Fırat Üniversitesi Müh. Bil. Dergisi
36(2), 957-963, 2024

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1535830

COVID Geliştirme Duyarlılığını Tahmin Etmek için Makine Öğrenimi Modellerinin Geliştirilmesi

Zeynep Ertem[*1]
[1] School of Industrial and Systems Engineering, Watson School of Engineering, State University of New York Binghamton, Binghamton, USA
[*1] zeynep@binghamton.edu

(Geliş/Received: 19/08/2024; Kabul/Accepted: 18/09/2024)

Öz: COVID-19 genomiklerinin karmaşıklıklarını çözmek son derece önemli bir sorundur. Virüsün genetik yapısında meydana gelen mutasyonlar, ilerlemesini ve semptomatolojisini doğal olarak öngörülemez kılmaktadır. Özellikle, "Uzun COVID" terimi, COVID-19'un kalıcı sonuçlarını belirtmek için ortaya çıkmış olup, etiyolojisini anlamak için yoğun çabaları tetiklemiştir. Devam eden çalışmalar, Uzun COVID'i ve belirleyicilerini titizlikle araştırmaktadır. Yapay zekâ (YZ) ve makine öğrenimi (MO) bu amaçla vazgeçilmez varlıklar olarak ortaya çıkmış olup, COVID-19 krizi ortamında hastalık duyarlılığını açıklığa kavuşturma ve öngörme konusunda dikkate değer etkinlik sergilemektedirler. Bu çerçevede, çabamız, Uzun COVID'in başlangıç olasılığını öngörmek için MO metodolojilerini kullanmaya yöneliktir. Bu amaçla, birden fazla MO modeli titizlikle eğitilmiştir. Ampirik bulgular, en yetkin modelin Uzun COVID'in meydana gelme olasılığını tahmin etmede takdir edilecek bir doğruluk oranı olan %80'e ulaştığını ortaya koymaktadır.

Anahtar kelimeler: Sınıflandırma, özellik mühendisliği, özellik seçimi, makine öğrenimi.

Development of Machine Learning Models to Predict the Susceptibility of Developing COVID

Abstract: Unraveling the intricacies of COVID-19 genomics is a very important problem. The mutations occurring within the virus's genetic makeup render its progression and symptomatology inherently unpredictable. Notably, the term "Long COVID" has surfaced to delineate the enduring repercussions of COVID-19, prompting concerted efforts to comprehend its etiology. Ongoing studies are meticulously investigating Long COVID and its determinants. Artificial intelligence (AI) and machine learning (ML) have emerged as indispensable assets in this pursuit, demonstrating remarkable efficacy in elucidating underlying factors and forecasting disease susceptibility amidst the COVID-19 crisis. Within this framework, our endeavor aims to harness ML methodologies to prognosticate the likelihood of Long COVID onset. Multiple ML models have been meticulously trained for this purpose. The empirical findings reveal that the most proficient model attains a commendable accuracy rate of 80% in predicting Long COVID occurrence.

Key words: Classification, feature engineering, feature selection, machine learning.

1. Giriş

Son zamanlarda, makine öğrenimi uygulamalarının yaygınlaşması ve hasta verilerinin artan erişilebilirliği ile, sağlık alanında makine öğrenimi tekniklerinin önemi profesyoneller arasında sürekli olarak artmakta ve bu alanın geleceği üzerinde önemli bir etki yapmaktadır. Önceden belirlenmiş algoritmalar, matematiksel fonksiyonlar ve istatistiksel analizlere dayanarak, makine öğrenimi hasta verilerini kullanarak sonuçlar, sonuçlar ve öneriler üretir. Sağlık alanı, farklı bireyler arasında geniş bir hastalık ve semptom yelpazesi nedeniyle kazalara ve hatalara maruz kalmaktadır. Son yıllarda, makine öğrenimi teşhis ve hastalık ilerlemesini tahmin etme, klinik karar destek sistemlerinde ve optimal hasta bakımında önemli ilerlemeler kaydetmiştir.

Makine öğreniminin sağlık alanındaki rolünü, özellikle son COVID-19 pandemisi gibi sağlık krizleri sırasında, uygulamasının giderek daha hayati hale geldiği durumlarda incelemektedir. Kritik halk sağlığı sorunlarını, özellikle son pandemi, ele almak için çeşitli makine öğrenimi tekniklerini modelledik, makine öğrenimi modellerinin sağlık hizmetlerini iyileştirmedeki ve küresel sağlık krizlerinin etkilerini hafifletmedeki etkinliğini göstermeyi amaçladık. Makine öğrenimi terimi, önceden belirlenmiş matematiksel fonksiyonlar ve istatistiksel analizlere dayanan ve hasta verilerini giriş olarak kullanan yapay zekâ kategorisini ifade eder. Makine öğrenimi algoritmaları dört kategoriye ayrılır: denetimli, denetimsiz, yarı-denetimli ve pekiştirmeli öğrenme [1]. Makine öğrenimindeki son gelişmeler, sağlık hizmetleri uygulama ve karar verme alanlarında çeşitli ilerlemeler kaydedilmesine yardımcı olmuştur. Örneğin, COVID-19 pandemisi sırasında, virüsün yayılmasını sınırlamak için önemli çabalar gösterilmiştir; ancak bu önlemlere rağmen pek çok hayat kaybedilmiş ve ekonomik etkiler meydana

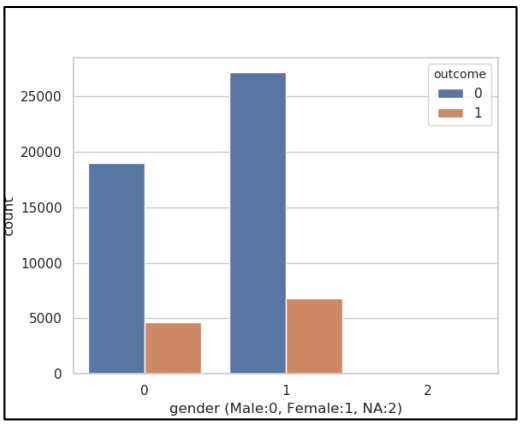
[*] Sorumlu yazar: zeynep@binghamton.edu. Yazarların ORCID Numarası: [1] 0000-0003-0632-0905

gelmiştir. Bu tür koşullarda, makine öğreniminin COVID-19'u izleme ve tanımlamada kullanılması, onun kontrol altına alınmasına yardımcı olabilir.

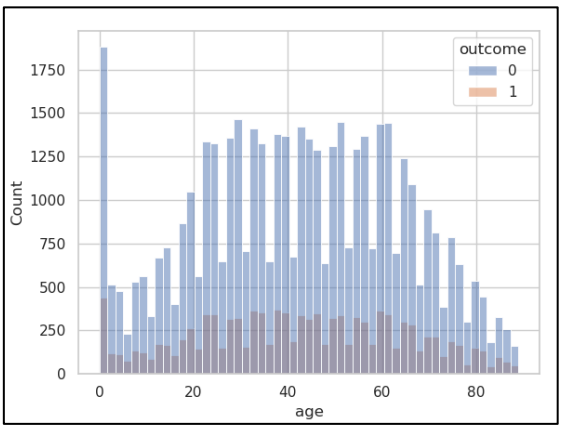
COVID-19 pandemisi hayatlarımızı dramatik bir şekilde etkiledi. COVID-19'un genomlarını anlamak ve pandeminin etkilerini hafifletmek için sürekli ve önemli çabalar olmasına rağmen, virüsün genlerinde meydana gelen mutasyonlar, gelişimini ve semptomlarını öngörülemez hale getiriyor. Bilim insanları, virüsün gelişimini izleyerek her mutasyondaki yayılma düzeyini ve potansiyel gelişmiş semptomları belirlemeye çalışıyor. COVID-19'un gelişimi ve mutasyonlarının yanı sıra, bu virüsün enfeksiyonu uzun vadeli etkiler yaratabilir; buna post-akut SARS-CoV-2 sekelleri (PASC) ya da uzun COVID denir. PASC/Uzun COVID semptomları göğüs ağrısı, öksürük, nefes almada zorluk, döküntü ve diğerlerini içerebilir. Ayrıca, son raporlar PASC/Uzun COVID'in semptom çeşitliliği içerdiğini göstermektedir. Şu anda PASC/Uzun COVID'in genel yayılma durumu kesin olarak bilinmemektedir. Ancak, son çalışmalar COVID-19'a yakalanmış kişilerin %50'den fazlasının en az bir PASC/Uzun COVID semptomu geliştirdiğini göstermektedir [2,3]. PASC/Uzun COVID'in yayılımını, süresini ve semptomlarını anlamak için çalışmalar devam etmektedir. Bazı çalışmalar, Uzun COVID'li hastaları tanımlamak için farklı yaklaşımlar önermiştir. Genel olarak, Uzun COVID tanısı hastalığın akut aşamasından sonra başlar ve semptomlar, kısa ve uzun COVID semptomlarını karakterize eden eski ve yeni semptomlar olmak üzere iki kategoriye ayrılır.

Yapay Zekâ (YZ), insanlığın bilişsel işlevlerini ve akıllı davranışlarını taklit eden güçlü bir araçlar yelpazesidir. YZ'nin iyi bilinen alt alanlarından biri, veri setlerinden karmaşık ilişkileri incelemek ve gizli desenleri tanımlamak için özel olarak tasarlanmış geniş bir algoritma yelpazesi içeren makine öğrenimidir. Sağlık alanında, özellikle COVID krizi sırasında, makine öğrenimi teknikleri, tıp endüstrisini ve sağlık uzmanlarını çeşitli hastalıkları tahmin etme, sağlık alanındaki gizli desenleri belirleme ve hastaları kümelendirme konusunda desteklemek için yaygın olarak uygulanmıştır. COVID pandemisi sırasında, makine öğrenimi teknikleri hastalığın gelişme olasılığını tahmin etmek ve COVID-19'un gelişimine neden olan faktörleri belirlemek için hızla uyarlanmıştır. Bu projede özel olarak, farklı makine öğrenimi tekniklerini kullanmayı amaçlıyoruz [4,5].

Bu makalenin geri kalanı şu şekilde düzenlenmiştir: Bölüm 2, COVID-19 hastalarının teşhis, tespit ve sınıflandırılmasında Yapay Zekâ ve makine öğrenimi tekniklerinin kullanılmasına yönelik yapılan farklı araştırmaları özetlemektedir. Makalenin metodolojisi Bölüm 3'te sunulmuştur. Bölüm 4, deneysel sonuçları tartışmaktadır. Son olarak, Bölüm 5 makaleyi sonuçlandırmakta ve birkaç gelecekteki yönü vurgulamaktadır.



(a) Sonuç değişkenine göre cinsiyet dağılımı



(b) Sonuç değişkenine göre yaş dağılımı

Şekil 1.

2. Materyal ve Metot

Bu bölüm önce veri setinin kısa bir tanımını sunar, ardından uzun COVID gelişme olasılığını tahmin etmek için makine öğrenimi modelleri geliştirme metodolojisini tanıtır. Ana adımlar, giriş verisi işleme, özellik mühendisliği ve seçimi, model eğitimi ve model değerlendirmesi olarak ayrılabilir. Bir sonraki alt bölümler bu adımları daha detaylı bir şekilde açıklamaktadır.

2.1. Giriş verisi işleme

Giriş verisi, National COVID Cohort Collaborative (N3C) Veri Bölgesi tarafından sağlanmaktadır. N3C platformu, COVID-19 vakalarıyla doğrulanmış 6,3 milyon hasta ve olası COVID-19 vakalarıyla 168,937 hasta olmak üzere toplamda 16 milyonun üzerinde klinik kayıt sunmaktadır. Toplamda 19 milyardan fazla veri satırı bulunmaktadır ve 75 farklı siteye yayılmaktadır [6]. N3C, uyumlu EHR (Elektronik Sağlık Kayıtları) verilerine geniş erişim ve analiz imkanları sunarak, mevcut ve gelecekteki sağlık acil durumlarını aşabilecek işbirlikçi veri paylaşımı için yenilikçi bir model sunar. N3C'nin başlıca özellikleri arasında ulusal iş birliği ve yönetim, düzenleyici stratejiler, topluluk tarafından geliştirilen fenotiplere dayalı COVID-19 kohort tanımları, dört farklı veri modelinde veri uyumlaştırması ve ABD'den toplanan verilerin yenilikçi algoritmalarla işlenmesini destekleyen bir işbirlikçi analiz platformunun geliştirilmesi yer alır. N3C, COVID-19 verileri ile topluluk destekli, yeniden üretilebilir ve şeffaf analizler sunarak, sonuçların hızla paylaşılmasını ve atomik atamaların yapılmasını teşvik eder. Ayrıca, açık bilimin EHR verileri üzerinde büyük ölçekli olarak etkili bir şekilde uygulanabileceğini gösterir. Analitik platform veya N3C Enclave, Ulusal Çevirisel Bilimler Merkezi (NCATS) tarafından yönetilen güvenli bir bulut ortamında barındırılmakta olup, ABD genelindeki çeşitli merkezlerden Ocak 2018'den itibaren N3C COVID-19 fenotip kriterlerine uyan hastalardan alınan klinik verileri içermektedir. Gizliliği koruyan kayıt bağlantıları, görüntüleme, genomik veya klinik deneme verileri gibi diğer veri setleri ile ek düzenleyici onaylarla ilişkilendirilmek üzere geliştirilecektir. Ayrıca, N3C, algoritmalar kullanılarak türetilmiş sentetik veri setlerinin oluşturulmasını pilot olarak gerçekleştirecektir. N3C verileri, araştırmacılara COVID-19 ile ilgili geniş çaplı analizler yapma fırsatı sunar [13].

Her bir site, hasta demografisi, tıbbi geçmişi ve laboratuvar ölçüm sonuçları gibi bilgileri temsil eden Kişi, Gözlem ve Ölçüm tabloları gibi tekil tablolardan oluşmaktadır. Bu adımda, şekilde belirtilen veri sitelerini dahil ettik ve COVID-19 aşı durumu, hastalık durumları ve ilaç geçmişi gibi bazı temel koşullara dayalı olarak tabloları kademeli olarak birleştirmeye başladık.

2.2. Özellik mühendisliği ve seçimi

Giriş tabloları işlendikten sonra, 51 yeni özellik çıkarılmıştır. Bu özellikler, yaklaşık 60,000 hastanın yaş ve cinsiyet gibi bazı statik bilgilerini, Kardiyomiyopatiler ve Diyabet gibi çeşitli hastalık göstergelerini ve antibiyotikler ve antiviral ilaçlar gibi hastaların ilaç geçmişini içermektedir. Bu özelliklerin çoğu, bazı hastalıklar veya durumlar için ikili değerler alır. Bu adımın amacı, modelimize mümkün olduğunca ilgili bilgi sağlamaktır. Ayrıca, modelimizin hastaların statik bilgileri, sağlık ve hastalık durumu ve ilaç geçmişinin Uzun COVID hastalığı geliştirme olasılığı üzerindeki etkilerini yakalamasını hedeflemekteyiz.

2.3. Model eğitimi ve değerlendirme

Model eğitimi aşamasında, Uzun COVID olasılığını tahmin etmek için çeşitli makine öğrenimi modelleri oluşturulmuştur. Problemin doğası, konuları post-akut semptomlar geliştirip geliştirmeyecekleri olarak sınıflandırmak olduğundan, denetimli öğrenme algoritmalarına, özellikle sınıflandırma algoritmalarına odaklandık. Kullandığımız algoritmalar şunlardır: Destek Vektör Makineleri (SVM), Lojistik Regresyon sınıflandırıcısı (LR) ve Karar Ağacı Sınıflandırıcısı (DT). Bu algoritmalar ayrıca giriş verilerindeki değişkenlerin önemini de belirtebilir. Bu algoritmalar makine öğrenmesinde en sık kullanılan tekniklerdir.

D**estek Vektör Makinesi** (SVM), [14,16] sınıflandırma ve regresyon görevleri için kullanılan denetimli bir makine öğrenimi algoritmasıdır. SVM'nin temel fikri, farklı sınıflara ait veri noktalarını maksimum marj ile ayıran optimal hiper düzlemi bulmaktır.

- **Hiper Düzlem:** n-boyutlu bir uzayda, hiper düzlem n-1 boyutunda düz bir afine alt uzaydır. Örneğin, 2-boyutlu bir uzayda hiper düzlem, farklı sınıfları ayıran bir doğrudur.
- **Marj:** Marj, hiper düzlem ile her sınıftan en yakın veri noktaları (destek vektörleri olarak bilinir) arasındaki mesafedir. Bu marjın maksimize edilmesi, modelin genelleme kapasitesini artırır.
- **Kernel Kurnazlığı:** SVM, doğrusal olarak ayrılamayan verileri daha yüksek boyutlu bir uzaya dönüştürerek bu verilerle daha etkili bir şekilde başa çıkabilir. Bu dönüşüm, polinom, radyal temel fonksiyon (RBF) ve sigmoid gibi Kernel fonksiyonları kullanılarak gerçekleştirilir.

Lojistik Regresyon [15, 17], sonucu iki olası sınıftan biri olan ikili sınıflandırma problemleri için kullanılan istatistiksel bir yöntemdir. Verilen bir girişin belirli bir sınıfa ait olma olasılığını lojistik fonksiyon kullanarak tahmin eder.

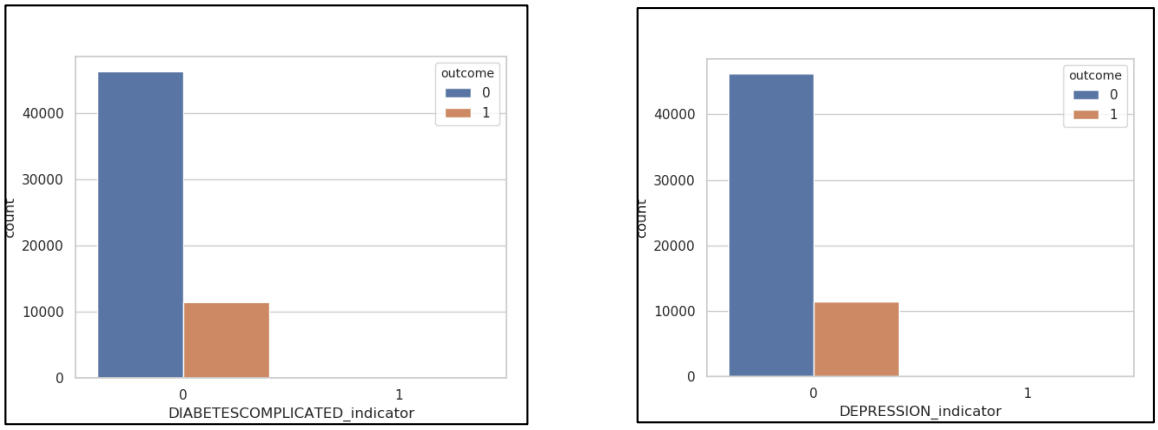
Karar Ağaçları [16, 19, 20], sınıflandırma ve regresyon görevleri için kullanılan parametrik olmayan bir denetimli öğrenme yöntemidir. Bir karar ağacı modeli, veri kümesini giriş özelliklerinin değerlerine göre alt kümelere ayırarak, her iç düğümün bir özellik testini, her dalın bir test sonucunu ve her yaprak düğümün bir sınıf etiketini veya regresyon değerini temsil ettiği ağaç benzeri bir yapı oluşturur. Özetle, **SVM** sınıfları ayırmak için maksimum marj ile bir hiper düzlem bulmaya odaklanır, **Lojistik Regresyon** lojistik fonksiyon kullanarak olasılıkları tahmin eder ve genellikle ikili sınıflandırma için kullanılır, **Karar Ağaçları** ise kararların ve olası sonuçlarının ağaç benzeri bir modelini sunar. Her teknik kendi güçlü yönlerine sahiptir ve veri ve sorunların özel özelliklerine göre seçilir.

Modellerin performansını değerlendirmek için birkaç ölçüt kullandık: doğruluk, kesinlik, özgüllük, duyarlılık ve F1 skoru [7]. Doğruluk, modelin genel tahmin gücünü ölçer. Ancak, her zaman modelin performansını iyi bir gösterge olmayabilir. Bu nedenle, Uzun COVID modellerini doğru bir şekilde değerlendirmek için ek metrikler göz önünde bulundurduk. Yani, duyarlılığı, özgüllüğü, kesinliği ve F1 skorunu dikkate aldık.

3. Bulgular

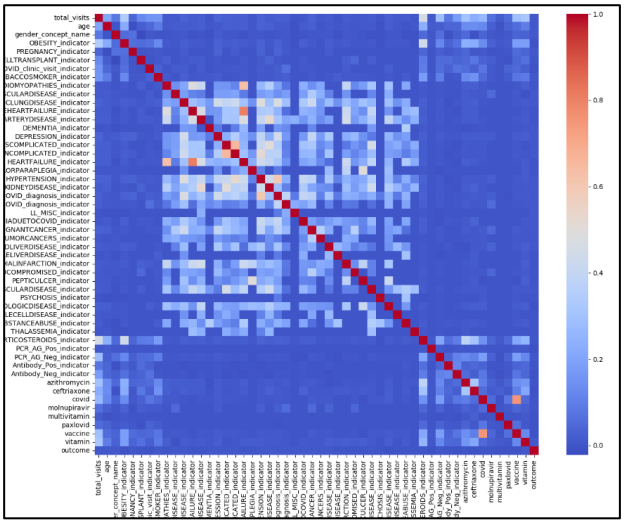
Tahmin modellerimizin deneysel sonuçları bu bölümde sunulmuştur. İlk olarak, giriş özelliklerinin sonuç değişkeni üzerindeki etkisini incelemek için bazı açıklayıcı veri analizleri gerçekleştirdik. Ardından, modellerin performansını daha önce belirtilen performans ölçütleri kullanarak değerlendirdik.

Şekil 1, cinsiyetin sayısını ve yaşın dağılımını sonuç değişkeni açısından göstermektedir. Şekilden de görebileceğimiz gibi, verilerde erkeklerden daha fazla kadın bulunmaktadır ve bu durum, Uzun COVID tanısı konan kadın sayısının erkeklerden fazla olduğunu göstermektedir. Yaş açısından, çok genç yaş grubunun (2 yaşa kadar) diğer yaş aralıklarına göre baskın olduğunu görebiliyoruz. Ancak, bu iki özelliğin gerçekten sonuç değişkenini etkileyip etkilemediğini henüz çıkaramıyoruz.



Şekil 2. (a) Sonuç değişkenine göre şeker hastalığı dağılımı (b) Sonuç değişkenine göre depresyon dağılımı.

Şekil 2, şeker hastalığı ve depresyon hastalık göstergelerinin sonuç değişkenine göre dağılımını göstermektedir. Şekilde de belirtildiği gibi, yüksek veri dengesizliği nedeniyle herhangi bir sonuca varamıyoruz. Şekil 3, tüm değişkenler arasındaki korelasyon grafiğini göstermektedir. Bazı hastalık göstergelerinin yüksek korelasyon gösterdiği açıktır. Ancak, bizim esas ilgilendiğimiz nokta, giriş özellikleri ile sonuç değişkenleri arasındaki korelasyonu yakalamaktır. Verinin büyük boyutu ve veri dengesizliği problemi nedeniyle bu etkileri yakalayamıyoruz ve bu, şekilde de açıkça görülmektedir.



Şekil 3. Korelasyon grafiği.

Verilerimiz, Uzun COVID hastalarının sayısı açısından yüksek derecede dengesiz olduğu için doğruluk, modellerin performansını iyi bir gösterge olarak kabul edilemez. Bu nedenle, Uzun COVID hastalarını tanımlama açısından modelin performansını daha doğru bir şekilde açıklamak için duyarlılık, özgüllük, kesinlik ve F1 skoru da kullanılmıştır. Tablo 1, Uzun COVID modellerinin performansını özetlemektedir. Eğitim veri seti açısından, DT en iyi performansı sergileyerek %85 geri çağırma (Recall) oranı elde etmiştir. Test veri seti açısından ise, %80 geri çağırma oranı ile SVC en iyi performansı göstermiştir.

Tablo 1. Performans Değerlendirmesi.

		Doğruluk	Kesinlik	Geri Çağırma (Recall)	F1 Skoru
Özellik seçimi olmadan	**Lojistik Regresyon**	0,8	0,69	0,49	0,54
	Karar Agaçlari	0,85	0,89	0,85	0,86
	Karar agaci siniflandirmasi (SVC)	0,82	0,8	0,82	0,86
Özellik seçimli	**Lojistik Regresyon**	0,48	0,68	0,48	0,54
	Karar Agaçlari	0,62	0,69	0,62	0,65
	Karar agaci siniflandirmasi (SVC)	0,8	0,67	0,8	0,72

4. Tartışma

Son araştırmalara göre [8, 9], COVID-19 vakalarının tespiti ve uç bilgi işlem alanında yürütülen üç farklı çalışma akışı bulunmaktadır. İlk akış, tıp görüntüleme analizi ile COVID-19'u tanımlamak için içgörümü desenler çıkarmaya çalışan radyologlar için yararlı olan görüntü verilerinin analizini içermektedir. İkinci akış, COVID-19'un yeni fenotip sınıflandırmasının tespitini, tanı koymayı ve ölüm oranını nasıl minimize edebileceğine dair içgörüler sağlayan modeller geliştirmeyi kapsamaktadır. Üçüncü akış ise, özellik çıkarımı ve derin öğrenme tekniklerini kullanan çerçeveler aracılığıyla COVID-19'un otomatik sınıflandırılması için teknikleri içermektedir.

[10] araştırmacıları, COVID-19 hastalarının sonunda ventilasyon ihtiyacı olup olmayacağını tahmin etmek için bir lojistik regresyon sınıflandırma modeli kullanmıştır. Geliştirilen metodoloji, çok merkezli klinik ortamlarda uygulanmış ve COVID-19 hastalarının 24 saatlik bir süre içinde invaziv mekanik ventilasyon ihtiyaçlarını tahmin etme doğruluğuna göre değerlendirilmiştir. Deney, COVID-19 tanısı almış ve beş farklı Amerika Birleşik Devletleri sağlık sistemine kabul edilmiş 197 hastayı içermektedir. Araştırmacılar, geliştirilen modeli (MLA) Modified Early Warning Score (MEWS) adı verilen bir erken uyarı sistemi ile karşılaştırmıştır. Deneysel sonuçlar, makine öğrenme algoritmasının (MLA) MEWS'e göre %78 olan duyarlılık oranını %90'a çıkardığını göstermiştir. Ayrıca, MLA %58 spesifiklik oranı elde etmişken, MEWS'de bu oran %40'tır.

COVID-19'daki makine öğreniminin bir diğer yönünde, [11] araştırmacıları, COVID-19 tanısı almış veya şüphelenilen hastaların entübasyon gerektirip gerektirmeyeceğini tahmin etmek için bir derin öğrenme modeli geliştirmiştir. Çalışma, 2020 yılında beş farklı hastaneye kabul edilen 4087 hastayı kapsamaktadır. Bu hastalardan %11'i entübasyon gerektirmiş ve bu durum dengesiz sınıflara yol açmıştır. Hastanelerde ventilatör eksikliği nedeniyle, çalışma sağlık çalışanlarının risk yönetimini iyileştirmek ve entübasyon gerektirecek COVID-19 hastalarını tahmin etmek için bir makine öğrenme modeli geliştirmeyi hedeflemiştir. Geliştirilen makine öğrenme modeli, mekanik ventilasyon tahmini için iyi bir şekilde yerleşmiş bir araç olan ROX indeksi ile karşılaştırılmıştır. Deneysel sonuçlar, geliştirilen modelin ROX indeksini geride bıraktığını ve geliştirilen modelin 0,84 AUC, ROX indeksinin ise 0,64 AUC elde ettiğini göstermiştir. Ayrıca, deneysel sonuçlar, geliştirilen modelin precision-recall eğrisinin 0,30, ROX indeksinin ise 0,13 olduğunu ortaya koymuştur.

PASC/Uzun COVID açısından, uzun COVID gelişme olasılığını tahmin etmek için makine öğrenme modellerinin geliştirilmesini teşvik eden araştırmalar henüz gelişmektedir. NIH, bu özel amaç için bir yarışma başlatmıştır. NCATS Ulusal COVID Kohort İş birliği (N3C) Veri Bölgesi, COVID-19 ve PASC/Uzun COVID hastalarını içeren veri setini sağlayacaktır [6]. N3C Veri Bölgesi, merkezi, güvenli ve ulusal klinik veriler sunan bir platformdur. Ayrıca, N3C Veri Bölgesi, COVID-19 risk faktörlerini ve uzun vadeli sonuçlarını incelemeyi amaçlayan araştırma topluluğu için güçlü analitik araçlarla donatılmıştır. Veri henüz sağlanmamıştır ve yalnızca N3C Veri Bölgesi platformu aracılığıyla erişilebilir. Veri seti, COVID-19 hastalarına ait bilgileri, demografik veriler, işlemler, semptomlar, laboratuvar sonuçları, ilaçlar, fiziksel ölçümler ve daha fazlasını içerecektir [6]. Veri setinde 15 milyonun üzerinde hastaya ait bilgi bulunmaktadır ve bu, 17,5 milyar veri satırını temsil etmektedir. Ayrıca, veriler 5,8 milyon COVID pozitif hastayı içermekte olup, hastaların gizliliğini korumak amacıyla kimlik bilgileri çıkarılmıştır. Verinin genişlik ve derinlik karmaşıklığını karşılamak için, mevcut veri çeşitliliğinden yararlı içgörüler elde etmek için ileri düzey analitik teknikler gerekmektedir. Bu projede, ana motivasyonumuz, daha önce COVID-19 tanısı almış hastaların PASC/Uzun COVID geliştirme olasılığını belirleyebilen makine öğrenimi modelleri geliştirmektir. Geliştirilen modellerin sağlık sektöründe karar destek mekanizması olarak entegresi sağlık politikalarını geliştirecek ve hastalıklarla mücadelede daha öngörülü bir politika sağlayacaktır.

5. Sonuçlar

Bu makale ile Makine öğreniminin sağlık hizmetlerini devrim niteliğinde değiştirmedeki anahtar rolünü, özellikle COVID-19 pandemisi gibi sağlık krizleri bağlamında ortaya koyuyoruz. Sonuç olarak, Yapay Zekâ ve makine öğrenimi, COVID krizi sırasında güçlü araçlar olarak kendilerini kanıtlamıştır. Bu projede, Uzun COVID hastalığını geliştirme olasılığını tahmin etmek için çeşitli makine öğrenimi modelleri geliştirdik. En iyi model, Uzun COVID'i tahmin etme konusunda %80 doğruluk oranı elde etmiştir. Ancak, hala iyileştirme için bazı alanlar bulunmaktadır. Gelecek çalışmalarda, eksik veri problemini ele almak için daha eğitimli stratejiler kullanmayı planlıyoruz. Ayrıca, özellik mühendisliği metodolojimizi daha bilgilendirici özellikleri içerecek şekilde genişletmeyi düşünüyoruz. Son olarak, veri dengesizliği sorununu ele almak için bazı teknikleri de entegre etmeyi planlıyoruz.

Sağlık krizleri sırasında kritik halk sağlığı sorunlarını ele almadaki makine öğreniminin hayati rolü incelenmektedir. Makine öğrenimi teknikleri, büyük miktarda hasta verisini analiz etme yetenekleri ve teşhis, hastalık ilerlemesi tahmini ve klinik karar destek sistemleri için değerli içgörüler üretme kapasiteleri nedeniyle sağlık alanında giderek daha fazla benimsenmiştir. Bu gelişmeler, çeşitli hastalıklar ve semptomların karmaşıklıklarıyla ilişkili hataları azaltarak ve hasta bakımını optimize ederek sağlık hizmetlerinin sunumunu geliştirmede kritik öneme sahiptir. Farklı makine öğrenimi tekniklerini modelleyerek, bu yetenekleri küresel sağlık krizleri sırasında halk sağlığı yönetimindeki gerçek hayattaki sorunları ele almak için kullanmayı amaçlamaktadır. Bu konuda yapılan çalışmalar [20] Lojistik Regresyon, Karar Ağaçları ve Karar Ağacı Sınıflandırması tekniklerinin en çok kullanılan metotlar olduğunu göstermiştir [21-24]. Bu sebeple bu 3 algoritma seçilmiş ve bu araştırma, hastalık tahmininde farklı denetimli makine öğrenimi algoritmalarının karşılaştırmalı performanslarını incelemeyi amaçladı. Klinik veriler ve araştırma kapsamı hastalık tahmini çalışmalarında geniş ölçüde farklılık gösterdiğinden, yalnızca veri seti ve kapsam üzerinde ortak bir kıyaslama kriteri belirlendiğinde karşılaştırma

yapılabilmiştir. Bu nedenle, karşılaştırma için aynı veri ve hastalık tahmini üzerinde birden fazla makine öğrenimi yöntemi uygulayan çalışmalar seçildi.

Kaynaklar

[1] Ahsan M. M., Luna S. A., Siddique Z. Machine-learning-based disease diagnosis: A comprehensive review. In Healthcare, 2022, 10: 541.

[2] Silva Andrade B, Siqueira S, de Assis Soares WR, de Souza Rangel F, Santos NO, dos Santos Freitas A, Ribeiro da Silveira P, Tiwari S, ve diğerleri. Long-COVID and post-COVID health complications: an up-to-date review on clinical conditions and their possible molecular mechanisms. Viruses, 2021; 13(4): 700.

[3] https://portal.challenge.gov/public/previews/

[4] Raveendran AV, Jayadevan R, Sashidharan S. Long COVID: An overview. Diabetes Metab Syndr 2021; 15(3): 869-875.

[5] Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, Prior F, Yu Jr F. Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. JMIR Med Inform 2021; 9(1): 23811.

[6] https://covid.cd2h.org/enclave

[7] Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput & Applic 2018; 29(10): 685-693.

[8] Kwekha-Rashid AS, Abduljabbar HN, Alhayani B. Coronavirus disease (COVID-19) cases analysis using machine-learning applications. Appl Nanosci 2021.1-13.

[9] Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. Physiological genomics, 2020; 52(4): 200-202.

[10] Burdick H, Lam C, Mataraso S, Siefkas A, Braden G, Dellinger RP, McCoy A, Vincent JL, ve diğerleri. Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. Comput Biol and Med 2020; 124: 103949.

[11] Arvind V, Kim JS, Cho BH, Geng E, Cho SK. Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. J Crit Care, 2021; 62:.25-30.

[12] Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, Payne PR, Pfaff ER, Robinson, PN, Saltz JH and Spratt H, The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association, 2021, 28(3): 427-443

[13] Vishwanathan, SVM, Murty MN, May. SSVM: a simple SVM algorithm. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02, 2002, 3: 2393-2398

[14] Kleinbaum, DG, Dietz K, Gail M, Klein, M. and Klein, M., Logistic regression, 2002, New York: Springer-Verlag.

[15] Rokach L and Maimon O. Decision trees. Data mining and knowledge discovery handbook, 2005.

[16] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Wiley; 2013.

[17] Joachims T. Making large-scale SVM learning practical. SFB 475: Komplexitätsreduktion Multivariaten Datenstrukturen, Univ. Dortmund, Dortmund, Tech. Rep. 1998.

[18] Quinlan JR. Induction of decision trees. Mach Learn. 1986, 1(1):81–106.

[19] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informat. 2006, **2**:59–77.

[20] Uddin S, Khan A, Hossain ME and Moni MA, 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1):1-16.

[21] Aneja S, Lal S. International Conference on Parallel, Distributed and Grid Computing (PDGC) 2014. Effective asthma disease prediction using naive Bayes—Neural network fusion technique.

[22] Ahmad LG, Eshlaghy A, Poorebrahimi A, Ebrahimi M, Razavi A. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013, 4(124):3.

[23] Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access. 2017, **5**:8869–8879.

[24] Yang J, Yao D, Zhan X, Zhan X. International Symposium on Bioinformatics Research and Applications. 2014. Predicting disease risks using feature selection based on random forest and support vector machine.