T.C.
NECMETTİN ERBAKAN
ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİ İLE TAM KAN SAYIMI SONUÇLARINDAN COVID-19 TEST SONUÇLARININ TAHMİNİ

Aybüke BOZKURT

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Temmuz/2021
KONYA

TEZ KABUL VE ONAYI

Aybüke Bozkurt tarafından hazırlanan "Veri Madenciliği Yöntemleri ile Tam Kan Sayımı Sonuçlarından COVID-19 Test Sonuçlarının Tahmini" adlı tez çalışması 12/07/2021 tarihinde aşağıdaki jüri tarafından oy birliği ile Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Dr. Öğr. Üyesi Onur İNAN	……..
Danışman Dr. Öğr. Üyesi Ayşe Merve ACILAR	……..
Üye Dr. Öğr. Üyesi Cengiz SERTKAYA	……..

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …./…/20.. gün ve …….. sayılı kararıyla onaylanmıştır.

Prof. Dr. İbrahim KALAYCI
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Aybüke BOZKURT

Tarih: 28 Haziran 2021

ÖZET

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ YÖNTEMLERİ İLE TAM KAN SAYIMI SONUÇLARINDAN COVID-19 TEST SONUÇLARININ TAHMİNİ

Aybüke BOZKURT

Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Ayşe Merve ACILAR,
Dr. Öğr. Üyesi Cengiz SERTKAYA

2021, 65 Sayfa

Jüri
Dr. Öğr. Üyesi Ayşe Merve ACILAR
Dr. Öğr. Üyesi Cengiz SERTKAYA
Dr. Öğr. Üyesi Onur İNAN

2019 yılında Çin'in Wuhan kentinde ilk vakaları görülen COVID-19 hastalığı tüm dünyayı etkisi altına almıştır. Başlangıçta sebebi belli olmadığı ve grip, soğuk algınlığı gibi hastalıklarla benzer etkileriyle karşılaşıldığı için hızla yayılmış ve Dünya Sağlık Örgütü tarafından pandemi olarak ilan edilmiştir.

Hastalığın hızla yayılımının önüne geçmek ve teşhisini hızlandırmak için yeni yöntemler aranarak, makine öğrenmesi algoritmalarından faydalanılmıştır. COVID-19 şüphesi ile hastanelere ulaşan bireyler içerisinde hastane verileri bir araya getirilerek veri setleri oluşturulmuştur.

Brezilya'daki Albert Einstein Hastane' sini ziyaret eden bireylere ait rutin kan sayımı sonuçları ve COVID-19 test sonuçları kullanılarak oluşturulan bu tez çalışmasında, eksik verilerin tamamlanması için K-En Yakın Komşu(KNN) algoritması, dengesiz veri problemi için SMOTE algoritması, gürültülü verilerin temizlenmesi için dağılım grafikleri ve özellik seçimi için Temel Bileşen Analizi (TBA) kullanılarak veri seti oluşturulmuştur. Oluşturulan veri seti makine öğrenmesi algoritmalarından Destek Vektör Makineleri, Rastgele Orman ve Naive Bayes algoritmalarıyla sınıflandırılarak test edilmiştir. Elde edilen sonuçlar ışığında Rastgele Orman algoritması %99.2 genel doğruluk ile en yüksek başarıyı elde etmiştir.

Anahtar Kelimeler: COVID-19, Destek Vektör Makineleri (DVM), K-En Yakın Komşu(KNN), Naive Bayes, SMOTE

ABSTRACT

MS THESIS

PREDICTION OF COVID-19 TEST RESULTS FROM WHOLE BLOOD COUNT RESULTS BY DATA MINING METHODS

Aybüke BOZKURT

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF NECMETTİN ERBAKAN UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE
IN COMPUTER ENGINEERING

Advisor: Assist. Prof. Ayşe Merve ACILAR,
Assist. Prof. Cengiz SERTKAYA

2021, 65 Pages

Jury
Assist. Prof. Ayşe Merve ACILAR
Assist. Prof. Cengiz SERTKAYA
Assist. Prof. Onur İNAN

The COVID-19 disease, the first cases of which were seen in Wuhan, China in 2019, has affected the whole world. It spread rapidly and was declared a pandemic by the World Health Organization, as the cause was not clear at the beginning and similar effects were encountered with diseases such as flu and colds.

In order to prevent the rapid spread of the disease and to accelerate its diagnosis, new methods were sought and machine learning algorithms were used. Data sets were created by bringing together hospital data among individuals who reached hospitals with the suspicion of COVID-19.

In this thesis study, which was created using routine blood count results and COVID-19 test results of individuals visiting Albert Einstein Hospital in Brazil, the K-Nearest Neighbor (KNN) algorithm to complete the missing data, the SMOTE algorithm for the unbalanced data problem, the noisy data. The data set was created using scatter plots to clean up the data and Principal Component Analysis (PCA) for feature selection. The generated data set has been tested by classifying with Support Vector Machines, Random Forest and Naive Bayes algorithms from machine learning algorithms. In the light of the results obtained, the Random Forest algorithm achieved the highest success with an overall accuracy of 99.2%.

Keywords: COVID-19, Support Vector Machine (DVM), K-Nearest Neighbor(KNN), Naive Bayes, SMOTE

ÖNSÖZ

Tezimin hazırlanması sırasında en başından beri yardımlarını ve bilgilerini esirgemeyen danışmanım Dr. Öğr. Üyesi Ayşe Merve Acılar' a ve aynı zamanda ilk çalışma ortamımda bana destek olan, ikinci danışmanım olarak tezimin ilerlemesinde katkıda bulunan Dr. Öğr. Üyesi Cengiz Sertkaya' ya, son olarak ise desteklerinden dolayı aileme ve başından beri beni yüreklendiren iş arkadaşım Arş. Gör. Sema Çiftçi' ye teşekkür ederim.

Aybüke BOZKURT
KONYA-2021

İÇİNDEKİLER

ŞEKİLLER LİSTESİ

ÇİZELGELER LİSTESİ

SİMGELER VE KISALTMALAR

Kısaltmalar

CPK: Creatine Phosphokinase (Keratinin Fosfokinaz)

CRISP-DM: Veri Madenciliği için Çapraz Endüstri Standart Süreci

DN: Doğru Negatif

DP: Doğru Pozitif

DSÖ: Dünya Sağlık Örgütü

DTX: Karar Ağaçları Açıklayıcı

DVM: Destek Vektör Makineleri

GBDT: Gradyan Artırılmış Karar Ağaçları

GBT: Gradyan Artırıcı Ağaçlar

GLMNET: Kement-Elastik Net Düzenlenmiş Genelleştirilmiş Doğrusal Ağ Algoritması

INR: International Normalized Ratio (Uluslararası Normalleştirilmiş Oran)

KNN: K-En Yakın Komşu

LASSO: En Az Mutlak Büzülme Ve Seçim Operatörü Lojistik Regresyon Modeli

LR: Lojistik Regresyon

MCH: Mean Corpuscular Hemoglobin (Ortalama Korpüsküler Hemoglobin)

MCHC: Ortalama Korpüsküler Hemoglobin Konsantrasyonu

MCV: Mean Corpuscular Volüme (Ortalama Korpüsküler Hacim)

mRMR: Maksimum Alaka Düzeyi Minimum Artıklık Algoritması

NN: Sinir Ağları

PT: Prothrombin Time (Protrombin Zamanı)

PTT: Partial Thromboplastin Time (Kısmi Tromboplastin Zamanı)

RDW: Red Blood Cell Distribution Width (Kırmızı Kan Hücresi Dağılım Genişliği)

RF Rastgele Orman Algoritması

RT-PCR: Ters Transkripsiyon Polimeraz Zincir Reaksiyonu

SHAP: Shapely Additive Explanations

SMOTE: Sentetik Hazırlık Yüksek Hızla Örnekleme Yöntemi

TBA: Temel Bileşenler Analizi

XGBoost: Aşırı Gradyan Artırma

YN: Yanlış Negatif

YP: Yanlış Pozitif

YSA: Yapay Sinir Ağları Algoritması

1. GİRİŞ

Geçmişten günümüze gelinceye kadar insanlık birçok hastalıkla baş etmeye çalışmıştır. Bu hastalıklar ile başa çıkmak kimi zaman kolay olsa da, kimi zaman ise tüm insanoğlunu etkileyecek şekilde büyüyerek salgın boyutuna ulaşmışlardır. Salgınların üstesinden gelebilmek için tüm dünya birlik olmuş ve salgının etkilerini en aza indirecek ve kurtulmayı sağlayacak aşı çalışmaları ve tedavi yöntemleri aramışlardır. 2019 yılının son periyodunda, ilk olarak Çin'in Wuhan kentinde başlayıp ardından büyüyerek tüm dünyaya yayılan nedeni bilinmeyen pnömoni vakaları ortaya çıkmıştır (Huang vd., 2020). Dünya Sağlık Örgütü yapılan incelemeler sonucunda bu vakaların daha önce insanlarda karşılaşılmayan bir tür olduğunu vurgulayarak 7 Ocak 2020 tarihinde bu virüsü 2019-nCoV olarak isimlendirmiştir (Chen vd., 2020). İlerleyen dönemlerde bu virüsün Şiddetli Akut Solunum Yetmezliği Sendromu olan SARS-CoV'a benzerliği sebebiyle, virüsü SARS-CoV-2 olarak adlandırmıştır (Culp, 2020a). SARS-CoV-2' nin sebep olduğu bu hastalığa ise Koronavirüs Hastalığı (COVID-19) ismi verilmiştir. Vakaların önlenemez olması ve ortalığa çıktığı günden itibaren çok fazla insana yayılması sebebiyle Dünya Sağlık Örgütü(DSÖ) 11 Mart 2020 tarihinde pandemi ilan etmiştir (Culp, 2020b).

COVID-19 salgınının belirtilerinin diğer yaygın hastalıklar (grip, soğuk algınlığı vb.) ile benzerliği sebebiyle teşhis edilmesi zorlaşmıştır. Vaka sayılarının ortalama %40' ında hafif (öksürük, ateş vb.), %40' ında orta, %15' inde şiddetli ve %5' inde ise kritik seviyede hastalık geçireceği tahmin edilmektedir (World Health Organization, 2020). DSÖ'nün ortaya çıkan bu vakalar için önerisi, enfekte olan vakaları belirleyerek, bu vakaları izole edip izlemek ve bulaşmasını önlemek için hastaların erken taramasını yapmak olmuştur (World Health Organization, 2020). Bu nedenle, virüsün hızını en aza indirmek ve erken teşhisi hızlandırmak amacıyla tüm dünyada büyük bir çaba gösterilmiştir.

COVID-19 hastalığı teşhisi için kullanılan ilk tanı testi, Ters transkripsiyon Polimeraz Zincir Reaksiyonudur (RT-PCR)(Döhla et al., 2020; Jin et al., 2020). Amerika Birleşik Devletleri Hastalık Kontrol ve Önleme Merkezi, bu tanı testinin yeterli olduğunu bildirmiş ve üst solunum yolundan alınan örnekler bu şekilde toplanmıştır (Interim Guidelines for Clinical Specimens for COVID-19 | CDC, 2021). İlk test sonucu negatif gelen hastalarda COVID-19 şüphesi devam ediyorsa test belirli zaman aralıkları ile tekrarlanabilmektedir. Çin'de 51 COVID-19 belirtisine sahip olan hastalar ile yapılan bir

çalışmada; RT-PCR testinin ilk test sonucunda negatif çıktığı, ancak seri testler sonucunda hastaların COVID-19 tanısının konduğu belirtilmiştir (Fang vd., 2020). RT-PCR testinin kesin tanı ortaya koyamaması, maliyetinin yüksek olması, zaman alıcı olması ve birçok ülkede mevcut olmaması dezavantajları sebebiyle daha ucuz, erişimi kolay ve her yerde kullanılabilen bir yöntemin bulunma zorunluluğu ortaya çıkmıştır (Alves et al., 2021; Li et al., 2020).

COVID-19 hastalığının teşhisinde rutin kan sonuçları önemli bir rol oynamaktadır. Bu alanda yapılan birçok çalışmada hastalığa sahip vakaların kan değerlerinin önemli bir değişiklik gösterdiği ve bu değerlerin COVID-19 teşhisi için ilk aşamada önemli olduğu ifade edilmiştir (Chen vd., 2020; Fan vd., 2020; Liu vd., 2020; Tan vd., 2020). Bu nedenle, RT-PCR testine ek olarak bölüm 2' de görüldüğü gibi makine öğrenmesi algoritmalarının rutin kan sayımı değerlerini öğrenerek ayırt edebilmesi sayesinde COVID-19 hastalığının erken teşhisine yönelik çalışmalar yapılmıştır. (Alves vd., 2021; Bhandari vd., 2020; Meng vd., 2020; Schwab vd., 2020; Soltan vd., 2020; N. Zhang vd., 2020).

1.1. Tezin Amaç ve Kapsamı

Bu tez çalışması, rutin laboratuvar verilerini kullanarak COVID-19 hastalığının erken teşhisi için farklı makine öğrenmesi algoritmalarının kullanıldığı modeller oluşturmuştur. Modeller için, öncelikle veri setinin sahip olduğu problemlere (eksik, dengesiz ve gürültülü veri) çözüm önerisi sunulmuş ve TBA ile özellik seçimi yapılarak COVID-19 vakalarında etkili olan rutin laboratuvar değerlerinin hangileri olduğu ortaya çıkarılmış ve oluşturulan veri seti ile COVID-19 hastalığının tahmini için Rastgele Orman, Destek Vektör Makineleri ve Naive Bayes makine öğrenmesi algoritmalarından yararlanılmıştır.

1.2. Tez Organizasyonu

Bu tez çalışmasının bölümleri aşağıdaki gibi organize edilmiştir.

Birinci bölümde; teze genel bir bakış açısı kazandırmak için temel bilgilere giriş yapılmış ve tezin amaç ve kapsamı hakkında bilgi verilmiştir.

İkinci bölümde; literatürde bulunan COVID-19 rutin laboratuvar test sonuçları kullanılarak yapılan mevcut çalışmalar ile bu tezin literatüre yapacağı katkılardan bahsedilmiştir.

Üçüncü bölümde; veri madenciliğinin tanımından başlanarak tez çalışmasında kullanılan veri seti tanımlanması, tez çalışmasında kullanılacak veri setinin oluşturulması için kullanılan veri ön işlem adımları, veri setinin sınıflandırılması ve modelleri oluşturmak için kullanılan Destek Vektör Makineleri, Rastgele Orman ve Naive Bayes algoritmaları ile ilgili bilgiler ile model değerlendirme ölçütleri hakkında bilgiler verilmiştir.

Dördüncü bölümde; COVID-19 veri seti analizinden başlayarak, veri setinin hazırlanması ve modellenmesi için yapılan deneysel çalışmalardan ve değerlendirme için ise makine öğrenmesi algoritmalarının başarılarından elde edilen bulgular açıklanmıştır.

Son bölümde ise; tez çalışmasında yapılan uygulamaların sonuçları, tez çalışmasında uygulanan adımların literatürle karşılaştırılması ve öneriler hakkında bilgi verilmiştir.

2. KAYNAK ARAŞTIRMASI

COVID-19' un erken teşhisini gerçekleştirmek için rutin laboratuvar test verileri ve klinik veriler kullanılarak oluşturulan makine öğrenmesi tabanlı literatürde dikkat çeken bazı uygulamalar şu şekildedir:

AlJame et al. , (2020), COVID-19' un erken tespiti için rutin kan testlerini kullanarak toplu bir öğrenme modeli geliştirmişlerdir. Bu tahmini gerçekleştirmek için ekstra ağaç, rastgele orman ve lojistik regresyon sınıflandırma algoritmaları ile aşırı gradyan artırma algoritması olan XGBoost sınıflandırma algoritmasını birleştiren ERLX adında bir model geliştirmişlerdir. Geliştirilen model, Brezilya'daki Albert Einstein Hastanesi'nden toplanan 559 tanesi COVID-19 hastası olan 5644 bireye ait rutin kan testi kayıtlarını içeren veri seti üzerinde %99.88 genel doğruluk oranı elde etmiştir.

Alves et al., (2021), rutin kan testleri kullanarak COVID-19 erken teşhisi için makine öğrenmesi tekniklerine dayalı bir model geliştirmişlerdir. Brezilya'da bulunan Albert Einstein Hastanesi'ndeki 608 hastaya ait rutin kan testlerini kullanarak farklı makine öğrenmesi algoritmaları ile sınıflandırma yapmışlardır. Bu sınıflandırma için karar ağaçları açıklayıcı (DTX) ve rastgele orman algoritmalarını kullanmışlardır. Oluşturulan modelde rastgele orman algoritması %88 sınıflandırma doğruluğu en iyi başarıyı elde etmiştir.

Meng et al., (2020), COVID-19 teşhisinin kaynak eksikliğini giderebilmek için laboratuvar sonuçlarını kullanarak makine öğrenmesi tabanlı bir model geliştirmişlerdir. Geliştirdikleri modeli kullanarak COVID-19 teşhis yardım uygulaması adında bir uygulama tasarlamışlardır. Bu sınıflandırma işlemi için kullanılan algoritma çok değişkenli lojistik regresyon olup, veri seti ise Batı Çin Hastanesi'nden alınan 602 hastaya ait veriyi içermektedir. Elde edilen sonuçlar değerlendirildiğinde, pozitif hasta tahmin sınıflandırma doğruluğu %86,25 iken, negatif hasta tahmin sınıflandırma doğruluğu ise %84,62' dir.

J. Wu et al., (2020), birden fazla kaynaktan topladıkları rutin laboratuvar test sonuçlarını kullanarak COVID-19 tespitini gerçekleştirmek için makine öğrenmesi tabanlı bir model oluşturmuşlardır. Bu tespiti gerçekleştirmek için kullandıkları algoritma rastgele orman algoritmasıdır. Oluşturulan model, Çin'deki farklı hastanelerden toplanan

169 şüpheli hastaya ait toplamda 253 örnekten oluşan veri seti üzerinde %96.95 genel doğruluk oranı elde etmiştir.

G. Wu et al., (2020), COVID-19 hastalığının tespiti için laboratuvar bulgularını kullanarak makine öğrenmesi tabanlı bir model geliştirdiler. Geliştirilen model için maksimum alaka düzeyi minimum artıklık (mRMR) algoritması ve en az mutlak büzülme ve seçim operatörü (LASSO) lojistik regresyon modeli kullanılmıştır. COVID-19 virüsüne sahip 110 hastanın (59 taburcu edilen ve 51 hayatta kalmayan hasta dâhil) verileri kullanılarak yapılan tahmin sonucunda model %98 duyarlılık ve %91 özgüllük elde etmiştir.

Yan et al., (2020),epidemiyolojik ve klinik verilere dayalı olarak en yüksek riske sahip COVID-19 hastalarını hızlı bir şekilde tahmin edebilmek için aşırı gradyan artırma (XGBoost) makine öğrenmesi metodunu kullanmışlardır. Çin'in Wuhan Tongji Hastanesi'nden elde ettikleri 375 hastaya ait veriyi kullanarak %90' dan fazla başarı elde etmişlerdir.

Feng et al., (2020), COVID-19 erken teşhisi için tanısal bir yardım modeli geliştirmişlerdir. Bu model, klinik belirtiler, rutin laboratuvar testleri ve hastaneye yatışla ilgili diğer klinik bilgilerin de dâhil olduğu Çin'in Pekin Halk Kurtuluş Ordusu Genel Hastanesi'nden toplanan 132 hastaya ait veri seti kullanılarak test edilmiştir. Geliştirilen modelde, en az mutlak büzülme ve seçim operatörü (LASSO) ile lojistik regresyon modeli, ridge regülasyonlu lojistik regresyon ve karar ağaçları sınıflandırma algoritmalarını kullanılmıştır. Elde edilen sonuçlar değerlendirildiğinde, LASSO ile lojistik regresyon modeli %93,8 sınıflandırma doğruluğu ile diğer sınıflandırma algoritmalarına göre daha yüksek sınıflandırma doğruluğu elde etmiştir.

Soares, (2020), şüpheli COVID-19 vakalarının tespiti için kan incelemelerine dayanan makine öğrenmesi tabanlı bir çerçeve tasarlamıştır. Tasarlanan modelde, destek vektör makineleri (DVM), SMOTEBoost ve topluluk algoritmalarının birleşiminden oluşan ER-CoV isimli karma bir model kullanılmıştır. Bu tahmini gerçekleştirmek için kullanılan veri seti 81 doğrulanmış COVID-19 hastasına ait 599 kan örneğinden oluşan Brezilya'daki Albert Einstein Hastanesi'ne ait olup, ER-CoV modeli %86,78 sınıflandırma doğruluğu elde etmiştir.

Banerjee et al., (2020), COVID-19' un erken teşhisi için hastaların kan testlerini kullanarak dört makine öğrenimi modelini test etmişlerdir. Bu modelde, rastgele orman

(RF), yapay sinir ağları (YSA), lojistik regresyon (LR) ve Kement-elastik net düzenlenmiş genelleştirilmiş doğrusal ağ (GLMNET) algoritmaları kullanılmıştır. Algoritmalar, 81 doğrulanmış vakaya ait 598 kan örneğinden oluşan Brezilya'da bulunan Albert Einstein Hastanesi'ne ait veri seti ile test edilmiştir. Elde edilen sonuçlar değerlendirildiğinde, yapay sinir ağları algoritması normal servisteki hastalar için %95 sınıflandırma doğruluğu , hastaneye kabul edilmeyen hastalar için ise %80-86 sınıflandırma doğruluğu ile diğer algoritmalardan daha yüksek sınıflandırma doğruluğu elde etmiştir.

Brinati et al., (2020), COVID-19 tespiti için rutin kan örneklerini kullanarak farklı makine öğrenmesi sınıflandırma algoritmalarını test etmişlerdir. Kullanılan modeller, karar ağaçları, aşırı derecede rastgele ağaçlar, K-en yakın komşular, lojistik regresyon, naive bayes, rastgele orman ve destek vektör makinaları algoritmalarıdır. İtalya'daki San Raffaele Hastanesi'ne kabul edilen 279 hastaya ait rutin kan sonuçları kullanılarak yapılan değerlendirme sonucunda Rastgele Orman algoritması %86 doğruluk oranı ile en iyi sınıflandırma doğruluğunu elde etmiştir.

Batista et al.,(2020), COVID-19 tespiti tahmini için acil bakım kan örneklerini kullanarak makine öğrenmesi tabanlı bir model geliştirmişlerdir. Bu tahmini gerçekleştirmek için sinir ağları, rastgele orman, gradyan artırıcı ağaçlar (GBT), lojistik regresyon ve destek vektör makineleri algoritmalarını kullanmışlardır. Sınıflandırma için kullanılan veri seti Brezilya'daki Albert Einstein Hastanesi'ne ait olup 102 doğrulanmış COVID-19 vakası ile 235 kan örneğinden oluşmaktadır. Elde edilen sonuçlar değerlendirildiğinde, destek vektör makineleri %85 sınıflandırma doğruluğu ile en yüksek başarıyı elde etmiştir.

Bao et al., (2020), COVID-19 vakalarının erken tespiti için rutin kan testlerinden yararlanarak makine öğrenmesi tabanlı bir model geliştirmişlerdir. Bu tespit için kullanılan algoritmalar rastgele orman ve destek vektör makineleri algoritmalarıdır. Modeli geliştirmek için kullanılan veri seti Çin'deki Kunshan Halk Hastanesi'nden ve Wuhan Birlik Hastanesi'nden toplanan 294 kan örneğinden oluşmaktadır. Elde edilen sonuçlar değerlendirildiğinde, destek vektör makinelerinin %84 başarı ile en yüksek başarı elde ettiği ifade edilmiştir.

Kukar et al., (2020), COVID-19 tespiti için Slovenya Üniversitesi Tıp Merkezi'nden toplanan çeşitli bakteriyel ve enfeksiyona sahip 5333 kan örneğini

kullanmışlardır. Bu tespit için kullandıkları algoritma aşırı gradyan artırma (XGBoost) makine öğrenmesi algoritması olup %97 sınıflandırma doğruluğu elde etmişlerdir.

de Freitas Barbosa et al., (2020), COVID-19' un erken teşhisi için Brezilya'daki Albert Einstein Hastanesi'nden toplanan 559 doğrulanmış hastaya ait 5644 veri örneğini kullanarak bir model geliştirmişlerdir. Bu tahmini gerçekleştirmek için kullanılan algoritmalar, çok katmanlı algılayıcı, destek vektör makineleri, rastgele orman, rastgele ağaç, bayes ağları ve naive bayes algoritmalarını kullanmışlardır. Elde edilen sonuçlar, bayes ağlarının %95,159 sınıflandırma doğruluğu ile diğer algoritmalara göre daha yüksek başarı elde ettiğini göstermiştir.

Yang et al., (2020), COVID-19 tespiti için 27 kan örneğine ek olarak hastaların demografik özelliklerini(yaş, cinsiyet vb.) kullanarak makine öğrenmesi tabanlı bir model oluşturmuşlardır. Bu tespit için lojistik regresyon, karar ağaçları, rastgele orman ve gradyan artırılmış karar ağaçları (GBDT) algoritmaları kullanılmıştır. Modelde kullanılan veri seti New York Presbiteryen Hastanesi'nden toplanan 3346 hastaya ait olup, GBDT algoritması %85,3 sınıflandırma doğruluğu ile diğer sınıflandırıcılar arasında en iyi sonucu vermiştir.

Sun et al., (2020), COVID-19 'un erken teşhisi için en iyi modeli belirlemek için destek vektör makineleri, lojistik regresyon, karar ağaçları, rastgele orman ve derin sinir ağları algoritmalarını kullanmışlardır. Bu tahmini gerçekleştirmek için kullandıkları veri seti Zhejang Eyaletindeki 18 hastaneden toplanan 912 hastaya ait klinik bulguları içermektedir. Elde edilen sonuçlar değerlendirildiğinde, lojistik regresyon modeli %91 sınıflandırma doğruluğu ile diğer algoritmalardan daha yüksek sınıflandırma doğruluğu elde etmiştir.

Langer et al., (2020), acil servislerdeki klinik, radyolojik ve rutin laboratuvar verilerini kullanarak COVID-19 hastalarının teşhisi için makine öğrenmesi temelli bir model geliştirdiler. Bu model, yapay sinir ağları, karar ağaçları, rastgele orman ve lojistik regresyon algoritmalarını kullanmıştır. İtalya Milano'da bulunan ana hastanelerden birinden toplanan 127 doğrulanmış vakaya sahip 199 veri örneği tahmin için kullanılmıştır. Elde edilen sonuçlar değerlendirildiğinde, yapay sinir ağları algoritması %91,4 sınıflandırma doğruluğu ile diğer algoritmalardan daha yüksek başarı elde etmiştir.

Soltan et al., (2020), rutin laboratuvar verilerini kullanarak COVID-19 'un erken teşhisi için iki model geliştirmişlerdir. Modeller, Birleşik Krallık Oxford Üniversitesi

Hastanesi'ne ait laboratuvar kan testleri, rutin kan sonuçları ve hasta başı kan gazı ölçümlerini kullanmışlar. Modellerden bir tanesi acil servislerdeki hastaların vaka tahminini yaparken, diğer model ise doğrulanmış vakaların hastaneye kaldırılıp kaldırılmayacağını belirler. Bu tahmini gerçekleştirmek için, lojistik regresyon, rastgele orman ve XGBoost algoritmaları kullanılmıştır. Elde edilen sonuçlara bakıldığında XGBoost algoritması %92,3 ile en iyi sınıflandırma doğruluğunu elde etmişlerdir.

Zhang et al., (2020), Huazhong Bilim ve Teknoloji Üniversitesi'ne bağlı Tongji Hastanesi'nden alınan doğrulanmış 137 vakaya ait klinik, kan ve idrar sonuçlarını değerlendirerek, COVID-19 vakalarının ağır hastalarını hafif belirtileri olanlar arasından tahmin etmek için makine öğrenmesi tabanlı bir model geliştirmişlerdir. Bu model, lojistik regresyon, destek vektör makineleri, rastgele orman, k en yakın komşu ve AdaBoost algoritmalarını kullanmıştır. Elde edilen sonuçlar destek vektör makinelerinin %81.48 sınıflandırma doğruluğu ile diğer algoritmalardan daha yüksek performans elde ettiğini göstermiştir.

Bhandari et al., (2020), COVID-19 hastalarında mortalite riskini tahmin etmek için makine öğrenmesi tabanlı bir model geliştirmişlerdir. Bu model, Hindistan'daki SMS Tıp Koleji'nde bulunan 70 hayatta kalan hastaların yaş, cinsiyet, belirtiler, rastgele kan şekeri ve tam kan sayımı değerlerini kullanarak lojistik regresyon modeli ile sınıflandırma gerçekleştirmiştir. Elde edilen sonuçlar değerlendirildiğinde lojistik regresyon modeli %70 sınıflandırma doğruluğu elde etmiştir.

Assaf et al., (2020), COVID-19 hastaları arasında, hastanede kaldıkları süre içerisinde kötüleşme riski taşıyan hastaları tahmin edebilmek için bir yöntem geliştirmişlerdir. Bu yöntem, Çin'deki Sheba Tıp Merkezi'nde bulunan 6995 hastaya ait laboratuvar sonuçlarını kullanarak rastgele orman, sinir ağları ve sınıflandırma ve regresyon karar ağacı modelini tahmin için kullanmıştır. Elde edilen sonuçlar, rastgele orman algoritmasının başarısının %92,9 ile en yüksek başarı elde ettiğini ifade etmiştir.

Turlapati & Prusty, (2020), COVID-19' un erken tespiti için makine öğrenmesi algoritmalarının sınıflandırma performansını artıracak Outlier-Smote adında bir yöntem geliştirmişlerdir. Geliştirilen model Brezilya'daki Albert Einstein Hastanesi'nde bulunan COVID-19 hastalığına sahip hastaların laboratuvar bulguları verisi ile test edilmiştir. Geliştirilen model, SMOTE(Sentetik Azınlık Yüksek Hızla Örnekleme Tekniği) ve

ADASYN algoritmalarıyla karşılaştırıldığında, sınıflandırma doğruluğunun daha yüksek olduğu ifade edilmiştir.

Yavaş vd., (2020), COVID-19 hastalığı şüphesi ile hastaneye başvuran vakaların laboratuvar test sonuçlarını kullanarak hastalığın erken teşhisini tahmin edebilmek için bir model geliştirmişlerdir. Geliştirilen model, Brezilya'daki Albert Einstein Hastanesi'nde bulunan COVID-19 şüphesi ile hastaneye başvuran 602 hastaya ait sonuçlar üzerinde test edilmiştir. Veri setinin dengesiz olması sebebiyle model, SMOTE algoritması ile veri dengesizliği problemine çözüm bularak yapay sinir ağları algoritması ile sınıflandırma gerçekleştirmiştir. Elde edilen sonuçlar değerlendirildiğinde, orijinal veri seti %86 sınıflandırma doğruluğu gösterirken, SMOTE ile dengelendikten sonra oluşturulan yeni veri seti %90 sınıflandırma doğruluğu göstermiştir.

Literatür incelemesi sonucunda elde edilen bilgiler ışığında, rutin laboratuvar sonuçları ile yapılan çalışmalarda veri setlerinin boyutları ve özellikleri az sayıdadır. Tez çalışması için seçilen Brezilya'daki Albert Einstein Hastanesi'nden alınan veri seti ile yapılan çalışmalarda veri setinin en çok dengesizlik problemine çözüm bulmak için çalışmalar yapılmıştır. Veri setinin sahip olduğu eksik veri problemi önemsenmeyerek veri seti küçültülerek %10' dan daha az bir hale getirilmiştir. Bu sebeple bu tez çalışması veri setinin her problemine çözüm bulabilmek için dengesizlik problemi için SMOTE ve eksik verilerin tamamlanması için KNN kullanarak bir veri analizi gerçekleştirilmiştir. Aynı zamanda veri setinde bulunan gürültünün giderilmesi için dağılım grafikleri kullanılmıştır. Literatürde çok fazla değinilmeyen COVID-19 hastalığı için önemli olan kan değerlerinin neler olduğunu belirleyebilmek amacı ile özellik seçimi için TBA kullanılarak, ortaya çıkarılan özelliklerdeki önem sıraları da RandomizedSearchCV ve SHAP ile ifade edilerek, COVID-19 erken teşhisi için geliştirilen modelde yüksek sınıflandırma doğruluğu elde edilmiştir.

3. MATERYAL VE YÖNTEM

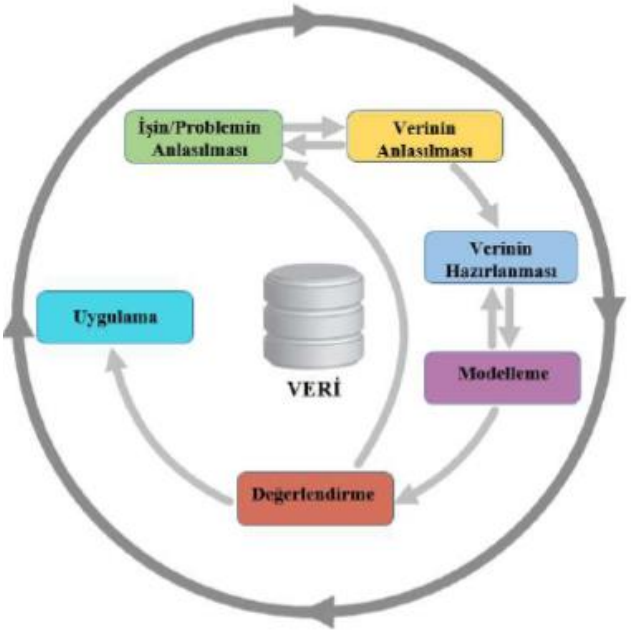
Bu bölümde öncelikle veri madenciliğinin tanımı verilecek, devamında ise CRISP-DM veri madenciliği metodolojisine uygun süreç tanımları bu tez çalışması özelinde açıklanacaktır.

3.1. Veri Madenciliği

Bilgisayarlar tarafından üretilen veriler, tek başına bir anlam ifade etmezler (Savaş vd., 2012). Belli bir amaç doğrultusunda işlenerek bir anlam ifade eden veriye ise bilgi denir (Tüzüntürk, 2010). Veri madenciliği ise veriden, bilgiye ulaşabilmek için kullanılan yöntemdir(Demircioğlu, 2019). Böylelikle, veriler arasındaki ilişkiler ve değişiklikler keşfedilerek, geleceğe yönelik çıkarımlarda bulunmak mümkündür. Veri Madenciliği bankacılık, pazarlama, tıp ve endüstri gibi alanlarda kullanılmaktadır.

Veri madenciliği projelerinde en çok tercih edilen süreç yönetim modeli olan CRISP-DM (Cross- Industry Standard Process for Data Mining - Veri Madenciliği için Çapraz Endüstri Standart Süreci) ilk olarak 2000 yılında SPSS, NCR, OHRA ve Daimler-Chrysler ortak çalışması ile geliştirilmiştir. Bu metodolojinin temel amacı, ele alınan konunun yaşam döngüsüne genel bakışı sağlamaktır. Bu yöntem, diğer yöntemlerden daha hızlı, daha az maliyetli ve daha çok verimli olarak veri madenciliği işlemlerini gerçekleştirebilmektedir.

CRISP-DM metodolojisi altı aşamadan oluşan bir süreci ifade etmektedir. Bu adımlar Şekil 3.1’ de gösterilmiştir.



Şekil 3.1. CRISP-DM Metodolojisi (Koçoğlu, 2017)

CRISP- DM metodolojisinde bulunan aşamalar altında bu tez çalışmasında kullanılan metotlar kısaca açıklanmıştır.

3.1.1. Tez Çalışmasında Kullanılan Veri Seti

Bu çalışmada, Kaggle çevrimiçi web tabanlı makine öğrenmesi platformunda paylaşıma açık olan bir COVID-19 veri seti kullanılmıştır. Veri seti, 28 Mart 2020 - 3 Nisan 2020 tarihleri arasında São Paulo, Brezilya'daki Albert Einstein Hastanesi'nde yapılan SARS-CoV-2 RT-PCR testi ve ek laboratuvar testlerini yapmak için numuneler toplanan 559'u doğrulanmış toplam 5644 hastaya ait 111 nitelikten oluşan anonimleştirilmiş verileri içermektedir. Veri setindeki tüm klinik veriler, ortalama sıfıra ve birim standart sapmaya uygun olacak şekilde standardize edilmiş şekilde paylaşılmıştır (Diagnosis of COVID-19 and Its Clinical Spectrum | Kaggle, 2020). Çizelge 3.1, 3.2 ve 3.3' te ise veri setinde bulunan özelliklerin türleri gösterilmiştir. Veri setinin analizinin ayrıntıları Bölüm 4.1' de anlatılacaktır**.**

Çizelge 3.1. Kategorik Değişkenler

Değişken Adı	Türü
Patient ID (Hasta ID)	Kategorik
SARS-Cov-2 exam result (SARS-Cov-2 test sonucu)	Kategorik
Respiratory Syncytial Virus (Solunum sinsityal virüsü)	Kategorik
Influenza A (Grip Virüsü A)	Kategorik
Influenza B (Grip Virüsü B)	Kategorik
Parainfluenza 1	Kategorik
CoronavirusNL63	Kategorik
Rhinovirus/Enterovirus	Kategorik
Coronavirus HKU1	Kategorik
Parainfluenza 3	Kategorik
Chlamydophila pneumoniae	Kategorik
Adenovirus	Kategorik
Parainfluenza 4	Kategorik
Coronavirus229E	Kategorik
CoronavirusOC43	Kategorik
Inf A H1N1 2009	Kategorik
Bordetella pertussis	Kategorik
Metapneumovirus	Kategorik
Parainfluenza 2	Kategorik
Influenza B, rapid test	Kategorik
Influenza A, rapid test	Kategorik
Strepto A	Kategorik
Urine – Esterase (İdrar – Esteraz)	Kategorik
Urine – Aspect (İdrar – Görünüm)	Kategorik
Urine – Hemoglobin (İdrar – Hemoglobin)	Kategorik
Urine - Bile pigments (İdrar - Safra pigmentleri)	Kategorik
Urine - Ketone Bodies (İdrar - Keton Cisimleri)	Kategorik
Urine – Urobilinogen (İdrar – Ürobilinojen)	Kategorik
Urine – Protein (İdrar – Protein)	Kategorik
Urine – Crystals (İdrar – Kristaller)	Kategorik
Urine - Hyaline cylinders (İdrar - Hiyalin silindirleri)	Kategorik
Urine - Granular cylinders(İdrar - Granül silindirler)	Kategorik
Urine – Yeasts (İdrar – Mayalar)	Kategorik
Urine - Color	Kategorik

Çizelge 3.2. Numerik Değişkenler

Değişken Adı	Türü
Patient age quantile	Numerik
Patient addmited to regular ward (1=yes, 0=no)	Numerik
Patient addmited to semi-intensive unit (1=yes, 0=no)	Numerik

Patient addmited to intensive care unit (1=yes, 0=no)	Numerik
Hematocrit (Hematokrit)	Numerik
Hemoglobin (Hemoglobin)	Numerik
Platelets (Trombosit)	Numerik
Mean platelet volüme (Ortalama Trombosit Hacmi)	Numerik
Red blood Cells (Kırmızı Kan Hücreleri)	Numerik
Lymphocytes (Lenfositler)	Numerik
Mean corpuscular hemoglobin concentration (MCHC)	Numerik
Leukocytes (Lökositler)	Numerik
Basophils (Basofiller)	Numerik
Mean corpuscular hemoglobin (MCH)	Numerik
Eosinophils (Eozinofiller)	Numerik
Mean corpuscular volume (MCV)	Numerik
Monocytes (Monositler)	Numerik
Red blood cell distribution width (RDW)	Numerik
Serum Glucose (Serum Glikoz)	Numerik
Neutrophils (Nötrofiller)	Numerik
Urea (İdrar)	Numerik
Proteina C reativa mg/dL	Numerik
Creatinine (Keratinin)	Numerik
Potassium (Potasyum)	Numerik
Sodium (Sodyum)	Numerik
Alanine transaminase	Numerik
Aspartate transaminase	Numerik
Gamma-glutamyltransferase	Numerik
Total Bilirubin	Numerik
Direct Bilirubin	Numerik
Indirect Bilirubin	Numerik
Alkaline phosphatase (Alkalin fosfataz)	Numerik
Ionized calcium (İyonize Kalsiyum)	Numerik
Magnesium (Magnezyum)	Numerik
pCO2	Numerik
Hb saturation (Hb Saturasyon)	Numerik
Base excess (Baz fazlalığı)	Numerik
pO2	Numerik
Fio2	Numerik
Total CO2	Numerik
pH	Numerik
HCO3	Numerik
Rods #	Numerik
Segmented	Numerik
Promyelocytes (Promyelositler)	Numerik
Metamyelocytes (Metamyelositler)	Numerik
Myelocytes (Miyelositler)	Numerik
Myeloblasts (Miyeblostlar)	Numerik
Urine – pH	Numerik
Urine – Density	Numerik
Urine – Leukocytes	Numerik
Urine - Red blood cells	Numerik
Relationship(Patient/Normal)	Numerik
International normalized ratio (INR)	Numerik
Lactic Dehydrogenase (Laktik Dehidrasyon)	Numerik
Vitamin B12	Numerik
Creatine phosphokinase (CPK)	Numerik
Ferritin	Numerik
Arterial Lactic Acid	Numerik
Lipase dosage	Numerik
Albumin	Numerik

Hb saturation (arterial blood gases)	Numerik
pCO2 (arterial blood gas analysis)	Numerik
Base excess (arterial blood gas analysis)	Numerik
pH (arterial blood gas analysis)	Numerik
Total CO2 (arterial blood gas analysis)	Numerik
HCO3 (arterial blood gas analysis)	Numerik
pO2 (arterial blood gas analysis)	Numerik
Phosphor	Numerik
ctO2 (arterial blood gas analysis)	Numerik

Çizelge 3.3. Tamamı Boş Değerler

Değişken Adı
Mycoplasma pneumoniae
Urine - Sugar
Urine-Nitrite
Arteiral Fio2
Partial thromboplastin time (PTT)
Prothrombin time (PT), Activity
D-Dimer

3.1.2. Verinin Hazırlanması

Veri madenciliği uygulamalarında birçok problemle karşılaşılabilir. Bunun başlıca sebebi, veri tabanlarında bulunan verinin eksik ya da sağlıksız bilgiler içermesinden kaynaklanmaktadır. Veri tabanlarında bulunan verinin net, eksiksiz ve dinamik olması gerekmektedir. Bu durum gerçekleşemediğinde, yapılan analizler yanlış stratejilerin oluşmasına sebep olacaktır. Veri madenciliğinde karşılaşılabilecek ilk problemler gürültülü ve eksik veri problemleridir.

Veri toplanması ya da veri girişi sırasında oluşabilecek sistem dışı hatalar gürültü olarak adlandırılmaktadır. Veri tabanları büyüdükçe pek çok niteliğin değeri yanlış girilebilir. Aynı zamanda veri toplanması sırasında oluşan ölçüm hataları da yanlışlıklara sebep olabilir (Savaş et al., 2012). Bu problemler veri madenciliğinin amacına tam olarak ulaşamamasına sebep olabilir. Bu sebeple ortaya çıkan gürültülü verilerin tespit edilmesi ve ihmal edilmesi gerekmektedir. Gürültülü verinin sınıflandırma üzerindeki etkisini araştıran çalışmalar sonucunda, algoritmaların başarısının doğrudan kötü etkilendiği ortaya konmuştur. Ancak, çalışmalarda sadece %10 oranında gürültü, veri setinden elenebilmektedir (Sever & Oğuz, 2002).

Eksik veri, veri tabanlarının büyüklüğünden ya da doğasından kaynaklanmaktadır. Bu veriler, istatiksel analizler veya sınıflandırma analizlerinde önemli sorunlar ortaya çıkarmaktadır. Bunun sebebi bu analizlerin yapılması için oluşturulan programlar veya algoritmalar, veri setinde bulunan verilerin tamamının dolu olduğu durumlarda çalışmaktadır (Savaş vd., 2012).

Verinin hazırlanması aşamasında, veri ön işlemesi gerçekleştirilmiştir. Bu tez çalışmasında gerçekleştirilen veri ön işleme adımları; eksik değerlerin tespiti, veri dengesizliği probleminin giderilmesi, gürültülü verilerin tespiti ve giderilmesi ve özellik çıkarımı olmak üzere dört bölümden oluşmaktadır.

3.1.2.1. Veri Setindeki Eksik Değerlerin Tespiti

KNN algoritması makine öğrenmesi sınıflandırma algoritmalarından biridir. Aynı zamanda, Troyanskaya et al., (2001), tarafından ortaya çıkarılmış bir eksik değer tamamlama yöntemidir. Bu en yakın komşu tabanlı yöntem, eksik değerleri tamamlamak için eksik veriye en yakın k örneği bularak, hesaplamayı gerçekleştirir ve veriyi doldurur (S. Zhang vd., 2018). KNN algoritması eksik verileri tahmin etmek için gerçek veri noktalarını kullanır. Aynı zamanda hem ayrık hem de sürekli değişkenlerle çalışır, bu özellikler de eksik veri tamamlanması için bu algoritmanın tercih edilirliğini artırmaktadır. KNN algoritması, veri noktaları arasındaki mesafeleri kullanarak çalışır ve elde edilen sonuçları benzerlik ölçüsü olarak kullanır. Öklid, Minkowski, Manhattan gibi farklı mesafe ölçüleri kullanılabilir. Ancak genel olarak en çok tercih edilen mesafe ölçümü Öklid'dir (Yılmaz & Aydın, 2019). Öklid uzaklık formülü eşitlik 3.1' deki gibidir:

$$\text{Öklid uzaklığı} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{3.1}$$

Burada a ve b iki noktayı temsil etmektedir. Veri setinde bulunan her eksik özellik, en yakın komşularından alınan değerler ile hesaplanır. Komşuların her birinin ortalaması alınarak ya da komşulara ait mesafeler hesaplanarak bir ağırlık değeri ortaya çıkarılır. Bu değere göre verilerde doldurma işlemi yapılmış olur. KNN algoritması, birden fazla eksik değeri olan özellikleri de tahmin edebilir (Choudhury vd., 2020).

3.1.2.2. Veri Dengesizliğinin Giderilmesi

İstatiksel analizlerde veya sınıflandırma analizlerinde kullanılan veri setlerinde her sınıftan hemen hemen aynı oranda örneğin olması beklenir. Fakat bunun olmadığı durumlarda veri setlerinde bir dengesizlik oluşur ve sınıflandırma gerçekleştirilirken örneğin fazla olduğu tarafa doğru bir eğilim gerçekleşir. Aynı zamanda azınlıkta olan sınıf ile sınıflandırma algoritması yeterince eğitilemediği için doğru ve başarılı bir sınıflandırma mümkün olmamaktadır. Bu durum istenen bir sonuç değildir. Sınıflandırıcının her bir sınıf için yüksek başarı elde etmesi amaçlanmaktadır (Bulut, 2016).

Veri setinde meydana gelen bu dengesizliği ortadan kaldırabilmek için kullanılabilecek çeşitli yöntemler bulunmaktadır. SMOTE (Synthetic Minority Over Sampling Technique) metodu bu sorunu çözmek için uygulanabilecek yöntemlerden birisidir.

Chawla et al., (2002), yaptıkları bir çalışmada veri setinde bulunan dengesizlik problemine çözüm bulabilmek amacı ile SMOTE algoritmasını önermişlerdir. Bu algoritma, azınlık olarak bulunan sınıfa ait kayıtların sayısını artırmak içim sentetik veri üreterek veri setinde bulunan dengesizliği ortadan kaldırmaya yönelik bir yöntemdir.

Sentetik veri üretimi şu şekilde gerçekleşmektedir (Chawla vd., 2002):

- İncelenen özellik vektörü ile en yakın komşusu arasındaki fark hesaplanır.
- Bu fark 0 ile 1 arasında rastgele bir sayı ile çarpılır ve söz konusu olan özellik vektörüne eklenir.
- Bu durum, iki belirli özellik arasındaki çizgi parçası boyunca rastgele bir noktanın seçilmesine sebep olur.
- Bu yaklaşım, azınlık sınıfının karar bölgesini daha genel olmaya zorlamaktadır, yani yapay örneklemler oluşturmaktadır.

3.1.2.3. Gürültülü Verilerin Tespiti ve Elenmesi

Gürültülü verilerin tespiti için her özelliğe ait dağılım grafikleri oluşturulmuş ve oluşturulan grafikler sayesinde özelliklerde bulunan mevcut gürültüler giderilmiştir. Ayrıntıları Bölüm 4.2.3’ te ifade edilmiştir.

3.1.2.4. Özellik Çıkarımı

Temel Bileşen Analizi (TBA), veri setinin boyutunu azaltırken birbirleri ile yüksek korelasyona sahip değişkenleri bir araya getirmektedir. Aynı zamanda, veri analizi için kullanılan büyük boyutlu veri setlerini mantıksal çerçevede küçültmeye yarayan yöntemdir. Bu yöntemde, orijinal veri setinde bulunan her bir temel bileşene lineer dönüşüm uygulanarak varyans hesaplamaları yapılmaktadır (Johnson & Wichern, 2007). İki değişkenin birbiri ile ilişkili olup olmadığını öğrenebilmek için bu iki değişkenin birbirlerine göre değişimlerini gösteren kovaryans değerleri hesaplanmaktadır. Kovaryans, sıralı iki veri kümesinden karşılık gelen elemanların aynı yönde hareket etmesinin bir ölçüsüdür. X ve Y olarak verilen iki değerin ilişkili olması pozitif kovaryans olarak isimlendirilirken, negatif kovaryans ise zıt ilişkiyi ifade etmektedir. Bu yöntemin amacı, çıkış parametresi için yüksek korelasyona sahip olan girdi parametrelerini seçmektir (Sertkaya & Yurtay, 2015).

TBA analizi için kullanılan kovaryans matrisinin hesaplanması denklemi eşitlik 3.2' deki gibidir.

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x}) - (Y_i - \bar{y}) \quad (3.2)$$

Tüm veri setine ait kovaryans matrisleri oluşturulduktan sonra matrisin özvektörleri ve öz değerlerinin hesaplanma işlemi gerçekleştirilir.

$$\det(B - \lambda I) = 0 \quad (3.3)$$

Eşitlik 3.3' de B veri setini yani bir kare matrisi göstermektedir. λ özdeğerleri, ve I ise birim matris anlamına gelmektedir. Özdeğerler hesaplandıktan sonra özvektörlere erişilir. Lineer dönüşüm yapıldıktan sonra yönü değişmeyen vektörleri özvektörleri göstermektedir. En büyük değere sahip özvektörler seçilerek temel bileşenler bulunmuş olur.

3.1.3. Modelleme

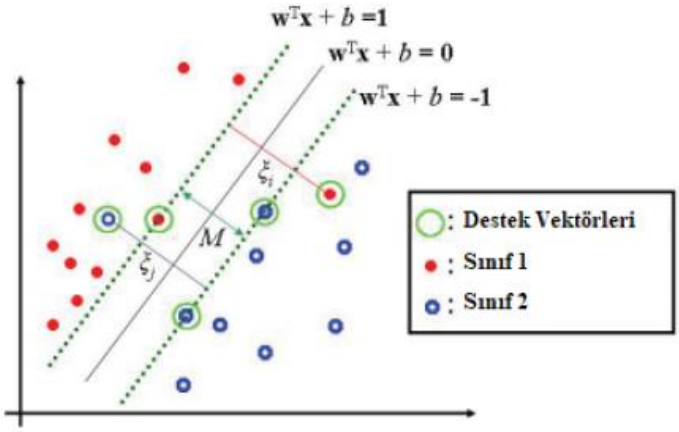
Sınıflandırma, bir veri seti içindeki verilerin ortak özellikleri kullanılarak sınıflara ayrılması yöntemidir. Aynı zamanda, eldeki verilere bakılarak, geleceğe yönelik

tahminlerde bulunmak için ve yeni eklenen bir veri ögesinin var olan sınıflara atamasının yapılabilmesi için kullanılır.

Sınıflandırma algoritmaları ise bu sınıflandırma işlemini gerçekleştirebilmek için geliştirilmiş algoritmalardır (Çınar 2019). Bu tez çalışmasında veri madenciliği modelleme aşamasında kullanılan makine öğrenmesi sınıflandırma algoritmaları Destek Vektör Makineleri, Rastgele Orman ve Naive Bayes algoritmalarıdır. Bu algoritmaların açıklamaları aşağıdaki gibidir.

3.1.3.1. Destek Vektör Makineleri

Destek Vektör Makinaları (DVM), 1992 yılında Boser, Guyan ve Vapnik tarafından ortaya atılan istatiksel teoriler üzerine kurulan bir makine öğrenmesi algoritmasıdır (Boser vd., 1992). Algoritma, bir düzlemde bulunan iki farklı grubu ayırmak için bir sınır çizer. Bu sınır, iki gruba ait olan verilere de en uzak yere çizilir. Test verisine ait değerler hangi gruba daha yakınsa, o değer grubun yeni üyesi olur. Sınır çizilirken, iki gruba da yakın iki farklı sınır çizgileri çizilir. Bu sınır çizgileri birbirlerine yaklaştırılarak ortak bir sınır belirlenir (Khorraminezhad vd., 2020). Sınır doğrusunun en uygun ve doğru yere konumlandırılması çok önemlidir. Şekil 3.2 destek vektörlerini ve sınıfları göstermektedir.



Şekil 3.2. Destek Vektör Makineleri (Yetginler, 2019)

Sınıflama için kullanılacak eğitim seti $\{(x_1, y_1), \dots, (x_n, y_n)\}$ olmak üzere $x_i \epsilon R^P$ ve $y_i \epsilon \{-1,1\}$ 'dir. y_i, x_i örneğinin sahip olduğu sınıf etiketidir. İki sınıfı ayıran sınır denklemi eşitlik 3.4' deki gibidir:

$$w^T x + b = 0 \tag{3.4}$$

Burada w ağırlık vektörünü ifade etmektedir.

Destek Vektör Makinelerinin sahip olduğu hiper parametreler ise aşağıdaki gibidir.

- Çekirdek Türü: Algoritmada kullanılan çekirdek türünü ifade etmektedir. 'Doğrusal', 'rbf', 'sigmoid' gibi çekirdek türleri bulunmaktadır. Varsayılan çekirdek türü 'rbf 'tir.
- C: düzenlileştirme parametresidir. Kesinlikle pozitif bir değer olmalıdır. Varsayılan değeri 1' dir.

3.1.3.2. Rastgele Orman

Rastgele Orman algoritması bir topluluk öğrenme yöntemidir ve sınıflandırma işlemini gerçekleştirirken çok sayıda karar ağacı kullanmaktadır. Burada amaç, sınıflandırma değerini yükseltmektir. Sınıflandırmada kullanılacak karar ağaçları, veri kümesinden rastgele seçilerek bir orman oluşturmaktadır (Farnaaz & Jabbar, 2016). Sınıflandırma sırasında farklı ağaçlar oluşturarak ağaçlar arasındaki korelasyonu düşük tutması ve standart sapması düşük olan sonuçlar ortaya çıkarması sebebiyle başarılı ve performansı yüksek bir algoritmadır.

Rastgele Orman algoritmasında dallara ayırma özelliğine sahip değişken, bütün değişkenler arasından rastgele olarak seçilen m adet değişken içerisinden belirlenmektedir. Her ağaçta bulunan m değeri sabit olup, genellikle √p olarak hesaplanmaktadır. Burada p değeri değişken sayısını göstermektedir.

Ormanda bulunan ağaçlar, dallar ve yapraklardan oluşmaktadır. Her bir özellik düğüm olarak nitelendirilmektedir. En sonda bulunan yapı 'yaprak' olarak isimlendirilirken en üstte bulunan yapı 'kök' olarak ve yaprak ve kök arasında kalan yapılar ise 'dal' olarak isimlendirilmektedir. Rastgele Orman algoritmasında, ağaç bütün veriye ait tek bir düğümle başlar ve eğer veri setinde bulunan örneklerin hepsi aynı sınıfta

bulunuyorsa düğüm, yaprak olarak bitmekte ve sınıf etiketi verilmektedir (Korkem, 2013).

Dallara ayırma ise gini indeksi ile ifade edilmektedir. Örneğin; dallara ayırmada kullanılan değişken bireyin trombosit düzeyi olarak seçildiğinde, ayırıcı kriter ise trombosit değerinin alt ve üst değeri olarak ayrılmaktadır. Bu işlemler yaprak düğümü elde edilene kadar devam etmektedir.

Rastgele Orman algoritmasının sahip olduğu hiper parametreler ise aşağıdaki gibidir:

- n_estimator: Bu değer ormandaki ağaç sayısını göstermektedir. 10' dan 100' e kadar değişen bir değer aralığı vardır. Varsayılan değeri 100' dür.
- criterion: Bu özellik bir bölünmenin kalitesini ölçmek için kullanılmaktadır. Gini indeksi ve entropi olarak iki kriteri mevcuttur. Varsayılan değeri gini değeridir.
- max_depth: Bu değişken ağaçtaki maksimum derinliği göstermektedir.
- min_samples_split: Bir düğümü bölmek için gereken minimum örnek sayısını ifade etmektedir. Varsayılan değeri 2' dir.
- min_samples_leaf: Bir yaprak düğümünde bulunması gereken minimum örnek sayısını ifade etmektedir. Varsayılan değeri 1' dir.
- max_features: Bu değişken en iyi bölünmeyi ararken bilinmesi gerekli olan özelliklerin sayısını ifade etmektedir. 'auto' , 'sqrt' ve 'log2' olarak belirlenebilmektedir. Varsayılan değeri 'auto' ' dur. 'Auto' değeri ve 'sqrt' seçilirse max_features değeri özellik sayısının karekökü'dür. Ancak 'Log2' seçilirse o zaman max_features değeri özellik sayısının log2' sinin hesaplanması ile elde edilmektedir.

3.1.3.3. Naive Bayes

Naive Bayes algoritması bir olasılık sınıflandırıcı algoritması olup, temelde bayes teoremine dayanmaktadır. Veri seti üzerinde yapılacak bir sınıflandırmada, verinin hangi sınıfa ait olabileceği olasılığı hesaplanır. Olasılık değerleri karşılaştırıldığında, verinin hangi sınıfa ait değeri daha yüksekse, veri o sınıfın üyesi olur (Martínez Torres vd., 2019). Eşitlik 3.5 A ve B olmak üzere 2 olay için Bayes teoreminin formülünü göstermektedir.

$$P(A\backslash B) = \frac{P(B\backslash A)\text{x}P(A)}{P(B)} \quad (3.5)$$

Bu denklemde;

- P(A\B): B olayı gerçekleştiğinde, A olayının olma ihtimalini göstermektedir.
- P(B\A): A olayı gerçekleştiğinde, B olayının olma ihtimalini göstermektedir.
- P(A): A olayının gerçekleşme ihtimalini göstermektedir.
- P(B): B olayının gerçekleşme ihtimalini göstermektedir.

3.1.4. Değerlendirme

Sınıflandırma modellerinin başarı oranları hesaplanırken, doğru sınıfa ataması yapılan örnek sayısı ile yanlış sınıfa ataması yapılan örnek sayısı karşılaştırılır (Özlüer Başer et al., 2021). Çizelge 3.4' de verilen karmaşıklık matrisinden (confusion matris) yararlanarak hesaplanırlar. Karmaşıklık matrisinde bulunan sütunlar, modelin tahmini sonucunda elde edilen sınıf sayılarını gösterirken, satırlar ise test kümesine ait gerçek sayıları göstermektedir.

Çizelge 3.4 Karmaşıklık Matrisi

		TAHMİN EDİLEN SINIF	
		Sağlıklı	**COVID-19**
GERÇEK SINIF	**Sağlıklı**	DN	YP
	COVID-19	YN	DP

Karmaşıklık matrisinde;

Doğru Pozitif (DP) değeri doğru tahmin edilen pozitif hasta sayını,

Yanlış Negatif (YN) değeri, yanlış tahmin edilen negatif hasta sayısını,

Yanlış Pozitif (YP) değeri, yanlış tahmin edilen pozitif sınıf değerini,

Doğru Negatif (DN) değeri ise doğru tahmin edilen negatif hasta sayını temsil etmektedir.

Modellerin performans değerlendirmede kullanılan kavramlar ve hesaplanma şekilleri aşağıdaki gibidir:

Doğruluk (Accuracy) : Sınıflandırma modellerinin genel başarısıdır. Eşitlik 3.6 sınıflandırma başarısının hesaplanması için kullanılmaktadır.

$$Doğruluk = \frac{DP+DN}{DP+DN+YP+YN} \tag{3.6}$$

Duyarlılık (Recall) : Pozitif olan sınıfı doğru tespit edebilme ölçütüdür. Eşitlik 3.7 duyarlılık değerinin hesaplanması için kullanılmaktadır.

$$Duyarlılık = \frac{DP}{DP+YN} \tag{3.7}$$

Kesinlik (Precision) : Pozitif olan sınıfın ne kadar doğru olduğunu belirleyen ölçüttür. Eşitlik 3.8 kesinlik değerinin hesaplanması için kullanılmaktadır.

$$Kesinlik = \frac{DP}{DP+YP} \tag{3.8}$$

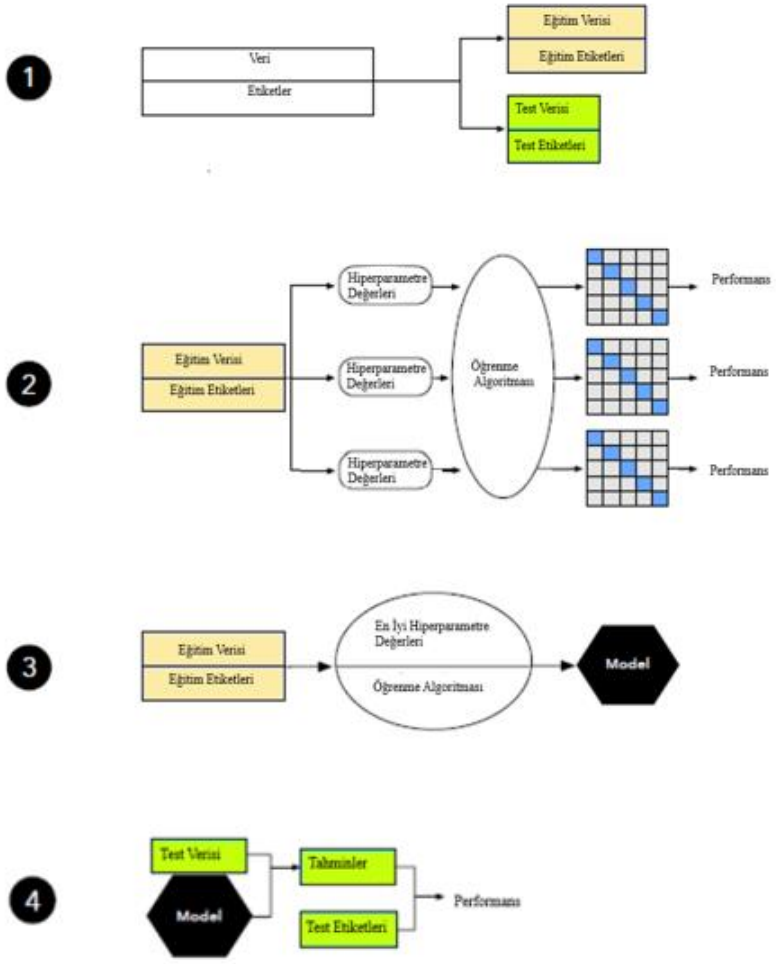
F1- Skor (F1- Score) : Duyarlılık ve kesinliğin harmonik ortalamasıdır. Eşitlik 3.9 ise f1-skor değerinin hesaplanması için kullanılmaktadır.

$$\text{F1- Skor= 2 x } \left(\frac{Kesinlik x Duyarlılık}{Kesinlik+Duyarlılık}\right) \tag{3.9}$$

3.1.5. K- Katlamalı Çapraz Doğrulama

K-Katlamalı Çapraz Doğrulama(K-fold Cross Validation) yöntemi, modelin test edilmesi sırasında en iyi modeli oluşturabilmek için kullanılmaktadır. Öncelikle modelin eğitim sürecinde kullanılacak eğitim kümesi karıştırılır. Ardından eğitim verisi ile hiper parametre değerlerini bulabilmek için algoritma çalışmaya başlattırılır. Bu süreç, seçilen k sayısı kadar tekrarlanarak her seferinde sıradaki alt küme eğitim veri setinden çıkarılarak test kümesi olarak kullanılır. Değerlendirme süreci bittiğinde, bu model tüm veriler için bir performans ölçütü ve sınıflandırma doğruluğu üretir (Wiens vd., 2008). En

iyi hiper parametreler bulunduktan sonra eğitim verisi ile algoritma tekrar çalıştırıldığında model oluşturulur. Oluşturulan modelden elde edilen tahmin test etiketleri ile karşılaştırılarak modelin performansı elde edilir. Şekil 3.3 k katlamalı çapraz doğrulama yöntemini göstermektedir.



Şekil 3.3. Model Seçimi için K kat Çapraz Doğrulama (Raschka, 2018)

Veri Madenciliği çalışmalarında, uygulamada kullanılacak yöntemin başarının karşılaştırılabilmesi için, kullanılan veri seti eğitim ve test olarak ikiye ayrılmaktadır. Bu şekilde eğitim ve test olarak veri setinin rastgele olarak parçalanması yöntemi de farklı bir yöntemdir. Ancak, k-katlamalı çapraz doğrulama yönteminde, eğitim ve test kümeleri kendi aralarında belli bir sırayla değişmektedir. Literatürde en çok tercih edilen k değeri 5 ve 10’ dur.

4. UYGULAMA

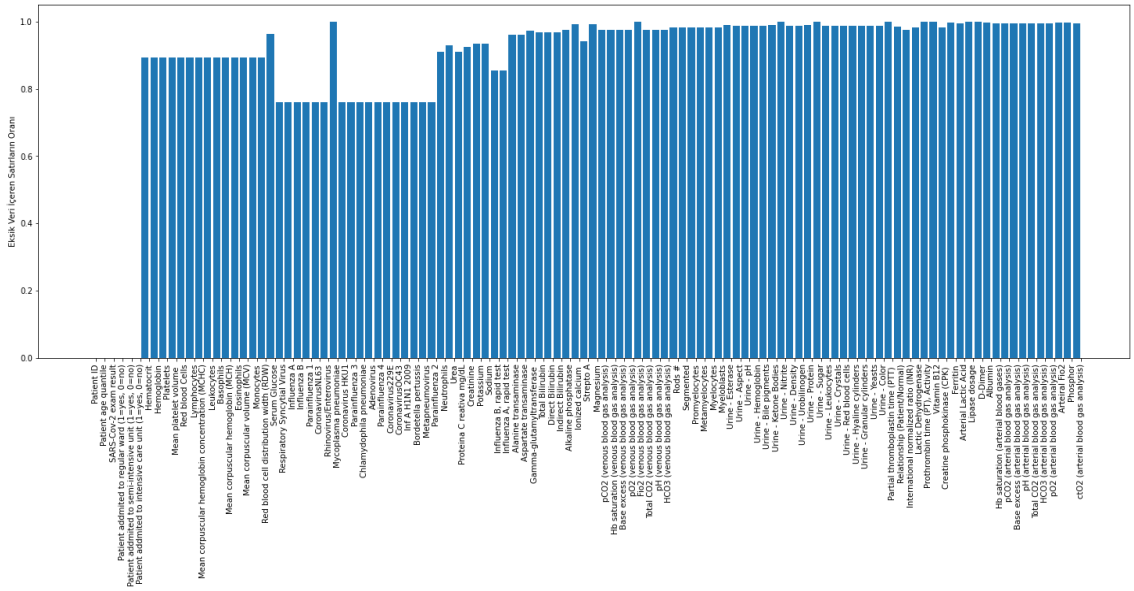
Bu bölümde COVID-19 veri setinin analizinden başlayarak, veri setinin hazırlanması ve modellenmesi için kullanılan yöntemlerin uygulamalarından ve model hazırlandıktan sonra yapılan sınıflandırma sonucunda elde edilen algoritmaların başarılarından bahsedilecektir.

4.1. COVID-19 Veri Setinin Analizi

Verinin analiz edilerek anlaşılması, CRISP-DM veri madenciliği modelinin ilk adımıdır. Bu aşamada, verinin en doğru şekilde anlaşılmaması ya da yeterince analiz edilmemesi oluşturulacak modelin performansını etkilemektedir.

Bu tez çalışmasında, Bölüm 3.1.1 ‘de anlatılan veri seti kullanılmıştır. Bu veri setinde Albert Einstein Hastanesi’ne ait COVID-19 veri setinde 559 COVID-19 pozitif hastası bulunan toplam 5644 hastaya ait 111 nitelik bulunmaktadır.

Veri setinde bulunan her değişken için eksik veri analizi sonuçları Şekil 4.1’ de gösterilmektedir. Şekilde x ekseni veri setinde bulunan özellikleri gösterirken, y ekseni ise veri setinde her özelliğe ait boş verilerin miktarını göstermektedir.



Şekil 4.1. Eksik Veri Analizi Sonuçları

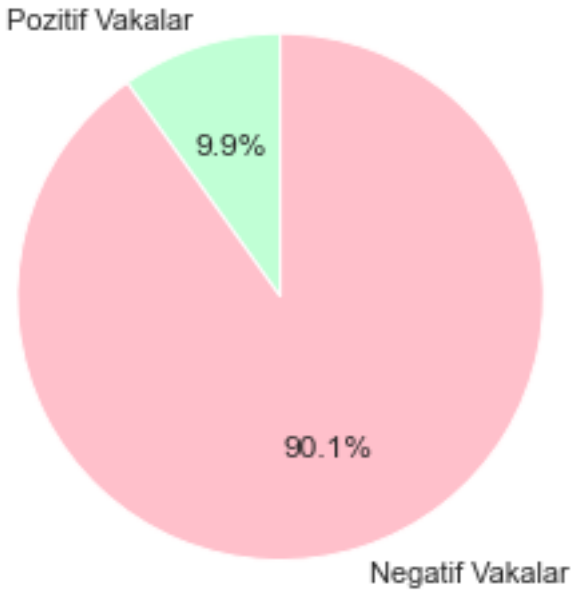
Şekil 4.1’ de de görüldüğü veri setinde çok fazla eksik veri bulunmaktadır. %95 üzerinde boş veriye sahip değişkenler ve boş değer oranları Çizelge 4.1’ de gösterilmektedir.

Çizelge 4.1. Özelliklerin Boş Değer Oranları

Özellikler	**Boş Değer Oranları**
Mycoplasma pneumoniae	1
Urine - Sugar	1
Partial thromboplastin time (PTT)	1
Prothrombin time (PT), Activity	1
D-Dimer	1
Fio2 (venous blood gas analysis)	1
Urine - Nitrite	1
Vitamin B12	0.999
Lipase dosage	0.999
Albumin	0.998
Arteiral Fio2	0.996
Phosphor	0.996
Ferritin	0.996
Arterial Lactic Acid	0.995
Hb saturation (arterial blood gases)	0.995
pCO2 (arterial blood gas analysis)	0.995
Base excess (arterial blood gas analysis)	0.995
pH (arterial blood gas analysis)	0.995
Total CO2 (arterial blood gas analysis)	0.995
HCO3 (arterial blood gas analysis)	0.995
pO2 (arterial blood gas analysis)	0.995
ctO2 (arterial blood gas analysis)	0.995
Magnesium	0.993
Ionized calcium	0.991
Urine - Ketone Bodies	0.990
Urine - Esterase	0.989
Urine - Protein	0.989
Urine - Hyaline cylinders	0.988
Urine - Urobilinogen	0.988
Urine - Granular cylinders	0.988
Urine - Aspect	0.988
Urine - pH	0.988
Urine - Hemoglobin	0.988
Urine - Bile pigments	0.988
Urine - Density	0.988
Urine - Leukocytes	0.988
Urine - Crystals	0.988
Urine - Red blood cells	0.988
Urine - Yeasts	0.988
Urine - Color	0.988
Relationship (Patient/Normal)	0.984
Rods #	0.983
Segmented	0.983
Promyelocytes	0.983

Metamyelocytes	0.983
Myelocytes	0.983
Myeloblasts	0.983
Lactic Dehydrogenase	0.982
Creatine phosphokinase (CPK)	0.982
International normalized ratio (INR)	0.976
pCO2 (venous blood gas analysis)	0.976
Hb saturation (venous blood gas analysis)	0.976
Base excess (venous blood gas analysis)	0.976
pO2 (venous blood gas analysis)	0.976
Total CO2 (venous blood gas analysis)	0.976
pH (venous blood gas analysis)	0.976
HCO3 (venous blood gas analysis)	0.976
Alkaline phosphatase	0.974
Gamma-glutamyltransferase	0.973
Total Bilirubin	0.968
Direct Bilirubin	0.968
Indirect Bilirubin	0.968
Serum Glucose	0.963
Alanine transaminase	0.960
Aspartate transaminase	0.960

Şekil 4.2’ de görüldüğü gibi 5644 hastanın 558 tanesi SARS-CoV-2 ile enfekte olmuşken 5086 tanesinin SARS-CoV-2 test sonucu negatiftir.



Şekil 4.2. Veri Setinin Pozitif ve Negatif Vaka Dağılımı

Veri analizi aşamasında elde edilen bulgular doğrultusunda;

- Veri setinde bulunan ve %95' in üzerinde eksik veri içeren 65 özellik modellemeye doğru etkisi olmayacağından veri setinden çıkarılmıştır. Boş değerlerin fazlalığı sebebiyle geriye kalan nitelik sayısı 46 olmuştur.
- Şekil 4.2' de görüldüğü gibi veri setinde dengesizlik problemi mevcuttur.
- Veri setinde kategorik olarak bulunan niteliklerden, 'positive' ve 'negative' olarak sınıflandırılan nitelikler 'pozitif=1' 'negatif=0' olarak numerik hale getirilmişken, 'not_detected' ve 'detected' olarak sınıflandırılan nitelikler ise 'not_detected=0' ve 'detected=1' olarak 'Label Encoder' yöntemi ile numerik hale getirilmiştir.
- Veri setinde birincil anahtar olarak bulunan 'PatientID' modelde kullanılmayacağı için veri setinden çıkarılmıştır.

4.2. Veriyi Hazırlama

Tez çalışmasının bu aşamasında veri setinin yapılacak çalışmaya uygun hale dönüştürülmesi sağlanmıştır. Şekil 4.3' te verinin hazırlanması süreci gösterilmiştir.

Şekil 4.3. Veri Setinin Oluşturulma Evreleri

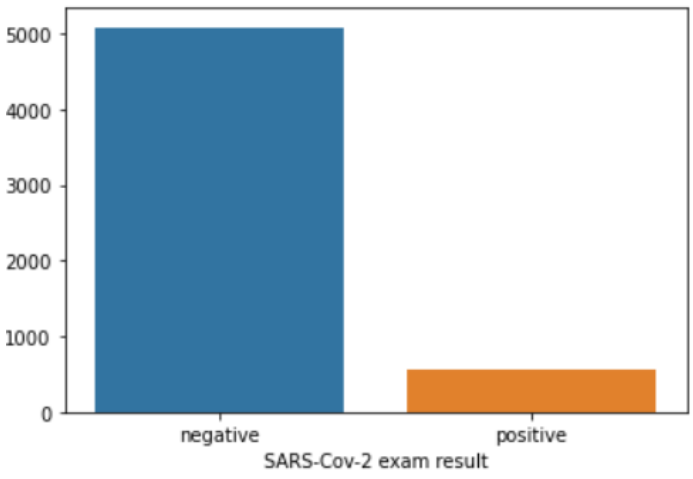
4.2.1. KNN ile Eksik Verilerin Giderilmesi

Şekil 4.1' de görüldüğü gibi neredeyse her niteliğe ait eksik değerlerin bulunması, veri setindeki eksik değerlerin tamamlanması gerektiğini göstermektedir. Bunun sebebi

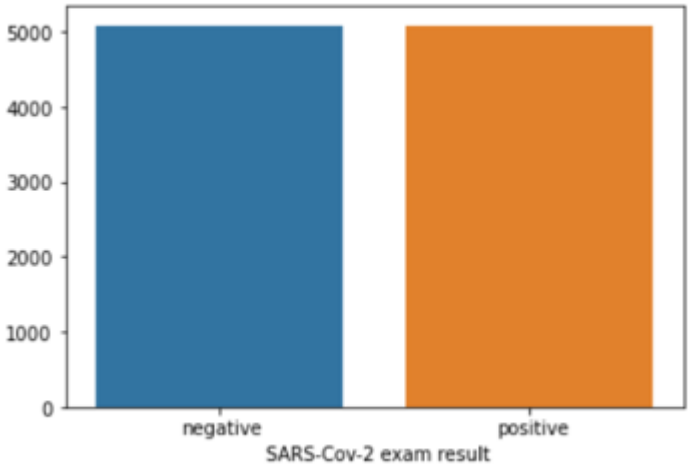
veri setinden eksik değerlerin tamamını sildiğimizde veri setinde toplamda 598 hastaya ait veri kalmaktadır. Bu da veri setinin %10.6' lık kısmını oluşturmaktadır. %89.4 oranında veriyi silmek veri setinde ciddi derecede veri kaybına sebep olacağından ve analiz sonuçlarının etkilenmemesi için, yaygın olarak kullanılan k en yakın komşu (kNN) algoritması ile tamamlanmıştır. K değeri literatürde en çok tercih edilen ve varsayılan değer olan 5 olarak seçilmiş ve kNN yöntemi uygulanmıştır (Soares, 2020).

4.2.2. Veri Dengesizliğinin Giderilmesi

Şekil 4.2' de görüldüğü gibi veri setinin %90.1' lik kısmını negatif vakalar oluştururken, ancak %9.9' luk kısmını pozitif vakalar oluşturmaktadır. Bu da veri setinde sınıf dengesizliği olduğunu göstermektedir. Bu dengesizlik durumu, veri seti üzerinde yapılacak olan çalışmaların sonuçlarını yanlış yönlendirilecek etkilere sebep olabileceğinden bu problemi çözmek için SMOTE algoritması kullanılmıştır. Python' un Imbalanced-Learn kütüphanesinde olan SMOTE fonksiyonu ile örneklem arttırma yapılmıştır. Böylece Şekil 4.4' te de görüldüğü gibi toplamda 5086 negatif vaka ve 5684 toplam hasta varken, Şekil 4.5' te olduğu gibi SMOTE algoritması sayesinde 5086 negatif vaka ve 5086 pozitif vaka ile toplamda 10172 hastaya ait veri olmuştur.



Şekil 4.4. SMOTE Öncesi Veri Dağılımı

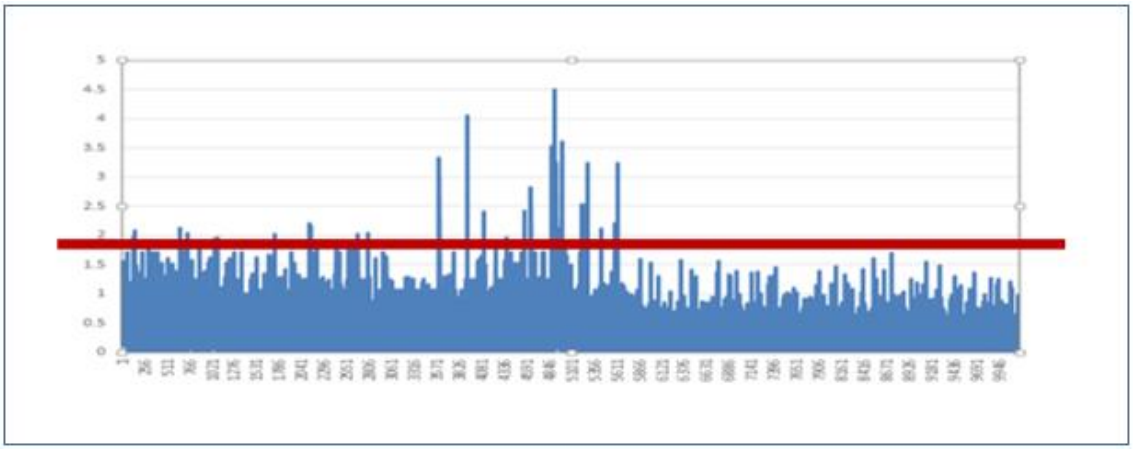


Şekil 4.5. SMOTE Sonrası Veri Dağılımı

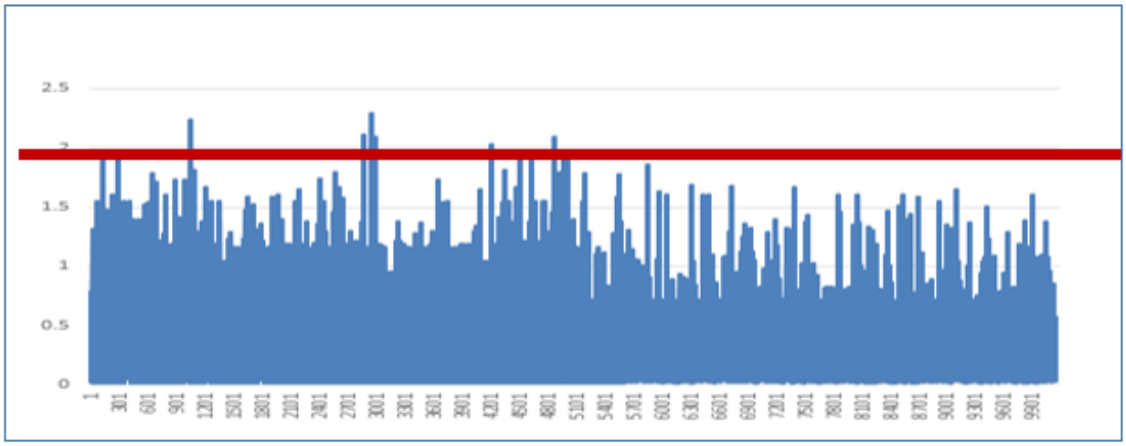
4.2.3. Gürültülü Verileri Elenmesi

Veri seti dengelendikten sonra veri setinde bulunan gürültülü verilerin elenebilmesi için her niteliğe ait dağılım grafikleri oluşturuldu. Oluşturulan grafikler sayesinde niteliklere ait gürültünün görüldüğü yerde eşik değerleri belirlenerek gürültüye sebep olan veriler elendi. Her bir dağılım grafiği için x ekseni özelliğe ait hastayı temsil ederken, y ekseni ise hastaya ait kan değerinin miktarını göstermektedir.

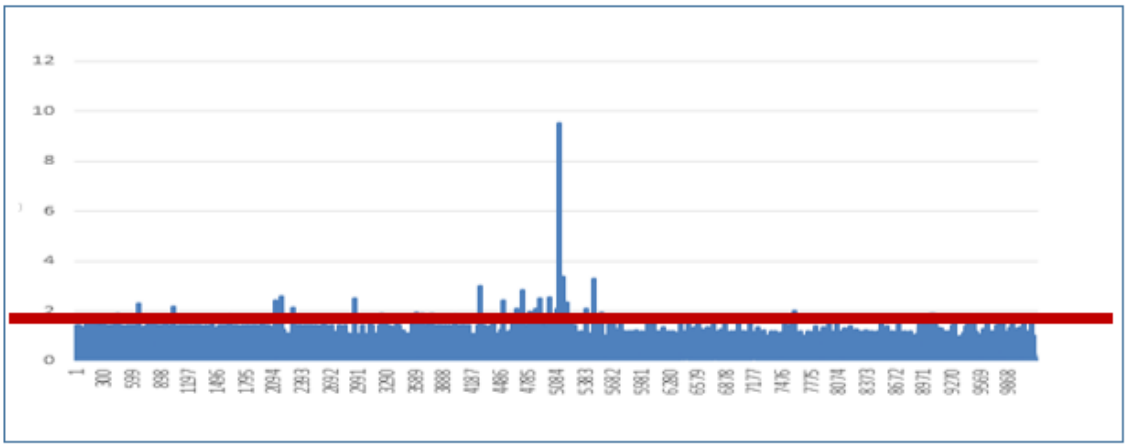
Hematocrit, Hemoglobin ve Platelets kan değerinin SMOTE sonrası dağılım grafiği Şekil 4.6, Şekil 4.7 ve Şekil 4.8' deki gibidir.



Şekil 4.6. Hematocrit kan değerinin SMOTE sonrası dağılım grafiği

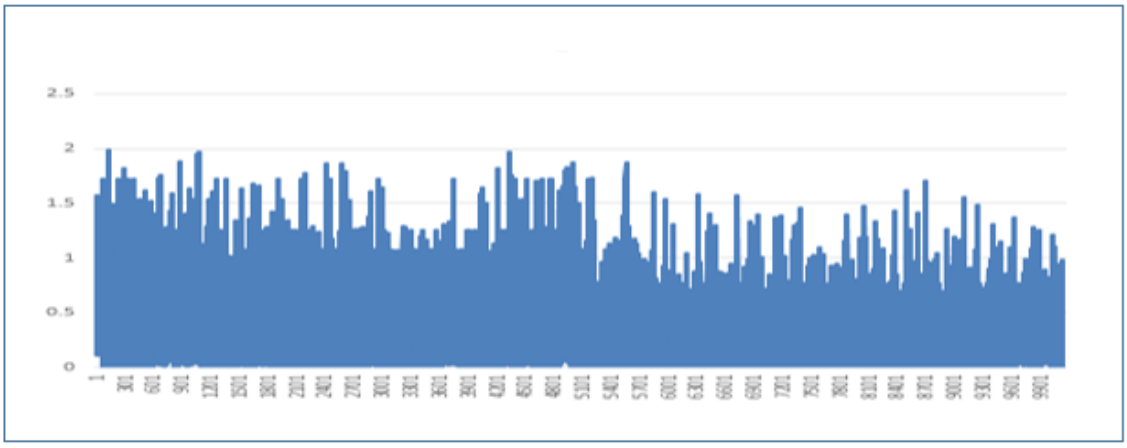


Şekil 4.7. Hemoglobin kan değerinin SMOTE sonrası dağılım grafiği

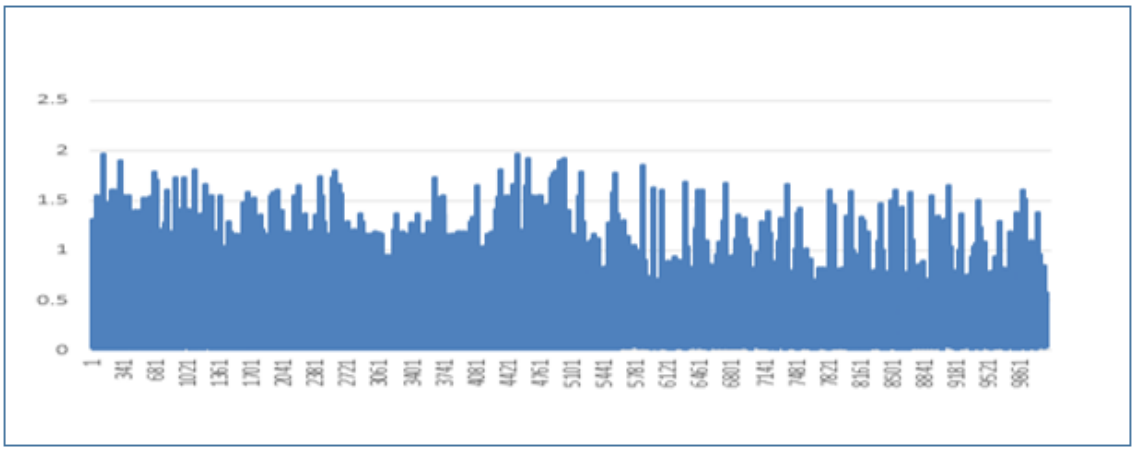


Şekil 4.8. Platelet kan değerinin SMOTE sonrası dağılım grafiği

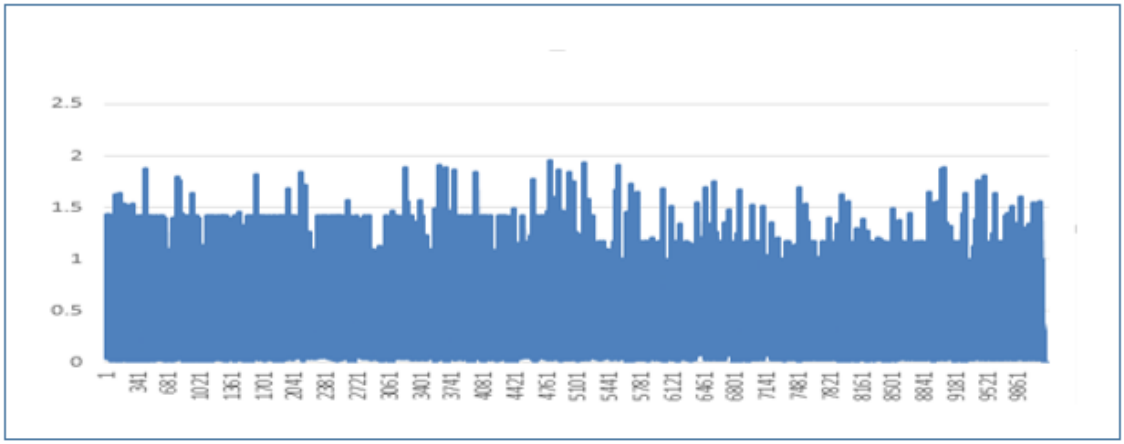
Şekil 4.6, 4.7 ve 4.8' de bulunan kırmızı çizgi eşik değeri çizgisini göstermektedir. Görüldüğü gibi 2 ve üzerinde yer alan Hematocrit, Hemoglobin ve Platelets değerleri normal dağılımı bozduğu için gürültü yaratmaktadır. Bu sebeple 2 ve üzerinde Hematocrit, Hemoglobin ve Plateletes sonucu bulunan hastalar gürültülü veri olduğu için eleme yapılmalıdır. Eleme sonucunda elde edilen dağılım grafikleri Şekil 4.9, Şekil 4.10 ve Şekil 4.11' deki gibidir.



Şekil 4.9. Hematocrit kan değerinin eleme sonrası dağılım grafiği

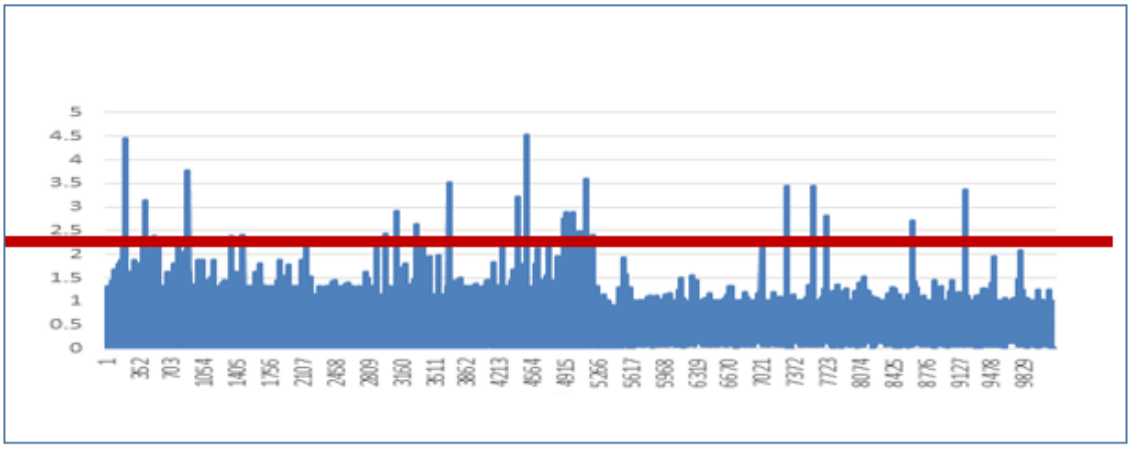


Şekil 4.10. Hemoglobin kan değerinin eleme sonrası dağılım grafiği

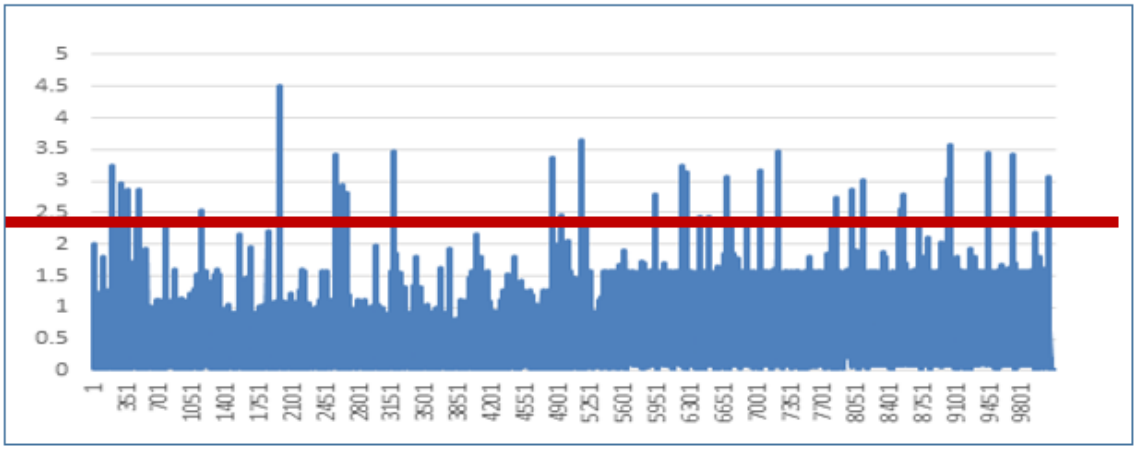


Şekil 4.11. Platelet kan değerinin eleme sonrası dağılım grafiği

Leukocytes ve Monocytes kan değerlerinin SMOTE sonrası dağılım grafikleri Şekil 4.12 ve Şekil 4.13' deki gibidir.

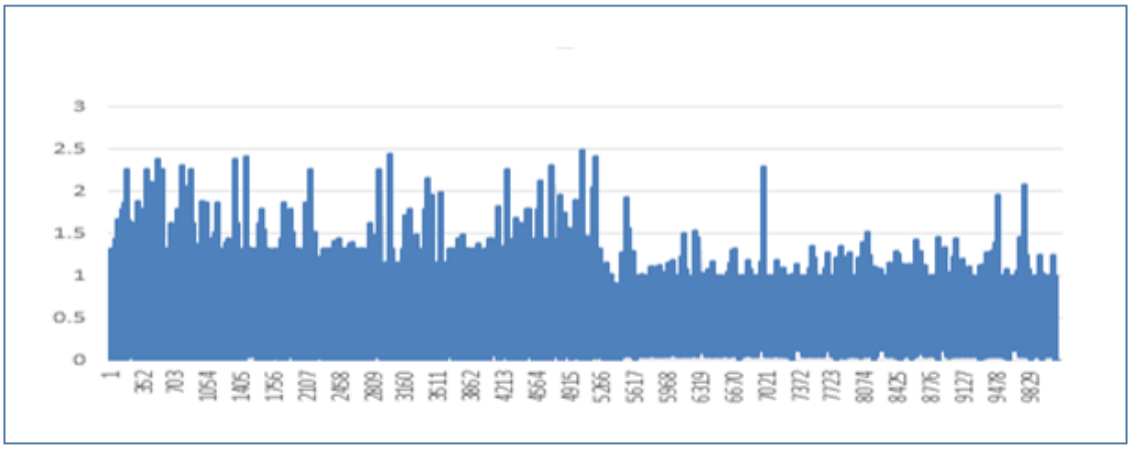


Şekil 4.12. Leukocytes kan değerinin SMOTE sonrası dağılım grafiği

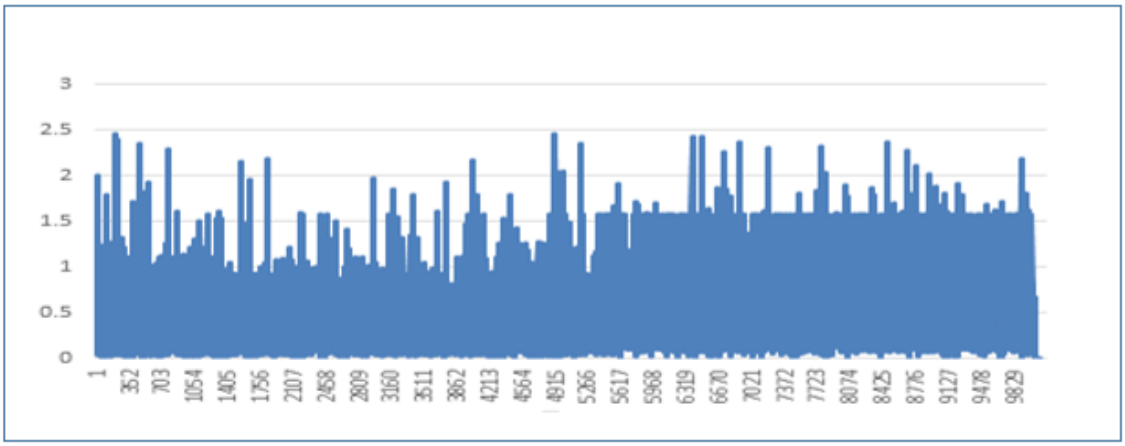


Şekil 4.13. Monocytes kan değerinin SMOTE sonrası dağılım grafiği

Şekil 4.12 ve 4.13' de bulunan kırmızı çizgi eşik değeri çizgisini göstermektedir. Görüldüğü gibi 2.5 ve üzerinde yer alan Leukocytes ve Monocytes değerleri normal dağılımı bozduğu için gürültü yaratmaktadır. Bu sebeple 2.5 ve üzerinde Leukocytes ve Monocytes sonucu bulunan hastalar gürültülü veri olduğu için eleme yapılmalıdır. Eleme sonucunda elde edilen dağılım grafikleri Şekil 4.14 ve Şekil 4.15' deki gibidir.



Şekil 4.14. Leukocytes kan değerinin eleme sonrası dağılım grafiği



Şekil 4.15. Monocytes kan değerinin eleme sonrası dağılım grafiği

Bu grafikler her kan değeri sonucu için tekrar edilmiş olup veri setinde gürültüye sebep olan verilerin tamamı elenmiştir. Eleme sonucunda toplamda 10172 hastaya ait veriden geriye kalan hasta sayısı 9736 olmuştur. Bu da %4.286 hastaya ait verinin silindiğini göstermektedir. Bu hastaların 4997 tanesi pozitif vakaya sahip iken 4739 tanesi ise negatif vakaya sahiptir.

4.2.4. Özellik Çıkarımı

Son olarak elde edilen modelde her niteliğin TBA sonuçları incelenerek sınıflandırmada kullanılacak nitelikler belirlenmiştir.

Her niteliğin COVID-19 sonuçlarına ait TBA sonuçları Çizelge 4.2' deki gibidir.

Çizelge 4.2. Özelliklerin TBA Sonuçları

ÖZELLİKLER	TBA DEĞERİ
Potassium	**0.009**
Proteina C reativa mg/dL	**0.009**
Parainfluenza 1	**0.018**
Inf A H1N1 2009	**0.026**
Influenza B, rapid test	**0.035**
Bordetella pertussis	**0.054**
Sodium	**0.055**
Basophils	**0.06**
Rhinovirus/Enterovirus	**0.074**
Red blood Cells	**0.099**
Influenza B	**0.106**
Lymphocytes	**0.11**
Metapneumovirus	**0.116**
Adenovirus	**0.127**
Influenza A, rapid test	**0.127**
CoronavirusOC43	**0.129**
Creatinine	**0.132**
Parainfluenza 3	**0.134**
Red blood cell distribution width (RDW)	**0.139**
Mean corpuscular hemoglobin concentration (MCHC)	**0.142**
Parainfluenza 4	**0.188**
Chlamydophila pneumoniae	**0.189**
Influenza A	**0.195**
Respiratory Syncytial Virus	0.214
Mean corpuscular volume (MCV)	0.238
Urea	0.243
Strepto A	0.246
Coronavirus229E	0.256
Mean platelet volume	0.259
Hematocrit	0.262
Hemoglobin	0.263
Coronavirus HKU1	0.27
Mean corpuscular hemoglobin (MCH)	0.278
Neutrophils	0.289
Eosinophils	0.323
CoronavirusNL63	0.415
Monocytes	0.515
Leukocytes	0.577
Platelets	0.632

Yapılan testler sonucunda en iyi eşik değeri 0.21 olarak seçilmiş olup (en yüksek değerin 1/3 oranı) , seçilen eşik değeri sonucunda sınıflandırmada kullanılacak özellik sayısı 16 olarak belirlenmiştir**.**

Yapılan işlemler sonucunda sınıflandırma için kullanılacak model 16 niteliğe sahip olup toplamda 9736 hastaya ait veriyi içermektedir.

4.3. Model Oluşturma ve Değerlendirme

Tez çalışmasında oluşturulan modelde kullanılan ve %80 eğitim ve %20 test olarak ayrılan veri seti üzerinde hiper parametre değerlerinin ayarlanması için scikit learn kütüphanesinde bulunan 'GridSearchCV' kullanılmıştır. Algoritmalara uygun parametreleri bulabilmek için GridSearchCV yöntemi kullanılırken her algoritmaya ait hiper parametre değerleri için değer aralıkları oluşturulmuş ve bu oluşturulan değer aralıkları 10 katlamalı çapraz doğrulama ile değerlendirilmiştir. En iyi hiper parametreler elde edilen en yüksek sınıflandırma doğruluğuna sahip modelden alınmıştır. Bu değerlendirme sonucunda elde edilen en iyi modele ait değerlendirme ölçütleri aşağıdaki gibidir.

Destek Vektör Makineleri için yapılan değerlendirme sonucunda 'kernel' (çekirdek türü) parametresi için en iyi parametre 'rbf' iken, 'C' değeri (düzenlilik parametresi) ise 10 olarak belirlenmiştir. Destek Vektör Makineleri algoritmasına ait karmaşıklık matrisi ve değerlendirme ölçütleri Çizelge 4.3 ve Çizelge 4.4' te olduğu gibidir.

Çizelge 4.3. Destek Vektör Makineleri Test Kümesine ait olan Karmaşıklık Matrisi

		TAHMİN EDİLEN SINIF	
		Sağlıklı	**COVID-19**
GERÇEK SINIF	**Sağlıklı**	912	36
	COVID-19	29	971

Destek Vektör Makinelerine ait doğruluk değeri Eşitlik 3.6’ ye göre hesaplanmış olup, 0.967’ dir. Çizelge 4.3’ deki değerler ise Bölüm 3’ de bulunan Çizelge 3.4’ teki karmaşıklık matrisine uygun olarak oluşturulmuştur.

Çizelge 4.4. Destek Vektör Makineleri Başarı Değerlendirme Ölçütleri

	Kesinlik	**Duyarlılık**	**F1-Score**
Sağlıklı	0.97	0.96	0.96
COVID-19	0.96	0.97	0.96

Çizelge 4.4’ te bulunan değerler ise Bölüm 3’ de bulunan Eşitlik 3.7, 3.8 ve 3.9 ile hesaplanmıştır.

Rastgele Orman algoritması için GridSearchCV ile yapılan değerlendirme sonucunda ‘criterion’ (ölçüt değeri) parametresi için en iyi parametre ‘gini’ ,’n_estiamtors’ (ormandaki ağaç sayısı) parametresi için en iyi değer ‘150’ ve ‘max_depth’ (ağacın maksimum derinliği) parametresi için en iyi değer ‘20’ olmuştur. Aynı zamanda, ‘min_sample_leaf’(Bir yaprak düğümünde olması gereken minimum örnek sayısı) için en iyi değer ‘1’ iken, ‘max_features’ (en iyi bölünmeyi ararken göz önünde bulundurulması gereken özelliklerin sayısı) parametresi için belirlenen en iyi parametre ‘auto’ olarak belirlenmiştir. Son olarak ‘min_samples_split’ (bir düğümü bölmek için gereken minimum örnek sayısı) parametresi için en iyi değer ise ‘2’ olarak belirlenmiştir.

Rastgele Orman algoritmasına ait karmaşıklık matrisi ve değerlendirme ölçütleri Çizelge 4.5 ve Çizelge 4.6’ daki gibidir.

Çizelge 4.5. Rastgele Orman Test Kümesine ait olan Karmaşıklık Matrisi

		TAHMİN EDİLEN SINIF	
		Sağlıklı	**COVID-19**
GERÇEK SINIF	**Sağlıklı**	939	9
	COVID-19	3	997

Rastgele Orman algoritmasına ait sınıflandırma doğruluğu Eşitlik 3.6' ya göre hesaplandığında elde edilen sonuç 0.992' dir.

Çizelge 4.6. Rastgele Orman Başarı Değerlendirme Ölçütleri

	Kesinlik	**Duyarlılık**	**F1-Score**
Sağlıklı	1.00	0.99	0.99
COVID-19	0.99	1.00	0.99

Çizelge 4.6' da bulunan duyarlılık değeri Bölüm 3' de bulunan Eşitlik 3.7' ye göre, kesinlik değeri Eşitlik 3.8' ya göre ve F1 Score ise Eşitlik 3.9' e göre hesaplanmıştır.

Naive Bayes algoritması için yapılan değerlendirme sonucunda, algoritmanın sahip olduğu parametreler mevcut olmadığından, yalnızca k değerinin 10 olarak kullanımı ile elde edilen karmaşıklık matrisi ve değerlendirme ölçütleri Çizelge 4.7 ve Çizelge 4.8' deki gibidir.

Çizelge 4.7. Naive Bayes Test Kümesine ait olan Karmaşıklık Matrisi

		TAHMİN EDİLEN SINIF	
		Sağlıklı	**COVID-19**
GERÇEK SINIF	**Sağlıklı**	781	170
	COVID-19	41	956

Naive Bayes algoritmasına ait sınıflandırma doğruluğu Bölüm 3' de bulunan Eşitlik 3.6' ya göre hesaplandığında elde edilen sonuç 0.892' dir.

Çizelge 4.8. Naive Bayes Başarı Değerlendirme Ölçütleri

	Kesinlik	**Duyarlılık**	**F1-Score**
Sağlıklı	0.95	0.82	0.88
COVID-19	0.85	0.96	0.90

Bölüm 3' de hesaplamaları gösterilen kesinlik, duyarlılık ve F1- Score eşitlik değerlerine göre Çizelge 4.8 oluşturulmuştur.

Elde edilen sonuçlar değerlendirildiğinde k fold ve GridSearchCV kullanılarak test kümesinden elde edilen algoritmaların performansı değerlendirildiğinde en iyi sınıflandırma doğruluğuna sahip algoritma %99.2 genel sınıflandırma doğruluğu ile ile Rastgele Orman algoritması olmuştur. Aynı şekilde Rastgele Orman algoritması bireylerin COVID-19 hastalığına sahip olmasını en iyi tespit eden algoritma olmuştur. Bu değerin tespit edilmesinde kullanılan değerlendirme ölçütü 'Duyarlılık' değeridir. Naive Bayes algoritması bu değeri 0.96 olarak bulmuşken, Destek Vektör Makineleri 0.97 olarak bulmuş ve Rastgele Orman ise 1.00 değeri ile en iyi sonucu elde etmiştir.

Veri seti oluşturulma aşamasında kullanılan ön işlem adımlarının sınıflandırma doğruluğu üzerindeki etkisinin gözlemlenmesi için, ön işlem adımı uygulanmadan önceki ve sonraki test kümesine ait sınıflandırma doğruluk değerleri Çizelge 4.9' da verilmiştir.

Çizelge 4.9. Yapılan Ön İşlem Adımlarının Sınıflandırma Doğruluğu Üzerine Etkisi

	Test Verisi Başarı Sonuçları		
	Naïve Bayes	Rastgele Orman	Destek Vektör Makineleri
SMOTE Öncesi (KNN ile Tamamlanmış Veri Seti)	0.765	0.923	0.924
SMOTE Sonrası(Gürültülü Veriler Elenmeden Önceki Veri Seti)	0.726	0.976	0.945
TBA Öncesi (Gürültülü Verilerin Elenmesi Sonundaki Veri Seti)	0.830	0.985	0.951
TBA Sonrası (Eksik Verisi Tamamlanmış, Dengelenmiş, Gürültülü verisi elenmiş ve Özellik Seçimi Gerçekleşmiş Veri Seti)	0.892	0.992	0.967

Çizelge 4.9'da görüldüğü gibi, öncelikle veri setinde %95 üzerinde eksik verisi bulunan özellikler çıkarıldığında elde edilen 5644 hastaya ait veri seti Knn algoritması ile tamamlandığında elde edilen başarı oranları Naive Bayes için 0.765, Rastgele Orman için 0.923 ve Destek Vektör Makineleri için ise 0.924' tür. Bu durumda en yüksek başarıya sahip algoritma Destek Vektör Makineleri olmuştur. Eksik veriler tamamlandıktan sonra dengesizlik problemi için kullanılan SMOTE algoritması sonucu 10172 veri ile yapılan sınıflandırma sonucunda, Naive Bayes algoritması 0.726, Rastgele Orman 0.976 ve Destek Vektör Makineleri ise 0.945 genel başarı elde etmiştir. Bu aşamada en yüksek başarıya sahip algoritma Rastgele Orman algoritması olmuştur ve Naive Bayes algoritmasının başarısı ilk duruma göre daha düşük olmuştur. SMOTE sonrasında elenen gürültülü veri ile veri setinde kalan 9736 hastaya ait veri seti ile yapılan çalışmada Naive Bayes algoritması 0.83, Rastgele Orman algoritması 0.985 ve Destek Vektör Makineleri ise 0.951 genel sınıflandırma doğruluğu elde etmiştir. Son olarak TBA sonrasında oluşturulan veri seti ile gerçekleştirilen sınıflandırma sonucunda Rastgele Orman algoritması 0.992 sınıflandırma doğruluğu ile en yüksek performansı göstermiştir. Çizelge 4.9' da da görüldüğü gibi veri ön işlem adımları algoritmaların sınıflandırma doğruluklarını artırmaktadır.

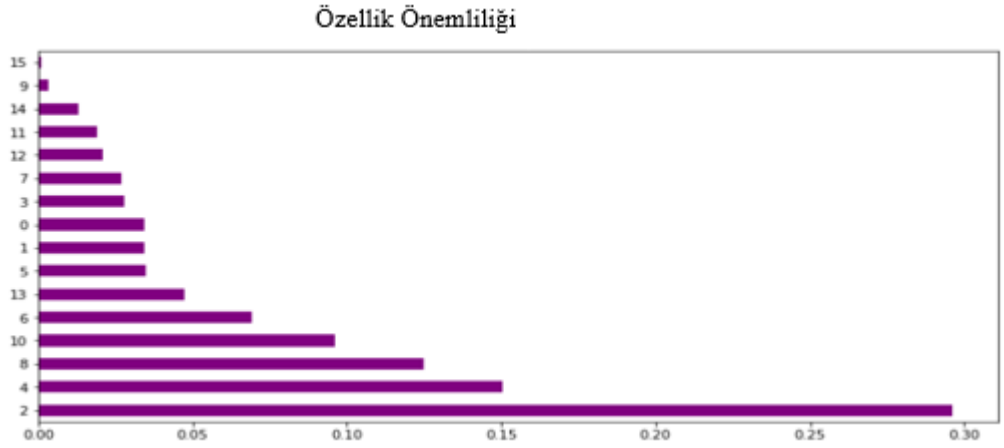
Yukarıda verilen sonuçların tesadüfi olarak elde edilmediğini göstermek adına, 10 katmanlı çapraz doğrulama işlemi, algoritmaların belirlenen hiper parametre değerleri ile 30 kez tekrarlanmıştır. 30x10 katmanlı çapraz doğrulama sonuçlarına ait ortalama, en büyük ve en küçük sınıflandırma doğrulukları Çizelge 4.10' da sunulmuştur.

Çizelge 4.10. 30x10 Katmanlı Çapraz Doğrulama Sonuçlarına Ait Ortalama, En Büyük Ve En Küçük Sınıflandırma Doğrulukları

		Ortalama Sınıflanma Doğruluğu	En Büyük Sınıflandırma Doğruluğu	En Küçük Sınıflandırma Doğruluğu
Rastgele Orman	Eğitim Kümesi	1.000	1.000	1.000
	Doğrulama Kümesi	0.991	0.992	0.990
	Test Kümesi	0.991	0.992	0.989
DVM	Eğitim Kümesi	0.962	0.965	0.959
	Doğrulama Kümesi	0.960	0.963	0.957
	Test Kümesi	0.960	0.967	0.951
NB	Eğitim Kümesi	0.878	0.881	0.873
	Doğrulama Kümesi	0.877	0.880	0.870
	Test Kümesi	0.878	0.892	0.870

Çizelge 4.10 incelendiğinde, Naive Bayes algoritmasına ait sınıflandırma sonucunda eğitim kümesi en yüksek 0.881 sınıflandırma doğruluğu, doğrulama kümesi en yüksek 0.880 sınıflandırma doğruluğu ve test kümesi ise en yüksek 0.892 sınıflandırma doğruluğu göstermiştir. Destek Vektör Makineleri ise eğitim kümesinde en yüksek 0.965, doğrulama kümesinde en yüksek 0.963 sınıflandırma doğruluğu ve test kümesinde ise en yüksek 0.967 sınıflandırma doğruluğu elde etmiştir. En yüksek sınıflandırma doğruluğuna sahip olan Rastgele Orman algoritması eğitim kümesinde %100 başarı gösterirken, doğrulama ve test kümelerinde en yüksek 0.992 sınıflandırma doğruluğu elde etmiştir.

En yüksek başarıya sahip Rastgele Orman algoritması kullanılarak modelde bulunan her özelliğin önemlilik derecesini gösteren Özellik Önemliliği (Feature Importance) tablosu Şekil 4.16' daki gibidir.



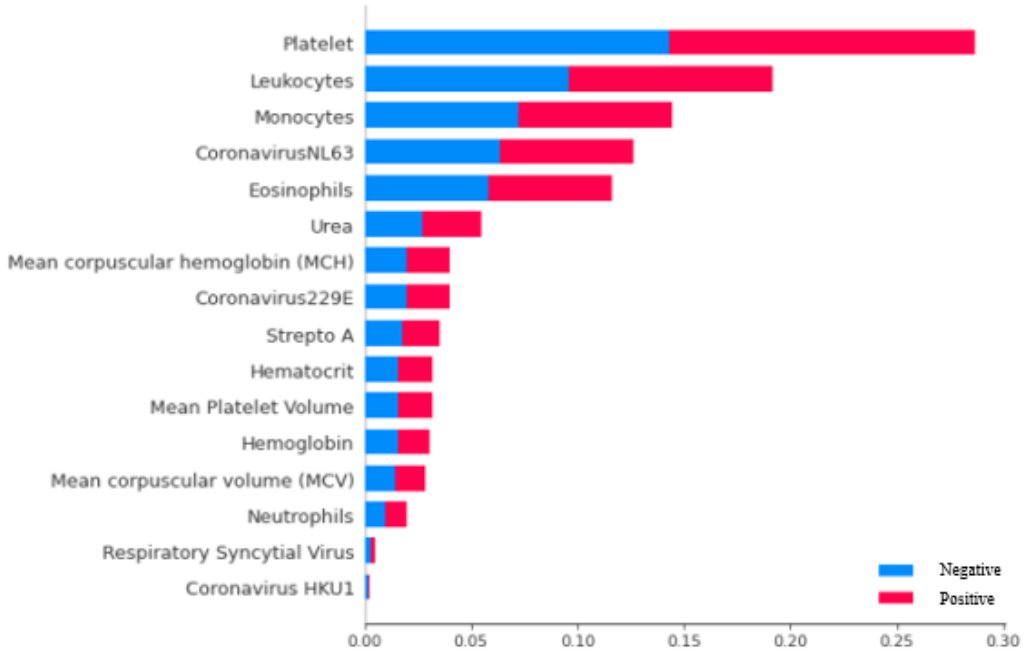
Şekil 4.16. Rastgele Orman Algoritmasına göre Özellik Önemliliği

Özellik önem grafiğine baktığımızda sırasıyla en çoktan en aza doğru önemlilik sırası aşağıdaki gibidir.

1. Platelets
2. Leukocytes
3. Monocytes
4. CoronavirusNL63
5. Eosinophils
6. Urea
7. Mean corpuscular hemoglobin (MCH)
8. Hemoglobin
9. Hematocrit
10. Mean Platelet Volume
11. Mean corpuscular volume (MCV)
12. Neutrophils
13. Coronavirus229E
14. Strepto A
15. Respiratory Syncytial Virus
16. Coronavirus HKU1

Sıralamaya bakıldığında en önemli özellikler Trombositler ve Lökositler olmuştur. Bu önemlilik sıralaması, COVID-19 teşhisinde çok önemli olan niteliklerin

isimlerini vermektedir. Bu özelliklerin değerlerinin COVID-19 için daha yüksek ya da daha düşük olduğunu görebilmek için ise SHAP (Shapely Additive Explanations) yöntemi kullanılmıştır. Bu yöntem, herhangi bir makine öğrenmesi algoritmasının verdiği sonuçları açıklamak için kullanılan oyun teorik bir yaklaşımdır. Bu grafik, özeliklerin hangisinin daha fazla katkıda bulunduğunu ayırt etmek için önemlidir (Turlapati & Prusty, 2020). Şekil 4.17' de görüldüğü gibi COVID-19 hastalığının teşhisi için kullanılan niteliklerin hangisinin baskın veya önemli olduğu görülmektedir.



Şekil 4.17. SHAP Yöntemine Göre Özellik Önemliliği

Şekil 4.16 ve 4.17' ye bakıldığında en önemli 8 özellik aynı iken sonraki özelliklerde bazı değişiklikler olmaktadır. Ancak hastalık teşhisi için en önemli özellikler Trombositler, Lökositler, Monositler, KoronavirüsNL63 ve Eozinofiller olmuştur.

5. SONUÇLAR

5.1. Sonuçlar

2019 yılının son zamanlarında, Çin'in Wuhan kentinde nedeni belli olmayan COVID-19 vakaları ortaya çıkmıştır. Virüsün hızla yayılması sebebiyle pandemi ilan edilmiş ve ölüm oranları gün geçtikçe daha da fazla artmıştır. Bu durumun yaşanmasının en büyük sebebi virüs ile etkileşimin ilk kez yaşanması ve belirtilerinin grip vb. hastalıklarla karıştırılabilir olmasından kaynaklanmaktadır. Bu sebeple, virüsün erken teşhisi için yeni yöntemler aranmaya başlamıştır.

COVID-19 hastalığının erken teşhisi için literatürde makine öğrenmesi kullanılarak birçok çalışma bulunmaktadır. Ancak COVID-19 hastalığı teşhisi için yapılan rutin kan sayımı sonuçları kullanılarak yapılan çalışmalarda hastanelerden elde edilen veri setleri üzerinde COVID-19 teşhisi için önemli olan kan sayımı sonuçları kesin olarak belirlenmediği için, bu tez çalışmasında yapılan değerlendirmeler sonucunda COVID-19 hastalığını etkileyen en önemli kan değerleri belirlenerek bir çalışma yapılmıştır.

Bu tez çalışmasında kullanılan veri seti, Brezilya'da bulunan Albert Einstein Hastanesi'ni ziyaret eden 5644 bireye ait rutin kan sayımı sonuçları ve COVID-19 test sonuçlarını içermektedir. Hastalığın hızla yayılması ve gelen bireylerin her biri için aynı testlerin yapılmaması sebebi ile veri setinde çok miktarda eksik veriler bulunmaktadır. Aynı zamanda COVID-19 pozitif ve negatif hasta sayılarının miktarı da aynı olmadığından veri setinde dengesizlik problemi bulunmaktadır. Bu sebeple bu tez çalışmasında veri setinin her bir problemi tek tek ele alınarak bir çalışma gerçekleştirilmiştir.

Literatürde, Albert Einstein Hastanesi'ne ait veriler kullanılarak yapılan çalışmalar Bölüm 2 'de bahsedilmiştir (AlJame et al., 2020; Alves et al., 2021; Banerjee et al., 2020; Batista et al., 2020; de Freitas Barbosa et al., 2020; Soares, 2020). Yapılan incelemeler sonucunda elde edilen karşılaştırma sonuçları Çizelge 5.1' deki gibidir.

Çizelge 5.1. Literatürdeki Modellerin Karşılaştırılması

Referans	Veri Seti	Özellik Sayısı	Sınıflandırma Algoritması	Sınıflandırma Doğruluğu
(AlJame vd., 2020)	559 doğrulanmış hasta ile 5644 kan örneği	18	Extra Ağaç, RF, LR, XGBoost **(Tek bir model oluşturulmuş)**	%99,88
(Alves vd., 2021)	84 doğrulanmış hasta ile 608 kan örneği	20	DTX, **RF**	%88
(Banerjee vd., 2020)	81 doğrulanmış hasta ile 598 kan örneği	14	RF, YSA, LR, GLMNET	%81-%87
(Batista vd., 2020)	102 doğrulanmış hasta ile 235 kan örneği	15	NN, RF, GBT, LR, **DVM**	%85
(de Freitas Barbosa vd., 2020)	559 doğrulanmış hasta ile 5644 kan örneği	24	DVM, **Bayes Ağları**, Karar Ağaçları	%95,15
(Soares, 2020)	81 doğrulanmış hasta ile 599 kan örneği	16	DVM, SmoteBoost **(Tek bir model oluşturulmuş)**	%86,78
Tez Çalışması	559 doğrulanmış hasta ile 5644 kan örneği	16	DVM, **RF**, Naive Bayes	%99,2

- Kalın olanlar en yüksek başarıya sahip algoritmaları göstermektedir.

Önerilen model, diğer araştırmalarla karşılaştırıldığında daha iyi performans ortaya koymuştur (Alves vd., 2021; Banerjee vd., 2020; Batista vd., 2020; de Freitas Barbosa vd., 2020; Soares, 2020). Alves et al., (2021), öncelikle veri setinde bulunan özelliklerin %95 ve üzerinde boş verisi bulunan özellikleri veri setinden çıkarmıştır. Ardından veri setinde bulunan kan dışı değerler de veri setinden çıkarılarak, toplamda 20 nitelik ve 84 doğrulanmış hastaya ait 608 kan örneği ile veri ön işlemesi gerçekleştirilmiştir. Veri setinde bulunan dengesizlik problemi sebebi ile SMOTE ile dengelenen veri seti %80 eğitim ve %20 test verisi olarak ayrıldığında en yüksek başarı Rastgele Orman algoritmasına ait olup genel sınıflandırma doğruluğu %88 olmuştur. Bu tez çalışması ile yapılan karşılaştırma sonucunda veri setinin sahip olduğu eksik veri problemine çözüm bulunamamış olup aynı zamanda herhangi bir özellik seçimi algoritması da kullanılmamıştır. Aynı zamanda Rastgele Orman algoritmalarının başarıları karşılaştırıldığında tez çalışması daha yüksek sınıflandırma doğruluğu elde etmiştir. Banerjee et al., (2020), Brezilya'da bulunan Albert Eintein Hastanesi'ne ait veri setinden 598 kan örneğini kullanarak istatitiksel analizini gerçekleştirmiştir. Geriye kalan 5046 veri eksik değerler içerdiği için veri setinden çıkarılmıştır. 14 özellik kullanılarak yapılan çalışmada 10 kat çapraz doğrulama kullanılmış ve veri setinde algoritmaların

genel başarıları %81 ve %87 arasında değişmiştir. Aynı zamanda yapılan çalışmada RandomizedSearchCV ve GLMNET ile özellik önemliliği uygulanmıştır. Elde edilen sonuçlar değerlendirildiğinde en önemli iki özellik Eozinofiller ve Lökositler olmuştur. Banerjee et al., (2020), yapmış oldukları çalışmada veri setinde bulunan eksik veri problemine ve dengesizlik problemine çözüm bulmadan, veri setinde bulunan tüm eksik verileri çıkartarak yaptıkları çalışmada, bu tez çalışmasından daha düşük sınıflandırma doğruluğu elde etmişlerdir. Soares, (2020), Albert Einstein Hastanesi'ne ait toplamda 5644 hastası bulunan veri setinde 16 yaygın kan özelliklerinin olduğu en az eksik değere sahip 81' i doğrulanmış 599 hastaya ait veriyi kullanmıştır. Veri setinde bulunan eksik değerler knn algoritması ile tamamlanmıştır. Knn algoritması için seçilen en yakın komşu sayısı ise 5 olarak belirlenmiştir. Veri setinde bulunan dengesizlik SMOTEBoost ile giderilerek veri seti oluşturulmuştur. Veri seti %90 eğitim ve %10 test verisi olarak ayrılarak eğitim süreci 100 kez tekrarlanmıştır. Elde edilen sonuçlar ışığında geliştirilen model %86,78 genel sınıflandırma doğruluğu elde etmiştir. Soares, (2020), veri setinin bütününde bulunan eksik veri problemine çözüm bulmak yerine belirlemiş olduğu 16 niteliğe ait veri setinde kalan eksik verileri Knn algoritması ile tamamlamıştır. Veri setinde bulunan dengesizlik problemi için tez çalışması özelinde kullanılan SMOTE algoritması yerine SMOTEBoost kullanmış olsa bile elde edilen sonuçlar ışığında, tez çalışmasından daha az sınıflandırma doğruluğu elde etmiştir. de Freitas Barbosa et al., (2020), Albert Einstein Hastanesin' den alınan 5644 hastaya ait 111 nitelikten oluşan veriyi kullanarak, öncelikle Parçacık Sürü Optimizasyonu ile 111 nitelikten 24 tanesini veri setinde kullanmak için belirlemiştir. Veri setinde bulunan eksik değerler niteliklere ait ortalama değerler ile tamamlanmış ve veri dengesizliği için ise SMOTE yöntemi kullanılmıştır. Oluşturulan veri seti Destek Vektör Makineleri, Bayes Ağları ve Karar Ağaçları ile sınıflandırıldığında en yüksek başarıya sahip algoritma %95,15 ile Bayes Ağları olmuştur. de Freitas Barbosa et al., (2020), diğer çalışmaların aksine özellik seçimi için Parçacık Sürü Optimizasyonu algoritması sonucunda elde ettiği özellikleri kullanmıştır. Ancak veri setinde bulunan eksik değerleri değerlerin ortalaması ile tamamlayarak gerçekleştirmesi ve devamında SMOTE ile veri dengesizliğine çözüm bulması ile oluşturduğu veri setinden bu tez çalışmasında kullanılan veri seti ile yapılan sınıflandırmadan daha az başarı elde etmiştir. Batista et al.,(2020), Brezilya'daki Albert Einstein Hastanesi'ne ait veri setinden 102' si doğrulanmış toplamda 235 hastaya ait veri setini kullanmıştır. 15 nitelik kullanarak ve %70 eğitim %30 test verisi olarak rastgele ayırdıkları veri seti üzerinden yapılan çalışmada Destek Vektör Makineleri %85 ile en

yüksek başarıyı elde etmiştir. Batista et al.,(2020), Albert Einstein Hastanesi'ne ait veriyi ciddi anlamda küçülterek kullanması ve herhangi bir ön işlem gerçekleştirmeden yaptığı sınıflandırma sonucunda bu tez çalışmasından daha düşük başarı elde etmiştir.
AlJame et al., (2020), geliştirdikleri ERLX modeli ile %99,88 genel doğruluk oranı ile tez çalışmasında gerçekleştirilen modelden daha yüksek genel doğruluk oranı elde etmiştir. Modeller karşılaştırıldığında ERLX modeli başlangıçta manuel olarak seçtiği 18 niteliği kullanarak modeli oluşturmaya başlamıştır. 18 niteliğe sahip tüm hastaları veri setine dâhil edebilmek amacı ile boş verileri KNN algoritması ile doldurmuş ve en yakın komşu sayısı olarak 7 belirlemiştir. Ardından aykırı verileri eleyebilmek için iForest algoritmasını kullanmıştır. Veri setinin sahip olduğu dengesizlik problemi için SMOTE algoritmasını kullanarak veri setini dengeleyerek sınıflandırma için ise %80 eğitim ve %20 test verisi olarak ayırdıkları veri setini kullanarak, Ekstra Ağaç, Rastgele Orman ve Lojistik Regresyon algoritmaları ile ilk seviyedeki sınıflandırmayı yapıp, ikinci seviyede ise XGBoost algoritması ile performansı artırmak istemiştir. Bu modelin sınıflandırma doğruluğu %99,88 olarak belirlenmiştir. ERLX modelinde sınıflandırma için kullanılan 18 özellik herhangi bir özellik seçimi veya özellik çıkarımı algoritmasına bağlı olmaksızın el ile seçilmiştir.

Bu tez çalışmasında, öncelikle veri setinde %95 üzerinde boş olan özellikler veri setinin başarısını olumsuz yönde etkileyeceğinden çıkarılarak veri setinde geri kalan 46 nitelik kullanılarak model oluşturulmaya başlanmıştır. Ardından 46 niteliğe ait boş veriler KNN algoritması ile doldurulmuş ve en yakın komşu sayısı 5 olarak seçilmiştir. Ancak, ERLX modeli veri doldurma işleminden sonra aykırı verileri eleme işlemi gerçekleştirmişken veri setindeki dengesizlik sebebiyle pozitif hastaların veri setindeki miktarı %9.9 oranında olduğu için eleme sırasında orijinal verideki pozitif hastaların elenmesi çok yüksek bir ihtimaldir. Bu sebeple bu tez çalışması gürültülü verileri elemeden önce SMOTE ile veri setini dengelemiş ve bu veri seti üzerinden eleme işlemini gerçekleştirmiştir. Son olarak ise, ERLX modelinin aksine özellik seçimini manuel olarak değil Temel Bileşen Analizi algoritması ile gerçekleştirerek sınıflandırmada 16 nitelik kullanmıştır.

Tez çalışmasında oluşturulan modelde sınıflandırma işlemi için kullanılan algoritmalar Destek Vektör Makineleri, Rastgele Orman ve Naive Bayes algoritmaları olup, Rastgele Orman algoritması %99,2 başarı ile en yüksek başarıyı elde etmiştir. Bu sebeple geliştirilen model ERLX modelinden % 0.68 daha az oranda sınıflandırma

doğruluğu elde etmiş olsa bile modelin geliştirilmesinde kullanılan her adım daha sağlamdır ve COVID-19' un erken ve hızlı teşhisi için kullanılabilir.

5.2. Öneriler

Bu tez çalışmasında elde edilen sonuçlar değerlendirildiğinde COVID-19 hastalığı gibi salgın oluşturabilecek durumlarda kullanılmak üzere hastalığın hızlı teşhisi için makine öğrenmesi algoritmalarından yararlanılabilir.

Geliştirilen sistem değişimlere adapte edilerek, COVID hastalığının varyantlarında kullanılabilecek hale getirilebilir.

Gelecek çalışmalarda, uzman bir doktor eşliğinde hastalardan elde edilen kan değerleri üzerinden hastalık teşhisini yapmaya yönelik doktorlara karar desteği sağlayacak bir karar destek sistemi oluşturulabilir.

7. KAYNAKLAR

AlJame, M., Ahmad, I., Imtiaz, A., & Mohammed, A. (2020). Ensemble learning model for diagnosing COVID-19 from routine blood tests. *Informatics in Medicine Unlocked*, *21*, 100449. https://doi.org/10.1016/j.imu.2020.100449

Alves, M. A., Zanon de Castro, G., Soares Oliveira, B. A., Ferreira, L. A., Ramírez, J. A., Silva, R., & Guimarães, F. G. (2021). Explaining Machine Learning based Diagnosis of COVID-19 from Routine Blood Tests with Decision Trees and Criteria Graphs. *Computers in Biology and Medicine*, *132*, 104335. https://doi.org/10.1016/j.compbiomed.2021.104335

Assaf, D., Gutman, Y., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., Shilo, N., Epstein, A., Mor-Cohen, R., Biber, A., Rahav, G., Levy, I., & Tirosh, A. (2020). Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Internal and Emergency Medicine*, *15*(8), 1435–1443. https://doi.org/10.1007/s11739-020-02475-0

Banerjee, A., Ray, S., Vorselaars, B., Kitson, J., Mamalakis, M., Weeks, S., Baker, M., & Mackenzie, L. S. (2020). Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. *International Immunopharmacology*, *86*, 106705. https://doi.org/10.1016/j.intimp.2020.106705

Bao, F. S., He, Y., Liu, J., Chen, Y., Li, Q., Zhang, C. R., Han, L., Zhu, B., Ge, Y., Chen, S., Xu, M., & Ouyang, L. (2020). Triaging moderate COVID-19 and other viral pneumonias from routine blood tests. *arXiv*. http://arxiv.org/abs/2005.06546

Batista, A. F. de M., Miraglia, J. L., Donato, T. H. R., & Chiavegatto Filho, A. D. P. (2020). COVID-19 diagnosis prediction in emergency care patients: A machine learning approach. Içinde *medRxiv* (s. 2020.04.04.20052092). medRxiv. https://doi.org/10.1101/2020.04.04.20052092

Bhandari, Shaktawat, A. S., Tak, A., Patel, B., Shukla, J., Singhal, S., Gupta, K., Gupta, J., Kakkar, S., & Dube, A. (2020). Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. *Ibnosina Journal of Medicine and Biomedical Sciences*, *12*(2), 123. https://doi.org/10.4103/IJMBS.IJMBS_58_20

Boser, B. E., Laboratories, T. B., Guyon, I. M., Laboratories, T. B., & Vapnik, V. N. (1992). *SVM-A training algorithm for optimal margin classifiers.pdf*.

Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *Journal of Medical Systems*, *44*(8), 1–12. https://doi.org/10.1007/s10916-020-01597-4

Bulut, F. (2016). Sınıflandırıcı Topluluklarının Dengesiz Veri Kümeleri Üzerindeki Performans Analizleri Faruk BULUT. *Bilişim Teknolojileri Dergisi*, *9*(2), 153. https://doi.org/10.17671/btd.81137

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., & Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, *395*(10223), 507–513. https://doi.org/10.1016/S0140-6736(20)30211-7

Choudhury, A., Kosorok, M. R., Bach, F., Blei, D., & Scholkopf, B. (2020). *Missing Data Imputation for Classification Problems*.

Culp, W. C. (2020a). Coronavirus Disease 2019. *A & A Practice*, *14*(6), e01218. https://doi.org/10.1213/xaa.0000000000001218

Culp, W. C. (2020b). Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020; published online Feb 3. DOI:10.1038/s41586-020-2008-3. *A & A Practice*, *14*(6), e01218. https://doi.org/10.1213/xaa.0000000000001218

de Freitas Barbosa, V. A., Gomes, J. C., de Santana, M. A., de Almeida Albuquerque, J. E., de Souza, R. G., de Souza, R. E., & dos Santos, W. P. (2020). Heg.IA: An intelligent system to support diagnosis of Covid-19 based on blood tests. Içinde *medRxiv* (s. 2020.05.14.20102533). medRxiv. https://doi.org/10.1101/2020.05.14.20102533

Demircioğlu, M. (y.y.). *COVID-19 SALGINI İLE MÜCADELEDE KÜMELEME ANALİZİ İLE ÜLKELERİN SINIFLANDIRILMASI*.

Diagnosis of COVID-19 and its clinical spectrum | Kaggle. (y.y.). Tarihinde 31 Ocak 2021, adresinden erişildi https://www.kaggle.com/einsteindata4u/covid19

Döhla, M., Boesecke, C., Schulte, B., Diegmann, C., Sib, E., Richter, E., Eschbach-Bludau, M., Aldabbagh, S., Marx, B., Eis-Hübinger, A. M., Schmithausen, R. M., &

Streeck, H. (2020). Rapid point-of-care testing for SARS-CoV-2 in a community screening setting shows low sensitivity. *Public Health*, *182*, 170–172. https://doi.org/10.1016/j.puhe.2020.04.009

Fan, B. E., Chong, V. C. L., Chan, S. S. W., Lim, G. H., Lim, K. G. E., Tan, G. B., Mucheli, S. S., Kuperan, P., & Ong, K. H. (2020). Hematologic parameters in patients with COVID-19 infection. *American Journal of Hematology*, *95*(6), E131–E134. https://doi.org/10.1002/ajh.25774

Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., & Ji, W. (2020). Sensitivity of chest CT for COVID-19: Comparison to RT-PCR. Içinde *Radiology* (C. 296, Sayı 2, ss. E115–E117). Radiological Society of North America Inc. https://doi.org/10.1148/radiol.2020200432

Farnaaz, N., & Jabbar, M. A. (2016). Random Forest Modeling for Network Intrusion Detection System. *Procedia Computer Science*, *89*, 213–217. https://doi.org/10.1016/j.procs.2016.06.047

Feng, C., Huang, Z., Wang, L., Chen, X., Zhai, Y., Zhu, F., Chen, H., Wang, Y., Su, X., Huang, S., Tian, L., Zhu, W., Sun, W., Zhang, L., Han, Q., Zhang, J., Pan, F., Chen, L., Zhu, Z., … Li, T. (2020). A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in Fever Clinics. Içinde *medRxiv* (s. 2020.03.19.20039099). medRxiv. https://doi.org/10.1101/2020.03.19.20039099

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Interim Guidelines for Clinical Specimens for COVID-19 | CDC. (y.y.). Tarihinde 15 Nisan 2021, adresinden erişildi https://www.cdc.gov/coronavirus/2019-nCoV/lab/guidelines-clinical-specimens.html

Jin, Y. H., Cai, L., Cheng, Z. S., Cheng, H., Deng, T., Fan, Y. P., Fang, C., Huang, D., Huang, L. Q., Huang, Q., Han, Y., Hu, B., Hu, F., Li, B. H., Li, Y. R., Liang, K., Lin, L. K., Luo, L. S., Ma, J., … Wang, X. H. (2020). A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version). Içinde *Military Medical Research* (C. 7, Sayı 1, s. 4). BioMed Central Ltd. https://doi.org/10.1186/s40779-020-0233-6

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*.

Khorraminezhad, L., Leclercq, M., Droit, A., Bilodeau, J. F., & Rudkowska, I. (2020). Statistical and machine-learning analyses in nutritional genomics studies. Içinde *Nutrients* (C. 12, Sayı 10, ss. 1–19). MDPI AG. https://doi.org/10.3390/nu12103140

Koçoğlu, F. Ö. (2017). *Müşteri Kayıp Analizi Probleminin Çözümünde Analitik Yaklaşımlar*. İstanbul Üniversitesi.

Korkem, E. (2013). *MİKROARRAY GEN EKSPRESYON VERİ SETLERİNDE RANDOM FOREST VE NAIVE BAYES SINIFLAMA YÖNTEMLERİ YAKLAŞIMI*. Hacettepe Üniversitesi.

Kukar, M., Gunčar, G., Vovko, T., Podnar, S., Černelč, P., Brvar, M., Zalaznik, M., Notar, M., Moškon, S., & Notar, M. (2020). COVID-19 diagnosis by routine blood tests using machine learning. *arXiv*. http://arxiv.org/abs/2006.03476

Langer, T., Favarato, M., Giudici, R., Bassi, G., Garberi, R., Villa, F., Gay, H., Zeduri, A., Bragagnolo, S., Molteni, A., Beretta, A., Corradin, M., Moreno, M., Vismara, C., Perno, C. F., Buscema, M., Grossi, E., & Fumagalli, R. (2020). *Use of Machine Learning to Rapidly Predict Positivity to Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2) Using Basic Clinical Data*. https://doi.org/10.21203/rs.3.rs-38576/v1

Li, D., Wang, D., Dong, J., Wang, N., Huang, H., Xu, H., & Xia, C. (2020). False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: Role of deep-learning-based ct diagnosis and insights from two cases. *Korean Journal of Radiology*, *21*(4), 505–508. https://doi.org/10.3348/kjr.2020.0146

Liu, J., Li, S., Liu, J., Liang, B., Wang, X., Wang, H., Li, W., Tong, Q., Yi, J., Zhao, L., Xiong, L., Guo, C., Tian, J., Luo, J., Yao, J., Pang, R., Shen, H., Peng, C., Liu, T., … Zheng, X. (2020). Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. *EBioMedicine*, *55*. https://doi.org/10.1016/j.ebiom.2020.102763

Martínez Torres, J., Iglesias Comesaña, C., & García-Nieto, P. J. (2019). Review: machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, *10*(10), 2823–2836. https://doi.org/10.1007/s13042-018-00906-1

Meng, Z., Wang, M., Song, H., Guo, S., Zhou, Y., Li, W., Zhou, Y., Li, M., Song, X., Zhou, Y., Li, Q., Lu, X., & Ying, B. (2020). Development and utilization of an

intelligent application for aiding COVID-19 diagnosis. Içinde *medRxiv* (s. 2020.03.18.20035816). medRxiv. https://doi.org/10.1101/2020.03.18.20035816

ÖZLÜER BAŞER, B., YANGIN, M., & SARIDAŞ, E. S. (2021). Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*. https://doi.org/10.19113/sdufenbed.842460

Raschka, S. (2018). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*.

Savaş, S., Topaloğlu, N., & Yilmaz, M. (y.y.). *VER MADENCLĞĞ VE TÜRKYE'DEK UYGULAMA ÖRNEKLER*.

Savaş, S., Topaloğlu, N., & Yilmaz, M. (2012). *VERİ MADENCİLİĞİ VE TÜRKİYE'DEKİ UYGULAMA ÖRNEKLERİ*.

Schwab, P., Schütte, A. D., Dietz, B., & Bauer, S. (2020). Clinical Predictive Models for COVID-19: Systematic Study. *Journal of Medical Internet Research*, *22*(10). http://arxiv.org/abs/2005.08302

Sertkaya, C., & Yurtay, N. (2015). *Artificial immune system based wastewater parameter estimation*. https://doi.org/10.3906/elk-1503-206

Sever, H., & Oğuz, B. (2002). *Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım:Kısım I: Eşleştirme Sorguları ve Algoritmalar*.

Soares, F. (2020). A novel specific artificial intelligence-based method to identify COVID-19 cases using simple blood exams. Içinde *medRxiv* (s. 2020.04.10.20061036). medRxiv. https://doi.org/10.1101/2020.04.10.20061036

Soltan, A. A. S., Kouchaki, S., Zhu, T., Kiyasseh, D., Taylor, T., Hussain, Z. B., Peto, T., Brent, A. J., Eyre, D. W., & Clifton, D. (2020). Artificial intelligence driven assessment of routinely collected healthcare data is an effective screening test for COVID-19 in patients presenting to hospital. Içinde *medRxiv* (s. 2020.07.07.20148361). medRxiv. https://doi.org/10.1101/2020.07.07.20148361

Sun, N. N., Yang, Y., Tang, L. L., Dai, Y. N., Gao, H. N., Pan, H. Y., & Ju, B. (2020). A prediction model based on machine learning for diagnosing the early COVID-19 patients. Içinde *medRxiv* (s. 2020.06.03.20120881). medRxiv. https://doi.org/10.1101/2020.06.03.20120881

Tan, L., Wang, Q., Zhang, D., Ding, J., Huang, Q., Tang, Y. Q., Wang, Q., & Miao, H. (2020). Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. Içinde *Signal Transduction and Targeted Therapy* (C. 5, Sayı 1).

Springer Nature. https://doi.org/10.1038/s41392-020-0148-4

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Turlapati, V. P. K., & Prusty, M. R. (2020). Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. *Intelligence-Based Medicine*, *3–4*, 100023. https://doi.org/10.1016/j.ibmed.2020.100023

Tüzüntürk, S. (2010). *VERİ MADENCİLİĞİ VE İSTATİSTİK*.

Wiens, T. S., Dale, B. C., Boyce, M. S., & Kershaw, G. P. (2008). Three way k-fold cross-validation of resource selection functions. *Ecological Modelling*, *212*(3–4), 244–255. https://doi.org/10.1016/j.ecolmodel.2007.10.005

World Health Organization. (2020). Co V I D - 19 Strategy Up Date. *Covid-19 Strategy Update*, *3*(April), 18. https://www.who.int/docs/default-source/coronaviruse/covid-strategy-update-14april2020.pdf?sfvrsn=29da3ba0_19

Wu, G., Zhou, S., Wang, Y., & Li, X. (2020). *Machine learning: a predication model of outcome of SARS-CoV-2 pneumonia*. https://doi.org/10.21203/rs.3.rs-23196/v1

Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., Li, Y., Cai, J., Yang, Z., Zhu, J., Zhao, M., Huang, H., Xie, X., & Li, S. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. Içinde *medRxiv* (s. 2020.04.02.20051136). medRxiv. https://doi.org/10.1101/2020.04.02.20051136

Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jin, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., … Yuan, Y. (2020). A machine learning-based model for survival prediction in patients with severe COVID-19 infection. Içinde *medRxiv* (s. 2020.02.27.20028027). medRxiv. https://doi.org/10.1101/2020.02.27.20028027

Yang, H. S., Hou, Y., Vasovic, L. V., Steel, P., Chadburn, A., Racine-Brzostek, S. E., Velu, P., Cushing, M. M., Loda, M., Kaushal, R., Zhao, Z., & Wang, F. (2020). Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. Içinde *medRxiv* (s. 2020.06.17.20133892). medRxiv. https://doi.org/10.1101/2020.06.17.20133892

YAVAŞ, M., GÜRAN, A., & UYSAL, M. (2020). Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması. *European Journal of*

Science and Technology, 258–264. https://doi.org/10.31590/ejosat.779952

Yetginler, B. (2019). *Rahim Ağzı Kanserinin Veri Madenciliği Yöntemleri ile Sınıflandırılması* (C. 8, Sayı 5). Kırıkkale Üniversitesi.

YILMAZ, E., & AYDIN, D. (2019). Estimation of Right Censored Nonparametric Regression Solved by kNN Imputation: A Comparative Study. *Turkiye Klinikleri Journal of Biostatistics*, *11*(2), 83–92. https://doi.org/10.5336/biostatic.2019-66285

Zhang, N., Zhang, R., Yao, H., Xu, H., Duan, M., Xie, T., Pan, J., Peng, E., Huang, J., Zhang, Y., Xu, X., Zhou, F., & Wang, G. (2020). Severity Detection For the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3564426

Zhang, S., Cheng, D., Deng, Z., Zong, M., & Deng, X. (2018). A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, *109*, 44–54. https://doi.org/10.1016/j.patrec.2017.09.036