T.C.
YALOVA ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE KOVİD-19 TRANSKRİPTOMİK BİYOBELİRTEÇLERİN BELİRLENMESİ

YÜKSEK LİSANS TEZİ

HATİCE YILDIZ

DANIŞMAN: PROF. DR. MURAT GÖK

YALOVA
NİSAN 2022

T.C.
YALOVA ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE KOVİD-19 TRANSKRİPTOMİK BİYOBELİRTEÇLERİN BELİRLENMESİ

YÜKSEK LİSANS TEZİ

HATİCE YILDIZ
185105014

DANIŞMAN: PROF. DR. MURAT GÖK

YALOVA
NİSAN

ETİK BEYAN

Yalova Üniversitesi Lisansüstü Eğitim Enstitüsü Tez Yazım Kuralları'na uygun olarak hazırladığım "MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE KOVİD-19 TRANSKRİPTOMİK BİYOBELİRTEÇLERİN BELİRLENMESİ" başlıklı bu tez çalışmasında; tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi, tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi, kullanılan verilerde herhangi bir değişiklik yapmadığımı, bu tezde sunduğum çalışmanın özgün olduğunu bildirir, aksinin tespiti halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi taahhüt ve beyan ederim.

Hatice YILDIZ

ÖNSÖZ

Tez çalışmamda değerli bilgileriyle bana yol gösteren, yardımlarını ve desteğini hiçbir zaman esirgemeyen tez danışmanım Prof. Dr. Murat Gök hocama, Çalışmam boyunca bana hem maddi hem manevi yardımlarını asla esirgemeyen başta annem, babam ve canım ablalarım olmak üzere tüm aileme sonsuz teşekkürlerimi sunarım.

Nisan 2021

Hatice Yıldız
(Bilgisayar Mühendisi)

İÇİNDEKİLER

TERİMLER

Artık Değerli Nöral Ağlar	: Residual Neural Network
Artımlı Özellik Seçimi	: Incremental Feature Selection
Bayes Ağları	: Bayes Networks
Çok Katmanlı Algılayıcı	: Multilayer Perceptron
Destek Vektör Makineleri	: Support Vector Machines
k-En Yakın Komşuluk	: k-Nearest Neighborhood
Lojistik Regresyon	: Logistic Regression
Maksimum Uygunluk ve Minimum Yedeklilik	: Maxi-Relevance and Min-Redundancy
Matthews Korelasyon Katsayısı	: Matthews Correlation Coefficient
Naif Bayes	: Naive Bayes
Parçaçık Sürü Optimizasyonunu	: Particle Swarm Optimization
Radyal Tabanlı Fonksiyon	: Radial Based Function
Rastgele Orman	: Random Forest
Yapay Sinir Ağı	: Artificial neural network

KISALTMALAR

ÇKA	: Çok Katmanlı Algılayıcı
DDVM	: Doğrusal Destek Vektör Makineleri
DN	: Doğru Negatif
DP	: Doğru Pozitif
DVM	: Destek Vektör Makineleri
DSÖ	: Dünya Sağlık Örgütü
GEO	: Gene Expression Omnibus
IFS	: Artımlı özellik seçimi
k-EYK	: k En Yakın Komşuluk
MCC	: Matthews Korelasyon Katsayısı
mRMR	: Maksimum Uygunluk ve Minimum Yedeklilik
PSO	: Parçaçık Sürü Optimizasyonunu
Resnet50	: Artık Değerli Nöral Ağlar
RTF DVM	: Radyal Tabanlı Fonksiyon Destek Vektör Makineleri
YN	: Yanlış Negatif
YP	: Yanlış Negatif
YSA	: Yapay Sinir Ağ

ÇİZELGE LİSTESİ

ŞEKİLLER LİSTESİ

Makine Öğrenmesi Yöntemleri ile Kovid-19 Transkriptomik Biyobelirteçlerin Belirlenmesi

ÖZET

Aralık 2019'da Çin'de ortaya çıkan, SARS-CoV-2 virüsünün neden olduğu salgın, dünya çapında kritik bir tehdit olmaya devam etmektedir. SARS-CoV-2 virüsü, insanlarda ölümcül bir hastalık olan kovid-19'a neden olmaktadır.

Diğer viral göğüs hastalıklarında görülen semptomlar Kovid-19 hastalığını taşıyan kişilerde de görülebilmektedir. Kovid-19 için iki önemli konu ele alınabilir. Bunlar sırasıyla şunlardır; Birincisi hastada herhangi bir semptom olmayabilir fakat yine de diğer insanlara virüs bulaştırabilir, ikincisi de Kovid-19'lu hastalar diğer solunum yolu enfeksiyonları ile aynı semptomları gösterebilir. Bu sebeple Kovid-19'un teşhisi için geniş çaplı tarama ve çalışmalara ihtiyaç vardır. Günümüzde, kovid-19'un tanısı ve tedavisi için bilim insanları yoğun bir şekilde çalışmaktadır.

Kovid-19'un tanısında, makine öğrenmesi yöntemleri ile yapılan çalışmalar önemli bir yere sahiptir. Bu konuda birçok çalışma yapılmıştır. Biz bu çalışmada, Kovid-19'un makine öğrenmesi yöntemleri ile tanısı için GEO veri tabanında elde ettiğimiz enfeksiyonlu ve sağlıklı hastalara ait kan ekspresyon verileri üzerinde Genetik Algoritma, Parçacık Sürü Enyinileme Yöntemi, En İyi Öncelikli öznitelik seçimi yöntemleri ile öznitelik sayısını düşürdük ve sonrasında çeşitli sınıflandırma algoritmaları (Naif Bayes, Bayes Ağları, k-EYK, Rastgele Orman, Lojistik Regresyon, Doğrusal DVM, Radyal Tabanlı Fonksiyon DVM, Poliynomal DVM, Çok Katmanlı Algılayıcı) ile hastalığı tahmin ettik.

Anahtar Kelimeler: Makine öğrenmesi, Koronavirus, Kovid-19, Öznitelik seçim yöntemleri, Sınıflandırma

Determination of Kovid-19 Transcriptomic Biomarkers using Machine Learning Methods

ABSTRACT

The epidemic caused by the SARS-CoV-2 virus, which emerged in China in December 2019, remains a critical threat worldwide. The SARS-CoV-2 virus causes the deadly disease, covid-19, in humans.

Symptoms seen in other viral chest diseases can also be seen in people with Kovid-19 disease. Two important issues can be addressed for Covid-19. These are as follows; First, the patient may not have any symptoms but can still infect other people, and secondly, patients with covid may show the same symptoms as other respiratory infections. For this reason, large-scale screening and studies are needed for the diagnosis of Covid-19. Today, scientists work intensively for the diagnosis and treatment of covid-19.

Studies with machine learning methods have an important place in the diagnosis of Kovid-19. Many studies have been done on this subject. In this study, we used Genetic Algorithm, Particle Swarm Optimization Method, Best Priority feature selection methods on blood express data of infected and healthy patients that we obtained in Gene Expression Omnibus (GEO) database for the diagnosis of Kovid-19 with machine learning methods. We reduced the number of cases and then predicted the disease with various classification algorithms (Naive Bayes, Bayesian Networks, k-NN, Random Forest, Logistic Regression, Linear SVM, Radial Based Function SVM, Polynomial SVM, Multilayer Perceptron).

Keywords: Machine learning, Coronavirus, Covid-19, Feature selection methods, Classification

1. GİRİŞ

2019 yılı sonunda Çin'in Wuhan eyaletinde ortaya çıkan Kovid-19 hastalığı hızlı bir biçimde yayılarak tüm dünyayı etkisi altına aldı [1]. Hastalığın milyonlarca kişiye bulaşması karşısında DSÖ, koronavirüs salgınını bir pandemi olarak ilan etti [2]. Kovid-19'a, koronavirüsün (SARS-CoV-2) yeni bir türünün sebep olmuştur [3].

Kovid-19, bulaştığı kişilerde ölüme kadar götüren ciddi bir solunum yolu enfeksiyonuna neden olmaktadır. 2021'in sonlarına doğru, dünya genelinde yaklaşık dört milyondan fazla kovid-19 kaynaklı can kaybı olduğu bildirilmiştir. Kovid-19, geçmişten günümüze en yaygın ve ölümcül bulaşıcı hastalıklardan biri olarak kabul edilmektedir [4].

Kovid 19, diğer virüs yoluyla bulaşan hastalıklardan farkı, tanı için kullanılabilecek tipik semptomlara sahip olmamasıdır. Diğer solunum yolu hastalıklarında görülen ateş, baş ağrısı, öksürük, kas, ishal ve/veya vücut ağrıları gibi pek çok çeşit belirtilerin Kovid-19 ile ilişkili olduğu ortaya konulmuştur [5]. Bununla beraber, Kovid-19 hastası herhangi bir belirti göstermeyebilir de. Ancak yine de virüsü yayabilir.

Kovid-19'un tedavisi için günümüzde hala bilim çevreleri aşı ve ilaç geliştirilmesi için yoğun bir çalışma içindedir. Kovid-19'un teşhis ve tedavi seyrinde, biyobelirteç molekülleri kullanılabilir. Biyobelirteçler, kanda, diğer vücut sıvılarında ya da dokularda bulunan, hastalığın işareti olan, normal veya anormal bir sürece tepki veren biyolojik moleküllerdir [6]. Vücudun Kovid-19 gibi bir hastalık veya durum için bir tedaviye ne kadar iyi yanıt verdiğini görmek için yeni biyobelirteçler bulunabilir [4].

In vivo, in vitro ve in silico ortamlarda yapılan çalışmalar tedavi için birbirini tamamlayıcısı olmaktadır. Makine öğrenmesi yöntemleri ile *in silico* ortamda yapılan çalışmalar laboratuvar şartlarına göre zaman ve mekan açısından üstünlüklere sahiptirler. Biyobelirteçlerin makine öğrenmesi öğrenmesi yöntemleri ile in silico ortamda tahmin edilmesi Kovid-19 tedavisi açısından kritik öneme sahiptir.

1.1. Tezin Amacı

Bilgisayar ortamında biyobelirteçler makine öğrenmesi yöntemleri ile hızlı bir şekilde tahmin edilebilmektedir. Biz bu tez çalışmasında; sınıflandırıcı algoritmanın başarımını artırmak için Kovid-19 enfeksiyonlu ve sağlıklı hastalara ait kan

ekspresyon verileri üzerinde yeni bir makine öğrenmesi modeli geliştirmeyi amaçlamaktayız.

1.2 Literatür Çalışması

Makine öğrenmesi yöntemleriye Kovid-19'un makine öğrenmesi yöntemleri ile tespiti literatürde sıklıkla çalışılmaktadır. Chen ve arkadaşları [6], yaptıkları çalışmada akut solunum yolu enfeksiyonu örneklerinde 15.379 genin kan ekspresyon profilleri bulunan hastalar üzerinde optimize edilmiş makine öğrenimi modellerini uyguladılar. Demografik bilgileri olan toplam 195 örnek üzerinde çalıştılar. Veri seti sırasıyla 19 sağlıklı kontrol, 23 pnömoni hastası, 17 grip hastası, 59 mevsimsel enfeksiyonlu hasta ve 77 Kovid-19 hastası içermektedir. Bu veri seti üzerinde iki güçlü öznitelik seçim yöntemi (mRMR ve Boruta) tek tek uygulanmıştır. Artımlı özellik seçimi (IFS) yöntemiyle beslenen bir özellik listesi oluşturuldu. Daha sonra IFS yönteminde dört klasik sınıflandırma algoritması denenmiştir.

Dairi ve arkadaşları [7], derin öğrenme ve makine öğrenmesi modellerinin performansını, Kovid-19'dan etkilenen yedi ülkeden (ABD, Brezilya, Fransa, Meksika, Hindistan, Rusya ve Suudi Arabistan) elde ettikleri hasta verileri üzerinde karşılaştırmışlardır.

Arpaci ve arkadaşlarının [8] çalışmasında, altı farklı sınıflandırıcı (J48, Bayes Ağları, Lojistik regresyon, k-EYK, CR ve PART) kullanarak Kovid-19 teşhisi için modeller geliştirmişlerdir. Bu çalışma, Çin'deki Zhejiang Eyaletindeki Taizhou hastanesinden 114 vakayı içeren veriler üzerinde gerçekleşmiştir.

Muhammad ve arkadaşları [9], Meksika'nın pozitif ve negatif Kovid-19 vakaları içeren veri seti üzerinde lojistik regresyon, karar ağacı, DVM, Naif Bayes ve Yapay Nötr Ağı içeren öğrenme algoritmaları kullanarak denetimli makine öğrenimi modelleri geliştirmişlerdir. Modeller geliştirilmeden önce, veri setinin her bir bağımlı özelliği ile bağımsız özelliği arasındaki kuvvet ilişkisini belirlemek için korelasyon katsayı analizi yapılmıştır.

Zhang ve arkadaşları [10], akut solunum yolu hastalıkları olan üst solunum yolu dokusunun transkriptomik verilerini kullanarak, SARS-CoV-2'nin diğer bulaşıcı hastalıklardan ayırt etmek için çoklu makine öğrenimi yöntemlerini geliştirmişlerdir. Veriler ilk olarak Boruta tarafından analiz edilmiş ve bir özellik listesi üretilmiştir. Daha sonra bu liste üzerinden makine öğrenme yöntemleri uygulanmıştır. Bu liste,

nitel biyobelirteç genlerini çıkarmak ve nicel kurallar oluşturmak için bazı sınıflandırma algoritmalarını içeren artımlı özellik seçimiyle beslenmiştir.

Kukar ve arkadaşları [11], çeşitli bakteriyel ve viral enfeksiyonları olan 5333 hastanın ve 160 pozitif Kovid-19 hastanın rutin kan testlerini içeren veri üzerinde Kovid-19 teşhisi için bir makine öğrenimi modeli oluşturmuşlardır.

McClain ve arkadaşlarının [12], 46 Kovid-19 hastasından aldıkları kan örneklerini RNA dizilimi yoluyla hastalığın diğer solunum yolu hastalıklarıyla karşılaştırmak için bir çalışma gerçekleştirmişlerdir. Farklı şekilde ifade edilen genlere dayalı sınıflandırıcılar, SARS-CoV-2 enfeksiyonunu diğer akut hastalıklardan doğru bir şekilde ayırt ettiğini göstermişlerdir.

Abed ve arkadaşları [13], 400 sağlıklı ve 400 Kovid-19 hastasını içeren veriler üzerinde makine öğrenme ve derin öğrenme modelleri kullanarak hastalığı tahmin etmişlerdir. Çalışmada elde edilen sonuçlara dayanarak, tüm modeller arasında, derin öğrenme modellerinden ResNet50 modelinin en yüksek doğruluğu verdiği belirtilmiştir. Buna karşılık, geleneksel makine öğrenimi tekniklerinde DVM'nin en yüksek doğruluğu verdiği gözlemlenmiştir.

Zoabi ve arkadaşlarının [14] yaptığı çalışmada, İsrail Sağlık Bakanlığı tarafından kamuya açıklanan ülke çapındaki verilere dayanarak 51.831 kişiden (4769'unun Kovid-19'lu vaka) oluşan kayıtlar üzerinde eğitim alan bir makine öğrenimi yaklaşımı oluşturmuşlardır. Test seti, sonraki haftanın verilerini içermektedir (47.401 test edilmiş kişiden 3624'ünün Kovid-19 olduğu doğrulanmıştır). Geliştirdikleri modelleri, Kovid-19 test sonuçlarını yüksek doğrulukla öngördüğü belirtilmiştir.

Kassania ve arkadaşlarının [15] yaptığı çalışmada, Kovid-19 hastalığının tahmini için derin öğrenme ve makine öğrenmesi kullanılarak yöntemlerin karşılaştırılması gerçekleştirilmiştir. İlk olarak çalışmada derin öğrenme yöntemleri uygulanmıştır. Daha sonra çıkarılan özellikler Kovid-19 vakasını sınıflandırmak için birkaç makine öğrenmesi sınıflandırıcısına aktarılmıştır. Önerilen yöntemin başarımı, halka açık bir Kovid-19 veri seti üzerinde doğrulanması gerçekleştirilmiştir.

Sujath ve arkadaşları [16], Kovid-19'un olası ilerlemesini tahmin etmek için farklı sayısal modeller kullanarak bir çalışma gerçekleştirmişlerdir. Farklı faktörlere ve araştırmalara bağlı bu sayısal modeller, potansiyel eğilime bağlı olduğu belirtilmiştir. Hindistan'daki Kovid-19 vakalarının epidemiyolojik örneğini ve hızını tahmin etmek için Kovid-19 Kaggle verilerinde doğrusal regresyon ve ÇKA regresyon yöntemlerini uygulamışlardır.

Patterson ve arkadaşları [17], Kavid-19'un etkili tedavi stratejilerinin tasarlanabilmesi ve izlenebilmesi için bir biyoenformatik yaklaşımı geliştirmişlerdir. Bu çalışma, Kovid-19 hastaları ve sağlıklı bireyler olmak üzere toplam 224 kişiyi kapsamaktadır.

1.3 Hipotez

Biz bu tez çalışmasında; öznitelik seçim yöntemlerine dayanan bir makine öğrenmesi modeli ile kan ekspresyon verileri üzerinde öznitelik sayısını azaltma yoluyla sınıflandırma performansını artırmayı hedeflemekteyiz. Böylelikle kan ekspresyon verileri ile Kovid-19 teşhisinin hızlı bir biçimde yapılabileceğini öngörmekteyiz.

1.4 Tezin İçeriği

Tez çalışmamız toplamda dört bölümden oluşmaktadır.

Bölüm 1'de, Kovid-19 un tanımı, kan ekspres verileri üzerinden tespiti, tezin amacı, literatür çalışması, hipotez ve tezin içeriği yer almaktadır.

Bölüm 2'de, bu tez için kullanılan yöntemler tek tek ele alınmıştır.

Bölüm 3'de, tezimiz için geliştirilen makine öğrenmesi modeli anlatılmıştır.

Bölüm 4'de ise bulgular anlatılmaktadır.

Son olarak Bölüm 5'de ise sonuçlar ele alınmaktadır.

2.YÖNTEMLER

2.1. Biyobelirteç

Biyobelirteçler, kanda, diğer vücut sıvılarında ya da dokularda bulunan, hastalığın işareti olan, normal veya anormal bir sürece tepki veren biyolojik moleküllerdir [6].

Biyobelirteçlerin kullanım amacı; popülasyonun taramasını gerçekleştirmek, tanıyı ortaya koymak, hastalığı evrelendirmek, tedavinin boyutunu değerlendirmek, riski hesaplamak, tedavinin izlenmesini gerçekleştirmek olarak belirtilebilir. Biyobelirteçler, ilaç geliştirme sürecinin iyileştirilmesinde olduğu kadar daha büyük biyomedikal araştırma kuruluşunda da kritik bir rol oynamaktadır. Ölçülebilir biyolojik süreçler ve klinik sonuçlar arasındaki ilişkiyi anlamak, tüm hastalıklar için tedavi cephanemizi genişletmek ve normal, sağlıklı fizyoloji anlayışımızı derinleştirmek için hayati önem taşımaktadır [18].

Biyobelirteçler, protein profili ya da gen ekspresyonu ile belirlenebilir. Daha sonra belirlenen biyobelirteçler prognostik, terapötik ve diagnostik değerlerine göre önceliklendirilerek klinik kullanıma geçişi önerebilecek hale getirilir.

İdeal bir biyobelirteçte bulunması gereken özellikler:

1. Biyolojik yapıdadırlar.
2. Biyokimyasal olarak küçük kan hacminde ve stabil olarak çalışılabiliyor olması gerekmektedir.
3. Uygun maliyetli, hızlı, kolay ve otomatize analiz yöntemleri ile saptanabilmelidir.
4. Biyobelirteçlerin eşik değeri iyi belirlenmeli ve sonuçlar karşılaştırılabilmelidir.
5. Viral patojeni ya da bakteriyeli ayırt edebilmelidir.
6. Organ disfonksiyonuyla veya inflamatuar yanıt ile bağıntısı olmalıdır.
7. Erken teşhis yaptırabilme özelliği göstermelidir.
8. İlgili hastalık veya hasar durumu için prognostik olmalı ve terapötik yanıtın yönlendirmesini gerçekleştirmelidir.

9. Ayrıca karakteristiği ya da ilgili yanıt istatistiksel olarak güçlü olmalıdır [19,20].

Çizelge 2'de biyobelirteçlerin avantaj ve dezavantajları verilmektedir [21].

Çizelge 2. Biyobelirteçlerin Üstün ve Kısıtları

Üstünlükleri	Kısıtları
Objektif değerlendirme	Zamanlama kritiği
Ölçüm hassasiyeti	Pahalı (analiz maliyetleri)
Güvenirliği	Depolama (numunelerin uzun ömürlü olması)
Anketlerden daha az önyargılı	Laboratuvar hataları
Hastalık mekanizmaları sıklıkla incelemesi	Normal aralığın oluşturulma zorluğu

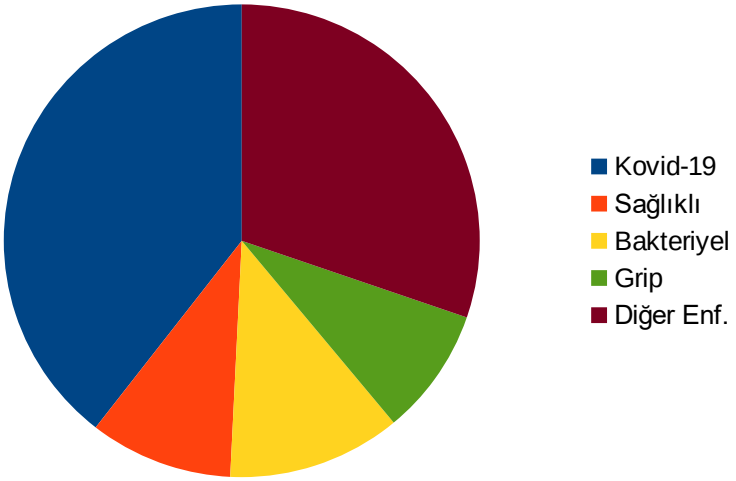
Kovid-19'un teşhis ve tedavi seyrinde, biyobelirteç molekülleri kullanılabilir. Canlı vücudunun, Kovid-19 gibi bir hastalık ya da durum için bir tedaviye ne kadar iyi yanıt verdiğini görmek için yeni biyobelirteçler tespit edilebilir [4]. Ayrıca Çizelge 2'de belirtildiği gibi biyobelirteçler, labaratuvar şartlarında analizleri maliyetlidir ve laboratuvar koşullarından kaynaklanan hatalara sahiptirler. Bu kısıtlar da göz önünde bulundurulduğunda in silico ortamda bilgisayar ile hesaplamalı analizler biyobelirteçler için ideal bir ortam oluşturmaktadır. Kandan elde edilen gen ekspresyon profil verileri biyobelirteç olarak kullanılabilir.

2.2. Kan Ekspresyon Verisi

Kan, farklı fonksiyonlara ve karşılık gelen gen ekspresyon profillerine sahip çok sayıda hücre tipinden oluşan karmaşık bir dokudur. Kan, T-hücreleri, B-hücreleri, monositler, NK hücreleri ve granülositler dahil olmak üzere çeşitli hücre tiplerini içerir ve bunların her biri daha da alt bölümlere ayrılabilir. Bu hücre tiplerinin her birinin nispi oranı, bireyler arasında sağlık ve hastalık durumlarıyla ve uyaranlara yanıt olarak büyük ölçüde değişebilir. Ekspresyon, genetik bir yapıdan meydana gelen aktivite ve yapıyı ifade etmektedir [22].

Kan numunelerinden, homojen hücre popülasyonlarının gen ekspresyon profillerini karşılaştırarak, hücre tipine özgü gen ekspresyon örüntülerine sahip genleri tanımlamak mümkündür. Bu gen kümeleri, belirli hücre türlerinin bolluğunu tahmin etmek için "biyobelirteçler" olarak hizmet edebilir ve hücresel işleyiş hakkında bilgi sağlayabilir [22].

Biz çalışmamızda GEO'dan elde ettiğimiz, Kovid-19 enfeksiyonlu ve sağlıklı kişilere ait 15.379 tane kan ekspresyon transkriptomik verilerini kullandık. Şekil 2.1'de görüldüğü gibi veri seti, 19 sağlıklı, 23 bakteriyel pnömonili hasta, 17 grip hastası, 59 diğer mevsimsel enfeksiyon hastası ve 77 Kovid-19 hastası olmak üzere toplam 195 örnekten oluşmaktadır. Makine öğrenmesi modellerini, ikili ve çoklu sınıf açısından bu veriler üzerinde test ettik.



Şekil 2.1. Veri Kümesinde Çoklu Sınıfın dağılımı

2.3 Önişleme Yöntemleri

Verinin kalitesi, veri madenciliğinde çok önemli bir konudur. Verinin güvenilirliğini ve kalitesini artırılması için, veri ön işleme uygulanmalıdır. Aksi takdirde hatalı veriler hatalı sonuçlara sebep olacaktır. Ön işleme verileri genellikle, her biri belirli bir yapıyı düzelten birden fazla adım içerir. Her biri belirli bir yapıyla ilgili olan birkaç bireysel ön işleme yönteminin, verilerde bulunan tüm yapaylıklara karşı koymak için bir ön işleme stratejisinde art arda uygulanması gerekecektir [23].

Günümüzde artan veriler ve buna bağlı olarak verilerin ön işlemeden geçirilmesinin gerekliliği büyük bir önem arz etmektedir. Veri ön işleme yöntemleri şu durumlarda uygulanmaktadır: Veri analizlerini engelleyecek veri sorunlarının çözümünü gerçekleştirmede, verilerin anlaşılmasını sağlamada ve tüm veri içinden daha anlamlı verinin çıkarılması işlemlerinde. Burada önemli olan veri ön işleme türünü belirlemektir [24]. Bunun için çeşitli veri ön işleme tekniği bulunmaktadır. Biz tez çalışmamızda standartlaştırma ve ayrıklaştırma yöntemlerini kullanmaktayız.

Ayrıklaştırma işleminde, bazı veri madenciliği algoritmalarında sürekli verilerin kesikli değerlere dönüştürülmesi gerekebilir. Böylelikle sürekli verilerin kesikli değer aralıklarına dönüştürülmesiyle oluşturulan kategorik değer verileri, orijinal veri değerlerinin yerine kullanma imkânı elde ederler. Burada bir kavram hiyerarşisi, sürekli değişkenler içinden verilerin ayrıştırılması olarak tanımlanabilir. Kavram hiyerarşileri, düşük seviyeli verilerin yüksek seviyeli verilerle değiştirilmesiyle bu verilerin indirgenmesinde kullanılır. Bu yöntem ile veri indirgemede genelleştirilmiş veriler daha kolay yorumlanabilecek, daha anlamlı ve daha düşük hacimli yer kaplayacaktır [24].

Standartlaştırma ise veri setindeki değişkenin kendi içindeki varyans ve bilgi yapısını bozmadan, değerler üstünde oynamalar yapıp belli bir format oluşturup bu formata göre işlenmesini sağlamak için veriyi standart hale getirme işlemine denir. Veri standartlaştırmada taşınan bilginin dağılımı yayılımı değiştirmediği görülmektedir [24].

2.4 Öznitelik Seçim Yöntemleri

Özellik Seçimi, verilerin anlaşılmasına, hesaplama gereksiniminin azaltılmasına, boyutsallık sorunlarının etkisinin azaltılmasına ve tahmin performansının iyileştirilmesine yardımcı olur. Öznitelik seçiminin odak noktası, girdiden, gürültüden veya alakasız değişkenlerden gelen etkileri azaltırken girdi verilerini verimli bir şekilde tanımlayabilen ve en iyi tahmin sonuçları sağlayan bir değişken alt kümesi seçmektir [25]. Çeşitli Öznitelik seçme yöntemleri mevcuttur. Tez çalışmamızda sırasıyla Genetik Algoritmalar Yöntemi, Parçacık Sürü Eniyileme Yöntemi ve En İyi Öncelikli Arama yöntemlerini kullandık.

2.4.1 Genetik Algoritma Yöntemi

Genetik Algoritmalar, en iyileme problemlerinin çözümünde tercih edilen bir yöntemdir. Bu yöntemde kromozom, seçilim, çaprazlama, gen, mutasyon ve uygunluk gibi biyolojik terimler en iyileme problemlerinin tasarım ve çözüm bileşenlerine eşlenmekte ve olabilecek en iyi çözümün bulunmasında kullanılmaktadır. Genetik Algoritmalar, problemden bağımsız olarak çalışmaktadır. Şöyle ki, Genetik Algoritmalar için oluşturulmuş prosedürler başka bir en iyileme probleminin çözümünde kullanılabilmektedir. Genetik Algoritmalarda her bir çözüm bileşeni bir gen olarak ifade edilmektedir. Genlerin bir araya gelmesiyle ile elde edilen kromozomlar ise birer çözüm adayını temsil etmektedir. Uygunluk fonksiyonu ile kromozomların kalitesine karar verilmektedir [26].

İlk başta popülasyon belirli sayıda kromozoma sahip olacak şekilde oluşturulmaktadır. Ardından, ebeveyn kromozomlar uygunluk değerine göre seçilip çaprazlanmakta ve bunun sonucunda yeni yavru kromozomlar oluşturulmaktadır. Sonlanma koşulu sağlanıncaya kadar bu işlem böyle devam etmektedir. Çaprazlamadan sonra mutasyon ile belirli oranda yavru kromozomların genleri üzerinde rastgele değişiklikler gerçekleştirilmektedir. Bu işlemden sonra, elde edilen yavru kromozomlar yeni popülasyona aktarılmaktadır. Önceki popülasyonun büyüklüğüne ulaşan kromozomlar yeni popülasyononun yerine geçmektedir. Sonlanma kriteri gerçekleştiği durumunda ise dönen değer en iyi çözüm kabul edilerek algoritma sonlandırılmaktadır [26].

Genetik Algoritma yöntemi adımları:

1. İlk populasyon için rastgele bir değer üret.

2. Populasyondaki tüm kromozomlar için amaç fonksiyon değerlerini oluştur.

3. Yeniden üreme, çaprazlama ve mutasyon işlemlerini uygula.

4. Elde edilen her yeni kromozomun amaç fonksiyonunu hesapla.

5. Kötü değerleri popülasyondan çıkar.

6. 3-5’deki adımları tekrar et [27, 28].

2.4.2 Parçacık Sürü Eniyileme Yöntemi

PSO, genellikle balık ve kuş gibi sürülerden görülen sürü zekâsı kavramından ilham alan bir algoritmadır. PSO'nun çalışma mantığı örnekleyecek olursak; Bir yerin üzerinde uçan bir kuş sürüsü, karaya çıkmak için bir nokta bulmalıdır ve bu durumda, tüm sürünün hangi noktaya inmesi gerektiğinin tanımı karmaşık bir problemdir. Çünkü burada gıda mevcudiyetini en üst düzeye çıkarmak ve yırtıcıların var olma riskinin en aza indirilmesi gerekmektedir. Bu bağlamda kuşların hareketi bir koreografi olarak anlaşılabilir. Kuşlar, konacak en iyi yer belirlenene ve tüm sürü aynı anda inene kadar bir süre eş zamanlı olarak hareket eder. Verilen örnekte, sürünün hareketi ancak tüm sürü üyeleri kendi aralarında bilgi paylaşabildiğinde anlatıldığı gibi gerçekleşir; aksi takdirde, her hayvan büyük olasılıkla farklı bir noktaya ve farklı bir zamanda inerdi. Kuşların sosyal davranışlarından esinlenerek, bu davranışı taklit edebilecek PSO adlı algoritma geliştirildi [29].

Sürüdeki her bir üye sürü parçacık optimizasyonunda bir çözümdür. Bu üyelerin her biri pozisyon ve hız bilgilerine sahiptir. Parçacıklar daha sonra en iyi pozisyona göre kendini yeniler.

Denklem (1) ve (2)'de Y_{id} hız değerini, X_{id} ise pozisyon değerini, c1 ve c2 öğrenme katsayılarını, r_1 ve r_2 0,1 aralığında olan rastgele sayılar ve W ağırlık değerini ifade eden PSO denklemi aşağıda gösterilmektedir [30].

$$Y_{id} = W x Y_{id} + c_1 x r_1 x (P_{id} - X_{id}) + c_2 x r_2 x (P_{jd} - X_{jd}) \quad (1)$$

$$X_{id} = X_{id} + Y_{id} \quad (2)$$

PSO adımları sırasıyla şöyledir:

1. Popülasyonun tanımlanması.

2. Her bir bireysel parçacığın uygunluk değerinin ölçülmesi.

3. Parçacıkların hızlarını en iyi önceki ve en iyi global değere göre değiştirilmesi. Konum ve hız değerleri, Denklem (1) ve Denklem (2)'ye göre kendini güncellemesi.

4. Belirlenen koşula göre kendini sonlandırma.

5. 2. adıma git [31].

2.4.3 En İyi Öncelikli Arama Yöntemi

En iyi öncelikli arama, iyi bilinen ve yaygın bir buluşsal arama algoritmasıdır [32]. En iyi öncelikli arama, bir değerlendirme işlevine dayalı olarak genişletme için sonraki düğümü seçen genel grafik ve ağaç arama algoritmalarının bir örneğidir. Bu nedenle bilgilendirilmiş bir arama stratejisidir.

Algoritma, genişletme için değerlendirme işlevi verilen en iyi düğüme erişmesi gerektiğinden, arama düğümlerini depolamak için bir öncelik sırası kullanır. Öncelik sırası genellikle, O(1) zamanında en yüksek (veya en düşük) önceliğe sahip nesneye erişim sağlayan karmaşık ancak verimli ağaç tabanlı bir veri yapısı olan Heaps ile uygulanır. En iyi düğüm seçildikten sonra, uygulanabilir her operatör, eylem, öncelik sırasına geri eklenen alt düğümler üretir, eklemeler O(log n) zaman alır, burada n, kuyruğun boyutudur. Algoritma, bir hedef durumu bulunana kadar arama düğümlerini seçmeye ve genişletmeye devam eder [33].

2.5 Sınıflandırma Yöntemleri

Sınıflandırma yöntemleri, sınıflandırıcı makine öğrenmesi algoritmalarını kullanarak kategorisi bilinmeyen verilerinin sınıflarının tahmin edilmesidir. Regresyon analizi gibi sınıflandırma yöntemleri de danışmanlı öğrenme yöntemidir [34].

Çeşitli sınıflandırma yöntemleri mevcuttur. Biz bu tez çalışmasında Naif Bayes, Bayes Ağları, k-EYK, Rastgele Orman, Lojistik Regresyon, DVM, ÇKA kullanmaktayız.

2.5.1 Naif Bayes

Naif Bayes, niteliklerin sınıfa göre koşullu olarak bağımsız olduğuna dair güçlü bir varsayımla birlikte Bayes kuralını kullanan basit bir öğrenme algoritmasıdır. Bu bağımsızlık varsayımı uygulamada sıklıkla ihlal edilse de Naif Bayes yine de yüksek sıklıkla sınıflandırma doğruluğu sağlar. Hesaplama verimliliği ve diğer arzu edilen birçok özelliği ile birleştiğinde bu, Naif Bayes'in pratikte yaygın olarak uygulanmasına yol açar [35].

Çok sık kullanılan Naif Bayes, olasılık tabanlı bir algoritma yöntemidir. Bu algoritma ile mevcut veri kümesi üzerinden her bir verinin sistem üzerinde koşullu olasılığı hesaplanabilmektedir. Gauss Dağılımı gibi metotlar yardımı ile koşullu olasılığın hesaplanması için istenen özelliğe bağlı olasılıkların her birinin değeri

hesaplanarak çarpma fonksiyonunun uygulanması gerekir. Son değer ise işleme uygun hale getirilmesi için normalleştirilir [36].

Naif Bayes sınıflandırıcısında önemli iki özellik vardır. Bunlar;

1. Tüm niteliklerin aynı derecede önem arz etmesi,

2. Niteliklerin her biri birbirinden bağımsız olarak kabul edilmesi [37].

Naif Bayes bayes teoremini kullanır. Koşullu olasılıklar arasında bağıntı bayes teoremi ile kurulur. a özniteliğinin bilinmesi koşulu ile C sınıfının bilinmesi koşulunda a özniteliğinin var olma olasılığında Bayes teoremindeki eşitlik denklem (3)'deki gibi bir durum mevcuttur.

$$P(C|a) = \frac{P(C)P(a|C)}{P(a)} \tag{3}$$

Tüm Özniteliklerin birbirlerinden bağımsız olduğu kabul edilmesi durumunda Bayes bağıntısı çok değişkenli uzayda denklem (4) de olduğu gibi yalın olur.

$$P(C|a_1,\ldots.a_n) = \frac{P(C).\prod_{k-1}^{n} P(a_k|C)}{P(a_1,\ldots.a_n)} \tag{4}$$

Denklem (4) de bağıntı aracılığıyla otomatik sınıflandırma yapılabilir. Öznitelik bilinen bir vektörün hangi C sınıfına ait olduğu hesaplanabilir. Naif Bayes, karar sınıfı için en iyi olasılık sınıfı hangisi hesaplanmışsa onu seçer ve işlemine devam eder [37].

2.5.2 Bayes Ağları

Bayes ağları her düğüm ayrı bir değişkeni ifade eden yönlü dönüşsüz ağlardır. Bayes ağları ile bu değişkenler arasındaki sıralama gösterilir. Kısaca bir düğümden diğer bir düğüme geçiş sırasını gösterir [38]. Judea Pearl (1985) tarafından önerilen bu ağlar, veri modelemede çok sık kullanılarak özellikle geçmiş olayları analiz ederek gelecek ile ilgili olaylar hakkında çıkarım yapabilmemize imkân sağlar. Bu ağlar bayes teoremi'nin üzerine kurulmuştur. Kısacası bu ağlar olasılık ilişkisi mevcut küme elemanlarının ilişkilerini modellemeye yardımcı olan, kararların ya da belirsiz değerlerin grafikselleştiren, kompleks matematiksel ifadeleri ve denklemlerin analizini kolaylaştıran ağlardır. Bu ağlar düğümler arasındaki yollarda bulunan değişkenler arasındaki olasılık ilişkilsini hesaplar. Bu ilişkiler, hem ilişkisiz düğümlerden hem de ilişkili düğümlerden oluşan bir yönlendirilmiş döngüsüz diyagram ile ifade edilir [39].

Bayes ağları olaylar arasındaki neden sonuç ilişkisini açıklamak için kullanılan bir yöntemdir. Örnek verecek olursak bir hastalık ve o hastalığın gösterdiği semptomlar arasındaki bağlantıyı inceleyip arasındaki ilişkiyi ortaya koyar [40].

Bayes ağları, bağımsızlıkla ilgili yersiz varsayımlardan kaçınarak naif Bayes sınıflandırıcılarının performansını iyileştirir. Bu sorunu etkili bir şekilde ele almak için, bağımsızlık iddialarını manipüle etmek için uygun bir dile ve verimli bir makineye ihtiyac vardır. Her ikisi de Bayes ağları tarafından sağlanmaktadır [41].

Bir Bayes ağı, etki alanı değişkenleri arasındaki koşullu bağımlılıkları gösteren bir yapıdır ve etki alanı değişkenleri arasındaki olasılık temelli ilişkileri grafiksel olarak göstermek için de kullanılabilir. Bir Bayes ağı, yönlendirilmiş bir döngüsel olmayan grafik ve olasılık tablolarından oluşur. Ağın düğümleri değişkenlerini temsil eder ve iki düğüm arasındaki bir yay, bu iki düğüm arasında temel bir ilişkinin veya bağımlılığın varlığını gösterir [42].

2.5.3 k-En Yakın Komşuluk Algoritması

k-EYK, veri madenciliğinde kullanılan bir sınıflandırma yöntemidir. k-En Yakın Komşu sınıflandırıcısı öznitelik uzayında yer alan en yakın eğitim verilerine dayanarak verileri sınıflandıran en kolay ve en sık kullanılan sınıflandırma algoritmalarından biridir. Bu yöntem sınıflandırma işlemini verilen k değeri kadar en

yakın komşunun sınıfına göre gerçekleştirmektedir. k-EYK algoritması, sınıfı bilinen vektörler kullanılarak bir vektörün sınıflandırılmasını yapmaktadır.

Test setindeki veriler eğitim setindeki her bir veri ile tek tek işleme alımı gerçekleştirilir. Test setindeki verilerin sınıfını bulabilmek için eğitim setindeki o veriye en yakın k adet örnek veri seçilir. Test edilecek olan veri seçilen örnek veriler arasında hangi sınıfta en çok veri varsa o sınıfa ait olduğu kararlaştırılır. Veriler arasındaki uzaklık öklid yöntemi ile belirlenir [43].

Kısacası bilinmeyen bir örneği sınıflandırmak için k-EYK sınıflandırıcısı belgenin komşularını eğitim örnekleri arasında sıralar ve giriş örneğinin sınıfını tahmin etmek için en çok benzerlik gösteren k komşunun sınıf etiketlerini kullanır [44].

k-EYK yönteminde test verisinin eğitim verilerine uzaklığı belirtildiği gibi öklid yöntemi ile hesaplanır. Test verisinin en yakınında bulunan k adet eğitim verisi en yüksek oranla hangi sınıfa aitse test veriside aynı sınıfa aittir. A=(y_1,y_2,y_3,.....y_n) ve B=(t_1,t_2,t_3,.....t_n) arasındaki öklid uzaklığı aşağıda gösterilmektedir.

$$(AB) = \sqrt{\sum_{i=1}^{n} (y_i - t_i)^2} \tag{5}$$

k-En Yakın Komşuluk yöntemi eğitim verisinin fazla olduğu durumlarda ve sayısal veriler üzerinde işlem yapılması gibi durumlarda avantaj sağlamaktadır [45].

2.5.4 Rastgele Orman

Rastgele Orman, birleştirilmiş karar ağaçları topluluğu olarak tanımlanabilir. Torbalama yöntemine rastgelelik özelliği eklenerek oluşturulmuştur. Sınıflandırılacak veri setinde, orman içindeki tüm ağaçta sınıflandırma işlemini gerçekleştirmektedir. Verinin sınıfını belirlemek için daha sonra oylama işlemi gerçekleştirilir ve sınıf belirlenir. Rastgele orman sınıflandırıcılarının en büyük avantajı birden fazla karar ağacından oluşan veri içinden orman adına tek bir karar vererek yüksek güvenilir tahminler ortaya koymasıdır. Ayrıca Rastgele Orman yönteminin aşırı öğrenme problemi yoktur [46].

Rastgele orman sınıflandırıcısının aşamaları aşağıda verilmektedir:

1. Karar ağacı adedi n'i belirlemek için veri setinde bulunan öznitellikler kullanılır .

2. En iyi dalı belirlemek için karar ağaçları arasında her düğümde rastgele t adet değişken seçilir.

3. Belirlenen en iyi dal tekrar kendi içinde iki alt dala ayrılır. Ve daha sonra her bir yaprak düğümde bir sınıf kalıncaya kadar yani gini indeksi sıfıra ulaşınca ağaç dallanma işlemi devam eder [47].

Sınıf homojenliğini ölçmek için gini indeksi kullanılır. Burada X_i rastgele seçilen pikselin ait olduğu sınıfı, T eğitim veri kümesini, X_i ve $f(X_i,T)/|T|$, seçilen verinin X_i sınıfına ait olma olasılığını gösteren denklem (6) daki gibi ifade edilmektedir [48].

$$\sum \sum i \neq j \left(f\left(X_i, T\right)/|T|\right)\left(f\left(X_i, T\right)/|T|\right) \tag{6}$$

4. Ayrı ayrı yapılan tüm tahminler sonucu en çok oy alan sınıf seçilir ve son karar olarak belirlenir.

Rastgele orman sınıflandırıcıda en önemli seçim, her düğümde geliştirilecek ağaç sayısı ve kullanılacak değişken sayısı parametrelerinin seçimidir [47].

2.5.5 Lojistik Regresyon

Bir gözetimli algoritma olan lojistik regresyon, ikili problemlerde kullanılan istatistiksel bir algoritmadır. Regresyon analiz sonucunda veri setinin iki olasılıklı sonucunu belirtir [49]. Lojistik Regresyon, veri kümesindeki değişkenlerin birbirlerini hangi yönden etkilediğini ortaya koyar. Lojistik Regresyon bağımlı değişken ile bağımsız değişken arasındaki olası ilişkiyi araştırır. Burada bağımsız değişkenler kategorik veya sürekli bir yapıya sahip iken bağımlı değişkenlerde iki veya ikiden fazla sınıfı içeren bir yapıya sahiptir [50]. Lojistik regresyon, karmaşık veri kümelerini anlamak için kullanışlı bir istatistiksel tekniktir. Lojistik regresyon, hastanın yaşaması veya ölmesi veya hastanın tedaviye yanıt vermesi veya yanıt vermemesi gibi ikili bağımlı bir değişkenin arasındaki ilişkiyi test eder [51].

Lojistik regresyon analizi logit dönüştürmeden adını almaktadır. Logit dönüştürme işlemi bağımlı değişkenlere uygulanmaktadır. Lojistik regresyonda en çok olabilirlik yöntemi en küçük kareler yöntemine tercih edilerek model oluşturulmaktadır. Lojistik regresyonda olasılık, odds ve odds'un yöntemine dayanmaktadır. Kısaca

odds bir olayın oluşma olasılığının oluşmama olasılığına bölünmesi olarak ifade edilir.

Lojistik regresyon analizi çok değişkenli regresyon analizi ve diskriminant analizinden farklı yönü bağımsız değişkenin dağılımı için varsayımlara ihtiyaç duymamasıdır. Lojistik regresyon analizinde doğrusallık, bağımsız değişkenlerin normal dağılması ve kovaryans-varyans matrislerinin eşitliği gibi varsayımların karşılaması gerekmemektedir [52].

A : Veri kümesinde araştırılan durumun gözlenme olasılığını,

β_0 : Bağımsız değişkenin 0 olması durumunda bağımlı değişkenin değerinin sabiti,

$\beta_1\beta_2$.....βn : Bağımsız değişkenlerin regresyon katsayılarını,

Y_1Y_2.....Yn: Bağımsız değişkenleri,

k : Bağımsız değişkenlerin sayısını,

e : Sabit 2.71 sayısı olarak verilen Lojistik Regresyon denklem (7) formüle edilmektedir.

Lojistik regresyon denklem (7)'de A ele alınan olayın gözlenme olasılığını belirtmektedir [53].

$$A = \frac{e^{\beta_0+\beta_1 Y_1+\ldots..+\beta_n Y_n}}{1+e^{\beta_0+\beta_1 Y_1+\ldots..+\beta_n Y_n}} \quad (7)$$

2.5.6 Destek Vektör Makineleri

Destek vektör makineleri ilk olarak sınıflandırma ve regresyon analizi problemlerini çözmek için önerildi. DVM çeşitli disiplinlerden gelen farklı veri kategorilerini sınıflandırmak için eğitilmiş bir denetimli öğrenme tekniğidir. DVM'ler hem doğrusal hem de doğrusal olmayan veri sınıflandırma problemlerinde uygulanabilir. DVM, yüksek boyutlu bir uzayda bir düzlem veya çoklu düzlemler oluşturur ve bunların içindeki en iyi düzlemi bularak verileri en uygun şekilde farklı sınıflara böler [54].

Destek vektör makineleri yüksek boyutlu uzayda doğrusal olmayan sınırları ikinci dereceden optimizasyon problemlerini çözerek tanımlama özelliğine sahiptir. DVM teorisi, yüksek boyutlu uzayda ki bir veri seti için sınıfları birbirinden ayıran sonsuz

çizginin olduğunu ifade eder. Destek vektörleri olarak tanımlanan sadece eğitim verilerinin bir alt kümesindeki veriler kullanılarak iki sınıfı birbirinde en iyi şekilde ayıran çizgi seçilir. Optimal karar sınırını temsil etmek adına destek vektörleri arasındaki en yüksek sınır mesafesi alınır. Bazı sınıfların doğrusal olarak ayrılamadığı durumlarda mevcuttur. DVM bu durumda girdi değişkenlerinin örtülü bir dönüşümü için kernel işlevini kullanır. Kernel işlevleri kısaca destek vektör makinelerinin doğrusal düzlem üzerinde doğrusal olmayan destek vektörlerinin ayrılmasına olanak sağlar. Birçok uygulamada performansı en yüksek değerde tutmak için uygun bir çekirdek genişliği ve işlevinin belirlenmesi gerekir [55].

Destek vektör makineleri çok sık olarak kullandığı başlıca dört çekirdek fonksiyonu mevcuttur. Bu çekirdek fonsiyonları şunlardır:

a. Doğrusal Fonksiyon,

b. Polinomiyal Fonksiyon

c. Radyal Tabanlı Fonksiyon,

d. Sigmoid Fonksiyon [56].

Destek vektör makinesi, nesnelere etiket atamayı öğrenen bir bilgisayar algoritmasıdır. Örneğin, bir DVM, yüzlerce veya binlerce hileli ve hileli olmayan kredi kartı faaliyet raporunu inceleyerek hileli kredi kartı faaliyetini tanımayı öğrenebilir. Alternatif olarak bir DVM, elle yazılmış sıfırlar, birler ve benzerlerinden oluşan geniş bir taranmış görüntü koleksiyonunu inceleyerek el yazısı rakamları tanımayı öğrenebilir. DVM'ler ayrıca giderek daha geniş çeşitlilikte biyolojik uygulamalara başarıyla uygulanmaktadır [57].

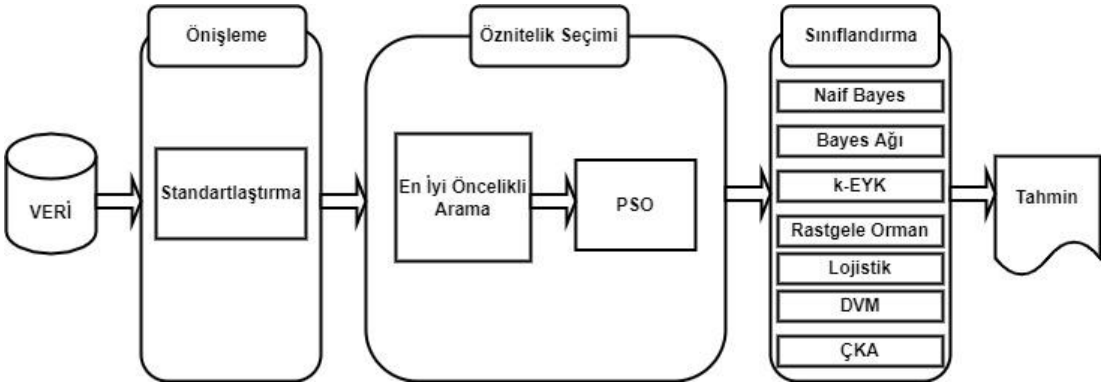
2.5.7 Çok Katmanlı Algılayıcı

Çok katmanlı algılayıcı, katmanlar halinde düzenlenmiş nöronlardan oluşan tam bağlantılı, ileri beslemeli bir yapay sinir ağı türüdür. ÇKA modeli en az üç katmandan oluşur: tek bir girdi katmanından, bir çıktı katmanından ve son olarak bir veya daha fazla gizli (ara) katmandan meydana gelir. ÇKA modelinin tercih edilme sebebi, bu yöntemlerin uygulama kolaylığından kaynaklanmaktadır. ÇKA'nın ayrıca yüksek kaliteli modeller sağladığı ve eğitim süresini daha karmaşık yöntemlere kıyasla nispeten düşük tuttuğu bilinmektedir [58].

Çok katmanlı algılayıcı modellerde problemle ilişkili bilgiler giriş katmanında tutulur ve bu bilgilerin sayısı kadar nöron bu katmanda bulundurur. Birden fazla ara katman sayısı olabilir ve nöron sayısı ile katman sayısı deneme yanılma yöntemiyle bulunur. Bilgiler işlendikten sonra çıkış katmanına gönderilir yani çıkış katmanı işlenen verileri barındırır. Ve çıkış katmanı ne kadar sınıflandırma çeşidi içerirse o kadar nörondan oluşur [59].

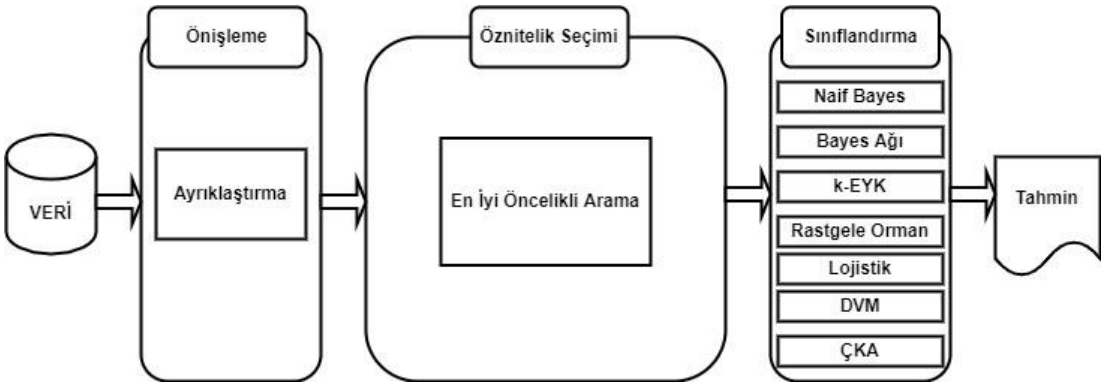
3. GELİŞTİRİLEN MAKİNE ÖĞRENMESİ MODELİ

Bu çalışmada, Kovid-19 hastalığının tanısının kandaki biyobelirteçler üzerinden yüksek doğrulukla yapılabilmesi için Şekil 3.1 ve Şekil 3.2'de görülen iki makine öğrenmesi modeli geliştirdik. İlk olarak kan ekspresyon verilerini önişlemden geçirdik. Sonra öznitelik seçme işlemini uygulayarak hastalığın yüksek doğrulukla teşhisi için gerekli öznitelik sayısını düşürdük. Son aşamada Naif Bayes, Bayes Ağları, k-EYK, Rastgele Orman, Lojistik Regresyon, DVM ve ÇKA algoritmaları ile tahmin gerçekleştirdik.



Şekil 3.1. Geliştirilen Makine Modeli 1

Şekil 3.1'de gösterilen ilk modelimizde, GEO'dan elde ettiğimiz enfeksiyonlu ve sağlıklı olan hastalara ait kan ekspresyon verilerini içeren transkriptomik verileri stadartlaştırma uyguladık. İkinci aşamada, en iyi öncelikli arama ve parçaçık sürü optimizasyonu yöntemlerini ard arda uygulayarak öznitelik sayısını 15.379 öznitelikten 121'e düşürdük. Son aşamada, elde ettiğimiz veriler üzerinde Naif Bayes, Bayes Ağları, k-EYK, Rastgele Orman, Lojistik Regresyon, DVM ve ÇKA algoritmaları ile sınıflandırma yaptık.



Şekil 3.2. Geliştirilen Makine Modeli 2

Şekil 3.2 de ikinci modelimiz görülmektedir. Burada çoklu sınıflandırılma işlemi ile beş adet sınıfa dair sınıflandırma yaptık. İlk modelden farklı olarak önişlem adımında standartlaştırma yerine ayrıklaştırma yöntemi uyguladık. Öznitelik safhasında öznitelik sayısı 15.379'dan 200'e düştü. Son aşamada ilk modelde kullandığımız algoritmalar ile sınıflandırma yaptık.

4. BULGULAR

4.1 Başarı Metrikleri

Başarım ölçmek için karmaşıklık matrisinden elde ettiğimiz doğruluk, kesinlik, duyarlılık, F-skor, MCC ve AUC metriklerini kullandık.

	Tahmin		
Gerçek		P	N
	P	DP	YN
	N	YP	DN

Şekil 4.1. Karmaşıklık Matrisi

Karmaşıklık matrisinde;

DP: Pozitif sınıf verilerinden doğru sınıflandırılan verilerin sayısını;

DN: Negatif sınıf verilerinden doğru sınıflandırılan verilerin sayısını;

YN: Gerçekte pozitif olan ancak negatif olarak sınıflandırılan verilerin sayısını;

YP: Gerçekte negatif olan ancak pozitif olarak sınıflandırılan verilerin sayısını verir [60].

Doğruluk, doğru tahmin edilen toplam pozitif ve negatif örneklerin oranını verir [60]:

$$Doğruluk = \frac{(DP + DN)}{(DP + YP + DN + YN)} \tag{8}$$

Kesinlik, tahmin edilen pozitif örnekler içerisinde doğru tahmin edilen pozitif örneklerin oranını verir [61]:

$$Kesinlik = \frac{DP}{(YP + DP)} \tag{9}$$

Duyarlık, gerçek pozitif örneklerden doğru tahmin edilen pozitif örneklerin oranını verir [61]:

$$Duyarlılık = \frac{DP}{(YN + DP)} \tag{10}$$

F-Skor, duyarlılık ve kesinlik metriklerinin birlikte değerlendirilebilmesini sağlar. F-skor duyarlılık ve kesinlik metriklerinin harmonik ortalamasıdır [60]:

$$F = \frac{2 * D * K}{K + D} \tag{11}$$

Kappa metriği, +1 ile -1 arasında değer alır. K<0 durumunda gözlemlenen durumun şansa bağlı durumdan küçük olduğunu ve K≥0 olduğunda gözlemlenen durumun şansa bağlı durumdan büyük ya da ona eşit olduğunu gösterir. Kappa değeri formülü aşağıda verilmiştir. Burada T_c kabul edilmesi gereken ve T_o kabul edilen oranı değeri olmak üzere,

$$K = \frac{T_0 - T_c}{1 - T_c} \tag{12}$$

formülü ifade edilir [62].

MCC, diğer metriklerden farklı olarak karışıklık matrisinde bulunan tüm değerleri kullanır. +1 ile -1 değerleri arasında yer alır. MCC=0 ise bu rastgele tahminde bulunduğunu, aldığı değerler -1'e yakınsa yanlış tahminde bulunduğunu ve aldığı değer +1'e yakınsa bu da doğru tahminde bulunduğunu göstermektedir [60]:

$$MCC = \frac{(DP * DN - YP * YN)}{\sqrt{(DP + YP) * (DP + YN) * (DN + YP) * (DN + YN)}} \tag{13}$$

AUC, sınıflandırma işlemlerinde başarımı test etmek için kullanılan bir diğer yöntemdir. 1 ve 0 arasında değerler üretir. AUC değeri, 1'e yakın bir değerse bu doğru tahmin etiğini ve aldığı değer 0.5'in altındaysa bu da yanlış tahminde bulunduğunu göstermektedir [63]:

$$AUC = \frac{1}{2} * \left(\frac{DP}{DP + YN} + \frac{DN}{DN + YP} \right) \tag{14}$$

4.2 Deneysel Sonuçlar

Kovid-19 hastalığının teşhisi için geliştirdiğimiz en iyi başarımı veren Model-1 ve Model-2'ye ait sonuçlar Çizelge 4.1 ve 4.2'de görülmektedir.

Çizelge 4.1. Model 1 için Sınıflandırma algoritmalarının başarım metrikleri

Sınıflandırıcılar	Doğruluk (%)	Kesinlik	Duyarlık	AUC	MCC	F-Skor	Kappa
Naif Bayes	88,71	0,802	0,948	0,972	0,779	0,869	0,771
Bayes Ağı	98,46	0,963	1	0,999	0,969	0,981	0,968
k-EYK	93,33	0,881	0,961	0,977	0,865	0,919	0,863
RO	97,44	0,95	0,987	0,997	0,947	0,968	0,947
Lojistik	89,74	0,861	0,883	0,951	0,787	0,872	0,786
Doğrusal DVM	95,38	0,936	0,948	0,953	0,904	0,942	0,904
RBF DVM	94,36	0,934	0,922	0,94	0,882	0,928	0,882
Poly DVM	67,18	0,933	0,182	0,587	0,318	0,304	0,202
ÇKA	95,9	0,937	0,961	0,986	0,915	0,949	0,915

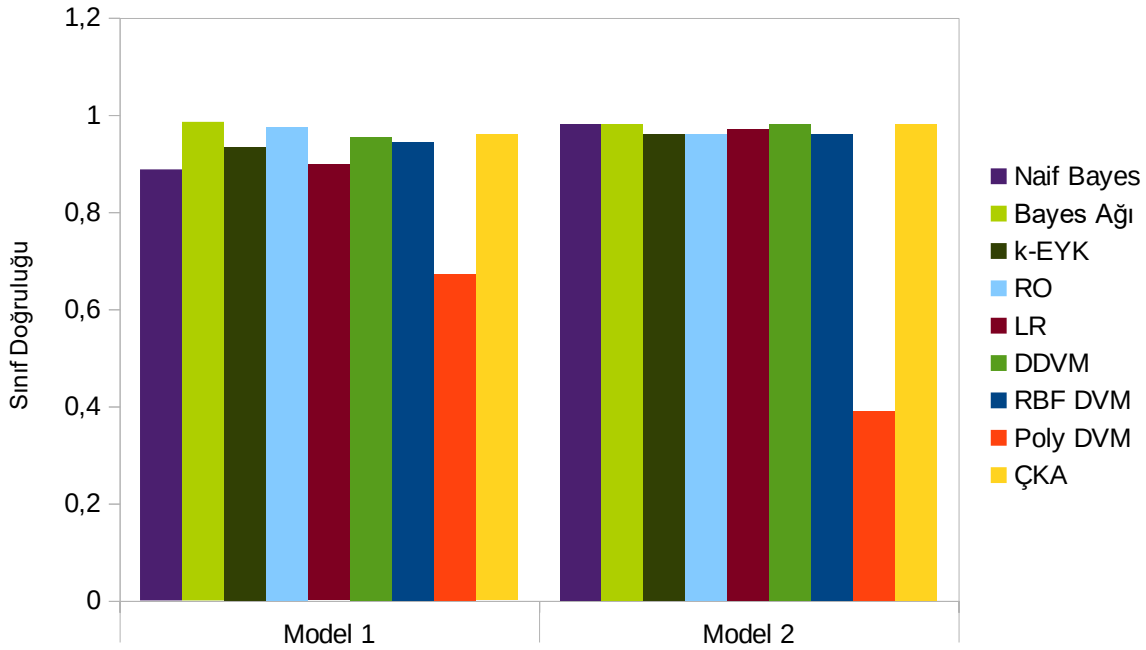
Çizelge 4.1 de bizim geliştirdiğimiz model 1 için sınıflandırma algoritmalarının başarım metrikleri verilmiştir. Model 1 de daha önce belirtiğimiz gibi ikili sınıflandırılmış veri seti kullanıldı. Veri setinden elde ettiğimiz öznitelik listesine sırasıyla Naif Bayes, Bayes Ağı, k-EYK, Rastgele Orman, Lojistik Regresyon, DVM, Polinomal DVM, Radyal Tabanlı DVM ve son olarak ÇKA sınıflandırma algoritmaları uygulandı. Burada biz başarı metrikleri olarak doğruluk, kesinlik, duyarlılık, AUC, MCC, f-ölçütü ve kappa değerlerini kullandık. Sınıflandırma sonucunda en yüksek değeri Bayes Ağılarının verdiği gözlemlenmiştir. Bayes Ağı için sınıf doğruluğu %98,46, kesinlik 0,963, duyarlık 1, AUC 0,999, MCC 0,969, f-skor 0,981 ve son olarak kappa 0.968 ile en yüksek değerler elde edilmiştir. Bayes Ağından sonra en yüksek sınıf doğruluğu veren ikinci sınıflandırma algoritması Rastgele Orman algoritmasıdır. Rastgele orman sınıflandırma algoritması için sınıf doğruluğu %97,44, kesinlik 0,95, duyarlılık 0,987, roc 0,997, mcc 0,947, f-skor 0,968 ve son olarak kappa 0,947 ile ikinci en yüksek değerleri elde edilmiştir. En düşük sınıf doğruluk değerini %67,18 ile Polinomal DVM'nin verdiği gözlemlenmiştir. Sınıflandırma algoritmaları sonucunda sınıf doğruluğu metrik değerleri Bayes Ağı (%98,46), Rastgele Orman (%97,44), ÇKA (%95,9), Doğrusal DVM (%95,38), RBF DVM (%94,36), k-EYK (%93,33), Lojistik (%89,74), Naif Bayes (%88,71) ve Polinomal DVM (%67,18) olarak elde edilmiştir.

Çizelge 4.2. Model 2 için Sınıflandırma algoritmalarının başarım metrikleri

Sınıflandırıcılar	Doğruluk (%)	MCC	Kappa
Naif Bayes	98,46	0,979	0,979
Bayes Ağı	98,46	0,979	0,979
k-EYK	96,41	0,95	0,95
RO	96,41	0,95	0,95
Lojistik	96,5	0,953	0,96
Doğrusal DVM	98,46	0.979	0,979
RBF DVM	96,92	0,958	0,957
Poly DVM	39,49	NA	0,362
ÇKA	98,46	0,979	0,979

Çizelge 4.2 de geliştirdiğimiz model 2 için sınıflandırma algoritmalarının başarım metrikleri verilmiştir. Model 2 de çoklu sınıflandırılmış veri seti kullanıldı. Veri setinden elde ettiğimiz öznitelik listesine sırasıyla sınıflandırma algoritmaları uygulandı. Burada biz başarı metrikleri olarak sınıf doğruluğu, MCC ve kappa değerlerini kullandık. Sınıflandırma sonucunda en yüksek değeri burada Bayes Ağı, Naif Bayes, ÇKA ve Doğrusal DVM'nin verdiği gözlemlenmiştir. Bayes Ağı, Naif Bayes, ÇKA ve Doğrusal DVM için sınıf doğruluğu %98,46 MCC 0,979 ve son olarak kappa 0,979 ile en yüksek değerler elde edilmiştir. En düşük sınıf doğruluk değerini %39,49 ile Polinomal DVM'nin verdiği gözlemlenmiştir. En düşük sınıf doğruluğunu veren Polinomal DVM'nin MCC değeri hesaplanamamaktadır bu yüzden not applicable (NA) değerini almıştır. Sınıflandırma algoritmaları sonucunda sınıf doğruluğu metrik değerleri Bayes Ağı (%98,46), Rastgele Orman (%96,41), ÇKA (%98,46), Doğrusal DVM (%98,46), RBF DVM (%96,92), k-EYK (%96,41), Lojistik Regresyon (%96,5), Naif Bayes (%98,46) ve Polinomal DVM (%39,49) olarak elde edilmiştir.

Şekil 4.2'de iki modelimiz için sınıflandırma algoritmalarının doğruluk değerlerinin karşılaştırılmalı sonucu grafiksel olarak görülmektedir.



Şekil 4.2. Geliştirilen Modellerin Sınıf Doğruluk değerleri

Çizelge 4.3 ve Çizelge 4.4 de çalışmamızın diğer çalışmalarla karşılaştırılması verilmektedir.

Çizelge 4.3. Çalışmamızın Farklı Veri Setini Kullanan Diğer Çalışmalarla Kıyaslanması

	Doğruluk (%)	MCC
Çalışmamız	98,46	0,98
Zhang ve arkadaşları	89,3	0,83
Arpaci ve Arkadaşları	84,21	NA
Muhammad ve arkadaşları	94,99	NA

Zhang ve arkadaşları, Makine öğrenme yöntemlerinden en yüksek değeri Rastgele Orman ile sınıf doğruluğu %89,03 ve MCC 0,83 olarak elde edilmiştir [10]. Arpaci

ve arkadaşlarının çalışmasında, %84,21 doğrulukla tahmin etmekte olduğunu göstermişlerdir [8]. Muhammad ve arkadaşları, karar ağacı modelinin %94,99 ile en yüksek doğruluğa sahip olduğunu göstermişlerdir [9].

Çizelge 4.4. Çalışmamızın Çoklu Sınıflandırılmış Veri Setini Kullanan Diğer Çalışmalarla Kıyaslanması

	Doğruluk (%)	MCC
Çalışmamız	98,46	0,98
Chen ve arkadaşları	93,8	0,92

Çalışmamızda çoklu sınıflandırılmış veri setini kullanarak geliştirilen modelin sınıflandırma sonucunda en yüksek değeri Bayes Ağı, Naif Bayes, ÇKA ve Doğrusal DVM'nin verdiği gözlemlenmiştir. Bayes Ağı, Naif Bayes, ÇKA ve Doğrusal DVM için sınıf doğruluğu %98,46 ve MCC 0,979 ile en yüksek değerler elde edilmiştir. Aynı şekilde çoklu sınıflandırılmış veri setini kullanan Chen ve arkadaşları, en yüksek değeri DVM algoritması ile % 93,08 sınıf doğruluğu ve 0,92 MCC değerleri elde etmişlerdir [6].

5. SONUÇLAR

Kovid-19'un teşhis ve tedavi seyrinde, kandan elde edilen gen ekspresyon profil verileri biyobelirteç olarak kullanılabilir. Kovid-19 için, kanda, diğer vücut sıvılarında ya da dokularda bulunan, hastalığın işareti olan, normal veya anormal bir sürece tepki veren biyolojik moleküller olan biyobelirteçler belirlenebilmektedir. Biyobelirteçlerin makine öğrenmesi öğrenmesi yöntemleri ile in silico ortamda tahmin edilmesi Kovid-19 tedavisi açısından kritik öneme sahiptir. Makine öğrenmesi yöntemleri ile *in silico* ortamda yapılan çalışmalar laboratuvar şartlarına göre zaman ve mekan açısından üstünlüklere sahiptirler.

Bu tez çalışmasında, Kovid-19 hastalığının teşhisinde önemli rol oynayabilecek biyobelirteçlerinin tahmini için iki makine öğrenmesi modeli geliştirdik. GEO'dan elde ettiğimiz sağlıklı ve Kovid-19 hastalarına ait kan ekspresyon verilerini içeren transkriptomik veriler üzerinde modellerimizi test ettik. Test işlemini verilerin ikili ve çoklu sınıflandırmaya tabi tuttuk. Geliştirdiğimiz iki makine öğrenmesi modelleri ile en iyi sonuçları elde etik.

İkili sınıflandırma için geliştirdiğimiz model, Bayes Ağları algoritması ile %98,46 sınıf doğruluğu, 0,963 kesinlik, 1 duyarlılık, 0,999 AUC, 0,969 MCC, 0,981 F-skor ve 0,968 kappa değerlerini vermiştir. Bu değerler yüksek başarım göstermişlerdir.

Çoklu sınıflandırma için geliştirdiğimiz ikinci model de Bayes Ağları, Naif Bayes, ÇKA algoritmaları ile literatürdeki en iyi başarımı %98,46 sınıf doğruluğu, 0,979 MCC, 0,981 F-skor ve 0,979 kappa değerleri ile vermiştir.

Gelecekte yapılacak çalışmalarda, biyobelirteçlerin tahmini başarımını artırmak için birleştirilmiş sınıflandırıcılar kullanmayı planlamaktayız. Ayrıca makine öğrenmesi modellerimizin çevrim içi uygulanabileceği bir web sunucu geliştirmeyi hedeflemekteyiz.

KAYNAKLAR

[1] Sevli, Onur, and Vesile Gül BAŞER GÜLSOY. "Covid-19 Salgınına Yönelik Zaman Serisi Verileri ile Prophet Model Kullanarak Makine Öğrenmesi Temelli Vaka Tahminlemesi." Avrupa Bilim ve Teknoloji Dergisi 19 (2020): 827-835.

[2] Kwekha-Rashid, Ameer Sardar, Heamn N. Abduljabbar, and Bilal Alhayani. "Coronavirus disease (COVID-19) cases analysis using machine-learning applications." Applied Nanoscience (2021): 1-13.

[3] Velavan, Thirumalaisamy P., and Christian G. Meyer. "The COVID-19 epidemic." Tropical medicine & international health 25.3 (2020): 278.

[4] Dong, Ensheng, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time." The Lancet infectious diseases 20.5 (2020): 533-534.

[5] Rothan, Hussin A., and Siddappa N. Byrareddy. "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak." Journal of autoimmunity 109 (2020): 102433.

[6] Chen, L., Li, Z., Zeng, T., Zhang, Y. H., Feng, K., Huang, T., & Cai, Y. D. (2021). Identifying COVID-19-Specific Transcriptomic Biomarkers with Machine Learning Methods. BioMed Research International, 2021.

[7] Dairi, Abdelkader, et al. "Comparative study of machine learning methods for COVID-19 transmission forecasting." Journal of Biomedical Informatics 118 (2021): 103791.

[8] Arpaci, Ibrahim, et al. "Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms." Multimedia Tools and Applications 80.8 (2021): 11943-11957.

[9] Muhammad, L. J., et al. "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset." SN computer science 2.1 (2021): 1-13.

[10] Zhang, Yu-Hang, et al. "Identifying transcriptomic signatures and rules for SARS-CoV-2 infection." Frontiers in Cell and Developmental Biology 8 (2021): 1763.

[11] Kukar, Matjaž, et al. "COVID-19 diagnosis by routine blood tests using machine learning." Scientific reports 11.1 (2021): 1-9.

[12] McClain, Micah T., et al. "Dysregulated transcriptional responses to SARS-CoV-2 in the periphery." Nature communications 12.1 (2021): 1-8.

[13] Abed, Mazin, et al. "A comprehensive investigation of machine learning feature extraction and classification methods for automated diagnosis of COVID-19 based on X-ray images." Computers, Materials, & Continua (2021): 3289-3310.

[14] Zoabi, Yazeed, Shira Deri-Rozov, and Noam Shomron. "Machine learning-based prediction of COVID-19 diagnosis based on symptoms." npj digital medicine 4.1 (2021): 1-5.

[15] Kassania, Sara Hosseinzadeh, et al. "Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach. " Biocybernetics and Biomedical Engineering 41.3 (2021): 867-879.

[16] Sujath, R., Jyotir Moy Chatterjee, and Aboul Ella Hassanien. "A machine learning forecasting model for COVID-19 pandemic in India." Stochastic Environmental Research and Risk Assessment 34 (2020): 959-972.

[17] Patterson, Bruce K., et al. "Immune-based prediction of COVID-19 severity and chronicity decoded using machine learning." Frontiers in immunology 12 (2021): 2520.

[18] Strimbu, Kyle, and Jorge A. Tavel. "What are biomarkers?." Current Opinion in HIV and AIDS 5.6 (2010): 463.

[19] Sönmezer, Meliha Çağla, and Necla Tülek. "Bakteriyel infeksiyonlarda ve sepsiste biyobelirteçler." Klimik Dergisi 28.3 (2015): 96-102.

[20] Bozkaya, Tijen Alkan. "Kalp cerrahisi sonrasında organ hasarının erken belirteçleri olarak biyo-belirteçler." Clinical and Experimental Health Sciences 5.1 (2015): 65-74.

[21] Mayeux, Richard. "Biomarkers: potential uses and limitations." NeuroRx 1.2 (2004): 182-188.

[22] Palmer, Chana, et al. "Cell-type specific gene expression profiles of leukocytes in human peripheral blood." BMC genomics 7.1 (2006): 1-15.

[23] Engel, Jasper, et al. "Breaking with trends in pre-processing?." TrAC Trends in Analytical Chemistry 50 (2013): 96-106.

[24] Oğuzlar, Ayşe. "Veri ön işleme." Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 21 (2003).

[25] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." Computers & Electrical Engineering 40.1 (2014): 16-28.

[26] SAĞBAŞ, Ensar Arif, Osman GÖKALP, and U. Ğ. U. R. Aybars. "Yüz İfadesi Tanıma için Mesafe Oranlarına Dayalı Öznitelik Çıkarımı ve Genetik Algoritmalar ile Seçimi." Veri Bilimi 2.1 (2019): 19-29.

[27] Öznur, İ. Ş. Ç. İ., and Serdar Korukoğlu. "Genetik algoritma yaklaşımı ve yöneylem araştırmasında bir uygulama." Yönetim ve Ekonomi: Celal Bayar Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 10.2 (2003): 191-208.

[28] Kramer, Oliver. "Genetic algorithms." Genetic algorithm essentials. Springer, Cham, 2017. 11-19.

[29] de Almeida, Bruno Seixas Gomes, and Victor Coppo Leite. "Particle swarm optimization: A powerful technique for solving engineering problems." Swarm Intelligence-Recent Advances, New Perspectives and Applications (2019).

[30] AYDIN, İlhan, Mehmet Umut SALUR, and Fatma BAŞKAYA. "Duygu Analizi için Çoklu Populasyon Tabanlı Parçacık Sürü Optimizasyonu." Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi 11.1 (2018): 52-64.

[31] AYDIN, İlhan, and Büşran AŞICI. "İnsan Hareketlerinin Tanınması için Parçacık Sürü Optimizasyonu Tabanlı Topluluk Sınıflandırıcı Yöntemi." Fırat Üniversitesi Mühendislik Bilimleri Dergisi 32.2 (2020): 381-390.

[32] Felner, Ariel, Sarit Kraus, and Richard E. Korf. "KBFS: K-best-first search." Annals of Mathematics and Artificial Intelligence 39.1 (2003): 19-39.

[33] El Baz, Didier, et al. "Parallel best-first search algorithms for planning problems on multi-core processors." The Journal of Supercomputing (2021): 1-30.

[34] Murat, G. Ö. K. "Makine öğrenmesi yöntemleri ile akademik başarinin tahmin edilmesi."Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji5.3 (2017): 139-148.

[35] Webb, Geoffrey I., Eamonn Keogh, and Risto Miikkulainen. "Naïve Bayes." Encyclopedia of machine learning 15 (2010): 713-714.

[36] Koşan, Muhammed Ali, Oktay YILDIZ, and Hacer Karacan. "Kimlik avı web sitelerinin tespitinde makine öğrenmesi algoritmalarının karşılaştırmalı analizi." Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi 24.2 (2018): 276-282.

[37] Solmaz, Ramazan, Mücahid Günay, and Ahmet Alkan. "Fonksiyonel Tiroit Hastalığı Tanısında Naive Bayes Sınıflandırıcının Kullanılması." Akademik Bilişim Konferansı, Mersin (2014): 891-897.

[38] KAYA, Alkan, and Aslı UYAR. "YAPAY ÖĞRENME METODLARI KULLANILARAK PETROL ÜRÜNLERINDE KALİTENİN ARTTIRILMASI."

[39] HIZLISOY, Serhat, and Zekeriya TÜFEKCİ. "Türkçe Müzikten Duygu Tanıma." Avrupa Bilim ve Teknoloji Dergisi: 6-12.

[40] ELBAHADIR, Hamza, and Ebubekir ERDEM. "Kablosuz Algılayıcı Ağlarda Hibrit Saldırı Tespit Sistemi Geliştirme." Computer Science Special: 162-174.

[41] Öztürk, Övünç, Didem Abidin, and Tuğba Özacar. "Using classification algorithms for Turkish music makam recognition." Selçuk Üniversitesi Mühendislik, Bilim Ve Teknoloji Dergisi 6.3 (2018): 377-393.

[42] Parsania, Vaishali, Navneet Bhalodiya, and N. N. Jani. "Applying Naïve bayes, BayesNet, PART, JRip and OneR algorithms on hypothyroid database for comparative analysis." (2014).

[43] Küçük, Hanife, Cengiz Tepe, and İlyas Eminoğlu. "Classification of EMG signals by k-Nearest Neighbor algorithm and Support vector machine methods."2013 21st Signal Processing and Communications Applications Conference (SIU). IEEE, 2013.

[44] Samuk, Doğan Can, and Fatih Mehmet Nuroğlu. "Seri kapasitör içeren şebekelerde k-en yakın komşuluk (k-EYK) sınıflandırma yöntemi kullanılarak geniş bölge izleme tabanlı yeni bir arızalı hat belirleme algoritması." Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi 36.2 (2021): 871-882.

[45] Tan, Songbo. "An effective refinement strategy for KNN text classifier." Expert Systems with Applications 30.2 (2006): 290-298.

[46] Breiman, L., "Random Forests", Machine Learning, Cilt 45, No 1, 5-32, 2001.

[47] Okumuş, Hatice, and Önder Aydemır. "Random forest classification for brain computer interface applications." 2017 25th Signal Processing and Communications Applications Conference (SIU). IEEE, 2017.

[48] Akar, Özlem, and Oğuz Güngör. "Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması." Jeodezi ve Jeoinformasyon Dergisi 1.2 (2012): 139-146.

[49] SAYGIN, Emre, and Muhammet BAYKARA. "Karaciğer Yetmezliği Teşhisinde Özellik Seçimi Kullanarak Makine Öğrenmesi Yöntemlerinin Başarılarının Ölçülmesi." Fırat Üniversitesi Mühendislik Bilimleri Dergisi 33.2 (2021): 367-377.

[50] Coşar, Mustafa, and Emre Deniz. "Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi." Avrupa Bilim ve Teknoloji Dergisi 28 (2021): 1112-1116.

[51] Connelly, Lynne. "Logistic regression." Medsurg Nursing 29.5 (2020): 353-354.

[52] AKSOY, Barış. "Finansal Tablo Hileleri'nin Makine Öğrenmesi Yöntemleri ve Lojistik Regresyon Kullanılarak Tahmin Edilmesi: Borsa İstanbul Örneği." Maliye ve Finans Yazıları 115: 79-104.

[53] Girginer, Nuray, and Bülent Cankuş. "Tramvay yolcu memnuniyetinin lojistik regresyon analiziyle ölçülmesi: Estram örneği." Yönetim ve Ekonomi: Celal Bayar Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 15.1 (2008): 181-193.

[54] AYDIN, Can. "Makine öğrenmesi algoritmaları kullanılarak itfaiye istasyonu ihtiyacının sınıflandırılması." Avrupa Bilim ve Teknoloji Dergisi 14 (2018): 169-175.

[55] Ahmad, Iftikhar, et al. "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection." IEEE access 6 (2018): 33789-33795.

[56] Yakut, Yöntemleriyle Borsa Endeksi Tahmini Yr, Emre YAKUT, and Selahattin Yavuz. "Yapay Sinir Ağlari Ve Destek Vektör Makineleri Yöntemleriyle Borsa Endeksi Tahmini." Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 19.1 (2014): 139-157.

[57] Noble, William S. "What is a support vector machine?." Nature biotechnology 24.12 (2006): 1565-1567.

[58] Car, Zlatan, et al. "Modeling the spread of COVID-19 infection using a multilayer perceptron." Computational and mathematical methods in medicine 2020 (2020).

[59] SABANCI, Kadir. "Kentsel Alanın Arazi Örtüsünün Veri Madenciliği Algoritmaları ile Sınıflandırılması." Honor Committee: 471.

[60] Gök, Murat. "A novel machine learning model to predict autism spectrum disorders risk gene." Neural Computing and Applications 31.10 (2019): 6711-6717.

[61] Dalianis, Hercules. "Evaluation metrics and evaluation."Clinical text mining. Springer, Cham, 2018. 45-53.

[62] Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment." IEEE Access 9 (2021): 78368-78381.

[63] Abdullah, A. L. A. N., and Murat KARABATAK. "Veri Seti-Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi." Fırat Üniversitesi Mühendislik Bilimleri Dergisi 32.2 (2020): 531-540.

ÖZGEÇMİŞ

Lisans mezuniyetimi Harran Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümünde tamamladım. Yalova Üniversitesi Bilgisayar Mühendisliği Bölümünde 2019 yılında Yüksek Lisans Eğitimime başladım. Aynı zamanda 2017 yılında beri özel sektörde çalışmaktayım

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

[1] Hatice Yıldız, Murat Gök, (2022). "*Prediction of Covid-19 Disease Using Transcriptomic Biomarkers Data in Silico*" 5th International Conference on Data Science and Applications 2022, Fethiye, Türkiye.