T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
AKILLI SİSTEMLER MÜHENDİSLİĞİ ANA BİLİM DALI

VERİ MADENCİLİĞİNDE SINIFLANDIRMA VE KÜMELEME ALGORİTMALARI İLE COVID-19 ŞÜPHESİ TAŞIYAN HASTALARIN DEĞERLENDİRİLMESİ

Yüksek Lisans Tezi

Beyza DURMAZ

Danışman
Doç. Dr. Aslı ÇALIŞ BOYACI

SAMSUN
2024

T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
AKILLI SİSTEMLER MÜHENDİSLİĞİ ANA BİLİM DALI

VERİ MADENCİLİĞİNDE SINIFLANDIRMA VE KÜMELEME ALGORİTMALARI İLE COVID-19 ŞÜPHESİ TAŞIYAN HASTALARIN DEĞERLENDİRİLMESİ

Yüksek Lisans Tezi

Beyza DURMAZ

Danışman

Doç. Dr. Aslı ÇALIŞ BOYACI

SAMSUN
2024

TEZ KABUL VE ONAYI

Beyza DURMAZ tarafından, **Doç. Dr. Aslı ÇALIŞ BOYACI** danışmanlığında hazırlanan **"VERİ MADENCİLİĞİNDE SINIFLANDIRMA VE KÜMELEME ALGORİTMALARI İLE COVID-19 ŞÜPHESİ TAŞIYAN HASTALARIN DEĞERLENDİRİLMESİ"** başlıklı bu çalışma, jürimiz tarafından 10.5.2024 tarihinde yapılan sınav sonucunda oy birliği ile başarılı bulunarak Yüksek Lisans Tezi olarak kabul edilmiştir.

	Unvanı Adı Soyadı **Üniversitesi** **Ana Bilim/Ana Sanat Dalı**	**Sonuç**
Başkan	Doç. Dr. Naci MURAT Ondokuz Mayıs Üniversitesi Endüstri Mühendisliği Ana Bilim Dalı	☒ Kabul ☐ Ret
Üye	Dr. Öğr. Üyesi Aycan PEKPAZAR Samsun Üniversitesi Endüstri Mühendisliği Ana Bilim Dalı	☒ Kabul ☐ Ret
Üye	Doç. Dr. Aslı ÇALIŞ BOYACI Ondokuz Mayıs Üniversitesi Yöneylem Araştırması Ana Bilim Dalı	☒ Kabul ☐ Ret

Bu tez, Enstitü Yönetim Kurulunca belirlenen ve yukarıda adları yazılı jüri üyeleri tarafından uygun görülmüştür.

Prof. Dr. Ahmet TABAK
Enstitü Müdürü

BİLİMSEL ETİĞE UYGUNLUK BEYANI

Hazırladığım Yüksek Lisans tezinin bütün aşamalarında bilimsel etiğe ve akademik kurallara riayet ettiğimi, çalışmada doğrudan veya dolaylı olarak kullandığım her alıntıya kaynak gösterdiğimi ve yararlandığım eserlerin Kaynaklar'da gösterilenlerden oluştuğunu, her unsurun enstitü yazım kılavuzuna uygun yazıldığını ve TÜBİTAK Araştırma ve Yayın Etiği Kurulu Yönetmeliği'nin 3. bölüm 9. maddesinde belirtilen durumlara aykırı davranılmadığını taahhüt ve beyan ederim.

Etik Kurul Gerekli mi ?

Evet ☒ (Gerekli ise ekler kısmına ekleyiniz)

Hayır ☐

5/1/ 2024
Beyza DURMAZ

TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI

Tez Başlığı : VERİ MADENCİLİĞİNDE SINIFLANDIRMA VE KÜMELEME ALGORİTMALARI İLE COVID-19 ŞÜPHESİ TAŞIYAN HASTALARIN DEĞERLENDİRİLMESİ

Yukarıda başlığı belirtilen tez çalışması için şahsım tarafından 5/1/2024 tarihinde intihal tespit programından alınmış olan özgünlük raporu sonucunda;

Benzerlik oranı : % 15

Tek kaynak oranı : % 3 çıkmıştır.

5 /1/ 2024
Doç. Dr. Aslı ÇALIŞ BOYACI

ÖZET

VERİ MADENCİLİĞİNDE SINIFLANDIRMA VE KÜMELEME ALGORİTMALARI İLE COVID-19 ŞÜPHESİ TAŞIYAN HASTALARIN DEĞERLENDİRİLMESİ

Beyza DURMAZ
Ondokuz Mayıs Üniversitesi
Lisansüstü Eğitim Enstitüsü
Akıllı Sistemler Mühendisliği Ana Bilim Dalı
Yüksek Lisans, Mayıs/2024
Danışman: Doç. Dr. Aslı ÇALIŞ BOYACI

Veri madenciliği büyük hacimdeki karmaşık verilerin işlenebilmesine olanak sağlayarak, anlamlı bilgi ve örüntüleri ortaya çıkararır. Böylece karar vericilerin karar süreçlerine destek olur. Günümüzde karar verme sürecine ihtiyaç duyulan birçok sektörde olduğu gibi sağlık sektöründe de veri madenciliği teknikleri yaygın bir şekilde kullanılmaktadır.

Koronavirüs pandemisi ilk günden itibaren tüm dünyayı fiziksel, mental ve ekonomik olarak tehdit etmiştir. Türkiye'de etkileri ağırlıklı olarak 2020-2022 yıllarında hissedilmiş olsa da, Koronavirüs hastalığı yok olmamış olup etkilerini sürdürmektedir. Koronavirüs hastalığı ve pandemi dönemi hala anlaşılmaya çalışılmakta, konuyla ilgili güncel çalışma ve yayımlar devam etmektedir. Bu bağlamda hastalığı anlamak; teşhis konulması, bireylere ve ülkelere olan etkilerin azaltılması ve alınacak önlemler açısından son derece kritik bir rol oynamaktadır.

Bu çalışmada, COVID-19 şüphesi ile hastaneye başvurarak test yaptırmış olan hastaların sınıflandırma algoritmaları kullanılarak test sonuçlarının tahmin edilmesi ve kümeleme analizi ile test sonucu pozitif ve negatif olan hastalara ilişkin değerlendirmelerin yapılması amaçlanmaktadır. Sınıflandırma için oluşturulan veri seti karar ağacı, K-en yakın komşu, Naive Bayes, lojistik regresyon, rastgele orman ve destek vektör makineleri algoritmalarıyla KNIME 5.2.2 kullanılarak modellenmiştir. En yüksek doğruluk oranı 0,796 ve en yüksek tanısal üstünlük oranı 15,340 ile rastgele orman modelinde elde edilmiştir. En düşük doğruluk 0,690 ve en düşük tanısal üstünlük oranı 5,729 değeri ile K-en yakın komşu modelinde elde edilmiştir. Kümeleme aşamasında veri seti, COVID-19 test sonucu negatif ve pozitif olan hastalar için ayrılmış, iki ayrı iki aşamalı kümeleme analizi gerçekleştirilmiştir. Uygulama için SPSS Clementine 12.0 programı kullanılmıştır. Her iki durumda da iki küme elde edilmiştir.

Anahtar Sözcükler: COVID-19, Veri Madenciliği, Sınıflandırma Yöntemleri, İki Aşamalı Kümeleme Analizi

ABSTRACT

EVALUATION OF PATIENTS WITH SUSPECTED OF COVID-19 USING CLASSIFICATION AND CLUSTERING ALGORITHMS IN DATA MINING

Beyza DURMAZ
Ondokuz Mayıs University
Institute of Graduate Studies
Department of Intelligent Systems Engineering
Master, May/2024
Supervisor: Assoc. Prof. Dr. Aslı ÇALIŞ BOYACI

Data mining enables the processing of large volumes of complex data and reveals meaningful information and patterns. Thus, it supports the decision processes of decision makers. Today, data mining techniques are widely used in the health sector, as in many sectors where decision-making is needed.

The coronavirus pandemic has threatened the whole world physically, mentally and economically since day one. Although its effects were felt mainly in 2020-2022 in Turkey, the coronavirus disease has not disappeared and continues its effects. The coronavirus disease and the pandemic period are still being understood, and current studies and publications on the subject continue. In this context, understanding the disease plays a critical role in terms of diagnosis, reducing the effects on individuals and countries, and the measures to be taken.

This study aims to predict the test results of patients who have been tested by applying to the hospital with suspicion of COVID-19 by using classification algorithms and to make evaluations of patients with positive and negative test results using cluster analysis. For classification, the created data set was modeled using KNIME 5.2.2 with decision tree, K-nearest neighbor, Naive Bayes, logistic regression, random forest and support vector machines algorithms. The highest accuracy rate of 0.796 and the highest diagnostic superiority rate of 15.340 were obtained in the random forest model. The lowest accuracy was obtained in the K-nearest neighbor model with a value of 0.690 and the lowest diagnostic superiority rate with a value of 5.729. For clustering, the data set was separated for patients with negative and positive COVID-19 test results and two different two-stage clustering analyzes were performed. SPSS Clementine 12.0 program was used for the application. In both cases, two clusters were obtained.

Keywords: COVID-19, Data Mining, Classification Methods, Two-Step Cluster Analysis

TEŞEKKÜR

Tezim boyunca bana rehberlik eden ve değerli katkılarıyla çalışmama yön veren danışman hocam Doç. Dr. Aslı ÇALIŞ BOYACI'ya en içten teşekkürlerimi sunarım. Onun bilgi birikimi ve tecrübesi, araştırmamı başarıyla tamamlamamda büyük bir rol oynamıştır. Ayrıca, her zaman yanımda olan ve süreç boyunca bana destek veren Doç. Dr. Hakan TAŞKIN hocama minnettarlığımı iletmek isterim. Onun yardımları ve yol göstericiliği benim için çok kıymetliydi. Tezimin nihai haline ulaşmasında önemli katkılarda bulunan jüri üyelerim Doç. Dr. Naci MURAT ve Dr. Öğr. Üyesi Aycan PEKPAZAR'a teşekkürlerimi sunarım. Onların değerli geri bildirimleri ve eleştirileri, tezimin kalitesini artırmada büyük bir etkiye sahip olmuştur. Bu süreçte bana her zaman destek olan ve beni cesaretlendiren aileme sonsuz teşekkür ederim.

Beyza DURMAZ

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

AIC	: Akaike Information Criterion
ALT	: Alanin Aminotransferaz
APTT	: Alanin Aminotransferaz
AST	: Aspartat Aminotransferaz
BIC	: Bayesian Information Criterion
CRP	: C-reaktif Protein
FP-Growth	: Frequent Pattern Growth Algorithm
GGT	: Gama Glutamil Transferaz
HGB	: Hemoglobin
LYM	: Lenfosit
MONO	: Monosit
PT	: Protrombin Time
RBC	: Eritrosit
SMOTE-NC	: Synthetic Minority Oversampling Technique - Nominal Continous
WBC	: Lökosit

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

1. GİRİŞ

2019 yılının son günlerinde, Çin'in Wuhan şehrinde bir deniz ürünleri pazarında nedeni belirsiz birçok pnömoni hastası olduğu kayıtlara geçmiştir. 2020 yılının ilk günlerinde hastalık sebebinin yeni tip koronavirüs (2019-nCoV) olduğu Dünya Sağlık Örgütü (DSÖ) tarafından açıklanmıştır. Salgın Aralık 2019'da Çin'de başlayıp Ocak 2022'de ilk kez Çin sınırlarını açmıştır. Mart 2022'de virüsün 50'den fazla ülkede görüldüğü DSÖ tarafından bildirilmiş ve COVID-19 küresel salgın olarak ilan edilmiştir (DSÖ, 2020a; TÜBA, 2020).

COVID-19, dünyadaki bireylere ve toplumlara büyük zarar vermiştir. Bu salgın günlük yaşam şeklini değiştirmiş, ekonomileri ve sosyal hayatı baskı altına almıştır (DSÖ, 2020b). Bununla beraber bireylerin psikolojik sağlamlık düzeyleri düşmüş (Killgore vd., 2020), depresyon, anksiyete ve stres gibi ruh sağlığı sorunlarının arttığı tespit edilmiştir (Wang vd., 2020).

COVID-19, ilk günden bu yana tüm dünyayı fiziksel, mental ve ekonomik olarak tehdit etmiştir. Bu bağlamda hastalığı anlamak; teşhis konulması, bireylere ve ülkelere olan etkilerinin azaltılması ve alınacak önlemler açısından son derece kritik bir rol oynamaktadır.

Bu çalışmada, COVID-19 şüphesi taşıyan hastalara ilişkin verilerin veri madenciliğinde sınıflandırma ve kümeleme yöntemleri ile değerlendirilmesini amaçlanmaktadır. Sınıflandırma algoritmaları ile hastaların laboratuvar test sonuçlarından COVID-19 test sonucunun tahmin edilmesi, kümeleme analizi ile çok sayıda özniteliğe göre vakaların değerlendirilerek ilginç örüntülerin keşfedilmesi hedeflenmektedir. Bu bağlamda öncelikle COVID-19 testi yaptırmış olan hastalara ait cinsiyet, yaş, WBC, RBC, HGB, LYM%, MONO%, ALT, AST, CRP, Kreatin Kinaz, Kreatinin ve Üre öznitelikleri kullanılarak test sonucu karar ağacı, K-en yakın komşu, Naive Bayes, lojistik regresyon, rastgele orman ve destek vektör makineleri algoritmalarıyla tahmin edilmektedir. KNIME 5.2.2 kullanılarak oluşturulan modeller doğruluk oranı, hata oranı, duyarlılık, belirleyicilik, yanlış pozitif oranı, yanlış negatif oranı, kesinlik, negatif öngörü değeri, F, pozitif olabilirlik oranı, negatif olabilirlik oranı ve tanısal üstünlük oranı bakımından değerlendirilmektedir. Çalışmanın ikinci aşamasında ise veri seti COVID-19 test sonucu negatif ve pozitif olan hastalar için ayrılmış ve iki aşamalı kümeleme analizi gerçekleştirilmiştir. Uygulama için SPSS

Clementine 12.0 programı kullanılmıştır. Çalışma sonuçlarının sağlık kuruluşları ve bilim için yol gösterici olabileceği düşünülmektedir.

Çalışmanın geri kalanı şu şekilde organize edilmektedir: İkinci bölümünde veri madenciliğinden bahsedilmektedir. Üçüncü bölümde çalışmada kullanılan veri madenciliği yöntemleri açıklanmaktadır. Dördüncü bölümde literatür araştırmasına yer verilmektedir. Beşinci bölümde COVID-19 şüphesi taşıyan hastalar için gerçekleştirilen veri madenciliği uygulamasına ve altıncı bölümde elde edilen bulgulara yer verilmektedir. Yedinci bölümde ise sonuçlar değerlendirilmekte ve ileriki çalışmalar için öneriler sunulmaktadır.

2. VERİ MADENCİLİĞİ

Veri madenciliği, veri yığınlarının içerisinden anlamlı olan bilgilerin ortaya çıkarılması sürecidir.

Bilgi sistemleri, artık geleneksel yöntemlerle anlaşılmayacak kadar büyük ve belirsiz bilgi kütleleri içermektedir. Günümüzde elde edilen verileri analiz edebilmek, işleyebilmek çok daha önemli hale gelmiştir. Artık veri, mal veya hizmet üretimi etkenlerinden biri olarak tanımlanmaktadır.

Veri madenciliğinin en değerli özelliği, çok büyük miktarlardaki verinin işlenebilmesini sağlamasıdır. Böylece karar vericilerin daha iyi karar vermelerine yardımcı olur, rekabetçiliği arttırır.

Veri madenciliği günümüz rekabet dünyasında, mühendislik, üretim, sağlık sektörü, güvenlik, bankacılık, lojistik, istihbarattan spor alanlarına kadar birçok farklı sektörde aktif kullanılmaktadır (Ersöz, 2019).

2.1. Veri Madenciliği ve İstatistik İlişkisi

Veri madenciliği kavram olarak ilk defa istatistikçiler tarafından verinin araştırılması gereken durumlar için kullanılmıştır (Ersöz, 2019).

Veri madenciliği ve istatistiğin ortak özellikleri;

• Verinin bilgiye dönüştürülmesi hedeflenir.

• Verilerin anlamlandırılması ve rakamları okuyabilmek hedeflenir.

• Belirsizliklerin açığa kavuşturulması ve geleceğe dair bilgi vermeyi amaçlar.

• Bir olayı etkileyen önemli faktörlerin saptanmasını ve daha iyi tahminleme hedeflenir.

Bahsedilen ortak özelliklerle beraber istatistik ve veri madenciliği arasında güven düzeyi, hipotez testi, genellenebilirlik ve teorinin rolü olma üzere dört temek farklılık olduğu Zhao ve Luan tarafından ortaya koyulmuştur.

İstatistik bilimi çoğunlukla bir örneklemden anakütle hakkında çıkarım yaparken, veri madenciliği istatistiğin tersine tümevarımla değil tümdengelim ile ilgilenir.

İstatistiksel analizler bilgiye dayalı teorilerle başlar ve teorinin reddedilmesi ve onaylanmasıyla kanıtlanması üzerine doğrulayıcı bir süreçtir. Bununla berbaber veri madenciliği, teorilerin doğrulanması üzerinden ilerlemez.

Veri madenciliği çoğunlukla anakütle ile çalışıldığından istatistikte kullanılan anlamlılık süzeyinde çıkarımlara ihtiyaç duymaz, hipotez testi kullanılmaz.

İstatistik alanında veriler belirli sorular için toplanır. Veri madenciliğinde ise toplanan veri üzerinden anlamlı ilişkiler keşfedilir (Ersöz, 2019).

2.2. Veri Madenciliği Aşamaları

Veri madenciliği; problemin tanımlanması, verilerin hazırlanması, modelin kurulması, modelin değerlendirilmesi, model çıktılarının karar verici tarafından kullanılması ve modelin izlenmesi olmak üzere ana olarak altı aşamaya ayrılabilir (Ersöz, 2019).

2.2.1. Problemin Tanımlanması

Bu aşamada yapılan çalışmanın amacı tanımlanır. İhtiyaçlar açık bir şekilde tanımlanır, performans ölçütleri belirlenir.

2.2.2. Verilerin Hazırlanması

Veri hazırlığı aşamasında veri kaynaklarının belirlenmesi, veriyi tanımlama, veri kalitesinin değerlendirilmesi, veri gruplarının değerlendirilmesi gibi çalışmaları içermektedir. Veri hazırlama süreci veri toplama, temizleme, bütünleştirme, dönüştürme ve indirgenmesi olmak üzere beş aşamaya ayrılmaktadır.

2.2.2.1. Verilerin Toplanması

Bu aşamada belirlenen amaçlar doğrultusunda güvenilir kaynaklardan veriler çekilir.

2.2.2.2. Verilerin Temizlenmesi

Elde edilen veri setinde uç değerler, hatalı veya uygunsuz veriler çıkartılır, eksik veriler temizlenir veya uygun şekilde doldurulur. Bu aşama modelin güvenilirliği ve kalitesi açısından kritik rol oynamaktadır.

2.2.2.3. Verilerin Bütünleştirilmesi

Bu aşamada farklı veri kaynaklarından elde edilen veri kümeleri bütünleştirilir. Bütünleştirilmiş veri tekrarlı kayıtlardan ve ilgisiz niteliklerden arındırılır.

2.2.2.4. Verilerin Dönüştürülmesi

Veriler, özellikleri korunarak veri madenciliğine uygun forma dönüştürülür. Bu süreçte düzeltme, birleştirme, genelleştirme ve normalleştirme işlemleri yer alır.

2.2.2.5. Verilerin İndirgenmesi

Bu aşamada model, daha sağlıklı sonuçlar verebilmesi için yalınlaştırılır. Değişken sayısının çok fazla olduğu veya değişkenlerin önem derecesine göre sıralandırılması gereken durumlarda uygun yöntemlere başvurularak boyut indirgeme işlemi gerçekleştirilir. Faktör analizi, temel bileşenler analizi bu yöntemlere örnektir.

2.2.3. Modelin Kurulması

Problem için en uygun modelin bulunabilmesi için fazla sayıda modelin kurulup denenmesiyle gerçekleştirilir. Bu sebeple veri hazırlama ve model kurma aşamaları yinelenen süreçlerdir.

2.2.4. Modelin Değerlendirilmesi

Bu adımda kurulan modeller değerlendirilir ve en iyi performans gösteren model seçilir. Modellerin performanslarının değerlendirilmesi için kullanılan teknikler, modelde kullanılan veri madenciliği yöntemlerine göre değişkenlik gösterir.

2.2.5. Modelin Çıktılarının Karar Verici Tarafından Kullanılması

Modelin çıktılarının uygulanabilirliği ve kullanılabilirliği daha çok karar vericilerin, modeli kullanacak kişilerin karra vermesi gereken bir konudur. Model seçeneği sayısı birden fazla işe karar vericiler gereksinimlerine göre tercihlerini yaparlar. Tercih edilen model doğrudan bir uygulama olarak kullanılabileceği gibi, başka bir uygulamanın bir alt parçası olarak da kullanılabilir.

2.2.6. Modelin İzlenmesi

Model bu aşamada kullanım esnasındayken sürekli izlenmeye devam edilir. Eğer gerekliyse modelde düzenlemeler yapılır (Ersöz, 2019).

2.3. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri üç ana sınıfta incelenebilir. Bunlar sınıflandırma yöntemleri, kümeleme yöntemleri ve birliktelik kurallarıdır. Sınıflandırma tahmin edici, kümeleme ve birliktelik kuralları ise tanımlayıcı yöntemlerdir.

Tahmin edici yöntemlerin amacı, sonuçları bilinen veri kümelerinden model geliştirilerek sonuçları bilinmeyen verilerin sonuç değerlerinin tahminlenmesidir.

Tanımlayıcı yöntemlerin amacı ise mevcut verilerdeki örüntülerin keşfedilerek karar vermeye yardımcı olmasını sağlamaktır (Dağaslanı, 2022).

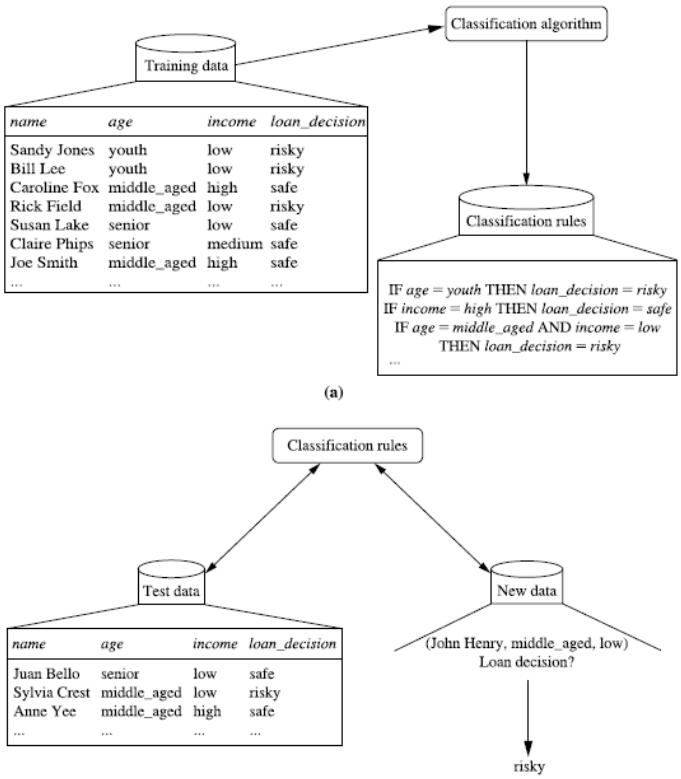
Birliktelik kuralları, olayların birlikte gerçekleşme durumlarını inceler. Olayların birlikte gerçekleşme kuralları tespit edilerek olasılıklarla ifade edilir. Birliktelik kurallarının kullanım alanlarına müşteri satın alma alışkanlıkları, satış analizleri, katalog tasarımları vb. problemler örnek verilebilir (Dağaslanı, 2022). Bu çalışmada kullanılan sınıflandırma ve kümeleme yöntemlerinden üçüncü bölümde detaylıca bahsedilecektir.

3. YÖNTEM

3.1. Sınıflandırma Yöntemleri

Sınıflandırma yöntemleri öğrenme ve sınıflandırma olmak üzere iki adımlı bir süreçlerdir. Öğrenme aşamasında eğitim verileri ile sınıflandırma algoritmaları kullanılarak sınıf kuralları çıkartılır. İkinci aşamada ise test verileri, ilk aşamada çıkartılan sınıf kurallarının doğruluğunu tahmin etmek için kullanılır (Han vd., 2023). Şekil 2.1'de bir sınıflandırma algoritmasının çalışma şekli bir örnek üzerinden görselleştirilmiştir.

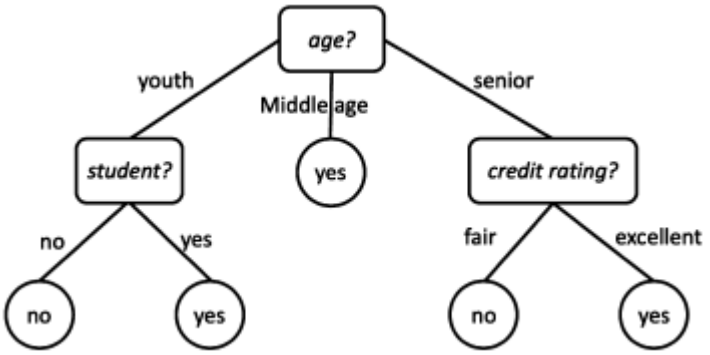
Bu yöntem bir gözlemin veya nesnenin nitelik ve niceğini değerlendirerek önceden belirlenmiş olan bir sınıfa atar. Bu yönteme örnekler; karar ağacı, yapay sinir ağı, genetik algoritma, K-en yakın komşu, rassal orman olarak özetlenebilir (Yalçıner Çal, 2023).



Şekil 3. 1 Sınıflandırma Algoritması Örneği Grafiksel Gösterim (Han vd., 2023)

3.1.1. Karar Ağaçları

Bir karar ağacı, her iç düğümün bir özniteliğe ilişkin bir testi gösterdiği, her dalın bir test sonucunu temsil ettiği ve her yaprak düğümünün bir sınıf etiketini taşıdığı akış diyagramı benzeri bir ağaç yapısıdır. Bir ağaçtaki en üst düğüm kök düğüm olarak adlandırılır. Tipik bir karar ağacı şekil 2.2'de gösterilmiştir. Bu yöntem bir sınıflandırma algoritması olarak sınıf etiketsiz bir veri kümesinin özniteliklerine göre karar ağacı kökten yaprak düğüme kadar yol izlenerek test edilir. Son yaprak düğümü ilgili veri kümesi için elde dilen tahmindir. (Han vd., 2023)



Şekil 3.2. Karar Ağacı Örneği (Han vd., 2023)

Karar ağaçları keşifsel bilgi için uygun olması, çok boyutlu verilerle başa çıkabilmesi, bilginin ağaç formunda temsil edilmesinin anlaşılırlığını kolaylaştırması bu yöntemin çok popüler olmasını sağlamıştır.

3.1.1.1. Karar Ağacı Algoritmaları

Karar ağaçlarının, düğümlerde gelişen yeni düğümlerin sayısı, ağacın büyümesini durdurma kriteri, en iyi bölen özniteliğin seçilmesi ve budama süreci gibi özellikler açısından farklılık gösteren farklı algoritmaları mevcuttur (Toprak, 2017).

Bu algoritmalar aşağıdaki gibi özetlenebilir;

• ID3 (Iterative Dichotomiser 3), C4.5 ve C5.0.

• SLIQ (Supervised Learning in Quest)

• SPRINT (Scalable Parallelizable Induction of Decision Trees)

• CART (Classification and Regression Trees)

• CHAID (Chi-Squared Automatic Interaction Detector)

ID3 algoritması, Ross Quinlan tarafından 1986 yılında açıklanmıştır. Bu algoritmada bölen özniteliğin seçimi için entropi veya gini indeksi kullanılır. Her düğüm, kökten itibaren hiç dikkate alınmamış özelliklerden en kazançlısı seçilerek belirlenir. Bu algoritma C4.5 algoritmasının öncülüdür.

C4.5 algoritması, Ross Quinlan tarafından 1993 yılında açıklanmıştır. Bu algoritma ID3 algoritmasından farklı olarak hem kategorik hem de sürekli öznitelik değerlerini işleyebilmektedir. Kayıp verileri hesaba katmaz, böylelikle daha duyarlı ve daha anlamlı kurallar çıkartabilir.

C5.0 algoritması, ID3 ve C4.5 algoritmalarına göre çok daha hızlıdır. Bu algoritmanın diğerlerine göre ayırt edici olan özelliği faydasız öznitelikleri eleyen winnowing özelliğidir.

SPRINT algoritması, değişkenlere ait sınıf ve sıra numaralarını belirlemek için ayrı bir liste kullanır, bu liste sınıf ve kayıt numaralarını saklar. Ağaç oluşturma sürecinde, eğitim kümelerinden elde edilen başlangıç listeleri sınıflandırma köküyle ilişkilendirilir. Ağaç büyüdükçe ve düğümler yeni dallara bölündükçe, düğümlere ait değişken listeleri de bölünerek yeni dallarla ilişkilendirilir.

CART algoritması, ikili ağaçlar üreterek hangi düğümün kök veya düğüm olacağına karar verir. Bununla beraber, belirlenen düğümün hangi noktadan ikiye ayrılacağını da hesaplar. Algoritma bu süreçte, en uygun değişkeni bulurken hem dallara ayırma işlemini gerçekleştirir hem de değişkenin iki veya daha fazla farklı türde değer taşıyorsa bu değerleri nasıl iki gruba ayıracağını belirler. CART algoritması, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi araştırmakla kalmaz, aynı zamanda bağımsız değişkenlerin birbirleriyle olan etkileşimlerini de saptar. Bu sayede, bağımsız değişkenlerin bağımlı değişkenle olan ilişkisini daha iyi değerlendirir ve model içindeki etkileşim durumlarını ortaya koyar. Bu algoritma C&RT adıyla da bilinir.

CHAID algoritması, parametrik olmaması, kullanılan verilerin normal dağılması zorunluluğunun olmaması, eksik değerleri ayrı bir kategoride analiz edebilmesi gibi esnekliklere sahiptir. ID3, C4.5, C5.0 ve CART algoritmalarından çoklu ağaçlar türetebilmesi özelliğiyle ayrılır. Hem sürekli hem kategorik değişken tipleriyle çalışabilir. Benzer kategorilerin adım adım birleştirilmesinde ki-kare istatistiği kullanılır ve bu süreç, değişkenler arasında istatistiksel olarak daha fazla birleştirme

yapılamayacağına karar verilene kadar devam eder. Değişkenlerin bölünmeye uygun olup olmadığı, Bonferroni düzeltilmiş p değeri kullanılarak belirlenir (Toprak, 2017).

3.1.1.2. Ayrıştırma Kriterleri

Ayrıştırma kriteri, sınıf etiketli eğitim veri setini en iyi şekilde ayrıştırmak için kullanılan yöntemlerdir. Veri ayrıştırma kriterine göre ayrı gruplara ayırıldığında, her bir grubun kendi içinde benzer ve diğer gruplardan benzersiz olmasını sağlanır. Dolaysıyla en iyi bölme kriteri, gruplar arasındaki benzersizliği maksimize eden kriterdir.

Ayrıştırma kriteri, eğitim veri setindeki öznitelikler için bir sıralama sağlar. Skoru en yüksek olan öznitelik, bölme özniteliği olarak seçilir. Eğer bölme özniteliği sürekli bir değere sahipse veya ikili ağaçlara sınırlıysak, sırasıyla, bölme kriterinin bir parçası olarak ya bir bölme noktası ya da bir bölme alt kümesi de belirlenir (Han vd., 2023).

Ayrıştırma kriterleri aşağıdaki gibi özetlenebilir;

• Bilgi Kazancı (Information Gain)

• Kazanç Oranı (Gain Ratio)

• Gini İndeksi (Gini Impurity)

3.1.1.2.1. Bilgi Kazancı Kriteri

Bilgi kazancı kriteri, Claude Shannon'ın bilgi teorisi üzerine yaptığı çalışmalara dayanır. En yüksek bilgi kazancına sahip olan öznitelik, düğüm N için bölme özniteliği olarak seçilir. Bu öznitelik, sonuçta ortaya çıkan bölümlerdeki veri gruplarını sınıflandırmak için gereken bilgiyi en aza indirir ve bu bölümlerdeki en az rastlantısallığı veya "kirliliği" yansıtır. Bu yaklaşım, belirli bir veri grubunu sınıflandırmak için gerekli beklenen test sayısını en aza indirir.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \mathbf{3.1}$$

Burada D sınıf etiketli eğitim veri seti, p_i D'deki rastgele bir veri grubunun C_i sınıfına ait olma olasılığının, $|C_i,D|/|D|$ ile tahmin edilen, sıfır olmayan bir olasılıktır. Bilgi bitleriyle kodlandığı için 2 tabanında bir logaritma fonksiyonu kullanılır. Info(D), D'deki bir veri grubunun sınıf etiketini tanımlamak için gereken ortalama bilgi miktarıdır (Denklem 2.1). Info(D) aynı zamanda D'nin entropisi olarak da bilinir.

Eğitim verilerinden gözlemlenen v farklı değere sahip A adlı bir özniteliği baz alarak D içindeki veri gruplarının bölündüğü kabul edilsin. A niteliği D' yi $\{D_1,D_2,\ldots,D_v\}$ şeklinde v adet bölüme ayrılır. (D_j' de A niteliği a_j değerini almış gözlemler bulunmaktadır), bununla beraber amaç en iyi sınıflandırmanın oluşturulmasıdır. Bir sınıflandırmaya kesin olarak ulaşmak için (bölümleme sonrasında) ne kadar daha fazla bilgiye ihtiyaç olacağı Denklem 3.2 ile hesaplanır.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \, x \, Info(D_j) \qquad \mathbf{3.2}$$

|Dj| / |D| terimi, j. bölümün ağırlığını temsil eder. $Info_A(D)$, A tarafından bölümlendirilmiş D'nin bir veri grubunu sınıflandırmak için gereken beklenen bilgidir. Beklenen bilgi ne kadar küçükse, bölümlerin saflığı o kadar yüksektir. $Info_A(D)$ aynı zamanda A'ya bağlı olarak D'nin koşullu entropisi olarak da bilinir.

$$Gain\ (A) = Info(D) - \ Info_A(D) \qquad \mathbf{3.3}$$

Bilgi kazancı, orijinal bilgi gereksinimi ile (sadece sınıf oranlarına dayalı olarak) ve yeni gereksinim (öznitelik A'ya göre bölümlendirme yapıldıktan sonra elde edilen) arasındaki fark olarak tanımlanır (Denklem 2.3). Diğer bir deyişle, Gain(A), A üzerinde dallanmanın ne kadar kazanç getireceğini belirtir. En yüksek bilgi kazancına sahip olan öznitelik A, Gain (A), düğüm N'de bölünme özniteliği olarak seçilir (Han vd., 2023).

3.1.1.2.2. Kazanç Oranı

Kazanç oranı, çok sayıda sonuca sahip testlere yöneliktir.

$$SplitInfo_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} x \log_2\left(\frac{|D_j|}{|D|}\right) \qquad \mathbf{3.4}$$

Bu değer, eğitim veri kümesini, A özniteliği üzerinde yapılan bir testin v çıktısına karşılık gelecek şekilde bölmenin potansiyel bilgisini temsil eder. Her çıktı için, D'deki ilgili çıktıya sahip olan veri grubu sayısını, toplam D veri grubu sayısına göre düşünür. Bu, aynı bölünme üzerinden elde edilen sınıflandırmaya göre kazanılan bilgiyi ölçen bilgi kazancından farklıdır.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \qquad \mathbf{3.5}$$

En yüksek kazanç oranına sahip öznitelik, bölünme özniteliği olarak seçilir. Ancak, bölünme bilgisi 0'a yaklaştıkça, oran dengesiz hale gelir. Bunu önlemek için bir kısıtlama eklenir; seçilen testin bilgi kazancı büyük olmalıdır, en az tüm testlerin ortalamasına eşit olmalıdır (Han vd., 2023).

3.1.1.2.3. Gini İndeksi

Gini indeksi, CART'ta kullanılır. Daha önce tanımlanan gösterimleri kullanarak, Gini, D'nin (bir veri bölümü veya eğitim örnekleri kümesi) belirsizliğini ölçer (Denklem 2.6).

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \qquad \mathbf{3.6}$$

Bu formülde m sınıf olmak üzere, p_i'nin D içindeki bir örneğin C_i sınıfına ait olma olasılığı olduğu ve $|C_i,D|/|D|$ ile tahmin edildiği belirtilmektedir. Gini indeksi, her bir değişken için ikili bir bölünmeyi dikkate alır. D sınıf etiketli eğitim verisi içinde bulunan v farklı değere sahip A özniteliği için; A'da en iyi ikili bölünmeyi belirlemek için, A'nın bilinen değerlerini kullanarak oluşturulabilecek tüm olası alt kümeler incelenir. Veri setini iki bölüme ayırmak için A üzerinde ikili bir bölünmeye dayanan $(2^v - 2)/2$ olası yol vardır. İkili bir bölünme göz önüne alındığında, her bir oluşan bölümün karışıklığının ağırlıklı toplamını hesaplanır Denklem 2.7, A üzerindeki ikili bir bölünme, D'yi D_1 ve D_2 olarak bölerse, bölünme durumunda D'nin Gini indeksi şöyle hesaplanır:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \qquad \mathbf{3.7}$$

Her bir öznitelik için, mümkün olan her ikili bölünme göz önüne alınır. Bir kategorik öznitelik için, o özniteliğin en az Gini indeksine sahip olan alt kümesi, bölünme alt kümesi olarak seçilir. Sürekli değerli öznitelikler için her olası bölünme noktası dikkate alınır. Yöntem, bilgi kazancı için yöntemine benzerdir, burada her bir (sıralanmış) bitişik değer çiftinin orta noktası, bir olası bölünme noktası olarak alınır. Belirli bir (sürekli değerli) özniteliğin minimum Gini karışıklığına sahip noktası, o özniteliğin bölünme noktası olarak alınır.

Bir kesikli veya sürekli değerli bir öznitelik A için ikili bir bölünmeyle sağlanacak indeks azalması denklem 2.8'de belirtilmiştir.

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \qquad \mathbf{3.8}$$

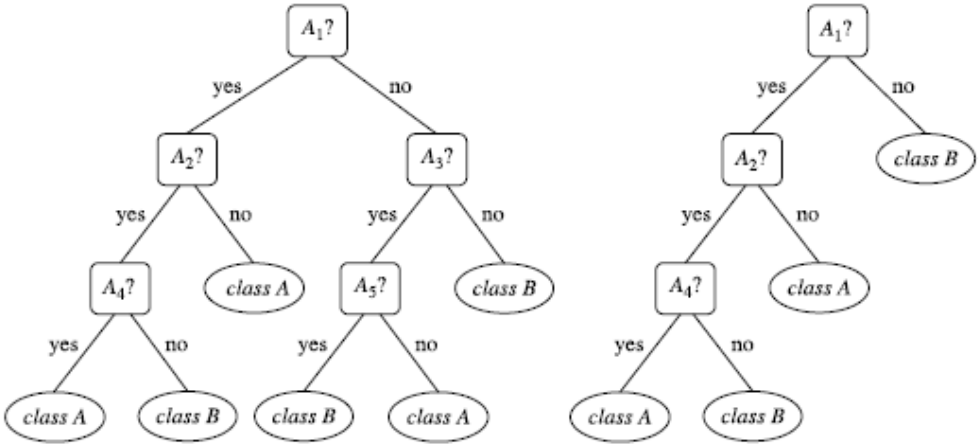
İndeksin azalmasını maksimize eden öznitelik bölünme özniteliği olarak seçilir. Bu öznitelik veya bir bölünme özniteliği için bölünme alt kümesi (kesikli değerli bir bölünme özniteliği için) ya da bölünme noktası (sürekli değerli bir bölünme özniteliği için) birlikte bölünme kriterini oluşturur (Han vd., 2023).

3.1.1.3. Ağaç Budama

Karar ağacı oluşturulduğunda, bazı dallar gürültü veya aykırı veriler nedeniyle eğitim verilerindeki anormallikleri yansıtabilir. Ağaç budama yöntemleri, bu tür veriye aşırı uydurma sorununu ele alır. Bu yöntemler genellikle en güvenilmez dalları kaldırmak için istatistiksel ölçümler kullanır. Budanmış ağaçlar genellikle daha küçük ve daha az karmaşık olup, bu nedenle daha anlaşılması daha kolaydır. Ayrıca, genellikle bağımsız test verilerini (yani önceden görülmemiş örnekleri) daha hızlı ve daha doğru bir şekilde sınıflandırır. Ağaç budama, ağacın aşırı uyumunu önlemek ve genelleme yeteneğini artırmak için belirli dalları veya düğümleri kaldırarak çalışır. Ağaç budamanın iki yaygın yaklaşımı vardır: ön budama(prepruning) ve sonradan budama (postpruning) (Han vd., 2023).

Ön budama yaklaşımında, bir ağaç yapısı oluşturulması erken durdurularak budanır. Örneğin, belirli bir düğümde eğitim örneklerinin alt kümesini daha fazla bölmeme veya bölememe kararı alınır. Durduktan sonra, düğüm bir yaprak haline gelir. Yaprak, alt küme örnekleri arasında en sık görülen sınıf etiketini veya bu örneklerin sınıf etiketlerinin olasılık dağılımını içerebilir. Ağaç yapısı oluşturulurken, istatistiksel önem, bilgi kazancı, Gini indeksi gibi ölçüler, bir bölünmenin kalitesini değerlendirmek için kullanılır. Bir düğümdeki örnekleri bölme önceden belirlenmiş bir eşiğin altına düşerse, verilen alt kümenin daha fazla bölünmesi durdurulur. Bununla beraber uygun bir eşik seçmek zor kritiktir. Yüksek eşikler basitleştirilmiş ağaçlara yol açabilirken, düşük eşikler çok az basitleştirmeye neden olabilir (Han vd., 2023).

İkinci ve daha yaygın olan yaklaşım sonradan budamadır. Bu yaklaşım, tamamen büyümüş bir ağaçtan alt ağaçları kaldırır. Bir düğümdeki bir alt ağaç, dalları kaldırılarak ve yerine bir yaprak eklenerek budanır. Yaprak, değiştirilen alt ağacın içindeki en sık sınıf etiketi ile etiketlenir (Han vd., 2023).



Şekil 3.3. Budanmamış(sol) ve Budanmış(sağ) Karar Ağacı Örneği (Han vd., 2023)

3.1.2. K-En Yakın Komşu

K-en yakın komşu yöntemi ilk kez 1950'lerin başında tanımlanmıştır. Yöntem, büyük bir eğitim seti verildiğinde işgücü yoğun olup, 1960'larda artan hesaplama gücü kullanılabilir hale gelinceye kadar popülerlik kazanmamıştır. O zamandan beri desen tanıma alanında geniş çapta kullanılmıştır (Han vd., 2023).

Veri uzayında birbirine yakın olan aynı tür gözlemler, birbirlerinin komşusu durumundadır. Bu anlayışa dayanarak, çok basit ancak güçlü bir k - en yakın komşu algoritması geliştirilmiştir. Özetle k-en yakın komşu algoritmasının temel mantığı, komşunun ne yaptığına bakarak karar vermektir. Belirli bir gözlemin davranışı tahmin etmek istenirse, veri uzayında o gözleme yakın olan k adet komşunun davranışı incelenir. Bu k komşunun davranışlarının ortalaması hesaplanır ve bu hesaplanan ortalama, gözlemin tahmini olur (Dolgun, 2006).

K-en yakın komşu algoritmasının çalışma mantığı dört ana madde ile özetlenebilir (Harrington, 2012);

- Sınıfı tahminlemek istenen gözlemin, veri setindeki tüm gözlemlere olan uzaklığı hesaplanır.
- Hesaplanan uzaklık değerleri artan şekilde sıralanır.
- En küçük uzaklık değerine sahip olan, belirli k adet komşu gözlem alınır.
- K adet gözlemin içindeki en sık tekrarlanan sınıf, tahminlenerek istenen gözlemin sınıfı olarak tahminlenir.

Bu yöntemde, belirli bir test gözlemi benzer olan eğitim gözlemiyle karşılaştırılır. Eğitim gözlemleri n öznitelikle tanımlanmıştır. Her gözlem, n boyutlu bir uzaydaki bir noktayı temsil eder. Bu sayede, tüm eğitim gözlemleri n boyutlu bir öznitelik uzayında saklanır. Bilinmeyen bir gözlem verildiğinde, k-en yakın komşu algoritması, bilinmeyen gözleme en yakın k eğitim gözlemini arar. Bu k eğitim gözlemi, bilinmeyen gözlemin k "en yakın komşusu" olarak kabul edilir. K-en yakın komşu algoritması, bilinmeyen gözlemin tahmin edilen sınıf etiketini belirlemek için k-en yakın komşular arasında en yaygın olan sınıf etiketini seçer (Han vd., 2023).

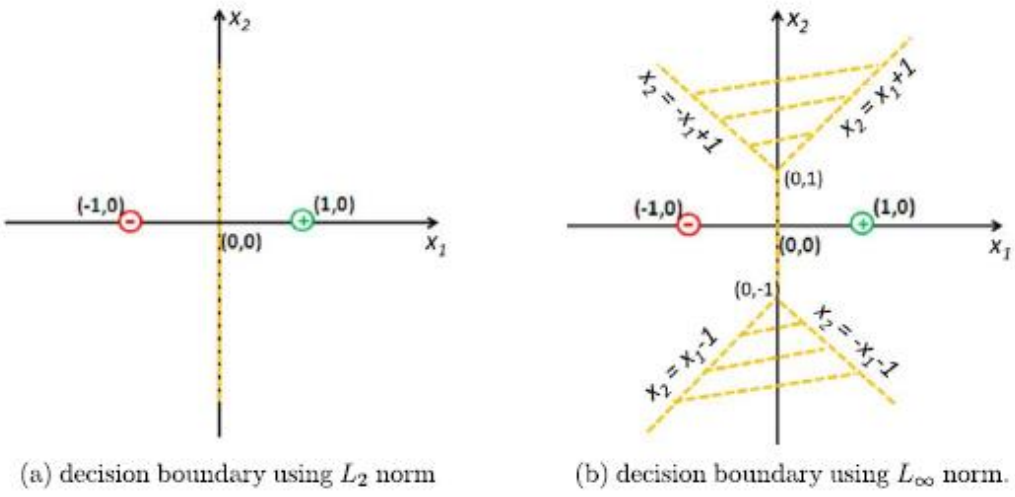
"Yakınlık" örneğin, Öklid mesafesi gibi bir mesafe ölçütü açısından tanımlanır. İki nokta veya gözlem arasındaki Öklid mesafesi, diyelim ki, $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ ve $X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, denklem X.X'deki gibi hesaplanır:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \qquad \mathbf{3.9}$$

Her sayısal öznitelik için, gözlem X_1'deki bu özniteliğin karşılık gelen değeri ile gözlem X_2'deki bu özniteliğin karşılık gelen değeri için formül uygulanır. Tipik olarak, uzaklık formülü kullanmadan önce her özniteliğin değerleri normalize edilir. Bu, başlangıçta büyük aralıklara sahip özniteliklerin (örneğin, gelir) başlangıçta daha küçük aralıklara sahip öznitelikleri (örneğin, ikili öznitelikler) domine etmesini önlemeye yardımcı olur.

K-en yakın komşu algoritmasında bilinmeyen gözlem, k-en yakın komşuları arasında en yaygın sınıf etiketi ile atanır. $k = 1$ olduğunda, bilinmeyen gözleme, öznitelik uzayında en yakın olan eğitim demetinin sınıfı atanır. $k > 1$ olduğunda, k-en yakın komşuları arasında sınıf etiketleri üzerinde (ağırlıklı) çoğunluk oylaması yapılır.

Nominal nitelikler için, gözlem X_1'deki niteliğin karşılık gelen değerini gözlem X_2'dekiyle karşılaştırmaktır. İki değer birbirine eşitse (örneğin, gözlemler X_1 ve X_2 her ikisi de mavi renkteyse), o zaman ikisinin arasındaki fark 0 olarak alınır. İki değer farklıysa (örneğin, gözlem X_1 mavi iken gözlem X_2 kırmızıdır), o zaman fark 1 olarak kabul edilir. Başka yöntemler daha karmaşık derecelendirme şemalarını (örneğin, mavi ve beyaz için mavi ve siyah gibi bir fark puanından daha büyük bir fark puanı atanır) içerebilir.



Şekil 3.4. K-En Yakın Komşu Algoritmasının Farklı Mesafe Ölçütleri (Han vd., 2023)

Şekil X.X 1-en yakın komşu sınıflandırıcısının farklı mesafe ölçütleri üzerindeki etkisini göstermektedir. İki eğitim örneği verildiğinde, pozitif bir örnek (1, 0) ve negatif bir örnek (−1, 0) içerir. Farklı mesafe ölçütlerini kullanan 1-en yakın komşu sınıflandırıcısının karar sınırları oldukça farklıdır. L_2 normunu kullanırken (solda), karar sınırı $x_2 = 0$ olan dikey bir çizgidir. L_∞ normunu kullanırken (sağda), karar sınırı (0,−1) ve (0, 1) arasında bir çizgi parçasını ve taralı alanları içerir.

K komşu sayısı deneysel olarak tespit edilebilir. K = 1'den başlayarak, bir test seti kullanarak sınıflandırıcının hata oranını tahmin edilir. Bu işlem, her seferinde bir komşuya izin vermek için k değerini artırarak tekrarlanır. En düşük hata oranını sağlayan k değeri seçilir. Genel olarak, eğitim verilerinin sayısı arttıkça, k değeri de artar (Han vd., 2023)

3.1.3. Naive Bayes

Naive Bayes, veri sınıflandırmasında istatistiksel bir yöntem olarak kullanılır. Bu model, sınıflandırılacak olayları birbirinden bağımsız olarak ele alır ve her bir özniteliğin sonuca etkisini olasılık olarak hesaplar. Bu özellikleri sayesinde, araştırmacılar tarafından sıkça tercih edilen bir sınıflandırma yöntemi haline gelmiştir. Bunun sebeplerinden biri de uygulamasının kolaylığı ve hızlı hesaplama performansıdır (Kurnaz, 2019).

Naive bayes algoritması şu şekilde işler;

D'nin, içerdiği gözlemlerin ve bunlara ilişkin sınıf etiketlerinin eğitim seti olduğu varsayılsın. Her gözlem, sırasıyla A_1, A_2, ..., A_n olarak adlandırılan n

özniteliğin ölçümlerini içeren bir n boyutlu öznitelik vektörüyle temsil edilir, X = (x_1, x_2, ..., x_n) (Han vd., 2023).

C_1, C_2, ..., C_m olmak üzere m adet sınıf olduğu varsayılsın. Bir gözlem olan X verildiğinde, sınıflandırıcı X'e koşullanmış en yüksek sonsal olasılığa sahip sınıfa ait olduğunu tahmin eder. Yani, Naive Bayes algoritması, gözlem X'in yalnızca şu koşulu sağladığında X'in C_i sınıfına ait olduğunu tahmin eder:

$$P(C_i|X) > P(C_j|X), \quad 1 \leq j \leq m \; için, \qquad j \neq i \qquad \mathbf{3.10}$$

Bu durumda $P(C_i|X)$ maksimize edilir. $P(C_i|X)$ maksimize edilen sınıf C_i, maksimum olasılıklı hipotez olarak adlandırılır.

$$P(C_i|X) = \frac{P(C_i|X)P(C_i)}{P(X)} \qquad \mathbf{3.11}$$

P(X) tüm sınıflar için sabittir, bu yüzden yalnızca $P(X|C_i)P(C_i)$ 'nin maksimize edildiği sınıfın bulunması yeterlidir. Sınıf öncelik olasılıkları bilinmiyorsa, genellikle sınıfların eşit olasılıklı olduğu varsayılır, yani $P(C_1) = P(C_2) = \cdots = P(C_m)$, bu durumda $P(X|C_i)$ 'yi maksimize edilir. Aksi halde, $P(X|C_i)P(C_i)$ 'i maksimize edilir. Sınıf öncelik olasılıkları, $|C_i,D|$, D içindeki C_i sınıfına ait eğitim demetlerinin sayısı olduğunda tahmin edilebilir.

Önceden belirtilen nedenlerden dolayı $P(X|C_i)$ değerini hesaplamak son derece maliyetlidir. $P(X|C_i)$ değerini değerlendirme işleminde hesaplamanın azaltılması için sınıf şartlı bağımsızlık varsayımı yapılır. Bu, gözlemin sınıf etiketine bağlı olarak özellik değerlerinin birbirinden koşullu olarak bağımsız olduğunu varsayar diğer bir deyişle gözlemin hangi sınıfa ait olduğunu bildiğimizde özellikler arasında bağımlılık ilişkileri bulunmaz. Böylece;

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \qquad \mathbf{3.12}$$

$$= P(x_1|C_i)xP(x_2|C_i)x \ldots xP(x_n|C_i) \qquad \mathbf{3.13}$$

Eğitim gözlemlerinden olasılıklar $P(x_1|C_i)$,$P(x_2|C_i)$, . . . , P $(x_n|C_i)$ kolayca tahmin edilebilir. Burada x_k, gözlem X için özniteliğin A_k değerine işaret eder. Her özniteliğin kategorik veya sürekli değer olduğuna bakılır. Örneğin, $P(X|C_i)$'yi hesaplamak için;

Eğer A_k kategorik ise, $P(x_k|C_i)$, A_k için x_k değerine sahip D'deki C_i sınıfı gözlemlerinin sayısı, $|C_i,D|$, yani D'deki C_i sınıfı gözlemlerinin sayısına bölünür.

Eğer A_k sürekli bir değere sahipse, bir sürekli değerli öznitelik genellikle bir ortalama μ ve bir standart sapma σ ile Gauss dağılımına sahip olduğu varsayılır. Böylece;

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \mathbf{3.14}$$

$$P(x_k|C_i) = g\left(x_k, \mu_{C_i}, \sigma_{C_i}\right) \qquad \mathbf{3.15}$$

μ_{Ci} ve σ_{Ci}'i sırasıyla sınıf C_i'nin eğitim gözlemlerinin özniteliği A_k için değerlerinin ortalaması ve standart sapmasıdır. Daha sonra, bu ortalama ve standart sapma değeri denklem 2.14'de x_k ile birlikte kullanılarak $P(x_k|C_i)$'nin tahmin edilmesi için kullanılır.

X gözleminin sınıf etiketini tahmin etmek için, her bir sınıf C_i için $P(X|C_i)P(C_i)$ değeri hesaplanır. Algoritma, X gözleminin sınıf etiketini şu koşullar altında tahmin eder (Denklem 2.15):

$$P(X|C_i)P(C_i) > P(X|C_i)P\left(C_j\right), \qquad 1 \leq j \leq m \; için, \qquad j \neq i \qquad \mathbf{3.16}$$

Diğer bir deyişle, tahmin edilen sınıf etiketi, $P(X|C_i)P(C_i)$'nin maksimum olduğu sınıf C_i'dir (Han vd., 2023).

3.1.4. Lojistik Regresyon

İlk olarak, lojistik regresyonun biyolojik deneylerin analizinde kullanılması 1944, 1953 ve 1955 yıllarında Berkson tarafından incelenmiştir. Daha sonra, 1972'de Finley, probit analizine bir alternatif olarak lojistik regresyonu önermiştir. Son 20 yılda, bu yöntem askeri, meteoroloji ve sağlık alanlarında yaygın bir şekilde kullanılmaktadır.

Eğer bağımlı değişken iki kategoriye sahipse, bu durumda ikili (binary) lojistik regresyon kullanılırken, bağımlı değişkenin kategorisi ikiyi aşarsa, bu çok kategorili (multinomial) lojistik regresyon olarak adlandırılır (Doğançay, 2023).

İkili lojistik regresyon modelinde, bağımlı (açıklanan) değişkenin gözlenen değeri iki olası durumu temsil eder. Olay gerçekleşirse 1, gerçekleşmezse 0 değerini alır. Bu algoritmada kullanılan bağımsız değişkenlerin sürekli veya kategorik olması

gibi bir zorunluluk bulunmamaktadır. Bağımsız değişkenlerin tümü veya bir kısmı, sürekli veya kategorik tipte olabilirler. Bununla birlikte bağımsız değişkenlerin ikili veya üçlü etkileşimleri ortak değişken olarak modele dahil edilebilir (Kurnaz, 2019).

Bağımsız değişken X, bağımlı değişken Y olmak üzere, ikili lojistik regresyon modelinin açıklanmasında lojistik dağılım fonksiyonundan yararlanılır (Denklem 3.17).

$$P_i = E(Y = 1 \backslash X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \qquad \mathbf{3.17}$$

P_i bağımsız değişkeni, X_i i. birey için Y'nin 1-0 değerini alma olasılığını ifade eder. Bu fonksiyonda, P_i'nin hem X'e hem de β'lara göre doğrusal olmayan bir şekilde değiştiği görülmektedir. Doğrusal olmayan bir yapıya sahip olan lojistik regresyon modeli uygun dönüşümlerle doğrusallaştırılabilir ve bu model "Logit" olarak adlandırılır. Bu şekilde bahsedilen modeller "Logit Model" olarak ifade edilir.

İki veya daha fazla bağımsız değişkenin olduğu durumlarda, bu model ikili çoklu logit regresyon modeli olarak adlandırılır. İkili çoklu logit regresyon modeli için, çoklu lojistik dağılım fonksiyonundan yararlanılır.

$$P_i = E(Y = 1 \backslash X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}} \qquad \mathbf{3.18}$$

Odds (üstünlük) olasılık oranı, meydana gelen olay sayısının meydana gelmeyen olay sayısına oranıdır. 1'den büyük bir Odds oranı olayın gerçekleşme olasılığının arttığını ifade eder. 1'den küçü bir Odds oranı ise olayın gerçekleşme olasılığının düştüğünü ifade eder. P ilgili olayın gözlenme olasılığını ifade eder. Üstünlük değerlerinin birbirine oranı olan üstünlük oranı, lojistik regresyon denkleminde Exp (B) olarak ifade edilir. Exp (B), bir olayın meydana gelme olasılığının meydana gelmeme olasılığına oranı olan Odds'un, X_k değişkeninin Y değişkeni üzerindeki etkisiyle ne kadar veya ne kadar yüksek bir olasılık artışı olduğunu veya azalışı olduğunu açıklar.

Çoklu lojistik regresyon modelinde, bağımlı (açıklanan) değişkenin gözlenen değeri ikiden fazla olası durumu temsil eder. Örneğin, bağımlı değişkenin 0, 1, 2 gibi üç farklı kategoriye sahip olduğu varsayılsın. Bu durumda, 0 kategorisi göz önüne alındığında, 2 numaralı kategorinin 1 numaralı kategori ile karşılaştırılması için aşağıdaki Denklem 3.19 ve Denklem 3.20 kullanılır.

$$g_1(X) = Log\left(\frac{P(Y=1\backslash X)}{P(Y=0\backslash X}\right) = \beta_{10} + \beta_{11}X_1 + \cdots + \beta_{1p}X_p \qquad \mathbf{3.19}$$

$$g_2(X) = Log\left(\frac{P(Y=2\backslash X)}{P(Y=0\backslash X}\right) = \beta_{20} + \beta_{21}X_1 + \cdots + \beta_{2p}X_p \qquad \mathbf{3.20}$$

Bu denklemlerden hareketle üç kategori için koşullu olasılıklar k=0,1,2, için aşağıda yer alan Denklem 3.21 ile hesaplanır.

$$Pk(X) = \frac{\exp\big(gk(X)\big)}{\sum_{t=0}^{2}\exp\big(gt(X)\big)} \qquad \mathbf{3.21}$$

Bu denklemden hareketle, 0 kategorisi sabitken, 1 kategorisinin gerçekleşme olasılığının, 2 kategorisinin gerçekleşme olasılığına göre yüzde x kadar daha fazla ya da az olduğu yorumu yapılabilir (Kurnaz, 2019).

3.1.5. Rastgele Orman

Rastgele orman algoritması Leo Brieman tarafından 2001 yılında geliştirilmiştir. Bu algoritma birden fazla karar ağacı üreterek sınıflandırıcının değerini yükseltmeyi hedefler. Hem gözlem hem de değişken seçimini rastgele yapması sebebiyle rassal (random) adını almıştır (Doğançay, 2023).

Rastgele orman algoritması, birden fazla karar ağacından oluşan bir topluluk yöntemidir. Her karar ağacı, veri setinden rastgele seçilen bir öznitelik alt kümesine dayanarak oluşturulur. Bu durum, her ağacın daha farklı bir karar kuralı öğrenmesini sağlar. Sınıflandırma sırasında, her karar ağacı yeni bir veri noktası için bir sınıf tahmininde bulunur. Son olarak, en çok oy alan sınıf, nihai tahmin olarak döndürülür. Bu yaklaşım, tek bir karar ağacına kıyasla daha sağlam ve daha az aşırı uyuma sahip modeller oluşturmaya yardımcı olur (Han vd., 2023).

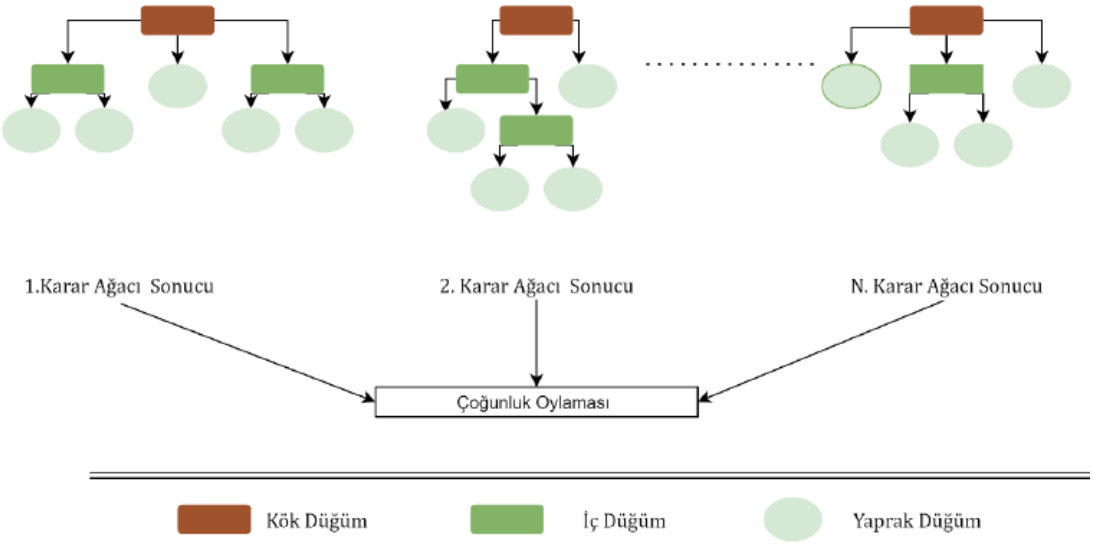
Rastgele ormanlar, rastgele öznitelik seçimi için bootstrap kullanılarak oluşturulabilir. D'nin d adet gözlemden oluşan bir eğitim veri seti olduğu varsayılsın. Topluluk için k adet karar ağacı oluşturacak proses şöyledir (her ağaç için tekrarlanır i=1,2,…,k);

- D eğitim setinden iadeli olarak(seçilen örnek tekrar veri setine konularak bir sonraki seçimde de seçilebilme ihtimaline sahiptir) d adet örnek içeren bir alt eğitim seti oluşturulur. Bu durum, bazı örneklerin Di'de birden fazla yer almasına, bazılarının ise hiç yer almamasına neden olur.

- Bir düğümde bölünme için aday olacak F adet öznitelik, rastgele seçilir. F, her düğümde bölünme kararını belirlemek için kullanılacak öznitelik sayısıdır, toplam öznitelik sayısından çok daha küçük olmalıdır.
- CART yöntemi kullanılarak karar ağacı büyütülür. Ağaçlar maksimum boyutlarına kadar büyütülür ve budama yapılmaz.

Bu şekilde oluşturulan, rastgele girdi seçimi kullanılan rastgele ormanlar, Forest-RI olarak adlandırılır.

Rastgele ormanların bir başka formu olan Forest-RC, geleneksel yaklaşımdan farklı olarak giriş özniteliklerinin rastgele doğrusal kombinasyonlarını kullanır. Özniteliklerin bir alt kümesini rastgele seçmek yerine, var olan özniteliklerin doğrusal kombinasyonlarından oluşan yeni öznitelikler oluşturur. Bir öznitelik, birleştirilecek orijinal öznitelik sayısı olan L'yi belirterek oluşturulur. Belirli bir düğümde, L adet öznitelik rastgele seçilir ve [-1, 1] aralığında eşit dağılmış rastgele sayılar olan katsayılarla birlikte toplanır. F adet doğrusal kombinasyon oluşturulur ve bunlar arasında en iyi bölünmeyi bulmak için bir arama yapılır. Bu tür rastgele orman, az sayıda öznitelik olduğunda, bireysel sınıflandırıcılar arasındaki korelasyonu azaltmak için faydalıdır. Forest-RC, geleneksel yaklaşıma kıyasla daha az sayıda öznitelik olduğunda daha faydalı bir seçenek olabilir (Han vd., 2023).



Şekil 3.5 Rastgele Orman (Gök, 2021)

3.1.6. Destek Vektör Makineleri

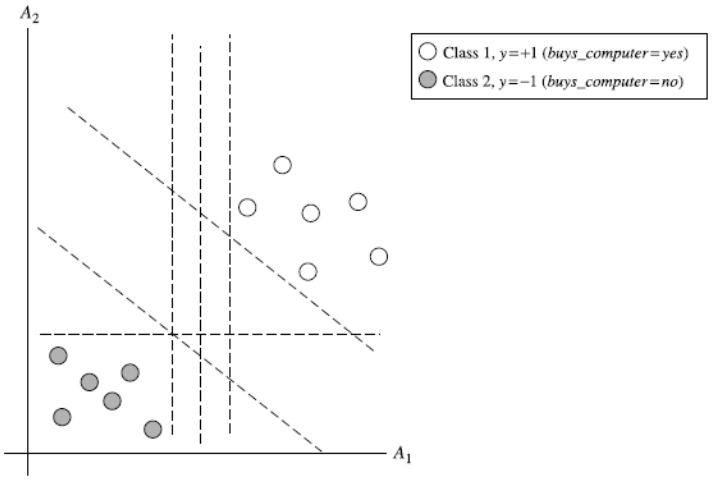
Destek vektör makineleri, Vladimir Vapnik ve Alexey Chervonenski tarafından 1960'lı yıllarda geliştirilmiştir. Aykırı değerlere karşı dayanıklı olması, büyük veri setleri ve öznitelik sayısının fazla olduğu veri setleri için kullanışlı olması bu yöntemin önemli özelliklerindendir (Doğançay, 2023).

Kısaca, bir destek vektör makinesi eğitim verilerini daha yüksek boyutlu bir uzaya dönüştürmek için doğrusal olmayan bir eşleme kullanır. Bu yeni uzayda, bir sınıftaki veri noktalarını diğer sınıftaki veri noktalarından ayıran doğrusal optimal ayırma hiper düzlemini (bir karar sınırı) arar. Yeterince yüksek boyutlu bir uzaya uygun bir doğrusal olmayan eşleme ile iki sınıftaki veriler her zaman bir hiper düzlemle ayrılabilir. Destek vektör makinesi bu hiper düzlemi destek vektörleri (support vectors) (önemli eğitim gözlemleri) ve marjinler (margins) (destek vektörleriyle tanımlanan) kullanarak bulur (Han vd., 2023).

En hızlı destek vektör makinelerinin bile eğitimi yavaş olsa da, karmaşık doğrusal olmayan karar sınırlarını modelleme yetenekleri sayesinde oldukça doğrudurlar. Diğer yöntemlere göre aşırı öğrenmeye çok daha az eğilimlidirler.

Bu algoritma, veriye dair herhangi bir birleşik dağılım fonksiyonu bilgisine ihtiyaç duymaz. Bu sebeple dağılımdan bağımsız çalışan bir öğrenme algoritmasıdır. Destek vektör makineleri, veri setinin doğrusal ayrılıp ayrılmama durumuna göre doğrusal ve doğrusal olmayan destek vektör makineleri olmak üzere ikiye ayrılır (Kurnaz, 2019).

Sınıfların doğrusal olarak ayrılabildiği iki sınıflı bir problem için; D, bir veri seti olmak üzere (X1, y1), (X2, y2), ..., (X|D|, y|D|) şeklinde olduğu varsayılsın. Burada X_i, sınıf etiketleriyle birlikte eğitim gözlemlerini ve y_i, her bir gözleme ait sınıf etiketlerini temsil etmektedir. Her y_i, +1 veya -1 değerlerinden birini alabilir ($y_i \in \{+1, -1\}$). Şekil 2.6'da görüldüğü üzere, 2 boyutlu verilerin doğrusal olarak ayrılabilmektedir, +1 sınıfındaki tüm gözlemleri -1 sınıfındaki tüm gözlemlerden ayıran düz bir çizgi çizilebilmektedir. Şekil 3.6'da kesikli çizgiler halinde gösterilenler de dahil olmak üzere sonsuz sayıda olası ayırıcı hiper düzlem (karar sınırları) vardır (Han vd., 2023).



Şekil 3.6 Doğrusal Olarak Ayrılabilen İki Öznitelikli Bir Eğitim Veri Seti Örneği (Han vd., 2023)

Bir ayrıştırıcı hiperdüzlem aslında bir doğrusal sınıflayıcıdır. Denklem 3.22'de W bir ağırlık vektörüdür. W = {w1, w2, . . . , wn}, n öznitelik sayısıdır ve b bir skalerdir, genellikle bir sapma olarak adlandırılır.

$$W.X + b = 0 \qquad \mathbf{3.22}$$

İki öznitelikli (A_1 ve A_2) bir veri seti için Denklem 2.21 şu şekilde yazılabilir;

$$b + w_1x_1 + w_2x_2 \ = 0 \qquad \mathbf{3.23}$$

Böylece, ayırıcı hiper düzlemin üzerinde bulunan herhangi bir nokta aşağıdaki denklemi sağlar (Denklem 3.24). Bu nokta sınıfı +1 olarak sınıflandırılacakken, 3.25 numaralı denklemi sağlayan herhangi bir nokta ayırıcı hiper düzlemin altında yer alacak ve sınıfı -1 olarak sınıflandırılacaktır.

$$b + w_1x_1 + w_2x_2 > 0 \qquad \mathbf{3.24}$$

$$b + w_1x_1 + w_2x_2 < 0 \qquad \mathbf{3.25}$$

Ağırlıklar, marjın sınırlarını tanımlayan hiper düzlemlerin şu şekilde yazılabilmesi için ayarlanabilir (Denklem 3.26 ve 3.27):

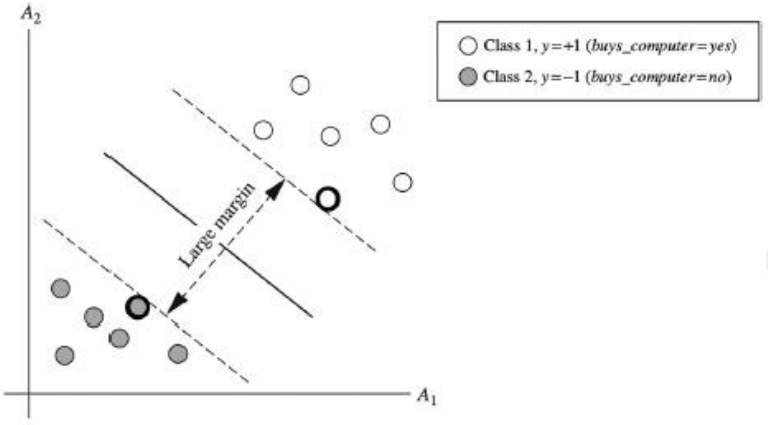
$$H_1: b + w_1x_1 + w_2x_2 \ \geq 1, \qquad y_i = +1 \ için \qquad \mathbf{3.26}$$

$$H_2: b + w_1x_1 + w_2x_2 \ \leq -1, \qquad y_i = -1 \ için \qquad \mathbf{3.27}$$

H_1 üzerinde veya üstünde yer alan herhangi bir gözlem +1 sınıfına aitken, H_2 üzerinde veya altında yer alan herhangi bir gözlem −1 sınıfına aittir. Eşitsizliklerin (2.25) ve (2.26) birleştirilmesiyle, Denklem 3.28 elde edilir.

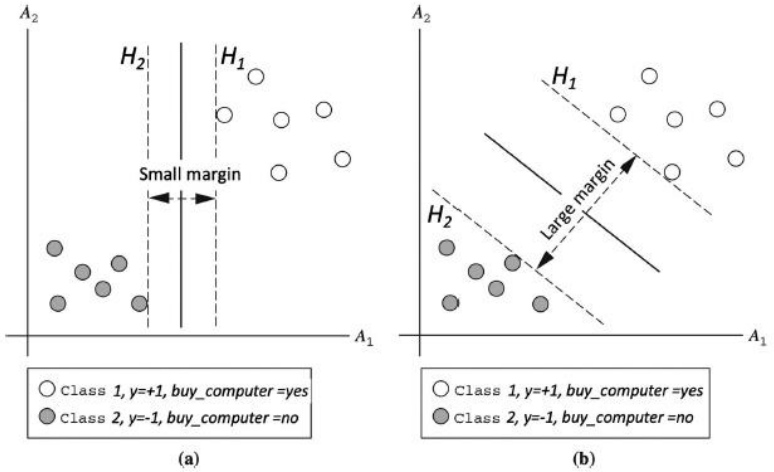
$$y_i(b + w_1x_1 + w_2x_2) \geq 1, \qquad \forall i \qquad \textbf{3.28}$$

H_1 veya H_2 hiper düzlemlerine düşen herhangi bir eğitim gözlemi destek vektörleri olarak adlandırılır. Destek vektörleri maksimum marjin ayırıcı düzlemine eşit uzaklıktadırlar. Destek vektör makineleri maksimum marjin ayırıcı düzlemini (en yakın eğitim örnekleri arasındaki mesafeyi maksimize eden düzlemi) bulur (Şekil 3.7)



Şekil 3.7 Destek Vektörleri (Han vd., 2023)

Aynı veri setine ait iki olası marjin(küçük(a) ve büyük(b)) Şekil 3.8'de gösterilmiştir. Daha büyük marjine sahip olan(b), daha yüksek genelleme doğruluğuna sahiptir.



Şekil 3.8 Küçük ve Büyük Marjin Örneği (Han vd., 2023)

W = {w1, w2, . . . , wn} ise, $\sqrt{W.W}$ denklem 3.29'daki gibi olacaktır.

$$\sqrt{W.W} = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2} \qquad \textbf{3.29}$$

Böylece, $\sqrt{W.W}$ tanım gereği H_2 üzerindeki herhangi bir noktadan ayırıcı hiper düzleme olan mesafeye eşittir. Dolayısıyla, maksimum marjin $\frac{2}{\|W\|}$ olacaktır. Böylece marjini maksimize edilmesi için $\|W\|^2$'nin minimize edilmesi gerekir. Eğer gözlemler n boyutlu bir uzayda ise eşitlik şu şekilde olur (Denklem 3.30):

$$y_i(W'X_i + b) \geq 1 \qquad \mathbf{3.30}$$

Böylece, destek vektör makinesinin matematiksel formülü Denklem 3.31'daki gibidir.

$$\min\|W\|^2,$$

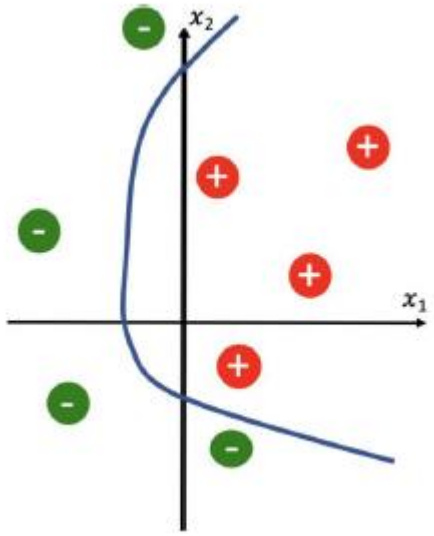
$$y_i(W'X_i + b) \geq 1, \qquad \forall i \qquad \mathbf{3.31}$$

Destek vektör makineleri yeni bir gözlemi sınıflandırırken ise 3.32 numaralı denklemi kullanır.

$$d(X) = \sum_{i=1}^{l} y_i a_i X'X_i + b \qquad \mathbf{3.32}$$

Burada y_i, X_i destek vektörünün sınıf etiketini, X sınıfı tahminlenecek bir gözlemi, ' bir vektörün transpozunu gösterir. α_i ve b, önceden belirtilen optimizasyon veya SVM algoritması tarafından otomatik olarak belirlenen sayısal parametrelerdir. l destek vektörlerinin sayısıdır, genellikle toplam eğitim örneklerinin sayısından çok daha küçüktür. Verilen bir gözlem, X, Denklem 2.32'ye yerleştirilir ve ardından sonucun işareti kontrol edilir. Bu, gözlemin hiper düzlemin hangi tarafında olduğunu gösterir (Han vd., 2023).

Doğrusal br hiper düzlem ile ayrılamayan veri setlerinde çekirdek hileleri (Kernel tricks) kullanılır (Şekil 3.9). Çekirdek hileleri aynı zamanda çekirdek numaraları olarak da bilinir. Polynomial kernel, Gaussian RBF (Radial Basis Function) en sık kullanılan çekirdek fonksiyonlarına örnek verilebilir (Doğançay, 2023).



Şekil 3.9 Doğrusal Olmayan Destek Vektör Makinesi - Kernel Hilesi (Han vd., 2023)

3.2.7 Model Değerlendirme ve Seçimi

Karışıklık matrisi, sınıflandırıcının farklı sınıflara ait örnekleri ne kadar iyi tanıdığını analiz etmek için kullanışlı bir araçtır.

Tablo 3. 1 Karmaşıklık Matrisi

		Tahmin Sınıfı		
		Evet	Hayır	Toplam
Gerçek Sınıf	Evet	DP	YN	P
	Hayır	YP	DN	N
	Toplam	P'	N'	P+N

Tablo 3.1'deki karmaşıklık matrisinde kullanılan terimlerin açıklaması şu şekildedir;

Doğru Pozitifler (DP): Sınıflandırıcı tarafından doğru bir şekilde etiketlenen pozitif örneklerdir.

Doğru Negatifler (DN): Sınıflandırıcı tarafından doğru bir şekilde etiketlenen negatif örneklerdir.

Yanlış Pozitifler (YP): Sınıflandırıcı tarafından yanlış bir şekilde pozitif olarak etiketlenen fakat gerçekte negatif olan örneklerdir.

Yanlış Negatifler (YN): Sınıflandırıcı tarafından yanlış bir şekilde negatif olarak etiketlenen fakat gerçekte pozitif olan örneklerdir.

DP ve DN, sınıflandırıcının doğru tahminler yaptığı durumları belirtirken, YP ve YN sınıflandırıcının yanlış tahminler yaptığı durumları gösterir. İdeal olarak, YP ve YN sıfıra yakın olmalıdır. P' sınıflandırıcı tarafından pozitif olarak etiketlenen örneklerin sayısıdır (DP+YP), N' sınıflandırıcı tarafından negatif olarak etiketlenen örneklerin sayısıdır (YN+DN). Örneklerin toplam sayısı DP+DN+YP+YN veya P'+ N' veya P+N'dir. Gösterilen karışıklık matrisinin ikili sınıflandırma problemi için olduğu gibi, karışıklık matrisleri çoklu sınıflar için benzer şekilde kolayca çizilebilir (Han vd., 2023).

Değerlendirme ölçütlerinden doğruluk (accuracy), bir sınıflandırıcının doğru bir şekilde sınıflandırdığı test seti örneklerinin yüzdesini iade eder. Desen tanıma literatüründe, bu aynı zamanda sınıflandırıcının genel tanıma oranı olarak da adlandırılır; sınıflandırıcının çeşitli sınıfların örneklerini ne kadar iyi tanıdığını yansıtır. Doğruluk, sınıf dağılımı nispeten dengeli olduğunda en etkilidir. Denklem 3.33 ile hesaplanır.

$$Doğruluk = \frac{DP + DN}{P + N} \qquad \mathbf{3.33}$$

Değerlendirme ölçütlerinden hata oranı (error rate), bir sınıflandırıcının yanlış bir şekilde sınıflandırdığı test seti örneklerinin yüzdesini iade eder. Hata oranı (1-Doğruluk) şeklinde hesaplanabilir veya denklem 3.34 ile hesaplanır.

$$Hata\ Oranı = \frac{YP + YN}{P + N} \qquad \mathbf{3.34}$$

Veri setindeki dağılım önemli ölçüde negatif sınıfın çoğunluğunu ve azınlık pozitif sınıfından oluştuğunda sınıf dengesizlik problemi ortaya çıkar. Tıbbi verilerde "kanser" gibi nadir bir sınıf olabilir.

Tıbbi veri örneklerini sınıflandırmak için bir sınıflandırıcı eğitildiğini, burada sınıf etiketi özelliği "kanser" ve olası sınıf değerleri "evet" ve "hayır" olduğu varsayılsın. Bu durumda %97'lik bir doğruluk oranı, sınıflandırıcının oldukça doğru görünmesini sağlar, ancak eğer eğitim örneklerinin sadece %3'ü gerçekten kanser ise %97'lik bir doğruluk oranı kabul edilemez olabilir. Sınıflandırıcı yalnızca kanser olmayan örnekleri doğru bir şekilde etiketleyebilir ve tüm kanser örneklerini yanlış

etiketleyebilir. Bunun yerine, sınıflandırıcının pozitif örnekleri (kanser = evet) ve negatif örnekleri (kanser = hayır) ne kadar iyi tanıyabildiğini değerlendiren diğer değerlendirme ölçütlerine ihtiyaç duyulur.

Duyarlılık (sensitivity) ve belirleyicilik (specificity) ölçümleri bu amaçla kullanılılır. Duyarlılık aynı zamanda doğru pozitif (tanıma) oranı olarak da adlandırılır (doğru şekilde tanımlanan pozitif örneklerin oranı), belirleyicilik ise doğru negatif oranıdır (doğru şekilde tanımlanan negatif örneklerin oranı). Bu ölçütler Denklem 3.35 ve Denklem 3.36 ile tanımlanır.

$$Duyarlılık = \frac{DP}{P} \qquad \mathbf{3.35}$$

$$Belirleyicilik = \frac{DN}{N} \qquad \mathbf{3.36}$$

Doğruluk (accuracy) duyarlılık ve belirleyicilik fonksiyonu olarak gösterilebilir (Denklem 3.37):

$$Doğruluk = duyarlılık\ \frac{P}{P+N} + belirleyicilik\frac{N}{P+N} \qquad \mathbf{3.37}$$

Kesinlik/pozitif öngörü değeri (precision / positive predictive value) ve geri çağırma (recall) ölçütleri de sınıflandırmada yaygın olarak kullanılır. Kesinlik, doğruluğun ölçüsü, pozitif olarak etiketlenen örneklerin ne kadarının gerçekten pozitif olduğunu ifade eder. Geri çağırma ise eksiksizlik ölçüsüdür, gerçek pozitif örneklerin ne kadarının doğru bir şekilde pozitif etiketlendiğini ifade eder. Kesinlik ve geri çağırma Denklem 3.38 ve Denklem 3.39 ile hesaplanır.

$$Kesinlik = \frac{DP}{DP + YP} = \frac{DP}{P'} \qquad \mathbf{3.38}$$

$$Geri\ Çağırma = \frac{DP}{DP + YN} = \frac{DP}{P} \qquad \mathbf{3.39}$$

İyi bir kesinlik (precision) skoru, örneğin C sınıfı için 1.0 değeri, sınıf C'ye ait olduğunu belirttiği her örneğin gerçekten sınıf C'ye ait olduğu anlamına gelir. Ancak bu, sınıf C örneklerinin kaçının yanlış etiketlendiğini ifade etmez.

İyi bir geri çağırma (recall) skoru, örneğin C için 1.0 değeri, sınıf C'den her öğenin bu şekilde etiketlendiği anlamına gelir, ancak sınıf C'ye ait olduğu yanlış etiketlenen diğer örneklerin sayısını ifade etmez.

Kesinlik ve geri çağırma arasında genellikle ters bir ilişki bulunur, birini artırmanın diğerini azaltma maliyetiyle mümkün olduğu durumlar mevcuttur. Örneğin, tıbbi sınıflandırıcı, belirli bir şekilde sunulan tüm kanser örneklerini kanser olarak etiketleyerek yüksek bir kesinlik elde edebilir, ancak birçok başka kanser örneğini yanlış etiketlediği için düşük bir geri çağırma skoru elde edebilir. Kesinlik ve geri çağırma skorları genellikle birlikte kullanılır, kesinlik değerleri sabit bir geri çağırma değeri için karşılaştırılır veya tersi yapılır.

Kesinlik (precision) ve geri çağırma (recall) kullanmanın alternatif bir yolu, bunları tek bir ölçütte birleştirmektir. Bu, F ve F_β ölçütü yaklaşımıdır. F, F1 skoru veya F-skoru olarak da bilinir. Bunlar şu şekilde tanımlanır (Denklem 3.40 ve Denklem 3.41):

$$F = \frac{2\ x\ kesinlik\ x\ geri\ çağırma}{kesinlik +\ geri\ çağırma} \qquad \mathbf{3.40}$$

$$F_\beta = \frac{(1+\beta^2) x\ kesinlik\ x\ geri\ çağırma}{\beta^2\ x\ kesinlik +\ geri\ çağırma} \qquad \mathbf{3.41}$$

Burada β, negatif olmayan bir gerçek sayıdır. F ölçütü, kesinlik ve geri çağırmanın harmonik ortalamasıdır. Kesinlik ve geri çağırmaya eşit ağırlıklar verir. F_β ölçütü ise, kesinlik ve geri çağırmanın ağırlıklı bir ölçüsüdür. Geri çağırmaya, kesinliğe göre β katı ağırlık verir. Yaygın olarak kullanılan F_β ölçüleri F_2 (geri çağırma, kesinliğin iki katı ağırlıkta) ve $F_{0.5}$ (kesinlik, geri çağırmanın iki katı ağırlıkta)'dir (Han vd., 2023).

Yanlış pozitif oranı (false positive rate), yanlış pozitif etiketli örneklerin tüm negatif etiketli örneklere oranıdır. Denklem 3.42 ile hesaplanır.

$$Yanlış\ Pozitif\ Oranı = \frac{YP}{N} \qquad \mathbf{3.42}$$

Yanlış negatif oranı (false negative rate), yanlış negatif etiketli örneklerin tüm pozitif etiketli örneklere oranıdır. Denklem 3.43 ile hesaplanır.

$$Yanlış\ Negatif\ Oranı = \frac{YN}{P} \qquad \mathbf{3.43}$$

Negatif öngörü değeri (negative predictive value), doğru negatif örneklerin toplam negatif tahmin edilen örneklere oranıdır. (Denklem 3.44).

$$Negatif\ \text{Öngörü}\ Değeri = \frac{DN}{DN + YN} = \frac{DN}{N'} \qquad \mathbf{3.44}$$

Pozitif olabilirlik oranı (positive likelihood ratio (LR+)), duyarlılık değerinin yanlış pozitif oranı değerine oranıdır. Belirleyicilik kullanılarak da formüle edilebilir (Denklem 3.45).

$$Pozitif\ Olabilirlik\ Oranı\ (LR+) = \frac{Duyarlılık}{(1 - Belirleyicilik)} \qquad \mathbf{3.45}$$

Negatif olabilirlik oranı (negative likelihood ratio (LR+)), yanlış negatif oranı değerinin negatif öngörü değerine oranıdır. Belirleyicilik ve duyarlılık kullanılarak da formüle edilebilir (Denklem 3.46).

$$Negatif\ Olabilirlik\ Oranı\ (LR-) = \frac{(1 - Duyarlılık)}{Belirleyicilik} \qquad \mathbf{3.46}$$

LR+ değerinin yüksek olması gerçekte pozitif sınıf etiketi almış örneklerin iyi sınıflandırıldığını, LR- değerinin çok düşük olması ise gerçekte sınıf etiketine sahip örneklerin iyi ayrıldığını ifade eder.

Tanısal Üstünlük Oranı (Diagnostic Odds Ratio), tahmin edilen pozitif sınıfın üstünlüğünün negatif sınıfın üstünlüğe oranıdır (Denklem 3.47) (Kurnaz, 2019).

$$Tanısal\ \text{Üstünlük}\ Oranı\ (DOR) = \frac{LR+}{LR-} \qquad \mathbf{3.47}$$

3.2. Kümeleme Yöntemleri

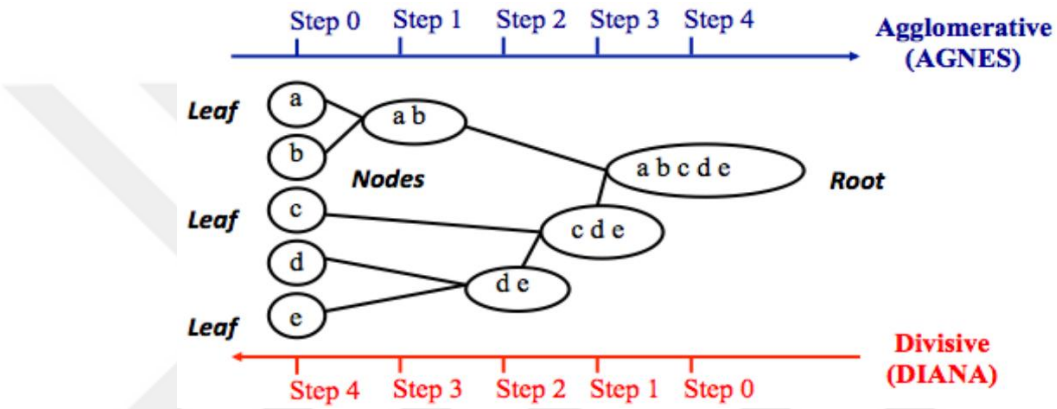
Kümeleme analizi, gözlemleri veya nesneleri aralarındaki uzaklıklara göre küme olarak adlandırılan gruplara bölme işlemidir. Bu yöntemde amaç, bir küme içinde yer alan üyelerin niteliklerinin birbirine çok benzerken (homojen), farklı kümelere ait üyelerin niteliklerinin birbirine benzemeyecek (heterojen) şekilde gruplara bölmektir (Yalçıner Çal, 2023).

Kümeleme analizi ilişkilerin görüntülenebilmesi, anormalliklerin tespiti için kullanılabileceği gibi diğer veri madenciliği yöntemleri için bir ön hazırlık süreci olarak da kullanılabilir. (Taşatan, 2018)

Veri madenciliği alanında kümeleme yöntemleri hiyerarşik ve hiyerarşik olmayan olmak üzere iki ana grupta incelenebilir.

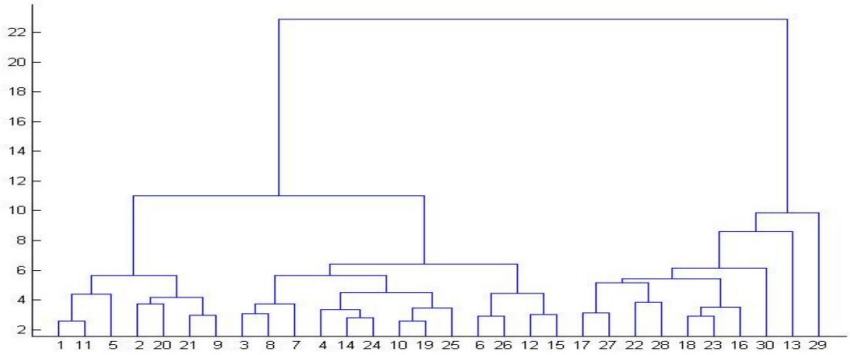
3.2.1. Hiyerarşik Kümeleme Yöntemleri

Hiyerarşik kümeleme yöntemleri ayrıştırıcı ve birleştirici teknikler olmak üzere ikiye ayrılabilir. Birleştirici teknikler, gözlemleri veya nesneleri sürekli bir biçimde birleştirerek kümeler. Birleşerek yeni bir küme oluşturan gözlem sonraki adımlarda ayrılamaz. Analizin başında her bir gözlem birer küme olarak ele alınır. Her bir adımda birbirine en yakın olan kümeler birleştirilerek yeni bir küme elde edilir. Bu işlem "gözlem -1" defa yani tüm kümeler birleşip tek bir küme oluşturuncaya dek devam eder. Analizin sonucunda bu süreç dendogram adı verilen diyagram ile gösterilir. Ayrıştırıcı tekniklerde ise birleştirici tekniklerin adımları tersten izlenir (Taşatan, 2018).



Şekil 3.10. Hiyerarşik Kümeleme Grafiksel Gösterim (Takaoğlu ve Takaoğlu, 2019)

Hiyerarşik kümeleme adımları Şekil 3.10'da incelenebilir. Grafikte birleştirici tekniklerde adımlar mavi işaretli ok ile gösterildiği gibi soldan sağa izlenirken, ayrıştırıcı teknikler tam tersi, kırmızı k ile gösterildiği gibi sağdan sola izlenir (Taşatan, 2018).



Şekil 3.11. Dendogram (Takaoğlu ve Takaoğlu, 2019)

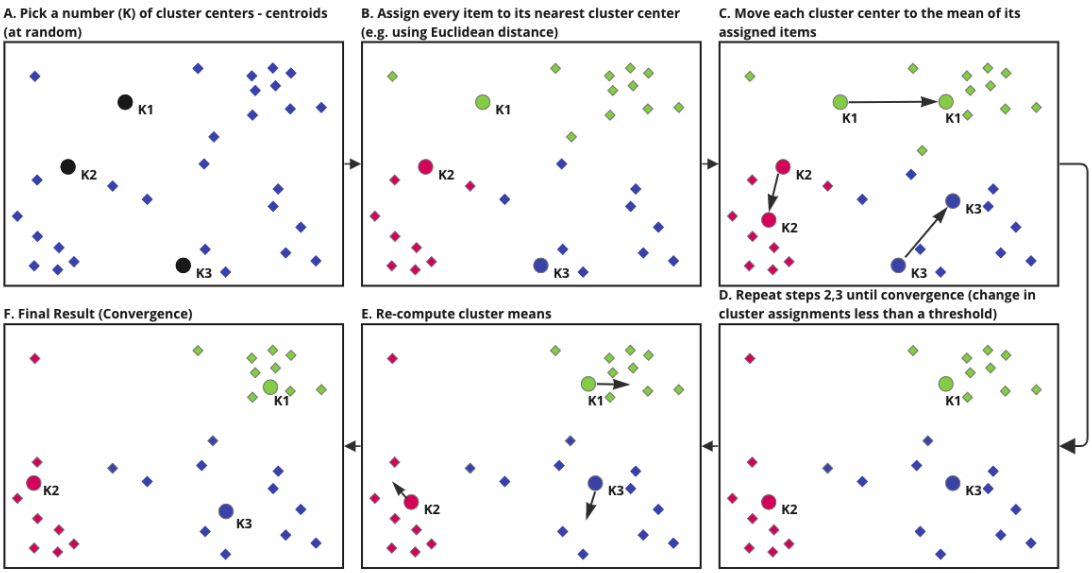
Şekil 3.11'de bir dendogram örneği gösterilmiştir.

Bu tekniklerde en büyük problem gözlemlerin birleşerek kümelere dâhil olma sürecinde analizi nerede durdurmak gerektiğidir. İdeal küme sayısı küme içinde üyeler arası uzaklıklarının minimum, kümeler arasındaki uzaklığın ise maksimum olduğu noktadır. İdeal küme sayısı için farklı yaklaşımlar mevcuttur (Taşatan, 2018).

Hiyerarşik kümeleme yöntemlerine Ward tekniği, Centroid tekniği, tek bağlantı tekniği, tam bağlantı tekniği örnek gösterilebilir (Taşatan, 2018).

3.2.2. Hiyerarşik Olmayan Kümeleme Yöntemleri

Hiyerarşik olmayan kümeleme yöntemlerinde küme sayısı önceden bellidir. Amaç analizle gözlemleri belirlenen K adet kümeye bölmektir. Analizin başında K adet küme için küme merkezleri rastgele seçilir. Her bir gözlem, belirlenen küme merkezlerine göre uzaklıkları hesaplanarak en yakın küme merkezinin kümesine atanır. Bu işlem sonrası oluşturulan kümelerin küme merkezleri tekrar hesaplanır. Hesaplanan yeni küme merkezlerine göre tüm gözlemlerin uzaklıkları tekrar değerlendirilir ve en yakın küme merkezinin olduğu kümelere atamaları yapılır. Bu işlemler, kümeler arası üye geçişi olmayana dek sürdürülür. Hiyerarşik kümeleme yönetiminin aksine bu yöntemlerde gözlemler atandıkları kümelerden ayrılabilirler. Ayrıca hiyerarşik olmayan kümeleme yöntemleri hiyerarşik yöntemlere göre daha büyük veri setlerine uygulanmaya elverişlidir (Taşatan, 2018).



Şekil 3.12. Hiyerarşik Olmayan Kümeleme Yöntemi Adımları (Pandey vd., 2022)

Şekil 3.12'de hiyerarşik olmayan kümeleme adımlarının görsel örneği mevcuttur. İlk adımda seçilen rastgele küme merkezlerinin her adımda nasıl değiştiği gözükmektedir.

Hiyerarşik olmayan kümeleme yöntemlerinde en yaygın kullanılan teknik K-ortalamalar tekniğidir (Taşatan, 2018).

3.2.3. İki Aşamalı Kümeleme Analizi

İki aşamalı kümeleme analizi, hiyerarşik yöntemlerden Ward ile hiyerarşik olmayan yöntemlerden K-ortalamalar tekniklerinin birleşiminden oluşan hibrit bir yaklaşımdır. Bu yöntemin başlıca özellikleri;

- Büyük veri setlerine uygulanabilir
- Hem kategorik hem de sürekli değişkenlerle uygulanabilir
- Küme sayısı konusunda ön bilgi olmadığında kullanılabilir, optimum küme sayını otomatik belirler (Giray, 2016).

İki aşamalı kümeleme analizi, ön kümeleme ve kümeleme olmak üzere iki adımdan oluşur.

Ön kümeleme adımında, gözlemler küçük alt kümelere ayrılır ve ön kümeleme işlemi gerçekleşmiş olunur. Mesafe kriterine dayanarak mevcut kaydı önceden oluşturulmuş kümelerden birine ekleyip eklemeyeceğini veya yeni bir küme başlatıp başlatmayacağını belirler. Burada mesafe kriteri için Öklid veya log-olabilirlik uzaklık ölçüsü kullanılır (Şchiopu, 2010).

İki aşamalı kümeleme analizinde en az bir kategorik değişken mevcutsa log-olabilirlik uzaklık ölçüsü kullanılır. Bu uzaklık ölçüsü olasılığa dayalı mesafeyi temsil eder, kategorik değişkenler için multinomial dağılım, sürekli değişkenler için ise normal dağılım varsayımı yapılır. Bununla beraber değişkenlerin birbirinden bağımsız olduğu varsayılır (Şchiopu, 2010). Eğer tüm değişkenler sürekli ise Öklid uzaklığı kullanılır. Log-olabilirlik uzaklığı, (2.36), (2.37) ve (2.38) numaralı denklemler ile hesaplanır:

$$d(i,j) = \xi_i + \xi_j - \xi_{\langle i,j \rangle} \qquad \mathbf{3.48}$$

Denklem 3.48'de i ve j kümeleri, ξ_i i kümesi içindeki varsayansı, $d(i,j)$ i ve j kümeleri arasındaki log-olabilirlik uzaklık ölçüsünü, $< i,j >$ i ve j kümelerinin birleşiminden oluşan kümeyi ifade eder (Ceylan vd., 2017).

$$\xi_j = -N_j \left(\sum_{k=1}^{K^A} \frac{1}{2} \log\left(\sigma_k^2 + \sigma_{jk}^2\right) + \sum_{k=1}^{K^B} E_{jk} \right) \qquad \mathbf{3.49}$$

Denklem 3.49'da j küme, k kategorik veya sürekli olan değişken, N_j j kümesinin gözlem sayısını, K^A sürekli değişkenlerin toplam adetini, K^B kategorik değişkenlerin toplam adetini, σ_k^2 k sürekli değişkeninin tahminlenmiş varyansını (tüm veri seti bazında), σ_{jk}^2 k sürekli değişkeninin j kümesindeki tahminlenmiş varyansını ifade eder.

$$E_{jk} = -\sum_{l=1}^{L^K} \frac{N_{jkl}}{N_j} \log \frac{N_{jkl}}{N_j} \qquad \mathbf{3.50}$$

Denklem 3.50'de bir kategori olmak üzere, L^K k. kategorik değişkenin kategori sayısını, N_j j kümesinin gözlem sayısını, N_{jkl} bir kategorili k değişkeninin bulunduğu j kümesindeki gözlem sayısını ifade etmektedir.

İşi aşamalı kümele analizi, optimum küme sayısını otomatik belirler. Bunun için Bayesçi Bilgi Ölçütü (Bayesian Information Criterion-BIC) veya Akaike Bilgi Ölçütü (Akaike Information Criterion-AIC) kullanılır. İki bilgi ölçütü kıyaslandığında genel bir kanı olarak kümeleme ve benzeri sınıflandırma tekniklerinde Bayesçi bilgi ölçütünün daha tutarlı kümeler oluşturduğu kabul edilmektedir (Kayri, 2007). İlgili deklemler (3.51), (3.52) ve (3.53) numaraları ile gösterilmiştir:

$$BIC(j) = -2 \sum_{j=1}^{J} \xi_j + m_j \log(N) \qquad \mathbf{3.51}$$

$$AIC(j) = -2 \sum_{j=1}^{J} \xi_j + 2m_j \qquad \mathbf{3.52}$$

$$m_j = J \left\{ 2K^A + \sum_{k=1}^{K^B} L^K - 1 \right\} \qquad \mathbf{3.53}$$

Denklem 2.39, 2.40 ve 2.41'de j küme, k değişken, K^A sürekli değişkenlerin toplam adetini, K^B kategorik değişkenlerin toplam adetini, L^K k kategorik değişkenin kategori sayısını ifade eder.

İki aşamalı kümeleme analizinde kümeleri oluşturan değişkenlerin göreli katkısı (önemi), sürekli ve kategorik değişken türleri için ayrı ayrı hesaplanır. Bu önem değerleri, 0 ile 1 arasında bir derecelendirme ile ifade edilir. 0, kümelerin

belirlenmesinde en az öneme sahip olan değişkeni temsil ederken, 1 ise son derece önemli olan değişkeni ifade eder. Önem ölçütü sürekli değişkenler için t testine dayanır. İlgili denklem (3.54) numarası ile gösterilmiştir:

$$t = \frac{\mu_k - \mu_{jk}}{\sigma_{jk}}\sqrt{N_k} \qquad \mathbf{3.54}$$

Denklem 2.42'de j küme ve k sürekli değişken, N_K veri seti içindeki sürekli değişken sayısını, μ_k k sürekli değişkeninin ortalamasının tahmincisini, μ_{jk} k sürekli değişkeninin j kümesindeki tahminlenmiş ortalamasını, σ_{jk} k sürekli değişkeninin j kümesindeki tahminlenmiş varyansını temsil eder.

Önem ölçütü kategorik değişkenlerde chi-square anlamlılık testine dayanır. İlgili denklem (3.55) numara ile gösterilmiştir (Ceylan vd., 2017):

$$x^2 = \sum_{l=1}^{L^K}\left(\frac{N_{jkl}}{N_{kl}} - 1\right) \qquad \mathbf{3.55}$$

Denklem 2.43'de bir kategori, k kategorik değişken, j küme, L^K k kategorik değişkenin kategori adetini, N_{jkl} bir kategorili k kategorik değişkenin bulunduğu j kümesinin gözlem sayısını, N_{kl} bir kategorili k kategorik değişkenin sayısını ifade eder.

4. LİTERATÜR ARAŞTIRMASI

Veri madenciliği alanında COVID-19 ile ilgili yapılmış çalışmalar aşağıdaki gibi özetlenebilir:

Gök (2021), Brezilya Einstein Hastanesi'nden toplanan laboratuvar sonuçlarını (Hasta yaşı, Ortalama Korpüsküler Hacim, Koronavirus hku1, SarsCov2 Test Sonucu, Monositler, Parainfluenza3, Hematokrit, Ortalama Korpüsküler Hemoglobin, Chylamydophila Pneumoniae, Hemoglobin, Eozinofiller, Adenovirüs, Trombositler, Kırmızı kan hücresi dağılım genişliği, Parainfluenza4, Ortalama trombosit hacmi, Solunum sinsityal virüsü, Koronavirus 229e, Kırmızı kan hücreleri, Influenza a virüsü, Koronavirusoc43, Lenfositler, Influenza b virüsü, inf a h1n1 2009, Ortalama Korpüsküler Hemoglobin Konsantrasyonu, Parainfluenza1, Bordetella pertusis, Lökositler, Koronavirüsni63, Metapneumovirüs, Bazofiller, Rinovirüs enterovirüs) kullanarak makine öğrenmesi yöntemlerini uygulamıştır. Birinci aşamada veri madenciliği yöntemlerini kullanmamış, ikinci aşamada veri madenciliği yöntemlerini kullanarak aynı makine öğrenmesi yöntemlerini uygulamış ve iki aşamayı karşılaştırmıştır. Çalışmanın ilk aşamasında boş gözlem ve nitelikler silinerek çok katmanlı algılayıcı, naive bayes, gradyan artırma, destek vektör makinesi, rastgele orman, karar ağacı ve k-en yakın komşu yöntemleri uygulanmıştır. Bu şamada en iyi sonuç veren grandyan artırma yöntemi veri ön hazırlık aşamasında eksik verileri doldurmak için kullanılmış, veri dengeleme için SMOTE-NC kullanılmıştır. Tekrarlanan uygulamada en başarı sonucu %97 doğruluk ile rastgele orman algoritması vermiştir.

Naseef (2022), Irak Sağlık Bakanlığı tarafından COVID-19 hastaları için oluşturulan özel veri tabanındaki verileri (cinsiyet, yaş, akut flask felç durumu, kardiyovasküler hastalık durumu, nefes darlığı, diyabet durumu, yüksek tansiyon durumu, kanser, karaciğer hastalığı durumu, AIDS, böbrek hastalığı, sigara kullanımı, bronşiyal solunum durumu, gebelik sayısı, ateş, hırıltılı solunum, yutak yanması, bulantı veya kusma, göğüs sıkışması, öksürük, baş ağrısı, siyanoz, burun akıntısı, konvülsiyonlar, şiddetli karışıklık/kontrol kaybı, zayıflık/güçsüzlük, ishal, kronik obstrüktif akciğer hastalıkları, COVID-19 test sonucu) kullanarak hastaları tespit etmeyi amaçlamıştır. Çalışmasında karar ağacı, k-en yakın komşu, lojistik regresyon, doğrusal diskriminant analizi, gaussian naive bayes, destek vektör makinesi, rassal orman, çok katmanlı algılayıcı, gradyan artırma yöntemlerini uygulamış ve en başarılı

sonucu %86,3 doğruluk oranı ile rastgele orman yöntemiyle elde etmiştir.

Tekin (2023), COVID-19 olan ve olmayan kişilerin alışkanlıklarını ve davranışsal niteliklerini sınıflandırma ve birliktelik kuralları yöntemlerini kullanarak incelemiştir. Sınıflandırma çalışmasında random committe (rastgele komite), regularized random forest (düzenlenmiş rassal orman), rassal orman, düzenlenmiş diskriminant analizi, artırma, K-en yakın komşu, sıralı minimal optimizasyon – destek vektör makineleri yöntemleri uygulanarak kişilerin COVID-19 olup olmadıkları tahminlenmiştir. En başarılı sonucu %81 doğruluk oranı ile düzeltilmiş rastgele orman algoritmasıyla elde etmiştir. Birliktelik kuralları çalışmasında ise apriori algoritması uygulanarak COVID-19 hastası olan ve olmayan kişilerin davranışsal kalıpları elde edilmiştir.

Doğançay (2023), çalışmasında COVID-19 hastalarının hastalık evresini tahminlemeyi amaçlamıştır. İzmir'deki bir hastanede COVID-19 tanısı mevcut olan kişilerin laboratuvar (akyuvar, netrofil, lenfosit, monosit, hemoglobin, hemotokrit, ortalama korpisküler hacim, trombosit, procalsitonin, glukoz, üre, kreatinin, aspartat aminotranferaz, sodyum, potasyum, kanser antijen proteini, d-dimer, laktat dehidrogenaz, ferritin, fibrinojen, CRP(C-reaktif protein) ve klinik (kaharitmi, hipotirodi, kronik böbrek yetmezliği, Alzheimer–Parkinson, malignite, epilepsi, koah astım, romatolojik, viral hepatit, siroz, svortoemboli, talasemi taşıyıcı-hersferositoz, öksürük, ateş, nefes darlığı, göğüs ağrısı, baş ağrısı, boğaz ağrısı, bilinç kaybı, halsizlik, tat veya koku kaybı, ishal, miyalji, HIV, cinsiyet, yaş) verileri işlenerek lojistik regresyon, rastgele orman ve destek vektör makineleri algoritmaları uygulanmıştır. Rastgele orman modelinde Gini indeksinden faydalanarak değişken azaltmış ve ikici bir rastgele orman modeli daha kurmuştur. Uygulama sonucunda en başarılı olan model %85,3 doğruluk ile ikinci kurulan rastgele orman modeli olmuştur.

Çotoy (2022), COVID-19 hastalarının yoğun bakım ünitesi gerekliliklerini tahminlemeyi amaçlamıştır. Çalışma için Meksika'daki hastanelere başvuran hastaların verileri (yaş, cinsiyet, semptom günü, entübe durumu, zatürre durumu, hamilelik durumu, diyabet, kronik obstrüktif akciğer hastalığı, astım, inmsupre, hipertansiyon, diğer hastalıklar, kardiyovasküler, obezite, kronik böbrek hastalıkları, sigara kullanma durumu, COVID-19 test sonucu) işlenerek yapay sinir ağı ve SMOTE yöntemleri uygulanmıştır. Uygulama sonucu oluşturulan modelde %79 doğruluk oranıyla tahminleme gerçekleştirilebilmiştir.

Chimbunde vd. (2023), Güney Afrika için yaptıkları çalışmada yoğun bakım ünitesindeki ölüm oranını tahminlemeyi amaçlamıştır. Çalışma için yaş, cinsiyet, hipertansiyon, diyabet, entübasyon durumu, astım, HIV durumu, başvuru sırasındaki semptomların şiddeti (şiddetli/şiddetli değil), CRP (C-reaktif protein), yüksek hassasiyetli troponin T (hs-TnT), Nterminal pro-beyin natriüretik peptid (NT-proBNP), prokalsitonin (PCT), glikolize hemoglobin (HbA1c), D-dimer ve nötrofilmenfosit oranı değişkenleri kullanılmıştır. Veri setine yapay sinir ağı, rastgele orman ve K-ortalamalar tabanlı küme değişkeni de dahil olmak üzere yarı parametrik bir lojistik regresyon algoritmasını uygulayarak yöntemleri karşılaştırmıştır. Uygulama sonucunda %71 doğrulukla en iyi performansı yapay sinir ağı algoritması göstermiştir.

Rai vd. (2023), Hindistan'da yaptıkları çalışmada FP-Growth algoritmasını kullanarak COVID-19 hastalığının bulaşmasının en önemli altı faktörünün solunum sorunu, ateş, kuru öksürük, boğaz ağrısı, yurtdışına seyahat ve kalabalık toplantılara katılmak olduğunu tespit etmiştir. Aynı veri setiyle COVID-19 hastalarını tahminlemek için doğrusal regresyon modeli de geliştirmiştir.

Shakhovska vd. (2021), COVID-19'dan etkilenenlerin sık görülen kalıpları ve parametreleri bulmayı amaçlamıştır. Ukrayna, Beyaz Rusya ve Almanya'dan alınan veriler ile veri madenciliği yöntemlerinden kümeleme ve sınıflandırma çalışmalarında bulunmuşlardır. Kümeleme yöntemlerinden K-ortalamalar, sınıflandırma yöntemlerinden karar ağaçları, yapay sinir ağı, rastgele orman, ekstrem gradyan artırma ve naive bayes yöntemleri kullanılmıştır. Yapılan çalışmada yaş, cinsiyet, bölge(ülke), sigara içme durumu, COVID-19 geçirme durumu, IgM seviyesi, IgG seviyesi, kan grubu, grip aşısı kullanma durumu, tüberküloz aşısı kullanma durumu, ilgili yılda grip olma durumu ve ilgili yılda tüberküloz olma durumu değişkenleri kullanılmıştır. Uygulama sonucunda en başarılı olan model %89,9 doğruluk ile ekstrem gradyan artırma modeli olmuştur.

Kumar vd. (2024), yaptıkları çalışmada hastaların COVID-19 olup olmadığını tahminlemeyi hedeflemişlerdir. Çalışmadaki veriler Kaggle'dan elde edilmiş ve değişken olarak nefes alma problemleri, ateş, kuru öksürük, boğaz ağrısı, burun akıntısı, astım, kronik akciğer hastalığı, baş ağrısı, kalp hastalıkları, diyabet, hipertansiyon, halsizlik, mide ve bağırsak problemleri, yurtdışı seyahat, COVID-19 hastalarıyla temas, kalabalık toplantılara katılım, kamuya açık alanlarla etkileşim,

kamuya açık alanlarda çalışan aile, maske takma durumu ve dezenfektan kullanım durumu kullanılmıştır. Sınıflandırma çalışmaları için lojistik regresyon, destek vektör makineleri, naive bayes, rastgele orman ve K-en yakın komşu algoritmaları kullanılmıştır. Bahsi geçen algoritmalar özellik seçilmeden ve özellik seçimi uygulanarak iki defa kurulmuş, özellik seçimi sonrası naive bayes ve rastgele orman modellerinde dikkate değer değişimler elde edilmiştir. Çalışma sonucunda %98 doğruluk ile özelik seçimi sonrası kurulan rastgele orman modeli olmuştur.

Literatürde veri madenciliği yöntemleri kullanılarak COVID-19 üzerine gerçekleştirilmiş bazı çalışmalar Tablo 4.1'de özetlenmiştir. Literatürde veri madenciliği yöntemlerinden sınıflandırma yöntemleri kullanılarak COVID-19 konusunda yapılmış bazı çalışmalar ise Tablo 4.2'de özetlenmiştir.

Veri madenciliğinde kümeleme analizi alanında COVID-19 ile ilgili yapılmış çalışmalar aşağıdaki gibi özetlenebilir:

Paşin ve Paşin (2020), ülkelerin toplam nüfus büyüklüğünü dikkate alarak toplam ölüm sayısı ve toplam vaka sayısı göstergeleri ile 191 adet ülkeyi kümelendirmeyi amaçlamıştır. Yöntem olarak iki aşamalı kümeleme analizi ve k-ortalamalar yöntemi tercih edilmiş ve kıyaslanmıştır. K-ortalamalar yöntemi sonucunda 4 küme elde edilirken iki aşamalı kümeleme analizinde 3 küme elde edilmiştir.

Chan vd. (2021), Umman'daki sağlık çalışanlarını demografik özellikler ve ruh sağlığı ölçümlerine göre profillerini belirleyerek kümelemeyi amaçlamıştır. Çalışmada iki aşamalı kümeleme analizi kullanılmıştır. Çalışma sonucunda A(düşük riskli ve en az etkilenmiş), B (yüksek riskli ve orta düzeyde etkilenmiş olmak üzere) ve C(yüksek riskli ve yüksek düzeyde etkilenmiş olmak üzere) 3 küme elde edilmiştir. C kümesini daha genç ve daha az tecrübeli sağlık çalışanlarının oluşturması dikkat çekmiştir.

Abdullah vd. (2022), Endonezya'daki 34 ili ölümle sonuçlanmış ve iyileşmiş vaka verilerine göre K-ortalamalar yöntemi ile kümeleyerek hükümete hastalığın yayılmasını önleyici politikalar için girdi oluşturmayı amaçlamışlardır. Çalışma sonucunda 5 üye, 28 üye ve 1 üye sahibi toplamda 3 grup elde edilmiştir.

Demircioğlu ve Eşiyok (2020), yaptıkları çalışmada COVID-19 salgını sonuçlarını AB ve OECD üyesi olan 36 ülkenin verileri ile ülkeler bazında incelemeyi amaçlamışlardır. Araştırmada; vaka/nüfus oranı, ölüm/nüfus oranı, iyileşen/nüfus

oranı, test sayısı/nüfus oranı, doktor sayısı, hemşire sayısı, sağlık harcamaları, yatak sayısı, yoğun bakım yatak sayısı ve yaşlı nüfus/nüfus oranı değişkenleri kullanılmış ve kümeleme yöntemlerinden K-ortalamalar yöntemi uygulanmıştır. Uygulamada 36 ülkelerin 2, 3 ve 4 kümeye ayrıldığı üç farklı varyasyon incelenmiştir.

Kumar (2020), Hindistan şehirlerini COVID-19 durumlarına göre değerlendirmeyi amaçlamıştır. Çalışmada teyit edilmiş vakalar, ölüm vakaları ve tedavi edilmiş vaka değişkenleri kullanılarak veri madenciliği yöntemlerinden kümeleme analizi uygulanmıştır. Kümeleme tekniklerinden hiyerarşik kümeleme yöntemi olan Ward yöntemi tercih edilmiştir.

Gençer Çelik ve Öngel (2021), yaptıkları çalışmada sağlık çalışanlarının online market alışveriş deneyimlerini meydana getiren alt faktörler arası ilişkileri ve profilleri belirlemeyi amaçlamışlardır. Çalışmada müşteri hizmetleri kalitesi, sipariş sistemi kalitesi, güven faktörü, kurye kalitesi ve zorlayıcı psikolojik faktörler dikkate alınmıştır. Alt faktörler arası ilişkilerin saptanması için korelasyon analizi, profillerin saptanması için iki aşamalı kümeleme analizi yöntemleri tercih edilmiştir. Çalışma sonucunda sağlık çalışanlarının iki kümeye ayrılmıştır. Her iki kümede de sipariş sistemi kalitesi algısı ok yüksek olarak saptanmıştır. İlk kümedeki sağlık çalışanlarının müşteri hizmet kalitesi, güven düzeyi ve zorlayıcı psikolojik faktörler algısının da çok yüksek olarak saptanmasına rağmen kurye kalitesi algısının kararsızlık düzeyinde kaldığı görülmüştür. İkinci kümedeki sağlık çalışanları ise müşteri hizmet kalitesi ve zorlayıcı psikolojik faktörler algısı düşük iken güven düzeyi ve kurye kalitesi algısının düşük olduğu saptanmıştır.

Kartal vd. (2021), güncel COVID-19 verilerine göre 209 ülkenin özet durum ve analizini oluşturan çevrimiçi dinamik bir uygulama geliştirmiştir. Çalışmada, Doğrusal Değişim Oranı (DDO), Üstel Büyüme Katsayısı (ÜBK) ve vaka sayısının ikiye katlanması için gereken gün sayısı gibi değişkenler hesaplanarak K-ortalamalar kümeleme analizi yöntemi uygulanmıştır. Kümeleme analizi ile ülkeler COVID-19 durumlarına göre iyi, kritik ve kötü olmak üzere üç kümeye ayrılmıştır.

Peñas vd. (2023), hastalığın seyrini öngörmek için COVID-19 sonrası semptomları sergileyen, COVID-19'u atlatmış olan kişileri gruplandırmayı amaçlamıştır. İspanya'daki beş farklı hastaneden iyileşenler hasta kayıtları toplanmış ve hastaların taburcu olmasından ortalama 8,4 ay sonra kişilerin COVID-19 sonrası

semptomları değerlendirilmiştir. Kümeleme için K-ortalamalar yöntemi kullanılmış ve uygulama sonunda üç küme elde edilmiştir.

Literatürde COVID-19 konusunda yapılmış bazı kümeleme çalışmalarının, hangi kümeleme yöntemleri kullanılarak yapıldığı Tablo 4.3'de özetlenmiştir.

Tablo 4. 1 COVID-19 Üzerine Gerçekleştirilmiş Veri Madenciliği Çalışmaları

Yazar(lar)	Tür	Yıl	Kümeleme	Sınıflandırma	Birliktelik Kuralları	Konu	Ülke
Tekin	Yüksek Lisans Tezi	2023		X	X	COVID-19 Hastalarının Tespiti, COVID-19 Hasta Olan ve Olmayan Kişilerin Davranışsal Kalıplarının Keşfedilmesi	Türkiye
Doğançay	Yüksek Lisans Tezi	2023		X		COVID-19 Hastalarının Hastalık Evresini Tahminlemek	Türkiye
Naseef	Yüksek Lisans Tezi	2022		X		COVID-19 Hastalarının Tespiti	Türkiye
Çotoy	Yüksek Lisans Tezi	2022		X		COVID-19 Hastalarının Yoğun Bakım Ünitesi Gerekliliklerinin Tahmini	Türkiye
Gök	Yüksek Lisans Tezi	2021		X		Kan Değerleri ile COVID-19 Enfekte Düzeyinin Tahminlenmesi	Türkiye
Chimbunde vd.	Makale	2023		X		Yoğun Bakım Ünitelerindeki Ölüm Oranlarının Tahminlenmesi	Güney Afrika
Rai vd.	Makale	2023			X	COVID-19 Hastalığının Bulaşmasının İncelenmesi	Hindistan
Peñas vd.	Makale	2023	X			İyileşen Hastaları COVID-19 Sonrası Semptomlarına Göre Kümeleme	İspanya
Abdullah vd	Makale	2022	X			Şehirleri Ölümle Sonuçlanmış ve İyileşmiş Vakalara Göre Kümeleme	Endonezya

Tablo 4.1 (Devam)

Yazar(lar)	Tür	Yıl	Kümeleme	Sınıflandırma	Birliktelik Kuralları	Konu	Ülke
Shakhovska vd.	Makale	2021	X	X		COVID-19 Vakalarının İncelenmesi	Ukrayna
Kumar vd.	Makale	2024		X		COVID-19 Hastalarının Tespiti	Hindistan
Chan vd	Makale	2021	X			Sağlık Çalışanlarını Demografik Özellikler ve Ruh Sağlığı Ölçümlerine Göre Kümeleme	Umman
Gençer Çelik ve Öngel	Makale	2021	X			Sağlık Çalışanarının Online Market Alışveriş Deneyimlerini Meydana Getiren Faktörler Arası İlişkileri ve Profillerin Belirlenmasi	Türkiye
Pasin ve Pasin	Makale	2020	X			Ülkeleri Ölüm Oranı ve COVID-19 Vakalarına Göre Kümeleme	İtalya
Demircioğlu ve Eşiyok	Makale	2020	X			COVID–19 Salgını ile Mücadelede Ülkelerin Sınıfladırılması	Türkiye
Kartal vd.	Makale	2020	X			COVID-19 Salgının Dünyada ve Türkiye'de Değişen Durumunun İncelenmesi	Türkiye
Kumar	Makale	2020	X			Hindstan Şehirlerinin COVID-19 Durumlarına Göre Değerlendirimesi	Hindistan

Tablo 4.2 COVID-19 Üzerine Gerçekleştirilmiş Sınıflandırma Çalışmaları

Yazar (lar)	Tür	Yıl	Çok Katmanlı Algılayıcı	Destek Vektör Makinesi	Diskriminant Analizi	Gradyan Artırma	Karar Ağacı	K-En Yakın Komşu	Lojistik Regresyon	Naive Bayes	Rastgele Orman	Yapay Sinir Ağı	Konu	Ülke
Tekin	Yüksek Lisans Tezi	2023		X	X	X		X			X		COVID-19 Hastalarının Tespiti, COVID-19 Hasta Olan ve Olmayan Kişilerin Davranışsal Kalıplarının Keşfedilmesi	Türkiye
Doğançay	Yüksek Lisans Tezi	2023		X					X		X		COVID-19 Hastalarının Hastalık Evresini Tahminlemek	Türkiye
Naseef	Yüksek Lisans Tezi	2022	X	X	X	X		X	X	X	X		COVID-19 Hastalarının Tespiti	Türkiye

Tablo 4.2 (Devam)

Yazar (lar)	Tür	Yıl	Çok Katmanlı Algılayıcı	Destek Vektör Makinesi	Diskriminant Analizi	Gradyan Artırma	Karar Ağacı	K-En Yakın Komşu	Lojistik Regresyon	Naive Bayes	Rastgele Orman	Yapay Sinir Ağı	Konu	Ülke
Çotoy	Yüksek Lisans Tezi	2022										X	COVID-19 Hastalarının Yoğun Bakım Ünitesi Gerekliliklerinin Tahmini	Türkiye
Gök	Yüksek Lisans Tezi	2021	X	X		X	X	X		X	X		Kan Değerleri ile COVID-19 Enfekte Düzeyinin Tahminlenmesi	Türkiye
Chimbunde vd.	Makale	2023							X		X	X	Yoğun Bakım Ünitelerindeki Ölüm Oranlarının Tahminlenmesi	Güney Afrika
Shakhovska vd.	Makale	2021				X				X	X	X	COVID-19 Vakalarının İncelenmesi	Ukrayna
Kumar vd.	Makale	2024		X				X	X	X	X		COVID-19 Hastalarının Tespiti	Hindistan

Tablo 4.3. COVID-19 Üzerine Gerçekleştirilmiş Kümeleme Çalışmaları

Yazar(lar)	Tür	Yıl	K-Ortalamalar	İki Aşamalı Kümeleme	Ward	Konu	Ülke
Peñas vd.	Makale	2023	X			İyileşen Hastaları COVID-19 Sonrası Semptomlarına Göre Kümeleme	İspanya
Abdullah vd	Makale	2022	X			Şehirleri Ölümle Sonuçlanmış ve İyileşmiş Vakalara Göre Kümeleme	Endonezya
Shakhovska vd.	Makale	2021	X			COVID-19 Vakalarının İncelenmesi	Ukrayna
Chan vd	Makale	2021		X		Sağlık Çalışanlarını Demografik Özellikler ve Ruh Sağlığı Ölçümlerine Göre Kümeleme	Umman
Gençer Çelik ve Öngel	Makale	2021		X		Sağlık Çalışanarının Online Market Alışveriş Deneyimlerini Meydana Getiren Faktörler Arası İlişkileri ve Profillerin Belirlenmasi	Türkiye
Pasin ve Pasin	Makale	2020	X	X		Ülkeleri Ölüm Oranı ve COVID-19 Vakalarına Göre Kümeleme	İtalya
Demircioğlu ve Eşiyok	Makale	2020	X			COVID–19 Salgını ile Mücadelede Ülkelerin Sınıfladırılması	Türkiye
Kartal vd.	Makale	2020	X			COVID-19 Salgının Dünyada ve Türkiye'de Değişen Durumunun İncelenmesi	Türkiye
Kumar	Makale	2020			X	Hindstan Şehirlerinin COVID-19 Durumlarına Göre Değerlendirimesi	Hindistan

Veri madenciliği kümeleme analizi alanında yapılmış tez çalışmaları aşağıdaki gibi özetlenebilir:

Kişi (2021), yaptığı çalışmada Avrupa Birliği ülkeleri ve aday ülkeleri olmak üzere toplamda otuz bir ülkeyi sürdürülebilir kalkınma hedefleri göstergelerine göre kümelemeyi amaçlamıştır. Uygulamada veri madenciliği yöntemlerinden K-ortalamalar tekniği kullanılmıştır. Analiz sonucunda dört küme elde edilmiştir.

Baca (2022), çalışmasında Türkiye illerini sektörel bazda kümelemeyi amaçlamıştır. İstihdam verilerini kullanarak büyüklük, başatlık ve uzmanlaşma verilerine ayrı analizler uygulamıştır. Uygulamada hiyerarşik kümeleme tekniklerinden Ward ve hiyararşik olmayan kümeleme tekniklerinden K-ortalamalar kullanılmıştır.

Dağaslanı (2022), çalışmasında perakende sektöründe faaliyet gösteren bir firmanın müşteri profillerinin oluşturulmasını amaçlamıştır. Müşterilerin günlük alışveriş hareketleri verisi kullanılarak veri madenciliği yöntemlerinden K-ortalamalar tekniği uygulanmıştır. Analiz sonucunda dört müşteri kümesi oluşmuştur.

Doğan (2023), çalışmasında 2021-2022 yıllarında piyasa hakimiyeti yüksek olan kripto paraları kümelemeyi amaçlamıştır. Uygulamada hiyerarşik kümeleme yöntemlerinden tek bağlantı yöntemi, tam bağlantı yöntemi ve Ward, hiyerarşik olmayan kümeleme yöntemlerinden K-ortalamalar kullanılmıştır.

Aylıkcı (2023), çalışmasında filo firmalarının hasarlarını, hasar özelliklerine göre kümelendirmeyi amaçlamıştır. Çalışma kullanım şekli hususi otomobil olan araçlar üzerine yapılmıştır. Uygulama için K-ortalamalar yöntemi kullanılmıştır. Analiz sonucunda iki küme elde edilmiştir.

Sarı (2023), çalışmasında Türkiye illerini havadaki partiküler madde ve Kükürt Dioksit değerlerine göre kümelendirmeyi amaçlamıştır. Ülke genelindeki hava izleme istasyonlarına ait 2016-2022 yılı verilerini kullanarak hiyerarşik kümeleme yöntemlerinden Ward uygulanmıştır.

Eser (2019), çalışmasında PISA sonuçlarına göre fen bilgisi öğretimine ilişkin öğrencileri kümelemeyi amaçlamıştır. PISA 2015 veri seti kullanılarak K-ortalamalar, iki aşamalı kümeleme analizi ve Kohonen'in öz örgütlemeli harita yöntemi uygulanarak yöntemler karşılaştırılmıştır.

Literatürde veri madenciliği yöntemlerinden kümeleme teknikleri kullanılarak hazırlanmış bazı tez çalışmaları Tablo 4.4'de özetlenmiştir.

Literatürde COVID-19 hasta özniteliklerine göre yapılan iki aşamalı kümeleme analizi uygulaması örneği bulunamamıştır. COVID-19 hasta öznitelikleri kullanılarak COVID-19 pozitiflik durumunu tahminleyen sınıflandırma çalışmaları mevcut olsa da bu çalışmada kullanılan özniteliklerle yapılmış bir çalışma bulunamamıştır. Böylece çalışmanın özgünlüğü ile sağlık kuruluşlarının yanı sıra literatüre de katkı sağlaması hedeflenmektedir.

Tablo 4.4. Kümeleme Alanında Yapılmış Tez Çalışmaları

Yazar(lar)	Tür	Yıl	K-Ortalamalar	İki Aşamalı Kümeleme	Ward	Tek Bağlantı	Tam Bağlantı	Konu
Kişi	Yüksek Lisans Tezi	2021	X					Ülkelerin Sürdürülebilir Kalkınma Hedeflerine Göre Kümelenmesi
Baca	Yüksek Lisans Tezi	2022	X		X			Türkiye İllerinin Sektörel Bazda Kümelenmesi
Dağaslanı	Yüksek Lisans Tezi	2022	X					Perakende Sektöründe Müşteri Profillerinin İncelenmesi
Doğan	Yüksek Lisans Tezi	2023	X		X	X	X	Kriptoparalarda Kümeleme Analizi
Aylıkcı	Yüksek Lisans Tezi	2023	X					Filo Araçlarının Kümeleme Yöntemiyle Hasar Analizi
Sarı	Yüksek Lisans Tezi	2023			X			Türkiye İllerini Havadaki Partiküler Madde ve Kükürt Dioksit Değerlerine Göre Kümelendirilmesi
Eser	Doktora Tezi	2019	X	X				PISA Sonuçlarına Göre Fen Bilgisi Öğretimine İlişkin Öğrencilerin Kümelenmesi

5. UYGULAMA

Veri madenciliği yöntemleri ile COVID-19 şüphesi taşıyan hastaların değerlendirilmesinin hedeflendiği bu çalışmada öncelikle ayakta tedavi görmüş hastaların verileri kullanılarak sınıflandırma algoritmaları uygulanmıştır. Hastaların laboratuvar test sonuçları kullanılarak COVID-19 test sonucunu tahminlemeyi hedefleyen modeller oluşturulmuştur.

Uygulamanın ikinci bölümünde ise veri seti COVID-19 test sonucu negatif ve pozitif olan hastalar için ayrılmış ve iki aşamalı kümeleme analizi ile hasta profillerinin ortaya çıkarılması hedeflenmiştir.

Bu çalışmada, sınıflandırma ve kümeleme uygulaması için toplam 35 öznitelik kullanılmaktadır. COVID-19 test sonucu, hasta türü ve cinsiyet öznitelikleri kategorik değişken, kalan tüm öznitelikler ise sürekli değişkendir. Uygulamada kullanılan öznitelikler aşağıda açıklanmıştır.

- Albümin, kan plazmasında bulunan bir proteindir.
- Alkalen Fosfataz, bir karaciğer enzim türüdür.
- ALT (Alanin Aminotransferaz), bir karaciğer enzim türüdür.
- APTT (Aktif Parsiyel Tromboplastin Zamanı), kanın pıhtılaşmasını karakterize eden bir kan testidir.
- AST (Aspartat Aminotransferaz), bir karaciğer enzim türüdür.
- Bilirubin, kanda doğal olarak bulunan, eski kırmızı kan hücrelerinin parçalanması ile oluşan sarımsı bir pigmenttir. Direkt Bilirubin, suda çözünebilen Bilirubindir. İnirekt Bilirubin, suda çözünemeyen Bilirubindir.
- Cinsiyet, COVID-19 testi yapılmış olan hastanın cinsiyetini belirtir.
- CRP, “C-reaktif protein” ifadesinin kısaltmasıdır, bu protein türünün kandaki seviyesini ifade eder.
- Ferritin, demiri hücrelerde depolayan ve salınımını yapan demir içeren kristal yapıda bir kan proteinidir.
- Fibrinojen, karaciğer tarafından üretilen, kan pıhtılarının oluşmasına yardımcı olarak kanamayı durdurmaya yardımcı olan proteindir.
- GGT, “Gama Glutamil Transferaz” ifadesinin kısaltmasıdır, çoğunlukla karaciğerde bulunan bir enzimdir.

- Glukoz, bir karbonhidrat türüdür, bu öznitelik glukozun kandaki seviyesini ifade eder.
- Hasta türü, hastanın tedavi türünü ifade eder. Ayaktan tedavi gören, yatarak tedavi gören ve günübirlik tedavi gören olmak üzere iki kategoriden oluşur.
- HDL-Kolesterol, iyi kolesterol olarak tanımlanan yağ türü olan kolesterolü kanda taşımakla görevli hem yağ hem de proteinlerden oluşan bir yapıdır.
- HGB, hemoglobinin kısaltmasıdır, kandaki hemoglobin miktarını ifade eder.
- Kalsiyum, bir mineral çeşididir, bu öznitelik kalsiyumun kandaki miktarını ifade eder.
- Kreatin Kinaz, bir protein türüdür, bu öznitelik Kreatin Kinaz proteinin kandaki seviyesini ifade eder.
- Kreatinin bir amino asit türüdür, bu öznitelik Kreatinin'in kandaki seviyesini ifade eder.
- LYM%, kandaki lenfosit oranını ifade eder.
- MONO%, kandaki monosit oranını ifade eder.
- PO2, kan gazlarından biridir, bu öznitelik PO2 kan gazının kandaki seviyesini ifade eder.
- Procalcitonin, bir protein türüdür, bu öznitelik Procalcitonin proteininin kandaki seviyesini ifade eder.
- PT(Protrombin Time), kanın pıhtılaşmasında etkin olan Protrombin proteini ile ilgili zaman testidir. PT/%, PT/INR, PT/sn bu testin çeşitleridir.
- RBC, "red blood cells" ifadesinin kısaltmasıdır, kandaki kırmızı kan hücrelerini (eritrosit) ifade eder.
- Sonuç, hastanın COVID-19 test sonucunu ifade eder.
- Total Demir Bağlama Kapasitesi, serumdaki demir bağlayan bölgelerin demire ne dereceye kadar doyurulabildiğinin ölçütüdür.
- Total Protein, kan plazmasında var olan proteinlerin miktarını ifade eder.
- Troponin I, kalp ile ilşkili bir çeşit proteindir.
- Üre, vücuttaki bir atık türüdür, kandaki üre seviyesini ifade eder.

- WBC, “white blood cells” ifadesinin kısaltmasıdır, kandaki beyaz kan hücrelerini (lökosit) ifade eder.
- Yaş değişkeni hastanın test yapıldığı zamandaki yaşını ifade eder.

5.1 Sınıflandırma Yöntemleri ile Uygulama

Bu bölümde ayakta tedavi gören hastaların laboratuvar sonuçlarından hastanın COVID-19 test sonucunun tahminlenmesi hedeflenmiştir. Çalışmada, Samsun ilinde bir hastaneye ait 2021 yılında COVID-19 testi yaptırmış olan hastaların verileri kullanılmıştır. Bu uygulama için açık kaynak kodlu KNIME programı kullanılmıştır.

5.1.1 Veri Ön Hazırlık İşlemleri

Veri toplama aşamasında hastanenin bilgi yönetim sistemi kullanılarak 2021 yılına ait veriler çekilmiştir. Veri birleştirme adımı için, COVID-19 testi yaptırmış olan hasta listesi elde edildikten sonra diğer test sonuçlarına ilişkin veriler hastane bilgi yönetim sisteminden çekilerek birleştirilmiştir. Bu aşamada 44x10429’luk bir veri matrisi elde edilmiştir. Matris; hastanın COVID-19 test sonucu, cinsiyet, yaş, Ferritin, Troponin I, Procalcitonin, Kortizol, CRP (Nefolometrik), PT/sn, PT/INR, PT/%, APTT, Fibrinojen, WBC, RBC, HGB, LYM%, MONO%, PO2, ALT, Albümin, Alkalen, Fosfataz, Amilaz, AST, Bilirubin Direk, Bilirubin İndirek, Bilirubin Total, CRP, D-Dimer, GGT, Glukoz, HDL-Kolesterol, Kalsiyum (Ca), Klor (Cl), Kreatin Kinaz (CK), Kreatinin, Magnezyum, Potasyum (K), Sodyum (Na), Total Demir Bağlama Kapasitesi, Total Protein, Trigliserid, Üre, Ürik Asit özniteliklerinden oluşmaktadır.

Veri setindeki uygunsuz veriler tespit edilmiştir. Uygunsuz veriler dokuz gruba ayrılmıştır. Sınıflandırılmış veri, x> veya >x şeklinde ifade edilen sonuçlardır. İlgili öznitelikler sürekli veri olarak analiz edileceğinden uygunsuzdur. Hatalı sonuç sonuç yok, hatalı test, negatif test değerleri, “*” veya “-” değerlerini ifade eder.

Tablo 5. 1 Uygunsuz Verilerin Tespiti

Öznitelik / Uygunsuz Veri	Hatalı Sonuç	Hatalı Test istemi	Hemolizli Numune	Lipemik Numune	Pıhtılı Numune	Sınıflandırılmış Veri	Tekrarı Uygundur	Uygun Olmayan Kap	Yetersiz Numune	*TOPLAM*
Ferritin	9		3			17				***29***
Troponin I		1	1			1.880				***1.882***
Procalcitonin						9				***9***
Kortizol		1				1				***2***
CRP (Nefolometrik)		1				372			1	***374***
PT/sn	2		7	1	12				15	***37***
PT/INR	1		8	1	12				13	***35***
PT/%		1	8	1	13				15	***38***
APTT	2	2	7	1	12	3	4	1	17	***49***
Fibrinojen	1		2		6	1			6	***16***
WBC		1			1			1	5	***8***
RBC		1			1			1	5	***8***
HGB		1			1			1	5	***8***
LYM%	1	1			1			1	5	***9***
MONO%	1	1			1			1	5	***9***
PO2					17					***17***
ALT			11	1	1				1	***14***
Albümin			12						1	***13***
Alkalen Fosfataz			12							***12***
Amilaz			13		1				1	***15***
AST			12	1	1				1	***15***
Bilirubin Direk			14						1	***15***

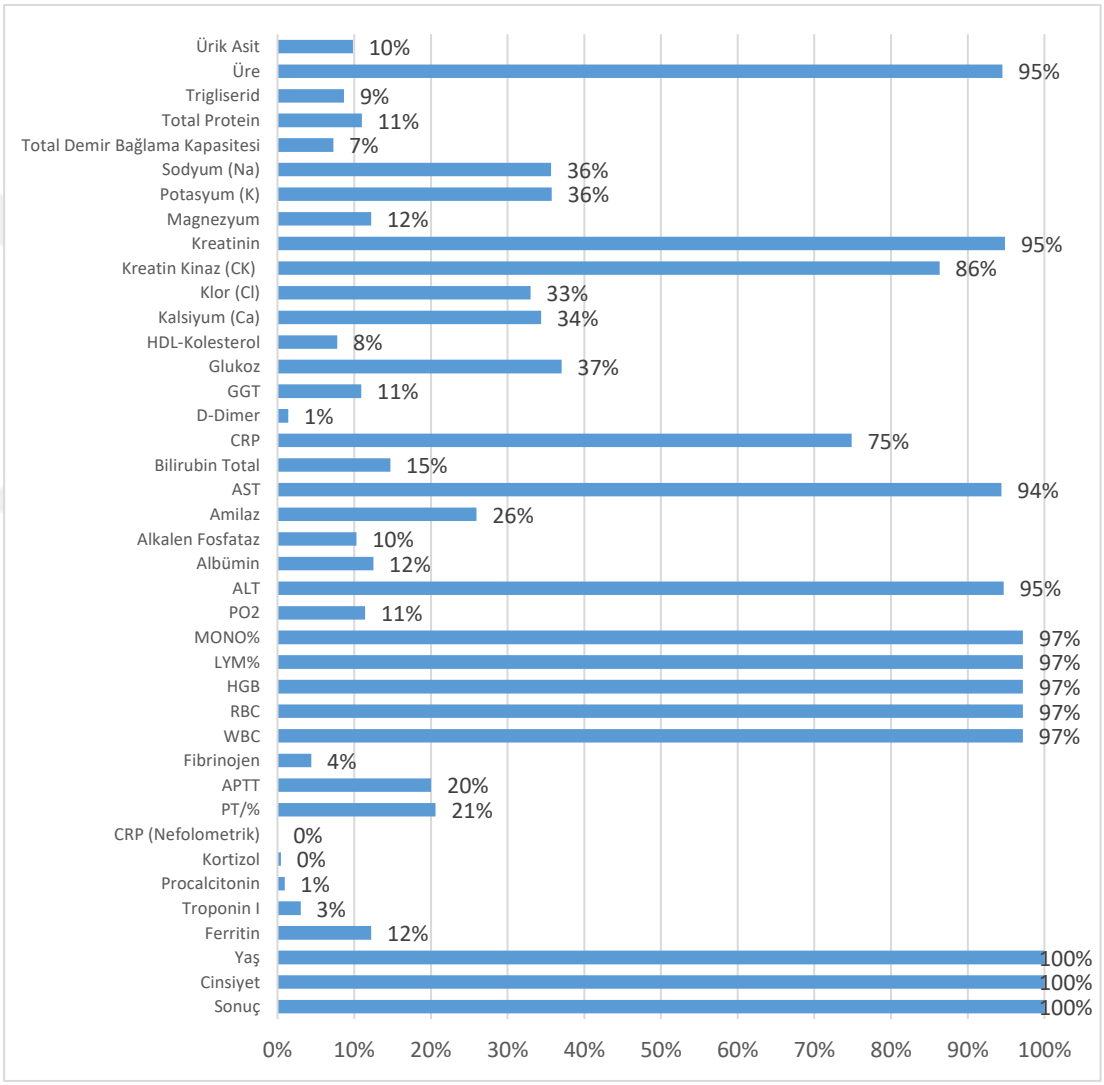
Tablo 5.1 (Devam)

Öznitelik / Uygunsuz Veri	Hatalı Sonuç	Hatalı Test istemi	Hemolizli Numune	Lipemik Numune	Pıhtılı Numune	Sınıflandırılmış Veri	Tekrarı Uygundur	Uygun Olmayan Kap	Yetersiz Numune	*TOPLAM*
Bilirubin İndirek	15	1	7							***23***
Bilirubin Total	7		15						1	***23***
CRP	8		9						1	***18***
D-Dimer					1	1			1	***3***
GGT	12		11							***23***
Glukoz			14		1				1	***16***
HDL-Kolesterol		1	13							***14***
Kalsiyum (Ca)	1		11		1				3	***16***
Klor (Cl)	2		16	1	1				1	***21***
Kreatin Kinaz (CK)	1		11		1				1	***14***
Kreatinin			12		1				1	***14***
Magnezyum	10		12							***22***
Potasyum (K)	2		22	1	1				1	***27***
Sodyum (Na)	2		15	1	1				1	***20***
Total Demir Bağlama Kapasitesi			8			2				***10***
Total Protein	11	1	16							***28***
Trigliserid		1	13							***14***
Üre	1		12	1	1				1	***16***
Ürik Asit	3		10							***13***
TOPLAM	***92***	***15***	***337***	***10***	***89***	***2.286***	***4***	***6***	***110***	***2.949***

Hemolizli numune, kan hücresi elemanlarının seruma ya da plazmaya geçişidir. Klinik kimyasal analizde önemli bir hata kaynağı oluşturmaktadır. Lipemik numune, makroskobik bulanıklık bulunması halidir. Pıhtılı numune, alınan kan numunesinin pıhtılaşmış olduğunu ifade eder. Tekrarı uygundur, tekrar edilmesi gereken test

durumunu ifade eder. Uygun olmayan kap, kap uygunsuzluğundan gerçekleştirilemeyen testi ifade eder. Yetersiz numune, alınan numunenin teşhis için yetersiz olduğunu ifade eder. Uygunsuz veri miktarları öznitelik bazında Tablo 5.1'de paylaşılmıştır. Toplamda 2.949 öznitelik değeri veri setinden temizlenmiştir.

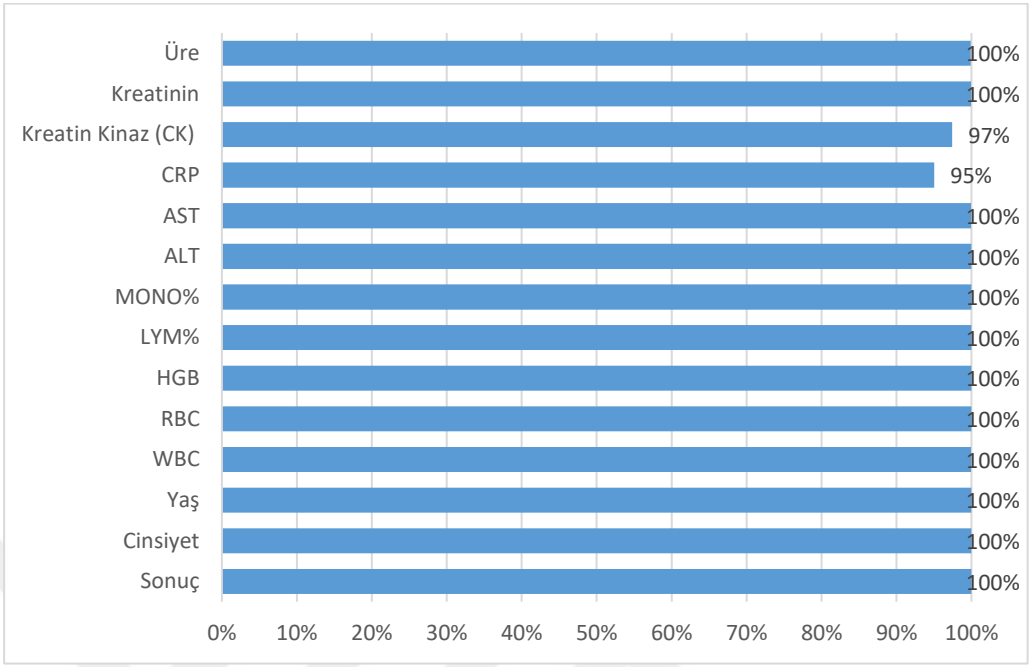
Bu işlem sonrasında veri setinde ayakta tedavi gören hastalar filtrelenmiştir. Bu aşamada 44x6427'lik bir veri matrisi elde edilmiştir. Belirtilen öznitelikler bazında gözlemlerin doluluk oranları %7 ile %93 arasında değişmektedir. Özniteliklerin doluluk oranı ise Şekil 5.1'de gösterilmiştir.



Şekil 5. 1 Öznitelik Doluluk Oranları

Veri setinde %80'den az doluluk oranına sahip olan gözlemler ve öznitelikler kaldırılmıştır. Oluşturulan yeni veri seti sonuç, cinsiyet, yaş, WBC, RBC, HGB, LYM%, MONO%, ALT, AST, CRP, Kreatin Kinaz, Kreatinin ve Üre

özniteliklerinden oluşmaktadır. Yeni veri seti 14x4.765'lik bir matristir. Yeni veri setindeki öznitelik doluluk oranları Şekil 5.2'de gösterilmiştir.



Şekil 5. 2 Öznitelik Doluluk Oranları

Bahsi geçen özniteliklerden COVID-19 test sonucu ve cinsiyet öznitelikleri kategorik değişken, kalan özniteliklerin tümü sürekli değişkendir.

Veri setinin bir kesiti şekil 5.3'te paylaşılmıştır.

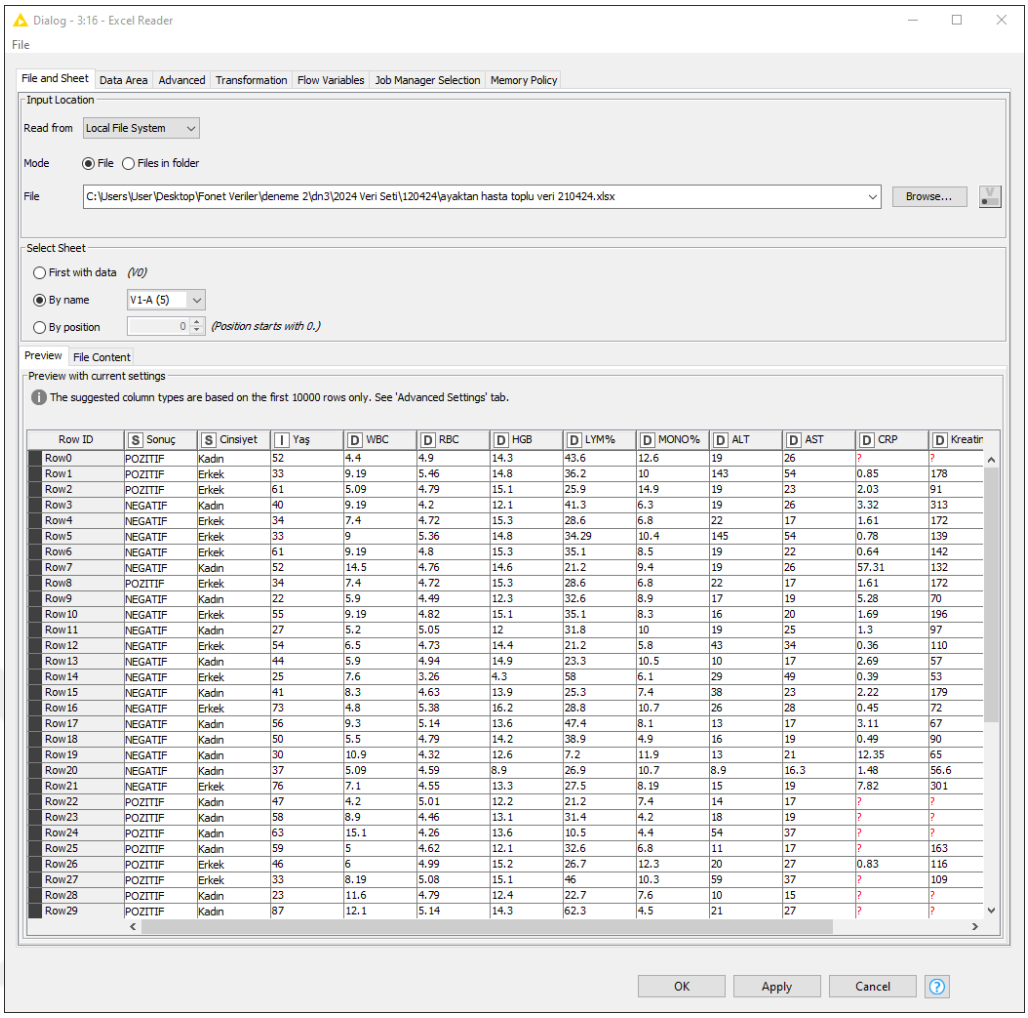
Sonuç	Cinsiyet	Yaş	WBC	RBC	HGB	LYM%	MONO%	ALT	AST	CRP	Kreatin Kinaz (CK)	Kreatinin	Üre
P	K	52	4,4	4,9	14,3	43,6	12,6	19	26			0,5	28
P	E	33	9,19	5,46	14,8	36,2	10	143	54	0,85	178	0,9	47
P	E	61	5,09	4,79	15,1	25,9	14,9	19	23	2,03	91	0,8	26
N	K	40	9,19	4,2	12,1	41,3	6,3	19	26	3,32	313	0,6	31
N	E	34	7,4	4,72	15,3	28,6	6,8	22	17	1,61	172	1,1	40
N	E	33	9	5,36	14,8	34,29	10,4	145	54	0,78	139	0,8	40
N	E	61	9,19	4,8	15,3	35,1	8,5	19	22	0,64	142	0,8	28
N	K	52	14,5	4,76	14,6	21,2	9,4	19	26	57,31	132	0,6	25
P	E	34	7,4	4,72	15,3	28,6	6,8	22	17	1,61	172	1,1	40
N	K	22	5,9	4,49	12,3	32,6	8,9	17	19	5,28	70	0,7	22
N	E	55	9,19	4,82	15,1	35,1	8,3	16	20	1,69	196	0,9	37
N	K	27	5,2	5,05	12	31,8	10	19	25	1,3	97	0,6	28
N	E	54	6,5	4,73	14,4	21,2	5,8	43	34	0,36	110	1	24
N	K	44	5,9	4,94	14,9	23,3	10,5	10	17	2,69	57	0,8	28
N	E	25	7,6	3,26	4,3	58	6,1	29	49	0,39	53	0,8	22
N	K	41	8,3	4,63	13,9	25,3	7,4	38	23	2,22	179	0,7	52
N	E	73	4,8	5,38	16,2	28,8	10,7	26	28	0,45	72	0,6	28
N	K	56	9,3	5,14	13,6	47,4	8,1	13	17	3,11	67	0,7	20
N	K	50	5,5	4,79	14,2	38,9	4,9	16	19	0,49	90	0,9	39
N	K	30	10,9	4,32	12,6	7,2	11,9	13	21	12,35	65	0,6	16

Şekil 5.3 Veri Seti

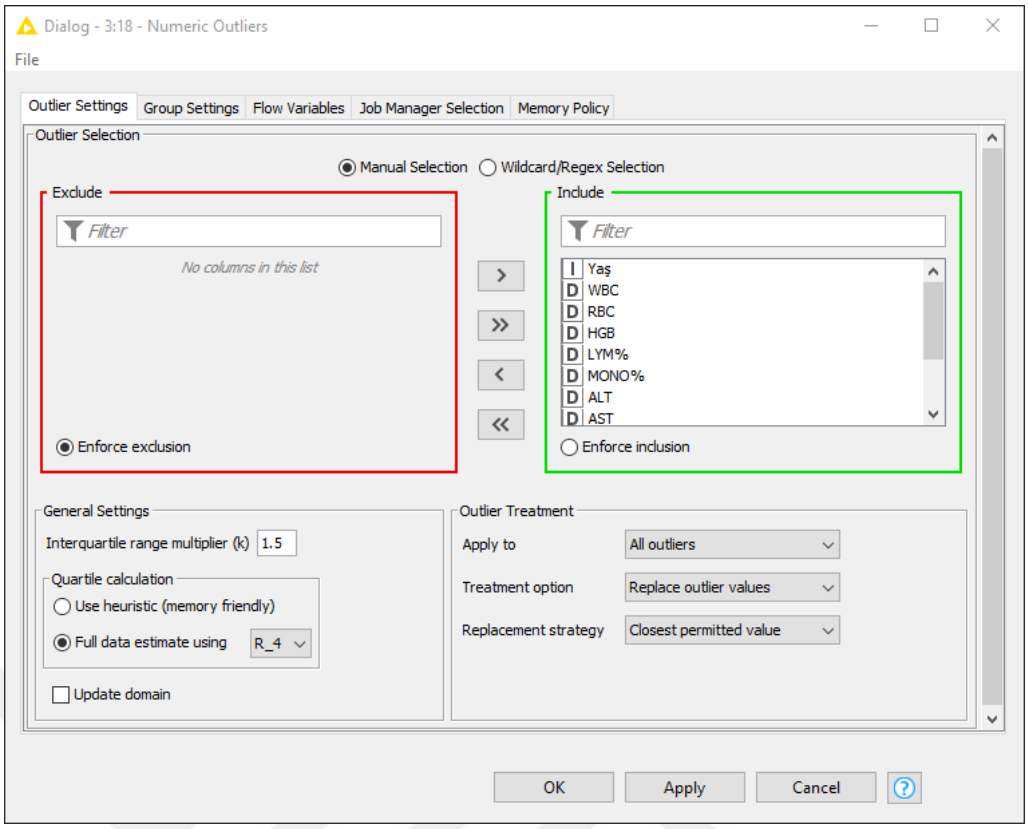
Oluşturulan veri seti KNIME (5.2.2 sürümü) programına aktarılmıştır. İş akışına ilk olarak excel okuyucu operatörü eklenmiştir. Operatörün düzenleme penceresindeki dosya (file) kısmından ilgili veri dosyası konumu belirtilerek programa yüklenmiştir. Sayfa seç (select sheet) bölümü isim ile (by name) seçeneği seçilerek ilgili excel

dosyasının kullanılmak istenen sayfası seçilmiştir. Ön izleme (preview) kısmında ilgili ayarlar yapıldıktan sonra dosyanın bir ön izlemesi görüntülenecektir. İlgili düzenlemeler Şekil 5.4'te gösterilmiştir.

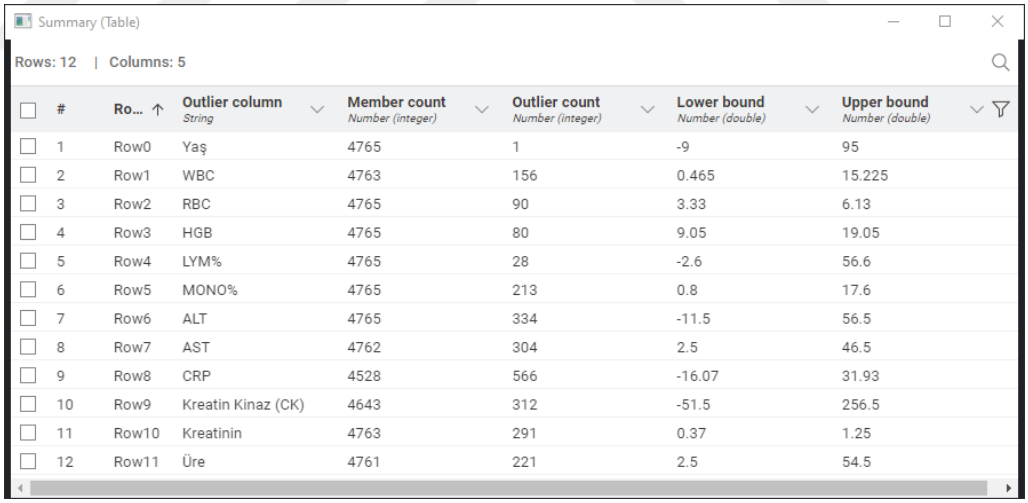
Dosya programa yüklendikten sonra iş akışına sayısal aykırı değerler (numeric outliers) operatörü eklenmiştir. Bu operatör sayesinde veri setindeki aykırı değerler saptanabilir ve tercih edilen işlemler yaptırılabilir. Düzenleme penceresinden aykırı değerler için işlem yaptırılmak istenen değişkenler aykırı değer seçimi (outlier selection) kısmındaki include bölümüne seçilir. Bu uygulama için sürekli olan değişkenlerin tamamı dahil edilmiştir. Aykırı değer müdahalesi (outlier treatment) bölümünden saptanan aykırı değerler için yapılacak işlemler belirtilir. KNIME yalnızca alt sınırın altındaki aykırı değerlere veya yalnızca üst sınırın üzerindeki aykırı değerlere işlem yapılabilmesine olanak sağlamaktadır. Müdahale opsiyonları arasında aykırı değerleri değiştirme, aykırı değer bulunan satırları kaldırma veya aykırı değer bulunmayan satırları kaldırmaya olanak sağlanmaktadır. Değiştirme stratejisi için saptanan aykırı değerleri izin verilen en yakın değer ile değiştirebilmekte veya eksik veri olarak işleyebilmektedir. Bu uygulama için saptanan tüm aykırı değerlerin izin verilen en yakın değer ile değiştirilmesi stratejisi uygulanmıştır. İlgili işlemler Şekil 5.5'te gösterilmektedir. Uygulanan işlemin özet çıktısı ise Şekil 5.6'da verilmektedir. Veri setinde saptanan aykırı değerler öznitelik bazında %12,5 ile %0,02 arasında değişmektedir. En fazla aykırı değer saptanan öznitelik CRP değeri iken en az aykırı değer saptanan öznitelik yaş değişkenidir. Tüm özniteliklerin aykırı değer oranları Tablo 5.2'de sunulmaktadır.



Şekil 5.4 Excel Okuyucu Operatörü



Şekil 5.5 Sayısal Aykırı Değerler Operatörü



#	Ro...	Outlier column (String)	Member count (Number (integer))	Outlier count (Number (integer))	Lower bound (Number (double))	Upper bound (Number (double))
1	Row0	Yaş	4765	1	-9	95
2	Row1	WBC	4763	156	0.465	15.225
3	Row2	RBC	4765	90	3.33	6.13
4	Row3	HGB	4765	80	9.05	19.05
5	Row4	LYM%	4765	28	-2.6	56.6
6	Row5	MONO%	4765	213	0.8	17.6
7	Row6	ALT	4765	334	-11.5	56.5
8	Row7	AST	4762	304	2.5	46.5
9	Row8	CRP	4528	566	-16.07	31.93
10	Row9	Kreatin Kinaz (CK)	4643	312	-51.5	256.5
11	Row10	Kreatinin	4763	291	0.37	1.25
12	Row11	Üre	4761	221	2.5	54.5

Şekil 5.6 Aykırı Değerler Özet Tablo

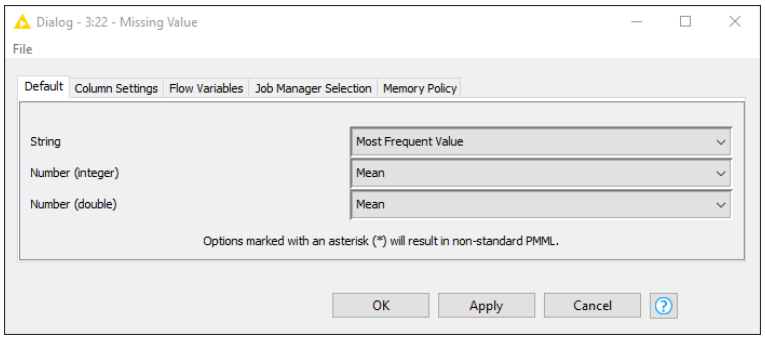
Tablo 5. 2 Öznitelik Bazında Aykırı Değer Oranları

Öznitelik	Aykırı Değer Oranı
Yaş	0,02%
WBC	3,28%
RBC	1,89%
HGB	1,68%
LYM%	0,59%
MONO%	4,47%
ALT	7,01%
AST	6,38%
CRP	12,50%
Kreatin Kinaz (CK)	6,72%
Kreatinin	6,11%
Üre	4,64%

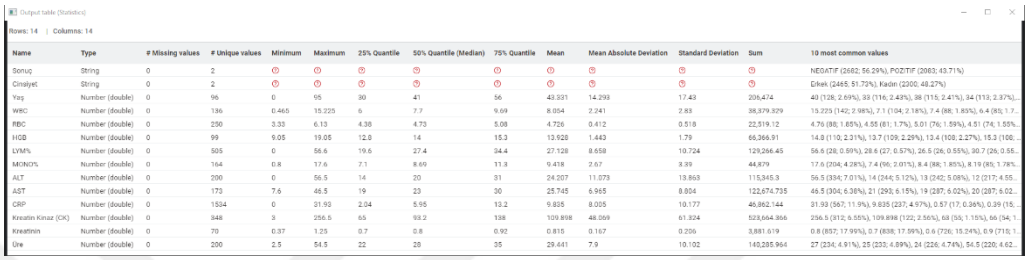
Aykırı değerlerin saptanıp izin verilen en yakın değer ile değiştirilmesi sonrası iş akışına eksik veri (missing value) operatörü eklenmiştir. Bu operatör eksik verilere müdahale edilmesini sağlar. Metin (string) veri tipindeki değişkenler için sabit bir değer, en sık tekrarlanan değer, bir sonraki dolu değer, bir önceki dolu değeri atama veya satırı kaldırma seçenekleri mevcuttur. Sayısal (integer ve double) veri tipindeki değişkenler için ortalama enterpolasyon, lineer enterpolasyon, sabit değer, minimum, maksimum, ortalama, medyan, en sık tekrarlanan, hareketli ortalama, bir sonraki dolu değer, bir önceki dolu değer, yuvarlanmış ortalama değerini atama veya satırı kaldırma seçenekleri mevcuttur (Şekil 5.7). Ayrıca ikinci sekmede bulunan satır ayarları (column settings) sekmesinden her bir satır/değişken bazında uygulanacak eksik veri stratejisi ayrı seçilebilmektedir. Bu uygulama için metin veri tipindeki değişkenler için en sık tekrarlanan değer, sayısal veri tipindeki değişkenler için ortalama değerin eksik veri yerine atanması stratejisi uygulanmıştır.

Uygulanan işlemler sonrası oluşturulan veri setinin minimum, maksimum, ortalama değerleri, ortalama mutlak sapma değeri ve standart sapma değerleri Şekil 5.8'de gösterilmektedir. Bu işlemler sonrası veri ön hazırlık aşaması tamamlanmıştır.

Sınıflandırma algoritmalarında kullanılacak eğitim ve test veri gruplarını oluşturabilmek için iş akışına x-bölümleyici (x-partitioner) operatörü eklenmiştir. Bu operatör verilerin eğitim ve test verisi olarak bölümlenmesini, bununla beraber grupların birleştirilerek hem eğitim hem de test aşamasında kullanılabilmesini sağlamaktadır.



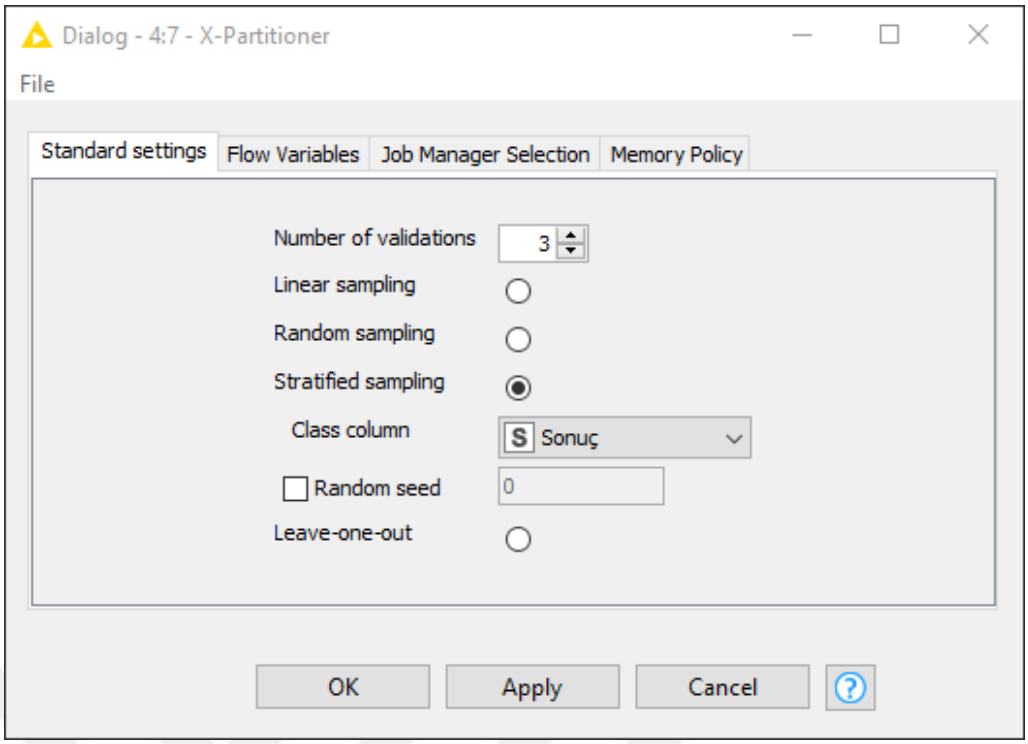
Şekil 5. 7 Eksik Veri Operatörü



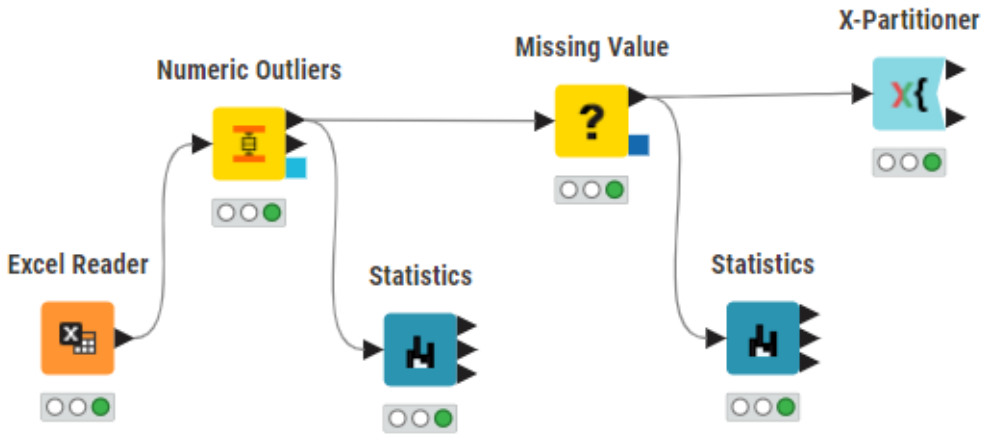
Şekil 5. 8 Oluşturulan Veri Setinin İstatistik Verileri

COVID-19 test sonuçlarının tahminlenebilmesi için kullanılan her bir değişken için 4.765 veri bulunmaktadır. Bu değerlerin %66' sının eğitim %34' ünün test aşamasında kullanılmak üzere bölünebilmesi için x-bölümleyici penceresindeki "number of validations" kısmına 3 değeri yazılmıştır. Aynı penceredeki "class column" kısmında yer alan "sonuç" etiketine göre hem eğitim hem de test setindeki yüzdeliklerin korunması için "stratified sampling" seçeneği işaretlenmiştir. Böylece test sonuçlarında %56,29'unda negatif test sonucuna sahip , %43,71'inde ise pozitif test sonucuna sahip veri setinin hem eğitim hem de test veri setindeki oranları sabit kalır (Şekil 5.9). X-bölümleyici operatörünün düzenlenmesi sonrası sınıflandırma algoritmalarının her biri sırasıyla modele eklenmiştir.

Oluşturulan iş akışı şekil 5.10'da gösterilmektedir. Sınıflandırma algoritmalarından karar ağacı, K-en yakın komşu, naive bayes, lojistik regresyon, rastgele orman, destek vektör makineleri kullanılarak oluşturulan tüm modellerde bu veri seti kullanılmıştır.



Şekil 5. 9 X-Bölümleyici Operatörü



Şekil 5. 10 Veri Ön Hazırlık İşlemleri İçin Oluşturulan İş Akışı

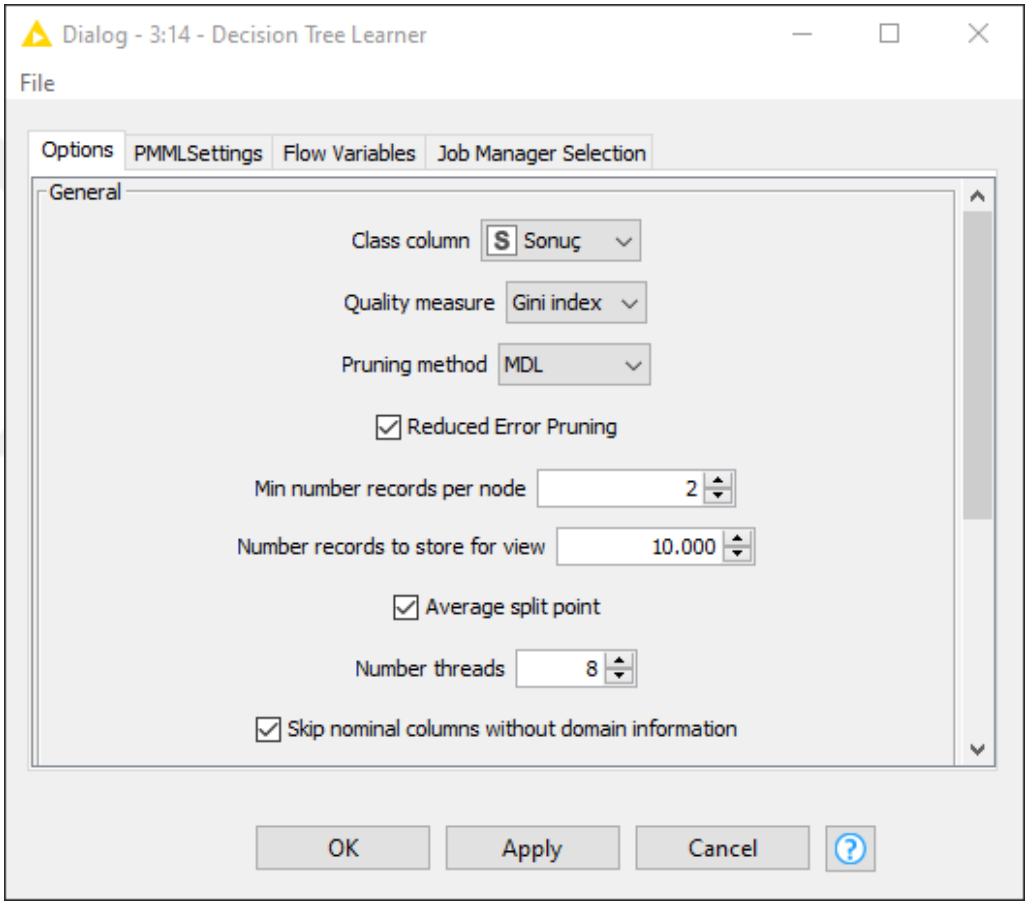
5.1.2 Karar Ağacı ile Modelleme

Bir önceki bölümde gösterilen iş akışına karar ağacı öğrenici (desicion tree learner) ve karar ağacı öngörücü (desicion tree predictor) operatörleri eklenmiştir.

Karar ağacı öğrenici (desicion tree learner) operatörü düzenleme penceresinde sınıf sütunu (class column) kısmında "sonuç" sütunu seçilmiştir. Kalite ölçütü (quality measure) kısmında "Gini indeksi seçilmiştir. Azaltılmış hata budaması (reduced error pruning) varsayılan seçenektir, aynı şekilde bırakılmıştır. Bu seçenek eğer modelin doğruluğu azalmıyorsa, ağacın yapraklarından başlayarak her bir düğümü en popüler

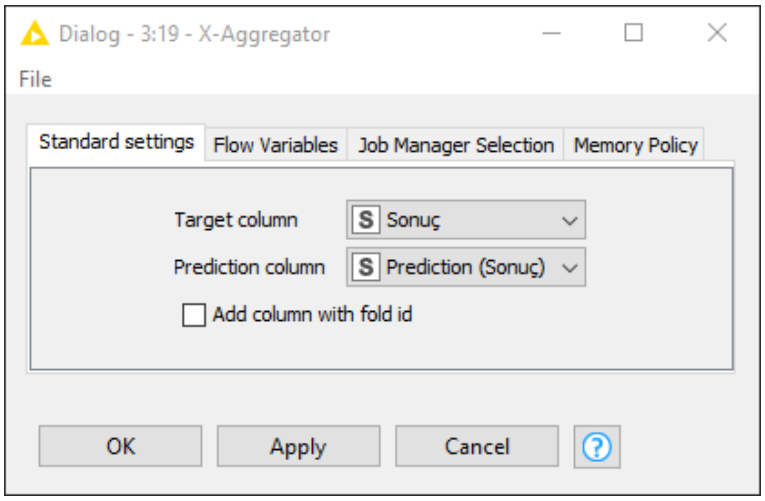
sınıfıyla değiştirir. Bu işlem gereksiz dalların kaldırılmasına veya ağacın daha basit bir versiyonunun oluşturulmasına yardımcı olur. Böylece model daha az karmaşık hale gelir ve aşırı uyum riski azalır.

KNIME karar ağacı öğrenici operatörü, budama yöntemi (pruning method) olarak MDL (Minimal Description Length) minimum açıklama uzunluğu metodunu sunar. Bu uygulama, karar ağacı budama yöntemi seçilmeden ve budama yöntemi olarak MDL seçilerek gerçekleştirilmiş olup MDL metodunun modelin doğruluğunu artırdığı tespit edildiğinden budama yöntemi olarak tercih edilmiştir. Bu operatörde yapılan işlemler Şekil 5.11'de gösterilmektedir.



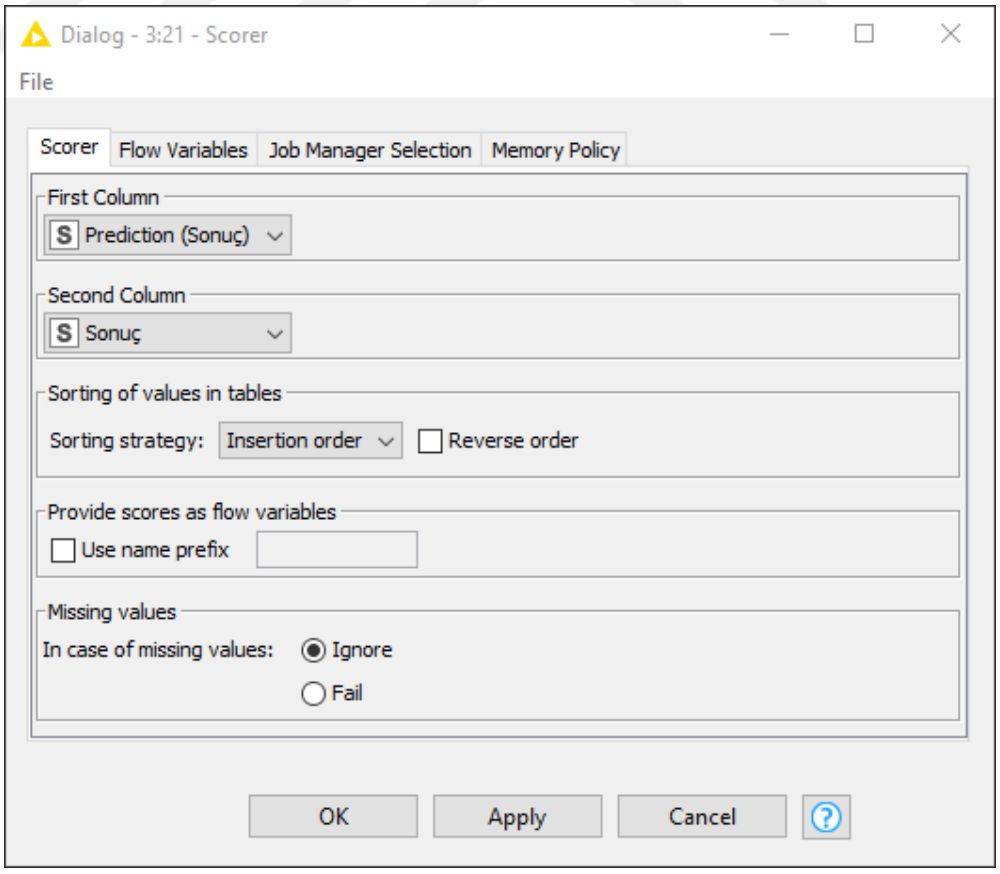
Şekil 5. 11 Karar Ağacı Öğrenici Operatörü

X-bölümleyici operatörü test verisi çıktısı ve karar ağacı öğrenici operatörü modeli karar ağacı öngörücü operatörüne bağlanmıştır. Önceki bölümde bölünen veri setinin tekrar birleştirilebilmesi için iş akışına x-toplayıcı (x-aggregator) eklenmiş, karar ağacı öğrenici operatörü x-toplayıcı operatörüne bağlanmıştır. X-toplayıcı operatörü penceresinde hedef sütun (target column) kısmında sonuç, tahmin sütun (prediction column) kısmına prediction (sonuç) sütunları seçilmiştir (Şekil 5.12).



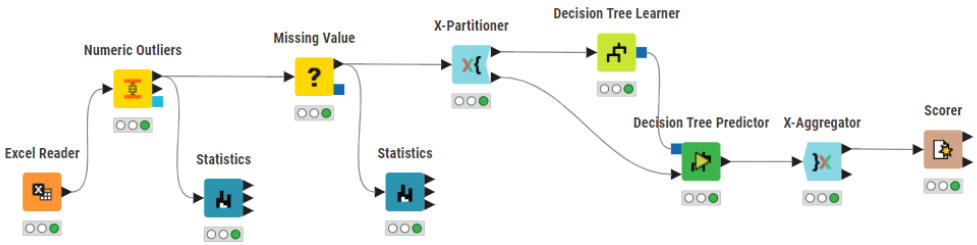
Şekil 5. 12 X-Toplayıcı Operatörü

Son olarak iş akışına skorcu (scorer) operatörü eklenmiştir. Skorcu operatörü kurulan modelin değerlendirilebilmesi için doğru pozitifler, yanlış pozitifler, doğru negatifler, yanlış negatifler, geri çağırma, duyarlılık, belirleyicilik, kesinlik, geri çağırma, F-skoru, doğruluk ölçütlerini rapor eder. Skorcu operatörü penceresinden ilk sütun (first column) kısmında prediction (sonuç), ikinci sütun (second column) kısmında sonuç sütunları seçilmiştir. İlgili işlem Şekil 5.13'te gösterilmektedir.



Şekil 5. 13 Skorcu Operatörü

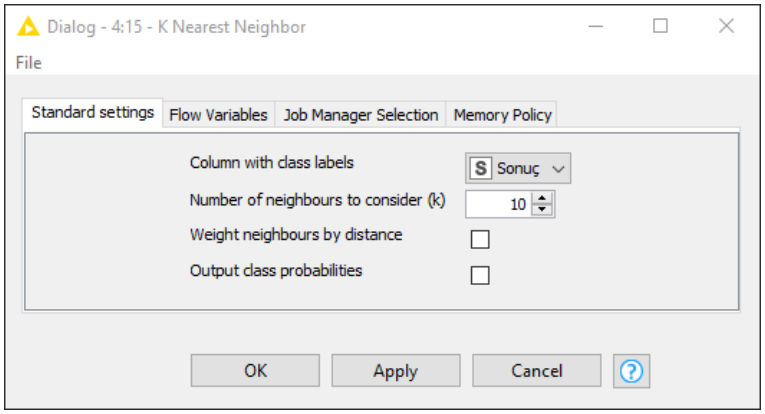
Karar ağacı modeli için oluşturulan iş akışı Şekil 5.14'te gösterilmektedir.



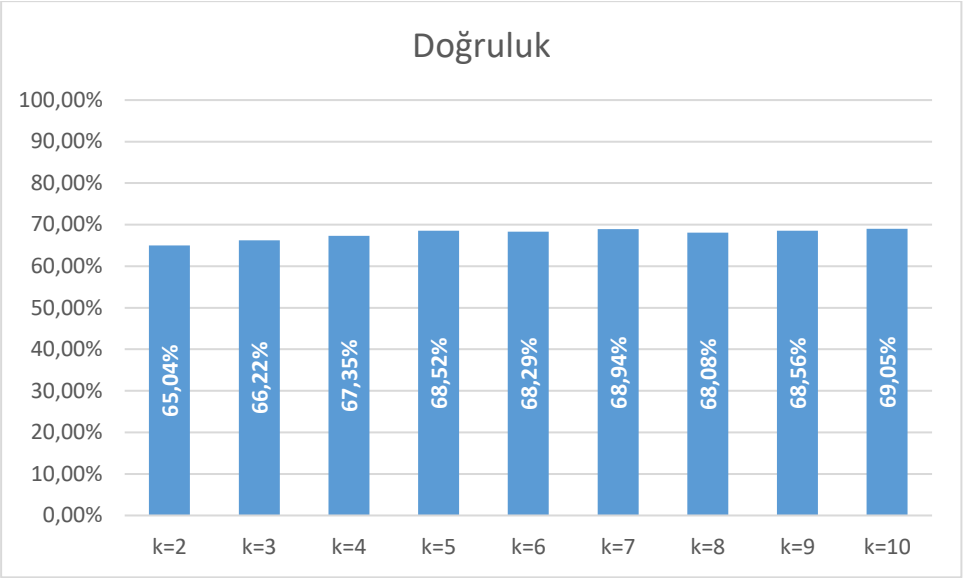
Şekil 5. 14 Karar Ağacı Modeli İş Akışı

5.1.3 K-En Yakın Komşu ile Modelleme

Şekil 5.10'daki iş akışına K-en yakın komşu (k-nearest neighbor) operatörü eklenmiştir. K-en yakın komşu düzenleme penceresinde sınıf etiketi sütunu (column with class labels) kısmında "sonuç" sütunu seçilmiştir. Dikkate alınacak komşu sayısı (number of neighbours to consider (k)) kısmı için literatürde en sık kullanılan değerler olan 2 ile 10 arasında denemeler yapılmıştır. Denemeler sonucunda en iyi çıkan doğruluk değeri %69,05 ile 10 elde edilmiş olup k-en yakın komşu değeri 10 olarak kabul edilmiştir. Bu pencerede yapılan işlemler Şekil 5.15'te gösterilmektedir. Komşu sayısı için 2'den 10'a kadar yapılan denemelerin sonuçları Şekil 5.16'te gösterilmektedir.

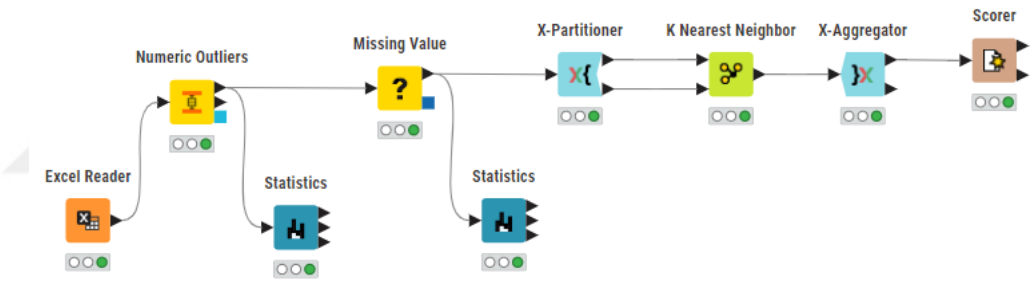


Şekil 5. 15 K-En Yakın Komşu Operatörü



Şekil 5. 16 Komşu Sayısı İçin Doğruluk Değerleri

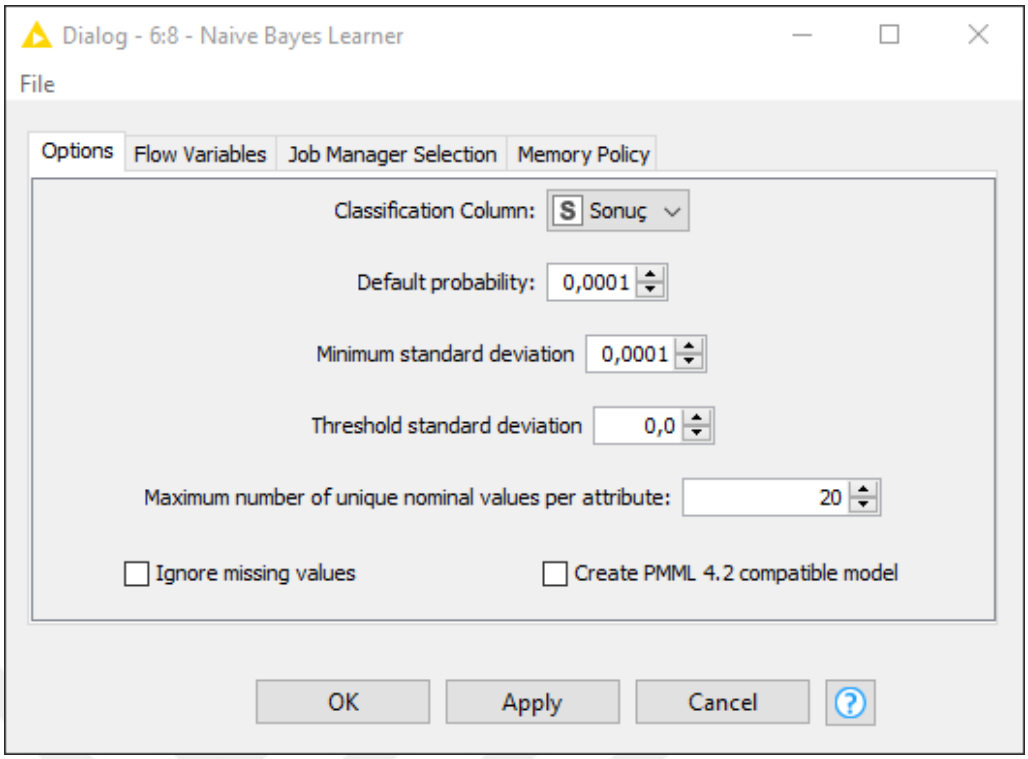
İş akışına sırayla x-toplayıcı ve skorcu operatörler eklenerek K-en yakın komşu için kurulan model tamamlanmıştır. Model için iş akışı Şekil 5.17'de gösterilmektedir.



Şekil 5. 17 K-En Yakın Komşu Modeli İş Akışı

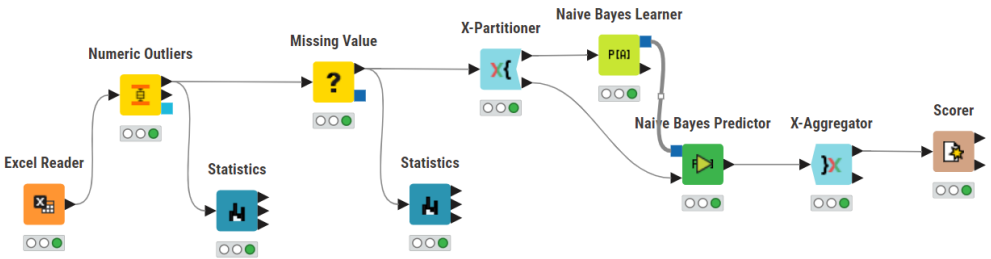
5.1.4 Naive Bayes ile Modelleme

Şekil 5.10'daki iş akışına Naive Bayes öğrenici (Naive Bayes learner) ve Naive Bayes öngörücü (Naive Bayes predictor) operatörleri eklenmiştir. Naive bayes öğrenici operatörü düzenleme penceresinde sınıflandırma sütunu (classification column) kısmında "sonuç" sütunu seçilmiştir (Şekil 5.18).



Şekil 5. 18 Naive Bayes Öğrenici Operatörü

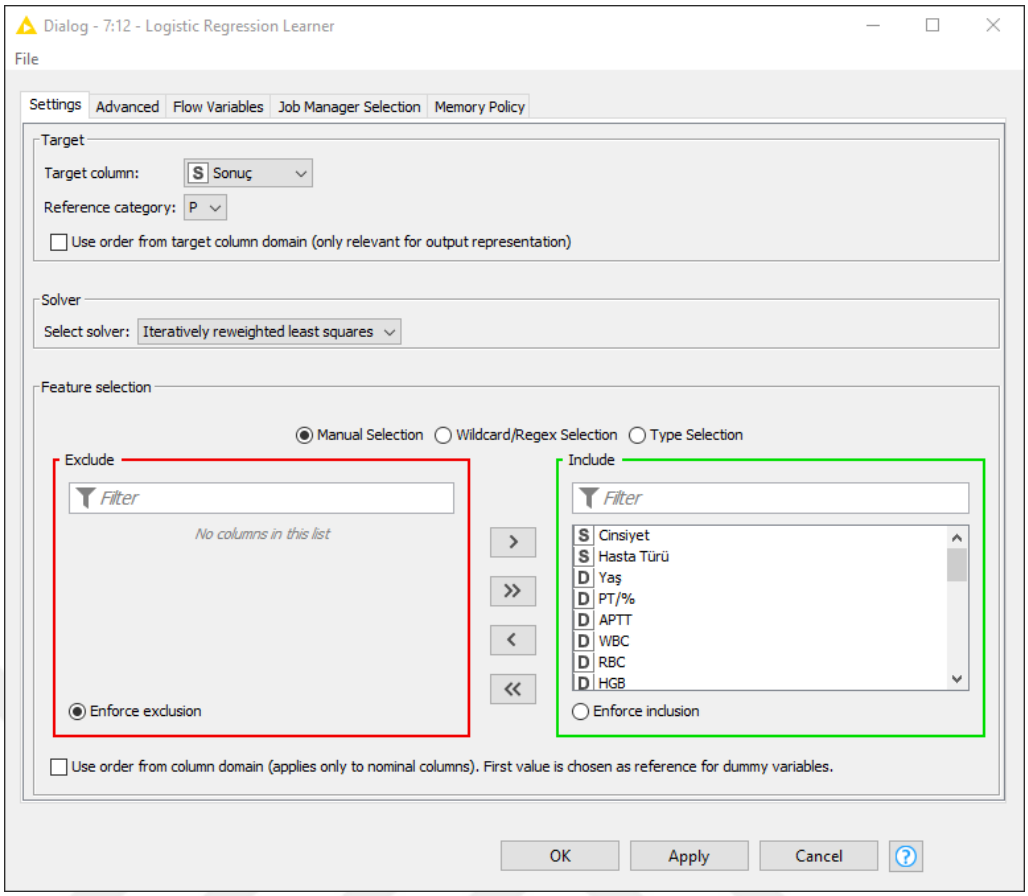
İş akışına sırayla x-toplayıcı ve skorcu operatörler eklenerek Naive Bayes için kurulan model tamamlanmıştır. Model için iş akışı Şekil 5.19'da gösterilmektedir.



Şekil 5. 19 Naive Bayes Modeli İş Akışı

5.1.5 Lojistik Regresyon ile Modelleme

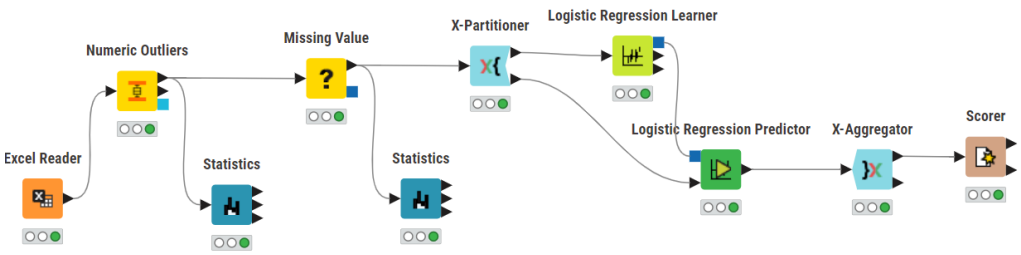
Şekil 5.10'daki iş akışına lojistik regresyon öğrenici (logistic regression learner) ve lojistik regresyon öngörücü (logistic regression predictor) operatörleri eklenmiştir.



Şekil 5. 20 Lojistik Regresyon Öğrenici Operatörü

Lojistik regresyon öğrenici operatörü düzenleme penceresinde hedef sütun (target column) kısmında sonuç sütunu seçilmiştir. Aynı pencerede referans kategori (referance category) kısmına COVID-19 test sonucunun pozitif olduğunu gösteren "P" değeri atanmıştır. Çözücü seçimi (select solver) kısmında iteratif olarak yeniden ağırlıklandırılmış en küçük kareler (iteratively reweighted least squares) seçeneği seçilmiştir (Şekil 5.20).

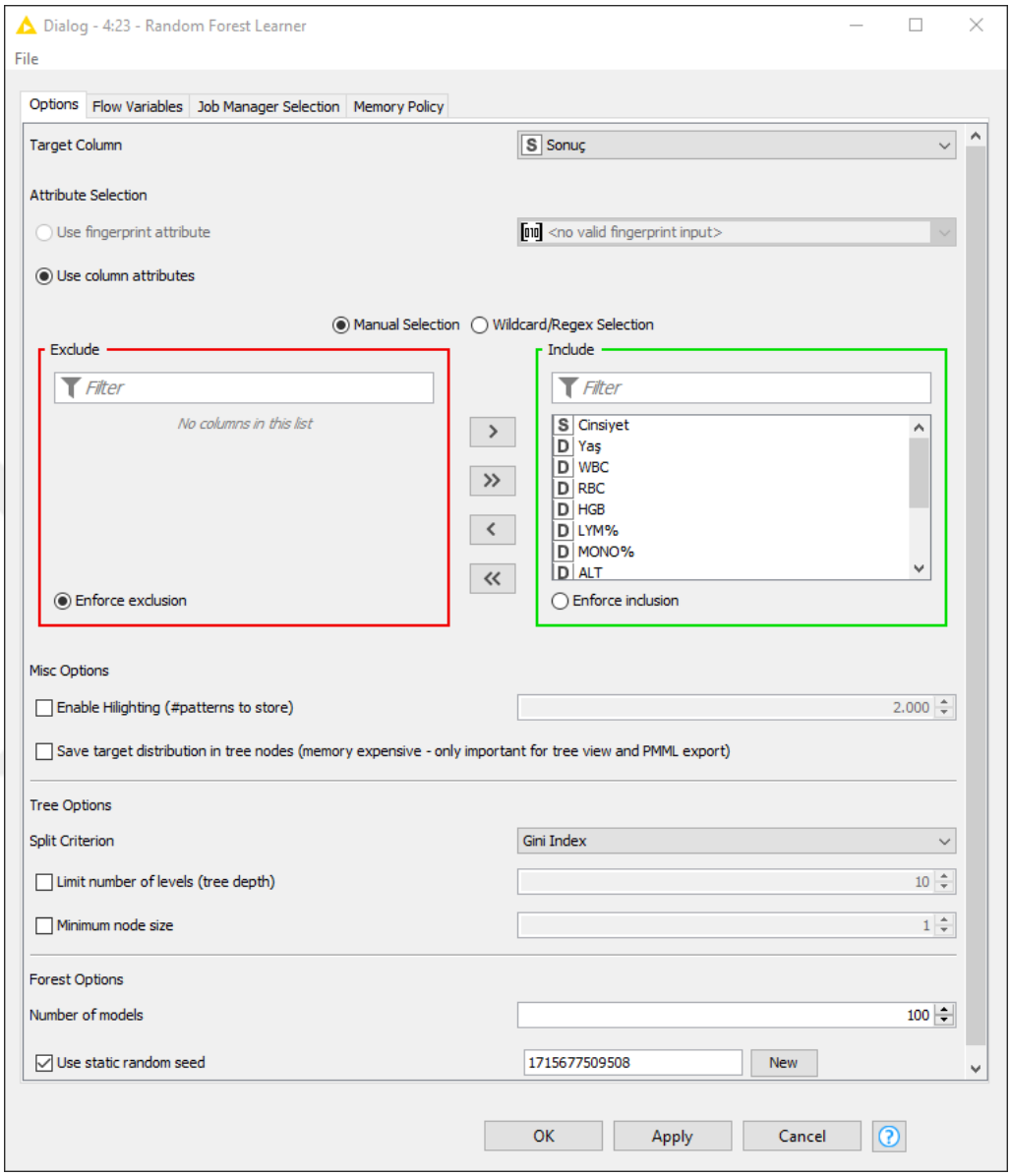
İş akışına sırayla x-toplayıcı ve skorcu operatörler eklenerek lojistik regresyon için kurulan model tamamlanmıştır. Model için iş akışı Şekil 5.21'de gösterilmektedir.



Şekil 5. 21 Lojistik Regresyon Modeli İş Akışı

5.1.6 Rastgele Orman ile Modelleme

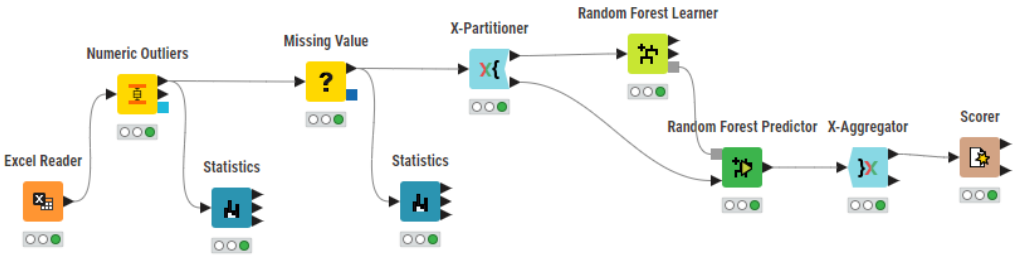
Şekil 5.10'daki iş akışına rastgele orman öğrenici (random forest learner) ve rastgele orman öngörücü (random forest predictor) operatörleri eklenmiştir.



Şekil 5. 22 Rastgele Orman Öğrenici Operatörü

Rastgele orman öğrenici operatörü düzenleme penceresinde hedef sütun (target column) kısmında sonuç sütunu seçilmiştir (Şekil 3.22).

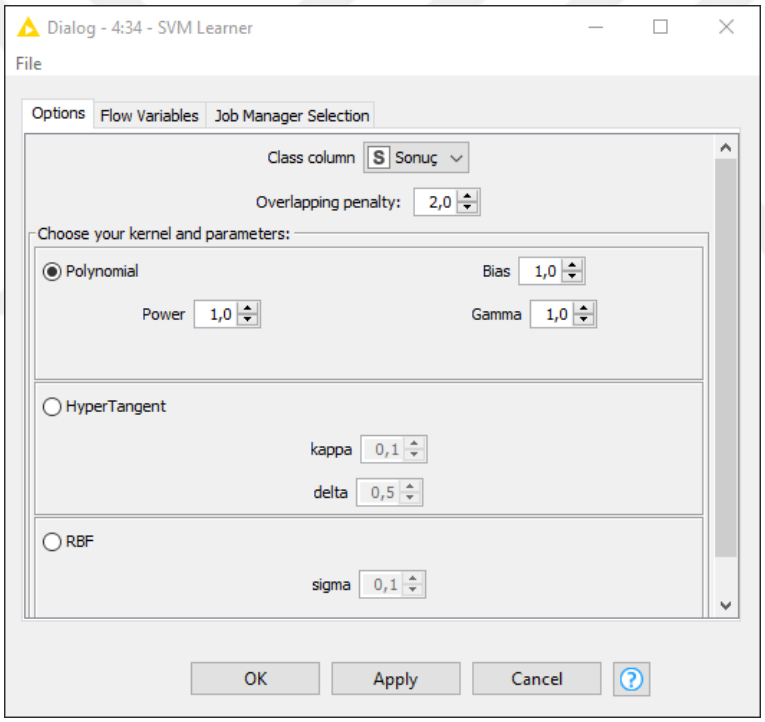
İş akışına sırayla x-toplayıcı ve skorcu operatörler eklenerek rastgele orman için kurulan model tamamlanmıştır. Model için iş akışı Şekil 3.23'te gösterilmektedir.



Şekil 5. 23 Rastgele Orman Modeli İş Akışı

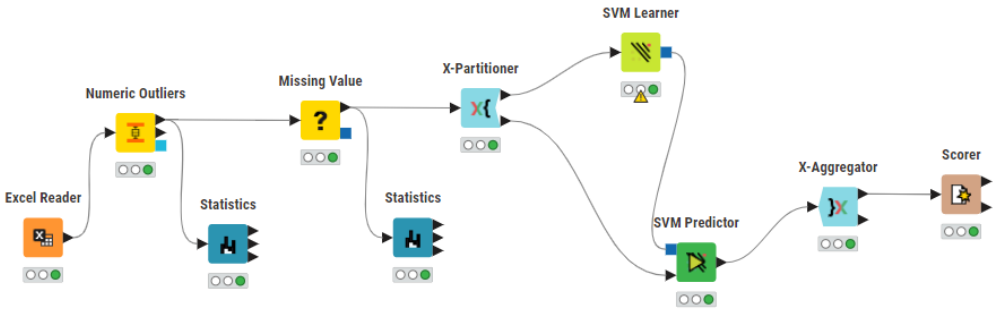
5.1.7 Destek Vektör Makineleri ile Modelleme

Şekil 5.10'daki iş akışına destek vektör makineleri öğrenici (SVM learner) ve destek vektör makineleri öngörücü (SVM predictor) operatörleri eklenmiştir. Destek vektör makineleri öğrenici operatörü düzenleme penceresinde sınıf sütunu (class column) kısmında "sonuç" sütunu seçilmiştir (Şekil 3.24).



Şekil 5.24 Destek Vektör Makineleri Öğrenici Operatörü

İş akışına sırayla x-toplayıcı ve skorcu operatörler eklenerek destek vektör makineleri için kurulan model tamamlanmıştır. Model için iş akışı Şekil 3.25'te gösterilmektedir.



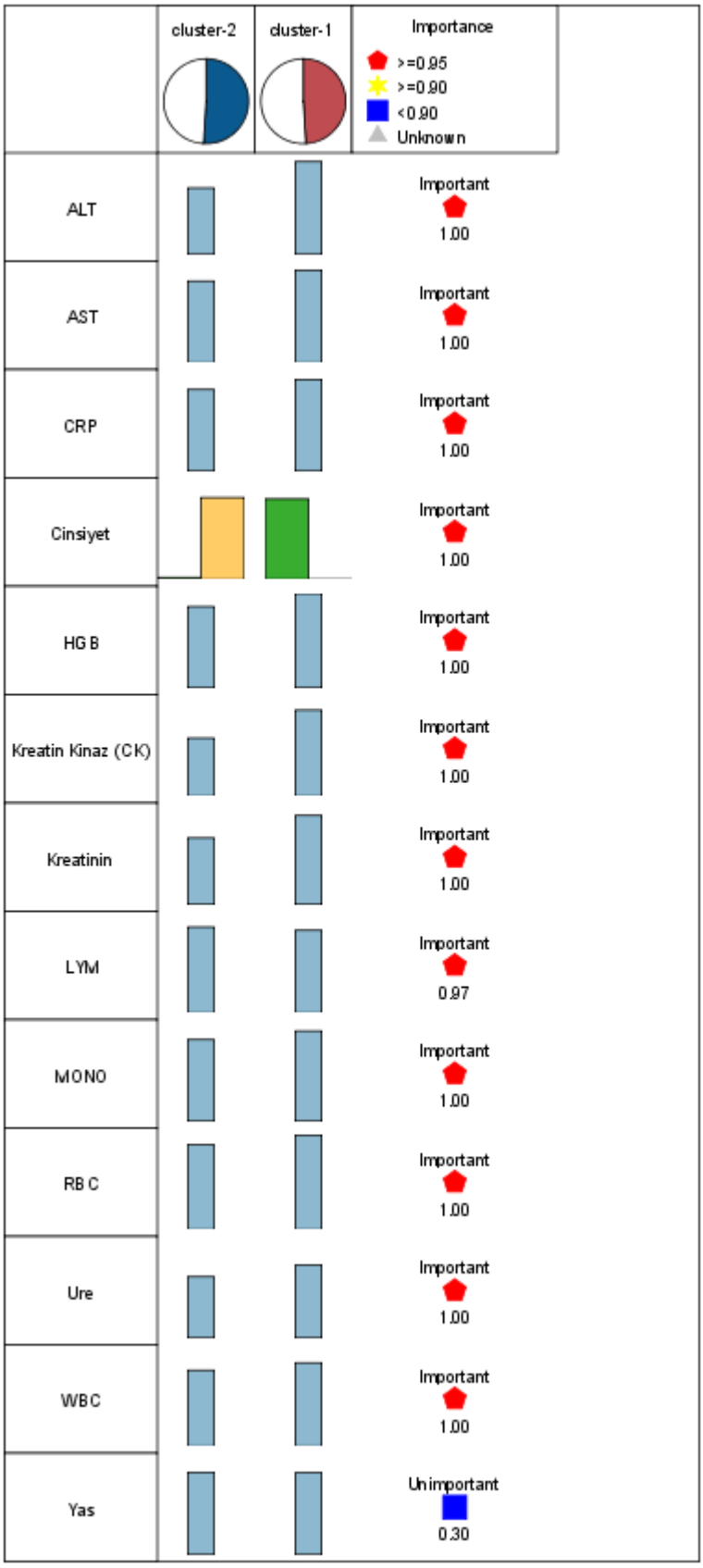
Şekil 5.25 Destek Vektör Makineleri Modeli İş Akışı

5.2 İki Aşamalı Kümeleme Analizi ile Uygulama

Bu bölümde COVID-19 testi yaptırmış olan hastaların laboratuvar sonuçlarından hasta profillerinin incelenmesi, ilginç örüntülerin keşfedilmesi amaçlanmaktadır. Bu doğrultuda, şekil 5.10'daki iş akışına excel yazıcı (excel writer) operatörü eklenerek veri ön hazırlık işlemleri tamamlanmış olan veri seti çıktısı alınmıştır. Yöntem olarak hem kategorik hem de sürekli verilerle çalışabilmesi, büyük veri setleri için uygun olması ve küme sayısını otomatik belirleyebilmesi sebebiyle iki aşamalı kümeleme analizi seçilmiş olup uygulama için SPSS Clementine 12.0 programı kullanılmıştır. Veri seti COVID-19 test sonucu negatif ve pozitif çıkan hastalar için ayrılarak iki ayrı kümeleme analizi gerçekleştirilmiştir.

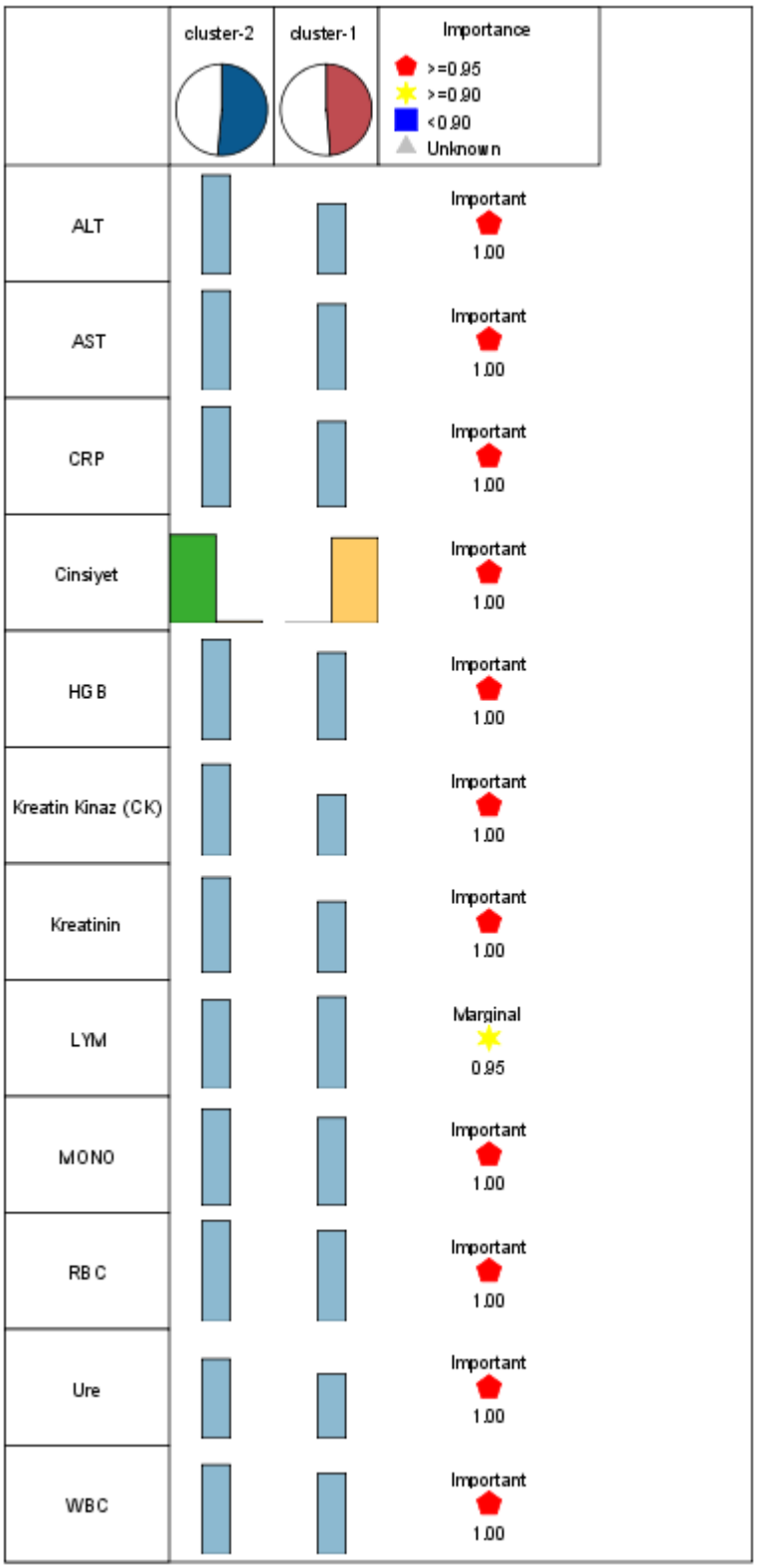
5.2.1 Pozitif Hastalar İçin İki Aşamalı Kümeleme Analizi Uygulaması

Test sonucu pozitif olan hastalar için verilere iki aşamalı kümeleme analizi uygulanmıştır. İlk analiz sonucunda yaş değişkeninin kümeler üzerinde önemli etkisi olmadığı tespit edilmiştir. Analiz sonuçları Şekil 5.26'da gösterilmektedir. Modelden yaş değişkeni çıkarılarak analiz tekrarlanmıştır.



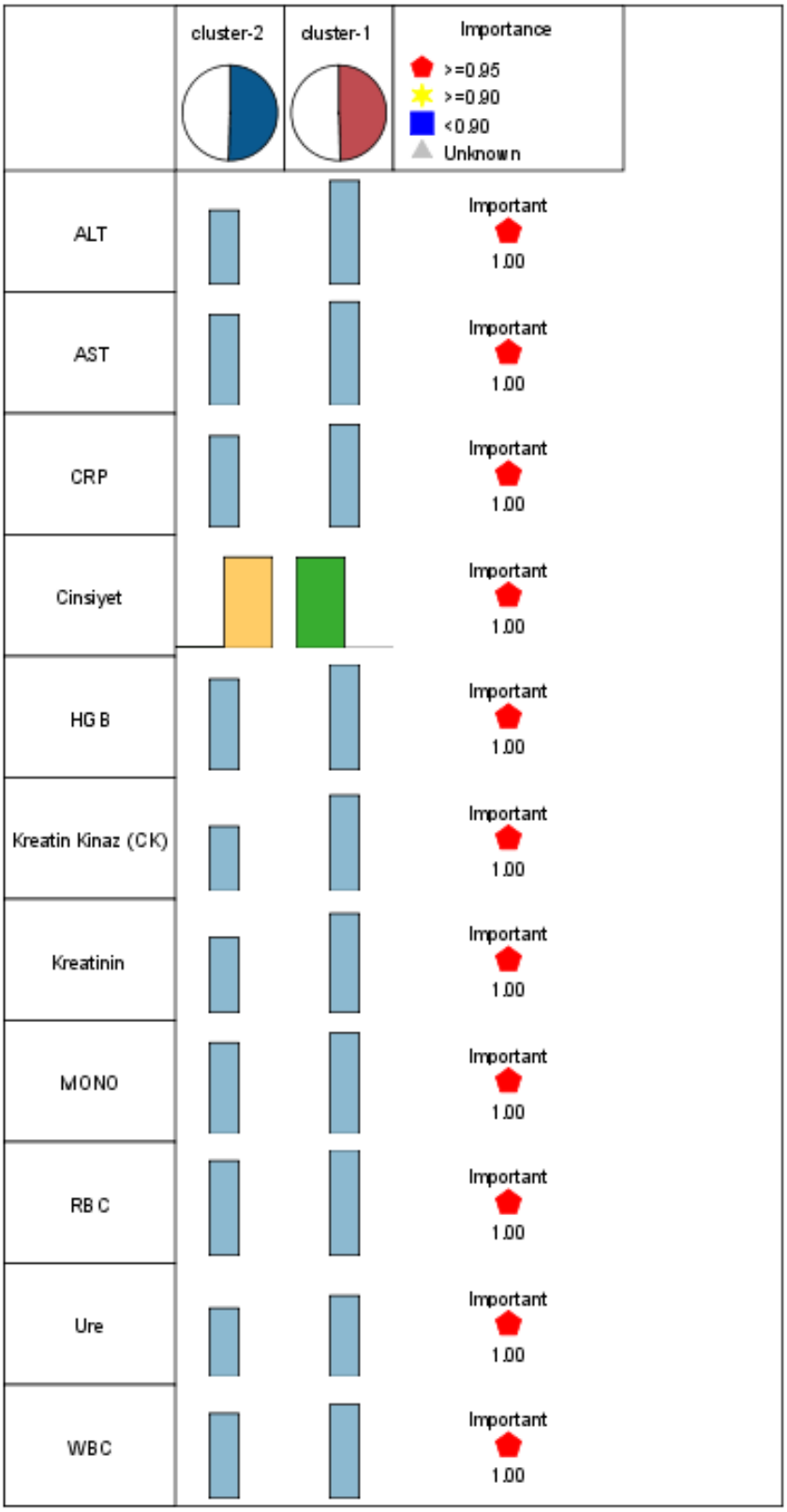
Şekil 5.26 Pozitif Hastalar İçin İki Aşamalı Kümeleme Analizi

İkinci analizde elde edilen sonuçlara göre LYM% değişkeninin önem düzeyinin "Marginal" olduğu tespit edilmiştir. Analiz sonuçları Şekil 5.27'de gösterilmektedir. Analiz, LYM% değişkeni çıkartıldıktan sonra tekrarlanmıştır.



Şekil 5.27 Yaş Değişkeni Çıkarılarak Gerçekleştirilen Analiz

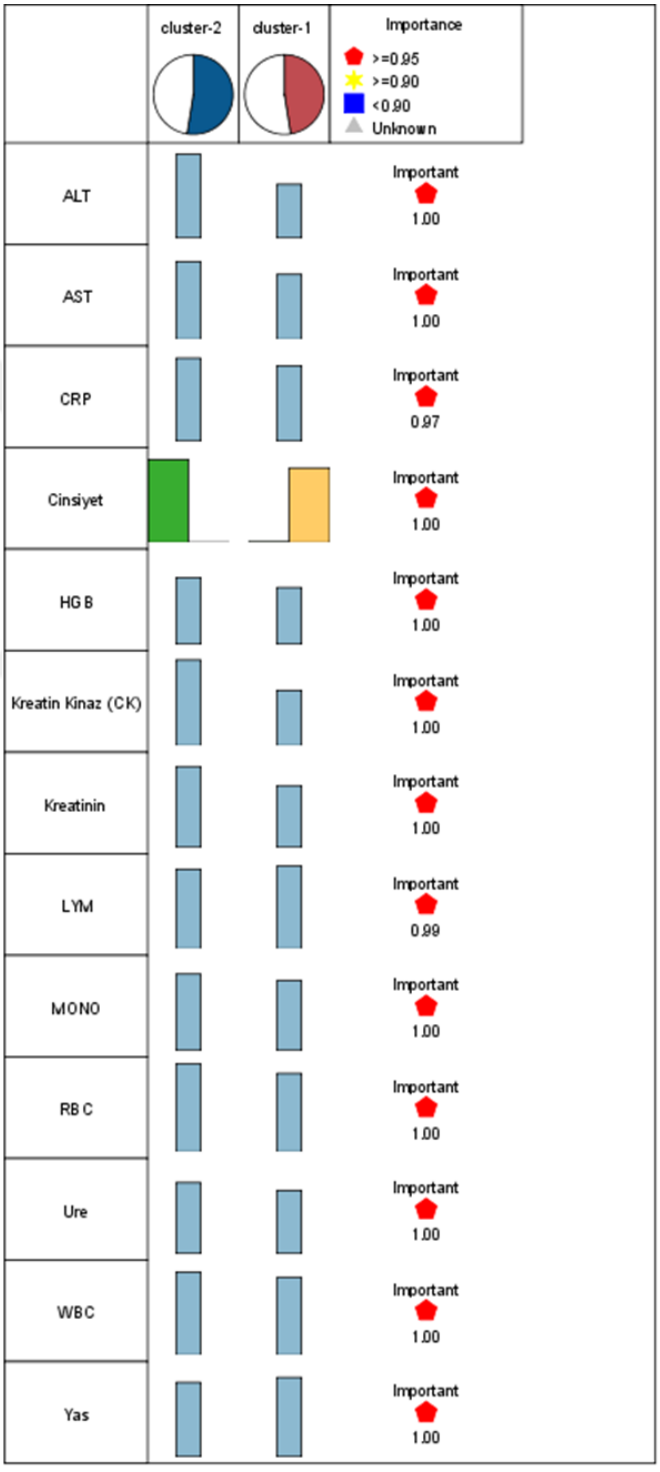
Üçüncü analizde elde edilen sonuçlara göre kalan on bir değişkenin tamamının kümeler üzerinde önemli etkisi olduğu tespit edilmiştir. Analiz sonucunda iki küme elde edilmiştir. Analiz sonuçları şekil 5.28'de gösterilmektedir.



Şekil 3.28 LYM Değişkeni Çıkarılarak Gerçekleştirilen Analiz

5.2.2 Negatif Hastalar İçin İki Aşamalı Kümeleme Analizi Uygulaması

Test sonucu negatif olan hastalar için verilere iki aşamalı kümeleme analizi uygulanmıştır. Analizde elde edilen sonuçlara göre on üç değişkenin tamamının kümeler üzerinde önemli etkisi olduğu tespit edilmiştir. Analiz sonucunda iki küme elde edilmiştir (Şekil 3.29).



Şekil 5.29 Negatif Hastalar İçin İki Aşamalı Kümeleme Analizi

6. BULGULAR

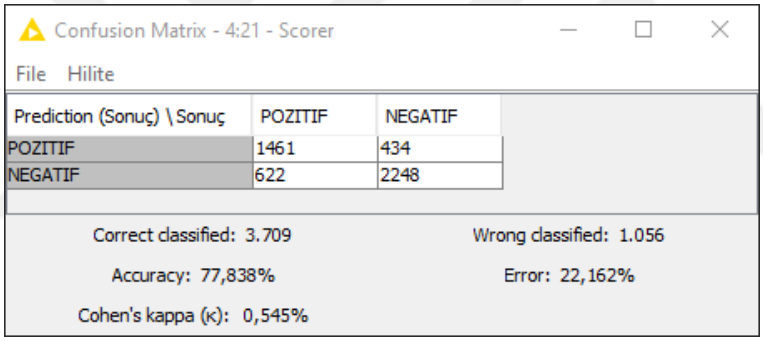
Çalışmanın sınıflandırma uygulaması kısmında laboratuvar test sonuçlarından hastanın COVID-19 test sonucu tahminlenmeye çalışılmış, karar ağacı, Naive Bayes, K-en yakın komşu, lojistik regresyon, rastgele orman ve destek vektör makineleri modelleri oluşturulmuştur.

Çalışmanın kümeleme analizi kısmında ise pozitif ve negatif hastalar için iki ayrı veri setine iki aşamalı kümeleme analizi uygulanmıştır.

6.1 Sınıflandırma Yöntemlerine İlişkin Bulgular

6.1.1 Karar Ağacı ile Elde Edilen Bulgular

Oluşturulan karar ağacı modelinin çalıştırılması sonrası hastanın COVID-19 test sonucunu tahminleyen modelin başarısını gösteren karmaşıklık matrisi Şekil 6.1'de verilmektedir.



Şekil 6. 1 Karar Ağacı Karmaşıklık Matrisi

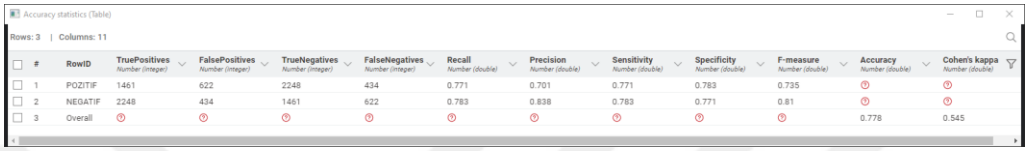
Elde edilen sonuçlar doğrultusunda COVID-19 test sonuçlarına göre gerçekte 2083 pozitif hastadan 1461 hastanın pozitif, 622 hastanın ise negatif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu pozitif olan hastaların doğru tahmin edilme oranı %70,1 olarak saptanmıştır. COVID-19 test sonuçlarına göre gerçekte 2682 negatif sonuçlu hastadan 2248 hastanın negatif, 434 hastanın ise pozitif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu negatif olan hastaların doğru tahmin edilme oranı %83,8 olarak saptanmıştır. Karmaşıklık matrisi köşegeninde yer alan doğru tahmin sayılarına dair oran ise %77,8 olarak saptanmıştır (Tablo 6.1).

Tablo 4.1 Karar Ağacı Modeline İlişkin Bulgular

	Gerçek

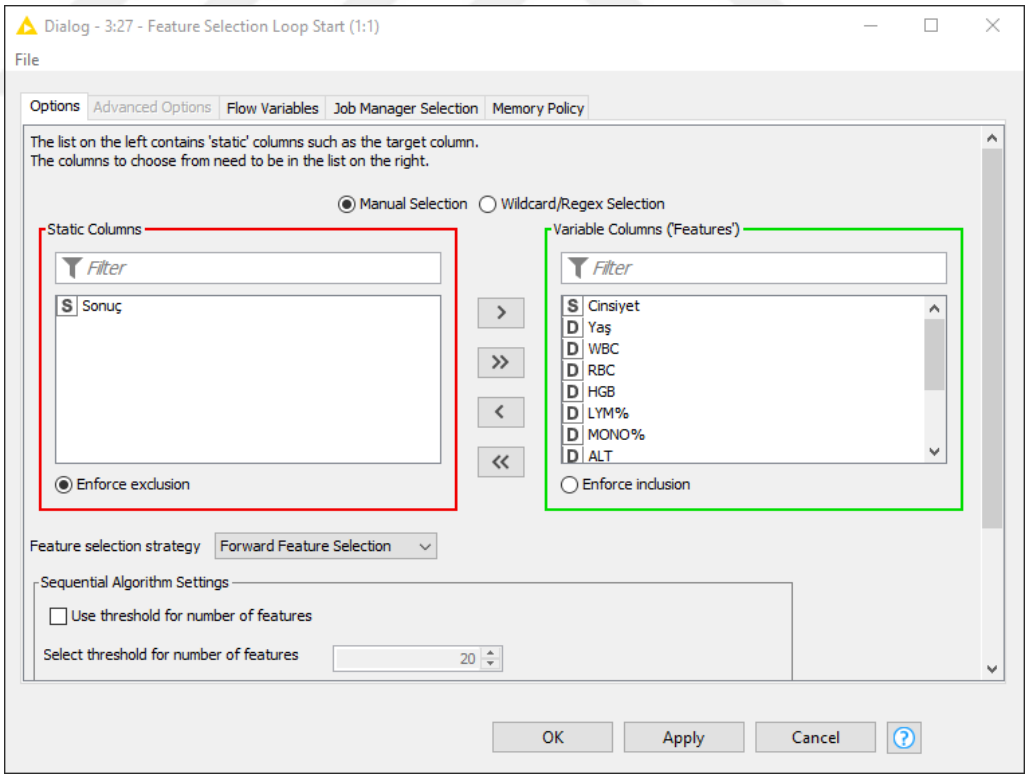
		Test Sonucu Pozitif	Test Sonucu Negatif	Toplam	Teşhis (%)
Tahmin	Test Sonucu Pozitif	1461	434	1895	70,1%
	Test Sonucu Negatif	622	2248	2870	83,8%
	Toplam	2083	2682	4765	77,8%

Karar ağacı modeline ilişkin skorcu operatörü çıktısı şekil 6.2'te gösterilmektedir.



Şekil 6.2 Karar Ağacı Modeli Skorcu Çıktısı

Model doğruluğunu arttırmak için iş akışına özellik seçimi döngüsü başlangıç operatörü (feature selection loop start (1:1)) ve özellik seçimi döngüsü bitiş operatörü eklenmiştir (Şekil 6.3).

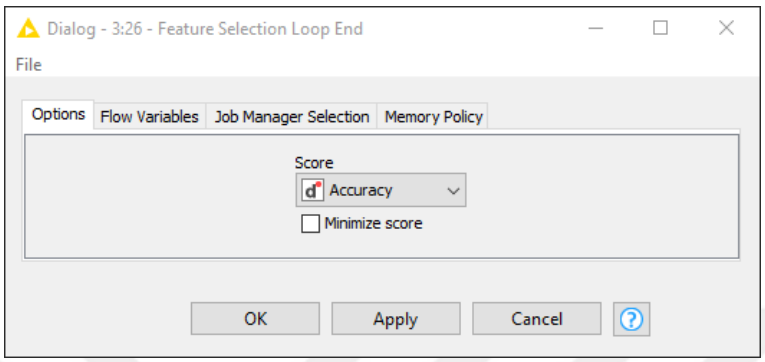


Şekil 6.3 Özellik Seçimi Döngüsü Başlangıç Operatörü

Özellik seçimi döngüsü başlangıç operatörü penceresinde sınıf etiketi olan sonuç sütunu statik sütunlar (static columns) kısmına alınmıştır. Özellik seçimi stratejisi için

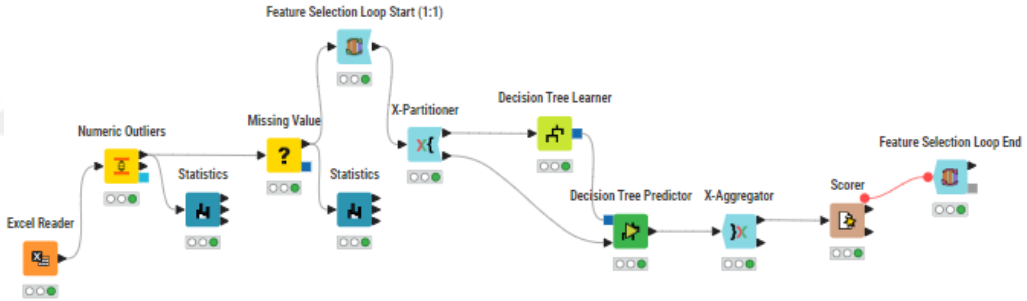
ileri özellik seçimi (forward feature selection) stratejisi tercih edilmiştir. İleri özellik seçimi stratejisi yinelemeli bir yaklaşımdır. Başlangıçta hiçbir değişken seçilmez. Her yinelemede, modele en fazla katkıyı sağlayan öznitelik, öznitelik setine eklenir.

Özellik seçimi döngüsü bitiş operatörü penceresinde skor olarak doğruluk (accuracy) seçilmiştir (Şekil 6.4).



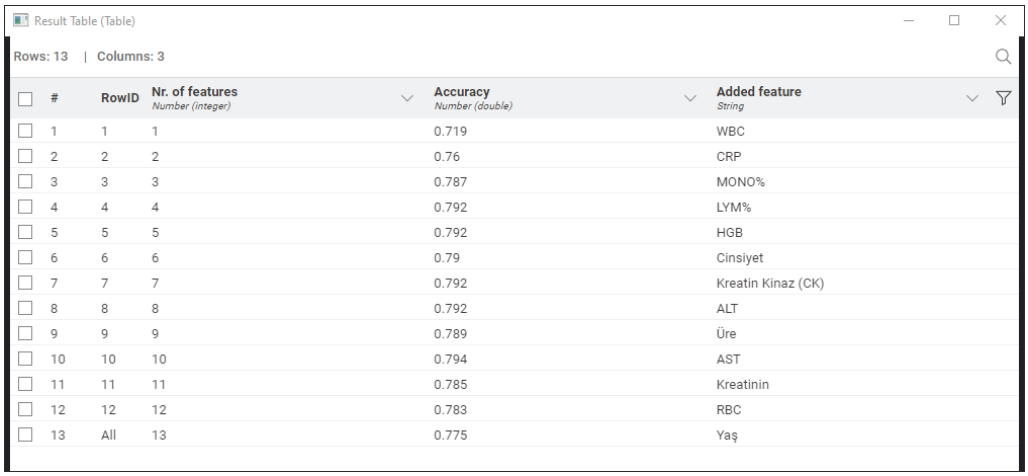
Şekil 6.4 Özellik Seçimi Döngüsü Bitiş Operatörü

İş akışının son hali şekil 4.5'te gösterilmektedir.



Şekil 6.5 Karar Ağacı - Özellik Seçimi İçin İş Akışı

Döngü, WBC değişkeni eklenerek %71,9 ile başlamıştır. En yüksek doğruluk oranı 10. yinelemede modelde WBC, CRP, MONO%, LYM%, HGB, cinsiyet, Kreatin Kinaz (CK), ALT, Üre, AST öznitelikleri mevcutken elde edilmiştir. Özellik seçimi analizi sonucunda modelden Kreatinin, RBC ve yaş değişkenlerinin çıkarılmasıyla %79,4 doğruluk değerinin elde edilebileceği görülmüştür (Şekil 4.6 ve Şekil 4.7).

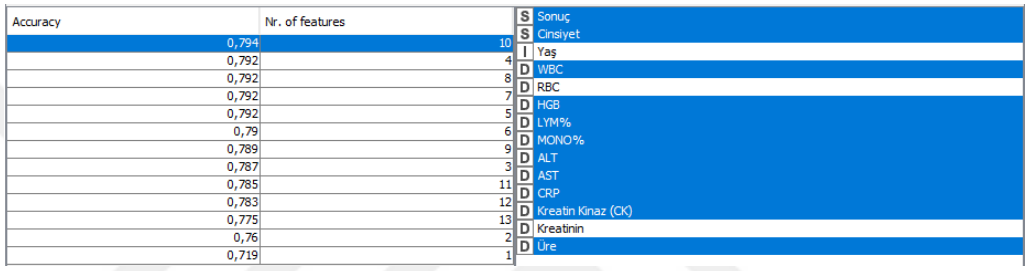


Result Table (Table)

Rows: 13 | Columns: 3

#	RowID	Nr. of features (Number (integer))	Accuracy (Number (double))	Added feature (String)
1	1	1	0.719	WBC
2	2	2	0.76	CRP
3	3	3	0.787	MONO%
4	4	4	0.792	LYM%
5	5	5	0.792	HGB
6	6	6	0.79	Cinsiyet
7	7	7	0.792	Kreatin Kinaz (CK)
8	8	8	0.792	ALT
9	9	9	0.789	Üre
10	10	10	0.794	AST
11	11	11	0.785	Kreatinin
12	12	12	0.783	RBC
13	All	13	0.775	Yaş

Şekil 6.6 Karar Ağacı Özellik Seçimi Analizi Sonucu



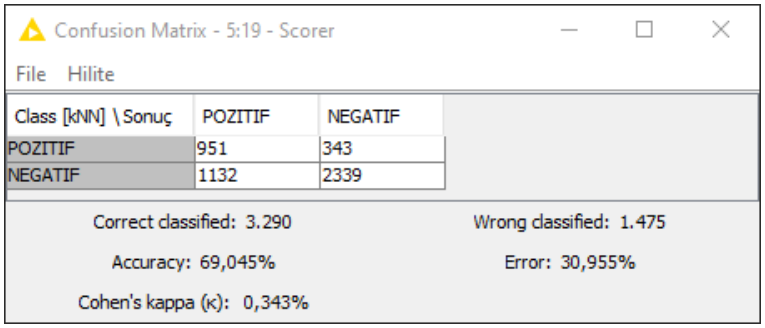
Accuracy	Nr. of features
0,794	10
0,792	4
0,792	8
0,792	7
0,792	5
0,79	6
0,789	9
0,787	3
0,785	11
0,783	12
0,775	13
0,76	2
0,719	1

S Sonuç
S Cinsiyet
I Yaş
D WBC
D RBC
D HGB
D LYM%
D MONO%
D ALT
D AST
D CRP
D Kreatin Kinaz (CK)
D Kreatinin
D Üre

Şekil 6.7 Karar Ağacı Özellik Seçimi Analizi - Maksimum Doğruluk Değerini Sağlayan Kombinasyon

6.1.2 K-En Yakın Komşu ile Modelleme

Oluşturulan K-en yakın komşu modelinin çalıştırılması sonrası hastanın COVID-19 test sonucunu tahminleyen modelin başarısını gösteren karmaşıklık matrisi Şekil 6.8'de verilmektedir.



Confusion Matrix - 5:19 - Scorer

File Hilite

Class [kNN] \ Sonuç	POZITIF	NEGATIF
POZITIF	951	343
NEGATIF	1132	2339

Correct classified: 3.290 Wrong classified: 1.475

Accuracy: 69,045% Error: 30,955%

Cohen's kappa (κ): 0,343%

Şekil 6. 8 K-En Yakın Komşu Karmaşıklık Matrisi

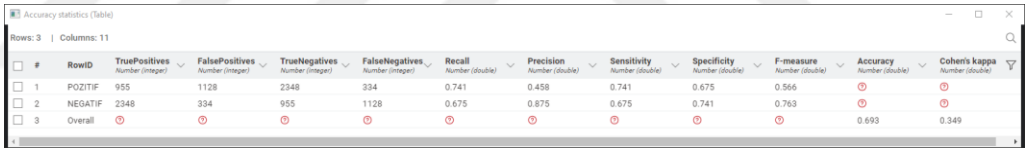
Elde edilen sonuçlar doğrultusunda COVID-19 test sonuçlarına göre gerçekte 2083 pozitif hastadan 951 hastanın pozitif, 1132 hastanın ise negatif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu pozitif olan hastaların doğru tahmin edilme oranı %45,7 olarak saptanmıştır. COVID-19 test sonuçlarına göre gerçekte 2682

negatif sonuçlu hastadan 2339 hastanın negatif, 343 hastanın ise pozitif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu negatif olan hastaların doğru tahmin edilme oranı %87,2 olarak saptanmıştır. Karmaşıklık matrisi köşegeninde yer alan doğru tahmin sayılarına dair oran ise %69,0 olarak saptanmıştır. Modelin teşhis oranları Tablo 6.2'te sunulmaktadır.

Tablo 6. 2 K-En Yakın Komşu Modeline İlişkin Bulgular

		Gerçek			
		Test Sonucu Pozitif	Test Sonucu Negatif	Toplam	Teşhis (%)
Tahmin	Test Sonucu Pozitif	951	343	1294	45,7%
	Test Sonucu Negatif	1132	2339	3471	87,2%
	Toplam	2083	2682	4765	69,0%

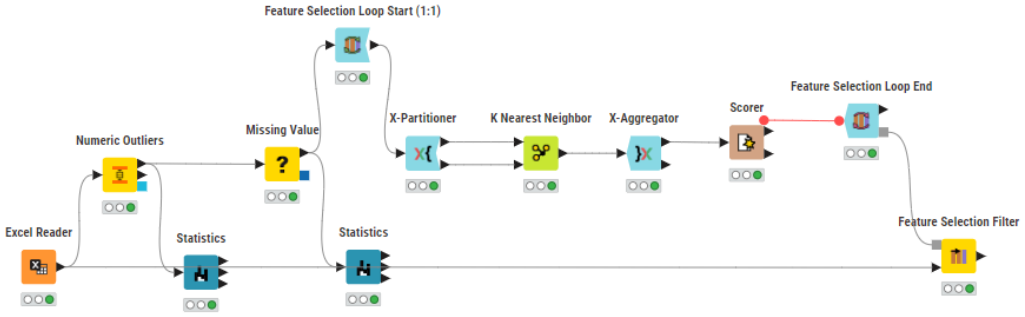
K-en yakın komşu modeline ilişkin skorcu operatörü çıktısı Şekil 6.9'da gösterilmektedir.



Şekil 6. 9 K-En Yakın Komşu Modeli Skorcu Çıktısı

Model doğruluğunu arttırmak için iş akışına özellik seçimi döngüsü başlangıç operatörü (feature selection loop start (1:1)) ve özellik seçimi döngüsü bitiş operatörü eklenmiştir. Özellik seçimi döngüsü başlangıç operatörü penceresinde sınıf etiketi olan sonuç sütunu statik sütunlar (static columns) kısmına alınmıştır. Özellik seçimi stratejisi için ileri özellik seçimi (forward feature selection) stratejisi tercih edilmiştir. Özellik seçimi döngüsü bitiş operatörü penceresinde skor olarak doğruluk (accuracy) seçilmiştir.

İş akışının son hali Şekil 6.10'da gösterilmektedir.



Şekil 6.10 K-En Yakın Komşu - Özellik Seçimi İçin İş Akışı

Döngü, WBC değişkeni eklenerek %71,4 ile başlamıştır. En yüksek doğruluk oranı 4. yinelemede modelde WBC, CRP, MONO%, Kreatinin öznitelikleri mevcutken elde edilmiştir. Özellik seçimi analizi sonucunda modelden cinsiyet, yaş, RBC, HGB, LYM%, ALT, AST, Kreatin Kinaz (CK) ve Üre değişkenlerinin çıkarılmasıyla %77,5 doğruluk değerinin elde edilebileceği görülmüştür (Şekil 6.11 ve Şekil 6.12).

Result Table (Table)

Rows: 13 | Columns: 3

#	RowID	Nr. of features *Number (integer)*	Accuracy *Number (double)*	Added feature *String*
1	1	1	0.714	WBC
2	2	2	0.746	CRP
3	3	3	0.774	MONO%
4	4	4	0.775	Kreatinin
5	5	5	0.774	HGB
6	6	6	0.772	RBC
7	7	7	0.772	Cinsiyet
8	8	8	0.77	LYM%
9	9	9	0.758	Yaş
10	10	10	0.747	AST
11	11	11	0.737	Üre
12	12	12	0.721	ALT
13	All	13	0.685	Kreatin Kinaz (CK)

Şekil 6.11 K-En Yakın Komşu Özellik Seçimi Analizi Sonucu

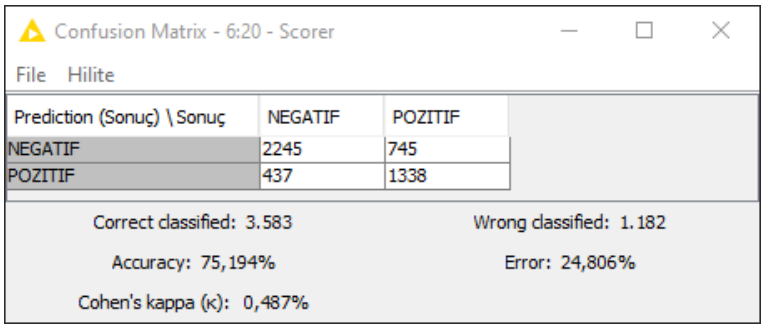
Accuracy	Nr. of features
0,775	4
0,774	5
0,774	3
0,772	7
0,772	6
0,77	8
0,758	9
0,747	10
0,746	2
0,737	11
0,721	12
0,714	1
0,685	13

	Sonuç
I	Cinsiyet
I	Yaş
D	WBC
D	RBC
D	HGB
D	LYM%
D	MONO%
D	ALT
D	AST
D	CRP
D	Kreatin Kinaz (CK)
D	Kreatinin
D	Üre

Şekil 6.12 K-En Yakın Komşu Özellik Seçimi Analizi - Maksimum Doğruluk Değerini Sağlayan Kombinasyon

6.1.3 Naive Bayes ile Modelleme

Oluşturulan Naive Bayes modelinin çalıştırılması sonrası hastanın COVID-19 test sonucunu tahminleyen modelin başarısını gösteren karmaşıklık matrisi Şekil 6.13'te verilmektedir.



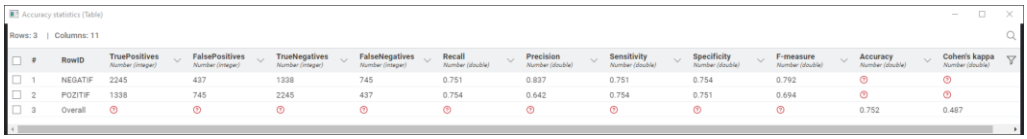
Şekil 6. 13 Naive Bayes Karmaşıklık Matrisi

Elde edilen sonuçlar doğrultusunda COVID-19 test sonuçlarına göre gerçekte 2083 pozitif hastadan 1338 hastanın pozitif, 745 hastanın ise negatif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu pozitif olan hastaların doğru tahmin edilme oranı %64,2 olarak saptanmıştır. COVID-19 test sonuçlarına göre gerçekte 2682 negatif sonuçlu hastadan 2245 hastanın negatif, 437 hastanın ise pozitif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu negatif olan hastaların doğru tahmin edilme oranı %83,7 olarak saptanmıştır. Karmaşıklık matrisi köşegeninde yer alan doğru tahmin sayılarına dair oran ise %75,2 olarak saptanmıştır. Modelin teşhis oranları Tablo 6.3'te gösterilmektedir.

Tablo 6. 3 Naive Bayes Modeline İlişkin Bulgular

		Gerçek			
		Test Sonucu Pozitif	Test Sonucu Negatif	Toplam	Teşhis (%)
Tahmin	Test Sonucu Pozitif	1338	437	1775	64,2%
	Test Sonucu Negatif	745	2245	2990	83,7%
	Toplam	2083	2682	4765	75,2%

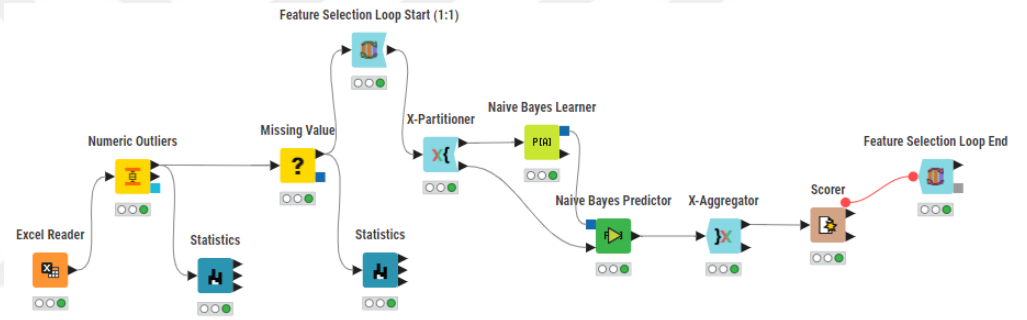
Naive Bayes modeline ilişkin skorcu operatörü çıktısı Şekil 6.14'te gösterilmektedir.



Şekil 6. 14 Naive Bayes Modeli Skorcu Çıktısı

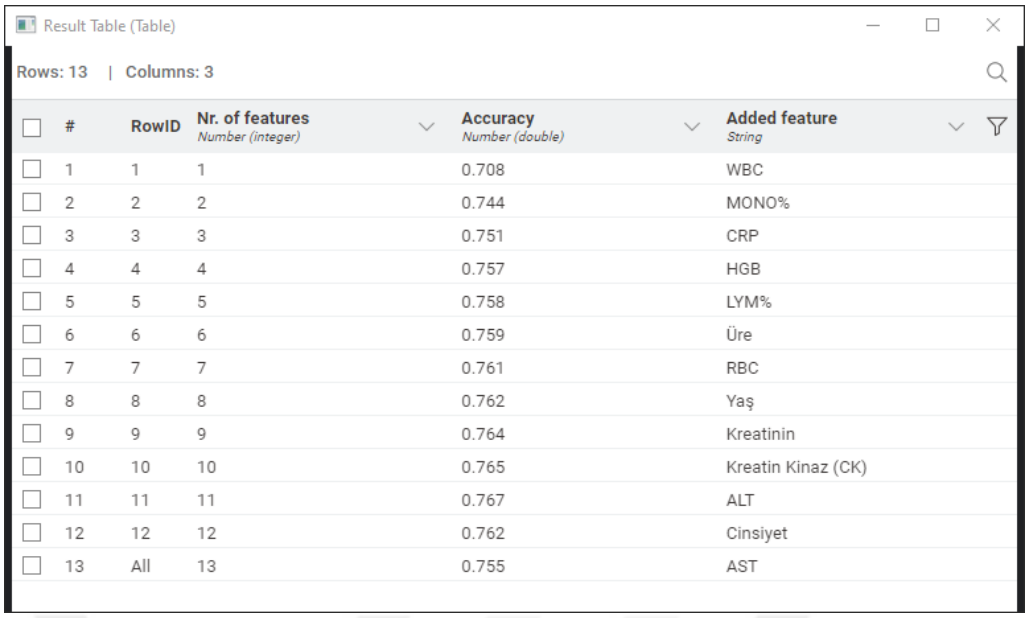
Model doğruluğunu arttırmak için iş akışına özellik seçimi döngüsü başlangıç operatörü (feature selection loop start (1:1)) ve özellik seçimi döngüsü bitiş operatörü eklenmiştir. Özellik seçimi döngüsü başlangıç operatörü penceresinde sınıf etiketi olan sonuç sütunu statik sütunlar (static columns) kısmına alınmıştır. Özellik seçimi stratejisi için ileri özellik seçimi (forward feature selection) stratejisi tercih edilmiştir. Özellik seçimi döngüsü bitiş operatörü penceresinde skor olarak doğruluk (accuracy) seçilmiştir.

İş akışının son hali Şekil 6.15'te gösterilmektedir.



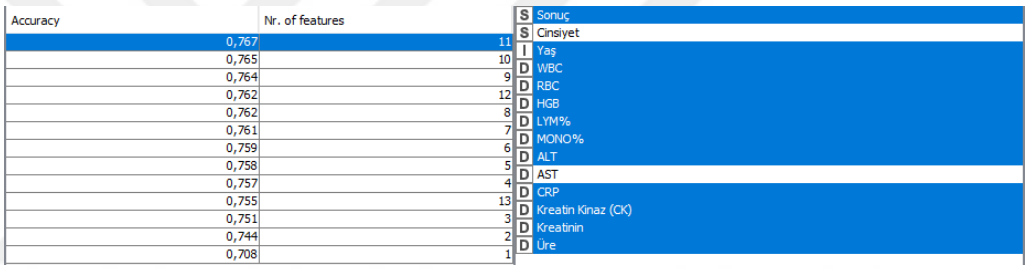
Şekil 6.15 Naive Bayes - Özellik Seçimi İçin İş Akışı

Döngü, WBC değişkeni eklenerek %70,8 ile başlamıştır. En yüksek doğruluk oranı 11. yinelemede modelde WBC, yaş, RBC, HGB, LYM%, ALT, CRP, MONO%, Kreatinin, Kreatinin Kinaz (CK), Üre öznitelikleri mevcutken elde edilmiştir. Özellik seçimi analizi sonucunda modelden cinsiyet ve AST değişkenlerinin çıkarılmasıyla %76,7 doğruluk değerinin elde edilebileceği görülmüştür (Şekil 6.16 ve Şekil 6.17).



#	RowID	Nr. of features *Number (integer)*	Accuracy *Number (double)*	Added feature *String*
1	1	1	0.708	WBC
2	2	2	0.744	MONO%
3	3	3	0.751	CRP
4	4	4	0.757	HGB
5	5	5	0.758	LYM%
6	6	6	0.759	Üre
7	7	7	0.761	RBC
8	8	8	0.762	Yaş
9	9	9	0.764	Kreatinin
10	10	10	0.765	Kreatin Kinaz (CK)
11	11	11	0.767	ALT
12	12	12	0.762	Cinsiyet
13	All	13	0.755	AST

Şekil 6.16 Naive Bayes Özellik Seçimi Analizi Sonucu



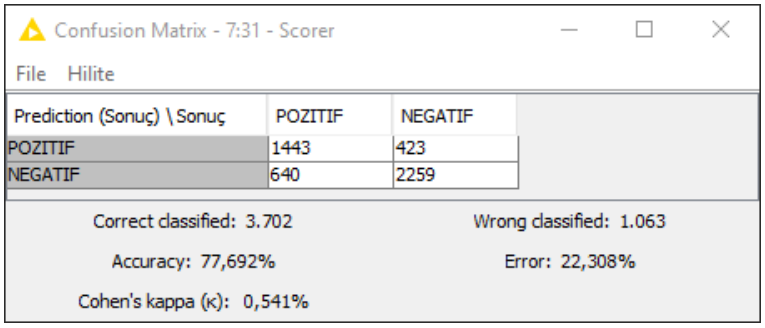
Accuracy	Nr. of features
0,767	11
0,765	10
0,764	9
0,762	12
0,762	8
0,761	7
0,759	6
0,758	5
0,757	4
0,755	13
0,751	3
0,744	2
0,708	1

S Sonuç
S Cinsiyet
I Yaş
D WBC
D RBC
D HGB
D LYM%
D MONO%
D ALT
D AST
D CRP
D Kreatin Kinaz (CK)
D Kreatinin
D Üre

Şekil 6.17 Naive Bayes Özellik Seçimi Analizi - Maksimum Doğruluk Değerini Sağlayan Kombinasyon

6.1.4 Lojistik Regresyon ile Modelleme

Oluşturulan lojistik regresyon modelinin çalıştırılması sonrası hastanın COVID-19 test sonucunu tahminleyen modelin başarısını gösteren karmaşıklık matrisi Şekil 6.18'de verilmektedir.



Prediction (Sonuç) \ Sonuç	POZITIF	NEGATIF
POZITIF	1443	423
NEGATIF	640	2259

Correct classified: 3.702 Wrong classified: 1.063
Accuracy: 77,692% Error: 22,308%
Cohen's kappa (κ): 0,541%

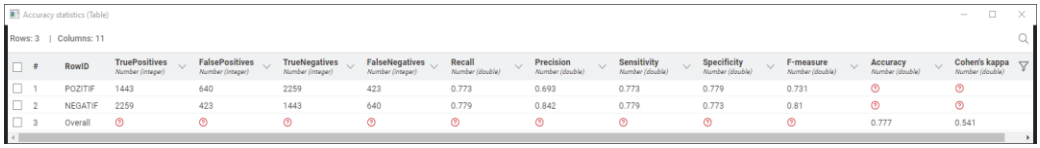
Şekil 6. 18 Lojistik Regresyon Karmaşıklık Matrisi

Elde edilen sonuçlar doğrultusunda COVID-19 test sonuçlarına göre gerçekte 2083 pozitif hastadan 1443 hastanın pozitif, 640 hastanın ise negatif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu pozitif olan hastaların doğru tahmin edilme oranı %69,3 olarak saptanmıştır. COVID-19 test sonuçlarına göre gerçekte 2682 negatif sonuçlu hastadan 2259 hastanın negatif, 423 hastanın ise pozitif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu negatif olan hastaların doğru tahmin edilme oranı %84,2 olarak saptanmıştır. Karmaşıklık matrisi köşegeninde yer alan doğru tahmin sayılarına dair oran ise %77,7 olarak saptanmıştır. Modelin teşhis oranları Tablo 6.4'te gösterilmektedir.

Tablo 6. 4 Lojistik Regresyon Modeline İlişkin Bulgular

		Gerçek			
		Test Sonucu Pozitif	Test Sonucu Negatif	Toplam	Teşhis (%)
Tahmin	Test Sonucu Pozitif	1443	423	1866	69,3%
	Test Sonucu Negatif	640	2259	2899	84,2%
	Toplam	2083	2682	4765	77,7%

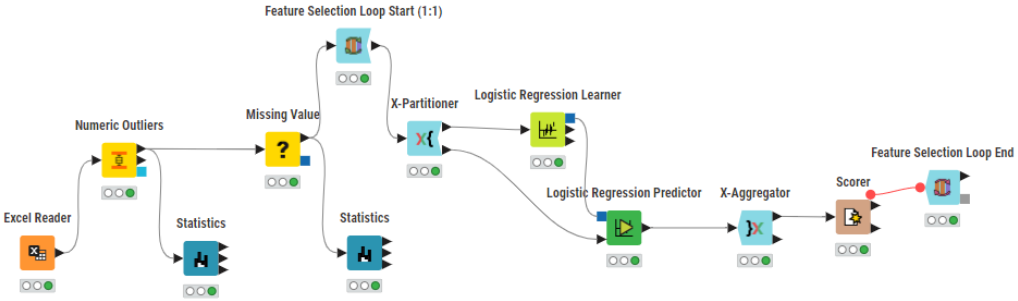
Lojistik regresyon modeline ilişkin skorcu operatörü çıktısı Şekil 6.19'da gösterilmektedir.



Şekil 6. 19 Lojistik Regresyon Modeli Skorcu Çıktısı

Model doğruluğunu arttırmak için iş akışına özellik seçimi döngüsü başlangıç operatörü (feature selection loop start (1:1)) ve özellik seçimi döngüsü bitiş operatörü eklenmiştir. Özellik seçimi döngüsü başlangıç operatörü penceresinde sınıf etiketi olan sonuç sütunu statik sütunlar (static columns) kısmına alınmıştır. Özellik seçimi stratejisi için ileri özellik seçimi (forward feature selection) stratejisi tercih edilmiştir. Özellik seçimi döngüsü bitiş operatörü penceresinde skor olarak doğruluk (accuracy) seçilmiştir.

İş akışının son hali Şekil 6.20'de gösterilmektedir.



Şekil 6.20 Lojistik Regresyon - Özellik Seçimi İçin İş Akışı

Döngü, WBC değişkeni eklenerek %71,4 ile başlamıştır. En yüksek doğruluk oranı 9. yinelemede modelde WBC, LYM%, ALT, AST, CRP, MONO%, Kreatinin, Kreatinin Kinaz (CK), Üre öznitelikleri mevcutken elde edilmiştir. Özellik seçimi analizi sonucunda modelden cinsiyet, yaş, RBC ve HGB, değişkenlerinin çıkarılmasıyla %77,8 doğruluk değerinin elde edilebileceği görülmüştür (Şekil 6.21 ve Şekil 6.22).

Result Table (Table)

Rows: 13 | Columns: 3

#	RowID	Nr. of features *Number (integer)*	Accuracy *Number (double)*	Added feature *String*
1	1	1	0.714	WBC
2	2	2	0.741	MONO%
3	3	3	0.759	CRP
4	4	4	0.762	Üre
5	5	5	0.766	AST
6	6	6	0.77	LYM%
7	7	7	0.776	Kreatin Kinaz (CK)
8	8	8	0.777	Kreatinin
9	9	9	0.778	ALT
10	10	10	0.777	RBC
11	11	11	0.776	Yaş
12	12	12	0.776	HGB
13	All	13	0.772	Cinsiyet

Şekil 6.21 Lojistik Regresyon Özellik Seçimi Analizi Sonucu

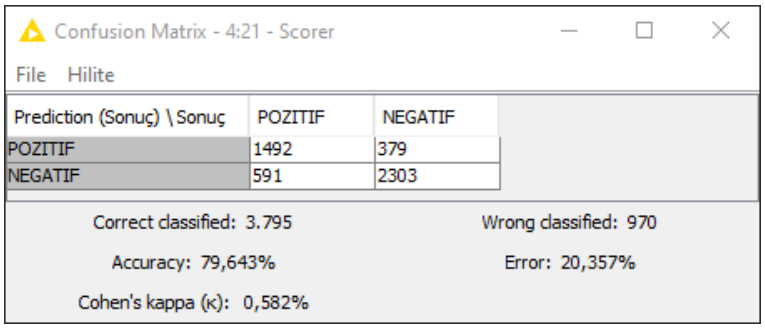
Accuracy	Nr. of features
0,778	9
0,777	10
0,777	8
0,776	7
0,776	12
0,776	11
0,772	13
0,77	6
0,766	5
0,762	4
0,759	3
0,741	2
0,714	1

	Soniuç
S	Cinsiyet
I	Yaş
D	WBC
D	RBC
D	HGB
D	LYM%
D	MONO%
D	ALT
D	AST
D	CRP
D	Kreatin Kinaz (CK)
D	Kreatinin
D	Üre

Şekil 6.22 Lojistik Regresyon Özellik Seçimi Analizi - Maksimum Doğruluk Değerini Sağlayan Kombinasyon

6.1.5 Rastgele Orman ile Modelleme

Oluşturulan rastgele orman modelinin çalıştırılması sonrası hastanın COVID-19 test sonucunu tahminleyen modelin başarısını gösteren karmaşıklık matrisi şekil 6.23'te gösterilmiştir.



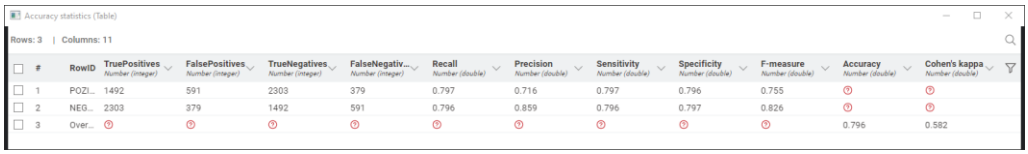
Şekil 6. 23 Rastgele Orman Karmaşıklık Matrisi

Elde edilen sonuçlar doğrultusunda COVID-19 test sonuçlarına göre gerçekte 2083 pozitif hastadan 1492 hastanın pozitif, 591 hastanın ise negatif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu pozitif olan hastaların doğru tahmin edilme oranı %71,6 olarak saptanmıştır. COVID-19 test sonuçlarına göre gerçekte 2682 negatif sonuçlu hastadan 2303 hastanın negatif, 379 hastanın ise pozitif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu negatif olan hastaların doğru tahmin edilme oranı %85,9 olarak saptanmıştır. Karmaşıklık matrisi köşegeninde yer alan doğru tahmin sayılarına ilişkin oran ise %79,6 olarak saptanmıştır. Modelin teşhis oranları Tablo 6.5'te gösterilmektedir.

Tablo 6. 5 Rastgele Orman Modeline İlişkin Bulgular

		Gerçek			
		Test Sonucu Pozitif	Test Sonucu Negatif	Toplam	Teşhis (%)
Tahmin	Test Sonucu Pozitif	1492	379	1871	71,6%
	Test Sonucu Negatif	591	2303	2894	85,9%
	Toplam	2083	2682	4765	79,6%

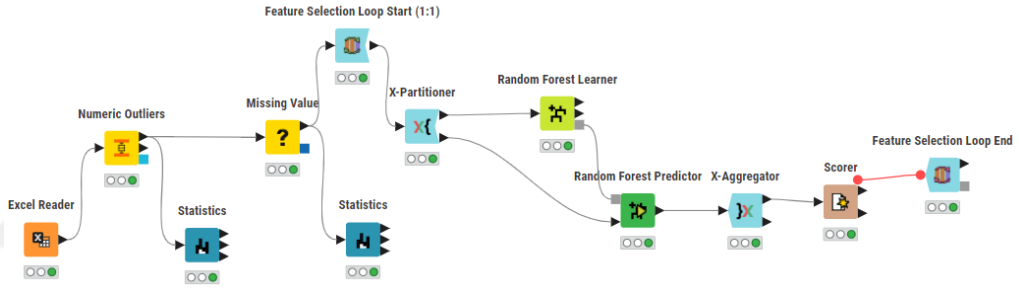
Rastgele orman modeline ilişkin skorcu operatörü çıktısı Şekil 6.24'te gösterilmektedir.



Şekil 6. 24 Rastgele Orman Modeli Skorcu Çıktısı

Model doğruluğunu arttırmak için iş akışına özellik seçimi döngüsü başlangıç operatörü (feature selection loop start (1:1)) ve özellik seçimi döngüsü bitiş operatörü eklenmiştir. Özellik seçimi döngüsü başlangıç operatörü penceresinde sınıf etiketi olan sonuç sütunu statik sütunlar (static columns) kısmına alınmıştır. Özellik seçimi stratejisi için ileri özellik seçimi (forward feature selection) stratejisi tercih edilmiştir. Özellik seçimi döngüsü bitiş operatörü penceresinde skor olarak doğruluk (accuracy) seçilmiştir.

İş akışının son hali Şekil 6.25'te gösterilmektedir.



Şekil 6.25 Rastgele Orman - Özellik Seçimi İçin İş Akışı

Döngü, WBC değişkeni eklenerek %71 ile başlamıştır. En yüksek doğruluk oranı 11. yinelemede modelde WBC, CRP, LYM%, MONO%, yaş, HGB, Kreatinin, Kreatinin Kinaz (CK), AST, RBC, Üre öznitelikleri mevcutken elde edilmiştir. Özellik seçimi analizi sonucunda modelden cinsiyet ve ALT değişkenlerinin çıkarılmasıyla %80,2 doğruluk değerinin elde edilebileceği görülmüştür (Şekil 6.26 ve Şekil 6.27).

Result Table (Table)

Rows: 13 | Columns: 3

#	RowID	Nr. of features *Number (integer)*	Accuracy *Number (double)*	Added feature *String*
1	1	1	0.71	WBC
2	2	2	0.723	CRP
3	3	3	0.767	MONO%
4	4	4	0.778	LYM%
5	5	5	0.79	Yaş
6	6	6	0.797	HGB
7	7	7	0.798	Kreatin Kinaz (CK)
8	8	8	0.799	Kreatinin
9	9	9	0.8	AST
10	10	10	0.798	RBC
11	11	11	0.802	Üre
12	12	12	0.801	ALT
13	All	13	0.796	Cinsiyet

Şekil 6.26 Rastgele Orman Özellik Seçimi Analizi Sonucu

Accuracy	Nr. of features
0,802	11
0,801	12
0,8	9
0,799	8
0,798	7
0,798	10
0,797	6
0,796	13
0,79	5
0,778	4
0,767	3
0,723	2
0,71	1

S Sonuç
S Cinsiyet
I Yaş
D WBC
D RBC
D HGB
D LYM%
D MONO%
D ALT
D AST
D CRP
D Kreatin Kinaz (CK)
D Kreatinin
D Üre

Şekil 6.27 Rastgele Orman Özellik Seçimi Analizi - Maksimum Doğruluk Değerini Sağlayan Kombinasyon

6.1.6 Destek Vektör Makineleri ile Modelleme

Oluşturulan destek vektör makineleri modelinin çalıştırılması sonrası hastanın COVID-19 test sonucunu tahminleyen modelin başarısını gösteren karmaşıklık matrisi Şekil 6.28'de gösterilmiştir.

Confusion Matrix - 4:29 - Scorer

File Hilite

Prediction (Sonuç) \ Sonuç	NEGATIF	POZITIF
NEGATIF	2242	624
POZITIF	440	1459

Correct classified: 3.701 Wrong classified: 1.064

Accuracy: 77,671% Error: 22,329%

Cohen's kappa (κ): 0,542%

Şekil 6.28 Destek Vektör Makineleri Karmaşıklık Matrisi

Elde edilen sonuçlar doğrultusunda COVID-19 test sonuçlarına göre gerçekte 2083 pozitif hastadan 1459 hastanın pozitif, 624 hastanın ise negatif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu pozitif olan hastaların doğru tahmin edilme oranı %70 olarak saptanmıştır. COVID-19 test sonuçlarına göre gerçekte 2682 negatif sonuçlu hastadan 2242 hastanın negatif, 440 hastanın ise pozitif sonuçlu olarak tahmin edildiği görülmektedir. Test sonucu negatif olan hastaların doğru tahmin edilme oranı %83,6 olarak saptanmıştır. Karmaşıklık matrisi köşegeninde yer alan doğru tahmin sayılarına ilişkin oran ise %77,7 olarak saptanmıştır. Modelin teşhis oranları Tablo 6.6'da sunulmaktadır.

Tablo 6. 6 Destek Vektör Makineleri Modeline İlişkin Bulgular

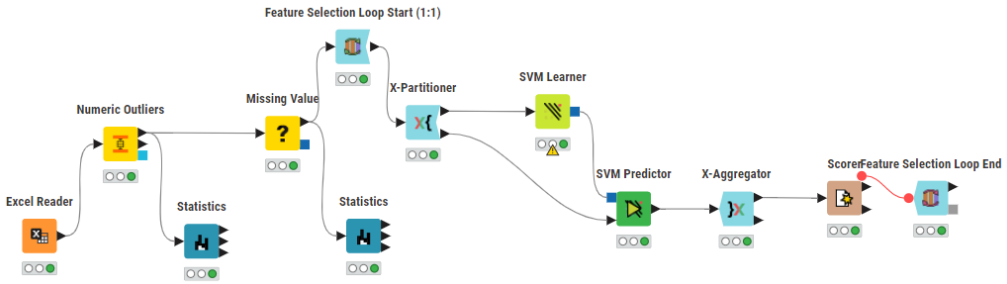
		Gerçek			
		Test Sonucu Pozitif	Test Sonucu Negatif	Toplam	Teşhis (%)
Tahmin	Test Sonucu Pozitif	1459	440	1899	70,0%
	Test Sonucu Negatif	624	2242	2866	83,6%
	Toplam	2083	2682	4765	77,7%

Destek vektör makineleri modeline ilişkin skorcu operatörü çıktısı Şekil 6.29'da gösterilmektedir.

Şekil 6.29 Destek Vektör Makineleri Modeli Skorcu Çıktısı

Model doğruluğunu arttırmak için iş akışına özellik seçimi döngüsü başlangıç operatörü (feature selection loop start (1:1)) ve özellik seçimi döngüsü bitiş operatörü eklenmiştir. Özellik seçimi döngüsü başlangıç operatörü penceresinde sınıf etiketi olan sonuç sütunu statik sütunlar (static columns) kısmına alınmıştır. Özellik seçimi stratejisi için ileri özellik seçimi (forward feature selection) stratejisi tercih edilmiştir. Özellik seçimi döngüsü bitiş operatörü penceresinde skor olarak doğruluk (accuracy) seçilmiştir.

İş akışının son hali Şekil 6.30'da gösterilmektedir.



Şekil 6.30 Destek Vektör Makineleri - Özellik Seçimi İçin İş Akışı

Döngü, WBC değişkeni eklenerek %71,4 ile başlamıştır. En yüksek doğruluk oranı 13. yinelemede modelde başlangıçta olan tüm öznitelikleri mevcutken elde edilmiştir. Özellik seçimi analizi sonucunda modelden öznitelik çıkarılmaması önerilmektedir (Şekil 6.31 ve Şekil 6.32).

Result Table (Table)

Rows: 13 | Columns: 3

#	RowID	Nr. of features *Number (integer)*	Accuracy *Number (double)*	Added feature *String*
1	1	1	0.714	WBC
2	2	2	0.742	MONO%
3	3	3	0.76	CRP
4	4	4	0.763	LYM%
5	5	5	0.766	Üre
6	6	6	0.769	ALT
7	7	7	0.774	AST
8	8	8	0.775	HGB
9	9	9	0.776	Cinsiyet
10	10	10	0.775	Kreatinin
11	11	11	0.776	RBC
12	12	12	0.775	Yaş
13	All	13	0.777	Kreatin Kinaz (CK)

Şekil 6.31 Destek Vektör Makineleri Özellik Seçimi Analizi Sonucu

Accuracy	Nr. of features
0,777	13
0,776	11
0,776	9
0,775	10
0,775	8
0,775	12
0,774	7
0,769	6
0,766	5
0,763	4
0,76	3
0,742	2
0,714	1

S Sonuç
I Cinsiyet
I Yaş
D WBC
D RBC
D HGB
D LYM%
D MONO%
D ALT
D AST
D CRP
D Kreatin Kinaz (CK)
D Kreatinin
D Üre

Şekil 6.32 Destek Vektör Makineleri Özellik Seçimi Analizi - Maksimum Doğruluk Değerini Sağlayan Kombinasyon

6.1.7 Sınıflandırma Modellerinin Performans Değerlendirmesi

Sınıflandırma modellerinin performanslarına ilişkin göstergeler Tablo 6.7'de sunulmaktadır.

Tablo 6.7 Sınıflandırma Modelleri Performans Değerlendirme Tablosu

	Karar Ağacı	K- En Yakın Komşu	Naive Bayes	Lojistik Regresyon	Rastgele Orman	Destek Vektör Makineleri
Doğruluk	0,778	0,690	0,752	0,777	**0,796**	0,777
Hata Oranı	0,222	0,310	0,248	0,223	**0,204**	0,223
Duyarlılık	0,701	0,457	0,642	0,693	**0,716**	0,700
Belirleyicilik	0,838	0,872	0,837	0,842	**0,859**	0,836
Yanlış Pozitif Oranı	0,162	0,128	0,163	0,158	**0,141**	0,164
Yanlış Negatif Oranı	0,299	0,543	0,358	0,307	**0,284**	0,300
Kesinlik	0,771	0,735	0,754	0,773	**0,797**	0,768
Negatif Öngörü Değeri	0,783	0,674	0,751	0,779	**0,796**	0,782
F	0,735	0,563	0,694	0,731	**0,755**	0,733
Pozitif Olabilirlik Oranı	4,334	3,570	3,942	4,392	**5,069**	4,269
Negatif Olabilirlik Oranı	0,356	0,623	0,427	0,365	**0,330**	0,358
Tanısal Üstünlük Oranı	12,167	5,729	9,226	12,041	**15,340**	11,914

Tablo 6.7'ye göre modeller arasında en yüksek doğruluk, 0,796 değeri ile rastgele orman modeliyle sağlanmıştır. En düşük doğruluk değeri ise 0,690 ile K-en yakın komşu modeline aittir. Tahmin modellerinin doğruluk değerleri en yüksekten en düşüğe doğru sırasıyla rastgele orman, karar ağacı, lojistik regresyon ve destek vektör makineleri, Naive Bayes, K-en yakın komşu şeklindedir.

En yüksek duyarlılık değerine 0,716 değeri ile rastgele orman modeli sahiptir. K-en yakın komşu modeli 0,457 ile en düşük duyarlılığa sahiptir. Tahmin modellerinin duyarlılık (doğru pozitif (tanıma) oranı) değerleri en yüksekten en düşüğe doğru

sırasıyla rastgele orman, karar ağacı, destek vektör makineleri, lojistik regresyon, Naive Bayes, K-en yakın komşu şeklindedir.

En yüksek belirleyicilik değerine 0,872 değeri ile K-en yakın komşu modeli sahiptir. Destek vektör makineleri modeli 0,836 ile en düşük belirleyiciliğe sahiptir. Tahmin modellerinin belirleyicilik (doğru şekilde tanımlanan negatif örneklerin oranı) değerleri en yüksekten en düşüğe doğru sırasıyla K-en yakın komşu, rastgele orman, lojistik regresyon, karar ağacı, Naive Bayes, destek vektör makineleri şeklindedir.

En yüksek tanısal üstünlük oranına 15,340 değeri ile rastgele orman modeli sahiptir. En düşük tanısal üstünlük oranına ise 5,729 değeri ile K-en yakın komşu modeli sahiptir. Tahmin modellerinin tanısal üstünlük oranına değerleri en yüksekten en düşüğe doğru sırasıyla rastgele orman, karar ağacı, lojistik regresyon, destek vektör makineleri, Naive Bayes, K-en yakın komşu şeklindedir.

İş akışlarına özellik seçimi döngüsü eklendikten sonra çıkarılması önerilen öznitelikler ve model doğrulukları Tablo 6.8'de sunulmaktadır.

Karar ağacı modeli için yaş ve RBC özniteliklerinin çıkarılması önerilmektedir. Böylece %1,6 artış sağlanarak doğruluk değerinin %79,4'e ulaşabileceği görülmüştür.

K-en yakın komşu modeli için cinsiyet, yaş, RBC, HGB, LYM%, ALT, AST, Kreatin Kinaz (CK) ve Üre özniteliklerinin çıkarılması önerilmektedir. Böylece %8,5 artış sağlanarak doğruluk değerinin %77,5'e ulaşabileceği görülmüştür.

Naive Bayes modeli için cinsiyet ve AST özniteliklerinin çıkarılması önerilmektedir. Böylece %1,5 artış sağlanarak doğruluk değerinin %76,7'ye ulaşabileceği görülmüştür.

Lojistik regresyon modeli için cinsiyet, yaş, RBC ve HGB özniteliklerinin çıkarılması önerilmektedir. Böylece %0,1 artış sağlanarak doğruluk değerinin %77,8'e ulaşabileceği görülmüştür.

Rastgele orman modeli için cinsiyet ve ALT özniteliklerinin çıkarılması önerilmektedir. Böylece %0,6 artış sağlanarak doğruluk değerinin %80,2'ye ulaşabileceği görülmüştür.

Destek vektör makineleri modeli için öznitelik çıkarılması önerilmemektedir.

Tablo 6.8 Özellik Seçimi Döngüsü Çıktıları

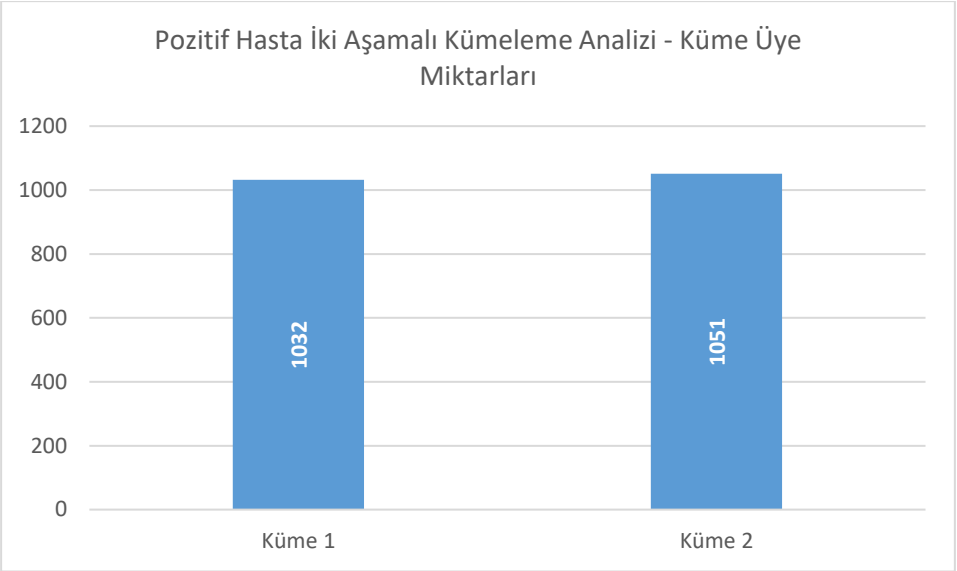
Çıkarılması Önerilen Öznitelik	Karar Ağacı	K- En Yakın Komşu	Naive Bayes	Lojistik Regresyon	Rastgele Orman	Destek Vektör Makineleri
Cinsiyet		X	X	X	X	
Yaş	X	X		X		
WBC						
RBC	X	X		X		
HGB		X		X		
LYM%		X				
MONO%						
ALT		X			X	
AST		X	X			
CRP						
Kreatin Kinaz (CK)		X				
Kreatinin	X					
Üre		X				
Yeni Doğruluk Oranı	**79,4%**	**77,5%**	**76,7%**	**77,8%**	**80,2%**	**77,7%**
Eski Doğruluk Oranı	**77,8%**	**69,0%**	**75,2%**	**77,7%**	**79,6%**	**77,7%**
Fark	**1,6%**	**8,5%**	**1,5%**	**0,1%**	**0,6%**	**0,0%**

Önerilen özniteliklerin çıkarılması sonrası elde edilen doğruluk değerleri en yüksekten en düşüğe doğru sırasıyla rastgele orman, karar ağacı, lojistik regresyon, destek vektör makineleri, K-en yakın komşu, Naive Bayes şeklindedir.

6.2 İki Aşamalı Kümeleme Analizine İlişkin Bulgular

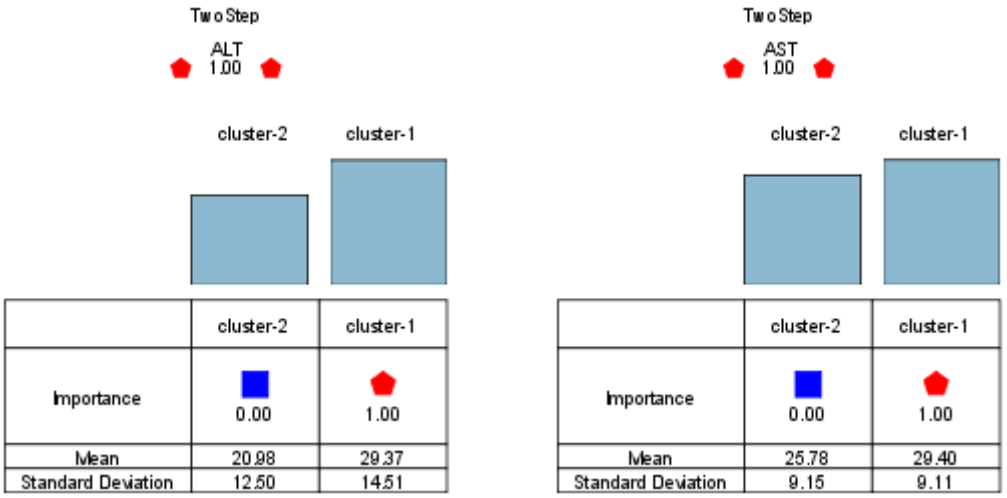
6.2.1 Pozitif Hastalar İçin İki Aşamalı Kümeleme Analizi

Test sonucu pozitif olan hastalara ilişkin analiz sonucunda iki küme elde edilmiştir. Oluşan kümelere ait üye sayıları Şekil 6.33'teki grafikte gösterilmektedir. Birinci küme 1032 hastadan, ikinci küme 1051 hastadan oluşmaktadır.



Şekil 6.33 Pozitif Hastalar İki Aşamalı Kümeleme Analizi - Küme Üye Sayıları

Değişkenlerin kümeler üzerindeki etkisi aşağıdaki şekiller üzerinden incelenmektedir. ALT değişkeni yalnızca Küme 1 için önemli etkiye sahiptir. Küme 1'deki hastaların ALT değeri ortalaması 29,37; Küme 2'deki hastaların ise 20,98 olarak elde edilmiştir. AST değişkeni yalnızca Küme 1 için ayırt edici etkiye sahiptir. Küme 1'deki hastaların AST değeri ortalaması 29,40; Küme 2'deki hastaların ise 25,78 olarak elde edilmiştir (Şekil 6.34).

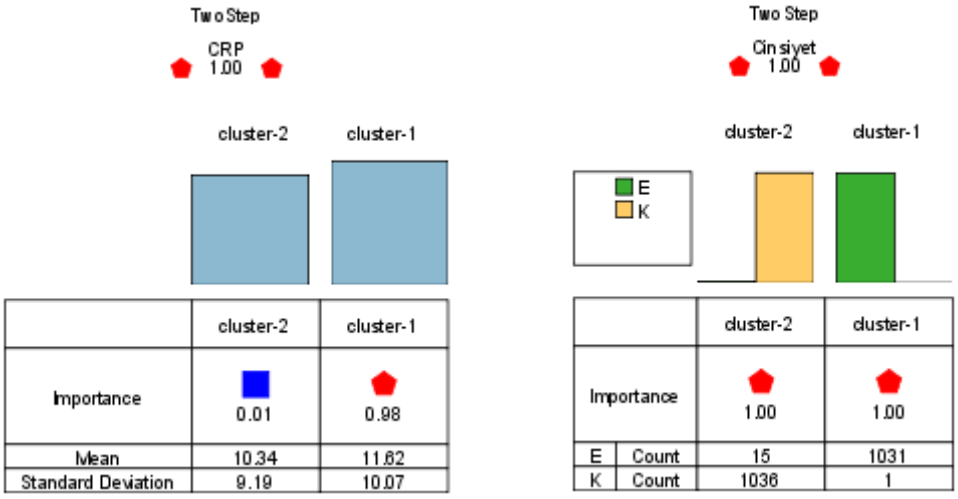


	cluster-2	cluster-1
Importance	0.00	1.00
Mean	20.98	29.37
Standard Deviation	12.50	14.51

	cluster-2	cluster-1
Importance	0.00	1.00
Mean	25.78	29.40
Standard Deviation	9.15	9.11

Şekil 6.34 Pozitif Hastalar ALT, AST - Kümeler Üzerindeki Etkisi

CRP değişkeni yalnızca Küme 1 için önemli etkiye sahiptir. Küme 1'deki hastaların CRP değeri ortalaması 11,62; Küme 2'deki hastaların ise 10,34 olarak elde edilmiştir. Cinsiyet değişkeni her iki küme için de ayırt edici etkiye sahiptir. Küme 1'deki hastaların %99,9'u erkeklerden, Küme 2'deki hastaların %98,6'sı kadınlardan oluşmaktadır (Şekil 6.35).

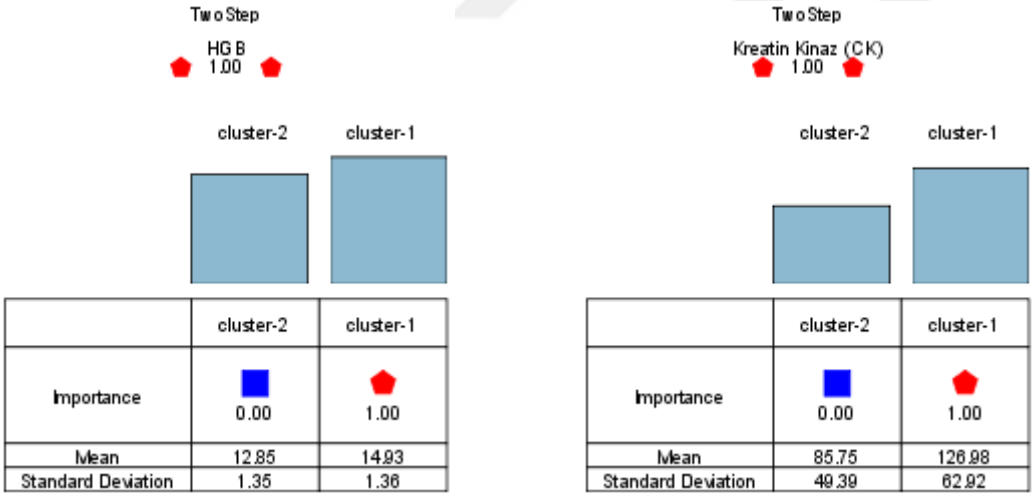


	cluster-2	cluster-1
Importance	0.01	0.98
Mean	10.34	11.62
Standard Deviation	9.19	10.07

		cluster-2	cluster-1
Importance		1.00	1.00
E	Count	15	1031
K	Count	1036	1

Şekil 6.35 Pozitif Hastalar CRP, Cinsiyet - Kümeler Üzerindeki Etkisi

HGB değişkeni yalnızca Küme 1 için önemli etkiye sahiptir. Küme 1'deki hastaların HGB değeri ortalaması 14,93; Küme 2'deki hastaların ortalaması ise 12,85 olarak elde edilmiştir. Kreatin Kinaz değişkeni yalnızca Küme 1 için önemli etkiye sahiptir. Küme 1'deki hastaların Kreatin Kinaz değeri ortalaması 126,98; Küme 2'deki hastaların ise 85,75 olarak elde edilmiştir (Şekil 6.36).

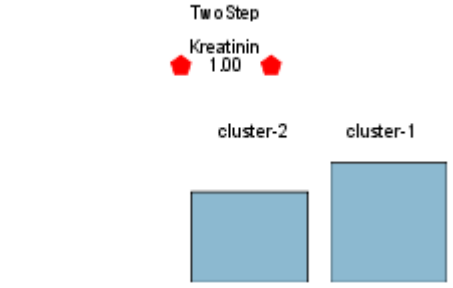


	cluster-2	cluster-1
Importance	0.00	1.00
Mean	12.85	14.93
Standard Deviation	1.35	1.36

	cluster-2	cluster-1
Importance	0.00	1.00
Mean	85.75	126.98
Standard Deviation	49.39	62.92

Şekil 6.36 Pozitif Hastalar HGB, Kreatin Kinaz - Kümeler Üzerindeki Etkisi

Kreatinin değişkeni yalnızca Küme 1 için ayırt edici etkiye sahiptir. Küme 1'deki hastaların Kreatinin değeri ortalaması 0,94; Küme 2'deki hastaların ise 0,71 olarak elde edilmiştir. MONO% değişkeni yalnızca Küme 1 için önemli etkiye sahiptir. Küme 1'deki hastaların MONO% değeri ortalaması 11,44 iken Küme 2'deki hastaların ortalaması ise 10,36 olarak elde edilmiştir (Şekil 6.37).



	cluster-2	cluster-1
Importance	0.00	1.00
Mean	0.71	0.94
Standard Deviation	0.18	0.19

	cluster-2	cluster-1
Importance	0.00	1.00
Mean	10.36	11.44
Standard Deviation	3.75	3.65

Şekil 6.37 Pozitif Hastalar Kreatinin, MONO% - Kümeler Üzerindeki Etkisi

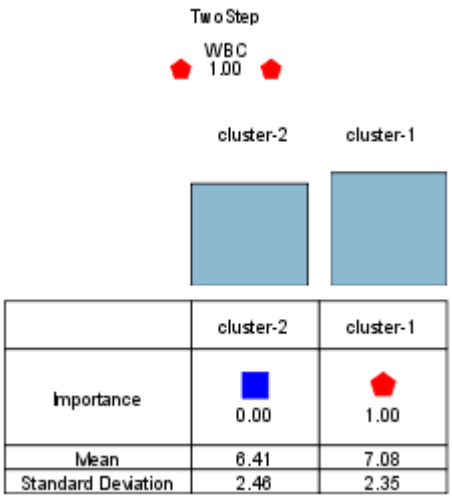
RBC değişkeni yalnızca Küme 1 için ayırt edici etkiye sahiptir. Küme 1'deki hastaların RBC değeri ortalaması 4,99 iken Küme 2'deki hastaların ortalama değeri 4,48 olarak elde edilmiştir. Üre değişkeni yalnızca Küme 1 için önemli etkiye sahiptir. Küme 1'deki hastaların Üre değeri ortalaması 30,72; Küme 2'deki hastaların ise 25,95 olarak elde edilmiştir (Şekil 6.38).

	cluster-2	cluster-1
Importance	0.00	1.00
Mean	4.48	4.99
Standard Deviation	0.42	0.44

	cluster-2	cluster-1
Importance	0.00	1.00
Mean	25.95	30.72
Standard Deviation	9.98	9.35

Şekil 6.38 Pozitif Hastalar RBC, Üre - Kümeler Üzerindeki Etkisi

WBC değişkeni yalnızca Küme 1 için önemli etkiye sahiptir. Küme 1'deki hastaların WBC değeri ortalaması 7,08; Küme 2'deki hastaların ise 6,41 olarak elde edilmiştir (Şekil 6.39).



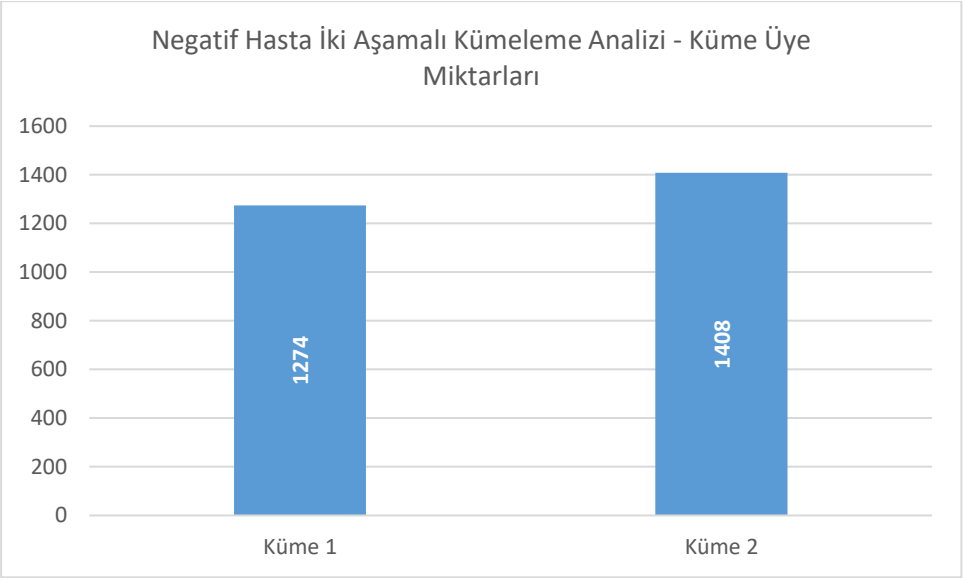
Şekil 6.39 Pozitif Hastalar WBC - Kümeler Üzerindeki Etkisi

Küme 1 için genel bir değerlendirme yapacak olursak bu kümenin çoğunlukla erkek hastalardan oluştuğunu ve bu hastaların değişkenlere ilişkin ortalama değerlerinin ALT için 29,37; AST için 29,40; CRP için 11,62; HGB için 14,93; Kreatin Kinaz için 126,98; Kreatinin için 0,94; MONO% için 11,44; RBC için 4,99; Üre için 30,72 ve WBC için 7,08 olduğunu söylemek mümkündür.

Küme 2 ise çoğunlukla kadın hastalardan oluşmaktadır. Bu kümenin ALT değerleri ortalaması 20,98; AST değerleri ortalaması 25,78; CRP değerleri ortalaması 10,34; HGB değerleri ortalaması 12,85; Kreatin Kinaz değerleri ortalaması 85,75; Kreatinin değerleri ortalaması 0,71; MONO% değerleri ortalaması 10,36; RBC değerleri ortalaması 4,48; Üre değerleri ortalaması 25,95 ve WBC değerleri ortalaması 6,41'dir.

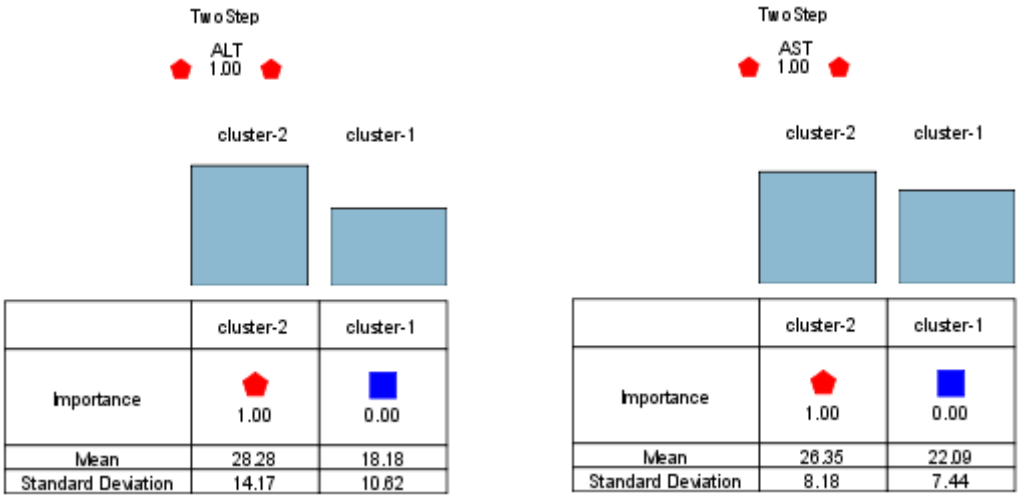
6.2.2 Negatif Hastalar İçin İki Aşamalı Kümeleme Analizi

Test sonucu negatif olan hastalara ilişkin analize göre iki küme elde edilmiştir. Oluşan kümelere ait üye sayıları Şekil 4.40'taki grafikte gösterilmektedir. Birinci küme 1274 hastadan, ikinci küme 1408 hastadan oluşmaktadır.



Şekil 6.40 Negatif Hastalar İki Aşamalı Kümeleme Analizi - Küme Üye Sayıları

Değişkenlerin kümeler üzerindeki etkisi aşağıda incelenmektedir. ALT değişkeni yalnızca Küme 2 için önemli etkiye sahiptir. Küme 1'deki hastaların ALT değeri ortalaması 18,18; Küme 2'deki hastaların ise 28,28 olarak elde edilmiştir. AST değişkeni yalnızca Küme 2 için ayırt edici etkiye sahiptir. Küme 1'deki hastaların AST değeri ortalaması 22,09; Küme 2'deki hastaların ise 26,35 olarak elde edilmiştir (Şekil 6.41).

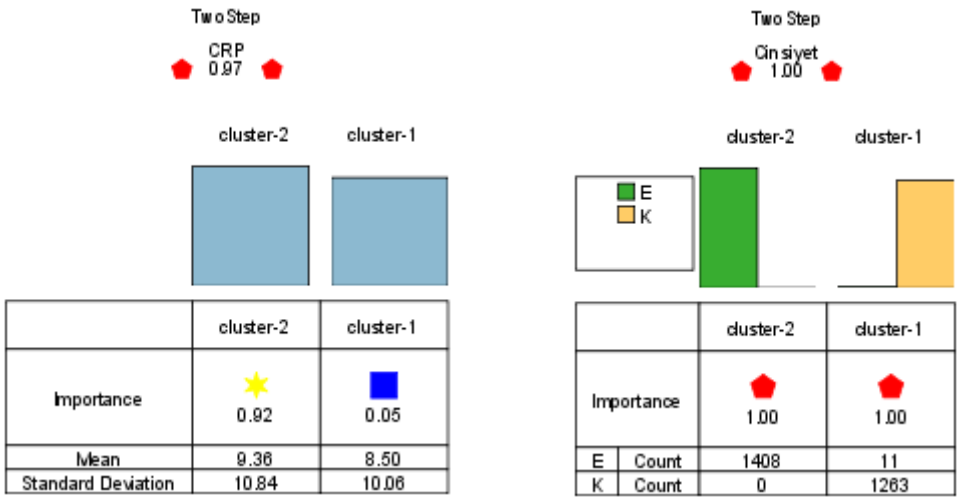


	cluster-2	cluster-1
Importance	1.00	0.00
Mean	28.28	18.18
Standard Deviation	14.17	10.62

	cluster-2	cluster-1
Importance	1.00	0.00
Mean	26.35	22.09
Standard Deviation	8.18	7.44

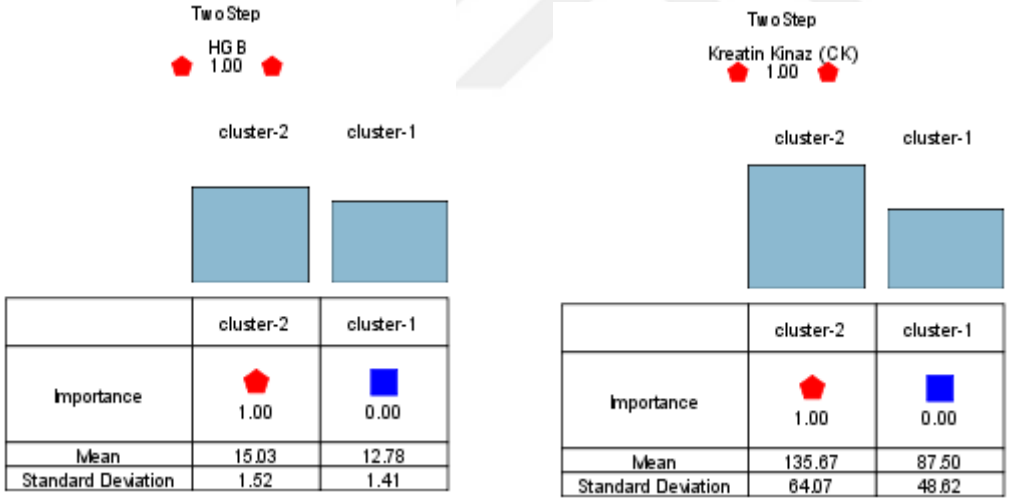
Şekil 6.41 Negatif Hastalar ALT, AST- Kümeler Üzerindeki Etkisi

Küme 1'deki hastaların CRP değeri ortalaması 8,50; Küme 2'deki hastaların ise 9,36 olarak elde edilmiştir. Cinsiyet değişkeni her iki küme için de ayırt edici etkiye sahiptir. Küme 1'deki hastaların %99,1'i kadınlardan, Küme 2'deki hastaların tamamı erkeklerden oluşmaktadır (Şekil 6.42).



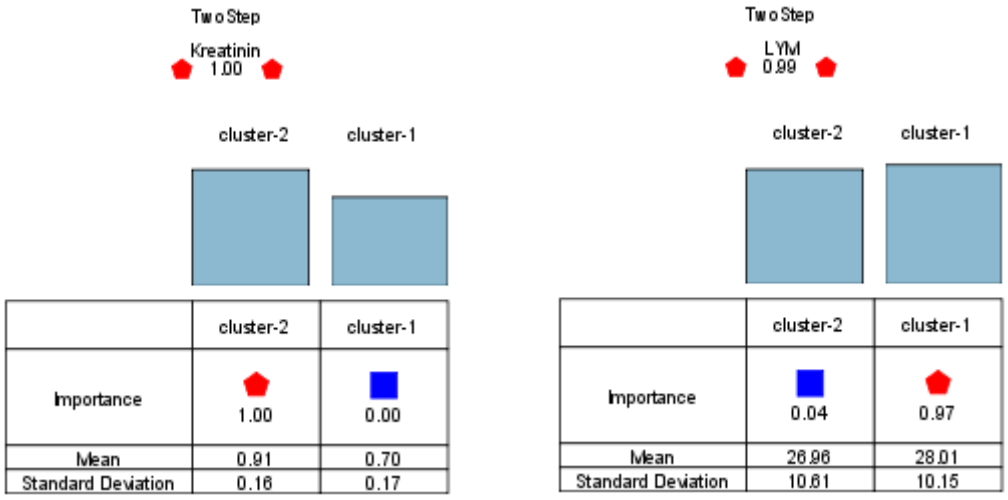
Şekil 6.42 Negatif Hastalar CRP, Cinsiyet - Kümeler Üzerindeki Etkisi

HGB değişkeni yalnızca Küme 2 için önemli etkiye sahiptir. Küme 1'deki hastaların HGB değeri ortalaması 12,78; Küme 2'deki hastaların ise 15,03 olarak elde edilmiştir. Kreatin Kinaz değişkeni yalnızca Küme 2 için önemli etkiye sahiptir. Küme 1'deki hastaların Kreatin Kinaz değeri ortalaması 87,50; Küme 2'deki hastaların ise 135,67 olarak elde edilmiştir (Şekil 6.43).



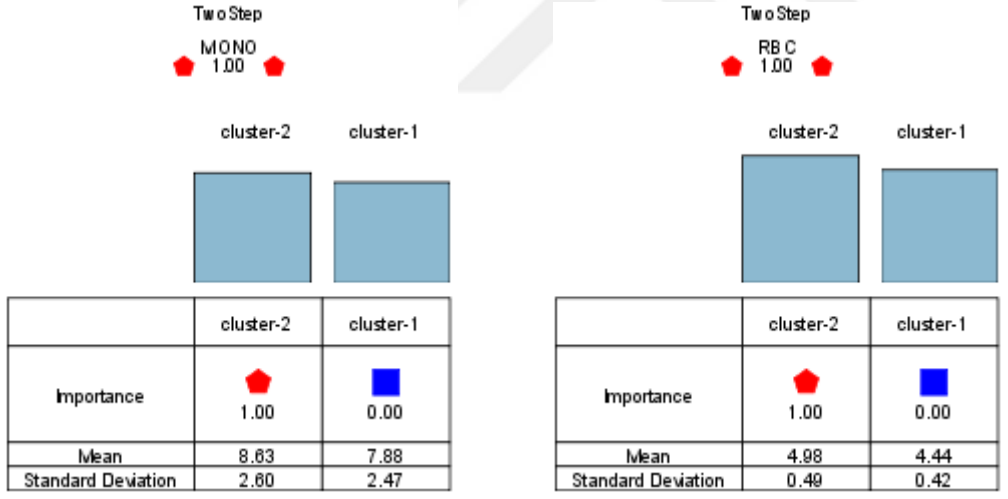
Şekil 6.43 Negatif Hastalar HGB, Kreatin Kinaz (CK) - Kümeler Üzerindeki Etkisi

Kreatinin değişkeni yalnızca Küme 2 için önemli etkiye sahiptir. Küme 1'deki hastaların Kreatinin değeri ortalaması 0,70 iken Küme 2'deki hastaların ortalama değeri 0,91 olarak elde edilmiştir. LYM% değişkeni yalnızca Küme 1 için ayırt edici etkiye sahiptir. Küme 1'deki hastaların LYM% değeri ortalaması 28,01; Küme 2'deki hastaların ise 26,96 olarak elde edilmiştir (Şekil 6.44).



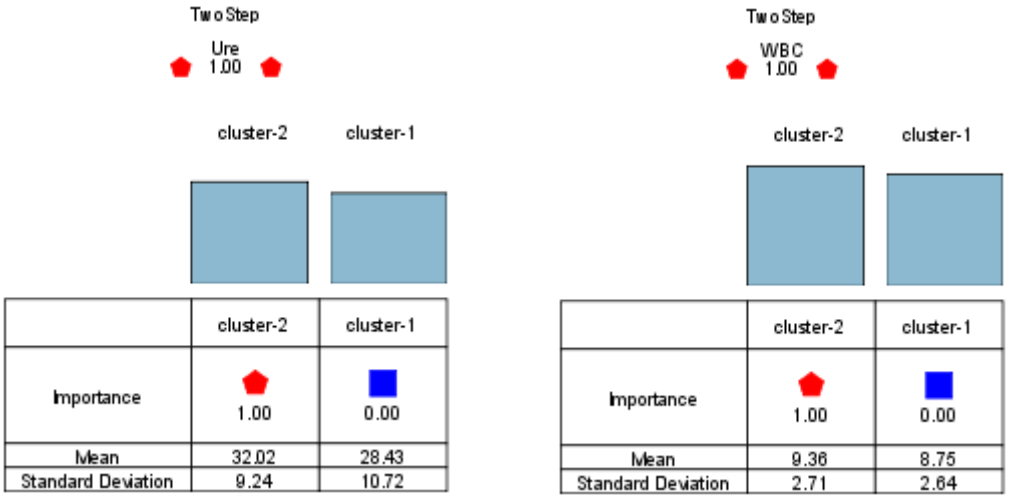
Şekil 6.44 Negatif Hastalar Kreatinin, LYM% - Kümeler Üzerindeki Etkisi

MONO% değişkeni yalnızca Küme 2 için önemli etkiye sahiptir. Küme 1'deki hastaların MONO% değeri ortalaması 7,88; Küme 2'deki hastaların 8,63 olarak elde edilmiştir. RBC değişkeni yalnızca Küme 2 için ayırt edici etkiye sahiptir. Küme 2'deki hastaların RBC değeri ortalaması 4,98 iken Küme 1'deki hastaların ortalama değeri 4,44 olarak elde edilmiştir (Şekil 6.45).



Şekil 6.45 Negatif Hastalar MONO%, RBC - Kümeler Üzerindeki Etkisi

Üre değişkeni yalnızca Küme 2 için ayırt edici etkiye sahiptir. Küme 1'deki hastaların Üre değeri ortalaması 28,43; Küme 2'deki hastaların 32,02 olarak elde edilmiştir. WBC değişkeni yalnızca Küme 2 için önemli etkiye sahiptir. Küme 1'deki hastaların WBC değeri ortalaması 8,75; Küme 2'deki hastaların 9,36 olarak elde edilmiştir (Şekil 6.46).

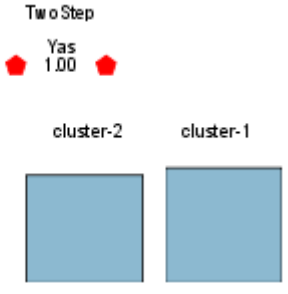


	cluster-2	cluster-1
Importance	1.00	0.00
Mean	32.02	28.43
Standard Deviation	9.24	10.72

	cluster-2	cluster-1
Importance	1.00	0.00
Mean	9.36	8.75
Standard Deviation	2.71	2.64

Şekil 6.46 Negatif Hastalar Üre, WBC - Kümeler Üzerindeki Etkisi

Yaş değişkeni yalnızca Küme 1 için ayırt edici etkiye sahiptir. Küme 1'in yaş ortalaması 45 iken Küme 2 ortalama 42 yaşındaki hastalardan oluşmaktadır (Şekil 6.47).



	cluster-2	cluster-1
Importance	0.00	1.00
Mean	41.93	44.76
Standard Deviation	16.81	17.82

Şekil 6.47 Negatif Hastalar Yaş - Kümeler Üzerindeki Etkisi

Küme 1 için genel bir değerlendirme yapacak olursak bu kümenin çoğunlukla yaş ortalaması 45 olan kadın hastalardan oluştuğunu ve diğer değişkenler için ortalama değerlerin sırasıyla ALT için 18,18; AST için 22,09; CRP için 8,50; HGB için 12,78; Kreatin Kinaz için 87,50; Kreatinin için 0,70; LYM% için 28,01; MONO% için 7,88; RBC için 4,44; Üre için 28,43 ve WBC için 8,75 olduğunu söylemek mümkündür.

Küme 2 ise çoğunlukla yaş ortalaması 42 olan erkek hastalardan oluşmaktadır. Bu kümedeki hastaların ALT değerleri ortalaması 28,28; AST değerleri ortalaması 26,35; CRP değerleri ortalaması 9,36; HGB değerleri ortalaması 15,03; Kreatin Kinaz değerleri ortalaması 135,67; Kreatinin değerleri ortalaması 0,91; LYM% değerleri

ortalaması 26,96; MONO% değerleri ortalaması 8,63; RBC değerleri ortalaması 4,98; Üre değerleri ortalaması 32,02 ve WBC değerleri ortalaması 9,36'dır.

7. SONUÇ VE ÖNERİLER

COVID-19, ortaya çıktığı ilk günden itibaren tüm dünyayı fiziksel, mental ve ekonomik olarak tehdit etmiştir. Ülkemizdeki etkileri ağırlıklı olarak 2020-2022 yıllarında hissedilmiş olsa da, COVID-19 yok olmamış olup etkilerini sürdürmektedir. Bu tez çalışması, COVID-19 şüphesi taşıyan hastalara ilişkin verilerin veri madenciliğinde sınıflandırma ve kümeleme yöntemleri ile değerlendirilmesini içermektedir.

Çalışmada kullanılan veri seti ön hazırlık aşaması için %20'den fazla eksik değere sahip olan gözlem ve özniteliklerin kaldırılması, aykırı değerlerin izin verilen en yakın değer ile değiştirilmesi ve eksik verilerin metin veri tipindeki değişkenler için en sık tekrarlanan değer, sayısal veri tipindeki değişkenler için ortalama değerin eksik veri yerine atanması stratejileri uygulanmıştır. Veri seti, COVID-19 testi yaptırmış olan ayaktan tedavi gören hastalara ait sonuç, cinsiyet, yaş, WBC, RBC, HGB, LYM%, MONO%, ALT, AST, CRP, Kreatin Kinaz, Kreatinin ve Üre olmak üzere toplamda on dört öznitelikten oluşmaktadır.

Sınıflandırma çalışması bölümünde COVID-19 testi yaptırmış olan hastalara ait cinsiyet, yaş, WBC, RBC, HGB, LYM%, MONO%, ALT, AST, CRP, Kreatin Kinaz, Kreatinin ve Üre olmak üzere toplamda on üç öznitelik kullanılarak test sonucunun tahminlenmesi hedeflenmiştir. Özniteliklerin seçimi ve özniteliklerin elenmesi aşamalarında uzman görüşünden yararlanılmıştır. Oluşturulan veri seti karar ağacı, K-en yakın komşu, Naive Bayes, lojistik regresyon, rastgele orman ve destek vektör makineleri algoritmalarıyla KNIME 5.2.2 kullanılarak modellenmiştir. Oluşturulan modeller doğruluk oranı, hata oranı, duyarlılık, belirleyicilik, yanlış pozitif oranı, yanlış negatif oranı, kesinlik, negatif öngörü değeri, F, pozitif olabilirlik oranı, negatif olabilirlik oranı, tanısal üstünlük oranı bakımından değerlendirilmiştir. En yüksek doğruluk 0,796 ve en yüksek tanısal üstünlük oranı 15,340 ile rastgele orman modelinde elde edilmiştir. En düşük doğruluk 0,690 ve en düşük tanısal üstünlük oranı 5,729 değeri ile K-en yakın komşu modelinde elde edilmiştir.

Oluşturulan modellere özellik seçim döngüsü eklenerek modelden çıkarılması önerilen öznitelikler saptanmış ve doğruluk değerlerine etkisi incelenmiştir. Elde edilen çıktılara göre özellik seçimi sonrası en yüksek doğruluk oranını rastgele orman modeli vermiştir. Cinsiyet ve ALT özniteliklerinin çıkarılması sonrası rastgele orman

modelinin %80,2 doğruluğa ulaşabileceği görülmüştür. Özellik seçimi ile doğruluk oranının en çok arttığı model %8,5 ile k-en yakın komşu modeli olmuştur. K-en yakın komşu aynı zamanda modelden en çok öznitelik çıkarılması önerilen model olmuştur. Cinsiyet, yaş, RBC, HGB, LYM%, ALT, AST, Kreatin Kinaz (CK) ve Üre olmak üzere toplam dokuz özniteliğin çıkarılması sonrası elde edilen doğruluk oranı %77,5 olmuştur. Modelden öznitelik çıkarılması önerilmeyen tek model ise destek vektör makineleri olmuştur. Bununla beraber cinsiyet değişkeni uygulanan altı modelden dördünde çıkarılması önerilerek, en çok çıkarılması önerilen öznitelik olmuştur.

Kümeleme çalışması bölümünde veri seti, COVID-19 test sonucu negatif ve pozitif olan hastalar için ayrılmış, iki ayrı iki aşamalı kümeleme analizi gerçekleştirilmiştir. Uygulama için SPSS Clementine 12.0 programı kullanılmıştır.

Test sonucu pozitif olan hastalar için gerçekleştirilen analizde yaş ve LYM% değişkeninin kümeler üzerinde önemli etkisi olmadığı tespit edilerek modelden çıkartılmış ve analiz tekrarlanmıştır. Analiz sonucunda iki küme elde edilmiştir. Küme 1 çoğunlukla erkek hastalardan oluşmakta olup 1032 üyeye sahiptir. Bu kümenin ALT değerleri ortalaması 29,37; AST değerleri ortalaması 29,40; CRP değerleri ortalaması 11,62; HGB değerleri ortalaması 14,93; Kreatin Kinaz değerleri ortalaması 126,98; Kreatinin değerleri ortalaması 0,94; MONO% değerleri ortalaması 11,44; RBC değerleri ortalaması 4,99; Üre değerleri ortalaması 30,72 ve WBC değerleri ortalaması 7,08 olarak elde edilmiştir. Küme 2 ise çoğunlukla kadın hastalardan oluşmakta olup 1051 üyeye sahiptir. Bu kümenin özniteliklere ilişkin ortalama değerleri ALT için 20,98; AST için 25,78; CRP için 10,34; HGB için 12,85; Kreatin Kinaz için 85,75; Kreatinin için 0,71; MONO% için 10,36; RBC için 4,48; Üre için 25,95 ve WBC için 6,41 olarak elde edilmiştir.

Test sonucu negatif olan hastalar için gerçekleştirilen analizde tüm değişkenlerin kümeler üzerinde önemli etkisi olduğu tespit edilmiştir. Analiz sonucunda iki küme elde edilmiştir. Küme 1 yaş ortalaması 45 olan kadın hastalardan oluşmaktadır. Bu kümenin ALT değerleri ortalaması 18,18; AST değerleri ortalaması 22,09; CRP değerleri ortalaması 8,50; HGB değerleri ortalaması 12,78; Kreatin Kinaz değerleri ortalaması 87,50; Kreatinin değerleri ortalaması 0,70; LYM% değerleri ortalaması 28,01; MONO% değerleri ortalaması 7,88; RBC değerleri ortalaması 4,44; Üre değerleri ortalaması 28,43 ve WBC değerleri ortalaması 8,75 olarak elde edilmiştir. Küme 2 ise yaş ortalaması 42 olan erkek hastalardan oluşmaktadır. Bu kümenin

özniteliklere ilişkin ortalama değerleri sırasıyla ALT için 28,28; AST için 26,35; CRP için 9,36; HGB için 15,03; Kreatin Kinaz için 135,67; Kreatinin için 0,91; LYM% için 26,96; MONO% için 8,63; RBC için 4,98; Üre için 32,02 ve WBC için 9,36 olarak bulunmuştur.

İki aşamalı kümeleme analizi sonuçları Tablo 7.1'de özetlenmiştir.

Tablo 7.1 İki Aşamalı Kümeleme Analizi Sonuçları

	POZİTİF	NEGATİF	POZİTİF	NEGATİF			
	KÜME 1	KÜME 2	KÜME 2	KÜME 1	Min.	Mak.	Birim
Hasta Adet	1032	1408	1051	1274			
Cinsiyet	%99 Erkek	Erkek	%98 Kadın	%99 Kadın			
Yaş	-	42	-	45			
ALT	29,37	28,28	20,98	18,18	0	50	U/L
AST	29,4	26,35	25,78	22,09	5	50	U/L
CRP	11,62	9,36	10,34	8,50	0	5	mg/L
HGB	14,93	15,03	12,85	12,78	12	17,4	g/dL
Kreatin Kinaz (CK)	126,98	135,67	85,75	87,50	0	171	U/L
Kreatinin	0,94	0,91	0,71	0,70	0,67	1,17	mg/dL
LYM%	-	26,96	-	28,01	10	50	%
MONO%	11,44	8,63	10,36	7,88	0	12	%
RBC	4,99	4,98	4,48	4,44	4,1	5,7	10^12/L
Üre	30,72	32,02	25,95	28,43	17	43	mg/dL
WBC	7,08	9,36	6,41	8,75	4,5	10,5	10^9/L

Tablo 7.1'de toplamda dört kümeye ait sonuçlar ve kullanılan özniteliklerin minimum ve maksimum değerleri gösterilmiştir. ALT değeri hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha yüksektir. ALT değeri pozitif kadın hastalarda, erkek hastalara göre ortalama 2,5 kat daha fazla artış göstermiştir. AST değeri hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha yüksektir. Tüm kümelerdeki hastaların CRP değerleri sınır değerin üzerindedir. Bununla beraber CRP değeri hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha yüksektir. Cinsiyet bazında pozitif ve negatif hastaların Hemoglobin (HGB) değerleri birbirine yakın olmakla beraber erkek hastalarda negatif hastaların Hemoglobin değeri daha yüksek, kadınlarda ise pozitif hastaların Hemoglobin değeri daha yüksek elde edilmiştir. Kadın hastaların Hemoglobin değerlerinin alt sınıra yakın olduğu dikkat

çekmiştir. Kreatin Kinaz (CK) değeri hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha düşüktür. Kreatin Kinaz değeri pozitif erkek hastalarda, kadın hastalara göre ortalama 5 kat daha fazla düşüş göstermiştir. Cinsiyet bazında pozitif ve negatif hastaların Kreatinin değerleri birbirine yakın olmakla beraber, hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha yüksektir. MONO% değeri hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha yüksektir. Pozitif olan erkek hastaların MONO% değerlerinin üst sınıra yakın olduğu dikkat çekmiştir. Cinsiyet bazında pozitif ve negatif hastaların Eritrosit (RBC) değerleri birbirine yakın olmakla beraber, hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha yüksektir. Eritrosit değeri pozitif kadın hastalarda, erkek hastalara göre ortalama 4 kat daha fazla artış göstermiştir. Üre değeri hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha düşüktür. Üre değeri pozitif kadın hastalarda, erkek hastalara göre ortalama 2 kat daha fazla düşüş göstermiştir. Lökosit (WBC) değeri hem kadın hem de erkek COVID-19 hastalarında, test sonucu negatif olan hastalara göre daha düşüktür.

Gelecek çalışmalar için uygulama alanı Samsun'un farklı hastanelerine ya da Türkiye'nin farklı şehirlerindeki hastaneler de dahil edilerek uygulama kapsamı genişletilebilir. Uygulamaya farklı öznitelikler dahil edilerek çalışma genişletilebilir. Bununla beraber literatürdeki farklı yöntemler kullanılarak sonuçlar mukayese edilebilir.

KAYNAKLAR

Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2022). The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. *Quality & Quantity*, 56(3), 1283-1291.

Aylıkcı, Burak, Filo Araçlarının Kümeleme Yöntemiyle Hasar Analizi, Üsküdar Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul 2023.

Baca, Hakan, Kümeleme Analizi: Türkiye İllerinin Sektörel Bazda Kümelenmesi, Kafkas Üniversitesi Sosyal Bilimler Enstitüsü Yüksek Lisans Tezi, Kars 2022.

Ceylan, Z., Gürsev, S., & Bulkan, S. (2017). İki aşamalı kümeleme analizi ile bireysel emeklilik sektöründe müşteri profilinin değerlendirilmesi. *Bilişim Teknolojileri Dergisi*, 10(4), 475-485.

Chan, M. F., Al-Shekaili, M., Al-Adawi, S., Hassan, W., Al-Said, N., Al-Sulaimani, F., ... & Al-Mawali, A. (2021). Mental health outcomes among health-care workers in Oman during COVID-19: A cluster analysis. *International Journal of Nursing Practice*, 27(6), e12998.

Chimbunde, E., Sigwadhi, L. N., Tamuzi, J. L., Okango, E. L., Daramola, O., Ngah, V. D., & Nyasulu, P. S. (2023). Machine learning algorithms for predicting determinants of COVID-19 mortality in South Africa. Frontiers in Artificial Intelligence, 6.

Çelik, G. G., & Öngel, G. (2021). COVİD-19 Döneminde Sağlık Çalışanlarının Marketten Eve Online Alışverişe Yönelik Deneyimlerini Oluşturan Faktörler: İki Aşamalı Kümeleme Analizi. *Akademik Hassasiyetler*, 8(17), 453-471.

Çotoy, Yeliz, Predicting ICU Requirements of COVID-19 Patients Using Artificial Neural Network, Galatasaray University Graduate School Of Science And Engineering Master's Thesis, İstanbul 2022.

Dağaslanı, Hatice, Güncellik, Sıklık, Parasallık Analizi ve K- Ortalamalar Kümeleme Analizi ile Müşteri Bölümlendirme: Perakende Sektöründe Bir Uygulama, T.C. İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul 2022.

Demircioğlu, M., & Eşiyok, S. (2020). COVID–19 Salgını ile Mücadelede Kümeleme Analizi İle Ülkelerin Sınıflandırılması. *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi*, 19(37), 369-389.

Doğan, Ezgi, COVID-19 Hastaları İçin Makine Öğrenmesi Sınıflandırma Yöntemleriyle Hastalık Evresinin Tahmini, Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İzmir 2023.

Doğançay, Melis Merve, Kriptoparalarda Kümeleme Analizi Uygulamaları, İstanbul Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Yüksek Lisans Tezi, İstanbul 2023.

Dolgun, M. Ö. (2006). Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Yüksek Lisans Tezi, Ankara.

Ersöz, F. (2019). *Veri Madenciliği Teknikleri ve Uygulamaları Kavram – Teori – Modeller – Yöntem*. Ankara: Seçkin Yayıncılık

Eser, Mehmet Taha, Uluslararası Öğrenci Değerlendirme Programı 2015 Verilerinin Veri Madenciliğinde Kümeleme Yöntemleriyle İncelenmesi, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Doktora Tezi, Ankara 2019.

Fernández-de-Las-Peñas, C., Martín-Guerrero, J. D., Florencio, L. L., Navarro-Pardo, E., Rodríguez-Jiménez, J., Torres-Macho, J., & Pellicer-Valero, O. J. (2023). Clustering analysis reveals different profiles associating long-term post-COVID symptoms,

COVID-19 symptoms at hospital admission and previous medical co-morbidities in previously hospitalized COVID-19 survivors. *Infection*, 51(1), 61-69.

Han, J., & Pei J. & Tong H. (2023). *Data Mining Consepts and Techniques*. Cambridge, United States: Elsevier Inc.

Harrington, P. 2012. *Machine Learning in Action* (First Edition). Shelter Island, New York: Manning Publications

Giray, S. (2016). İki Aşamalı Kümeleme Analizi ile Hükümlü Verilerinin İncelenmesi. *İstanbul Üniversitesi İktisat Fakültesi Ekonometri ve İstatistik Dergisi*, 25:1-31.

Gök, Elif Ceren, Kan Değerleri ile COVID-19 Enfekte Düzeyinin Rassal Orman Sınıflandırıcı ile Tahmin Edilmesi, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Isparta 2021.

Kartal, E., Balaban, M. E., & Bayraktar, B. (2021). Changing Status of Global COVID-19 Outbreak in The World And in Turkey And Clustering Analysis/Küresel COVID-19 Salgının Dünyada ve Türkiye'de Değişen Durumu ve Kümeleme Analizi. *Journal of Istanbul Faculty of Medicine*, 84(1), 9-20.

Kayrı, M. (2007). Araştırmalarda İki Aşamalı Kümeleme (Two-Step Clustering) Analizi ve Bir Uygulaması. *Eurasian Journal of Educational Research (EJER)*, (28).

Kişi, Esin, Sürdürülebilir Kalkınma Göstergelerine Göre K-Ortalamalar Yöntemi ile Kümeleme Analizi Uygulaması, Süleyman Demirel Üniversitesi sosyal Bilimler Enstitüsü Yüksek Lisans Tezi, Isparta 2021.

Kumar, N. K., Rajkumar, G., Poojitha, N., Prasad, K. R., Reddy, P. S., & Kalyani, C. (2024, March). Enhancing COVID-19 Prediction: Optimized Random Forest Classifier with Feature Selection. In 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) (Vol. 2, pp. 1-6). IEEE.

Kumar, S. (2020). Monitoring Novel Corona Virus (COVID-19) Infections in India by Cluster Analysis. *Annals of Data Science*, 7(3), 417-425.

Kurnaz, Gülşah, Kablo Takımı Üretim Süresinin ve Kusurlu Ürün Oluşumuna Yönelik Risk Faktörlerinin Makine Öğrenmesi Algoritmaları İle Belirlenmesi, Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Samsun 2019.

Naseef, Ruslan Salim, Using Data Mining Techniques to Determine The Evidence of Infection with Coronavirus "Covid-19" in Iraq, Kirsehir Ahi Evran University Institute Of Science Master's Thesis, Kırşehir 2022.

Pandey, A., Hidden, K., Chandra, A., *K-Means Clustering An unsupervised learning algorithm*, Retrieved December1, 2022, from https://kmean.scrib.ink/

Pasin, O., & Pasin, T. (2020). Clustering of countries in terms of deaths and cases of COVID-19. *J Health Soc Sci*, 5(4), 587-94.

Rai, V. K., Chakraborty, S., & Chakraborty, S. (2023). Association rule mining for prediction of COVID-19. *Decision Making: Applications in Management and Engineering*, 6(1), 365-378.

Sarı, Tamer, Türkiye'deki İllerin Havadaki Partiküler Madde ve Kükürt Dioksit Değerlerine Göre Kümeleme Analizi İle Sınıflandırılması, Gazi Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Ankara 2023.

Shakhovska, N., Izonin, I., & Melnykova, N. (2021). The Hierarchical Classifier for COVID-19 *Resistance Evaluation*. Data, 6(1), 6.

Şchiopu, D. (2010). Applying TwoStep cluster analysis for identifying bank customers' profile. Buletinul, 62(3), 66-75.

Taşatan, Akif, Hiyerarşik Kümeleme Tekniklerinde Küme Eleman Sayısının Eşitlenmesine Yönelik Bir Yaklaşım Önerisi ve Gerçek Karayolu Uzaklık Verilerine Dayalı Kümeleme Analizi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Kocaeli 2018.

Takaoğlu, M., & Takaoğlu, F. (2019). K-Means ve Hiyerarşik Kümeleme Algoritmanın Weka ve Matlab Platformlarında Karşılaştırılması. *İstanbul Aydın Üniversitesi Dergisi*, 11(3), 303-317.

Tekin, İrem Sena, COVID-19'un Yayılma Sürecinde Yaşam Alışkanlıklarının Etkisinin Yapay Zeka ile Analizi, Tekirdağ Namık Kemal Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Tekirdağ 2023.

Yalçıner Çal, Damla, Kümeleme ve Birliktelik Kuralları Analizi ile Borsa İstanbul 100 Endeksinde Yer Alan Hisse Senetlerinin İncelenmesi, Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü Doktora Tezi, Isparta 2023.

Zhao, C. M., & Luan, J. (2006). Data Mining: Going beyond Traditional Statistics. *New Directions for Institutional Research*, 131, 7-16.

ETİK KURUL KARARI

SAĞLIK BİLİMLERİ ÜNİVERSİTESİ
SAMSUN EĞİTİM VE ARAŞTIRMA HASTANESİ
GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU
KARAR FORMU

ARAŞTIRMANIN AÇIK ADI	**İki Aşamalı Kümeleme Analizi ile COVID-19 Vakalarının Değerlendirilmesi**
ARAŞTIRMANIN PROTOKOL KODU	GOKA/2021/13/6
ARAŞTIRMANIN BAŞLAMA-BİTİŞ TARİHİ	01.07.2021/01.11.2021

BAŞVURU BİLGİLERİ	**KOORDİNATÖR/SORUMLU ARAŞTIRMACI UNVANI/ADI/SOYADI**	UZM.DR.MEHMET HAKAN TAŞKIN			
	KOORDİNATÖR/SORUMLU ARAŞTIRMACININ UZMANLIK ALANI	TIBBİ MİKROBİYOLOJİ			
	KOORDİNATÖR/SORUMLU ARAŞTIRMACININ BULUNDUĞU MERKEZ	SBÜ SAMSUN EĞİTİM VE ARAŞTIRMA HASTANESİ			
	ARAŞTIRMACI UNVANI/ADI/SOYADI	DR.ÖĞRT.ÜYESİ ASLI ÇALIŞ BOYACI			
	ARAŞTIRMACI UNVANI/ADI/SOYADI	ENDÜSTRİ MÜHENDİSİ BEYZA DURMAZ			
	ARAŞTIRMACI UNVANI/ADI/SOYADI	-			
	ARAŞTIRMACI UNVANI/ADI/SOYADI	-			
	ARAŞTIRMACI UNVANI/ADI/SOYADI	-			
	ARAŞTIRMACI UNVANI/ADI/SOYADI	-			
	ARAŞTIRMACI UNVANI/ADI/SOYADI	-			
	ARAŞTIRMAYA KATILAN MERKEZLER	TEK MERKEZ ☒	ÇOK MERKEZLİ ☐	ULUSAL ☐	ULUSLARARASI ☐

Etik Kurul Başkanının
Unvanı/Adı Soyadı: Doç. Dr. Mahcube ÇUBUKÇU
İmzası

SAĞLIK BİLİMLERİ ÜNİVERSİTESİ
SAMSUN EĞİTİM VE ARAŞTIRMA HASTANESİ
GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU
KARAR FORMU

ARAŞTIRMANIN AÇIK ADI	**İki Aşamalı Kümeleme Analizi ile COVID-19 Vakalarının Değerlendirilmesi**
ARAŞTIRMANIN PROTOKOL KODU	GOKA/2021/13/6
ARAŞTIRMANIN BAŞLAMA-BİTİŞ TARİHİ	01.07.2021/01.11.2021

	Belge Adı	Tarihi	Versiyon Numarası	Dili
Değerlendirilen Belgeler	ARAŞTIRMA PROTOKOLÜ/PLANI			Türkçe ☒ İngilizce ☐ Diğer ☐
	BİLGİLENDİRİLMİŞ GÖNÜLLÜ OLUR FORMU		☐	Türkçe ☒ İngilizce ☐ Diğer ☐
Karar Bilgileri	**Karar No: 2021/13/6**	**Tarih:07.07.2021**		
	Yukarıda bilgileri verilen Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu başvuru dosyası ile ilgili belgeler araştırmanın gerekçe, amaç, yaklaşım ve yöntemleri dikkate alınarak incelenmiş ve araştırmanın etik ve bilimsel yönden uygun olduğuna **"salt çoğunluğu"** ile karar verilmiştir.			

SAĞLIK BİLİMLERİ ÜNİVERSİTESİ SAMSUN EĞİTİM VE ARAŞTIRMA HASTANESİ GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU

BAŞKANIN UNVANI / ADI / SOYADI	Doç. Dr. Mahcube ÇUBUKÇU

Unvanı/Adı/Soyadı	Uzmanlık Alanı	Kurumu	Araştırma ile ilişki		Katılım *		İmza
Doç. Dr. Mahcube ÇUBUKÇU (Başkan)	Aile Hekimliği	SBÜ. Samsun EAH	E☐	H☒	E☐	H☐	
Doç. Dr. Yonca SEMET	Çocuk Nefrolojisi	SBÜ. Samsun EAH	E☐	H☒	E☐	H☐	
Doç. Dr. Murat GÜZEL	Acil Tıp	SBÜ. Samsun EAH	E☐	H☒	E☐	H☐	
Doç.Dr.Vaner KÖKSAL	Beyin ve Sinir Cerrahisi	SBÜ. Samsun EAH	E☐	H☒	E☐	H☐	
Dr.Öğrt.Üyesi Gökhan ÖZGÜR	Göz Hastalıkları	SBÜ. Samsun EAH	E☐	H☒	E☐	H☐	
Dr.Öğr.Üyesi Ahmet Burak ÇİFTÇİ	Genel Cerrahi	SBÜ. Samsun EAH	E☐	H☒	E☐	H☐	
Uzm.Dr. Melek BİLGİN	Tıbbi Mikrobiyoloji	SBÜ. Samsun EAH	E☐	H☒	E☐	H☐	
Uzm.Dr.Düriye Sıla KARAGÖZ ÖZEN	İç Hastalıkları	SBÜ. Samsun EAH	E☐	H☒	E☐	H☐	

* :Toplantıda Bulunma

ÖZ GEÇMİŞ

Beyza DURMAZ, Samsun Tülay Başaran Anadolu Lisesi'ni bitirdikten sonra Ondokuz Mayıs Üniversitesi Mühendislik Fakültesi, Endüstri Mühendisliği bölümünden 2019 yılında mezun oldu. 2020 yılında OMÜ LEE Akıllı Sistemler Mühendisliği Ana Bilim Dalı Yüksek Lisans programına girdi.

İletişim Bilgileri

ORCID ID : 0009-0001-4813-4054

Yayınlar:

1. İki Aşamalı Kümeleme Analizi İle Pandemi Döneminde Hasta Profilinin Değerlendirilmesi - 12. Uluslararası Mühendislik Mimarlık Ve Tasarım Kongresi, 23-25 Aralık, İstanbul / Türkiye

2.

Kazanılan Ödüller, Teşvikler ve Burslar

1.

2.