bu yöntem, dengesiz verileri örnekleme için RF algoritmasının avantajlarını SMOTE ve ENN algoritmalarıyla birleştirir. Deneysel çalışmalarda, önerilen yöntemin performansı ve geleneksel algoritmalar, model performans kriterlerine göre 10 farklı veri tıbbi veri kümesi üzerinde karşılaştırılmıştır. Elde edilen deneysel sonuçlara göre, önerilen metodoloji tıbbi teşhis alanında geleneksel algoritmalar ve diğer yeniden örnekleme yöntemlerine göre üstün performans göstermiştir.

Devarriya ve diğ. [14], dengesiz meme kanseri veri kümesi sınıflandırma sorununu ele almak için iki yeni uygun maliyetli uygunluk fonksiyonu (F2-skoru ve Mesafe skoru) önermiştir. Performans sonuçları, D-skoru ve F2-skorunu literatürdeki diğer yöntemlerle karşılaştırarak elde edilmiştir. Deneysel çalışmalarda, önerilen ölçütlerin Wisconsin Meme Kanseri veri kümesi üzerindeki performansı diğer mevcut model performans metrikleriyle karşılaştırılmıştır. Deneysel bulgulara göre, önerilen ölçütler, dengesiz veri kümeleriyle ilgili sınıflandırma sorununa yönelik tüm alternatif yöntemlere göre üstün performans sergilemektedir.

Khan ve Hoque [15], eksik veri doldurma için hem tekli hem de çoklu doldurma stratejilerini içeren bir hibrit strateji önermiştir. Bu strateji, yaygın MICE algoritmasının bir uzantısını kullanmaktadır. Önerilen algoritmaların performansı, UCI makine öğrenme deposundan alınan üç kamu veri kümesi ve Bangladeş'teki çeşitli hastanelerden ve tanı enstitülerinden toplanan 65.000 gerçek sağlık kaydının orijinal veri kümesi kullanılarak değerlendirilmiştir. Deneysel sonuçlar, önerilen yöntemin literatürdeki benzer yaklaşımları geride bıraktığını göstermekte olup, ikili veri doldurma için %20 daha yüksek bir F-skor ve sayısal veri doldurma için %11 daha düşük bir hata elde edilmiştir.

Raja ve Thangavel [16], eksik değerlerle başa çıkmak için kaba K-ortalama merkezi tabanlı bir doldurma tekniği önermiştir. Bu strateji, K-ortalama merkezi tabanlı, bulanık C-ortalama merkezi tabanlı, K-ortalama parametre tabanlı, bulanık C-ortalama parametre tabanlı ve kaba K-ortalama parametre tabanlı yöntemler de dahil olmak üzere çeşitli doldurma yaklaşımlarıyla karşılaştırılmıştır. Çeşitli tekniklerin performansı, Kök Ortalama Kare Hata (RMSE) ve Ortalama Mutlak Hata (MAE) ölçütleri kullanılarak incelenmiştir. Deneysel çalışmada,

Dermatoloji, PID, Wisconsin ve Yeast veri kümeleri kullanılmıştır. Önerilen yöntemin kullanışlılığı, birden fazla veri kümesinde gösterilmiş ve umut verici sonuçlar elde edilmiştir.

Madhu, Bharadwaj, Nagachandrika ve Vardhan [17], tıbbi veri kümelerinde eksik verilerin doldurulması için XGBoost kullanarak yeni bir yaklaşım önermiştir. Bu önerilen yöntem, sürekli ve kesikli veri özelliklerini uygun şekilde ele alabilmek için geliştirilmiştir. Sağlık veri kümelerindeki eksik değerlerle yapılan deneylerde, %1,98 ila %50,65 arasında değişen eksik değerlere sahip veri kümeleri kullanılmıştır. Önerilen yöntem, tekrarlı doldurma, KNN doldurma ve missForest doldurma yöntemleriyle karşılaştırılmıştır. Sonuçlara göre, önerilen missXGBoost tekniği, özellikle sürekli türlerde eksik veri özniteliklerini başarılı bir şekilde yönetmiş ve alternatif doldurma yöntemlerinden daha iyi performans göstermiştir.

Rani, Kumar ve Jain [18], tıbbi veri kümelerindeki eksik değerleri tahmin etmek için sınıflandırıcıyı optimize eden hibrit bir doldurma yöntemi olan HIOC'u önermiştir. Bu yaklaşım, eksik verileri doldurmak için MICE, KNN, ortalama ve mod doldurma yöntemlerini bir sınıflandırıcıyla entegre etmektedir. HIOC'un performansı, MICE, KNN, ortalama ve mod doldurma tekniklerini kullanarak SVM, NB, RF ve DT sınıflandırıcılarıyla karşılaştırılmıştır. Deneysel bulgular, HIOC yaklaşımının RMSE metriğini önemli ölçüde azalttığını göstermektedir. Kalp hastalığı veri kümesi için %17,32'lik bir RMSE azalması ve meme kanseri veri kümesi için %34,73'lük daha dikkate değer bir azalma görülmüştür. Ayrıca, önerilen teknik, önemli miktarda eksik veriyle karşı karşıya kaldığında bile iyi performans göstermektedir. Kalp hastalığı veri kümesinde sınıflandırma doğruluğu %18,61 artarken, meme kanseri veri kümesinde %6,20 artmıştır.

Wang ve diğ. [19], dengesiz veri ve eksik değerleri ele almak için DMP_MI yöntemini önermiştir. Bu yöntemde, eksik değerleri doldurmak için NB sınıflandırıcısı kullanılmış, dengeli sınıflar için ADASYN yöntemi uygulanmış ve son olarak bir bireyin diyabetik veya sağlıklı olma durumu RF sınıflandırıcısı kullanılarak tahmin edilmiştir. Deneysel bulgular, önerilen DMP_MI algoritmasının PID veri kümesinde diğer sınıflandırma yöntemlerine göre daha iyi performans elde ettiğini göstermektedir.

Yadav, Maravi, Agrawal ve Mishra [20], diyabet hastalığını tespit etmek için makine öğrenme yönteminin etkinliğini değerlendirmiştir. Bu çalışmada, sağlık veri kümelerinde sıklıkla karşılaşılan eksik veri, aykırı değerler ve dengesiz sınıf dağılımı gibi zorluklar ele alınmıştır. Bu zorlukları aşmak için ADASYN yöntemi ile sınıf dengesizliği azaltılmış ve eksik değerler

özniteliklerin ortalama değeri ile doldurulmuştur. Ayrıca, öznitelik seçimi teknikleri ve Çok Katmanlı Algılayıcı (MLP) sınıflandırıcıları kullanılarak genelleme ve tahmin doğruluğu artırılmıştır. Deneysel sonuçlar, nöral ağ modeli ile %84'lük bir doğruluk oranı elde edildiğini göstermektedir.

Alruily ve diğ. [21], SİT veri kümesinde inme tahmini için RF, XGBoost ve Hafif Gradyan Arttırma Makinesi (LightGBM) makine öğrenme algoritmalarını entegre eden bir ensemble RXLM önermişlerdir. Veri hazırlığı sırasında, veri kümesindeki eksik değerler KNN tekniği kullanılarak doldurulmuş, aykırı değerler silinmiş, öznitelikler tek yönlü kodlama yaklaşımıyla kodlanmış ve çeşitli değer aralıklarına sahip öznitelikler standart hale getirilmiştir. Veri dengesizliğini gidermek için SMOTE yöntemi kullanılmıştır. Hiper-parametreler rastgele arama yaklaşımıyla ayarlanmıştır. Ensemble RXLM'nin performansı doğruluk, AUC, duyarlılık, kesinlik, F1-skoru, Kappa ve MCC metrikleri kullanılarak değerlendirilmiştir. Sonuçlar sırasıyla %96,34, %99,38, %96,12, %96,55, %96,33, %92,68 ve %92,69 olarak belirlenmiştir.

Jing [22], dengesiz SİT veri kümesindeki azınlık sınıfı sınıflandırması üzerine bir çalışma yapmış ve serebral inme ile ilişkili risk faktörlerine odaklanmıştır. Ön işleme aşamasında, veri kümesindeki eksik değerler ortalama ve mod doldurma yöntemleriyle doldurulurken, dengesiz sınıf dağılımı SMOTE yöntemiyle dengelenmiştir. Deneysel sonuçlar, SMOTE ile birlikte Focal Loss ve k-ortalama (k-means) ile Temel Bileşen Analizi (PCA)'nin dengesiz veri kümeleriyle başa çıkmada üstün performans sergilediğini göstermektedir. Ayrıca, LR, Stokastik Gradyan İniş (SGD) ve DT algoritmalarının SMOTE ile birleştirildiğinde umut verici sonuçlar verdiği belirlenmiştir. DNN Focal Loss, diğer modellere göre PCA-k-means ile benzer performans sergilemesine rağmen, SMOTE ile LR ile benzer sonuçlar elde etmiştir. PCA-k-means Focal Loss ile DNN yaklaşımı %92 doğruluk, %77 F1-skoru ve %90 AUC elde etmiştir.

Rana, Chitre, Poyekar ve Bide [23], karmaşık makine öğrenmesi ve derin öğrenme tekniklerini kullanarak inme riskini tahmin etmek için bir sistem geliştirmişlerdir. SİT veri kümesindeki eksik değerler KNN doldurma yöntemleriyle doldurulurken, dengesiz sınıf dağılımı SMOTE ve Tomek-link yöntemiyle dengelemiştir. Önerilen model, %84 ROC skoru ile en iyi performansı sergilemiştir. Ayrıca, popülasyon tabanlı, ağaç tabanlı ve Naive Bayes tabanlı makine öğrenmesi algoritmaları da değerlendirilmiş ve önerilen Yapay Sinir Ağı (ANN) modelinin diğer algoritmaların önünde olduğu görülmüştür.

Maisha, Biswangri, Hossain ve Andersson [24], KBH veri kümesinde gürültülü ve tutarsız değerlerle başa çıkabilen bir yaklaşım önermişlerdir. Ön işleme aşamasında, sayısal özniteliklerdeki eksik değerleri ele almak için ortalama doldurma uygulanmış, kategorik eksik değerler için ise AWS laboratuvarı tarafından geliştirilen "datawig" adlı gelişmiş bir DNN modeli uygulanmıştır. Ayrıca, dengesiz sınıf dağılımı SMOTE yöntemiyle dengelenmiştir. Deneysel bulgular, RF sınıflandırıcısının %98,5 doğruluk elde ettiğini ve SVM, KNN, DT, NB ve LR sınıflandırıcılarına göre daha iyi performans elde ettiğini göstermektedir.

Wibowo ve Palupi [25], eksik değerlerle ve dengesiz veri kümeleriyle başa çıkmak için çeşitli ön işleme tekniklerini ve tahmin modellerini entegre eden bir yaklaşım önermişlerdir. Bu çalışmada, KBH veri kümesindeki eksik değerler doğrusal interpolasyonla doldurulurken, dengesiz sınıf dağılımı SMOTE yöntemiyle dengelenmiştir. Ayrıca, öznitelik seçimi için PCA ve Pearson Korelasyon Katsayısı (PCC) yaklaşımları uygulanmış, sağlıklı veya hasta bireyleri tahmin etmek için SVM ve LR sınıflandırıcıları kullanılmıştır. Deneysel sonuçlar, SMOTE yöntemi ve PCA'e dayalı öznitelik seçiminin en umut verici sonuçları verdiğini göstermektedir. LR algoritması, SVM sınıflandırıcısı ile karşılaştırıldığında %98,8 doğruluk, %100 kesinlik ve %98,77 F1-skoru ile en yüksek performansı elde etmiştir.

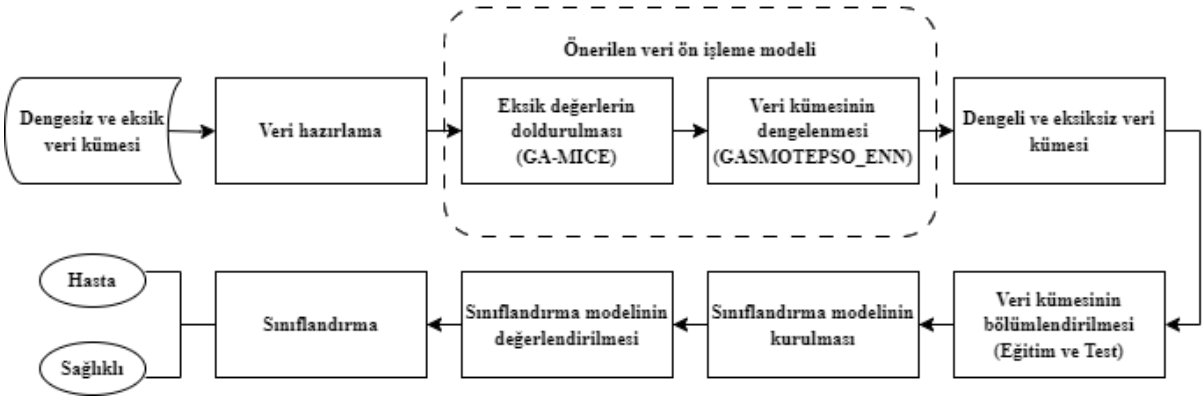
Imran, Amin ve Johora [26], KBH veri kümesindeki bireyleri hasta ve sağlıklı olarak sınıflandırmak için LR, İleri Beslemeli Sinir Ağı (FNN) ve Geniş+Derin modeli gibi üç makine öğrenme yöntemi uygulamışlardır. Veri ön işleme, eksik değerlerin doldurulması, öznitelik ölçeklendirme, dengesiz şekilde dağılmış sınıfların dengelenmesi ve veri kümesinin bölünmesi gibi aşamalar için deneysel çalışmalar yapılmıştır. Makine öğrenmesi sınıflandırma yöntemlerinin performansı, dengesiz ve dengeli veri kümelerinde F1-skoru, kesinlik, duyarlılık ve AUC skoru üzerinden değerlendirilmiştir. FNN modeli hem dengesiz hem de dengeli verilerde en iyi sonuçları vermiştir, %99 F1-skoru, %97 kesinlik, %99 duyarlılık ve %99 AUC skoru elde etmiştir. Ayrıca, Geniş+Derin modelin çoğu durumda FNN modeli ile benzer sonuçlar verdiği, LR'nin ise tüm modeller arasında en düşük skorları ürettiği gözlemlenmiştir.

Sağlık veri kümelerindeki dengesiz sınıf dağılımı ve eksik veri sorunları, makine öğrenmesi modellerinin performansını ciddi şekilde etkileyerek teşhis ve tedavi süreçlerini zorlaştırmaktadır. Literatürde, bu problemlere yönelik birçok yöntem ve çözüm önerilmiştir. Örneğin, dengesiz sınıf dağılımını dengelemek için SMOTE, RUS ve maliyet-duyarlı öğrenme teknikleri, eksik verileri doldurmak için ise ortalama doldurma, KNN, MICE ve gelişmiş

modeller kullanılmıştır. Ancak, bu yaklaşımlar hala bazı eksiklikler ve iyileştirme potansiyeli taşımaktadır. Sağlık verilerinin karmaşıklığı ve hassasiyeti göz önünde bulundurulduğunda, bu alanda yapılan çalışmaların önemi büyüktür ve daha fazla araştırmaya ihtiyaç duyulmaktadır. Bu tez çalışmasında, bu problemlere yenilikçi bir çözüm sunmak amacıyla, eksik verilerin doldurulması ve sınıf dengesizliğinin giderilmesi için sezgisel yöntemler ile literatürde başarıyla kullanılan teknikleri bir araya getiren bir ön işleme modeli önerilmiştir. Bu modelin, sağlık verilerindeki makine öğrenmesi modellerinin doğruluğunu ve güvenilirliğini artırarak, daha etkili teşhis ve tedavi planları oluşturulmasına katkıda bulunması hedeflenmektedir. Bu nedenle, sağlık veri analizi alanındaki bu konu, hala çalışmaya değer ve geliştirilmesi gereken önemli bir araştırma alanıdır.

3. YÖNTEM

Sağlık veri kümeleri genellikle eksik değerler ve dengesiz sınıf dağılımları gibi önemli problemlerle karşılaşmaktadır. Bu tür veri kümelerinin analizi, eksik veya dengesiz verilerin doğru bir şekilde işlenmesini ve makine öğrenmesi sınıflandırma algoritmalarıyla etkili bir şekilde kullanılmasını gerektirir. Bu nedenle, literatürde bu tür çalışmalara olan ihtiyaç oldukça belirgindir. Tez kapsamında, sağlık veri kümelerinde sıkça görülen eksik değerler ve dengesiz sınıf dağılımları gibi zorlukların üstesinden gelmek için bir veri ön işleme modeli önerilmiştir. Bu model, eksik değerlerin doldurulması, dengesiz sınıf dağılımlarının dengelenmesi ve makine öğrenmesi sınıflandırma algoritmaları kullanılarak bireylerin hasta veya sağlıklı olarak sınıflandırılması aşamalarından oluşmaktadır. Veri kümesindeki eksik değerler sorununu çözmek için literatürde başarıyla kullanılan MICE yöntemi GA sezgisel yöntemiyle birleştirilerek GA-MICE yaklaşımı geliştirilmiştir. Bu yaklaşım, eksik değerlerin veri kümesindeki sorununu çözmek için daha etkili bir strateji sunar. Dengesiz veri kümesi sorununu çözmek için GASMOTEPSO_ENN adlı sezgisel tabanlı hibrit bir örnekleme yaklaşımı geliştirilmiştir. Bu yöntem, SMOTE aşırı örnekleme ve ENN eksik örnekleme yaklaşımlarını etkili bir şekilde PSO ve GA sezgisel yöntemleriyle entegre ederek geliştirilmiştir. Önerilen veri ön işleme modelinin akış diyagramı Şekil 3.1'de verilmiştir.



Şekil 3.1: Önerilen veri ön işleme modelinin akış diyagramı.

Bu akış diyagramında, önerilen model PID, SİT ve KBH gibi sağlık veri kümelerindeki eksik değerleri doldurmak ve dengesiz sınıfları dengelemek için uygulanmıştır. Daha sonra, önerilen yöntemin performansı LR, RF, SVM, DT, NB ve XGBoost olmak üzere altı makine öğrenmesi

sınıflandırma algoritması kullanılarak doğruluk, kesinlik, duyarlılık, F1-skor, MCC ve AUC gibi model değerlendirme metrikleriyle değerlendirildi. Bu önerilen ön işleme modeli, sağlık veri kümelerindeki eksik değerlerin doldurulması ve dengesiz sınıfların dengelenmesi gibi önemli problemleri ele alırken, bireylerin doğru bir şekilde sınıflandırılmasını sağlamak için çeşitli makine öğrenmesi sınıflandırma algoritmalarıyla birlikte kullanılabilir. Bu yaklaşım, sağlık alanında veri analizi ve karar destek sistemlerinin geliştirilmesinde önemli bir adım olabilir.

Tez çalışması kapsamında kullanılan PID, SİT ve KBH veri kümeleri Bölüm 3.1'de ayrıntılı olarak tanıtılmaktadır. Bu bölümde, her bir veri kümesinin içeriği, özellikleri ve kullanılan değişkenler hakkında bilgi verilmiştir. Veri hazırlama süreci Bölüm 3.2'de ele alınmıştır. Bu bölümde, veri kümelerinin temizlenmesi, özelliklerin standartlaştırılması gibi işlemler ayrıntılı bir şekilde açıklanmıştır. Eksik veri ve dengesiz dağılımlı sınıf problemleri ile başa çıkmak için geliştirilen veri ön işleme modeli Bölüm 3.3'te detaylı olarak incelenmiştir. Bu bölümde, önerilen modelin temel prensipleri, sözde kodu, kullanılan yöntemler ve algoritmalar ayrıntılı bir şekilde açıklanmıştır. Önerilen modelin hasta ve sağlıklı bireylerin sınıflandırılmasında kullanılan çeşitli makine öğrenmesi sınıflandırma algoritmalarının parametre uzayı Bölüm 3.4'te verilmiştir.

3.1. VERİ KÜMELERİ

Tez çalışması kapsamında, veri analizi sürecinde karşılaşılan kritik iki önemli problem olan eksik değerler ve sınıf dengesizliği, Kaggle [68] platformundan elde edilen PID, SİT ve KBH veri kümeleri kullanılarak incelenmiştir. Her bir veri kümesinin detaylarına alt başlıklarda detaylı olarak yer verilmiştir.

3.1.1. PIMA Indians Diyabet (PID) Veri Kümesi

PID veri kümesi, Arizona'daki PIMA Kızılderili kabilesinden gelen 21 yaşından büyük kadınların diyabet durumunu belirlemek için kullanılan bir diyabet tahmini veri kümesidir [69]. Veri kümesi, toplamda 768 gözlem içermekte olup bunların 268'i diyabet hastası ve 500'ü sağlıklı bireylerden oluşmaktadır. PID veri kümesi, hasta ve sağlıklı bireyleri belirlemeyi mümkün kılan belirgin teşhis parametreleri ve ölçümler içermektedir. Diyabeti tahmin etmeye önemli katkı sağlayan 8 etkili özellik arasında hamilelik sayısı, Vücut Kitle İndeksi (BMI),

insulin seviyesi, yaş, kan basıncı, deri kalınlığı, glikoz ve diyabet aile geçmişi fonksiyonu bulunmaktadır. PID veri kümesine ait ilk beş örnek Tablo 3.1'de sunulmuştur.

Tablo 3.1: PID veri kümesine genel bakış.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Veri kümesi, hasta ve sağlıklı bireyler arasındaki dengesiz dağılımı ifade eden sınıf dengesizliği ile bazı biyolojik özelliklere ait ölçümlerde eksik veri değerleri olmak üzere veri analizi açısından dikkate değer iki probleme sahiptir. Veri kümesindeki eksik değerler ve sınıf dengesizliği problemleri, veri analizi ve makine öğrenmesi modellerinin genelleme sürecini etkileyebilir, bu da modelin doğru sonuçlar elde etme yeteneğini engelleyebilir. Tez çalışması kapsamında, eksik değerlerin yanı sıra, makine öğrenmesi algoritmaları üzerinde olumsuz etkilere neden olabilecek yanlış ölçümler "nan" eksik değer olarak değiştirilmiştir.

3.1.2. Serebral İnme Tahmini (SİT) Veri Kümesi

SİT veri kümesi, çeşitli giriş parametrelerine dayanarak bir hastanın inme geçirme olasılığını tahmin etmek için kullanılan bir veri kümesidir [70]. Veri kümesi, 249 hasta ve 4.861 sağlıklı bireylerden oluşan toplam 5.110 gözlem içermektedir. Bu veri kümesi, çeşitli demografik özellikler, tıbbi geçmiş ve yaşam tarzı alışkanlıkları gibi çeşitli özniteliklerin yanı sıra hastaların inme geçirme durumlarını içerir. Özellikle, cinsiyet, yaş, hipertansiyon, kalp hastalığı öyküsü, evlilik durumu, iş türü, ikamet türü, sigara içme alışkanlığı, ortalama glukoz seviyesi, BMI gibi faktörler bu veri kümesinde yer alır. SİT veri kümesine ait ilk beş örnek Tablo 3.2'de sunulmuştur.

Tablo 3.2: SİT veri kümesine genel bakış.

id	gender	age	hypertension	heart disease	ever married	work type	Residence type	avg glucose level	bmi	smoking status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1

Veri kümesi, dengesiz sınıf dağılımı ve eksik değerler gibi iki önemli problemi içermektedir. Tez kapsamında, modelin doğruluğunu ve güvenilirliğini artırmak için eksik değerler ve dengesiz sınıflarla başa çıkma stratejileri geliştirilmiştir.

3.1.3. Kronik Böbrek Hastalığı (KBH) Veri Kümesi

KBH veri kümesi, kronik böbrek hastalığını tahmin etmek için kullanılan veri kümesidir [71]. Bu veri kümesi, 150 sağlıklı ve 250 hasta bireylerden oluşan toplamda 400 gözlem içermektedir. Veri kümesi içerisinde bireylerin kan testleri ve diğer ölçümlerini kapsayan klinik özellikleri ile demografik bilgilerini içeren toplam 25 özellik bulunmaktadır. Bu özellikler arasında yaş, kan basıncı, özgül ağırlık, albümin, kan şekeri, kırmızı kan hücreleri, irin hücresi, irin hücresi pıhtıları, bakteriler, rastgele kan glukozu, kan üre, serum kreatinin, sodyum, potasyum, hemoglobin, paketlenmiş hücre hacmi, beyaz kan hücresi sayısı, kırmızı kan hücresi sayısı, hipertansiyon, diyabet, koroner arter hastalığı, iştah, pedal ödemi, anemi ve sınıflandırma etiketi bulunmaktadır. KBH veri kümesine ait ilk beş örnek Tablo 3.3'te sunulmuştur.

Tablo 3.3: KBH veri kümesine genel bakış.

id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	48	80	1.02	1	0		normal	notpresent	notpresent	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	7	50	1.02	4	0		normal	notpresent	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
2	62	80	1.01	2	3	normal	normal	notpresent	notpresent	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	51	80	1.01	2	0	normal	normal	notpresent	notpresent	106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd

KBH veri kümesi, sınıf dengesizliği ve eksik değerler gibi iki önemli problemi içermektedir. Tez çalışması kapsamında, bu iki problem detaylı bir şekilde ele alınmıştır.

3.2. VERİ HAZIRLAMA

Veri analizinde veri hazırlama, analiz sürecinin önemli bir aşamasıdır ve genellikle "veri ön işleme" olarak bilinmektedir. Bu aşama, veri analizi için kullanılacak veri kümesinin temizlenmesi, dönüştürülmesi, düzenlenmesi ve hazırlanması süreçlerini kapsamaktadır. Veri hazırlama işlemi, veri kümesinin eksik veya hatalı değerlerin ele alınması, dengesiz dağılımlı sınıfların dengelenmesi, gereksiz veya tekrarlayan özelliklerin çıkarılması, özelliklerin standartlaştırılması veya ölçeklendirilmesi gibi işlemleri içerir.

Tez çalışması kapsamında, PID, SİT ve KBH sağlık veri kümelerinin makine öğrenmesi sınıflandırma algoritmalarına uygun hale getirilmesi için bir dizi ön işleme adımı

gerçekleştirilmiştir. İlk olarak, veri kümesindeki hasta kimliğini temsil eden "id" gibi gereksiz özellikler tespit edilerek veri kümesinden çıkarılmıştır. Ardından, eksik değerlerin ve sınıf dengesizlik oranlarının analizi yapılmıştır. Veri kümesindeki kategorik özellikler scikit-learn kütüphanesinde bulunan kodlama teknikleri kullanılarak nümerik değerlere dönüştürülmüştür. Tüm özellikler, veri kümesindeki değerlerin dağılımını koruyacak şekilde standart ölçekleme yöntemiyle ölçeklendirilmiştir. Veri kümesindeki dengesiz dağılımlı sınıflar ve eksik değerler tez kapsamında geliştirilen veri ön işleme modeli kullanılarak dengelenmiş ve tamamlanmıştır. Son olarak, veri kümesindeki gereksiz özellikler IQR yöntemiyle belirlenmiş ve bu özellikler, tezde önerilen eksik değer doldurma yöntemiyle yeniden doldurulmuştur. Bu aşama, veri kümesinin eğitim ve test veri kümelerine bölünmesiyle tamamlanmıştır. Veri kümesi, eğitim için %67 ve test için %33 oranında bölünmüştür. Bu bölünme işlemi, modelin eğitiminde kullanılacak verilerin ayrılması ve geri kalan verilerin ise modelin performansının değerlendirilmesi için kullanılması amacıyla gerçekleştirilmiştir. Veri hazırlığında gerçekleştirilen tüm ön işleme adımları, veri kümelerinin makine öğrenmesi modelleri için analiz edilmeye hazır hale getirilmesini sağlayarak bu modellerin daha güvenilir sonuçlar elde etmeyi amaçlamaktadır.

3.3. GELİŞTİRİLEN VERİ ÖN İŞLEME MODELİ

Bu bölümde, çeşitli sağlık veri kümelerinde eksik değerlerden ve dengesiz dağıtılmış sınıflardan kaynaklanan sınıflandırma zorluklarını ele almak amacıyla bir hibrit ön işleme modeli geliştirilmiştir. Bu model, eksik değerlerin ataması için literatürde başarılı olarak kullanılan MICE yönteminin GA ile iyileştirilmesiyle oluşturulan GA-MICE yöntemini içermektedir. Ayrıca, dengesiz dağılıma sahip sınıfların ele alınması için SMOTE aşırı örnekleme ve ENN eksik örnekleme yöntemlerinin GA ve PSO sezgiselleri ile entegre edilmesiyle oluşturulan GASMOTEPSO_ENN yöntemi geliştirilmiştir. Önerilen ön işleme modeli aşağıda detaylı bir şekilde açıklanmaktadır.

3.3.1. GA-MICE Eksik Değer Atama Yöntemi

Tez kapsamında, sağlık veri kümelerinde sıklıkla karşılaşılan eksik veri problemini ele almak için yeni bir metodoloji sunulmaktadır. Önerilen yöntem, eksik değerleri tahmin etmek için GA ile MICE yöntemini entegre ederek eksik verilerin etkin bir şekilde tahmin edilmesini amaçlamaktadır. Geliştirilen GA-MICE yöntemi, eksik veri atama sürecindeki doğruluğu ve verimliliği artırmayı hedeflemektedir. Bu tez çalışması, eksik veri analitiği alanında yeni bir

bakış açısı sunmakta ve pratik uygulamalarda değerli bir katkı sağlamayı amaçlamaktadır. Önerilen GA-MICE eksik değer atama yönteminin sözde kodu Şekil 3.2'de sunulan Algoritma 2'de verilmiştir.

Algoritma 2, klasik MICE eksik değer atama yönteminden iki açıdan farklılık gösterir. İlk olarak, Algoritma 1 veri kümesindeki eksik değerlerin oranlarını belirleyerek başlar ve ardından değişkenleri eksik değer oranlarına göre en küçükten en büyüğe doğru sıralar. Bu sıralama işlemi, eksik değerlerin doldurulması sürecinde değişkenlerin öncelik sırasını belirlemeye yardımcı olur ve atanan değerlerin veri kümesinin geneline daha tutarlı bir şekilde yayılmasını sağlar. İkinci olarak, eksik değerlerin başlangıç parametreleri GA kullanılarak belirlenir. Bu adım, eksik verilerin doldurulması sürecinde temel bir noktayı oluşturur ve GA'nın optimizasyon gücünden yararlanarak eksik değerler için uygun başlangıç değerlerinin bulunmasını sağlar. Bu adımda, GA, eksik verileri doldurmak için rastgele oluşturulan bireylerin bir popülasyonunu oluşturarak başlar. Ardışık nesiller boyunca, her bireyin performansı, her iterasyonda bir uygunluk fonksiyonu kullanılarak değerlendirilir. Uygunluk fonksiyonu, her birey tarafından oluşturulan veri kümesi ile, eksik veriler için doldurulan değerleri dikkate alarak orijinal veri kümesi arasındaki benzerliği ölçer. Çaprazlama ve mutasyon operatörlerinin uygulanması, en başarılı bireylerin sonraki nesillere yayılmasını sağlar. Bu tekrarlayan süreç, belirli bir iterasyon sayısı tamamlanana kadar devam eder. Son olarak, GA, uygunluk fonksiyonuna dayanarak en uygun bireyi belirler ve bu bireyin genetik materyali, eksik verileri doldurmak için MICE algoritmasının optimal başlangıç değerlerini belirlemek için kullanılır. Daha sonra, eksik değerlerin doldurulma süreci MICE algoritması uygulanarak devam eder. Algoritmanın temel tekrarlayıcı yapısı, her değişken için eksik değerlerin doldurulmasını sağlamak üzere tasarlanmıştır. Her iterasyonda, birincil değişken (V) seçilir ve bu değişkenin eksik değerleri, önceki iterasyondan elde edilen doldurulmuş değerlerle diğer değişkenlerdeki doldurulmuş değerlerle değiştirilir. Daha sonra, her biri bağımsız olarak seçilen diğer değişkenler (U) üzerinde doğrusal regresyon modelleri oluşturulur. Bu regresyon modelleri, değişken U'nun eksik değerlerini tahmin etmek için kullanılır ve tahmin edilen değerler eksik değerlerle değiştirilir. Bu aşama, değişken ilişkilerinin anlaşılmasını geliştirir ve eksik değerlerin doldurulmasını kolaylaştırır. Yakınsama kontrolü, her iterasyonda GA-MICE tarafından doldurulan veriler ile bir önceki iterasyondan elde edilen doldurulmuş veriler arasındaki farkın değerlendirilmesiyle gerçekleştirilir. Bu fark, belirli bir önceden belirlenmiş eşik değerinden küçükse, yakınsama sağlanır ve iterasyonlar sona erer. Bu, algoritmanın kararlı

bir çözüme ulaştığını ve artık değişikliklere ihtiyaç duymadığını gösterir. Algoritma 2'de kullanılan parametre uzayı Tablo 3.4'te verilmiştir.

Algorithm 2 GA-MICE Eksik Değer Atama Algoritması

```
function GA_MICE(veri, max_iterasyon=100, yakınsama_eşiği=1e-4)
    Eksik değer oranlarını belirle ve değişkenleri küçükten büyüğe sırala.
    Belirli başlangıç değerlerini kullanarak eksik değerleri GA ile doldur.
    iterasyon = 0
    while iterasyon < max_iterasyon do
        for her değişken V için veri do
            for her değişken U için veri (V dışında) do
                Bağımlı değişken olarak V'yi seç.
                U içindeki eksik değerleri önceki iterasyondan
                doldurulan değerlerle değiştir.
                U ve V'nin eksik olmayan değerleri üzerinde bir
                regresyon modeli oluştur.
                U içindeki eksik değerleri tahmin et.
            end for
        end for
        Eksik değerleri V içinde tahmin edilen değerlerle güncelle.
        if iterasyon > 0 then
            fark = |GA ile doldurulan_veri − önceki_doldurulan|
            if Farkın maksimum değeri < yakınsama_eşiği then
                yazdır("Yakınsama sağlandı, iterasyon: iterasyon + 1")
                break
            end if
        end if
        Güncel iterasyondan elde edilen doldurulan veriyi sakla.
        Önceki doldurulan veri = GA ile doldurulan verinin kopyası
        Iterasyon sayacını artır.
        iterasyon+ = 1
    end while
    return Eksik verileri tamamlanmış veri
end function
```

Şekil 3.2: GA-MICE eksik değer atama algoritmasının sözde kodu.

Tablo 3.4: GA-MICE yönteminin parametre değerleri.

Parametre	Değer
Popülasyon Boyutu	100
Çaprazlama Olasılığı	0.8
Mutasyon Olasılığı	0.2
Nesil Sayısı	100

GA-MICE algoritması, GA ve regresyon modellerinin etkin bir şekilde entegrasyonuyla eksik değer atama için güçlü bir çözüm sağlar.

3.3.2. GASMOTEPSO_ENN Yeniden Örnekleme Yöntemi

Tez kapsamında, sağlık veri kümelerinde sıkça karşılaşılan dengesiz sınıf dağılımlarını ele almak için sezgisel tabanlı GASMOTEPSO_ENN isimli bir hibrit örnekleme yöntemi geliştirilmiştir. Bu yöntem, SMOTE aşırı örnekleme ve ENN eksik örnekleme yaklaşımlarını etkin bir şekilde GA ve PSO sezgisel yöntemleri ile entegre eder [72]. Bu yöntemde, ilk olarak GASMOTEPSO yöntemiyle veri kümesinde sentetik örnekler üretilir. Ardından, ENN eksik örnekleme yöntemi ile gürültülü veriler veri kümesinden çıkarılır. Son olarak, IQR yöntemi ile veri kümesindeki aykırı değerler tespit edilerek bu değerler Algoritma 2 ile yeniden atanır. Böylece, dengeli gürültüden arındırılmış bir veri kümesi elde edilir. Önerilen GASMOTEPSO_ENN yönteminin sözde kodu Şekil 3.3'te sunulan Algoritma 3'te verilmiştir.

Algoritma 3'te dengesiz veriler GASMOTE ve SMOTE-PSO aşırı örnekleme yöntemlerinin birleştirilmesiyle iki aşamada dengelenir. İlk aşamada, Jiang ve diğ. [73] tarafından önerilen GASMOTE ile veri aşırı örnekleme yapılır. Bu yöntemde, GA sezgiseli ve SMOTE aşırı örnekleme birleştirilerek dengeli bir veri oluşturulur. Bu işlem sırasında, optimal örnekleme oranını bulmak için GA algoritmasından yararlanılırken, SMOTE yöntemi optimize edilmiş örnekleme hızına dayalı olarak aşırı örnekleme yoluyla yeni bir veri kümesi oluşturur. Bu yaklaşım, GA algoritmasını kullanarak P boyutunda bir popülasyon üretir. Popülasyondaki her birey, tüm örnekler için N_i ile gösterilen örnekleme oranlarının bir kombinasyonunu yansıtır. Kromozomal uzunluk M azınlık sınıfı örneklerinin sayısına karşılık gelirken, P popülasyon büyüklüğünü temsil eder. Kromozomun her düğümüne, tanımlanmış parametreler dahilinde rastgele bir tamsayı değeri atanarak başlatılır. Daha sonra her bireyin uygunluk fonksiyonu değerleri hesaplanır ve popülasyon azalan düzende sıralanır. En üst sıradaki bireyler yeni popülasyonu oluşturmak için kopyalanırken, en düşük sıradaki bireyler bir seçim olasılığı P_r kullanılarak silinir. Daha sonra popülasyondan rastgele seçilen bireylerin kromozomları rastgele belirlenen bir düğümde çaprazlanır ve popülasyona P_m olasılığıyla mutasyon uygulanır.

Algorithm 3 GASMOTEPSO_ENN Algoritması

Başla
Giriş: Eğitim kümesi $Egitim$
Veriyi normalize edin ve ön işleme yapın.
Veriyi dengeli hale getirmek için GASMOTE algoritmasını uygulayın.
    GASMOTE için P bireylik Popülasyon Pop'u başlatın.
    $EnIyiSmotedEgitim$'i boş bir küme olarak ayarlayın.
while Sonlandırma koşulu sağlanana kadar **do**
    X_i'ye dayanarak $SmotedEgitim_i$ üretin.
    Geçerlilik kullanarak uygunluk değerini değerlendirin.
    Eğer uygunluk geliştirildiyse, $EnIyiSmotedEgitim$'i güncelleyin.
    Seçim, çaprazlama ve mutasyon gerçekleştirin.
end while
$EnIyiSmotedEgitim$'i çıkış olarak alın.
Daha fazla dengeleme için SMOTE-PSO algoritmasını uygulayın.
    k-katlamalı çapraz doğrulama kullanarak ayırın: $EnIyiSmotedEgitim$, $Test$.
    $EnIyiSmotedEgitim$ üzerinde SVM eğitin, SVs elde edin.
    PSO'yu SVs ile başlatın, popülasyon oluşturun.
    PSO ile SVM parametrelerini optimize edin.
    Son SVM performansını değerlendirin.
Gürültüyü kaldırmak için ENN algoritmasını uygulayın.
    Giriş: Eğitim kümesi $EnIyiSmotedEgitim$.
for her örnek $EnIyiSmotedEgitim$ içinde **do**
    k en yakın komşuyu bulun.
    if Çoğunluk sınıf etiketi aynı ise **then**
        Örneği kaldırın.
    end if
end for
Çıkış olarak $EnIyiDengelenmisEgitim$'i alın.
Aykırı örnekleri kaldırmak için IQR kullanın.
    Her özellik için IQR hesaplayın.
for Her bir özellik **do**
    IQR sınırlarında dışarıda kalan örnekleri tanımlayın.
    Aykırı örnekleri $EnIyiDengelenmisEgitim$'den GA-MICE ile yeniden tanımlayın.
end for
Son olarak, dengelenmiş ve gürültü filtrelenmiş eğitim kümesini çıkış olarak alın: $EnIyiDengelenmisEgitim$.
Bitir

Şekil 3.3: GASMOTEPSO_ENN yönteminin sözde kodu.

Bir mutasyon meydana gelirse, bireyin kromozomundaki rastgele bir düğüm seçilir ve değeri buna göre değiştirilir. Son olarak sonlandırma kriterleri açısından incelenir. Mevcut nesil maksimum nesilden fazlaysa, algoritma optimal bireyi çıkarır; aksi takdirde uygunluk fonksiyonu adımına geri döner. Sonlandırmanın ardından, SMOTE aşırı örnekleme kullanılarak yeni bir veri kümesi oluşturmak için uygun örnekleme oranları kullanılır. İkinci aşamada, GASMOTE ile dengeli bir veri kümesi elde edilirken sınıflandırıcı doğruluğunu ve verimliliğini artırmak için SMOTE aşırı örnekleme yöntemini PSO sezgiseliyle birleştirilen Cervantes ve

diğ. [74] tarafından tanıtılan SMOTE-PSO algoritması ile veri aşırı örnekleme yapılır. Öncelikle, GASMOTE yöntemiyle elde edilen veri kümesi azınlık ve çoğunluk sınıflarına ayrılır. Daha sonra, eğitim veri kümesiyle eğitilmiş bir SVM kullanılarak bir hiper düzlem oluşturulur ve destek vektörler belirlenir. SMOTE algoritması ile yeni sentetik örnekler üretilirken, bu örnekler PSO algoritması kullanılarak optimize edilmiş bir hiper düzlemle ilişkilendirilir. Böylece, SMOTE-PSO yöntemi, SVM'in kenar bölgesinde sentetik örnekler üreterek dengeli bir veri kümesi elde edilir. GASMOTEPSO yöntemiyle veri kümesinde sentetik örneklerin oluşturulmasının ardından veri kümesindeki gürültülü örnekler ENN eksik örnekleme yöntemi ile veri kümesinden çıkarılır. ENN yöntemi, veri kümesindeki her bir örneği inceleyerek komşuları tarafından yanlış sınıflandırılmış örnekleri belirler ve veri kümesinden kaldırır. Bu adım, makine öğrenmesi yöntemlerinin öğrenme sürecine gürültü veya önyargı ekleyebilecek gürültülü örnekleri silme yoluyla veri kümesinin genel kalitesini artırmaya yardımcı olur. Son olarak, veri kümesindeki aykırı değerler IQR yöntemiyle tespit edilir ve Algoritma 2 ile yeniden belirlenerek dengeli ve kaliteli bir veri kümesi elde edilir. Algoritma 3'te kullanılan parametre uzayı Tablo 3.5'te verilmiştir.

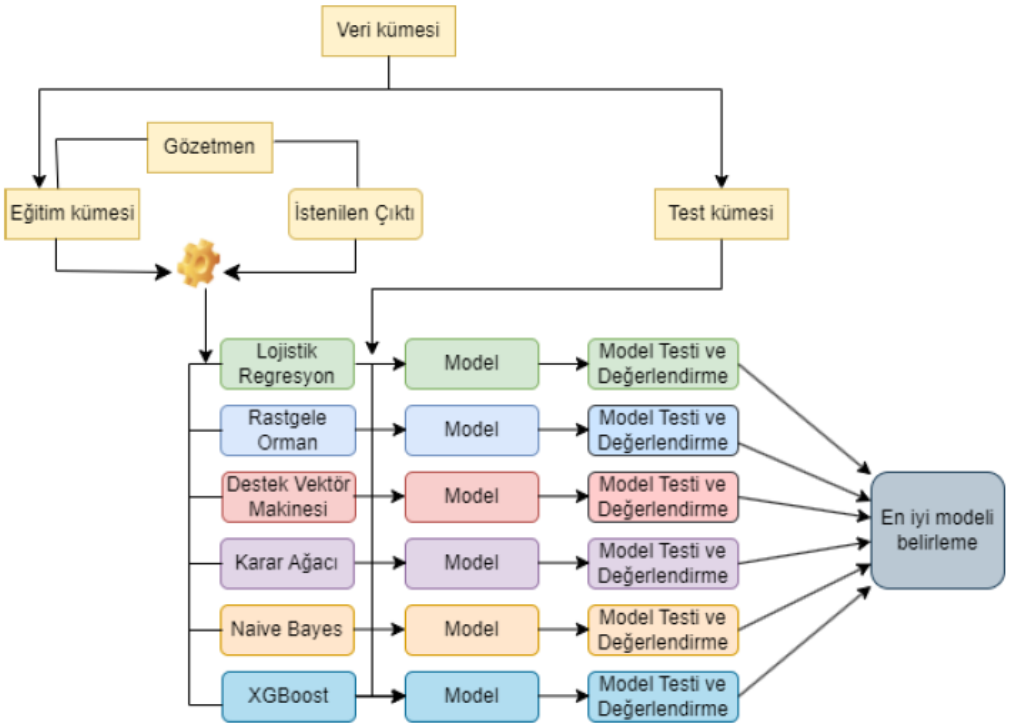
Tablo 3.5: GASMOTEPSO_ENN yönteminin parametre değerleri.

GASMOTE	**SMOTEPSO**	**ENN**
k_neighbors = 3	k = 3	sampling_strategy = 'auto'
population_size = 50	eps = 0.01	n_neighbors = 10
generations = 20	n_pop = 10	kind_sel = 'all'
crossover_rate = 0.8	w = 1.0	n_jobs = None
mutation_rate = 0.1,	c1 = 2.0	
maxn = 2	c2 = 2.0	
random_state = 42	num_iteration = 15	
	n_jobs = 1	
	random_state = 42	

GASMOTEPSO_ENN yöntemi, aşırı örnekleme ve eksik örnekleme tekniklerini bir araya getirerek azınlık sınıfının daha iyi temsil edilmesini sağlar ve aynı zamanda aşırı örnekleme bölgelerindeki gereksiz verileri azaltır. Bu, sınıflandırma modelinin dengesizlikle başa çıkma yeteneğini artırır ve daha dengeli ve güvenilir sonuçlar elde etmeyi mümkün kılar. Bu nedenle, özellikle sağlık veri kümeleri gibi dengesiz sınıf dağılımlarına sahip veri kümelerinde kullanıldığında, GASMOTEPSO_ENN yöntemi önemli bir rol oynar.

3.4. SINIFLANDIRMA MODELİ

Tez kapsamında, sağlık veri kümelerindeki bireylerin hasta veya sağlıklı olarak sınıflandırılması amacıyla LR, RF, SVM, DT, NB ve XGBoost makine öğrenmesi sınıflandırma algoritmaları kullanılmıştır. Bu algoritmalar, sağlık veri kümelerindeki bireylerin hastalık teşhisi veya sağlıklı olma durumunu sınıflandırmak için geniş bir uygulama yelpazesine sahip güçlü ve yaygın olarak kullanılan makine öğrenmesi sınıflandırma algoritmalarıdır. Genellikle, LR algoritması, basitlik ve yorumlanabilirlik açısından avantaj sağlarken, RF, SVM, DT, NB ve XGBoost gibi algoritmalar karmaşık ilişkileri yakalamak ve yüksek performans elde etmek için tercih edilir. Bu algoritmalar, farklı veri yapılarına uyum sağlayabilme ve sağlık veri kümelerindeki özellikleri etkin bir şekilde modelleyebilme yetenekleri nedeniyle bu tez çalışmasında seçilmiştir. Bu algoritmaların bir arada kullanılması, sağlık veri analizinde çeşitli perspektiflerin ve yaklaşımların birleştirilmesini sağlar, böylece daha kapsamlı sonuçlar elde edilir. Şekil 3.4, tez çalışmasında kullanılan denetimli öğrenme sürecini gösteren sınıflandırma modelini içermektedir.



Şekil 3.4: Sınıflandırma modelinin akış diyagramı.

Sınıflandırma algoritmalarının performansını optimize etmek için, her bir algoritmanın parametrelerinin doğru şekilde ayarlanması kritik bir öneme sahiptir. Bu bağlamda, her bir sınıflandırma algoritmasının en uygun parametrelere değerlerinin belirlenmesinde kapsamlı bir

parametre ayarlama süreci gerçekleştirilmiştir. Sınıflandırma algoritmaları için en uygun parametrelerin seçimi için scikit-learn kütüphanesinde bulunan GridSearchCV yaklaşımı kullanılmıştır. Tablo 3.6 'da her bir algoritmada kullanılan parametrelerin ve bu parametrelere atanmış değerlere ilişkin bilgiler sunulmuştur. Bu parametreler, modelin performansını etkileyen faktörler arasında yer alır ve uygun şekilde ayarlanmaları, en iyi sonuçların elde edilmesine yardımcı olur.

Tablo 3.6: Sınıflandırma algoritmalarında kullanılan parametreler ve değerleri.

Algoritma	Parametre	Değerler
LR	scaler	StandardScaler(), MinMaxScaler(), Normalizer(), MaxAbsScaler()
	penalty	'l1', 'l2', 'elasticnet', 'none'
	C	0.001, 0.01, 0.1, 1.0, 10, 100
	class_weight	'dict', 'balanced', 'None'
	random_state	42, 50, 100
	solver	'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga2'
	max_iter	10, 50, 100
RF	scaler	StandardScaler(), MinMaxScaler(), Normalizer(), MaxAbsScaler()
	n_estimators	10, 30, 50, 100
	criterion	'gini', 'entropy'
	max_features	'auto', 'sqrt', 'log2'
	random_state	42, 50, 100
	class_weight	'balanced', 'balanced_subsample', 'dict', 'None'
	bootstrap	0, 1
	max_depth	2, 3
SVM	scaler	StandardScaler(), MinMaxScaler(), Normalizer(), MaxAbsScaler()
	C	0.1, 1, 10, 100, 1000
	gamma	'scale', 'auto', 1, 0.1, 0.01, 0.001, 0.0001
	random_state	42, 50, 100
	kernel	'linear', 'rbf'
	class_weight	'dict', 'balanced', 'None'
	decision_function_shape	'ovo', 'ovr'
	degree	2, 3, 5
DT	scaler	StandardScaler(), MinMaxScaler(), Normalizer(), MaxAbsScaler()
	criterion	'gini', 'entropy'
	splitter	'best', 'random'
	max_features	'auto', 'sqrt', 'log2', 'None'
	random_state	42, 50, 100
	class_weight	'dict', 'balanced', 'None'
	ccp_alpha	0.0, 0.005, 0.010, 0.015, 0.030, 0.035
	max_depth	1, 2, 3, 4, 5, 6, 7, 'None'
NB	scaler	StandardScaler()
XGBoost	scaler	StandardScaler(), MinMaxScaler(), Normalizer(), MaxAbsScaler()
	max_depth	2, 3, 5
	n_estimators	10, 50, 100
	learning_rate	0.1, 0.01, 0.05
	colsample_bytree	0.5, 1, 2
	reg_alpha	0, 0.5, 1, 5
	reg_lambda	0, 0.5, 1, 5

Her bir algoritmanın parametrelerinin ayarlanmasının ardından, sınıflandırma performansı belirli model değerlendirme metrikleri kullanılarak titizlikle değerlendirilmiştir. Doğruluk, kesinlik, duyarlılık, F1-skoru, MCC ve AUC gibi ölçümler bu değerlendirme metrikleri arasında yer almaktadır. Her bir algoritmanın performansı bu metrikler ışığında detaylı bir şekilde analiz edilmiş ve birbirleriyle karşılaştırılmıştır. Yapılan karşılaştırmalar sonucunda, veri kümesindeki bireyleri en etkili şekilde sınıflandıran algoritma en iyi model olarak belirlenmiştir.

4. BULGULAR

Bu tez çalışmasında, sağlık veri kümelerinde yaygın olarak karşılaşılan eksik veri ve dengesiz sınıf dağılımı sorunlarını etkin bir şekilde çözmek için bir ön işleme yöntemi geliştirilmesi hedeflenmiştir. GA-MICE yöntemi ile eksik değerlerin ataması yapılırken, GASMOTEPSO_ENN yöntemi ile dengesiz sınıfların yeniden dengelenmesi sağlanmıştır. Önerilen yöntemin etkinliği, literatürde sıkça kullanılan LR, RF, SVM, DT, NB ve XGBoost gibi makine öğrenmesi sınıflandırma algoritmaları kullanılarak PID, SİT ve KBH sağlık veri kümeleri üzerinde değerlendirilmiştir. Bu veri kümeleri, çeşitli büyüklüklerde, eksik değer miktarlarına ve sınıf dengesizlik oranlarına sahiptir.

Bu tez kapsamında önerilen yöntemin performansını değerlendirmek için kapsamlı deneyler gerçekleştirilmiştir. Deneyler, AMD Ryzen 7 5700U işlemcili, 1.80 GHz hızında çalışan (8 çekirdekli), 8 GB RAM ve 64 bit Windows 11 Pro işletim sistemine sahip bir dizüstü bilgisayar üzerinde gerçekleştirilmiştir. Hesaplamalar, Jupyter Notebook ortamında Python programlama dili ve scikit-learn makine öğrenmesi kütüphanesi kullanılarak yapılmıştır. Bu bölümde, farklı sağlık veri kümelerindeki bireylerin hasta veya sağlıklı olarak sınıflandırılmasına ilişkin deneysel çalışmaların sonuçları detaylı bir şekilde sunulmaktadır. İlk olarak, farklı özelliklere sahip veri kümelerinden elde edilen bulgular, model değerlendirme metriklerini içeren bir tablo ve ROC eğrisini içeren grafiklerle sunulmuştur. Ayrıca, literatürde benzer çalışmaların sonuçları, tez kapsamında geliştirilen yöntemle karşılaştırılmıştır.

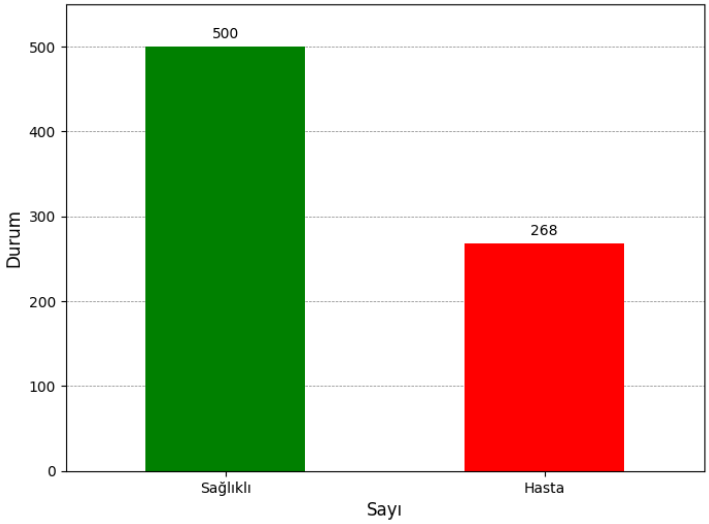
Sağlık veri kümelerinde eksik değer ve dengesiz sınıf dağılımı sorunlarını çözmek amacıyla önerilen yaklaşımların makine öğrenmesi yöntemleri kullanılarak bireylerin hasta veya sağlıklı olarak sınıflandırılması için üç farklı çalışma gerçekleştirilmiştir. İlk çalışmada, sağlıklı ve hasta bireylerin sınıflandırılmasında eksik değerler ve dengesiz sınıf dağılımı ele alınmamıştır. İkinci çalışmada, sağlıklı ve hasta bireylerin sınıflandırılmasında eksik değerlerin tamamlanması için MICE yöntemi, sınıfların dengelenmesi için tez kapsamında önerilen GASMOTEPSO_ENN yöntemi kullanılarak elde edilen sonuçlar, literatürdeki benzer çalışmalarla karşılaştırılmıştır. Üçüncü çalışmada ise hem eksik değerlerin tamamlanması hem de dengesiz sınıf dağılımının ele alınması için geliştirilen ön işleme yaklaşımı kullanılarak elde edilen sonuçlar, literatürdeki benzer çalışmalarla karşılaştırılmıştır. Önerilen yöntemin

etkinliğini değerlendirmek için doğruluk, kesinlik, duyarlılık, F1-skoru, MCC ve AUC gibi performans metrikleri kullanılmıştır.

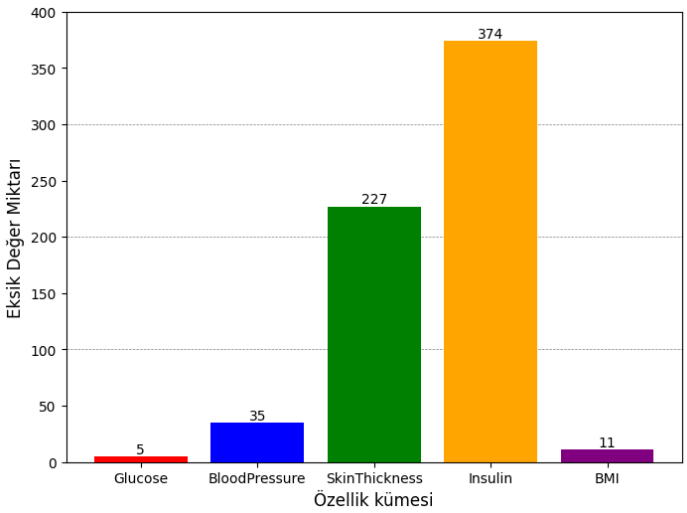
4.1. VERİ ÖN İŞLEME YÖNTEMLERİ KULLANILMADAN FARKLI SINIFLANDIRICILARIN HASTALIK TESPİTİ ÜZERİNDE PERFORMANSLARININ KARŞILAŞTIRILMASI

Sağlık veri kümelerinde dengesiz dağılımlı sınıflar ve eksik değerler sıklıkla karşılaşılan zorluklardır. Bu deneyde, bu zorlukların üstesinden gelmek için herhangi bir ön işleme yöntemi kullanılmadan altı farklı makine öğrenmesi sınıflandırma yönteminin performansı analiz edilmiştir. Bu deney, dengesiz dağılımlı sınıfların ve eksik verilerin olduğu veri kümelerinde, sağlıklı ve hasta bireylerin doğru bir şekilde sınıflandırılmasında ön işleme yöntemlerinin etkisini araştırmayı hedeflemektedir. Sınıflandırma algoritmaların hasta ve sağlıklı bireyleri ayırt edebilme yetenekleri model değerlendirme metriklerine göre ölçülmüş ve sonuçlar değerlendirilmiştir.

PID, SİT ve KBH veri kümelerine ait her sınıf için veri dağılımı sırasıyla Şekil 4.1, Şekil 4.4 ve Şekil 4.7'de gösterilmiştir. Eksik değer miktarlarına ait özellik kümeleri ise Şekil 4.2, Şekil 4.5 ve Şekil 4.8'de sunulmuştur. Bu şekiller, her bir veri kümesinin özelliklerini ve veri bütünlüğünü açıkça gösterirken, sınıflar arasındaki dengesizliği ve eksik veri durumunu da vurgulamaktadır. Veri kümelerinin hazırlığı sürecinde, kullanılan veri kümelerindeki kategorik değişkenlerin sayısal değerlere dönüştürülmesi için scikit-learn kütüphanesinde bulunan bir-etiket kodlama tekniği kullanılmıştır. Sürekli değişkenler ise standart ölçekleme tekniğiyle ölçeklendirilmiştir. Ancak, dengesiz sınıfları düzeltmek, eksik ve aykırı değerleri belirlemek ve kaldırmak için herhangi bir ön işleme yöntemi uygulanmamıştır. Ardından, tüm veri kümeleri %33 test oranıyla eğitim ve test alt kümelerine bölünmüş ve bu alt kümeler makine öğrenmesi sınıflandırma modellerine giriş olarak sağlanmıştır. Doğruluk, kesinlik, duyarlılık, F1-skor ve MCC model değerlendirme metriklerine göre elde edilen test sonuçları Tablo 4.1, Tablo 4.2 ve Tablo 4.3'te sunulmuştur. MCC metriğine göre en iyi sonuçlar her bir veri kümesi için kalın yazı tipiyle vurgulanmıştır. Bu tablolara ek olarak, sınıflandırıcıların PID, SİT ve KBH veri kümeleri üzerindeki performansının ROC eğrisi altında kalan alan (AUC) değerlerine ait grafikler sırasıyla Şekil 4.3, Şekil 4.6 ve Şekil 4.9 'da verilmiştir.



Şekil 4.1: PID veri kümesindeki bireylerin dağılımı.

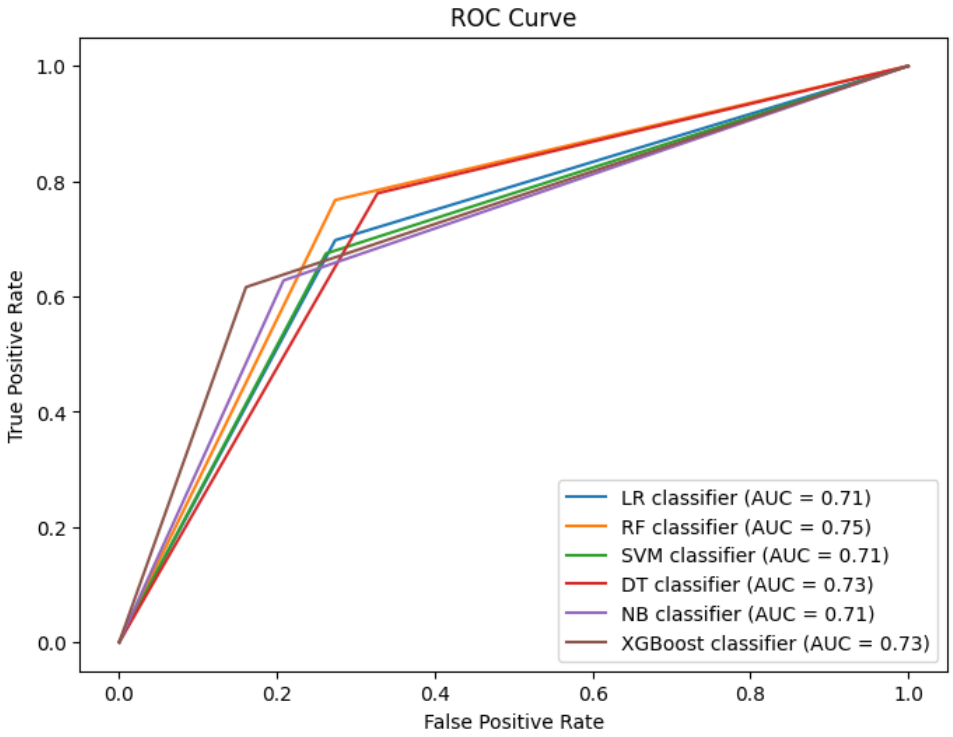


Şekil 4.2: PID veri kümesindeki eksik özelliklerin miktarı.

Tablo 4.1: Sınıflandırıcıların PID veri kümesi üzerindeki performansının değerlendirilmesi.

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	0.72	0.57	0.70	0.62	0.41
RF	**0.74**	**0.59**	**0.77**	**0.67**	**0.47**
SVM	0.72	0.57	0.67	0.62	0.40
DT	0.71	0.55	0.78	0.64	0.43
NB	0.74	0.61	0.63	0.62	0.42
XGBoost	0.76	0.66	0.62	0.64	0.46

Tablo 4.1'de analiz edilen verilere dayanarak, PID veri kümesi üzerinde farklı sınıflandırıcıların performansı değerlendirilmiştir. Sınıflandırıcıların doğruluk, kesinlik, duyarlılık, F1-skor ve MCC gibi metrikler üzerinden performansları ölçülmüştür. RF sınıflandırıcısı, diğer yöntemlere kıyasla daha yüksek bir doğruluk (0,74) ve MCC (0,47) değerleri sergilemiştir. Bu bulgular, RF'nin PID veri kümesi üzerinde daha üstün bir genel performans sergilediğine işaret etmektedir. Ayrıca, RF'nin yüksek bir duyarlılık değeri (0,77) ile hasta bireyleri daha etkin bir şekilde tanıma eğiliminde olduğu belirlenmiştir. XGBoost sınıflandırıcısı da dikkate değer yüksek bir doğruluk (0,76) ve kesinlik (0,66) değerleri ile RF'ye yakın bir performans sergilemektedir. Ancak, diğer metriklerde (duyarlılık, F1-skor ve MCC) RF'nin biraz gerisinde kalmıştır. Ayrıca, sınıflandırıcıların PID veri kümesi üzerindeki performansının AUC değerlerine ait sonuçları Şekil 4.3'te verilmiştir.

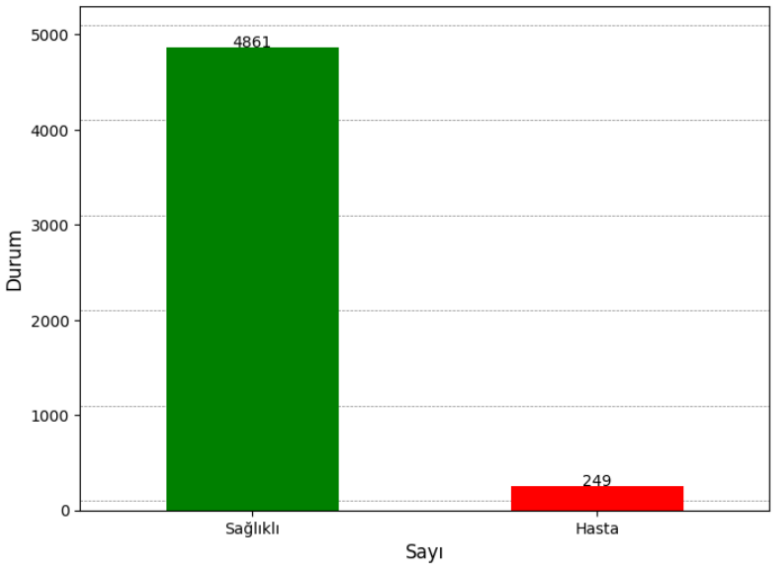


Şekil 4.3: PID veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.

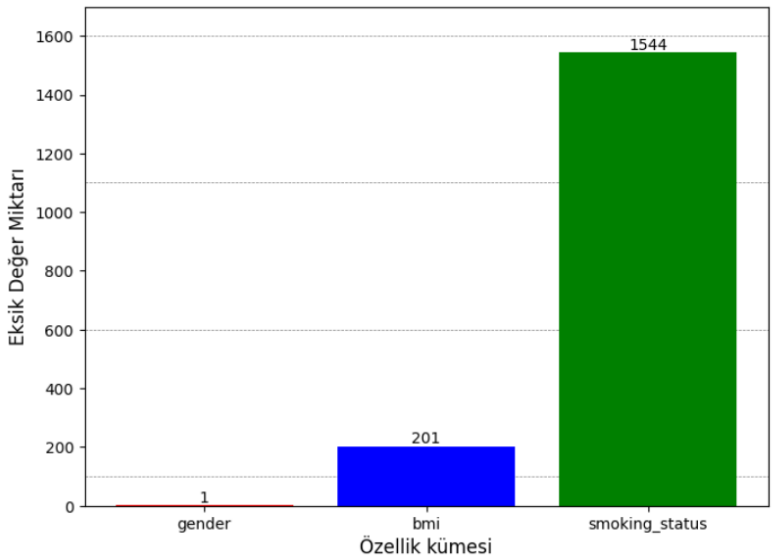
Sonuçlar incelendiğinde, RF sınıflandırıcısı en yüksek AUC değerine sahiptir (0,75), bu da modelin PID veri kümesi üzerindeki performansının diğer modellere kıyasla daha iyi olduğunu göstermektedir. Diğer sınıflandırıcılar arasında ise XGBoost ve DT, RF'ye yakın AUC değerleri elde etmiştir. Sınıf dengesizliği göz önüne alındığında, sınıflandırıcıların performansları yanıltıcı olabilmekte ve çoğunluk sınıfına ağırlık verilmesi ile azınlık sınıfın performansı olumsuz etkilenebilmektedir. Bu nedenle, sınıf dengesizliği ile başa çıkabilen metriklerin

(örneğin, MCC) kullanılması daha sağlıklı sonuçlar elde etmeye yardımcı olabilir. Eksik verilerin varlığı da model performansını etkileyebilir. Özellikle, "skinthickness" ve "insulin" gibi önemli özelliklerdeki yüksek eksik değer miktarı, modelin bu özellikleri dikkate alırken zorluklarla karşılaşabileceğini işaret etmektedir.

SİT veri kümesine ait sınıf dağılımı Şekil 4.4'te ve eksik değer miktarları Şekil 4.5'te verilmiştir.



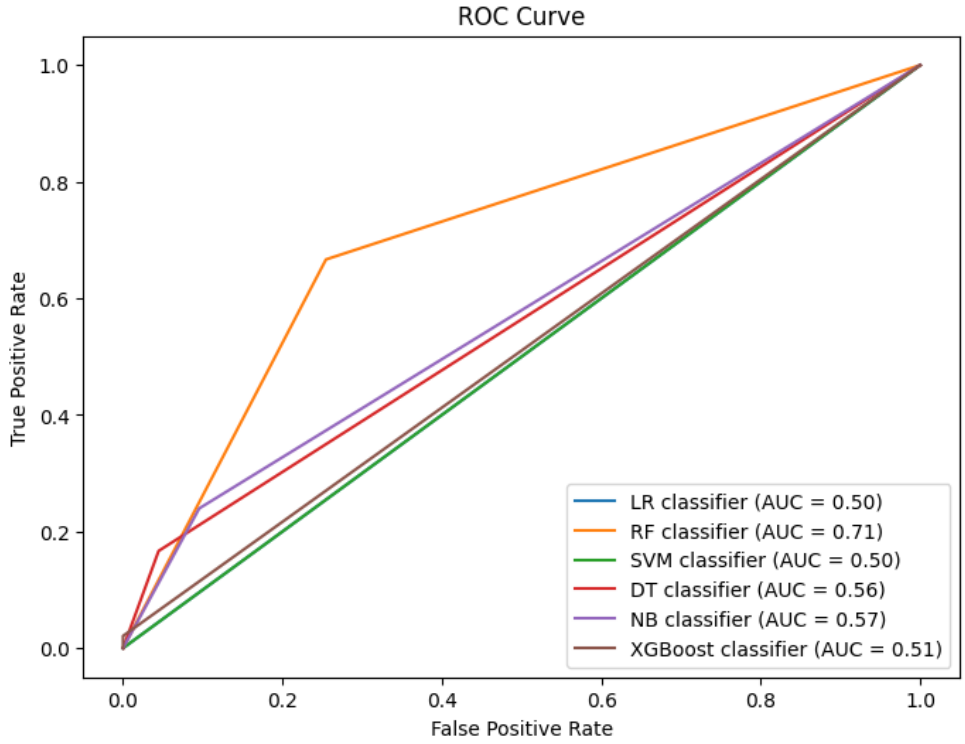
Şekil 4.4: SİT veri kümesindeki bireylerin dağılımı.



Şekil 4.5: SİT veri kümesindeki eksik özelliklerin miktarı.

Tablo 4.2: Sınıflandırıcıların SİT veri kümesi üzerindeki performansının değerlendirilmesi.

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	0.94	0.00	0.00	0.00	0.00
RF	**0.74**	**0.14**	**0.67**	**0.23**	**0.21**
SVM	0.94	0.00	0.00	0.00	0.00
DT	0.91	0.18	0.17	0.17	0.13
NB	0.87	0.13	0.24	0.17	0.11
XGBoost	0.94	1.00	0.02	0.04	0.14

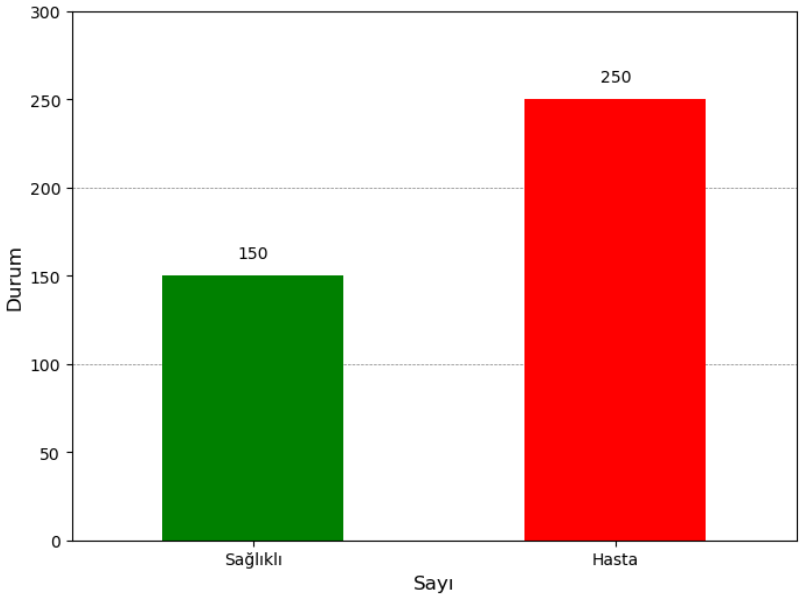


Şekil 4.6: SİT veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.

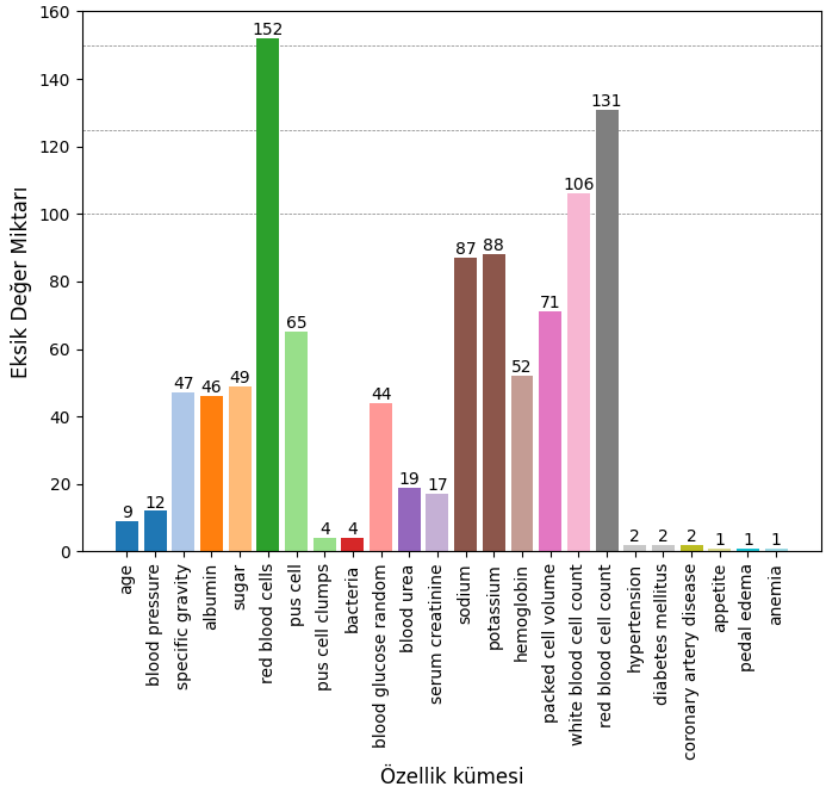
Tablo 4.2'de analiz edilen verilere göre, SİT veri kümesi üzerinde farklı sınıflandırıcıların performansı değerlendirilmiştir. Doğruluk, kesinlik, duyarlılık, F1-skor ve MCC gibi metrikler kullanılarak yapılan değerlendirme, sınıflandırıcıların sağlık durumunu tanıma performansları ölçülmüştür. RF sınıflandırıcısı, diğer yöntemlere kıyasla daha iyi bir performans sergilemiştir. Özellikle, RF'nin doğruluk (0,74), duyarlılık (0,67) ve MCC (0,21) değerleri, diğer sınıflandırıcılardan daha yüksek olmuştur. Bu sonuçlar, RF'nin SİT veri kümesi üzerinde daha etkili bir şekilde hasta ve sağlıklı bireyleri ayırt edebilme yeteneğine sahip olduğunu göstermektedir. Diğer sınıflandırıcılar arasında, XGBoost sınıflandırıcısının kesinlik açısından (1.00) yüksek bir performans sergilediği görülmektedir. Ancak, bu yüksek kesinlik değeri,

düşük duyarlılık (0,02) ve düşük F1-skor (0,04) ile sonuçlanmıştır, bu da XGBoost'un sağlık durumunu doğru bir şekilde tanıma yeteneğinin sınırlı olduğunu göstermektedir. Ayrıca, sınıflandırıcıların SİT veri kümesi üzerindeki performanslarının AUC değerlerine ait sonuçları incelendiğinde, RF sınıflandırıcısının, 0,71 AUC en yüksek değeriyle, orijinal veri kümesi üzerindeki performansının diğer modellere kıyasla daha iyi olduğu gözlemlenmiştir. Diğer sınıflandırıcılar arasında ise DT ve NB, RF'ye yakın AUC değerlerine sahiptir. RF'nin en yüksek AUC değerine sahip olması, bu sınıflandırıcının orijinal veri kümesi üzerinde hasta ve sağlıklı bireyleri doğru bir şekilde ayırt etme yeteneğinin daha yüksek olduğunu göstermektedir. Sınıf dengesizliği göz önüne alındığında, RF'nin performansının diğer sınıflandırıcılardan belirgin şekilde daha iyi olması dikkat çekicidir. Ancak, sınıf dengesizliği nedeniyle modelin yanıltıcı olabileceği göz önüne alındığında RF'nin düşük bir kesinlik değeri (0,14) elde ettiği gözlemlenmiştir. Eksik değerlerin varlığı da model performansını etkileyebilir, özellikle "smoking_status" özelliğindeki yüksek eksik değer miktarı modelin bu özelliği dikkate alırken zorluklarla karşılaşabileceğini gösterebilmektedir.

KBH veri kümesine ait sınıf dağılımı Şekil 4.7'de ve eksik değer miktarları Şekil 4.8'de verilmiştir.



Şekil 4.7: KBH veri kümesindeki bireylerin dağılımı.



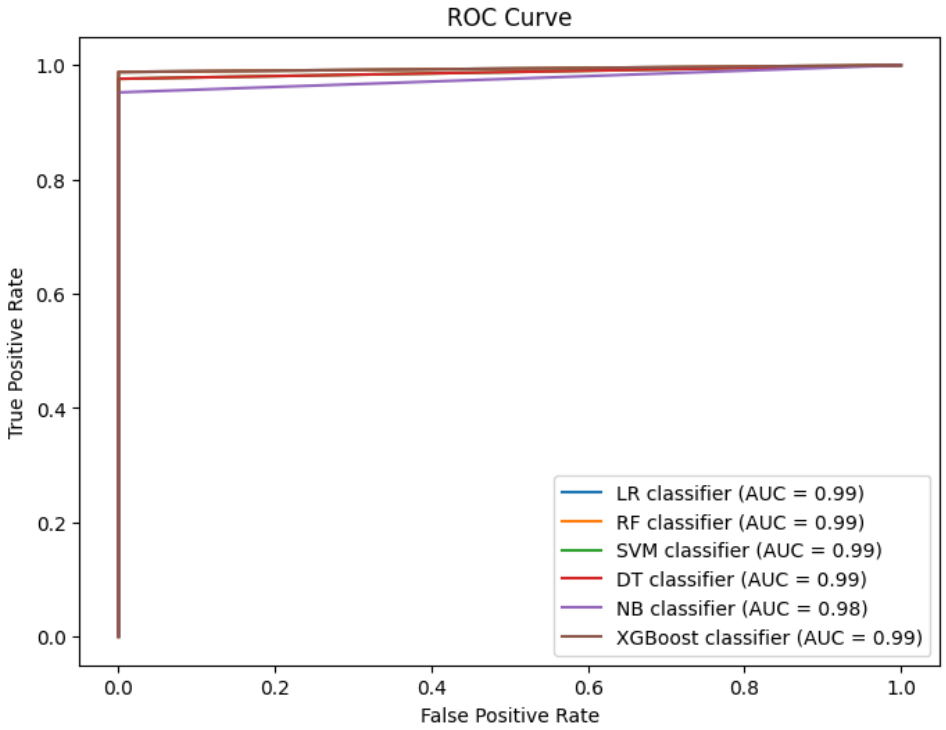
Şekil 4.8: KBH veri kümesindeki eksik özelliklerin miktarı.

Tablo 4.3: Sınıflandırıcıların KBH veri kümesi üzerindeki performansının değerlendirilmesi.

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	**0.99**	**1.00**	**0.99**	**0.99**	**0.98**
RF	**0.99**	**1.00**	**0.99**	**0.99**	**0.98**
SVM	0.98	1.00	0.98	0.99	0.97
DT	0.98	1.00	0.98	0.99	0.97
NB	0.97	1.00	0.95	0.98	0.94
XGBoost	**0.99**	**1.00**	**0.99**	**0.99**	**0.98**

Tablo 4.3 'te verilen bulgulara göre, KBH veri kümesi üzerinde farklı sınıflandırıcıların performansı değerlendirilmiştir. Doğruluk, kesinlik, duyarlılık, F1-skor ve MCC gibi metrikler kullanılarak yapılan değerlendirme, sınıflandırıcıların böbrek hastalığını tanıma performanslarını ölçmektedir. LR, RF ve XGBoost sınıflandırıcıları, yüksek doğruluk (0,99) ve kesinlik (1.00) değerlerine sahiptir. Bu sınıflandırıcılar, KBH veri kümesi üzerindeki hasta ve sağlıklı bireyleri doğru bir şekilde sınıflandırmada yüksek bir başarı elde etmiştir. Ayrıca, bu modellerin duyarlılık, F1-skor ve MCC değerleri de oldukça yüksektir, bu da modelin performansının dengeli olduğunu göstermektedir. Diğer sınıflandırıcılar, LR, RF ve XGBoost'a kıyasla biraz daha düşük performans sergilemektedir. Ancak, tüm sınıflandırıcılar yüksek

kesinlik ve F1-skor değerlerine sahiptir. Ayrıca, sınıflandırıcıların KBH veri kümesi üzerindeki performanslarının AUC değerlerine ait sonuçları Şekil 4.9'da verilmiştir.



Şekil 4.9: KBH veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.

Sonuçlar incelendiğinde, özellikle, LR, RF, SVM, DT ve XGBoost sınıflandırıcıları için AUC değerleri 0,98 ve üzerindedir, bu da bu modellerin oldukça güvenilir olduğunu göstermektedir. NB sınıflandırıcısı için AUC değeri biraz daha düşüktür, ancak hala kabul edilebilir bir seviyededir. Bu sonuçlar, sınıflandırıcıların orijinal veri kümesi üzerinde yüksek performans sergilediğini ve hastalıklı ve sağlıklı bireyleri başarılı bir şekilde ayırt etme yeteneklerine sahip olduğunu göstermektedir. Sınıf dengesizliği göz önüne alındığında, KBH veri kümesi düşük oranda dengesizdir. Ayrıca, birçok özellikte eksik değerler bulunmaktadır. Bu eksiklikler, özellikle "red blood cells", "white blood cell count" ve "red blood cell count" gibi önemli özelliklerde yoğun bir şekilde bulunmaktadır. KBH veri kümesindeki yüksek başarının arkasında birden çok faktör bulunmaktadır. İlk olarak, veri kalitesi bu başarının temelini oluşturur. Sınıf dengesizliği ve eksik değerlere rağmen, veri kümesindeki örneklerin doğru ve temsilci olması, sınıflandırıcıların doğru sonuçlar elde etmesini sağlamıştır. Özelliklerin ayrım gücü, bir başka önemli etkendir. KBH veri kümesinde kullanılan özellikler, hasta ve sağlıklı bireyleri ayırt etme konusunda yüksek ayrım gücüne sahip olabilir. Bu özellikler, hasta ve sağlıklı bireyleri açıkça ayırt edebilen belirgin özellikler olabilir. Model uyumu da başarıda

önemli bir rol oynar. Kullanılan sınıflandırıcılar, KBH veri kümesine uygun olarak seçilmiş ve eğitilmiş olmalıdır. Özellikle, modelin karmaşıklığını iyi dengeleyen ve veriye uyum sağlayabilen modeller tercih edilmelidir.

Genel olarak, PID, SİT ve KBH veri kümeleri üzerinde gerçekleştirilen deneysel analiz, sınıflandırıcıların performansını değerlendirirken dengesiz sınıfların ve eksik verilerin dikkate alınmasının önemini vurgulamaktadır. Üç veri kümesi üzerinde yapılan analizler, sınıflandırıcıların performansını değerlendirirken karşılaşılan çeşitli zorlukları ve başarıları açıkça ortaya koymaktadır. İlk olarak, PID veri kümesinde, sınıf dengesizliği ve eksik verilerin bulunması, sınıflandırıcıların performansını olumsuz etkileyebilir. Ancak, RF sınıflandırıcısının göreceli olarak daha yüksek başarı elde etmesi, bu zorlukların üstesinden gelme potansiyelini göstermektedir. SİT veri kümesindeki ciddi düzeydeki sınıf dengesizliği oranı ve büyük miktarda eksik verinin bulunması, sınıflandırıcıların doğruluğunu ve güvenilirliğini zorlayabilir. Ancak, RF'nin diğer sınıflandırıcılara kıyasla daha yüksek bir performans göstermesi, doğru model seçimi ve uygun ön işleme adımlarının önemini vurgular. KBH veri kümesinde, dengesiz sınıfların yanı sıra eksik verilerin de bulunması, sınıflandırıcıların performansını etkileyebilir. Ancak, LR, RF ve XGBoost gibi sınıflandırıcıların yüksek doğruluk ve diğer performans metriklerinde başarılı olması, modelin veri kümesine uygun bir şekilde uyarlanmasının önemini gösterir. Sınıflandırıcıların performansı üzerinde sınıf dengesizliği ve eksik veriler gibi zorlukların etkisi göz önüne alındığında, uygun ön işleme yöntemlerinin ve model seçiminin kritik öneme sahip olduğu açıktır.

4.2. GASMOTEPSO_ENN ÖRNEKLEME YÖNTEMİNİN FARKLI SINIFLANDIRICILAR ÜZERİNDE PERFORMANS KARŞILAŞTIRMASI

Bu deneyde, dengesiz dağılımlı üç farklı sağlık veri kümesinde, önerilen GASMOTEPSO_ENN yeniden örnekleme yönteminin uygulanmasıyla dengelenmiş veri kümelerinde hastalıkların sınıflandırılması performansı incelenmiştir. Literatürde sıkça kullanılan altı farklı makine öğrenmesi sınıflandırma algoritmalarının sınıflandırma performansı GASMOTEPSO_ENN yöntemiyle dengelenmiş veri kümeleri üzerinde test edilmiştir. Bu deneyin amacı, dengesiz sınıflara sahip sağlık veri kümelerindeki hastalık sınıflandırma performansının, önerilen yöntem ile nasıl etkilendiğini belirlemektir. Ayrıca, bu

deneyde önerilen GASMOTEPSO_ENN yöntemi literatürde dengesizlik problemlerini çözen ilgili çalışmaların sonuçlarıyla karşılaştırılmıştır.

Veri kümelerinin hazırlığı sürecinde, eksik verilerin ele alınması için literatürde başarıyla kullanılan MICE yöntemiyle eksik değerler tamamlanmıştır. Kategorik değişkenler, scikit-learn kütüphanesinde bulunan bir-etiket kodlama tekniği kullanılarak sayısal değerlere dönüştürülmüş; sürekli değişkenler ise standart ölçekleme tekniğiyle ölçeklendirilmiştir. Veri kümelerindeki dengesiz sınıfları düzeltmek için GASMOTEPSO_ENN yöntemi uygulanmış ve dengelenen veri kümelerinde aykırı değerleri belirlemek ve kaldırmak için IQR yöntemi kullanılmıştır. Sonrasında, tüm veri kümeleri %33 test oranıyla eğitim ve test alt kümelerine bölünmüş ve bu alt kümeler makine öğrenmesi sınıflandırma modellerine giriş olarak sağlanmıştır. Doğruluk, kesinlik, duyarlılık, F1-skor ve MCC model değerlendirme metriklerine göre elde edilen test sonuçları Tablo 4.4, Tablo 4.5 ve Tablo 4.6'da sunulmuştur. MCC metriğine göre en iyi sonuçlar her bir veri kümesi için kalın yazı tipiyle vurgulanmıştır.

Tablo 4.4: PID veri kümesinde sınıflandırıcıların GASMOTEPSO_ENN yöntemi sonrası performansının analizi.

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	0.73	0.76	0.69	0.72	0.47
RF	0.81	0.76	0.90	0.83	0.63
SVM	**0.94**	**0.95**	**0.92**	**0.93**	**0.87**
DT	0.82	0.78	0.89	0.83	0.64
NB	0.76	0.76	0.75	0.76	0.51
XGBoost	0.90	0.87	0.94	0.91	0.81

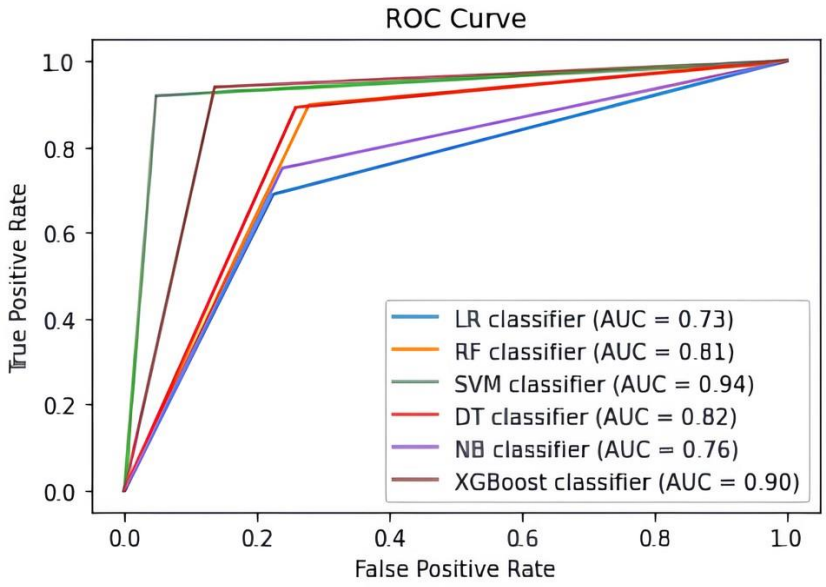
Tablo 4.5: SİT veri kümesinde sınıflandırıcıların GASMOTEPSO_ENN yöntemi sonrası performansının analizi.

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	0.96	0.99	0.95	0.97	0.92
RF	0.97	1.00	0.95	0.97	0.93
SVM	0.96	0.97	0.97	0.97	0.92
DT	0.97	1.00	0.95	0.97	0.93
NB	0.97	1.00	0.95	0.97	0.93
XGBoost	**0.97**	**0.99**	**0.96**	**0.98**	**0.94**

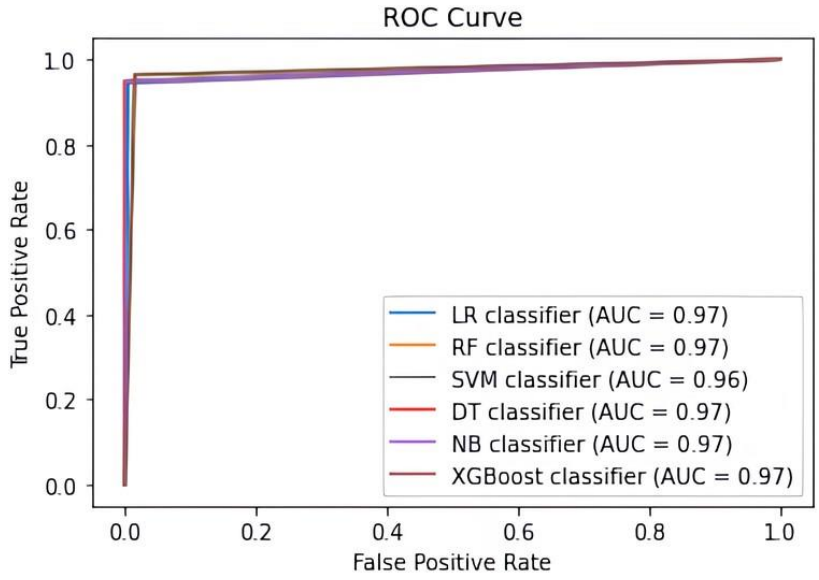
Tablo 4.6: KBH veri kümesinde sınıflandırıcıların GASMOTEPSO_ENN yöntemi sonrası performansının analizi.

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
RF	0.99	1.00	0.99	0.99	0.99
SVM	0.99	1.00	0.99	0.99	0.99
DT	0.98	1.00	0.95	0.98	0.96
NB	0.98	0.95	1.00	0.98	0.96
XGBoost	0.99	1.00	0.99	0.99	0.99

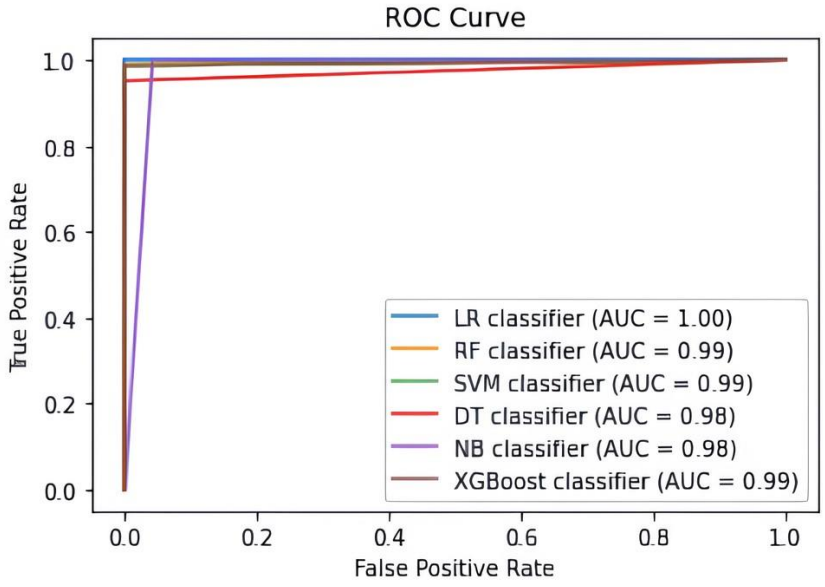
Elde edilen sonuçlar, önerilen GASMOTEPSO_ENN yönteminin dengesiz veriler ile başa çıkmada ve tahmin performansını artırmada etkili olduğunu göstermektedir. Model değerlendirme metriklerine dayanarak, GASMOTEPSO_ENN yönteminin kabul edilmesinin, tüm analiz edilen veri kümelerinde üstün sonuçlar elde ederek model değerlendirme metriklerinde belirgin bir iyileşme sağladığı açıktır. GASMOTEPSO_ENN yöntemi kullanılan sınıflandırma modelleri, orijinal veri kümelerinde çoğu model değerlendirme metriği için %90'ın üzerinde performans elde etmiştir. Bu bulgular, GASMOTEPSO_ENN yönteminin yüksek kaliteli sentetik örnekler üretmede başarılı olduğunu göstermektedir, böylelikle sınıflandırma modellerinin yeteneklerini artırmaktadır. Ayrıca, PID, SİT ve KBH veri kümeleri için sınıflandırıcıların ROC eğrilerini ve ilgili AUC değerlerini gösteren grafikler Şekil 4.10 Şekil 4.11 ve Şekil 4.12'de verilmiştir.



Şekil 4.10: GASMOTEPSO_ENN yöntemi sonrası PID veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.



Şekil 4.11: GASMOTEPSO_ENN yöntemi sonrası SİT veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.



Şekil 4.12: GASMOTEPSO_ENN yöntemi sonrası KBH veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.

PID, SİT ve KBH veri kümeleri için sınıflandırıcıların ROC eğrilerini ve ilgili AUC değerlerini gösteren grafikler, hastalıkların tahmininde model değerlendirme metrikleriyle elde edilen iyi performansı görsel olarak da desteklemektedir. Bu grafikler, sınıflandırıcıların hastalık teşhisindeki duyarlılık ve özgüllüğünü belirlemeye yardımcı olan ROC eğrilerini göstermektedir. AUC değerleri, sınıflandırıcının hastalık teşhisindeki genel performansını ölçer ve yüksek AUC değerleri, modelin hastalık ve sağlık durumunu doğru bir şekilde ayırt etme

yeteneğini yansıtır. Farklı özelliklere sahip sağlık veri kümelerine ait ROC eğrileri incelendiğinde, hafif dengesizlikler gösteren veri kümelerindeki sınıflandırıcıların ortalama AUC performansının %90'ı aştığını göstermektedir. Öte yandan, önemli dengesizlikler sergileyen veri kümelerinde ortalama AUC performansının %97 olduğu belirlenmiştir. Veri kümesinin dengelenmesi için önerilen GASMOTEPSO_ENN yöntemi kullanıldığında, hastaların ve sağlıklı bireylerin tahmininde dikkate değer bir iyileşme gözlemlenmiş ve önceki deneysel çalışmanın sonuçlarını aşmıştır. Bu nedenle, tez kapsamında önerilen GASMOTEPSO_ENN yönteminin daha yüksek bir geçerlilik ve doğruluğa sahip olduğu sonucuna varılmıştır. Özellikle, SİT gibi önemli dengesizliklere sahip veri kümeleri ile uğraşılırken, GASMOTEPSO_ENN yönteminin daha büyük bir stabilite ve sağlamlık sergilediği gözlemlenmiştir. GASMOTEPSO_ENN yöntemi, ön işlenmiş veri kümesindeki sınıf dengesizliğini ele alarak, dengesizlik sorunlarında etkili olmayan makine öğrenmesi algoritmalarının eksikliklerini gidermiştir.

GASMOTEPSO_ENN yöntemi ile literatürdeki benzer çalışmalar arasında karşılaştırmalı bir analiz PID, SİT ve KBH veri kümeleri için Tablo 4.7 Tablo 4.8 ve Tablo 4.9'da sunulmaktadır. Tez kapsamında, önerilen GASMOTEPSO_ENN yöntemi ile literatürdeki benzer çalışmaların karşılaştırıldığı tablolarda, MCC model değerlendirme ölçütü sonuçlarına göre her bir veri kümesinde en iyi performans gösteren makine öğrenmesi modelinin sonuçları yer almaktadır. Böylece, her bir veri kümesinin değerlendirilmesinde en etkili modelin vurgulanması sağlanarak sonuçların daha açık ve anlaşılır olmasını amaçlamaktadır.

Tablo 4.7: Önerilen GASMOTEPSO_ENN yönteminin literatürdeki PID tahmin modelleriyle karşılaştırılması.

Yöntem	Doğruluk	Kesinlik	Duyarlılık	F1-skor	AUC
SMOTE+LOF [7]	0.75	0.71	0.82	0.76	0.79
SMOTE+RF [8]	0.79	0.68	0.68	0.68	0.76
CS+XGBoost [10]	0.79	0.77	0.84	0.80	0.83
GASMOTEPSO_ENN+SVM	**0.94**	**0.95**	**0.92**	**0.93**	**0.94**

Tablo 4.8: Önerilen GASMOTEPSO_ENN yönteminin literatürdeki SİT tahmin modelleriyle karşılaştırılması.

Yöntem	Doğruluk	Kesinlik	Duyarlılık	F1-skor	AUC
Undersampling+NB [3]	0.82	0.79	0.86	0.82	0.82
SMOTE+TOMEK+DNN [23]	0.86	0.99	0.87	0.92	0.84
Hybrid ML+RF [4]	0.94	0.94	0.94	0.95	0.95
GASMOTEPSO_ENN+XGBoost	**0.97**	**0.99**	**0.96**	**0.98**	**0.97**

Tablo 4.9: Önerilen GASMOTEPSO_ENN yönteminin literatürdeki KBH tahmin modelleriyle karşılaştırılması.

Yöntem	Doğruluk	Kesinlik	Duyarlılık	F1-skor	AUC
SMOTE+ANN [9]	0.96	0.98	0.91	0.94	0.99
CS+RF [10]	0.98	0.99	1.00	0.99	1.00
SMOTE+PSO+MetaCost [6]	-	1.00	0.99	0.99	1.00
GASMOTEPSO_ENN+LR	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Elde edilen bulgular incelendiğinde, önerilen GASMOTEPSO_ENN yönteminin, mevcut literatürdeki yeniden örnekleme yöntemlerine kıyasla benzer veya üstün performans sergilediği gözlemlenmiştir. Ayrıca, GASMOTEPSO_ENN yöntemi ile elde edilen performans seviyelerinin, SİT ve KBH veri kümeleri için alternatif yeniden örnekleme yöntemleriyle karşılaştırılabilir olduğunu vurgulamaktadır. Öte yandan, PID veri kümesini kullanan diğer ilgili çalışmalara kıyasla olağanüstü performans sergilemiştir. Bu, GASMOTEPSO_ENN yönteminin genel etkinliğini ve farklı veri kümelerindeki başarı potansiyelini daha belirgin bir şekilde ortaya koymaktadır.

Genel olarak, GASMOTEPSO_ENN yöntemi ile elde edilen sonuçlar, hastalık teşhisinde artan performansın pratik uygulamaları için önemli sonuçları vurgulamaktadır. Özellikle, önemli dengesizliklere sahip veri kümelerindeki artan doğruluk oranı, yanlış teşhisleri azaltma ve tedavi planlamasını optimize etme potansiyeline işaret etmektedir. Bu durum, daha etkin kaynak tahsisi, iyileştirilmiş hasta sonuçları ve sağlık hizmetleri kalitesinde artış anlamına gelmektedir. Yöntemin sağlamlığı, sağlık veri kümelerinde sıkça karşılaşılan sınıf dengesizliği gibi yaygın zorlukları ele almak için değerli bir araç oluşturmakta ve bu da araştırmacılar ve klinisyenler için hastalık teşhisi ve tedavisinde pratik faydalar sunar.

4.3. ÖNERİLEN VERİ ÖN İŞLEME YÖNTEMİNİN FARKLI SINIFLANDIRICILAR ÜZERİNDE PERFORMANS KARŞILAŞTIRMASI

Bu deneyde, sağlık veri kümelerindeki ikili sınıflandırma görevlerinde, dengesiz dağılımlı sınıfların ve eksik değerli verilerin belirgin olduğu durumlar için tez kapsamında önerilen ön işleme yönteminin etkisi sistemli bir şekilde analiz edilmiştir. Bu tez çalışmasında, önerilen ön işleme yönteminin performansı, literatürde başarıyla kullanılan altı farklı makine öğrenmesi sınıflandırma algoritması ile test edilmiştir. Bu testler, önerilen ön işleme yöntemi ile dengesiz dağılımlı sınıfların dengelendiği ve eksik değerlerin tamamlandığı veri kümesindeki bireylerin doğru ve etkili bir şekilde hasta veya sağlıklı olarak ayırt edilme yeteneğini değerlendirmeyi amaçlamıştır. Yöntemin başarısı, bu algoritmaların kullanımıyla elde edilen sonuçlarla ölçülmüş ve analiz edilmiştir. Ayrıca, önerilen ön işleme yönteminin ikili sınıflandırma problemlerindeki performansı, eksik değerlerin tamamlanması ve dengesiz dağılımlı sınıfların dengelenmesi üzerine literatürde yapılan ilgili çalışmaların sonuçlarıyla karşılaştırılmıştır. Bu karşılaştırma, önerilen yöntemin etkinliğini daha geniş bir bağlamda değerlendirme imkanı sunmaktadır.

Veri kümelerinin hazırlığı sürecinde, eksik verilerin ele alınması için önerilen GA-MICE yöntemiyle eksik değerler tamamlanmıştır. Kategorik değişkenler, scikit-learn kütüphanesinde bulunan bir-etiket kodlama tekniği kullanılarak sayısal değerlere dönüştürülmüş; sürekli değişkenler ise standart ölçekleme tekniğiyle ölçeklendirilmiştir. Veri kümelerindeki dengesiz sınıfları düzeltmek için önerilen GASMOTEPSO_ENN yöntemi uygulanmış ve dengelenen veri kümelerinde aykırı değerleri belirlemek ve kaldırmak için IQR yöntemi kullanılmıştır. Sonrasında, tüm veri kümeleri %33 test oranıyla eğitim ve test alt kümelerine bölünmüş ve bu alt kümeler makine öğrenmesi sınıflandırma modellerine giriş olarak sağlanmıştır. Doğruluk, kesinlik, duyarlılık, F1-skor ve MCC model değerlendirme metriklerine göre elde edilen test sonuçları Tablo 4.10, Tablo 4.11 ve Tablo 4.12'de sunulmuştur. MCC metriğine göre en iyi sonuçlar her bir veri kümesi için kalın yazı tipiyle vurgulanmıştır.

Tablo 4.10: PID veri kümesinde sınıflandırıcıların önerilen veri ön işleme yöntemi sonrası performansının analizi.

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	0.78	0.74	0.86	0.79	0.57
RF	0.84	0.77	0.96	0.85	0.70
SVM	**0.93**	**0.92**	**0.94**	**0.93**	**0.86**
DT	0.85	0.82	0.88	0.85	0.70
NB	0.73	0.73	0.73	0.73	0.47
XGBoost	0.92	0.92	0.89	0.96	0.84

Tablo 4.11: SİT veri kümesinde sınıflandırıcıların önerilen veri ön işleme yöntemi sonrası performansının analizi.

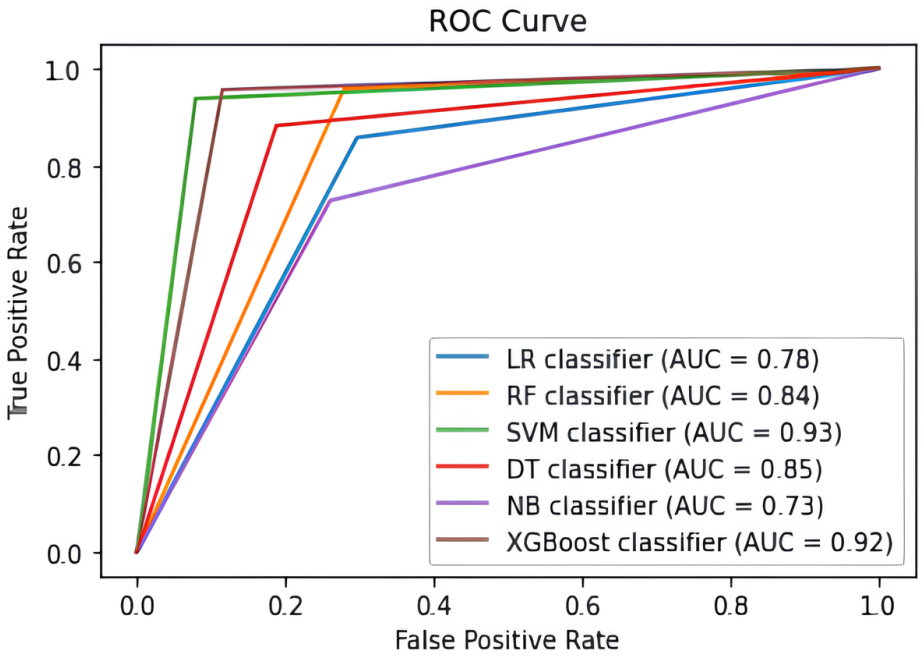
Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	0.92	0.92	0.90	0.91	0.84
RF	0.92	0.91	0.93	0.92	0.85
SVM	0.96	0.95	0.96	0.96	0.92
DT	0.94	0.93	0.95	0.94	0.88
NB	0.92	0.91	0.92	0.92	0.84
XGBoost	**0.97**	**0.97**	**0.98**	**0.97**	**0.95**

Tablo 4.12: KBH veri kümesinde sınıflandırıcıların önerilen veri ön işleme yöntemi sonrası performansının analizi.

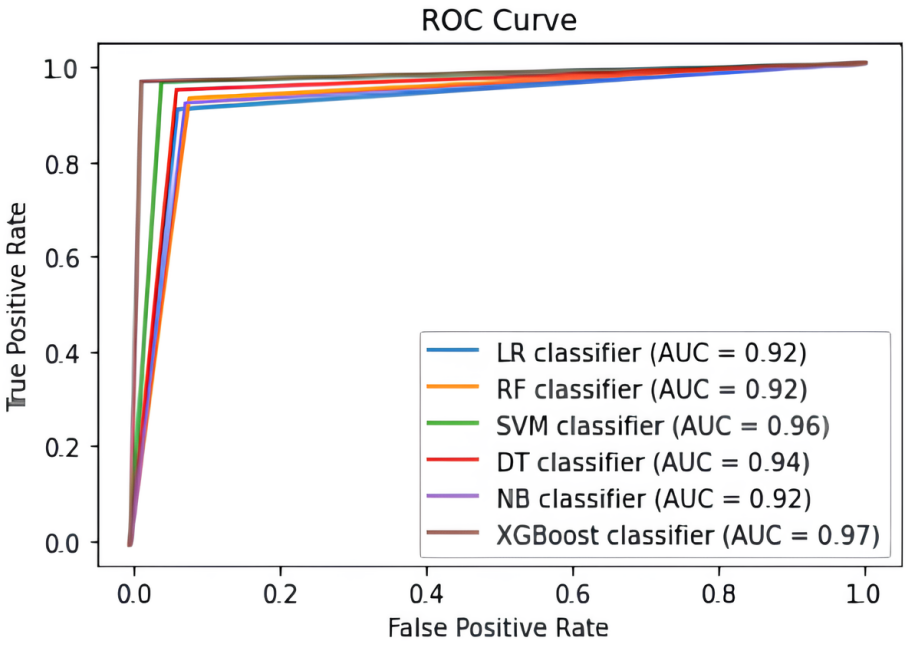
Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F1-skor	MCC
LR	0.99	1.00	0.99	0.99	0.99
RF	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
SVM	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
DT	0.98	0.95	1.00	0.98	0.95
NB	0.98	0.98	0.99	0.98	0.96
XGBoost	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Deneysel sonuçlar incelendiğinde, tez kapsamında geliştirilen ön işleme yönteminin, makine öğrenmesi sınıflandırıcılarının performansını önemli ölçüde artırdığını gözlemlenmiştir. Öncelikle, PID veri kümesi üzerinde yapılan değerlendirmede, SVM ve XGBoost gibi önde gelen algoritmaların, GASMOTEPSO_ENN ve GA-MICE yöntemleri uygulandıktan sonra elde edilen tam ve dengeli veri kümelerinde yüksek doğruluk, kesinlik, duyarlılık, F1-skoru ve

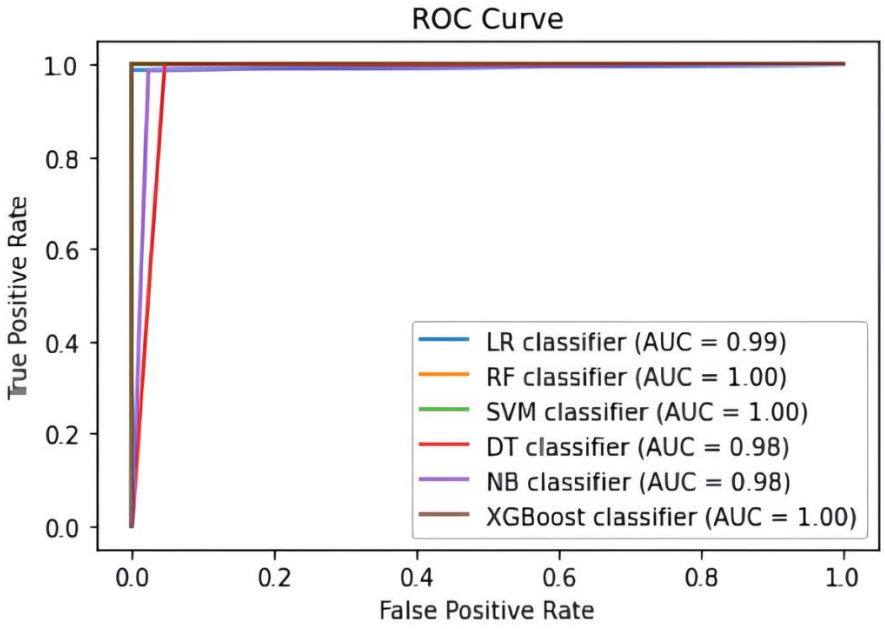
MCC model değerlendirme metriklerinde dikkate değer bir başarıya ulaştığı gözlenmiştir. Bu sonuçlar, veri yeniden örnekleme ve eksik değer atama süreçlerinin, sınıflandırıcıların doğruluğunu ve güvenilirliğini artırmada etkili olduğunu göstermektedir. SİT veri kümesi üzerinde yapılan değerlendirmede, SVM ve XGBoost gibi algoritmaların, doğruluk, kesinlik, duyarlılık, F1-skoru ve MCC model değerlendirme metriklerinde oldukça yüksek sonuçlar elde ettiği gözlemlenmiştir. Bu bulgular, önerilen ön işleme yönteminin, veri kümesinin oldukça yüksek oranda sınıf dengesizliği göstermesine rağmen, makine öğrenmesi sınıflandırıcılarının performansını artırdığını açıkça ortaya koymaktadır. KBH veri kümesinde yapılan değerlendirme ise, tüm sınıflandırıcıların yüksek performans sergilediğini gözlemlenmektedir. Bu bulgular, eksik değerlerin atanması ve sınıf dengesizliğinin düzeltilmesi gibi önemli veri ön işleme adımlarının, sınıflandırıcıların istikrarlı ve güvenilir bir şekilde yüksek performans sergilemesine katkı sağladığını vurgulamaktadır. Bu bağlamda, deney sonuçları önerilen ön işleme yönteminin etkinliğini güçlü bir şekilde desteklemekte ve gerçek dünya verileri üzerinde sınıflandırma performansını artırmada önemli bir adım olarak öne çıkmaktadır. Ayrıca, önerilen ön işleme yönteminin performansı, ikili sınıflandırma modelinin farklı eşiklerdeki pozitif ve negatif sınıfları ayırt etme yeteneğini ölçen AUC metriği üzerinden de analiz edilmiştir. Sınıflandırıcıların sırasıyla PID, SİT ve KBH veri kümelerinde analiz edilen ROC eğrileri ve ilgili AUC değerleri Şekil 4.13, Şekil 4.14 ve Şekil 4.15 'te sunulmaktadır.



Şekil 4.13: Önerilen veri ön işleme yöntemi sonrası PID veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.



Şekil 4.14: Önerilen veri ön işleme yöntemi sonrası SİT veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.



Şekil 4.15: Önerilen veri ön işleme yöntemi sonrası KBH veri kümesiyle eğitilmiş çeşitli sınıflandırıcıların ROC eğrileri.

PID veri kümesi için Şekil 4.13'te elde edilen bulgulara göre, SVM ve XGBoost sınıflandırıcılarının pozitif ve negatif sınıflar arasında diğer sınıflandırıcılara göre daha iyi bir ayrım yeteneği sergilediği, NB'nin ise nispeten daha zayıf bir ayrım yeteneği sergilediği açıkça gözlemlenmektedir. SİT veri kümesi için, Şekil 4.14 'te elde edilen bulgulara göre, LR, RF,

SVM, DT, NB ve XGBoost sınıflandırıcıları AUC değerlerine dayalı olarak dikkate değer bir performans sergilemiştir. XGBoost sınıflandırıcısı 0,97 en yüksek AUC değeri ile öne çıkmaktadır. Bu bulgular SİT veri kümesindeki pozitif ve negatif sınıfları ayırt etme konusunda tüm sınıflandırıcıların yüksek yeteneklere sahip olduğunu göstermektedir. KBH veri kümesine ilişkin olarak elde edilen sonuçlara göre, Şekil 4.15 'teki bulgular, tüm sınıflandırıcıların yüksek AUC değerleri elde ettiğini açıkça göstermektedir. Tüm sınıflandırıcılarda 0,98'in üzerinde AUC değerleri gözlemlenmiştir. RF, SVM ve XGBoost sınıflandırıcıları, %100 AUC değerleriyle mükemmel ayırt etme yeteneği sergilemiştir. Bu bulgular KBH veri kümesindeki pozitif ve negatif sınıfları ayırt etmede tüm sınıflandırıcıların son derece yüksek performans gösterdiği elde edilen sonuçlar arasındadır

Tez kapsamında önerilen ön işleme yönteminin başarısı, veri dengesini sağlamak için GASMOTEPSO_ENN ve eksik değerleri ele almak için GA-MICE yöntemlerinin yeteneklerinden kaynaklandığını açıkça göstermektedir. GASMOTEPSO_ENN yöntemi, veri kümesindeki dengesizliği gidermek için SMOTE örnekleme tekniğini GA ve PSO sezgisel yöntemleriyle birleştirir. Ardından, ENN yöntemi, örnekler arasındaki mesafeleri ve sınıf etiketlerini kullanarak aşırı örnekleme sonrası veri kümesini düzenler, bu da aşırı örnekleme sonrası oluşturulan sentetik örneklerin kalitesini artırır. GASMOTEPSO_ENN yöntemi, azınlık sınıfında yeni örnekler oluşturarak makine öğrenmesi sınıflandırıcıları için eğitim verilerinin daha dengeli bir şekilde dağıtılmasını sağlar. Öte yandan, GA-MICE yöntemi eksik değerlerin güvenilir bir şekilde tespit edilmesini ve atanmasını sağlayarak sınıflandırma algoritmalarının önyargılı olmasını engeller. GA-MICE yöntemi, eksik değerleri doldurmak için MICE atama yaklaşımını GA ile birleştirir ve eksik değerleri etkili bir şekilde tamamlar. Bu yöntem, eksik verilerin tespit edilmesini ve doğru şekilde doldurulmasını sağlayarak sınıflandırma algoritmalarının önyargılı olmasını önler. Bu deneysel çalışma, GASMOTEPSO_ENN ve GA-MICE yaklaşımlarının gerçek dünya verileriyle çalışan sınıflandırma modellerinin performansını artırmada etkili olduğunu göstermektedir. Ayrıca, önerilen ön işleme yönteminin performansı, eksik değer atama ve sınıf dengesizliği konularını ele alan literatürdeki benzer çalışmaların bulgularıyla karşılaştırılmıştır. Her sınıflandırıcının göreceli performansını değerlendirmek için Bölüm 2.2.9'da sunulan model değerlendirme metriklerine dayalı olarak karşılaştırmalı bir analizi gerçekleştirilmiştir. Deneysel bulgular, Tablo 4.13, Tablo 4.14 ve Tablo 4.15'te sunulmaktadır. Her veri kümesi için F1-skoru açısından en iyi sonuçlar kalın yazı tipi kullanılarak vurgulanmıştır.

Tablo 4.13: Önerilen ön işleme yönteminin literatürdeki PID tahmin modelleriyle karşılaştırılması.

Yöntem	Doğruluk	Kesinlik	Duyarlılık	F1-skor	AUC
DMP_MI [19]	0.87	0.80	0.85	0.83	0.92
ADASYN+MLP [20]	0.84	0.91	0.85	0.88	0.83
CS+XGBoost [10]	0.79	0.77	0.84	0.80	0.83
MissForest+XGBoost [17]	0.77	-	-	-	-
Önerilen yöntem	**0.93**	**0.92**	**0.94**	**0.93**	**0.93**

Tablo 4.14: Önerilen ön işleme yönteminin literatürdeki SİT tahmin modelleriyle karşılaştırılması.

Yöntem	Doğruluk	Kesinlik	Duyarlılık	F1-skor	AUC
Ensemble RXLM [21]	0.96	0.96	0.96	0.96	0.99
PCA+DNN+Focal Loss [22]	0.92	-	-	0.77	0.90
SMOTE+TOMEK+DNN [23]	0.86	0.99	0.87	0.92	0.84
Undersampling+NB [3]	0.82	0.79	0.86	0.82	0.82
Önerilen yöntem	**0.97**	**0.97**	**0.98**	**0.97**	**0.97**

Tablo 4.15: Önerilen ön işleme yönteminin literatürdeki KBH tahmin modelleriyle karşılaştırılması.

Yöntem	Doğruluk	Kesinlik	Duyarlılık	F1-skor	AUC
CS+RF [10]	0.99	0.99	1.00	0.99	1.00
Datawig+RF [24]	0.98	-	-	-	-
LR+SMOTE [25]	0.98	1.00	-	0.98	-
SMOTE+FNN [26]	-	0.97	0.99	0.99	0.99
Önerilen yöntem	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Sağlık veri kümelerinde eksik değer ve dengesiz dağılımlı sınıf problemlerine yönelik literatür çalışmaları incelendiğinde, tez kapsamında önerilen ön işleme yönteminin alternatif yöntemlere kıyasla üstün performans sergilediği açıkça görülmektedir. PID veri kümesi üzerinde yapılan çalışmalarda, Wang ve diğ. [19] tarafından kullanılan DMP_MI yöntemi kabul edilebilir bir doğruluk oranına sahip olsa da önerilen yöntemimiz ile karşılaştırıldığında daha düşük performans göstermektedir. Benzer şekilde Yadav ve diğ. [20] ile Mienye ve Sun [10] tarafından önerilen yöntemler, önerilen yöntemimizle karşılaştırıldığında daha düşük performans metriklerine sahiptir. Özellikle, önerilen yöntemimiz %93 oranında önemli bir doğruluk oranı elde etmiştir, bu da Madhu ve diğ. [17] tarafından önerilen modelin performansını aşmaktadır. Önerilen yöntemimiz, PID veri kümesinde yüksek doğruluk,

kesinlik, duyarlılık, F1-skoru ve AUC model değerlendirme sonuçları elde etmiştir. Önerilen yöntemimizin kesinlik, duyarlılık ve F1-skoru sırasıyla %92, %94 ve %93 olarak hesaplanmıştır. Ayrıca, bu yöntemin AUC değeri 0,93 olarak ölçülmüştür. Bu bulgular, önerilen yöntemin PID veri kümesinde daha etkili çalıştığını ve daha güvenilir sonuçlar üretebileceğini göstermektedir. Dolayısıyla, önerilen yöntemin diyabet teşhisi için daha güvenilir bir araç olduğu elde edilen sonuçlar arasındadır. SİT veri kümesi üzerinde yapılan literatür çalışmaları incelendiğinde, önerilen yöntemimizin diğer yöntemlere kıyasla üstün performans sergilediği açıkça görülmektedir. Alruily ve diğ. [21] tarafından önerilen yöntem %96 doğruluk elde ederken, önerilen yöntemimiz %97 doğruluk oranı elde ederek daha iyi bir performans göstermiştir. Jing [22]'in yöntemi %92 doğruluk oranına sahipken, önerilen yöntemimiz hem daha yüksek doğruluk hem de daha yüksek bir F1-skoru performansı elde etmiştir. Rana ve diğ. [23]'in yöntemi %86 doğruluk, %99 kesinlik ve %87 duyarlılık sonucu elde etmiştir. Önerilen yöntemimiz ise daha yüksek doğruluk, kesinlik ve duyarlılık değerleriyle daha etkili bir performans sergilemiştir. Benzer şekilde, Sailasya ve Kumari [3]'nin yöntemi %82 doğruluk, %79 kesinlik ve %86 duyarlılık sonucu gösterirken, önerilen yöntemimiz daha yüksek doğruluk, kesinlik ve duyarlılık metrik sonuçları sağlamıştır. Genel olarak, sonuçlar önerilen sistemimizin SİT veri kümesinde daha etkili ve güvenilir bir sınıflandırma performansı sunduğunu göstermektedir. SİT veri kümesi üzerinde yapılan literatür çalışmalarının incelenmesi, önerilen sistemin diğer yöntemlere kıyasla üstün performans sergilediğini göstermektedir. KBH veri kümesi üzerinde yapılan literatür çalışmaları incelendiğinde, Mienye ve Sun [10] tarafından önerilen yöntem %99 doğruluk oranı elde ederken, önerilen yöntemimiz %100 oranında doğruluk etmiştir. Maisha ve diğ. [24] ile Wibowo ve Pulupi [25]'nin modelleri %98 doğruluk oranına sahipken, önerilen yöntemimiz %100 doğruluk sonucu elde ederek olağanüstü bir sınıflandırma yeteneği göstermiştir. Ayrıca, önerilen yöntemimiz, kesinlik, duyarlılık, F1-skoru ve AUC açısından mükemmel sonuçlar elde etmiştir, bu da yöntemimizin güçlü ve güvenilir olduğunu göstermektedir. Imran ve diğ. [26]'nin yöntemi %97 kesinlik ve %99 duyarlılık gibi etkileyici sonuçlar elde etmiş olsa da doğruluk metriklerini sunmamıştır. Genel olarak, sonuçlar, önerilen sistemimizin KBH veri kümesindeki bireyleri hasta ve sağlıklı olarak doğru bir şekilde sınıflandırmada olağanüstü etkinlik ve üstünlüğünü vurgulamaktadır. Tüm veri kümelerinde elde edilen sonuçlar, önerilen yöntemimizin çeşitli dengesiz ve eksik değerli sağlık veri kümelerinde etkili bir şekilde kullanılabileceğini ve literatürde önerilen diğer yöntemlere kıyasla daha iyi performans sergilediğini göstermektedir.

5. TARTIŞMA

Dünya genelinde, kronik hastalıklardan kaynaklanan ölümlerin oranları WHO'nun yayınladığı rapora göre belirtilmiştir. Bu hastalıklar içerisinde kalp hastalığı, felç, kanser, diyabet ve kronik solunum yolu hastalıkları dahil birçok kronik rahatsızlık bulunmaktadır. Kronik hastalıkların oluşması riski yüksek olan bireylerin erken tanısı, hastalığın ilerlemesinin yavaşlatılmasında ve tedavisinde önemli bir klinik değere sahiptir [2]. Son zamanlarda klinik çalışmalarda hastalıkların erken teşhisi ve tahmini için makine öğrenmesi sınıflandırma modellerinde dikkate değer bir artış yaşanmaktadır. Bu modeller ile sağlık sektöründeki bakım maliyetleri azaltılmakta ve hekimlerin iş gücünü azaltarak her hastaya ayrılan sürenin daha etkili kullanımı sağlanmaktadır. Artan küresel nüfusla birlikte hekimlerin, çeşitli hastalıkların erken tespitini ve takibini sağlayacak, makine öğrenmesi ve yapay zeka teknikleri ile geliştirilmiş yardımcı araçlara ihtiyaç duyduğu açıktır.

Makine öğrenmesi yöntemleri genellikle dengeli sınıf dağılımlı, eksiksiz ve tutarlı veri kümeleriyle uyumlu olarak tasarlanmıştır. Ancak, doğası gereği sağlık veri kümeleri genellikle dengesiz sınıf dağılımlarına ve eksik değerlere sahiptir. Bu durumda, veri kümesindeki eksik veya hatalı veriler makine öğrenmesi modellerinin öğrenme sürecini olumsuz yönde etkileyebilir. Sınıflar arasındaki dengesizlik, nadir sınıfların tanınmasını zorlaştırabilir ve modelin çoğunluk sınıfına yanlış şekilde eğilimli olmasına neden olabilir. Literatürde, her iki problemin çözümüne yönelik yeterli sayıda modelin bulunmadığı gözlemlenmiştir. Bu durum, sağlık alanında eksik değer ve dengesiz sınıf problemlerini ele alan yeni araştırmalara olan ihtiyacı vurgulamaktadır.

Tez çalışması kapsamında, sağlık verilerinin analizi ve hastalıkların tahmini üzerine yapılan araştırmalarda sıkça karşılaşılan eksik veri ve dengesiz sınıf dağılımı problemlerini ele alan özgün bir hibrit ön işleme modeli geliştirilmiştir. Bu model, eksik değerlerin ataması için literatürdeki çalışmalarda başarılı bir şekilde kullanılan MICE yönteminin GA sezgisel yöntemini birleştirerek geliştirilen GA-MICE yöntemini içermektedir. Dengesiz dağılıma sahip sınıfların ele alınması için SMOTE aşırı örnekleme ve ENN eksik örnekleme yöntemlerinin GA ve PSO sezgisel yöntemlerinin birleştirilmesiyle oluşturulan GASMOTEPSO_ENN yöntemini kapsamaktadır. Bu şekilde, veri kümesindeki eksik değerlerin ve sınıf dengesizliğinin

giderilmesiyle birlikte, makine öğrenmesi sınıflandırma modelinin performansı artırılarak daha güvenilir sonuçlar elde edilmesi amaçlanmıştır. Önerilen yöntemin etkinliğinin değerlendirilmesinde, WHO verilerine dayanarak belirlenen ölüm oranlarının başında gelen diyabet, inme ve böbrek hastalıklarının tespit edilmesine odaklanan üç farklı halka açık veri kümesi kullanılmıştır. Bu veri kümeleri seçilirken, farklı sınıf dengesizlik oranları, eksik değer miktarları, veri kümesi büyüklüğü ve hastalıkların yaygınlığı gibi çeşitli kriterler dikkate alınmıştır. Bu çok yönlü veri kümesi seçimi, önerilen yöntemin genel geçerliliğini ve gerçek dünya koşullarında performansını daha etkili bir şekilde değerlendirmeye olanak tanımıştır. Veri kümesi hazırlığı aşamasında, eksik değerler ve sınıf dengesizliği oranları tespit edilmiş ve tez kapsamında önerilen hibrit ön işleme yöntemi kullanılarak eksik değerler doldurulmuş ve sınıflar dengelenmiştir. Ardından, dengelenmiş ve tamamlanmış veri kümesi eğitim ve test alt kümelerine bölünerek makine öğrenmesi sınıflandırma modellerine uygun hale getirilmiştir. Veri kümelerindeki bireylerin hasta ve sağlıklı olarak sınıflandırılmasında, literatürde sıkça kullanılan LR, RF, SVM, DT, NB ve XGBoost makine öğrenmesi sınıflandırma algoritmaları kullanılmıştır. Sınıflandırma modelinin performansı, doğruluk, kesinlik, duyarlılık, F1-skoru, MCC ve AUC model değerlendirme metrikleri kullanılarak ölçülmüş ve önerilen hibrit ön işleme yönteminin sınıflandırma performansına olan etkisi değerlendirilmiştir.

Bu çalışmada, üç farklı deney tasarlanmıştır. İlk deneyde, sağlık veri kümelerinde sıkça görülen eksik değerler ve sınıf dengesizliği gibi önemli veri problemleri göz ardı edilerek makine öğrenmesi sınıflandırıcılarının doğrudan hasta ve sağlıklı bireyleri ayırt etme performansı değerlendirilmiştir. Elde edilen analiz sonuçlarına göre, PID veri kümesinde sınıf dengesizliği ve eksik verilerin bulunması, sınıflandırıcıların performansını olumsuz etkileyebildiği gözlemlenmiştir. Ancak, RF sınıflandırıcısının göreceli olarak daha yüksek başarı elde etmesi, bu zorlukların üstesinden gelme potansiyelini göstermektedir. SİT veri kümesindeki ciddi düzeydeki sınıf dengesizliği oranı ve büyük miktarda eksik verinin bulunması, sınıflandırıcıların doğruluğunu ve güvenilirliğini zorlayabilir. Ancak, RF'nin diğer sınıflandırıcılara kıyasla daha yüksek bir performans göstermesi, doğru model seçimi ve uygun ön işleme adımlarının önemini vurgulamaktadır. KBH veri kümesinde, dengesiz sınıfların yanı sıra eksik verilerin de bulunması, sınıflandırıcıların performansını etkileyebildiği gözlemlenmiştir. Ancak, LR, RF ve XGBoost gibi sınıflandırıcıların yüksek doğruluk ve diğer performans metriklerinde başarılı olması, modelin veri kümesine uygun bir şekilde uyarlanmasının önemini göstermektedir. Sınıflandırıcıların performansı üzerinde sınıf

dengesizliği ve eksik veriler gibi zorlukların etkisi göz önüne alındığında, uygun ön işleme yöntemlerinin ve model seçiminin kritik öneme sahip olduğu açıkça görülmüştür.

İkinci deneyde, sınıf dengesizliği için bir çözüm olarak tez kapsamında önerilen GASMOTEPSO_ENN yöntemi kullanılmış ve eksik değerler literatürde başarıyla kullanılan MICE yöntemiyle doldurulmuştur. Bu sayede, daha dengeli ve eksiksiz bir veri kümesi elde edilerek sınıflandırıcıların performansı literatürdeki diğer yöntemlerle karşılaştırılarak değerlendirilmiştir. Elde edilen analiz sonuçları incelendiğinde, önerilen GASMOTEPSO_ENN yönteminin sağlık veri kümelerindeki sınıf dengesizliği ve eksik veriler gibi önemli veri problemlerini ele almak için etkili olduğunu göstermektedir. Bu yöntem sayesinde, makine öğrenmesi sınıflandırıcılarının performansında belirgin bir iyileşme sağlanmıştır. Özellikle, GASMOTEPSO_ENN yöntemi sonrası kullanılan sınıflandırma modelleri, orijinal veri kümelerinde yüksek performans elde etmiş ve model değerlendirme metriklerinde belirgin bir iyileşme göstermiştir. Ayrıca, literatürdeki diğer benzer çalışmalarla karşılaştırıldığında, önerilen yöntemin benzer veya üstün performans sergilediği gözlemlenmiştir. Özellikle, PID, SİT ve KBH gibi farklı sağlık veri kümelerinde elde edilen sonuçlar, GASMOTEPSO_ENN yönteminin genel etkinliğini ve farklı veri kümelerindeki başarı potansiyelini açıkça ortaya koymaktadır. Bu bulgular ışığında, GASMOTEPSO_ENN yöntemi, sağlık veri kümelerinde sınıf dengesizliği gibi yaygın zorlukları ele almak için değerli bir araç olarak kullanılabilir. Bu yöntem, hastalık teşhisinde artan doğruluk oranıyla yanlış teşhisleri azaltma ve tedavi planlamasını optimize etme potansiyeline sahiptir. Böylece, bu yöntemle daha etkin kaynak tahsisi, iyileştirilmiş hasta sonuçları ve sağlık hizmetleri kalitesinde artış sağlanabilir.

Üçüncü deneyde hem eksik değerlerin hem de sınıf dengesizliğinin giderilmesine yönelik tez kapsamında önerilen hibrit ön işleme yönteminin etkinliği incelenmiştir. Bu yeni yöntem, literatürdeki diğer yöntemlerle karşılaştırılarak değerlendirilmiştir. Bu deney, daha kapsamlı bir veri ön işleme stratejisinin sınıflandırıcıların performansını nasıl etkilediğini anlamak için önemli bir adım olmuştur. Elde edilen bulgulara dayanarak, önerilen hibrit ön işleme yönteminin sağlık veri kümelerindeki sınıf dengesizliği ve eksik veriler gibi önemli veri problemlerini ele almak için etkili olduğu görülmektedir. Özellikle, PID, SİT ve KBH veri kümelerinde yapılan değerlendirmelerde, önde gelen sınıflandırma algoritmalarının GASMOTEPSO_ENN ve GA-MICE yöntemleri uygulandıktan sonra belirgin bir şekilde yüksek performans sergilediği gözlemlenmiştir. PID veri kümesi üzerinde yapılan çalışmalarda,

SVM ve XGBoost gibi önde gelen algoritmaların, önerilen ön işleme yöntemleri uygulandıktan sonra yüksek doğruluk, kesinlik, duyarlılık, F1-skoru ve MCC gibi model değerlendirme metriklerinde önemli bir başarı elde ettiği belirlenmiştir. Benzer şekilde, SİT ve KBH veri kümelerinde de tüm sınıflandırıcıların yüksek performans sergilediği gözlemlenmiştir. Bu bulgular, önerilen ön işleme yönteminin sınıflandırıcıların performansını artırmada etkili olduğunu ve gerçek dünya verileri üzerinde önemli bir adım olduğunu vurgulamaktadır. Deneysel sonuçlar, GASMOTEPSO_ENN ve GA-MICE yöntemlerinin, veri dengesini sağlamak ve eksik değerleri ele almak için etkili olduğunu açıkça göstermektedir. GASMOTEPSO_ENN yöntemi, veri kümesindeki dengesizliği gidermek için SMOTE örnekleme tekniğini GA ve PSO sezgisel yöntemleriyle birleştirirken, GA-MICE yöntemi eksik değerlerin güvenilir bir şekilde tespit edilmesini ve atanmasını sağlayarak sınıflandırma algoritmalarının önyargılı olmasını engeller. Ayrıca, literatürdeki benzer çalışmalarla karşılaştırıldığında, önerilen ön işleme yönteminin üstün performans sergilediği görülmektedir. PID, SİT ve KBH veri kümelerinde elde edilen sonuçlar, önerilen yöntemin diğer yöntemlere kıyasla daha etkili ve güvenilir olduğunu göstermektedir. Özellikle, önerilen yöntemin %93 ile %100 arasında değişen doğruluk, kesinlik, duyarlılık, F1-skoru ve AUC değerleri elde ettiği gözlemlenmiştir.

Bu çalışmanın bulguları, eksik değerlerin ve sınıf dengesizliğinin sınıflandırıcıların performansını önemli ölçüde etkileyebileceğini göstermektedir. Ayrıca, bu tür veri problemlerini ele almak için etkili ön işleme yöntemlerinin kullanılmasının önemli olduğunu vurgulamaktadır. Sonuç olarak, bu çalışma, sağlık veri analizinde karşılaşılan yaygın veri problemlerini ele almak için çeşitli ön işleme yöntemlerinin etkinliğini araştırmıştır. Bulgular, daha sağlıklı ve güvenilir sonuçlar elde etmek için uygun ön işleme stratejilerinin benimsenmesinin gerekliliğini ortaya koymaktadır. Tez kapsamında önerilen hibrit ön işleme yöntemi, gerçek dünya verileri üzerinde başarılı bir şekilde uygulanabilir ve sağlık sektöründe karar alma süreçlerini iyileştirmeye yönelik önemli bir katkı sağlayabilir.

6. SONUÇ VE ÖNERİLER

Hızla artan veriler sağlık verilerinin makine öğrenmesi yöntemlerine ihtiyacını da beraberinde getirmiştir. Bu tez çalışması, sağlık veri analizinde sıkça karşılaşılan eksik değerler ve sınıf dengesizliği gibi önemli veri problemlerini ele alan özgün bir hibrit ön işleme modelinin geliştirilmesini ve değerlendirilmesini amaçlamıştır. Çalışmanın bulguları, sağlık sektöründeki veri analizi ve hastalıkların tahmini gibi alanlarda sıkça karşılaşılan bu tür zorlukların üstesinden gelmek için etkili bir yaklaşım sunmaktadır.

Sağlık alanında, eksik değerlerle ve dengesiz veri kümeleriyle sıkça karşılaşılmasına rağmen, bu tür verilerle yapılan çalışmaların sonuçlarının doğruluğu ve pratik uygulanabilirliği genellikle sınırlıdır. Araştırmacılar ve hekimler, dengesiz ve eksik değerlerle başa çıkmak için analiz yöntemleri ve araçlarının geliştirilmesine ihtiyaç duymaktadırlar. Bu tez çalışmasında ilk olarak, sağlık verilerinin analizinde bu iki sorunun üstesinden gelmek için birçok problemin optimize edilmesinde başarıyla kullanılan sezgisel yöntemler ile makine öğrenmesi algoritmalarını birleştiren literatürdeki çeşitli çalışmalar incelenmiş ve bu çalışmaların sonuçları bir kitap bölümünde yayınlanmıştır [2]. İncelenen çalışmaların ışığında, sezgisel yöntemlerin sağlık veri analizi alanında da yapılan araştırmalara önemli katkılar sağladığı gözlemlenmiştir.

Sağlık veri analizindeki eksik veri ve dengesiz sınıf dağılımı gibi temel zorlukların etkili bir şekilde ele alınması, makine öğrenmesi sınıflandırıcılarının performansını önemli ölçüde etkileyebilir. Bu tez çalışmasında ikinci olarak, sağlık veri kümelerinde sıkça karşılaşılan bu eksik veri ve dengesiz sınıf dağılımı problemlerinin makine öğrenmesi sınıflandırıcılarının performansını nasıl etkilediğini detaylı bir şekilde incelenmiştir. Bu bağlamda, literatürde başarıyla kullanılan örnekleme yöntemleri ile dengesiz dağılımlı sınıfların dengelemesine odaklanılmış ve eksik değerlerin tamamlanmasında etkili olan stratejilerin kapsamlı bir analizi gerçekleştirilmiştir. Bu analiz, sağlık veri kümelerinde eksik verilerin ve dengesiz sınıfların yaygınlığını ve etkilerini detaylı bir şekilde araştırmıştır. Özellikle, sağlık veri analizi gibi hassas alanlarda, eksik verilerin sınıflandırıcıların doğruluğu ve güvenilirliği üzerinde önemli bir etkiye sahip olduğu gözlemlenmiştir. Ayrıca, dengesiz sınıf dağılımının nadir sınıfların tanınmasını zorlaştırdığı ve modelin çoğunluk sınıfına yanlış şekilde eğilimli olmasına neden

olduğu belirlenmiştir. Bu analizin bulguları, makale formatında yayınlanmış [28] ve aynı zamanda poster bildirisi olarak 9. Tıp Dünyası Kongre'sinde sunulmuştur [75].

Sağlık veri analizinde sınıf dengesizliği, özellikle azınlık ve çoğunluk sınıfların doğru sınıflandırılması açısından önemli bir zorluk teşkil etmektedir. Bu tez çalışmasında üçüncü olarak, dengesiz dağılımlı sınıfların düzeltilmesine yönelik bir ön işleme yöntemi olarak GA ve PSO sezgisel yöntemleri ile SMOTE ve ENN yöntemlerini birleştiren GASMOTEPSO_ENN yöntemini geliştirilmiştir. Bu analiz, GASMOTEPSO_ENN yöntemi ile dengelenen veri kümesindeki bireylerin hasta veya sağlıklı olarak teşhisinde artan performansın pratik uygulamaları için önemli bir potansiyeli vurgulamaktadır. Özellikle, dengesizliklere sahip veri kümelerindeki artan doğruluk oranı, yanlış teşhisleri azaltma ve tedavi planlamasını optimize etme fırsatını işaret etmektedir. Bu yöntem, sağlık veri kümelerinde sıkça karşılaşılan sınıf dengesizliği gibi yaygın zorlukları ele almak için değerli bir araç oluşturmakla birlikte araştırmacılar ile hekimler için hastalık teşhisi ve tedavisinde pratik faydalar sunmaktadır. Bu analizin bulguları, makale formatında yayınlanmıştır [72].

Sağlık veri kümelerinde özellikle nadir hastalıklarda sıkça karşılaşılan aşırı dengesiz veri kümeleri, makine öğrenmesi yöntemlerinin etkinliğini önemli ölçüde etkileyebilir. Bu tez çalışmasında ayrıca, yüksek oranda sınıf dengesizliğinin XGBoost algoritmasının başarısına etkisi ve literatürde en sık kullanılan eksik değer doldurma ve sınıf dengeleme yöntemlerinin araştırılması yapılmıştır. Bu çalışma, yüksek oranda sınıf dengesizliği içeren inme tahmini veri kümesinde eksik değerler ve dengesiz veri kümeleri gibi zorlukları aşmak için literatürde başarıyla kullanılan yöntemlerin etkinliğini XGBoost makine öğrenmesi yöntemiyle kapsamlı bir karşılaştırmalı analizini sunmaktadır. Elde edilen bulgular, SMOTEENN yönteminin sınıf dengesizliğinin etkisini azaltma ve eksik veri sorunlarını yönetme konusundaki etkinliğini vurgulamıştır. Bu analiz, uygun örnekleme ve eksik veri doldurma yöntemlerinin XGBoost algoritmasının performansını artırmak için önemli olduğunu göstermiştir. Bu çalışma, eksik veri ve dengesiz veri kümeleri gibi temel zorluklara odaklanan detaylı bir analiz sunması ve inme tahmini modellerinin doğruluğunu ve güvenilirliğini artırmak için uygun yöntemlerin önemini vurgulamaktadır. Bu çalışmanın bulguları, bildiri formatında hazırlanarak uluslararası konferansta sunulmuştur [76].

Tez çalışmamızın temel odak noktası, sağlık verilerinin analizinde yaygın olarak görülen hem eksik değer hem de sınıf dengesizliği problemlerine yönelik bir hibrit ön işleme yönteminin

geliştirilmesidir. Bu yöntem, GASMOTEPSO_ENN ve GA-MICE yöntemlerini birleştirerek sınıf dengesizliğini düzeltmeyi ve eksik değerleri doldurmayı hedeflemektedir. Geliştirilen hibrit ön işleme yönteminde, literatürde farklı problemlerin optimize edilmesinde başarıyla kullanılan GA ve PSO sezgisel yöntemlerinden yararlanılmıştır. Yapılan deneyler, önerilen hibrit ön işleme yönteminin sınıflandırıcıların performansını önemli ölçüde artırdığını göstermektedir. Özellikle, bu yöntemin kullanılmasıyla, sınıflandırıcıların doğruluğunda ve diğer performans metriklerinde belirgin bir iyileşme sağlanmıştır. Bu bağlamda, önerilen yöntemin, sağlık verilerindeki eksik değerler ve sınıf dengesizliği gibi yaygın zorlukları ele almak için etkili bir araç olduğu ve literatürdeki benzer çalışmalara göre üstün performans sergilediği belirlenmiştir. Bu analizin bulguları, makale formatında düzenlenerek dergiye gönderilmiştir ve şu an değerlendirme aşamasındadır.

Bu tez çalışması, sağlık veri analizi alanında çalışan araştırmacıların ve uygulayıcıların, veri problemlerini dikkatlice ele alarak daha sağlıklı ve güvenilir sonuçlar elde etmelerine yardımcı olması açısından önemlidir. Ayrıca, önerilen hibrit ön işleme yönteminin, sağlık sektöründe karar alma süreçlerini iyileştirmeye yönelik önemli bir katkı sağlayabileceği düşünülmektedir. Bu çalışmanın bulguları, sağlık verilerinin daha etkin bir şekilde analiz edilmesi ve hastalıkların daha doğru bir şekilde tahmin edilmesi için önemli bir adım olarak değerlendirilebilir ve araştırmacılara rehberlik edebilir. Gelecekteki çalışmalar, sağlık veri analizi ve sınıflandırma yöntemlerinin daha kapsamlı bir biçimde incelenmesine odaklanabilir. Özellikle, ikili sınıflandırma senaryolarından çok sınıflı ve dengesiz veri kümelerine genişletme çabaları büyük önem taşımaktadır. Bu genişleme, tıbbi durumların ve sınıflandırma karmaşıklıklarının daha iyi yansıtılmasına ve bu alandaki karar alma süreçlerinin iyileştirilmesine katkı sağlayabilir. Gelecekteki araştırmaların, özellik seçimi, yeniden örnekleme, eksik değer atama ve makine öğrenmesi yöntemlerinin entegrasyonunu daha fazla vurgulayabileceği düşünülebilir. Bu, klinik uygulamalarda daha etkili karar almayı teşvik ederek tahmin doğruluğunu ve model genellemesini artırabilir. Ayrıca, yöntemlerin pratikliği ve gerçek tıbbi senaryolardaki etkinliği daha fazla değerlendirilebilir ve klinik uzmanlardan geri bildirim alınabilir. Bu öneriler, gelecek araştırmacıların dengesiz veri ve eksik değer sorunlarını ele almak ve klinik uygulamalarda daha gelişmiş çözümler geliştirmek için rehberlik sağlayabilir.

KAYNAKLAR

[1]. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, [Ziyaret Tarihi:20 Şubat 2024].

[2]. Hatice Nizam Ozogur and Zeynep Orman. The effect of heuristic methods toward performance of health data analysis. In Next Generation Healthcare Informatics, pages 147–171. Springer, 2022.

[3]. Gangavarapu Sailasya and Gorli L Aruna Kumari. Analyzing the performance of stroke prediction using ml classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6), 2021.

[4]. Hamza Al-Zubaidi, Mohammed Dweik, and Amjed Al-Mousa. Stroke prediction using machine learning classification methods. In 2022 International Arab Conference on Information Technology (ACIT), pages 1–8. IEEE, 2022.

[5]. Tianyu Liu, Wenhui Fan, and Cheng Wu. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial intelligence in medicine, 101:101723, 2019.

[6]. Yun-Chun Wang and Ching-Hsue Cheng. A multiple combined method for rebalancing medical data with class imbalances. Computers in Biology and Medicine, 134:104527, 2021.

[7]. Nur Ulfa Maulidevi, Kridanto Surendro, et al. Smote-lof for noise identification in imbalanced data classification. Journal of King Saud University-Computer and Information Sciences, 34(6):3413–3423, 2022.

[8]. Badiuzzaman Pranto, Sk Maliha Mehnaz, Sifat Momen, and Syed Maruful Huq. Prediction of diabetes using cost sensitive learning and oversampling techniques on bangladeshi and indian female patients. In 2020 5th international Conference on information technology research (ICITR), pages 1–6. IEEE, 2020.

[9]. Pankaj Chittora, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat, Tulika Chakrabarti, Zbigniew Leonowicz, Micha l Jasinski, Lukasz Jasinski, Radomir Gono, Elzbieta Jasinska, et al. Prediction of chronic kidney disease - a machine learning perspective. IEEE Access, 9:17312–17334, 2021.

[10]. Ibomoiye Domor Mienye and Yanxia Sun. Performance analysis of costsensitive learning methods with application to imbalanced medical data. Informatics in Medicine Unlocked, 25:100690, 2021.

[11]. Dan Gan, Jiang Shen, Bang An, Man Xu, and Na Liu. Integrating tanbn with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Computers & Industrial Engineering, 140:106266, 2020.

[12]. Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta, Saqib Hakak, Wazir Zada Khan, Ali Kashif Bashir, Alireza Jolfaei, and Usman Tariq. Antlion re-sampling based deep

neural network model for classification of imbalanced multimodal stroke dataset. Multimedia Tools and Applications, pages 1–25, 2020.

[13]. Zhaozhao Xu, Derong Shen, Tiezheng Nie, and Yue Kou. A hybrid sampling algorithm combining m-smote and enn based on random forest for medical imbalanced data. Journal of Biomedical Informatics, 107:103465, 2020.

[14]. Divyaansh Devarriya, Cairo Gulati, Vidhi Mansharamani, Aditi Sakalle, and Arpit Bhardwaj. Unbalanced breast cancer data classification using novel fitness functions in genetic programming. Expert Systems with Applications, 140:112866, 2020.

[15]. Shahidul Islam Khan and Abu Sayed Md Latiful Hoque. Sice: an improved missing data imputation technique. Journal of big Data, 7(1):1–21, 2020.

[16] PS Raja and KJSC Thangavel. Missing value imputation using unsupervised machine learning techniques. Soft Computing, 24(6):4361–4392, 2020.

[17]. G Madhu, B Lalith Bharadwaj, G Nagachandrika, and K Sai Vardhan. A novel algorithm for missing data imputation on machine learning. In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), pages 173–177. IEEE, 2019.

[18]. Pooja Rani, Rajneesh Kumar, and Anurag Jain. Hioc: a hybrid imputation method to predict missing values in medical datasets. International Journal of Intelligent Computing and Cybernetics, 14(4):598–616, 2021.

[19]. Qian Wang, Weijia Cao, Jiawei Guo, Jiadong Ren, Yongqiang Cheng, and Darryl N Davis. Dmp mi: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. IEEE access, 7:102232–102238, 2019.

[20]. Shivani Yadav, Yogendra PS Maravi, Jitendra Agrawal, and Nishchol Mishra. A neural network based diabetes prediction on imbalanced data. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), pages 515–521. IEEE, 2021.

[21]. Meshrif Alruily, Sameh Abd El-Ghany, Ayman Mohamed Mostafa, Mohamed Ezz, and AA Abd El-Aziz. A-tuning ensemble machine learning technique for cerebral stroke prediction. Applied Sciences, 13(8):5047, 2023.

[22]. Yuru Jing. Machine learning performance analysis to predict stroke based on imbalanced medical dataset. In CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms, pages 1–7. VDE, 2022.

[23]. Chirag Rana, Nikita Chitre, Bhargavi Poyekar, and Pramod Bide. Stroke prediction using smote-tomek and neural network. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–5. IEEE, 2021.

[24]. Sabrina Jahan Maisha, Ety Biswangri, Mohammad Shahadat Hossain, and Karl Andersson. An approach to detect chronic kidney disease (ckd) by removing noisy and inconsistent values of uci dataset. In Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering: TCCE 2021, pages 457–472. Springer, 2022.

[25]. Muhammad Raihan Wibowo and Irma Palupi. Enhancing accuracy on chronic-kidney disease detection using machine learning with technique of resampling and missing

value treatment. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 2023.

[26]. Abdullah Al Imran, Md Nur Amin, and Fatema Tuj Johora. Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning. In 2018 International Conference on Innovation in Engineering and Technology (ICIET), pages, 1–6. IEEE, 2018.

[27]. Ethem Alpaydin. Introduction to machine learning. MIT press, 2020.

[28]. Hatice Nizam Özoğur and Zeynep Orman. Sağlık verilerinin analizinde veri on işleme adımlarının makine öğrenmesi yöntemlerinin performansına etkisi. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 16(1):23–33, 2023.

[29]. Thomas R Vetter and Patrick Schober. Regression: the apple does not fall far from the tree. Anesthesia & Analgesia, 127(1):277–283, 2018.

[30]. David G Kleinbaum, K Dietz, M Gail, M Klein, and M Klein. Logistic regression. A Self-Learning Tekst, 2008.

[31]. Leo Breiman. Random forests. Machine learning, 45:5–32, 2001.

[32]. Steven J Rigatti. Random forest. Journal of Insurance Medicine, 47(1):31–39, 2017.

[33]. Arti Patle and Deepak Singh Chouhan. Svm kernel functions for classification. In 2013 International conference on advances in technology and engineering (ICATE), pages 1–9. IEEE, 2013.

[34]. Chee Sun Lee, Peck Yeng Sharon Cheang, and Massoud Moslehpour. Predictive analytics in business analytics: decision tree. Advances in Decision Sciences, 26(1):1–29, 2022.

[35]. Lior Rokach and Oded Maimon. Decision trees. Data mining and knowledge discovery handbook, pages 165–192, 2005.

[36]. Daniel Berrar. Bayes' theorem and naive bayes classifier. Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics, 403:412, 2018.

[37]. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.

[38]. Reza Piraei, Seied Hosein Afzali, and Majid Niazkar. Assessment of xgboost to estimate total sediment loads in rivers. Water Resources Management, 37(13):5289–5306, 2023.

[39]. Weinan Dong, Daniel Yee Tak Fong, Jin-sun Yoon, Eric Yuk Fai Wan, Laura Elizabeth Bedford, Eric Ho Man Tang, and Cindy Lo Kuen Lam. Generative adversarial networks for imputing missing data for big data clinical research. BMC medical research methodology, 21:1–10, 2021.

[40]. Rajni Rajni and Amandeep Amandeep. Rb-bayes algorithm for the prediction of diabetic in pima indian dataset. International Journal of Electrical and Computer Engineering, 9(6):4866, 2019.

[41]. Aixia Guo, Randi E Foraker, Robert M MacGregor, Faraz M Masood, Brian P Cupps, and Michael K Pasque. The use of synthetic electronic health record data and deep learning

to improve timing of high-risk heart failure surgical intervention by predicting proximity to catastrophic decompensation. Frontiers in digital health, 2:576945, 2020.

[42]. Suraj Kumar Gupta, Aditya Shrivastava, SP Upadhyay, and Pawan Kumar Chaurasia. A machine learning approach for heart attack prediction. International Journal of Engineering and Advanced Technology (IJEAT), 10(6):1–11, 2021.

[43]. Boseong Seo, Jaekyung Shin, Taejin Kim, and Byeng D Youn. Missing data imputation using an iterative denoising autoencoder (idae) for dissolved gas analysis. Electric Power Systems Research, 212:108642, 2022.

[44]. Robert W Krause, Mark Huisman, Christian Steglich, and Tom AB Snijders. Missing network data a comparison of different imputation methods. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 159–163. IEEE, 2018.

[45]. Jinyan Li, Lian-sheng Liu, Simon Fong, Raymond K Wong, Sabah Mohammed, Jinan Fiaidhi, Yunsick Sung, and Kelvin KL Wong. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. PloS one, 12(7):e0180830, 2017.

[46]. Koichi Fujiwara, Yukun Huang, Kentaro Hori, Kenichi Nishioji, Masao Kobayashi, Mai Kamaguchi, and Manabu Kano. Over-and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. Frontiers in Public Health, 8, 2020.

[47]. Vanaja Ramaswamy and Saswati Mukherjee. An effective clinical decision support system using swarm intelligence. The Journal of Supercomputing, 76(9):6599–6618, 2020.

[48]. Tince Etlin Tallo and Aina Musdholifah. The implementation of genetic algorithm in smote (synthetic minority oversampling technique) for handling imbalanced dataset problem. In 2018 4th international conference on science and technology (ICST), pages 1–4. IEEE, 2018.

[49]. Apurva Sonak, Ruhi Patankar, and Nitin Pise. A new approach for handling imbalanced dataset using ann and genetic algorithm. In 2016 International Conference on Communication and Signal Processing (ICCSP), pages 1987–1990. IEEE, 2016.

[50] Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web, 16(4):449–475, 2013.

[51]. Andrea Dal Pozzolo, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. Expert systems with applications, 41(10):4915–4928, 2014.

[52]. Haizhou Wang, Anoop Singhal, and Peng Liu. Tackling imbalanced data in cybersecurity with transfer learning: a case with rop payload detection. Cybersecurity, 6(1):1–15, 2023.

[53]. Hong Qian, Xiantao Zhang, Chaofan Zhang, and Cheng Jiang. A novel cyber intrusion detection model based on improved hybrid sampling. Transactions of the Institute of Measurement and Control, page 01423312231158422, 2023.

[54]. Sanjeev Rao, Anil Kumar Verma, and Tarunpreet Bhatia. Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data. Expert Systems with Applications, 217:119594, 2023.

[55]. Manoj Sethi, Naman Tyagi, Parmeet Singh Kalsi, and Parupalli Atchuta Rao. Deep learning-based binary classification for spam detection in sms data: Addressing imbalanced data with sampling techniques. In 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), pages 1–9. IEEE, 2023.

[56]. Chensu Zhao, Yang Xin, Xuefeng Li, Yixian Yang, and Yuling Chen. A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data. Applied Sciences, 10(3):936, 2020.

[57]. Mulyana Saripuddin, Azizah Suliman, Sera Syarmila Sameon, and Bo Norregaard Jorgensen. Random undersampling on imbalance time series data for anomaly detection. In Proceedings of the 2021 4th International Conference on Machine Learning and Machine Intelligence, pages 151–156, 2021.

[58]. Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems (ICICS), pages 243–248. IEEE, 2020.

[59]. Firuz Kamalov, Ho-Hon Leung, and Aswani Kumar Cherukuri. Keep it simple: random oversampling for imbalanced data. In 2023 Advances in Science and Engineering Technology International Conferences (ASET), pages 1–4. IEEE, 2023.

[60]. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[61]. Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pages 1322–1328. IEEE, 2008.

[62]. Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1):20–29, 2004.

[63]. Gustavo EAPA Batista, Ana LC Bazzan, Maria Carolina Monard, et al. Balancing training data for automated annotation of keywords: a case study. Wob, 3:10–8, 2003.

[64]. James Kennedy and Russell Eberhart. Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks, volume 4, pages 1942–1948. ieee, 1995.

[65] Federico Marini and Beata Walczak. Particle swarm optimization (pso). A tutorial. Chemometrics and Intelligent Laboratory Systems, 149:153–165, 2015.

[66]. Hatice Nizam Ozogur, Gokhan Ozogur, and Zeynep Orman. Blood glucose level prediction for diabetes based on modified fuzzy time series and particle swarm optimization. Computational Intelligence, 37(1):155–175, 2021.

[67]. John H Holland. Genetic algorithms. Scientific american, 267(1):66–73, 1992.

[68]. Kaggle veri kümeleri. https://www.kaggle.com/datasets, [Ziyaret Tarihi: 22 Şubat 2023].

[69]. Pima indians diyabet veri k¨umesi. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, [Ziyaret Tarihi: 22 Şubat 2023].

[70]. Serebral inme tahmini veri kümesi. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset, [Ziyaret Tarihi: 01 Nisan 2023].

[71]. Kronik böbrek hastalığı veri kümesi. https://www.kaggle.com/datasets/mansoordaku/ckdisease, [Ziyaret Tarihi: 22 Şubat 2023].

[72]. Hatice Nizam-Ozogur and Zeynep Orman. A heuristic-based hybrid sampling method using a combination of smote and enn for imbalanced health data. Expert Systems, page e13596, 2024.

[73]. Kun Jiang, Jing Lu, and Kuiliang Xia. A novel algorithm for imbalance data classification based on genetic algorithm improved smote. Arabian journal for science and engineering, 41:3255–3266, 2016.

[74]. Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodriguez, Asdrubal Lopez, Jose Ruiz Castilla, and Adrian Trueba. Pso-based method for svm classification on skewed data sets. Neurocomputing, 228:187–197, 2017.

[75]. NİZAM ÖZOĞUR, H., & ORMAN, Z., (2023). Sağlık Verilerinin Makine Öğrenimi Tabanlı Analizinde Veri Ön İşleme Stratejileri ve Performans Analizinin Rolü. 9. Türk Tıp Dünyası Kongresi (pp.1-5). İstanbul, Turkey.

[76]. Nizam Özoğur, H., & Orman, Z., (2024). A Comparative Study of Preprocessing Techniques for Stroke Prediction using XGBoost Classifier. 3rd International Conference On Advanced Engineering, Technology And Applications, Catania, Italy.

İNTİHAL RAPORU İLK SAYFASI

Hatice NİZAM ÖZOĞUR

ORİJİNALLİK RAPORU

%5	%4	%3	%0
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	acikbilim.yok.gov.tr İnternet Kaynağı	%1
2	toad.halileksi.net İnternet Kaynağı	%1
3	dergipark.org.tr İnternet Kaynağı	<%1
4	acikerisim.omu.edu.tr İnternet Kaynağı	<%1
5	acikerisim.iku.edu.tr İnternet Kaynağı	<%1
6	acikerisim.bartin.edu.tr İnternet Kaynağı	<%1
7	smote-variants.readthedocs.io İnternet Kaynağı	<%1
8	Gümüştaş, Enis. "Kayıp gözlem içeren dengesiz veri setlerinin topluluk öğrenme algoritmaları ile sınıflandırılması", Mimar Sinan Fine Arts University (Turkey), 2024 Yayın	<%1