TOPLULUK ÖĞRENME YÖNTEMLERİNİN HASTALIKLARIN TEŞHİSİNDE KULLANIMI

2023
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ

Emin FARZALIYEV

Tez Danışmanı
Dr. Öğr. Üyesi Emrullah SONUÇ

TOPLULUK ÖĞRENME YÖNTEMLERİNİN HASTALIKLARIN TEŞHİSİNDE KULLANIMI

Emin FARZALIYEV

Tez Danışmanı
Dr. Öğr. Üyesi Emrullah SONUÇ

T.C.
Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalında
Yüksek Lisans Tezi
Olarak Hazırlanmıştır

KARABÜK
Temmuz 2023

Emin FARZALIYEV tarafından hazırlanan "TOPLULUK ÖĞRENME YÖNTEMLERİNİN HASTALIKLARIN TEŞHİSİNDE KULLANIMI" başlıklı bu tezin Yüksek Lisans Tezi olarak uygun olduğunu onaylarım.

Dr. Öğr. Üyesi Emrullah SONUÇ .....
Tez Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

Bu çalışma, jürimiz tarafından Oy Birliği ile Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir. 10/07/2023

Ünvanı, Adı SOYADI (Kurumu)	İmzası
Başkan : Doç. Dr. Rafet DURGUT (BANÜ)	ONLİNE
Üye : Dr. Öğr. Üyesi Emrullah SONUÇ (KBÜ)	.....
Üye : Dr. Öğr. Üyesi Kürşat Mustafa KARAOĞLAN (KBÜ)	.....

KBÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulu, bu tez ile, Yüksek Lisans derecesini onamıştır.

Prof. Dr. Müslüm KUZU .....
Lisansüstü Eğitim Enstitüsü Müdürü

"Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim."

Emin FARZALIYEV

ÖZET

Yüksek Lisans Tezi

TOPLULUK ÖĞRENME YÖNTEMLERİNİN HASTALIKLARIN TEŞHİSİNDE KULLANIMI

Emin FARZALIYEV

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Emrullah SONUÇ
Temmuz 2023, 55 sayfa

Bu çalışmada, Tiroit ve çocuklardaki Anemi hastalığını tahmin etmek için topluluk öğrenme (ensemble learning) tekniklerinin kullanımı araştırılmıştır. Bu iki hastalığı tahmin etmek için Karar Ağacı, Destek Vektör Makineleri, Rastgele Orman, Lojistik Regresyon, K-En Yakın Komşu gibi çeşitli makine öğrenmesi algoritmaları test edilmiştir. Daha sonra bu sınıflandırıcılar, Torbalama, Artırma, İstifleme gibi öğrenme teknikleri kullanılarak daha doğru ve güçlü bir tahmin modeli oluşturulması amaçlanmıştır. Bu çalışmada, Anemi hastalığının tahmini için kullanılan veri seti, Haditha Genel Hastanesi ve kliniklerinde toplanan 600 örneği içermektedir. Bu örneklerin 429'u anemi hastası iken, 171'i anemi hastası değildir. Veri setinde her bir örneğe ait 31 özellik bulunmaktadır. Tiroit hastalığının tahmini için kullanılan veri seti ise 1 ila 90 yaş arası Iraklı erkek ve kadınlardan alınan 1250 örneği içermektedir. İlgili veri setinde her bir örneğe ait 17 özellik bulunmaktadır. Farklı topluluk tekniklerinin performansı veri setleri üzerinde değerlendirilmiştir. Sonuçlara göre,

Anemi hastalığının tahmini için topluluk öğrenme teknikleri, bireysel sınıflandırıcılara göre daha düşük doğrulukla tahminde bulunmuştur. Ayrıca, topluluk öğrenme teknikleri arasında artırma yönteminin en yüksek doğruluk oranına (%100) ulaştığı gözlemlenmiştir. Diğer taraftan, Tiroit hastalığı için olan sonuçlara göre topluluk öğrenme teknikleri ve bireysel sınıflandırıcılar, birbirine yakın doğrulukta tahminde bulunmuştur. Fakat topluluk öğrenme teknikleri arasında yine artırma yönteminin en yüksek doğruluğa (%89,6) eriştiği görülmüştür. Bu çalışma, tiroit ve çocuklardaki anemi hastalığını tahmin etmek için topluluk öğrenme tekniklerinin farklı bir yaklaşım olabileceğini göstermektedir. Ancak, gelecekteki araştırmalarda veri ön işleme, özellik seçimi gibi yöntemlerin topluluk öğrenme modellerinin performansını artırmada etkili olabileceği düşünülmektedir.

Anahtar Sözcükler : Makine öğrenmesi, Topluluk öğrenme, Anemi hastalığı, Tiroit hastalığı.

Bilim Kodu **:**92431

ABSTRACT

Master Thesis

THE USE OF ENSEMBLE LEARNING METHODS IN THE DIAGNOSIS OF DISEASES

Emin FARZALIYEV

Karabük University
Institute of Graduate Programs
Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Emrullah SONUÇ
July 2023, 55 pages

In this study, the utilization of ensemble learning techniques has been investigated for predicting Thyroid disease and Anemia disease in children. Various machine learning algorithms such as Decision Trees, Support Vector Machines, Random Forests, Logistic Regression, and K-Nearest Neighbors were tested to predict these two diseases. Subsequently, these classifiers were used in ensemble learning techniques such as bagging, boosting, and stacking to create a more accurate and robust prediction model. The dataset used for predicting Anemia in this study comprises 600 samples collected from Haditha General Hospital and clinics. Among these samples, 429 are Anemic patients, while 171 are non-Anemic. Each sample in the dataset contains 31 features. The dataset used for predicting Thyroid disease consists of 1250 samples obtained from Iraqi males and females aged between 1 and 90. Each sample in the relevant dataset contains 17 features. The performance of different ensemble techniques was evaluated on the datasets.

According to the results, ensemble learning techniques for predicting Anemia yielded lower accuracy compared to individual classifiers. Additionally, it was observed that the boosting method achieved the highest accuracy rate (100%). On the other hand, the results for predicting Thyroid disease showed that ensemble learning techniques and individual classifiers yielded similar accuracies. However, once again, the boosting method achieved the highest accuracy (89,6%) among the ensemble learning techniques. This study demonstrates that ensemble learning techniques can be a different approach for predicting Thyroid disease and Anemia in children. However, it is believed that future research should focus on methods such as data preprocessing and feature selection to improve the performance of ensemble learning models.

Key Word **:** Machine learning, Ensemble learning, Anemia disease, Thyroid disease.

Science Code : 92431

TEŞEKKÜR

Bu tez çalışmasının planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, Dr. Öğr. Üyesi Emrullah SONUÇ'a ve çalışmada kullanılan veri setlerini bulmamda yardımcı olan Qusay Luay SAIHOOD ve Khalid SALMAN'a tüm kalbimle teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

KISALTMALAR

LR : Lojistik Regresyon

RO : Rastgele Orman

KEYK : K-En Yakın Komşu

KA : Karar Ağacı

DVM : Destek Vektör Makineleri

YSA : Yapay Sinir Ağları

ESA : Evrişimsel Sinir Ağları

UCI : University of California Irvine

ROC : Receiver Operating Characteristic

SİMGELER

T3 : Tri-iyodotironin

T4 : L-tiroksin

BÖLÜM 1

GİRİŞ

1.1. PROBLEMİN TANIMLANMASI

Yapay Zeka, hastalıkların tahmini ve teşhisi için önemli bir araç haline gelmiştir. Çeşitli yapay zeka yöntemleri, hastalıkların erken teşhisi, tedavi planlaması ve hastalık tahmini gibi konularda yardımcı olabilir. Makine öğrenmesi, yaygın kullanılan yapay zeka yöntemlerindendir.

Makine öğrenmesi, büyük veri kümelerindeki desenleri tanıyarak ve istatistiksel modeller oluşturarak hastalıkların tahmininde kullanılabilir. Örneğin, kanser tanısı veya kalp hastalıklarının riskini belirleme gibi konularda makine öğrenmesi yöntemleri kullanılabilir [1,2]. Bu yöntemlerin avantajları arasında şunlar yer alır: Verilerin daha hızlı ve daha doğru bir şekilde analiz edilmesi, daha iyi anlaşılması ve daha iyi bir şekilde yorumlanması, hastalıkların daha erken teşhis edilmesi, tedavi edilmesi, daha iyi bir şekilde takip edilmesi ve yönetilmesi.

Makine öğrenmesi yöntemleri, hastalıkların tahmininde yardımcı olmak için bir araç olarak kullanılabilir. Ancak, kesin bir teşhis koyma veya tedavi planlama sürecinde uzman bir sağlık profesyonelinin değerlendirmesi ve onayı gereklidir. Ayrıca, bu yapay zeka yöntemlerinin doğruluğunu artırmak için sürekli olarak eğitilmeleri ve doğrulama verileriyle güncellenmeleri gerekmektedir.

Bu tezin konusu olan iki hastalıktan biri Anemi hastalığıdır. Anemi hastalığı, kırmızı kan hücrelerinin sayısının azalması veya hemoglobin proteininin normal seviyenin altına düşmesi durumunda ortaya çıkan bir hastalıktır. Aneminin üç önemli nedeni vardır: kan kaybı, yetersiz beslenme ve vücudun kırmızı kan hücrelerini üretme

yeteneğindeki bozukluklar. Anemi belirtileri arasında halsizlik, çabuk yorulma, uyku isteği ve soluk cilt yer alır [3].

Özellikle okul çağı altındaki çocukların anemi olma olasılığı daha yüksektir. Bu yaş grubundaki çocukların %42'sinin anemi hastalığına sahip olduğu tahmin edilmektedir. Aneminin birkaç çeşidi vardır, ancak özellikle beş yaşın altındaki çocuklar arasında demir eksikliği anemisi en yaygın anemi türlerinden biridir [3,4,5].

Bu tezde çalışılan bir diğer hastalık ise tiroit hastalığıdır. Tiroit, boyunda bulunan bir bezdir ve vücutta metabolizmayı düzenlemekle görevlidir [6]. Tiroit bezinin düzgün çalışması, vücuttaki enerji seviyelerini, kalp atış hızını, sindirimi ve diğer önemli işlevleri dengelemeye yardımcı olur. Ancak bazı durumlarda tiroit bezindeki fonksiyonlar bozulabilir ve tiroit hastalıklarına neden olabilir.

Tiroit hastalıkları, genellikle tiroit bezinin fazla çalışmasına (hipertiroidi), yetersiz çalışmasına (hipotiroidi) [6] veya nodüllerin oluşmasına bağlı olarak ortaya çıkar. Bununla birlikte, tiroit hastalıklarının belirtileri ve tedavileri hastalığın türüne bağlı olarak değişebilir.

Hipertiroidi, tiroit bezinin aşırı aktif olduğu durumlarda ortaya çıkar. Tiroit bezinden fazla miktarda tiroit hormonu salgılanır ve vücut metabolizması hızlanır. Hipertiroidi belirtileri arasında kilo kaybı, hızlı kalp atışı, sinirlilik, terleme, ellerde titreme, uykusuzluk ve yorgunluk yer alabilir. Hipertiroidi genellikle Graves hastalığı, tiroit nodülleri veya tiroidit gibi durumlarla ilişkilidir [7].

Hipotiroidi ise tiroit bezinin yetersiz çalıştığı durumlarda meydana gelir. Tiroit bezinin salgıladığı tiroit hormonu miktarı azalır ve vücut metabolizması yavaşlar [8]. Hipotiroidi belirtileri arasında yorgunluk, kilo alma, depresyon, soğuğa karşı duyarlılık, kabızlık, kuru cilt ve saçlar yer alabilir.

Tiroit nodülleri, tiroit bezinde oluşan küçük kitlelerdir. Nodüller genellikle iyi huyludur (kanserli değiller), ancak bazı durumlarda kanserle ilişkili olabilirler.

Nodüller genellikle ağrısızdır ve belirtiler göstermeyebilir, bu yüzden çoğu zaman tesadüfen tespit edilir [9].

Tiroit hastalıklarının teşhisi, genellikle tıbbi öykü, fizik muayene ve kan testleriyle yapılır. Tedavi, hastalığın türüne bağlı olarak değişir. Hipertiroidi tedavisi ilaçlar, radyoaktif iyot tedavisi veya cerrahi olabilir. Hipotiroidi tedavisi ise genellikle tiroit hormonu takviyesiyle sağlanır.

Tiroit bezi, birçok farklı kanserin gelişiminin gelişebileceği bir bölgedir. Tri-iyodotironin (T3) ve L-tiroksin (T4) olmak üzere iki hormon, tiroit tarafından üretilir. Bu hormonlar, metabolizmanın birçok yönünü kontrol eder [6]. Hipofiz bezinden salgılanan Tirotropin-Uyarıcı Edici Hormon, vücut daha fazla tiroit hormonuna sahip olduğunda tiroit bezine yönlendirilir. Ardından, Tirotropin-Uyarıcı Edici Hormon, tiroit bezine T4 ve T3 hormonlarının kontrolünü sağlar.

Çocuklarda anemi hastalığının erken evrede teşhis edilmesi son derece önemlidir. Anemi, vücutta yeterli miktarda sağlıklı kırmızı kan hücrelerinin üretilememesi veya yok edilmesi sonucu ortaya çıkar. Bu durum, çocuğun büyüme, gelişme ve genel sağlık durumu üzerinde olumsuz etkilere neden olabilir.

Erken teşhis, çocuğun tedavi sürecine hızla başlamasını sağlar, anemiye bağlı komplikasyonların önlenmesine yardımcı olur ve çocuğun demir eksikliği veya diğer anemi türleri gibi sorunları hızlı bir şekilde tespit etmeyi sağlar. Bu da tedavinin daha etkili olmasını ve çocuğun sağlıklı bir şekilde büyümesini destekler. Aynı zamanda, erken teşhis sayesinde aneminin altında yatan nedenler de belirlenebilir ve gerektiğinde ek tedavi yöntemleri uygulanabilir.

Aynı şekilde Tiroit hastalığının da erken evrede teşhis edilmesi son derece önemlidir. Tiroid, vücudun metabolizma, enerji düzenlemesi, büyüme ve gelişme gibi önemli işlevlerini kontrol eden bir bezdir. Erken teşhis, hastanın tedavi sürecine hızla başlamasını sağlar, olası komplikasyonları önlemeye yardımcı olur ve hastanın tiroit problemlerini hızlı bir şekilde tespit etmeyi ve uygun tedaviyi başlatmayı sağlar.

Tiroid hormon düzeylerinin kontrol edilmesi, ilaç tedavisi veya diğer yöntemlerle tiroid fonksiyonunun düzeltilmesi gereken durumlar tespit edilebilir.

1.2. ÇALIŞMANIN AMACI

Bu çalışmanın amacı, Tiroit ve çoçuklarda Anemi hastalığının tahmini için topluluk öğrenme yöntemlerinin performansını test etmektedir. Bunun için Karar Ağacı (KA), Destek Vektör Makineleri (DVM), Rastgele Orman (RO), Lojistik Regresyon (LR), K-En Yakın Komşu (KEYK) gibi çeşitli makine öğrenmesi algoritmaları test edilmiş, daha sonra bu sınıflandırıcılar, torbalama, artırma, istifleme gibi öğrenme teknikleri kullanılarak daha doğru ve güçlü bir tahmin modeli oluşturulmuş, doğruluk, kesinlik, duyarlılık ve F1-skoru gibi performans ölçümlerine bakılmış, sınıflandırıcıların performans ölçümleri karşılaştırılmıştır.

1.3. ORGANİZASYON ve YOL HARİTASI

Çalışmanın ilk bölümünde problemin tanımlanması ve çalışmanın amacı ile ilgili bilgiler yer almaktadır. Problemin tanımlanması kısmında, yapay zeka yöntemleriyle hastalıkların tahmini, anemi ve tiroit hastalığıyla ilgili temel bilgiler bahsedilmiştir.
İkinci bölümde, çalışmanın konusuyla ilgili son yıllarda yapılmış bazı çalışmalar incelenmiştir.

Bir sonraki bölümdeyse, hastalığın tahmini için kullanılan makine öğrenme ve topluluk öğrenme yöntemlerinden, veri önişlemeden, kullanılan veri setinden ve performans ölçümünden bahsedilmiştir.

Dördüncü bölümde, kullanılan yöntemlerin sonuçları karışıklık matrisleri ve tablolar halinde sunulmuş ve analiz edilmiştir. Son bölüm ise sonuç ve önerileri içermektedir.

BÖLÜM 2

LİTERATÜR ÇALIŞMALARINA GENEL BAKIŞ

2.1. ÇOCUKLARDA ANEMİ HASTALIĞI

Sarıbacak [10] yaptığı çalışmada, istatistiksel analizler Bağımsız Örneklemler T Testi, Mann-Whitney U Testi ve Lojistik Regresyon gibi yöntemlerin yanı sıra, makine öğrenme ve sınıflandırma algoritmaları da kullanılmıştır. KA, KEYK, Yapay Sinir Ağları (YSA) ve DVM gibi makine öğrenme sınıflama algoritmaları tanıtılmış ve performansları karşılaştırılmıştır. DVM algoritması 89,92 ile en düşük oranı elde etmiştir. Onu sırasıyla 94,14 sonucu ile YSA, 94,43 sonucu ile KEYK, 96,21 sonucu ile KA sınıflandırıcıları izlemektedir.

Khan vd. [11] ortak risk faktörlerini kullanarak makine öğrenimi algoritmalarını kullanarak çocuklarda (beş yaş altı) anemi durumunun tahmin edilmesi amaçlanmıştır. Çalışmada kullanılan veriler, 2011 yılında gerçekleştirilen ulusal temsili bir kesitsel araştırma olan Bangladeş Nüfus ve Sağlık Araştırması'ndan elde edilmiştir. Bu çalışmada, seçilen tüm değişkenlerle ilgili verilere sahip 2013 çocuktan oluşan bir örneklem seçilmiştir. Lineer diskriminant analizi, sınıflandırma ve regresyon ağaçları, KEYK, DVM, RO ve LR gibi birçok makine öğrenme algoritması kullanarak anemi durumunu tahmin etmek için sistematik bir değerlendirme yapılmıştır. RO algoritmasının en iyi sınıflandırma doğruluğunu (%68,53) göstermiştir. Öte yandan, klasik LR algoritması %62,75 sınıflandırma doğruluğuna sahip olmuş ve KEYK en düşük doğruluğu sağlamıştır.

Appiahene vd. [12] yaptıkları çalışmada, avuç içi kullanılarak anemiyi tespit etmek için Evrişimsel Sinir Ağları (ESA), KEYK, KA, Naif Bayes ve DVM'nin performansı karşılaştırılmıştır. Çalışmada 527 temel veri kümesi kullanılmış ve küçük veri kümeleri kullanmanın neden olduğu aşırı öğrenmeyi önlemek için

görüntü artırma tekniği kullanarak veri setlerinin boyutunu 2635'e çıkartılmıştır. Çalışmada kullanılan tüm modeller önemli sonuçlar üretmiştir. Naif Bayes modeli tüm modeller arasında en yüksek olan %99,96 doğruluk elde etmiştir. KEYK ve ESA her ikisi de %99,92 doğruluk sağlamıştır. KA ise %97,32 doğruluk elde etmiştir. DVM, tüm modeller arasında en düşük olan %94,94 doğruluğa sahip olmuştur.

El-kenawy [13] yapılan çalışmada, kan testi parametrelerine dayalı olarak hemoglobin seviyesinin tahmin edilmesi ve anemi sınıflandırması için makine öğrenme modeli önerilmiş, önerilen modelin sonuçları, farklı makine öğrenme modellerinden elde edilen sonuçlarla karşılaştırılmıştır. Araştırmada, iki makine öğrenimi görevi gerçekleştirilmiştir: regresyon ve sınıflandırma. Regresyon için, hematolojik parametreleri kullanarak Hemoglobin değerini tahmin edilmiştir. RO, Doğrusal Regresyon ve YSA makine öğrenimi algoritmaları kullanılmış. Çalışmada önerilen model olan birleştirilmiş model daha iyi bir doğruluk elde etmiştir (%99,7).

Asare vd. [14] yapılan çalışmada, Naif Bayes, ESA, DVM, KEYK ve KA algoritmalarının uygulanmasıyla demir eksikliği anemisini tespit etmek için bir makine öğrenimi yaklaşımı kullanılmıştır. Bu, konjonktiva, hissedilebilir avuç içi ve tırnak rengi görüntülerini çocuklarda anemi tespiti için hangi yöntemin daha yüksek doğruluğa sahip olduğunu belirlemek için karşılaştırma yapılmıştır. Kullanılan yöntem, veri seti toplama, veri seti ön işleme ve anemi tespiti için model geliştirme olmak üzere üç farklı aşamaya kategorize edilmiştir. ESA modeli ile (%99,12) gibi en yüksek doğruluğu elde ederken, DVM modeli ile (%95,4) en düşük doğruluk elde edilmiştir.

Dejene vd. [15] yapılan çalışmada kullanılan veriler, Etiyopya Demografik Sağlık Araştırması'ndan toplanmış ve makine öğrenimi algoritması için uygun olan kaliteli veriler elde etmek için ön işlemden geçirilmiştir. Hamile kadınlar arasındaki anemi seviyelerini tahmin eden bir model geliştirmek için KA, RO, CatBoost ve Aşırı Gradyan Artırma gibi modeller kullanılmıştır. Önerilen modelin oluşturulması için toplamda 29.104 örnek ve 23 özellik içeren on iki deney yapılmış ve eğitim ve test veri seti ayrım oranı 80/20 olarak belirlenmiştir. Sınıf ayrıştırma olmadan RO, Aşırı Gradyan Artırma ve CatBoost'ın genel doğruluk oranı sırasıyla %91,34, %94,26 ve

%97,08'dir. Bir birine karşı sınıf ayrıştırmasıyla RO, Aşırı Gradyan Artırma ve CatBoost'ın genel doğruluk oranı sırasıyla %94,4, %95,21 ve %97,44'tür. Geri kalanına karşı sınıf ayrıştırmasıyla RO, Aşırı Gradyan Artırma ve CatBoost'ın genel doğruluk oranı sırasıyla %94,4, %94,54 ve %97,6'dır.

Asare vd. [16] yapılan çalışmada, sağlık hizmetlerindeki makine öğrenimi trendlerini ve kavramlarını sistemik bir şekilde inceleyerek aneminin tespiti için uygun yaklaşımları belirlemek amacıyla bir sistemik derleme yapılmıştır. Makine öğrenimi algoritmaları, değerlendirme metrikleri, görüntü artırma yöntemleri ve kullanılan veri setinin kaynağı ve boyutu gibi konularda mevcut en başarılı makine öğrenimi algoritmaları karşılaştırılmıştır. Çalışmada DVM modelinin anemi tespiti için en çok kullanılan makine öğrenimi algoritmalarından biri olduğu ve onu sırasıyla KEYK, KA, ESA ve YSA'ın takip ettiği kanıtlanmıştır.

Sarsam vd. [17] yapılan çalışmada, Twitter platformunda hastalık belirtileri ile hastaların duyguları arasındaki ilişkilere dayanan bir anemi tanı mekanizması önerilmiştir. Benzer tweet'leri gruplandırmak ve gizli hastalık konularını belirlemek için k-means ve Gizli Dirichlet Ayırımı algoritmaları kullanılmıştır. Hastalık duyguları ve belirtileri, Apriori algoritması kullanılarak eşleştirilmiştir. Önerilen yaklaşım bir dizi sınıflandırıcı kullanılarak değerlendirilmiştir. Sıralı Minimal Optimizasyon (Sequential Minimal Optimization) kullanılarak %98,96 daha yüksek bir tahmin doğruluğu elde edilmiştir. Sonuçlar, korku ve üzüntü duygularının anemik hastalar arasında baskın olduğunu ortaya koymuştur.

Saputra vd. [18] yapılan çalışmada kullanılan veriler, Endonezya'nın Yogyakarta şehrinde bulunan Gadjah Mada Üniversitesi Tıp, Halk Sağlığı ve Hemşirelik Fakültesi Klinik Patoloji ve Laboratuvar Tıbbı Bölümü Laboratuvarı'ndan elde edilmiştir. Anemiyi güvenilir bir şekilde tespit etmek ve teşhis etmek için Aşırı Öğrenme Makinesi modeli (Extreme Learning Machine) kullanılmıştır. Toplamda 127 eğitim ve 63 test verisi kullanılmış ve Aşırı Öğrenme Makinesinin yaklaşımının RO, KEYK ve DVM ile kıyasla %99,21 doğruluk elde ederek çok daha iyi performans gösterdiği keşfedilmiştir.

Saihood ve Sonuç [5] yapılan çalışmada, sosyal faktörler kullanılarak çocuklarda anemi tahmin etmek için sekiz farklı makine öğrenme tekniğinin performansı karşılaştırılmıştır ve en uygun yöntem bulunmaya çalışılmıştır. Makine öğrenme teknikleri, çocukluk anemisini tahmin etme ve ilişkili faktörleri belirleme konusunda umut verici sonuçlar elde etmiştir. Çok Katmanlı Algılayıcı, tüm özelliklerle %81,67 ile en yüksek doğruluk oranına sahip olmuştur. Öte yandan KA, özellik seçimi yöntemleri uygulandığında %82,50 ile en yüksek doğruluk oranına sahip olmuştur.

2.2. TİROİT HASTALIĞI

Yıldız [6] yapılan çalışmada, tiroit hastalığının teşhisinde yapay sinir ağları teknolojisinin kullanımını incelemiştir. Çalışmada, farklı kaynaklardan elde edilen iki ayrı veri grubu kullanılmıştır. Veri grupları, Weka programı ve Matlab programının Yapay Sinir Ağları Toolbox'ı kullanılarak farklı öğrenme yöntemleriyle eğitilmiş ve sonuçlar analiz edilmiştir. Matlab ve Weka'da KEYK algoritmasıyla yapılan uygulamaların düşük doğruluk oranlarına karşılık, çok katmanlı yapay sinir ağlarıyla yapılan uygulamaların %100'e yakın doğruluk oranları elde etmesi, çok katmanlı yapay sinir ağlarının karmaşık, doğrusal olmayan ve düzensiz ilişkilere sahip giriş ve çıkış değerleri arasında iyi bir öğrenme gerçekleştirdiğini göstermiştir.

Akgül vd. [8] yapılan çalışmada, LR, KEYK ve DVM algoritmaları kullanılarak hipotiroidi hastalığının tanısında farklı örnekleme teknikleri kullanılmıştır. Bu yöntemlerle elde edilen başarı oranları %92'nin üzerinde olmuştur. LR ve fazla örnekleme tekniğinin birlikte kullanıldığı durumda diğer sınıflandırıcılardan daha iyi sonuçlar elde edilmiştir. LR ile elde edilen en iyi sonuçlar şunlardır: doğruluk oranı %97,8, F-Skor değeri %82,26 ve Matthews korelasyon katsayısı %81.8.

Alsaadawi ve Şehirli [19,20] tarafından yapılan çalışmada, altı geleneksel model (KEYK, DVM, KA, Naive Bayes, LR, Çok Katmanlı Algılayıcılar) ve beş Topluluk modeli (RO, Xgboost (eXtreme gradient boosting), Soft Vote, İstifleme, Torbalama) kullanarak tiroit hastalığını tespit etmek için kapsamlı bir yaklaşım sunulmuştur. Önerilen yöntemler iki aşamada test edilmiştir. İlk adımda veri setinin tüm özellikleri, eksik verilerin eklenmesinden sonra kullanılmış, veri seti işlenmiş ve

dengelenmiştir. Geleneksel modellerin en yüksek doğruluk oranı KA ve Çok Katmanlı Algılayıcılar modellerinde sırasıyla %99,92 ve %97,30 olarak bulunmuştur. Topluluk modelleri olan XGboost ve Torbalama modelleri %100 doğruluk oranına ulaşmıştır. İkinci adımda ise, tahmin için en iyi ilişkili özellikleri belirlemek için Özyinelemeli Öznitelik Eliminasyonu (Recursive Feature Elimination) modeli kullanılmıştır. Bu model, geleneksel modellere uygulanmış ve KA ve Naive Bayes modellerinde sırasıyla %100 ve %98,06 doğruluk oranına ulaşılmıştır. Topluluk modelleri olarak XGboost ve Torbalama de %100, İstifleme modeli ise %99,53 doğruluk oranına ulaşmıştır.

Salman ve Sonuç [21] yapılan çalışmada, hipertiroidizm ve hipotiroidizm tiroit hastalığı, makine öğrenme algoritmalar kullanılarak sınıflandırılmıştır. Kullanılan algoritmalar iyi sonuçlar vermiş ve iki model şeklinde oluşturulmuştur. İlk modelde, 16 giriş ve bir çıkıştan oluşan tüm özellikler kullanılmış ve RO algoritmasının doğruluk sonucu diğer algoritmalara göre en yüksek olan %98,93 olarak bulunmuştur. İkinci modelde ise daha önceki bir çalışmaya dayanarak aşağıdaki özellikler atılmıştır: query_thyroxine, query_hypothyorid ve query_hyperthyroid. Burada, bazı algoritmaların doğruluğunun arttığı, bazılarının ise doğruluğunu koruduğu görülmüştür. Naive Bayes algoritmasının doğruluğu %90,67'e artmıştır. Çok Katmanlı Algılayıcılar algoritması ise en yüksek hassasiyeti %96,4 doğrulukla elde etmiştir.

Aversanoa vd. [22] yapılan çalışmada, Napoli hastanesinde tedavi gören toplamda 247 hastaya ait 2211 örnek içeren bir veri setinde 10 farklı makine öğrenimi sınıflandırıcısı test edilmiş ve Ekstra Ağaçlar Sınıflandırıcısı (ExtraTreesClassifier) modelinin en iyi performansı sergilediği, doğruluk, kesinlik, duyarlılık ve F-Skorunun sırasıyla %84, %85, %84 ve %84 olduğu görülmüştür.

Alyas vd. [23] yapılan çalışmada, KA, RO algoritması, KEYK ve YSA gibi çeşitli makine öğrenimi algoritmaları veri kümesindeki parametrelere dayalı olarak hastalığı daha iyi tahmin etmek için karşılaştırmalı bir analiz oluşturmuştur. Veri kümesinin karşılaştırılması için örnekleme yapılmış ve örnekleme yapılmamış veri kümeleri üzerinde sınıflandırma yapılmıştır. RO, sınıflandırmada en etkili yöntemken (%94,8),

KEYK en düşük performans sergilemiş, öte yandan, YSA ve Naive Bayes, KEYK'in ortalama seviyesinin üzerinde performans sergilemiştir.

Mir ve Mittal [24] yapılan çalışmada, 1464 Hintli hastadan elde edilen birincil veri setine dayanan üç yeni model önerilmiştir. Bu modellerde, literatür araştırması sırasında bulunan en iyi beş makine öğrenme algoritması (DVM, Naive Bayes, J48, Torbalama, Artırma) karşılaştırılmıştır. Deneyler üç bölüme ayrılmış: patolojik gözlemler, serolojik testler ve bu iki parametrenin kombinasyonu. İlk modelde, her iki parametre üzerinde de Torbalama ile %98,56 doğruluk elde edilmiştir. Hastanın patolojik gözlemlerine dayanan ikinci modelde, DVM ile %99,08 doğruluk elde edilmiştir. Üçüncü modelde ise, serolojik testler üzerinde J48 sınıflandırıcısı ile %92,07 doğruluk elde edilmiştir.

Chaganti vd. [25] yapılan çalışmada, ileri özellik seçimi, geriye doğru özellik elemesi, çift yönlü özellik elemesi ve ekstra ağaç sınıflandırıcıları kullanarak makine öğrenimi tabanlı özellik seçimi benimsenmiştir. Yapılan kapsamlı deneyler, Ekstra Ağaç Sınıflandırıcıları temelli seçilen özelliklerin, RO sınıflandırıcısı ile kullanıldığında en iyi sonuçları sağladığını göstermiştir (%0,99 doğruluk ve F1 skoru). Sonuçlar, sağlanan doğruluk ve hesaplama karmaşıklığı açısından makine öğrenimi modellerinin tiroit hastalığı tespiti için daha iyi bir seçenek olduğunu göstermiştir.

Raghurama vd. [26] yapılan çalışmada, makine öğrenme teknikleri olan DVM, Çoklu Doğrusal Regresyon ve KA kullanılarak karşılaştırmalı tiroit hastalığı teşhisi yapılmıştır. Bu amaçla, UCI makine öğrenimi veritabanından elde edilen tiroit hastalığı veri seti kullanılmıştır. Karar ağacına dayalı sınıflandırma modeli en iyi doğruluğu (%97,35) elde etmişken, SVM ise en zayıf sınıflandırmayı gerçekleştirmiştir.

Shivastuti vd. [27] yapılan çalışmada, tiroit bozukluğu teşhisi için destek vektör makinesi (SVM) ve rastgele orman gibi iki ayrı makine öğrenimi (ML) tekniği değerlendirilmiştir. Deney, UCI (University of California Irvine) makine öğrenimi deposundan elde edilen Tiroit Veri Seti üzerinde gerçekleştirilmiştir. Bu iki makine

öğrenimi tekniği, doğruluk, hassasiyet, hatırlama ve F-skor gibi dört performans metriği kullanılarak karşılaştırılmıştır. Deneysel bulgulara göre, her iki yöntem de tiroit bozukluklarını tahmin etmek için kullanılabilir fakat DVM, tiroit bozukluğu teşhisi için RO'dan daha iyi performans göstermiştir. İncelenen her iki sınıflandırıcı arasında, DVM, doğruluk (%93), kesinlik (%89) ve duyarlılık (%93) değerlerine ulaşarak en iyi sonuçları göstermiştir. RO ise %92 doğruluk, %85 kesinlik, %92 duyarlılık ve 0,88 F1-skoru ile daha düşük sonuca sahip olmuştur.

BÖLÜM 3

MATERYAL VE YÖNTEM

3.1. VERİ SETİ

3.1.1. Anemi Hastalığı İçin Veri Seti

Bu çalışmada, Anemi hastalığı ile ilgili kullanılan veri seti 600 örneği içermektedir [4,5]. Bu örneklerden 429'u anemi hastası iken, 171'i anemi hastası değildir. Verilerin %80'i eğitim seti olarak, %20'i ise test seti olarak kullanılmıştır. İlgili veri setinde her bir örneğe ait 31 özellik bulunmaktadır. Veri setinde kullanılan tüm özellikler Çizelge 3.1'de açıklanmıştır.

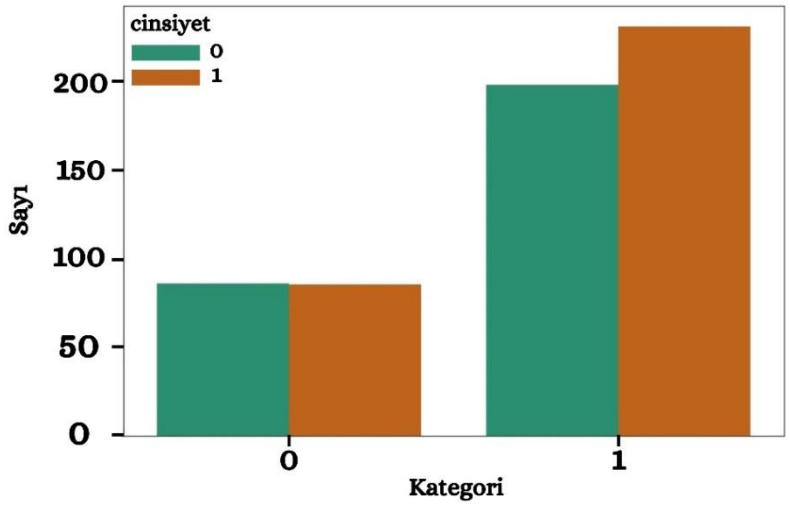
Çizelge 3. 1. Anemi hastalığı veri setinin özelliklerinin açıklaması.

No	Özellik adı	Türü	Açıklama
1	Id	Sayısal	Birincil anahtar 600 satırdan oluşur
2	Çocuğun yaşı	Sayısal	6 aydan 6 yaşa kadar
3	Cinsiyet	Sayısal	0 = Kadın, 1 = Erkek
4	Annenin yaşı	Sayısal	0 = Yaş < 30, 1 = Yaş ≥ 30
5	Annenin eğitim düzeyi	Sayısal	0 = Okur yazarlık yok, 1 = İlkokul, 2 = Ortaokul, 3 = Üniversite ve üzeri
6	Annenin mesleki düzeyi	Sayısal	0 = İşsiz, 1 = Çalışan
7	Babanın eğitim düzeyi	Sayısal	0 = Okur yazarlık yok, 1 = İlkokul, 2= Ortaokul, 3 = Üniversite ve üzeri
8	Babanın mesleki düzeyi	Sayısal	0 = İşsiz, 1 = Çalışan
9	Konut	Sayısal	0 = kırsal, 1 = kentsel
10	Sosyo-ekonomik durum	Sayısal	0 = Yoksul, 1 = Orta düzey, 2 = İyi
11	Hemoglobin seviyesi	Sayısal	
12	Yaşına göre kısa boy	Sayısal	0 = Anormal, 1 = Normal
13	Son 15 gün içinde ateş	Sayısal	0 = Hayır, 1 = Evet
14	Anemiye ilişkin önceki tıbbi geçmiş.	Sayısal	0 = Hayır, 1 = Evet
15	Son 15 gün içinde ishal	Sayısal	0 = Hayır, 1 = Evet

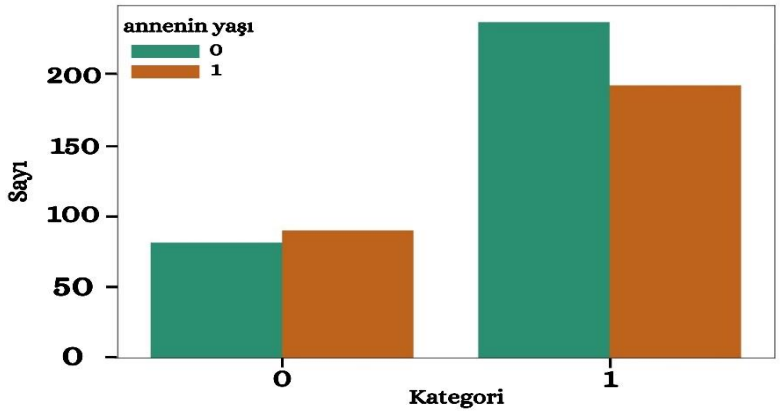
Çizelge 3.1. Anemi hastalığı veri setinin özelliklerinin açıklaması(devam ediyor)

16	Emzirme türü	Sayısal	0 = Anormal, 1 = Normal, 3 = Maksimum
17	Süt tozu tüketim	Sayısal	0 = Hayır, 1 = Evet
18	Şekerli içecek tüketim	Sayısal	0 = Hayır, 1 = Evet
19	Yoğurt tüketim	Sayısal	0 = Hayır, 1 = Evet
20	Katı/yarı katı gıda tüketim	Sayısal	0 = Hayır, 1 = Evet
21	Emzirme süresi	Sayısal	0 = Anormal, 1 = Normal
22	Et tüketimi	Sayısal	0 = Hayır, 1 = Evet
23	Koyu yeşil yapraklı sebzelerin tüketimi.	Sayısal	0 = Hayır, 1 = Evet
24	Demir kaynağı gıdaların tüketimi.	Sayısal	0 = Hayır, 1 = Evet
25	Karaciğer tüketimi.	Sayısal	0 = Hayır, 1 = Evet
26	Tamamlayıcı beslenmenin optimal zamanını bilmek.	Sayısal	0 = Bilinmiyor, 1 = Biliniyor, 2 = Emin Değilim
27	İlk tamamlayıcı gıdayı bilmek.	Sayısal	0 = Bilinmiyor, 1 = Biliniyor, 2 = Emin Değilim
28	Takviye demirin optimal gıdasını bilmek.	Sayısal	0 = Bilinmiyor, 1 = Biliniyor, 2 = Emin Değilim
29	Anemi ile ilişkili besinleri bilmek.	Sayısal	0 = Bilinmiyor, 1 = Biliniyor, 2 = Emin Değilim
30	Emzirmenin optimal zamanını bilmek.	Sayısal	0 = Bilinmiyor, 1 = Biliniyor, 2 = Emin Değilim
31	Anemi durumu	Sayısal	0 = Anemi Hastası Değil (171 örnek), 1 = Anemi hastası (429 örnek)

Şekil 3.1'de veri setinin cinsiyete göre dağılımını, Şekil 3.2'de ise annenin yaşına göre dağılımını gösteren grafik verilmiştir.



Şekil 3.1. Veri setinin cinsiyete göre dağılımı.



Şekil 3.2. Veri setinin yaşa göre dağılımı.

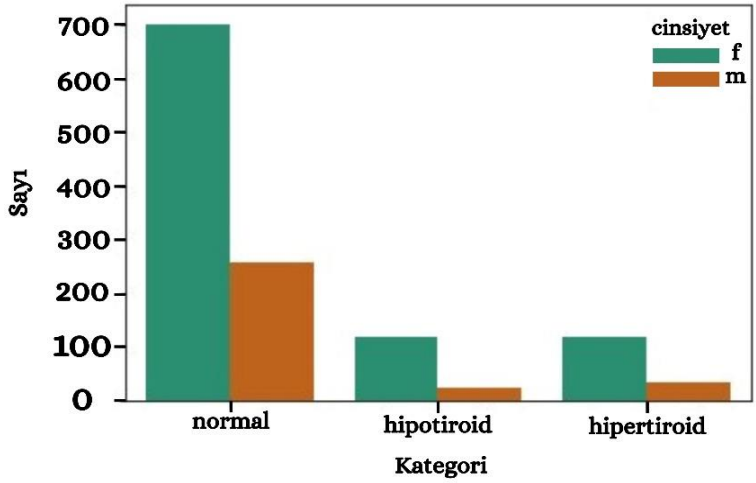
3.1.2. Tiroit Hastalığı İçin Veri Seti

Bu çalışmada, Tiroit hastalığı ile ilgili kullanılan veri seti 1 ila 90 yaş arası Iraklı erkek ve kadınlardan alınan 1250 örneği içermektedir [21]. İlgili veri setinde her bir örneğe ait 17 özellik bulunmaktadır. Veri setinde kullanılan tüm özellikler Çizelge 3.2'de açıklanmıştır.

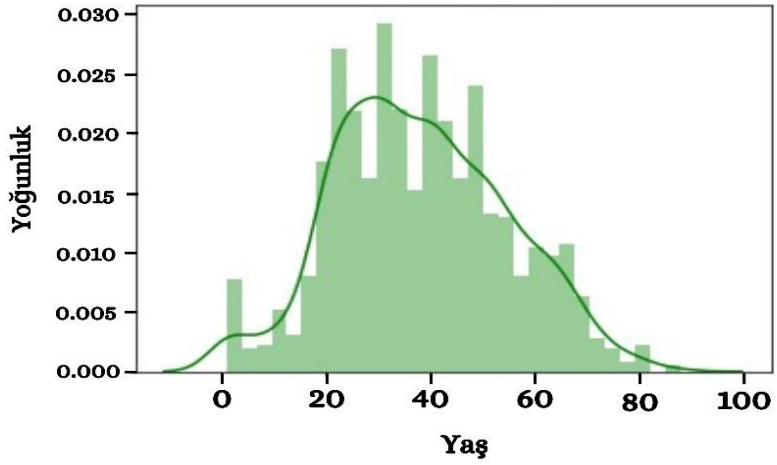
Çizelge 3. 2. Tiroit hastalığı veri setinin özelliklerinin açıklaması.

No	Özellik adı	Türü	Açıklama
1	id	Sayısal	1,2,3…
2	yaş	Sayısal	1,10,20, 50…
3	cinsiyet	1 , 0	1=E, 0=K
4	tiroit hormonu	1 , 0	1=Evet, 0=Hayır
5	antitiroid ilaçlar	1 , 0	1=Evet, 0=Hayır
6	hasta	1 , 0	1=Evet, 0=Hayır
7	hamile	1 , 0	1=Evet, 0=Hayır
8	tiroit cerrahisi	1 , 0	1=Evet, 0=Hayır
9	hipotiroidi	1 , 0	1=Evet, 0=Hayır
10	hipertiroidi.	1 , 0	1=Evet, 0=Hayır
11	TSH ölçüldü	1 , 0	1=Evet, 0=Hayır
12	TSH	Analiz oranı	Sayısal değer
13	T3 ölçüldü	1 , 0	1=Evet, 0=Hayır
14	T3	Analiz oranı	Sayısal değer
15	T4 ölçüldü	1 , 0	1=Evet, 0=Hayır
16	T4	Analiz oranı	Sayısal değer
17	kategori	0,1,2	0 'Normal', 1 'Hipotiroid', 2 'Hipertiroid'

Şekil 3.3'de veri setinin cinsiyete göre dağılımını , Şekil 3.4'de ise yaşa göre dağılımını gösteren grafik verilmiştir.



Şekil 3.3. Veri setinin cinsiyete göre dağılımı.



Şekil 3.4. Veri setinin yaşa göre dağılımı.

3.2. VERİ ÖN İŞLEME

Yapay zeka algoritmalarının doğru çalışmasını olumsuz etkileyen ve düşük performans seviyesine neden olan birçok yaygın sorun vardır. Bu sorunlar arasında kayıp gözlem, sınıf dengesizliği, ilgisiz değişken ve aykırı gözlem gibi sorunlar yer almaktadır [28].

Veri ön işleme, makine öğrenime teknikleri için önemli bir aşamadır. Bu algoritmaların temel hedefi, veriden bilgi elde etmektir ve bunu hesaplama yöntemlerini kullanarak yapmaktalar [29]. Verilerin kalitesini ve analiz sonuçlarının doğruluğunu artırarak veri odaklı uygulamaların daha etkili bir şekilde çalışmasına yardımcı olur ve veri analizi ve makine öğrenmesi gibi veri odaklı uygulamalar için önemli bir adımdır. Veri ön işleme, ham veri kümesini temizlemek, dönüştürmek ve hazırlamak amacıyla çeşitli teknikleri içeren bir dizi işlemi ifade eder. Bu, veri setinin analize uygun hale getirilmesini sağlar ve sonuçların daha doğru ve güvenilir olmasını sağlar.

Veri temizleme, veri dönüşümü, aykırı değer tespiti, özellik seçimi, normalizasyon ve ölçeklendirme veri ön işlemenin temel bileşenleridir.

3.2.1. Veri Temizleme

Veri temizleme, bir veri kümesindeki eksik, hatalı veya yanlış verilerin tespit edilmesi, aykırı değerleri ve gereksiz verilerin kaldırılması ve bu verilerle başa çıkılması sürecidir [30]. Bu adım, veri setinin analiz veya makine öğrenmesi modelleri gibi veri odaklı uygulamalar için uygun hale getirilmesini sağlar. Veri temizleme aşağıdaki temel işlemleri içerir:

Eksik verilerin işlenmesi: Veri setlerinde eksik veriler yaygın bir sorundur. Eksik veriler, boş hücreler veya belirsiz değerler şeklinde ortaya çıkabilir. Bu adımda, eksik verilerin tespit edilmesi ve bunlarla başa çıkılması gerekmektedir. Eksik veriler, silinerek, istatistiksel yöntemlerle (ortalama, medyan, en yakın komşu) tamamlanarak veya daha karmaşık yöntemlerle (makine öğrenmesi algoritmaları) doldurularak işlenebilir.

Aykırı değerlerin işlenmesi: Aykırı değerler, diğer verilere kıyasla önemli ölçüde farklı olan değerlerdir. Aykırı değerler, veri analizini ve model performansını olumsuz etkileyebilir. Bu nedenle, aykırı değerlerin tespit edilmesi ve işlenmesi önemlidir. Aykırı değerlerin silinmesi, değiştirilmesi veya daha gerçekçi değerlerle değiştirilmesi gibi işlemler yapılabilir.

Tutarsız veya çelişkili verilerin işlenmesi: Veri setleri bazen tutarsız veya çelişkili veriler içerebilir. Örneğin, bir kişinin yaşının negatif olması veya bir kişinin iki farklı cinsiyete sahip olması gibi durumlar söz konusu olabilir. Bu tür verilerin tespit edilmesi ve düzeltilmesi veya çıkarılması gerekmektedir.

Gürültülü verilerin işlenmesi: Gürültülü veriler, ölçüm hataları, veri toplama hataları veya veri kaynaklarındaki diğer sorunlar nedeniyle oluşabilir. Bu tür veriler, analiz sonuçlarını bozabilir. Veri temizleme sürecinde, gürültülü verilerin tespit edilmesi ve bunlarla başa çıkılması önemlidir. Gürültülü verilerin filtrelenmesi, düzeltilmesi veya çıkarılması gibi yöntemler kullanılabilir.

Veri temizleme işlemleri, veri setinin kalitesini artırır, analiz sonuçlarını doğrulukla ilişkilendirir ve daha güvenilir sonuçlar elde etmeyi sağlar. Bu nedenle, veri ön işleme sürecinin önemli bir bileşeni olarak kabul edilir.

3.2.2. Veri Dönüştürme

Veri dönüştürme, veri kümesindeki verilerin orijinal formatından farklı bir formata dönüştürülmesi işlemidir [3]. Bu işlem, veri setinin analiz, görselleştirme veya makine öğrenmesi gibi uygulamalara uygun hale getirilmesini sağlar. Veri dönüştürme aşağıdaki temel işlemleri içerebilir:

Kategorik verilerin dönüştürülmesi: Veri setinde kategorik (nominal veya ordinal) veriler bulunabilir. Bu tür veriler, sınıflandırmaya yönelik modellerde veya analizde sorunlara neden olabilir. Kategorik veriler, sayısal değerlere dönüştürülerek işlenmesi daha kolay hale getirilebilir. Bu dönüşüm, sınıf etiketlerinin sayısal değerlere atanması veya sınıf etiketlerinin "one-hot encoding" gibi kodlama teknikleriyle dönüştürülmesi gibi yöntemlerle gerçekleştirilebilir.

Zaman verilerinin dönüştürülmesi: Veri setinde zamanla ilgili veriler bulunabilir. Zaman verileri, analiz ve modelleme için uygun bir formatta olmalıdır. Zaman verileri, tarih ve saat bilgisi olarak temsil edilebilir ve daha spesifik bileşenlere

ayrılabilir (yıl, ay, gün, saat, dakika, saniye). Ayrıca, zaman aralıkları, zaman damgaları veya zaman bazlı özellikler gibi farklı şekillerde dönüştürülebilir.

Ölçeklendirme ve normalizasyon: Veri setindeki farklı özellikler farklı ölçeklerde olabilir. Bu durum, bazı modellerin veya algoritmaların etkili bir şekilde çalışmasını engelleyebilir. Ölçeklendirme ve normalizasyon işlemleri, veri setindeki özelliklerin benzer bir ölçeğe getirilmesini sağlar. Böylece, farklı özelliklerin doğru bir şekilde karşılaştırılabilmesi ve etkili bir şekilde işlenebilmesi sağlanır. Örneğin, Min-Max ölçeklendirme veya Z-skoru normalizasyonu gibi teknikler kullanılabilir.

Özellik türetme: Var olan veri özelliklerinden yeni özelliklerin türetilmesi, analiz ve modelleme performansını artırabilir. Özellik türetme, veri setindeki mevcut özelliklerin birleştirilmesi, dönüşümleri veya yeni istatistiksel özetlerin oluşturulması gibi işlemleri içerebilir. Bu, daha etkili ve özgün özellik setleri oluşturarak veri setinin ifade gücünü artırır.

Veri dönüştürme işlemleri, veri setinin analiz ve modelleme için uygun hale getirilmesini sağlar. Bu işlemler, veri setindeki farklı veri türlerinin ve özelliklerinin işlenebilir hale getirilmesini sağlar ve sonuçların daha doğru ve güvenilir olmasına yardımcı olur.

3.2.3. Aykırı Değer Tespiti

Aykırı değer tespiti, bir veri kümesinde diğer verilere göre belirgin bir şekilde farklı olan veya beklenmeyen değerleri tanımlama ve analiz etme sürecidir. Aykırı değerler, genellikle diğer değerlerden önemli ölçüde farklı olan ve modelin performansını etkileyebilecek nadir veri noktalarıdır. Bu nedenle, aykırı değerlerin tespit edilmesi ve uygun şekilde işlenmesi son derece önemlidir. Aykırı değerlerin işlenmesi, geliştirilen modelin performansını belirleyici ölçüde etkileyebilmektedir [30]. Aykırı değerler, veri analizinde veya istatistiksel modelleme süreçlerinde önemli sorunlara yol açabilir ve sonuçları yanıltabilir.

Aykırı değerlerin tespiti, veri kümesindeki anormallikleri belirlemek ve analiz sürecini iyileştirmek için önemlidir. Bununla birlikte, aykırı değerlerin ne yapılacağına karar vermek de önemlidir. Aykırı değerlerin neden kaynaklandığını anlamak ve analiz veya modelleme sürecine olan etkisini değerlendirmek önemlidir. Aykırı değerlerin çıkarılması, değiştirilmesi veya analiz sürecinde dikkate alınmaması gibi farklı yaklaşımlar kullanılabilir. Ancak, bu kararlar verilirken veri setinin bağlamı ve analiz hedefleri göz önünde bulundurulmalıdır.

3.2.4. Özellik Seçimi

Özellik seçimi (feature selection), bir veri kümesindeki özelliklerin (değişkenlerin) analiz veya modelleme için seçilmesi veya sıralanması sürecidir. Özellik seçimi, veri kümesindeki özelliklerin önem derecesini belirleyerek yüksek öneme sahip olan özellikleri seçmeyi ve gereksiz özellikleri etkisiz hale getirmeyi amaçlayan bir yöntemdir. Bu şekilde, veri boyutunu azaltmak ve analiz sürecini daha etkili hale getirmek mümkün oluyor [30]. Veri setinde bulunan tüm özelliklerin kullanılması, zaman ve kaynaklar açısından maliyetli olabilir ve aşırı uyum (overfitting) sorununa neden olabilir. Bu nedenle, özellik seçimi, daha az karmaşık, daha az gürültülü ve daha anlamlı özelliklere odaklanmayı amaçlar.

Özellik seçimi, gereksiz veya gürültülü özelliklerin çıkarılmasını sağlar, modelin performansını iyileştirir ve aşırı uyum sorununu azaltır. Ancak, doğru özelliklerin seçilmesi, analiz hedeflerine ve veri setinin bağlamına uygun olmalıdır. Ayrıca, seçilen özelliklerin anlamlılığının ve etkisinin doğrulanması ve doğrulama veri setleriyle test edilmesi önemlidir.

3.2.5. Normalizasyon ve Ölçeklendirme

Normalizasyon ve ölçeklendirme, veri işleme süreçlerinde kullanılan temel tekniklerdir ve veri özelliklerinin farklı ölçeklerde olmasından kaynaklanan sorunları gidermeyi amaçlar. Normalizasyon, her veri seti için zorunlu bir adım değildir. Ancak özellikler farklı aralıklara sahip olduğunda normalizasyon yapılması önerilir [31].

Normalizasyon, veri özelliklerini belirli bir aralığa veya dağılıma getirerek ölçeklendirme işlemidir. Normalizasyon, veri setinin tüm özelliklerinin aynı ölçeğe sahip olmasını sağlar ve farklı özelliklerin birbirlerine göre daha adil bir şekilde karşılaştırılmasını sağlar. En yaygın kullanılan normalizasyon yöntemleri arasında Min-Max normalizasyonu ve Z-skoru normalizasyonu bulunur.

Min-Max normalizasyonu: Bu yöntemde, veri özellikleri belirli bir aralığa dönüştürülür, genellikle [0, 1] veya [-1, 1]. Her bir veri noktası, minimum ve maksimum değerler kullanılarak aşağıdaki formülle dönüştürülür:

$$x' = (x - min) / (max - min) \quad (3.1)$$

Burada, x, orijinal değer; x', normalleştirilmiş değer; min, özellik için minimum değer; max, özellik için maksimum değerdir.

Z-skoru normalizasyonu: Bu yöntemde, veri özellikleri ortalama değerlerine ve standart sapmalarına göre dönüştürülür. Her bir veri noktası, aşağıdaki formülle dönüştürülür:

$$x' = (x - mean) / std \quad (3.2)$$

Burada, x, orijinal değer; x', normalleştirilmiş değer; mean, özellik için ortalama değer; std, özellik için standart sapmadır.

Ölçeklendirme: Ölçeklendirme, veri özelliklerini farklı ölçeklerde yeniden ölçeklendirme işlemidir. Ölçeklendirme, veri setinin farklı özelliklerinin benzer bir aralığa sahip olmasını sağlar ve analiz veya modelleme süreçlerinde doğru sonuçlar elde etmeyi kolaylaştırır. En yaygın kullanılan ölçeklendirme yöntemleri arasında standart ölçeklendirme ve normalizasyon bulunur. Standart ölçeklendirme, veri özellikleri ortalama değerlerine ve standart sapmalarına göre ölçeklendirilir.

Normalizasyon ve ölçeklendirme, veri setinin özelliklerini aynı ölçekte tutarak modelleme veya analiz süreçlerinde daha doğru sonuçlar elde etmeyi sağlar. Hangi yöntemin kullanılacağı, veri setinin yapısına, özelliklerinin dağılımına ve analiz hedeflerine bağlı olarak belirlenmelidir.

3.3. MAKİNE ÖĞRENMESİ

Makine öğrenmesi, bilgisayar sistemlerinin verilerden öğrenme yapmasını ve deneyimlerden bilgi çıkarmasını sağlayan bir yapay zeka alanıdır. Günümüzde makine öğrenmesi, sanal gerçeklik, nesne tanıma, yüz tanıma, ses tanıma, pazarlama, yer bilimi ve birçok farklı alanlarda kullanılmaktadır [32]. Bu çalışmadaysa, sağlık alanındaki hastalık tahmini için kullanıldı. Makine öğrenmesi, bir modelin veriye dayalı olarak öğrenmesini ve gelecekteki verileri analiz etmek, tahmin yapmak, desenleri tanımak veya kararlar vermek için kullanmasını amaçlar. Hastalık tahmininde ve tedavisinde, makine öğrenmesi yöntemleri kullanılabilir.

Gözetimli Öğrenme, Gözetimsiz Öğrenme ve Takviyeli Öğrenme gibi makine öğrenmesi türleri mevcuttur. Bu yöntemlerin dışında, denetimli öğrenme de, makine öğrenmesi alanında yaygın olarak kullanılan bir öğrenme yöntemlerindendir [3]. Denetimli öğrenme, bir modelin veriye dayalı olarak örüntüler öğrenmesini sağlar. Veri setinde hem girdi (input) hem de hedef (output) değerleri bulunur ve model, girdi verilerinden hedef değerleri tahmin etmeyi öğrenir.

Denetimli öğrenme genellikle iki farklı probleme uygulanır: sınıflandırma ve regresyon.

Sınıflandırma: Sınıflandırma, bir girdi verisini belirli bir sınıfa atama problemini çözmeyi hedefler. Örnek olarak, bir e-posta mesajının spam veya spam olmayan olarak sınıflandırılması gibi bir problem verilebilir. Sınıflandırma problemlerinde, hedef değerler genellikle kategorik (örneğin, etiketler veya sınıflar) olarak temsil edilir.

Regresyon: Regresyon, bir girdi verisine karşılık gelen bir hedef değeri tahmin etme problemini çözmeyi hedefler. Örnek olarak, bir evin özelliklerine dayanarak evin fiyatını tahmin etme gibi bir problem verilebilir. Regresyon problemlerinde, hedef değerler genellikle sürekli sayısal değerlerdir.

Denetimli öğrenme, veri setindeki girdi ve hedef değerlerini kullanarak bir modeli eğitir ve bu modeli yeni verilerle tahmin yapmak için kullanır. Eğitim süreci, veri setindeki örnekler üzerinde modelin performansını optimize etmek için gerçekleştirilir. Bu genellikle bir hata fonksiyonunun (kayıp fonksiyonu) minimize edilmesiyle yapılır.

Denetimli öğrenmenin avantajları arasında genel olarak iyi performans, tahminlerin yorumlanabilirliği, çeşitli algoritma ve model seçenekleri ve geniş bir uygulama alanı bulunur. Bununla birlikte, denetimli öğrenme için yeterli ve temsil edici bir eğitim veri setinin bulunması önemlidir. Ayrıca, aşırı uyum (overfitting) riski ve sınıf dengesizliği gibi zorluklarla da karşılaşılabilir.

Hastalık tahmininde kullanılabilecek birçok farklı makine öğrenmesi yöntemi bulunmaktadır. Bu yöntemlerin her biri farklı özelliklere sahiptir ve hangi yöntemin kullanılacağına karar vermek için veri setinin boyutu, veri özellikleri ve hastalığın türü gibi faktörler dikkate alınmalıdır. Bazı yaygın kullanılan algoritmalar şunlardır: RO, DVM, KA, LR, KEYK.

Makine öğrenmesi, verilerdeki desenleri ve ilişkileri yakalamak, tahmin yapmak, sınıflandırma yapmak ve kararlar vermek için kullanılan güçlü bir araçtır. Makine öğrenmesi modelleri, sağlam bir eğitim süreci ve doğru parametre ayarları ile yüksek performans sağlayabilir.

Bu çalışmada Anemi ve Tiroit hastalığını tahmin etmek için KA, DVM, RO, LR, KEYK, gibi çeşitli makine öğrenme teknikleri kullanılmıştır.

3.3.1. Lojistik Regresyon (LR)

LR, bir sınıflandırma yöntemi olup, bir girişin belirli bir sınıfa ait olma olasılığını tahmin etmek için kullanılan istatistiksel bir modeldir. Bu yöntem, tahmini etiket değerlerini, çıkış değerlerinin sınırlı olmasını sağlamak için lojistik bir fonksiyondan geçirir [3].

LR, bağımlı değişkenin kategorik olduğu durumlarda kullanılır. Örneğin, bir hastanın bir hastalığa sahip olup olmadığını tahmin etmek için kullanılabilir. Bu yöntem, lojistik fonksiyon (sigmoid fonksiyon) kullanarak giriş verileri ile sınıflandırma yapar. Lojistik fonksiyon, çıktıyı [0, 1] aralığına sıkıştırarak olasılık değerini temsil eder. Modelin eğitim süreci, maksimum olabilirlik yöntemi kullanılarak gerçekleştirilir ve verilere uyum sağlamak için parametreleri (ağırlıklar) ayarlar. Algoritma, sınıflandırma sırasında, bağımsız ve bağımlı değişkenler arasındaki ilişkiyi analiz eder [33]. Yöntemin işleyiş prensibi:

Veri Hazırlığı: Eğitim veri seti hazırlanır, her veri noktası bir vektör olarak temsil edilir ve sınıflarla birlikte verilir. Veri seti, aynı boyutta ve sayısal değerlere sahip özelliklerden oluşur.

Lojistik Fonksiyonu: LR, sınıflandırma için lojistik fonksiyonunu kullanır. Lojistik fonksiyonu, girdi değerlerini 0 ile 1 arasında bir olasılık değerine dönüştürür. Bu olasılık değeri, veri noktasının sınıfa ait olma olasılığını temsil eder.

Model Parametreleri: LR, sınıflandırma modelini belirlemek için bir dizi parametre kullanır. Bu parametreler, veri noktalarının özellikleriyle ağırlıklandırılır ve sınıf olasılığını hesaplamak için kullanılır.

Model Eğitimi: Eğitim veri seti kullanılarak LR modeli eğitilir. Eğitim sürecinde, model parametreleri iteratif olarak güncellenir ve en iyi uyumu sağlamak için optimize edilir. Genellikle en küçük kareler yöntemi veya maksimum olabilirlik yöntemi kullanılır.

Tahmin ve Sınıflandırma: Oluşturulan LR modeli, yeni veri noktalarının sınıflandırılması için kullanılır. Yeni veri noktası, özellik değerleri kullanılarak modeldeki lojistik fonksiyonuna uygulanır. Sonuç, 0 ile 1 arasında bir olasılık değeri olarak elde edilir. Elde edilen olasılık değeri, belirlenen bir eşik değeriyle karşılaştırılarak sınıflandırma kararı verilir. Yöntemin avantajları:

Basit ve anlaşılır: LR modeli, basit matematiksel formülasyonlara dayanır ve yorumlanması kolaydır.

İyi performans: LR, doğru eğitildiğinde iyi performans gösterir ve iyi sınıflandırma sonuçları sağlar.

Değişken önemi: LR, değişkenlerin etkisini değerlendirmek için katsayıları sağlar. Bu sayede, hangi değişkenlerin sınıflandırmada daha önemli olduğunu belirlemek mümkün olabilir. Yöntemin dezavantajları:

Lineer ayrılabilirlik varsayımı: LR, verilerin lineer olarak ayrılabilir olduğunu varsayar. Eğer veri seti doğrusal olarak ayrılamazsa, LR düşük performans gösterebilir.

Aşırı uyum: LR modeli, aşırı uyuma eğilimli olabilir. Özellikle, aşırı karmaşık modellerde veya aşırı fazla değişken içeren durumlarda aşırı uyum riski vardır.

Aykırı değerlere hassasiyet: LR, aykırı değerlere hassas olabilir. Aykırı değerler, modelin parametrelerini etkileyerek tahminleri yanıltabilir.

LR, sınıflandırma problemlerinde yaygın olarak kullanılan bir yöntemdir. Ancak, veri setinin özelliklerine ve problem bağlamına bağlı olarak avantajları ve dezavantajları dikkate alınmalıdır. Ayrıca, LR doğru uygulanması için veri ön işleme adımları, özellik seçimi ve hiperparametre ayarlamaları gibi faktörler de göz önünde bulundurulmalıdır.

3.3.2. Rastgele Orman (RO)

RO, makine öğrenmesi alanında yaygın olarak kullanılan bir makine öğrenme yöntemidir. Tek bir ağaç kullanıldığında oluşan ezberleme sorununu çözmek için geliştirilmiş bir yöntemdir [34]. RO, birden fazla karar ağacının bir araya gelerek oluşturduğu bir modeldir. Bu model, hem sınıflandırma hem de regresyon problemleri için kullanılabilir. Model oluşturmak için birçok karar ağacı kullanılır. Her bir ağaç, rastgele örneklerle ve rastgele seçilen özelliklerle eğitilir. RO, her bir ağacın tahminini birleştirerek sonuç oluşturur. Sınıflandırma problemlerinde çoğunluk oylaması, regresyon problemlerinde ise tahminlerin ortalaması kullanılır. RO yöntemi, yüksek boyutlu verilerde iyi performans gösterir [3]. Yöntemin çalışma prensibi:

Veri Hazırlığı: Eğitim veri seti hazırlanır, her veri noktası bir vektör olarak temsil edilir ve sınıflar veya hedef değerlerle birlikte verilir. Veri seti, aynı boyutta ve sayısal değerlere sahip özelliklerden oluşur.

Ağaç Oluşturma: RO, birden fazla karar ağacının birleşiminden oluşur. Her ağaç, rastgele seçilen alt veri kümesi ve rastgele seçilen özellikler üzerinde eğitilir. Alt veri kümesi, orijinal veri setinden tekrarlı örneklemelerle oluşturulur.

Karar Ağacı Oluşturma: Her ağaç için, bir karar ağacı algoritması kullanılır. Karar ağacı, veriyi bölerek en iyi ayrıştırmayı sağlayan karar düğümlerini ve yaprakları oluşturur. Her düğümde, özellikler arasında en iyi ayrıştırmayı sağlayan bir ayrıştırma kriteri (örneğin, gini impurity veya entropi) kullanılır.

Topluluk Oluşturma: Oluşturulan karar ağaçları, bir ensemble olarak birleştirilir. Sınıflandırma durumunda, toplulukdaki ağaçların oyları kullanılarak sınıf tahmini yapılır. Regresyon durumunda, toplulukdaki ağaçların tahminleri ortalaması veya medyanı alınarak tahmin yapılır.

Tahmin ve Sınıflandırma: Oluşturulan RO modeli, yeni veri noktalarının sınıflandırılması veya regresyon tahmini için kullanılır. Yeni veri noktası,

toplulukdaki tüm ağaçlara uygulanır ve sınıf tahmini veya tahmin değeri elde edilir. Sınıflandırma durumunda, toplulukdaki ağaçların oyları kullanılarak sınıf tahmini yapılır. Regresyon durumunda, toplulukdaki ağaçların tahminleri ortalaması veya medyanı alınarak tahmin yapılır. Yöntemin avantajları:

Yüksek performans: RO, genellikle yüksek doğruluk ve genelleme performansı sağlar. Birden fazla ağacın birleşimi, aşırı uyuma karşı daha dirençli bir model oluşturur.

Değişken önemi: RO, her bir özelliğin sınıflandırmadaki önemini değerlendirmek için bir özellik önemi ölçütü sağlar. Bu sayede, hangi özelliklerin sınıflandırmada daha etkili olduğunu belirlemek mümkündür.

Eş zamanlı çoklu sınıflandırma: RO, birden fazla sınıfın olduğu çoklu sınıflandırma problemlerini başarıyla çözebilir. Yöntemin dezavantajları:

Hesaplama maliyeti: RO modeli, birden fazla karar ağacının bir araya gelmesiyle oluştuğu için hesaplama maliyeti yüksektir, özellikle büyük veri setleri veya karmaşık modeller için.

Yorumlanabilirlik: RO modelleri, karmaşık yapısı nedeniyle bazen yorumlanması zor olabilir. Her bir ağacın nasıl tahmin yaptığını anlamak ve açıklamak daha zor olabilir.

Ağaç sayısı: RO modelinin performansı, ağaç sayısı ile ilişkilidir. Yeterli sayıda ağaç kullanılmazsa, modelin performansı düşebilir.

RO, genel olarak sınıflandırma ve regresyon problemlerinde etkili bir yöntemdir. Ancak, veri setinin özelliklerine ve problem bağlamına bağlı olarak avantajları ve dezavantajları dikkate alınmalıdır. Ayrıca, RO modelinin uygun parametre ayarlamaları ve doğru eğitim verisi ile iyi performans sağlanabilir.

3.3.3. K-En Yakın Komşu (KEYK)

KEYK, makine öğrenmesi alanında sınıflandırma ve regresyon problemlerini çözmek için kullanılan bir algoritmadır. KEYK, yeni bir veri noktasının sınıfını veya değerini, en yakınındaki eğitim veri noktalarının etrafındaki çoğunluk sınıfının veya ortalamasının etkisiyle tahmin eder. Bu algoritma, verilerin uzaklık metriğine dayalı olarak yakınlık ilişkisini kullanır. Genellikle Euclidean veya Manhattan uzaklık metrikleri kullanılır. Bu yöntem, her veri noktasını bir vektör olarak temsil eder ve yeni bir veri noktasının sınıfını veya değerini tahmin etmek için en yakınındaki k veri noktasını kullanır. KEYK yöntemi kolay anlaşılır olmasının yanı sıra, veri setinin hacmine göre maliyet ve işlem yükü artabiliyor [3]. Yöntemin çalışma prensibi:

Veri Hazırlığı: Eğitim veri seti hazırlanır, her veri noktası bir vektör olarak temsil edilir ve sınıflar veya hedef değerlerle birlikte verilir. Veri seti, aynı boyutta ve sayısal değerlere sahip özelliklerden oluşur.

Uzaklık Ölçümü: KEYK, veri noktaları arasındaki benzerlik veya uzaklık ölçümünü kullanır. Genellikle Euclidean veya Manhattan mesafesi kullanılır, ancak farklı uzaklık metrikleri de kullanılabilir.

K Değerinin Belirlenmesi: KEYK algoritması, yakın komşuların sayısını temsil eden bir K değeri gerektirir. K değeri, kullanıcının belirlemesi gereken bir parametredir ve en iyi sonuçları elde etmek için deneme yanılma yöntemiyle belirlenir.

Sınıflandırma veya Regresyon: Veri noktasını sınıflandırmak veya bir regresyon tahmini yapmak için KEYK, K en yakın komşuyu seçer. K en yakın komşu, veri noktasına en yakın olan K eğitim veri noktalarıdır. Sınıflandırma durumunda, en sık görülen sınıf etiketi veri noktasının sınıf tahminini belirler. Regresyon durumunda, K en yakın komşunun hedef değerlerinin ortalaması veri noktasının tahmin değerini belirler.

Tahmin ve Sınıflandırma: Oluşturulan KEYK modeli, yeni veri noktalarının sınıflandırılması veya regresyon tahmini için kullanılır. Yeni veri noktası, uzaklık

ölçümü kullanılarak K en yakın komşularının belirlenir. Sınıflandırma durumunda, en sık görülen sınıf etiketi yeni veri noktasının sınıf tahminini belirler. Regresyon durumunda, K en yakın komşunun hedef değerlerinin ortalaması yeni veri noktasının tahmin değerini belirler. Yöntemin avantajları:

Basitlik: KEYK algoritması, basit bir yapıya sahiptir ve kolayca anlaşılabilir [3].

Esneklik: KEYK algoritması, sınıflandırma ve regresyon problemlerine uygulanabilir ve çeşitli veri tipleriyle kullanılabilir.

İyi performans: KEYK, doğru şekilde uygulandığında iyi bir performans gösterebilir, özellikle veri setinin yapısı ve boyutu uygun olduğunda. Yöntemin dezavantajları:

Hesaplama maliyeti: KEYK algoritması, yeni bir örneğin tahminini yaparken tüm eğitim veri noktalarının uzaklıklarını hesaplamayı gerektirir. Bu, büyük veri setleri için hesaplama maliyetini artırabilir [8].

Boyut problemleri: KEYK, yüksek boyutlu veri setlerinde iyi performans göstermeyebilir. "Boyut laneti" olarak adlandırılan durumda, veri noktalarının uzaklık hesaplaması daha zor hale gelir.

Veri dengesizliği: Eğitim veri setindeki sınıfların dengesiz dağılımı, KEYK algoritmasının performansını etkileyebilir. Azınlık sınıflarının yanlış sınıflandırılma olasılığı daha yüksek olabilir.

KEYK, basit ve anlaşılır yapısıyla tercih edilen bir algoritmadır. Ancak, veri setinin özelliklerine ve problem bağlamına bağlı olarak avantajları ve dezavantajları dikkate alınmalıdır. Ayrıca, uygun k değeri seçimi ve veri önişleme teknikleri gibi faktörler, algoritmanın performansını etkileyebilir.

3.3.4. Karar Ağacı (KA)

KA, makine öğrenmesi alanında sınıflandırma ve regresyon problemlerini çözmek için kullanılan bir algoritmadır. KA, bir veri kümesini hedef değişkenine göre bölerek kararlar veren bir ağaç yapısı oluşturur, veri setindeki özellikleri bir ağaç yapısı şeklinde sıralar [3]. Bu model, veri setinin özelliklerine dayalı olarak bir dizi karar kuralı ve ağaç yapısı oluşturur ve sınıflandırma problemlerinde veri noktalarını sınıflara ayırırken, regresyon problemlerinde ise hedef değişkenin tahminini yapar. Bu yöntem, veri setinin yapısını ve özelliklerini anlamak için kullanıcı tarafından yorumlanabilir ve görselleştirilebilir bir model sağlar. Yöntemin çalışma prensibi:

Veri Hazırlığı: Eğitim veri seti hazırlanır, her veri noktası bir vektör olarak temsil edilir ve hedef değerler veya sınıflarla birlikte verilir. Eğitim verisi üzerinde kullanılacak özellikler ve hedef değişken belirlenir.

Karar Kriteri: KA, veri kümesini bölerek kararlar alır. Bu bölme işleminin temelinde bir karar kriteri bulunur. Genellikle Gini impurity veya bilgi kazancı gibi ölçütler kullanılır. Bu ölçütler, veri kümesinin ne kadar homojen veya heterojen olduğunu ölçer.

Kök Düğüm: İlk adımda, KA bir kök düğüm oluşturur. Kök düğüm tüm veri kümesini temsil eder. Kök düğümde bir karar kuralı oluşturulur. Bu kural, veri kümesini en iyi şekilde bölerek sınıflara veya tahmin değerlerine ayrılmasını sağlar.

Dallanma: Kök düğümde oluşturulan karar kuralına göre veri kümesi alt kümelerine ayrılır. Her alt küme, bir düğüm veya yaprak düğüm olarak adlandırılır ve bir karar kuralını veya karar noktasını temsil eder. Dallanma işlemi, özellik değerlerine ve karar kriterlerine göre tekrarlanır. Böylece ağaç yapısı oluşturulur.

Yaprak Düğümler: Dallanma işlemi, belirli bir durumu veya sınıfı temsil eden yaprak düğümlere ulaşılıncaya kadar devam eder. Yaprak düğümler, tahminlerin veya sınıfların belirlendiği son düğümlerdir.

Tahmin ve Sınıflandırma: Oluşturulan KA, yeni bir veri noktasının sınıfını veya tahmin değerini belirlemek için kullanılır. Yeni veri noktası, ağaç yapısında kök düğümden başlayarak uygun dallara yönlendirilir ve sonunda bir yaprak düğümünde tahmin yapılır. Yöntemin avantajları:

Anlaşılabilirlik: KA, basit ve anlaşılır bir yapısı olduğu için kolayca yorumlanabilir [3]. Karar kuraları ve ağaç yapısı, veri setinin yapısını anlamak için kullanıcıya bilgi sağlar.

Veri ön işleme gereksinimi: KA, veri ön işleme adımlarına (örneğin, veri normalizasyonu veya ölçeklendirme) ihtiyaç duymaz [3]. Ayrıca, eksik verilerle başa çıkma yeteneğine sahiptir.

Özellik önemi: KA, her özelliğin sınıflandırma veya tahmindeki önemini değerlendirebilir. Bu, özellik seçimi veya önem sıralaması için değerli bir bilgi sağlar. Yöntemin dezavantajları:

Aşırı uyum (overfitting): KA, eğitim verilerine çok fazla uyum sağlayabilir ve bu nedenle yeni verilere kötü tahminler yapabilir [3]. Bu, ağaç yapısının karmaşık veya derin olduğunda ortaya çıkabilir.

Duyarlılık: KA, veri kümesindeki küçük değişikliklere hassas olabilir. Bu, ağaç yapısının değişebileceği ve tahmin sonuçlarının farklılık gösterebileceği anlamına gelir.

Sınıf dengesizliği: KA, sınıflar arasındaki dengesizlik durumunda yanlılık oluşturabilir. Eğer veri setinde bir sınıf diğerlerinden çok daha fazla ise, ağaç yapısı bu sınıfa odaklanabilir ve diğer sınıfları yanlış sınıflandırabilir.

KA, basitliği ve anlaşılabilirliği nedeniyle yaygın olarak kullanılan bir algoritmadır. Ancak, aşırı uyum riski ve sınıf dengesizliği gibi dezavantajlarını göz önünde bulundurmak önemlidir. Ayrıca, uygun parametre ayarlamaları ve ağaç büyüklüğü kontrolü ile daha iyi performans elde edilebilir.

3.3.5. Destek Vektör Makineleri (DVM)

DVM, makine öğrenmesi alanında sınıflandırma ve regresyon problemlerini çözmek için kullanılan güçlü bir algoritmadır. DVM bir ayrım metodudur ve fazla veriye sahip sınıflandlrma problemlerinde sıkça kullanılmaktadır [35]. Bu algoritma, bir veri kümesini sınıflara ayıran bir hiper düzlem bulmak için öğrenme algoritması kullanır ve sınıfları bölmek için en iyi ayrıştırıcı hiper düzlemi bulmaya çalışır. Bu yöntem, aşırı uyuma karşı dirençlidir ve yüksek boyutlu verilerde de etkili bir şekilde çalışabilir [3]. Yöntemin çalışma prensibi:

Veri Hazırlığı: Eğitim veri seti hazırlanır, her veri noktası bir vektör olarak temsil edilir ve sınıflar veya hedef değerlerle birlikte verilir. Veri seti, aynı boyutta ve sayısal değerlere sahip özelliklerden oluşur.

Özellik Uzayı ve Çekirdek Fonksiyonu: Veri seti, yüksek boyutlu bir özellik uzayına projeksiyon yapılır. DVM, farklı çekirdek fonksiyonlarını kullanarak veriyi bu özellik uzayına dönüştürür. Çekirdek fonksiyonu, veri noktalarının benzerliklerini veya iç ürünlerini hesaplar.

Sınıflandırma veya Regresyon: DVM, veri noktalarını sınıflara ayırmak veya bir regresyon problemini çözmek için bir hiper düzlem bulmaya çalışır. Hiper düzlem, veri noktalarını sınıflar arasında en iyi şekilde ayıran bir karar sınırıdır. DVM, sınıflar arasındaki en geniş marjı (destek vektörler arasındaki uzaklığı) maksimize etmeyi hedefler.

Destek Vektörler: Destek vektörler, marjın kenarında veya marjın üzerinde yer alan veri noktalarını temsil eder. Bu veri noktaları, sınıflandırma veya regresyon kararını etkileyen kritik örneklerdir. Destek vektörler, hiper düzleme en yakın olan ve marjı belirleyen noktalardır.

Optimizasyon: DVM, matematiksel optimizasyon yöntemlerini kullanarak en iyi ayrıştırıcı hiper düzlemi bulmaya çalışır. Genellikle Lagrange çarpanları kullanılarak

çözülür ve ikili veya çok sınıflı sınıflandırma problemlerine uygulanır. Optimizasyon işlemi, marjın maksimizasyonunu ve sınıflar arasındaki hata miktarını minimize etmeyi amaçlar.

Tahmin ve Sınıflandırma: Oluşturulan DVM modeli, yeni veri noktalarının sınıflandırılması veya regresyon tahmini için kullanılır. Yeni veri noktası, özellik uzayında hiper düzlemle ilişkilendirilerek sınıf veya tahmin değeri belirlenir. Yöntemin avantajları:

Etkili sınıflandırma: DVM, doğru parametreler ve çekirdek fonksiyonları kullanıldığında yüksek doğruluk sağlar.

Veri boyutu ve özellik sayısı: DVM, yüksek boyutlu veri setlerinde iyi performans gösterir [3] ve veri boyutu arttıkça diğer algoritmalara kıyasla daha az etkilenir.

Destek vektörleri: DVM, sınıflandırmada destek vektörlerini kullanarak modele odaklanır. Bu sayede model daha genelleştirilebilir ve gereksiz veri noktalarından etkilenmez. Yöntemin dezavantajları:

Parametre ayarı: DVM'nin bazı parametreleri (C ve çekirdek fonksiyonu gibi) doğru şekilde ayarlanmalıdır. Bu parametrelerin yanlış ayarlanması, performansı etkileyebilir.

Hesaplama yükü: DVM, büyük veri setleri üzerinde çalışırken hesaplama yükü yüksek olabilir. Özellikle çok sayıda veri noktası ve özellik olduğunda eğitim süresi uzayabilir.

Sınıf dengesizliği: Eğer sınıflar arasında dengesizlik varsa (bir sınıf diğerinden çok daha fazla örneğe sahipse), DVM'nin performansı etkilenebilir.

DVM, güçlü bir sınıflandırma ve regresyon algoritmasıdır. Ancak, parametre ayarının önemi ve hesaplama yükünün yüksek olabileceği göz önünde

bulundurulmalıdır. Ayrıca, sınıf dengesizliği durumunda dikkatli bir şekilde kullanılmalıdır.

3.4. TOPLULUK ÖĞRENME YÖNTEMLERİ

Topluluk öğrenme, makine öğrenmesi alanında birden fazla modelin bir araya gelerek daha iyi bir performans elde etmek için kullanıldığı bir yöntemdir. Topluluk öğrenme, farklı modellerin farklı öğrenme stratejilerini birleştirerek daha güçlü ve genelleyici tahminler yapabilme yeteneğini sağlar [36]. Topluluk öğrenme yöntemleri, birden fazla öğrenme modelinin birleştirilmesiyle oluşturulan bir topluluk modelidir. Bu yöntemlerde, farklı öğrenme algoritmaları veya aynı algoritmanın farklı ayarları kullanılabilir. Topluluk öğrenme yöntemlerinde, her model aynı veri kümesini veya farklı alt kümesini kullanarak eğitilir ve sonuçlar birleştirilerek tahmin yapılır. Topluluk öğrenme yöntemleri genellikle daha iyi tahmin performansı, daha iyi genelleme yeteneği ve daha az aşırı uyum (overfitting) riski sağlar.

Topluluk öğrenme yöntemleri, farklı model kombinasyonları ve stratejileri kullanarak daha iyi tahmin performansı ve genelleme yeteneği sağlayabilir. Bu yöntemler, veri setinin çeşitliliğini ve karmaşıklığını artırarak daha güvenilir tahminler yapma yeteneği sunar. Ancak, topluluk öğrenme yöntemlerinin kullanılması, daha fazla hesaplama gücü gerektirebilir ve modelin karmaşıklığını artırabilir. Bu çalışmada, torbalama, artırma ve istifleme gibi topluluk öğrenme yöntemleri denenmiştir.

3.4.1. Torbalama

Torbalama yöntemi, makine öğrenmesinde sınıflandırma ve regresyon problemlerini çözmek için kullanılan topluluk öğrenme tekniklerinden biridir [3]. Bu yöntemde, birden fazla model aynı eğitim veri setine uygulanır ve her bir modelin tahminleri birleştirilerek son tahmin yapılır. Torbalama topluluk öğrenme yöntemi, Bootstrap aggregating kısaltması olarak ifade edilir ve bu yöntemde, bootstrap örnekleme yöntemi esas alınmaktadır [36]. Torbalama yöntemi aşağıdaki şekilde çalışır:

Veri Hazırlığı: Eğitim veri seti, rastgele örnekleme (bootstrap) yöntemiyle farklı alt kümelere ayrılır. Her alt küme, orijinal veri setinden, aynı boyutta ancak bazı örneklerin yinelenerek seçildiği bir şekilde oluşturulur.

Model Eğitimi: Her alt küme üzerinde ayrı bir model eğitilir. Her model, farklı bir öğrenme algoritması veya aynı algoritmanın farklı ayarları kullanabilir.

Tahmin Birleştirme: Eğitilen modeller, test veri seti veya yeni veri noktaları üzerinde tahmin yapar. Her modelin tahmini, genellikle bir oylama (voting) veya ortalama yöntemiyle birleştirilir. Sınıflandırma problemleri için en yaygın birleştirme yöntemi, çoğunluk oylamasıdır. Regresyon problemleri için ise tahminlerin ortalaması kullanılabilir. Torbalama yönteminin avantajları şunlardır:

Daha İyi Genelleme: Torbalama, birden fazla modelin birleştirilmesiyle daha iyi bir genelleme yeteneği sağlar. Farklı alt küme ve modellerin kullanılması, veri setindeki çeşitliliği artırır.

Daha İyi Stabilite: Tek bir modele kıyasla daha fazla modelin kullanılması, tahminlerin daha kararlı ve güvenilir olmasını sağlar. Her modelin tek başına yaptığı hataların etkileri birbirini dengeleyebilir.

Aşırı Uyum Riskinin Azalması: Torbalama, aşırı uyum (overfitting) riskini azaltır [3]. Her alt kümenin rastgele örneklemeyle oluşturulması ve farklı modellerin kullanılması, aşırı uyumun önlenmesine yardımcı olabilir. Torbalama yönteminin dezavantajları şunlardır:

Artan Hesaplama Maliyeti [3]: Birden fazla modelin eğitilmesi ve tahmin yapılması, hesaplama maliyetini artırabilir. Torbalama yöntemi, daha fazla işlem gücü gerektirir.

Yüksek Depolama Alanı İhtiyacı [3]: Her bir eğitilmiş modelin saklanması gerektiği için daha fazla depolama alanı gerektirir.

Yorumlanabilirlik Zorluğu: Torbalama yöntemi, tahminlerin birleştirilmesiyle elde edildiği için tek bir modelin yaptığı gibi açık bir şekilde yorumlanması zor olabilir.

Genel olarak, torbalama yöntemi, farklı modellerin birleştirilmesiyle daha iyi bir tahmin performansı ve genelleme yeteneği sağlar. Ancak, hesaplama maliyeti ve depolama alanı gibi zorlukları da göz önünde bulundurmak önemlidir.

3.4.2. Artırma

Artırma yöntemi, topluluk öğrenme tekniklerinden biridir. Bu yöntemde, zayıf öğreniciler olarak adlandırılan birçok zayıf model (örneğin, karar ağaçları) bir araya getirilerek daha güçlü bir model oluşturulur [3]. Artırma yöntemi aşağıdaki şekilde çalışır:

Veri Ağırlıklandırma: İlk olarak, veri setinin her bir örneği eşit ağırlığa sahiptir. Her örnek, modelin hatalı tahmin ettiği veya doğru tahmin ettiği verilere göre bir ağırlıkla temsil edilir.

Zayıf Model Eğitimi: İlk zayıf model, ağırlıklandırılmış veri seti üzerinde eğitilir. Model, veri setindeki hataları minimize etmek için çalışır.

Ağırlık Güncelleme: İlk modelin eğitimi tamamlandıktan sonra, hatalı tahmin edilen örneklerin ağırlıkları artırılır. Doğru tahmin edilen örneklerin ağırlıkları ise azaltılır.

Yeni Zayıf Model Eğitimi: Ağırlıklandırılmış veri seti üzerinde ikinci bir zayıf model eğitilir. Bu model, önceki modelin hatalı tahmin ettiği örnekler üzerinde odaklanır.

Süreç Tekrarı: Ağırlıklar güncellenir ve bir sonraki zayıf model eğitilir. Bu süreç, belirli bir sayıda zayıf modelin eğitilene kadar tekrarlanır.

Tahmin Birleştirme: Tüm zayıf modellerin tahminleri ağırlıklı bir şekilde birleştirilerek final tahmin yapılır. Artırma yönteminin avantajları şunlardır:

Yüksek Tahmin Performansı: Artırma, zayıf modellerin birleştirilmesiyle daha güçlü bir model oluşturarak yüksek tahmin performansı sağlar.

Daha Az Aşırı Uyum: Artırma yöntemi, aşırı uyum riskini azaltır [3]. Zayıf modeller, önceki modellerin hatalarına odaklandığı için hataların düzeltilmesi sağlanır.

Esneklik: Artırma yöntemi, farklı zayıf öğrenicilerin (örneğin, farklı algoritmalar veya farklı parametre ayarları) kullanılmasına olanak tanır. Artırma yönteminin dezavantajları şunlardır:

Hesaplama Gücü: Artırma yöntemi, birden fazla modelin eğitilmesini gerektirir, bu da hesaplama gücü gerektirir.

Daha Hassas Veriye Duyarlılık: Artırma yöntemi, veri setindeki ağırlıklı örneklerin doğru bir şekilde işlenmesine dayanır. Eğer veri setindeki ağırlıklar yanlış verilirse, performans etkilenebilir.

Veri Dengesizliği: Veri setinde dengesizlik varsa (örneğin, bir sınıf diğerinden çok daha fazla örneğe sahipse), artırma yöntemi performansını düşürebilir.

Genel olarak, artırma yöntemi, zayıf modellerin bir araya gelmesiyle daha güçlü bir model oluşturarak yüksek tahmin performansı sağlar. Ancak, hesaplama gücü ve veri dengesizliği gibi zorlukları göz önünde bulundurmak önemlidir.

3.4.3. İstifleme

İstifleme yöntemi, topluluk öğrenme tekniklerinden biridir. Bu yöntemde, farklı öğrenme algoritmaları veya farklı ayarları olan aynı algoritmanın birleşimi kullanılarak birden fazla model bir araya getirilir. Bu modeller, alt modeller olarak adlandırılır ve bir meta model tarafından birleştirilir. İstifleme yöntemi, sınıflayıcı tahminlerini meta sınıflayıcı için girdi kabul ederek, farklı türden sınıflayıcıların tahminleri ile yüksek doğruluk performansı sağlanan bir yöntemdir [2]. İstifleme topluluk öğrenmesi, özellikle büyük veri kümeleri için yararlıdır ve aşırı uyum

problemlerini önlemeye yardımcı olmaktadır [3]. İstifleme yöntemi aşağıdaki şekilde çalışır:

Veri Hazırlığı: Eğitim veri seti, k-folds çapraz doğrulama yöntemiyle k parçaya bölünür. Her bir parça, k-1 parça kullanılarak eğitim, k. parça ise doğrulama için kullanılır.

Alt Modellerin Eğitimi: Her bir alt model, farklı bir öğrenme algoritması veya aynı algoritmanın farklı ayarları kullanılarak eğitilir. Eğitim veri setinin k-1 parçası üzerinde eğitim yapılır.

Meta Modelin Eğitimi: Alt modellerin çıkışları (tahminleri), doğrulama seti üzerinde kullanılır. Bu çıkışlar, doğrulama seti üzerindeki hedef değişkenin tahminlenmesi için meta modelin eğitiminde kullanılır.

Tahmin Yapma: Test veri seti üzerindeki örnekler, alt modeller tarafından tahmin edilir. Alt modellerin tahminleri, meta model tarafından birleştirilerek final tahmin yapılır. İstifleme yönteminin avantajları şunlardır:

Yüksek Tahmin Performansı: İstifleme yöntemi, farklı modellerin birleştirilmesiyle daha güçlü bir model oluşturarak yüksek tahmin performansı sağlar.

Esneklik: İstifleme yöntemi, farklı öğrenme algoritmaları veya farklı parametre ayarları kullanarak çeşitli alt modellerin oluşturulmasına olanak tanır.

Daha İyi Genelleme: Çapraz doğrulama yöntemiyle eğitim ve doğrulama setleri üzerindeki performans ölçülerek, modelin genelleme yeteneği artırılabilir. İstifleme yönteminin dezavantajları şunlardır:

Hesaplama Gücü: Birden fazla modelin eğitilmesi ve birleştirilmesi, hesaplama gücü gerektirir.

Veriye Duyarlılık: İstifleme yöntemi, alt modellerin çıkışlarını kullanarak meta modeli eğittği için, alt modellerin tahminlerindeki hatalar veya veriye ilişkin özellikler meta model performansını etkileyebilir.

Yüksek Düzeyde Ayar Gerektirebilir: İstifleme yöntemi, farklı öğrenme algoritmalarını veya ayarlarını birleştirdiği için, uygun bir yapılandırma ve hiperparametre ayarı gerektirebilir.

Genel olarak, istifleme yöntemi, farklı alt modellerin birleştirilmesiyle daha güçlü bir model oluşturarak yüksek tahmin performansı sağlar. Ancak, hesaplama gücü ve veriye duyarlılık gibi zorlukları göz önünde bulundurmak önemlidir.

3.5. PERFORMANS ÖLÇÜMÜ

Karışıklık matrisi, sınıflandırma problemlerinde modelin gerçek ve tahmin edilen sınıflar arasındaki ilişkiyi gösteren bir tablodur. Karışıklık matrisi, modelin sınıflandırma performansını değerlendirmek için kullanılır.

Karışıklık matrisi genellikle 2x2 veya daha büyük boyutlarda olabilir, ancak temel olarak 2x2 boyutlu bir karışıklık matrisine odaklanalım. Bu matrisde, sınıflandırma sonuçları dört farklı kategoriye ayrılır: Gerçek Pozitifler (GP), Gerçek Negatifler (GN), Yanlış Pozitifler (YP), Yanlış Negatifler (YN).

Gerçek Pozitifler (GP): Hasta olan, hasta olarak tahmin edilen.
Gerçek Negatifler (GN): Hasta olmayan, hasta olmayan olarak tahmin edilen.
Yanlış Pozitifler (YP): Hasta olmayan, hasta olarak tahmin edilen.
Yanlış Negatifler (YN): Hasta olan, hasta olmayan olarak tahmin edilen.

Bu dört kategori, sınıflandırma modelinin performansını değerlendirmek için kullanılan çeşitli metriklerin hesaplanmasında kullanılır. Bazı yaygın olarak kullanılan metrikler şunlardır:

Doğruluk (Accuracy): GP ve GN değerlerinin toplam veri sayısına oranıdır. Genel tahmin doğruluğunu gösterir.

Doğruluk = (GN + GP)/(GN + GP + YN + YP) (3.4)

Kesinlik (Precision): GP değerinin, pozitif olarak tahmin edilen tüm örneklerin toplamına oranıdır. Yanlış pozitif tahminleri minimize etmeye odaklanır.

Kesinlik = GP/(GP + YP) (3.5)

Duyarlılık (Recall/Sensitivity): GP değerinin, gerçek pozitif sınıfın toplam sayısına oranıdır. Yanlış negatif tahminleri minimize etmeye odaklanır.

Duyarlılık = GP/(GP + YN) [37] (3.6)

F1 Skoru: Hassasiyet ve duyarlılığın harmonik ortalamasını temsil eder. Hem hassasiyeti hem de duyarlılığı dengelemeye çalışır.

F1-skoru = 2*(kesinlik*duyarlılık)/(kesinlik + duyarlılık) (3.7)

Karışıklık matrisi, sınıflandırma modelinin performansını daha ayrıntılı bir şekilde anlamak için kullanışlı bir araçtır. Modelin hangi sınıfları doğru tahmin ettiğini ve hangi sınıfları yanlış tahmin ettiğini gösterir.

BÖLÜM 4

DENEYSEL SONUÇLAR

4.1. GELİŞTİRİLEN ORTAM

Çalışma için Python programlama dili kullanılmıştır. Python, genel amaçlı bir yüksek seviye programlama dilidir ve birçok alanda kullanılmaktadır. Özellikle veri analizi, yapay zeka, bilimsel hesaplama, web geliştirme ve otomasyon gibi alanlarda tercih edilmektedir.

Python'un avantajları şunlardır: Kolay okunabilirlik, geniş kütüphane desteği, platform bağımsızlığı ve büyük bir toplululuğa sahip olması.

Python'un dezavantajları şunlardır: Performans, mobil uygulama geliştirme, bellek kullanımı ve geliştirme süresi.

Çalışmadaki model esas olarak Windows, Linux ve Mac işletim sistemlerinde çalışmaktadır fakat bu çalışma Windows işletim sisteminde kurulmuştur.

Geliştirilen ortam, donanım olarak Jupyter Lab, çalışma zamanı kullanılan platform için ise Jupyter Notebook kullanılmıştır.

Çizelge 4.1'de, çalısmada kullanılan kütüphaneler ve kullanım amaçları belirtilmiştir.

Çizelge 4.1. Çalısmada kullanılan kütüphaneler ve kullanım amaçları.

Kütüphaneler	**Kullanım amaçları**
Sklearn	Çeşitli makine öğrenimi problemlerini çözme
Statsmodels	İstatistiksel modellerin keşfi, tahmini ve analizi

Çizelge 4.1. Çalısmada kullanılan kütüphaneler ve kullanım amaçları.(devam ediyor)

Matplotlib	Veri görselleştirmesi
Seaborn	Veri görselleştirmesi
Pandas	Veri analizi ve veri işleme
NumPy	Bilimsel hesaplamalar ve sayısal işlemler
Pickle	Nesnenin bellek veya dosya gibi veri depolama alanlarına dönüştürülmesi ve depolanan verinin orijinal nesneye dönüştürülmesi

4.2. ANEMİ TEŞHİSİ İÇİN DENEYSEL SONUÇLAR

Test edilen modeller, test verilerinin %20'lik bir örneğinde 5 kat çapraz doğrulama kullanılarak değerlendirilmiştir. Modellerin parametrelerinin konfigürasyonları Çizelge 4.2'te verilmiştir.

Veri seti üzerinde farklı topluluk tekniklerinin performansları değerlendirilmiştir. Çizelge 4.3'deki sonuçlara göre topluluk öğrenme teknikleri bireysel sınıflandırıcılara göre daha az doğrulukla tahminde bulunmuştur. Ayrıca, topluluk öğrenme teknikleri arasında Artırma yönteminin en yüksek doğruluğa (%90,8) eriştiği görülmüştür.

Sonuçlara göre, LR 0,908 doğruluk, 0,952 kesinlik, 0,919 duyarlılık, 0,935 F1-skoru ve 0,964 AUC skoru, RO 0,958 doğruluk, 0,945 kesinlik, 1 duyarlılık, 0,972 F1-skoru ve 0,998 AUC skoru, KEYK 0,825 doğruluk, 0,842 kesinlik, 0,930 duyarlılık, 0,884 F1-skoru ve 0,848 AUC skoru, KA 1 doğruluk, 1 kesinlik, 1 duyarlılık 1 F1-skoru ve 1 AUC skoru, DVM 0,908 doğruluk, 0,921 kesinlik, 0,954 duyarlılık, 0,937 F1-skoru ve 0,962 AUC skoru, Torbalama 0,908 doğruluk, 0,903 kesinlik, 0,977 duyarlılık, 0,939 F1-skoru ve 0,971 AUC skoru, İstifleme 0,800 doğruluk, 0,787 kesinlik, 0,988 duyarlılık 0,876 F1-skoru ve 0,869 AUC skoru, son olarak Artırma ise, 1,000 doğruluk, 1,000 kesinlik, 1,000 duyarlılık 1,000 F1-skoru ve 1,000 AUC skoru elde etmiştir.

Çizelge 4.2. Çalısmada kullanılan modellerin parametre değerleri.

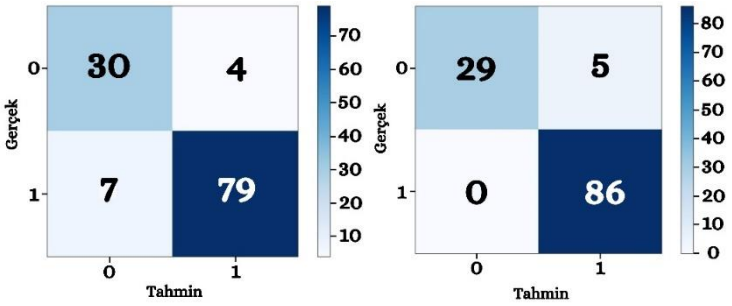
Model	**Parametre Adı**	**Parametre Değeri**
LR	C	6.576748562456391
	solver	saga
	tol	0.00017253203723844928
	fit_intercept	True
	class_weight	balanced
	max_iter	170
RO	min_samples_split	4
	criterion	entropy
	bootstrap	False
	class_weight	balanced
	n_estimators	99
KEYK	leaf_size	22
	p	2
	algorithm	kd_tree
	weights	distance
	n_neighbors	9
KA	min_samples_split	4
	criterion	gini
	min_samples_leaf	4
	class_weight	balanced
DVM	C	0.619165036139557
	kernel	linear
	degree	3
	gamma	scale
	tol	0.0051483760635355185
	class_weight	None
	probability	True

Sonuçlar incelendiğinde bireysel sınıflandırıcıların daha doğru bir tahmin yaptığı açıkça görülmektedir. Bunun başlıca nedenlerinden birisi, veri setinin eğitim kısmında öğrenmeyi gerçekleştirirken RO gibi algoritmaların önemli özellikleri tespit etmesi gösterilebilir.

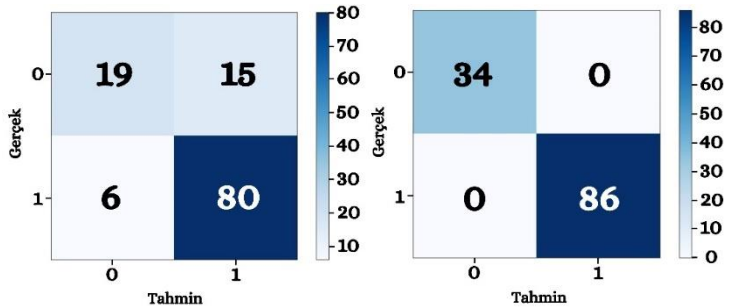
Çizelge 4.3. Anemi hastalığı için sınıflandırma yöntemlerinin sonuçlarının karşılaştırılması.

Sınıflandırma Modeli	Doğruluk	Kesinlik	Duyarlılık	F1-skoru	AUC skoru
Lojistik Regresyon	0,908	0,952	0,919	0,935	0,964
Rastgele Orman	0,958	0,945	1,000	0,972	0,998
K-En Yakın Komşu	0,825	0,842	0,930	0,884	0,848
Karar Ağacı	1,000	1,000	1,000	1,000	1,000
Destek Vektör Makineleri	0,908	0,921	0,954	0,937	0,962
Torbalama	0.908	0,903	0,977	0,939	0,971
İstifleme	0,800	0,787	0,988	0,876	0,869
Artırma	1,000	1,000	1,000	1,000	1,000

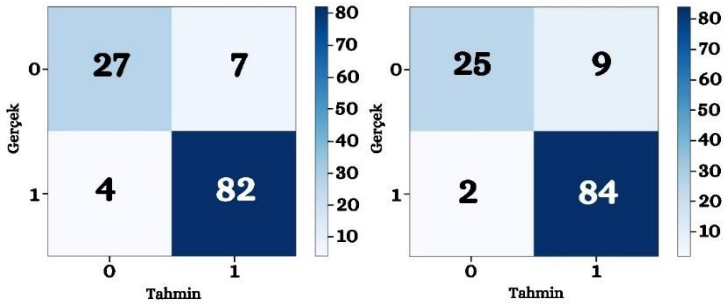
Şekil 4.1-4.4'de LR, RO, KEYK, KA, DVM, Torbalama, İstifleme ve Artırma yöntemlerinin karışıklık matrisi verilmiştir.



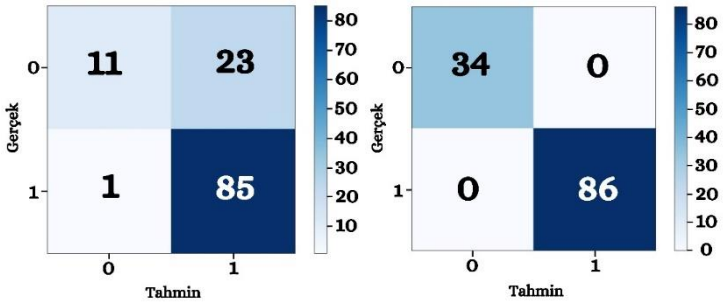
Şekil 4.1. LR (solda) ve RO (sağda) yöntemlerinin karışıklık matrisleri.



Şekil 4.2. KEYK (solda) ve KA (sağda) yöntemlerinin karışıklık matrisleri.

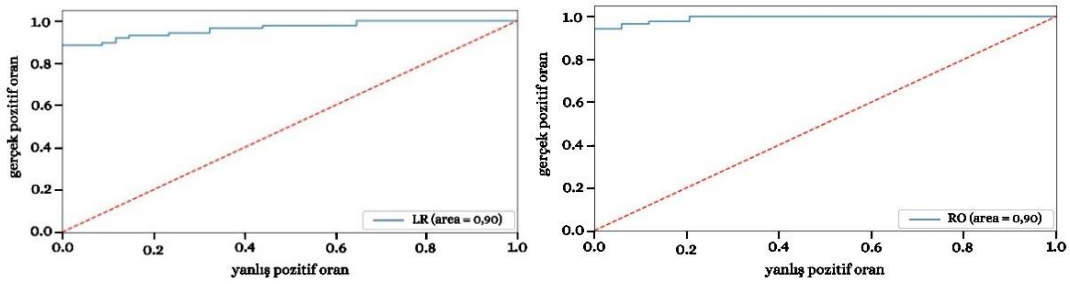


Şekil 4.3. DVM (solda) ve Torbalama (sağda) yöntemlerinin karışıklık matrisleri.

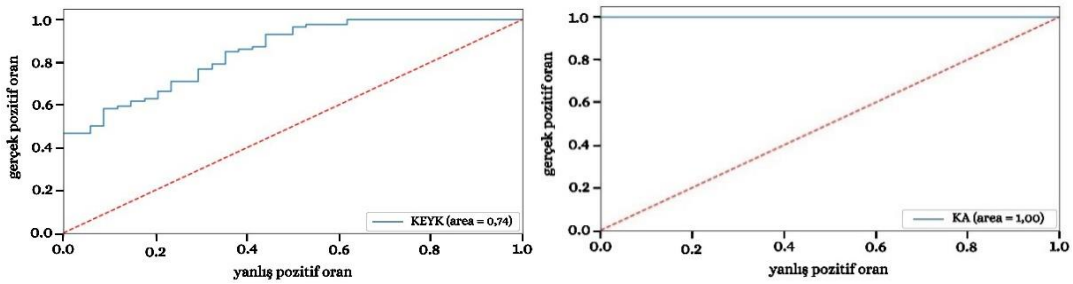


Şekil 4.4. İstifleme (solda) ve Artırma (sağda) yöntemlerinin karışıklık matrisleri.

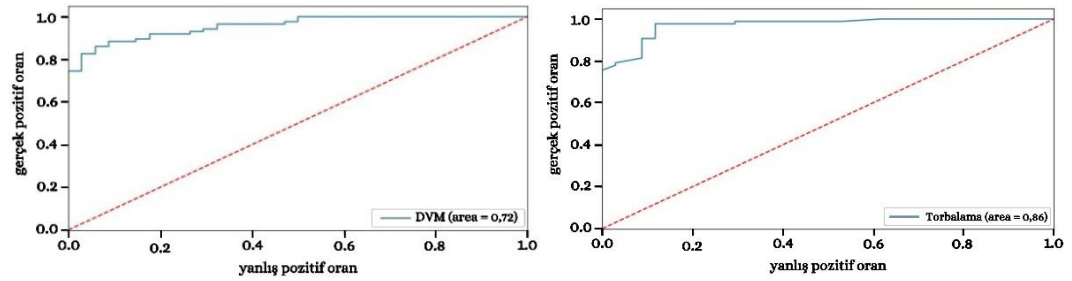
Şekil 4.5-4.8'de LR, RO, KEYK, KA, DVM, Torbalama, İstifleme ve Artırma yöntemlerinin ROC eğrileri verilmiştir.



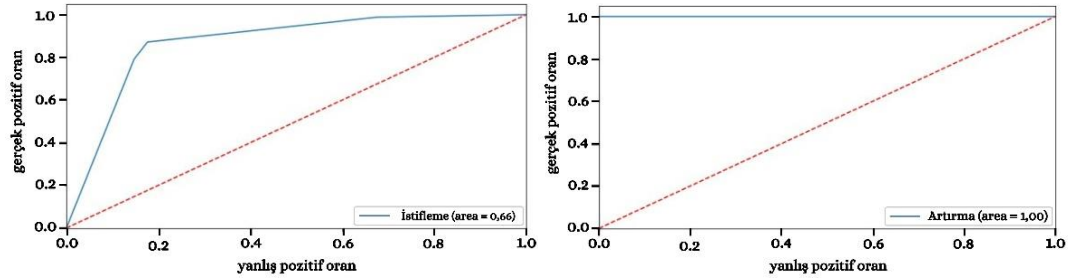
Şekil 4.5. LR (solda) ve RO (sağda) yöntemlerinin ROC eğrileri.



Şekil 4.6. KEYK (solda) ve KA (sağda) yöntemlerinin ROC eğrileri.



Şekil 4.7. DVM (solda) ve Torbalama (sağda) yöntemlerinin ROC eğrileri.



Şekil 4.8. İstifleme (solda) ve Artırma (sağda) yöntemlerinin ROC eğrileri.

4.3. TİROİT TEŞHİSİ İÇİN DENEYSEL SONUÇLAR

Test edilen modeller, test verilerinin %20'lik bir örneğinde 5 kat çapraz doğrulama kullanılarak değerlendirilmiştir. Modellerin parametrelerinin konfigürasyonları Çizelge 4.4'de verilmiştir.

Çizelge 4.4. Çalısmada kullanılan modellerin parametre değerleri.

Model	**Parametre Adı**	**Parametre Değeri**
LR	C	0.4928325931903053
	solver	liblinear
	tol	0.0031356724293372337
	fit_intercept	True
	max_iter	430
RO	min_samples_split	4
	criterion	entropy
	bootstrap	True
	class_weight	balanced
	n_estimators	93
KEYK	leaf_size	32
	p	1
	algorithm	ball_tree
	weights	uniform
	n_neighbors	10
KA	min_samples_split	2
	criterion	entropy
	min_samples_leaf	1
	class_weight	balanced
DVM	kernel	poly
	degree	6
	gamma	scale
	tol	0.07201901435698717
	class_weight	balanced
	probability	True

Veri seti üzerinde farklı topluluk tekniklerinin performansları değerlendirilmiştir. Çizelge 4.5'deki sonuçlara göre topluluk öğrenme teknikleri ve bireysel sınıflandırıcılar, birbirine yakın doğrulukta tahminde bulunmuştur. Fakat topluluk öğrenme teknikleri arasında Artırma yönteminin en yüksek doğruluğa (%89,1) eriştiği görülmüştür. Sonuçlar için, mikro F1 yöntemi kullanılmıştır.

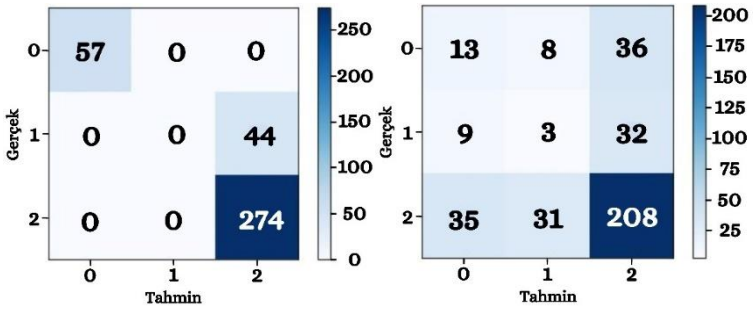
Sonuçlara göre, LR 0,883 doğruluk, 0,883 kesinlik, 0,883 duyarlılık, 0,883 F1-skoru ve 0,910 AUC skoru, RO 0,597 doğruluk, 0,597 kesinlik, 0,597 duyarlılık, 0,597 F1-skoru ve 0,540 AUC skoru, KEYK 0,717 doğruluk, 0,717 kesinlik, 0,717 duyarlılık, 0,717 F1-skoru ve 0,565 AUC skoru, KA 0,835 doğruluk, 0,835 kesinlik, 0,835 duyarlılık, 0,835 F1-skoru ve 0,781 AUC skoru, DVM 0,741 doğruluk, 0,741 kesinlik, 0,741 duyarlılık, 0,741 F1-skoru ve 0,767 AUC skoru, Torbalama 0,885 doğruluk, 0,885 kesinlik, 0,885 duyarlılık, 0,885 F1-skoru ve 0,866 AUC skoru, İstifleme 0,883 doğruluk, 0,883 kesinlik, 0,883 duyarlılık, 0,883 F1-skoru ve 0,870

AUC skoru, son olarak Artırma ise, 0,896 doğruluk, 0,896 kesinlik, 0,896 duyarlılık, 0,896 F1-skoru ve 0,919 AUC skoru elde etmiştir.

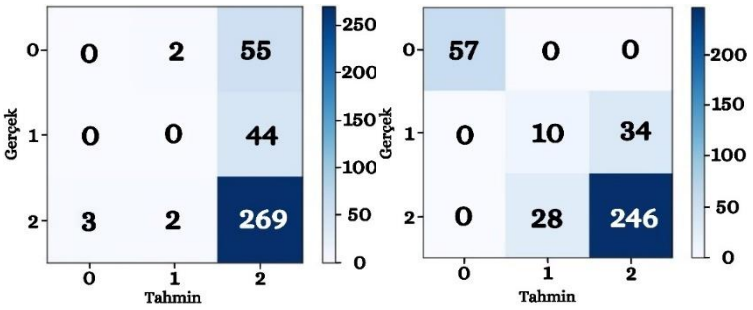
Çizelge 4.5.Tiroit hastalığı için sınıflandırma yöntemlerinin sonuçlarının karşılaştırılması.

Sınıflandırma Modeli	Doğruluk	Kesinlik	Duyarlılık	F1-skoru	AUC skoru
Lojistik Regresyon	0,883	0,883	0,883	0,883	0,910
Rastgele Orman	0,597	0,597	0,597	0,597	0,540
K-En Yakın Komşu	0,717	0,717	0,717	0,717	0,565
Karar Ağacı	0,835	0,835	0,835	0,835	0,781
Destek Vektör Makineleri	0,741	0,741	0,741	0,741	0,767
Torbalama	0,885	0,885	0,885	0,885	0,866
İstifleme	0,883	0,883	0,883	0,883	0,870
Artırma	0,896	0,896	0,896	0,896	0,919

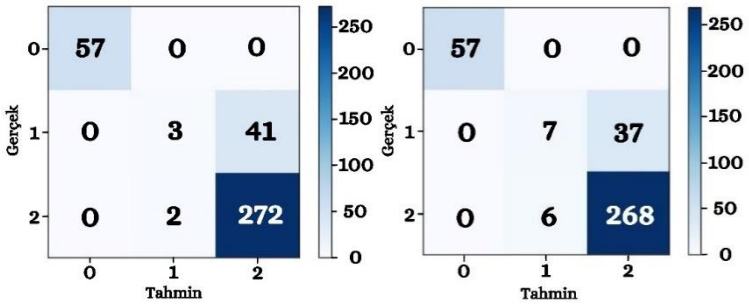
Şekil 4.9-4.12'de LR, RO, KEYK, KA, DVM, Torbalama, İstifleme ve Artırma yöntemlerinin karışıklık matrisi verilmiştir.



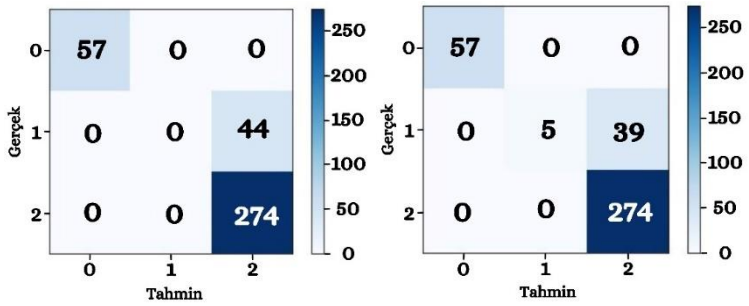
Şekil 4.9. LR (solda) ve RO (sağda) yöntemlerinin karışıklık matrisleri.



Şekil 4.10. KEYK (solda) ve KA (sağda) yöntemlerinin karışıklık matrisleri.



Şekil 4.11. DVM (solda) ve Torbalama (sağda) yöntemlerinin karışıklık matrisleri.



Şekil 4.12. İstifleme (solda) ve Artırma (sağda) yöntemlerinin karışıklık matrisleri.

4.4. TARTIŞMA

Bu çalışmada, Anemi ve Tiroit hastalıklarının tahmininde topluluk öğrenme yöntemleri ve bazı makine öğrenme yöntemleri denenmiştir. İlk olarak Anemi hastalığı, okul öncesi çocuklar arasında, özellikle gelişmekte olan ülkelerde en yaygın görülen hastalıklardan biridir. Anemi, genellikle yetersiz beslenmeyle ilişkilidir ve demografik ve sosyal faktörlerle de ilişkilendirilmektedir. Diğer taraftan Tiroit, boyunda bulunan bir bezdir ve vücutta metabolizmayı düzenlemekle görevlidir. Tiroit bezinin düzgün çalışması, vücuttaki enerji seviyelerini, kalp atış hızını, sindirimi ve diğer önemli işlevleri dengelemeye yardımcı olur. Ancak bazı durumlarda tiroit bezindeki fonksiyonlar bozulabilir ve tiroit hastalıklarına neden olabilir.

Bu iki hastalığı tahmin etmek için veri setleri üzerinde Karar Ağacı, Destek Vektör Makineleri, Rastgele Orman, Lojistik Regresyon, En Yakın Komşu gibi çeşitli

makine öğrenmesi algoritmaları test edilmiştir. Daha sonra bu sınıflandırıcılar, bagging, boosting, stacking gibi öğrenme teknikleri kullanılarak daha doğru ve güçlü bir tahmin modeli oluşturulması amaçlanmıştır. Veri seti üzerinde farklı topluluk tekniklerinin performansları değerlendirilmiştir. Deneysel sonuçlara göre, anemi hastalığı için topluluk öğrenme teknikleri bireysel sınıflandırıcılara göre daha az doğrulukla tahminde bulunmuştur. Ayrıca, topluluk öğrenme teknikleri arasında Artırma yöntemi en yüksek doğruluğa (%100) sahip olmuştur.

Diğer taraftan, tiroit hastalığı için yapılan deneysel sonuçlara göre topluluk öğrenme teknikleri ve bireysel sınıflandırıcılar, birbirine yakın doğrulukta tahminde bulunmuştur. Fakat tiroit hastalığı için de, topluluk öğrenme teknikleri arasında Artırma yöntemi en yüksek doğruluğa (%89,6) sahip olmuştur.

BÖLÜM 5

SONUÇLAR VE ÖNERİLER

Bu çalışmada, Anemi ve Tiroit hastalıklarının tahmininde topluluk öğrenme yöntemleri ve bazı makine öğrenme yöntemleri denendi. Anemi hastalığı için kullanılan Lojistik Regresyon, Rastgele Orman, En Yakın Komşu, Karar Ağacı, Destek Vektör Makineleri yöntemlerinin doğruluk oranları sırasıyla %90,8, %95,8, %82,5, %100 ve %90,8 olmuştur. Diğer taraftan, anemi hastalığı için kullanılan torbalama, istifleme ve artırma topluluk öğrenme yöntemlerinin doğruluk oranları sırasıyla %90,8, %80 ve %100 olmuştur.

Tiroit hastalığı için kullanılan Lojistik Regresyon, Rastgele Orman, En Yakın Komşu, Karar Ağacı, Destek Vektör Makineleri yöntemlerinin doğruluk oranları sırasıyla %88,3, %59,7, %71,7, %83,5 ve %74,1 olmuştur. Diğer taraftan, tiroit hastalığı için kullanılan torbalama, istifleme ve artırma topluluk öğrenme yöntemlerinin doğruluk oranları sırasıyla %88,5, %88,3 ve %89,6 olmuştur. Bu çalışma, Tiroit ve çocuklardaki anemi hastalığını tahmin etmek için topluluk öğrenme tekniklerinin farklı bir yaklaşım olabileceğini göstermiştir.

İleriki çalışmalarda veri ön işleme, özellik seçimi gibi yöntemlerin topluluk öğrenme modellerinin performansını artırmada etkili olabileceği düşünülmektedir. Daha iyi bir performans için daha büyük veri kümeleri üzerinde eğitim modellerine odaklanılabilir ve derin öğrenme tekniklerinin topluluk yöntemlerine entegrasyonu kullanılarak daha efektik modeller geliştirilebilir.

KAYNAKLAR

1. Coşar, M. ve Deniz, E., "Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi," Avrupa Bilim ve Teknoloji Dergisi, (28): 1112–1116 (2021).

2. Kıvrak, M., "Sınıflandırma Problemlerinde Topluluk Öğrenme Yöntemlerinin İncelenmesi ve Küçük Hücreli Dişi Akciğer Kanseri Verileri Üzerine Bir Uygulaması", Doktora Tezi, T.C. İnönü Üniversitesi Sağlık Bilimleri Enstitüsü, Malatya, 1-2 (2021).

3. Farzaliyev, E., Saihood, Q. ve Sonuç E., "Çocuklarda Anemi Hastalığının Teşhisinde Topluluk Öğrenme Yöntemlerinin Kullanılması," 1 st International Conference on Recent Academic Studies, 129-135 (2023).

4. Saihood, Q. and Sonuç, E., "The Efficiency of Classification Techniques in Predicting Anemia Among Children: A Comparative Study", International Conference on Emerging Technology Trends in Internet of Things and Computing, 167–181 (2022).

5. Saihood, Q. L. and Sonuç, E., "Exploration of Machine Learning Techniques in Predicting the Childhood Anemia", M. Sc. Thesis, T.C. Karabuk University Institute of Graduate Programs, Karabuk, 1-2 (2021).

6. Yıldız, A., "Makine Öğrenmesi Yöntemleri ile Tiroit Hastalığının teşhisi", Yüksek Lisans Tezi, T.C. Sakarya Uygulamalı Bilimler Üniversitesi Lisansüstü Eğitim Enstitüsü, 29-31 (2019)

7. Cindoğlu, Ç., Güler, M. S., Eren, M. A. ve Sabuncu, T., "Hipertiroidi Hastalarında Tedavi Öncesi ve Sonrası Trombosit/Lenfosit ve Nötrofil/Lenfosit Oranlarının Değerlendirilmesi" Harran Üniversitesi Tıp Fakültesi Dergisi, 104-107 (2020)

8. Akgül, G., Çelik, A. A., Ergül Aydın, Z. and Kamışlı Öztürk, Z., "Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı" Bilişim Teknolojileri Dergisi, 13 (3): 255–268 (2020)

9. Kant, R., Davis, A. and Verma, V., "Thyroid Nodules: Advances in Evaluation and Management", American Family Physician, 102 (5): 298-304 (2020).

10. Sarıbacak, B., "Sınıflandırma Yöntemleri İle Hematoloji Hastalıklarından Demir Eksikliği Anemisinin Erken Teşhis Edilmesi", Doktora Tezi, T.C. Ondokuz Mayıs Üniversitesi Lisansüstü Eğitim Enstitüsü, Samsun, 50-51 (2021).

11. Khan, R.J., Chowdhury, S., Islam, H. and Raheem, E., “Machine Learning Algorithms to Predict the Childhood Anemia in Bangladesh” Journal of Data Science, 17 (1): 195–218 (2019).

12. Appiahene, P., Asare, J. W., Donkoh, E. T., Dimauro, G. and Maglietta, R., “Detection of Iron Deficiency Anemia by Medical Images: A Comparative Study of Machine Learning Algorithms”, BioData Mining, 16 (1): 1-20 (2023).

13. Towfek El-Kenawy, E.-S. M., “A Machine Learning Model for Hemoglobin Estimation and Anemia Classification”, International Journal of Computer Science and Information Security, 17 (2): 100-108 (2019).

14. Asare, J. W., Appiahene, P., Donkoh, E. T. and Dimauro, G., “Iron Deficiency Anemia Detection Using Machine Learning Models: A Comparative Study of Fingernails, Palm and Conjunctiva Of The Eye Images”, Engineering Reports, 1-21 (2023).

15. Dejene, B. E., Abuhay, T. M. and Bogale, D. S., “Predicting the Level of Anemia Among Ethiopian Pregnant Women Using Homogeneous Ensemble Machine Learning Algorithm”, BMC Medical Informatics and Decision Making, 22 (1): 1-11 (2022).

16. Asare, J. W., Appiahene, P. and Donkoh, E. T., “Detection of Anaemia Using Medical Images: A Comparative Study of Machine Learning Algorithms – A Systematic Literature Review”, Informatics in Medicine Unlocked, 1-10 (2023).

17. Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I. and Shibghatullah, A. S., “A Non-Invasive Machine Learning Mechanism for Early Disease Recognition on Twitter: The Case of Anemia”, Artificial Intelligence in Medicine, 1-12 (2022).

18. Saputra, D. C. E., Sunat, K. and Ratnaningsih T., “A New Artificial Intelligence Approach Using Extreme Learning Machine as the Potentially Effective Model to Predict and Analyze the Diagnosis of Anemia”, Healthcare, 11 (5): 1-25 (2023).

19. Alsaadawi M., and Şehirli E., “The Efficiency of Ensemble Techniques in Predicting Thyroid Disorder: A Comparative Study”, International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 834–840 (2022).

20. Alsaadawi M. A. W., “Detection of Thyroid Disease Using Machine Learning Models”, M. Sc. Thesis, T.C. Karabuk University Institute of Graduate Programs, Karabuk, 91-92 (2023).

21. Salman K. and Sonuc, E., “Thyroid Disease Classification Using Machine Learning Algorithms” Journal of Physics: Conference Series, 1963(1): 1-12 (2021).

22. Aversano L., Bernardi, M. L., Cimitile, M., Iammarino, M., Macchia, P. E., Nettore, I. C. and Verdone, C., "Thyroid Disease Treatment Prediction with Machine Learning Approaches" Procedia Computer Science, 192, 1031–1040 (2021).

23. Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N. and Ahmad, A., "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach", BioMed Research International, 1-10 (2022).

24. Mir, Y. I. and Mittal, S., "Thyroid Disease Prediction Using Hybrid Machine Learning Techniques: An Effective Framework", International Journal of Scientific & Technology Research, 9 (2): 2868-2874 (2020).

25. Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J. L. V., Rodríguez, C. L. and Ashraf, I., "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques," Cancers, 14 (16): 1-23 (2022).

26. Raghuraman, M. T., Sailatha, E., and Gunasekaran, S., "Efficient Thyroid Disease Prediction and Comparative Study Using Machine Learning Algorithms", International Journal of Information and Computing Science, 6 (6): 617-624 (2019).

27. Shivastuti, Kour, H., Sharma, V. and Manhas, J., "Performance Evaluation of SVM and Random Forest for the Diagnosis of Thyroid Disorder", International Journal for Research in Applied Science & Engineering Technology (IJRASET), 9 (5): 945-947 (2021).

28. Mert Demirarslan, A. S., "Rutin Kan Testleriyle COVID-19 Tanı Tahmininde Makine Öğrenmesi Yöntemleriyle Bir Mobil Uygulama Geliştirilmesi", Ege Tıp Dergisi, 60 (4): 384-393 (2021).

29. Karadağ, K., "Kan Vermeye Elverişli Donörlerin Bilgisayar Destekli Tespiti", Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi, 508-514 (2021).

30. Emeç, M. ve Özcanhan, M. H., "Veri Ön İşleme ve Öznitelik Mühendisliğinin Yapay Zekâ Yöntemlerine Uygulanması", Mühendislikte Öncü ve Çağdaş Çalişmalar, 33-54 (2023).

31. Yüce, H., "Normalizasyon Tekniklerinin Biyomedikal Verilerde Siniflama Başarisina Etkisi", Yüksek Lisans Tezi, T.C. Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü, Konya, 6-27 (2021).

32. Karslı, Ö. B., "Makine Öğrenme Yöntemleri İle Karaciğer Hastaliğinin Teşhisi", Yüksek Lisans Tezi, T.C. Ağrı İbrahim Çeçen Üniversitesi Fen Bilimleri Enstitüsü, Ağrı, 3-5 (2019).

33. Kaba, G. ve Bağdatli Kalkan, S., "Kardiyovasküler Hastalık Tahmininde Makine Öğrenmesi Sınıflandırma Algoritmalarının Karşılaştırılması", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21 (42): 183-193 (2022).

34. Akyol, K. ve Karacı, A., "Diyabet Hastalığının Erken Aşamada Tahmin Edilmesi İçin Makine Öğrenme Algoritmalarının Performanslarının Karşılaştırılması", Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 9 (6): 123–134 (2021).

35. Bilgin, G. ve Çifci, A. "Eritematöz Skuamöz Hastalıkların Teşhisinde Makine Öğrenme Algoritmalarının Etkisi", Journal of Intelligent Systems: Theory and Applications, 4 (2): 195–202 (2021).

36. Doğaner, A., "Topluluk Öğrenme Yöntemleri İle Renal Hücreli Karsinom'un Tahmin Edilmesi", Doktora Tezi, T.C. İnönü Üniversitesi Sağlık Bilimleri Enstitüsü, Malatya, 7-8 (2020).

37. EKREM, Ö., SALMAN, O. K. M., AKSOY, B. ve İNAN, S. A., "Yapay Zekâ Yöntemleri Kullanilarak Kalp Hastaliğinin Tespiti", Mühendislik Bilimleri ve Tasarım Dergisi, 8 (5): 241–254 (2020).

ÖZGEÇMİŞ

2020 yılında, Azerbaycan Cumhuriyetinin Bakü ilindeki Azerbaycan Devlet Petrol ve Sanayi Üniversitesinin Sistem Mühendisliği bölümünün lisans programından mezun oldu. Aynı yıl, Karabük Üniversitesi Bilgisayar Mühendisliği bölümü Yüksek Lisans programını kazandı. 2023 yılında Emrullah SONUÇ ve Qusay SAIHOOD ile birlikte yazdığı "Çocuklarda Anemi Hastalığının Teşhisinde Topluluk Öğrenme Yöntemlerinin Kullanılması" isimli bildirisi yayınlandı. Şu an Yüksek Lisans eğitimine Karabük Üniversitesinde devam etmektedir.