

T.C.
İSTANBUL ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

**COVID-19 ENFEKSİYONUNUN BİYOKİMYASAL
YANSIMALARININ DEĞERLENDİRİLMESİ**

MAİDE HACER ALAGÖZ

**TIBBİ BİYOKİMYA ANABİLİM DALI
TIBBİ BİYOKİMYA PROGRAMI**

**DANIŞMAN
PROF. DR. EVİN ADEMOĞLU**

İSTANBUL-2023

TEŞEKKÜR

Bilgi ve deneyimlerini benimle paylaşan, değerli katkılarıyla beni yönlendiren ve tezime göstermiş olduğu destek ve özen için tez danışmanım sayın Prof. Dr. Evin ADEMOĞLU'na, katkılarıyla destek olan tez izleme komitesi öğretim üyeleri sayın Doç. Dr. Ayşem KAYA ve sayın Doç. Dr. A. Fatih AYDIN'a,

Doktora eğitimime destek ve katkıda bulunan İstanbul Üniversitesi Tıbbi Biyokimya Anabilim Dalı'nın değerli öğretim üyeleri ve doktora arkadaşlarıma,

Tez çalışmalarına klinik katkıları için İstanbul Prof. Dr. Cemil Taşcıoğlu Şehir Hastanesi, Kulak Burun Boğaz Hastalıkları Kliniğinden Doç. Dr. Ayşe Enise GÖKER'e ve tezimin veri analizlere yaptığı katkıları için İstanbul Tıp Fakültesi Biyoistatistik Anabilim Dalı'ndan Dr. Öğr.Üyesi Mehmet Güven GÜNVER'e,

Tüm doktora eğitim boyunca yanımda olan, sevgi, destek ve katkılarını esirgemeyen kıymetli aileme ve sevgili eşime sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

TEŞEKKÜR.....	İİ
İÇİNDEKİLER	İİİ
TABLolar LİSTESİ.....	İİV
ŞEKİLLER LİSTESİ.....	İV
KISALTMALAR LİSTESİ	V
ÖZET	V
ABSTRACT.....	İVİİ
1. GİRİŞ VE AMAÇ.....	1
2. GENEL BİLGİLER	3
3. GEREÇ VE YÖNTEM.....	223
4. BULGULAR.....	33
5. TARTIŞMA	445
KAYNAKLAR	51
İNTİHAL RAPORU İLK SAYFASI	58

TABLolar LİSTESİ

Tablo 2.1. COVID-19’da laboratuvar parametrelerde ortaya çıkan başlıca deęişiklikler

Tablo 3.1. Sonuçların özetlendięi karışıklık matrisi

Tablo 4.1. COVID-19 tanısı ile takip edilen hastaların demografik bulguları

Tablo 4.2. COVID-19 tanısı ile takip edilen hastaların rutin biyokimya ve hemostaz testlerinin sonuçları

Tablo 4.3. Ayaktan/yatarak takip prognoz öngörüsünde etkili parametre(ler)nin belirlenmesi

Tablo 4.4. COVID-19 tanısı almış tüm hastaların mortalite öngörüsünde etkili parametre(ler)nin belirlenmesi

Tablo 4.5. Hastane içi mortalite öngörüsünde etkili parametre(ler)nin belirlenmesi

Tablo 4.6. COVID-19 tanısı almış hastaların ayaktan/yatarak takip prognoz öngörüsü için Model-1’in karışıklık matrisi

Tablo 4.7. COVID-19 tanısı almış tüm hastaların mortalite öngörüsü için Model-2’nin karışıklık matrisi

Tablo 4.8. Yatarak takip edilen COVID-19 hastalarının hastane içi mortalite öngörüsü için Model-3’ün karışıklık matrisi

ŞEKİLLER LİSTESİ

Şekil 2.1. Yapay zeka teknikleri

Şekil 2.2. Lineer (Doğrusal) Regresyon grafiği

Şekil 2.3. Lojistik Regresyon grafiği

Şekil 2.4. K-En Yakın Komşu (K-NN) grafiği

Şekil 2.5. Karar Ağacı Modeli

Şekil 2.6. SARS-CoV-2 virüsünün yapısı

Şekil 3.1. Hastaların istatistiksel sonucu

Şekil 3.2. Çalışmada kullanılacak karar ağacı algoritmasının seçimi

Şekil 4.1. COVID-19 tanısı almış tüm hastalarda ayakta/yatarak takip prognoz öngörüsünde etkili parametre(ler)nin belirlenmesinde CART karar ağacı modeli (Model-1)

Şekil 4.2. COVID-19 tanısı almış hastaların mortalite öngörüsünde etkili parametre(ler)nin belirlenmesinde CART karar ağacı modeli (Model-2)

Şekil 4.3. Hastane içi mortalite öngörüsünde etkili parametre(ler)nin belirlenmesinde CART karar ağacı modeli (Model-3)

KISALTMALAR LİSTESİ

- AID:** Automatic Interaction Detector (Otomatik Etkileşim Belirleyicisi)
- CART:** Classification and Regression Trees (Sınıflama ve Regresyon Ağaçları)
- CHAID:** Chi-Squared Automatic Interaction Detector (Otomatik Ki-Kare Etkileşim Belirleme)
- CI:** Confidence Interval (Güven Aralığı)
- COVID-19:** Coronavirus Disease 2019 (Koronavirüs Hastalığı 2019)
- FITAS:** Filyasyon ve İzolasyon Takip Sistemi
- GCV:** Generalized Cross Validation (Genelleştirilmiş Çapraz Geçerlilik)
- GDPR:** General Data Protection Regulation (Genel Veri Koruma Yönetmeliği)
- HBIS:** Hastane Bilgi İşletim Sistemi
- ID3:** Iterative Dichotomiser 3 (Yinelemeli dikotomatör 3)
- K-NN:** K-Nearest Neighbours (K-En Yakın Komşular)
- LAD:** Least Absolute Deviation (En Küçük Mutlak Sapma)
- LSD:** Least Squared Deviation (En Küçük kareler sapması)
- MARS:** Multivariate Adaptive Regression Splines (Çok Değişkenli Uyarlanabilir Regresyon Uzanımları)
- MDL:** Minimum Description Length (En Küçük Uzunluk Tanımlaması)
- MLE:** Maximum Likelihood Estimation (Maksimum Olasılık tahmini)
- MSE:** Mean Squared Error (Ortalama Karesel Hata)
- QDA:** Qualitative Data Analysis (kuadratik diskriminant analizi)
- OLS:** Ordinary Least Squares (en küçük kareler)
- QUEST:** Quick, Unbiased, Efficient Statistical Tree (Hızlı, Yansız, Etkin İstatistiksel Ağaç)
- ROC:** Receiver Operating Characteristic (Alıcı İşlem Karakteristiği)
- rRT-PCR:** gerçek-zamanlı reverse transkripsiyon polimeraz zincir reaksiyonu
- SARS-CoV-2:** Severe Acute Respiratory Syndrome Coronavirus 2 (Şiddetli Akut Solunum Sendromu Koronavirüs 2)
- SLIQ:** Supervised Learning in Quest (Araştırmada Denetimli Öğrenme)
- SPRINT:** Scalable Parallelizable Induction of Decision Trees (Ölçeklenebilir, Parellenebilir Tümevarım Karar Ağacı)

ÖZET

Alagöz MH. COVID-19 Enfeksiyonunun Biyokimyasal Yansımalarının Değerlendirilmesi. İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü, Tıbbi Biyokimya ABD. Doktora Tezi, İstanbul, 2023.

Yapay zeka teknolojileri sayısı ve çeşitliliği gün geçtikçe artan elektronik kayıtların büyük bir doğrulukla hızla değerlendirebilme yetenekli sayesinderobotik, otomasyon, görüntü ve doğal dil işlemeyi kapsayan birçok sektörde giderek benimsenmektedir. Yapay zeka destekli tıbbi teknolojiler, özellikle de görüntü analizi uygulamaları klinik pratikte hızla uygulanabilir çözümlere dönüşmektedir. Yakın gelecekte de klinik laboratuvar verilerini yapay zeka ile birleştiren uygulamaların sağlık alanındaki çalışmalara yeni bir ufuk açması beklenmektedir.

Bu kapsamda çalışmamızda COVID-19 tanısı rRT-PCR ile doğrulanmış hastaların rutin biyokimyasal test sonuçları retrospektif olarak bir geleneksel istatistik yöntem olan lojistik regresyon analizi ve makine öğrenmesi temelli karar ağacı algoritması olan Sınıflandırma ve Regresyon Ağacı (CART) ile değerlendirilerek hastalığın başlangıcından itibaren ayaktan/yatarak takip prognozunun tahmin edilmesi, tüm vakalarda başlangıç ve hastane içi mortalite risk profillerinin belirlenmesinde kullanılacak modeller oluşturulması amaçlanmıştır.

Toplam 8036 COVID-19 vakasının başlangıç rutin tıbbi laboratuvar test sonuçları, hastalar prognozlarına göre ayaktan/yatarak takip edilenler, tüm hastalar sağkalan/vefat edenler ve hastanede yatan hastalar sağkalan/vefat edenler olarak etiketlenip gruplandırılarak geleneksel lojistik regresyon analizi ve CART algoritması ile modellenmiştir.

COVID-19 tanısı almış hastaların ayaktan-yatarak takip öngörüsünde en etkili parametrelerin sırasıyla CRP >31.45 mg/L, LDH >305.50I U/L ve $Ca^{+2} \leq 8.225$ mg/dL olduğun saptanmış ve elde edilen modelin doğruluğunun %80.8, duyarlılığının %91.4 ve özgüllüğünün % 48.8 olduğu hesaplanmıştır. Başlangıç mortalite öngörüsünde en etkili parametrelerin ise üre >62.65 mg/dL, LDH >359.5 IU/L, CRP >73.45 mg/dL ve cTnI >33.39 pg/mL olduğun saptanmış ve elde edilen modelin doğruluğu %96.8, duyarlılığı %99.5 ve özgüllüğü %30.1 olarak hesaplanmıştır. Hastane içi mortalite

öngörüsü için etkili parametrelerin sırasıyla cTnI >33.06 pg/mL, üre >65.5mg/dL, LDH >435.50 IU/L ve nötrofil sayısı >5.78 x10³/µL olduğu saptarak modelin doğruluğu %87.7, duyarlılığı %96.2 ve özgüllüğü %42.1 olarak hesaplanmıştır.

Elde ettiğimiz bulgular makine öğrenmesi yöntemlerinin değişken çeşitliği fazla olan büyük veri setlerindeki doğrusal olmayan ilişkileri modellemede geleneksel istatistiki yöntemlerden daha başarılı olduğu görüşünü doğrular niteliktedir. Makine öğrenmesi yöntemi ile elde ettiğimiz tüm karar ağacı modelleri ayaktan-hastanede yatarak takip prognozunun, başlangıç ve hastane içi mortalite riskinin belirlenmesinde yüksek riskli hastaları büyük bir doğruluk oranı ile belirleyebilmiş ve yüksek risk ile ilişkili parametreleri kesim değerleri ile birlikte ortaya koyabilmiştir.

Anahtar Kelimeler: COVID-19, Yapay zeka, Makine öğrenmesi, CART algoritması

ABSTRACT

Alagöz MH. Evaluation of Biochemical Reflections in COVID-19 Infection. Istanbul University Institute of Health Sciences, Department of Medical Biochemistry. PhD Thesis, Istanbul, 2023.

Artificial intelligence technologies are increasingly integrated to a wide range of industries related to robotics, automation, image processing and natural language processing for their ability to evaluate growing quantity and diversity electronic records with high accuracy rapidly. Artificial intelligence-powered medical technologies, especially the applications on image processing, are rapidly evolving into applicable solutions in clinical practice. Applications combining clinical laboratory data and artificial intelligence are also expected to open a new horizon in medical sciences in near future.

In this context, we developed models to predict the out- and inpatient follow-up prognosis along with all case and in-hospital mortality risk profiles by a conventional statistical method, logistic regression analysis, and by a decision tree-based machine learning algorithm, Classification and Regression Tree (CART) using routine biochemical test results at admission. For this aim routine biochemical test results of a total of 8036 rRT-PCRconfirmed COVID-19 cases weremodeled using traditional logistic regression analysis and CART algorithm after grouped according to prognosis and labeled as out- or in patient; in all cases as survivor or death; in hospital patients survivor or death.

CART decision tree revealed thatin patients with COVID-19 the most effective parameters for the prediction of out-/in-patient follow-up prognosis of were CRP >31.45 mg/L, LDH >305.50 IU/L, and $Ca^{+2} \leq 8.225$ mg/dL, respectively with accuracy 80.8%, sensitivity 91.4% and specificity 48.8%. The most effective parameters for the prediction of all case mortality were urea >62.65 mg/dL, LDH >359.5 IU/L, CRP >73.45 mg/dL and cTnI >33.39 pg/mL, respectively with accuracy 96.8%, sensitivity 99.5% and specificity 30.1%. The effective parameters for the prediction of in-hospital mortality were cTnI >33.06 pg/mL, urea >65.5 mg/dL, LDH >435.50 IU/L, and neutrophil count $>5.78 \times 10^3/\mu\text{L}$, respectively with accuracy 87.7%, sensitivity 96.2%, and specificity 42.1%.

Our results confirm the opinion that machine learning applications seem to be more successful than conventional statistical methods in modeling nonlinear relationships in large data sets with various variables. All machine learning based decision tree models developed by CART algorithm were able to predict high-risk patients with a great accuracy for of outpatient-inpatient follow-up prognosis, all case and in-hospital mortality risk and, also, revealed the most effective parameters and their cut-off values.

Keywords: COVID-19, Artificial intelligence, Machine learning, CART algorithm



1. GİRİŞ VE AMAÇ

Bilgisayar teknolojisindeki hızlı ilerlemeler ve veri madenciliğinin birçok alandaki etkileyici performansı göz önüne alındığında, hastaların demografik bilgileri, tanı ve tedavi geçmişleri, tıbbi laboratuvar ve görüntüleme sonuçlarını içeren elektronik tıbbi kayıtların da büyük ölçekli analizler gerektiren değerli bir kaynak olduğu görülmektedir. Bilgisayar teknolojileri ve veri madenciliği biliminin bir bileşkesi olan yapay zeka uygulamaları ile elektronik tıbbi kayıt verileri kullanılarak ön tanı ve prognoz taminlemesi ya da risk belirlemesi yapabilen modeller oluşturulabilmektedir. Yapay zekanın alt çalışma alanları olan makine öğrenmesi ve derin öğrenme ile sayısı ve çeşitliliği giderek artan sağlık verileri büyük bir doğrulukla hızlıca değerlendirilip yorumlanarak hasta ile geçirilen kısıtlı sürede klisyenlere tıbbi kararlar için yeterli zaman kalması sağlanabilir. Bu görüşten yola çıkarak son yıllarda uygulaması nispeten basit, sunduğu modelin karar verme kuralları açık ve kolay anlaşılabilir olduğu için bankacılık, pazarlama, finans, sigortacılık, üretim, endüstri, turizm gibi birçok alanda yaygın olarak kullanılmakta olan karar ağacı algoritmaları, sağlık alanında da kullanılmaya başlamıştır. Bir makine öğrenmesi algoritması olan karar ağaçları, genel olarak sınıflandırma ve regresyon temeline dayanmaktadır. Bir problemin veriye dayalı olarak tanımlanması, açıklanması ya da sınıflandırılması için tüm olası seçenekleri ve bu seçenekleri etkileyebilecek tüm değişkenleri olası her bir sonuç için değerlendirerek karar verme sürecinde izlenecek yolu grafiksel olarak ortaya koyarlar.

Diğer taraftan Dünya Sağlık Örgütü tarafından 11 Mart 2020'de resmen pandemi olarak sınıflandırılan Corona virüs hastalığı 2019 (COVID-19), hastalıkla etkin mücadele, sınırlı hastane ve yoğun bakım kaynaklarının rasyonel kullanımı için yüksek riskli vakaların erkenden belirlenmesinde kanıta dayalı prognoz ve risk belirleme sisteminin gerekliliğini ortaya koymuştur. Küresel çaptaki bu hastalığı anlamak, mücadele etmek ve kontrol altına almak için gösterilen çabalar birçok bilimsel çalışmaya da konu olmuştur. Bu çalışmaların bir kısmında hastalılık şiddetini belirlemek ve mortalite riskin öngörmek için makine öğrenmesi yöntemleri kullanılarak tek başına veya klinik bulgular ile birlikte tıbbi laboratuvar sonuçlarının değerlendirilmesinden de yararlanılmıştır.

Bu kapsamda çalışmamızda COVID-19 tanısı gerçek-zamanlı reverse transkripsiyon polimeraz zincir reaksiyonu (rRT-PCR) ile doğrulanmış hastaların rutin

biyokimyasal test sonuçları retrospektif olarak bir geleneksel istatistik yöntem olan lojistik regresyon analizi ve bir karar ağacı algoritması olan Sınıflandırma ve Regresyon Ağacı (CART) ile değerlendirilerek hastalığın başlangıcından itibaren hastalık şiddetinin tahmin edilmesi, başlangıç ve hastane içi mortalite risk profillerinin belirlenmesinde kullanılacak modeller oluşturulması amaçlanmıştır.



2. GENEL BİLGİLER

2.1. YAPAY ZEKA VE MAKİNE ÖĞRENMESİ

Yapay zeka görsel işleme, dil işleme, hesaplama, öğrenme, problem çözme, karar verme gibi insan zekasını gerektiren karmaşık bilişsel işlevleri yapabilecek bilgisayar sistemleri ve teknolojilerin geliştirilmesi girişimlerinin tümünü içeren bir bilgisayar bilimleri alanıdır. Bu bilim alanının başlıca çalışma alanlarından biri büyük veri kümelerinin içerdiği birçok parametre arasındaki olası gizli bağlantı ve örüntüleri açığa çıkartmaktır. İnsan zekasına benzer şekilde akılcı düşünebilen, öğrenebilen bilgisayar ve yazılımların (makinelere) bir toplamı olan yapay zeka, uygulama prensipleri, uygulama alanları ve sağladığı katkılara göre makine öğrenmesi, derin öğrenme gibi çeşitli alt çalışma alanları bulunmaktadır (Şekil 2.1) [1- 3].



Şekil 2.1. Yapay zeka teknikleri

Bunlardan makine öğrenmesi yapay zeka çalışmalarının temelini oluşturur. Bilgisayarların insanlara benzer şekilde öğrenmesini sağlamak için bir görevi, bu görevle ait veriye göre bilgisayar algoritmaları ile modelleyerek yüksek performanslı çıkarımlar ve/veya tahminler yapabilen modeller geliştirilmesi çalışmalarını kapsar. Derin öğrenme ise bir veya daha fazla gizli katman içeren çok büyük veri setlerinin insan beyninin yapı ve işlevinden ilham alan yapay sinir ağları ve diğer makine öğrenme algoritmaları eğitilmesi konusunda çalışır [1,3].

2.1.1. Yapay Zeka ve Makine Öğrenmesi

Genel olarak makinelerin insanlar gibi düşünmesini amaçlayan bilgisayar algoritmalarını ifade eden makine öğrenmesi, örnek verilerin (eğitim verisinin) bir dizi algoritma ile işlenerek elde edilen bilgiden öğrenebilen ve anlamlı çıkarım/tahminler yapabilen bir model oluşturulmasıdır. Makine öğrenmesi ile bir görevi gerçekleştirmek için her seferinde bir dizi talimat kullanarak kod yazmak yerine, makineler öğrenme yeteneği kazandıran bir dizi eğitim verisi ve algoritma ile bu görevi yürütecek şekilde eğitilir. Veri, kullanılan algoritma ve verilerin işlendiği platformun hesaplama gücü, makine öğrenmesi çalışmalarının üç temel bileşenini oluşturur [1,2]. Ayrıca, analitik sorgularda verilerden yüksek performanslı çıktılar elde edebilmek için makine öğrenmesinde izlenecek yöntem de önemlidir. Makine öğrenme yöntemleri denetimli (supervised) öğrenme, denetimsiz (unsupervised) öğrenme ve destekli/pekiştirmeli (reinforcement) öğrenme olarak üç temel grupta sınıflandırılmaktadır [4].

Bu makine öğrenmesi yöntemleri en uygun tahmine ulaşmak için farklı yollar izler ancak, tümünde ortak özellik oluşan modelin insanlar tarafından programlanmayıp kendi sahip olduğu özellikleri baz alarak veriden öğrenmesi ve eğitilmesidir [1,5]. Bunlardan denetimli öğrenme, tahmin edilmeye çalışılan sonuç için bilinen (etiketlenmiş) girdi ve çıktı verisi varsa kullanabilir. Etiketlenmiş girdi verisi ve bilinen yanıtları genellikle sınıflandırma ve regresyon temelli bir öğrenme algoritması ile işlenir ve oluşturulan model belli bir belirsizlik seviyesinde kanıta dayalı tahmin(ler) yapmak üzere yeni (etiketlenmemiş) veriler ile test edilerek geliştirilir [6,7]. Denetimsiz öğrenme ise tahmin edilmeye çalışılan sonuç için etiketlenmiş girdi ve çıktı verisinin olmadığı durumlarda tercih edilir. Etiketlenmemiş verilerden oluşan veri seti genellikle kümeleme algoritmaları ile işlenerek setinin açıklanması, veri içerisindeki benzer ve farklı örüntü/ilişkilerin tanımlanması yapılabilir [6,7]. Etiketlenmiş ve etiketlenmemiş verinin bir arada bulunması gibi durumlarda ise sınıflandırma performansı sağlamak için denetimli ve denetimsiz öğrenmenin ardışık kullanıldığı yarı denetimli öğrenme kullanılır [2]. Pekiştirmeli öğrenme, denetimli-denetimsiz öğrenmeden farklıdır. Bu tip öğrenmede makinenin deneme yanılma yoluyla istenilen uygun tahmine ulaşmasında bir ödül-ceza sistemi uygulanır. Modelin en az hata ile doğru çıktıya ulaşırken ödül-ceza sistemiyle hatalardan çıkarımlar yaparak kendini sürekli olarak eğittiği bir ortam kurgulanır [8].

2.1.2. Makine Öğrenmesi Algoritmaları

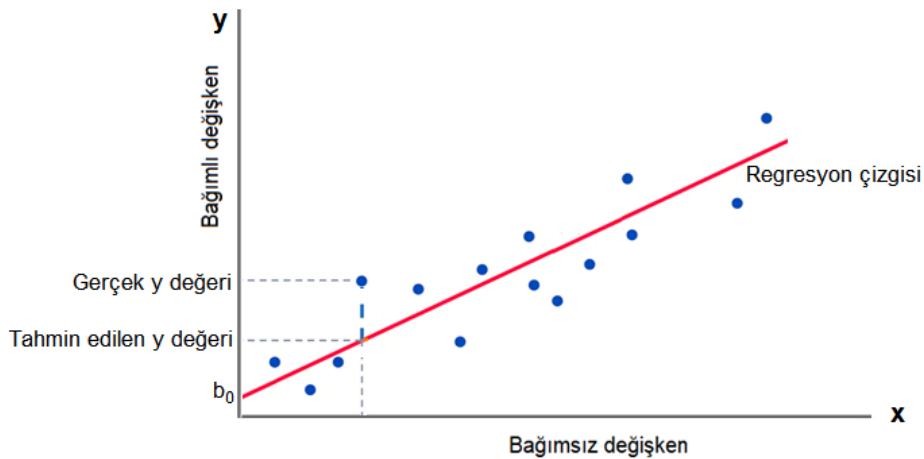
Bir veri madenciliğinde iyi tanımlanmış bir veriyi girdi olarak alan ve sonuçta modeller ya da örüntüler formunda bir çıktı üretilmesini sağlayan yol ve yöntemlerin tümü algoritma olarak isimlendirilmektedir. Makine öğrenmesi algoritmaları ise insan müdahalesi olmadan verilerden öğrenebilen ve geliştirebilen bilgisayar programlarıdır. Bilgisayarlara istatistiksel ve mantıksal işlemler aracılığıyla öğrenme yeteneği kazandırılması gerçekte bir yöntemler bütünüdür [2,5]. Yapılmak istenen görev, veri setinin büyüklüğü, parametrelerin sayısı ve özelliklerine göre sayısal veriler için lineer regresyon, ikili sınıflama için lojistik regresyon ve destek vektör makinesi, çoklu sınıflama için lineer diskriminant analizi, sınıflandırma ve regresyon için karar ağaçları ya da vektör nicemlemenin öğrenilmesi, kümeleme için K-en yakın komşuluk, olasılık tabanlı istatistikî sınıflamalar için Naive Bayes ya da birden çok modelin üretilip ortalamadan elde edilen sonucun değerlendirildiği rastgele ormanlar makine öğrenmesinde en yaygın kullanılan algoritmalarıdır [1,2].

2.1.2.1. Lineer (Doğrusal) Regresyon Algoritması

Denetimli bir öğrenme algoritması olan lineer regresyon, veri bilimi ve makine öğrenmesinde iki ya da daha çok değişken arasında doğrusal bir ilişki bulunup bulunmadığını araştırmak için sık başvurulan en basit algoritmadır [2]. Yaş, cinsiyet, hastalık süresi, uygulanan tedavi gibi bir bağımsız değişkenin tip 2 diyabetiklerin sağkalım öngörüsüne katkısının incelenmesinde olduğu gibi bağımsız bir değişkendeki artış ya da azalmanın bağımlı değişkeni nasıl etkilediğini doğru olarak öngörebilecek katsayılar elde etmek ve yeni vakalar için tutarlı tahminlerde bulunan bir model oluşturmak amacıyla kullanılır. Değişkenler arasındaki ilişki $y = ax + b$ denklemi ile temsil edilir ve bağımsız değişkenlerin parametreleri en küçük kareler (Ordinary Least Squares, OLS) yöntemiyle tahmin edilir. Burada a regresyon çizgisinin eğimini, y bağımlı değişkeni, x bağımsız değişkeni, b ise doğrunun ne kadar kaydırılacağını yani x sıfırken y 'nin değerini gösterir. Bağımlı ve bağımsız değişkenler arasındaki ilişki ait tahminler grafiksel olarak regresyon çizgisi olarak isimlendirilen doğrusal bir çizgi ile gösterilir (Şekil 2.2) [9]. Bağımlı ve bağımsız değişken arasında pozitif yönlü bir ilişki varsa eğim pozitif ve regresyon çizgisi yukarı doğrudur, negatif yönlü bir ilişki varsa da tam tersi olur. Lineer regresyonla oluşturulan modelden elde edilen y değeri ile gerçek y

değeri arasındaki fark ne kadar az ise modelin tahmin gücü o kadar gerçeğe yakındır. Elde edilen tahminlerin (regresyon çizgisinin) gerçek duruma ne kadar yakın olduğunun matematiksel ölçüsü determinasyon katsayısı R^2 ile ifade edilir [9].

Lineer regresyon değişkenler arasındaki ilişkileri analiz etmek için anlaşılması ve uygulanması en kolay algoritmalarından biri olmakla birlikte, değişkenler arasında doğrusal bir ilişki olduğunu varsayması ve değişkenler arasındaki ilişkileri açıklamada yetersizdir. Bu nedenle değişkenler arasında doğrusal bir ilişkisi olup olmadığını değerlendirmek dışında lineer regresyon pratik uygulamada genellikle önerilmemektedir [2,10].



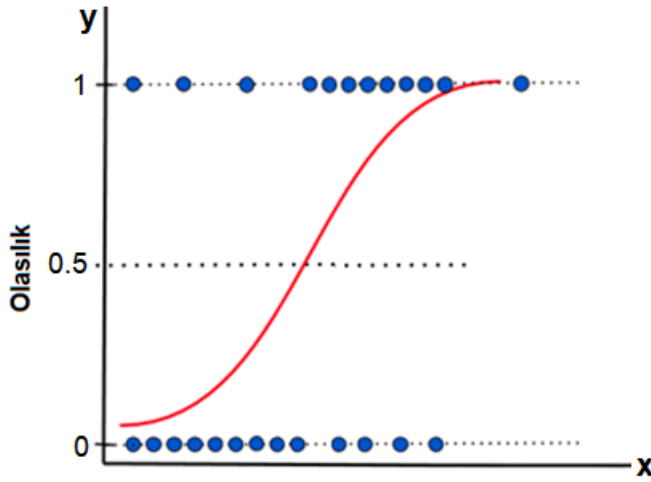
Şekil 2.2. Linear (doğrusal) regresyon grafiği

2.1.2.2. Lojistik Regresyon Algoritması

Lojistik regresyon isminde regresyon barındırmasına rağmen bir regresyon algoritması değil, sınıflandırma algoritmasıdır. Lojistik regresyonun temel hedefi, birey/durum/olayların farklı gruplara ayrılarak sınıflanmasıdır.

Sonuçları kanser-kanser değil, etkili-etkili değil, sağ-ölü gibi iki sınıflı (ikili/ dikotomik /binary) olarak ifade edilen bağımlı değişkenlerin modellenmesinde kullanılır. Bu modellerde ikili sonuç genellikle 0 veya 1 olarak tanımlanır ve bu ikili bağımlı değişken sınıfına üye olma durumunu en iyi açıklayan bağımsız değişkenler tahmin edilir. Amaç, bağımlı ile bağımsız değişkenler arasındaki ilişkiyi en az değişkeni kullanarak en iyi uyuma sahip olacak şekilde tanımlayan, kabul edilebilir bir model kurmaktır. Modelde kullanılan veri setinin gerçek hayattaki sonsuz parametre havuzundan fark edilme olasılığını en üst düzeye çıkaran en uygun parametreleri seçmek amacıyla lojistik

regresyonda parametreler maksimum olabilirlik çıkarımı (Maximum Likelihood Estimation,(MLE) yöntemiyle hesaplanır. Artı ve eksi sonsuz arasında $(-\infty, +\infty)$ sonuç veren lineer regresyonun aksine, lojistik regresyonda elde edilmek istenen tek bir olasılık değeri vardır. Bu nedenle değerler 0 ile 1 arasında alınarak S şeklindeki lojistik fonksiyonundan (sigmoid fonksiyon) geçirilir, elde edilen olasılık değeri 0.5'den (eşik değer) büyük ise sonuç 1, küçük ise 0 olarak etiketlenir (Şekil 2.3) [2,11,12].

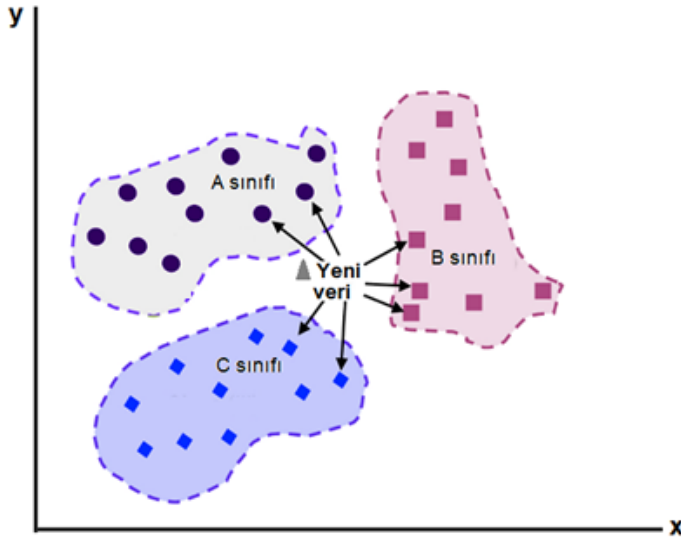


Şekil 2.3. Lojistik regresyon grafiği

2.1.2.3. K-En Yakın Komşu (K-NN) Algoritması

En basit makine öğrenmesi algoritmalarından birisi olan K-en yakın komşu (K-Nearest Neighbours) algoritması, hem sınıflama hem de regresyon işlemleri için kullanılabilen bir denetimli makine öğrenmesi yöntemidir. K-NN, eğitim setindeki sınıfı bilinmeyen verilerin diğer veriler ile karşılaştırılarak bir uzaklık hesaplaması yapıldıktan sonra hesaplanan uzaklığa göre en uygun sınıfa atanarak sınıflandırılması temeline dayanır [2,13]. Uzaklık hesaplamasında Öklid ve Pisagor bağıntılarını kullanan Öklidyen yöntem ya da uyarlanabilir eksenler boyunca dik açılarda ölçülen iki nokta arasındaki mesafeyi hesaplayan Manhattan yöntemi başta olmak üzere Chebyshev, Hamming, Minkowski, Jacckard uzaklık hesaplama yöntemi gibi çok sayıda yöntem bulunmaktadır. K-NN'de ilk olarak verilen bir noktaya en yakın komşuların sayısını ifade eden k parametresi belirlenir. Örneğin k üç olarak belirlendiğinde en yakın üç komşuya göre sınıflama yapılır. Daha sonra eğitim setine katılacak yeni verinin diğer verilere göre uzaklığı, kullanılan uzaklık hesaplama yöntemi ile tek tek hesaplanır. Elde

edilen değerlere göre yeni veri k'nın bir tam sayı olabileceği en yakın k komşu veya komşuların sınıfına atanarak sınıflandırılmış yani etiketlenmiş olur (Şekil 2.4) [2,13,14]. K-NN gürültülü eğitim verilerine karşı dirençli olması nedeniyle en popüler makine öğrenme algoritmalarından biridir. Ancak, diğer denetimli öğrenme algoritmalarının aksine, K-NN eğitim aşamasına sahip değildir, eğitim verilerini öğrenmek yerine eğitim veri setini ezberlediğinden örüntü tabanlı öğrenme sunar. Ayrıca, tekrarlanan algoritmik işlemler sınıflandırma sürecinin uzamasına yol açtığından tembel bir öğrenme türüdür, çok büyük veri setlerini işlemek için önerilmemektedir. Bir diğer dezavantajı da değişkenlerin nominal, ordinal veya orantılı ölçek gibi farklı ölçme türlerinde ya da sayısal ve kategorik değişkenlerin bir karışımı olması durumunda ortaya çıkan zorluklardır (2,14).



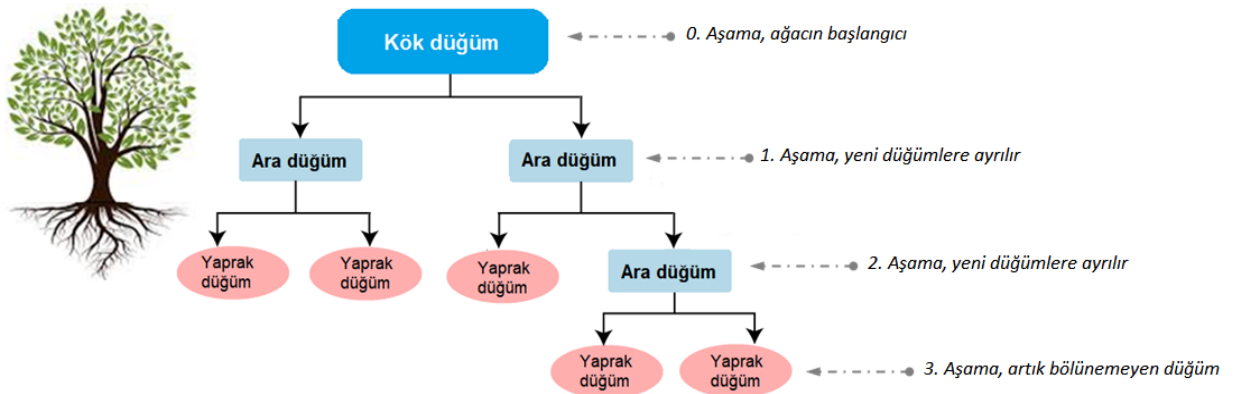
Şekil 2.4. K-En Yakın Komşu (K-NN) grafiği

2.1.2.4. Karar Ağaçları

Karar ağaçları genel olarak sınıflandırma ve regresyon temeline dayanan ve veri biliminde hemen her tür sınıflandırma problemlerinde kullanılan gözetimli öğrenme algoritmalarıdır. Bir problemin veriye dayalı olarak tanımlanması, açıklanması ya da sınıflandırılması için tüm olası seçenekleri ve bu seçenekleri etkileyebilecek tüm değişkenleri olası her bir sonuç için değerlendirerek karar verme sürecinde izlenecek yolu grafiksel olarak ortaya koyarlar. Karar ağaçlarının sunduğu modelin karar verme kuralları açık ve kolay anlaşılabilir olduğundan bankacılık, pazarlama, finans,

sigortacılık, mühendislik, üretim, endüstri, sağlık ve turizm gibi birçok alanda yaygın olarak kullanılmaktadır [2,5,15].

Karar ağacı algoritmaları, bir problemi çözmek için araştırma evrenini her aşamada bir ağacın kökünden başlayıp bir yaprağa doğru ilerleyen bir strateji ile tarayarak, bir ağaca benzer yapıda akış diyagramları oluşturur. Bu akış diyagramlarında bir ağaçta olduğu gibi kök, dal ve yapraklar bulunur (Şekil 2.5) [1,16]. Kök düğüm ile başlayan diyagram, veri dizisinin belli kriterlere göre ara düğümlere ayrılması ile devam eder ve yaprak düğümler (karar düğümleri ya da yapraklar) ile sonlanır, düğümler arası bağlantılar da dallar ile sağlanır. Her düğüm bağımsız değişkenlerin nitelik ya da nicelik değeri üzerindeki ölçümü, her dal o ölçüme ait bir çıktıyı ve yapraklar da sınıf ya da sınıf dağılımlarını gösterir [1,16]. Akış diyagramında da her düğüm test ettiği bağımsız değişkenle etiketlenir ve dallara karşılık gelen değerler belirtilir. Sonuçta elde edilen karar ağacı modeli ile araştırma evreni, en önemli nitelik ya da bağımsız değişken(ler)e dayanan bir dizi hiyerarşik karara göre o gruptaki tüm elemanlar aynı sınıf etiketine sahip olana kadar tekrar tekrar gruplanarak iki veya daha fazla sayıda homojen sınıfa ayrılır. Bu sınıflara ulaşmak için izlenen kural kümesi, anlaşılabilirliği ve uygulanabilirliği basitleştirilerek benzer alanlarda çalışan kullanıcılar için kurallara/talimatlara/önerilere dönüştürülebilir [1,2,16].



Şekil 2.5. Karar ağacı modeli

Karar ağaçları temeline dayanan makine öğrenmesi çalışmaları 20. yüzyılda başlayıp kök, düğüm ve dallanma kriterlerinin seçiminde izlenen yola göre farklılık gösteren birçok karar ağacı algoritması geliştirilmiştir [1,5]. Kayıtlara geçen karar ağacı

tabanlı ilk algoritma 1963 yılında Morgan ve Sonquist tarafından geliştirilen, klasik istatistiki varsayımlara dayanarak değişken değerlerine göre en iyi tahmini gerçekleştirilmeye ve en kuvvetli ilişkiye sahip bağımsız değişkeni bulmayı hedefleyen AID (Automatic Interaction Detector) algoritmasıdır [17]. Bunu 1970'lerde J. Ross Quinlan tarafından geliştirilen, Iterative Dichotomiser 3 (ID3) algoritması izlemektedir. ID3 sınıflandırma problemlerinin çözümünde temel olarak entropiden faydalanır ve en iyi bölünmeyi belirlemek için bilgi kazancını kullanır. Bilgi kazancı entropi ile yakından ilişkilidir, veriler ile ilgili belirsizliğin bir ölçüsü olan entropi ne kadar az ise bilgi kazancı o kadar fazladır [18]. Veri seti bu algoritma ile işlendiğinde, her aşamada veri kümesinin belirsizliğinin en az, bilgi kazancının ise en yüksek olduğu bölünme ölçütü dallanma kriteri olarak kullanılır. C4.5 ve C5.0 de yine Quinlan tarafından ID3'ün hız, bellek kullanımı, kural oluşturma, budanabilme, değişkenlerin ve sınıflandırma türlerinin ağırlıklandırılma performansları açısından geliştirilmesi ile ortaya çıkan entropi temelli algoritmalarıdır [5,19].

Günümüzde CHAID (Chi-Squared Automatic Interaction Detector), CART (Classification and Regression Trees), QUEST (Quick, Unbiased, Efficient Statistical Tree), Random Forest (Rassal Karar Ormanları), LMDT, MARS (Multivariate Adaptive Regression Splines) gibi çok sayıdaki makine öğrenmesi algoritmaları birçok sektörde yaygın olarak kullanılmakta ve bu listeye hızla yeni algoritmalar eklenmektedir [2,16]. Bunlardan yaygın olarak kullanılmakta olan CHAID algoritmasında, oluşturulacak regresyon denklemi normal dağılım, doğrusallık ve homojenlik gibi klasik istatistiki varsayımlardan farklı bir davranış sergiler. İstatistiksel olarak yarı parametrik bir özellik taşıyan ve bağımsız değişkenlerin birbirleriyle olan ilişki, etkileşim ve kombinasyonlarını da inceleyen bu algoritmada, dallanma için bağımlı değişken kategorik ise ki-kare testini, sürekli ise F testini kullanarak her adımda düğüm noktası için en uygun değişken belirlenir ve bu değişkene ait homojen değerler bir araya toplayarak bir grup oluşturulur [20,21].

SLIQ ise büyük veri kümelerine seri ve paralel modelde uygulanabilen hızlı ve ölçeklenebilir bir algoritma üretmek amacıyla 1996 yılında IBM Almaden Research Center araştırmacıları tarafından geliştirilen bir algoritmadır. Hem sürekli hem de kategorik değişkenlerin sınıflandırılmasında kullanılabilen bu algoritma verileri sıraya dizme işlemini her düğümde yapmak yerine başlangıcında sadece bir kere yapılmasına dayanmaktadır. Bu nedenle SLIQ algoritmasında en iyi dallanmayı en az işleme

sağlamak için verilerin önceden sıraya dizilmesi önemlidir. Dallanma için Gini ayırma kriterini kullanarak önce genişlik ilkesine dayanan açgözlü bir teknikle aynı anda birçok yaprağı oluşturmaktadır. Hızlı olması ve iyi sonuçlar veren karar ağaçları üretmesi nedeniyle tercih edilen SLIQ algoritmasının en önemli dezavantajlarının bellekte yerleşik bir veri sınıfı listesi kullandığından verilere bellek kısıtlamaları getirmesidir [20,22].

Karar ağacı algoritmasının ölçeklenebilir paralelleştirilebilir induksiyonu anlamına gelen SPRINT algoritması 1996 yılında Shafer ve arkadaşları tarafından ortaya atılmıştır. SLIQ algoritmasına benzer şekilde büyük veri kümelerinde hem sürekli hem de kategorik değişkenleri işleyebilen, veri yerleştirme ve yük dengeleme için hem seri hem de paralel düzende uygulanabilen hızlı ve ölçeklenebilir bir karar ağacı algoritmasıdır. SPRINT algoritması öncelikle tüm değişkenler için ayrı bir değişken listesi ve değişken sayısı kadar tablo oluşturur ve veri yapısı olarak bu listeleri ve her bir değişkene ait verilerin bir kez sıraya dizildiği bellekte yerleşik olmayan tabloları kullanır. Daha sonra da karar ağacı oluşumunda Gini ayırma kriterini kullanarak veri setini, her bölüm aynı yaprak düğüm veya sınıfa ait oluncaya kadar önce genişlik ilkesine dayanan açgözlü bir teknikle özyinelemeli olarak bölümlere ayırır. SPRINT'in kullandığı farklı veri yapısı ve bellekte yerleşik olmayan tablolar sayesinde başta SLIQ olmak üzere diğer benzer algoritmalarda görülen bellek kısıtlamalarını ortadan kaldırdığı bildirilmektedir [20,23].

QUEST ise kapsamlı karar ağacı algoritmalarının daha fazla bölme sağlayan değişkenleri seçmeye eğilimli olduğu görüşünden yola çıkılarak, ihmal edilebilir önyargıya sahip, daha hızlı bir algoritma olarak 1997 yılında Loh ve Shih tarafından geliştirilmiştir. Bölünmüş bir seçim stratejisi izleyen QUEST'de karar verilirken bir seri istatistiksel testlerden yararlanır. Gerçek ve simüle edilmiş veriler kullanarak bağımsız değişken kategorik ise Pearson ki-kare testi, sürekli ise Anova F-testi ile en iyi dallanmayı sağlayacak bağımsız değişken ve karar noktası ayrı ayrı hesaplanır. Bağımsız değişkenin hatalı sınıflandırma olasılığını en aza indirmek için gruplar arası farklılıklar kuadratik diskriminant analizi (QDA) ile değerlendirir. Dallanmayı sağlayan bu değerlendirme süreci bir durdurma kuralına rastlayıncaya kadar tekrarlanır. Büyük ve karmaşık bir veri setlerinde oldukça hızlı ve doğru karar ağacı modelleri üretebilen QUEST algoritmasının en önemli dezavantajı sadece bağımlı değişkenin kategorik olması durumunda kullanılabilmesidir [21,24].

MARS algoritması büyük boyutlu verilerin esnek regresyon modellenmesi için 1990'ların başında Jerome Friedman tarafından yeni bir yöntem olarak sunulmuştur. MARS'da ağaç, temel fonksiyonların sayısının yanı sıra ilişkili parametrelerin önem derecesi ve düğüm konumlarının otomatik olarak veriler tarafından belirlendiği bir fonksiyonla genişler. Makine öğrenmesine yeni bir yaklaşım sunan bu algoritma, mevcut yöntemler ile ortaya çıkarılması oldukça güç olan, çok büyük boyutlu veri setleri içinde gizlenmiş karmaşık veri ilişkilerinin ve karşılıklı etkileşimlerin belirlenmesine olanak sağlayan bir yaklaşımdır. Ayrıca, algoritma değişkenlerin toplam katma değeri ve çok değişkenli farklı etkileşimleri ayrı ayrı tanımlayacak şekilde de sunulabilmektedir. Bu sayede değişkenlerin toplam katma değerini veya en önemli birkaç değişkendeki etkileşimi içeren ilişkileri modelleyebilecek güç ve esnekliğe sahiptir. Temel olarak CART algoritmasının geliştirilmiş bir modifikasyonu olan MARS algoritması nonparametrik bir yaklaşıma sahiptir. Regresyonun tüm cazip özelliklerini kullanarak regresyona özgü özyinelemeli bölümlenme tarafından motive edilir ve çeşitli dönüştürme teknikleri ile bağımlı ve bağımsız değişkenler arasındaki doğrusal olmayan ilişkileri doğrusal hale getirir. Ancak tipik özyinelemeli bölümlenmeden farklı olarak, büyük veri setleri için sürekli yeni türevler üretebilen devamlı modeller geliştirir. Hem kategorik hem de sürekli değişkenleri işleyebilen bu algoritma, bağımlı ve bağımsız değişkenlerin dağılımları üzerine herhangi bir ön yargı gerektirmeden, bağımlı değişken sürekli ise kestirim, kategorik ise sınıflandırma amacı ile kullanılma seçeneklerine sahiptir. Yöntem, regresyon düzenlemelerindeki performansı geliştirmek için, adimsal regresyonun ve tekrarlamalı ayırma mantığına dayanan CART yönteminin modifikasyonunun genelleştirilmiş bir hali olarak düşünülebilmektedir [25,26].

2.1.2.4.1. CART Algoritması

Breiman ve arkadaşları tarafından 1984 yılında hem sınıflandırma hem de regresyon amacıyla önerilen CART algoritmasında ise temel hedef, ağacın her adımında oluşan sınıfın kendinden daha homojen iki alt sınıfa bölünerek büyütülmesidir. Düğümün saflığı yani homojenliği hedef değişkenlere göre değişeceğinden karar ağacı algoritmalarında bölünme kararının nasıl verileceği ağacın doğruluğunu etkileyen önemli bir faktördür[27]. Bir düğümü iki veya daha fazla alt düğüme bölmeye karar

vermek için genel olarak sınıflandırma ve regresyon problemleri için bağımlı değişkenin kategorik ya da sürekli olmasına göre Gini, twoing (ikili ayırma) veya en küçük kareler sapması (LSD) teknikleri gibi farklı bölünme kriterleri uygulanmaktadır [28].

Twoing ayırma kriterinin işlem süresi daha uzun olmakla birlikte tek bir sınıfı diğerinden ayırmak yerine, bir düğüme ait verilerin %50'sini içeren ve birbirine benzemeyen sınıflar oluşturduğundan Gini kriterine göre daha dengeli bir yapı sunar [29]. Her ikisi de karar noktaları oluşturmada safsızlık fonksiyonundan yararlanan bu bölünme kurallarından Gini kriteri her adımda en büyük veri kümesini oluşturarak en iyi bölünmeyi hedefler. Bu amaçla, bölünme noktalarında bağımsız değişken değerlerini iki kısma ayırarak *Ginisağ* ve *Ginisol* değerleri hesaplanır ve dallanma ölçütü olarak Gini değeri en küçük olan değişken seçilir. Gini değeri 0 ile 1 arasında bir değerdir, 0 olması ideal bir sınıflandırmayı, 1 olması ise kötü bir bölünme sonucunu gösterir [27,30]. Twoing ayırma kriterinde ise en iyi dallanmayı sağlayan değişken değerinin seçimi ve sınıflandırma yapılırken her bir veri sağ ve sol dalda yer alabileceğinden önce tüm aday veriler için ayrı ayrı sağ ya da sol dalda bulunma olasılığı hesaplanır. Sonra, en büyük sonuç uygunluk ölçüsü olarak seçilerek değişkenler, sahip oldukları değere göre ilgili düğüme ait verilerin %50'sini içeren ve birbirine benzemeyen sınıflar oluşturmak üzere *tsol* ve *tsağ* olacak şekilde iki dala ayrılır. Aynı işlemler karar ağacının tüm yapraklarına ulaşıncaya kadar aşağıya doğru tekrarlanır [27]. Varyans azaltma olarak da isimlendirilen en küçük kareler sapması yönteminde ise en iyi dallanmayı sağlayan değişken ve değişken değeri seçilirken bağımsız değişkenlerin gözlenen ve öngörülen değerler arasındaki sapmaların (kalıntıların) karesinin toplamını en aza indiren bölünmeler aranır [31].

Karar ağaçlarının başlıca avantajları, hem kategorik hem de sayısal değişkenlerin işlenmesine olanak sağlaması; çok çıktılı problemler için kullanılabilmesi; tutarlı olması için verilerin matematiksel fonksiyonlarla normalleştirilmesi ya da dönüştürülmesini gerektirmemesi; diğer algoritmalara göre nispeten daha az veri ön hazırlığı gerektirmesi; doğrusal olmayan ilişkilerin ağaç performansını etkilememesi; elde edilen modelin kolay anlaşılabilir, yorumlanabilir ve görselleştirilebilir olmasıdır [2,5,21].

Karar ağaçlarının dezavantajları topluca ele alındığında, verileri iyi genelleştirmeyen, anlaşılması ve yorumlaması güç aşırı büyük ve karmaşık ağaçlar

oluşturma eğilimleri en önemli dezavantajlarından birisidir. Ağacın büyüklüğü arttıkça hata oranı da artmakta, doğruluğu ise azalmaktadır. Sıklıkla dengesiz veri setleri ile çalışıldığında ortaya çıkan *taşma* ya da *aşırı uyum gösterme* (over fitting) olarak isimlendirilen bu durumu önlemek için uygulamadan önce veri setinin dengelenmesi önerilmektedir. Ancak over fitting dengeli bir veri seti ile çalışıldığında da olasıdır. Böyle bir sonucun nedenlerinden biri de veri setinin gürültülü olması yani veri setinde çok sayıda tutarsız, eksik ya da hatalı veriler bulunması olabilir. Gürültü veri yığınının gereksiz kurallar türetmesine ve ağacın çok büyük ve karmaşık olmasına yol açabilmektedir. Diğer taraftan ağaç oluşturma algoritmalarının temel hedefi eğitim setindeki en ufak detayları bile değerlendirerek kurallaştırırken eğitim setini tutarlı yapmaya çalışmaktır. Over fitting sorununun yol açan nedenlerden biri de bu değerlendirip kurallaştırma işlemidir çünkü önemli olan, ağacın eğitim kümesindeki tüm verileri tutarlı kılması değil doğruluğudur [2,5,21].

Over fitting sorunu ile baş etmek için sınıflamaya katkısı olmayan düğümlerin karar ağacından çıkartılması işlemi yani *budama* (pruning) yöntemlerine başvurulmaktadır. Budama işlemi daha ağaç oluşturulurken *önden budama* (erken durdurma, pre-pruning) ya da ağaç oluşumunu takiben *sonradan budama* (post-pruning) şeklinde uygulanabilir [32].

Önden budamada temel yaklaşım başlangıçta çapraz doğrulama hatası için kabul edilebilir minimum eşik değer belirlenerek dallanmanın gerçekleşmesi ya da durdurulması kararının bu önceden belirlenmiş kurala göre verilmesidir. Ağaç, bölmenin her aşamasında bu kurala uygunluk açısından kontrol edilir, bölünme yeteri kadar fayda sağlamayacak aksine sonraki bölünmeler hatayı arttırıcı nitelikte olacaksa bölünme işlemi durdurulur ve o kümedeki baskın sınıf yaprağa dönüştürülür. Önden budama ağaca erken müdahale edilmesine olanak sağlayan, hız ve verimliliği arttıran bir yöntem olduğundan çok büyük ve karmaşık veri setlerinin incelenmesinde tercih edilen bir yöntemdir. Ancak önden budamanın açgözlü olmak gibi önemli bir sorunu vardır; sonraki bir bölüm son derece değerli olsa bile erken durdurma kuralları algoritmanın bir bölmeden kaçınmasına ve karar ağacının yeterli olgunluğa ulaşmadan durdurulmasına yol açabilir. Ayrıca örnek sayısı az olan veri setlerinde durdurma kurallarının hata oranını arttırması gibi sakıncaları da bulunmaktadır. Önden budamaya göre daha yavaş ancak matematiksel olarak daha titiz ve doğruluğu daha yüksek olan sonradan budama yönteminde ise oluşturulmuş karar ağacının dalları ortalama kareler

hatası ve/veya yanlış sınıflandırma oranı ölçütlerini kullanarak budanır. Budama sonrası orijinal ağaçla aynı köke sahip ancak bazı dalları olmayan alt ağaçlardan oluşan bir budama kümesi (prune set) elde edilir. Daha sonra bu kümedeki alt ağaçlardan her biri çapraz doğrulanmış bir ölçüt kullanılarak değerlendirilir ve kesildiğinde en az hata üreten dallar budanarak yaprağa dönüştürülür [2,32].

Karar ağaçlarının bir diğer önemli dezavantajı da kararsız/dengesiz olabilmeleridir. Verilerdeki küçük bir değişiklik özellikle de köke yakın bir bölünmedeki küçük bir değişiklik karar ağacının yapısında büyük bir değişikliğe yol açtığından tamamen farklı bir ağaç oluşumuna neden olabilmektedir. Kararsızlığın çoğu kez over fitting'in göstergesi olduğu gösterilmiştir. Gürültü verilerinden öğrenme ve/veya uygun algoritma seçilmemesi çapraz doğrulama sonuçlarının yüksek varyanslı tahmin ediciler haline gelmesine, bunun da algoritmanın verilerdeki kalıcı ve rastgele modeller arasında net bir ayırım yapamamasına yol açtığı bildirilmektedir. Kararsızlık sorunuyla başa çıkmak her biri ilave hesaplamalar ve çaba gerektiren birçok yöntem önerilmektedir. Bunlar arasında, ihtiyaç duyulan özelliklerin ve örneklemelerin değiştirilmesi ile bir dizi alternatif ağaç eğitilerek en iyi ağacın kendi alanında uzman kullanıcılar tarafından seçilmesi ya da her çalıştırmada yalnızca bir ağaç üreten C4.5 algoritması kullanılarak birkaç örneğe dayalı alternatif ağaçlardan tek bir sınıflandırıcı üreten *pencereleme* yaklaşımının kullanılması gibi yöntemler bulunmaktadır [5,33].

Makine öğrenmesi çalışmalarında bir diğer önemli konuda veri setindeki eksik/kayıp verilerdir. Karar ağaçları diğer algoritmalara kıyasla eksik verilerden daha az etkilense de eksik veriler öğrenme aşamasında uygulanan algoritmaları ezberle yöneltebilir hatta çok sayıda eksik veri olması algoritmanın çalışmasını engelleyebilir. Veri setinde eksik veri bulunması durumunda ya eksik verisi olan değişkenin veri setinden çıkarılması ya da eksik doldurulması seçeneklerinden birisine başvurulmaktadır. Genel olarak veri setindeki değişkenlerden birinde %50'ye varan bir eksiklik varsa kullanırken çok dikkat edilmesi, eksiklik %70– 80 katarsa o değişkenin veri setinden çıkarılması önerilmektedir. Eksik veriyi doldururken de herhangi bir sapmaya yol açmamak, verinin karakterini ve içerdiği bilgileri bozmadan yansız ve etkin bir doldurma yapılması önemlidir. Eksik veriden kaynaklanacak hata toleransını azaltmak için istatistikî teorilerle geliştirilmiş zincirleme denklemlerle çoklu atama (MICE), complete case analysis (CCA), cold/hot deck imputation gibi birçok algoritma bulunmaktadır. Ancak, eksik veri sayısı çok fazla ise, özellikle de eksiklik bağımlı

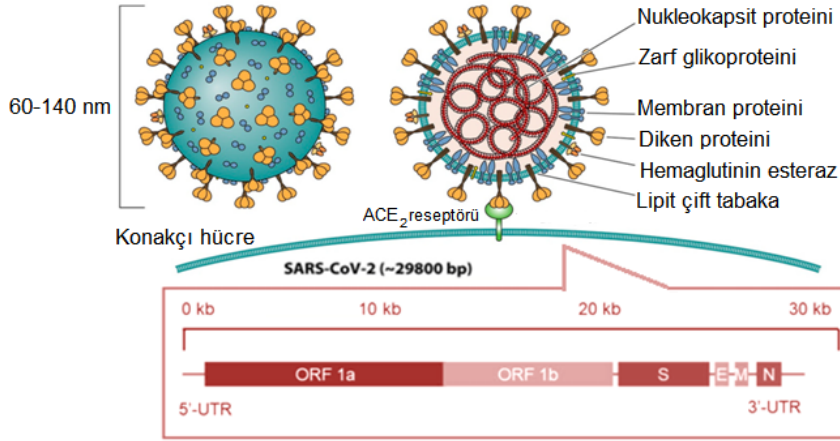
değişken üzerinde açıklayıcılığı yüksek bir parametrede ise bu yöntemlerden hangisi kullanılırsa kullanılacak karar ağacının tahmin potansiyeli ve doğruluğu olumsuz yönde etkilenir. Ayrıca, karar ağaçlarının dezavantajları arasında bazen diğer makine öğrenmesi algoritmalarına göre modeli eğitmek için daha uzun zaman gerektirmeleri ve bir ağaç oluşturabilmek için hesaplamaların daha karmaşık olması da sayılabilir [5,34].

2.2. Corona Virus Hastalığı 2019 (COVID-19)

Aralık 2019'da Çin Halk Cumhuriyeti'nin Hubei eyaleti Vuhan şehrinde bir deniz ürünleri ve kümes hayvanı pazarında ortaya çıkan ve hızla birçok ülkeye yayılan nedeni bilinmeyen pnömoni ve şiddetli akut solunum sendromu vakaları ortaya çıkmıştır. Ocak 2020'de bu duruma yol açan etkenin betacoronavirus tipinde yeni bir virus olduğu saptanmıştır. Genom yapısı, doku tropizmi ve viral patogenezi açısından SARS CoV'ye yakın benzerlik gösteren bu yeni virus "Şiddetli Akut Solunum Sendromu Corona virus 2" (SARS-CoV-2) olarak isimlendirilmiştir [35].

İnsandan insana çok kolay bulaşan ve bilinen koronavirüs enfeksiyonlarından daha farklı immün reaksiyonlara yol açan SARS-CoV-2 enfeksiyon tablosu, Corona virus hastalığı 2019'un kısaltması olarak COVID-19 olarak adlandırılmıştır [36]. Dünya Sağlık Örgütü'nün (DSÖ) 11 Mart 2020'de resmen pandemi olarak sınıflandırdığı COVID-19, morbidite ve mortalitesi yüksek bir hastalıktır. DSÖ verilerine göre Kasım 2022 itibarıyla dünya çapında 629 milyonu aşan bildirilen vaka sayısı ve 6.5 milyonu aşkın ölüme neden olmuştur Ülkemizde de güncel toplam vaka sayısının yaklaşık 17 milyon, COVID-19 kaynaklı ölüm sayısının ise 101.203 olduğu bildirilmektedir [37].

SARS-CoV-2 virus yapısal olarak 60–140 nm çapında olup başlıca diken protein, zarf glikoprotein, nükleokapsit, membran proteinini içeren yapısal proteinlerden oluşur. SARS-CoV-2 'nin hücrel enfeksiyonu sırasında viral yüzeyde bulunan diken proteinleri, konakçı hücreye anjiyotensin dönüştürücü enzim-2 (ACE-2) reseptörüne bağlanma yoluyla girer. Genomun 5' ucunda tüm genlerin en büyüğü olan ORF1ab, 3' terminalinde yapısal proteinleri içerir (Şekil 2.6) [38].



Şekil 2.6. SARS-CoV-2 virusunun yapısı

COVID-19 2003 yılında Hong Kong'da başlayıp 37 ülkeye yayılan ve şiddetli akut solunum yolu sendromuna yol açan SARS-CoV enfeksiyonu ile benzerlikler göstermekle birlikte daha bulaşıcı ve daha farklı immün reaksiyonlara yol açtığı bildirilmektedir [35]. Genellikle damlacık yoluyla presemptomatik, asemptomatik veya semptomatik bireylerden bulaşan COVID-19 enfeksiyonu, vakalarının büyük kısmında hafif bir klinik seyir göstererek birkaç hafta içinde iyileşmektedir [39]. Hastaların %20'sinde ise hastalığın başlangıcından 5-7 günlük bir süre sonra bağışıklık sisteminin hiperaktivasyonu ve kontrolsüz sitokin salınımına bağlı olarak hastaneye yatmayı hatta yoğun bakım ünitesine alınmayı gerektiren koagülopati, masif alveoler hasar, ilerleyici solunum yetersizliği gelişmektedir. Çoklu organ yetmezliği ve ölümlerle sonuçlanabilen bu durum sitokin fırtınası ile ilişkilendirilmekte olup $IFN\gamma$, $IL-1\beta$, $IL-6$, $TNF-\alpha$ gibi proinflatuvar sitokin ve kemokinlerin ekspresyonunun artışı ve doğal bağışıklık sisteminin hiperaktivasyonu ile ortaya çıktığı ileri sürülmektedir [35,40-42].

Diğer viral enfeksiyonlarda olduğu gibi SARS-CoV-2 de viral nükleik asit paterninin konakçı hücre tanıma reseptörlere (PRR'ler) bağlanması ile başlayarak adaptör proteinler, ubiquitin ligazlar ve kinazlar aracılığıyla aktive olan sinyal yollarının transkripsiyon faktörleri ve interferonlar (IFN), sitokinler ve kemokinler dahil çeşitli immün bileşiklerin ekspresyonunu uyarması ile sonuçlanmaktadır. IFN yolağı, viral enfeksiyonu ortadan kaldırmadaki hızlı ve etki potansiyeli nedeniyle genellikle ilk aktive olan yoldur. Ancak virüsün sinyal iletimine müdahale etmek ya da IFN ile uyarılan gen ürünlerinin inhibisyonu için bu yollara farklı seviyelerde etki edebildiği gösterilmiştir. SARS-CoV-2'nin doğal immüniteden kaçınmasında etkili stratejilerin viral nükleik asidin PRR'ler tarafından tanınmasını engelleyen çift membranlı

veziküllerinin olması ve bağışıklık sinyal moleküllerinin işlevselliğinin viral proteinler tarafından doğrudan ablasyonu olduğu ileri sürülmektedir [43,44]. SARS-CoV-2'nin diğer çeşitli solunum virusleri arasında insan bronşiyal epitel hücresinde tip I ve tip III IFN ekspresyonunu en güçlü şekilde baskılayan etken olduğu, IFN sinyalizasyonundan kaçınma ve periferik kandaki immün hücrelerde bozulmuş IFN üretiminin COVID-19 sırasında üretken viral replikasyona, transmisyona ve şiddetli patogeneze katkıda bulunduğu ileri sürülmektedir. Bu durumun klinik yansımaları kendini hastalık şiddetindeki değişkenliklerle belli etmektedir. Şiddetli vakalarda dolaşımda enflamatuvar sitokin/kemokin konsantrasyonlarının yüksek buna karşın IFN- β veya IFN- λ düzeylerinin çok düşük olması, dengesiz bir immün yanıt sergilemeleri ve bu durumun kalıcı viremi ile sonuçlanması IFN yolağı üzerindeki etkilerle ilişkilendirilmektedir [45,46].

COVID-19'lu hastalarda hastalık şiddeti genellikle klinik olarak gözlenen klinik semptomlar ve bulgular, tıbbi laboratuvar bulguları, radyografik bulgular ve organ disfonksiyonu gelişimine göre belirlenmektedir [39]. Buna göre;

- SARS-CoV-2 testi pozitif olup COVID-19 ile uyumlu herhangi bir klinik semptomu bulunmayan kişiler *asemptomatik veya preasemptomatik enfeksiyon*,
- Ateş, öksürük, boğaz ağrısı, halsizlik, baş ağrısı, kas ağrısı, bulantı, kusma, ishal, koku ve tat alma bozukluğu gibi COVID-19 semptomları olan ancak nefes darlığı veya akciğer grafisinde anormal bulgusu bulunmayan kişiler *hafif hastalık*,
- Alt solunum yolu hastalığına ilişkin klinik semptom veya radyolojik bulguları olan ancak oksijen saturasyonu (SpO_2) \geq %94 olan kişiler *orta şiddette hastalık*,
- $SpO_2 \leq$ %94 olup parsiyel arteriyel oksijen basıncının (PaO_2) solunan havadaki oksijen fraksiyonuna (FiO_2) oranı (PaO_2/FiO_2) $<$ 300 ya da dakikadaki solunum sayısı \geq 30 olan kişiler veya akciğer infiltrasyonu $>$ %50 olan kişiler *şiddetli hastalık*,
- Akut solunum yetmezliği, septik şok ve/veya çoklu organ disfonksiyonu olan kişiler *kritik hastalık* olarak sınıflandırılmaktadır. Şiddetli hastalık grubunda olan hastalar, semptomların başlamasından yaklaşık bir hafta sonra ortaya çıkma eğiliminde olan akut solunum sıkıntısı sendromunun (ARDS) gelişmesiyle kritik hastalık sınıfına geçebilmektedir.

COVID-19'un immün sistemdeki bu etkileri birçok laboratuvar parametresinin düzeyinde çoğu kez hastalık şiddeti ile ilişkilendirilebilen bir seri değişikliğe yol açmaktadır [42,47,48]. COVID-19'lu hastalarda laboratuvar bulgulara yansıyan başlıca

değişiklikler tam kan sayımı parametrelerinde özellikle lökosit ve lökosit alt gruplarının sayısında, üre, kreatinin ALT, AST, LDH, total CK, kardiyak troponin, d-dimer ve kalsiyum düzeyleri ile başta CRP ve ferritin olmak üzere bazı akut faz proteinlerinin düzeylerinde gözlenmektedir. Şiddetli vakalarda lökosit sayısındaki belirgin artış, buna karşılık lenfositlerdeki azalma dikkat çekicidir. Nötrofil sayısı ve nötrofil-lenfosit oranındaki (NLR) belirgin artış kritik hastalık ve kötü prognozla ilişkilendirilmektedir. COVID-19'un platelet sayısında da değişikliğe yol açabildiği, ciddi vakalarda sayılarının anlamlı olarak azaldığı bildirilmektedir. LDH, CK, AST, ALT, üre, kreatinin ve kardiyak troponin düzeylerindeki değişikliklerin hafif vakalarda çoğu kez belirgin olmadığı ancak hastalığın ciddiyeti ile ilişkili olarak yükseldikleri gözlenmektedir. Bu parametrelerdeki artışın süreklilik göstermesi ise genellikle şiddetli hastalık ve kötü prognoz riski ile ilişkilendirilmektedir. COVID-19'da ortaya çıkan laboratuvar bulguların bir kısmı da pıhtılaşma ve fibrinolitik sistem değişiklikleridir ve kendini genellikle d-dimer düzeylerinde artış ile belli etmektedir. Şiddetli vakalarda d-dimer düzeylerindeki artışın sürekliliği mortalite ile ilişkilendirilmiştir. COVID-19'lu hastalarda kötü prognoz ve mortalite riski ile ilişkilendirilen diğer bir önemli parametrede CRP'dir. Yüksek CRP düzeylerine sahip vakaların mortalite riskinin yaklaşık dört kat fazla olduğunu bildirilmektedir. COVID-19'da laboratuvar parametrelerde ortaya çıkan başlıca değişiklikler bunların hastalığın prognozu ile ilişkileri Tablo 2.1'de özetlenmiştir [49].

Diğer taraftan COVID 19 ile ilgili bilgi ve tecrübeler arttıkça hastalığın takip ve risk belirlenmesinde kullanılacak potansiyel bazı yeni parametreler de önerilmektedir. Bunların arasında dikkat çekenlerden birisi homosistein düzeyidir. Homosistein düzeyinin hastalığın başlangıcından hastaneye yatışa kadar geçen süre içinde yaş ve monosit-lenfosit oranı birlikte değerlendirilmesinin şiddetli pnömoni için spesifik bir kardiyovasküler risk öngörücü etken olabileceği bildirmektedir. Diğer bir parametre ise SARS-CoV-2'nin viral bağlanma bölgesi olarak ACE₂ reseptörleri ile ilişkisinden kaynaklanmaktadır. Bağlanma sonucu ACE₂'de ortaya çıkan işlev kaybının artmış Anjiotensin (Ang) II ve azalmış Ang (1-7), Ang (1-9) ve alamandin düzeylerine yol açtığı, bu durumun hastalık şiddeti ile de doğru ilişkili olduğu ileri sürülmektedir [42].

Tablo 2.1. COVID-19’da laboratuvar parametrelerde ortaya çıkan başlıca değişiklikler

Biyobelirteç		Hastalık Şiddeti			Prognoz	
		Hafif	Şiddetli	Kritik	Sağ kalım	Vefat
Hematolojik	Lenfosit	N/↓	↓↓	↓↓↓	N/↓	N/↓↓
	WBC	N/↓/↑	N/↓↓/↑	N/↓/↑↑	N/↓/↑	N/↓/↑/↑↑
	Nötrofil	N [#]	N/↑*			
Biyokimyasal	LDH	N/↑	N/↑/↑↑	↑/↑↑/↑↑↑	N/↑	↑↑↑
	AST	N/↑	N/↑	N/↑↑		
	ALT	N/↑	N/↑	N/↑↑	N/↑	N/↑↑
	CK	N/↑	N/↑	N/↑↑	N/↑	N/↑↑
	Kreatinin	N	N	N/↑	N	N/↑
	cTnI	N [#]	N/↑*		N	N/↑↑
	Üre	N [#]	N/↑*			
Hemostatik	Albümin				N/↓	↓↓
	D-dimer	N/↑	N/↑↑	N/↑↑↑	N/↑	↑/↑↑/↑↑↑
	Fibrinojen				N/↑	N/↑/↑↑
	PT				N	N/↑
Enflamatuvar	APTT				N	N/↑
	CRP	N/↑	N/↑/↑↑	↑/↑↑/↑↑↑	↑↑	↑↑↑
	PCT	N/↑	N/↑/↑↑		N/↑	N/↑/↑↑/↑↑↑
	Ferritin	↑/↑↑	↑↑/↑↑↑		N/↑/↑↑	↑↑/↑↑↑
	IL-6	N/↑	↑/↑↑		N/↑	↑/↑↑
	IL-2R	N/↑	N/↑/↑↑			
TNF-α	N/↑	N/↑/↑↑				

N: Normal; ↑: Hafif artış; ↑↑: Orta düzey artış; ↑↑↑: Şiddetli artış; ↓: Hafif azalma; ↓↓: Orta düzey azalma; ↓↓↓: Şiddetli azalma; [#]: Yoğun bakım hastası olmayan; *: Yoğun bakım hastası

2.3. COVID-19’DA TIBBİ LABORATUVAR ve YAPAY ZEKA

COVID-19 enfeksiyonunun ortaya çıkması ve küresel çapta yayılması ile hastalığı anlamak, mücadele etmek ve kontrol altına almak için gösterilen çabalar birçok bilimsel çalışmaya da konu olmuştur. Bu çalışmaların bir kısmında hastalılık şiddetini belirlemek ve mortalite riskin öngörmek için makine öğrenmesi yöntemleri kullanılarak tek başına veya klinik bulgular ile birlikte tıbbi laboratuvar sonuçlarının

değerlendirilmesinden de yararlanılmıştır. Pandeminin en şiddetli olduğu dönemde COVID-19 için güvenilir bir risk sınıflandırması ve prognoz tahmini yapmaya yardımcı olacak nicel bir araç geliştirmeye yönelik yapay zeka destekli bu çalışmaların çoğu, hastanede yatan kısıtlı sayıda hastanın dahil edildiği küçük veri setlerinde yürütülmüş olsa da, klinik pratikte yararlı olabilecek sonuçlara ulaştığı görülmektedir. İleri yaş, lenfosit ve trombosit sayılarının düşük, lökosit ve nötrofil sayılarının yüksek olması, CRP, LDH, BUN ve cTn düzeylerinin yüksekliği çoğu çalışmada kötü prognoz ve mortalite ile ilişkilendirilmiştir [50-52].

Bu çalışmalardan birinde Booth AL ve arkadaşları hastane içi mortalite riskini tahmin etmek için yaklaşık 400 hastanın CRP, BUN, serum kalsiyum, albümin ve laktik asit düzeylerini destek vektör makinesi yöntemi ile retrospektif olarak değerlendirerek %91 duyarlılık ve %91 özgüllük gösteren bir model geliştirmiştir. Bu modelde COVID-19 enfeksiyonunda prognozu etkileyen önemli laboratuvar bulgularının yüksek CRP, BUN ve laktik asit ile azalmış kalsiyum ve albümin düzeyleri olduğunu saptamışlardır [50]. Yine mortalite riskinin öngörülmesinde COVID-19 tanısı almış 306 hastanın klinik laboratuvar sonuçlarının sekiz farklı makine öğrenme algoritması ile değerlendirildiği bir başka çalışmada ise serum CRP ve LDH yüksekliğinin mortalite riskini arttırdığı, lenfosit sayısı ile CRP ve LDH düzeyleri arasında ters bir ilişki olduğunu saptamıştır [53]. Assaf ve arkadaşları ise COVID-19 hastalarında hastaneye yatış sırasında solunum yetmezliği, yoğun bakım ünitesine yatma, çoklu organ yetmezliği ve/veya ölüm gelişme riskin doğru tahmin etmek ve hastane içi tıbbi kaynakları daha iyi önceliklendirmek için 162 COVID-19 hastasının CRP ve ferritin dahil rutin biyokimya, tam kan sayımı ve koagülasyon test sonuçlarını üç farklı makine öğrenmesi algoritması (yapay sinir ağları, random forest ve CART) ile retrospektif olarak değerlendirmiş ve elde edilen modellerin başarısını APACHE II, SOFA ve NEWS2 gibi halihazırda kullanılmakta risk öngörüsünün puanlama ile yapıldığı yöntemler ile karşılaştırmıştır. Elde edilen sonuçlar kritik hastaların ön görülmesinde en etkili bulguların yüksek lökosit ve nötrofil sayıları, yüksek CRP, AST, LDH ve düşük albümin düzeyleri olduğunu ve makine öğrenmesi modellerinin mevcut risk puanlama yöntemlerine göre daha gelişmiş tahmin ve yüksek doğruluk sağladığını ortaya koymuştur [54]. Benzer şekilde Yan ve arkadaşları da COVID-19'da mortaliteyi azaltmak için yüksek riskli hastaları doğru ve hızlı bir şekilde öngörebilen basit ve uygulanabilir bir karar kuralı önermek amacıyla makine

öğrenmesinden yararlanmıştır. Durum bir sınıflandırma görevi olarak formüle edilmiş ve COVID-19 tanısı konmuş 485 hastada mortalite için en ayırt edici biyobelirteçler makine öğrenmesi algoritmalarına dayalı iki aşamalı bir modelleme yaklaşımı ile saptanmıştır. Yaş, cinsiyet, başlangıç karaciğer ve böbrek fonksiyonu, elektrolitler, enflamatuvar belirteçler ve koagülasyon testlerinin sonuçları, klinik semptom ve bulgular ölüm ve sağkalım ile ilişkilendirilerek XGBoost sınıflama algoritması ile retrospektif olarak değerlendirilmiş. Sonra da yine aynı algoritma kullanılarak optimizasyon yoluyla, LDH, lenfosit sayısı ve hs-CRP'nin ileriki on gün içinde ölüm riski olan hastaları %90'dan fazla doğrulukla ayırt edebilen biyobelirteçler olduğu saptanmıştır. Hatta nispeten yüksek LDH düzeylerinin acil tıbbi müdahale gerektiren vakaların büyük çoğunluğunu ayırt etmede tek başına çok önemli bir rol oynadığı sonucuna varılmıştır [55].



3. GEREÇ VE YÖNTEM

3.1. Çalışmanın Tasarımı

3.1.1. Çalışma Popülasyonu

Vaka serisinde retrospektif olarak yürütülen tanımlayıcı tipteki bu çalışma 01 Mart 2020-31 Aralık 2020 tarihleri arasında COVID-19 şüphesi ile Bursa Çekirge Devlet Hastanesi'ne başvuran ve COVID-19 tanısı nazofarengal ve/veya orofarengal sürüntü örneklerinde SARS-CoV-2 RNA'sının özgül dizileri gerçek-zamanlı reverse transkripsiyon polimeraz zincir reaksiyonu (rRT-PCR) testi ile doğrulanmış 18 yaş ve üzerindeki hastalarda yürütülmüştür.

3.1.2. Kullanılan Biyokimyasal Parametreler

COVID-19 tanısı doğrulanmış hastalarda Bursa Çekirge Devlet Hastanesi'nde ilk başvuruda rutin olarak istenen biyokimyasal, hematolojik ve immünokimyasal laboratuvar testleri aşağıdaki yer almaktadır.

Rutin Biyokimyasal Parametreler: Serum glikoz (mg/dL), Üre (mg/dL), kreatinin(mg/dL), alanin amiotransferaz (ALT, U/L), aspartat aminotransferaz (AST, U/L), laktat dehidrogenaz (LDH, U/L), sodyum (Na, mmol/L), potasyum (K, mmol/L), klor (Cl, mmol/L), Kalsiyum (Ca, mg/dL), C-reaktif protein (CRP, mg/L), ferritin (µg/L) ve kardiyak troponin-I (cTnI, pg/mL).

Tam Kan sayımı Parametreleri: Eritrosit sayısı (RBC, $\times 10^6/\mu\text{L}$), hemoglobin (HGB, g/dL), hematokrit (HCT, %), ortalama eritrosit hacmi (MCV, fL) ortalama eritrosit hemoglobini (MCH, pg) ortalama eritrosit hemoglobin konsantrasyonu (MCHC, g/dL), eritrosit dağılım genişliği (RDW, fL), lökosit sayısı (WBC, $\times 10^3/\mu\text{L}$), lökosit alt gruplarının mutlak sayısı ($\times 10^3/\mu\text{L}$) ve yüzdeler (%) ; nötrofil (NE), lenfosit (LYM), monosit (MONO), eozinofil (EOS), bazofil (BASO) ve trombosit sayısı (PLT, $\times 10^3/\mu\text{L}$), ortalama trombosit hacmi (MPV, fL), trombosit dağılım genişliği (PDW, fL), plateletkrit (PCT, %)

Hemosataz Parametreleri: protrombin zamanı (PT, saniye), aktive parsiyel tromboplastin zamanı (aPTT, saniye) ve d-dimer (µg/L).

Tüm biyokimyasal testler hastadan kan örneğinin alındığı gün çalışılmış ve test sonuçları biyokimya uzmanı tarafından onaylanmıştır. Rutin biyokimyasal parametreler serum örneklerinde Cobas 6000 ve Roche Cobas 8000 (Roche Diagnostics, Mannheim, Almanya) kimya analizöründe çalışılmıştır.

CRP düzeyi nefelometrik yöntemle Immage 800 (Beckman Coulter, Kaliforniya, ABD) spesifik protein analizörü ile tayin edilmiştir.

Ferritin ve cTnI düzeyleri immünokemilüminesans yöntemle ADVIA Centaur XP (Siemens Healthcare Diagnostics, Malburg, Almanya) analizör ile ölçülmüştür.

PT ve aPTT BCS® XP System (Siemens Healthcare Diagnostics, Malburg, Almanya) tam otomatik hemostaz analizörü ile saptanmıştır.

d-Dimer düzeyleri partikül bazlı immünotürbidimetrik yöntemle INNOVANCE® (Siemens Healthcare Diagnostics, Malburg, Germany) analizörde ölçülmüştür.

Tam kan sayımı LH 780 otomatik hematoloji analizöründe (Beckman Coulter Pasadena, Kaliforniya, ABD) yapılmıştır.

3.1.3. Verilerin Toplanması, Birleştirilmesi ve Temizlenmesi

Bursa İl Sağlık Müdürlüğü Halk Sağlığı Yönetim Sistemi'nin Filyasyon ve İzolasyon Takip Sistemi (FİTAS) ve Bursa Çekirge Devlet Hastanesi Hastane Bilgi İşletim Sistemi (HBİS) birlikte incelenerek 01 Mart 2020-31 Aralık 2020 tarihleri arasında Bursa Çekirge Devlet Hastanesine başvuran ve COVID-19 tanısı doğrulanmış 18 yaş ve üzerinde 25998 hasta olduğu görülmüştür.

COVID-19 tanısı doğrulanmış bu 25998 hastanın yaş, cinsiyet ve uyruğu gibi demografik özellikleri, diyabet, hipertansiyon, kardiyovasküler hastalık, astım, kanser, karaciğer ve böbrek yetersizliği gibi eşlik eden hastalıkları ve biyokimyasal test sonuçları HBİS'den alınmıştır. Bursa Çekirge Devlet Hastanesi veri tabanında hasta verileri kimlik ve demografik bilgiler, klinik ve tedavi bilgileri, laboratuvar bilgileri, taburcu ve ölüm kayıtları şeklinde farklı başlıklar altında kaydedilmektedir. Çalışma verilerinin elde edilebilmesi için HBİS'deki bu farklı başlıklara her hasta için ayrı ayrı girilerek verilerin sağlanması gerekmiştir.

Hasta mahremiyetinin korunması amacıyla verilerin toplanması aşamasından itibaren Kişisel Verilerin Korunması Kanunu (KVKK) ve Avrupa Birliği'ne uyum çalışmaları kapsamındaki General Data Protection Regulation (GDPR) gerekliliklerine uyulmuştur. Bu amaçla elde edilen her hasta verisi, verinin biçimsel ve anlamsal bütünlüğü bozulmayacak şekilde ad-soyad, T.C kimlik numarası gibi doğrudan tanımlayıcı veriler, hastalık prognozu gibi hassas veriler ve uyruk, cinsiyet ya da yaş gibi dolaylı tanımlayıcı verilerin bazı özellikleri veya tamamı belirlenen bir kod sistemine göre kodlanarak anonimleştirilmiştir.

Hastane veri tabanındaki farklı başlıklar altından elde edilen hasta verileri tek bir tabloda birleştirildiğinde yabancı uyruklu hastaların çoğunda demografik bilgiler dahil birçok verinin eksik olduğu, özellikle ayaktan takip edilen hastalarda prognoz bilgisinin ya da eşlik eden hastalık bilgilerinin bulunmadığı ve/veya laboratuvar testlerinin sonuçlarının çoğunun eksik olduğu görülmüştür. Bu eksik verilere için uygun birer değer atanması için çeşitli yöntemler bulunmaktadır. Ancak, geliştirilecek modelde atanan bu değerlerin anlamlı bir sonuç ya da ayırt edici bir özellik olarak karşımıza çıkması durumunda modelin yorumlanması aşamasında güçlüklerle ve/veya hatalı sonuçlara yol açabileceği bildirilmektedir. Veri setinin dağılımında değişikliğe ve hatalı bir eğilime yol açmaması için yabancı uyruklu hastalar, prognoz bilgisi bulunmayan hastalar ve ilk başvuruda rutin olarak istenen laboratuvar testlerine ait sonuçları eksik olan hastalar çalışmaya dahil edilmemiştir. Veri setinin çoğunu oluşturan ayaktan hastaların nerede ise tamamında eşlik eden hastalık bilgileri bulunmadığı için bu parametrelerde çalışmada kullanılmamıştır. Sonuç olarak çalışmaya HBİS'de ilk başvuruda bulunması istenen biyokimyasal parametrelerin tamamı için kayıtlı test sonucu ve hastalığın prognoz bilgisi bulunan 18 yaş ve üzerindeki 8036 Türkiye Cumhuriyeti vatandaşı COVID-19 tanısı doğrulanmış hasta dahil edilmiştir. Çalışma grubunu oluşturan 8036 hastanın 6035 tanesi ayaktan izlenmiş, 2001 tanesi hastanede yatarak tedavi görmüş ve bunlardan 316 tanesi COVID-19 nedeniyle vefat etmiştir (Şekil 3.1.3).



Şekil 3.1. Hastaların istatistiksel sonucu

3.2. Yasal İzinler ve Etik Onay

Çalışma T.C Sağlık Bakanlığı Sağlık Hizmetleri Genel Müdürlüğü bünyesindeki COVID-19 Bilimsel Araştırma Değerlendirme Komisyonu'nun izni takiben Bursa İl Sağlık Müdürlüğü'nün izni ve Bursa Şehir Hastanesi Klinik Araştırmalar Etik Kurulunun 19.08.2020 tarih ve 2020-5/9 no'lu etik onayı ile başlatılmıştır. Çalışmanın her aşaması Helsinki Deklerasyonu ve bu deklereyona dayanan İyi Klinik Uygulamalar ilkelerine uygun olarak yürütülmüştür. Ayrıca, yayınladığımız verilerin farklı kaynaklardan elde edilen verilerle eşleştirilerek kişilerin mahremiyetinin ifşasına yol açmaması için tüm sonuçlar prognoza ve/veya parametrelerin kesim değerlerine göre gruplandırılmış yüzdeler veya ortalamalar olarak sunulmuştur.

3.3. Verilerin Analizi

İstatistiksel analizler anonimleştirilmiş veriler kullanılarak IBM SPSS Statistics for Windows, version 28.0 (IBM Corp., Armonk, NY, ABD) paket programı ile gerçekleştirilmiştir. Kategorik değişkenler sayı ve yüzde (%) olarak, sürekli değişkenler ortalama \pm standart sapma (SD) ve minimum-maksimum olarak ifade edilmiştir.

COVID-19 tanısı almış hastalarda prognozun öngörülmesinde biyokimyasal parametrelerin rolünün araştırıldığı bu çalışmada veri seti çoğunlukla kesiksiz (sürekli) değer alan laboratuvar test sonuçları ve kesikli (süreksiz) değer alan cinsiyet, ayaktan takip, yatarak takip, taburcu olma ve vefat gibi kategorik verilerden oluşmaktadır. Bu veriler lojistik regresyon analizi ve denetimli makine öğrenmesi yönteminin bir türü olan Karar Ağaçları ile incelenmiştir.

3.3.1. Lojistik Regresyon Analizi

Yeterli örneklem (8036 hasta) ve veri (321520) büyüklüğüne sahip olan çalışmamızda (i) bağımsız değişkenlerin normal dağılması, doğrusallık ve varyans-kovaryans matrislerinin eşitliği gibi varsayımların karşılanmasını gerektirmemesi, (ii) bağımsız değişkenlerin sürekli ya da süreksiz (kategorik) olmasına yönelik hiçbir kısıtlama getirmemesi, (iii) bağımlı değişkenin ikiden fazla kategoride olmasına olanak sağladığı için ve (iv) katsayıların tahmininde yanlılığa ve modelin yetersizliğine yol açmamak için verileri eksik olan hastalar çalışmaya alınmadığı için çok değişkenli lojistik regresyon analizi uygulanmıştır.

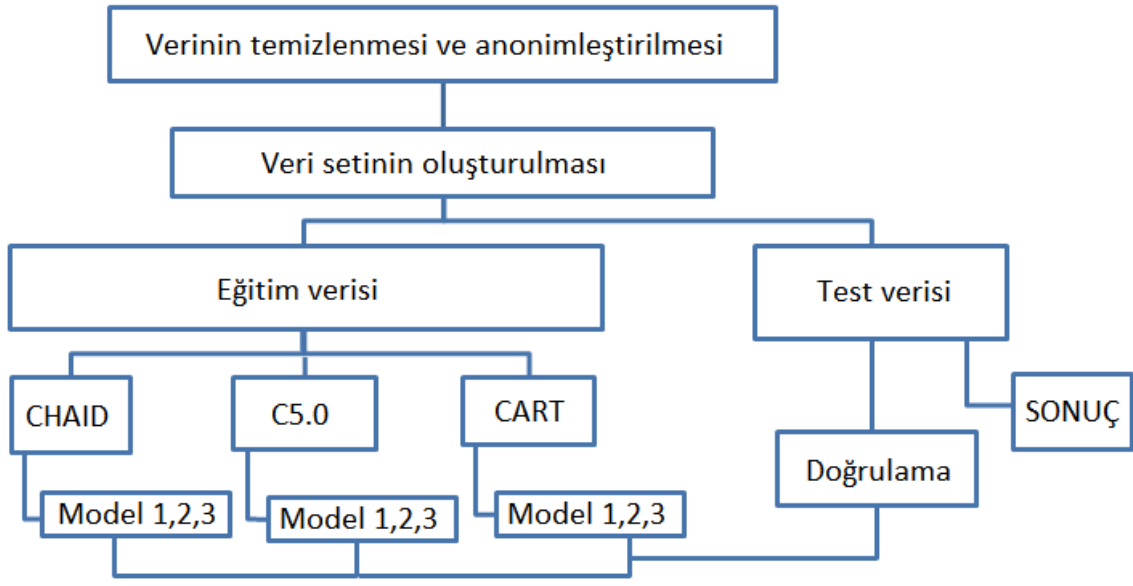
Lojistik regresyon modelinde, COVID-19 tanısı doğrulanmış vakalar prognozlarına göre ayaktan takip edilen, yatarak tedavi edilen, yatarak tedavi edilenler de iyileşerek taburcu olan ve vefat eden şeklinde gruplanarak bağımlı değişkenleri; yaş, cinsiyet ve biyokimyasal test sonuçları bağımsız değişkenleri oluşturmuştur. COVID-19 enfeksiyonunun prognozunu belirlemede ayırt edici güce sahip parametre bilinmediği için, tüm değişkenlerin analize dahil edildiği ve daha fazla atılacak değişken kalmayana kadar her adımda ayırt edici güce katkı sağlamayan ya da en az katkıyı sağlayan değişkeninin çıkarılması prensibine dayanan geriye-doğru seçim (backward stepwise) lojistik regresyon yöntemi uygulanmıştır. Lojistik fonksiyonda yer alan “ayırt edici olmanın”, “ayırt edici olmamaya” oranı $(\pi(x_i) / (1-\pi(x_i)))$ ile gösterilip “odds” olarak bilinmektedir. Lojistik regresyon modelinde değişkenlere ait katsayıların yorumu için odds oranından $(OR = \exp(\beta))$ yararlanıldı. OR değeri 0 ile 1 arasında olursa risk faktörünün sonuç değişkeni için “koruyucu” olduğu, OR değeri 1 olursa risk faktörü ve sonuç değişkeni arasında bir fark olmadığı ve OR değeri 1’den büyük olursa risk faktörü ve sonuç değişkeni arasında bir fark olduğu ve bu farkın matematiksel olarak bir kat ile açıklanacağı şeklinde yorumlanmaktadır. Bağımlı değişkenlere ait katsayıların

prognoz üzerindeki etkisinin değerlendirilmesinde %95 güven aralığı (CI) içinde OR değerleri kullanıldı.

Modelde bulunması gereken tüm değişkenler modele alındıktan sonra, lojistik regresyon analizi sonucu elde edilen modelin sonuç değişkenini açıklamadaki etkinliği (uyum iyiliği) model ki-kare testi ile, her bir bağımsız değişkenin modelde bulunmasının anlamlı olup olmadığı ise Wald istatistiği ile test edilmiştir. Modelin uygunluğunun değerlendirilmesinde bir sonraki adımda bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin derecesini gösteren Cox-Snell R^2 ve Nagelkerke R^2 istatistikleri kullanılmıştır. Ki-kare $p < 0.05$, Cox-Snell R^2 ve Nagelkerke $R^2 > 0.2$ olması durumunda model anlamlı kabul edilmiştir.

3.3.2. Karar Ağacı Analizleri

COVID-19 tanısı almış hastalarda prognozun öngörülmesinde biyokimyasal parametrelerin rolünü incelemek için öncelikle mevcut karar ağacı algoritmaları arasından veri setimize ve amacımıza en uygun olası algoritmaların belirlenmesi için ön çalışmalar yapılmıştır. Kategorik ve sürekli tüm değişken tipleriyle çalışabilmesi ve ağaçtaki her düğümü ikiden fazla alt gruba ayırabilmesi, geniş örnek büyüklüklerinden yararlanarak sağlık verileri gibi çoğu kez doğrusal olmayan karmaşık ilişkilerin varlığında bile güçlü öngörü potansiyeline sahip olması, çıktıları kolay ve anlaşılır şekilde yorumlaya olanak sağlaması nedeni ile Otomatik Ki-Kare Etkileşim Belirleme (Chi-Squared Automatic Interaction Detector, CHAID), C5.0 ve Sınıflama ve Regresyon Ağacı (Classification and Regression Tree, CART) algoritmaları ile ön çalışmalar yapılmıştır. Çalışmada kullanılacak karar ağacı algoritmasının belirlenmesine yönelik bu çalışmalar için veri setinin %70'i eğitim verisi, %30'u test verisi olarak ayrıldıktan sonra CHAID, C5.0 ve CART algoritmaları ile farklı modeller oluşturulmuştur (Şekil 3-2). CART algoritması ile elde edilen sonuçların daha uygun düğüm sayıları, dal/yaprak kombinasyonu içerdiği ve modelin çıktısının en yüksek doğruluk oranına (overall confidence) sahip olduğu görülmüş ve çalışmanın CART algoritması kullanılarak yürütülmesine karar verilmiştir.



Şekil 3.2. Çalışmada kullanılacak karar ağacı algoritmasının seçimi

Deneysel çalışmalar, tüm veri kümesinden elde edilen tek bir karar ağacının doğruluğunun eğitim ve test verisi olarak bölünmeden daha fazla olduğunu bildirdiğinden [21] çalışmada CART algoritmasının kullanılması kararına varıldıktan sonra incelemeler veri setinin tamamı kullanılarak yürütülmüştür.

Hastalığın henüz başlangıç aşamasındayken, COVID-19 prognozunun doğru öngörüsülmesini sağlayabilecek biyokimyasal değişken(ler) ve bu değişkenlerin takibinde izlenecek yolu belirlemek üzere üç ayrı model oluşturulmuştur:

Model-1: Hastaların ayaktan veya yatarak takip (prognoz) öngörüsü

Model-2: Tüm hastalarda hastalığın başlangıcında mortalite öngörüsü

Model-3: Yatarak takip edilen hastaların mortalite (hastane içi mortalite) öngörüsü.

CART algoritması ile karar ağacı oluşumunda en iyi dallanmayı sağlayan değişken(ler) için en doğru değer seçilmesi ve hastaların minimum sınıflandırma hatasıyla yerleştirilmesi için çalışmamızda dallanmalar safsızlık fonksiyonuna dayalı dallandırma kriteri olan twoing ayırma kriterine (ikili ayırma kuralı) göre yapılmıştır. Bu kuralda, bir düğüme ait verilerin %50'sini içeren ve birbirine benzemeyen sınıflar oluşturmak üzere değişkenler sahip oldukları değere göre eğitim kümesi ve aday bölünme olarak iki dala ayrılır. Bu sayede bir t düğümünde *sağtsol* ve *sağtsağ* olacak şekilde iki dal oluşturulur. Karar ağacı oluşumunda kullanılacak her bir veri sağ ve sol

dalda yer alabileceğinden önce her bir adayın sağ vesol dalda bulunma olasılığı ($P_{sağ}$ ve (P_{sol}) hesaplanır, $P(j|t_{sağ})$ ve $P(j|t_{sol})$ olarak ifade edilir.

t : twoing kriterine göre dallanmanın oluşacağı düğümü

P_{sol} ve $P_{sağ}$: veri setindeki değerlerin solda ya da sağda olma olasılığını

$(j|t_{sol})$: j sınıf değerinin sol tarafındaki sınıfta olma olasılığını

$(j|t_{sağ})$: j sınıf değerinin sağ tarafındaki sınıfta olma olasılığını gösterir.

Buna göre n sayıda hasta içeren t düğümündeki s aday bölünmelerinin uygunluk ölçüsü,

$(s|t) = 2P_{sol}P_{sağ} \sum_{j=1}^n |P(j|t_{sol}) - P(j|t_{sağ})|$ işlemi ile hesaplanır. Bulunan en büyük sonuç uygunluk ölçüsü olarak seçilerek kök düğümde yer alacak parametrenin hangisi olacağı belirlenir ve dallandırma yapılır. Karar ağacının tüm yapraklarına ulaşıncaya kadar aynı işlem aşağıya doğru tekrarlanır.

CART algoritması ikili ağaçların iyileştirilmesi ve saflığı arttıran yeni bölünmeler bulunduğu bölünmeye devam etmesini ve yaprak başına en düşük öngörülede bulunan dalların budanmasını da sağlayan bir düzeneğe sahiptir. Buna göre algoritmanın işleyiş sürecinde her biri, model karmaşıklığı ve yanlış sınıflandırma oranı arasında farklı bir dengeyi tanımlayan bir grup aday alt ağaç oluşturulur. Bu aday alt ağaçlar validasyon grubu olarak kullanılır ve düzeltilmiş hata oranına göre tekrarlanan budama prosedürü ile zayıf dallar budanır. Sonuçta, doğrulama seti minimum yanlış sınıflandırma oranına sahip ağaç son model olarak seçilir. Bu nedenle çalışmamızda elde edilen karar ağaçları için ayrıca bir budama algoritması uygulanmamıştır.

Modellerden tarafından doğru veya yanlış öngörülen hasta sayıları 2×2 'lik karışıklık matrisine yerleştirilerek modellerin performansları aşağıdaki ölçütler kullanılarak hesaplanmıştır [55]. Karışıklık matrisinde CART algoritması elde edilen modelin;

Doğru olarak öngördüğü pozitif sınıf örneklerinin sayısı, Gerçek Pozitif (GP),

Doğru olarak öngördüğü negatif sınıf örneklerinin sayısı, Gerçek Negatif (GN),

Yanlış olarak öngördüğü negatif sınıf örneklerinin sayısı, Yanlış Pozitif (YP),

Yanlış olarak öngördüğü pozitif sınıf örneklerinin sayısı Yanlış Negatif (YN) olarak sınıflanmıştır (Tablo 3.1).

Tablo 3.1. Sonuçların özetlendiği karışıklık matrisi

		GERÇEK DURUM	
		İSTENEN DURUM	İSTENMEYEN DURUM
MODELİN ÖNGÖRÜSÜ	POZİTİF	GP	YP
	NEGATİF	YN	GN

3.3.2.1. Modellerin performanlarının hesaplanması

Geri Çağırma (Recall): Duyarlılık (sensitivity) olarak da bilinen bu ölçüt modelin hedeflenen durumu saptama kabiliyeti olarak da tanımlanabilir. Modelin doğru olarak öngördüğü pozitif hastaların, tüm gerçek pozitif hastalara oranıdır. Recall değeri $0 < R < 1$ arasında bir değişim gösterir; 1'e yakın bir recall değeri modelin gerçek pozitifleri kaçırmadığını, doğru ve yanlış etiketleme arasında iyi bir sınıflandırma yapabildiğini gösterirken düşük sonuçlar (< 0.5), dengesiz sınıf veya iyi ayarlanmamış model hiperparametrelerine bağlı olarak çok sayıda yanlış negatif öngöründe bulunduğunu göstermektedir. Geri çağırma aşağıdaki formül ile hesaplanmış ve sonuçlar yüzde (%) olarak ifade edilmiştir.

$$\text{Geri Çağırma} = \text{GP} / (\text{GP} + \text{YN})$$

Özgüllük (Specificity): Gerçek negatif oranı (true negatif rate) olarak da bilinen bu ölçüt, modelin istenmeyen durumları saptama kabiliyeti olarak da tanımlanabilir. Modelin doğru olarak öngördüğü negatif hastaların, tüm gerçek negatif hastalara oranıdır, sonuçlar duyarlılığa benzer şekilde değerlendirilir. Özgüllük aşağıdaki formül ile hesaplanmış ve sonuçlar yüzde (%) olarak ifade edilmiştir.

$$\text{Özgüllük} = \text{GN} / (\text{GN} + \text{YP})$$

Kesinlik (Precision): Hassasiyet ya da Pozitif Prediktif Değer (positive predictive value) olarak da bilinir, doğru pozitif öngörülerin, tüm pozitif öngörülere oranıdır. Yanlış pozitiflere odaklanan bu ölçüt, modelden elde edilen sonuçların ne kadar geçerli olduğunun yani modelin pozitif olarak öngördüğü hastaların gerçekte kaç tanesinin pozitif olduğunun göstergesidir. Kesinlik değeri $0 < P < 1$ arasında bir değişim gösterir;

1'e yakın bir kesinlik değeri modelin hiçbir gerçek pozitif hastayı kaçırmadığını, doğru ve yanlış etiketleme arasında iyi bir sınıflandırma yapabildiğini gösterirken düşük sonuçlar (<0.5), modelin çok sayıda yanlış pozitif öngöründe bulunduğunu gösterir. Kesinlik aşağıdaki formül ile hesaplanmış ve sonuçlar yüzde (%) olarak ifade edilmiştir.

$$\text{Kesinlik} = \text{GP} / (\text{GP} + \text{YP})$$

Doğruluk (Accuracy): Bir modelin başarısını ölçmek için çok kullanılan bir ölçüt olan doğruluk, modelin doğru öngördüğü sonuçların toplam sonuçlara oranıdır. Aşağıdaki formül ile hesaplanmış ve sonuçlar yüzde (%) olarak ifade edilmiştir.

$$\text{Doğruluk} = (\text{GP} + \text{GN}) / (\text{GP} + \text{YP} + \text{YN} + \text{GN})$$

Dengelenmiş Doğruluk (Balanced Accuracy): Bir ikili sınıflandırmada sınıfların dağılımı dengesiz ise, yani verilerin çoğu iki sınıftan birinde yığılım gösteriyor ise sonuçların tek başına doğruluk ile değerlendirilmesi yeterli olmamaktadır. Bu durumda modelin başarısı Dengeli Doğruluk ile belirlenmiştir. Dengeli Doğruluk aşağıdaki formül ile hesaplanmış ve sonuçlar yüzde (%) olarak ifade edilmiştir.

$$\text{Dengelenmiş Doğruluk} = \text{Geri Çağırma} + \text{Özgüllük} / 2$$

F₁Skoru: Kesinlik ve geri çağırmanın harmonik ortalamasıdır. Sınıfların dengesiz dağıldığı durumlarda kesinlik ve geri çağırma arasında bir denge sağlayarak modelin doğruluğu hakkında karar vermek için daha uygun bir ölçüt olduğu bildirilmektedir. Ancak, yüksek bir F₁skoru, hem kesinlik hem de geri çağırmanın yüksek olduğu konusunda bilgi verirken düşük bir F₁skoru sorunun ne olduğu (keskinlik mi düşük, geri çağırma mı düşük) konusunda bilgi sağlamaz. F₁skoru aşağıdaki formül ile hesaplanmış ve sonuçlar yüzde (%) olarak ifade edilmiştir.

$$F_1 \text{ Skoru} = 2[(\text{Kesinlik} \times \text{Duyarlılık}) / (\text{Kesinlik} + \text{Duyarlılık})]$$

4. BULGULAR

4.1. Demografik Bulgular ve Laboratuvar Test Sonuçları

Çalışma grubunu 4140 kadın, 3896 erkek olmak üzere toplam 8036 hasta oluşturmuştur. Hastaların yaş ortalaması 51.38 ± 17.10 'dur. Yatarak tedavi edilen hastaların %94'ü iyileşerek taburcu edilirken %16'sı vefat etmiştir. Çalışma grubunu oluşturan tüm COVID-19 tanısı konulmuş hastalardan %3.9'u vefat etmiştir (Tablo 4.1).

Tablo 4.1. COVID-19 tanısı ile takip edilen hastaların demografik bulguları

	Ayaktan (n= 6035)	Yatan (n= 2001)	Taburcu (n= 1685)	Vefat eden (n= 316)	Tüm iyileşen (n= 7720)
Yaş (yıl)	48.65±16.45	59.63±16.41	57.16±16.10	72.76±10.90	50.51±16.74
Cinsiyet (E/K) (%)	2781/3254 (46/54)	1115/886 (56/44)	925/760 (55/45)	190/126 (60/126)	3706/4014 (48/52)

Çalışma grubunu oluşturan ayaktan ve yatarak takip edilen hastalar ile vefat edenlerin biyokimyasal, hematolojik, hemostatik ve enflamatuvar parametrelerine ait test sonuçları normal dağılıma uyduğundan bağımsız örnekler için Student t testi ile karşılaştırıldı. Yatan hastaların ayaktan takip edilen hastalar arasında RBC, MCV, MCHC, PCT, lenfosit sayıları ve potasyum düzeyi dışında tüm sonuçların anlamlı olarak farklı olduğu (hepsi için $p < 0.05$) saptanmıştır (Tablo 4.2).

Tablo 4.2. COVID-19 tanısı ile takip edilen hastaların rutin biyokimya ve hemostaz testlerinin sonuçları (Ort. ± SD)

	Ayaktan (n= 6035)	Yatan (n= 2001)	Yatan sağkalan (n= 1685)	Vefat eden (n= 316)	Toplam sağkalan (n= 7720)
Glukoz (mg/dL)	125.3±55.8	147.9±79.2 ^a	142.9±74.5	174.7±96.6 ^{bc}	129.16±60.9
Üre (mg/dL)	30.5±15.5	43.7±35.2 ^a	37.6±26.2	76.0±54.2 ^{bc}	32.1±18.6
Kreatinin (mg/dL)	0.89±0.5	1.11±0.8 ^a	0.99±0.6	1.71±1.5 ^{bc}	0.92±0.5
AST (U/L)	25.41±19.1	37.78±88.2 ^a	34.26±68.8	56.51±153.7 ^{bc}	27.35±36.5
ALT (U/L)	25.68±22.7	31.15±50.9 ^a	30.95±51.6	32.26±47.5 ^b	26.83±31.5
LDH (U/L)	221.7±78.8	315.1±179.3 ^a	288.8±124.7	454.9±312.5 ^{bc}	236.38±94.9
Na (mmol/L)	138.63±3.1	137.24±4.4 ^a	137.37±3.8	136.53±6.6 ^{bc}	138.35±3.3

K (mmol/L)	4.33±0.42	4.33±0.55	4.31±0.49	4.47±0.73 ^{bc}	4.32±0.44
Ca (mg/dL)	9.156±0.53	8.65±0.61 ^a	8.71±0.57	8.31±0.69 ^{bc}	9.06±0.57
CRP (mg/L)	20.37±38.3	67.47±78.8 ^a	55.92±67.1	128.86±103.8 ^{bc}	28.14±48.4
Ferritin (µg/L)	127.2±189.9	331.5±377.7 ^a	295.6±342.5	523.2±485.2 ^{bc}	164.1±242.3
cTnI (pg/mL)	13.5±116.5	78.5±703.1 ^a	37.0±344.0	299.8±1565.5 ^{bc}	18.6±191.2
D-Dimer (µg/L)	0.65±1.41	1.52±4.27 ^a	1.13±2.97	3.61±7.94 ^{bc}	0.75±1.88
RBC (x10 ⁶ /µL)	4.81±0.52	4.72±6.34	4.64±0.59	5.15±15.94 ^b	4.78±0.54
HGB (g/dL)	13.74±1.63	13.01±1.92 ^a	13.18±1.80	12.12±2.24 ^{bc}	13.62±1.69
HCT (%)	41.03±4.60	38.97±5.46 ^a	39.43±5.11	36.52±6.52 ^{bc}	40.68±4.76
MCV (fL)	85.405±6.11	85.364±6.73	85.17±6.50	86.41±7.75 ^{bc}	85.35±6.20
MCH (pg)	28.62±2.47	28.54±2.98	28.47±2.63	28.93±4.39 ^{bc}	28.58±2.51
MCHC (g/dL)	33.477±1.053	33.362±1.194 ^a	33.40±1.19	33.153±1.21 ^{bc}	33.46±1.08
RDW (fL)	40.92±3.98	41.96±5.38 ^a	41.29±4.51	45.60±7.74 ^{bc}	41.00±4.10
WBC (10 ³ /µL)	6.43±5.48	7.052±5.82 ^a	6.57±2.99	9.66±12.64 ^{bc}	6.46±5.05
NE (10 ³ /µL)	4.038±3.91	4.92±3.34 ^a	4.48±2.85	7.22±4.57 ^{bc}	4.14±3.71
NE (%)	61.43±12.35	67.45±15.40 ^a	65.51±14.80	77.76±14.41 ^{bc}	62.32±13.03
LYM (10 ³ /µL)	1.67±1.43	1.57±4.67	1.48±0.98	2.04±11.55 ^b	1.63±1.35
LYM (%)	27.14±11.01	23.31±13.08 ^a	24.88±12.67	14.90±11.96 ^{bc}	26.64±11.43
MONO (10 ³ /µL)	0.60±0.60	0.51±0.39 ^a	0.50±0.25	0.55±0.8 ^c	0.58±0.8
MONO (%)	9.80 ±4.3	7.94±3.95 ^a	8.24 ±3.79	6.35±4.43 ^{bc}	9.46±4.24
EOS (10 ³ /µL)	0.07±0.10	0.055±0.21 ^a	0.06±0.21	0.05±0.25 ^b	0.07±0.13
EOS (%)	1.17±1.41	0.89±3.50 ^a	0.97±3.80	0.44±0.71 ^{bc}	1.13±2.17
BASO (10 ³ /µL)	0.025±0.05	0.021±0.05 ^a	0.02±0.05	0.02±0.06	0.02±0.05
BASO (%)	0.46±0.59	0.38±0.55 ^a	0.4±0.56	0.26±0.43 ^{bc}	0.45±0.59
PDW (fL)	16,78 ±0,61	16,96±0,76 ^a	16.91±0.63	17.20±1.22 ^{bc}	16.81±0.62
MPV (fL)	8.038±0.97	8.20±1.07 ^a	8.18±1.04	8.31±1.21 ^{bc}	8.07±0.98
PLT (10 ³ /µL)	230±71.7	223.1±90.1 ^a	224.9±90.3	213.6±88.9 ^{bc}	228.9±76.1
PCT (%)	0.18±0.05	0.17±0.07	0.18±0.07	0.17±0.07 ^b	0.18±0.06
RDW-CV	13.98 ±1.66	14.39±2.09 ^a	14.19 ±1.88	15.48±2.71 ^{bc}	14.03±1.71

^ap< 0.05: Ayaktan takip edilen hastalar yatan hastalarla karşılaştırıldığında

^bp<0.05 : Toplam sağ kalan hastalar vefat edenlerle karşılaştırıldığında

^cp< 0.05: Hastane içi sağ kalan hastalar vefat edenlerle karşılaştırıldığında

4.2. Lojistik Regresyon Analizi Bulguları

Student t testi ile anlamlı fark olduğu saptanan parametrelerin dahil edildiği regresyon analizinde COVID-19 enfeksiyonunun prognozunu belirlemede ayırt edici güce sahip parametreler üç model ile araştırılmıştır. (i) COVID-19 tanısı doğrulanmış tüm vakalar prognozlarına göre ayakta ve yatarak tedavi edilen hastalar; (ii) COVID-19 tanısı almış tüm hastalarda sağkalan ve vefat edenler; (iii) yatarak takip edilen hastalar için sağkalan ve vefat edenler şeklinde (hastane içi mortalite) gruplanarak bağımlı değişkenleri oluşturmuştur. Yaş, cinsiyet ve student-t testinde anlamlı bulunan biyokimyasal parametreler ise bağımsız değişkenleri oluşturmuştur.

4.2.1. COVID-19 tanısı doğrulanmış tüm vakaların ayakta/yatarak takip prognoz öngörüsünde etkili parametre(ler)in belirlenmesi

COVID-19 tanısı konmuş hastaların ilk başvuru sırasındaki başlangıç laboratuvar sonuçları kullanılarak uygulanan multivaryant regresyon analizinde, analize dahil edilen laboratuvar parametrelerinin hemen tamamı anlamlı sonuçlar verdiği saptanmıştır (Tablo 4.3). Odds oranları dikkate alındığında yaş ile beraber hastanede yatarak takip edilme öngörüsü için en etkili faktörlerin sırasıyla MPV, ALT, üre, LDH, CRP ve aynı odds oranına sahip ferritin ve PLT olduğu görülmüştür

Tablo 4.3. Ayaktan/yatarak takip prognoz öngörüsünde etkili parametre(ler)nin belirlenmesi

	OR	%95 CI	p
Yaş (yıl)	1.017	1.013-1.022	.000
Üre (mg/dL)	1.005	1.001-1.010	.013
Kreatinin (mg/dL)	0.799	0.694-0.919	.002
AST (U/L)	0.993	0.991-0.996	.000
ALT (U/L)	1.007	1.004-1.010	.000
LDH (U/L)	1.004	1.003-1.005	.000
Na (mmol/L)	0.974	0.957-0.991	.004
Ca (mg/dL)	0.398	0.350-0.453	.000
CRP (mg/L)	1.004	1.002-1.006	.000
Ferritin (µg/L)	1.001	1.000-1.001	.000
HCT (%)	0.938	0.923-0.953	.000

RDW (fL)	0.979	0.964-0.994	.005
WBC ($10^3/\mu\text{L}$)	0.982	0.965-1.000	.045
NE ($10^3/\mu\text{L}$)	0.879	0.792-0.976	.016
LYM ($10^3/\mu\text{L}$)	0.890	0.801-0.989	.030
MONO ($10^3/\mu\text{L}$)	0.837	0.752-0.931	.001
EOS ($10^3/\mu\text{L}$)	0.186	0.035-0.999	.050
BASO ($10^3/\mu\text{L}$)	0.834	0.726-0.960	.011
MPV (fL)	1.169	1.095-1.248	.000
PLT ($10^3/\mu\text{L}$)	1.001	1.000-1.002	.020

CI: Güven aralığı; OR: Odds oranı

4.2.2. COVID-19 tanısı almış tüm hastaların mortalite öngörüsünde etkili parametre(ler)nin belirlenmesi

COVID-19 tanısı almış hastaların ilk başvuru sırasındaki biyokimyasal test sonuçları aracılığıyla mortalitenin öngörülmesinde de multivaryant regresyon analizine dahil edilen laboratuvar parametrelerinin eosinofil sayısı hariç hepsinin anlamlı sonuçlar verdiği saptanmıştır (Tablo 4.4). Odds oranlarına göre COVID-19'da başlangıç mortalite riskinin öngörülmesinde en belirleyici faktörlerin yaş ile beraber lökosit, üre, LDH, CRP ve glukoz olduğu görülmüştür.

Tablo 4.4. COVID-19 tanısı almış tüm hastaların mortalite öngörüsünde etkili parametre(ler)nin belirlenmesi

	OR	%95 CI	p
Yaş (yıl)	1.058	1.046-1.071	.000
Glukoz (mg/dL)	1.002	1.000-1.003	.017
Üre (mg/dL)	1.009	1.005-1.013	.000
AST (U/L)	0.996	0.993-0.998	.001
LDH (U/L)	1.004	1.003-1.005	.000
Na (mmol/L)	0.972	0.945-1.000	.049
Ca (mg/dL)	0.596	0.469-0.759	.000
CRP (mg/L)	1.002	1.000-1.004	.016
HCT (%)	0.940	0.914-0.967	.002
MCV (fL)	0.682	0.538-0.866	.002
MCH (pg)	3.280	1.584-6.791	.001

MCHC (g/dL)	0.339	0.183-0.628	.001
WBC ($10^3/\mu\text{L}$)	3.531	1.673-7.454	.001
NE (%)	0.267	0.122-0.585	.001
LYM ($10^3/\mu\text{L}$)	0.966	0.946-0.986	.001
LYM (%)	0.283	0.130-0.618	.001
MONO ($10^3/\mu\text{L}$)	0.889	0.830-0.953	.001
EOS ($10^3/\mu\text{L}$)	0.875	0.717-1.069	.190
BASO ($10^3/\mu\text{L}$)	0.768	0.572-1.032	0.08
MPV (fL)	0.837	0.718-0.975	.023
PLT ($10^3/\mu\text{L}$)	0.996	0.994-0.998	.000

CI: Güven aralığı; OR: Odds oranı

4.2.3. Hastane içi mortalite öngörüsünde etkili parametre(ler)nin belirlenmesi

Hastanede yatarak takip edilen hastalarda hastane içi mortalitenin öngörülmesinde etkili olan faktörler incelendiğinde multivaryant regresyon analizine dahil edilen laboratuvar parametrelerinden lenfosit sayısı ve CRP düzeyi hariç diğerlerinin anlamlı sonuç verdiği saptanmıştır (Tablo 4.5). Elde edilen odds oranlarına göre hastane içi mortalitenin öngörülmesinde en etkili faktörlerin sırasıyla MCH, lökosit sayısı, RDW, kreatinin, yaş, nötrofil sayısı, LDH ve glukoz olduğu görülmüştür.

Tablo 4.5. Hastane içi mortalite öngörüsünde etkili parametre(ler)nin belirlenmesi

	OR	%95 CI	p
Yaş (yıl)	1.059	1.045-1.073	.000
Glukoz (mg/dL)	1.002	1.001-1.004	.010
Kreatinin (mg/dL)	1.345	1.149-1.575	.000
AST (U/L)	0.996	0.994-0.998	.001
LDH (U/L)	1.004	1.003-1.006	.000
Ca (mg/dL)	0.752	0.576-0.983	.037
CRP (mg/L)	1.002	1.000-1.004	.057
HCT (%)	0.968	0.938-0.999	.043
MCV (fL)	0.632	0.429-0.930	.020
MCH (pg)	4.140	1.259-13.62	.019
MCHC (g/dL)	0.286	0.102-0.802	.017
RDW (fL)	1.033	1.002-1.065	.035

WBC ($10^3/\mu\text{L}$)	1.851	1.089-3.145	.023
NE (%)	0.534	0.305-0.935	.028
NE ($10^3/\mu\text{L}$)	1.040	1.018-1.063	.000
LYM ($10^3/\mu\text{L}$)	0.596	0.349-1.020	.059
MPV (fL)	0.809	0.687-0.952	.011
PLT ($10^3/\mu\text{L}$)	0.995	0.993-0.997	.000

CI: Güven aralığı; OR: Odds oranı

4.3. Karar Ağacı Analizi Bulguları

CART yöntemi uygulanarak yapılan karar ağacı analizinde COVID-19 tanısı almış hastaların prognoz öngörüsünde etkili parametre(ler)in belirlenmesinde lojistik regresyon analizine benzer bir gruplama yapılmış ve 3 farklı model oluşturulmuştur.

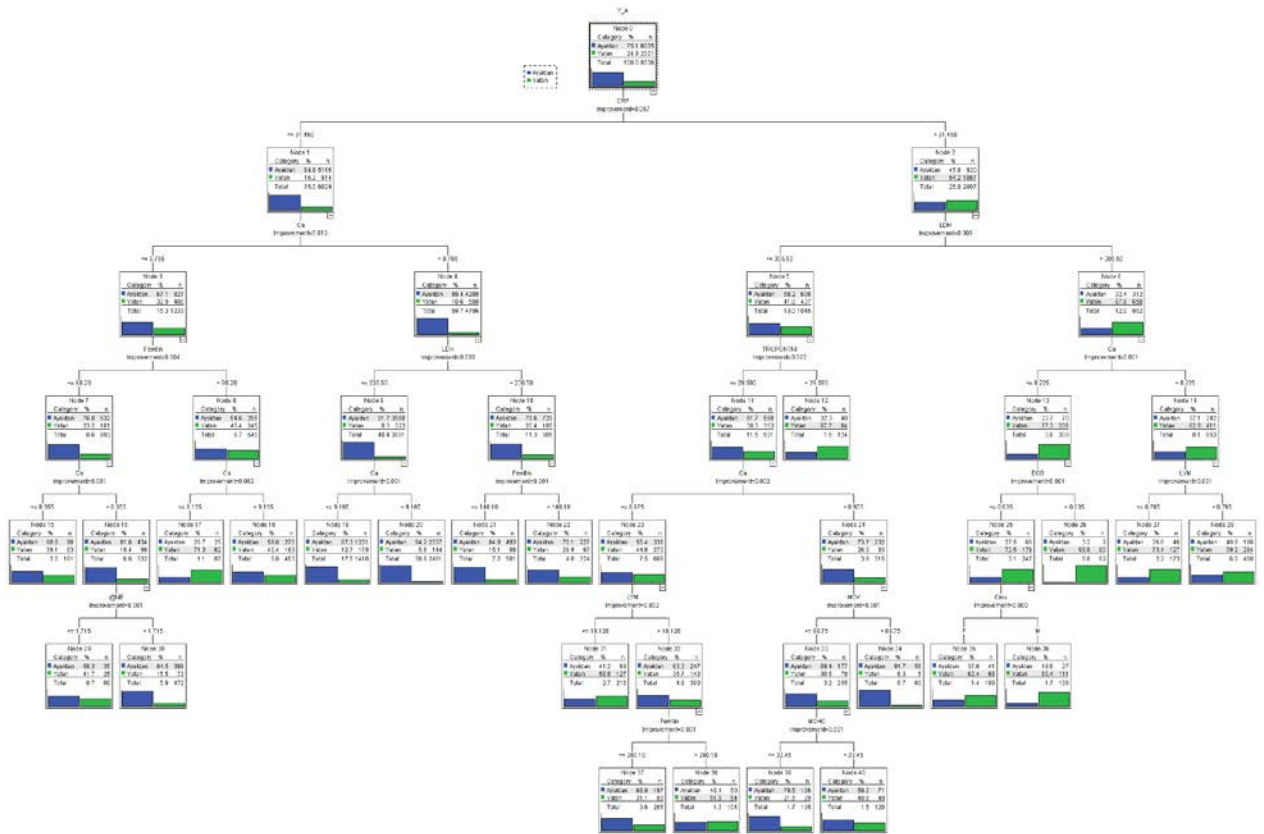
4.3.1. Model-1: Ayaktan/yatan prognoz öngörüsünde etkili parametre(ler)nin belirlenmesi

Model-1'den elde edilen sonuçlar incelendiğinde ayaktan/yatarak öngörüsünü etkileyen başlıca değişkenlerin CRP, Ca^{+2} ve LDH düzeyleri olduğu görüldü.

Karar ağacında yatarak takip edilme oranı ortalamasının yani kök düğümünde 0.057 olduğu saptandı. Hastaların sınıflandırması için en iyi ilk belirleyici CRP düzeyidir. CRP düzeyi 31.40 mg/L'ye eşit ya da küçük olan hastalar (hastaların %75'i) 1.alt düğümde (n=6029), CRP düzeyi 31.40 mg/L'den büyük olan hastalar ise 2.alt düğümde (%25) yer aldı. Birinci düğümdeki hastaların %15.5'i, 2. düğümdeki hastaların ise %54.2'sinin yatan hasta olduğu saptandı. Bu aşamada CRP düzeyi yükseldikçe hastanede yatış oranının da arttığı görüldü (Şekil 4.1). Karar ağacının sonraki seviyesinde CRP düzeylerini Ca^{+2} ve LDH değişkenlerinin etkilediği saptandı; 1. düğüm yatarak takip edilme oranı ortalaması 0.012 olarak kalsiyum düzeylerine göre 3. ve 4. alt düğümlere ayrıldı. Ca^{+2} düzeyi ≤ 8.765 mg/dL olan hastalar 3. alt düğümde, üzerinde olanlar ise 4. Alt düğümde yer aldı. Üçüncü alt düğümde yatan hasta oranı %32.9 iken 4. alt düğümde bu oranın %10.6'ya indiği gözlemlendi. Bu gözlem Ca^{+2} düzeyi azaldıkça hastanede yatış oranının arttığını ortaya koymuştur. CRP düzeyi ile ilişkili 2. alt düğümde LDH düzeylerine göre yatarak takip edilme oranı ortalaması 0.009 olarak 5. ve 6. alt düğümlere ayrılmıştır. Beşinci alt düğümde LDH düzeyi ≤ 305.6 U/L olan hastalar, 6. alt düğümde ise LDH düzeyi 305.6 U/L'nin üzerinde olanların yer aldığı

görüldü. Beşinci düğümdeki yatan hastaların oranı % 41.8 iken bu oranın 6. düğümde %67.6'ya yükselmesi LDH düzeylerindeki artışın da hastanede yatarak takip ile ilişkili olduğunu ortaya koymuştur. , Bu aşamadan sonra karar ağacının 40. alt düğümüne kadar devam etmiştir (Şekil 4.1).

Model-1 karar ağacının sonraki aşamalarında etkili olan parametreler için yatarak takip oranı ortalamaları sırasıyla ferritin için 0.004, cTnI için 0.002, lenfosit sayısı için 0.001 olduğu görüldü (Şekil 4.1).



Şekil 4.1. COVID-19 tanısı almış tüm hastalarda ayaktan/yatarak takip prognoz öngörüsünde etkili parametre(ler)nin belirlenmesinde CART karar ağacı modeli (Model-1)

4.3.1.1. Model-1'in performans bulguları

Ayaktan hastaların pozitif (istenen durum), yatan hastaların negatif (istenmeyen durum) olarak sınıflandığı CART algoritması ile elde edilen 1. modelin karışıklık matrisi Tablo 4.6'da verilmiştir.

Tablo 4.6. COVID-19 tanısı almış hastaların ayaktan/yatarak takip prognoz öngörüsü için Model-1'in karışıklık matrisi

		Modelin öngörüsü	
		Pozitif	Negatif
Gerçek durum	Ayaktan	5519 GP	516 YN
	Yatan	1024 YP	977 GN

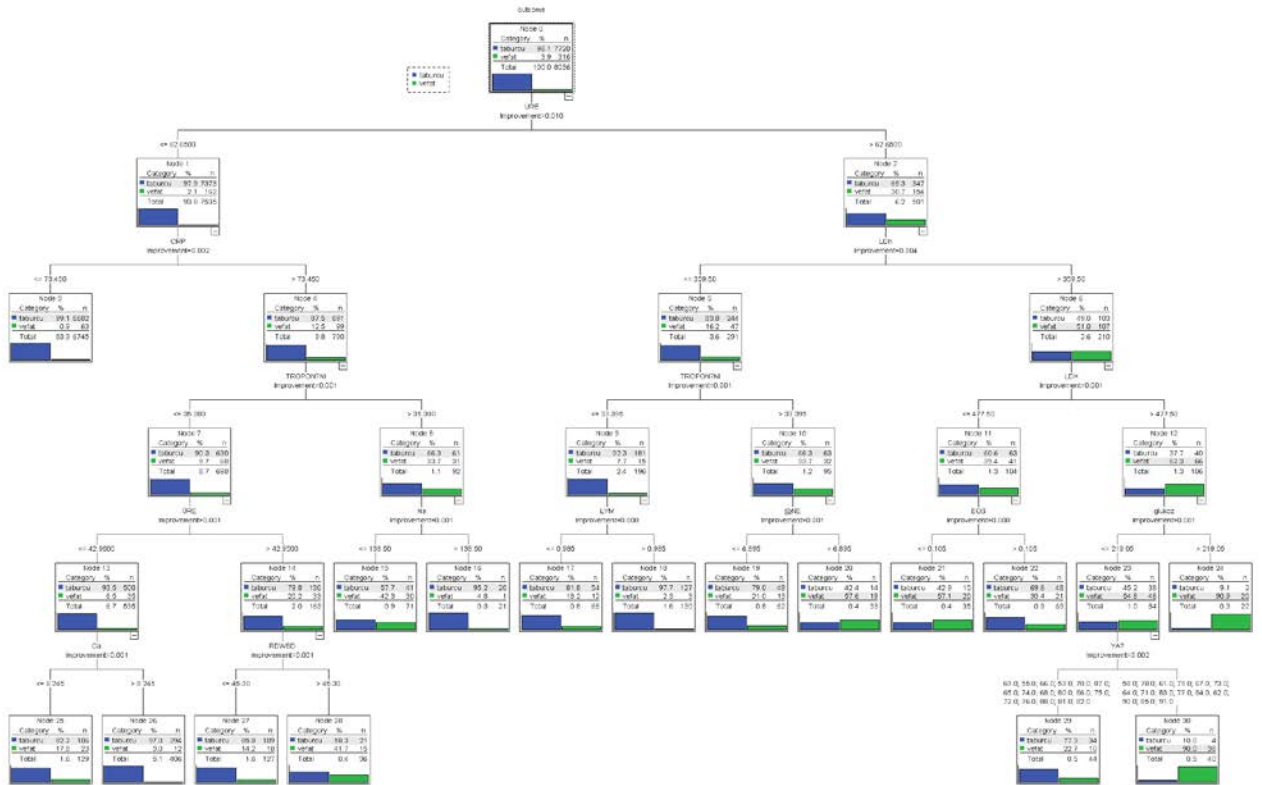
Modelin yanlış ve doğru öngörülerine göre performans kriterleri; geri çağırma %91.4, özgüllük %48.8, kesinlik %84.3, doğruluk %80.8, dengelenmiş doğruluk %70.1 ve F₁ Skoru %87.7 olarak hesaplanmıştır.

4.3.2. Model-2: COVID-19 tanısı almış tüm hastaların mortalite öngörüsünde etkili parametre(ler)nin belirlenmesi

Model-2'den elde edilen sonuçlar incelendiğinde COVID-19 tanısı alan tüm hastalarda hastalığın başlangıcında mortalite riskinin belirlenmesinde etkili değişkenlerin üre, LDH, CRP ve cTnI düzeyleri olduğu görüldü.

Hastaların sınıflandırması için en etkili ilk belirleyici üre düzeyidir. Üre düzeyi 62.65'e eşit ya da küçük olan hastalar (%93.8) 1. alt düğümde; üre düzeyi 62.65'den büyük olan hastalar ise 2. alt düğümde yer aldı. Birinci alt düğümde taburcu olan hastaların oranı %97.9 iken vefat eden hastaların oranının %2.1 olduğu saptandı. İkinci alt düğümde ise vefat eden hasta oranının %30.7'ye yükseldiği görüldü. Bu durum üre düzeyi arttıkça mortalite risk oranının da arttığı göstermektedir. Mortalite öngörüsü için önemle izlenmesi gereken sonraki parametrelerin sırası ile CRP ve LDH düzeyleri olduğu görülmüştür. Birinci düğüm mortalite risk oranı ortalaması 0.002 olan CRP düzeyi ≤ 73.45 mg/L olan hastalar 3. alt düğümde, CRP düzeyi > 73.45 mg/L olan hastalar 4. alt düğümde olacak şekilde ayrıldı. Üçüncü alt düğümde vefat eden hasta oranı %0.9 iken 4. alt düğümde bu oranın %12.5'e yükseldiği gözlemlendi. Bu gözlem CRP

düzeyi arttıkça mortalite oranının arttığını ortaya koymuştur. Üre düzeyi ile ilişkili 2. alt düğüm de LDH düzeylerine göre mortalite risk oranı ortalaması 0.004 olarak 5. ve 6. alt düğümlere ayrılmıştır. Beşinci alt düğümde LDH düzeyi ≤ 359.50 U/L olan hastalar, 6. alt düğümde ise LDH düzeyi 359.50 U/L'nin üzerinde olanların yer aldığı görüldü. Beşinci düğümdeki vefat eden hastaların oranı % 16.2 iken bu oranın 6. düğümde %51'e yükselmesi LDH düzeylerindeki artışın mortalite riski ile ilişkili olduğunu ortaya koymuştur. Bu aşamadan sonra karar ağacının 30. alt düğüme kadar devam ettiği görülmüştür (Şekil 4.2).



Şekil 4.2. COVID-19 tanısı almış hastaların mortalite öngörüsünde etkili parametre(ler)nin belirlenmesinde CART karar ağacı modeli (Model-2)

4.3.2.1. Model-2'nin performans bulguları

COVID-19 tanısı almış ve iyileşmiş tüm hastaların pozitif (istenen durum), vefat eden hastaların ise negatif (istenmeyen durum) olarak sınıflandığı CART modelinden elde edilen 2.Modelin karışıklık karışıklık tablosu Tablo 4.7'de verilmiştir.

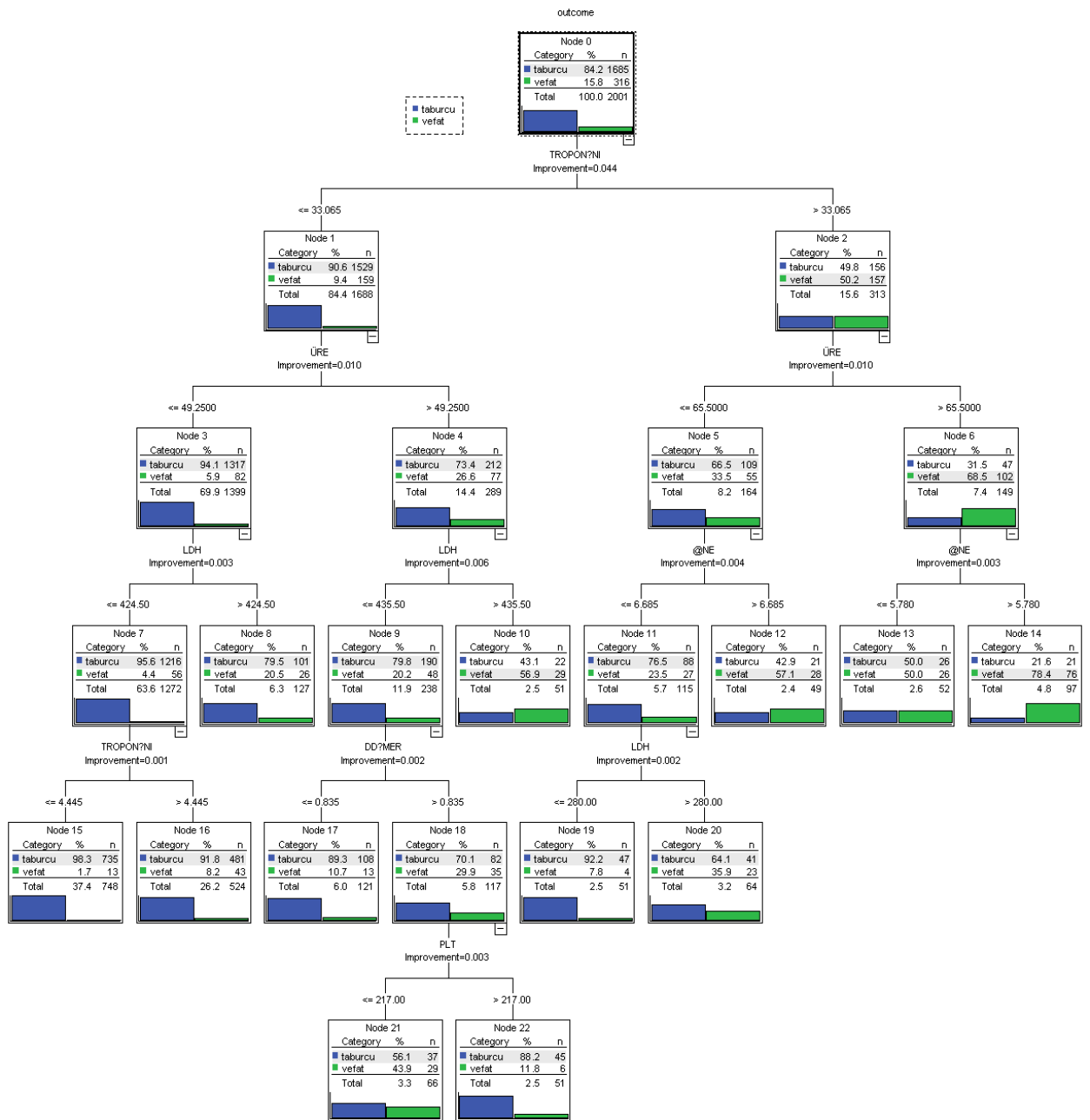
Tablo 4.7. COVID-19 tanısı almış tüm hastaların mortalite öngörüsü için Model-2'nin karışıklık matrisi

		Modelin öngörüsü	
		Pozitif	Negatif
Gerçek durum	İyileşen	7685 GP	35 YN
	Vefat eden	221 YP	95 GN

Modelin yanlış ve doğru öngörülerine göre performans kriterleri; geri çağırma %99.5, özgüllük %30.1, kesinlik %97.2, doğruluk %96.8, dengelenmiş doğruluk %64.8 ve F₁ Skoru %98.6 olarak hesaplanmıştır.

4.3.3. Model-3: Hastane içi mortalite öngörüsünde etkili parametre(ler)nin belirlenmesi

Model-3'den elde edilen sonuçlara göre COVID-19'da hastane içi mortalitenin öngörülmesinde etkili ilk değişken cTnI'dır. Karar ağacının cTnI düzeyinin 33.065 pg/mL'ye eşit ya da az olması durumunda 1.alt düğüm (%84.4), 33.065 pg/mL'den büyük olması durumunda ise 2.alt düğüm (%15.6) olarak ikiye ayrıldığı görülmüştür. Birinci alt düğümde taburcu olan hastaların oranı %90.6 ve vefat eden hastaların oranı %9.4 iken 2. alt düğümde taburcu olan hasta oranının %49.8'e düştüğü ve vefat eden hasta oranının %50.2'ye yükseldiği saptanmıştır. Bu durum cTnI düzeyinin artışı ile mortalite oranının artışı arasındaki ilişkiyi ortaya koymuştur. Hastane içi mortalite öngörülmesinde önemle izlenmesi gereken sonraki parametrelerin sırasıyla üre, LDH düzeyleri ve nötrofil sayısı olduğu görülmüştür. Model-3 karar ağacının sonraki aşamalarda 22. alt düğüme kadar bölündüğü görülmüştür (Şekil 4.3).



Şekil 4.3. Hastane içi mortalite öngörüsünde etkili parametre(ler)nin belirlenmesinde CART karar ağacı modeli (Model-3)

Model-3 karar ağacının sonraki aşamalarında etkili olan parametreler için hastane içi mortalite oranı ortalamaları sırasıyla cTnI için 0.044, üre için 0.010, LDH için 0.006, nötrofil sayısı için 0.004 olduğu görüldü (Şekil 4.3).

4.3.3.1. Model-3'ün performans bulguları

Taburcu olan hastaların pozitif (istenen durum), vefat eden hastaların negatif (istenmeyen durum) olarak sınıflandığı CART modelinden elde edilen 3.Modelin karışıklık tablosu Tablo 4.8'de verilmiştir.

Tablo 4.8. Yatarak takip edilen COVID-19 hastalarının hastane içi mortalite öngörüsü için Model-3'ün karışıklık matrisi

		Modelin öngörüsü	
		Pozitif	Negatif
Gerçek durum	Taburcu	1621 GP	64 YN
	Vefat eden	183 YP	133 GN

Modelin yanlış ve doğru öngörülerine göre performans kriterleri; geri çağırma %96.2, özgüllük %42.1, kesinlik %89.9, doğruluk %87.7, dengelenmiş doğruluk %69.2 ve F_1 skoru %92.9 olarak hesaplanmıştır.

5. TARTIŞMA

Geleneksel istatistik yöntemler ve bu yöntemler kullanılarak geliştirilen makine öğrenmesi uygulamalarının her ikisi de prensip olarak hem öngörü hem de çıkarım yapmak için kullanılabilir de bazı önemli farklılıkları bulunmaktadır [56]. İstatistikî yöntemler daha çok mevcut bir duruma özgü olasılık modelinin oluşturulması ve uyumunun sağlanması ile elde edilen çıkarıma odaklanmaktadır. İstatistikî model ile saptanan ilişkinin gerçek bir etkiyi ortaya koyduğu nicel bir güven ölçüsü ile belirtilebilir ve yeterli veri mevcutsa bir hipotez açıkça doğrulanabilir. Ancak, geleneksel istatistikî modelleme birkaç düzine değişken ve küçük-orta örneklem genişliğine sahip veriler için tasarlanmıştır ve bu tür parametreler çoğu kez güvenilir kararlar almak için yeterli bilgi sağlamaz. Değişken sayısı, çeşitliliği ve aralarındaki olası ilişkiler arttıkça bu ilişkileri yakalamak için model daha karmaşık ve sağladığı çıkarımlar daha az kesin hale gelir (57). Buna karşılık makine öğrenmesi genel amaçlı öğrenme algoritmalarını kullanarak büyük ve karmaşık veri setlerinde heterojen değişken türlerini tek bir modele dahil ederek çıkarım yapmaya odaklanır. Geniş örneklemlili büyük veri setleri ile çalışıldığında bile girdiler ve çıktı arasında karmaşık doğrusal olmayan ilişkiler bulunsada etkili olabilirler. Ayrıca, verileri etiketlemeden işleyen denetimsiz öğrenme yöntemleri kullanıldığında veri setindeki gizli örüntü ve kalıpları saptama avantajına da sahiptir [19,58]. Kısaca makine öğrenmesi modellerinin performansının risk sınıflaması ve prognoz öngörüsü bağlamında geleneksel istatistikî yöntemlerinden daha başarılı olduğu bildirilmektedir [54]. Bu görüş makine öğrenmesi algoritmalarının doğrusal olmayan ilişkileri ortaya çıkarma yetenekleri, çok faktörlü algoritmaların optimizasyonu imkanı ve modeli doğrulamak için yerleşik metodolojilerinin olması gibi nedenlerle desteklenmektedir [54,57].

Son yıllarda sağlık alanında da karar verme sürecini kolaylaştırmak ve/veya faydalı bilgileri ya da gizli örüntüleri ortaya çıkarmak için çeşitli yapay zeka ve makine öğrenmesi yöntemleri kullanılmaya başlamıştır [53,59]. Bu uygulamalar laboratuvar tıbbında hasta verilerine dayalı gerçek zamanlı kalite kontrolü, dış kalite değerlendirmesi, analitik doğruluğun artırılması, laboratuvar yönetimi ve verimliliği, çeşitli analitler için varyasyon kaynaklarının analizi, referans aralıklarının

oluşturulması, onay destek sistemleri gibi farklı perspektiflerde uygulama alanı bulmuştur [60-64]. Ancak, klinik tanının modele bağımlı değişken olarak dahil edilmesi gerektiğinden laboratuvar test sonuçlarını kullanarak tanı modelleri ve/veya prognostik modeller oluşturulması konusu nispeten özel bir alan olarak kalmıştır. Klinik laboratuvarların çoğunun işlevsel ve teknik sınırlamaları, sağlık sisteminin gereklilikleri, hasta mahremiyeti gibi nedenlerle klinik tanı ve hasta bilgilerini elde etmede yaşanan zorluklar bu tür araştırmaların sayısını oldukça kısıtlamıştır [64-65]. Ancak morbidite ve mortalite riskinin güvenilir ve kanıta dayanır yöntemlerle hızlıca belirlenerek mevcut sağlık altyapısının en rasyonel şekilde kullanılmasını gerektiren COVID-19 pandemisi, laboratuvar test sonuçlarının risk sınıflaması, tanı veya prognoz modellemesinde kullanılmasına yönelik çalışmalara hız kazandırmıştır.

Bu kapsamda klinik bulguları ve/veya laboratuvar bulguları kullanılarak hastanede yatmakta olan hastalarda farklı makine öğrenmesi yöntemleri kullanılarak hastalık şiddetinin ve mortalite riskinin belirlenmesine yönelik birçok COVID-19 çalışması bulunmaktadır. Genellikle pandeminin en yoğun olduğu dönemde kısıtlı sayıda hasta ile yürütülmüş olan bu çalışmaların çoğunda elde edilen model davranışının duyarlılığı ve özgüllüğü de açıkça ortaya konulamamıştır. Hastalığın başlangıcından itibaren COVID-19 şiddetini ve mortalite riskini belirlemeye yönelik çalışmaların sayısı ise oldukça kısıtlıdır. Tüm bu nedenlerle nispeten geniş bir örnekleme sahip hasta grubunda sadece prognoz bilgileri ve rutin biyokimyasal test sonuçlarını kullanarak yaptığımız bu çalışmada, hem geleneksel istatistik yöntemler hem de makine öğrenmesi yöntemi olan CART algoritması ile COVID-19'un başlangıcından itibaren hastaların prognoz öngörüsü yanısıra başlangıç ve hastane içi mortalite risk profillerinin belirlenmesi amaçlanmıştır.

Çalışmamızda geleneksel istatistik yöntem ile elde ettiğimiz sonuçlar hem hastalığın başlangıç aşamasında COVID-19 şiddetinin ve mortalite riskinin belirlenmesinde hem de hastanede yatan hastaların mortalite öngörüsünde çalışmaya dahil edilen laboratuvar parametrelerinin hemen hemen hepsinin anlamlı bir farklılık sergilediğini ortaya koydu. Anlamlı olarak farklı olan bu parametrelerin dahil edildiği lojistik regresyon modelinde de benzer şekilde, parametrelerin çoğunun öngörücü öneme sahip olduğu görüldü. COVID-19'lu hastalarda hastane içi mortalite risk faktörlerini lojistik regresyon analizi dahil diğer geleneksel yöntemler ile araştıran çalışmalar incelendiğinde, bunların çoğunun dar örnekleme kötü prognoz ile ilişkili

parametrelerin belirlenmesine yönelik olduğu görülmüştür. Ancak, bu çalışmalardan kötü prognoz ile ilişkili parametreler konusunda bir fikir edinilse de klinik durumu göstermede en ayırt edici öneme sahip parametre(ler) ve risk belirleyici kesim değerleri (cutoff) konusunda bir bilgi sağlanamamıştır [66-69]. Nispeten geniş bir örnekleme sahip olan çalışmamızdan geleneksel lojistik regresyon analizi ile bu çalışmalara benzer sonuçlar elde edilmiştir. Bu durum geleneksel yöntemler ile elde edilen sonuçların kolayca yorumlanıp klinik pratiğe uygulanabilecek çıkarımlar yapma olanağı sunmadığı yönündeki görüşü destekler niteliktedir.

Çalışmamızda bir makine öğrenmesi algoritması olan CART ile elde ettiğimiz Model-1'in sonuçları incelendiğinde ise COVID-19 tanısını takiben ayaktan-yatarak takip öngörüsü için en etkili parametrelerin sırasıyla CRP, LDH ve Ca^{+2} düzeyleri olduğu ve bunlar için kesim değerlerinin de >31.45 mg/L, >305.50 IU/L ve ≤ 8.225 mg/dL olduğu görüldü. Bu parametreler kullanılarak karar ağacının önerisi ile kolayca izlenip klinik pratiğe uygulanabilecek kural için modelin doğruluğunun %80.8, duyarlılığının %91.4 ve özgüllüğünün %48.8 olduğu hesaplanmıştır. Benzer şekilde başlangıç ve hastane içi mortalitenin öngörülmesi için elde ettiğimiz Model-2 ve Model-3 ile etkili parametreler ve kesim değerleri saptanmıştır. Buna göre başlangıç mortalite öngörüsünde en etkili parametreler üre >62.65 mg/dL, LDH >359.5 IU/L, CRP >73.45 mg/dl ve cTnI >33.39 pg/mL düzeyleridir. Bu parametreler kullanılarak Model-2 ile elde edilen kuralın doğruluğu %96.8, duyarlılığı %99.5 ve özgüllüğü %30.1 olarak hesaplanmıştır. Hastane içi mortalite öngörüsü için ise etkili parametrelerin sırasıyla cTnI >33.06 pg/mL, üre >65.5 mg/dL, LDH >435.50 IU/L ve nötrofil sayısı $>5.78 \cdot 10^3/\mu L$ olduğu ve Model-3 ile elde edilen kuralın doğruluğu %87.7, duyarlılığı %96.2 ve özgüllüğü %42.1 olarak hesaplanmıştır. Elde ettiğimiz bu bulgular makine öğrenmesi yöntemleri ile kolayca izlenip yorumlanabilecek ve bir kural haline dönüştürülerek rutin klinik uygulamalara entegre edilebilecek sonuçlar ortaya koyduğunu desteklemektedir. Lojistik regrasyon ve makine öğrenmesi yöntemleri ile elde ettiğimiz sonuçlar geleneksel istatistiki modellemenin parametre çeşitliliği ve sayısının nispeten az, küçük-orta örneklem büyüklüğündeki verilerle çalışmaya uygun olduğu, değişken çeşitliğinin fazla olan büyük veri setlerindeki doğrusal olmayan ilişkilerini modellemede daha az kesin sonuçlar verdiği görüşünü doğrulamaktadır.

Tıbbi literatürde COVID-19'un henüz başlangıcında prognoz ve/veya mortalitenin öngörülmesine yönelik makine öğrenmesi çalışmaları oldukça kısıtlı sayıdadır.

Çalışmamızdan elde edilen sonuçları COVID-19 konusunda yapılan benzer çalışmalarla karşılaştırdığımızda ise çoğu kez benzer parametrelerin öne çıktığı görülmektedir. Çalışmamıza benzer şekilde yapay zeka algoritmaları kullanılarak başlangıçta hastalık prognozu ve mortalite riskinin öngörülmesi konusundaki çalışma Liang W ve arkadaşları tarafından yürütülmüştür. Riskli vakaları saptamak amacı ile Çin Cumhuriyeti'nde yürütülen bu retrospektif çalışmada, COVID-19 tanısı PCR testi ile saptanmış 9252 vaka için kritik hastalık, yoğun bakım ünitesine yatış veya invaziv ventilasyon gerektiren bileşik bir olay veya ölüm olarak tanımlanmış. Bu vakalarda radyolojik bulgular, yaş, nefes darlığı, KOAH ve kanser öyküsü gibi eşlik eden hastalıklar ile nötrofil/lenfosit oranı, laktat dehidrojenaz, direkt bilirubin ve kreatin kinaz düzeyleri derin öğrenme algoritmaları ile oluşturulan Survival Cox modeli ve klasik Cox modeli kullanılarak riskli vakaları ayırt etme güçleri karşılaştırılmıştır. Yapay zeka ile geliştirilen Survival Cox modelinin kullanılan parametreler arasındaki karmaşık nonlineer ilişkileri çözmekte klasik Cox modeline göre üstün bir ayırt etme gücü gösterdiği saptanmıştır. Yapay zeka yöntemlerinin tıbbi personele öngörülen risk göstergeleri sayesinde hastaneye kabul için öncelik sıralaması yapabilme olanağı sağladığı sonucuna varılmıştır [70]. Hwangbo S. ve arkadaşları ise ilk başvuru kayıtlarındaki demografik, fiziksel ölçüm, klinik bulgu ve semptom, eşlik eden hastalıklar, radyolojik ve laboratuvar bilgilerini kullanarak 2.263 COVID-19 vakasında makine öğrenmesi ve derin öğrenme algoritmaları ile hastanede yatan hastalarda hastalık ciddiyetininin erken öngörüsü için tahmin modelleri geliştirmiştir. Hastaneye yatış sırasında hastalık şiddeti iki yaklaşım ile sınıflandırılmış; ilk yaklaşımda hafif, orta, şiddetli ve kritik olmak üzere dört gruba ayrılan hastalar GUIDE ve orantısız olasılık modeli, ikinci yaklaşımda ise şiddetli ya da değil şeklinde iki gruba ayrılan hastalar RandomForestSRC dahil olmak üzere üç değişken seçim yöntemi ile değerlendirmiştir. Geliştirilen üç modelde de yaş, akciğer infiltrasyonu, CRP ve AST'nin hastalık şiddetininin ortak belirleyicisi olduğu bildirilmiştir [71].

COVID-19'lu hastalarda hastane içi mortalite riskinin belirlendiği 452 vakayı içeren klinik bulgular ve bazı rutin biyokimya parametrelerinin ticari R 3.5.2 paket yazılımı kullanarak oluşturdukları karar ağacı ile değerlendirildiği bir çalışmada modelin doğruluğu %98 olarak hesaplanmış ve mortalite ile yakın ilişkili parametrelerin LDH, NLR ve CRP'nin olduğu saptanmış. LDH ≥ 330 IU/L, nötrofil-lenfosit oranı (NLR) ≥ 6.9 ve CRP ≥ 27 mg/L düzeylerine sahip COVID-19'lu hastaların yakından izlenmesi

gerektiği sonucuna varılmıştır [72]. Mazlumi ve ark.'larının 306 COVID-19'lu hastada yürüttüğü yakın tarihli bir çalışmada ise önceki çalışmalarda COVID-19'da belirgin değişiklik gösterdiği saptanan altı laboratuvar parametresi seçilerek mortalite öngörüsü için karar ağacı dahil sekiz farklı makine öğrenmesi algoritması ile mortalite riski tahminlemesi yapılmış. Karar ağacı modelinin en yüksek doğruluk yanı sıra veri setine uygun sensitivite ve spesifite sağladığı sonucuna varılarak mortalitenin öngörülmesinde en önemli faktörlerin sırasıyla yaş, LDH düzeyi ve lökosit sayısı olduğu bildirilmiştir [53]. Li ve arkadaşlarının 2924 hastalık daha geniş bir örnekleme ikili sınıflandırma problemini çözmek için gradyan artırıcı karar ağacı (GBDT) ve lojistik regresyon algoritmalarından yararlanarak geliştirdikleri CART regresyon ağacı modeli ile hastane içi mortalitenin öngörülmesinde etkili faktörlerin lenfosit sayısındaki azalma, nötrofil sayısı, NLR oranı, LDH, hs-CRP, BUN, IL-6, TNF- α ve IL-10 düzeylerindeki artış olduğu saptamıştır [73]. Elde ettiğimiz bu sonuçlar aynı konularda makine öğrenmesi algoritmalarıyla yürütülmüş çalışmalar sonuçları ile uyumludur. Bu çalışmalarda da mortalite riskini belirlemede öne çıkan laboratuvar parametreleri BUN, LDH ve nötrofil sayısıdır. Ancak çalışmamızda diğer çalışmalardan farklı olarak hastane içi mortalitenin öngörülmesinde CRP etkili bir faktör olarak bulunmazken cTnI düzeyleri en öncelikli faktör olarak saptanmıştır. Bu durumun büyük olasılıkla diğer çalışmalarda cTnI'nın çalışmaya dahil edilmemesi ile ilişkili olduğu görüşündeyiz. Ayrıca bu farklar çalışmamızın örneklem büyüklüğünün benzer çalışmalardan daha geniş olmasından kaynaklanıyor olabilir.

Diğer taraftan çalışmamız ve benzer şekilde elektronik sağlık verilerini kullanarak sağlık alanında yapılan diğer çalışmalar derin öğrenme ya da makine öğrenmesi gibi yapay zeka uygulamalarının klinisyenlere zaman kazandırılması ve hasta bakımının iyileştirmesi, maliyetlerin düşürülmesi gibi konularda mevcut uygulamaları yeni bir seviyeye taşıdığını ve sağladığı yararların sınırsız olabileceğini ortaya koymaktadır. Ancak tıbbi süreçler şeffaf, incelenebilir, değerlendirilebilir, açıklanabilir ve ölçülebilir olmalıdır. Yapay zeka uygulamalarının benimsenmesi de muhtemelen benzer süreçleri gerektirecektir. Tüm olumlu katkılarına rağmen yapay zeka uygulamalarının tıbbi problemlerin çözümü için sunduğu öneriler henüz ilişkileri net olarak açıklayamamakta ya da önerdiği çözümlerin mevcut tıbbi bilgilerle doğrudan ilişkilendirilmesinde bazı zorluklar yaşanmaktadır. Makine öğrenmesi yöntemlerinin klinik pratikte kullanılması ve benimsenmesinin önündeki engeller bu uygulamaların dezavantajları ile ilişkili

olabilir. Başlıca dezavantajlar makine öğrenmesi yöntemlerinin kapsayıcı, tarafsız ve iyi kalitede çok büyük veri setleri gerektirmesi, amaca uygun algoritma(lar)nın seçimi ve algoritmanın yeterince öğrenmesinin zaman alması olarak sıralanabilir. Diğer önemli zorluk ise algoritmaların ürettiği modelin yorumlama yeteneğidir. Çünkü algoritmalar otonom çalışır dışarıdan müdahaleye izin vermez bu da elde edilen modelleri hatalara karşı oldukça hassas kılar. Yeterli kapsayıcılıkta olmayan küçük veri kümeleri ile çalışıldığında bu önyargılı veri seti tek yönlü hatalı tahminlere neden olur. Böyle bir durumda hata(lar)nın saptanması ve düzeltilmesi uzun zaman alacağından, geri dönüşü mümkün olmayan zararlara yol açabilir [74]. Laboratuvar verileri kullanılarak klinik pratik uygulamalara destek sağlayacak modellerin geliştirilebilmesi için önerilerin dayanakları ve klinik sonuçları nasıl iyileştireceği kanıta dayalı olarak açıklanabilmelidir. Bu konuda yeterli sayıda tarafsız ve kaliteli veri toplanması, uygun algoritmaların seçimi ve sonuçların verifikasyonu ve yorumlanması konularında laboratuvar uzmanlarına ve klinisyenlere önemli görevler düşmektedir. Bu tip çalışmalar veri bilimciler, biyoinformatik, bilişim teknolojileri ve biyoistatistik alanında çalışan bilim insanları, laboratuvar uzmanları ve klinisyenlerin oluşturduğu ekip çalışmalarını ve sıkı bir işbirliğini gerektirmektedir.

Sonuç olarak COVID-19'un başlangıcından itibaren laboratuvar test sonuçları kullanarak hastaların prognoz öngörüsü yanısıra başlangıç ve hastane içi mortalite risk profillerinin belirlenmesini amaçlayan çalışmamızdan elde ettiğimiz bulgular makine öğrenmesi yöntemlerinin değişken çeşitliği fazla olan büyük veri setlerindeki doğrusal olmayan ilişkileri modellemede geleneksel istatistikî yöntemlerden daha başarılı olduğu görüşünü doğrular niteliktedir. Bir makine öğrenmesi yöntemi olan CART algoritması ile elde ettiğimiz tüm karar ağacı modelleri ayaktan-hastanede yatarak takip prognozunun, başlangıç ve hastane içi mortalite riskinin belirlenmesinde yüksek riskli hastaları büyük bir doğruluk oranı ile belirleyebilmiş ve yüksek risk ile ilişkili parametreleri kesim değerleri ile birlikte ortaya koyabilmiştir. Elde ettiğimiz sonuçların COVID-19 hastaları yanı sıra tıbbi laboratuvar sonuçlarını kullanarak diğer hastalık gruplarında yapılacak yapay zeka çalışmalarına da önemli katkılar sağlayacağı görüşündeyiz.

KAYNAKLAR

- [1] Han J, Pei J, Kamber M. Data Mining: Concepts and Techniques. Waltham, MA, USA: Elsevier Inc; 2011.
- [2] Kantardzic M. Data Mining: Concepts. Models. Methods. and Algorithms. 2nd ed. Hoboken, NY: Wiley-IEEE Press; 2011.
- [3] Harvey C. AI vs. Machine Learning vs. Deep Learning: Subsets of Artificial Intelligence. Eriřim Tarihi: 02.01.2023, <https://www.datamation.com/artificial-intelligence/ai-vs-machine-learning-vs-deep-learning/?nowprocket=1>
- [4] Canbay P. Saęlıkta Yapay Zeka: Makine Öğrenmesi Yöntemleri ve Uygulamaları. Saęlık Bilimlerinde Yapay Zeka. Ankara: Akademisyen Kitabevi; 2020.pp.11-23
- [5] Özdemir ř. Karar Aęaęları: Oluřturma ve Yorumlama. Rapidminer ile Uygulamalı Veri Madencilięi kitabı. İstanbul: Pusula Yayıncılık; 2017.
- [6] Koyuncugil A, Özgülbaş N. Veri Madencilięi: Tıp ve Saęlık Hizmetlerinde Kullanımı ve Uygulamaları. *Biliřim Teknolojileri Dergisi*, 2009;2(2):21-32
- [7] Shmueli G, Patel NR, Bruce PC. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. Hoboken, NJ: Wiley-Interscience;2007.
- [8] Sutton RS and Barto AG. Reinforcement learning: An introduction. 2nd ed. Cambridge, MA,USA:The MIT Pres;2018.
- [9] Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int.* 2010;107:776–82.
- [10] İnal ME, Topuz D, Uęan O. Doğrusal Olasılık ve Logit Modelleri ile Parametre Tahmini, *Sosyoekonomi Dergisi* 2006:1; 48-72.
- [11] Hilbe JM. Practical Guide to Logistic Regression. Chapman and Hall/CRC: Boca Raton, FL, USA: CRC Press; 2016.
- [12] Çokluk Ö. Lojistik regresyon analizi: Kavram ve uygulama. Kuram ve Uygulamada Eęitim Bilimleri. 2010;10(3):1357-1407.

- [13] Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. Comparing diferent supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* 2019;19:1–16.
- [14] Taşcı E, Onan A. K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi. *Akademik Bilişim*, 2016.
- [15] Savaş S, Topaloğlu N, Yılmaz M. "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri". *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi* 2012;11(21):1-23.
- [16] Jijo BT & Abdulazeez AM. Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends* 2021;2(01):20-28.
- [17] Loh WY. Fifty Years of Classification and Regression Trees. *International Statistical Review* 2014;82(3):329–348.
- [18] Berson, A, Smith S, Thearling K. *Building Data Mining Applications for CRM*, A Division of the McGraw- Hill Companies, 1999.
- [19] Quinlan JR. *C4.5: Programs For Machine Learning*. California, Morgan Kaufmann Publishers Inc; 1993.
- [20] Priyama A, Abhijeeta GR, G Rahul, Ratheeb A, Srivastavab S. Comparative Analysis of Decision Tree Classification Algorithms. *Int J Curr Eng Technol* 2013;3(2). 334-337.
- [21] Rokach L, Maimon O. editor. *Data Mining and Knowledge Discovery*. Boston. Handbook Springer; 2010. pp. 164-192
- [22] Mehta M, Agrawal R, Rissanen J. SLIQ: A fast scalable classifier for data mining *Advances in Database Technology-EDBT'96. 5th International Conference on Extending Database Technology Conference Avignon, France;1996. pp 18–32.*
- [23] Shafer JC, Agrawal R, Mehta M. SPRINT: A Scalable Parallel Classifier for Data Mining *Proceedings of the 22th International Conference on Very Large Data Bases. Bombay, India;1996. pp. 544–555*
- [24] Loh WY, Shih YS. Split selection methods for classification trees. *Statistica Sinica* 1997; 7:815-840

- [25] Friedman JH. Multivariate Adaptive Regression Splines. *Ann Stat* 1991;19(1):1-67
- [26] Tokalaçođlu E. Karar Ađacı Temelli Geri Bildirim Sistemi Tasarımı: Türkiye Mutluluk Ađacı Üzerine Bir Uygulama. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi. Sosyal Bilimler Enstitüsü; 2020.
- [27] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. 1st ed. Taylor and Francis, Chapman&Hall/CRC;1984.
- [28] Demirel Ş, Yakut SG. Karar Ađacı Algoritmaları ve Çocuk İşçiliđi Üzerine Bir Uygulama. *Sosyal Bilimler Araştırma Dergisi* 2019;8(4): 52–65.
- [29] Reddy S, Vasu V. A New Type Of Node Split Rule For Decision Tree Learning. *International Journal of Advanced Research in Computer Science*, 2013; 4(8), 75-81.
- [30] Tangirala S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *Int. J. Adv. Comput. Sci. Appl.* 2020; 11:612–619
- [31] Moisen G. Classification and Regression Tree. Encyclopedia of Ecology, Oxford, UK: Elsevier; 2008 pp.582-588.
- [32] Mansour Y. Pessimistic decision tree pruning based on tree size". Proc. 14th International Conference on Machine Learning; 1997 pp. 195–201.
- [33] Mark L, Maimon O, Minkov E. Improving Stability of Decision Trees. *Int J Pattern Recognit Artif Intell* 2002;16(02):145-159
- [34] Çınar UK. Eksik Veri-Kayıp Veride Kullanılabilecek Algoritmalar. Erişim Tarihi: 02.01.2023 <https://www.veribilimiokulu.com/eksik-veri-eksik-veride-kullanilabilecek-algoritmalar/>
- [35] Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z. et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan. China: a retrospective cohort study. *Lancet* 2020;395(10229):1054-62.
- [36] Naming the coronavirus disease (COVID-19) and the virus that causes it Available from: URL: ([https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)). Erişim Tarihi: 02.01.2023

- [37] WHO Coronavirus (COVID-19) Dashboard. Available from: URL: <https://covid19.who.int>. Erişim Tarihi: 07.11.2022
- [38] Li J, Jia H, Tian M, Wu N, Yang X, Qi J, Ren W, Li F, Bian H. SARS-CoV-2 and Emerging Variants: Unmasking Structure, Function, Infection, and Immune Escape Mechanisms. *Front. Cell Infect. Microbiol.* 2022; 12: 869832.
- [39] Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Napoli RD. Features, Evaluation and Treatment Coronavirus (COVID-19). Treasure Island, StatPearls Publishing; 2020.
- [40] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;395:497–506
- [41] Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA - J Am Med Assoc* 2020;323(11):1061-1069
- [42] Ponti G, Maccaferri M, Ruini C, Tomasi A, Ozben T. Biomarkers associated with COVID-19 disease progression. *Crit Rev Clin Lab Sci* 2020;57:389–99.
- [43] Lowery SA, Sariol A, Perlman S. Innate immune and inflammatory responses to SARS-CoV-2: implications for COVID-19. *Cell Host Microbe* 2021;29: 1052–1062.
- [44] Khalil BA, Elemam NM, Maghazachi AA. Chemokines and chemokine receptors during COVID-19 infection. *Comput Struct Biotechnol J* 2021; 19:976-988.
- [45] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China 2019. *N Engl J Med* 2020; 382:727-733.
- [46] Xiang N, Havers F, Chen T, Song Y, Tu W, Li L, et al. Use of national pneumonia surveillance to describe influenza A(H7N9) virus epidemiology, China, 2004–2013. *Emerg Infect Dis* 2013;19:1784–1790.
- [47] Henry BM, de Oliveira MHS, Benoit S, Plebani M, Lippi G. Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis. *Clin Chem Lab Med* 2020;58:1021-1028.

- [48] Malik P, Patel U, Mehta D, Patel N, Kelkar R, Akrmah M, et al. Biomarkers and outcomes of COVID-19 hospitalisations: systematic review and meta-analysis. *BMJ Evid Based Med* 2021;26(3):107-108.
- [49] Chen Z, Xu W, Ma W, Shi X, Li S, Hao M, Fang Y, Zhang L. Clinical laboratory evaluation of COVID-19. *Clin Chim Acta* 2021; 519:172-182.
- [50] Booth AL, Abels E, McCaffrey P. Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Mod Pathol* 2021;34:522–531.
- [51] Pan P, Li Y, Xiao Y, Han B, Su L, Su M, Li Y, Zhang S, Jiang D, Chen X, et al. Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation. *J Med Internet Res* 2020;22:e23128.
- [52] Yadaw AS, Li YC, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digit Health* 2020;2(10):516–525.
- [53] Mazlumi R, Abazari SR, Nafarieh F, Aghsami A, Jolai F. Statistical analysis of blood characteristics of COVID-19 patients and their survival or death prediction using machine learning algorithms. *Neural Computing and Applications* 2022;34:14729-14743.
- [54] Assaf D, Gutman YA, Neuman Y, Segal G, Amit S, Gefen-Halevi S, Tirosh A. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Intern Emerg Med* 2020;15(8):1435–1443.
- [55] Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020;2:283-288.
- [56] Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nature Methods* 2018;15:233–234.
- [57] Kourou K, Exarchos TP, Exarchos K, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comp Struc Biotech J* 2015;13: 8–17.
- [58] Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020;2:283-288.

- [59] Ahmad A, Garhwal S, Ray SK, Kumar G, Malebary SJ, Barukab OM. The number of confirmed cases of COVID-19 by using machine learning: Methods and challenges. *Arch Comput Meth Eng* 2021; 28(4):2645–2653.
- [60] Zierk J, Ganslandt T, RauhM, Metzler M, Strasser E. Data mining of reference intervals for coagulation screening tests in adult patients. *Clin Chim Acta* 2019; 499: 108-114.
- [61] Zhou R, Wang W, Padoan A, Wan Z, Feng X, Han Z, Chen C et al.. Traceable machine learning real-time quality control based on patient data. *Clin Chem Lab Med* 2022; 60(12):1994-2004.
- [62] Obstfeld AE, Patel K, Boyd JC, Drees J, Holmes DT, Ioannidis JPA, Manrai AK. Data mining approaches to reference interval studies. *Clin Chem* 2021;67:1175–81.
- [63] Randell EW, Yenice S, Khine Wamono AA, Orth M. Autoverification of test results in the core clinical laboratory. *Clin Biochem* 2019;73:11-25.
- [64] Ma C, Wang X, Wu J, Xinqi C, Liangyu X, Fang X, Qiu L. Real-world big-data studies in laboratory medicine: Current status, application, and future considerations. *Clin Biochem* 2020;84:21-30.
- [65] Zhang X, Pérez-Stable EJ, Bourne PE, Peparah E, Duru OK, Breen N, et al. Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. *Ethn Dis* 2017; 27(2): 95–106.
- [66] Zhang JJ, Cao YY, Tan G, Dong X, Wang BC, Lin J. et al. Clinical, radiological, and laboratory characteristics and risk factors for severity and mortality of 289 hospitalized COVID-19 patients. *Allergy* 2021; 76: 533– 550.
- [67] Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z. et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020;395:1054–1062.
- [68] Bastug A, Bodur H, Erdogan S, Gokcinar D, Kazancioglu S, Kosovali BD, Bahadır Orkun Ozbay BO, Gok G, et al. Clinical and laboratory features of COVID-19: Predictors of severe prognosis. *Int Immunopharmacol* 2020; 88:106950.

- [69] Leulseged TW, Hassen IS, Ayele BT, Tsegay YG, Abebe DS, Edo MG, et al. Laboratory biomarkers of COVID-19 disease severity and outcome: findings from a developing country. *PloS One* 2021;16(3):e0246087.
- [70] Liang W, Yao J, Chen A, Lv Q, Zanin M, Liu J, Wong S, Li Y, et al. Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun* 2020;11(1):3543.
- [71] Hwangbo S, Kim Y, Lee C, Lee S, Oh B, Moon MK, Park T. Machine learning models to predict the maximum severity of COVID-19 based on initial hospitalization record. *Front Public Health* 2022;10:1007205.
- [72] Yang Q, Li JX, Zhang ZJ, Wu XC, Liao TQ, Yu SY, et al. Clinical characteristics and a decision tree model to predict death outcome in severe COVID-19 patients. *Bmc Infect Dis* 2021;21:783.
- [73] Li S, Lin Y, Zhu T, Fan M, Xu S, Qiu W, Xu S. Development and external evaluation of prediction models for mortality of COVID-19 patients using machine learning method. *Neural Comput Appl* 2021; 1–10.
- [74] Khanzode CA, Sarode RD. Advantages and disadvantages of artificial intelligence and machine learning: a literature review. *Int J Lib Inf Sci* 2020;9(1):30-36.