MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

COVID-19 HASTALARINDA İLK BAŞVURUDAKİ VERİLERLE HASTALIK ŞİDDETİNİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ÖNGÖRÜLMESİ

KÜBRA KÖKSAL

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

DANIŞMAN

Dr.Öğr.Üyesi ZEHRA AYSUN ALTIKARDEŞ

EŞ-DANIŞMAN

Doç.Dr. BUKET DOĞAN

İSTANBUL, 2022

MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

COVID-19 HASTALARINDA İLK BAŞVURUDAKİ VERİLERLE HASTALIK ŞİDDETİNİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ÖNGÖRÜLMESİ

KÜBRA KÖKSAL
(523619016)

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

DANIŞMAN
Dr.Öğr.Üyesi ZEHRA AYSUN ALTIKARDEŞ

EŞ-DANIŞMAN
Doç.Dr. BUKET DOĞAN

İSTANBUL, 2022

TEŞEKKÜR

Veri Madenciliği alanında yüksek lisans eğitimimi tamamlamama imkân sağlayan Marmara Üniversitesi'ne, bu tezi hazırlamam için benden bilgisini ve anlayışını hiçbir zaman esirgemeyen bilgi ve deneyimleri ile yol gösteren hocalarım Dr.Öğr.Üyesi Zehra Aysun ALTIKARDEŞ ve Doç.Dr. Buket DOĞAN ve Arş. Gör. Abdulsamet Aktaş'a, sağlık sektörü üzerinde oluşan baskı ve yüke rağmen Marmara Üniversitesi Klinik Araştırma Etik Kurulunda 09.2020.487 protokol numarası ile onaylanan "Covid-19 hastalarında ilk başvurudaki verilerle hastalık şiddetinin makine öğrenmesi yöntemleri ile öngörülmesi" isimli proje kapsamında veri setinin hazırlamasından tez yazımına kadar her aşamada uzman görüşlerini ve bilgisini esirgemeyen, tıp alanındaki uzman görüşleri ile yol gösteren değerli sağlık çalışanları Prof. Dr Ali Serdar Fak, Prof. Dr. Lütfiye Mülazımoğlu Durmuşoğlu, Prof. Dr. Sait Karakurt, Doç. Dr. Elif Tükenmez Tigen, Doç. Dr. Buket Ertürk Şengel, Doç. Dr. Volkan Aydın, Doç. Dr. Gökhan Tazegül, Prof.Dr. Uluhan Sili, Uzm. Dr. Abidin Gündoğdu, Doç. Dr. Fethi Gül, Dr. Öğr.Üyesi Sena Tokay Tarhan, Prof. Dr. Semiha Emel Eryüksel, Hemşire Mümüne Topçu, Hemşire Berrin Aysevinç, Hemşire Songül Çeçen Düzel, Uzm. Dr. Tuba Güçtekin, Dr. Öğr. Üyesi Derya Kocakaya, Prof. Dr. Beste Özben Sadıç, Assoc. Prof. Halil Ataş, Prof. Dr. Mustafa Kürşat Tigen, Prof. Dr. Ahmet Altuğ Çinçin, Prof. Dr. Bülent Mutlu, Prof. Dr. Alper Kepez, Doç. Dr. Mehmet Baran Balcan, Uzm. Dr. Ayla Erdoğan, Uzm. Dr. Emre Çapar, Dr. Öğr. Üyesi Ömer Ataç, Doç. Dr. Beliz Bilgili, Prof. Dr. İsmail Cinel, Prof. Dr. Ahmet Akıcı, Prof. Dr. Rafi Haner Direskeneli'ye, öğrenim hayatım süresince her türlü konuda yardımlarını esirgemeyen arkadaşlarıma, yüksek lisans eğitimim sırasında zaman zaman ihmal etmek zorunda kalmama rağmen maddi ve manevi olarak beni sürekli destekleyen ve her zaman yanımda olan, motive edip bana güç veren değerli aileme yürekten teşekkür sunarım.

Aralık, 2022

Kübra KÖKSAL

İÇİNDEKİLER

TEŞEKKÜR i
İÇİNDEKİLER ii
ÖZET iv
ABSTRACT v
SEMBOLLER vi
KISALTMALAR vii
ŞEKİL LİSTESİ ix
TABLO LİSTESİ xi
1. GİRİŞ 1
1.1. İlgili Çalışmalar 6
1.1.1. Klinik veriler kullanılarak yapılan analizler 8
1.1.2. Görüntü verileri kullanılarak yapılan analizler 13
1.1.3. Zaman serisi verileri kullanılarak yapılan analizler 15
1.2. Amaç 16
2. MATERYAL VE YÖNTEM 17
2.1. Verisetinin Oluşturulması 18
2.1.1. Laboratuvar veri seti 19
2.1.2. Tıbbi bilgi veri seti 20
2.1.3. Nihai veri seti 21
2.2. Veri Önişleme 23
2.2.1. Veri temizleme 24
2.2.1.1. Gün aralıklarına göre eksik veri doldurma 24
2.2.1.2. Ortalama değer atama yöntemi kullanılarak eksik veri doldurma 26
2.2.2. Veri indirgeme 26
2.2.2.1. Temel bileşenler analizi (PCA) 27
2.2.3. Veri dönüştürme 28
2.2.3.1. Min-max normalizasyon 29
2.2.3.2. Z-skor normalizasyon 29
2.3. Özellik Seçimi (Feature Selection) 29
2.4. Dengesiz Veri Setlerinde Aşırı Örnekleme (Oversampling) 31
2.4.1. SMOTE (Synthetic Minority Oversampling Technique) 32
2.5. Makine Öğrenmesi 32
2.5.1. Sınıflandırma 34
2.5.1.1. K-En yakın komşu (KNN) 34

2.5.1.2. Bagging 35
2.5.1.3. Rastgele orman 36
2.5.1.4. Karar ağacı 37
2.5.1.5. Sınıflandırma modellerinin değerlendirilmesi 37
3. BULGULAR VE TARTIŞMA 39
3.1. Hastaneye İlk Yatış Aşamasında Oksijen İhtiyacının Tahmini 41
3.1.1. Analiz - 1 72 özellikle oluşturulan model sonuçları 43
3.1.1.1. Analiz - 1 72 özellikli rastgele orman modeli 44
3.1.2. Analiz - 1 16 özellikle oluşturulan model sonuçları 45
3.1.2.1. Analiz - 1 16 özellikli rastgele orman modeli 46
3.1.3. Analiz - 1 16 özellikli SMOTE uygulanmış model sonuçları 48
3.1.3.1 Analiz - 1 SMOTE 16 özellikli rastgele orman modeli 50
3.2. Hastanede Yatış Süresi Boyunca Oksijen İhtiyacının Tahmini 53
3.2.1. Analiz - 2 73 özellikle oluşturulan model sonuçları 55
3.2.1.1 Analiz - 2 73 özellikli karar ağacı modeli 56
3.2.2. Analiz - 2 18 özellikle oluşturulan model sonuçları 58
3.2.2.1 Analiz - 2 18 özellikli rastgele orman modeli 59
3.2.3. Analiz - 2 18 özellikli SMOTE uygulanmış model sonuçları 61
3.2.3.1 Analiz - 2 SMOTE 18 özellikli rastgele orman modeli 62
3.3. Hastanede Yatış Süresi Boyunca Yoğun Bakım İhtiyacının Tahmini 65
3.3.1. Analiz - 3 73 özellikle oluşturulan model sonuçları 68
3.3.1.1 Analiz - 3 73 özellikli karar ağacı modeli 69
3.3.2. Analiz - 3 12 özellikle oluşturulan model sonuçları 71
3.3.2.1 Analiz - 3 12 özellikli rastgele orman modeli 72
3.3.3. Analiz - 3 12 özellikli SMOTE uygulanmış model sonuçları 74
3.3.3.1 Analiz - 3 SMOTE 12 özellikli rastgele orman modeli 75
4. SONUÇLAR 78
5. ÖNERİLER 84
KAYNAKLAR 85
EKLER 95

ÖZET

COVID-19 HASTALARINDA İLK BAŞVURUDAKİ VERİLERLE HASTALIK ŞİDDETİNİN MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ÖNGÖRÜLMESİ

Covid-19 11 Mart 2020'de Dünya Sağlık Örgütü tarafından pandemi olarak ilan edilen, SARS-CoV-2 virüsünün neden olduğu bir bulaşıcı hastalıktır. İlk ortaya çıktığı tarihten itibaren hızlı bir şekilde yayılarak sağlık sektörü başta olmak üzere tüm alanlarda insan hayatını zorlayacak bir sürecin başlamasına neden olmuştur. Marmara Üniversitesi Hipertansiyon ve Ateroskleroz Eğitim, Uygulama ve Araştırma Merkezi koordinasyonunda, Etik Kurul onayı, 09.2020.487 sayılı protokol ve 27.04.2020 tarihli Sağlık Bakanlığı izni ile "Covid-19 hastalarında ilk başvurudaki verilerle hastalık şiddetinin makine öğrenmesi yöntemleri ile öngörülmesi" adlı proje kapsamında gerçekleştirilen bu tez kapsamında da sağlık kaynaklarının kullanımı için uygun bir strateji ve etkili bir triyaj, risk sınıflandırma yöntemi oluşturulmasında yardımcı olması hedeflenmektedir. 27 Nisan-1 Haziran 2020 tarihleri arasında İstanbul Marmara Üniversitesi Araştırma Hastanesi'ne başvuran 180 hastanın 28 günlük takip süreçlerini içeren demografik veriler, ilk laboratuvar sonuçlarından oluşan bir veriseti ile hedef değer olarak belirlenen WHO değeri arasındaki ilişki makine öğrenmesi yöntemleri kullanılarak incelenmiştir. Gün aralıklarına ve ortalamaya göre eksik veri doldurma teknikleri kullanılarak veriler tamamlanmıştır. Özellik Seçme aşamasında Rastgele Özellik Seçme algoritması kullanılarak özellik seçimi yapılmıştır. Veriseti üzerinde oksijen ihtiyacı ve yoğun bakım ihtiyacı olmak üzere belirlenen iki farklı sonlanım durumu ile ilk laboratuvar sonuçları arasındaki ilişki K-En yakın komşu, Torbalama, Rastgele Orman ve Karar Ağacı makine öğrenmesi yöntemleri kullanılarak analiz edilmiştir. Hastanın ilk yatış aşamasındaki oksijen ihtiyacı 16 özellik ile %91,67, yatış sırasındaki oksijen ihtiyacı 18 özellik ile %91,96 ve yatış sırasındaki yoğun bakım ihtiyacı 12 özellik ile %92,17 doğruluk değerlerine ulaşılarak tahmin edilmiştir.

Anahtar Kelimeler: Covid-19, makine öğrenmesi, Laboratuvar verisi, Prognoz

ABSTRACT

PREDICTION OF DISEASE SEVERITY WITH MACHINE LEARNING METHODS IN COVID-19 PATIENTS WITH THE DATA AT THE FIRST HOSPITALIZATION

Covid-19 is an infectious disease caused by the SARS-CoV-2 virus, which was declared a pandemic by the World Health Organization on March 11, 2020. Since its first appearance, it has spread rapidly and caused the start of a process that will force human life in all fields, especially in the health sector. This thesis study is carried out under the coordination of Marmara University Hypertension and Atherosclerosis Education, Application and Research Center, with the approval of the Ethics Committee, the protocol numbered 09.2020,487, and the permission of the Ministry of Health dated 27.04.2020. This thesis, which was carried out within the scope of the project titled "Prediction of disease severity with machine learning methods in Covid-19 patients with the data at the first hospitalization", aims to help create an appropriate strategy for the use of health resources and an effective triage and risk classification method. The relationship between a dataset consisting of initial laboratory results, and demographic data, including 28-day follow-up periods of 180 patients who applied to Istanbul Marmara University Research Hospital between April 27 and June 1, 2020 and WHO target value was examined using machine learning methods. Data were completed by using missing data-filling techniques according to day intervals and averages. In the Feature Selection stage, feature selection was made using the Random Feature Selection algorithm. The relationship between two different endpoints, namely oxygen need and intensive care need, and the first laboratory results on the dataset was analyzed using K-Nearest Neighbor, Bagging, Random Forest and Decision Tree machine learning methods. The oxygen requirement of the patient during the first hospitalization phase was estimated with 16 features with 91.67% accuracy, the oxygen requirement during hospitalization with 18 features and 91.96%, and the intensive care need during hospitalization with 12 features with an accuracy of 92.17%.

Keywords: Covid-19, machine learning, laboratory data, prognosis

SEMBOLLER

$\boldsymbol{\mu}$	**:** Ortalama
$\boldsymbol{\sigma}$	**:** Standart Sapma
z	**:** Z-skor
n	**:** Örnek sayısı
$\boldsymbol{d_{ij}}$	**:** i. ve j. noktanın aralarındaki uzaklık
$\boldsymbol{x_{ij}}$	**:** i. noktanın j. değişkeninin değerini
$\forall$	**:** Hepsi sembolü
$\in$	**:** Kümenin elemanı sembolü
$\sum$	**:** Toplam Sembolü
$\mathbf{r_p}$	**:** (j-1)'inci yinelemede özellik vektörünün alaka düzeyi
$\mathbf{r'_p}$	**:** Özellik vektörünün güncellenmiş geçerli alaka düzeyi değerleri
$(\mathbf{P.C})_\mathbf{j}$	**:** j. yineleme için P.C işlevinin değeri
E.C	**:** Beklenen kriter
d(a, b)	**:** a ve b noktanın aralarındaki uzaklığı
$\mathbf{a_i}$	**:** a noktasının i. değişken değeri
$\mathbf{b_i}$	**:** b noktasının i. değişken değeri

KISALTMALAR

ABD : Amerika Birleşik Devletleri
ALP : Alkalen Fosfataz
ALT : Alanin aminotransferaz
AST : Aminotransferaz testi
AUC : Area under the curve
BUN : Kan üre nitrojen
CO_2 : Karbondioksit
CRP : C-reaktif protein
CT : Computerized tomography
DBP : Diyastolik kan basıncı
DT : Decision Tree - Karar Ağacı
EGB : Extreme Gradient Boosting
EKG : Elektrokardiyografi
GBM : Gradient boosting machine
GGT : Gama glutamil transferaz
HCT : Hematokrit
HGB : Hemoglobin
IEEE : Institute of Electrical and Electronics Engineers
INR : International Normalized Ratio
KNN : K-Nearest Neighbors
MAPE : Mean absolute percentage error
MCH : Mean corpuscular hemoglobin
MCV : Mean corpuscular volume
MEWS : Modifiye Erken Uyarı Puanına
MLP : Multilayer neural networks
NB : Naive Bayes
NIV : Noninvaziv mekanik ventilasyon
$PaCO_2$: Kısmi karbondioksit basıncı
PaO_2 : Parsiyel Arteriyel Oksijen Basıncı
PCA : Principal component analysis
PDW : Platelet Distribution Width

PLT **:** Platelets, trombosit

PT **:** Protrombin zamanı

RBC **:** Red Blood Cell

RMSE **:** Root-mean-square deviation

RNA **:** Ribonükleik asid

RF **:** Random Forest - Rastgele Orman

RT-PCR **:** Ters transkriptaz polimeraz zincir reaksiyonuna

SBP **:** Sistolik kan basıncı

SpO_2 **:** Oksijen doygunluk

SVM **:** Support vector machine

WBC **:** White Blood Cell

WHO **:** World Health Organization - Dünya Sağlık Örgütü

YSA **:** Yapay Sinir Ağları

DP **:** Doğru Pozitif

DN **:** Doğru Negatif

YP **:** Yanlış Pozitif

YN : Yanlış Negatif

ŞEKİL LİSTESİ

Şekil 1.1. SARS-CoV-2 aşılarının kategorileri 2

Şekil 2.1. Tez kapsamında uygulanan işlem akışı 18

Şekil 2.2. Laboratuvar veri setindeki hastaların cinsiyete göre yaş dağılımı. 20

Şekil 2.3. 180 adet hastanın tıbbi geçmiş ve sigara içme durumu 22

Şekil 2.4. 180 adet hastaya ait WHO giriş-çıkış dağılımı 22

Şekil 2.5. Gün aralıklarına göre eksik veri doldurma örneği 26

Şekil 2.6. Örnek seçimi ve özellik seçimi 27

Şekil 2.7. SMOTE algoritması 32

Şekil 2.8. Makine öğrenmesi yöntemleri 33

Şekil 2.9. KNN algoritması (K=4) 35

Şekil 2.10. Rastgele Orman Algoritması 36

Şekil 3.1. Analiz – 1 seçilen 16 özelliğin CA algoritmasına göre katkıları 43

Şekil 3.2. Analiz – 1 72 özellik Rastgele Orman doğruluk değerleri 44

Şekil 3.3. Analiz – 1 72 özellikli Rastgele Orman modeli karmaşıklık matrisi 45

Şekil 3.4. Analiz – 1 16 özellik Rastgele Orman doğruluk değerleri 47

Şekil 3.5. Analiz – 1 16 özellikli Rastgele Orman modeli karmaşıklık matrisi 48

Şekil 3.6. Analiz – 1 SMOTE 16 özellik Rastgele Orman doğruluk değerleri 50

Şekil 3.7. Analiz – 1 SMOTE 16 özellikli Rastgele Orman modeli karmaşıklık matrisi 51

Şekil 3.8. Analiz – 1 SMOTE 16 özellikli Rastgele Orman modeli sınıflandırma grafiği 52

Şekil 3.9. Analiz – 2 seçilen 18 özelliğin CA algoritmasına göre katkıları 55

Şekil 3.10. Analiz – 2 73 özellikli Karar Ağacı modeli karmaşıklık matrisi 57

Şekil 3.11. Analiz – 2 73 özellikli Karar Ağacı Modeli 58

Şekil 3.12. Analiz – 2 18 özellik Rastgele Orman doğruluk değerleri 59

Şekil 3.13. Analiz – 2 18 özellikli Rastgele Orman modeli karmaşıklık matrisi 60

Şekil 3.14. Analiz – 2 SMOTE 18 özellik Rastgele Orman doğruluk değerleri 63

Şekil 3.15. Analiz – 2 SMOTE 18 özellikli Rastgele Orman modeli karmaşıklık matrisi 64

Şekil 3.16. Analiz – 2 SMOTE 18 özellikli Rastgele Orman modeli sınıflandırma grafiği 65

Şekil 3.17. Analiz – 3 seçilen 12 özelliğin CA algoritmasına göre katkıları 67

Şekil 3.18. Analiz – 3 73 özellikli Karar Ağacı modeli karmaşıklık matrisi 70

Şekil 3.19. Analiz – 3 73 özellikli Karar Ağacı Modeli 70

Şekil 3.20. Analiz – 3 12 özellik Rastgele Orman doğruluk değerleri 72

Şekil 3.21. Analiz – 3 12 özellikli Rastgele Orman modeli karmaşıklık matrisi..... 73

Şekil 3.22. Analiz – 3 SMOTE 12 özellik Rastgele Orman doğruluk değerleri..... 76

Şekil 3.23. Analiz – 3 SMOTE 12 özellikli Rastgele Orman modeli karmaşıklık matrisi 77

Şekil 3.24. Analiz – 3 SMOTE 12 özellikli Rastgele Orman modeli sınıflandırma grafiği 78

TABLO LİSTESİ

Tablo 1.1. WHO Klinik İlerleme Ölçeği 4

Tablo 1.2. Belirlenen anahtar kelimelere göre arama sonuçları 6

Tablo 2.1. Tıbbi bilgi veri seti nitelik açıklamaları 20

Tablo 2.2. 3 analizde kullanılan veri seti bilgileri 23

Tablo 2.3. Veri setindeki 75 özellik için belirlenen gün aralığı değerleri 24

Tablo 2.4. Karmaşıklık matrisi 38

Tablo 3.1. Her analiz için oluşturulan 16 sınıflandırma modelinin özellikleri..... 40

Tablo 3.2. Analiz – 1'de kullanılan 72 özellik ve RSFS algoritması ile seçilen 16 özellik 42

Tablo 3.3. Analiz – 1 72 özellik sınıflandırma performans sonuçları 43

Tablo 3.4. Analiz – 1 72 özellikli Rastgele Orman Modeli Performans Metrikleri 45

Tablo 3.5. Analiz – 1 16 özellik sınıflandırma performans sonuçları 46

Tablo 3.6. Analiz – 1 16 özellikli Rastgele Orman Modeli Performans Metrikleri 47

Tablo 3.7. Analiz – 1 72 özellik ve 16 özellik performans sonuçları..... 48

Tablo 3.8. Analiz – 1 SMOTE 16 özellik sınıflandırma performans sonuçları..... 49

Tablo 3.9. Analiz – 1 SMOTE 16 özellik Rastgele Orman Performans Metrikleri 51

Tablo 3.10. Analiz – 1 16 özellik ve SMOTE 16 özellik performans sonuçları 52

Tablo 3.11. Analiz – 2'de kullanılan 73 özellik ve RSFS algoritması ile seçilen 18 özellik .. 54

Tablo 3.12. Analiz – 2 73 özellik sınıflandırma performans sonuçları 55

Tablo 3.13. Analiz – 2 73 özellikli Karar Ağacı Modeli Performans Metrikleri 57

Tablo 3.14. Analiz – 2 18 özellik sınıflandırma performans sonuçları 58

Tablo 3.15. Analiz – 2 18 özellikli Rastgele Orman Modeli Performans Metrikleri 60

Tablo 3.16. Analiz – 2 73 özellik ve 18 özellik performans sonuçları..... 61

Tablo 3.17. Analiz – 2 SMOTE 18 özellik sınıflandırma performans sonuçları..... 62

Tablo 3.18. Analiz – 2 SMOTE 18 özellik Rastgele Orman Performans Metrikleri 63

Tablo 3.19. Analiz – 2 18 özellik ve SMOTE 18 özellik performans sonuçları 64

Tablo 3.20. Analiz – 3 73 özellik ve RSFS algoritması ile seçilen 12 özellik 66

Tablo 3.21. Analiz – 3 73 özellik sınıflandırma performans sonuçları 68

Tablo 3.22. Analiz – 3 73 özellikli Karar Ağacı Modeli Performans Metrikleri 69

Tablo 3.23. Analiz – 3 12 özellik sınıflandırma performans sonuçları 71

Tablo 3.24. Analiz – 3 12 özellikli Rastgele Orman Modeli Performans Metrikleri 73

Tablo 3.25. Analiz – 3 73 özellik ve 13 özellik performans sonuçları..... 74

Tablo 3.26. Analiz – 3 SMOTE 12 özellik sınıflandırma performans sonuçları..... 75

Tablo 3.27. Analiz – 3 SMOTE 12 özellik Rastgele Orman Performans Metrikleri 76

Tablo 3.28. Analiz – 3 12 özellik ve SMOTE 12 özellik performans sonuçları 77

Tablo 4.1. Analiz açıklaması ve doğruluk değerleri..... 80

Tablo 4.2. Analiz – 1, Analiz – 2 ve Analiz – 3 SMOTE karşılaştırma sonuçları 82

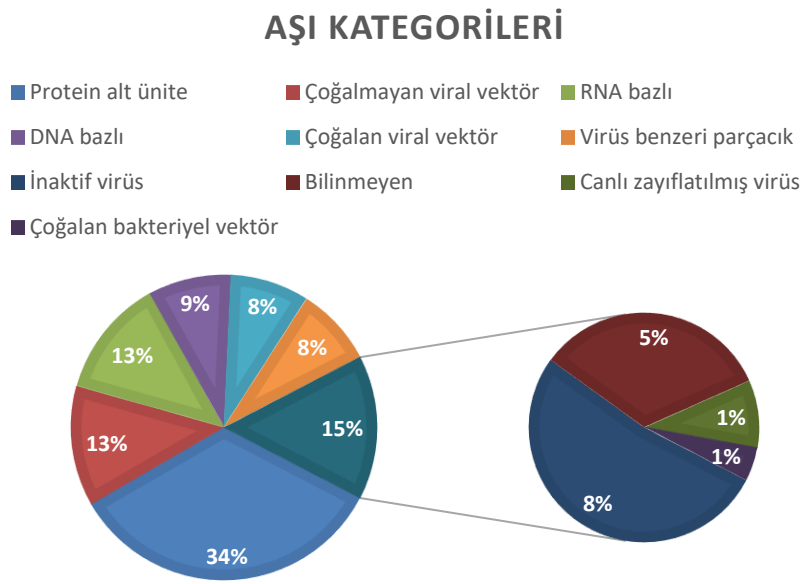
1. GİRİŞ

Şiddetli akut solunum sendromu koronavirüs 2 (SARS-CoV-2) virüsü ile ilişkili olan Covid-19 virüsü, dünya çapında bir pandemi haline gelmeden önce ilk olarak 2019 yılında Çin'in Wuhan kentinde ortaya çıkmıştır [1]. 2020 yılının Mart ayında Dünya Sağlık Örgütü tarafından pandemi olarak nitelendirilmiştir [2]. Covid-19'un ölümcüllük oranı %2 değeri ile koronavirüs ailesinden gelen diğer virüs çeşitlerine (Sars-Cov (%9,5), Mers-Cov (%35)) göre daha düşük [3] olmasına rağmen Aralık 2019'dan itibaren hızlı bir şekilde yayılarak tüm dünyayı etkisi altına almıştır. Covid-19 pandemisinin ilk ortaya çıktığı tarihten itibaren 2022 yılı Eylül ayına kadar, dünya çapında 614 milyondan fazla onaylanmış vaka ve altı milyondan fazla ölüm bildirilmiştir [4]. Bu eşi görülmemiş ve beklenmedik vaka artışı nedeniyle dünyanın dört bir yanındaki sağlık uzmanları ve sağlık tesisleri zorlu bir süreç geçirmek zorunda kalmıştır [5]. Covid-19'un heterojen klinik özellikler ile ortaya çıkabilmesi ve çoklu organ hasarına neden olmasından dolayı bu zorlu süreç boyunca etkili triyaj uygulanması, zamanında risk sınıflandırılmasının yapılması büyük bir önem teşkil etmektedir ve her bir hasta için hayati önem taşımaktadır [6,7].

Covid-19 virüsü ile enfekte olmuş kişilerde en sık görülen belirtiler öksürük, nefes darlığı, ateş, yorgunluk, kas-iskelet sistemi bozuklukları ve koku-tat duyusunun kaybedilmesidir [8,9]. Pandeminin başında elde edilen ilk raporlar en sık görülen semptomların yorgunluk ve nefes darlığı olduğunu ve kritik Covid-19 vakalarından %60-80'inin taburcu olduktan sonra da semptomlarının bir ay boyunca devam ettiğini göstermiştir [10]. Ayrıca yapılan çalışmalar erkek hastaların kadınlara göre daha yüksek ölüm ve yoğun bakım oranına sahip olduğunu ortaya koymuştur [11]. İncelenen vakalarda kardiyovasküler hastalık, kronik akciğer hastalığı gibi komorbiditeleri bulunan ve yaşlı hastaların ölüm oranlarının daha yüksek olduğu görülmüştür [12,13]. Bu yüzden pandemi boyunca bu koşullara uyan hastalar yüksek risk grubunda sayılmıştır.

Pandemi boyunca çoğunlukla ileri yaştaki hastaların yer aldığı Covid-19 şüphesi bulunan birçok insanın teşhis ve tedavi işlemleri için sağlık kuruluşlarına başvurması gerekmiştir. Artan vaka ve hasta sayısı ile birlikte sağlık sektörü üzerinde ciddi bir yük oluşmuştur. Bu da bazı ülkelerde teşhis, tedavi süreçlerinin uzamasına ve sağlık sektörünün görevini

yerine getiremeyecek hale gelmesine sebep olmuştur. Bu sorunlarla mücadele etmek için maske, mesafe ve temizlik önlemleri uygulanarak birçok farklı alanda çalışmalar yapılmıştır. Pandemiyi önlemek amacıyla çok fazla çaba sarf edilen savunma yöntemlerinden biri Covid-19 aşısının geliştirilmesidir. Aşılar hastalıklara karşı önemli bir savunma olarak büyük fayda sağlamaktadır ve günümüzde de kullanılan birçok aşı çeşidi bulunmaktadır. RNA (Ribonükleik asid) tabanlı, viral vektör tabanlı, inaktive, canlı ve alt ünite aşılar aşı çeşitlerine örnek olarak verilebilir [14]. Şekil 1.1 SARS-CoV-2 aşısı kapsamında geliştirilen farklı aşı kategorilerini göstermektedir [15]. Moderna, Pfizer/BioNTech, Sinovac, Johnson & Johnson gibi birçok aşı Covid-19 süresince geliştirilen aşılara örnek olarak verilebilir. BioNTech ve Sinovac aşıları Türkiye'de aktif olarak kullanılan aşılardandır. Ayrıca acil kullanım onayı alınarak seri üretimine başlanan TURKOVAC aşısı da aktif olarak kullanılan iki aşının yanında bir alternatif olarak vatandaşlara sunulmaktadır [16,17].



Şekil 1.1. SARS-CoV-2 aşılarının kategorileri

Aşı, test kitleri gibi sağlık temelli gelişmelerin yanında bilgisayar destekli sistemler veri bilimi, yapay öğrenme ve yapay zekâ gibi teknolojiler ile birlikte kullanılarak Covid-19 ile mücadelede tahmin, teşhis süreçlerinde büyük faydalar sağlamaktadır. İleri yaşta olan ve kalp rahatsızlığı, obezite, kanser gibi kronik rahatsızlıklara sahip kişiler Covid-19 süresince yüksek risk grubunda sayıldığından dolayı, bu veriler ile Covid-19 arasındaki ilişkinin incelenmesi bu alanda sağlık sektörüne ve sağlık bilimine çok faydalı bilgiler

sağlamaktadır. Çeşitli klinik veriler ile Covid-19 arasında tespit edilen bu ilişki bilgisi kullanılarak Covid-19 teşhisi erken bir aşamada gerçekleştirilebilir veya solunum desteğine ihtiyaç duyacak ya da durumu daha da kötüleşebilecek hastalar için risk sınıflandırılması yapılarak erkenden karar verilmesi sağlanabilir. Teşhis süreçlerini hızlandırabilecek veya çeşitli analizler ile birlikte solunum cihazı, yoğun bakım üniteleri gibi tıbbi kaynakların en doğru şekilde kullanılmasını sağlayacak her türlü bilimsel çalışma bu süreçte büyük önem arz etmektedir.

Tez kapsamında tüm analiz işlemleri WHO ölçüt değeri hedef değer olarak kullanılarak gerçekleştirilmiştir. WHO ölçüt değeri DSÖ (Dünya Sağlık Örgütü) liderliğinde kurulan Klinik Karakterizasyon ve Yönetim Çalışma Grubu tarafından geliştirilmiştir. Grup üyeleri epidemiyoloji, klinik araştırmalar, bulaşıcı hastalıklar ve viroloji bulaşıcı hastalıklar gibi alanlarda uzmanlığa sahiptir. WHO Klinik İlerleme Ölçeği, SARS-CoV-2 ile enfekte olmuş ve Covid-19'a yakalanmış hastalara fayda sağlamak amacıyla en uygun kaynak planlamasını sağlayabilmek ve bilgi alışverişini hızlandırabilmek için klinik araştırmalar ve kohort çalışmalar arasında veri havuzu oluşturma işlemini kolaylaştırmak üzere geliştirilmiştir. Bu ölçüt değeri Covid-19 gibi ortaya çıkabilecek bir bulaşıcı hastalık salgınında kullanılmak üzere avantajlı bazı özellikler içermektedir. Klinik kayıtlardan elde edilerek oluşturulmuş bu değer 0 (enfekte olmayan) ile 10 (ölüm) arasında bir hastalık şiddeti ölçü bilgisi sağlar. Enfeksiyon olmama durumundan ölüm durumuna kadar uzanan bu geniş spektrum bu ölçeğin geniş bir aralıkta kullanılmasına imkan verir [18]. WHO ölçüt değerinin yanında kısa vadede hastanın durumunun kötüleşmesini tanımlamaya yardımcı olan erken uyarı sistemleri de Covid-19 hastalarında kullanılmıştır [19] ancak son zamanlarda bu erken uyarı sistem modellerinin hiçbirinin SARS-CoV-2 virüsüne yakalanan hastalarda 24 saat içinde yoğun bakım ihtiyacını tahmin etmede yeterli olduğu bildirilmemiştir [20]. Bu yüzden Covid-19 hastalarında ilk yatış aşamasında risk sınıflandırılması önemli bir zorluk olmaya devam etmektedir.

Tablo 1.1 Çalışmada referans alınan ve DSÖ (Dünya Sağlık Örgütü) tarafından belirlenen WHO ölçüt değerlerini göstermektedir. 10 farklı WHO değeri bulunmaktadır ve bu 10 farklı WHO değeri kullanılarak hastalar beş farklı sınıfta değerlendirilmektedir [18]. 0 değeri hasta için herhangi bir enfekte olmama durumunu ve 10 değeri ise ölüm durumunu temsil etmektedir. Kalan WHO değerleri ise ayakta hafif hastalık (Ambulatory mild disease), hastanede yatış: orta derecede hastalık (Hospitalised: moderate disease) ve

hastanede yatış: ciddi derecede hastalık (Hospitalised: severe disease) olmak üzere üç farklı başlık altında değerlendirilebilir.

Tablo 1.1. WHO Klinik İlerleme Ölçeği

Hasta Durumu	Tanımlayıcı	Skor
Enfekte Değil	Enfekte değil; viral RNA saptanmadı	0
Ayakta hafif hastalık	Asemptomatik; viral RNA tespit edildi	1
	Semptomatik; bağımsız	2
	Semptomatik; yardım gerekli	3
Hastanede yatan: orta derecede hastalık	Hastanede yatan; oksijen tedavisi yok	4
	Hastanede; maske veya burun ucuyla oksijen	5
	Hastanede yatış; NIV veya yüksek akış ile oksijen	6
Hastanede yatış: ağır hastalık	Entübasyon ve mekanik ventilasyon, p0,/FiO, z150 veya Sp0,/FiO, a200	7
	Mekanik ventilasyon p0,/F10, .150 (SpOSFiO, ‹200) veya vazopressörler	8
	Mekanik ventilasyon pO,/FiO , <150 ve vazopressörler, diyaliz veya ECMO 9	9
Ölüm	Ölüm	10

Bahsedilen bu erken triyaj, kaynakların uygun kullanımı, erken risk sınıflandırmasındaki eksikliklerden ve bu eksiklerin sağlık sektörü ve sağlık çalışanları üzerindeki olumsuz etkilerinden dolayı bu tez çalışmasında hastaneye yatırılan COVID-19 ile enfekte olmuş hastaların hastaneye kabulde alınan temel klinik ve laboratuvar özellikleri kullanılarak WHO odaklı hastalık şiddetini ve/veya ilerlemesini tahmin etmeye yardımcı olan makine öğrenmesi modellerinin geliştirilmesi ve test edilmesi amaçlamıştır. Tez çalışması kapsamında nihai veri seti üzerinde üç farklı sorunun analiz edilmesi ve elde edilen bulguların paylaşılması hedeflenmiştir. Aşağıdaki listelenen üç araştırma sorusu bu analizlere karşılık olarak tez aşamasının başında oluşturulmuştur.

- Hastanın hastaneye ilk yatış aşamasında oksijen tedavisi ihtiyacı olma olasılığı nedir?

  Analizde hastanın ilk laboratuvar sonuçları ile yaş ve cinsiyet demografik bilgileri modele sunulacak parametreler olarak seçilmiştir. Hedef ise ilk yatış aşamasındaki oksijen ihtiyacını tespit etmek olduğundan dolayı WHO giriş olarak belirlenmiştir. WHO giriş sınır değeri dört olarak belirlenerek dört ve altında bir WHO giriş değerine sahip olan hastalar oksijen tedavisine ihtiyaç duymayacak, beş ve üstünde bir WHO giriş değerine sahip olan hastalar oksijen tedavisine ihtiyaç duyacak şekilde kabul edilmiştir.

- Hastanın hastanede yattığı süre boyunca oksijen tedavisi ihtiyacı olma olasılığı nedir?

  Analizde hastanın ilk laboratuvar sonuçları ile yaş, cinsiyet demografik bilgileri ve WHO giriş değeri modele sunulacak parametreler olarak seçilmiştir. Hedef ise hastanın hastanedeki tüm yatış süresi boyunca oksijen tedavisine olan ihtiyacını tespit etmek olduğundan dolayı WHO en kötü olarak belirlenmiştir. WHO en kötü sınır değeri dört olarak belirlenerek dört ve altında bir WHO en kötü değerine sahip olan hastalar oksijen tedavisine ihtiyaç duymayacak, beş ve üstünde bir WHO en kötü değerine sahip olan hastalar oksijen tedavisine ihtiyaç duyacak şekilde kabul edilmiştir.

- Hastanın hastanede yattığı süre boyunca yoğun bakım ihtiyacı olma olasılığı nedir?

  Analizde hastanın ilk laboratuvar sonuçları ile yaş, cinsiyet demografik bilgileri ve WHO giriş değeri modele sunulacak parametreler olarak seçilmiştir. Hedef ise hastanın hastanedeki tüm yatış süresi boyunca yoğun bakıma olan ihtiyacını tespit etmek olduğundan dolayı WHO en kötü olarak belirlenmiştir. WHO en kötü sınır değeri altı olarak belirlenerek altı ve altında bir WHO en kötü değerine sahip olan hastalar yoğun bakıma ihtiyaç duymayacak, yedi ve üstünde bir WHO en kötü değerine sahip olan hastalar yoğun bakıma ihtiyaç duyacak şekilde kabul edilmiştir.

  Ayrıca çalışmada yerli bir veri seti oluşturulup kullanılarak Covid-19 alanında yerli veri seti çözümüne katkı sağlanması hedeflenmiştir. Hastanın ilk laboratuvar sonuçları ile yoğun bakım ve oksijen tedavisi ihtiyacı arasındaki analiz sonucunda elde edilen anlamlı bilgiler sayesinde pandemi döneminde bu gibi tıbbi kaynakların daha yararlı ve etkili bir şekilde kullanılması, 24 saat içerisinde alınarak analiz edilen laboratuvar sonuçları sayesinde etkili triyaj ve zamanında risk sınıflandırılması işlemlerinin gerçekleştirilmesi daha kolay bir şekilde sağlanacaktır. Ayrıca tüm laboratuvar sonuçları yerine özellik seçme algoritması sonucunda elde edilen daha az sayıda özellik kullanılarak da yoğun

bakım ve oksijen ihtiyacının yüksek doğruluk ile tespit edilebileceği gösterilmiştir ve analizler kapsamında yoğun bakım ve oksijen ihtiyacı ile ilgili en ilişkili özelliklerin bulunması hedeflenmiştir.

1.1. İlgili Çalışmalar

İlgili çalışmalar kapsamında 2019-2021 tarihleri arasında IEEE ve Science Direct platformlarında beş farklı anahtar kelime kombinasyonu kullanılarak yapılan aramalar sonucunda 532 adet literatür çalışması elde edilmiştir. Tablo 1.2 belirlenen beş farklı anahtar kelime sonucunda her bir yayın platformunda bulunan toplam yayın sayısını göstermektedir.

Tablo 1.2. Belirlenen anahtar kelimelere göre arama sonuçları

Anahtar kelimeler	Veri tabanı sonuçları	Toplam yayın sayısı
Covid-19 AND machine learning AND diagnosis	Science Direct: 26 IEEE: 33	59
Covid-19 AND machine learning AND prediction	Science Direct: 55 IEEE: 29	84
Covid-19 AND machine learning AND classification	Science Direct: 25 IEEE: 46	71
Covid-19 AND machine learning	Science Direct: 131 IEEE: 136	267
SARS-CoV-2 AND machine learning	Science Direct: 38 IEEE: 13	51

532 adet çalışma içerisinden aşağıdaki belirtilen içe alma (inclusion) ve dışlama (exclusion) kriterleri kullanılarak 49 adet çalışma incelenmek üzere seçilmiştir.

İçe alma kriterleri aşağıdaki gibidir;

- Çalışmalar, Covid-19 alanında kullanılan makine öğrenmesi algoritmaları hakkında deneysel veriler sağlamalıdır.
- Araştırma çalışmaları 2019-2021 yılları arasında yayınlanmış olmalıdır.
- Çalışmalar İngilizce olmalıdır.
- Makine öğrenimi yöntemleri, Covid-19 tespit, teşhis veya tahmin alanlarıyla ilgili

olmalıdır.

Dışlama kriterleri aşağıdaki gibidir;

- Çalışmalar, makine öğrenimi algoritmaları ve sağladığı performans etkisinin kapsamı dışındadır.
- Ampirik veri sağlamayan çalışmalar,
- Virüsün hava koşulları, stres oranı, eğitim gibi alanlarla ilişkisini inceleyen Covid-19 tespiti veya teşhisi dışındaki çalışmalar kapsam dışı bırakılmıştır.

Seçilen 49 adet çalışmadan 13 tanesi klinik ve metin verileri, 14 tanesi görüntü verileri, 21 tanesi zaman serisi ve bir tanesi hem zaman serisi hem de klinik veriler üzerinde makine öğrenmesi algoritmalarını kullanarak analizler gerçekleştirmiştir. 49 adet çalışma Covid-19 hastalığının teşhisi, eğri uydurma ile hastalığın gidişatının tahmin edilmesi ve farklı sonlanım durumlarının tahmin edilmesi gibi üç farklı analiz türüne göre kategorilere ayrılmıştır.

Literatürdeki örneklerden yola çıkarak, farklı hayati değerler, laboratuvar ve demografik veriler kullanılarak Covid-19'un yüksek doğrulukta tespit edilebildiği, solunum ihtiyacı, solunum dekompansasyonu ve ölüm oranı gibi oranların başarı ile tahmin edilebildiği sonucuna varılmıştır. Çalışmalar ayrıntılı bir şekilde incelendiğinde 24 saat içerisinde alınan laboratuvar, test sonuçları kullanılarak solunum ve yoğun bakım ihtiyacını tespit eden çok az sayıda çalışma görülmüştür ve laboratuvar, test sonuçlarını kullanarak her iki durumu da değerlendiren, bu kaynakların daha etkili bir şekilde kullanılmasına yardımcı olan bir çalışma görülmemiştir. Etkili triyaj ve risk sınıflandırması yapılabilmesi için literatürden farklı olarak bu tez kapsamında 24 saat içerisinde alınan laboratuvar sonuçları kullanılarak hastaneye yatış aşamasındaki ve hastanede yatış süresi boyunca solunum ihtiyacı ile hastanede yatış süresi boyunca yoğun bakım ihtiyacı olmak üzere hem solunum ihtiyacı hem de yoğun bakım ihtiyacını değerlendiren ayrı makine öğrenmesi yöntemleri geliştirilmiştir. Literatürdeki çalışmalardan farklı olarak özgün, yerli bir yeni veri seti üzerinde analizler yapılmış ve hedef değer olarak DSÖ (Dünya Sağlık Örgütü) tarafından belirlenen WHO ölçüt değerleri sağlık çalışanları tarafından belirlenerek analizlerde kullanılmıştır. Literatür araştırması sonucunda incelenen çalışmalarda kullanılan makine öğrenmesi algoritmaları incelenerek yüksek doğrulukta modeller oluşturulabilen ve sıklıkla kullanılan RF (18), DT(7), KNN(7) ve Bagging makine

öğrenmesi yöntemleri tez çalışmasında kullanılmak üzere seçilmiştir. Rastgele Özellik Seçme algoritması kullanılarak WHO değeri ile ilişkili en önemli özellikler seçilmiştir ve daha az sayıda özellik ile yüksek doğruluklara ulaşılması, yoğun bakım, oksijen ihtiyacı ile en ilişkili özelliklerin seçilmesi hedeflenmiştir. Tez kapsamında oluşturulan bu algoritmaların sınırlı sayıda oksijen, yoğun bakım kaynaklarına sahip olan bölgeler için uygun bir kullanım stratejisi oluşturulmasında ve bir triyaj protokolü, risk sınıflandırma yöntemi oluşturulmasında yardımcı olması hedeflenmektedir.

1.1.1. Klinik veriler kullanılarak yapılan analizler

Arvind ve arkadaşları kısmi arteriyel CO_2 basıncı ($PACO_2$), oksijen satürasyonu (O_2 SAT), sıcaklık, kreatinin, trombosit sayısı, sistolik kan basıncı (sistolik), beyaz kan hücresi sayımı (WBC), kısmi arteriyel oksijen basıncı (PAO_2), nabız, solunum oranı (RR) ve diyastolik kan basıncı (diyastolik) gibi hayati değerleri, laboratuvar ve demografik verileri kullanarak Covid-19 şüphesi olan veya enfekte olmuş hastaların gelecekteki entübasyon oranını tahmin etmek için bir Rastgele Orman modeli geliştirilmiştir. Oluşturulan Rastgele Orman modelin performansı ventilasyon tahmini için kullanılan bir araç olan ROX indeksi ile karşılaştırılmıştır. Beş hastaneye başvuran toplam 4087 hasta üzerinde analiz işlemleri yapılmıştır. Kaplan-Meier analizi sonucunda hastaneye yatış sırasında entübasyon ihtiyacının daha yüksek olduğu görülmüştür. Kullanılan makine öğrenmesi algoritması 0,83 AUC ve 0,32 AUPRC değeri ile entübasyon riski açısından ROX endeksinden önemli ölçüde daha iyi performans göstermiştir [21].

Burdick ve arkadaşları diyastolik kan basıncı (DBP), beyaz kan hücresi (WBC), sistolik kan basıncı (SBP), kan üre nitrojen (BUN), kreatinin, trombosit sayımı, laktat, kalp hızı (HR), sıcaklık, solunum hızı (RR), oksijen doygunluk (SpO_2) ve bilirubin değerlerini kullanarak solunum dekompansasyonu tahmin etmiştir. XGBoost makine öğrenmesi yöntemi kullanılarak tahmin işlemi gerçekleştirilmiştir. Önerilen algoritma, 0,78 tanısal olasılık oranına sahip ventilasyonu tahmin etmek için kullanılan bir erken uyarı sistemi olan Modifiye Erken Uyarı Puanına (MEWS) göre 0,90 değeri ile daha yüksek bir performans ve duyarlılığa ulaşmıştır. Sonuçlar Covid-19 hastalarının 24 saat içinde mekanik ventilasyon ihtiyacının doğru bir şekilde tahmin edildiğini göstermiştir [22].

AlJame ve arkadaşları idrar testleri, kan testleri, RT-PCR testi ve SARS-CoV-2 testi dahil olmak üzere 100 laboratuvar testini kullanarak Covid-19 hastalığını teşhis etmeyi

hedeflenmiştir. Extra trees, RF, Logistic R. ve XGBoost algoritmaları kullanılarak oluşturulan bir topluluk öğrenim modeli (ERLX) kullanılmıştır. Brezilya'daki Albert Einstein Hastanesinden alınan 559 Covid-19 vakası ile 5644 veri örneği üzerinde analiz edilmiştir. Önerilen model %99,88 doğruluk, %99,38 AUC, %98,72 duyarlılık ve %99,99 özgüllük değerine ulaşmıştır. Ayrıca monositlerin Covid-19 pozitif vakalarını tahmin etmede daha performanslı olduğunu göstermiştir [23].

Soltan ve arkadaşları kan testlerini (hemoglobin, hematokrit, ortalama hücre hacmi, beyaz hücre sayısı vb.), bakım noktası kan gazı sonuçlarını (gerçek baz fazlalığı, standart baz fazlalığı, bikarbonat, kalsiyum, klorür vb.), yaşamsal belirtileri (kalp atış hızı, solunum hızı, oksijen satürasyonu, sistolik kan basıncı), kan testi sonuçlarını (albümin, alkalin fosfataz vb.) ve temel komorbidite verilerini (Charlson komorbidite indeksi) kullanarak Covid-19 hastalığını teşhis etmeyi hedeflenmiştir. Lojistik Regresyon, RF ve XGBoost algoritmaları kullanılarak üç farklı model oluşturulmuştur. Diğer sınıflandırıcılarla karşılaştırıldığında, XGBoost algoritması en yüksek sınıflandırma performansına ulaşmıştır. Geliştirilen ED (acil servisler) modeli 0,939 AUC, %77,4 duyarlılık, %95,7 özgüllüğe ulaşırken, kabul modeli 0,940 AUC, %77,4 duyarlılık, %94,8 özgüllüğe ulaşmıştır. ED modeli için en yüksek belirleyici özellikler olarak eozinofiller, bazofiller ve CRP seçilmiştir. Ayrıca kan gazı ölçümleri (kalsiyum ve metaemoglobin) ve yaşamsal belirtiler (oksijen gereksinimi ve solunum hızı) de tahmin için belirleyici değişkenler arasında yer alır [24].

Banerjee ve arkadaşları yaş, RT-PCR SARS-CoV-2 testi ve tam kan sayımı, hematokrit, hemoglobin, eozinofiller, ortalama hücresel hacim (MCV), lökositler, bazofiller, nötrofiller, trombositler, ortalama trombosit hacmi (MPV), kırmızı kan hücreleri (RBC), ortalama eritrosit hemoglobin (MCH), lenfositler, ortalama korpüsküler hemoglobin konsantrasyonu (MCHC), monositler ve kırmızı kan hücresi dağılımı gibi laboratuvar sonuçlarını kullanarak Covid-19 hastalığını teşhis etmeyi hedeflenmiştir. Logistic regression, RF, ANN ve Glmnet makine öğrenmesi yöntemleri kullanılarak dört farklı model oluşturulmuştur. Tam kan sayımını kullanan tanı prosedüründe YSA %95, RF %94 ve Glmnet %94 AUC değerine ulaşmıştır. Hastaneye kabul edilmeyen hastalarda tanı sürecinde YSA %82, RF %86 ve Glmnet %84 AUC değerlerine ulaşmıştır [25].

Mondal ve arkadaşları RT-PCR ve ek laboratuvar testlerini kullanarak Covid-19

hastalığını teşhis etmeyi hedeflenmiştir. SVM, KNN, XGBoost, MLP, Logistic R., DT, RF, ensemble methods, Linear R ve Polynomial R yöntemleri kullanılmıştır. Brezilya'daki İsrail Albert Einstein Hastanesi tarafından sağlanan 5644 veri üzerinde analiz işlemi gerçekleştirilmiştir. MLP, Logistic regression ve XGBoost %91 doğrulukla en iyi sınıflandırma modelleri olarak seçilmiştir. Ayrıca "Serum Glikoz" ilk, "Solunum Sinsitiyal Virüsü" ikinci ve "Influenza A" üçüncü en iyi ve belirleyici özellik olarak seçilmiştir [26].

Heydari ve arkadaşları klinik semptomlar, laboratuvar parametreleri (SPO2, LDH, Lymphopeni, CRP, D-dimmer) ve radyolojik görüntü bilgileri üzerinde kendi kendini organize eden haritalama tekniğini (Self Organized Mapping) kullanarak Covid-19'u tahmin eden bir yöntem geliştirmiştir. Önerilen yöntem %92,5 doğruluk değerine ulaşmıştır ve sonuçları hekim teşhisleri ile karşılaştırılmıştır. Ayrıca kümeleme sonucunda PCA yöntemi kullanılarak kümelerin sahip olduğu en önemli özellikler belirlenmiştir. Yapılan analizlere göre yaş, akciğer tutulumu ve SPO_2 özellikleri en önemli özellikler olarak seçilmiştir [27].

Yadaw ve arkadaşları yaş, cinsiyet ve etnik köken gibi demografik değişkenleri, diyabet ve astım gibi komorbiditeleri, hastaların sigara içme durumunu, vücut ısısı ve oksijen satürasyon düzeyini kullanarak ölüm oranını tahmin etmiştir. Denetimli makine öğrenmesi tekniklerinden Rastgele Orman, Lineer Regresyon, SVM ve XGBoost kullanılmıştır. Çalışmada kullanılan tahmin modelleri, iki bağımsız doğrulama veri setinde 0,91-0,94 AUC değerleri elde etmiştir. Yalnızca üç özellik kullanılarak elde edilen model, tüm özellikleri kullanan model kadar başarılı olmuştur. Ayrıca yaş ve minimum oksijen satürasyonu en belirleyici özellikler olarak seçilmiştir ve XGBoost algoritması diğer yöntemlere göre en doğru tahmin değerini üretmiştir [28].

Castelnuovo ve arkadaşları cinsiyet, yaş, kronik akciğer hastalığı, diyabet, kronik böbrek hastalığı, CRP, hipertansiyon, miyokard infarktüsü öyküsü, kalp yetmezliği, obezite ve sigara içme alışkanlığı özelliklerini kullanarak Covid-19 hastalarında ölüme yatkınlığı tahmin etmiştir ve en belirleyici faktörler analiz edilmiştir. RF modeli %95,2 hassasiyet, %30,8 özgüllük, %83,4 sınıflandırma doğruluğu ve %90,4 F1 değeri ile güçlü bir tahmin başarısı elde etmiştir. Ayrıca en tanımlayıcı özelliklerin bozulmuş böbrek fonksiyonu, yüksek C reaktif proteini ve ileri yaş olduğu bulunmuştur. Obezite, tütün kullanımı,

kardiyovasküler hastalık ve buna bağlı komorbiditelerle ilgili bir ilişki görülmemiştir. Özellikle 85 yaşındaki hastalar, 18-64 yaş arasındaki hastalara göre sekiz kat daha yüksek mortaliteye sahip olduğu sonucuna ulaşılmıştır. Bu sonuçlar çok değişkenli Cox hayatta kalma analizi kullanılarak doğrulanmıştır [29].

Ahamad ve arkadaşları cinsiyet, yaş, ateş, öksürük, zatürre, akciğer enfeksiyonu, burun akıntısı, kas ağrısı, ishal, seyahat geçmişi ve izolasyon özelliklerini kullanarak Covid-19 hastalığını teşhis etmeyi ve en ilişkili semptomları belirlemeyi hedeflemiştir. DT, RF, GBM, EGB ve SVM algoritmaları kullanılarak çeşitli modeller oluşturulmuştur. Çin'deki yedi farklı ilden (Anhui, Guangdong, Henan, Jiangsu, Shandong, Shanxi ve Zhejiang) alınan toplam 6512 hasta verisi kullanılmıştır. Covid-19 hastalığının teşhisinde yüksek tespit sağlayan özellikler kullanılarak bir makine öğrenimi modeli geliştirilmiştir. XGBoost algoritması, diğer modellere kıyasla tüm yaş gruplarındaki özellikleri seçmek için en yüksek doğruluğu (>%85) elde etmiştir. Analiz, en önemli semptomların ateş (%41,1), öksürük (%30,3), akciğer enfeksiyonu (%13,1) ve burun akıntısı (%8,43) olduğunu göstermiştir [30].

Li ve arkadaşları nüfus yoğunluğu, <10 yıllık yaşam beklentisi, cinsiyet oranı, tüberküloz prevalansı, diyabet prevalansı, kentsel nüfus yüzdesi, 12-23 aylık çocuklar arasında iki aşılama teminatı, iyileştirilmemiş su kaynakları kullanan nüfusun yüzdesi, yaygınlık HIV, sigara içme oranı, ortalama doğum yaşı, coğrafi konum ve sıcaklık gibi faktörleri inceleyerek Covid-19 vaka, ölüm oranlarını tahmin etmeyi ve en ilişkili özellikleri belirlemeyi hedeflemiştir. Lojistik regresyon yöntemi kullanılarak model oluşturulmuştur. Mortalite ve vaka oranlarıyla ilişkili faktörler 154 ülke ve 50 ABD eyaleti için belirlenmiştir. LASSO kullanılarak 27 değişken en ilişkili özellikler seçilmiştir ve test setinde vaka, ölüm ve vaka ölüm oranları tahminlerinde 0,80, 0,83 ve 0,81 AUC değerlerine ulaşılmıştır. Ekonomik eşitsizlik, alkol alımının artması Covid-19 riskinin artmasıyla ve HIV ve influenza ve pnömoni prevalansı, artan sebze, yemeklik yağ, protein, D vitamini ve K vitamini alımı Covid-19 riskinin azalmasıyla ilişkilendirilmiştir. Kan grubu B ve AB, Covid-19 için koruyucu bir faktörken kan grubu A bir risk faktörü olduğu sonucuna varılmıştır [31].

Otoom ve arkadaşları belirtiler, potansiyel olarak enfekte kişilerle temas geçmişi ve seyahat geçmişi bilgilerini kullanarak gerçek zamanlı bir IOT tabanlı Covid-19 algılama

sistemi önermiştir. Ateş için sıcaklık temelli sensörler, öksürük için ses tabanlı sensörler, yorgunluğu algılamak için hareket tabanlı kalp atış hızı sensörleri ve son olarak nefes darlığı için oksijen bazlı sensörler kullanılarak gerçek zamanlı olarak çeşitli belirtiler elde edilmiştir. SVM, NN, NB, KNN, Decision Table, Decision Stump, OneR, ve ZeroR algoritmaları kullanılarak çeşitli modeller oluşturulmuştur. Sonuçlar, beş algoritmanın da %90'ın üzerinde başarıya ulaştığını göstermiştir. SVM %92,95, NN %92,89, NB %90,58, KNN %92,89 doğruluk değeri elde etmiştir [32].

Althnian ve arkadaşları doğum tarihi, milliyet, cinsiyet, şehir ve tarama sonucu gibi demografik verileri kullanarak duyarlılık oranını tahmin etmiştir. MLP, SVM, DT ve RF algoritmaları kullanılarak dört farklı model elde edilmiştir. Duyarlılık, bir bireyin enfekte olma olasılığını ifade eder. Çalışma sonucunda DT %85,6, RF %85,3, MLP %77,2 ve SVM %74,2 doğruluk elde etmiştir. DT 0,76 ve RD 0,77 AUC değeri ile 0,77 AUC değerine ulaşan MLP 0,71 AUC değerine ulaşan SVM'den daha iyi performans göstermiştir. Ayrıca elde edilen sonuçlar bireysel duyarlılığının demografik verilere dayalı olarak yüksek doğrulukla tahmin edilebileceğini ve tıbbi geçmişe veya klinik verilere ihtiyaç duymadan yalnızca yaş, cinsiyet, milliyet ve konum kullanılarak tahmin edilebileceğini göstermiştir [33].

Klinik verileri kullanarak analizler yapan 13 çalışmadan 2 tanesi solunum, entübasyon ihtiyacını, 8 tanesi Covid-19 teşhisini ve 3 tanesi ölüm oranını tahmin etmiştir. Tüm çalışmalarda yüksek performanslı modeller oluşturulurken Rastgele Orman (RF), Karar Ağacı (DT), Destek Vektör Makineleri (SVM), XGBoost algoritmaları en çok kullanılan makine öğrenmesi yöntemlerindedir. Ayrıca kan gazı ölçümleri, yaşamsal belirtiler, ileri yaş, minimum oksijen satürasyonu, kan grubu gibi klinik parametreler bu çalışmalar sonucunda bulunan önemli niteliklere örnek olarak verilebilir.

Covid-19 hastalığının en önemli belirtisi olan solunum zorluğundan dolayı ortaya çıkan öksürük verisi de çalışmalarda Covid-19 analizi için kullanılan veri türlerinden biri olmuştur. Bansal ve arkadaşları açık olarak yayınlanan öksürük verisetini kullanarak CNN tabanlı bir sınıflandırıcı önermiştir. Öksürük verileri kullanarak hastalar Covid-19 veya değil olarak sınıflandırılmıştır. CNN ağına girdi olarak mfcc özellikleri ve spektrogram görüntüleri verilerek iki farklı yaklaşım kullanılmıştır. Mffc yaklaşımı spektrograma göre %81 hassasiyetle %70,58 test doğruluğu ile daha iyi bir performans

göstermiştir [34].

1.1.2. Görüntü verileri kullanılarak yapılan analizler

Demografik özellikler ve metinsel klinik verilerin yanında görüntü veri kümeleri de Covid-19 tespiti için sıklıkla kullanılan veri türlerinden biridir.

Yang ve arkadaşları CT görüntüleri ve biyobilgilerden yararlanarak oluşturulan bir Çok Modlu Füzyon Kademesi (MMFC) çerçevesi önermiştir. Önerilen çerçeve sırasıyla Özellik Öğrenme Modülü (BFL) ve Ortak Karar Modülü (JD)'nden oluşmaktadır. BFL kısmında, biyobilgi yardımıyla CT görüntülerinden ayırt edici özellikler öğrenilir. İkinci aşamada, biyobilgi tekrar kullanır ve ortak karar için BFL'nin sonucu birleştirilir. Elde edilen sonuçlar, çalışmada önerilen yöntemin temel yöntemlere kıyasla doğruluk ve hassasiyet açısından daha gelişmiş olduğunu göstermektedir [35].

Sun ve arkadaşları göğüs BT görüntülerine dayalı bir Covid-19 teşhisi için Uyarlanabilir Özellik Seçimi Rehberli Derin Orman (AFS-DF) modelini önermiştir. Eğitimli derin orman modeline dayalı bir yöntem özellik seçim işlemi için kullanılmıştır. 1495 Covid-19 ve 1027 toplum kökenli pnömoni (CAP) hastasına ait verilerden oluşan bir veriseti üzerinde analizler yapılmıştır. Önerilen yöntemin doğruluk (ACC), duyarlılık (SEN), özgüllük (SPE), EAA, duyarlılık ve F1 skoru sırasıyla %91,79, %93,05, %89,95, %96,35, %93,10 ve %93,07'dir. Ayrıca yöntem, CAP (toplum kaynaklı pnömoni) sınıflandırmasında yaygın olarak kullanılan dört makine öğrenimi modelinden daha iyi performans göstermiştir. Özellik seçme yöntemi ise Lasso ve ElasticNet özellik seçme yöntemleriyle karşılaştırıldığında, daha iyi bir sonuç elde etmiştir [36].

Hussain ve arkadaşları göğüs röntgeni ve CT tarama görüntülerini kullanarak otomatik olarak Covid-19'u tespit eden CoroDet adlı bir CNN modeli önermiştir. Önerilen model iki (COVID ve Normal) sınıf sınıflandırması için %99,1 doğruluğa, üç (COVID, Normal ve COVID olmayan pnömoni) sınıf sınıflandırması için %94,2 ve dört (COVID, Normal, COVID olmayan viral pnömoni ve COVID olmayan bakteriyel pnömoni) için %91,2 doğruluğa ulaşmıştır. Ayrıca önerilen CoroDet yöntemi, iki sınıf sınıflandırması ve dört sınıf sınıflandırması için mevcut tüm tekniklerden daha iyi doğruluk göstermiştir [37].

Lai ve arkadaşları Göğüs BT görüntüleri üzerinde Derin evrişimli sinir ağı (DCNN), MaskR-CNN-SVM modeli oluşturmuştur. Normal ve NCIP (yeni koronavirüs enfekte

pnömoni) tespiti için önerilen algoritma (NCIP-Net) 0,91 AUC değerine, NCIP ve akciğer nodülü tespiti için 0,95 AUC değerine ve NCIP ve diğer pnömoni tespiti için 0,83 AUC değerine ulaşmıştır [38].

Saha ve arkadaşları X-ray görüntülerini kullanarak Covid-19 teşhisi için derin öğrenme ve makine öğrenmesi tabanlı bir algılama sistemi önermiştir. Önerilen model %98,91 doğruluk, %100 kesinlik, %97,82 özgüllük, %98,89 F1 puanı ve 0,9894 ROC değeri ile diğer derin öğrenme sistemlerine göre daha iyi bir performans elde etmiştir [39].

Ohata ve arkadaşları göğüs röntgeni görüntülerini kullanarak Covid-19'un teşhisini hedeflemiştir. MobileNet öznitelik çıkarıcı ile kullanılan SVM algoritması %98,5 doğruluk ve F1 puan değerine, DenseNet201 ile kullanılan MLP algoritması %95,6 doğruluk ve F1 puan değerine ulaşmıştır [40].

Tuncer ve arkadaşları göğüs röntgeni görüntülerini kullanarak Covid-19'un teşhisini hedeflemiştir. SVM %100 sınıflandırma doğruluğu ve %99 genel başarı oranı ile en iyi sınıflandırıcı olarak seçilmiştir. Karar Ağacı ise en kötü performans gösteren makine öğrenmesi algoritması olarak seçilmiştir [41].

El-Kenawy ve arkadaşları BT görüntüleri üzerinde SVM, NN, KNN ve DT modelleri kullanılarak Covid-19'un tespit edilmesini hedeflemiştir. Öznitelik çıkarma aşamasında AlexNet kullanıldığında en yüksek sınıflandırma doğruluğuna ulaşılabileceği tespit edilmiştir. Kullanılan öznitelik seçim algoritması (SFS Kılavuzlu WOA), orijinal WOA algoritması ve LSH-SMOTE ön işleme ve önerilen oylama sınıflandırıcısı (PSO Kılavuzlu WOA) dahil olmak üzere karşılaştırılan algoritmalardan daha iyi performans göstermiştir ve ikili tahmin sonucu 0,995, 2,49569E- 05 MSE ile bir AUC değerleri ile son teknoloji topluluk öğrenimi yöntemlerinden daha iyi performans göstermiştir [42].

Javor ve arkadaşları BT görüntülerini kullanarak CNN tabanlı bir Covid-19 teşhis modeli önermiştir. Önerilen CNN tabanlı makine öğrenimi modeli sınıflandırıcısı %94,5 doğruluk ve 0,956 AUC'ye ulaşmıştır. Önerilen algoritma, 15 yıldan fazla deneyime sahip iki radyologdan daha üstün bir tahmin performansı elde etmiştir [43].

Togaçar ve arkadaşları X-ray görüntülerini kullanarak SVM tabanlı bir Covid-19 teşhis modeli önermiştir. Çalışmadan üç sınıf bilgisi kullanılarak (koronavirüs, pnömoni, normal) sınıflandırma analizleri yapılmıştır. Başarı oranı Covid-19 verilerinde %99,34, Pnömoni verilerinde %98,47 ve normal görüntülerde %98,68 olarak bulunmuştur. Genel

sınıflandırma oranı %99,27 olup, bu da SMO algoritmasının sınıflandırma başarısına katkı sağladığını göstermektedir [44].

1.1.3. Zaman serisi verileri kullanılarak yapılan analizler

Wadhwa ve arkadaşları Hindistan'daki vaka, ölüm ve iyileşme sayısını tahmin etmek için bir makine öğrenimi modeli geliştirmiştir. Bu model kullanılarak, kilitlenme süresi tahmin edilmektedir. Tüm önlemlerin doğru bir şekilde uygulanması halinde vakaların azalmaya başlayacağı ve 9 Eylül 2020'de sokağa çıkma yasağının kalkacağı önerilmektedir [45].

Ballı salgının gidişatını tahmin etmek için makine öğrenimi yöntemlerini kullanmıştır. Elde edilen sonuçlara göre en iyi tahmin performansı SVM ile elde etmiştir. SVM ve doğrusal regresyon çok yakın MAPE ve RMSE değerlerine sahiptir. MLP performansta üçüncü sırada ve RF en kötü performansa sahiptir. Edinilen sonuçlara göre salgının Ocak 2021'de zirveye ulaşacağı ve kümülatif olarak 80 milyon kişiye bulaşacağı tahmin edilmektedir [46].

Yeşilkanat Rastgele Orman algoritmasını kullanarak 190 ülke için vaka sayılarını tahmin etmiştir ve doğrulanmış vakaların sayısı ile karşılaştırmıştır. Çalışmanın sonunda RF algoritması test veri seti için 0,843 ve 0,995 R2 (ortalama R2 = 0,959), 141,76 ve 526,18 RMSE (ortalama RMSE = 259,38) elde etmiştir. Bu, RF algoritmasının yakın gelecek durum tahmini için uygun bir algoritma olduğunu göstermektedir [47].

Wang ve arkadaşları beş ülke arasında en hızlı büyüme noktası, salgının boyutu ve zirve zamanı gibi değerleri tahmin etmiştir. Küresel salgın boyutunun 21 Haziran'da 3.603.111 aktif enfeksiyonla zirveye ulaşacağı tahmin edilmektedir. Ekim sonunda 14.117.911 olacağı ve en hızlı büyüme noktasının 16 Mayıs olduğu belirlenmiştir [48].

Rustam ve arkadaşları Linear R., LASSO, SVM, ES algoritmaları üzerinde yeni enfekte vaka sayısı, ölüm sayısı ve önümüzdeki 10 gün içindeki iyileşme sayısı olmak üzere üç tür tahmin işlemi yapmıştır. Tüm modeller arasında ES en iyi performansı gösterirken, LR ve LASSO vakaları mortalite ve iyileşme oranını tahmin etmede ES'den sonra en iyi performansı göstermiştir. SVM algoritmasının tüm tahmin senaryolarında kötü performans gösterdiği gözlemlenmiştir [49].

1.2. Amaç

Bu tez çalışması, Marmara Üniversitesi Hipertansiyon ve Ateroskleroz Eğitim, Uygulama ve Araştırma Merkezi koordinasyonunda, Etik Kurul onayı, 09.2020.487 sayılı protokol ve 27.04.2020 tarihli Sağlık Bakanlığı izni ile "Covid-19 hastalarında ilk başvurudaki verilerle hastalık şiddetinin makine öğrenmesi yöntemleri ile öngörülmesi" adlı proje kapsamında gerçekleştirilmiştir. Marmara Üniversitesi Sağlık Bilimleri kampüsünde bulunan Marmara Üniversitesi Prof. Dr. Asaf Ataseven Hastanesi bu dönemde sadece Covid-19 hastalarını kabul eden pandemi hastanesi olarak görev yapmıştır. Bu çalışmada elde edilen veriler bu hastanede belirtilen araştırma grubunda yer alan tıp uzmanları tarafından 27 Nisan 2020 ile 1 Haziran 2020 tarihleri arasında toplanmıştır. Hastalara ait laboratuvar verileri Sağlık Bakanlığı sistemleri ile entegre olan ve sağlık işletmelerindeki tüm işlemleri elektronik ortamda depolayan Origo hastane bilgi yönetim sistemi üzerinden elektronik olarak çekilmiştir. Toplanılan 508 adet hasta sayısı veri ön işleme işlemlerinden geçirildikten sonra 180 adete indirgenmiş ve bu kayıtlar üzerinden analizler gerçekleştirilmiştir.

180 adet hasta kaydından oluşan nihai veri seti üzerinde özellik seçme işlemi yapılarak ilk laboratuvar sonuçları, demografik özellikler ile belirlenen üç farklı analiz hedefinin farklı hedef değerler kullanılarak analiz edilmesi, sonuçların görselleştirilmesi ve bulguların tartışılması amaçlanmaktadır.

Çalışmanın ilk adımı olan veri setinin oluşturulması aşamasında sağlık çalışanları tarafından Origo sistemine girilen laboratuvar sonuçları ve demografik bilgiler kullanılarak laboratuvar veri seti elde edilecektir. Ayrıca hastalara ait mevcut sağlık ve tıbbi geçmiş bilgileri ile WHO hedef değerlerinin bulunduğu hasta tıbbi bilgileri olmak üzere ikinci bir veri seti oluşturulacaktır. Sağlık çalışanları tarafından gerçek hastalardan alınarak oluşturulan bu veri seti ile literatürdeki yerli veri seti çözümüne katkı sağlanması hedeflenmektedir.

Laboratuvar sonuç veri setindeki ilk laboratuvar sonuçları ile hasta tıbbi bilgileri veri setindeki WHO değerleri birleştirilerek analizlerde kullanılacak olan nihai veri seti elde edilecektir. Veri setinde eksik veya niteliğin sahip olması gereken veri türüne uymayan tutarsız verilerin analiz işleminden önce ön işleme aşamasından geçirilmesi gerekmektedir. Böylece daha doğru ve etkili bir analiz işlemi gerçekleştirilebilir ve veri

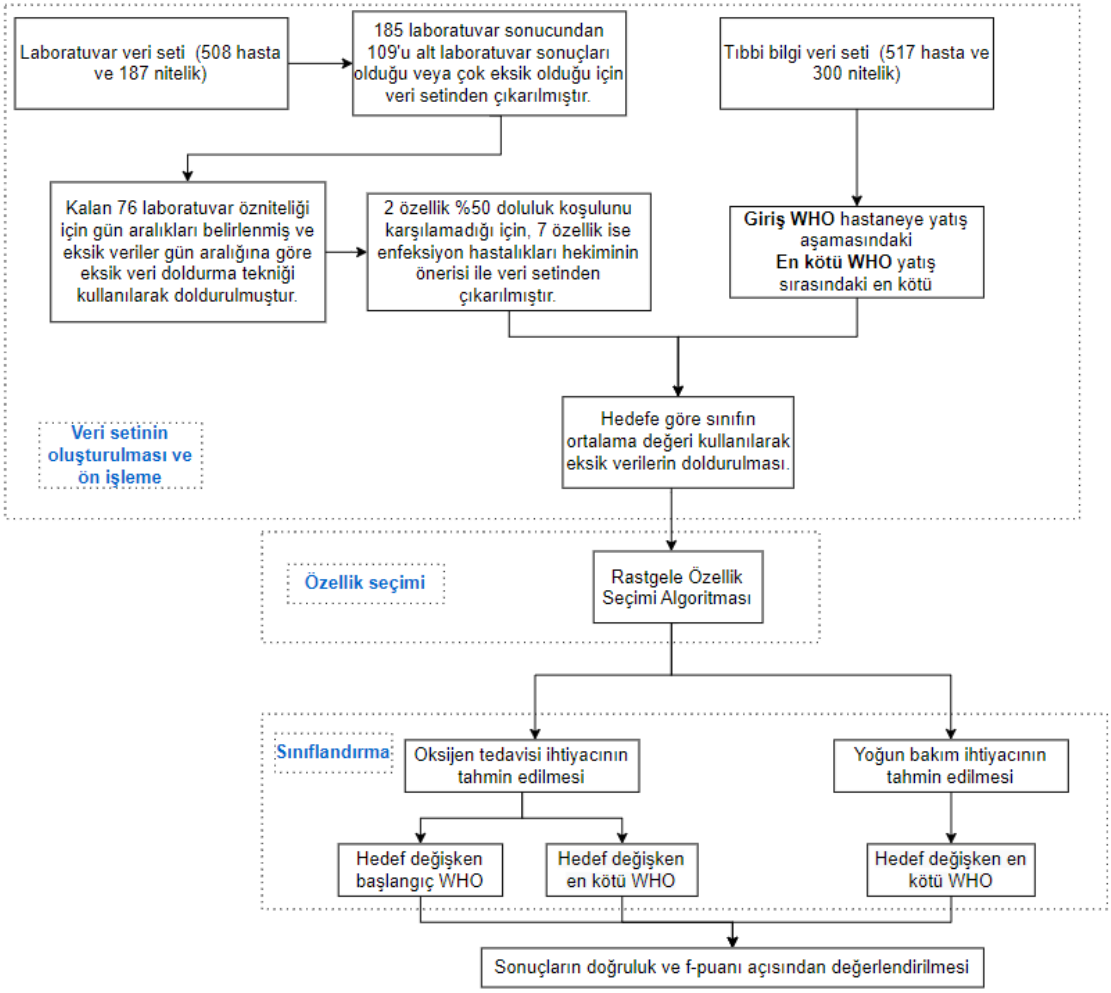
seti makine öğrenmesi yöntemlerinde uygulamaya hazır bir hale getirilebilir. Bu yüzden ikinci aşama olan ön işleme aşamasında tutarsız veriler silinerek farklı eksik veri doldurma tekniklerine göre veri setindeki eksik veriler doldurulacaktır. Gün aralıklarına göre eksik veri doldurma tekniği nihai veri seti üzerinde uygulandıktan sonra tez çalışmasında farklı hedef değerler belirlediğinden dolayı her analiz işleminden önce belirlenen hedef değere göre sınıf ortalaması yöntemi ile eksik veriler doldurulacaktır.

Üçüncü aşama olan özellik seçme aşamasında tüm laboratuvar sonuçları yerine analiz edilen sonlanım durumlarını en iyi ifade eden özellikler çıkartılarak daha az sayıda özellik ile yüksek doğrulukta sınıflandırma işleminin yapılabilmesi hedeflenmiştir. Özellik seçme aşaması için Rastgele Özellik Seçme algoritması kullanılacaktır.

Sınıflandırma aşamasında oksijen ihtiyacı ve yoğun bakım ihtiyacı olmak üzere iki farklı sonlanım durumu üzerinde üç farklı analiz yapılarak çok sayıda Covid-19 enfekte olan bu dönemlerde yoğun bakım ve oksijen ekipmanlarının daha etkili bir şekilde kullanılabilmesi ve hastanın gelecekteki durumu hakkında sağlık çalışanlarına yardımcı olarak tedavi sürecinin kolaylaştırması hedeflenmiştir. Oksijen ihtiyacı sonlanım durumu kapsamında WHO giriş ve WHO en kötü değerleri hedef değer olarak, yoğun bakım ihtiyacı için ise sadece WHO en kötü değeri hedef değer olarak kullanılacaktır. KNN, Bagging, Karar Ağacı ve Rastgele Orman olmak üzere dört farklı makine öğrenmesi algoritması kullanılarak elde edilen sınıflandırma sonuçları farklı performans metrikleri kullanılarak değerlendirilecek ve elde edilen sonuçlar grafikler yardımıyla görselleştirilerek sunulacaktır.

2. MATERYAL VE YÖNTEM

Tez Çalışması temelde beş farklı aşamadan oluşmaktadır. İlk aşamada sağlık çalışanları tarafından sisteme girilen hastaya ait laboratuvar sonuçları, demografik veriler, hastalık geçmişi gibi tüm bilgiler toplanarak nihai veri seti oluşturulmuştur. İkinci aşamada bu veri setindeki tutarsız değerler temizlenerek ve farklı eksik veri doldurma teknikleri kullanılarak veri setindeki eksik alanlar doldurulmuştur. Üçüncü aşamada özellik seçme işlemi gerçekleştirildikten sonra farklı sonlanım değerleri bakımından sınıflandırma ve tahmin işlemleri yapılarak sonuçlar değerlendirilmiştir. Şekil 2.1 tez kapsamında uygulanan metodolojinin kapsamlı bir akış şemasını temsil etmektedir.



Şekil 2.1. Tez kapsamında uygulanan işlem akışı

2.1. Verisetinin Oluşturulması

Sağlık çalışanları tarafından hastane İdaresi'nden veri kullanım izni alındıktan sonra, bu tek merkezli retrospektif çalışma, Helsinki Deklarasyonu ve Araştırma ve Yayın Etiği'ne göre yapılmıştır. Tez çalışmasında hastalara ait anonimleştirilmiş veriler kullanılmıştır. 27 Nisan 2020 ve 1 Haziran 2020 tarihleri arasında İstanbul Marmara Üniversitesi Araştırma Hastanesi'nde Covid-19 için belirlenen bölümlere başvuran tüm hastaların 28 günlük takip süreçlerini içeren bir veriseti bu çalışma kapsamında analiz edilmiştir. Çalışma Kurumsal Etik Kurulu tarafından 27.04.2020 tarihinde 09.2020.487 protokol numarası ile onaylanmıştır (70737436-050.01.04-E.2000138764). Yoğun bakım ünitelerine ilk elden kabul edilen hastalar dahil edilmemiştir. Dahil edilen hastalar, SARS-CoV-2 ribonükleik asit (RNA) için ters transkriptaz polimeraz zincir reaksiyonuna

(RT-PCR) dayalı Covid-19 enfeksiyonunu doğrulamıştır. Tüm hastalara tıbbi öykü, mevcut şikayetler, yaşamsal belirtilerin ölçümü ve kapsamlı fizik muayene, kan sayımı için başlangıç testi, böbrek ve karaciğer fonksiyon testleri ve serum ferritin ve D-dimer dahil inflamatuar belirteçler dahil olmak üzere rutin klinik değerlendirmeler sağlık çalışanları tarafından yapılmıştır ve elde edilen bilgiler analiz edilmek üzere çalışmaya yerli bir veri seti çözümü olarak sunulmuştur. Çalışma kapsamında iki farklı veri seti oluşturularak bu iki veri setinin birleştirilmesi ile nihai veri seti elde edilmiştir ve analiz işlemleri yapılmıştır.

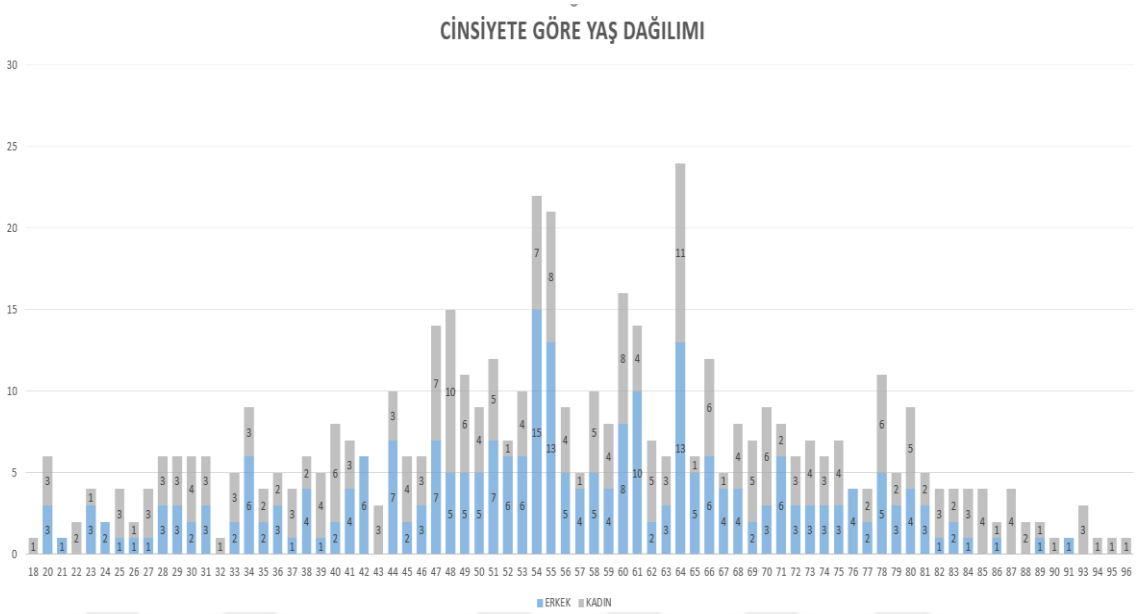
2.1.1. Laboratuvar veri seti

Çalışma kapsamında oluşturulan ilk veri setlerinden biri laboratuvar veri setidir. Hastalara ait laboratuvar verileri origo sistemi üzerinden elektronik olarak çekilmiştir. Sağlık çalışanları tarafından sisteme girilen laboratuvar sonuçları üzerinde aşağıdaki ön işleme adımları uygulanarak analizde kullanılmaya uygun bir veri seti dosyası oluşturulmuştur.

- Farklı formatta olan excel dosyaları tek bir formata getirildi.
- Python programlama dili ile hastalara ait tüm test adları taranarak tek bir dosyada toplandı. Referans değer oluşturuldu. Böylece dağınık halde olan veri tek bir excel dosyası haline getirildi.
- Test adlarına göre arama yapıp hastalara ait olan test değerleri ilgili sütunlara yazdırıldı. Bu sayede hem veri daha anlaşılır bir hale geldi hem de her bir hasta için yaklaşık olarak 3000 satır yer kaplayan veri, yaklaşık 10 satıra sığdırıldı.

Laboratuvar sonuçlarından oluşan veriseti üzerinde veri temizleme işlemleri de uygulandıktan sonra 508 hastanın toplamda 9528 adet 191 özellikten oluşan laboratuvar sonuçlarını içeren bir veriseti elde edilmiştir. Ek 2 Tablo 1 laboratuvar veri setindeki 191 özelliğin bir listesini göstermektedir.

Hastaların 259 (%51) tanesinin cinsiyeti erkek iken geri kalan 249 (%49) hasta kadındır. Yaş (56,22 ± 16,91) aralığı 18-96 arasındadır ve hemen her yaş aralığı için en az bir en fazla 24 hasta bulunmaktadır. Şekil 2.2 laboratuvar veri setindeki hastaların cinsiyete göre dağılımını temsil etmektedir.



Şekil 2.2. Laboratuvar veri setindeki hastaların cinsiyete göre yaş dağılımı.

2.1.2. Tıbbi bilgi veri seti

Tez kapsamında laboratuvar veri seti ile birlikte 517 adet hastaya ait tıbbi öykü, mevcut şikayetler, yaşamsal belirtilerin ölçümü, kapsamlı fizik muayene ve WHO değeri gibi özelliklerden oluşan 300 sütunlu bir tıbbi bilgi veri seti oluşturulmuştur. Tablo 2.1 veri setinde bulunan niteliklerin kategorilerini ve bu kategoride yer alan sütun sayısını göstermektedir.

Tablo 2.1. Tıbbi bilgi veri seti nitelik açıklamaları

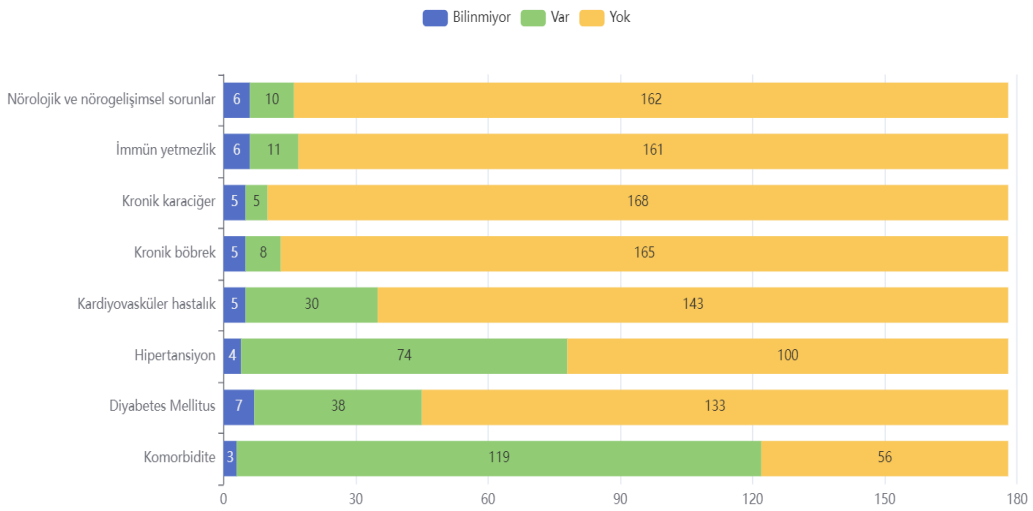
Kategori	Örnek nitelikler	Nitelik Sayısı
Demografik ve kişisel bilgiler	T.C., cinsiyet, dosya no	6
Tıbbi öykü	Komorbidite, kronik hastalıklar, sigara kullanımı	13
Mevcut şikayetler	Semptomlar, temas durumu, test sonuçları	58
Yaşamsal belirtiler ve kapsamlı fizik muayene	Hastaya ait 10 günlük takip bilgisi	220
WHO değeri	WHO giriş, çıkış, en kötü	3

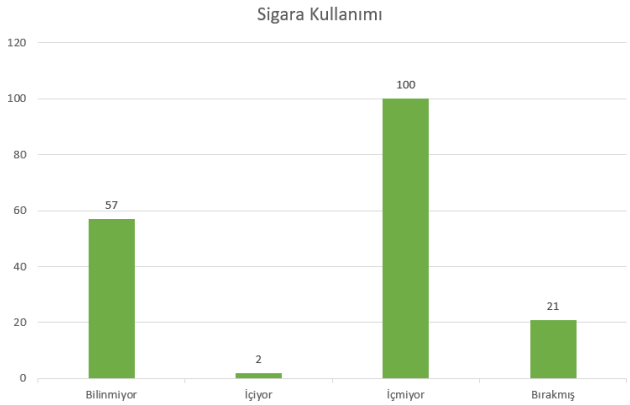
Sağlık çalışanları tarafından tek tek excel dosyalarına girilen hastaya ait tıbbi bilgiler toplanarak tek bir dosya haline getirilmiştir. Tekrar eden kayıtlar tespit edilerek veri setinden kaldırılmıştır ve her hastaya ait tek bir kayıt bulunacak şekilde nihai bir tıbbi bilgi veri seti elde edilmiştir.

2.1.3. Nihai veri seti

508 hasta kaydından oluşan laboratuvar veri seti ile 517 adet hasta kaydına ait tıbbi bilgileri içeren tıbbi bilgi veri seti T.C. kimlik numarası alanı kullanılarak birleştirilmiştir ve nihai veri seti elde edilmiştir. 508 hasta içerisinden 180 hastanın WHO giriş ve WHO en kötü değerleri bulunmaktadır bu yüzden kalan hastalar veri setinden çıkartılmıştır ve analize dahil edilmemiştir. Veri setindeki hastaların ortanca yaş değeri (medyan) 55 ve yaş ortalaması yaklaşık 56,01'dir. Hastaların 85 (47,2) tanesi kadın iken geri kalan 95 (52,8) hastanın cinsiyeti erkektir.

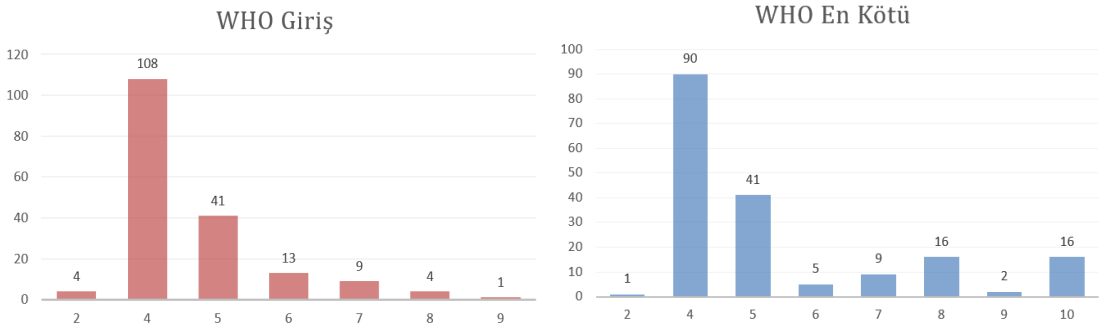
Şekil 2.3 verisetinde bulunan 180 adet hastanın tıbbi geçmişini ve sigara içme durumunu göstermektedir. Eştanı (komorbidite), belirli zaman diliminde birden fazla hastalığın birlikte bulunmasıdır. Kardiyovasküler hastalıklar kapsamında KAH, SVO, periferik VH ve immün yetmezlik kapsamında HIV, kanser, otoimmün gibi hastalıklar bulunmaktadır. Olguların %67,2'sinde en sık hastaneye yatış nedeni (%45,0), ardından ileri yaş (%37,7), dispne/hipoksi (%35,5) ve şiddetli pnömoninin radyolojik kanıtı (%34,4) olan komorbiditeler olarak belirlenmiştir.





Şekil 2.3. 180 adet hastanın tıbbi geçmiş ve sigara içme durumu

180 adet hastaya ait ilk laboratuvar sonucu ile hastaya ait WHO giriş-en kötü değeri birleştirilerek nihai veri seti oluşturulmuştur. Analizlerin amacına göre WHO giriş veya WHO en kötü değerleri veri setinde hedef değer olarak kullanılmıştır. Şekil 2.4 180 hastanın WHO giriş-en kötü değerlerinin dağılımını göstermektedir. WHO giriş için 10 farklı değerden yedi tanesine sahip hastalar bulunmaktadır ve en yüksek hasta sayısı dört değerinde yer almaktadır. WHO en kötü için ise 10 farklı değerden sekiz tanesine sahip hastalar bulunmaktadır ve en yüksek hasta sayısı dört değerinde yer almaktadır. İlk hastaneye başvuru sırasında hastaların %62,2'si oksijen tedavisi gerektirmezken, diğerleri ya non-invaziv ya da invaziv mekanik ventilasyon tedavisi almıştır. Yatış sırasında 32 hastada (%17,7) mekanik ventilasyon ihtiyacı gelişmiştir ve 16 hastada (%8,9) hayatını kaybetmiştir.



Şekil 2.4. 180 adet hastaya ait WHO giriş-çıkış dağılımı

180 adet hastadan oluşan nihai veri seti üç farklı analizde de veri setinde aşağıdaki işlemler yapılarak farklı şekillerde kullanılmıştır.

- WHO giriş hedef olarak kullanılarak yapılan ve hastaneye yatış aşamasındaki oksijen ihtiyacının tespit edilmesi hedeflenen analizde hastaların hepsi analize

dahil edilmiştir.

- WHO en kötü hedef olarak kullanılarak yapılan ve hastanede yatış süreci boyunca oksijen ihtiyacının tespit edilmesi hedeflenen analizde zaten hastaneye yatış aşamasında oksijen tedavisine ihtiyaç duyan hastalar WHO giriş değerine göre veri setinden çıkartılmıştır. 180 adet hastadan 68 tanesi çıkartılarak geriye kalan 112 hasta analizde kullanılmıştır.
- WHO en kötü hedef olarak kullanılarak yapılan ve hastanede yatış süreci boyunca yoğun bakıma ihtiyacın tespit edilmesi hedeflenen analizde zaten hastaneye yatış aşamasında yoğun bakıma ihtiyaç duyan hastalar WHO giriş değerine göre veri setinden çıkartılmıştır. 180 adet hastadan 14 tanesi çıkartılarak geriye kalan 166 hasta analizde kullanılmıştır.

Tablo 2.2 her bir analizde kullanılan veri setinin dağılımı temsil etmektedir.

Tablo 2.2. 3 analizde kullanılan veri seti bilgileri

Analiz	Veri seti Etiketi	Veri seti dağılımı (Sınıf 0 – Sınıf 1)	Toplam hasta sayısı
Analiz - 1	WHO giriş oksijen ihtiyacı tahmini	112 - 68	180
Analiz - 2	WHO en kötü oksijen ihtiyacı tahmini	91 - 21	112
Analiz - 3	WHO en kötü yoğun bakım ihtiyacı tahmini	137 - 29	166

Veri setinde eksik veriler bulunduğundan dolayı çalışmadaki ikinci aşama veri setinin temizlenmesi ve eksik verilerin doldurulmasıdır.

2.2. Veri Önişleme

Veri seti gerçek dünya verilerinden oluşmaktadır ve benzersiz bir veri seti olduğundan dolayı eksik veriler içermektedir. Toplanan verilerdeki eksik, gürültülü ve tutarsız veriler üzerinde ön işleme adımları gerçekleştirilerek bir sonraki adım olan sınıflandırma için daha kaliteli bir veri seti elde edilebilir ve sınıflandırmanın başarısı artırılır. Veri madenciliği işlemlerinde uygulanan eforun yaklaşık %80'i veri kalitesinin artırılmasına

harcanmaktadır [50].

Veri ön işleme aşaması veri temizleme, veri entegrasyonu, veri seçimi, veri dönüştürme vb. gibi etkileşimli adımlardan oluşur [51].

Tez kapsamında veri seti içerisinde yer alan eksik veriler gün aralıklarına göre ve ortalama değer atama yöntemleri kullanılarak tamamlanmıştır. İlk olarak özellikler için belirlenen gün aralıklarına göre eksik veri doldurma işlemi yapıldıktan sonra eksik veriler içeren özellikler ortalama değer yöntemi kullanılarak tamamlanmıştır.

2.2.1. Veri temizleme

2.2.1.1. Gün aralıklarına göre eksik veri doldurma

Veri setinde bulunan 191 özellikten 185 tanesi laboratuvar sonuçlarını oluşturmaktadır. Bu özelliklerde eksik veriler bulunduğundan dolayı sınıflandırma başarısını arttırmak için iki farklı eksik veri tamamlama yöntemi kullanılmıştır. Alanında uzman sağlık çalışanlarından alınan fikirler doğrultusunda uygulanan gün aralıklarına göre eksik veri doldurma tekniği veri setine ilk olarak uygulanan veri doldurma tekniğidir. Belirlenen özellikler için gün aralık değerleri sağlık çalışanları tarafından belirlenerek ön işleme aşamasında kullanılmıştır.

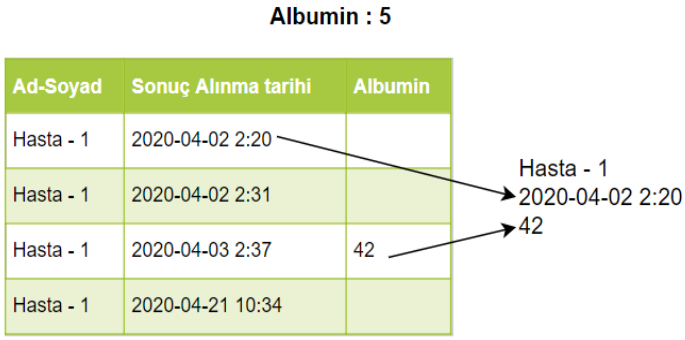
Bu eksik veri doldurma tekniği kapsamında sağlık çalışanları yardımı ile 185 laboratuvar sonucundan 75 tanesi için gün aralıkları tanımlanmıştır. Kalan 110 özellik ise alt laboratuvar sonuçları olduğundan veya bir dönem bakıldığından dolayı çok eksik veri içerdiği için uzman görüşü de dikkate alınarak veri setinden çıkartılarak analizlere dahil edilmemiştir. Tablo 2.3'te 75 özellik için belirlenen gün aralığı değerleri görülmektedir.

Tablo 2.3. Veri setindeki 75 özellik için belirlenen gün aralığı değerleri

Özellik	Gün aralığı	Özellik	Gün aralığı
Albumin	5	Fosfor	5
ALP	5	GGT	5
ALT	2	Glukoz	1
aPTT	2	HCO3-	48 saat
AST	2	HCT	1

BAS#	1	Hematokrit	1
BAS%	1	Hemoglobin	1
Baz Açığı	1	HGB	1
Bilirubin	2-3	INR	2
LYMPH%	1	PDW	3
Magnezyum	3	pH	1
MCH	1	PLT	1
MCHC	1	pO2	1
MCV	1	Potasyum	3
MDW	1	Prokalsitonin	1
Methemoglobin	2	PT	3
MON#	1	PT Zamanı %	3
MON%	1	RBC	2

Laboratuvar sonuçlarından oluşan veri setinde her hastaya ait bir veya birden daha fazla laboratuvar sonucu bulunmaktadır. Analizlerde sadece hastanın ilk laboratuvar sonucu kullanıldığından dolayı hastanın diğer laboratuvar sonuçları sağlık çalışanları tarafından belirlenen gün aralıkları ile birlikte kullanılarak hastanın ilk laboratuvar sonucundaki eksik veri oranının düşürülmesi hedeflenmiştir. Hastanın ilk laboratuvar sonucundaki x özelliği eksik ise hastanın diğer laboratuvar sonuçları sırayla kontrol edilmiştir. Hastanın ilk alınan sonucundan sonraki sonuçların tarihi ile ilk laboratuvar sonucu arasındaki süre gün aralığı değerinden az ve uygun bir değer var ise bu koşulu sağlayan laboratuvar sonucundaki değer kullanılarak ilk laboratuvar sonucundaki eksik veriler doldurulmuştur. Şekil 2.5 gün aralıklarına göre eksik veri doldurma işlemini göstermektedir.



Şekil 2.5. Gün aralıklarına göre eksik veri doldurma örneği

2.2.1.2. Ortalama değer atama yöntemi kullanılarak eksik veri doldurma

Ortalama değer atama yöntemi, her eksik değeri özniteliğin ait olduğu sınıfın ortalamasıyla değiştirir. Yalnızca sayısal nitelikler için ve örneklerin önceden sınıflandırıldığı, verinin sınıflara ayrıldığı veri setlerinde kullanılabilir [52]. Denklem 2.1'de görüldüğü gibi nitelikteki eksik veri, verinin sahip olduğu sınıf ile aynı değere ait olan ve boş olmayan değerlerin toplamının veri sayısına oranı kullanılarak doldurulur.

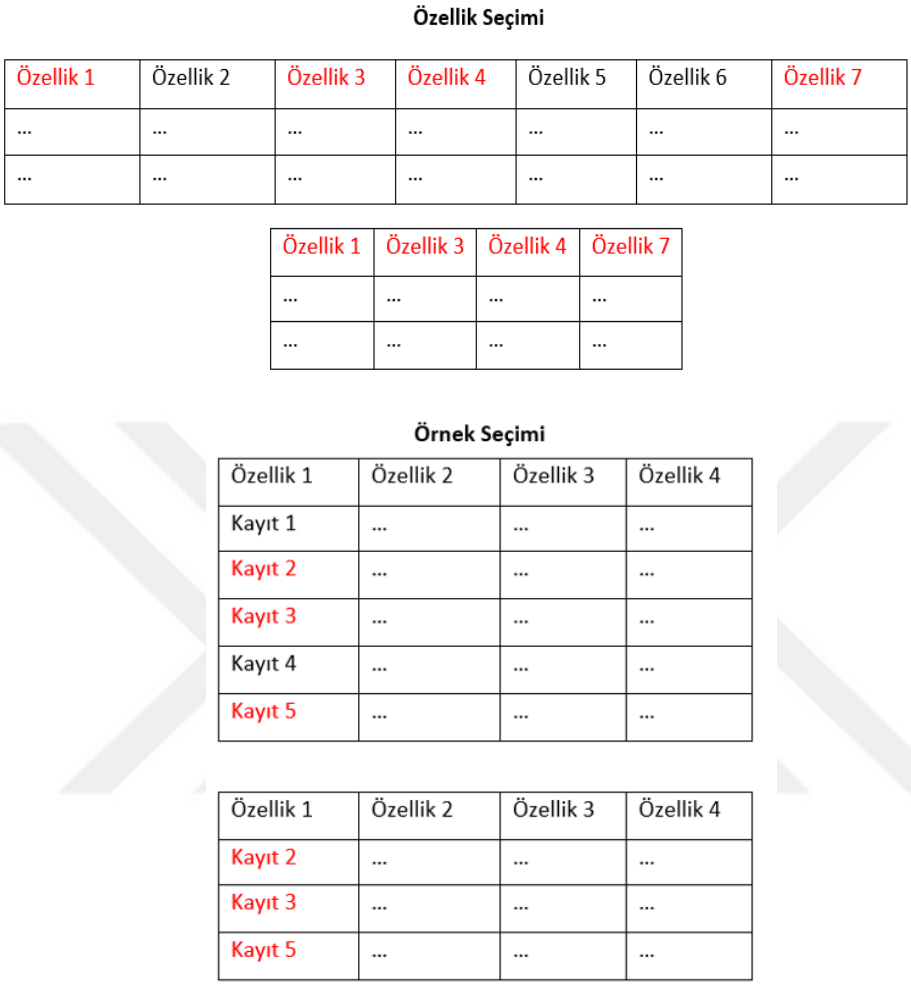
$$\text{Ortalama değer} = \frac{\text{(Eksik veri ile aynı sınıfa ait değerlerin toplamı)}}{\text{(Veri sayısı)}} \tag{2.1}$$

Gün aralıklarına göre eksik veri doldurma işleminden sonra veri setinde kalan eksik değerler her analizden önce belirlenen hedef değere göre ortalama değer atama yöntemi kullanılarak doldurulmuştur.

2.2.2. Veri indirgeme

Veri indirgeme, orijinal veri kümesinin boyutu azaltılmış bir halinin elde edilmesini sağlayan veri ön işleme tekniğidir. Veri azaltma işlemi sonrasında orijinal veri setindeki temel yapı ve bütünlük korunur fakat boyutu azaltılır. Boyut azaltma işlemi özellik seçimi veya örnek seçimi olarak ikiye ayrılır [53]. Şekil 2.6'da özellik seçim ve örnek seçim işlemi görülmektedir. Örnek seçim işleminde veri setindeki nitelik (kolon) sayısı değişmeden yatay olarak bir bölüm işlemi yapılarak daha az sayıda kayıt (satır) kullanılırken, özellik seçim işleminde veri setindeki kayıt (satır) sayısı değişmeden dikey olarak bir bölüm işlemi yapılır ve daha az sayıda nitelik (kolon) kullanılır. Çalışma kapsamında örnek seçimi yapılmamıştır ve tüm kayıtlar analizlere dahil edilmiştir.

Rastgele Özellik Seçme algoritması (Random Subset Feature Selection) kullanılarak özellik seçimi yapılmıştır ve veri kümesinin boyutu dikey olarak azaltılmıştır [54].



Şekil 2.6. Örnek seçimi ve özellik seçimi

2.2.2.1. Temel bileşenler analizi (PCA)

Temel bileşen analizinin ana fikri veri kümesinin sahip olduğu varyasyonu korurken veri fazlalığını azaltmaktır. Verilerin daha iyi ve kolay bir şekilde işlenmesi, depolanması ve görselleştirilmesi için orijinal veri kümesine göre daha az sayıda boyut kullanılır [55,56]. Bu yöntemde orijinal verilerin sahip olduğu özellikler mümkün olduğunca korunarak veriler maksimum varyansı sağlayacak şekilde oluşturulur. Tez kapsamında veri seti üzerindeki görselleştirme işlemleri PCA algoritması kullanılarak yapılmıştır ve veri 3 boyutlu olarak grafik üzerinde gösterilmiştir.

Ana bileşenlerin ortogonal olması, indirgemeden sonra oluşan verilerin birbiri üzerinde etkisinin olmaması ve hesaplanmasının kolay olması PCA algoritmasının avantajları arasında yer almaktadır. Fakat seçilen temel bileşenler haricindeki temel olmayan bileşenlerin kaybı, sonraki veri işlemeyi etkileyebilir [57].

Temel bileşen analizinin adımları aşağıdaki gibidir [58]:

1. Normalleştirme işlemi yapılarak ortalaması olan bir veri seti elde edilir. Normalleştirme işlemi her bir değerin ortalamadan çıkarılarak standart sapmaya bölünmesi ile yapılır. Normalleştirilmiş değer Denklem 2.2 kullanılarak elde edilir.

$$z = \frac{\text{value} - \mu}{\sigma} \tag{2.2}$$

2. Korelasyonları belirlemek için kovaryans matrisi hesaplanır.
3. Temel bileşenleri belirlemek için kovaryans matrisinin özvektörleri ve özdeğerleri hesaplanır.
4. Hangi temel bileşenlerin tutulacağına karar vermek için bir özellik vektörü oluşturulur.
5. Veriler ana bileşenler eksenleri boyunca yeniden düzenlenir. Yeni veri setindeki değerler Denklem 2.3 kullanılarak hesaplanır.

$$\text{Yeni Veri Seti} = (\text{Özellik Vektörü})^T * (\text{Normalleştirilmiş Veri Seti})^T \tag{2.3}$$

2.2.3. Veri dönüştürme

Veri setinde bulunan nitelikler farklı birim sistemlerinden oluştuğundan dolayı farklı standart sapma ve ortalama değerlerine sahiptir. Sayısal fark birçok niteliğin arasında bazı niteliklerin önemli görünmesine sebep olabilir [59]. Bu veri setleri tipik olarak K-En yakın komşu (KNN) gibi mesafe karşılaştırmasına dayalı modellerde veya nöron ağlardaki gradyan iniş gibi yöntemleri olumsuz anlamda etkileyerek sorun oluşturabilir. Bahsedilen bu sorunları çözmek için normalizasyon yöntemleri kullanılmıştır [50]. Min-max normalizasyon ve z-skor normalizasyon yöntemleri örnek olarak verilebilir.

2.2.3.1. Min-max normalizasyon

Verileri belirli bir minimum ve maksimum ile sınırlandırarak dönüştürme işlemidir. Genellikle [0-1] aralığı veya [-1,1] aralığı tercih edilebilmektedir. Min-max normalizasyonda veri seti içerisindeki sütundaki en büyük değer maksimum (x_{max}) ve en düşük değer minimum (x_{min}) kabul edilir. Aralıkta kalan değerler Denklem 2.4'teki gibi hesaplanır [53].

$$X_{\text{yeni}} = \frac{x - x_{\text{min}}}{x_{\text{min}} - x_{\text{maks}}} \tag{2.4}$$

Min-max normalizasyonu ile orijinal verileri belli sayısal bir aralığa çekmek veri madenciliğinde veri gizliliği için de tercih edilen bir yöntemdir [59]. R programlama dilinde verilerin min-max normalizasyonu için rescale() fonksiyonu kullanılır

2.2.3.2. Z-skor normalizasyon

Verisetindeki özelliğin aralığı tam olarak bilinemiyorsa veya aykırı değerler içeriyorsa min-max normalizasyon yöntemi etkilenebilir ve uygun değildir [53]. Z-skor normalizasyon yönteminde veriler ortalaması 0 ve standart sapması 1 olacak şekilde dönüştürülür. Denklem 2.5'te ortalama (μ) ve standart sapması (σ) değerleri kullanılarak veri dönüşüm işlemi gerçekleştirilebilir [60].

$$z(t) = \frac{v - \mu}{\sigma} \tag{2.5}$$

Ortalama ve standart sapma değerleri sadece Gauss dağılımı için en uygun ölçek parametresidir. Giriş skoru Gauss dağıtılmamışsa çıkış skoru giriş dağılımını korumaz [61].

2.3. Özellik Seçimi (Feature Selection)

Özellik alt küme seçimi, veri madenciliği sürecinde çok önemli bir rol oynar [62-64]. Bu çalışmada, RSFS (Random Subset Feature Selection) algoritması kullanılarak veri setindeki öznitelik sayısı azaltılmıştır [54]. RSFS algoritmasının yanı sıra çeşitli temel öznitelik seçim algoritmaları da bulunmaktadır. Örnek olarak CfsSubsetEval (CSE), Information Gain Attribute Assessment (IG), Correlation Attribute Eval (CA) gibi

algoritmalar gösterilebilir. Bu algoritmalar, GreedyStepWise, BestFirst ve Ranker gibi farklı yöntemler kullanarak özellik seçim işlemini gerçekleştirir. BestFirst, geri izleme olanağıyla zenginleştirilmiş açgözlü tepe tırmanışı ile özelliklerin alt kümelerini araştırır. GreedyStepwise, öznitelik alt kümelerinde açgözlü geriye veya ileriye doğru arama yapar. Ranker, nitelikleri bireysel değerlendirmelerine göre sıralar. RSFS algoritmasının bulduğu önemli öznitelikleri karşılaştırmak için analizde kullanılan veri setlerine farklı öznitelik seçim algoritmaları uygulanarak RSFS sonucunda elde edilen öznitelikler ile karşılaştırılmıştır. Bu sayede daha az özellik kullanılarak daha yüksek doğruluk değerleri elde edilmiştir. RSFS algoritması kullanılarak seçilen özelliklerin katkısı ve sıralaması Korelasyon Öznitelik Değerlendirmesi (CA) algoritması ve Ranker yöntemi ile değerlendirilmiştir. Korelasyon Öznitelik Değerlendirmesi bir niteliğin değerini, onunla sınıf arasındaki korelasyonu (Pearson's) ölçerek değerlendirir.

RSFS algoritması üç ana başlıktan oluşmaktadır. Bunlar; önişleme, rastgele altküme öznitelik seçimi ve sınıflandırma.

Ön işleme; Veri işlemeden önce tüm öznitelikler 0 ortalama ve birim standart sapma olacak şekilde normalize edilir.

RSFS; Özellik seçim işlemi yinelemeli olarak yapılır. Rastgele seçilen özellik alt kümeleri, KNN sınıflandırıcısı tarafından sınıflandırılır. Her adımda, rastgele seçilen özellikler, uygunluk değerlerine göre derecelendirilir. Her alt kümede, toplam öznitelik sayısının (özellik sayısı) karekökü (√) olarak rastgele seçilen öznitelikler vardır.

Sınıflandırma: Eğitim veri kümesi, rastgele oluşturulmuş alt kümedeki öznitelikler kullanılarak KNN sınıflandırıcısı ile sınıflandırılır. Rastgele oluşturulmuş her bir alt kümenin alaka düzeyi, P.C (performans kriteri) ile E.C beklenen kriteri arasındaki fark olarak hesaplanır. E.C, tüm yinelemelerdeki özgüllük değerlerinin ortalamasıdır. Denklem 2.6 kullanılarak EC değeri hesaplanır.

$$\mathrm{E(c)} = \frac{\sum(\text{Doğru sınıflandırılanlar})}{(\text{Doğru sınıflandırılanlar})+(\text{Yanlış sınıflandırılanlar})} x \frac{100}{\mathrm{n}} \tag{2.6}$$

P.C, mevcut yinelemeler için özgüllük değerinin ortalamasıdır. İlk yinelemede, P.C ve E.C aynıdır. Bu değerlerle her bir özelliğin alaka düzeyi denklem 2.7'deki gibi hesaplanır.

$$\text{Uygunluk(r)}=\big(\text{P.C(performans kriteri)-E.C(beklenen kriter)}\big) \tag{2.7}$$

Her yinelemede, doğru sınıflandırılmış özelliğin uygunluğu denklem 2.8'deki gibi güncellenir.

$$r_p^{'}=r_p+\left((P.C)_j\text{-}E.C\right) \tag{2.8}$$

r_p: (j-1)'inci yinelemede özellik vektörünün alaka düzeyi

r_p': özellik vektörünün güncellenmiş geçerli alaka düzeyi değerleridir

$(P.C)_j$: j. yineleme için P.C işlevinin değeridir

E.C: Beklenen kriter

Uygunluk matrisi, normal kümülatif fonksiyon (normcdf()) kullanılarak olasılık matrisine dönüştürülür. Bu dönüşüm, her bir veri kümesi türü için ortak bir anlayış oluşturulmasına yardımcı olur.

2.4. Dengesiz Veri Setlerinde Aşırı Örnekleme (Oversampling)

Veri setindeki sınıf oranı dengesiz ve bir sınıf diğerine göre çoğunluk (majority) konumunda ise sınıflandırıcı çoğunlukla baskın olan sınıf üzerinde tahminde bulunur. Örnek olarak sınıflandırıcı 99 adet pozitif kaydın ve 1 adet negatif kaydın bulunduğu veri kümesinde tüm kayıtları pozitif olarak tahmin ederse bu modelin başarısı %99 olacaktır. Fakat negatif sınıfın tahmin edilme başarısı sıfırdır bu durum makine öğrenmesinde dengesiz sınıf dağılım problemini göstermektedir. Dengesiz sınıf dağılımının bulunduğu veri kümelerinde sınıflandırıcı azınlık (minority) olan sınıfı çoğunluk (majority) olan sınıfa göre yüksek bir performans ile tahmin edemez. Sınıflandırma modeli çoğunluk olan sınıfa daha fazla eğilim gösterir. Bu da azınlık olan sınıfın büyük oranda hata ile tahmin edilmesi anlamına gelmektedir. Bu da modelin yüksek bir başarıya ulaşmasına rağmen azınlık olan sınıfı iyi bir şekilde tahmin edemediğinden dolayı kullanışlı ve faydalı bir model olmadığını göstermektedir [65].

Bu tez kapsamında da uygulanan üç analizde dengesiz bir veri kümesi oluştuğundan dolayı bu analizler için yüksek bir sınıflandırma doğruluğu elde edilmesine rağmen azınlık olan sınıf daha düşük bir sınıflandırma performansı ile tahmin edilmiştir. Bu yüzden veri seti üzerinde SMOTE algoritması uygulanarak daha dengeli bir veri seti elde edilmiştir ve sonuçlar SMOTE uygulanmadan önceki veri seti ile oluşturulan model ile karşılaştırılmıştır.

2.4.1. SMOTE (Synthetic Minority Oversampling Technique)

SMOTE algoritması dengesiz bir yapıya sahip olan veri kümelerindeki azınlık sınıfın sentetik bir şekilde arttırılmasını sağlayarak dengeli bir veri seti oluşturan yaklaşımdır [66].

SMOTE algoritmasının çalışma adımları aşağıdaki gibidir [67]:

- Azınlık sınıfa ait bir örneğin Öklid mesafesine göre en yakın k adet komşusu bulunur
- En yakın k adet komşudan rastgele biri seçilir
- Özellik uzayında bir çizgi parçası oluşturmak için a ve b noktaları birleştirilerek sentetik örnek oluşturulur. Oluşturulan bu sentetik örnekler a ve b örneğinin dışbükey bir kombinasyonunu temsil eder.

Böylece SMOTE algoritması oluşturulan bu sentetik örnekler ile birlikte azınlık sınıfı çoğunluk sınıfa eşitleyerek dengeli bir veri kümesi elde edilmesini sağlar. Şekil 2.7 adımları açıklanan SMOTE algoritmasının çalışma prensibini göstermektedir.



Şekil 2.7. SMOTE algoritması

2.5. Makine Öğrenmesi

Karmaşık veri setlerinin işlenerek anlamlı ve yararlı bilgilerin çıkarılma süreci veri madenciliği olarak adlandırılır. Veri madenciliği süreci sınıflandırma, kümeleme ve birliktelik kuralları gibi yöntemler kullanılarak gerçekleştirilir [68].

Veri madenciliği yöntemleri farklı boyutlardaki veri kümelerini analiz etmek, kategorilere ayırmak, özetlemek ve aradaki ilişkileri belirlemek amacıyla

kullanılmaktadır [69].

Veri madenciliği temelde denetimli ve denetimsiz olmak üzere iki farklı yöntemden oluşur. Denetimsiz öğrenme tekniğinde etiketsiz, sınıflara ayrılmamış veri kümesi kullanılarak öğrenme işlemi gerçekleştirilir [70]. Kümeleme ve birliktelik kuralları olmak üzere iki farklı denetimsiz öğrenme tekniği vardır ve kümeleme en yaygın kullanılan denetimsiz öğrenme tekniğidir [71]. Kümeleme yöntemleri veri azaltma, biyomedikal alanlarda ve kanser sınıflandırma gibi alanlarda kullanılabilmektedir [72].

Denetimsiz yöntemlerin aksine denetimli öğrenme tekniğinde etiketlenmiş veri kümeleri kullanılarak eğitim işlemi yapılır ve istenen çıktı değerleri tahmin edilir [73]. Sınıflandırma problemlerinde denetimli öğrenme en yaygın olarak kullanılan yöntemdir [74]. Veri analiz tekniklerindeki hızlı gelişim ve sağlık sektöründe verinin kullanılabilirliğindeki artış sağlık hizmetleri ile birlikte Yapay Zeka (AI) uygulamalarının kullanımını iyileştirmiştir [75]. Sağlık hizmetlerinde ilaç geliştirme, tedavi oluşturulması ve hastalık tahmini gibi birçok alanda yapay zeka uygulamaları bulunmaktadır [76]. Bu çalışmada da laboratuvar sonuçları ve demografik veriler kullanılarak Covid-19'un tahmini aşamasında denetimli öğrenme teknikleri kullanılmıştır. Şekil 2.8 veri madenciliği alanında kullanılan denetimli ve denetimsiz öğrenme yöntemlerini [77] göstermektedir.



Şekil 2.8. Makine öğrenmesi yöntemleri

2.5.1. Sınıflandırma

Denetimli bir öğrenme tekniği olan sınıflandırma, veri setindeki sınıf etiketi atanmış örnekleri kullanarak öğrenme işlemini yapan ve test verisinin ait olduğu sınıf etiketini tahmin eden bir makine öğrenmesi tekniğidir. Bu çalışmada denetimli sınıflandırma yöntemlerinden K-En Yakın Komşu, Bagging, Rastgele Orman ve Karar Ağacı makine öğrenmesi yöntemleri kullanılmak üzere seçilmiştir.

2.5.1.1. K-En yakın komşu (KNN)

KNN sınıflandırma algoritması basit ve temel sınıflandırma yöntemlerindendir. Parametrik olmayan bir yöntem olan KNN algoritması regresyon ve sınıflandırma için kullanılır. Algoritma çıktısı olarak bir sınıf etiketi üretilir. Sınıflandırma işlemi çoğunluk oylaması kullanılarak hesaplanır. Tahmin edilen nesne en yakın k komşusu arasında en fazla olan sınıfa atanır. Temelde iki değer kullanılarak tahmin işlemi gerçekleştirilir. Bunlar mesafe ve kullanıcı tanımlı k değeridir. K değeri pozitif bir tamsayıdır. Uzaklık değeri farklı uzaklık ölçme formülleri kullanılarak hesaplanabilir. Bu çalışmada Öklid Uzaklığı (Euclidean Distance) formülü mesafeyi ölçmek için kullanılmıştır. Öklid uzaklığı noktalar arasındaki doğrusal uzaklığı hesaplanan için kullanılan bir uzaklık hesaplamasıdır ve denklem 2.9'da verilmiştir.

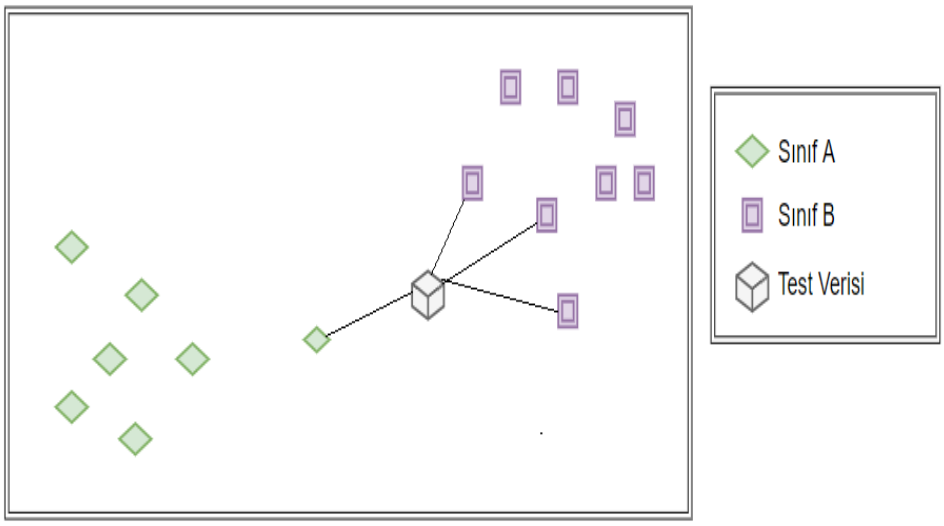
d(a, b) = a ve b noktanın aralarındaki uzaklığı,

a_i= a noktasının i. değişken değerini,

b_i= b noktasının i. değişken değerini belirtir.

$$d(a, b) = \sqrt{\sum_{i=1}^{n}(ai - bi)2} \tag{2.9}$$

KNN algoritması diğer makine öğrenme algoritmalarının arasında en basit makine öğrenmesi algoritmalarından birisidir. Şekil 2.9 K=4 değeri için örnek bir KNN sınıflandırma işlemini temsil etmektedir. Çalışma kapsamında yapılan denemeler sonucunda en yüksek doğruluk değerine ulaşan k değeri tespit edilerek analizler yapılmıştır ve sonuçlar diğer algoritmaların sınıflandırma doğrulukları ile karşılaştırılmıştır.



Şekil 2.9. KNN algoritması (K=4)

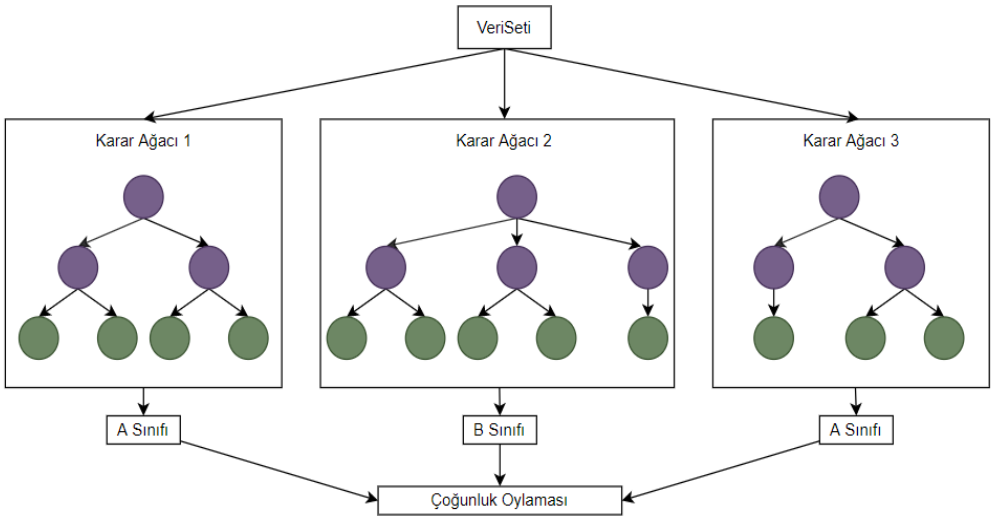
2.5.1.2. Bagging

Bootstrap toplama (Bootstrap aggregating) olarak da bilinen Torbalama (Bagging) sınıflandırıcısı ilk olarak Breiman tarafından 1996 yılında tanıtılmıştır. Torbalama sınıflandırıcısı nesnelerin örneklerini alır her bir örnek üzerinde bir sınıflandırıcı geliştirir. Çoğunluk oylama tekniği kullanılarak sınıflandırıcı oyları birleştirilir. Torbalama tekniği sayesinde zayıf sınıflandırıcılar kullanılarak optimalliğe doğru daha güçlü bir sınıflandırıcı geliştirilebilir. Torbalama yöntemi sosyal ve biyolojik birçok alanda tahmin ve sınıflandırma amaçlarıyla kullanılır [78,79].

Weka kütüphanesinde Torbalama yöntemi kapsamında DecisionStump, J48, LMT, RepTree, RandomTree olmak üzere farklı yöntemler mevcuttur. Tez çalışması kapsamında tüm bu yöntemlerin doğruluk performansları kendi aralarında karşılaştırılmış ve en yüksek sınıflandırma doğruluğuna ulaşan RandomTree algoritması kullanılarak Bagging modelleri oluşturulmuştur. Rastgele Ağaç (RandomTree), stokastik bir süreçle oluşturulan her düğümde rastgele seçilen K özniteliklerini dikkate alarak budama yapmadan bir ağaç oluşturur. Ek olarak, ağaçlardaki her parametre, örneğin yaprak sayısı veya çap gibi rastgele bir değişkendir. RandomTree ayrıca sınıf olasılıklarının tahminine izin verme seçeneğine veya bir gerileme setine dayalı olarak regresyon durumunda hedef ortalamaya izin verme seçeneğine sahiptir [80].

2.5.1.3. Rastgele orman

Rastgele Orman algoritması toplulukla öğrenme yöntemlerinden torbalama tekniğine örnek olarak verilebilir. Torbalama algoritması rastgele olarak oluşturulan verisetleri üzerinde paralel olarak sınıflandırma analizi yapan bir makine öğrenmesi modelidir. Temelinde karar ağaçları yer alan bir makine öğrenmesi modeli olan Rastgele Orman algoritmasında veri üzerinde aşırı öğrenme probleminin üstesinden gelinmesini sağlayan bir yapı mevcuttur. Temel olarak karar ağaçlarının aksine birden fazla karar ağacı yapısı oluşturularak daha yüksek bir sınıflandırma değerine ulaşılması hedeflenmiştir. Rastgele Orman algoritması veriseti N adet kayıta ve M adet özelliğe sahip olduğunda her bir ağaç yapısı için N adet kayı seçilerek ve m<M olacak şekilde m adet özellik seçilerek yapılır. Her bir karar ağacı yapısı m özellikli N adet kayda sahip alt verisetini kullanarak bir karar ağacı modeli oluşturur. Böylece her bir karar ağacı farklı alt eğitim setlerini kullanarak model oluşturur. Test aşamasında ise verilen örnek her bir ağaç tarafından tek tek değerlendirilir ve nihai sınıflandırma çoğunluk oylaması kullanılarak belirlenir. Böylece birkaç zayıf karar ağacı sınıflandırıcısı birleştirilerek güçlü bir sınıflandırıcı elde edilmiş olur. Şekil 2.10 Rastgele Orman algoritmasının çalışma adımları gösterilmektedir. [81,82]



Şekil 2.10. Rastgele Orman Algoritması

2.5.1.4. Karar ağacı

İlk olarak 1960 yılında tanıtılan karar ağacı algoritması yaygın olarak kullanılan bir makine öğrenmesi yöntemidir. Yöntem verisetindeki bilgileri kullanarak bir kök düğüm, iç düğümler ve yaprak düğümlerden oluşan bir ağaç modeli oluşturmaktadır. Algoritma parametrik değildir ve karmaşık verisetlerinde parametre kullanımı olmadan verimli sonuçlar elde edebilir. Hem ayrık değişkenler hem de sürekli değişkenler kullanılarak sınıflandırma işlemi gerçekleştirilebilir. CART, C4.5 ve CHAID sıklıkla kullanılan karar ağacı algoritmalarına örnek olarak verilebilir. [83]

Çalışmada C4.5 karar ağacı algoritması kullanılmıştır. Bu algoritmanın Weka'daki karşılığı J48'dir. J48 algoritması azaltılmış hata budaması ve açgözlü yöntemi kullanarak karar ağacını oluşturur. Temelinde bilgi entropisi kavramını kullanır ve entropi farkını inceler bu entropi farkı bilgi kazancı olarak adlandırılır. Bilgi kazancı nitelik kullanılarak veri bölme işlemi gerçekleştirildiğinde hedef değerdeki belirsizliği gösteren bir değerdir. En yüksek bilgi kazancına sahip nitelik kullanılarak ağaç kırımları oluşturulur ve J48 hem sürekli hem de ayrık verilerde iyi bir şekilde çalışan bir kara ağacı yöntemidir. Ayrıca ağaç oluşturma işlemi tamamlandıktan sonra budama seçeneği de sunar. Temelinde işlem adımları aşağıdaki gibidir [84]:

- Örnekler aynı sınıfa ait ise bu düğüm sınıf etiketi belirlenmiş tek yaprağı ifade eder. Aksi durumda özniteliğe göre bölerek kara ağacı oluşturulur ve tüm örnekler aynı sınıfa ait ise bölme işlemi sonlandırılır.
- Her bir nitelik için bilgi kazancı değeri hesaplanır.
- En yüksek bilgi kazancına sahip nitelik kullanılarak ağaç kırılımları oluşturulur.

2.5.1.5. Sınıflandırma modellerinin değerlendirilmesi

Sınıflandırma işleminin performans değerlendirilmesi Gerçek Pozitifler (TP), Gerçek Negatifler (TN), Yanlış Pozitif (FP) ve Yanlış Negatifler (FN) istatistiksel ölçüm değerleri hesaplanarak elde edilir. Bu hesaplanan değerler kullanılarak Tablo 2.4'teki gibi Karmaşıklık Matrisi elde edilebilir. Karmaşıklık Matrisi ikili bir veri seti üzerindeki sınıflandırıcı için oluşturulabilen ve sınıflandırmanın performansını değerlendirmek için kullanılan bir tablodur [85].

Tablo 2.4. Karmaşıklık matrisi

	Tahmin Edilen: Negatif	Tahmin Edilen: Pozitif
Gerçek: Negatif	Doğru Negatif	Yanlış Pozitif
Gerçek: Pozitif	Yanlış Negatif	Doğru Pozitif

Matrisi oluşturan temel değerler aşağıdaki gibidir:

- Doğru Negatifler (DN): tahmin edilen ve gerçek değer negatifse
- Doğru Pozitifler (DP): tahmin edilen ve gerçek değer pozitifse
- Yanlış Negatifler (YN): tahmin edilen değer negatif ve gerçek değer pozitifse
- Yanlış Pozitifler (YP): tahmin edilen değer pozitif ve gerçek değer negatifse

Hassasiyet (Precision): Gerçek değeri doğru olan verilerden doğru olarak tahmin edilen değer sayısının doğru pozitif ve yanlış pozitif değerlerinin toplamına bölümü ile elde edilir [86]. Bu istatistiksel hesaplama pozitif olarak sınıflandırılan değerlerden kaç tanesinin gerçekten pozitif olduğunu gösterir. Denklem 2.10'daki gibi hesaplanır.

$$\text{Hassasiyet (Precission)} = \frac{DP}{DP+YP} \tag{2.10}$$

Duyarlılık (Recall): Gerçek değeri doğru olan verilerden doğru olarak tahmin edilen değer sayısının doğru pozitif ve yanlış negatif değerlerinin toplamına bölümü ile elde edilir [86]. Bu istatistiksel hesaplama pozitif olarak sınıflandırılması gereken değerlerden kaç tanesinin pozitif olarak sınıflandırıldığını gösterir. Denklem 2.11'deki gibi hesaplanır.

$$\text{Duyarlılık (Recall)} = \frac{DP}{DP+YN} \tag{2.11}$$

Doğruluk (Accuracy): Doğru olarak tahmin edilen değerlerin sayısının tüm veri setindeki değer sayısına oranı kullanılarak elde edilir [87].

Fakat doğruluk bir modelin performansını ölçmek için yeterli olmayacaktır. Bu tez çalışmasında da olduğu gibi Covid-19 yoğun bakım ihtiyacı veya oksijen ihtiyacı olan hastalar diğer sınıfa göre daha az sayıdadır. Genel doğruluk değeri sadece performans ölçütü olarak kullanıldığında genel doğruluk yüksek olmasına rağmen az sayıda fakat önemli olan yoğun bakım, oksijen ihtiyacı sınıflarının sonucunu değerlendirmemiş oluruz. Bu yüzden duyarlılık, hassasiyet ve f1-skor değerleri de bu tez çalışmasında

değerlendirme ölçütü olarak kullanılmıştır. Denklem 2.12'deki gibi hesaplanır.

$$\text{Doğruluk (Accuracy)} = \frac{DP+DN}{DP+YP+DN+YN} \tag{2.12}$$

F1 Skor: F-Skoru Denklem 2.13'teki gibi hassasiyet ve duyarlılık değerlerinin harmonik ortalaması alınarak hesaplanır [87]. Harmonik ortalama kullanılarak uç durumların gözardı edilmemesi sağlanır. Eşit dağılıma sahip olmayan veri setlerinde daha doğru bir model seçim yapılmasını sağlaması ve tüm hata metriklerini içeren bir hesaplama değeri olması F1-Skoru'nun doğruluğa yerine kullanılmasının temel sebeplerindendir.

$$\text{F1-Skor} = \frac{2 * \text{duyarlılık} * \text{hassasiyet}}{\text{duyarlılık} + \text{hassasiyet}} \tag{2.13}$$

Özgüllük (Specifity): Gerçek negatif oranı olarak da adlandırılan bu metrik kullanılarak gerçekte negatif olarak sınıflandırılan değerlerden kaç tanesinin negatif olarak sınıflandırıldığı hesaplanır. Denklem 2.14'deki gibi hesaplanır.

$$\text{Özgüllük (Specifity)} = \frac{DN}{DN+YP} \tag{2.14}$$

3. BULGULAR VE TARTIŞMA

Materyal ve yöntem başlığı altında ayrıntılı olarak açıklanan veri seti üzerinde önişleme, özellik seçme ve farklı sonlanım durumlarına göre sınıflandırma işlemleri yapılmıştır. Laboratuvar ve demografik özellikler kullanılarak hastanın oksijene ve yoğun bakıma ihtiyacı olmak üzere iki farklı sonlanım durumu tez çalışması kapsamında analiz edilmiştir.

Her analizden önce belirlenen hedef değer referans alınarak ortalamaya göre eksik veri doldurma tekniği kullanılmıştır veri setindeki eksik veriler tamamlanmıştır. Aşağıdaki koşullar temel alınarak oluşturulan sınıflandırma modelleri değerlendirilerek elde edilen sonuçlar karşılaştırılmıştır:

- Tüm özellikler kullanılarak ve Rastgele Özellik Seçme algoritmasının bulduğu daha az sayıdaki özellik ile analizler yapılarak özellik seçim işleminin sınıflandırma performansına etkisi
- Hem Z-Skor standardizasyon işlemi yapılmış veri setinde hem de standardizasyon

işleminden önce sınıflandırma işlemi yapılarak standardizasyonun sınıflandırma performansına etkisi değerlendirilmiştir.

Belirtilen bu iki temel değerlendirme koşulu KNN, Bagging, Rastgele Orman ve Karar Ağacı olmak üzere dört farklı makine öğrenmesi üzerinde uygulandığı zaman özellik seçim ve z-skor standardizasyon işleminin olup olmaması durumuna göre toplamda 16 farklı sınıflandırma modeli oluşturulmuş ve sonuçlar karşılaştırılmıştır. Tablo 3.1 her bir analiz için oluşturulan bu 16 modelin sahip olduğu özellikleri göstermektedir.

Tablo 3.1. Her analiz için oluşturulan 16 sınıflandırma modelinin özellikleri

Algoritma	Özellik Seçme (Var Yok)	Standardizasyon (Var/Yok)	Model
KNN	Tüm özellikler	Var	1
		Yok	2
	RSFS ile seçilen özellikler	Var	3
		Yok	4
Bagging	Tüm özellikler	Var	5
		Yok	6
	RSFS ile seçilen özellikler	Var	7
		Yok	8
Rastgele Orman	Tüm özellikler	Var	9
		Yok	10
	RSFS ile seçilen özellikler	Var	11
		Yok	12
Karar Ağacı	Tüm özellikler	Var	13
		Yok	14
	RSFS ile seçilen özellikler	Var	15
		Yok	16

Oluşturulan bu 16 farklı modelden sadece aşağıdaki belirtilen iki model seçilerek performans metrikleri bulgular başlığı altında detaylı olarak incelenmiştir.

- Tüm özellikler kullanılarak oluşturulan modellerden en yüksek sınıflandırma doğruluğuna sahip olan

- RSFS ile seçilen özellikler kullanılarak oluşturulan modellerden en yüksek sınıflandırma doğruluğuna sahip olan

3.1. Hastaneye İlk Yatış Aşamasında Oksijen İhtiyacının Tahmini

Bu analizde Tablo 2.1'de "WHO giriş oksijen ihtiyacı tahmini" olarak isimlendirilen veri seti kullanılmıştır. Veri seti her bir hastaya ait ilk laboratuvar sonucundan alınan 67 özellik ile yaş ve cinsiyet niteliklerinden oluşmaktadır. Ayrıca Marmara Tıp Fakültesi'nden Kardiyoloji ve Dahiliye Uzmanı Prof. Dr. Ali Serdar Fak, Marmara Üniversitesi Tıp Fakültesi Göğüs Hastalıkları Anabilim Dalından Prof.Dr. Sait Karakurt, Marmara Üniversitesi Tıp Fakültesi Enfeksiyon Hastalıkları Anabilim Dalından Prof.Dr. Lütfiye Mülazımoğlu, Marmara Üniversitesi Tıp Fakültesi Enfeksiyon Hastalıkları Anabilim Dalından Doç.Dr. Elif Tigen'in önerileri ile belirlenen özelliklerin birbirlerine oranından elde edilen üç özellik de veri setine eklenerek toplamda 72 özellik elde edilmiştir. Veri setinde toplam 180 adet hasta bulunmaktadır. Bu hastaların 95'i erkek, geri kalan 85'i ise kadındır. 180 adet hastanın yaş ortalaması yaklaşık 56,01'dir ve 18-96 yaş aralığında Covid-19 şüphesi ile hastaneye başvuran hastalar bulunmaktadır.

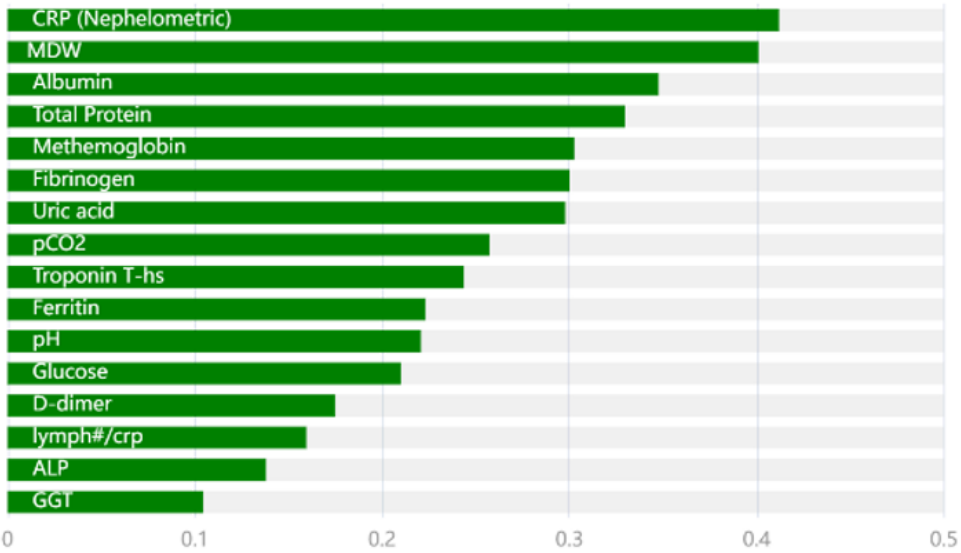
Bu analizde hastaneye ilk yatış aşamasındaki oksijen ihtiyacının tahmin edilmesi hedeflendiğinden dolayı hedef değer olarak WHO giriş değeri kullanılmıştır ve veri setindeki eksik nitelikler ortalamaya göre eksik veri doldurma tekniği kullanılarak doldurulmuştur.

Oksijen ihtiyacının tahmin edilmesi hedeflendiğinden dolayı sağlık çalışanlarının önerileri doğrultusunda WHO giriş limit değeri dört olarak belirlenmiştir. Böylece 180 adet hastadan WHO giriş değeri dört ve altında olan 112 hasta veri setinde "0" olarak etiketlenirken geri kalan 68 hasta ise "1" olarak etiketlenmiştir. Tüm sınıflandırma işlemleri KNN, Bagging, Rastgele Orman ve Karar Ağacı olmak üzere dört farklı makine öğrenmesi yöntemi kullanılarak Weka üzerinde yapılmıştır ve Tablo 3.1'de açıklanan 16 farklı model oluşturulmuştur. 72 özelliğin tümü kullanılarak elde edilen sınıflandırma sonuçları ile karşılaştırılmak üzere Rastgele Özellik Seçme algoritması ile seçilen özellikler üzerinde de analizler yapılmıştır. Tablo 3.2 analiz için kullanılan tüm özelliklerin bir listesini ve tüm özellikler içerisinden Rastgele Özellik Seçme algoritması kullanılarak seçilen 16 adet niteliği temsil etmektedir.

Tablo 3.2. Analiz – 1'de kullanılan 72 özellik ve RSFS algoritması ile seçilen 16 özellik

Özellikler				Seçilen Özellikler
Yaş	EOS#	Direkt Bilirubin	MCHC	CRP (Nefelometrik)
Cinsiyet	EOS%	Karboksihemoglobin	MCV	ALP
NEU%	Laktat	Kreatinin	PCT	Ferritin
Sodyum	PLT	Kütle CK-MB	MCH	Total Protein
ALT	Fosfor	Total Bilirubin	MON#	D-dimer
aPTT	WDOP	Baz Açığı	MON%	Troponin T-hs
AST	WBC	Prokalsitonin	MPV	Ürik asit
BAS#	HCO3-	Ozmolarite	NRBC#	Methemoglobin
BAS%	HCT	Magnezyum		GGT
LDH	HGB	plt/lymph#		Glukoz
BUN	PDW	neu#/lymph#		Fibrinojen
Klor	pO2	Kalsiyum		lymph#/crp
PT%	Üre	LYMPH#		Albumin
RBC	SO_2	LYMPH%		MDW
INR	NEU#	Potasyum		pCO_2
RDW	PT	NRBC		pH

Tabloda kırmızı bir şekilde belirtilen 16 özellik hastaneye yatış aşamasında hastanın oksijen tedavisine ihtiyaç duyup duymaması ile ilgili en önemli özellikler olarak bu tez kapsamında belirlenmiştir. Şekil 3.1 seçilen 16 özelliğin Korelasyon Öznitelik Algoritması tarafından belirlenen katkı değerlerini göstermektedir. Niteliklerin CA algoritması tarafından sıralaması, en önemli katkı sağlayan özelliklerin C-reaktif protein (CRP) (0,41) ve monosit dağılım genişliği (MDW) (0,40) olduğunu göstermiştir.



Şekil 3.1. Analiz – 1 seçilen 16 özelliğin CA algoritmasına göre katkıları

3.1.1. Analiz - 1 72 özellikle oluşturulan model sonuçları

Tablo 3.2'de belirtilen 72 özelliğin tamamı kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.3 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

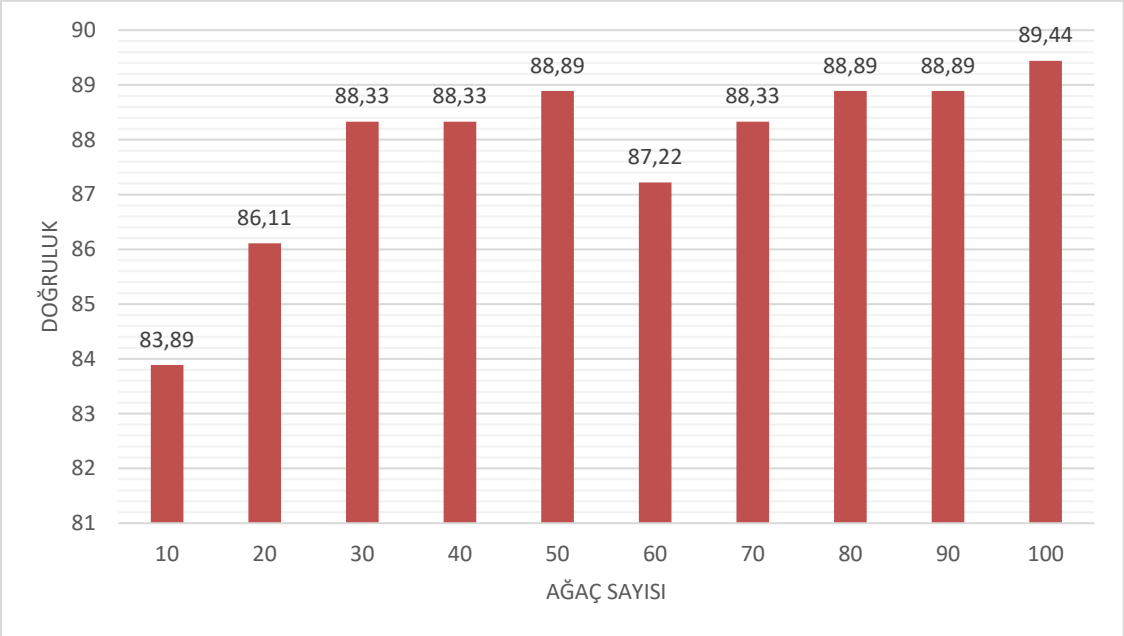
Tablo 3.3. Analiz – 1 72 özellik sınıflandırma performans sonuçları

Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	73,4	73,9	73,4	73,89
	Yok	73,4	73,9	73,4	73,89
Bagging	Var	86,8	86,7	86,4	86,67
	Yok	87,3	87,2	87	87,22
Rastgele Orman	**Var**	**89,5**	**89,4**	**89,3**	**89,44**
	Yok	**89,5**	**89,4**	**89,3**	**89,44**
Karar Ağacı	Var	84,9	85	84,9	85
	Yok	84,9	85	84,9	85

Sonuçlar detaylı olarak incelendiğinde en yüksek sınıflandırma doğruluğuna %89,44 ile Rastgele Orman algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman algoritmasının ardından Bagging ve Karar Ağacı da %85 ve üzeri bir sınıflandırma doğruluğuna ulaşırken KNN %73,89 genel doğruluk değeri ile diğer yöntemlere göre daha kötü bir sınıflandırma performansı göstermiştir. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan modeller üzerinde pozitif bir etki oluşturmadığı yani sınıflandırma performansına olumlu bir katkı sağlamadığı görülmektedir. Bu yüzden 72 özellik kullanılarak hastanın hastaneye yatış aşamasını en iyi tahmin eden yöntem standardizasyonun kullanılmadığı Rastgele Orman algoritması olarak seçilmiştir.

3.1.1.1. Analiz - 1 72 özellikli rastgele orman modeli

Weka üzerinde Rastgele Orman yöntemi için rastgele ormandaki ağaç sayısı, ağacın maksimum derinliği gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Ağacın maksimum derinliği sınırsız olarak belirlenerek farklı ağaç sayıları ile Rastgele Orman modelleri oluşturulmuştur. Şekil 3.2 Farklı ağaç sayıları kullanılarak elde edilen rastgele orman sınıflandırma doğruluk değerlerini temsil etmektedir. Standardizasyon işleminin uygulanmadığı veri setinde en yüksek sınıflandırma doğruluğuna %89,44 ile ağaç sayısının 100 olarak seçildiği modelde ulaşmıştır.



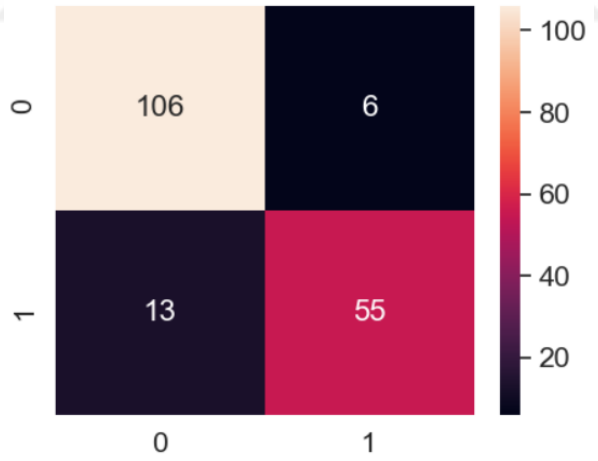
Şekil 3.2. Analiz – 1 72 özellik Rastgele Orman doğruluk değerleri

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.4 Rastgele Orman algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. "0" etiketli oksijen ihtiyacı olmayan hastalar için 0, 918 F1-skoru ile yüksek bir sınıflandırma modeli oluşturulurken, 0,853 ile 1 etiketli hastaneye yatışta oksijen ihtiyacı olan hastalar ortalama bir performans ile tahmin edilmiştir.

Tablo 3.4. Analiz – 1 72 özellikli Rastgele Orman Modeli Performans Metrikleri

Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,89	0,95	0,92	0,77	0,95	0,97
1	0,90	0,81	0,85	0,77	0,95	0,92

Şekil 3.3'te gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde de altı oksijen ihtiyacı olmayan hastanın ve 13 oksijen ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



Şekil 3.3. Analiz – 1 72 özellikli Rastgele Orman modeli karmaşıklık matrisi

3.1.2. Analiz - 1 16 özellikle oluşturulan model sonuçları

Tablo 3.2'de kırmızı olarak işaretlenen 16 özellik kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz

farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.5 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

Tablo 3.5. Analiz – 1 16 özellik sınıflandırma performans sonuçları

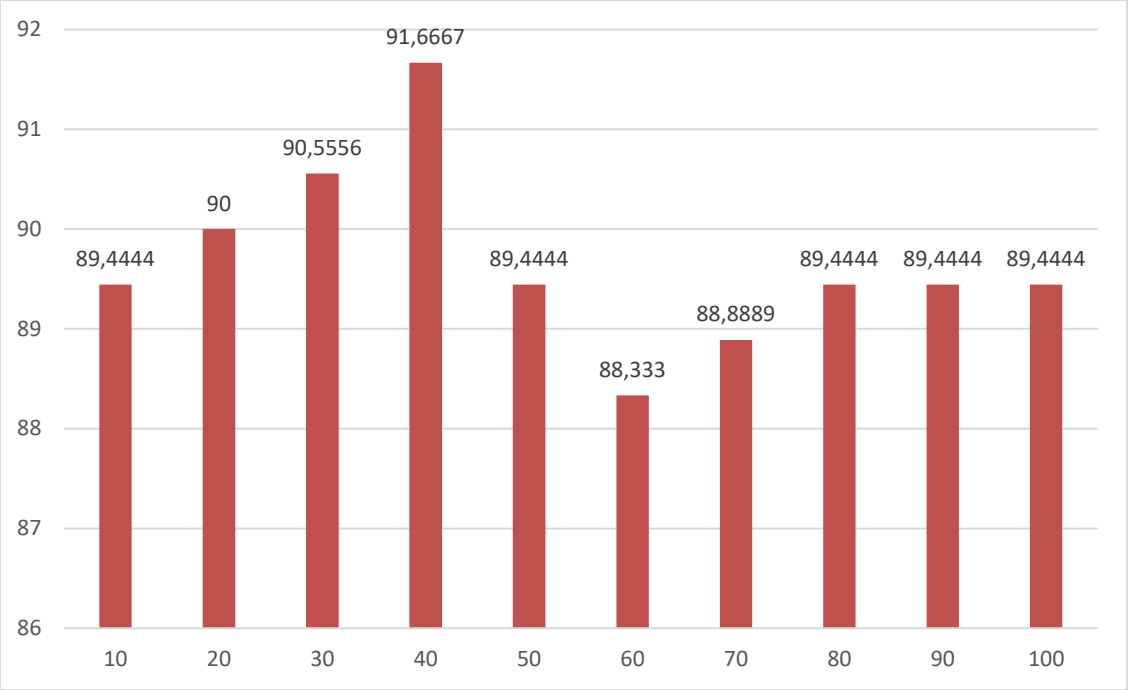
Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	82,2	82,2	81,8	82,22
	Yok	82,2	82,2	81,8	82,22
Bagging	Var	88,3	88,3	88,2	88,33
	Yok	88,3	88,3	88,3	88,33
Rastgele	Var	90,8	90,6	90,4	90,56
Orman	**Yok**	**91,6**	**91,7**	**91,7**	**91,67**
Karar Ağacı	Var	87,2	87,2	87,2	87,22
	Yok	87,2	87,2	87,2	87,22

Sonuçlar detaylı olarak incelendiğinde en yüksek sınıflandırma doğruluğuna %91,67 ile Rastgele Orman algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman algoritmasının ardından Bagging %87,22 ve Karar Ağacı da %88,33 sınıflandırma doğruluğuna ulaşırken KNN %82,22 genel doğruluk değeri ile diğer yöntemlere göre daha kötü bir sınıflandırma performansı göstermiştir. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan modeller üzerinde pozitif bir etki oluşturmadığı yani sınıflandırma performansına olumlu bir katkı sağlamadığı görülmektedir. Bu yüzden 16 özellik kullanılarak hastanın hastaneye yatış aşamasını en iyi tahmin eden yöntem standardizasyonun kullanılmadığı Rastgele Orman algoritması olarak seçilmiştir.

3.1.2.1. Analiz - 1 16 özellikli rastgele orman modeli

Weka üzerinde Rastgele Orman yöntemi için rastgele ormandaki ağaç sayısı, ağacın maksimum derinliği gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Ağacın maksimum derinliği sınırsız olarak belirlenerek farklı ağaç sayıları ile Rastgele Orman modelleri oluşturulmuştur. Şekil 3.4 Farklı ağaç sayıları kullanılarak elde edilen rastgele orman sınıflandırma doğruluk değerlerini temsil etmektedir. Standardizasyon işleminin uygulanmadığı veri setinde en yüksek sınıflandırma doğruluğuna %91,67 ile ağaç

sayısının 40 olarak seçildiği modelde ulaşmıştır.



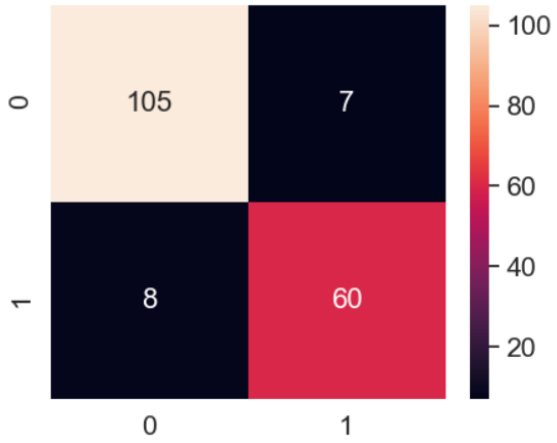
Şekil 3.4. Analiz – 1 16 özellik Rastgele Orman doğruluk değerleri

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.6 Rastgele Orman algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. "0" etiketli oksijen ihtiyacı olmayan hastalar için 0, 933 ve oksijen ihtiyacı olan hastalar için 0,889 F1-skoru ile iyi bir sınıflandırma performansı elde edilmiştir.

Tablo 3.6. Analiz – 1 16 özellikli Rastgele Orman Modeli Performans Metrikleri

Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,93	0,94	0,93	0,82	0,97	0,97
1	0,9	0,88	0,89	0,82	0,97	0,95

Şekil 3.5'de gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde de sekiz (0,07) oksijen ihtiyacı olmayan hastanın ve sekiz (0,12) oksijen ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



Şekil 3.5. Analiz – 1 16 özellikli Rastgele Orman modeli karmaşıklık matrisi

Tez kapsamında daha az sayıda özellik ile her iki sınıf için de yüksek sınıflandırma doğruluklarına ulaşılması hedeflendiğinden dolayı 72 özellik ve 16 özellik ile oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Tablo 3.7 72 özellikle oluşturulan Rastgele Orman modeli ve 16 özellikle oluşturulan Rastgele Orman modelinin her iki sınıf için de ulaştığı performans değerlerini göstermektedir. Sonuçlar incelendiğinde %91,67 genel doğruluk değeri ile sadece 16 özelik kullanılarak yüksek performans ile hastaneye yatış aşamasındaki oksijen ihtiyacının tespit edilebildiği görülmüştür. Ayrıca tez çalışması sonucunda elde edilen model oksijen tedavisine ihtiyaç duyan hastaları 0,933 F1-Skor ve oksijen tedavisine ihtiyaç duymayan hastaları 0,889 F1-Skor ile sınıflandırarak sınıf bazında da iyi bir tahmin performansı elde edilmiştir.

Tablo 3.7. Analiz – 1 72 özellik ve 16 özellik performans sonuçları

Model Adı	Doğruluk	Sınıf Bilgisi (Oksijen tedavisine ihtiyaç)	F1-Skor
72 özellik Rastgele Orman	%89,44	VAR	0,853
		YOK	0,918
16 özellik Rastgele Orman	**%91,67**	VAR	**0,889**
		YOK	**0,933**

3.1.3. Analiz - 1 16 özellikli SMOTE uygulanmış model sonuçları

Tablo 3.2'de kırmızı olarak işaretli 16 özellikten oluşan veri seti 121 adet "0" etiketli hastadan oluşurken 68 adet "1" etiketli yani hastaneye ilk yatış aşamasında oksijen ihtiyacı duyan hasta bulunmaktadır. Analiz kapsamında "1" sınıfı tahmin edilmesi

gereken önemli bir sınıftır fakat diğer sınıfa göre daha az sayıda hasta bulunduğundan dolayı yüksek bir sınıflandırma doğruluğu elde edilmesine rağmen bu sınıfa ait F1-Skor değeri 0,889 ile 0,933 F1-Skor değerine sahip "0" sınıfından daha düşüktür. Bu yüzden 16 özellikten oluşan veri seti üzerinde SMOTE tekniği kullanılarak "1" etiketli sınıftaki hasta sayısı da 121 olacak şekilde çoğaltılarak dengeli bir veri seti elde edilmiştir. Bu veri seti kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.8 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

Tablo 3.8. Analiz – 1 SMOTE 16 özellik sınıflandırma performans sonuçları

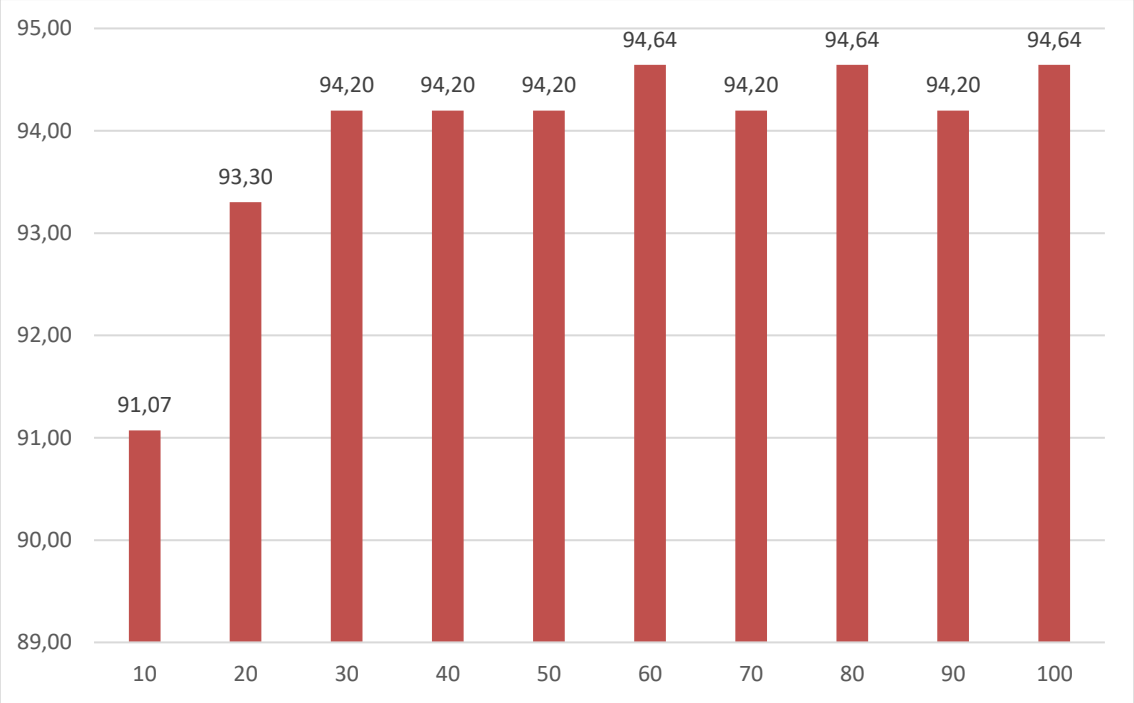
Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	84,8	84,8	84,8	84,82
	Yok	84,8	84,8	84,8	84,82
Bagging	Var	92,4	92,4	92,4	92,41
	Yok	92,9	92,9	92,9	92,86
Rastgele	**Var**	**94,7**	**94,6**	**94,6**	**94.64**
Orman	**Yok**	**94,7**	**94,6**	**94,6**	**94,64**
Karar Ağacı	Var	86,2	86,2	86,2	86,16
	Yok	86,2	86,2	86,2	86,16

Sonuçlar detaylı olarak incelendiğinde 121 adet "0" ve 121 adet "1" etiketli hastadan oluşan veri setinde en yüksek sınıflandırma doğruluğuna %94,64 ile Rastgele Orman algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman algoritmasının ardından Bagging %92,86 sınıflandırma doğruluğuna ulaşırken Karar Ağacı 86,16 ve KNN %84,82 genel doğruluk değeri ile diğer yöntemlere göre daha kötü bir sınıflandırma performansı göstermiştir. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan üzerinde pozitif bir etki oluşturmadığı yani sınıflandırma performansına olumlu bir katkı sağlamadığı görülmektedir. Nihai model olarak 16 özellik kullanılarak hastanın hastaneye yatış aşamasında oksijen ihtiyacını en iyi tahmin eden yöntem standardizasyonun

kullanılmadığı Rastgele Orman algoritması olarak seçilmiştir.

3.1.3.1 Analiz - 1 SMOTE 16 özellikli rastgele orman modeli

Weka üzerinde Rastgele Orman yöntemi için rastgele ormandaki ağaç sayısı, ağacın maksimum derinliği gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Ağacın maksimum derinliği sınırsız olarak belirlenerek farklı ağaç sayıları ile Rastgele Orman modelleri oluşturulmuştur. Şekil 3.6 Farklı ağaç sayıları kullanılarak elde edilen rastgele orman sınıflandırma doğruluk değerlerini temsil etmektedir. Standardizasyon işleminin uygulanmadığı veri setinde en yüksek sınıflandırma doğruluğuna %94,64 ile ağaç sayısının 60, 80 ve 100 olarak seçildiği modellerde ulaşmıştır. 60 ağaç sayısına sahip Rastgele Orman modeli nihai model olarak seçilmiştir.



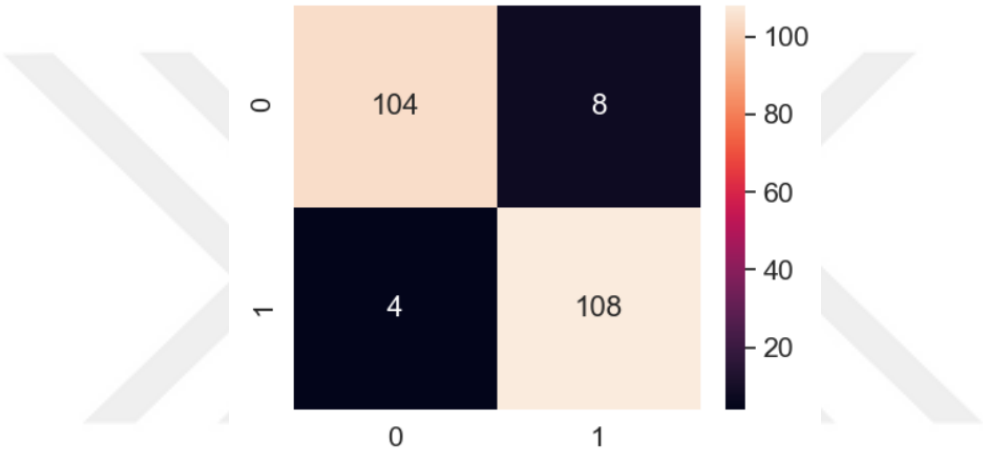
Şekil 3.6. Analiz – 1 SMOTE 16 özellik Rastgele Orman doğruluk değerleri

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.9 Rastgele Orman algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. "0" etiketli oksijen ihtiyacı olmayan hastalar 0, 95 F1-skoru ile iyi bir doğruluk ile tahmin edilirken oksijen ihtiyacı olan hastalar da SMOTE tekniği uygulandıktan sonra 0,95 ile yüksek bir F1-skoru performansı elde etmiştir.

Tablo 3.9. Analiz – 1 SMOTE 16 özellik Rastgele Orman Performans Metrikleri

Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,96	0,93	0,95	0,89	0,98	0,98
1	0,93	0,96	0,95	0,89	0,98	0,98

Şekil 3.7'de gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde de sekiz oksijen ihtiyacı olmayan hastanın ve dört oksijen ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



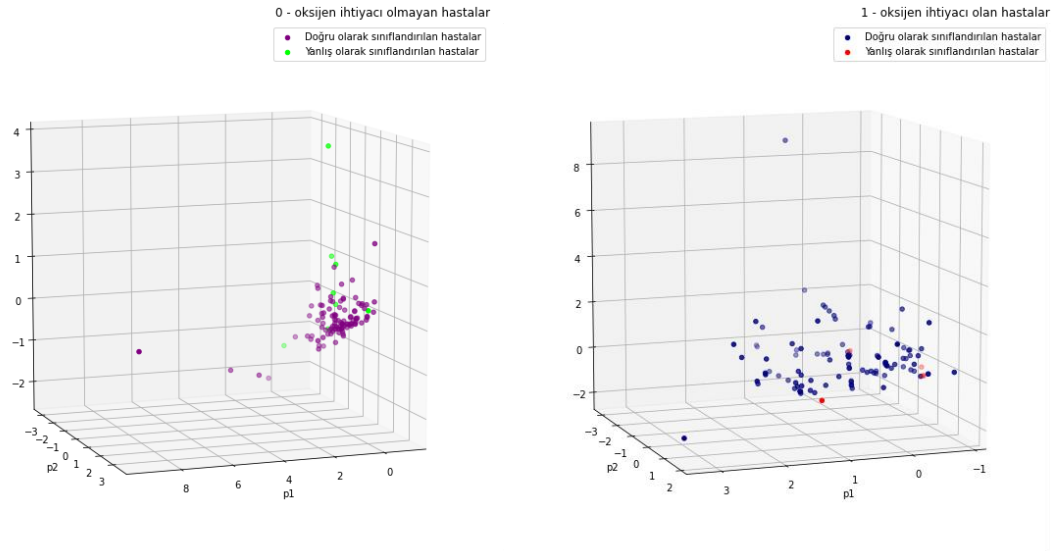
Şekil 3.7. Analiz – 1 SMOTE 16 özellikli Rastgele Orman modeli karmaşıklık matrisi

SMOTE tekniği ile "1" etiketli sınıf için de yüksek sınıflandırma doğruluklarına ulaşılması hedeflendiğinden dolayı SMOTE kullanılmayan veri seti ve SMOTE kullanılarak çoğaltılmış veri seti ile oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Tablo 3.10 16 özellikle oluşturulan Rastgele Orman modeli ve SMOTE uygulanarak 16 özellikle oluşturulan Rastgele Orman modelinin her iki sınıf için de ulaştığı performans değerlerini göstermektedir. Sonuçlar incelendiğinde %96,64 değeri ile SMOTE tekniği kullanıldıktan sonra genel doğruluk değerinde yükselme görülmüştür. Ayrıca SMOTE kullanılmadığı zaman 0,89 F1-Skoru ile tahmin edilen oksijen tedavisine ihtiyaç duyan hastalar 16 özellikli veri setine SMOTE uygulandıktan sonra 0,95 F1-Skor değerine yükselmiştir.

Tablo 3.10. Analiz – 1 16 özellik ve SMOTE 16 özellik performans sonuçları

Model Adı	Doğruluk	Sınıf Bilgisi (Oksijen tedavisine ihtiyaç)	F1-Skor
16 özellik Rastgele Orman	%91,67	VAR	0,89
		YOK	0,93
SMOTE - 16 özellik Rastgele Orman	**%96,64**	VAR	**0,95**
		YOK	**0,95**

SMOTE yöntemi uygulanan veri seti üzerinde PCA analizi yapılarak toplam varyansının yaklaşık %96,7'sinin korunabilmesini sağlayan ilk 3 bileşen seçilerek sınıflandırma sonucu üç boyutlu bir grafik yardımıyla görselleştirilmiştir. Grafik üzerinde görselleştirme işlemi yapılacağından dolayı tüm bileşenlerin yerine sadece ana bileşenleri kullanmak yüksek bir varyans oranı ile veri setinin gösterilmesinde yeterli olmuştur. Şekil 3.8 doğru ve yanlış olarak sınıflandırılan "0" ve "1" etiketli verilerin grafik üzerindeki dağılımını göstermektedir. SMOTE işlemi uygulandıktan sonra sekiz adet oksijen ihtiyacı olmayan hasta yanlış olarak sınıflandırılmıştır. Ayrıca SMOTE uygulanmadan önce yanlış olarak sınıflandırılan sekiz adet oksijen ihtiyacı olan hasta sayısı SMOTE uygulandıktan sonra dörde düşmüştür ve "1" etiketli sınıfın sınıflandırma performansında %89'dan %95'e artış gözlemlenmiştir.



Şekil 3.8. Analiz – 1 SMOTE 16 özellikli Rastgele Orman modeli sınıflandırma grafiği

3.2. Hastanede Yatış Süresi Boyunca Oksijen İhtiyacının Tahmini

Bu analizde Tablo 2.1'de "WHO en kötü oksijen ihtiyacı tahmini" olarak isimlendirilen veri seti kullanılmıştır. Veri seti her bir hastaya ait ilk laboratuvar sonucundan alınan 67 özellik ile yaş, cinsiyet ve WHO giriş niteliklerinden oluşmaktadır. Ayrıca Marmara Tıp Fakültesi'nden Kardiyoloji ve Dahiliye Uzmanı Prof. Dr. Ali Serdar Fak, Marmara Üniversitesi Tıp Fakültesi Göğüs Hastalıkları Anabilim Dalından Prof. Dr. Sait Karakurt, Marmara Üniversitesi Tıp Fakültesi Enfeksiyon Hastalıkları Anabilim Dalından Prof. Dr. Lütfiye Mülazımoğlu, Marmara Üniversitesi Tıp Fakültesi Enfeksiyon Hastalıkları Anabilim Dalından Doç. Dr. Elif Tigen'in önerileri ile belirlenen özelliklerin birbirlerine oranından elde edilen üç özellik de veri setine eklenerek toplamda 73 özellik elde edilmiştir. Analizde hastaneye yatış aşaması süresi boyunca oksijen ihtiyacının tespit edilmesi hedeflendiğinden dolayı ilk yatış sırasında zaten oksijen tedavisine ihtiyaç duyan 68 hasta sağlık uzmanlarının önerileri doğrultusunda veri setinden çıkartılmıştır. Veri setinde toplam 112 adet hasta bulunmaktadır. Bu hastaların 54'i erkek, geri kalan 58'i ise kadındır. 112 adet hastanın yaş ortalaması yaklaşık 51,72'dir ve 18-93 yaş aralığında Covid-19 şüphesi ile hastaneye başvuran hastalar bulunmaktadır.

Bu analizde hastanede yatış süreci boyunca hastanın oksijen ihtiyacının tahmin edilmesi hedeflendiğinden dolayı hedef değer olarak en kötü WHO değeri kullanılmıştır ve veri setindeki eksik nitelikler ortalamaya göre eksik veri doldurma tekniği kullanılarak doldurulmuştur.

Oksijen ihtiyacının tahmin edilmesi hedeflendiğinden dolayı sağlık çalışanlarının önerileri doğrultusunda WHO en kötü limit değeri dört olarak belirlenmiştir. Böylece 112 adet hastadan WHO en kötü değeri dört ve altında olan 91 hasta veri setinde "0" olarak etiketlenirken geri kalan 21 hasta ise "1" olarak etiketlenmiştir. 91 hasta ile "0" etiketli sınıf için iyi bir hasta kaydı var olmasına rağmen "1" olarak etiketlenen ve oksijen tedavisine ihtiyaç duyan hastaların sayısı azdır. Tüm sınıflandırma işlemleri KNN, Bagging, Rastgele Orman ve Karar Ağacı olmak üzere dört farklı makine öğrenmesi yöntemi kullanılarak Weka üzerinde yapılmıştır ve Tablo 3.1'de açıklanan 16 farklı model oluşturulmuştur.

73 özelliğin tümü kullanılarak elde edilen sınıflandırma sonuçları ile karşılaştırılmak üzere Rastgele Özellik Seçme algoritması ile seçilen özellikler üzerinde de analizler

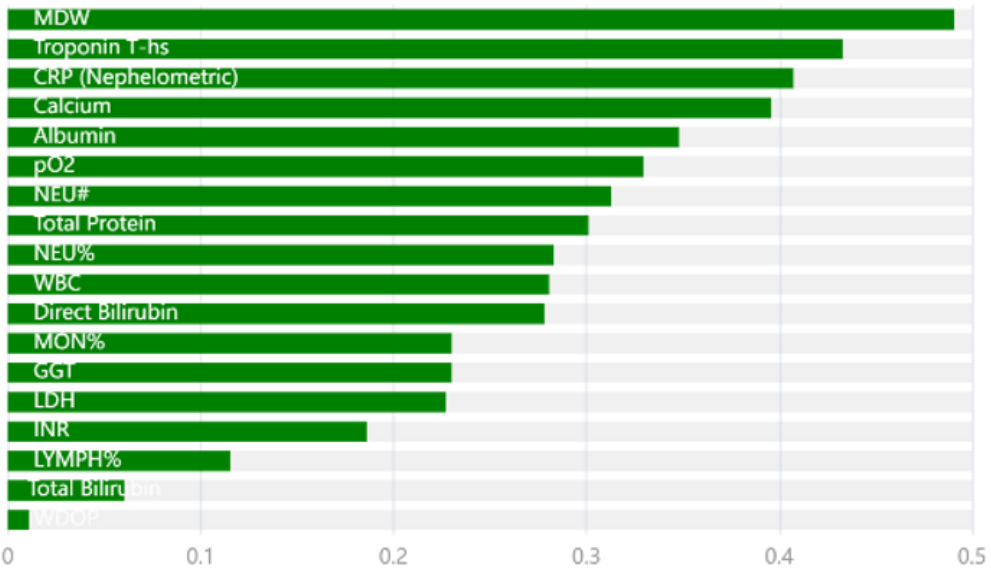
yapılmıştır. Tablo 3.11 analiz için kullanılan tüm özelliklerin bir listesini ve tüm özellikler içerisinden Rastgele Özellik Seçme algoritması kullanılarak seçilen 18 adet niteliği temsil etmektedir. Tablo 3.11'deki 73 özniteliğe Rastgele Özellik Seçim Algoritması uygulanarak 20 öznitelik seçilmiştir. 20 öznitelikten PT, %PT ve INR arasındaki yüksek korelasyon nedeniyle, PT ve %PT öznitelikleri veri setinden çıkarılmış ve sadece INR özniteliği bırakılmıştır.

Tablo 3.11. Analiz – 2'de kullanılan 73 özellik ve RSFS algoritması ile seçilen 18 özellik

Özellikler			Seçilen Özellikler
Sodyum	MCV	MPV	Total Protein
sO_2	MCHC	Kreatinin	INR
Laktat	MCH	LYMPH#	WBC
Üre	MON#	ALP	WDOP
Ürik asit	Kütle CK-MB	ALT	LYMPH%
plt/lymph#	Karboksihemoglobin	aPTT	pO2
neu#/lymph#	Cinsiyet	AST	MON%
lymph#/crp	NRBC	BAS#	Albumin
PCT	NRBC#	BAS%	Total Bilirubin
pCO_2	D-dimer	Baz Açığı	CRP (Nefelometrik)
PDW	EOS#	Prokalsitonin	Direkt Bilirubin
pH	EOS%	Ozmolarite	Troponin T-hs
PLT	Ferritin	BUN	MDW
PT	Fibrinojen	Klor	Kalsiyum
PT%	Fosfor	Glukoz	NEU#
RBC	Magnezyum	HCO3-	GGT
RDW	Methemoglobin	HCT	LDH
Potasyum	WHO giriş	HGB	NEU%
Yaş			

Tabloda kırmızı bir şekilde belirtilen 18 özellik hastanede yatış süresi boyunca hastanın oksijen tedavisine ihtiyaç duyup duymaması ile ilgili en önemli özellikler olarak bu tez kapsamında belirlenmiştir. Şekil 3.9 seçilen 18 özelliğin Korelasyon Öznitelik Algoritması tarafından belirlenen katkı değerlerini göstermektedir. MDW (0,49), yüksek

duyarlı troponin T (0,43), CRP (0,41) ve kalsiyum (0,40), başvuru sırasında oksijene ihtiyaç duymayan hastalar için oksijen tedavisi ihtiyacını belirlemede en kritik katkı sağlayanlar özelliklerdir.



Şekil 3.9. Analiz – 2 seçilen 18 özelliğin CA algoritmasına göre katkıları

3.2.1. Analiz - 2 73 özellikle oluşturulan model sonuçları

Tablo 3.11'de belirtilen 73 özelliğin tamamı kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.12 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

Tablo 3.12. Analiz – 2 73 özellik sınıflandırma performans sonuçları

Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	85,4	82,1	74,9	82,14
	Yok	85,4	82,1	74,9	82,14
Bagging	Var	86,9	87,5	85,9	87,5
	Yok	86,9	87,5	85,9	87,5

Rastgele	Var	87,7	87,5	85,4	87,5
Orman	Yok	89,2	87,5	84,8	87,5
Karar	**Var**	**88,6**	**88,4**	**88,5**	**88,39**
Ağacı	**Yok**	**88,6**	**88,4**	**88,5**	**88,39**

Sonuçlar detaylı olarak incelendiğinde en yüksek sınıflandırma doğruluğuna %88,39 ile Karar Ağacı algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman algoritmasının ardından Bagging ve Rastgele Orman %87,5 sınıflandırma doğruluğuna ulaşırken KNN %82,14 genel doğruluk değeri ile diğer yöntemlere göre daha kötü bir sınıflandırma performansı göstermiştir. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan modeller üzerinde pozitif bir etki oluşturmadığı yani sınıflandırma performansına olumlu bir katkı sağlamadığı görülmektedir. Bu yüzden 73 özellik kullanılarak hastanın hastanede yatış süresi boyunca oksijen ihtiyacını en iyi tahmin eden yöntem standardizasyonun kullanılmadığı Karar Ağacı algoritması olarak seçilmiştir.

3.2.1.1 Analiz - 2 73 özellikli karar ağacı modeli

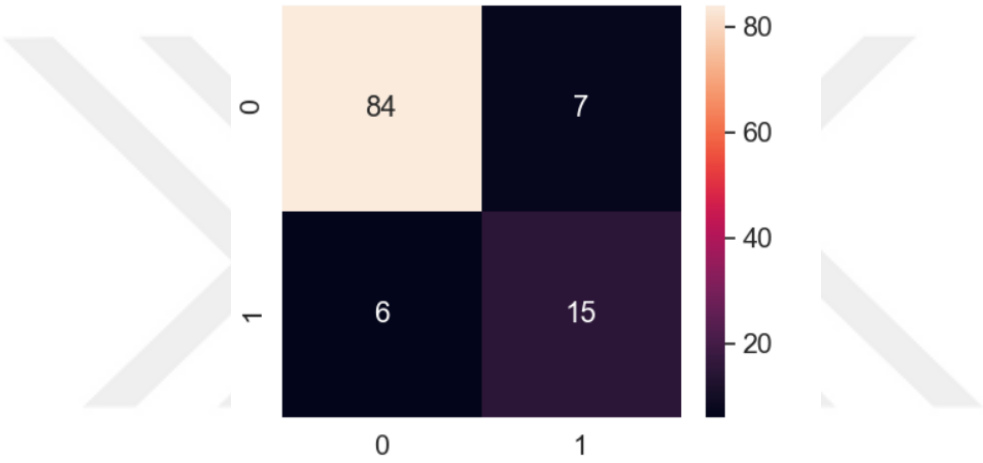
Weka üzerinde Karar Ağacı yöntemi için güven faktörü ve minimum obje sayısı gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Güven faktörü budama için kullanılan değeri temsil eder ve küçük değerler daha fazla budamaya neden olur. Minimum obje sayısı ise yaprak başına minimum örnek sayısı değerini gösterir. Farklı güven faktörü ve minimum obje sayısı değerleri ile karar ağacı modelleri oluşturularak en yüksek doğruluk değerine sahip olan karar ağacı modeli oluşturulmuştur. Minimum obje sayısı dört ve güven faktörü 0,25 olarak karar ağacı modeli %88,39 ile en yüksek sınıflandırma doğruluğuna sahip olan model olarak seçilmiştir.

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.13 algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. 0, 928 F1-skoru ile “0” etiketli oksijen ihtiyacı olmayan hastalar ve 0,698 ile “1” etiketli hastanede yatış süresi boyunca oksijen ihtiyacı olacak hastalar ortalama bir performans ile tahmin edilmiştir.

Tablo 3.13. Analiz – 2 73 özellikli Karar Ağacı Modeli Performans Metrikleri

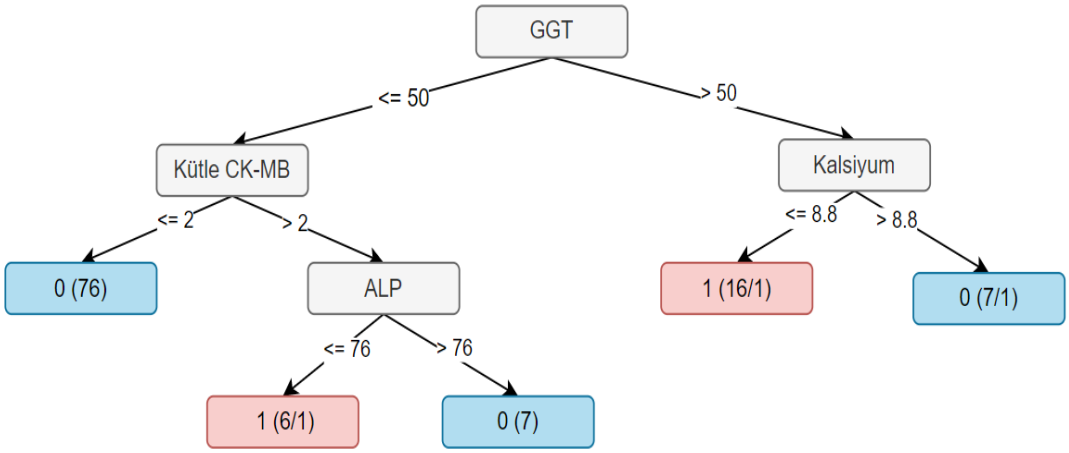
Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,93	0,92	0,93	0,63	0,87	0,95
1	0,68	0,71	0,7	0,63	0,87	0,59

Şekil 3.10'da gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde yedi oksijen ihtiyacı olmayan hastanın ve altı oksijen ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



Şekil 3.10. Analiz – 2 73 özellikli Karar Ağacı modeli karmaşıklık matrisi

Şekil 3.11 beş tanesi yaprak düğüm olan toplamda dokuz düğüm içeren karar ağacı modelini göstermektedir. Karar Ağacı modeli dört farklı nitelik kullanılarak oluşturulmuştur ve en yüksek bilgi kazancı sağlayan özellik GGT olarak seçilmiştir. GGT değerinin 0.1964'ten küçük, Kütle CK-MB değerinin 0,4120'den küçük olduğu durum 76 "0" etiketli hasta sayısı ile en fazla sayıda elemanın sınıflandırılmasını sağlayan yaprak düğümdür. Ayrıca karar ağacına göre "ggt <= 0,1964 && kütle ck-mb > 0,4120 && alp <= -3009" ve "ggt > 0,1964 && kalsiyum <= 0,1160" kuralları hastanede yatış süresi boyunca oksijen tedavisine ihtiyacı tek bir hasta hatası ile sınıflandıran kurallar olarak belirlenmiştir.



Şekil 3.11. Analiz – 2 73 özellikli Karar Ağacı Modeli

3.2.2. Analiz - 2 18 özellikle oluşturulan model sonuçları

Tablo 3.11'de kırmızı olarak işaretlenen 18 özellik kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.14 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

Tablo 3.14. Analiz – 2 18 özellik sınıflandırma performans sonuçları

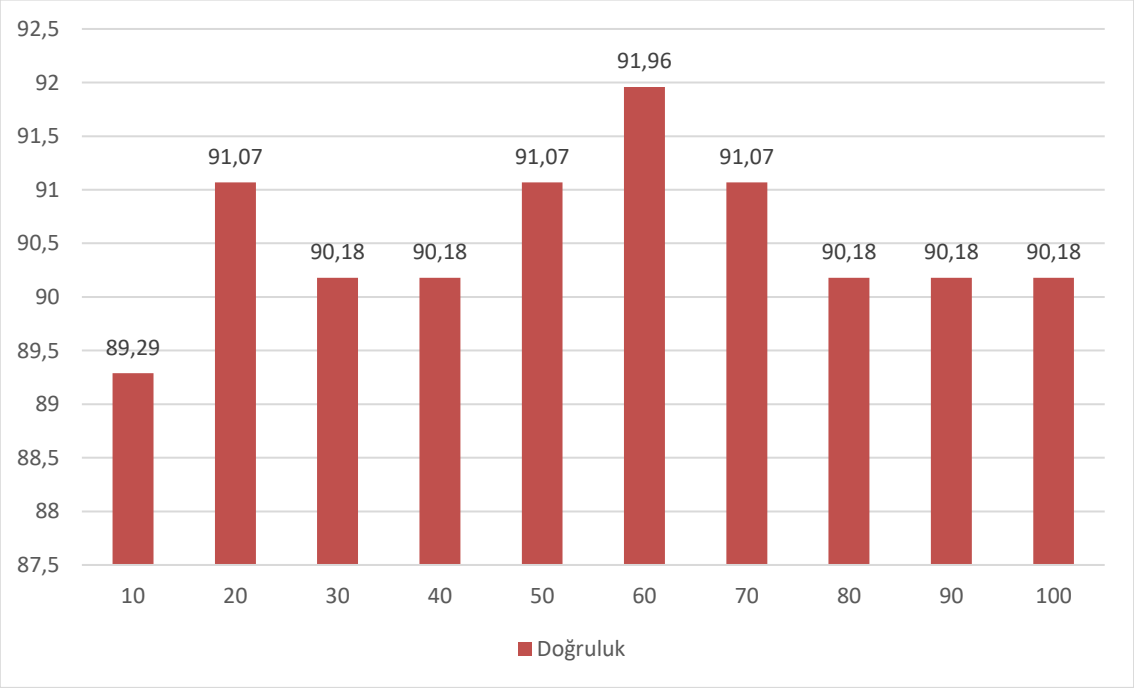
Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	87,2	84,8	80,3	84,82
	Yok	87,2	84,8	80,3	84,82
Bagging	Var	90,3	90,2	89,1	90,18
	Yok	86,9	87,5	85,9	87,5
Rastgele	**Var**	**92**	**92**	**91,3**	**91,96**
Orman	**Yok**	**92**	**92**	**91,3**	**91,96**
Karar Ağacı	Var	87,1	87,5	87,3	87,5
	Yok	87,1	87,5	87,3	87,5

Sonuçlar detaylı olarak incelendiğinde en yüksek sınıflandırma doğruluğuna %91,96 ile Rastgele Orman algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman

algoritmasının ardından Bagging %90,17 ve Karar Ağacı da %87,5 sınıflandırma doğruluğuna ulaşırken KNN %84,82 genel doğruluk değeri ile diğer yöntemlere göre daha kötü bir sınıflandırma performansı göstermiştir. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan modellerden sadece Bagging modeli üzerinde pozitif bir etki oluşturduğu yani sınıflandırma performansına olumlu bir katkı sağladığı görülmektedir. Nihai model olarak 18 özellik kullanılarak hastanın hastanede yatış süreci boyunca oksijen ihtiyacını en iyi tahmin eden yöntem standardizasyonun kullanılmadığı Rastgele Orman algoritması olarak seçilmiştir.

3.2.2.1 Analiz - 2 18 özellikli rastgele orman modeli

Weka üzerinde Rastgele Orman yöntemi için rastgele ormandaki ağaç sayısı, ağacın maksimum derinliği gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Ağacın maksimum derinliği sınırsız olarak belirlenerek farklı ağaç sayıları ile Rastgele Orman modelleri oluşturulmuştur. Şekil 3.12 Farklı ağaç sayıları kullanılarak elde edilen rastgele orman sınıflandırma doğruluk değerlerini temsil etmektedir. Standardizasyon işleminin uygulanmadığı veri setinde en yüksek sınıflandırma doğruluğuna %91,96 ile ağaç sayısının 60 olarak seçildiği modellerde ulaşmıştır.



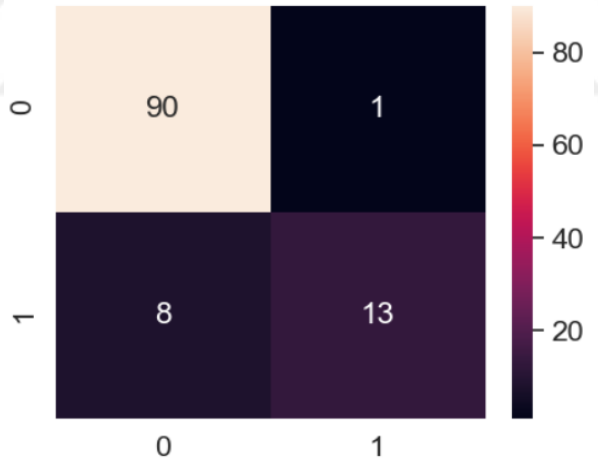
Şekil 3.12. Analiz – 2 18 özellik Rastgele Orman doğruluk değerleri

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.15 Rastgele Orman algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. "0" etiketli oksijen ihtiyacı olmayan hastalar 0, 989 F1- skoru ile iyi bir doğruluk ile tahmin edilirken oksijen ihtiyacı olan hastalar 0,619 ile düşük bir F1-skoru performansı elde etmiştir.

Tablo 3.15. Analiz – 2 18 özellikli Rastgele Orman Modeli Performans Metrikleri

Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,92	0,99	0,95	0,72	0,96	0,99
1	0,93	0,62	0,74	0,72	0,96	0,85

Şekil 3.13'te gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde de bir oksijen ihtiyacı olmayan hastanın ve sekiz oksijen ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



Şekil 3.13. Analiz – 2 18 özellikli Rastgele Orman modeli karmaşıklık matrisi

Tez kapsamında daha az sayıda özellik ile her iki sınıf için de yüksek sınıflandırma doğruluklarına ulaşılması hedeflendiğinden dolayı 73 özellik ve 18 özellik ile oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Tablo 3.16 73 özellikle oluşturulan Karar Ağacı modeli ve 18 özellikle oluşturulan Rastgele Orman modelinin her iki sınıf için de ulaştığı performans değerlerini göstermektedir. Sonuçlar incelendiğinde %91,96 genel doğruluk değeri ile sadece 18 özelik kullanılarak yüksek performans ile hastaneye yatış aşamasındaki oksijen ihtiyacının tespit edilebildiği görülmüştür. Fakat tez çalışması

sonucunda elde edilen model oksijen tedavisine ihtiyaç duyan hastaları 0,952 F1-Skoru ile iyi bir performans ile tahmin edebilirken oksijen tedavisine ihtiyaç duymayan hastalar veri sayısının az olmasından dolayı 0,743 F1-Skoru ile ortalama bir sınıflandırma performansına ulaşmıştır.

Tablo 3.16. Analiz – 2 73 özellik ve 18 özellik performans sonuçları

Model Adı	Doğruluk	Sınıf Bilgisi (Oksijen tedavisine ihtiyaç)	F1-Skor
73 özellik Karar Ağacı	%88,39	VAR	0,698
		YOK	0,928
18 özellik Rastgele Orman	**%91,96**	VAR	**0,743**
		YOK	**0,952**

3.2.3. Analiz - 2 18 özellikli SMOTE uygulanmış model sonuçları

Tablo 3.11'de kırmızı olarak işaretli 18 özellikten oluşan veri seti 91 adet "0" etiketli hastadan oluşurken 21 adet "1" etiketli yani hastanede yatış aşaması boyunca oksijen ihtiyacı duyan hasta bulunmaktadır. Analiz kapsamında "1" sınıfı tahmin edilmesi gereken önemli bir sınıftır fakat diğer sınıfa göre daha az sayıda hasta bulunduğundan dolayı yüksek bir sınıflandırma doğruluğu elde edilmesine rağmen bu sınıfa ait F1-Skor değeri 0,74 ile 0,95 F1-Skor değerine sahip "0" sınıfından daha düşüktür. Bu yüzden 18 özellikten oluşan veri seti üzerinde SMOTE tekniği kullanılarak "1" etiketli sınıftaki hasta sayısı da 91 olacak şekilde çoğaltılarak dengeli bir veri seti elde edilmiştir. Bu veri seti kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.17 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

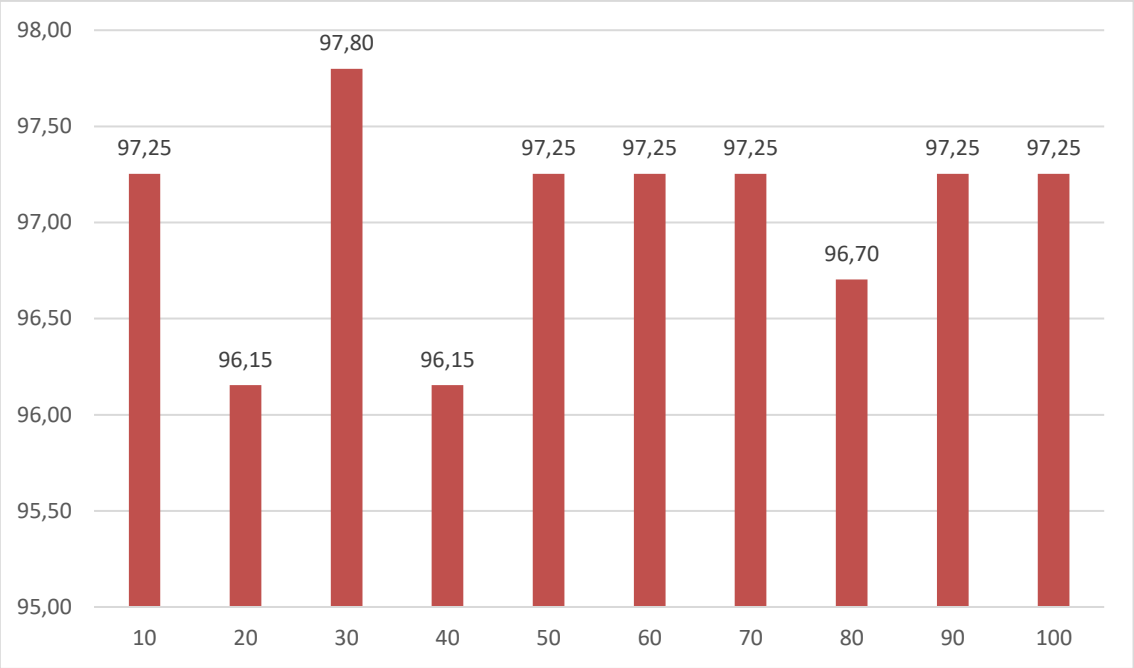
Tablo 3.17. Analiz – 2 SMOTE 18 özellik sınıflandırma performans sonuçları

Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	91,5	91,2	91,2	91,21
	Yok	91,5	91,2	91,2	91,21
Bagging	Var	95,1	95,1	95,1	95,06
	Yok	96,2	96,2	96,2	96,15
Rastgele	**Var**	**97,8**	**97,8**	**97,8**	**97,8**
Orman	**Yok**	**97,8**	**97,8**	**97,8**	**97,8**
Karar Ağacı	Var	93,4	93,4	93,4	93,41
	Yok	93,4	93,4	93,4	93,41

Sonuçlar detaylı olarak incelendiğinde 91 adet "0" ve 91 adet "1" etiketli hastadan oluşan veri setinde en yüksek sınıflandırma doğruluğuna %97,8 ile Rastgele Orman algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman algoritmasının ardından Bagging %96,15 sınıflandırma doğruluğuna ulaşırken Karar Ağacı %93,41 ve KNN %91,21 genel doğruluk değeri ile diğer yöntemlere göre daha kötü bir sınıflandırma performansı göstermiştir. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan üzerinde pozitif bir etki oluşturmadığı yani sınıflandırma performansına olumlu bir katkı sağlamadığı görülmektedir. Nihai model olarak 18 özellik kullanılarak hastanın hastanede yatış süreci boyunca oksijen ihtiyacını en iyi tahmin eden yöntem standardizasyonun kullanılmadığı Rastgele Orman algoritması olarak seçilmiştir.

3.2.3.1 Analiz - 2 SMOTE 18 özellikli rastgele orman modeli

Weka üzerinde Rastgele Orman yöntemi için rastgele ormandaki ağaç sayısı, ağacın maksimum derinliği gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Ağacın maksimum derinliği sınırsız olarak belirlenerek farklı ağaç sayıları ile Rastgele Orman modelleri oluşturulmuştur. Şekil 3.14 Farklı ağaç sayıları kullanılarak elde edilen rastgele orman sınıflandırma doğruluk değerlerini temsil etmektedir. Standardizasyon işleminin uygulanmadığı veri setinde en yüksek sınıflandırma doğruluğuna %97,8 ile ağaç sayısının 30 olarak seçildiği modellerde ulaşmıştır.



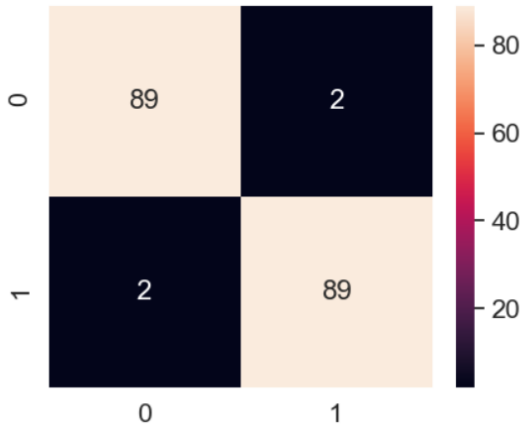
Şekil 3.14. Analiz – 2 SMOTE 18 özellik Rastgele Orman doğruluk değerleri

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.18 Rastgele Orman algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. "0" etiketli oksijen ihtiyacı olmayan hastalar 0, 98 F1-skoru ile iyi bir doğruluk ile tahmin edilirken oksijen ihtiyacı olan hastalar da SMOTE tekniği uygulandıktan sonra 0,98 ile yüksek bir F1-skoru performansı elde etmiştir.

Tablo 3.18. Analiz – 2 SMOTE 18 özellik Rastgele Orman Performans Metrikleri

Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,98	0,98	0,98	0,96	0,99	0,99
1	0,98	0,98	0,98	0,96	0,99	0,99

Şekil 3.15'de gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde de iki oksijen ihtiyacı olmayan hastanın ve iki oksijen ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



Şekil 3.15. Analiz – 2 SMOTE 18 özellikli Rastgele Orman modeli karmaşıklık matrisi

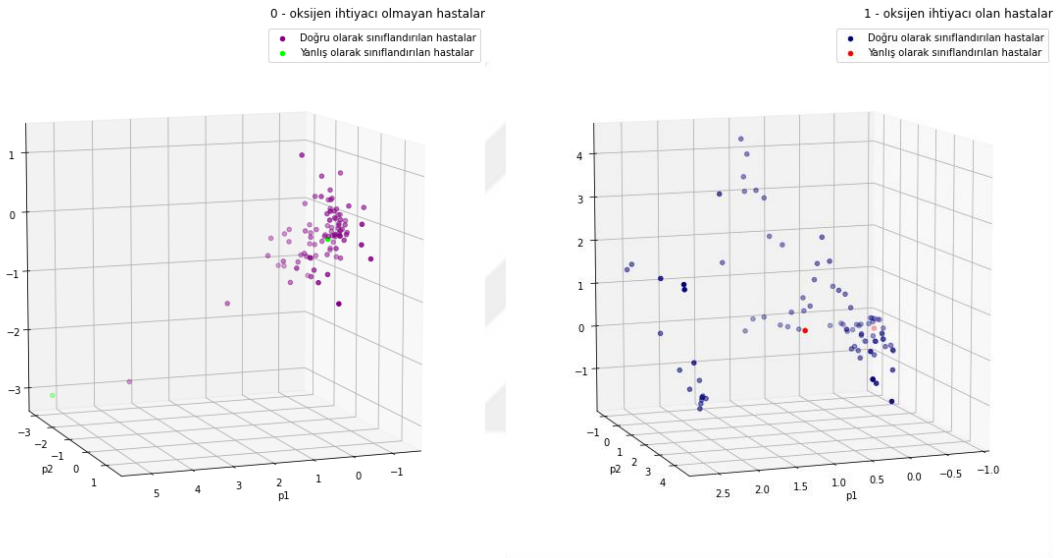
SMOTE tekniği ile "1" etiketli sınıf için de yüksek sınıflandırma doğruluklarına ulaşılması hedeflendiğinden dolayı SMOTE kullanılmayan veri seti ve SMOTE kullanılarak çoğaltılmış veri seti ile oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Tablo 3.19 18 özellikle oluşturulan Rastgele Orman modeli ve SMOTE uygulanarak 18 özellikle oluşturulan Rastgele Orman modelinin her iki sınıf için de ulaştığı performans değerlerini göstermektedir. Sonuçlar incelendiğinde %97,8 değeri ile SMOTE tekniği kullanıldıktan sonra genel doğruluk değerinde yükselme görülmüştür. Ayrıca SMOTE kullanılmadığı zaman 0,74 F1-Skoru ile tahmin edilen oksijen tedavisine ihtiyaç duyan hastalar 18 özellikli veri setine SMOTE uygulandıktan sonra 0,98 F1-Skor değerine yükselmiştir.

Tablo 3.19. Analiz – 2 18 özellik ve SMOTE 18 özellik performans sonuçları

Model Adı	Doğruluk	Sınıf Bilgisi (Oksijen tedavisine ihtiyaç)	F1-Skor
18 özellik Rastgele Orman	%91,96	VAR	0,74
		YOK	0,95
SMOTE - 18 özellik Rastgele Orman	**%97,8**	VAR	0,98
		YOK	0,98

SMOTE yöntemi uygulanan veri seti üzerinde PCA analizi yapılarak toplam varyansının yaklaşık %97,7'sinin korunabilmesini sağlayan ilk 3 bileşen seçilerek sınıflandırma sonucu üç boyutlu bir grafik yardımıyla görselleştirilmiştir. Grafik üzerinde görselleştirme işlemi yapılacağından dolayı tüm bileşenlerin yerine sadece ana bileşenleri

kullanmak yüksek bir varyans oranı ile veri setinin gösterilmesinde yeterli olmuştur. Şekil 3.16 doğru ve yanlış olarak sınıflandırılan “0” ve “1” etiketli verilerin grafik üzerindeki dağılımını göstermektedir. SMOTE işlemi uygulandıktan sonra 2 adet oksijen ihtiyacı olmayan hasta yanlış olarak sınıflandırılmıştır. Ayrıca SMOTE uygulanmadan önce yanlış olarak sınıflandırılan sekiz adet oksijen ihtiyacı olan hasta sayısı SMOTE uygulandıktan sonra ikiye düşmüştür ve “1” etiketli sınıfın sınıflandırma performansında %74’ten %98’e artış gözlemlenmiştir. Yanlış olan sınıflandırılan “1” etiketli hastalardan biri SMOTE ile oluşturulmuş yapay bir veridir ve diğeri ise veri setinde bulunan gerçek bir hasta kaydına aittir.



Şekil 3.16. Analiz – 2 SMOTE 18 özellikli Rastgele Orman modeli sınıflandırma grafiği

3.3. Hastanede Yatış Süresi Boyunca Yoğun Bakım İhtiyacının Tahmini

Bu analizde Tablo 2.1’de “WHO en kötü yoğun bakım ihtiyacı tahmini” olarak isimlendirilen veri seti kullanılmıştır. Veri seti her bir hastaya ait ilk laboratuvar sonucundan alınan 67 özellik ile yaş, cinsiyet ve WHO giriş niteliklerinden oluşmaktadır. Ayrıca Marmara Tıp Fakültesi’nden Kardiyoloji ve Dahiliye Uzmanı Prof. Dr. Ali Serdar Fak, Marmara Üniversitesi Tıp Fakültesi Göğüs Hastalıkları Anabilim Dalından Prof.Dr. Sait Karakurt, Marmara Üniversitesi Tıp Fakültesi Enfeksiyon Hastalıkları Anabilim Dalından Prof.Dr. Lütfiye Mülazımoğlu, Marmara Üniversitesi Tıp Fakültesi Enfeksiyon Hastalıkları Anabilim Dalından Doç.Dr. Elif Tigen’in önerileri ile belirlenen özelliklerin birbirlerine oranından elde edilen üç özellik de veri setine eklenerek toplamda 73 özellik

elde edilmiştir. Analizde hastaneye yatış aşaması süresi boyunca yoğun bakım ihtiyacının tespit edilmesi hedeflendiğinden dolayı ilk yatış sırasında zaten yoğun bakım tedavisine ihtiyaç duyan 14 hasta sağlık uzmanlarının önerileri doğrultusunda veri setinden çıkartılmıştır. Veri setinde toplam 166 adet hasta bulunmaktadır. Bu hastaların 84'i erkek, geri kalan 82'i ise kadındır. 166 adet hastanın yaş ortalaması yaklaşık 55,03'tür ve 18-96 yaş aralığında Covid-19 şüphesi ile hastaneye başvuran hastalar bulunmaktadır.

Bu analizde hastanede yatış süreci boyunca hastanın yoğun bakım ihtiyacının tahmin edilmesi hedeflendiğinden dolayı hedef değer olarak en kötü WHO değeri kullanılmıştır ve veri setindeki eksik nitelikler ortalamaya göre eksik veri doldurma tekniği kullanılarak doldurulmuştur.

Yoğun bakım ihtiyacının tahmin edilmesi hedeflendiğinden dolayı sağlık çalışanlarının önerileri doğrultusunda WHO en kötü limit değeri altı olarak belirlenmiştir. Böylece 166 adet hastadan WHO en kötü değeri altı ve altında olan 137 hasta veri setinde "0" olarak etiketlenirken geri kalan 29 hasta ise "1" olarak etiketlenmiştir. 137 hasta ile "0" etiketli sınıf için iyi bir hasta kaydı var olmasına rağmen "1" olarak etiketlenen ve oksijen tedavisine ihtiyaç duyan hastaların sayısı azdır. Tüm sınıflandırma işlemleri KNN, Bagging, Rastgele Orman ve Karar Ağacı olmak üzere dört farklı makine öğrenmesi yöntemi kullanılarak Weka üzerinde yapılmıştır ve Tablo 3.1'de açıklanan 16 farklı model oluşturulmuştur.

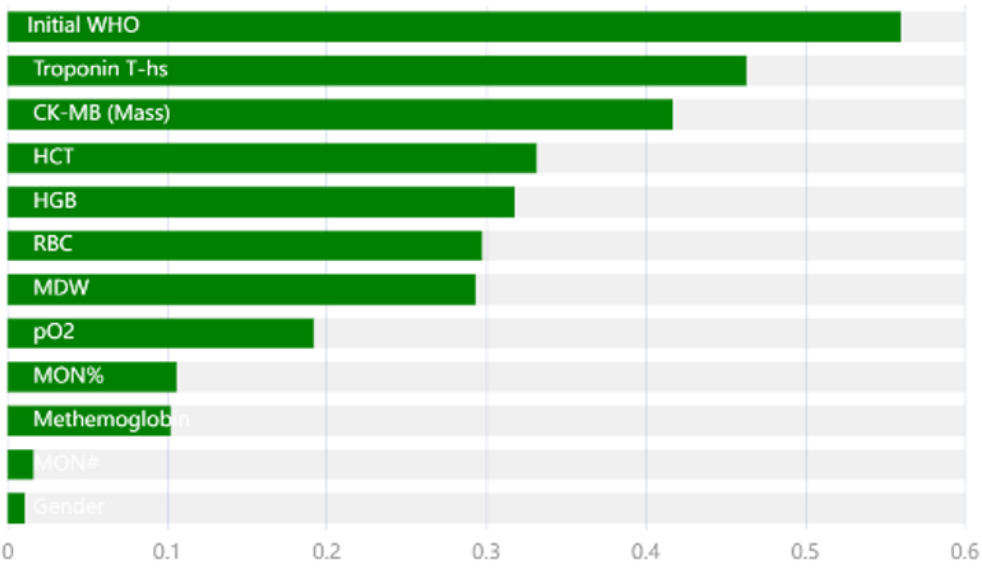
73 özelliğin tümü kullanılarak elde edilen sınıflandırma sonuçları ile karşılaştırılmak üzere Rastgele Özellik Seçme algoritması ile seçilen özellikler üzerinde de analizler yapılmıştır. Tablo 3.20 analiz için kullanılan tüm özelliklerin bir listesini ve tüm özellikler içerisinden Rastgele Özellik Seçme algoritması kullanılarak seçilen 12 adet niteliği temsil etmektedir. Tablo 3.20'deki 73 özniteliğe Rastgele Özellik Seçim Algoritması uygulanarak 12 öznitelik seçilmiştir.

Tablo 3.20. Analiz – 3 73 özellik ve RSFS algoritması ile seçilen 12 özellik

Özellikler					Seçilen Özellikler
NEU#	D-dimer	Kalsiyum	pCO_2	sO_2	Troponin T-hs
NEU%	MPV	Yaş	PCT	Sodyum	MDW
ALP	EOS#	Albumin	PDW	Laktat	MON#

ALT	EOS%	Prokalsitonin	pH	Total Protein	MON%
aPTT	Ferritin	Ozmolarite	PLT	LYMPH%	pO2
AST	BAS%	Fibrinojen	Üre	LYMPH#	Kütle CK-MB
BAS#	Fosfor	Baz Açığı	INR	Ürik asit	RBC
MCV	MCH	Direkt Bilirubin		WBC	Methemoglobin
LDH	Glukoz	Karboksihemoglobin	PT%	WDOP	WHO giriş
BUN	NRBC	CRP (Nefelometrik)	RDW	plt/lymph#	HCT
Klor	MCHC	Kreatinin		neu#/lymph#	HGB
PT	NRBC#	Total Bilirubin		lymph#/crp	Cinsiyet
GGT	HCO3-	Magnezyum		Potasyum	

Tabloda kırmızı bir şekilde belirtilen 12 özellik hastanede yatış süresi boyunca hastanın yoğun bakım tedavisine ihtiyaç duyup duymaması ile ilgili en önemli özellikler olarak bu tez kapsamında belirlenmiştir. Şekil 3.17 seçilen 12 özelliğin Korelasyon Öznitelik Algoritması tarafından belirlenen katkı değerlerini göstermektedir. Korelasyon Öznitelik Değerlendirmesine göre Giriş WHO (0,56), yüksek duyarlı troponin T (0,46) ve CK-MB (0,42) en önemli katkıda bulunan özellikler olmuştur.



Şekil 3.17. Analiz – 3 seçilen 12 özelliğin CA algoritmasına göre katkıları

3.3.1. Analiz - 3 73 özellikle oluşturulan model sonuçları

Tablo 3.20'de belirtilen 73 özelliğin tamamı kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.21 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

Tablo 3.21. Analiz – 3 73 özellik sınıflandırma performans sonuçları

Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	90,4	89,2	87,1	89,16
	Yok	90,4	89,2	87,1	89,16
Bagging	Var	88,9	89,2	87,7	89,16
	Yok	88,9	89,2	87,7	89,16
Rastgele Orman	Var	90,4	89,2	87,1	89,16
	Yok	90,4	89,2	87,1	89,16
Karar Ağacı	**Var**	**92,8**	**92,8**	**92,8**	**92,77**
	Yok	**92,6**	**92,8**	**92,7**	**92,77**

Sonuçlar detaylı olarak incelendiğinde en yüksek sınıflandırma doğruluğuna %92,77 ile Karar Ağacı algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman algoritmasının ardından Bagging, Rastgele Orman ve KNN modelleri %89,16 sınıflandırma doğruluğuna ulaşmıştır. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan modellerden sadece KNN modeli üzerinde pozitif bir etki oluşturduğu yani sınıflandırma performansına olumlu bir katkı sağladığı görülmektedir. Nihai model olarak 73 özellik kullanılarak hastanın hastanede yatış süreci boyunca yoğun bakım ihtiyacını en iyi tahmin eden yöntem standardizasyonun kullanılmadığı Karar Ağacı algoritması olarak seçilmiştir.

3.3.1.1 Analiz - 3 73 özellikli karar ağacı modeli

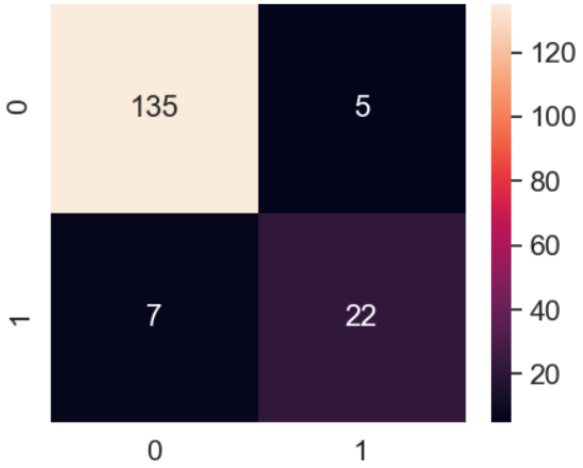
Weka üzerinde Karar Ağacı yöntemi için güven faktörü ve minimum obje sayısı gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Güven faktörü budama için kullanılan değeri temsil eder ve küçük değerler daha fazla budamaya neden olur. Minimum obje sayısı ise yaprak başına minimum örnek sayısı değerini gösterir. Farklı güven faktörü ve minimum obje sayısı değerleri ile karar ağacı modelleri oluşturularak en yüksek doğruluk değerine sahip olan karar ağacı modeli oluşturulmuştur. Minimum obje sayısı dört ve güven faktörü 0,1 olarak karar ağacı modeli %92,77 ile en yüksek sınıflandırma doğruluğuna sahip olan model olarak seçilmiştir.

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.22 Karar Ağacı algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. 0, 957 F1-skoru ile "0" etiketli yoğun bakım ihtiyacı olmayan hastalar iyi bir sınıflandırma performansı ile tahmin edilirken 0,786 ile "1" etiketli hastanede yatış süresi boyunca yoğun bakım ihtiyacı olacak hastalar ortalama bir performans ile tahmin edilmiştir.

Tablo 3.22. Analiz – 3 73 özellikli Karar Ağacı Modeli Performans Metrikleri

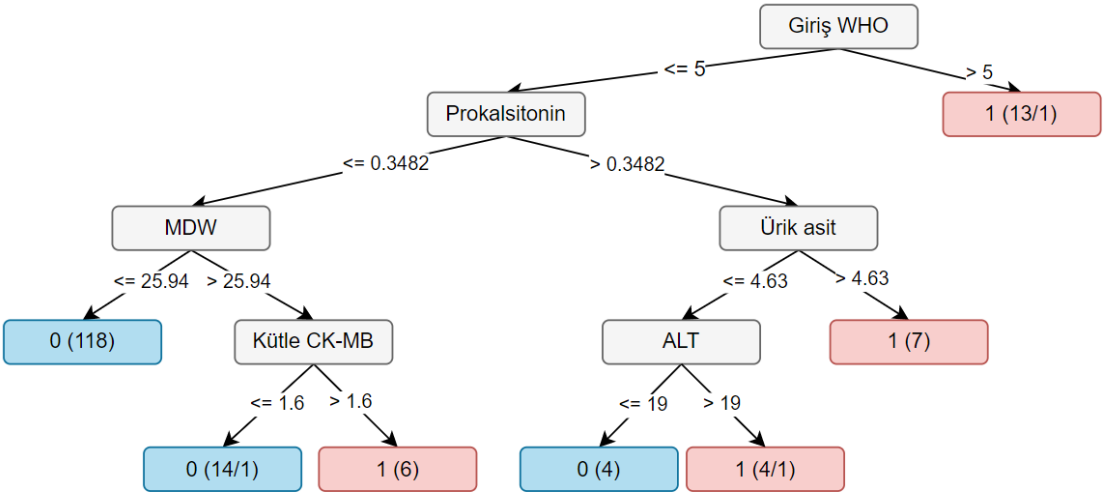
Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,95	0,96	0,96	0,74	0,91	0,98
1	0,82	0,76	0,79	0,74	0,91	0,69

Şekil 3.18'de gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde beş yoğun bakım ihtiyacı olmayan hastanın ve yedi yoğun bakım ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



Şekil 3.18. Analiz – 3 73 özellikli Karar Ağacı modeli karmaşıklık matrisi

Şekil 3.19 yedi tanesi yaprak düğüm olan toplamda 13 düğüm içeren karar ağacı modelini göstermektedir. Karar Ağacı modeli altı farklı nitelik kullanılarak oluşturulmuştur ve en yüksek bilgi kazancı sağlayan özellik WHO giriş olarak seçilmiştir. WHO giriş değerinin 5'ten küçük, Prokalsitonin değerinin 0,3482'den küçük ve mdw değerinin 25,94'den küçük olduğu durum 118 "0" etiketli hasta sayısı ile en fazla sayıda elemanın sınıflandırılmasını sağlayan yaprak düğümdür. Ayrıca karar ağacına göre WHO giriş değeri 5'ten büyük olan hastalar direkt olarak hastanede yatış süresi boyunca yoğun bakıma ihtiyaç duyacak hastalar olarak sınıflandırılmıştır.



Şekil 3.19. Analiz – 3 73 özellikli Karar Ağacı Modeli

3.3.2. Analiz - 3 12 özellikle oluşturulan model sonuçları

Tablo 3.20'de kırmızı olarak işaretlenen 12 özellik kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.23 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

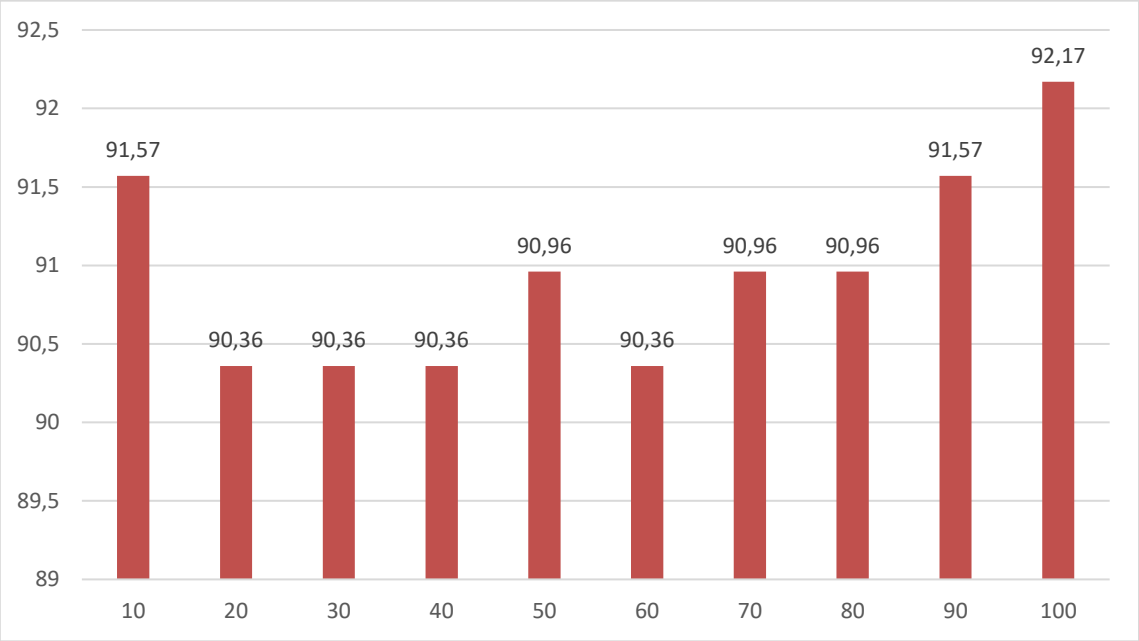
Tablo 3.23. Analiz – 3 12 özellik sınıflandırma performans sonuçları

Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	88,3	88,6	86,9	88,55
	Yok	88,3	88,6	86,9	88,55
Bagging	Var	90,6	91	90,3	90,96
	Yok	90,5	91	90,6	90,96
Rastgele Orman	**Var**	**91,9**	**92,2**	**91,7**	**92,17**
	Yok	**92**	**92,2**	**91,6**	**92,17**
Karar Ağacı	Var	91,2	91,6	91,3	91,57
	Yok	91,2	91,6	91,3	91,57

Sonuçlar detaylı olarak incelendiğinde en yüksek sınıflandırma doğruluğuna %92,17 ile Rastgele Orman algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman algoritmasının ardından Bagging %90,96 ve Karar Ağacı da %91,57 sınıflandırma doğruluğuna ulaşırken KNN %88,55 genel doğruluk değeri ile diğer yöntemlere göre daha kötü bir sınıflandırma performansı göstermiştir. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan modeller üzerinde pozitif bir etki oluşturmadığı yani sınıflandırma performansına olumlu bir katkı sağlamadığı görülmektedir. Bu yüzden 12 özellik kullanılarak hastanın hastanede yatış süreci boyunca yoğun bakım ihtiyacını en iyi tahmin eden yöntem standardizasyonun kullanılmadığı Rastgele Orman algoritması olarak seçilmiştir.

3.3.2.1 Analiz - 3 12 özellikli rastgele orman modeli

Weka üzerinde Rastgele Orman yöntemi için rastgele ormandaki ağaç sayısı, ağacın maksimum derinliği gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Ağacın maksimum derinliği sınırsız olarak belirlenerek farklı ağaç sayıları ile Rastgele Orman modelleri oluşturulmuştur. Şekil 3.20 Farklı ağaç sayıları kullanılarak elde edilen rastgele orman sınıflandırma doğruluk değerlerini temsil etmektedir. Standardizasyon işleminin uygulanmadığı veri setinde en yüksek sınıflandırma doğruluğuna %92,17 ile ağaç sayısının 100 olarak seçildiği modellerde ulaşmıştır.



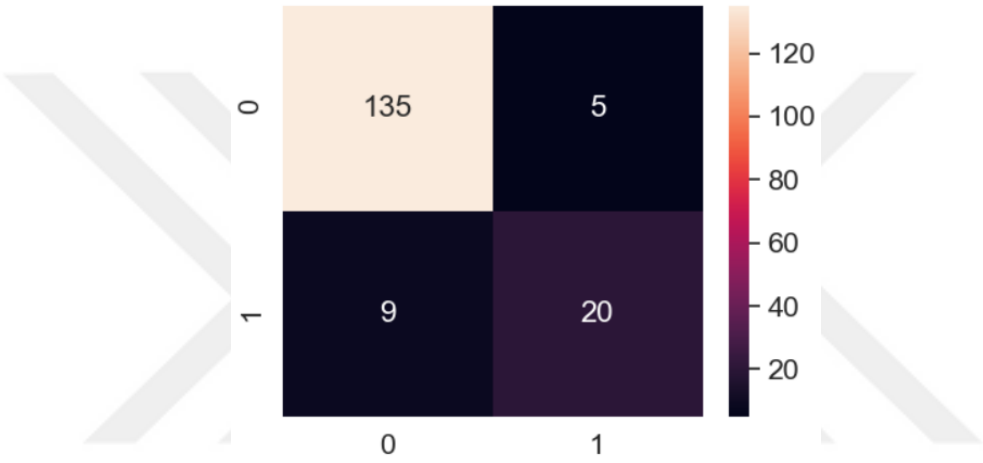
Şekil 3.20. Analiz – 3 12 özellik Rastgele Orman doğruluk değerleri

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.24 Rastgele Orman algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. "0" etiketli yoğun bakım ihtiyacı olmayan hastalar 0, 954 F1- skoru ile iyi bir doğruluk ile tahmin edilirken yoğun bakım ihtiyacı olan hastalar 0,735 ile ortalama bir F1-skoru performansı elde etmiştir.

Tablo 3.24. Analiz – 3 12 özellikli Rastgele Orman Modeli Performans Metrikleri

Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,93	0,99	0,95	0,71	0,96	0,99
1	0,90	0,62	0,74	0,71	0,96	0,86

Şekil 3.21'de gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde de beş yoğun bakım ihtiyacı olmayan hastanın ve dokuz yoğun bakım ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



Şekil 3.21. Analiz – 3 12 özellikli Rastgele Orman modeli karmaşıklık matrisi

Tez kapsamında daha az sayıda özellik ile her iki sınıf için de yüksek sınıflandırma doğruluklarına ulaşılması hedeflendiğinden dolayı 73 özellik ve 12 özellik ile oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Tablo 3.25 73 özellikle oluşturulan Karar Ağacı modeli ve 12 özellikle oluşturulan Rastgele Orman modelinin her iki sınıf için de ulaştığı performans değerlerini göstermektedir. Sonuçlar incelendiğinde %92,77 ile tüm özellikler kullanılarak daha yüksek bir doğruluk değeri elde edildiği fakat %92,17 genel doğruluk değeri ile sadece 12 özellik kullanılarak da yüksek performans ile hastanede yatış süresi boyunca yoğun bakım ihtiyacının tespit edilebildiği görülmüştür. Fakat tez çalışması sonucunda elde edilen model oksijen tedavisine ihtiyaç duyan hastaları 0,954 F1-Skoruyla iyi bir performans ile tahmin edebilirken oksijen tedavisine ihtiyaç duymayan hastalar veri sayısının az olmasından dolayı 0,735 F1-Skoru ile ortalama bir sınıflandırma performansına ulaşmıştır.

Tablo 3.25. Analiz – 3 73 özellik ve 13 özellik performans sonuçları

Model Adı	Doğruluk	Sınıf Bilgisi (Yoğun bakıma ihtiyaç)	F1-Skor
73 özellik Karar Ağacı	**%92,77**	VAR	**0,786**
		YOK	**0,957**
12 özellik Rastgele Orman	%92,17	VAR	0,735
		YOK	0,954

3.3.3. Analiz - 3 12 özellikli SMOTE uygulanmış model sonuçları

Tablo 3.20'de kırmızı olarak işaretli 12 özellikten oluşan veri seti 137 adet "0" etiketli hastadan oluşurken 29 adet "1" etiketli yani hastanede yatış aşaması boyunca oksijen ihtiyacı duyan hasta bulunmaktadır. Analiz kapsamında "1" sınıfı tahmin edilmesi gereken önemli bir sınıftır fakat diğer sınıfa göre daha az sayıda hasa bulunduğundan dolayı yüksek bir sınıflandırma doğruluğu elde edilmesine rağmen bu sınıfa ait F1-Skor değeri 0,735 ile 0,954 F1-Skor değerine sahip "0" sınıfından daha düşüktür. Bu yüzden 12 özellikten oluşan veri seti üzerinde SMOTE tekniği kullanılarak "1" etiketli sınıftaki hasta sayısı da 137 olacak şekilde çoğaltılarak dengeli bir veri seti elde edilmiştir. Bu veri seti kullanılarak KNN, Karar Ağacı, Rastgele Orman ve Bagging makine öğrenmesi yöntemleri üzerinde Z-Skor standardizasyon işlemlerinin uygulanıp uygulanmaması durumuna göre toplamda sekiz farklı model oluşturma işlemi gerçekleştirilmiştir. Tablo 3.26 Oluşturulan her bir modelin iki sınıf için de ağırlıklı ortalaması alınmış doğruluk, hassasiyet, duyarlılık ve F1-Skor performans metrikleri açısından karşılaştırmasını göstermektedir.

Sonuçlar detaylı olarak incelendiğinde 137 adet "0" ve 137 adet "1" etiketli hastadan oluşan veri setinde en yüksek sınıflandırma doğruluğuna %95,26 ile Rastgele Orman algoritması kullanılarak ulaşıldığı görülmektedir. Rastgele Orman algoritmasının ardından Bagging %92,7, Karar Ağacı %91,61 ve %KNN 94,16 genel doğruluk değeri ile yüksek bir sınıflandırma performansı göstermiştir. Ayrıca standardizasyon işleminin sınıflandırma performansına etkisi incelendiğinde z-skor standardizasyonun bu analizde oluşturulan modeller üzerinde pozitif bir etki oluşturmadığı yani sınıflandırma performansına olumlu bir katkı sağlamadığı görülmektedir. Nihai model olarak 12 özellik kullanılarak hastanın hastanede yatış süreci boyunca oksijen ihtiyacını en iyi tahmin eden

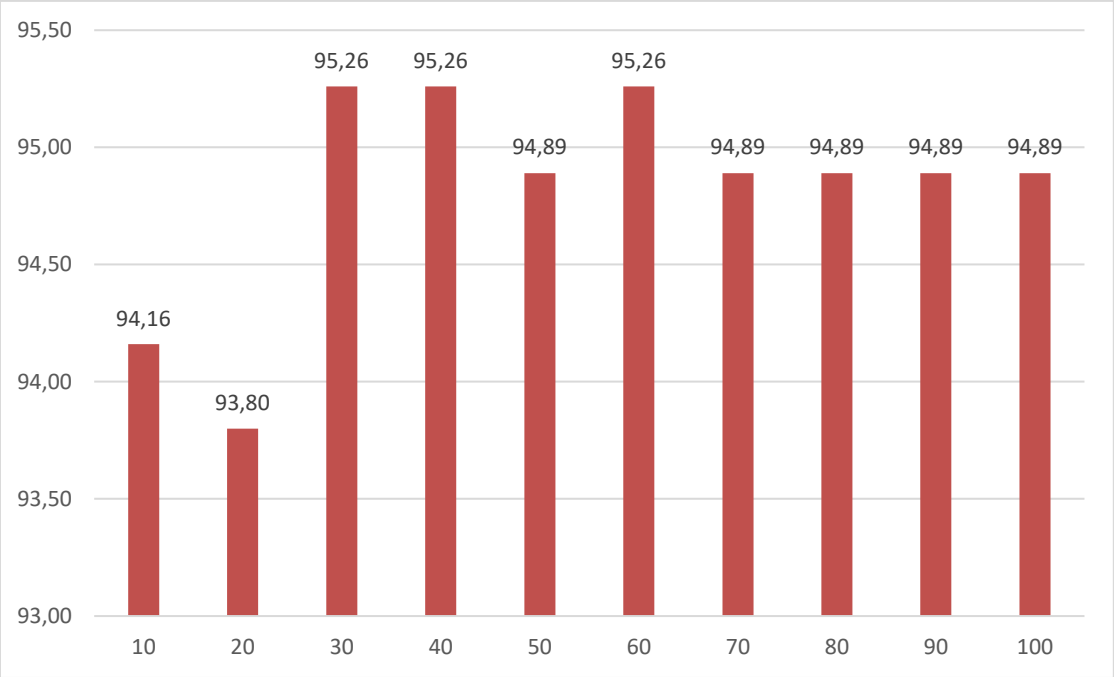
yöntem standardizasyonun kullanılmadığı Rastgele Orman algoritması olarak seçilmiştir.

Tablo 3.26. Analiz – 3 SMOTE 12 özellik sınıflandırma performans sonuçları

Algoritma	Standardizasyon (Var/Yok)	Hassasiyet (%)	Duyarlılık (%)	F1-Skor (%)	Doğruluk (%)
KNN	Var	94,2	94,2	94,2	94,16
	Yok	94,2	94,2	94,2	94,16
Bagging	Var	92,7	92,7	92,7	92,7
	Yok	92,7	92,7	92,7	92,7
Rastgele Orman	**Var**	**95,3**	**95,3**	**95,3**	**95,26**
	Yok	**95,3**	**95,3**	**95,3**	**95,26**
Karar Ağacı	Var	91,6	91,6	91,6	91,61
	Yok	91,6	91,6	91,6	91,61

3.3.3.1 Analiz - 3 SMOTE 12 özellikli rastgele orman modeli

Weka üzerinde Rastgele Orman yöntemi için rastgele ormandaki ağaç sayısı, ağacın maksimum derinliği gibi bazı kullanıcı tanımlı parametreler bulunmaktadır. Ağacın maksimum derinliği sınırsız olarak belirlenerek farklı ağaç sayıları ile Rastgele Orman modelleri oluşturulmuştur. Şekil 3.22 Farklı ağaç sayıları kullanılarak elde edilen rastgele orman sınıflandırma doğruluk değerlerini temsil etmektedir. Standardizasyon işleminin uygulanmadığı veri setinde en yüksek sınıflandırma doğruluğuna %95,26 ile ağaç sayısının 30,40 ve 60 olarak seçildiği modellerde ulaşmıştır. 30 ağaçtan oluşan Rastgele Orman modeli nihai model olarak seçilmiştir.



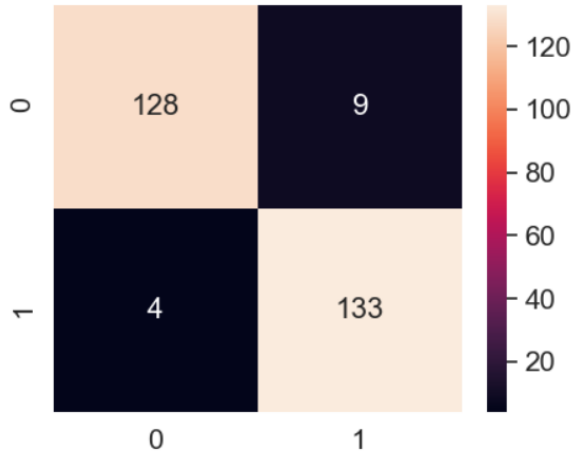
Şekil 3.22. Analiz – 3 SMOTE 12 özellik Rastgele Orman doğruluk değerleri

Sınıflandırma sonuçları doğruluk değerinin yanında hassasiyet, duyarlılık ve f1-skor ölçüm metrikleri de kullanılarak değerlendirilmiştir. Tablo 3.27 Rastgele Orman algoritması sonucunda oluşturulan modelin her bir sınıf için elde edilen performans metrik değerlerini göstermektedir. "0" etiketli oksijen ihtiyacı olmayan hastalar 0, 95 F1-skoru ile iyi bir doğruluk ile tahmin edilirken oksijen ihtiyacı olan hastalar da SMOTE tekniği uygulandıktan sonra 0,95 ile yüksek bir F1-skoru performansı elde etmiştir.

Tablo 3.27. Analiz – 3 SMOTE 12 özellik Rastgele Orman Performans Metrikleri

Sınıf	Hassasiyet	Duyarlılık	F1-Skor	MCC	ROC Alanı	PRC Alanı
0	0,97	0,93	0,95	0,91	0,99	0,99
1	0,94	0,97	0,95	0,91	0,99	0,99

Şekil 3.23'de gösterilen modele ait karmaşıklık matrisi değerlendirildiğinde de dört oksijen ihtiyacı olmayan hastanın ve bir oksijen ihtiyacı olan hastanın yanlış olarak sınıflandırıldığı görülmektedir.



Şekil 3.23. Analiz – 3 SMOTE 12 özellikli Rastgele Orman modeli karmaşıklık matrisi

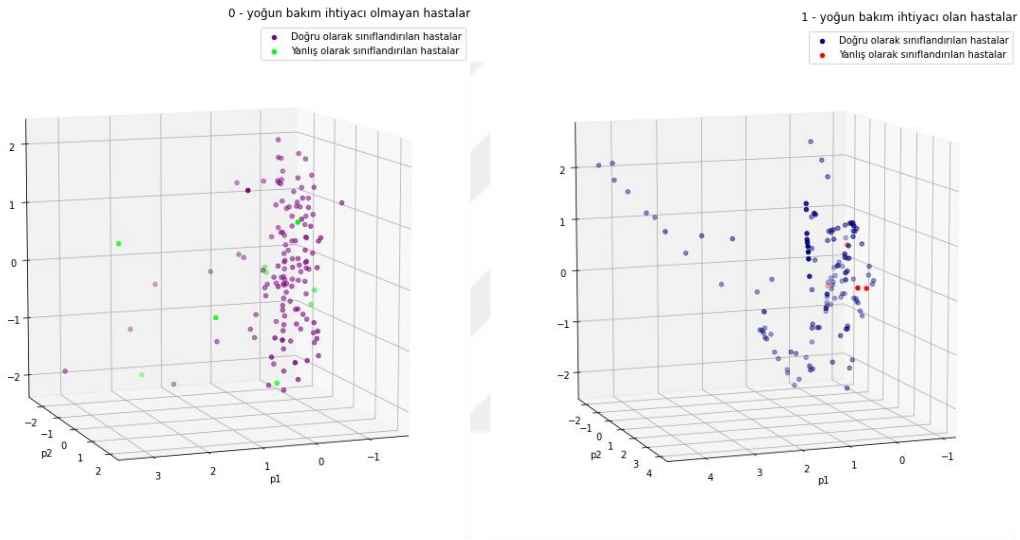
SMOTE tekniği ile "1" etiketli sınıf için de yüksek sınıflandırma doğruluklarına ulaşılması hedeflendiğinden dolayı SMOTE kullanılmayan veri seti ve SMOTE kullanılarak çoğaltılmış veri seti ile oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Tablo 3.28 12 özellikle oluşturulan Rastgele Orman modeli ve SMOTE uygulanarak 12 özellikle oluşturulan Rastgele Orman modelinin her iki sınıf için de ulaştığı performans değerlerini göstermektedir. Sonuçlar incelendiğinde %95,26 değeri ile SMOTE tekniği kullanıldıktan sonra genel doğruluk değerinde yükselme görülmüştür. Ayrıca SMOTE kullanılmadığı zaman 0,735 F1-Skoru ile tahmin edilen oksijen tedavisine ihtiyaç duyan hastalar 12 özellikli veri setine SMOTE uygulandıktan sonra 0,95 F1-Skor değerine yükselmiştir.

Tablo 3.28. Analiz – 3 12 özellik ve SMOTE 12 özellik performans sonuçları

Model Adı	Doğruluk	Sınıf Bilgisi (Yoğun bakıma ihtiyaç)	F1-Skor
12 özellik Rastgele Orman	%92,17	VAR	0,735
		YOK	0,954
SMOTE - 12 özellik Rastgele Orman	**%95,26**	VAR	0,95
		YOK	0,95

SMOTE yöntemi uygulanan veri seti üzerinde PCA analizi yapılarak toplam varyansının yaklaşık %96,8'sinin korunabilmesini sağlayan ilk 3 bileşen seçilerek sınıflandırma sonucu üç boyutlu bir grafik yardımıyla görselleştirilmiştir. Grafik üzerinde görselleştirme işlemi yapılacağından dolayı tüm bileşenlerin yerine sadece ana bileşenleri

kullanmak yüksek bir varyans oranı ile veri setinin gösterilmesinde yeterli olmuştur. Şekil 3.24 doğru ve yanlış olarak sınıflandırılan “0” ve “1” etiketli verilerin grafik üzerindeki dağılımını göstermektedir. SMOTE işlemi uygulandıktan sonra dokuz adet oksijen ihtiyacı olmayan hasta yanlış olarak sınıflandırılmıştır. Ayrıca SMOTE uygulanmadan önce yanlış olarak sınıflandırılan dokuz adet oksijen ihtiyacı olan hasta sayısı SMOTE uygulandıktan sonra dörde düşmüştür ve “1” etiketli sınıfın sınıflandırma performansında %74’ten %95’e artış gözlemlenmiştir. Yanlış olan sınıflandırılan “1” etiketli hastalardan üç tanesi SMOTE ile oluşturulmuş yapay verilerdir ve kalan 1 kayıt ise veri setinde bulunan gerçek bir hasta kaydına aittir.



Şekil 3.24. Analiz – 3 SMOTE 12 özellikli Rastgele Orman modeli sınıflandırma grafiği

4. SONUÇLAR

Covid-19; 11 Mart 2020’de Dünya Sağlık Örgütü tarafından pandemi olarak ilan edilmiştir ve SARS-CoV-2 virüsünün neden olduğu bir bulaşıcı hastalıktır. Enfekte hasta sayısındaki artış ile mücadele etmek için önlemler alınarak bilgisayar destekli sistemler de dahil olmak üzere birçok farklı alanda çalışmalar yapılmaktadır. Bu tez çalışması da bu kapsamda, Marmara Üniversitesi Hipertansiyon ve Ateroskleroz Eğitim, Uygulama ve Araştırma Merkezi koordinasyonunda, Etik Kurul onayı, 09.2020.487 sayılı protokol ve 27.04.2020 tarihli Sağlık Bakanlığı izni ile "Covid-19 hastalarında ilk başvurudaki verilerle hastalık şiddetinin makine öğrenmesi yöntemleri ile öngörülmesi" adlı proje kapsamında gerçekleştirildi. Hastanın ilk laboratuvar verileri ile WHO ölçüt değeri

arasındaki ilişki makine öğrenmesi teknikleri kullanılarak incelenerek bu zorlu süreçte sağlık sektörüne fayda sağlayabilecek anlamlı bilgiler elde edildi.

Çalışmanın ilk aşamasında makine öğrenmesi ve Covid-19 alanında yapılan çalışmalar detaylı bir şekilde incelenerek literatür araştırması yapıldı. Literatür araştırması sonucunda hangi makine öğrenmesi algoritmalarının kullanıldığı ve bu yöntemler kullanılarak hangi türde analizler yapıldığı ayrıntılı bir şekilde incelendi. Elde edilen bilgiler doğrultusunda tahmin problemlerinde sıklıkla kullanılan ve yüksek doğruluk değerlerine ulaşılmasını sağlayan KNN, Bagging, Rastgele Orman ve Karar Ağacı algoritmaları tez kapsamında kullanılacak makine öğrenmesi algoritmaları olarak belirlendi. Kullanılacak makine öğrenmesi yöntemleri belirlendikten sonra Nisan 2020 ve 1 Haziran 2020 tarihleri arasında İstanbul Marmara Üniversitesi Araştırma Hastanesi'nde Covid-19 için belirlenen bölümlere başvuran tüm hastaların 28 günlük takip süreçlerini içeren bilgiler kullanılarak laboratuvar ve tıbbi bilgi olmak üzere iki farklı veri seti oluşturuldu. İki veri seti T.C. alanı kullanılarak birleştirildi ve nihai veri seti elde edildi. 180 adet hastadan oluşan veri seti hastaların yaş, cinsiyet, tıbbi öykü ve sigara kullanma durumuna göre analiz edilerek sonuçlar görselleştirildi ve veri setinin kapsamının daha iyi bir şekilde anlaşılması sağlandı.

Çalışmanın ikinci aşaması olarak eksik veya tutarsız veriler içeren veri seti üzerinde veri temizleme işlemi yapılarak analize uygun bir hale getirildi. Tutarsız veriler manuel olarak veri setinden temizlenirken eksik veriler için gün aralıklarına göre doldurma ve sınıf ortalamasına göre doldurma olmak üzere iki farklı teknik kullanıldı. Gün aralıklarına göre eksik veri doldurma kapsamında alanında uzman sağlık çalışanları tarafından özelliklerin 75 tanesi için gün aralıkları değeri belirlendi ve bu değerler kullanılarak eksik veriler tamamlandı. Ortalamaya göre eksik veri doldurma işlemi ise her analizden önce belirlenen hedef değer kullanılarak uygulandı ve veri seti makine öğrenmesi yöntemlerinde kullanılmak üzere hazır hale getirildi.

İlk analizde hastanın hastaneye yatış aşamasındaki oksijen ihtiyacının tespit edilmesi hedeflendi. Bu kapsamda hastanın laboratuvar ve demografik bilgileri dahil olmak üzere 72 özellik ve WHO giriş hedef değerinden oluşan bir veri seti kullanıldı. RSFS algoritması sonucunda 72 özellik arasından 16 tanesi en önemli özellikler olarak seçildi. 72 özellik kullanılarak oluşturulan modellerden en yüksek doğruluk değerine %89,44 ile

Rastgele Orman algoritması ulaşırken, 16 özellik kullanılarak oluşturulan modellerden en yüksek doğruluk değerine %91,67 ile Rastgele Orman algoritması ulaştı. Sonuçlar RSFS algoritması sonucunda elde edilen 16 özellik kullanılarak daha yüksek bir doğruluk değeri elde edilebileceğini gösterdi.

İkinci analizde hastanın hastanede yatış süresi boyunca oksijen ihtiyacının tespit edilmesi hedeflendi. Bu kapsamda hastanın laboratuvar, demografik bilgileri ve WHO giriş dahil olmak üzere 73 özellik ve WHO en kötü hedef değerinden oluşan bir veri seti kullanıldı. RSFS algoritması sonucunda 73 özellik arasından 18 tanesi en önemli özellikler olarak seçildi. 73 özellik kullanılarak oluşturulan modellerden en yüksek doğruluk değerine %88,39 ile Karar Ağacı algoritması ulaşırken, 18 özellik kullanılarak oluşturulan modellerden en yüksek doğruluk değerine %91,96 ile Rastgele Orman algoritması ulaştı. Sonuçlar RSFS algoritması sonucunda elde edilen 18 özellik kullanılarak daha yüksek bir doğruluk değeri elde edilebileceğini gösterdi.

Üçüncü analizde hastanın hastanede yatış süresi boyunca yoğun bakım ihtiyacının tespit edilmesi hedeflendi. Bu kapsamda hastanın laboratuvar, demografik bilgileri ve WHO giriş dahil olmak üzere 73 özellik ve WHO en kötü hedef değerinden oluşan bir veri seti kullanıldı. RSFS algoritması sonucunda 73 özellik arasından 12 tanesi en önemli özellikler olarak seçildi. 73 özellik kullanılarak oluşturulan modellerden en yüksek doğruluk değerine %92,77 ile Karar Ağacı algoritması ulaşırken, 12 özellik kullanılarak oluşturulan modellerden en yüksek doğruluk değerine %92,17 ile Rastgele Orman algoritması ulaştı. Sonuçlar RSFS algoritması sonucunda elde edilen 12 özellik kullanılarak da aynı doğruluk değerine ulaşılabileceğini gösterdi.

Tablo 4.1. Analiz açıklaması ve doğruluk değerleri

Analiz	Özellik Sayısı	Hedef Değer	Algoritma	Doğruluk
Analiz - 1	72	WHO giriş	Rastgele Orman	89,44
Analiz - 1	16	WHO giriş	Rastgele Orman	**91,67**
Analiz - 2	73	WHO en kötü	Karar Ağacı	88,39
Analiz - 2	18	WHO en kötü	Rastgele Orman	**91,96**
Analiz - 2	73	WHO en kötü	Karar Ağacı	92,77
Analiz - 2	12	WHO en kötü	Rastgele Orman	**92,17**

Tablo 4.1 her bir analiz için tüm özellikler ve RSFS tarafından seçilen özellikler kullanılarak elde edilen doğruluk değerlerini göstermektedir. Her üç analizde RSFS özellik seçme algoritması tarafından seçilen ortak özellikler MDW ve Troponin T-hs idi. Ayrıca seçilen özellikler arasından Analiz-3 için WHO Giriş değeri yoğun bakım ihtiyacını tahmin etmek de önemli bir nitelik olarak seçilirken, oksijen ihtiyacını tahmin eden Analiz-3 kapsamında seçilen nitelikler arasında yer almamaktadır. WHO ölçüt değeri pandeminin başında ve öncesinde çoğunlukla yoğun bakım ihtiyacı olan ve hastalık şiddeti ciddi olan hastaları belirlemek üzere kullanılan bir ölçüt değeri olduğundan dolayı WHO ölçütünü oluşturan parametreler daha çok yoğun bakım durumunu belirleyen parametreler ile ilişkilidir. Oksijen ihtiyacını belirleyen parametrelerden daha bağımsız olması sonucunda da Analiz-2 kapsamında Rastgele Özellik seçme algoritması tarafından ilişkili özellikler arasında seçilmemesi fakat yoğun bakım ihtiyacını tahmin eden Analiz-3 kapsamında seçilmesi mantıklı ve sağlık uzmanları tarafından da tutarlı olarak kabul edilen bir sonuçtur.

Sonuçlar WHO ölçütü ile ilk laboratuvar değerlerinin kullanılarak üç analiz içinde yüksek doğruluk değerlerine ulaşıldığını göstermektedir. Ayrıca üç analiz içinde daha az sayıda özellik kullanılarak karşı karşıya kalınan bu zorlu dönemde sağlık sektörüne daha az bileşen kullanarak karar verme konusunda yardımcı olması sağlanmıştır. Elde edilen bu anlamlı bilgilerin yanında Covid-19 alanında yerli bir veri seti oluşturularak yerli veri seti çözümüne bir katkı sağlanmıştır.

Tüm analizlerde kullanılan veri setlerinde bulunan dengesiz veri kümesi dağılımı probleminden dolayı veri setine SMOTE algoritması uygulanmış ve dengeli bir veri kümesi elde edilmiştir. Elde edilen sonuçlar doğruluk ve her bir sınıf için elde edilen F1-Skor değerleri açısından karşılaştırılmıştır. Tablo 4.2 SMOTE algoritması kullanılmadan önce ve SMOTE algoritması uygulandıktan sonra elde edilen performans metrik değerlerini göstermektedir.

Sonuçlar incelendiğinde her üç analiz içinde genel sınıflandırma doğruluğunda SMOTE algoritması uygulandıktan gelişme olduğu görülmektedir. Ayrıca azınlık olan sınıfların F1-Skor değerleri karşılaştırıldığında Analiz – 1 için 89'ten 95'ye olacak şekilde %6'lik bir artış, Analiz – 2 için 74'ten 97'ye olacak şekilde %23'lik bir artış ve Analiz – 3 için 74'ten 95'e olacak şekilde %21'lik bir artış gözlendiği sonucuna varılmaktadır. Bu da

SMOTE algoritması uygulandıktan sonra azınlık olan sınıfın daha iyi bir sınıflandırma performansı ile tahmin edilebildiğini göstermektedir.

Tablo 4.2. Analiz – 1, Analiz – 2 ve Analiz – 3 SMOTE karşılaştırma sonuçları

Analiz Numarası	SMOTE	Doğruluk (%)	Sınıf Etiketi	F1- Skor (%)
1	YOK	91,67	Oksijen tedavisi ihtiyacı var	89
			Oksijen tedavisi ihtiyacı yok	93
	VAR	**96,64**	Oksijen tedavisi ihtiyacı var	**95**
			Oksijen tedavisi ihtiyacı yok	**95**
2	YOK	91,96	Oksijen tedavisi ihtiyacı var	74
			Oksijen tedavisi ihtiyacı yok	95
	VAR	**97,8**	Oksijen tedavisi ihtiyacı var	**98**
			Oksijen tedavisi ihtiyacı yok	**98**
3	YOK	92,17	Yoğun bakım ihtiyacı var	74
			Yoğun bakım ihtiyacı yok	95
	VAR	**95,26**	Yoğun bakım ihtiyacı var	**95**
			Yoğun bakım ihtiyacı yok	**95**

İlk olarak Aralık 2019 yılında ortaya çıkan ve küresel bir kriz haline gelen Covid-19 genellikle hafif ve asemptomatik bir hastalığa neden olurken, hastaların yaklaşık %10-15'i bu hastalığı ciddi bir şekilde geçirmiştir [7]. Bu yüzden hastalığı ciddi bir şekilde atlatan ve yüksek risk altındaki hastaların tespiti ve bakımı hastane ortamının ve kaynakların yönetimi için çok önemlidir.

CRP, ferritin, D-dimer, lenfosit sayısı gibi laboratuvar özellikleri ve WHO ölçeği gibi birçok parametre hastalığın şiddetini ve ölüm riskini öngörebilmesine rağmen klasik yöntemler ile birlikte bu parametrelerin kullanılması karmaşıktır ve uygun maliyetli değildir [88,89]. Bu çalışmada WHO ölçeğini ve laboratuvar sonuçlarını kullanarak yoğun bakım ve oksijen desteğinin gerekliliğini tahmin etmek için rastgele orman algoritması, KK, torbalama ve karar ağacı makine öğrenmesi yöntemleri kullanılarak farklı ML algoritmaları araştırıldı.

Uluslararası Klinik Kimya Federasyonu (IFCC), tek bir biyokimyasal veya hematolojik

belirtecin Covid-19 enfeksiyonunun sonucunu tahmin etmek için yeterince hassas veya spesifik olmadığını belgelemiştir [90]. Özellikle, IFCC, laboratuvar anormalliklerinin parametre gruplarına dayalı olarak yorumlanmasını önerir. Seçilen özellikler arasında CRP, albümin, toplam protein, ferritin, INR, ALP, D-dimer, troponin-T, RDW, GGT, MDW, pH, pCO2, fibrinojen bulunur. Ferritin ve D-dimer, Covid-19 hastalığının ilerlemesinde önemli parametreler olarak tanımlanmıştır ve bu nedenle bunlar, buradaki hemen hemen tüm modellere dahil edilmiştir [89,91,92]. D-dimer, şiddetli inflamatuar reaksiyonun bir sonucu olarak hiper-pıhtılaşabilirliğin göstergesidir [93]. D-dimer Covid-19 enfeksiyonunun ciddiyeti için umut verici bir belirteç gibi görünmektedir. aPTT, aynı nedenle ciddi Covid-19 enfeksiyonu için başka bir önemli öngörücüdür. Bu yüzden oluşturulan modellerin D-dimer ve aPTT gibi pıhtılaşabilirlik belirteçlerini tanıması umut verici bir sonuçtur.

Covid-19 teşhisi, ölüm oranı ve ciddiyet riski tahmininde makine öğrenimi yaklaşımlarını kullanarak yapılan çalışmalarda Covid-19 teşhisini tahmin etmek için sık bildirilen 5 özellik LYMPH, EOS, LDH, CRP ve yaş iken ölüm ve ciddiyet oranını tahmin etmek için sık bildirilen 5 özellik "Yaş, LDH, LYMPH, CRP ve oksijen satürasyonu" olmuştur. Çalışma kapsamında yapılan analizlerden seçilen özellikler ile literatürdeki çalışmalar sonucunda seçilen özellikler karşılaştırıldığında CRP, Albumin, Ferritin, D-dimer, Kalsiyum, NEU, WBC, Direkt bilirubin ve LDH niteliklerinin ortak olarak seçildiği görüşmüştür [94].

Analiz-1 ve Analiz-2'de hedeflenen solunum ihtiyacı ile en ilişkili özellikler arasında seçilen CRP ve LDH Olmedo ve arkadaşları tarafından yapılan çalışmada da en ilişkili olarak seçilen 4 özellik içerisinde yer almaktadır [95]. Ayrıca bu çalışma ile ortak olarak Arvind ve arkadaşları CRP, D-dimer ve lökosit/lenfosit sayısı (WBC) laboratuvar değişkenlerinin hastanede yatan COVID-19 hastalarında entübasyon riskini tahmin etmede daha yüksek bir öneme sahip olduğunu da göstermiştir [21].

Tez kapsamında oluşturulan bu algoritmaların sağlık kaynaklarının kullanımı için uygun bir strateji ve etkili bir triyaj protokolü, risk sınıflandırma yöntemi oluşturmaya izin vereceğini düşünüyoruz. Ek olarak yapılan analizler ve oluşturulan modeller temel laboratuvar testleri kullanılarak elde edildiğinden dolayı tıbbi kaynak sıkıntısı çeken gelişmekte olan bölgeler için değerli olacaktır. Ayrıca yoğun bakım ve oksijen gibi kritik

tıbbi kaynakların kullanımının planlanmasında faydalı bir çalışma olmaktadır. Tek bir hastanede yapılan retrospektif bir çalışma olması, klinik parametreler içerisinden sadece laboratuvar değişkenlerinin analiz edilmesi ve küçük örnek boyutu bu çalışmanın sahip olduğu sınırlandırmalara örnek olarak verilebilir.

5. ÖNERİLER

Çalışma tek bir hastanede yapılan retrospektif bir çalışma olduğundan dolayı bu kısıtlamaları ortadan kaldırmak ve daha iyi sonuçlar elde edebilmek için veri sayısı çoğaltılabilir böylece yapay veri çoğaltma algoritmaları kullanılmadan sınıflandırma işlemlerinde yüksek performanslı modeller oluşturulabilir. Özellik Seçme ve veri çoğaltmak için kullanılan RSFS ve SMOTE algoritmalarına alternatif olarak farklı teknikler de denenerek elde edilen sonuçlar daha net bir şekilde karşılaştırılabilir.

Ayrıca çalışma kapsamında ilgili çalışmalar incelendikten sonra seçilen 4 makine öğrenmesi yöntemi sınıflandırma işlemlerinde kullanılmıştır. Bu algoritmaların yanında farklı makine öğrenmesi, topluluk sınıflandırma yöntemleri ve derin öğrenme de eklenerek elde edilen sonuçlar karşılaştırılabilir.

KAYNAKLAR

[1] Cheruku, S.R., Barina, A., Kershaw, C.D., Goff, K., Reisch, J., Hynan, L.S., Ahmed, F., Armaignac, D.L., Patel, L., Belden, K.A. (2022) Palliative care consultation and end-of-life outcomes in hospitalized COVID-19 patients. Resuscitation, 170, 230-237.

[2] WHO characterizes COVID-19 as a pandemic, https://www3.paho.org/hq/index.php?option=com_content&view=article&id=15756:who-characterizes-covid-19-as-a-pandemic&Itemid=1926&lang=en, 23 Ocak 2022.

[3] Guarner, J. (2020) Three Emerging Coronaviruses in Two Decades. Am J Clin Pathol, 153, 420-421.

[4] World Health Organization. WHO coronavirus (COVID-19) emergency dashboard, https://covid19.who.int, 23 Ocak 2022.

[5] Fontanarosa, P.B., Bauchner, H. (2020) COVID-19—looking beyond tomorrow for health care and society. Jama, 323, 1907-1908.

[6] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395, 497-506.

[7] Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y. (2020) Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. Jama, 323, 1061-1069.

[8] Docherty, A.B., Harrison, E.M., Green, C.A., Hardwick, H.E., Pius, R., Norman, L., Holden, K.A., Read, J.M., Dondelinger, F., Carson, G., Merson, L., Lee, J., Plotkin, D., Sigfrid, L., Halpin, S., Jackson, C., Gamble, C., Horby, P.W., Nguyen-Van-Tam, J.S., Ho, A., Russell, C.D., Dunning, J., Openshaw, P.J., Baillie, J.K., Semple, M.G. (2020) Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. Bmj, 369, m1985.

[9] Petrilli, C.M., Jones, S.A., Yang, J., Rajagopalan, H., O'Donnell, L., Chernyak, Y., Tobin, K.A., Cerfolio, R.J., Francois, F., Horwitz, L.I. (2020) Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. Bmj, 369, m1966.

[10] Carfì, A., Bernabei, R., Landi, F. (2020) Persistent Symptoms in Patients After Acute COVID-19. Jama, 324, 603-605.

[11] Peckham, H., de Gruijter, N.M., Raine, C., Radziszewska, A., Ciurtin, C., Wedderburn, L.R., Rosser, E.C., Webb, K., Deakin, C.T. (2020) Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nat Commun, 11, 6317.

[12] Onder, G., Rezza, G., Brusaferro, S. (2020) Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. Jama, 323, 1775-1776.

[13] Wu, Z., McGoogan, J.M. (2020) Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. Jama, 323, 1239-1242.

[14] Kaur, S.P., Gupta, V. (2020) COVID-19 Vaccine: A comprehensive status report. Virus Res, 288, 198114.

[15] Milken Institute Covid-19 Tracker Vaccines, https://airtable.com/shrSAi6t5WFwqo3GM/tblEzPQS5fnc0FHYR/viweyymxOAtNvo7yH?blocks=bipZFzhJ7wHPv7x9z, 4 Ocak 2022.

[16] WORLD HEALTH ORGANIZATION (WHO) Covid 19 Track Vaccines, https://covid19.trackvaccines.org/agency/who/, 4 Ocak 2022.

[17] Türkiye'de COVID-19 aşılaması, https://tr.wikipedia.org/wiki/T%C3%BCrkiye%27de_COVID-19_a%C5%9F%C4%B1lamas%C4%B1, 4 Ocak 2022.

[18] (2020) A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis, 20, e192-e197.

[19] Su, Y., Ju, M.-j., Xie, R.-c., Yu, S.-j., Zheng, J.-l., Ma, G.-g., Liu, K., Ma, J.-f., Yu, K.-h., Tu, G.-w. (2021) Prognostic accuracy of early warning scores for clinical deterioration in patients with COVID-19. Frontiers in medicine, 7, 624255.

[20] Veldhuis, L., Ridderikhof, M.L., Schinkel, M., van den Bergh, J., Beudel, M., Dormans, T., Douma, R., van den Oever, N.G., de Haan, L., Koopman, K. (2021) Early

warning scores to assess the probability of critical illness in patients with COVID-19. Emergency Medicine Journal, 38, 901-905.

[21] Arvind, V., Kim, J.S., Cho, B.H., Geng, E., Cho, S.K. (2021) Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. J Crit Care, 62, 25-30.

[22] Burdick, H., Lam, C., Mataraso, S., Siefkas, A., Braden, G., Dellinger, R.P., McCoy, A., Vincent, J.L., Green-Saxena, A., Barnes, G., Hoffman, J., Calvert, J., Pellegrini, E., Das, R. (2020) Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. Comput Biol Med, 124, 103949.

[23] 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.

[24] Soltan, A., Kouchaki, S., Zhu, T., Kiyasseh, D., Taylor, T., Hussain, Z., Peto, T., Brent, A., Eyre, D., Clifton, D. (2020) Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. The Lancet Digital Health, 3.

[25] 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. Int Immunopharmacol, 86, 106705.

[26] Mondal, M.R.H., Bharati, S., Podder, P., Podder, P. (2020) Data analytics for novel coronavirus disease. Informatics in Medicine Unlocked, 20, 100374.

[27] Heydari, M.H., Kahrobaee, S., YazdanPanah, Y., Noori, A., Sarbisheie, G., (2020) Clustering of Infected Patients by COVID-19 Using Self-Organized Mapping and Extracting the Most Important Clinical Features, 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), 1-6.

[28] Yadaw, A.S., Li, Y.C., Bose, S., Iyengar, R., Bunyavanich, S., Pandey, G. (2020) Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health, 2, e516-e525.

[29] Di Castelnuovo, A., Bonaccio, M., Costanzo, S., Gialluisi, A., Antinori, A., Berselli,

N., Blandi, L., Bruno, R., Cauda, R., Guaraldi, G., My, I., Menicanti, L., Parruti, G., Patti, G., Perlini, S., Santilli, F., Signorelli, C., Stefanini, G.G., Vergori, A., Abdeddaim, A., Ageno, W., Agodi, A., Agostoni, P., Aiello, L., Al Moghazi, S., Aucella, F., Barbieri, G., Bartoloni, A., Bologna, C., Bonfanti, P., Brancati, S., Cacciatore, F., Caiano, L., Cannata, F., Carrozzi, L., Cascio, A., Cingolani, A., Cipollone, F., Colomba, C., Crisetti, A., Crosta, F., Danzi, G.B., D'Ardes, D., de Gaetano Donati, K., Di Gennaro, F., Di Palma, G., Di Tano, G., Fantoni, M., Filippini, T., Fioretto, P., Fusco, F.M., Gentile, I., Grisafi, L., Guarnieri, G., Landi, F., Larizza, G., Leone, A., Maccagni, G., Maccarella, S., Mapelli, M., Maragna, R., Marcucci, R., Maresca, G., Marotta, C., Marra, L., Mastroianni, F., Mengozzi, A., Menichetti, F., Milic, J., Murri, R., Montineri, A., Mussinelli, R., Mussini, C., Musso, M., Odone, A., Olivieri, M., Pasi, E., Petri, F., Pinchera, B., Pivato, C.A., Pizzi, R., Poletti, V., Raffaelli, F., Ravaglia, C., Righetti, G., Rognoni, A., Rossato, M., Rossi, M., Sabena, A., Salinaro, F., Sangiovanni, V., Sanrocco, C., Scarafino, A., Scorzolini, L., Sgariglia, R., Simeone, P.G., Spinoni, E., Torti, C., Trecarichi, E.M., Vezzani, F., Veronesi, G., Vettor, R., Vianello, A., Vinceti, M., De Caterina, R., Iacoviello, L. (2020) Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. Nutr Metab Cardiovasc Dis, 30, 1899-1913.

[30] Ahamad, M.M., Aktar, S., Rashed-Al-Mahfuz, M., Uddin, S., Liò, P., Xu, H., Summers, M.A., Quinn, J.M.W., Moni, M.A. (2020) A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Syst Appl, 160, 113661.

[31] Li, M., Zhang, Z., Cao, W., Liu, Y., Du, B., Chen, C., Liu, Q., Uddin, M.N., Jiang, S., Chen, C., Zhang, Y., Wang, X. (2021) Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. Sci Total Environ, 764, 142810.

[32] Otoom, M., Otoum, N., Alzubaidi, M.A., Etoom, Y., Banihani, R. (2020) An IoT-based framework for early identification and monitoring of COVID-19 cases. Biomed Signal Process Control, 62, 102149.

[33] Althnian, A., Abou Elwafa, A., Aloboud, N., Alrasheed, H., Kurdi, H. (2020)

Prediction of COVID-19 individual susceptibility using demographic data: A case study on Saudi Arabia. Procedia Computer Science, 177, 379-386.

[34] Bansal, V., Pahwa, G., Kannan, N., (2020) Cough Classification for COVID-19 based on audio mfcc features using Convolutional Neural Networks, 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 604-608.

[35] Yang, H., Zhang, M., Shen, L., Wang, Q., Chen, W., Liu, C., Hong, M., (2021) Mmfc: Multi-Modal Fusion Cascade Framework For Covid-19 Disease Course Classification, 2021 IEEE International Conference on Image Processing (ICIP), 190-194.

[36] Sun, L., Mo, Z., Yan, F., Xia, L., Shan, F., Ding, Z., Song, B., Gao, W., Shao, W., Shi, F. (2020) Adaptive feature selection guided deep forest for covid-19 classification with chest ct. IEEE Journal of Biomedical and Health Informatics, 24, 2798-2805.

[37] Hussain, E., Hasan, M., Rahman, M.A., Lee, I., Tamanna, T., Parvez, M.Z. (2021) CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images. Chaos, Solitons & Fractals, 142, 110495.

[38] Lai, Y., Li, G., Wu, D., Lian, W., Li, C., Tian, J., Ma, X., Chen, H., Xu, W., Wei, J. (2020) 2019 Novel coronavirus-infected pneumonia on CT: a feasibility study of few-shot learning for computerized diagnosis of emergency diseases. Ieee Access, 8, 194158-194165.

[39] Saha, P., Sadi, M.S., Islam, M.M. (2021) EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers. Informatics in medicine unlocked, 22, 100505.

[40] Ohata, E.F., Bezerra, G.M., Chagas, J.V.S.d., Neto, A.V.L., Albuquerque, A.B., Albuquerque, V.H.C.d., Filho, P.P.R. (2021) Automatic detection of COVID-19 infection using chest X-ray images through transfer learning. IEEE/CAA Journal of Automatica Sinica, 8, 239-248.

[41] Tuncer, T., Dogan, S., Ozyurt, F. (2020) An automated Residual Exemplar Local Binary Pattern and iterative ReliefF based COVID-19 detection method using chest X-ray image. Chemometrics and Intelligent Laboratory Systems, 203, 104054.

[42] El-Kenawy, E.-S.M., Ibrahim, A., Mirjalili, S., Eid, M.M., Hussein, S.E. (2020) Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. IEEE Access, 8, 179317-179335.

[43] Javor, D., Kaplan, H., Kaplan, A., Puchner, S., Krestan, C., Baltzer, P. (2020) Deep learning analysis provides accurate COVID-19 diagnosis on chest computed tomography. European journal of radiology, 133, 109402.

[44] Toğaçar, M., Ergen, B., Cömert, Z. (2020) COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. Computers in biology and medicine, 121, 103805.

[45] Wadhwa, P., Tripathi, A., Singh, P., Diwakar, M., Kumar, N. (2021) Predicting the time period of extension of lockdown due to increase in rate of COVID-19 cases in India using machine learning. Materials Today: Proceedings, 37, 2617-2622.

[46] Ballı, S. (2021) Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. Chaos, Solitons & Fractals, 142, 110512.

[47] Yeşilkanat, C.M. (2020) Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos, Solitons & Fractals, 140, 110210.

[48] Wang, P., Zheng, X., Li, J., Zhu, B. (2020) Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos, Solitons & Fractals, 139, 110058.

[49] Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.-W., Aslam, W., Choi, G.S. (2020) COVID-19 future forecasting using supervised machine learning models. IEEE access, 8, 101489-101499.

[50] Napoleon, D., Pavalakodi, S. (2011) A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set. International Journal of Computer Applications, 13, 41-46.

[51] Wei, J., (2010) Research on data preprocessing in supermarket customers data mining, 2010 2nd International Conference on Information Engineering and Computer Science. IEEE, 1-4.

[52] Kaiser, J. (2014) Dealing with Missing Values in Data. Journal of Systems Integration (1804-2724), 5.

[53] García, S., Luengo, J., Herrera, F. (2015) Data preprocessing in data mining. Springer.

[54] Pereira, R.B., Plastino, A., Zadrozny, B., Merschmann, L.H. (2018) Categorizing feature selection methods for multi-label classification. Artificial Intelligence Review, 49, 57-78.

[55] Jingze, X., Zuohong, W., Yan, X., Jianxing, Z. (2019) Face recognition based on PCA LDA and SVM algorithm. Computer engineering and application, 55, 34-37.

[56] Dong, H., Xuejun, W. (2018) Multi Pose Face recognition Based on Improved Convolutional Neural Network [J]. Journal of Jilin University: Information Science Edition, 36, 376-381.

[57] Wang, Y., Wu, Q., (2022) Research on Face Recognition Technology Based on PCA and SVM, 2022 7th International Conference on Big Data Analytics (ICBDA). IEEE, 248-252.

[58] Jaadi, Z.: A Step-by-Step Explanation of Principal Component Analysis (PCA), https://builtin.com/data-science/step-step-explanation-principal-component-analysis,

[59] Han, J., Pei, J., Kamber, M. (2011) Data mining: concepts and techniques. Elsevier.

[60] Li, C. (2019) Preprocessing methods and pipelines of data mining: An overview. arXiv preprint arXiv:1906.08510.

[61] Jain, A., Nandakumar, K., Ross, A. (2005) Score normalization in multimodal biometric systems. Pattern recognition, 38, 2270-2285.

[62] Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R. (2013) Data mining for scientific and engineering applications. Springer Science & Business Media.

[63] Padmaja, D.L., Vishnuvardhan, B., (2016) Comparative study of feature subset selection methods for dimensionality reduction on scientific data, 2016 IEEE 6th International Conference on Advanced Computing (IACC). IEEE, 31-34.

[64] Read, B.J. (1999) Data mining and science? Knowledge discovery in science as opposed to business.

[65] Pelayo, L., Dick, S., (2007) Applying novel resampling strategies to software defect prediction, NAFIPS 2007-2007 Annual meeting of the North American fuzzy information processing society. IEEE, 69-72.

[66] 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.

[67] Aydilek, İ.B. (2018) Yazılım hata tahmininde kullanılan metriklerin karar ağaçlarındaki bilgi kazançlarının incelenmesi ve iyileştirilmesi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24, 906-914.

[68] Başarslan, M.S., Argun, İ.D., (2018) Classification Of a bank data set on various data mining platforms, 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 1-4.

[69] Başarslan, M., Kayaalp, F. (2017) Customer churn analysis with classification algorithms in telecommunication sector.

[70] Rong, G., Mendez, A., Assi, E.B., Zhao, B., Sawan, M. (2020) Artificial intelligence in healthcare: review and prediction case studies. Engineering, 6, 291-301.

[71] Venkatkumar, I.A., Shardaben, S.J.K., (2016) Comparative study of data mining clustering algorithms, 2016 International Conference on Data Science and Engineering (ICDSE). IEEE, 1-7.

[72] Mazlan, A.U., binti Sahabudin, N.A., Remli, M.A., Ismail, N.S.N., Mohamad, M.S., Abd Warif, N.B., (2021) Supervised and unsupervised machine learning for cancer classification: recent development, 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS). IEEE, 392-395.

[73] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I. (2015) Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

[74] Nasteski, V. (2017) An overview of the supervised machine learning methods. Horizons. b, 4, 51-62.

[75] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H.,

Wang, Y. (2017) Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2.

[76] Murali[1], N., Sivakumaran, N. (2018) Artificial intelligence in healthcare–a review.

[77] Dalal, K.R., (2020) Analysing the role of supervised and unsupervised machine learning in iot, 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 75-79.

[78] Quinlan, J.R. (1987) Simplifying decision trees. International journal of man-machine studies, 27, 221-234.

[79] Novakovic, J., (2011) Bagging algorithm for pixel classification, 2011 19thTelecommunications Forum (TELFOR) Proceedings of Papers. IEEE, 1348-1351.

[80] Xing, B., Zhang, H., Zhang, K., Zhang, L., Wu, X., Shi, X., Yu, S., Zhang, S. (2019) Exploiting EEG signals and audiovisual feature fusion for video emotion recognition. IEEE Access, 7, 59844-59861.

[81] Sun, Y., Zhang, H., Zhao, T., Zou, Z., Shen, B., Yang, L. (2020) A new convolutional neural network with random forest method for hydrogen sensor fault diagnosis. IEEE Access, 8, 85421-85430.

[82] Breiman, L. (1996) Bagging predictors. Machine learning, 24, 123-140.

[83] Landron, C. (1981) Development of an amorphous film microanalysis method. Thin Solid Films, 84, 143-144.

[84] Bhargava, N., Sharma, S., Purohit, R., Rathore, P.S., (2017) Prediction of recurrence cancer using J48 algorithm, 2017 2nd International Conference on Communication and Electronics Systems (ICCES). IEEE, 386-390.

[85] Maria Navin, J., Pankaja, R. (2016) Performance analysis of text classification algorithms using confusion matrix. International Journal of Engineering and Technical Research (IJETR), 6, 75-78.

[86] Zhang, P., Su, W., (2012) Statistical inference on recall, precision and average precision under random selection, 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 1348-1352.

[87] Bahashwan, A.A., Anbar, M., Hasbullah, I.H., Alashhab, Z.R., Bin-Salem, A. (2021)

Flow-based approach to detect abnormal behavior in neighbor discovery protocol (NDP). IEEE Access, 9, 45512-45526.

[88] Gao, Y., Cai, G.-Y., Fang, W., Li, H.-Y., Wang, S.-Y., Chen, L., Yu, Y., Liu, D., Xu, S., Cui, P.-F. (2020) Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nature communications, 11, 1-10.

[89] Huang, I., Pranata, R., Lim, M.A., Oehadian, A., Alisjahbana, B. (2020) C-reactive protein, procalcitonin, D-dimer, and ferritin in severe coronavirus disease-2019: a meta-analysis. Therapeutic advances in respiratory disease, 14, 1753466620937175.

[90] Bohn, M.K., Lippi, G., Horvath, A., Sethi, S., Koch, D., Ferrari, M., Wang, C.-B., Mancini, N., Steele, S., Adeli, K. (2020) Molecular, serological, and biochemical diagnosis and monitoring of COVID-19: IFCC taskforce evaluation of the latest evidence. Clinical Chemistry and Laboratory Medicine (CCLM), 58, 1037-1052.

[91] Terpos, E., Ntanasis-Stathopoulos, I., Elalamy, I., Kastritis, E., Sergentanis, T.N., Politou, M., Psaltopoulou, T., Gerotziafas, G., Dimopoulos, M.A. (2020) Hematological findings and complications of COVID-19. American journal of hematology, 95, 834-847.

[92] Mudatsir, M., Fajar, J.K., Wulandari, L., Soegiarto, G., Ilmawan, M., Purnamasari, Y., Mahdi, B.A., Jayanto, G.D., Suhendra, S., Setianingsih, Y.A. (2020) Predictors of COVID-19 severity: a systematic review and meta-analysis. F1000Research, 9.

[93] Griffin, D.O., Jensen, A., Khan, M., Chin, J., Chin, K., Saad, J., Parnell, R., Awwad, C., Patel, D. (2020) Pulmonary embolism and increased levels of d-dimer in patients with coronavirus disease. Emerging infectious diseases, 26, 1941.

[94] Alballa, N., Al-Turaiki, I. (2021) Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: a review. Informatics in Medicine Unlocked, 24, 100564.

[95] Domínguez-Olmedo, J.L., Gragera-Martínez, Á., Mata, J., Álvarez, V.P. (2021) Machine learning applied to clinical laboratory data in Spain for COVID-19 outcome prediction: model development and validation. Journal of medical Internet research, 23, e26211.

EKLER

Ek 2 Tablo 1 Laboratuvar veri setinde bulunan 191 özellik

Özellikler			
ORNEKNO	TC_KIMLIKNO	ADSOYAD	CINSIYET
Albumin	ALP	ALT	aPTT
BAS%	Baz Açığı	Bilirubin	BUN
CRP (Nefelometrik)	D-dimer	Direkt Bilirubin	EGC
EOS#	EOS%	Ferritin	Fibrinojen
GGT	Glukoz	HCO3-	HCT
Hemoglobin	HGB	INR	K+
Karboksihemoglobin	Klor	Kortizol	Kreatinin
Laktat	LDH	LHD	Lipaz
LYMPH%	MAF	Magnezyum	MCH
MCV	MDW	Methemoglobin	MN-AL2-EGC
MN-AL2-LY	MN-AL2-MO	MN-AL2-NE	MN-AL2-NNRBC
MN-C-EGC	MN-C-EO	MN-C-LY	MN-C-MO
MN-C-NNRBC	MN-C-NRBC	MN-LALS-EGC	MN-LALS-EO
MN-LALS-MO	MN-LALS-NE	MN-LALS-NNRBC	MN-LALS-NRBC
MN-LMALS-EO	MN-LMALS-LY	MN-LMALS-MO	MN-LMALS-NE
MN-LMALS-NRBC	MN-MALS-EGC	MN-MALS-EO	MN-MALS-LY
MN-MALS-NE	MN-MALS-NNRBC	MN-MALS-NRBC	MN-UMALS-EGC
MN-UMALS-LY	MN-UMALS-MO	MN-UMALS-NE	MN-UMALS-NNRBC
MN-V-EGC	MN-V-EO	MN-V-LY	MN-V-MO
MN-V-NNRBC	MN-V-NRBC	MON#	MON%
Na+	NEU#	NEU%	NRBC
NT-proBNP	Ozmolarite	pCO2	PCT
pH	PLT	pO2	Potasyum
PT	PT Zamanı %	RBC	RDW
SD-AL2-EO	SD-AL2-LY	SD-AL2-MO	SD-AL2-NE

SD-AL2-NRBC	SD-C-EGC	SD-C-EO	SD-C-LY
SD-C-NE	SD-C-NNRBC	SD-C-NRBC	SD-LALS-EGC
SD-LALS-LY	SD-LALS-MO	SD-LALS-NE	SD-LALS-NNRBC
SD-LMALS-EGC	SD-LMALS-EO	SD-LMALS-LY	SD-LMALS-MO
SD-LMALS-NNRBC	SD-LMALS-NRBC	SD-MALS-EGC	SD-MALS-EO
SD-MALS-MO	SD-MALS-NE	SD-MALS-NNRBC	SD-MALS-NRBC
SD-UMALS-EO	SD-UMALS-LY	SD-UMALS-MO	SD-UMALS-NE
SD-UMALS-NRBC	SD-V-EGC	SD-V-EO	SD-V-LY
SD-V-NE	SD-V-NNRBC	SD-V-NRBC	Serbest T4
Sodyum	Tiroglobulin	Total Bilirubin	Total Demir Bağlama Kapasitesi
Troponin T-hs	TSH	Üre	Ürik Asit
WBC	WDOP	WNOP	CALISMA_TARIHI
Cl-	MN-AL2-NRBC	MPV	SD-UMALS-EGC
YAS	MN-C-NE	NRBC#	SD-UMALS-NNRBC
AST	MN-LALS-LY	PDW	SD-V-MO
Ca++	MN-LMALS-EGC	Prokalsitonin	sO2
EGC#	MN-LMALS-NNRBC	SD-AL2-EGC	Total Protein
Fosfor	MN-MALS-MO	SD-AL2-NNRBC	Vitamin B12
Hematokrit	MN-UMALS-EO	SD-C-MO	BAS#
Kalsiyum	MN-UMALS-NRBC	SD-LALS-EO	SD-LMALS-NE
Kütle CK-MB	MN-V-NE	SD-LALS-NRBC	SD-MALS-LY
LYMPH#	MCHC	MN-AL2-EO	


Ek 1.pdf


Ek 2.pdf


Ek 3.pdf

ÖZGEÇMİŞ

Adı Soyadı : Kübra KÖKSAL

Ana Dili : Türkçe

Yabancı Dili : İngilizce

Öğrenim Durumu

Derece	Bölüm/Program	Üniversite/Lise	Mezuniyet Yılı
Lise	Sayısal	Maltepe Anadolu Lisesi	2015
Lisans	Bilgisayar Mühendisliği	Kocaeli Üniversitesi	2019
Yüksek Lisans	Bilgisayar Mühendisliği	Marmara Üniversitesi	2022

İş Deneyimi

Yıl	Firma/Kurum	Görevi
Ocak 2020 - Nisan 2020	ETIYA	Yazılım Uzmanı
Nisan 2020 -	Huawei	Yazılım Mühendisi