T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARI İLE KARDİYOVASKÜLER HASTALIKLARININ TANI VE TAHMİNLERİ

Ebru ÖZTÜRK

YÜKSEK LİSANS TEZİ

Matematik Anabilim Dalı

Matematik Programı

Danışman

Doç. Dr. Mutlu AKAR

Temmuz, 2024

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Makine Öğrenmesi Sınıflandırma Algoritmaları ile Kardiyovasküler Hastalıklarının Tanı ve Tahminleri

Ebru ÖZTÜRK tarafından hazırlanan tez çalışması 01.07.2024 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Anabilim Dalı, Matematik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Mutlu AKAR
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Mutlu AKAR, Danışman
Yıldız Teknik Üniversitesi _____

Doç. Dr. Muttalip ÖZAVŞAR, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Fatih TEMİZ, Üye
Üsküdar Üniversitesi _____

Danışmanım Doç. Dr. Mutlu AKAR sorumluluğunda tarafımca hazırlanan “Makine Öğrenmesi Sınıflandırma Algoritmaları ile Kardiyovasküler Hastalıklarının Tanı ve Tahminleri” başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Ebru ÖZTÜRK

Aileme

TEŞEKKÜR

Yüksek lisans ögrenimim boyunca olumlu tavırlarıyla bana yardımcı olan, tavsiye ve deneyimleriyle yol gösteren beraber çalışmaktan ve ayrıca her zaman öğrencisi olmaktan gurur duyduğum değerli danışman hocam Doç. Dr. Mutlu AKAR'a çok teşekkür ederim.

Sadece tez çalışması sırasında değil tüm hayatım boyunca benim yanımda olan ve benden desteklerini esirgemeyen ve beni cesaretlendiren anneme ve babama sonsuz şükranlarımı sunar ve teşekkür ederim.

Ebru ÖZTÜRK

İÇİNDEKİLER

SİMGE LİSTESİ

C	Hard margin ve soft margin arasındaki denge hiperparametresi
Z	Z skoru
μ	Ortalama
σ	Standart sapma
ϵ	Hata terimi

KISALTMA LİSTESİ

ACC	Accuracy (Doğruluk Oranı)
AUC	Area Under Curve (Eğrinin Altında Kalan Alan)
FN	False Negative (Yanlış Negatif Değer)
FP	False Positive (Yanlış Pozitif Değer)
FPR	False Positive Rate (Yanlış Pozitif Oranı)
HDL	High Density Lipoprotein (İyi Kolesterol)
KB	Kan Basıncı
KKH	Kanjenital Kardiyovasküler Hastalık
KNN	K-Nearest Neighbors (K-En Yakın Komşuluk)
KVH	Kardiyovasküler Hastalık
LDL	Low Density Lipoprotein (Kötü Kolesterol)
MCC	Matthews Correlation Coefficient (Matthews Korelasyon Katsayı)
MLP	Multi Layer Perceptron (Çok Katmanlı Yapay Sinir Ağları)
PCA	Principal Component Analysis (Temel Bileşenler Analizi)
PCC	Pearson Correlation Coefficient (Pearson Korelasyon Katsayısı)
PR	Precision (Kesinlik)
RE	Recall (Duyarlılık)
ROC	Receiver Operating Characteristic (Alıcı İşletim Eğrisi)
SE	Sensitivity (Duyarlılık)
SP	Specificity (Özgüllük)
SVM	Support Vector Machine (Destek Vektör Makineleri)
TN	True Negative (Gerçek Negatif Değer)
TP	True Positive (Gerçek Pozitif Değer)
TPR	True Positive Rate (Doğru Pozitif Oranı)
TÜİK	Türkiye İstatistik Kurumu
WHO	World Health Organization (Dünya Sağlık Örgütü)

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

Makine Öğrenmesi Sınıflandırma Algoritmaları ile Kardiyovasküler Hastalıklarının Tanı ve Tahminleri

Ebru ÖZTÜRK

Matematik Anabilim Dalı

Matematik Programı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Mutlu AKAR

Kardiyovasküler hastalık, dünyada en fazla ölüme ve sakatlığa neden olan bulaşıcı olmayan hastalık ölümlerinin %37 sini oluşturmaktadır. Ölüm miktarının büyük bir çoğunluğuna neden olan kardiyovasküler hastalıklar için erken tanı önemli hale gelmektedir. Teknolojinin gelişmesi ile birlikte günümüzde sağlık sektöründe veri miktarı artış göstermiştir. Sağlık sektöründeki bu artan veri miktarı ile makine öğrenmesi tekniklerini kullanarak verilerin sınıflandırılması ve tahmine dayanan analizlerle anlamlı bilgiler çıkarılması kardiyovasküler hastalıkların erken tanısı için önem arz etmektedir. Bu çalışmada 70.000 gerçek hasta verilerinden oluşturulmuş 11 özelliğe sahip veri setinde makine öğrenmesi teknikleri uygulanarak en başarılı sınıflandırma tahmini yapan algoritmaya ulaşıp hastalığın erken teşhisine katkıda bulunulması hedeflenmiştir. İlk olarak veri seti analiz edilerek PCA yöntemi uygulanmıştır. Hangi özelliklerin kardiyovasküler hastalıklara neden olabileceği tespit edilerek özelliklerin karşılaştırmalı olarak görselleştirilmesi sağlanmıştır. Özellik mühendisliği ile hastanın tansiyonu kategorize edilerek normal, yüksek tansiyon, optimal, hipertansiyon olarak

sınıflandırılmıştır. Daha sonrasında lojistik regresyon, karar ağacı, rastgele orman, SVM, KNN, XGBoost, gradient boosting, LightGBM, AdaBoost ve ExtraTree yöntemleri kullanılarak kardiyovasküler hastalık tespiti yapılmıştır. XGBoost yönteminin %96.60 doğruluk oranı ile diğer yöntemlere göre daha başarılı olduğu tespit edilmiştir. Bu yöntemi sırasıyla %91.21 doğruluk oranı ile LightGBM ve %90.26 ile ExtraTree takip etmiştir. Sınıflandırma işleminden sonra modelin doğruluğunun gerçeği yansıtıp yansıtmadığını anlamak için karışıklık matrisi kullanılmıştır. Sınıflandırma modelleri başarısını değerlendirme yöntemlerinden biri olan ROC eğrisi ile XGBoost yönteminin doğruluk oranı grafik ile yorumlanmıştır. ROC eğrisinin altında kalan alan olan AUC skoru hesaplanmıştır.

Anahtar Kelimeler: Makine öğrenmesi, kardiyovasküler hastalık, data analizi, sınıflandırma algoritmaları.

ABSTRACT

The Diagnosis and Predictions of Cardiovascular Diseases with Machine Learning Classification Algorithms

Ebru ÖZTÜRK

Department of Mathematics

Master of Science Thesis

Supervisor: Assoc. Dr. Mutlu AKAR

Cardiovascular disease constitutes 37% of non-communicable disease deaths, leading to the highest number of deaths and disabilities worldwide. Early diagnosis of cardiovascular diseases, which cause a large majority of deaths, has become important. With the advancement of technology, the amount of data in the health sector has increased today. The classification of data using machine learning techniques and the extraction of meaningful information through predictive analyses with this increasing amount of data in the health sector is important for the early diagnosis of cardiovascular diseases. This study aims to contribute to the early diagnosis of the disease by reaching the most successful classification prediction algorithm applied to a dataset with 11 features created from 70,000 real patient data. Initially, the dataset was analyzed, and the PCA method was applied. It has been determined which features may cause cardiovascular diseases, and the features have been visualized comparatively. With feature engineering, the patient's blood pressure was categorized as normal, high blood pressure, optimal, and hypertension. Subsequently, cardiovascular disease detection was performed using

logistic regression, decision tree, random forest, SVM, KNN, XGBoost, gradient boosting, LightGBM, AdaBoost, and ExtraTree methods. The XGBoost method was found to be more successful than other methods with an accuracy rate of 96.60%. This method was followed by LightGBM with an accuracy rate of 91.21% and ExtraTree with 90.26%. After the classification process, a confusion matrix was used to understand whether the model's accuracy reflects reality. The accuracy rate of the XGBoost method has been interpreted graphically with the ROC curve, one of the methods for evaluating the success of classification models. The AUC score, which is the area under the ROC curve, has been calculated.

Keywords: Machine learning, cardiovascular disease, data analysis, classification algorithms.

1
GİRİŞ

1.1 Literatür Özeti

Bu tezde kullanılan veri seti ilk olarak 2019 yılında Jaouja Maiga, Gilbert Gutabaga Hungilo ve Pranowo [1] yayınladıkları makalede ortaya çıkmıştır. Bu makalede kardiyovasküler hastalığın öngörülmesi için kardiyovasküler risk faktörleri kullanılarak makine öğrenmesi algoritmalarından olan rastgele orman, Naïve Bayes, KNN ve lojistik regresyon kullanılarak karşılaştırılması yapılmıştır. Çalışmada rastgele orman algoritması %73 ile en yüksek doğruluk oranına ulaşmıştır. Daha sonra 2021 yılında Rachael Hagan, Charles J. Gillan ve Fiona Mallett [2] kardiyovasküler hastalıkların sınıflandırılması için SVM, MLP ve bütünleştirme yöntemleri olmak üzere makine öğrenmesi yöntemlerini uygularken oluşan belirsizliği incelemişlerdir. Son olarak, 2022 yılında Muhammad Owais Butt ve çalışma arkadaşları [3] kalp yetmezliği hastalarının yaşam sürelerini tahmin etmek ve erken teşhis etmek amacıyla bir karar destek sistemi oluşturmuşlardır. Karar ağacı ve KNN ile %84,11 en yüksek doğruluk oranına ulaşmışlardır.

1.2 Tezin Amacı

Günümüzde gelişen teknoloji ile birlikte sağlık, telekominasyon, güvenlik, mobil, finans, pazarlama gibi birçok alanda yüksek hızda üretilen büyük miktarda karmaşık ve belirsiz veri toplanmaktadır [4]. Bu biriken verilerin çeşitli uygulama alanlarında daha doğru kararlar verebilmek için ileri analiz teknikleri uygulanarak kullanılması gerekmektedir. Veri seti üzerinde makine öğrenmesi, regresyon analizi, sınıflandırma gibi ileri analiz yöntemleri uygulanarak veri setini daha anlaşılır hale getirebiliriz. Veri setinden anlamlı bilgiler elde etmek için uygulanacak ileri analiz tekniklerine verinin türü ve boyutuna, ulaşmak istediğimiz sonuca göre karar vermeliyiz [5].

KVH dünyada en çok ölüme neden olan hastalıklardan biridir. KVH ölüm istatistikleri ülkelere göre değişiklik göstermektedir. En yüksek ölüm istatistikleri Afrika ve Doğu Avrupa'da, en düşük ölüm istatistikleri Avrupa Birliği ülkeleri ve Kuzey Amerikada'dır. WHO'nun 2021 yılı verilerine göre, tahmini olarak KVH nedeniyle hayatını kaybeden insan oranı tüm ölümlerin

%32'sini oluşturmaktadır. 2021 yılında dünyada 17,9 milyon insan KVH nedeniyle hayatını kaybetmiştir [6].

Türkiye'de de aynı şekilde KVH en önde gelen sağlık sorunlarındandır. TÜİK'in 2021 yılı verilerine göre, KVH nedeniyle hayatını kaybeden insanların tüm ölümlere oranı %40,4'dür. Bu oran Türkiye'de erkeklerde %45,1 iken, kadınlarda %35,9'dur [7].

KVH kalp ve kan damarlarını, dolaşım sistemini etkileyen birçok hastalığı kapsamaktadır. KVH'ın koroner kalp hastalığı, kalp yetmezliği, inflamatuar kalp hastalığı gibi türleri vardır. KVH konjenital ve sonradan ortaya çıkan hastalıklar olarak iki ayrı gruba ayrılır. KKH kalp ve damarların anormal şekilde oluşması ile doğum kusurları olarak sınıflandırılır. KVH'a yakalanma riskini kan basıncı kontrolü, obeziteyi engelleyerek ve sigara kullanımını durdurarak yarıya indirilebilir. KVH konusunda riskli kişilerin erken tespiti ve korunmaları bu sebepten çok büyük önem arz etmektedir [8].

Bu tezin amacı hangi özelliklerin kardiyovasküler hastalığa daha fazla neden olduğu karşılaştırmalı bir şekilde özellik mühendisliği de kullanılarak sınıflandırılıp görselleştirilmiştir ve makine öğrenmesi teknikleri uygulanarak en başarılı sınıflandırma tahmini yapan algoritmaya ulaşıp hastalığın erken teşhisine katkıda bulunulması hedeflenmiştir.

Çalışmanın ikinci bölümünde, makine öğrenmesi alt başlıklarla detaylandırılarak anlatılmıştır. Keşifçi veri analizi, veri ön işleme, aykırı değer analizi, makine öğrenmesi türleri, makine öğrenmesi modelleri ve başarı ölçme parametreleri incelenmiştir. Çalışmanın üçüncü bölümünde veri seti açıklanarak yapılan sınıflandırma anlatılmıştır. Çalışmanın son bölümünde elde edilen sonuçlar karşılaştırılmıştır ve ortaya çıkan analizler yorumlanmıştır.

1.3 Hipotez

Her geçen gün artan veri miktarı ile birlikte veri analizi ve makine öğrenmesi tekniklerini sağlık sektöründe kullanmamak kaçınılmazdır. Makine öğrenmesi yöntemlerini sağlık sektöründe kullanarak hastalıkların teşhisinin hızlanması ve daha doğru teşhis konulması, hasta takibinin daha kolay yapılabilmesi, kişiselleştirilmiş tedavi süreçleri uygulanması, tedavi süreçlerinde daha hızlı aksiyon alınabilmesi ve daha etkin ilaçların kullanılması sağlanmış olur.

2
MAKİNE ÖĞRENMESİ

2.1 Keşifçi Veri Analizi

Veriye ilk aşamada analiz etmemiz ve bu analizi yaparken analitik ve metodik bir yaklaşım şeklimiz olması gerekmektedir. Keşifçi veri analizi büyük veya küçük ölçekli bir veride ilk keşfin ve analizin yapılmasıdır. Veri setindeki genel yapıyı anlamamıza olanak sağlar. Keşifçi veri analizi edindiğimiz bilgiler ile uygun bir tahmin modeli seçebilir, veriyi anlayabilir, değişkenler arasında ilişkiler kurabilir ve verideki hataları tespit etmemize yardımcı olur. Veri setinde eksik değer ve aykırı değer olup olmadığını saptamamıza olanak sağlar. Aynı zamanda verileri görselleştirmede de kullanılır [10].

Keşifçi veri analizi, 3 kategoride incelenir:

2.1.1 Tek Değişken Analizi (Univariate Analysis)

Tek bir sütunla uğraşıldığından en basit analizlerdendir. Tek bir sütun ile analiz yapıldığından sütunun dağılımı ve genel özellikleri ön plana çıkar. Kullanılan yöntemler arasında frekans, ortalama, standard sapma ve varyans gibi metodlar vardır. Aykırı gözlem konusunda karara varmadan önce çok değişkenli olarak incelenmesi gerekir.

2.1.2 İki Değişken Analizi (Bivariate Analysis)

Bu analiz türünde sebep sonuç ilişkisine ulaşılmak istenir. İki değişken arasındaki ilişki sorgulanır. Değişkenlerin birbiri ile olan ilişkileri incelenir ve sürekli değişkenler ile kategorik değişkenlerin durumları gözlemlenir.

2.1.3 Çok Değişken Analizi (Multivariate Analysis)

Üçden fazla değişken analiz edilir. Değişkenlerin birarada sergilediği durumlara odaklanmak gerekir [9].

Değişken analizleri yöntemlerine göre de 2'ye ayrılır:

2.1.4 Grafikli Analiz

Verideki ilişkiyi görmek için grafikler kullanılır. Boyut olarak büyük verilerde nicel değerler anlamsız kalabilir. Grafikler ile verileri görselleştirmek çok daha faydalı olabilir. Grafikli

analizlerde kullanılan türlere örnek olarak Bar Plot, Box Plot, Scatter Plot, Histogram, Pie Chart, Pair Plot ve Heatmap verilebilir.

2.1.5 Grafiksiz Analiz

İstatistiki metodlar kullanarak, grafikler olmadan analiz yapılır. Örnek olarak mod, medyan gibi yöntemleri verebiliriz.

2.2 Veri Ön İşleme

Veri ön işleme, makine öğrenmesi yöntemlerini uygulamadan önce veriyi temizleme ve dönüştürme işlemidir. Veri ön işleme adımı gerektirmeyen veri yoktur. Veri ön işleme adımının yanlış uygulanması modellerin tahmin performansını etkileyerek yanlış analitik sonuçlara yol açar.

Gürültülü veya dağınık verileri istatiktiksel veya görselleştirme teknikleri ile belirledikten sonra veri setine uygun veri temizleme teknikleri kullanılmalıdır. Veri temizleme adımında dağınık verilerdeki hatalar düzeltilir. Örneğin, veride boy bilgileri tutuluyor olsun. Bir kişinin 5 metre boyu olamayacağından verinin hatalı olduğunu anlayabiliriz. Aşağıda veri temizleme teknikleri verilmiştir.

- Aykırı değerleri belirlemek için uygun yöntemler kullanmak.
- Tekrarlayan satırların tespit edilmesi ve gerekiyorsa kaldırılması.
- Eksik verilerin tespit edilmesi ve gerekiyorsa doldurulması.
- Tek bir değer içeren sütunların varyansı 0'dır. Bu sütunlar tek bir değer içerdiğinden dolayı modelleme de kullanmak yanlış sonuçlara neden olabilir.

Şekil 2.1'de veri temizleme adımları görülmektedir.

Şekil 2.1 Veri Temizleme Çemberi [11]

2.3 Aykırı Değer Analizi

Veriyi sağlıklı şekilde analiz etmemize engel olan en önemli etkenlerden bir taneside veri setindeki aykırı değerlerdir. Aykırı veya uç değer veri setindeki diğer değerlerden farklı davranan veya diğer verilere belirgin şekilde uzak olan gözlemlere denir. Aykırı değerlerin çok farklı sebeplerden kaynaklandığını görebiliriz. Bu sebepler aşağıdaki gibidir:

- Veri girişinde veya veri ölçümlerindeki hatalardan kaynaklanan aykırı değerler olabilir.
- Verideki bozulmalardan kaynaklanan aykırı değerler.
- Veri setindeki herhangi bir veride üst veya alt düzey bir performans aykırı değerlere neden olabilir.

Her aykırı veya uç değer hatadan kaynaklı değildir. Bazı durumlarda aykırı değerler veri setindeki yüksek varyanstan kaynaklanıyor olabilir.

Aykırı değer türleri nokta aykırı değer türleri, bağlamsal aykırı değerler ve toplu aykırı değerler olarak 3'e ayrılır. Nokta aykırı değer türünde bir verinin diğer geri kalan veri türlerinden oldukça uzakta yer almasıdır. Örneğin kilo ortalaması 50 iken, bir veride kilonun 100 olarak görülmesi. Bu verinin, veri setinin ortalamasından oldukça saptığını ve analiz sonuçlarını etkilediğini söyleyebiliriz. Bağlamsal aykırı değer türünde ise bir verinin belirli bir nedenle veri setindeki diğer gözlemlerden farklı durumda olmasıdır. Bu veriler, tek tek incelediğinde nokta aykırı değerler olarak görünmezler. Ancak koşul veya nedenlere bakıldığında veri setindeki diğer değerlere göre aykırı oldukları görülebilir. Örneğin mağazada diğer aynı kategorideki veri setinde normalden çok daha yüksek fiyatlı ürün. Toplu aykırı değerler türünde veri setindeki veriler kendi aralarında gruplandırıldığında, bir grubun kendi içinde normal ancak tüm veri seti ile kıyaslandığında aykırı olarak görünmesidir. Bu veriler de tek başlarına değerlendirildiğinde nokta aykırı değer olarak görünmezler ancak gruplandırıldığında veri setinin analizini etkileyebilirler. Örneğin, bir şirketteki ekipte performans notlarının diğer ekiplere göre çok daha yüksek olması.

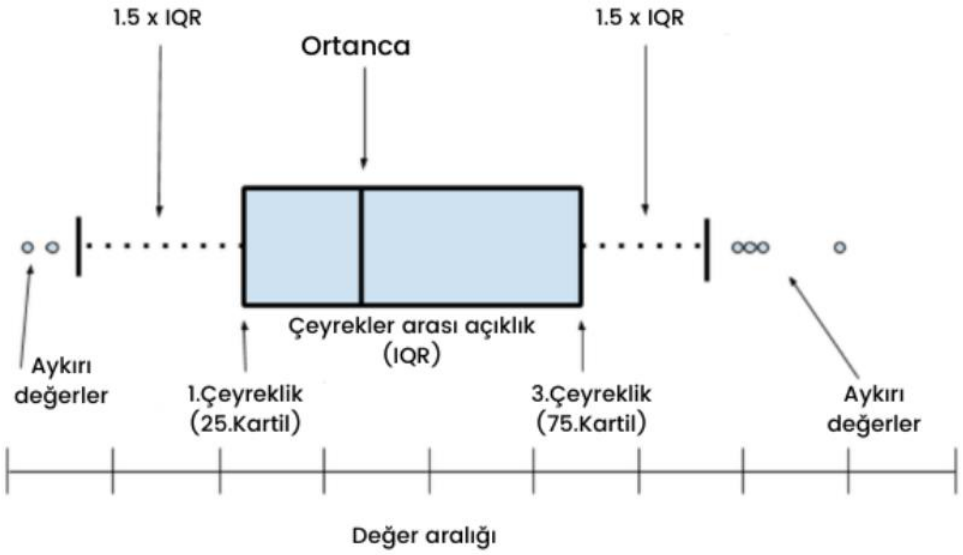
Aykırı değer analizinde kullanılacak yöntem seçimi, verideki değişken sayısına, verinin hacmine ve dağılımına bağlı olarak değişmektedir. Değişken sayısına göre tek değişkenli ve çok değişkenli yöntemler olarak ikiye ayrılır. Dağılımına göre ise parametrik yöntem ve istatistiksel veya grafiksel yöntemler olarak sınıflandırılır.

2.3.1 Grafiksel Tek Değişkenli Yöntemler

Bu yöntem bir veri setindeki tek değişkenli aykırı değerleri analiz etmek için kullanılır. Veri setindeki genel ortalamayı bozan uç değerleri analiz eder.

2.3.1.1 Kutu Grafiği (Box Plot)

Kutu grafiği veri setinin dağılımını çeyrek değerler üzerinden kutuda görselleştirerek gösteren bir grafik türüdür.

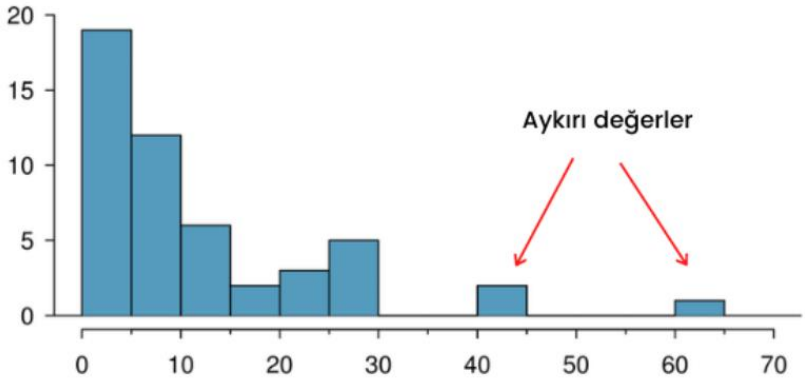


Şekil 2.2 Kutu grafiği yöntemi ile aykırı değer analizi [12]

Şekil 2.2'de görüldüğü üzere kutunun alt kenarı 1. çeyreklik, üst kenarı ise 3. çeyreklik değerlerini gösterir. 1. çeyreklik ve 3. çeyreklik arasındaki fark çeyrekler arası açıklığını bulmamızı sağlar. Aynı zamanda kutu grafiğinin ortasındaki çizgi ise ortanca değeri ifade eder. Alt ve üst kanca dışındaki değerler ise aykırı değerlerdir.

2.3.1.2 Histogram

Histogram grafiği, sayısal verilerin dağılımını etkili bir şekilde görmemizi sağlayan, veride değerlerin yayılımını gözlemleyebildiğimiz bir grafik türüdür. Bu grafik türünde, daha az bulunan ve uzakta kalan değerler ise aykırı değer olarak adlandırılır.



Şekil 2.3 Histogram grafiği yöntemi ile aykırı değer analizi [12]

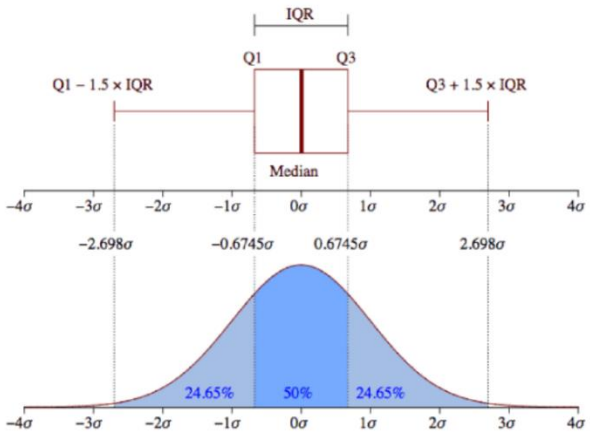
Şekil 2.3'de görüldüğü üzere aykırı değerler veri kümesinin normal dağılımını bozmuşlardır. Bu yöntemde doğrusal histogram grafiği dışında yoğunluk veya logaritmik histogram grafikleri de kullanılabilir.

2.3.2 İstatistiksel Tek Değişkenli Yöntemler

Aykırı değerleri sistematik bir şekilde belirlemek için istatistiksel yöntemler kullanılması gerekir.İstatistiksel yöntemlerin kullanılmasındaki genel amaç, gözlemleri tek tek incelemek yerine genel eğilimlerini ve dağılımlarını geniş bir çerçeveden görmektir. Bu yöntemler ile veriden daha doğru ve daha iyi anlaşılan değerler elde etmek mümkündür.

2.3.2.1 Çeyrekler Açıklığı (Interquartile Range-IQR)

Veri setindeki %75 ve %25'inci değerler arasındaki farka çeyrekler açıklığı denir. Kısaca ifade etmek gerekirse veri setinin orta yani %50'lik kısmına çeyrekler açıklığı denir.



Şekil 2.4 Çeyrekler açıklığı yöntemi ile aykırı değer analizi [12]

Şekil 2.4'de görüldüğü üzere çeyrekler açıklığını hesaplamak için önce veriler küçükten büyüğe doğru sıralanır. Ardından sıralanmış verilerde %25 ve %75'inci noktalar bulunur. Daha sonrasında bu noktalar arasındaki fark alınarak çeyrekler açıklığına ulaşılır. Çeyrekler açıklığının büyük olması veri setinin dağılımlı olduğunu, küçük olması ise veri setinin yoğun olduğunu gösterir. Alt ve üst sınırı belirlemek için 2.1 ve 2.2'deki denklemler kullanılır [12]. Bu alt ve üst sınırın dışında kalan değerlere aykırı değer denir.

$$Alt\ Sınır: Q_1 - 1.5 \times IQR \tag{2.1}$$

$$Üst\ Sınır: Q_3 + 1.5 \times IQR \tag{2.2}$$

Q_1: 1. Çeyreklik

Q_3: 3. Çeyreklik

IQR: Çeyrekler arası açıklık

2.3.2.2 Z-Skoru (Z-Score)

Z skoru, her bir gözlemin ortalamadan standart sapma uzaklığını gösteren bir yöntemdir. Z skorunu hesaplamak için 2.3'deki denklem kullanılır [12].

$$Z = (X - \mu)/\sigma \tag{2.3}$$

Z: Z skoru

X: Gözlem değeri

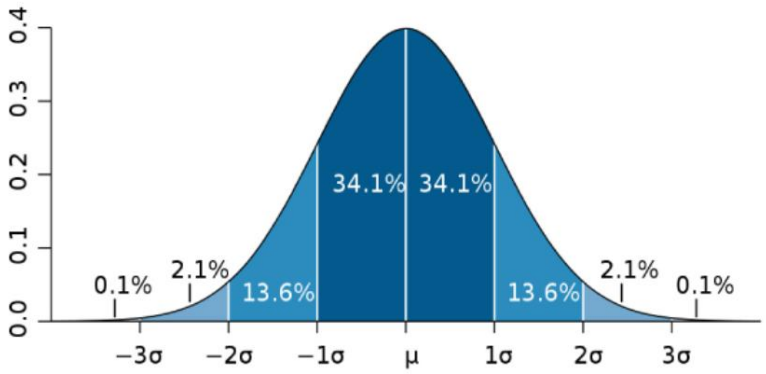
μ: Ortalama

σ: Standart sapma

Z skorunun yorumlanması aşağıdaki şekilde yapılabilir:

1. $|Z| < 2$: Veri setinin ortalama değere yakın olduğunu gösterir.
2. $2 \leq |Z| < 3$: Veri setinin ortalama değerden biraz uzak olduğunu gösterir.
3. $|Z| \geq 3$: Veri setinin ortalama değerden çok uzak olduğunu gösterir. Aykırı değer olarak kabul edilir.

Şekil 2.5'de görüldüğü üzere 3 standart sapma uzaklıkta olan veriler aykırı değer olarak kabul edilir. Veri setinin büyük olduğu durumlarda 4 standart sapma uzaklığı eşik değer, veri setinin az olduğu durumlarda ise 2 standart sapma eşik değer olarak kabul edilir.



Şekil 2.5 Z-skoru yönteminde normal dağılım eğrisi [12]

2.4 Makine Öğrenmesi Türleri

2.4.1 Denetimli Öğrenme

Verinin kategorisi ve temsil ettiği şeyler belirlenip etiketlendirilerek algoritmaların bu etiketler üzerinde eğitilmesi sağlanır. Daha sonrasında etiketli örneklerle eğitilen algoritma yardımıyla etiketsiz veriler için tahminler yapılır. Denetimli öğrenme yöntemine örnek olarak lineer regresyon, lojistik regresyon, CART, Naïve Bayes ve KNN verilebilir. Aynı zamanda toplama da bu öğrenme yönteminin türüdür. Toplamada, birçok makine öğrenmesi modeli tahmininden faydalanarak yeni bir model üzerinde tahminler yürütülür.

2.4.2 Denetimsiz Öğrenme

Verileri modellemek için etiketsiz örnekler kullanılır. Bu öğrenme türünde yalnızca girdi değişkeni vardır, girdi değişkenine karşılık gelen değişken yoktur. Algoritma verilerdeki ilişkileri bulmaya ve yeni veriler için tahminler çıkarmaya çalışır. Denetimsiz öğrenme algoritmaları birleştirme, kümeleme ve boyut azaltma olmak üzere 3 ana türe ayrılır. Birleştirmede herhangi bir analizde farklı ihtimallerin birarada olması durumu üzerinde çalışılır. Örneğin meyve sepetinde elma varsa %70 armut da alacaktır. Kümeleme de farklı kümelerdeki örneklere bakarak birbirine benzeyen örnekleri biraraya toplamadır. Boyut azaltma ise veri kümelerinin karmaşıklığını azaltmak için değişken sayısını azaltmaktır. Denetimsiz öğrenme yöntemine örnek olarak PCA verilebilir.

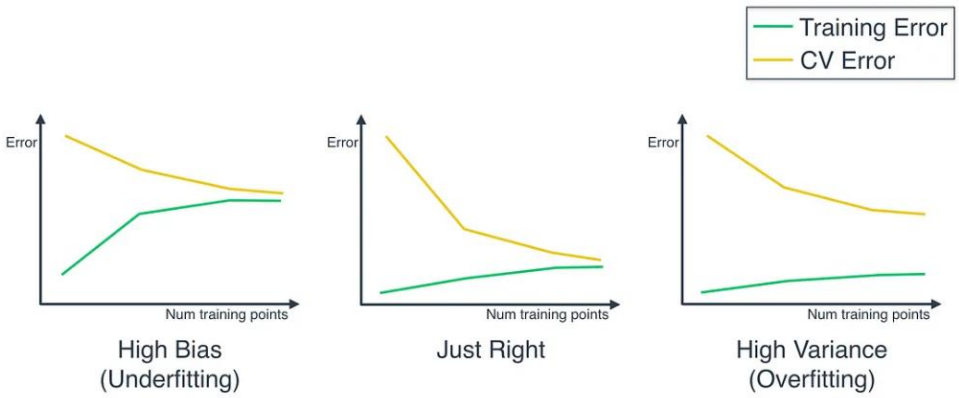
2.4.3 Yarı Denetimli Öğrenme

Hem etiketli hem etiketsiz veri kullanılarak modeler eğitilir. Genellikle etiketli veri daha az, etiketsiz veri daha çok olacak şekilde veriler seçilir. Bunun sebebi ise etiketli veriler kullanılarak modelin öğrenmesi gereken etmenler belirlenir ve etiketsiz veri kullanılarak da model sağlam hale getirilir. Yarı denetimli öğrenme türünde deneme yanılma yoluyla öğrenme sağlanır. Robotikte ve video oyunlarında kullanılır. Robot bir yere bir kere çarparak artık çarpışmalardan kaçması gerektiğini öğrenir. Video oyunlarında ise oyuncuların belirli hareketlerle ödüle ulaşması örnek verilebilir.

2.5 Makine Öğrenmesi Modelleri

Makine öğrenmesi, bilgisayarların detaylı olarak programlanmadan deneyimlerden ve eldeki verilerden öğrenerek gelişmelerini sağlar. Makine öğrenmesi modellerinde eğitilen modelin, yeni veriler üzerinde de yüksek doğruluklu sonuçlar üretmesi beklenir. Bu yeni verileri eğitmek için mevcuttaki verilerimizi test ve eğitim veri seti olarak ikiye ayırırız. Test veri setindeki veriler kullanılıp eğitilerek, eğitim veri seti üzerinde yüksek doğruluk oranına ulaşılıp ulaşılmadığına bakılır. Bu sebepten eğitim veri seti daha büyük bir alt küme olmalıdır. Buradaki amaç en az değişkenle en yüksek doğruluk oranına ulaşmaktır.

Aşırı öğrenme (overfitting) modelin eğitim verisini çok iyi öğrenmesi ancak yeni veriler üzerinde kötü sonuç sergilemesidir. Eğitim hatası düşük olmasına ragmen doğrulama hatası çok yüksektir. Modelin karmaşık olması veya eğitim verisinin az olması gibi sebepler aşırı öğrenmeye yol açabilir. Eksik öğrenme (underfitting) ise modelin eğitim verisini yeterince öğrenmemesi ve yeni veriler üzerinde kötü sonuç sergilemesidir. Eğitim verileri üzerinde düşük hata oranı olmasına rağmen test verileri üzerinde yüksek hata oranlarına sahiptir. Eğitim verisinin küçük ya da çeşitli olmaması ve modelin çok basit veya esnek olmaması eksik öğrenmeye yol açabilir. Şekil 2.6’da eksik ve aşırı öğrenmenin grafik üzerinde gösterimini inceleyebiliriz.



Şekil 2.6 Aşırı öğrenme ve eksik öğrenme [17]

2.5.1 Regresyon Analizi

İki ya da çoklu değişkenler arasındaki ilişkiyi anlamamıza yarayan istatistik metodudur. Tek bir değişken ile yapılan analizlere tek değişkenli regresyon, çok değişken kullanılan analizlere ise çok değişkenli regresyon denir. Matematiksel olarak regresyonun amacı x değişkenine bağlı olarak y değişkenini tahmin etmemize fayda sağlar.

Doğrusal regresyon analizi basit ve çoklu doğrusal regresyon olarak ikiye ayrılır. Basit regresyon, bağımlı değişken ve tek bir bağımsız değişken arasındaki bağlantıyı anlamamıza fayda sağlar. Çoklu doğrusal regresyon ise bağımlı değişken ve birden fazla bağımsız değişken arasındaki ilişkiyi anlamamıza yarar.

Basit doğrusal regresyon modeli [18]:

$$Y = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \qquad i = 1,2,\dots,n \tag{2.4}$$

Çoklu doğrusal regresyon modeli [18]:

$$Y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1,2,\dots,n \tag{2.5}$$

β_0, β_1: Regresyon katsayıları

ε: Hata terimi

X: Bağımsız değişken

Y: Bağımlı değişken

p: Bağımsız değişken sayısı

Gerçek değer ve tahmin modeli arasındaki farka hata denir. Aşağıdaki denklemdeki gibi hatayı hesaplayabiliriz.

Gerçek değer [18]:

$$y_i = b_0 + b_i x_i + e_i \tag{2.6}$$

Tahmin edilen değer [18]:

$$\hat{y}_i = b_0 + b_i x_i \tag{2.7}$$

Hata hesaplama [18]:

$$e_i = y_i - \hat{y}_i \tag{2.8}$$

$$e_i = y_i - b_0 - b_i x_i \tag{2.9}$$

Lojistik regresyon ise ikili sınıflandırmalarda bağımsız değişkenin herhangi bir kategoriye atanıp olasılığın tahmin edilmesidir. Yanıt değişkeni doğrudan modellenemez. Yapılan tahmin 0 ve 1 arasındadır. Amacı bağımlı değişken ve bağımsız değişken arasındaki ilişkiyi kurmaktır. Doğrusal regresyon analizi ile aralarındaki en önemli fark, doğrusal regresyon analizinde yanıt değişkeni süreklidir ancak lojistik regresyon analizinde yanıt değişkeni kesiklidir. Aynı zamanda doğrusal regresyon analizinde yanıt değişkeninin değeri tahmin edilmeye çalışılırken, lojistik regresyon analizinde ise yanıt değişkeninin alabileceği değerlerden yalnızca biri tahmin edilmeye çalışılır.

Lojistik regresyon modelinin denklemi aşağıdaki gibidir [18]:

$$E(Y \,/\, x) = \frac{e^{B_o + B_{1x}}}{1 + e^{\beta_0 + B_{1x}}} \qquad 0 \leq E(Y/x) \leq 1 \tag{2.10}$$

2.5.2 Sınıflandırma Yöntemleri

Sınıflandırma, verilerin belirlenmiş kategorilere veya sınıflara ayırmaktır. Sınıflandırmada verilere regresyondaki gibi sürekli değer atanmaz, bunun yerine sınıflar atanır. Veri test ve eğitim veri seti olarak ikiye ayrılır. Diğer verilerin eğitimi sırasında etiketli örnek veriler kullanılır. Bu eğitim verisinin çıktısı da sınıflandırmanın kuralına ulaşmamızı sağlar. Sınıflandırmadan sonra çalışmanın kalitesini ölçmek için hata oranı ölçülür. Hata oranını ölçmek için bir test veri seti gereklidir. Yanlış sınıflandırma gruplarına test hatası denir. Eğitim kümesinin hata oranına da eğitim hatası denir. Sınıflandırma yöntemlerinin performansı

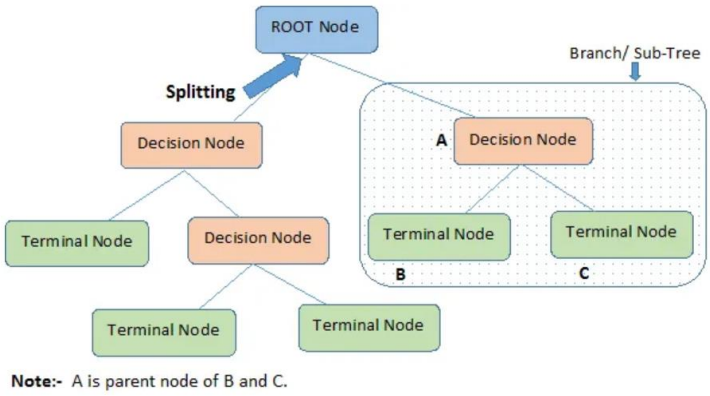
ölçülürken test sonuçları dikkate alınmaz. Ancak test ve eğitim seti arasındaki ilişki dikkate alınmalıdır.

2.5.2.1 Karar Ağacı

Hem regresyon hem de sınıflandırma algoritmalarını çözmeye yarayan denetimli öğrenme kategorisindeki bir yöntemdir. Ağaç temsili ile her yaprak yani düğüm bir sınıf etiketini temsil eder. Bu sınıflandırma algoritmasını kullanırken eğitim setinin tamamını kök olarak kabul ederiz.

Karar ağaçları hedef değişkenlerine göre kategorik değişken ve sürekli değişken karar ağaçları olarak ikiye ayrılır. Kategorik değişken karar ağacı, kategorik bir hedef değişkenine sahip olmalıdır. Örneğin bir ürünün fiyatını düşük, orta ve yüksek olarak tahmin etmek istiyorsak. Karar ağacı bu ürünün özelliklerine göre öğrenecek ve her verinin noktalarını düğümlerden geçirerek. Düşük, orta veya yüksek kategorik hedeflerinden birinin yaprak düğümüne ulaşacaktır. Sürekli değişken karar ağacı ise bir evin özelliklerini girdiğinde sürekli evin fiyatını tahmin etmemize olanacak tanıyacaktır.

Şekil 2.7'de görüldüğü üzere ağacın girdilerinin ilk alındığı kısma kök düğümü denir. Kök düğümdeki gözlemlerin koşullu şekilde ayrılmasına karar düğümleri denir. Tek bir düğüm birden fazla düğüme bölünebilir ve buna bölme denir. Eğer düğüm başka düğümlere bölünmezse yaprak veya terminal düğüm denir. Bölünmenin tam tersi budama olarak bilinir. Karar ağacı karmaşıklaşmış ve kaldırılması gereken karar kurallarının olmasına budama denir. Karar ağaçlarındaki karar düğümlerden ve terminal düğümlerden oluşan bir bölüme dal veya alt ağaç adı verilir. Şekil 2.7'de kutu içine alınmış kısım dal veya alt ağaca örnektir.



Şekil 2.7 Karar ağacı [20]

2.5.2.2 K-En Yakın Komşuluk

KNN algoritmasında benzer gözlemler birbirine yakın uzaklıktadır ve yeni gelen verilerde bu noktalara olan uzaklığı ile tahminlenir.

Büyük veri setlerinde kullanılırsa performansı iyi sonuç vermeyebilir. Çünkü yeni bir veri geldiğinde önceden hesaplanmış değerlere dayanır. Yani veriye özel bir hesaplama yapmaz ve eğitim verisindeki en yakın komşuların değerini kullanır. Bu yüzden tembel öğrenme algoritması (lazy learning) olarak bilinir.

KNN algoritmasında en yaygın kullanılan uzaklık hesaplama yöntemleri Öklid, Manhattan ve Minkowski uzaklığıdır. Öklid uzaklığı, pisagor bağlantısını kullanarak iki nokta arasında mesafe ölçümü bulunur. Aşağıdaki denklem kullanılarak Öklid uzaklığı hesaplanabilir [22].

$$D(x,y) = \sqrt{\sum_{i=1}^{\Pi} (x_i - yi)^2} \qquad i = 1,2,\dots,n \tag{2.11}$$

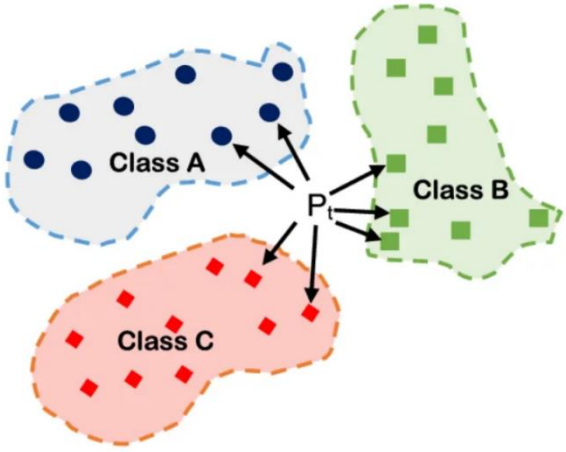
Manhattan uzaklığı, iki vektörün mutlak farklarının toplamını alarak hesaplanabilir. Aşağıdaki denklem uygulanarak Manhattan uzaklığı hesaplanabilir [22].

$$\sum_{i=1}^{k} |x_i - y_i| \tag{2.12}$$

Minkowski uzaklığı, Öklid uzayındaki bir metrik olarak tanımlanabilir. Öklid ve Manhattan uzaklığından oluşmuştur. p parametresine göre işlem yapılır. $p = 1$ ise Minkowski uzaklığı Manhattan uzaklığına eşittir, $p = 2$ ise Minkowski uzaklığı Öklid uzaklığına eşittir. Aşağıdaki denklem ile Minkowski uzaklığı hesaplanabilir [22].

$$\sum_{i=1}^{n} (|x_i - yi|^p)^{1/P} \tag{2.13}$$

Şekil 2.8'de $n = 8$ adet gözlem için P_t gözleminin sınıfı belirlenir.

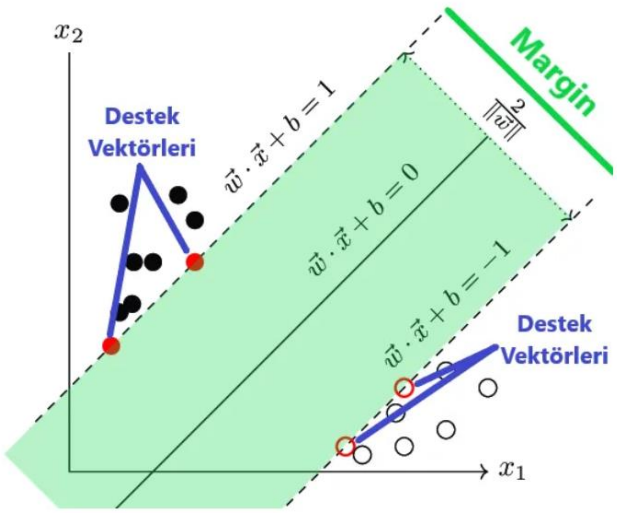


Şekil 2.8 KNN [22]

2.5.2.3 Destek Vektör Makinesi

SVM düzlem üzerinde noktaları ayırmak için doğru çizer ve bu doğru da destek vektörlerinden maksimum uzaklıkta olmalıdır. Maksimize etmemizin sebebi ise yeni gelen değerlerin doğru sınıf kategorisinde sınıflandırılmasını amaçlamaktır. Margin hiperdüzlemde sınıflar arasındaki boşluğu ifade eder. En geniş margin sınıflandırmanın doğruluğu ve genelliğini arttırır. SVM'nin temel amacı optimize olan hiperdüzlemi bulmaktır. Gözetimli öğrenme yöntemlerindendir.

Şekil 2.9'da görüldüğü üzere beyaz ve siyah olmak üzere iki veri grubu var. Sınıflandırmayı yapmamız için iki veri grubunu ayıran bir doğru çizilir ve ± 1 arasındaki yeşil bölgeye margin adı verilir. Marginin geniş olması sınıflandırmanın o kadar iyi yapıldığını ifade eder.



Şekil 2.9 SVM [23]

Aşağıdaki denklem ile hesaplama sağlanabilir. Yeni bir veri noktası geldiğinde çıkan sonuç 0'dan küçükse Şekil 2.9'daki beyaz noktalara, sonuç 0'a eşit veya büyükse siyah noktalara daha yakın olacaktır [24].

$$\hat{y} = \begin{cases} 0 \;\; if \; w^T \cdot x + b < 0, \\ 1 \;\; if \; w^T \cdot x + b \geq 0 \end{cases} \tag{2.14}$$

w: Ağırlık vektörü

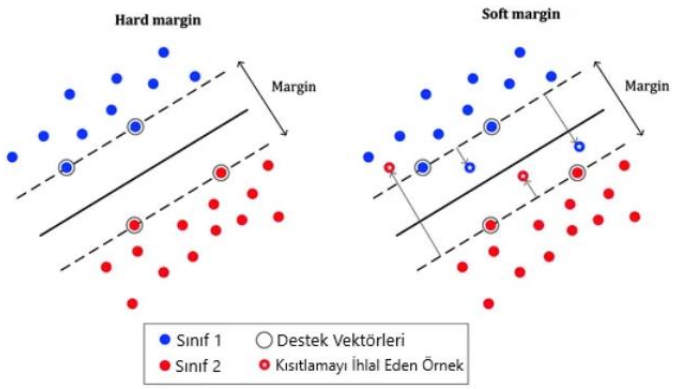
x: Girdi vektörü

b:Sapma

İki marjin hiperdüzlem arası uzaklık aşağıdaki denklem ile hesaplanır [25].

$$Marjin = \frac{2}{\|w\|} \tag{2.15}$$

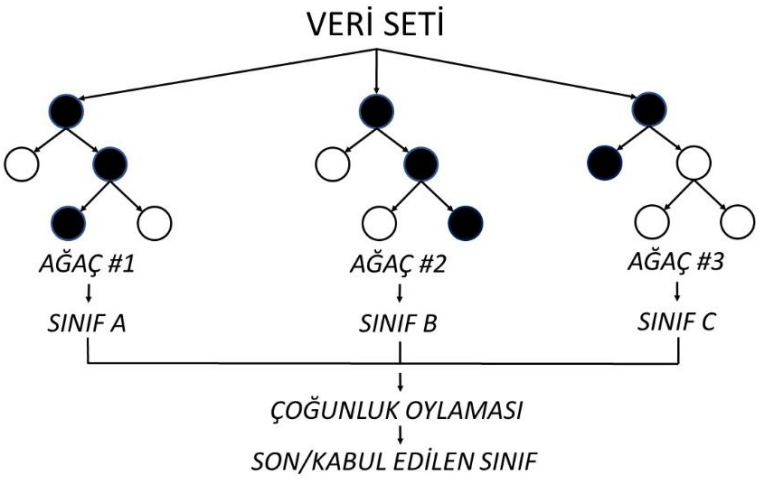
Veri noktalarımızın margin bölgesine girmesine soft margin denir. Hard margin aykırı değerlere karşı duyarlı olduğundan bazen soft margini tercih etmemiz gerekebilir. Hard margin ve soft margin arasındaki dengeyi C hiperparametresi ile kontrol ederiz. C büyüdükçe margin daralır [23]. Şekil 2.10'da soft marginde veri noktalarımızın margin bölgesine girdiğini görebiliriz.



Şekil 2.10 Hard margin ve soft margin [23]

2.5.2.4 Rastgele Orman

Rastgele bir orman yaratarak ağaç sayısı ve sonuç arasında ilişki kurar. Ağaç sayısı arttıkça doğruluk artar. Denetimli sınıflandırma algoritmasıdır. Hem regresyon hem de sınıflandırma da kullanılır. Algoritmanın en önemli özelliği yeterli ağaç bulunursa aşırı öğrenmenin ortaya çıkma ihtimali zayıftır.



Şekil 2.11 Rastgele orman [27]

Şekil 2.11'de görüldüğü üzere birçok karar ağacı üzerinde hepsini farklı gözlemlerle eğitir ve farklı modeller üretir. Birden fazla model oluşturarak verinin daha iyi şekilde analiz edilmesini sağlar. Algoritmada veri seti analize hazırlanır. Her değer için Şekil 2.11'de görüldüğü üzere karar ağacı oluşturulur ve karar ağaçlarından bir tahmin elde edilir. Tahmin sonucunda oluşan sınıflarda sınıflandırma algoritmaları için mod alınarak, regresyon algoritmaları için ortalama alınarak oylama gerçekleşir.

Karar ağacı algoritması ile aralarındaki fark ise rastgele orman algoritmasında kök düğümü bulmak ve bu düğümleri bölmek rastgele olur [27].

2.5.2.5 Gradient Boosting

Gradient Boosting algoritması, zayıf bir öğrenciyi güçlü bir öğrenciye dönüştürmedir. Bu dönüştürme işlemini iterasyonlar ile aşamalı olarak yapılmasını sağlar. Bu algoritmada ilk olarak yaprak (initial leaf) oluşur. Daha sonra tahmin yapılırkenki hatalar dikkate alınarak yeni ağaçlar oluşturulur. Oluşturulan modelde gelişme olmamasına kadar devam eder [28].

Algoritmada ilk adım olarak kayıp fonksiyon tanımlanmalıdır. Veriye uymada modelin katsayılarının iyi olup olmadığını gösteren ölçüye kayıp fonksiyon denir. Aşağıdaki fonksiyon ile kayıp fonksiyon hesaplanabilir [28].

$$L\big(y_i, F(x)\big) \tag{2.16}$$

y_i: Gözlemlenen Değer

$F(x)$: Tahmin Edilen Değer

İkinci adımda sabit değişken belirlenmelidir. Sabit değişkeni aşağıdaki formül ile belirlenebilir [29].

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L\,(y_i, \gamma) \tag{2.17}$$

$L(y_i, \gamma)$: Kayıp Fonksiyon

y_i: Gözlemlenen Değer

γ: Tahminlenen Değer

Kayıp fonksiyon toplanmalı ve minimum hali bütün gözlemler için bulunmalıdır. Bunu hesaplamak için kayıp fonksiyonun türevi alınır. Ardından değerler toplanır ve sıfıra eşitlenir. Bu hesaplamalar sonucunda yaprak değerini buluruz. Hedef değişkenindeki tüm değerlerin ortalamasına yaprak eşittir [32].

Son aşamada ise tüm ağaçlarda bir döngü uygulanır. Tahmine göre hatalar hesaplanır. Aşağıdaki formülde görüldüğü üzere parantez içerisindeki değer türevi alınmış kayıp fonksiyondur. Parantez dışındaki $F(x)$ ise önceki ağacın çıktısıdır. Yapılan tüm gözlemler için kalıntı hesaplamaları yapılır [29].

$$r_{im} = -\left[\frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad i = 1, \dots, n \tag{2.18}$$

r: Residual

i: Gözlem Numarası

m: Kurulan Ağacın Numarası

Karar ağacı residuallar için oluşturulur ve herbir yaprağın değeri bulunur. Aşağıdaki formül her yaprak için hatayı minimize eder. Kayıp fonksiyonun aynı şekilde türevi alınır ve ardından değerler toplanıp sıfıra eşitlenir. Sonuç ise yaprağın değeri olur. Herbir gözlem için tahmin oluşturulur [29].

$$\varphi_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{ij}} L\,(y_i, F_{m-1}(x_i) + \gamma) \;\; j = 1 \dots J_m \tag{2.19}$$

Aşağıdaki formülde görüldüğü üzere $F(x) =$ Önceki ağacın sonuçları + learning rate * yeni ağaç olduğunu gözlemleriz. Döngü aynı şekilde devam eder [29].

$$F_m(x) = F_{m-1}(x) + \zeta \sum_{j=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right) \tag{2.20}$$

2.5.2.6 XGBoost

XGBoost (eXtreme Gradient Boosting), Gradient Boosting algoritmasından üretilmiş daha yüksek performanslı halidir. Tianqi Chen ve Carlos Guestrin'in 2016 yılında yayınladıkları "XGBoost: A Scalable Tree Boosting System" makalesi ile XGBoost algoritması keşfedilmiştir. Tianqi'ye göre XGBoost diğer algoritmalara göre 10 kat daha hızlı çalışır. Algoritmanın yüksek tahmin yeteneği, aşırı öğrenmeye neden olmaması ve tüm bunları hızlı yapması en önemli özellikleridir [33].

	housing_median_age	households	median_house_value	median_income	initial_leaf	residual1
9491	22.0	285.0	118800.0	3.5917	0.5	3.0917
11843	15.0	316.0	118800.0	2.9559	0.5	2.4559
11271	21.0	683.0	213300.0	3.2857	0.5	2.7857
19219	19.0	433.0	190300.0	3.0568	0.5	2.5568
14356	11.0	1105.0	159800.0	3.3456	0.5	2.8456

Şekil 2.12 İlk tahminden sonra veri seti [33]

Çalışması Gradient Boosting ile benzerdir. Birinci adımı sonraki adımlarda yakınsama işlemi ile sonuca ulaşacağı için ilk tahmini (base score) 0.5 varsayılan olarak seçilir. Tahminin iyi veya kötü olup olmadığı hatalı tahmin (residual) ile bulunur. Şekil 2.12'de görülen kolonlardan median_income tahminlenmek istenen değişken, initial leaf ilk tahmin, residual1 tahmin hatası olarak değerlendirilir.

Daha sonrasında karar ağacı kurulur. Karar ağacında verilerin dallarda iyi gruplanıp gruplanmadığını anlamak için benzerlik skoru aşağıdaki denklem ile hesaplanır [30].

$$Benzerlik\ skoru = \frac{Kalanların\ toplamının\ karesi}{Kalanların\ sayısı + \lambda} \tag{2.21}$$

λ: Regülarizasyon parametresi

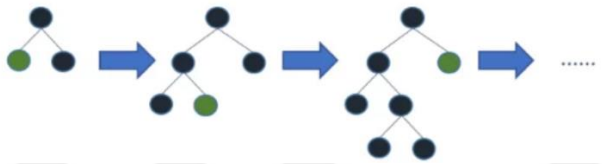
Bir sonraki aşamada, birçok ağaç kurulup hepsi için benzerlik skoru hesaplanır. Ağaçlardan doğru sonuca yaklaşanını bulmak için kazanç hesaplanır. Özetle dallar benzerlik skoru ile, ağaçlar ise kazançla değerlendirilir. Yüksek kazanç skoru olan ağaç kullanılır. Bu ağaç üzerinde

varsayılan gama değeri 0 seçilerek budama işlemi yapılır. Kazanç gamadan küçükse budanır. Budama yapmanın sebebi ise aşırı öğrenmeyi engellemektir [30].

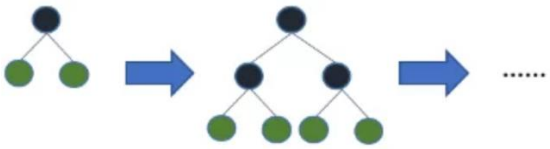
Benzerlik skorunda gördüğümüz lambda değeri modeli hassaslaştırır. Denklem 2.21‘e göre, benzerlik skoru dolaylı olarak da kazanç lambda artınca düşer. Lambda değeri ise varsayılan olarak 1’dir. Aşırı öğrenme lambda arttıkça azalır. Bu işlemler ile ilk ağaç oluşmuş olur. İkinci tahminlemeye geçilir. XGBoost algoritmasında dallar varsayılan olarak 6 kez bölünür.

2.5.2.7 LightGBM

LightGBM (Light Gradient Boosting Machine), yaprak odaklı (leaf-wise) ve seviye odaklı (level-wise veya depth-wise) olmak üzere iki yönteme ayrılır. Ağaç genişledikçe dengenin korunmasına seviye odaklı denir. Yaprak odaklı yöntemde ise hızlı öğrenme ve az hata oranına sahiptir. LightGBM, genellikle büyük verilerde kullanılmaya uygundur. Çünkü yaprak odaklı büyümede veri sayısı az ise aşırı öğrenmeye yatkındır. Şekil 2.13 ve Şekil 2.14’de yaprak ve seviye odaklı yöntemlerin ağaç yapılarını inceleriz.



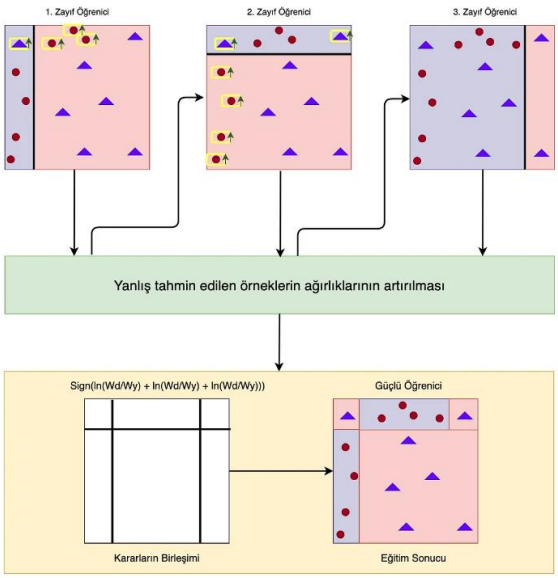
Şekil 2.13 Yaprak odaklı büyüme [37]



Şekil 2.14 Seviye odaklı büyüme [37]

2.5.2.8 AdaBoost

AdaBoost, çok sayıda zayıf verileri ardışık olarak birleştirerek güçlü bir veri oluşturur. Yanlış tahminlenen hatalı veriler her iterasyonda öncelik verilerek eğitilir. Son iterasyon adımında ise sonuçlar birleştirilip karar sınırları oluşturulur. Bu sınıflandırma yönteminde sıralama önemlidir. Şekil 2.15’de sınıflandırma yönteminin çalışma mantığını ayrıntılı olarak görebiliriz.



Şekil 2.15 AdaBoost yönteminin çalışma mantığı [38]

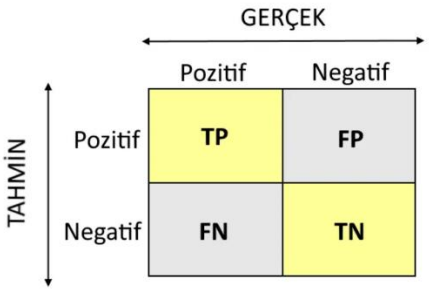
2.5.2.9 Extra Trees

Extra (Extremely Randomized) Trees, rastgele ağaçlar oluşturarak toplu öğrenme yöntemi ile ağaç tahminlerini birleştirip karar tahminini elde eder. Aşırı öğrenmeyi önlemek için rastgele bölme kullanılır. Büyük verilerde rastgele orman yöntemine göre doğruluk oranı düşüktür.

2.6 Başarı Ölçme Parametreleri

2.6.1 Karışıklık Matrisi

Kullanılan sınıflandırma modellerinin performansını değerlendirmemize yardımcı olur. Gerçek değer ve tahminler hata matrisi sayesinde karşılaştırılır. Matriste 1 değeri true, 0 değeri ise false olarak tanımlıdır. İkili sınıflandırma modelleri için karışıklık matrisi (Confusion Matrix) Şekil 2.6'daki gibi görülmektedir.



Şekil 2.16 Karışıklık matrisi [14]

TP: Gerçek değer 1 ve tahmin edilen değer de 1'dir. Örneğin hasta birine hasta denmiştir.

TN: Gerçek değer 0 ve tahmin edilen değer de 0'dır. Örneğin hasta olmayan birine hasta değil denmiştir.

FP: Gerçek değer 0 ve tahmin edilen değer 1'dir. Örneğin hasta olmayan birine hasta denmiştir.

FN: Gerçek değer 1 ve tahmin edilen değer 0'dır. Örneğin hasta olan birine hasta değil denmiştir.

2.6.2 ACC

ACC, karışıklık matrisindeki doğru tahmin toplamının doğru ve yanlış tahminlerin toplamına oranıdır. Doğru sınıflandırılan verilerin yüzdesini ifade eder. ACC, 0 ve 1 arasındadır. 0 en kötü tahmin, 1 ise en iyi tahmini ifade eder. ACC'nin denklemi aşağıdaki şekildedir [16]:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.22}$$

2.6.3 SE

SE, doğru tahmin edilen pozitif değerlerin gerçek değeri 1 olan verilere oranıdır. Modelin kesinliğini ölçer. RE ile aynı anlama gelir. SE aşağıdaki şekilde hesaplanır [16]:

$$SE = \frac{TP}{TP + FN} \tag{2.23}$$

2.6.4 SP

SP, doğru tahmin edilen negative değerlerin gerçek değeri 0 olan verilere oranıdır. Modelin iyi ayırt edilebildiğini gösterir. SP aşağıdaki şekilde hesaplanır [16]:

$$SP = \frac{TN}{TN + FP} \tag{2.24}$$

2.6.5 PR

PR, doğru tahmin edilen pozitif değerlerin tahmin edilen değeri 1 olan verilere oranıdır. Modelin kesinliğini gösterir. PR aşağıdaki şekilde hesaplanır [16]:

$$PR = \frac{TP}{TP + FP} \tag{2.25}$$

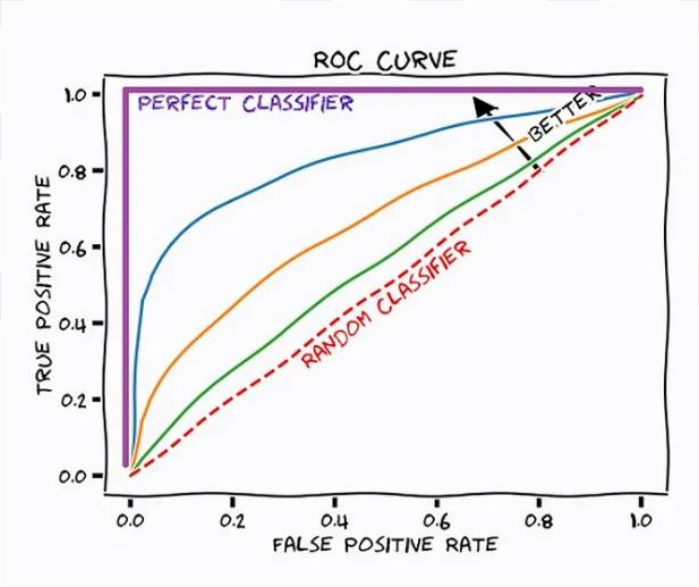
2.6.6 F_1 Skoru

Doğru tahmin edilen pozitif değerlerin gerçek ve tahmin edilen değeri 1 olan verilere oranını ifade eden recall ve hassasiyetin harmonik ortalamasıdır. Modelin performansının ne kadar iyi olduğunu ifade eder. F_1 skoru aşağıdaki şekilde hesaplanır [16]:

$$F_1 = 2 \times \frac{PR \times RE}{PR + RE} \tag{2.26}$$

2.6.7 ROC AUC

ROC eğrisi, olasılıkları gösteren bir grafiktir. Şekil 2.7'de X ekseninde FPR, Y ekseninde TPR görülmektedir. Karışıklık matrisinde FP verilerin negatif verilere oranını FPR gösterir. Aynı şekilde karışıklık matrisinde yer alan TP verilerin pozitif verilere oranını TPR gösterir.



Şekil 2.17 ROC eğrisi [16]

TPR, SE değerine eşittir. FPR, yanlışlıkla pozitif olarak sınıflandırılma yapılan örneklerdir. TPR ve FPR aşağıdaki şekilde hesaplanır [15]:

$$TPR = SE \tag{2.27}$$

$$FPR = 1 - SP \tag{2.28}$$

AUC ise eğrinin altında kalan alanı ifade eder. Bu alanın büyük olması modelin başarılı olduğunu, küçük olması ise modelin başarısız olduğunu ifade eder.

3
UYGULAMA VE ANALİZ

3.1 Veri Seti

Veri seti 2019 senesinde Kaggle adlı sitede Svetlana Ulianova adlı kullanıcı tarafından yayımlanmıştır [63]. Tüm değerler tıbbi muayene esnasında toplanmıştır. Veri seti 44.637 kadın, 23.898 erkek olmak üzere toplam 70.000 hastadan oluşmaktadır. Verilerin 11 özellikten oluşturulduğu görülmektedir ve özellikler gerçek bilgiler, muayene bilgileri, hasta tarafından verilen bilgiler olmak üzere 3 tür giriş özelliği taşır. Bu özellikler hastaların yaşı, boyu, kilosu, cinsiyeti, sistolik ve diyastolik kan basıncı, kolesterol, glikoz, sigara ve alkol kullanıp kullanılmama durumu, fiziksel aktivite gibi bilgilerdir.

Hastaların yaşı 35 ve 70 arasında değişmektedir ve yaş ortalamaları 53,3386'dır. Hastaların kanında sistolik kan basıncı, diyastolik kan basıncı, kolesterol, glikoz değerlerine bakılmıştır. Bu değişkenler veri setinde medikal özellikler olarak verilmiştir. Sigara, alkol, fiziksel aktivite veri setinde hasta tarafından verilen sübjektif özelliklerdir. Bu değişkenlere ek olarak, yaş, boy, kilo, cinsiyet olarak verilen değerler de gerçek özellikler olarak verilmiştir. Yaş değişkeni gün olarak verilmiştir.

Veri setinde eksik değer bulunmamaktadır. Verideki değişkenleri kullanarak hastaların KVH'a sahip olup olmama durumu doğru modellenmeyi amaçlıyoruz. Veri setinde Presence or absence of cardiovascular disease=1 KVH'a sahip olan hastayı, Presence or absence of cardiovascular disease=0 KVH'a sahip olmayan hastayı ifade etmektedir. Bu bilgiler doğrultusunda 55.065 KVH sahip, 13.470 KVH'a sahip olmayan hasta bulunmaktadır.

Sistolik kan basıncı, büyük tansiyon olarak da bilinir. Kalbin kasıldığı ve kan pompaladığı zaman damarlarda en yüksek kan basıncı değeri olarak tanımlanır. Sistolik kan basıncının yetişkinlerde normal değerleri 120-130 mmHg olarak ölçülür. Sistolik kan basıncı yüksekliği (hipertansiyon) kalp hastalığı, felç gibi birçok soruna yol açabilir [45].

Diyastolik kan basıncı, küçük tansiyon olarak da bilinir. Kalbin gevşediği zaman damarlardaki en düşük kan basıncı değeri olarak tanımlanır. Diyastolik kan basıncının yetişkinlerde normal değeri 70-80 mmHg olarak ölçülür. Diyastolik kan basıncı yüksekliği (hipertansiyon) de sistolik kan basıncı gibi kalp hastalığı, felç gibi sorunlara neden olabilir [46].

Kolesterol, karaciğerde üretilen bir yağ türüdür. Kötü ve iyi kolesterol olmak üzere ikiye ayrılır. LDL (kötü) kolesterol, kandan dokulara taşınır ve damarlarda birikip tıkanmaya neden olur. HDL (iyi) kolesterol, fazla kolesterolü dokulardan karaciğere taşır ve vücuttan atılmasına yardımcı olur. Yetişkin bireylerde normal kolesterol değeri, 170 mg/dL'den az olmalıdır [43].

Glikoz, kan şekeri olarak bilinir. Hücreler tarafından enerji tüketimi için kullanılır. Yetişkinler için normal kan şekeri açlıkta 70-100 mg/dL, toklukta ise 140 mg/dL'den azdır [44].

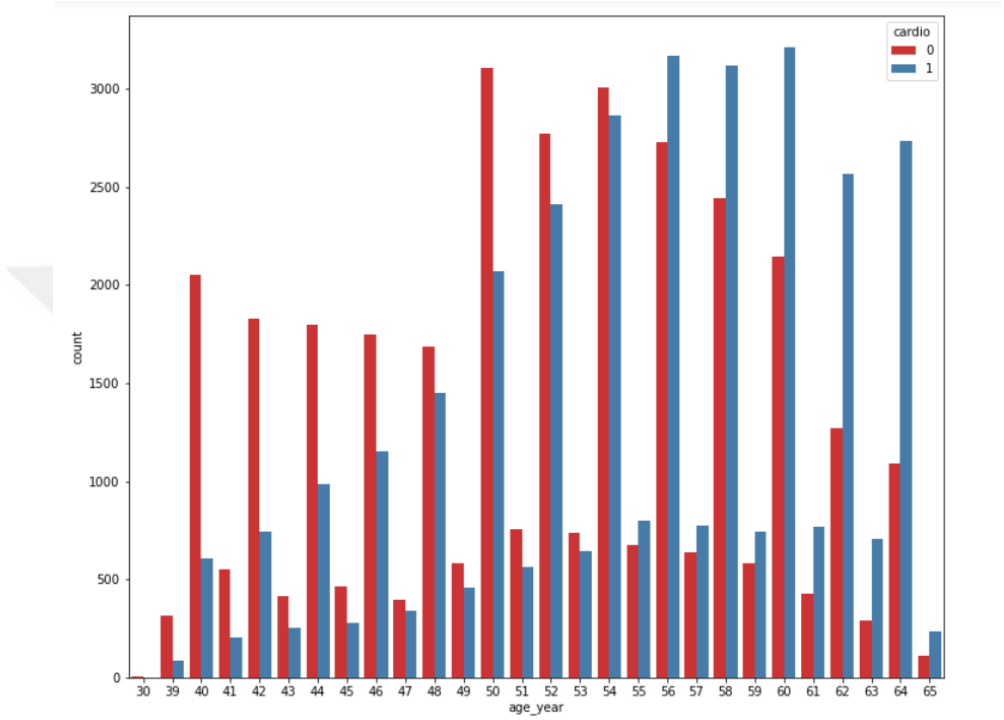
Tablo 3.1 Sübjektif özelliklerin hastalar üzerinde dağılımı

Sübjektif Özellikler	KVH'a Sahip Hasta Sayısı	KVH'a Sahip Olmayan Hasta Sayısı
Sigara Kullanılması (1: Var)	5035	990
Sigara Kullanılması (0: Yok)	50030	12480
Alkol Kullanılması (1: Var)	3097	570
Alkol Kullanılması (0: Yok)	51968	12900
Fiziksel Aktivite (1: Var)	26744	7186
Fiziksel Aktivite (0: Yok)	28321	6284
Cinsiyet (1: Kadın)	36516	8121
Cinsiyet (2: Erkek)	19745	4153
Kolesterol (1: Normal)	42032	10353
Kolesterol (2: Normalin Üstü)	7630	1919
Kolesterol (3: Normalin Çok Üstü)	6599	1467
Glikoz (1: Normal)	47895	11584
Glikoz (2: Normalin Üstü)	4099	1091
Glikoz (3: Normalin Çok Üstü)	4267	1064
Toplam Hasta Sayısı	55065	13470

Tablo 3.1'de gözlemlediğimiz üzere toplam KVH'a sahip hasta sayısı 55.065, KVH' a sahip olmayan hasta sayısı 13.470'dür. Bu sebepten dolayı veri setinin dengesiz olduğunu söyleyebiliriz. Tablo 3.2'de medikal ve gerçek özelliklerin KVH'a sahip olan, KVH'a sahip olmayan ve toplam hastalara göre ortalamaları görülmektedir.

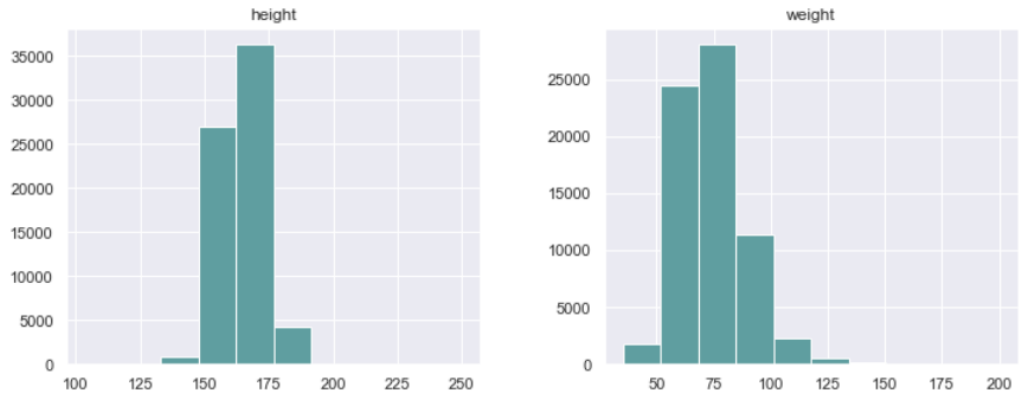
Tablo 3.2 Medikal ve gerçek özelliklerin ortalamasının hastalar üzerinde dağılımı

Medikal ve Gerçek Özellikler	KVH'a Sahip Hastaların Ortalaması	KVH'a Sahip Olmayan Hastaların Ortalaması	Tüm Hastaların Ortalaması
Yaş	53.3058	53.4731	53.33
Sistolik Kan Basıncı	128.814	128.827	128.817
Diyastolik Kan Basıncı	97.0756	94.8072	96.6304



Şekil 3.1 Yaşa göre fiziksel aktivite yapıp yapmama durumu

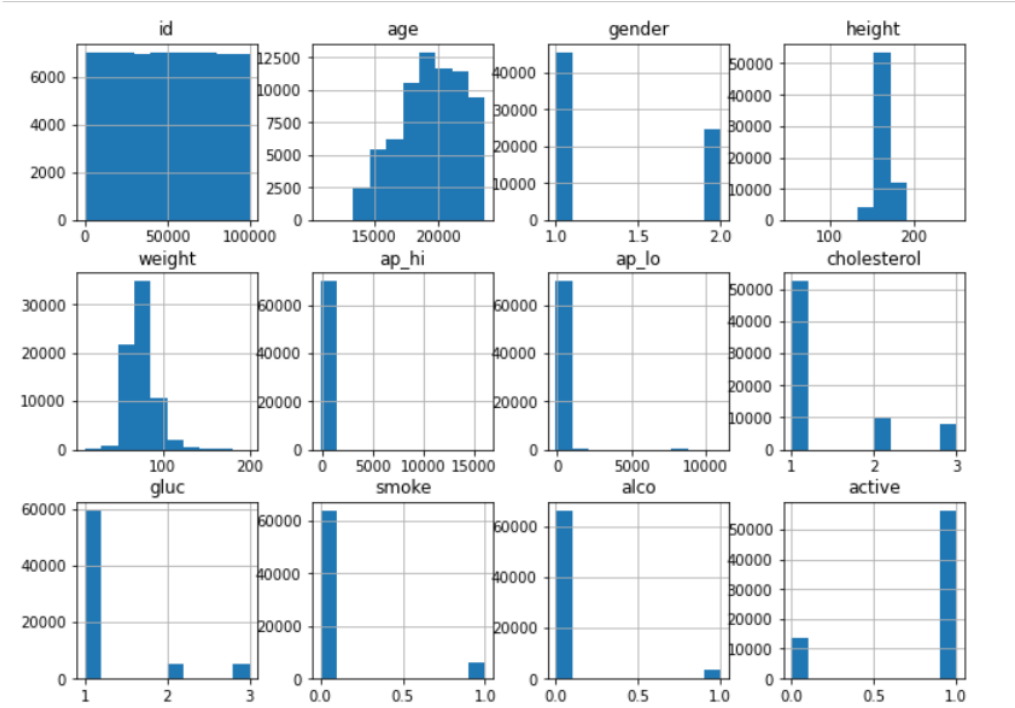
Şekil 3.1'de veri setinde yaşlara göre fiziksel aktivite yapıp yapmama durumunun dağılımını gözlemleyebiliriz. Şekil 3.2'de boy ve kilonun veri seti üzerinde dağılımını görebiliriz.



Şekil 3.2 Boy ve kilonun veri seti üzerinde dağılımı

3.2 Sınıflandırma

Veri seti, bir bağımlı değişken olan active yani KVH'a sahip olunup olunmaması değişkeninden ve 10 bağımsız değişkenden oluşur. Yaptığımız çalışmanın amacı, değişkenlerimize göre başarılı sınıflandırmayı yapacak model oluşturmayı sağlamaktır. Çalışmada JupyterLab program kullanılmıştır.

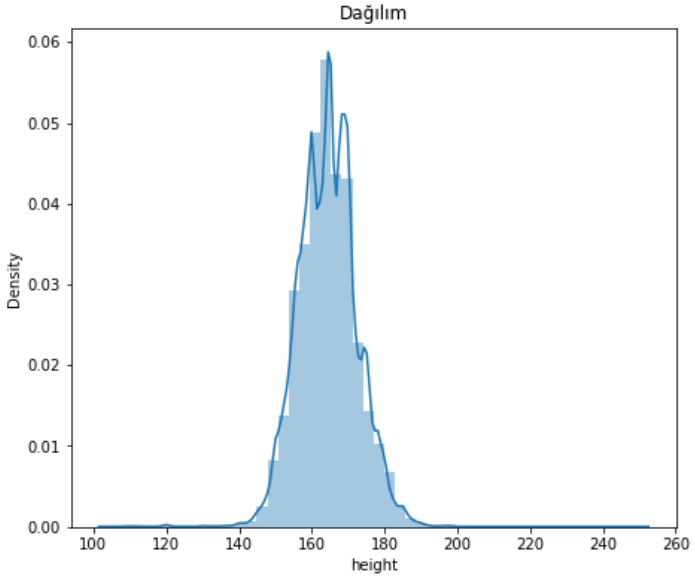


Şekil 3.3 Histogram grafiği ile aykırı değer analizi

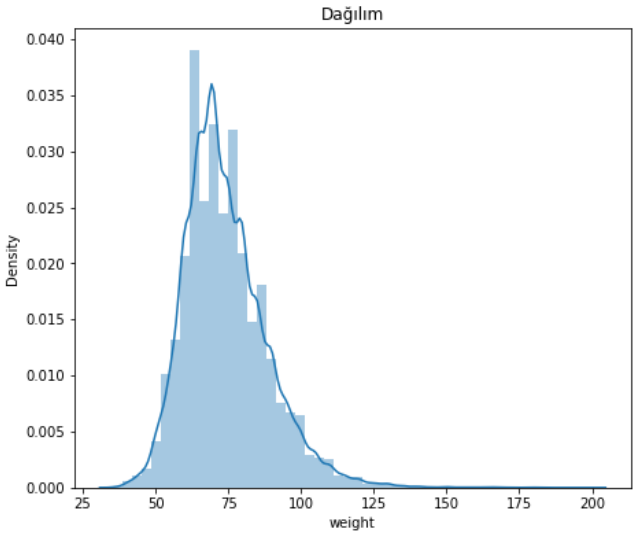
Veriyi analiz ederken kullanılan yazılım araçlarını ve Excel'in birbirinden ayrılmasının en temel noktası veriyi görmek için bazı komutlar kullanmamız gerekmesidir. Keşifçi veri analizi ile veri incelenir ve ilk adım olarak veriye bakış atılır. Aynı zamanda grafiksel tekniklerden olan Scatter Plot yöntemi de kullanılarak ilişkili veriler arasındaki bağlantılar incelenebilir. Şekil 3.3'de histogram grafiği ile değişken dağılımları tek tek incelenerek uçlarda yer alan değerlerin kontrolü sağlanabilir. Bu da aykırı değerlerin tespit edilmesini sağlar. Veri sayısını ise çubuğun yüksekliği belirler.

Veride tekrarlayan, eksik değerler var mı kontrol edilmiştir. Yaş değişkeninde gerekli tip dönüşümleri yapılmıştır. Veri setindeki id kolonu kaldırılmıştır. 1465 aykırı değer tespit edilerek bunlarda veri setinden kaldırılmıştır.

Yaş, boy, kilo, ap_hi ve ap_lo değişkenleri için örnek bir DataFrame oluşturulur. Standardizasyon fonksiyonu tanımlanarak DataFrame'e gelen değişkeni x girdi olarak algılar ve bu değişkenin bir standartlaşmış versiyonunu döndürür. DataFrame'deki bu 5 değişken standartlaştırılmış olur. Yaş, kilo, ap_hi ve ap_lo değişkenini normalleştirmek için 0.01'lik veri ortadan kaldırılmıştır. Şekil 3.4'de görüldüğü üzere height değişkeninin normalleştirilmesine ihtiyaç yoktur. Ancak Şekil 3.5'de görüldüğü üzere weight değişkeninin sağdan 0.01'lik olacak şekilde normalleştirilmesi gerekmektedir.



Şekil 3.4 Height değişkeninin dağılımı



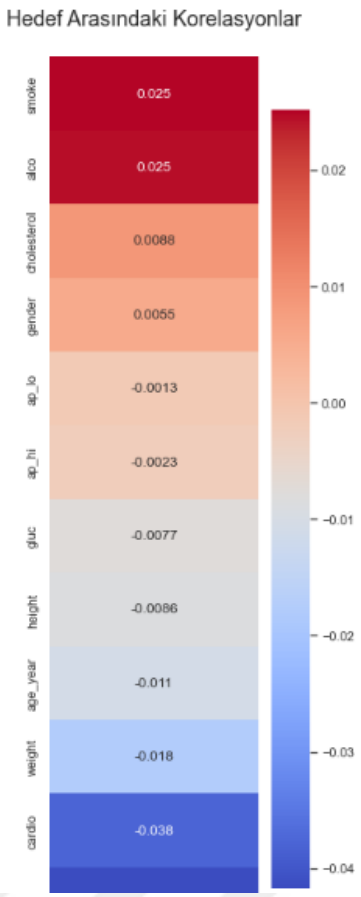
Şekil 3.5 Weight değişkeninin dağılımı

Veri setini normalleştirdikten sonra bağımlı değişken olan active yani KVH olup olmama durumunun bağımsız değişkenler ile arasındaki ilişkiyi incelemek için PCC kullanılır. Tablo 3.3'e bakıldığında active değişkeni için en önemli değişkenler sigara, alkol ve cardio olarak belirlenmiştir.

Tablo 3.3 Değişkenlerin active değişkenine göre PCC Hesaplanması

Değişken	PCC
Cinsiyet	0.0055
Boy	-0.0086
Kilo	-0.0176
ap_hi	-0.0023
ap_lo	-0.0013
Kolesterol	0.0087
Glikoz	-0.0077
Sigara	0.02517
Alkol	0.0245
Fiziksel Aktivite	-0.0379
Yaş	-0.0106

Veri setine özellikler arası korelasyon uygulanarak, veri setindeki değişkenler arasındaki bağlantı incelenir. Özelliklerin beraber nasıl değiştiğini ve bir değişkendeki değişimin diğer değişkeni nasıl etkilediğini incelememize olanak sağlar. Şekil 3.6'da KVH'a yakalanıp yakalanmama durumuna göre PCC değerlerini gösteren hedef arasındaki korelasyonu gözlemleyebiliriz.

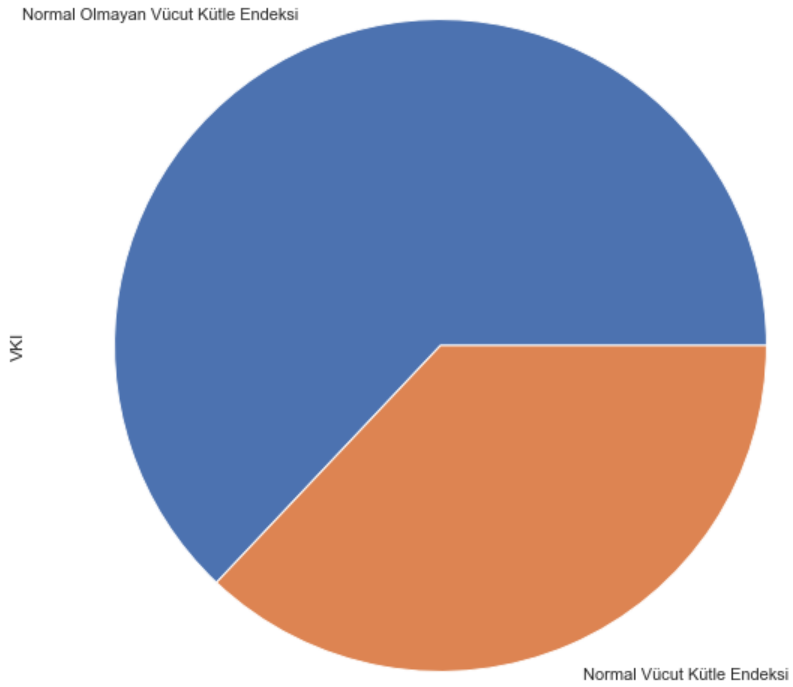


Şekil 3.6 Active değişkenine göre hedef arasındaki korelasyon

Özellik mühendisliği ile, verilerden yeni anlamlar ve özellikler türetilerek modelde kullanılması amaçlanmaktadır. İlk olarak özellik mühendisliği ile vücut kitle indekslerini (VKI) inceleyeceğiz. Şekil 3.7'de vücut kütle indeksi değerlerinin sınıflandırılmasını inceleyebiliriz.

Sınıflandırma	Vücut Kitle İndeksi
Zayıf	18,5'in altında
Normal	18,5-24,9
Fazla Kilolu	25-29,9
Obezite	30 ve üzeri
1. Derece Obezite	30-34,9
2. Derece Obezite	35-39,9
3. Derece Obezite	40 ve üzeri

Şekil 3.7 Vücut kitle indeksinin sınıflandırılması [47]



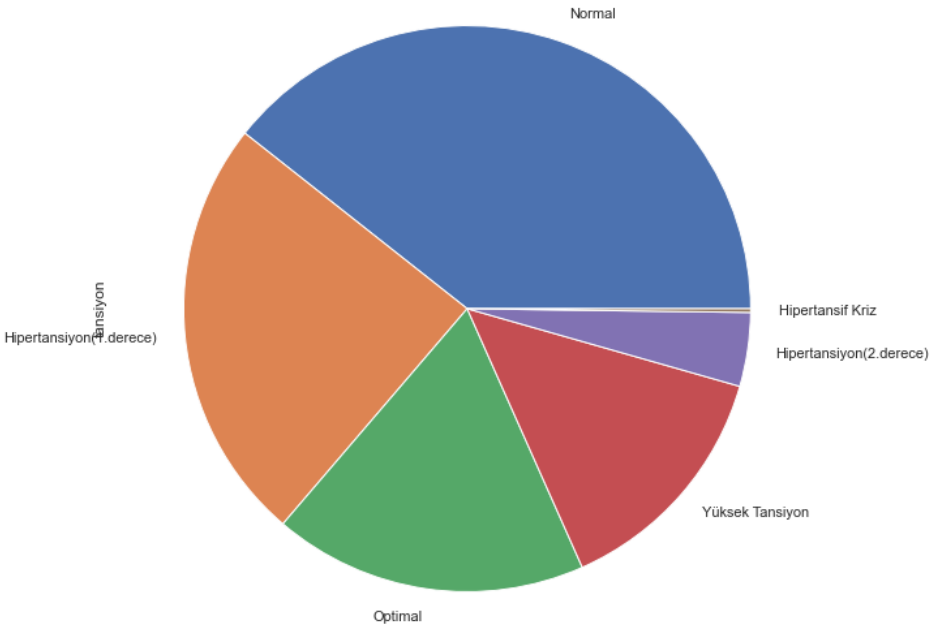
Şekil 3.8 Vücut kütle endeksinin veri seti üzerinde sınıflandırılması

Şekil 3.8'de veri seti üzerinde vücut kütle endeksinin normal ve normal olmayan şekilde sınıflandırılmasıdır.

Şekil 3.9 sistolik ve diyastolik kan basıncının Türk Kardiyoloji Derneği Ulusal Hipertansiyon Tedavi ve Takip Kılavuzuna göre [48] sınıflandırılma değerleridir.

	Kan Basıncı (mmHg)		
Kan Basıncı Derecesi	**Sistolik**		**Diyastolik**
Optimal	< 120	ve	<80
Normal	<130	ve	< 85
Yüksek- Normal	130-139	veya	85-89
Hipertansiyon			
Evre 1	140-159	veya	90-99
Evre 2	160-179	veya	100-109
Evre 3	≥180	veya	≥110
İzole sistolik hipertansiyon (sınırda)	140-160	< 90	
İzole sistolik hipertansiyon	≥160	< 90	

Şekil 3.9 Sistolik ve diyastolik kan basıncı değerlerinin sınıflandırılması [48]



Şekil 3.10 Ap_hi ve ap_lo değişkeninin veri seti üzerinde sınıflandırılması

Şekil 3.10'da ap_hi ve ap_lo değişkeninin Şekil 3.9'daki değerlere göre normal, yüksek tansiyon, optimal, hipertansiyon (2. derece), hipertansiyon (1. Derece) ve hipertansif kriz olarak sınıflandırılmasıdır.

Veri setinin %30'u test, %70'i eğitim seti olarak ayrılmıştır. Eğitim seti için 49.000, test seti için 21.000 hasta ayrılmıştır. Bu aşamalardan sonra verimize PCA yöntemi uygulanmıştır. PCA [49] yöntemi yüksek boyutlu veri setini daha düşük boyutlu bir veri setine dönüştürmede ve veri setini küçültmek için kullanılır. Bu yöntemi uygulayarak veri setimizde en önemli bilgileri belirlemeyi ve KVH'ın hastalık riskini tahmin etmede nasıl kullanabileceğimizi bulmayı amaçlarız. PCA'yı scikit-learn kütüphanesini kullanarak uyguladık. 10 bağımsız değişken 8'e indirgenmiştir.

Ardından lojistik regresyon [50], KNN [51], SVM[52], karar ağacı [53], Gradient Boosting [54], XGBoost [56], LightGBM [57], AdaBoost [58], rastgele orman [59] ve Extra Tree [60] sınıflandırıcı algoritmaları modele uygulanmıştır.

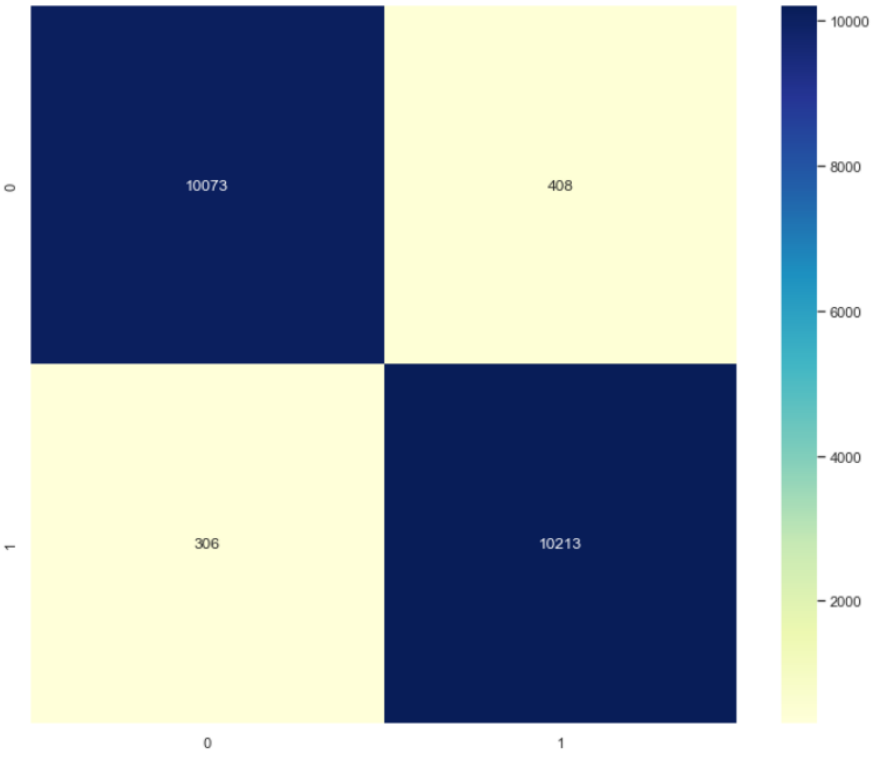
3.3 Sonuçlar

Algoritmada hiperparemetre optimizasyonu uygulanmıştır. En uygun parametreler RandomizedsearchCV [61] ile modelde elde edilmiştir ve daha sonrasında modelin performansı ölçülür.

Tablo 3.4 Modelin performans ölçümü

Yöntem	ACC	MCC	SP	PR	RE	F_1
Lojistik Regresyon	70.80	41.29	76.39	71.85	67.42	69.57
Karar Ağacı	88.62	77.90	89.21	88.42	89.54	88.98
Rastgele Orman	88.42	78.40	87.13	86.63	90.16	88.36
SVM	72.62	46.41	82.45	77.54	62.49	70.21
KNN	71.66	44.71	73.56	72.04	69.41	70.70
XGBoost	**96.60**	**93.84**	**96.40**	**95.96**	**96.89**	**96.42**
Gradient Boosting	80.40	62.31	83.30	81.36	78.00	80.13
LightGBM	91.21	81.17	90.52	89.58	91.74	90.57
AdaBoost	77.19	54.21	81.46	78.56	71.28	76.00
ExtraTree	90.26	80.89	88.26	88.88	92.31	90.41

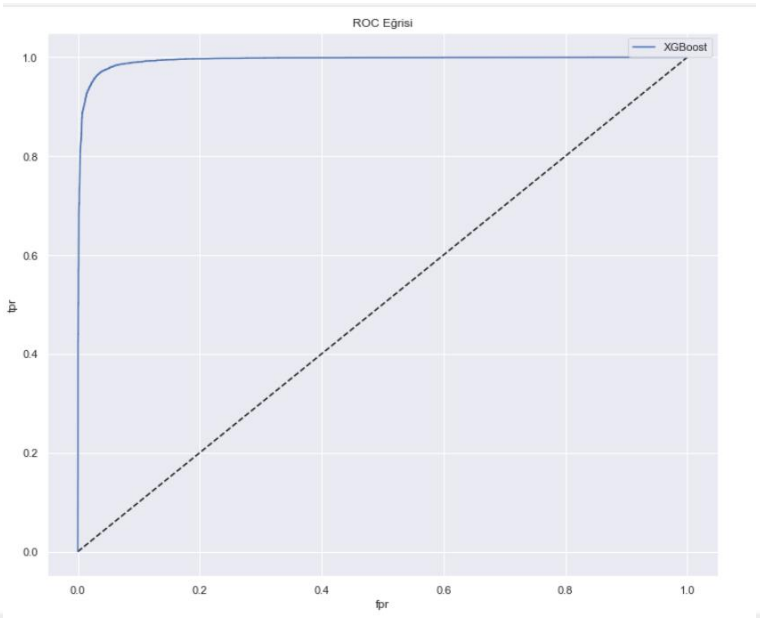
Modelin başarısını ölçmek için ACC, MCC [62], SP, PR, RE ve F_1 değerleri hesaplanmıştır. Tablo 3.4'de sınıflandırma yöntemlerinin performans ölçüm sonuçlarını göstermektedir. XGBoost yöntemi ACC, MCC, SP, PR, RE ve F_1 metriklerinde en başarılı sonucu vermiştir. XGBoost'dan sonra en başarılı metrikler LightGBM'den elde edilmiştir. ACC, SP, PR ve F_1 metriklerinde LightGBM'den sonra en başarılı sonucu veren karar ağacı olmuştur. Karar ağacı yönteminin RE metriği rastgele orman ile eşit sonuca ulaşmıştır.



Şekil 3.11 Karışıklık matrisi

Şekil 3.11'de modelimizin performansını değerlendirmek için oluşturduğum karışıklık matrisini göstermektedir. Şekil 3.12'de ROC eğrisini göstermektedir. Eğri, modelin hassasiyet ve özgüllük arasında ilişkiyi gösterir ve eğri 1'e ne kadar yakınsa modelin o kadar iyi oluşturulduğu anlamına gelir.

ROC eğrisinin altındaki alandan çeşitli metrikler hesaplayabiliriz. XGBoost yöntemi en başarılı sonucu verdiği için bu yöntemin AUC metriği hesaplanmıştır. AUC değeri 0.99 olarak bulunmuştur. Bu değer de modelin performansının iyi olduğu anlamına gelmektedir.



Şekil 3.12 ROC eğrisi

Tablo 3.5'de Butt, Rehman, Javaid, Ali ve Nawaz çalışmasında [3] kullanılan aynı veri seti için makine öğrenimi yöntemlerinden KNN ve karar ağacı ile bu çalışmada uyguladığımız ve elde ettiğimiz en başarılı sonucu elde eden XGBoost yöntemi ile karşılaştırılması gösterilmektedir.

Tablo 3.5 İki çalışma arasında en başarılı sonuçların karşılaştırılması

Yöntem	ACC	SP	PR	RE	F_1
KNN [3]	84.11	79.13	81.02	89.10	-
Karar ağacı [3]	84.11	81.62	82.49	86.60	-
Gradient Boosting [2]	75.00	-	-	-	-
Rastgele orman [1]	73.00	80.00	-	65.00	-
XGBoost	**96.60**	**96.00**	**96.00**	**97.00**	**97.00**

Butt, Rehman, Javaid, Ali ve Nawaz çalışmasında en başarılı sonucu KNN ve Karar ağacı vermiştir. Çalışmada örnek filtresi ile eksik değerler ve Sentetik Azınlık Örnek Verme Tekniği ile veri dengesizliği ortadan kaldırılmıştır. Daha sonrasında optimize edilmiş özellik seçimi kullanılmış ve normalleştirme tekniği ile veriler normalleştirilmiştir. Veri seti, rastgele örnekleme yöntemi ile %70 eğitim seti ve %30 test seti olmak üzere ikiye bölünmüştür.

Bu çalışmada ise veri seti %30'u test, %70'i eğitim seti olarak bölünmüştür. Butt, Rehman, Javaid, Ali ve Nawaz çalışmasından farklı olarak veriye PCA uygulanmış, z-skoru ile

normalleştirilme yapılmış, değişkenlerin PCC değerleri hesaplanmış, hedef arasındaki korelasyonlar incelenmiş ve özellik mühendisliği ile verilerden anlamlı sonuçlar çıkarılmıştır.

İki çalışmanın sonuçları karşılaştırıldığında, XGBoost yöntemi tüm performans metriklerinde daha iyi sonuç verdiği görülmüştür. Butt, Rehman, Javaid, Ali ve Nawaz çalışmasında doğruluk 84.11 iken bu çalışmada 96.60'dır. Sonuç olarak, XGBoost makine öğrenmesi yönteminin başarıyı arttırdığını görebiliriz.

4
SONUÇ

Kardiyovasküler hastalık, dünyada 17,9 milyon kişiyi etkilemiş bir sağlık problemidir [40]. KVH saptanarak hastaneye yatıştaki ölüm oranı %5 ile %30 arasında değişmektedir [41]. KVH tedavisinde ilerlemeler kaydedilmesine rağmen dünya genelinde morbidite ve mortalite oranı önde gelen hastalıklardandır. KVH, 510 milyon morbiditeye neden olmuştur. 1990'dan beri en fazla ölüme neden olan hastalıktır [42]. KVH'a yakalanma oranı ve KVH sebebi ile ölüm oranı her sene artış göstermektedir.

Bu çalışmada, kardiyovasküler hastalık riski tahminlenmesi amaçlanmıştır. Veri setindeki özelliklerin KVH'a neden olup olmayacağı ile ilgili analizler yapılmış ve bu özellikler karşılaştırılmalı olarak incelenmiştir. Türkiye'de erişken 3 kişiden 1'i hipertansiyon hastası ve hipertansiyon KVH'ın en önemli nedenlerinden biridir [42]. Bu sebepten çalışmada, özellik mühendisliği uygulanmış ve hastaların tansiyonları, yaşları ve obezite durumları analiz edilmiştir. Hipertansiyona sahip hastaların KVH durumları analiz edilmiştir. KVH'a neden olabilecek özelliklerin analiz edilmesi ve riskinin tahminlenmesi hastalığın erken teşhisine ve önlenmesine katkı sağlar.

KAYNAKÇA

[1] J. Maiga, G. G. Hungilo ve Pranowo, "Comparison of Machine Learning Models in Prediction of Cardiovascular Disease using Health Record Data," 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, ss. 45-48, 2019.

[2] R. Hagan, C. J. Gillan, F. Mallett, "Comparison of Machine Learning Methods for the Classification of Cardiovascular Disease,", Elsevier, cilt 24, 2021.

[3] M. O. Butt, A. Ur Rehman, S. Javaid, T. M. Ali ve A. Nawaz, "An Application of Artificial Intelligence for an Early and Effective Prediction of Heart Failure," 2022 Third International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT), Karachi, Pakistan, ss. 1-6, 2022

[4] E. Sirin ve H. Karacan, "A Review on Business Intelligence and Big Data," International Journal of Intelligent Systems and Applications in Engineering, cilt 5, sayı 4, 2017.

[5] I. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective, " SN Computer Science, cilt 2, ss. 377, 2021.

[6] Cardiovascular Diseases, World Health Organization, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 11.06.2021.

[7] Türkiye İstatistik Kurumu (TÜİK), "Ölüm ve Ölüm Nedenleri İstatistikleri, 2021", sayı 45715, 2023.

[8] A. Eray, T. Set ve E. Ateş, "Yetişkin Bireylerde Kardiyovasküler Hastalık Riskinin Değerlendirilmesi," Türkiye Aile Hekimliği Dergisi, cilt 22, sayı 1, 2018.

[9] K. Burhan, "Keşifçi Veri Analizi", Medium, 06.08.2023. [Online]. Available: Keşifçi Veri Analizi. Keşifçi veri analizi veya İngilizcesi… | by Burhan Kibar | Medium.

[10] G. Elif, "Keşifçi Veri Analizi", Medium, 23.03.2023. [Online]. Available: https://medium.com/@elifgafar/ke%C5%9Fif%C3%A7i-veri-analizi-exploratory-data-analysis-eda-f2e59a752659.

[11] U. Çolak, "Makine Öğrenmesi Veri Ön İşleme", Medium, 30.05.2021. [Online]. Available: https://ufukcolak.medium.com/makine-%C3%B6%C4%9Frenmesi-veri-%C3%B6n-i%CC%87%C5%9Fleme-5-58e1ce73c1fb.

[12] Dayanıklı A. S., "Aykırı Değer (Outlier) Analizi Nedir? Uç Değerler Nasıl Tespit Edilir?", Ravenfo, 11.02.2021. [Online]. Available: https://ravenfo.com/2021/02/11/aykiri-deger-analizi/.

[13] Cerebro, "Yeni Başlayanlar için Makine Öğrenmesi Algoritmaları", Medium, 13.02.2018. [Online]. Available: https://medium.com/t%C3%BCrkiye/yeni-ba%C5%9Flayanlar-i%C3%A7in-makine-%C3%B6%C4%9Frenmesi-algoritmalar%C4%B1-ae22f794af2f.

[14] Bilen B., "Karışıklık Matrisi (Confusion Matrix)", Medium, 04.02.2021. [Online]. Available: https://burhanbilen.medium.com/kar%C4%B1%C5%9F%C4%B1kl%C4%B1k-matrisi-confusion-matrix-990dfc718653.

[15] O. Gülcan, “ROC ve AUC”, Medium, 12.01.2020. [Online]. Available: https://medium.com/@gulcanogundur/roc-ve-auc-1fefcfc71a14.

[16] U. Çolak, “Makine Öğrenmesi Model Başarı Değerlendirme Ölçütleri”, Medium, 10.05.2021. [Online]. Available: https://ufukcolak.medium.com/makine-öğrenmesi-model-başarı-değerlendirme-ölçütleri-3-26014630d7ce.

[17] U. Çolak, “Makine Öğrenmesi Modellemeye Giriş”, Medium, 01.05.2021. [Online]. Available: https://ufukcolak.medium.com/makine-%C3%B6%C4%9Frenmesi-modellemeye-giri%C5%9F-2-405b4f3c8cfb.

[18] A. Arzu ve H. Önder, “Farklı Veri Yapılarında Kullanılabilecek Regresyon Yöntemleri”, Anadolu Tarım Bilim Dergisi, cilt 28, sayı 3, ss. 168-174, 2013.

[19] C. Can, “Eğitim Alanında Makine Öğrenimi Sınıflandırma Algoritmalarının İncelenmesi”, Akdeniz Üniversitesi Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Ana Bilim Dalı Tezli Yüksek Lisans Programı, Antalya, 2021.

[20] Kalim, “Decision Tree”, Medium, 19.08.2023. [Online]. Available: https://medium.com/@kalim7532/decision-tree-91797ab6e655.

[21] L. Rokach ve O. Maimon, “Decision Tree”, The Data Mining and Knowledge Discovery Handbook, Springer, Boston, bölüm 9, ss. 165-192.

[22] A. Atçılı, “KNN K-En Yakın Komşuluk”, Medium, 04.01.2022. [Online]. Available: https://medium.com/machine-learning-t%C3%BCrkiye/knn-k-en-yak%C4%B1n-kom%C5%9Fu-7a037f056116.

[23] M. F. Akça, ”Nedir Bu Destek Vektör Makineleri”, Medium, 25.08.2020. [Online]. Available: https://medium.com/deep-learning-turkiye/nedir-bu-destek-vekt%C3%B6r-makineleri-makine-%C3%B6%C4%9Frenmesi-serisi-2-94e576e4223e.

[24] K.-R. Müller, A.J.Smola, G.Ratsch,B. Schölkopf, J. Kohlmorgen ve V. Vapnik, “Predicting Time Series with Support Vector Machines”, Artificial Neural Networks, ICANN 1997, ss. 999-1004, 2005.

[25] S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang ve W. Xu, “Applications of Support Vector Machine (SVM) Learning in Cancer Genomics”, Cancer Genomics & Proteomics, cilt 15, sayı 1, ss. 41-51, 2018.

[26] L. Breiman, “Random Forests”, Machine Learning, cilt 45, ss. 5-32,2001.

[27] M. Öztürk, “Python ile Sınıflandırma Analizleri Rastgele Orman Algoritması”, Business Intelligence, 13.04.2022. [Online]. Available: https://miracozturk.com/python-ile-siniflandirma-analizleri-rastgele-orman-random-forest-algoritmasi/.

[28] A. Natekin ve A. Knoll, “Gradient Boosting Machines, A Tutorial”, Department of Informatics, Technical University Munich, Garching, Munich, Germany, cilt 7, 2013.

[29] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine”, The Annals of Statistics, cilt 29, sayı 5, ss. 1189-1232, 2001.

[30] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ss. 785–794, 2016.

[31] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia ve M. Apri, "An Optimized XGBoost-Based Machine Learning Method for Predicting Wave Run-Up on A Sloping Beach", School of Computing, Telkom University, Bandung, Indonesia, cilt 10, 2023.

[32] E. R. Muratlar, "Gradient Boosted Regresyon Ağaçları", Veri Bilimi Okulu, 24.01.2020. [Online]. Available: https://www.veribilimiokulu.com/gradient-boosted-regresyon-agaclari/.

[33] E. R. Muratlar, "XGBoost Nasıl Çalışır ? Neden İyi Performans Gösterir ?", Veri Bilimi Okulu, 22.03.2020. [Online]. Available: https://www.veribilimiokulu.com/xgboost-nasil-calisir/.

[34] J. Zhang, D. Mucs, U. Norinder ve F. Svensson, "LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets", Journal of Chemical Information and Modeling, cilt 59, sayı 10, ss. 4150-4158, 2019.

[35] T. Chengsheng, L. Huacheng ve X. Bing, "AdaBoost typical Algorithm and its application research", MATEC Web of Conferences, cilt 139, sayı 2, 2017.

[36] M. M. Hameed, M. K. AlOmar, F. Khaleel ve N. Al-Ansari, "An Extra Tree Regression Model for Discharge Coefficient Prediction: Novel, Practical Applications in the Hydraulic Sector and Future Research Directions", Mathematical Problems in Engineering, cilt 2021, sayı 1, ss. 1-19, 2021.

[37] E. R. Muratlar, "LightGBM", "Veri Bilimi Okulu", 29.04.2020. [Online]. Available: https://www.veribilimiokulu.com/lightgbm/.

[38] K. Güzel, "Boosting Nedir ? Adım Adım AdaBoost Algoritması", 28.12.2020. [Online]. Available:https://kadirguzel.medium.com/boosting-nedir-ad%C4%B1m-ad%C4%B1m-adaboost-algoritmas%C4%B1-439cce20ab9a.

[39] E. Asan, "Extra Trees", 10.01.2022. [Online]. Available: https://arinway.com/yapay-zeka-konulari/ekstra-agaclar-(extra-trees)-1225-konusu.

[40] Cardiovascular Diseases, World Health Organization, https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[41] "Dünyada ve Ülkemizdeki Birincil Ölüm Nedeni, Hala Kalp-Damar ve Dolaşım Sistemi Hastalıkları", Türk Kardiyoloji Derneği, https://tkd.org.tr/menu/202/turk-kardiyoloji-dernegi-dunyada-ve-ulkemizdeki-birincil-olum-nedeni-h-l-kal, 17.05.2022.

[42] S. Civek ve M. Akman, "Dünyada ve Türkiye'de kardiyovasküler hastalıkların sıklığı ve riskin değerlendirilmesi", Jour Turk Fam Phy, cilt 13, sayı 1, ss. 21-28, 2022.

[43] "Kolesterol nedir? Kolesterol Belirtileri Nelerdir?", Acıbadem Web ve Yayın Kurulu, [Online]. Available: https://www.acibadem.com.tr/ilgi-alani/kolesterol/, 07.08.2023.

[44] "Glukoz nedir? Glukozun düşmesi ve yükselmesi ne anlama gelir?", Medical Park Yayın Kurulu, [Online]. Available: https://www.medicalpark.com.tr/glukoz/hg-2179, 06.05.2021.

[45] “Hipertansiyonda Rakamlar Ne Anlama Geliyor?”, TEB, [Online]. Available: https://www.teb.org.tr/uploads/afis/hiper_tansiyon_11_2.pdf.

[46] “Kan Basıncı Nedir?”, Medical Park Yayın Kurulu, [Online]. Available: https://www.medicalpark.com.tr/kan-basinci-nedir/hg-2378, 22.06.2020.

[47] “Vücut Kütle İndeksi Değerleri Ne Anlama Gelir?”, [Online]. Available: https://www.diyabet.com/tr/diyabet-hakkinda/vucut-kitle-indeksi-hesaplama.html.

[48] “Kan Basıncının Ölçümü ve Klinik Değerlendirme”, [Online]. Available: https://tkd.org.tr/kilavuz/k03/3_2d304.htm#:~:text=Buna%20g%C3%B6re%20optimal%20kan%20bas%C4%B1nc%C4%B1,de%20hipertansiyon%20olarak%20kabul%20edilmektedir.

[49] S. Wold, K. Esbensen, ve P. Geladi, “Principal component analysis,” Chemometrics and Intelligent Laboratory Systems, cilt 2, sayı 1-3, ss. 37–52, 1987.

[50] S. Şenel ve B. Alatlı, “Lojistik Regresyon Analizinin Kullanıldığı Makaleler Üzerine Bir İnceleme”, DergiPark, cilt 5, sayı 1, ss. 35-52, 2014.

[51] P. Cunnigham ve S. J. Delany, ”K-Nearest Neighbors for Multi-Label Classification”, arXiv preprint arXiv:2004.04523, 2020.

[52] A. Tekerek, “Support Vector Machine Based Spam SMS Detection”, DergiPark, cilt 22, sayı 3, ss. 779-784, 2019.

[53] D. Atasoy ve H. Kara, “Karar Ağacı Optimizasyon Algoritması Üzerine Bir Çalışma”, DergiPark, cilt 13, sayı 2, ss. 1247-1255, 2023.

[54] A. Ramón, A. M. Torres, J. Milara, J. Cascón, P. Blasco ve J. Mateo, “eXtreme Gradient Boosting-based method to classify patients with COVID-19”, Journal of Investigative Medicine, cilt 70, sayı 7, ss. 1472-1480, 2022.

[55] A. Natekin ve A. Knoll, “Gradient boosting machines, a tutorial”, Front Neurorobot, cilt 7, sayı 21, 2013.

[56] I. L. Cherif ve A. Kortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification", 2019 Wireless Days (WD), ss. 1-6, 2019.

[57] M. Alkasassbeh, M. A. Abbadi ve A. M. Al-Bustanji, “LightGBM Algorithm for Malware Detection”, ss. 391-403, 2020.

[58] M. Bozuyla, “AdaBoost Ensemble Learning on Top of Naïve Bayes Algorithm to Discriminate Fake and Genuine News From Social Media”, sayı 28, ss. 459-462, 2021.

[59] Ö. Akar ve O. Güngör, “Rastgele Orman Algoritması Kullanarak Çok Bantlı Görüntülerin Sınıflandırılması”, cilt 1, sayı 2, ss. 139-146, 2012.

[60] M. S. H. Talukder, R. B. Sulaiman ve M. B. P. Angon, “Unleashing the Power Of Extra-Tree Feature Selection and Random Forest Classifier for Improved Survival Prediction in Heart Failure Patients”, 2023.

[61] J. Bergstra ve Y. Bengio, “Random Search for Hyper-Parameter Optimization”, ResearchGate, cilt 13, sayı 1, ss. 281-305, 2012.

[62] D. Chicco ve G. Jurman, “The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation”, ResearchGate, cilt 21, sayı 1, 2020.

[63] S. Ulianova, ”Cardiovascular Disease dataset”, [Online]. Available:

https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset, Kaggle, 2019.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirisi

1. E. Öztürk, M. Akar, “On the Prediction of Cardiovascular Diseases with Machine Learning Classification Algorithms”, 7th International Hybrid Conference on Mathematical Advances and Applications (ICOMAA-2024), İstanbul, pp. 71, May 8-11, 2024.