T.C.
KIRIKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
YÜKSEK LİSANS TEZİ

KORONER ARTER HASTALIĞI RİSKİNİN MAKİNE ÖĞRENMESİ İLE ANALİZ EDİLMESİ

Şeyma CİHAN

Haziran 2018

Bilgisayar Mühendisliği Anabilim Dalında Şeyma CİHAN tarafından hazırlanan KORONER ARTER HASTALIĞI RİSKİNİN MAKİNE ÖĞRENMESİ İLE ANALİZ EDİLMESİ adlı Yüksek Lisans Tezinin Anabilim Dalı standartlarına uygun olduğunu onaylarım.

Prof. Dr. Hasan ERBAY
Anabilim Dalı Başkanı

Bu tezi okuduğumu ve tezin **Yüksek Lisans Tezi** olarak bütün gereklilikleri yerine getirdiğini onaylarım.

Dr. Öğr. Üyesi Halil Murat ÜNVER
Danışman

Jüri Üyeleri

Başkan	: Doç. Dr. Necaattin BARIŞÇI	_____
Üye (Danışman)	: Dr. Öğr. Üyesi Halil Murat ÜNVER	_____
Üye	: Dr.Öğr. Üyesi B.Gürsel EMİROĞLU	_____

01/06 /2018

Bu tez ile Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onaylamıştır.

Prof. Dr. Mustafa YİĞİTOĞLU
Fen Bilimleri Enstitüsü Müdürü

ÖZET

KORONER ARTER HASTALIĞI RİSKİNİN MAKİNE ÖĞRENMESİ İLE ANALİZ EDİLMESİ

CİHAN, Şeyma
Kırıkkale Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi
Danışman: Dr. Öğr. Üyesi Halil Murat ÜNVER
Haziran 2018, 111 Sayfa

Kardiyovasküler hastalıklar, tüm dünyada ölüm nedenleri arasında ilk sırada yer almaktadır. Kardiyovasküler hastalıklar içerisinde ölümcül sonuçları olan en yaygın klinik tip koroner arter hastalığıdır. Bu nedenle, koroner arter hastalığının erken dönemde ve doğru bir biçimde saptanmasında tanı işlemlerinin iyileştirilmesi hayati önem taşımaktadır. Koroner anjiyografi yöntemi, koroner arter hastalığının tanısında ve hastalık sürecinin değerlendirilmesinde en yaygın kullanılan girişimsel yöntem olarak kabul edilmektedir. Ancak, koroner anjiyografi yöntemi yüksek maliyeti, ileri seviyede eğitimli personel gerektirmesi ve önemli klinik komplikasyonları olan girişimsel bir işlem olması sebebiyle, tarama amaçlı ya da tedavi altındaki hastaların takibi açısından kullanımı uygun değildir. Bu nedenle, birçok araştırmacı koroner arter hastalığının tanısında, makine öğrenmesi gibi alternatif yöntemler üzerinde çalışmalar yürütmektedir. Makine öğrenmesi yöntemlerinin klinik alanlarda kullanımı ile birlikte, hastalar için mevcut tüm değişkenlerin kolayca yorumlanarak değerlendirilmesi sağlanabilmekte ve bu şekilde her adımın tanısal doğruluğu artırılabilmektedir.

Bu çalışmada, Kaliforniya Üniversitesi, Irvine (UCI) veri kümesi koleksiyonundan alınan Cleveland, Macaristan, İsviçre ve VA Long Beach kalp hastalığı veri kümeleri üzerinde Rastgele Orman Algoritması kullanılarak koroner arter hastalığı riski analiz edilmiştir. Veri analizi aşamasında, belirtilen tüm veri kümeleri incelenmiştir. Ancak, sınıflama modeli daha az eksik veri içermeleri ve dengeli bir dağılıma sahip olmaları

nedeniyle Cleveland ve Macaristan veri kümeleri üzerinde kurulmuştur. Veri analizi, kardiyoloji alanında uzman bir hekimin rehberliğinde grafiksel ve istatistiksel yöntemlerle yapılmıştır. Uygulanan sınıflama modeli sonucunda, Cleveland veri kümesi üzerinde %86,13 doğruluk oranı ve Macaristan ve Cleveland veri kümelerinin birleştirilmesi ile elde edilen 596 hasta kaydından oluşan veri kümesi üzerinde ise %80 doğruluk oranı elde edilmiştir. Ayrıca, modelin uygulandığı her iki veri kümesinde de göğüs ağrısı tipi ve egzersizle tetiklenen ST depresyonu sınıflama açısından en önemli iki değişen olarak saptanmıştır.

Bu çalışmanın, koroner arter hastalığı olan bireylerin hastalık yönetiminde ve girişimsel klinik işlem uygulanacak hasta grubunun doğru bir biçimde belirlenmesinde sağlık çalışanlarına rehberlik edeceği düşünülmektedir. Bununla birlikte makine öğrenmesi yaklaşımı kullanılarak yapılan sınıflama sonucunda risk grubunun belirlenerek yalnızca gerekli hastalara girişimsel işlemlerin uygulanması sağlanabilecektir. Ayrıca, işlemden kaynaklanan medikal hatalar, sağlık bakım maliyeti ve sağlık uzmanı gereksinimi azaltılırken, hasta güvenliği ve klinik karar kalitesi artırılacaktır.

Anahtar Kelimeler: Makine Öğrenmesi, rastgele orman, koroner arter hastalığı

ABSTRACT

ANALYZING THE RISK OF CORONARY ARTERY DISEASE USING MACHINE LEARNING

CİHAN, Şeyma
Kırıkkale University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering, M. Sc. Thesis
Supervisor: Asst. Prof. Dr. Halil Murat ÜNVER
June 2018, 111 Pages

Cardiovascular diseases are the leading cause of death worldwide. Coronary artery disease is the most common clinical type of cardiovascular diseases with fatal outcomes. For this reason, improvement of diagnostic procedures for early identification of coronary heart disease has vital importance. Angiography procedure is accepted as the most common interventional method used in the diagnosis and evaluation process of coronary artery disease. However, angiography procedure is not suitable for screening patients because it is an interventional procedure with significant clinical complications, requiring high-cost, advanced educated medical personnel. For this reason, many researchers are working on alternative methods and models such as machine learning in the diagnosis process of coronary artery disease. With the clinical use of machine learning methods, it is possible to provide an effortless evaluation of all available variables for patients and thus improve the diagnostic accuracy of each step.

In this study, the risk of coronary artery disease was analyzed using a random forest algorithm on Cleveland, Hungary, Switzerland, and VA Long Beach heart disease data sets from the University of California, Irvine (UCI). All data sets were examined in data analysis process. However, the classification model is based on data sets in Cleveland and Hungary because they have less a missing value and balanced distribution. Data analysis was performed with graphical and statistical methods under

the guidance of cardiologist. As a result of the classification model, 86.13% accuracy rate was obtained on the Cleveland dataset, and 80% accuracy was obtained on the dataset of 596 patient records formed by combining the Hungarian dataset and Cleveland dataset. In addition, the two most important attributes in terms of classification were chest pain type and ST depression triggered by exercise in both data sets.

This study suggests that health professionals in terms of diagnosis and treatment process of individuals with coronary heart disease and determining the right patient group to perform an interventional clinical procedure. In addition, patient safety and clinical decision quality will be improved when the risk group is identified by classification process using the machine learning approach and only the necessary interventional medical procedures are applied, while the medical errors, health care costs and healthcare specialist need arising from the procedure are reduced.

Key Words: Machine learning, random forest, coronary artery disease.

TEŞEKKÜR

Yüksek lisans tezimin hazırlanması esnasında hiçbir desteğini esirgemeyen, tez yöneticisi hocam, Sayın Dr.Öğr. Üyesi Halil Murat ÜNVER ’e ve makine öğrenmesi alanında çalışmaya karar vermemi sağlayan ve tez süreci boyunca yardım aldığım hocam, Sayın Doç. Dr. Güvenç ARSLAN ’a ve çalışma arkadaşım Arş. Gör. Bergen KARABULUT ’a desteğinden dolayı teşekkür ederim.

Bu tezi, desteğini hep yanımda hissettiğim ve tez konusunda medikal danışmanlık aldığım, hayat arkadaşım, eşim Gökhan CİHAN’ a ve varlıklarıyla beni motive eden canım kızlarım İzem ve Eylül’ e ithaf ederim.

İÇİNDEKİLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

ÇİZELGE Sayfa

ŞEKİLLER DİZİNİ

ŞEKİL Sayfa

KISALTMALAR DİZİNİ

KVH	Kardiyovasküler hastalıklar
KAH	Koroner arter hastalığı
EKG	Elektrokardiyografi
LMA	Left Main Arter
RCA	Right Coronary Arter
Cx	Circumflex
LAD	Left Anterior Decending
LDL	Low Density Lipoprotein
HDL	High Density Lipoprotein
OOB	Out of Bag
UCI	University of California, Irvine
NRMSE	Normalized Root Mean Square Error
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
PFC	Percent of False Classified
ROC	Receiver Operating Characteristic

1. GİRİŞ

Kardiyovasküler hastalıklar (KVH) kalbi ve kan damarlarını etkileyen geniş yelpazedeki hastalıkların tümünü ifade etmektedir (Soni vd., 2011). KVH, koroner arter hastalığı, kalp yetmezliği, serebrovasküler hastalıklar, aort damarı hastalıkları ve periferik damar hastalıkları gibi hatalıklardan oluşmaktadır (Wong, 2014). Kardiyovasküler hastalıklar, ölüm nedenleri arasında ilk sırada yer almaktadır (Laslett vd., 2012; Smith vd., 2012). Dünya Sağlık Örgütü 2011 yılı raporuna göre her yıl yaklaşık 17,7 milyon kişi, kardiyovasküler hastalıklar nedeniyle yaşamını yitirmektedir. Bu sayı, küresel ölümlerin %31'ini oluşturmaktadır. Bununla birlikte, bu sayının 2030 yılına kadar 23,6 milyonu geçmesi beklenmektedir. (Mendis vd., 2011). Türkiye'de, erişkin bireylerde koroner kalp hastalığının risk faktörlerini inceleyen TEKHARF çalışması 2012 yılı sayısal verilerine göre, ölüm nedenleri arasında, koroner kalp hastalıkları %42 oranla ilk sırada yer almaktadır. Ayrıca, ülkemizde, 3,5 milyon kronik koroner kalp hastası bulunmakta ve bu sayının yılda 140 bin artış göstereceği tahmin edilmektedir. Bununla birlikte, koroner kalp hastalığına bağlı ölüm sayısı yıllık 215 bin civarındadır (Onat, 2017).

Kardiyovasküler hastalıkların en önemlilerinden biri koroner arter hastalığıdır (KAH) (Alizadehsani vd., 2013). Koroner arter hastalığının erken dönemde ve doğru tespit edilmesi hastalık yönetiminde kritik öneme sahiptir. Kalp hastalıklarında tanılama, dört tanılama seviyesinden oluşmaktadır. Bunlar; hastalık belirti ve bulgularının ve ayrıca istirahat Elektrokardiyografi (EKG) bulgularının değerlendirilmesi, kontrollü egzersiz sırasında çekilen ardışık EKG bulgularının değerlendirilmesi, myokard sintigrafisi ve son seviyede ise koroner anjiyografidir. Kalp hastalıklarının belirlenmesinde, rasyonel bir tanı algoritmasının amacı, kesin teşhisi koymak ve etkin hastalık yönetimi ve tedavisini yalnızca gerekli tanı adımlarını kullanarak planlamaktır (Kukar vd., 1999). Klinik muayene bulguları, EKG testi ve sintigrafi gibi girişimsel olmayan medikal prosedürler, koroner hastalıkların kesin bir biçimde tanılanmasında yetersiz kalabilmektedir. Bu nedenle, anjiyografi işlemi, kalp hastalıklarının tanı sürecinde ve koroner daralma oranının belirlenmesinde altın standart olarak yaygın bir biçimde kullanılmaktadır. Ancak, koroner anjiyografi prosedürü, maliyeti oldukça

yüksek ve ileri seviyede klinik bilgi ve beceri gerektiren bir yöntemdir (Alizadehsani vd., 2012). Ayrıca, koroner anjiyografi işlemi sırasında veya sonrasında, hastanın klinik durumuna, işlemi yapan sağlık personelin deneyimine ve yapılan işleme göre; ölüm, kalp krizi, beyin kanaması, ritim bozuklukları ve damarların görüntülenmesi amacıyla kullanılan opak maddeye bağlı böbrek yetmezliği gibi komplikasyonlar görülebilmektedir (Ökçün ve Gürmen, 2007). Bu nedenle, araştırmacılar koroner arter hastalığı tanısında, bilgisayar tabanlı ve klinik maliyeti daha az olan yöntemler üzerinde çalışmalar yürütmektedir (Soni vd., 2011; Alizadehsani vd., 2013; El-Bialy vd., 2015; Sharan ve Sathees, 2016). Bu yöntemlerden biri de makine öğrenmesidir. Son yıllarda biyomedikal verilerin analiz edilmesi, hastalıkların tanılanması ve saptanmasında makine öğrenmesinin kullanımı önemli ölçüde artış göstermiştir (Foster vd., 2014).

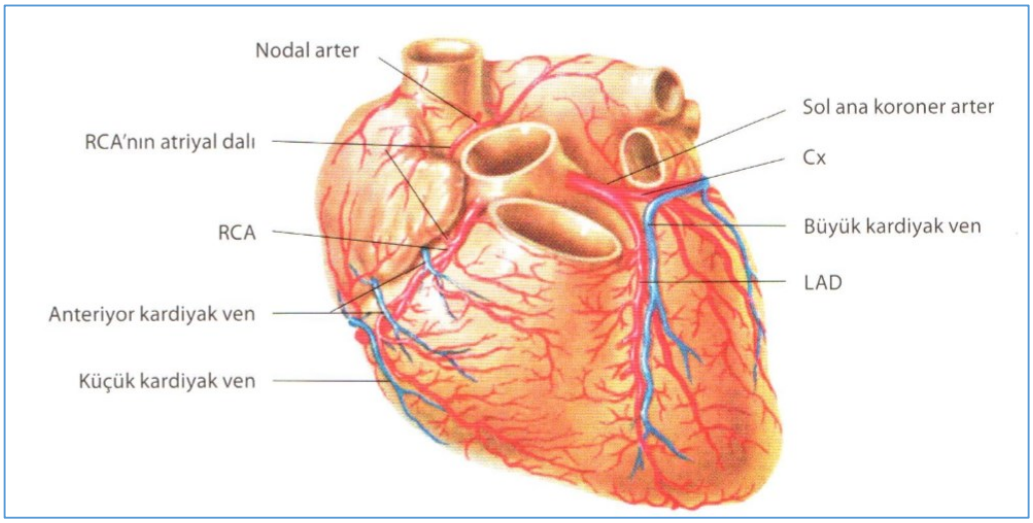
Makine öğrenmesi, bilgisayarların verilerden nasıl öğrendiğini araştıran bilimsel bir disiplin olarak tanımlanabilmektedir. Makine öğrenmesi, veriler arasındaki ilişkileri öğrenmeye çalışan istatistik ile etkili hesaplama algoritmalarının geliştirilmeye çalışıldığı bilgisayar bilimlerinin kesişiminden ortaya çıkmıştır (Deo, 2015). Matematik ve bilgisayar bilimlerinin teknolojilerini kullanan makine öğrenmesi, büyük miktarlardaki verinin rasyonel analizini sağlayan çok sayıda araç sunabilmektedir (Kononenko, 2011).

Makine öğrenmesi algoritmalarına dayalı hastalık tanılama araçları, sağlık alanında önemli karar destek sistemleri haline gelmiştir (Özçift, 2011). Makine öğrenmesi, klinik verileri analiz etmek ve bu verilerden tahminler üretmek için güçlü ve esnek bir araç olarak kullanılabilmektedir. Makine öğrenmesi modelleri, sağlık bakım kalitesini çeşitli yollarla iyileştirme potansiyeline sahiptir. Hastalık sürecini ve seyrini tahmin eden algoritmalar, sağlık çalışanlarının kaynakları en iyi şekilde tahsis etmelerini ve hastalar için daha iyi olabilecek tedavi seçeneklerine karar verebilmelerini sağlayabilmektedir. Bununla birlikte, makine öğrenmesi yaklaşımının klinik alanda kullanılması, sağlık çalışanlarının iş yükünü azaltmakta, hastanın sağlık bakımına erişimini hızlandırmakta ve artırmakta, klinik kaynakları korumakta ve sağlık bakım maliyetlerini düşürmektedir (Gui ve Chan, 2017).

Makine öğrenmesi algoritmaları kullanılarak geliştirilen yazılımlar karmaşık ve büyük miktardaki medikal veriyi kolayca yorumlayarak, hastalığın gerçek zamanlı analizini, tespitini ve sınıflandırılmasını sağlayabilmektedir. Bununla birlikte, Dünya Sağlık Örgütü'nün araştırmaları da makine öğrenmesi gibi yöntemler ile medikal veri kümelerinden elde edilen bilgi ve örüntülerin, hastalığın tanılanması ve tedavi sürecinin yönetimi, hastanın sağlık planlaması, sağlık bakım sistemlerinin izlenmesi ve planlanması, sağlık hizmetlerinin yönetimi ve hastalıkların önlenmesi açısından önemine dikkat çekmektedir (Nahar vd., 2013).

1.1. Koroner Arter Hastalığı

Kalbin beslenmesi, aort damarından ayrılan koroner arterler aracılığıyla sağlanmaktadır. Koroner arterler, aortanın ostium adı verilen bölgesinden iki ana dal olarak çıkarlar. Bunlar; sol ana koroner arter (Left Main Arter-LMA) ve sağ ana koroner arterdir (Right Coronary Arter-RCA). Sol ana koroner arter, atriyoventriküler olukta Sirkumfleks arter (Circumflex-Cx) ve sol ön inen arter (Left Anterior Decending-LAD) olarak ikiye ayrılmaktadır (Şekil 1.1.).



Şekil 1.1. Koroner arter anatomisi (Netter, 2001).

Koroner arter hastalığı, “damar sertliği” olarak bilinen, ateroskleroz sonucunda gelişir. Ateroskleroz, lipid ve fibröz dokudan oluşan aterom plakların koroner arter duvarının iç kısmında yer alan intima tabakasında birikmesi ile oluşan, koroner damarlarda kalınlaşma ve esneklik kaybı ile karakterize, kronik inflamatuvar bir süreçtir (Tokgözoğlu, 2009; Avşar vd., 2011).

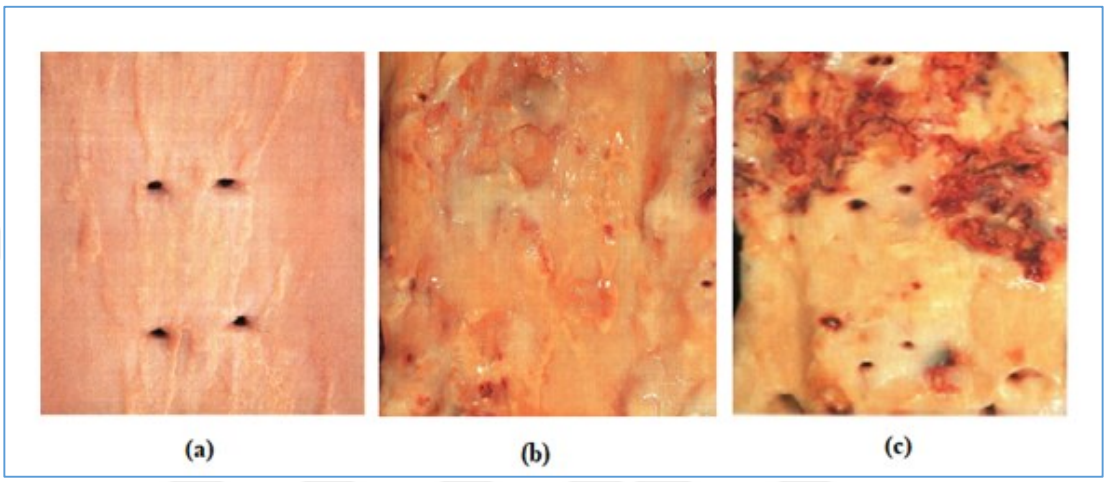
1.1.1. Aterosklerotik Plak Gelişimi

Literatürde aterosklerotik plak gelişimi 3 evrede tanımlanmaktadır (Harrison, 1997). Bunlar; yağlı çizgilenmeler, fibröz plaklar ve komplike lezyonlardır (Şekil 1.2.). Yağlı çizgilenmeler (Şekil 1.2.a.) ateroskleroz gelişimin erken safhasını oluşturur ve çocukluk döneminden itibaren görülmeye başlar. Yağlı çizgilenmeler, köpük hücreleri adı verilen, içerisinde fazla miktarda lipid damlacığı olan makrofajların damar intima tabakasında birikmeleri ile oluşur. Lipid damlacıkları, LDL (Low Density Lipoprotein) kolesterolden kaynaklanan kolesterol esterleridir. Ateroskerozun, bu safhasında kan akımını azaltacak herhangi bir daralma söz konusu değildir.

Aterosklerotik sürecin devamında oluşan fibröz plaklar (Şekil 1.2.b.), düz kas hücreleri ve bağ dokusundan oluşmaktadır. Fibröz plaklar, lipid içeriği lümenden ayıran fibröz bir kapsül ve lipid çekirdek içermektedirler. Lipid çekirdek bölgesi, inflamatuvar ve immün sistem hücreleri de içerir. Zamanla aterom plağı büyür ve intima kalınlaşır. Aterom plağı büyüdükçe damar lümenini daraltmaya başlar. Lümendeki daralma ciddi düzeye erişse bile plak bütünlüğü rüptür ya da erezyon ile bozulmadığı sürece klinik olarak belirti vermeyebilmektedir. Lipid çekirdek ve enflamatuvar hücrelerden zengin fibröz kapsülü ince olan plakların yırtılma ve komplike olma ihtimali fazladır (Stemmi vd., 1995; Hansson ve Nilsson, 2010). Lipid içeriği %40’dan fazla olan aterom plakları “hassas plak” olarak nitelendirilir ve bunlarda rüptür ihtimalinin daha fazla olduğu gösterilmiştir (Burke vd., 1997).

Komplike lezzyonlar ise (Şekil 1.2.c.) aterom plağının çeşitli nedenlerle erezyonu ve rüptüre olması sonucunda ortaya çıkar ve fibröz plak içeriğine ek olarak trombüs ya da hemoraji içerir. Plak rüptürü sonucu oluşan trombüse bağlı damar lümeninde kısmi

ya da tam tıkanıklık ortaya çıkar ve koroner ateroskleroza bağlı önemli kardiyovasküler olaylar ve ölüm daha çok bu lezyonlar nedeniyle meydana gelir (Stemmi vd., 1995; Hansson ve Nilsson, 2010). Komplike lezyonlarda yaşla birlikte kalsiyum depolanması izlenebilir. Plakta oluşan kalsiyum depolanması aterom plağının rüptüre olma ihtimalini önemli ölçüde artırır (Stary vd., 1995).



Şekil 1.2. Aterosklerotik plak gelişim süreci (**a.** Yağlı çizgilenme **b.** Fibröz plak **c.** Komplike plak (Davies, 1998))

1.1.2. Koroner Arter Hastalığında Risk Faktörleri

Koroner arter hastalığı, tedavi yöntemlerindeki gelişmelere rağmen gelişmiş ülkelerdeki mortalitenin (ölümün) en önemli nedeni olmaya devam etmektedir. Kardiyovasküler hastalıklar için risk faktörleri ilk kez 1948 yılında başlayan ve 1960'lı yıllarda sonuçları açıklanmaya başlanan ve günümüzde de halen devam eden Framingham kalp çalışmasında tanımlanmıştır. Bu çalışmada, başta hiperkolesterolemi ve hipertansiyon olmak üzere bazı risk faktörlerinin kardiyovasküler hastalık riskini artırdığı ortaya konmuş ve günümüze kadar birçok risk faktörü tanımlanmıştır (Kannel vd., 1961). Koroner arter hastalığının görülme sıklığının ve buna bağlı ölüm oranlarının azaltılabilmesi öncelikle risk faktörlerinin kontrol altına alınmasını gerektirmektedir. Koroner kalp hastalığında hastalık seyrinin

iyileştirilmesinin çoklu ilaç tedavisi ile birlikte temel kardiyovasküler risk faktörlerinin kontrol edilmesi ve yaşam tarzında yapılacak değişikliklerle mümkün olabileceği düşünülmektedir. Bu yaklaşım aterojenik süreci ve kronik inflamasyonu yavaşlatarak, kardiyovasküler olay sayısı ile stent ve cerrahi tedavi gereksinimin azaltılmasına yardımcı olur. 52 ülkede yürütülen ve büyük çaplı bir vaka kontrol çalışması olan "INTERHEART" çalışmasında akut kalp krizi riskinin düzeltilebilir risk faktörleri ile ilişkili olduğu belirtilmektedir. Bu çalışmada, belirtilen risk faktörleri; sigara, hipertansiyon, dislipidemi, psikososyal stres, diyabetes mellitus, fiziksel aktivitede yetersizlik, beslenme alışkanlığı problemleri ve artmış bel çevresi/kalça oranıdır (Yusuf vd., 2004). Ulusal Kolesterol Eğitim Programı (NCEP) yetişkin panelinde, koroner arter hastalığı için yapılan risk faktörleri sınıflandırması yaygın olarak kabul görmektedir (NCEP, 2002).

Koroner Arter Hastalığı İçin Tanımlanan Risk Faktörleri (NCEP, 2002)

1. **Lipid Risk Faktörleri**
    - LDL Yüksekliği
    - Trigliseridler
    - HDL Düşüklüğü
2. **Lipid Olmayan Risk Faktörleri**
    - **A. Değiştirilebilir Risk Faktörleri**
        - Hipertansiyon
        - Tütün Kullanımı
        - Diyabet
        - Obezite
        - Yetersiz Fiziksel Aktivite
        - Aterojenik Beslenme
        - Trombüs Eğilimi Oluşturan Durumlar
    - **B. Değiştirilemeyen Risk Faktörleri**
        - Yaş
        - Cinsiyet
        - Aile Öyküsü

Literatüre bakıldığında, son yıllarda, belirtilen risk faktörleri dışında yeni risk faktörlerinin de tanımlandığı görülmektedir (Adalet, 2013; Bonow vd., 2015; Onat, 2017).

Yeni Risk Faktörleri:

A. İnflamasyon Göstergeleri

- C-Reaktif Protein
- Diğer Göstergeler (adezyon molekülleri, IL miyeloperoksidaz)

B. Hiperkoagülasyon Göstergeleri

- PAI
- t-PA
- Fibrinojen

C. LDL Partikül Büyüklüğü

D. Homosistein

E. Lipoprotein (a)

Kolesterol: Lipid risk faktörlerinden olan kolesterol düzeyinin yüksekliği ile kardiyovasküler hastalıklar arasındaki ilişki çok uzun zaman önce gösterilmiştir ve günümüzde koroner arter hastalığı için önemli bir risk faktörü olarak kabul edilmektedir. Framingham çalışması, total kolesterol seviyesi ile koroner arter hastalığına bağlı ölüm oranları arasında önemli bir ilişki olduğunu göstermiştir (Kannel, 1961). Bu sonuçlar, MRFIT (Multiple Risk Factor Intervention Trial) çalışmasında da gösterilmiştir (Stamler vd., 1986). Ateroskleroz sürecinin en önemli safhası koroner damarın intima tabakası altında LDL (Low Density Lipoprotein) birikmesi ve oksidasyonudur. Bununla birlikte, kanda LDL düzeyi arttıkça, oluşmuş olan aterom plağının boyutu da artmaya devam eder (Libby ve Theroux, 2005). Buna karşın, HDL (High Density Lipoprotein) ise köpük hücrelerinden kolesterolü uzaklaştırarak, LDL'nin oksidasyonunu baskılar ve inflamasyonu sınırlandırarak aterosklerozis sürecinde koruyucu rol üstlenir (Barter, 2005). Ayrıca, düşük HDL seviyesine sahip olmak da koroner arter hastalığı için diğer önemli bir risk faktörüdür.

Hipertansiyon: Hipertansiyon toplumda yaygın olarak görülmektedir. Hipertansiyon; koroner arter hastalığı, ani kalp ölümü ve inme açısından majör ve bağımsız risk faktörlerinden birisidir. TEKHARF çalışmasında Türkiye'deki hipertansiyon sıklığı

erişkinlerde yaklaşık % 34'tür. Aynı çalışmada, kan basıncındaki her 20 mmHg'lık artışın koroner kaynaklı olay riskini de yaklaşık %30 artırdığı gösterilmiştir (Onat, 2017). Hipertansiyon, damar duvarındaki gerilimi artırarak endotel fonksiyonlarını bozar. Ayrıca, hem plak oluşumunu hem de plak rüptürünü artırarak ateroskleroz sürecinde rol oynar (Adalet, 2013). Framingham çalışmasında, hipertansiyonu olan hastalarda koroner kalp hastalığı riski, hipertansiyonu olmayan hastalara göre 2 kat arttığı saptanmıştır. Ayrıca, aynı çalışmada hipertansif koroner arter hastalarında hastalık seyrinin daha kötü olduğu da gösterilmiştir (Kannel, 1961).

Tütün Kullanımı: Tüm dünyada, önlenebilir ölüm nedenlerinden en önemlisi tütün kullanımıdır. Dünyada 1.3 milyarın üzerinde tütün kullanıcısı bulunmakta ve bunun 1 milyarını sigara kullanıcıları oluşturmaktadır (Shafey vd., 2009). Sigara kardiyovasküler hastalıklar için bağımsız bir risk faktörüdür. Ayrıca, diğer risk faktörleri ile etkileşerek toplam koroner hastalığı riskini de artırmaktadır. Sigaranın neden olduğu ölümlerin yaklaşık %35-40'ını koroner arter hastalığı oluşturmaktadır. Ayrıca, günde yaklaşık bir paket sigara içenlerde içmeyenlere göre koroner arter hastalığı riski 2-3 kat daha fazladır. Sigara LDL kolesterol oksidasyonunu ve inflamasyonu, kan basıncını ve kalp hızını artırır, endotel fonksiyonlarını bozar ve trombosit kümeleşmesini artırarak aterosklerotik sürece katkıda bulunur (Adalet, 2013). TEKHARF çalışmasında, ülkemizde en yaygın görülen risk faktörünün sigara olduğu ve erkeklerde sigara içiminde azalma gözlenirken kadınlarda ise artış olduğu gösterilmiştir (Onat, 2017). Buna ek olarak, sigara içimi bırakıldıktan sonra ise koroner arter hastalığı riski hemen azalmaya başlamaktadır.

Diyabet: Diyabet koroner arter hastalığı için hem erkeklerde hem de kadınlarda, bağımsız, önemli bir risk faktörüdür. Risk, erkeklerde 2 kat, kadınlarda ise 4 kat artmıştır. Diyabetik hastalarda gelişen koroner olayların seyri diyabeti olmayanlara göre daha kötüdür. Ülkemizde diyabetik kişi sayısı hızla artmaktadır. Diyabetik hasta sayısı, ülkemizde yaklaşık olarak her yıl 240 bin kişi civarında artış göstermektedir (Onat, 2017). Diyabetin, ateroskleroz sürecine etkisi damar sistemindeki metabolik etkileri ile beraber lipid metabolizmasındaki olumsuz etkileri ve diğer risk faktörleri ile daha sık birlikte olması ile açıklanabilir (Adalet, 2013). Henüz diyabet tanısı almamış, diyabetin gelişim sürecinde yer alan bozulmuş açlık glikozu ve bozulmuş

glikoz toleransına sahip hastalarda da kardiyovasküler riskin artma eğiliminde olduğu gösterilmiştir (Coutinho vd., 1999).

Obezite: Obezite**,** günümüz dünyasının en önemli sağlık problemlerinden biridir. Obezite hem kadın hem de erkeklerde koroner arter hastalığı açısından bağımsız risk faktörlerinden biridir. Obeziteye bağlı risk artışı direk obeziteye bağlı olabileceği gibi neden olduğu insülin direnci, hipertansiyon ve lipid metabolizması bozukluğuna bağlı olarak da ortaya çıkabilmektedir (Adalet, 2013). TEKHARF çalışmasında ülkemizde obesite prevelansının 30 yaş üzerindeki kadınlarda %44,2, erkeklerde %25,3 olduğu ve obesite prevalansının son 10 yıl içinde %20 oranında artış gösterdiği saptanmıştır (Onat, 2017). Abdominal obezite olarak da bilinen ve yağ dokusunun özellikle karında birikmesi sonucu oluşan obezite tipinde diğer risk faktörleri ile birliktelik daha fazladır. Bu obezite tipinde, hastalarda, kardiyovasküler olaylar ile arasında güçlü ilişki bulunan bel çevresi ölçümü artmıştır (Adalet, 2013).

Fiziksel Hareketsizlik (İnaktivite): Fiziksel hareketsizlik koroner arter hastalığı gelişimi açısından önemi giderek artan bir risk faktörü olarak görülmektedir. Sedanter yaşam tarzı, koroner kalp hastalığı riskini yaklaşık 2 kat artırmaktadır. Düzenli fiziksel aktivitenin kardiyovasküler risk faktörleri üzerinde olumlu etkiler göstermesi yanında bağımsız olarak da riski azalttığı bilinmektedir. Fiziksel aktivite ile birlikte kan basıncında, serum total kolesterol, LDL ve trigliserid düzeylerinde azalma, serum HDL düzeyinde ise artma ve kilo kaybı görülmektedir (Scrutinio vd., 2005; Onat, 2017).

Yaş ve Cinsiyet: Yaş artışı ile birlikte koroner arter hastalığında hem prevelans (yaygınlık) hem de insidans (sıklık) artışı olmaktadır. Bu durum, koroner arter hastalığında yaşı en önemli risk faktörü haline getirmektedir. Koroner arter hastalığında, erkeklerde 45 yaş, kadınlarda ise 55 yaş üzerinde olmak önemli bir risk faktörü olarak kabul edilmektedir (Bonow vd., 2015). TEKHARF çalışmasında, ülkemizde, her 11 yıllık yaşlanma ile koroner kalp hastalığı ihtimalinin yaklaşık 1,5 kat arttığı gösterilmiştir (Onat, 2017). Koroner arter hastalığının önemli risk faktörlerinden biri de cinsiyettir. Kadınlarda menopoz öncesi dönemde koroner kalp hastalığı erkeklere göre 4 kat daha az görülmektedir. Menopoz sonrası dönemde ise

prevalans eşitlenmektedir. Menopoz öncesindeki koroner arter hastalığının daha az görülmesi östrojenin lipid göstergeleri üzerindeki olumlu etkisi ile açıklanabilmektedir. Ancak, menopoz öncesinde sigara, hipertansiyon ve diyabet gibi risk faktörlerinin olması kadınlarda koroner kalp hastalığı insidansını erkeklerle benzer düzeye getirebilmektedir (Adalet, 2013).

Aile Öyküsü: Koroner kalp hastalığı şüphesi ile kliniğe başvuran hastaların birinci dereceden erkek akrabalarında 55 yaş öncesi, birinci dereceden kadın akrabalarında ise 65 yaş öncesi koroner arter hastalığına ait öykü bulunmasının, koroner arter hastalığı görülme riskini yaklaşık olarak 2 kat artırdığı gösterilmiştir. Ayrıca bu risk artışı diğer risk faktörlerinden bağımsızdır (Williams vd., 1994).

Risk faktörlerinin aynı kişide, bir arada bulunması hastalık oluşumu riski açısından sinerjistik bir etki oluşturabilmektedir. Bu nedenle, koroner arter hastalıklarının önlenmesinde risk faktörlerinin ayrı ayrı değerlendirilmesi yerine tamamının birlikte değerlendirilmesi önem taşımaktadır. Henüz, klinik olarak koroner arter hastalığı gelişmemiş sağlıklı görünen bireylerde, koroner arter hastalığı riskinin hesaplanması için çeşitli risk modelleri geliştirilmiştir. Bu risk modelleri, esas olarak orta ve yüksek riskli kişileri erken safhada tanısal işlemlere yönlendirmede önemli rol oynamaktadır. Ayrıca, bu bireylerde, risk faktörlerine yönelik yoğun tedavi yaklaşımı ile ileride gelişebilecek koroner arter hastalığının önlenebilmesi ve hastalık seyrinin yavaşlatılması hedeflenmektedir.

Koroner arter hastalığı riskinin değerlendirilmesinde farklı risk modelleri kullanılabilmektedir. Bunlar içerisinde en sık olarak Framingham Risk Modeli, SCORE Risk Modeli ve JOINT British Societies-2 Risk Modeli kullanılmaktadır Framingham Risk Modeli, temel olarak 10 yıllık koroner olay riskini göstermektedir. Bu model, bireyleri; düşük, orta ve yüksek riskli olmak üzere üç gruba ayırmaktadır. SCORE modeli ise, 10 yıllık ölümcül kardiyovasküler hastalık riskini değerlendirmektedir. Bu model, değerlendirmede; yaşı, cinsiyeti, sigara kullanımını, kan basıncını ve toplam kolesterolü temel almaktadır. Bireyler, bu modelde 10 yıllık kardiyovasküler hastalık riski açısından çok yüksek, yüksek, orta ve düşük riskli gruplara ayrılmaktadır. JOINT British Societies-2 modeli, 10 yıllık kardiyovasküler

hastalık riskini değerlendirmektedir. Risk hesaplanmasında, cinsiyet, sigara, kan basıncı, total kolesterolün HDL'ye oranı kullanılmaktadır. Risk grupları SCORE modeli ile benzerdir. Ayrıca, bu model ülkemiz için de en uygun risk modeli olarak kabul edilmektedir (Erkuş vd., 2013).

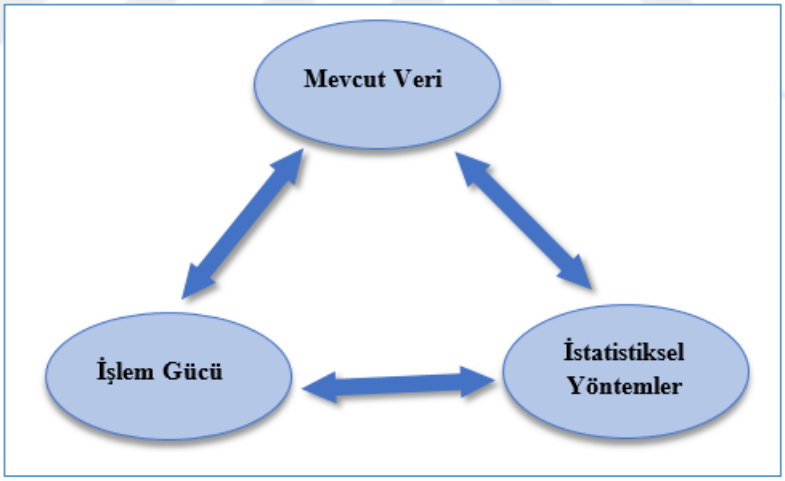
Koroner arter hastalığı riskinin analiz edilmesinde, risk faktörleri ve hasta sayısı arttıkça, risk modelleri ile analiz yapmak oldukça karmaşık ve zaman alıcı bir problem olabilmektedir. Bu nedenle, son yıllarda, makine öğrenmesi gibi, çok sayıda hasta verisinin, aynı anda ve kısa sürede, yüksek doğruluk oranı ile analizine imkân tanıyan yaklaşımlara ilgi artmaktadır. Makine öğrenmesi yaklaşımının klinik kullanımı, riskli hastaların erken dönemde belirlenerek ileri tanı işlemlerine yönlendirilmelerini sağlayabilmektedir.

1.2. Makine Öğrenmesi

Makine öğrenmesi, bilgisayarların verilerden öğrenmesine dayanan bir disiplin olarak tanımlanmaktadır. Makine öğrenmesinin temel çalışma alanı, bilgisayar programlarının karmaşık örüntüleri tanımayı otomatik olarak öğrenmesi ve edindiği bu verilere dayanan zeki kararlar alabilmesidir (Han vd., 2011). Makine öğrenmesi veriler arasındaki ilişkileri araştıran istatistik disiplini ile etkili hesaplama algoritmaları üzerinde çalışan bilgisayar bilimlerinin kesişim noktasında ortaya çıkmaktadır (Deo, 2015). Makine öğrenmesinin en önemli görevi örneklerden çıkarım yapabilmektir. Bu nedenle, matematiksel modelin kurulmasında istatistiksel kuramlardan yararlanır. Bilgisayar bilimlerine ise iki temel alanda gereksinim duyulmaktadır. Bunlar; büyük miktarlardaki veriyi işlemek kadar, eğitim aşamasında, optimizasyon problemlerini çözebilecek etkili algoritmaların oluşturulması ve model öğrenildikten sonra, bu modelden çıkarım yapabilmek için gereken algoritma performansının iyileştirilmesidir. Bazı özel uygulama alanlarında, öğrenme veya çıkarım algoritmasının verimliliği, yani uzay ve zaman karmaşıklığı, öngörü doğruluğu kadar önemli olabilmektedir (Alpaydın, 2010).

Makine öğrenmesi; bir bilgisayar programı, eğer P ile ölçülen, G görevindeki performansını, E deneyimi tarafından geliştirirse, bu bilgisayar programının deneyimlerinden öğrendiği söylenir şeklinde tanımlanmaktadır. Bu tanımlamadan yola çıkarak bir bilgisayar programının deneyimi ile birlikte performansı da artış gösteriyorsa makine öğrenmesinden söz edilebilmektedir (Mitchel, 1997) .

Makine öğrenmesi, verileri akıllı eylemlere dönüştürmek amacıyla bilgisayar algoritmalarının geliştirilmesi ile ilgilenen çalışma alanı olarak da tanımlanabilir Makine öğrenmesi alanı, mevcut verilerin, istatistiksel yöntemlerin ve işlem gücünün hızla ve eşzamanlı olarak geliştiği bir ortamda ortaya çıkmıştır. Verilerdeki büyüme, ek işlem gücünü gerektirmiş ve bu da büyük veri kümelerini analiz etmek için istatistiksel yöntemlerin geliştirilmesine zemin oluşturmuştur. Bu durum ise, daha büyük ve farklı alanlardan verilerin toplanmasını sağlayan bir ilerleme döngüsünün ortaya çıkmasını sağlamıştır (Şekil 1.3.).



Şekil 1.3. Makine öğrenmesinde bileşenler arasındaki ilişki (Lantz, 2013).

Makine öğrenmesi ile yakından ilişkili olan veri madenciliği, büyük veri tabanlarından yeni örüntülerin keşfedilmesi ile ilgilenmektedir. Makine öğrenmesi ve veri madenciliğinin ayrıldığı temel nokta; makine öğrenmesi, bilgisayarlara bir problemi çözmek için verileri nasıl kullanacaklarını öğretmeye odaklanırken, veri madenciliği

bilgisayarlara verilerdeki örüntüleri nasıl tanıyacaklarını öğretmeye odaklanır. Bu örüntüler daha sonra insanlar tarafından problemlerin çözümü için kullanılmaktadırlar. Veri madenciliği uygulamalarında, verideki örüntüleri keşfedebilmek için makine öğrenmesi algoritmaları kullanılmaktadır (Lantz, 2013). Makine öğrenmesi, aynı zamanda yapay zekânın bir parçasıdır. Bir sistemin zekiliğinden söz edebilmek için, değişen bir çevrede öğrenme yeteneğine sahip olması gerekmektedir. Sistem, öğrenir ve değişikliklere uyum sağlarsa, sistem tasarımcısı tüm olası durumlar için öngörü ve çözüm sunma gereksinimi duymaz (Alpaydın, 2010).

1.2.1. Makine Öğrenmesinin Tarihsel Gelişimi

Yıllar içerisinde makine öğrenmesi çalışmaları, farklı yaklaşım ve hedeflere sahip üç farklı dönemden geçmiştir. Bu dönemler:

- Nöral modelleme ve karar-teorik teknikler
- Sembolik kavram merkezli öğrenme
- Yoğun bilgi öğrenme sistemleri

Bunlardan ilk dönem olan nöral modelleme aşaması, genel amaçlı öğrenme sistemlerinin oluşturulmasına odaklanmıştır. Bu dönemdeki çalışmalarda daha çok rastsal ya da kısmi rastsal başlangıç yapısına sahip nöral model tabanlı nöral ağ (neural nets) ya da kendi kendini organize eden sistemler (self-organizing systems) olarak isimlendirilen makinalar oluşturulmuştur. İlk dönemin yaşandığı 1950'li yıllarda bilgisayar teknolojisinin ilkel doğası gereği, bu dönemde yapılan çalışmalar, kavramsal yapıda ya da Rosenblatt'ın tek katmanlı algılayıcısı (perceptron) gibi özel amaçlı deneysel donanım sistemlerinin kurulması yönünde planlanmıştır. Yapılan bu çalışmalardan elde edilen tecrübelerle örüntü tanıma disiplini ortaya çıkmış ve makine öğrenmesinde karar –teorik yaklaşımlar geliştirilmiştir.

1960'lı yılların başlarında, ikinci dönem olan sembolik kavram merkezli öğrenme, insan öğrenmesini modelleyen yapay zekâ çalışmaları ile birlikte ortaya çıkmıştır. Bu dönemde istatistiksel ya da numerik yöntemler yerine mantık ya da grafik yapısında ifadeler kullanılmaya başlanmıştır. Sistemler, daha yüksek düzeyde bilgiyi temsil eden sembolik tanımları öğrenmiş ve öğrenilecek kavram hakkında güçlü bir yapısal

varsayım sağlamaya başlamıştır. Çeşitli örüntü tanıma sistemleri bu aşamada geliştirilen sistemlere örnek olarak verilebilir.

Son aşama ise 1970'li yılların ortalarında başlayan güncel dönemi temsil etmektedir. Bu dönemde araştırmacılar öğrenme yöntemlerini geniş bir yelpazede ele alarak yoğun bilgi öğrenme sistemleri, alternatif öğrenme yöntemleri, öğrenme görevlerini oluşturmak ve seçmek için yetenekleri birleştirmek gibi alanlarda çalışmalarını yürütmektedirler (Michalski vd., 2013).

1.2.2. Makine Öğrenmesi Türleri

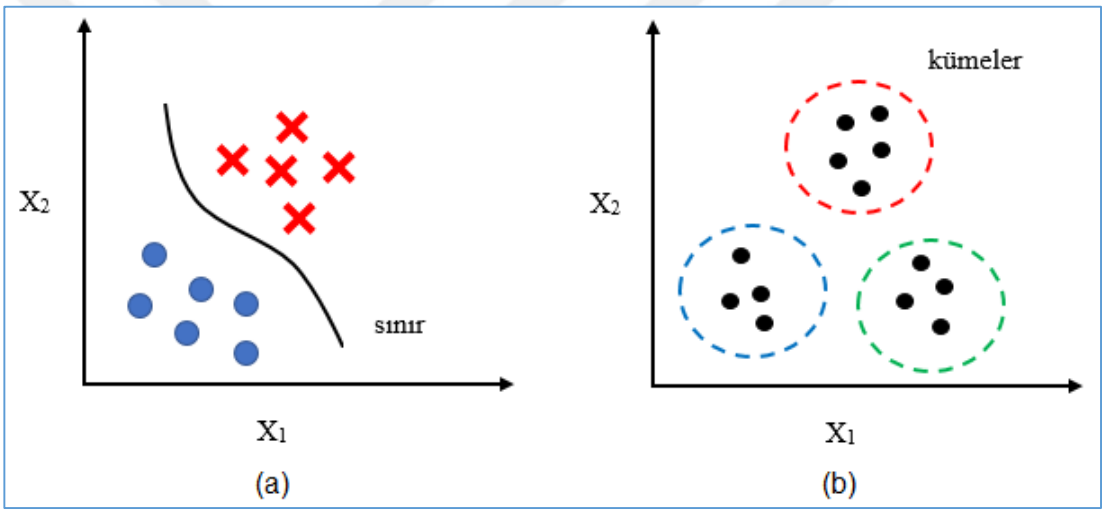
Öğrenme en genel haliyle bilgi edinme sürecidir. İnsanlar akıl yürütme yetenekleri sayesinde deneyimlerinden öğrenebilirler. Ancak, bilgisayarlar akıl yürütme yetenekleri olmadığı için algoritmalarla öğrenirler. Günümüzde, çok sayıda makine öğrenmesi algoritması bulunmaktadır. Bu algoritmalar kullandıkları öğrenme sürecine göre sınıflandırılabilmektedir (Portugal vd., 2017). Bu bağlamda, makine öğrenmesi algoritmaları dört ana sınıfta toplanabilmektedir. Bunlar; denetimli (supervised), denetimsiz (unsupervised), yarı denetimli (semi-supervised) ve takviyeli (reinforcement) öğrenmedir.

Denetimli (Supervised) Öğrenme:

Denetimli öğrenme, bilinen bir çıktının ya da hedefin tahmin edilmesiyle başlamaktadır (Deo, 2015). Denetimli öğrenme, makine öğrenmesi algoritmalarının eğitim verilerinden öğrenmesi ve gerçek veriler yoluyla kazandığı bu bilgiyi uygulamaya aktarmasına dayanmaktadır (Portugal vd., 2017). Bu öğrenme yaklaşımındaki "denetimli" öğrenme, eğitim veri kümesinde yer alan etiketli verilerden kaynaklanmaktadır (Şekil 1.4.a.). Modelin öğrenilmesi amacıyla kullanılan veri kümesine eğitim veri kümesi denir. Model eğitim veri kümesi üzerinde kurulur. Kurulan modelin performansının ölçüldüğü veri kümesine test veri kümesi ismi verilmektedir. Denetimli öğrenme yaklaşımı sıklıkla sınıflandırma, modelleme, sinyal işleme ve optimizasyon alanlarında kullanılmaktadır (Du ve Swamy, 2013).

Denetimsiz (Unsupervised) Öğrenme:

Denetimsiz öğrenmede, denetimli öğrenmenin aksine tahmin edilecek bir hedef değer ya da çıktı bulunmamaktadır. Bunun yerine veri kümesinde doğal olarak oluşmuş örüntü ve gruplanmalar bulunmaya çalışılmaktadır (Şekil 1.4.b.) (Deo, 2015). Denetimsiz öğrenmede herhangi bir eğitim veri kümesi bulunmamaktadır. Denetimsiz öğrenme algoritmaları sıklıkla veri kümesindeki gizli örüntüleri keşfetmeye odaklanırlar (Portugal vd., 2018). Denetimsiz öğrenme yaklaşımı sıklıkla, kümeleme, vektör kuantalama, öznitelik çıkarma, sinyal işleme ve veri analizi alanlarında kullanılmaktadır (Du ve Swamy, 2013).

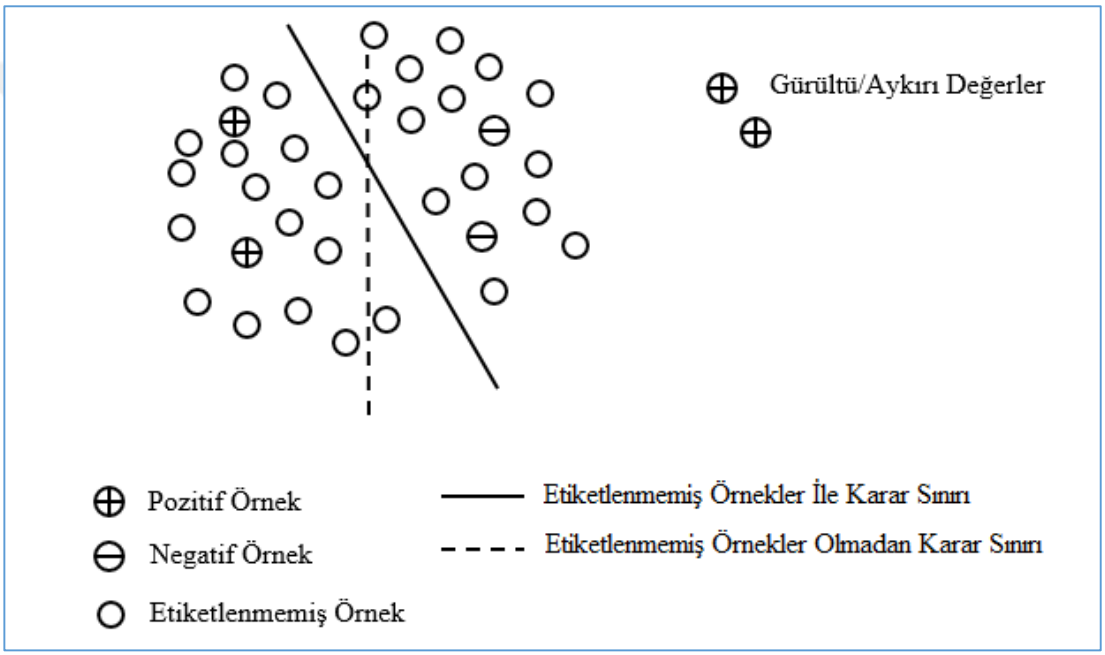


Şekil 1.4. Makine öğrenmesi türleri (**a**. Denetimli öğrenme **b.** Denetimsiz öğrenme)

Yarı Denetimli (Semi-Supervised) Öğrenme:

Biyoinformatik, web ve metin madenciliği, metin spam tespiti, yüz tanıma, metin kategorizasyonu gibi birçok makine öğrenmesi uygulaması, manuel etiketlemenin zaman alıcı bir işlem olması nedeniyle, çok miktarda etiketsiz veri içermektedir. Aynı zamanda, etiketsiz verilerin elde edilmesi, etiketli verilere göre daha kolay olmaktadır. Denetimli öğrenmenin aksine; yarı denetimli öğrenme hem etiketli hem de etiketsiz

verileri kullanabilmektedir (Şekil 1.5.). Yarı denetimli öğrenmenin amacı, az sayıda etiketli örnekle birlikte çok miktardaki etiketlenmemiş verinin bir araya getirilerek genelleme performansının iyileştirilmesidir (Du ve Swamy, 2013). Yarı denetimli öğrenme algoritmaları tamamlanmamış veri kümelerinden öğrenebilir ve sonuç çıkarabilirler (Portugal vd., 2018). Yarı denetimli öğrenme yaklaşımı sıklıkla beklendik en büyükleme (expectation maximization), kendi kendine eğitim (self-training), transdüktif destek vektör makineleri (transductive support vector machine) ve çizge tabanlı (graph-based) yöntemlerde kullanılmaktadır (Du ve Swamy, 2013).



Şekil 1.5. Yarı denetimli öğrenme yaklaşımı (Han vd., 2011).

Takviyeli (Reinforcement) Öğrenme:

Bazı uygulamalar için, sistemin çıkışı bir dizi eylemden oluşmaktadır. Bu durumda, önemli olan tek bir eylem değil, hedefe ulaşmada doğru eylemlerin sırası için izlenen politikadır. Herhangi bir ara aşamada en iyi eylem diye bir şey yoktur. Eylem iyi bir politikanın bir parçası ise iyidir. Böyle bir durumda, makine öğrenmesi algoritması, politikaların iyiliğini değerlendirebilmeli ve bir politika üretebilmek için geçmiş iyi

eylem dizilerinden öğrenebilmelidir. Bu tür öğrenme yöntemlerine takviyeli öğrenme denir (Alpaydın, 2010). Takviyeli öğrenme yaklaşımında, algoritma dışarıdan alınan geri bildirimlerle öğrenmektedir (Portugal vd., 2018). Takviyeli öğrenme, bir yapay ajanın (örneğin, gerçek veya simüle edilmiş bir robot) beklenen toplam ödülü maksimize etmek için eylemlerini seçmeyi nasıl öğrenebileceğini belirten bir hesaplama algoritmaları sınıfıdır. Takviyeli öğrenme, istenen çıktının kesin olarak bilinmediği bir denetimli öğrenme yaklaşımı olarak da tanımlanabilmektedir. Eğitici, yalnızca cevabın başarısı veya başarısızlığı hakkında geri bildirim sağlar. Gerçek hayatta her zaman, tanımlı, tam olarak doğru bir yanıt, öğrenici veya eğitici mevcut olamayacağından, denetimli öğrenmeden daha mantıklı bir yaklaşım olduğu söylenebilir. Takviyeli öğrenme yalnızca, gerçek çıktının tahmine yakın olup olmadığı bilgisine dayanır. Takviyeli öğrenme iyi öğrenme sonucu için sinir ağını ödüllendiren, kötü çıktılar için ise sinir ağını cezalandıran bir öğrenme yaklaşımıdır. Bu öğrenme yaklaşımında, açık hesaplama türevleri gerekli değildir. Ancak, daha yavaş bir öğrenme süreci sağlamaktadır. Takviyeli öğrenme yaklaşımı sıklıkla kontrol ve yapay zeka alanında kullanılmaktadır (Du ve Swamy, 2013).

Yukarıda verilen ve yaygın olarak kullanılan makine öğrenmesi yaklaşımlarının sınıflamasına ek olarak literatürde meta öğrenme algoritmaları sınıfı olarak farklı bir sınıf daha tanımlanmaktadır. Bu grupta yer alan ve meta-öğrenenler olarak bilinen algoritmalar, belirli bir öğrenme görevine odaklanmazlar. Bu yaklaşım daha çok, etkili öğrenmeyi öğrenmeye dayanmaktadır. Meta öğrenme yaklaşımında, algoritma diğer öğrenmelerinin sonuçlarını kullanır. Bu öğrenme türü, zor problemler ya da tahminin, algoritma performansının olabildiğince iyi olması gerektiği durumlarda faydalı sonuçlar üretebilmektedir. Bununla birlikte, makine öğrenmesi yaklaşımları içerisinde kullanılan tek bir modelin performansını arttırmak amacıyla, birkaç model birleştirilerek güçlü bir ekip oluşturulabilmektedir. Birden fazla modelin tahminlerini birleştirme ve yönetme yaklaşımı, öğrenmeyi öğrenme tekniklerini tanımlayan meta öğrenme yöntemi içerisinde yer almaktadır. Meta öğrenme yaklaşımı bu görevi yerine getirirken topluluk (ensemble) öğrenme yöntemlerini kullanır. Tüm topluluk öğrenme yöntemleri, daha zayıf öğrenenleri bir araya getirerek, daha güçlü bir öğrenenin oluşturulması fikrine dayanmaktadır (Lantz, 2013). Topluluk öğrenme yaklaşımında

en yaygın kullanılan algoritmalar; torbalama (bagging), hızlandırma (boosting) ve rastgele ormandır (random forest) (Yang vd., 2010).

1.3. Topluluk Öğrenme Yaklaşımları

Topluluk öğrenme yaklaşımları, tek olan modellere göre bir dizi performans avantajı sunarlar (Lantz, 2013). Bunlar;

Genellenebilirlik: Çok sayıda öğrenenin görüşü tek bir son tahmine dâhil edildiğinden, tahminde tek bir yargı (bias) baskın olamaz. Bu, öğrenme görevini ezberleme (overfitting) olasılığını azaltır. Bununla birlikte, topluluk öğrenme yöntemleri, eğitim verileri üzerinde daha doğru bir sınıflandırma, görünmeyen veriler üzerinde daha iyi bir genelleme yapılabilmesini sağlamaktadır (Yang vd., 2010).

Geliştirilmiş performans: Çok büyük miktardaki veri kümeleri üzerinde çalışıldığında, birçok model bellek veya karmaşıklık kısıtlarıyla karşılaşmaktadır. Bu gibi durumlarda küçük modellerin eğitilmesi, tek bir modelin eğitilmesinden daha iyi bir çözüm sunabilmektedir. Ayrıca, dağıtılmış hesaplama yöntemlerini kullanarak bir topluluğu paralel olarak eğitmek de mümkün olabilmektedir.

Farklı alanlardan gelen verileri sentezleyebilme: Tüm öğrenme algoritmalarına uygun tek bir boyutta veri kümesi olmadığından, birden fazla öğreniciden gelen bilgileri birleştirebilme kabiliyeti olan topluluk öğrenme algoritmalarının yeteneği, özellikle farklı alanlardan elde edilen verilere dayanan karmaşık durumlar için önem taşımaktadır.

Zorlu öğrenme görevlerinde daha detaycı bir yaklaşım: Gerçek yaşam, birçok faktörün etkileşim içinde olduğu son derece karmaşık durumları içermektedir. Öğrenme görevini küçük parçalara bölen modeller, tek bir küresel modelin gözden kaçırabileceği ince örüntüleri ve detayları daha doğru bir şekilde tespit edebilmektedir.

1.4. Literatür Çalışmaları

Literatürde son yıllarda koroner arter hastalığının makine öğrenmesi algoritmaları kullanılarak analiz edilmesi ile ilgili çalışmalar yer almaktadır. Çalışmalar incelendiğinde, literatürde daha çok, birden fazla makine öğrenmesi algoritmasının bir arada kullanılarak sınıflama doğruluğu açısından performanslarının karşılaştırıldığı çalışmalar olduğu görülmektedir. Shafique vd. (2015), çalışmalarında, 597 hasta kaydından oluşan UCI kalp hastalıkları veri kümesi üzerinde Yapay Sinir Ağları, Karar Ağacı ve Naive Bayes sınıflandırma algoritmaları ile koroner kalp hastalığı riskini belirlemişlerdir. Çalışmada, en yüksek sınıflama doğruluğu %82,914 oranı ile Naive Bayes algoritmasından elde edilmiştir. Marikani ve Shyamala (2017) çalışmalarında, kalp hastalığının varlığını tahmin etmek amacıyla denetimli öğrenme algoritmalarını kullanmışlardır. Çalışmada, Cleveland veri kümesinden kayıp verilerin olduğu kayıtların çıkarılması ile 297 hasta kaydından oluşan veri kümesi üzerinde sınıflandırma yapılmıştır. Araştırmacılar en yüksek doğruluk oranını Destek Vektör Makinesi algoritmasından elde etmişlerdir. Sharma vd. (2017) kalp hastalığı riskinin belirlenmesi amacıyla yaptıkları çalışmada Karar Ağacı, Naive Bayes ve Yapay Sinir Ağı algoritmaları kullanmışlardır. En yüksek doğruluk oranını 15 değişken ile uyguladıkları Yapay Sinir Ağından elde etmişlerdir.

Shamsollahi vd. (2018) çalışmalarında, kalp hastalıkları kliniğine başvuran 282 hasta ve 21 değişkenden oluşan veri kümesi üzerinde sınıflama ve kümeleme algoritmalarını uygulayarak kalp hastalığı varlığını tahmin etmişlerdir. Çalışmacılar, öncelikle veri kümesini k- ortalamalar yöntemi ile 3 kümeye ayırmışlardır. Kümeleme işleminde kullanılan k değeri için ortalama Siluet, Dunn İndeks, Dirsek (elbow) algoritması gibi yöntemler kullanılmıştır. Veri kümesi 90, 88 ve 104 kayıt içerecek biçimde 3 kümeye ayrılmıştır. Her bir kümeye Yapay Sinir Ağı ve Karar Ağacı algoritmaları uygulanmıştır. Çalışmada, en iyi sınıflama performansı Sınıflama ve Regresyon Karar Ağacı algoritmasından elde edilmiştir.

Koroner kalp hastalığının, makine öğrenmesi yaklaşımları ile tahmin edilmesi amacıyla yapılan çalışmaların çoğunda hastaların risk faktörlerine ilişkin bilgilerini içeren veri kümeleri kullanılmaktadır. Bu çalışmalarda, sınıflama ya da tahmin

performansını artırmak amacıyla sıklıkla tüm değişkenlerle bir model oluşturmak yerine önemli ya da sonuca daha fazla katkı sağlayan değişkenler seçilmektedir. Anbarasi vd. (2010) yaptıkları çalışmada, 13 değişken ve 909 hasta kaydından oluşan UCI kalp hastalıkları veri kümesi üzerinde kalp hastalığı tahmin sistemi geliştirmişlerdir. Araştırmacılar, Genetik Algoritma ile değişken sayısını azaltmışlardır. Bu değişkenler; göğüs ağrısı tipi, istirahat kan basıncı, egzersizle tetiklenen anjina, ST depresyon, floroskopide boyanan damar sayısı ve ulaşılan maksimum kalp hızıdır. Sonrasında, veri kümesi üzerinde, Naive Bayes, Kümeleme ve Karar Ağacı algoritmaları uygulanmıştır. Çalışmada, Karar Ağacı algoritması ile %99,2 doğruluk oranı elde edilmiştir. Karar Ağacı sınıflandırma algoritmasını, Naive Bayes ve Kümeleme algoritmaları izlemiştir.

Abdullah (2012), çalışmasında geliştirdiği sınıflama modeli ile Cleveland veri kümesi üzerinde koroner kalp hastalığı riskini incelemiştir. Veri kümesinde yer alan 14 değişken Parçacık Sürü Optimizasyon (PSO) algoritması ile 9 değişkene indirgenmiştir. Bunlar; hastanın yaşı, cinsiyeti, göğüs ağrısının tipi, serum kolesterolü ve açlık kan şekeri düzeyi, istirahat EKG bulguları, ulaşılan maksimum kalp hızı, floroskopide boyanan büyük damar sayısı ve defekt tipidir. Araştırmada, elde edilen veri kümesi üzerinde uygulanan J48 Karar Ağacı algoritması ile %60,74 sınıflama doğruluğu elde edilmiştir. Chaurasia (2013) yaptığı çalışmada, UCI Cleveland veri kümesindeki her bir değişkenin sınıflama performansı açısından önemini analiz etmiştir. Değişkenlerin analizi sırasında Ki-kare, Bilgi Kazanımı ve Kazanç Oranı testleri kullanılmıştır. Her bir değişken için üç testin ortalaması alınarak değişkenler önem derecesine göre sınıflandırılmıştır. Çalışmada, göğüs ağrısı tipi, ST segment eğimi, egzersizle tetiklenen anjina, istirahat EKG bulguları, cinsiyet, yaş, ulaşılan maksimum kalp hızı, açlık kan şekeri ve kolesterolün sınıflama açısından en önemli değişkenler olduğu belirlenmiştir. Nahar vd. (2013), yaptıkları çalışmada, Cleveland veri kümesi üzerinde birliktelik kurallarını kullanarak koroner kalp hastalığını etkileyen faktörleri cinsiyet açısından analiz etmişlerdir. Araştırmacılar, asemptomatik göğüs ağrısı ve egzersizle tetiklenen anjinanın hem kadın hem de erkekler için koroner kalp hastalığı varlığı açısından önemli bir gösterge olduğunu belirlemişlerdir. Bununla birlikte, kalp hastalığı varlığının tanımlanmasında istirahat EKG bulgularının kadınlar açısından önemli bir ayırıcı faktör olduğu saptanmıştır. Ayrıca, çalışmada yukarı ST

segment eğiminin varlığı, floroskopide boyanan damar sayısının 0 olması ve egzersizle tetiklenen ST depresyonunun 0,56'dan az olmasının her iki cinsiyet açısından sağlıklı koşulları ifade ettiği de gösterilmiştir.

Mukherjee vd. (2017) kalp hastalıklarının ve risk faktörlerinin tanımlanması amacıyla yaptıkları çalışmada, Destek Vektör Makinesi, Çok Katmanlı Yapay Sinir Ağı gibi iki sınıflama algoritması ve Genelleştirilmiş Katkı Modeli (GAM) gibi bir ileri regresyon yöntemini 270 kayıttan oluşan Statlog veri kümesi üzerinde uygulamışlardır. Çalışmada, Destek Vektör Makinesi ve GAM ile yapılan duyarlılık analizi sonucunda, ulaşılan maksimum kalp hızı, floroskopide boyanan damar sayısı, ST segment eğimi, göğüs ağrı tipi ve talyum tarama testi sonuçlarının en önemli faktörler olduğu saptanmıştır. Ahmadi vd. (2017) yaptıkları çalışmada, Cleveland veri kümesi üzerinde Sinir Ağı ve C5.0 Karar Ağacı algoritması uygulayarak kalp hastalıkları tahmini için bir karar modeli geliştirmişlerdir. Sinir ağı modelinin uygulanması sırasında değişkenler ve sonuca etkisi arasındaki ilişkiyi ortaya çıkarmak için ortalamaya dayalı Duyarlılık Analizi yapılmıştır. Duyarlılık Analizi sonuçlarına göre, kalp hastalığının tahmin edilmesinde, en fazla katkıyı yapan değişkenler floroskopide boyanan damar sayısı ve talyum sintigrafi sonucu iken, en az katkı sağlayan değişken kolesterol seviyesi olarak saptanmıştır. Bununla birlikte, C5.0 algoritması için yapılan öznitelik önemi ölçüm sonuçlarına göre, benzer şekilde kolesterol seviyesi en az öneme sahip değişken olarak saptanırken, floroskopide boyanan damar sayısı, talyum sintigrafi sonucu ve göğüs ağrısı tipi en önemli değişkenler olarak belirlenmiştir. Kolesterol seviyesi değişkeni çıkarılarak yapılan sınıflama sonuçlarına göre sinir ağı algoritmasının %95 güven aralığında C5.0 algoritmasından daha iyi bir performans gösterdiği belirlenmiştir.

Takcı (2018), öznitelik seçme yöntemleri ile kalp krizinin tahmininin iyileştirilmesi amacıyla yaptığı çalışmada öznitelik seçimi ve makine öğrenmesi algoritmalarını birlikte kullanarak en iyi performans gösteren makine öğrenmesi ve öznitelik seçimi algoritmalarını belirlemişlerdir. Çalışmada, UCI Statlog veri kümesi üzerinde, 12 farklı sınıflama ve 4 farklı öznitelik seçimi algoritması kullanılmıştır. Çalışmada kullanılan sınıflama algoritmaları dört ayrı kategoride incelenmiştir. Bunlar; Regresyon Analiz Modelleri, Destek Vektör Makineleri, Karar Ağaçları ve k- En

Yakın Komşuluk, Çok Katmanlı Yapay Sinir Ağı ve Naive Bayes algoritmalarıdır. Öznitelik seçmek amacıyla Fisher Filtreleme, Relieff, Backward-Logit Ve Forward-Logit algoritmaları kullanılmıştır. Araştırma bulgularına göre, en iyi performansa sahip makine öğrenme algoritması, Lineer Kernel ile Destek Vektör Makinesi algoritmasıdır, Öznitelik Seçim algoritması ise reliefF yöntemidir. Bu çiftten oluşan model % 84,81'lik oran ile en yüksek doğruluk değerini vermiştir.

Prakash vd. (2018) yaptıkları çalışmada UCI veri kümesi koleksiyonundan alınan kalp hastalıkları veri kümesinde yer alan değişkenlerin azaltılması amacıyla optimal kriterler geliştirmişlerdir. Çalışmada veri kümesinde yer alan değişkenler ile karar çizelgesi oluşturulmuş ve kalp hastalıkları riskinin belirlenmesinde gerekli olmayan değişkenler çıkarılarak işlem zamanının azaltılması sağlanmıştır. Göğüs ağrısı tipi ve istirahat EKG bulguları gibi değişkenler karar vermede dikkate alınan değişkenler olarak seçilmiştir.

Literatürde, koroner kalp hastalığı ile ilgili makine öğrenmesi çalışmalarında sıklıkla Cleveland veri kümesinin kullanıldığı görülmektedir. Ancak, farklı veri kümeleri üzerinde model geliştiren birçok çalışmaya da rastlanmıştır. Alizadehsani vd. (2013), çalışmalarında koroner arter hastalığı riskinin belirlenmesinde veri madenciliği algoritmalarını kullanarak bir sınıflama modeli geliştirmişlerdir. Çalışmada, 303 hasta ve 54 değişkenden oluşan Z-Alizadeh Zani veri kümesi kullanılmıştır. Veri kümesinde yer alan değişkenler medikal literatür incelemesi sonucunda belirlenen; demografik veriler, semptom ve muayene bulguları, EKG ve laboratuvar ve EKO bulguları olarak gruplandırılan değişkenlerden oluşturulmuştur. Destek Vektör Makinesi ve ağırlıklandırma yöntemi ile 0,6 ve daha fazla ağırlığa sahip 34 değişkenden oluşan veri kümesine, Sıralı Minimum Optimizasyon (SMO), Naive Bayes, Torbalama ve Sinir Ağları algoritmaları uygulanmıştır. Çalışmada en yüksek doğruluk oranı Sıralı Minimum Optimizasyon algoritmasından elde edilmiştir.

Masethe ve Masethe (2014) çalışmalarında, 108 hasta kaydı ve bu hastalara ilişkin cinsiyet, EKG bulguları, yaş, göğüs ağrısı tipi, kan basıncı, kalp hızı, kolesterol, sigara ve alkol tüketimi, diyet ve açlık kan şekeri seviyesi bilgilerinden oluşan veri kümesi üzerinde J48 Karar Ağacı, Bayes Net, Naive Bayes, Simple Cart ve REPTREE

algoritmaları uygulayarak sınıflama yapmışlardır. Çalışmada kullanılan algoritmaların sınıflama performansları birbirine benzer bulunmuştur. Tahminlerde yaklaşık % 97'nin üzerinde doğruluk oranı elde edilmiştir. Schlemmer vd. (2014) çalışmalarında EKG dalga özellikleri, yaş, cinsiyet, kalp hızı değişiklikleri gibi değişken bilgilerini içeren 261 hasta kaydı üzerinde makine öğrenmesi algoritmalarını uygulayarak kalp hastalığı tahmini yapan bir model geliştirmişlerdir. Araştırmada 15 ve daha fazla eksik veri içeren kayıtlar veri kümesinden çıkarılmış ve 87 değişkenden oluşan 227 hasta kaydı üzerinde k- En Yakın Komşuluk, Rastgele Orman ve Destek Vektör Makinası algoritmaları uygulanmıştır. Çalışmada, en yüksek doğruluk oranı Destek Vektör Makinesi algoritmasından elde edilmiştir.

Verma vd. (2016), koroner kalp hastalığını saptamak amacıyla yaptıkları çalışmada k-Ortalama Kümeleme ve Parçacık Sürü Optimizasyonu algoritmaları ile değişken alt kümesi seçimi yapmışlardır. Araştırmada, Yapay Sinir Ağı, Lojistik Regresyon, Bulanık Sırasız Kural Azaltma ve C4.5 algoritmaları kullanılarak karma bir model oluşturulmuştur. Çalışmacılar, geliştirdikleri karma modeli, 26 değişken ve 335 hasta kaydından oluşan veri kümesi üzerinde test etmişlerdir. En yüksek sınıflama doğruluğu %88,4 ile MLR algoritmasından elde edilmiştir. Arabasadi vd. (2017) yaptıkları çalışmada, koroner arter hastalığını klinik veriler üzerinden saptanmasına yönelik olarak Genetik Algoritma ve Yapay Sinir Ağlarından oluşan karma bir model önermişlerdir. Araştırmacılar, veri kümesi olarak 54 değişken ve 303 hasta kaydından oluşan Z-Alizadeh Sani veri kümesini kullanmışlardır. Çalışmada değişkenlerin seçimi Destek Vektör Makinesi yöntemi ile yapılmıştır. Araştırmacılar, veri kümesi üzerinde uyguladıkları karma model ile doğruluk oranı %93,85 olan bir sınıflama performansı elde etmişlerdir.

Koroner kalp hastalığının değerlendirilmesi ve riskinin belirlenmesi amacıyla yapılan çalışmalarda yapılandırılmış veri kümeleri dışında tanı işlemlerinden ya da hasta kayıtlarından ilgili bilgilerin çıkarılmasına dayanan yöntemleri kullanan çalışmalar da bulunmaktadır. Tantimongcolwat vd. (2008) çalışmalarında makine öğrenmesi yaklaşımlarını kullanarak manyetokardiyografi (MKG) kayıtlarından iskemik kalp hastalıkları örüntüsünün otomatik olarak yorumlanması için bir model önermişlerdir. Bu amaçla Geriye Yayılım Sinir Ağı ve Öz Düzenlemeli Harita (Self-Organizing Map-

SOM) algoritması olmak üzere iki tür makine öğrenmesi tekniği kullanmışlardır. Çalışmada, 125 hastadan oluşan veri kümesi, kalp kası tarafından yayılan manyetik alanın ardışık ölçümü ile elde edilmiştir. Veri kümesi 74 eğitim verisi ve 51 test verisi olarak ikiye bölünmüştür. Araştırmada, SOM makine öğrenmesi algoritmasının daha yüksek oranda bir tahmin performansı gösterdiği saptanmıştır. Jonnagaddala vd. (2015) elektronik ortamda bulunan hasta bilgilerinden kural tabanlı Metin Madenciliği yöntemi ile elde ettikleri bilgileri kullanarak, Framingham risk skoruna göre, 10 yıllık koroner arter hastalığı risk değerlendirmesi yapmışlardır. Çalışmada 296 diyabet hastasına ait 1304 sağlık kaydı üzerinde Metin Madenciliği uygulanmıştır. Veri kümesinden yaş, cinsiyet, diyabet hastalığı, sigara içme davranışı, kan basıncı, HDL kolesterol ve total kolesterol gibi risk faktörlerine ait bilgiler Metin Madenciliği ile çıkarılarak risk değerlendirmesi yapılmıştır. Kural tabanlı Metin Madenciliği sonucunda elde edilen sonuçlar, manuel olarak yapılan Framingham risk skoru ile tutarlılık göstermiştir. Veri kümesinde %10 ila %20 arasında değişen risk tahmini yapılmıştır.

Literatürde, makine öğrenmesi algoritmaları ile birlikte bulanık mantık yaklaşımının da kullanıldığı çalışmalar yer almaktadır. Muthukaruppan ve Er (2012) yaptıkları çalışmada, kalp hastalığının tanılanmasında kullanılmak üzere Parçacık Sürü Optimizasyon tabanlı bulanık bir uzman sistem geliştirmişlerdir. Geliştirilen sistem Cleveland ve Macaristan veri kümeleri üzerinde uygulanmıştır. Veri kümeleri birçok değişkenden oluştuğu için, tanıya katkıda bulunan değişkenleri ortaya çıkarmak için Karar Ağacı algoritması kullanılmıştır. Karar Ağacı çıktıları ise bulanık kural tabanlı modele dönüştürülmüştür. Geliştirilen bu modelle %93,27 sınıflama doğruluğu elde edilmiştir. Kim vd. (2015) yaptıkları çalışmada yaş, cinsiyet, total kolesterol, LDL, HDL, sistolik ve diyastolik kan basıncı, sigara kullanımı ve diyabet varlığı değişkenleri ile ilgili bilgilerin yer aldığı 748 hasta kaydı üzerinde Karar Ağacı ve Bulanık Mantık yöntemi kullanarak koroner kalp hastalığı tahmini yapmışlardır. 748 hastanın 525'i eğitim 223'ü test verisi olarak ayrılmıştır. Geliştirilen modelin değerlendirilmesinde doğruluk oranı ve ROC eğrisi (Alıcı işlem karakteristikleri, Receiver Operating Characteristic) analizi kullanılmıştır. Modelin doğruluk oranı %69,51, ROC eğrisi değeri 0,594 olarak saptanmıştır.

Uyar ve İlhan (2017) çalışmalarında, kalp hastalığının tahmin edilmesi amacıyla Genetik Algoritma tabanlı Tekrarlayan Bulanık Sinir Ağları (recurrent fuzzy neural networks -RFNN) ile bir model geliştirmişlerdir. RFNN 13 girdi, 7 gizli nöron ve 1 çıktı nöronu olacak şekilde uygulanmıştır. Ayrıca, ağırlık ve eşik değerleri 64 birim uzunlukta genlerle kodlanmıştır. UCI Cleveland veri kümesinden eksik verilerden oluşan kayıtlar çıkarıldıktan sonra elde edilen 297 hastanın 252'sinin eğitim ve 45'inin test olarak kullanıldığı veri kümesi üzerinde algoritma uygulanarak %97,78 sınıflama doğruluğu elde edilmiştir. Nazari vd. (2018) yaptıkları çalışmada kalp hastalığı varlığının değerlendirilmesi amacıyla bulanık Analitik Hiyerarşi Süreci (AHP) ve Bulanık Çıkarım Sistemi tabanlı klinik karar destek sistemi geliştirmişlerdir. Bulanık AHP yöntemi kalp hastalığı gelişiminde etkili risk faktörlerinin ağırlıklarını hesaplamak amacıyla, Bulanık Çıkarım Sistemi ise hastalarda kalp hastalığı gelişme riskini belirlemek ve değerlendirmek amacıyla kullanılmıştır. Araştırmada literatür inceleme ve uzman görüşü ile risk faktörleri; obezite, sigara içme, stres gibi değiştirilebilir risk faktörleri; artmış LDL ve trigliserid seviyesi, azalan HDL, yüksek kan basıncı, diyabet gibi kontrol edilebilir risk faktörleri; yaş, cinsiyet, genetik faktörler gibi değiştirilemeyen risk faktörleri olarak sınıflandırılmıştır. Geliştirilen klinik karar destek sistemi 100 hasta üzerinde değerlendirilmiştir. Kalp hastalıkları uzmanı, çalışmaya dahil edilen 100 hastadan 81'i için ileri kardiyolojik test önerirken, geliştirilen klinik karar destek sistemi modeli kalp hastalığı tespit edilen bu 20 hastayı da kapsayan 26 hasta için kalp hastalığı olma olasılığını yüksek olarak değerlendirmiştir.

Konu ile ilgili çalışmalar incelendiğinde, bazı çalışmalarda uygulanan modelin kullanıcı arayüzü ile ürün haline dönüştürüldüğü görülmektedir. Chen vd. (2011) yaptıkları çalışmada, koroner kalp hastalıklarının tanı sürecini desteklemek amacıyla, klinik karar destek sistemi geliştirmişlerdir. Karar destek sistemi iki aşamada geliştirilmiştir. İlk aşamada, Cleveland veri kümesi üzerinde Yapay Sinir Ağları algoritması kullanılarak sınıflama yapılmıştır. Sistemin ikinci aşamasında, kullanıcı arayüzü geliştirilmiştir. Çalışmada, kullanıcı arayüzü; hasta bilgileri, ROC eğrisi analizi, sınıflama performansı göstergeleri ve kalp hastalığı tahmin sonucu bölümlerini içerecek şekilde geliştirilmiştir.

Literatürde medikal alanda yapılan diğer çalışmalarda olduğu gibi koroner kalp hastalığı alanında da sağlık çalışanlarına yorumlama imkânı sunması açısından karar ağacı algoritmalarının sıklıkla kullanıldığı görülmektedir. Pandey vd. (2013) yaptıkları çalışmada Karar Ağacı algoritmasına dayalı Kalp Hastalığı Tahmin Sistemi geliştirmişlerdir. Sitemin geliştirilmesinde, farklı budama yaklaşımları ile Karar Ağacı Algoritması uygulanmış ve uygulama sonuçları karşılaştırılmalı olarak verilmiştir. Çalışmada, azaltılmış hata budaması yapılan Karar Ağacı algoritmasının %75,73 doğruluk oranında tahmin performansı gösterdiği belirlenmiştir. Sharan ve Sathees (2016) çalışmalarında, kalp hastalığı veri kümesine, Sınıflandırma Karar Ağacı (simple CART), J48 Karar Ağacı ve Naive Bayes (NB Tree) algoritmalarını uygulayarak sınıflama yapmışlardır. Karar Ağacı algoritmalarının uygulanmasında WEKA programı kullanılmıştır. Çalışmada, algoritmalar sınıflama doğruluğu ve işlem zamanı açısından karşılaştırılmıştır. Sınıflama doğruluğu en yüksek algoritma %92,2 oranı ile Simple CART algoritması, işlem zamanı en kısa olan algoritma ise 0.08 saniye ile J48 Karar Ağacı algoritması olarak belirlenmiştir.

Son yıllarda, araştırmacıların sınıflama performanslarının ve genelleme yapabilme özelliklerinin daha iyi olması nedeniyle Rastgele Orman gibi topluluk öğrenme algoritmaları ile ilgili çalışmalara odaklandığı görülmektedir. Abdullah, ve Rajalaxmi (2012) çalışmalarında UCI Cleveland veri kümesi üzerinde Rastgele Orman algoritması ile koroner kalp hastalığı tahmini için veri madenciliği modeli geliştirmişlerdir. Araştırmada, 10 karar ağacı ile rastgele orman algoritması uygulanmıştır. Rastgele orman algoritması %63,33 sınıflama doğruluğu ile karar ağacından daha iyi bir performans göstermiştir. Patil ve Kinariwala (2017) çalışmalarında, kalp hastalığının otomatik olarak tanımlanması amacıyla makine öğrenmesi algoritmalarından Rastgele Orman yöntemi ile bir karar destek sistemi geliştirmişlerdir. Çalışmada geliştirilen model Cleveland veri kümesi üzerinde test edilmiştir. Araştırmacılar veri kümesi üzerinde üç farklı Rastgele Orman algoritmasını uygulamışlardır. Bunlar; Klasik, Modifiye Edilmiş ve Ağırlıklandırılmış Rastgele Orman algoritmalarıdır. Araştırmada 14 değişkenli veri kümesi üzerinde bu üç Rastgele Orman algoritması uygulanmış ve sırasıyla % 74,19, %79,42 ve % 83,6 sınıflama doğruluğu elde edilmiştir. Bununla birlikte araştırmada, Modifiye Edilmiş ve Ağırlıklandırılmış Rastgele Orman algoritmalarının Klasik Rastgele Orman

algoritmasına göre daha kolay yorumlanabildiği ve Ağırlıklandırılmış Rastgele Orman Algoritmasının tüm anlamlı medikal değişkenleri tanımlayabilmesi açısından Modifiye Edilmiş Rastgele Orman algoritmasından daha iyi performans gösterdiği de saptanmıştır.

Liu vd. (2017) çalışmalarında, kalp hastalıklarının tanımlanması amacıyla reliefF ve Rough Set yöntemine dayanan Öznitelik Seçme ve Topluluk Öğrenme algoritmalarından olan C4.5 ile sınıflandırmaya dayanan karma bir sınıflama sistemi geliştirmişlerdir. Çalışmada UCI Statlog veri kümesi kullanılmıştır. Araştırmada, yaş, göğüs ağrısı tipi, istirahat EKG bulguları, ulaşılan maksimum kalp hızı, ST segment eğimi, floroskopide boyanan damar sayısı ve talyum sintigrafi sonucunun en yüksek sınıflama doğruluğu veren değişken kümesi olduğu saptanmıştır. Ayrıca, C4.5 sınıflama algoritması ile %92,59 sınıflama doğruluğuna ulaşılmıştır. Kinge ve Gaikwad (2018) yaptıkları çalışmada UCI Cleveland veri kümesi üzerinde J48 Karar Ağacı, Naive Bayes, Rastgele Orman, Adaboost, Torbalama, Çok Katmanlı Yapay Sinir Ağı ve Basit Lojistik Regresyon algoritmalarını uygulayarak koroner kalp hastalığı tahmini yapmışlardır. Araştırmada Rastgele Orman ve Basit Lojistik Regresyon algoritmaları % 83,15 doğruluk oranı ile en iyi performansı gösterirken, J48 Karar Ağacından % 78,15 sınıflama doğruluğu ile en düşük performans elde edilmiştir.

Bu çalışmanın amacı, makine öğrenmesi algoritmaları ile koroner arter hastalığı riskinin analiz edilmesidir. Bu amaç doğrultusunda, UCI veri kümesi koleksiyonundan alınan kalp hastalıkları veri kümeleri grafiksel ve istatistiksel yöntemlerle detaylı bir biçimde analiz edilmiştir. Veri analizi öncesinde, veri kalitesinin artırılması amacıyla veri kümeleri üzerinde gerekli ön işlemler yapılmıştır. Veri analizleri kalp hastalıkları uzmanının görüşleri doğrultusunda yapılmıştır. Analizler sonucunda, sınıflama modeli kurulmuştur. Makine öğrenmesi yaklaşımı kullanılarak geliştirilen modelin sonuçları, kalp hastalıklarının tanı ve tedavi sürecine yapacağı katkılar ve klinik etkileri açısından tartışılmıştır.

2. MATERYAL VE YÖNTEM

2.1. Kalp Hastalığı Veri Kümesi

Bu çalışmada, Kaliforniya Üniversitesi, Irvine (University of California, Irvine C.A-UCI) veri kümesi koleksiyonundan alınan ve 4 ayrı veritabanından oluşan kalp hastalığı veri kümesi kullanılmıştır. Bunlar; Cleveland, Macaristan, İsviçre ve VA Long Beach veri kümeleridir. Veri kümelerinin hepsi aynı formatta oluşturulmuştur ve aynı değişkenleri içermektedir. Veri kümeleri ve örnek sayıları Çizelge 2.1.'de verilmiştir. Kalp hastalıkları orijinal veri kümesi 76 ham değişkenden oluşmaktadır. Ancak, yayınlanan çalışmalarda, 14 değişkenden oluşan alt küme kullanılmıştır. Bu değişkenler ve özellikleri Çizelge 2.2.'de, veri kümelerinde koroner kalp hastalığına sahip bireylerin oranı Çizelge 2.3.'de verilmiştir.

Çizelge 2.1. Veri kümeleri ve örnek sayıları

Veri Kümesi	**Örnek Sayısı**
Cleveland	303
Macaristan	293
İsviçre	122
VA Long Beach	199
Toplam	**917**

Çizelge 2.2. Kalp hastalıkları veri kümesi değişkenleri

Değişken #	**Değişkenler**	**Değişken Tipi**	**Kısaltmalar**
1	Yaş	Sayısal	age
2	Cinsiyet	Kategorik	sex
3	Göğüs Ağrısı	Kategorik	cp
4	İstirahat Kan Basıncı (mm Hg)	Sayısal	trestbps
5	Serum Kolesterol Düzeyi (mg/dl)	Sayısal	chol
6	Açlık Kan Şekeri (> 120 mg/dl)	Kategorik	fbs
7	İstirahat EKG Sonuçları	Kategorik	restecg
8	Ulaşılan Maksimum Kalp Hızı	Sayısal	thalach
9	Egzersizle Tetiklenen Anjina	Kategorik	exang
10	Egzersizle Tetiklenen ST Depresyonu	Sayısal	oldpeak
11	Pik Egzersiz ST Segment Eğimi	Kategorik	slope
12	Floroskopide Boyanan Büyük Damar Sayısı	Kategorik	ca
13	Talyum Testi	Kategorik	thal
14	Koroner Arter Hastalığı Durumu	Kategorik (Hedef Değişken)	Num

Çizelge 2.3. Veri kümelerinde koroner kalp hastalığı olan hastaların oranı

Veri Kümesi	Kalp Hastalığı Olan Hastaların Sayısı	Sağlıklı Bireylerin Sayısı	Kalp Hastalığı Oranı
Cleveland	139	164	%45
Macaristan	106	187	%36
İsviçre	114	8	%92
VA Long Beach	148	51	%74

2.2. Verilerin Hazırlanması

Kalp hastalıkları veri kümesinin makine öğrenmesi yaklaşımı ile analiz edilmesinde CRISP-DM süreç modelinin adımları izlenmiştir. Makine öğrenmesi süreci açısından standart olarak kabul edilen CRISP_DM süreç modelinin üçüncü aşaması olan **verilerin hazırlanması** aşamasında veri analizi ve modelin oluşturulmasından önce veri kalitesi ve model performansının artırılması amacıyla veriler üzerinde bazı ön hazırlık işlemleri yapılmıştır. Öncelikle veri kümelerinde sayısal değerlerle ifade edilen kategorik değişkenler, verilerin incelenmesini ve anlaşılmasını kolaylaştırmak amacıyla etiketlenmiştir. Çizelge 2.4.'de veri kümesinin kategorik değişkenleri için yapılan dönüşümler verilmiştir.

Çizelge 2.4. Kategorik değişkenlerin dönüşüm değerleri

Değişken #	Değişkenler	Değerleri
1	Yaş	
2	Cinsiyet	1: Erkek 0: Kadın
3	Göğüs Ağrısı	1: Tipik Anjina 2: Atipik Anjina 3: Anjinal Olmayan Ağrı 4: Asemptomatik
4	İstirahat Kan Basıncı (mm Hg)	
5	Serum Kolesterol Düzeyi (mg/dl)	
6	Açlık Kan Şekeri (> 120 mg/dl)	1: Doğru 0: Yanlış
7	İstirahat EKG Sonuçları	0: Normal 1: ST-T dalga anormallikleri 2: Sol ventrikül hipertrofisi
8	Ulaşılan Maksimum Kalp Hızı	
9	Egzersizle Tetiklenen Anjina	1: Evet 0: Hayır
10	Egzersizle Tetiklenen ST Depresyonu	
11	Pik Egzersiz ST Segment Eğimi	1: Yukarı Eğimli 2: Düz 3: Aşağı Eğimli
12	Floroskopide Boyanan Büyük Damar Sayısı	
13	Talyum Testi	3: Normal 6: Fix defekt 7: Reversible defekt
14	Koroner Arter Hastalığı: Koroner Damar Çapında Daralma Durumu	0: < %50 Daralma 1: > %50 Daralma 2: > %50 Daralma 3: > %50 Daralma 4: > %50 Daralma

2.2.1. Kayıp Verilerin Yönetimi

Veri kümeleri üzerinde veri kalitesinin yükseltilmesi amacıyla yapılan işlemlerden biri de kayıp verilerin belirlenmesi ve yönetimidir. Veri kümelerinin eksik veri oranları R programlama aracılığıyla belirlenmiştir. Daha sağlıklı analiz yapabilme ve model oluşturma açısından veri kümelerinden %60 veya daha fazla oranda eksik veri içeren değişkenler çıkarılmıştır. %60'ın altında kayıp veri içeren değişkenler için bazı eksik veri tamamlama yöntemleri uygulanmıştır. Bunlar; literatürde klasik yöntem olarak kabul edilen sayısal değişkenler için ortanca ya da ortalama, kategorik değişkenler için mod işlemi yani sonucu en çok tekrar eden değerin kayıp verilerin yerine konması, Rastgele Orman ve k-En Yakın Komşuluk yöntemleridir. Çalışmada, kayıp verileri tamamlama yöntemlerinin etkinliğini değerlendirmek amacıyla Cleveland veri kümesi üzerinde yapay ve rastsal bir biçimde %10, %20 ve %40 oranlarında kayıp veri oluşturulmuştur. Oluşturulan bu yeni veri kümeleri üç yöntem ile tamamlanmıştır. Klasik yöntemde, değişkenler normal dağılıma sahip olmadığı için, veri kümesindeki kayıp sayısal veriler, ilgili değişkenlerin ortanca değerleri ile kayıp kategorik veriler ise ilgili değişkenlere mod işlemi ile uygulanarak tamamlanmıştır. Rastgele Orman yöntemi R programının "missForest" paketi, k-En Yakın Komşuluk ise R programının VIM (Visualization and Imputation of Missing Values) paketi kullanılarak uygulanmıştır.

2.2.2. Hata Parametreleri

Eksik verilerin tamamlanmasında kullanılan yöntemlerin performansları, literatürde en sık kullanılan hata parametreleri ile karşılaştırılmıştır. Bunlar:

Hata Karelerinin Ortalamasının Karekökü (Root Mean Square Error-RMSE): RMSE, veri kümelerindeki gözlenen değerler ve eksik veri tamamlama yönteminin tahminleri arasındaki hata miktarını belirlemekte en sık kullanılan yöntemlerden birisidir (Schmitt vd., 2015). RMSE sonucu sıfıra ne kadar yakın ise modelin tahmin yeteneği o kadar yüksektir şeklinde değerlendirilir. Eşitlik 2.1'de; $\boldsymbol{X_G}$ veri kümesinin

gerçek değerlerini, $\boldsymbol{X_T}$ veri kümesinin tahmin edilen değerlerini ve $\boldsymbol{N}$ veri kümesinin toplam kayıt sayısını ifade etmektedir.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(X_G - X_T\right)^2}{N}} \tag{2.1}$$

Normalize Hata Karelerinin Ortalamasının Karekökü (Normalized Root Mean Square Error-NRMSE): NRMSE**,** RMSE değerinin gerçek veri kümesinde yer alan değerlerin değişimi temel alınarak normalize edilmesi ile elde edilmektedir (Oba vd., 2003). RMSE değerinin normalize edilmesinde gerçek veri kümesinin standart sapması, en büyük değer ve en küçük değer arasındaki fark, ortalaması gibi değerler kullanılmaktadır. Kalp hastalıkları veri kümesinin değişkenleri normal dağılıma sahip olmadığı için NRMSE değerinin hesaplanmasında, ilgili değişken için veri kümesinin en büyük ve en küçük değeri arasındaki fark temel alınmıştır. Eşitlik 2.2'de; $\boldsymbol{X_G}$ veri kümesinin gerçek değerlerini, ***max*** en büyük değeri, ***min*** en küçük değeri ifade etmektedir.

$$NRMSE = \frac{RMSE}{X_{G\max} - X_{G\min}} \tag{2.2}$$

Ortalama Mutlak Hata (Mean Absolute Error- MAE): MAE, veri kümelerindeki gerçek değerler ile eksik veri tamamlama yönteminin tahminleri arasındaki ortalama mutlak hatayı ifade etmektedir (Zauniri vd., 2015). RMSE'ye benzer bir biçimde sonuç sıfıra yaklaştıkça modelin tahmin performansı da artar. Eşitlik 2.3'de; $\boldsymbol{X_G}$ veri kümesinin gerçek değerlerini, $\boldsymbol{X_T}$ veri kümesinin tahmin edilen değerlerini ve $\boldsymbol{N}$ veri kümesinin toplam gözlem sayısını ifade etmektedir.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|X_G - X_T\right| \tag{2.3}$$

Yanlış Sınıflandırılmış Verilerin Yüzdesi (Percent of False Classified Data-PFC): Veri kümesinde yer alan kategorik değişkenler için yanlış olarak sınıflandırılan verilerin yüzdesi hesaplanarak kullanılan yöntemin performansı değerlendirilmiştir (Malinowski vd., 2015).

2.3. Veri Analizi ve Modelinin Kurulması

Kategorik veri dönüşümünün yapılması ve kayıp verilerin tamamlanmasından sonra veri kümeleri analiz ve modelleme için uygun hale getirilmiştir. Bu tez çalışması kapsamında yapılan tüm veri analizleri R programlama dili ile gerçekleştirilmiştir. R kodlarını geliştirme aracı olarak entegre bir geliştirme ortamı olan RStudio'nun açık kaynak kod ürünü kullanılmıştır. Veri analizinde grafiksel yöntemlerin yanı sıra istatistiksel analizler de kullanılmıştır. Veri analizi medikal literatür incelemesi ve kardiyoloji uzmanının görüşleri doğrultusunda yapılmıştır.

Verilerin hazırlanması ve uzman görüşü doğrultusunda yapılan grafiksel ve istatistiksel analiz aşamalarından sonra kalp hastalıkları veri kümesi üzerinde sınıflama modeli kurulmuştur. Veri kümeleri analizi sonucunda kayıp veri oranının en az olduğu ve dengeli bir dağılıma sahip olduğu belirlenen Cleveland veri kümesi üzerinde sınıflama modeli kurulmuştur. Ayrıca, diğer veri kümeleri ile karşılaştırıldığında daha az kayıp veri içeren ve dengeli bir dağılıma sahip olan Macaristan veri kümesi, Cleveland veri kümesi ile birleştirilerek yeni bir veri kümesi oluşturulmuştur. 596 hasta kaydı ve %60'ın üzerinde eksik veri içeren 3 değişken çıkarıldıktan sonra kalan 11 değişkenden oluşan veri kümesi üzerinde de ayrıca bir sınıflama modeli oluşturularak sonuçlar karşılaştırılmıştır.

Cleveland veri kümesine iki ayrı sınıflama modeli uygulanmıştır. Bu sınıflama modellerinden birincisine Cihan vd. (2018) yaptıkları çalışmada budanmış J48 karar ağacından elde edilen *ca, exang, cp, thal, oldpeak ve age* değişkenleri temel alınarak algoritma uygulanmıştır. Diğer sınıflama modeli ise uzman görüşü rehberliğinde yapılan istatistiksel ve grafiksel analizler sonucunda belirlenen değişkenlerden oluşan veri kümesi üzerinde uygulanmıştır. Sınıflama modelinin kurulmasında bir topluluk

öğrenme algoritması olan Rastgele Orman algoritması kullanılmıştır. Rastgele Orman algoritmasının uygulanmasında R programının "randomForest" paketi kullanılmıştır. Rastgele Orman algoritmasının uygulanmasında, karar ağaçlarındaki her bir düğüm için girdi değişkenleri içerisinden rastgele seçilecek ***m*** adet değişken ve geliştirilecek ağaç sayısı olan ***N*** parametrelerinin seçimi sınıflama performansı üzerinde önemli etkilere sahiptir. Değişken sayısı olan ***m*** için başlangıçta toplam değişken sayısı ***M***' in karekökü değeri seçilmiştir (Breiman ve Cutler, 2004). Sonrasında ise OOB hata oranına göre optimum ***m*** değeri seçilerek model oluşturulmuştur. En uygun ağaç sayısı için algoritmanın doğru sınıflama oranı dikkate alınarak seçim yapılmıştır. Buna göre, sınıflama modelleri 500 adet karar ağacı ile oluşturulmuştur.

2.3.1. Rastgele Orman

Rastgele Orman, Leo Breiman ve Adele Cutler tarafından geliştirilmiş, çok sayıda karar ağacından oluşan bir topluluk öğrenme algoritmasıdır. Rastgele ormanlar, çok yönlülüğü ve gücü tek bir makine öğrenme yaklaşımıyla birleştirmesi ve kullanım kolaylığı sağlaması nedeniyle hızla en popüler makine öğrenme yöntemlerinden biri haline gelmiştir (Lantz, 2013).

Rastgele orman yönteminin güçlü ve zayıf yönleri aşağıda sıralanmıştır.

Rastgele Orman Algoritmasının Güçlü Yönleri:

1. Çoğu problemde iyi performans gösteren çok amaçlı bir modeldir.
2. Gürültülü veya eksik verilerin yanı sıra hem sayısal hem de kategorik veriler üzerinde iyi sonuçlar verebilir. Kayıp veri sayısının arttığı durumlarda da sınıflama başarısı yüksektir.
3. Yalnızca en önemli özellikleri seçer.
4. Çok sayıda özellik içeren veriler üzerinde uygulanabilir.
5. Ezberlemeye karşı güçlüdür.
6. Model işlemi sonrasında ortaya çıkan ağaçta budama işlemi yapmaya gerek duyulmamaktadır.
7. Büyük veya küçük boyutlardaki veri kümeleri üzerinde doğru sonuçlar elde edilebilmektedir.
8. Dengesiz dağılım gösteren veri kümeleri üzerinde de kullanılabilmektedir.

Rastgele Orman Algoritmasının Zayıf Yönleri:

1. Karar ağaçlarından farklı olarak, model kolay bir biçimde yorumlanamaz.
2. Modeli veriye göre ayarlamak için ön işlem gerektirmektedir.
3. Model sonucunda üretilen sonuç için güven aralığı verilememektedir.
4. Modelde karar ağaçları ile ilgili bilgiler tutulduğu için bellek gereksinimi artmaktadır (Akman vd., 2011; Lantz, 2013).

Rastgele Orman Algoritması

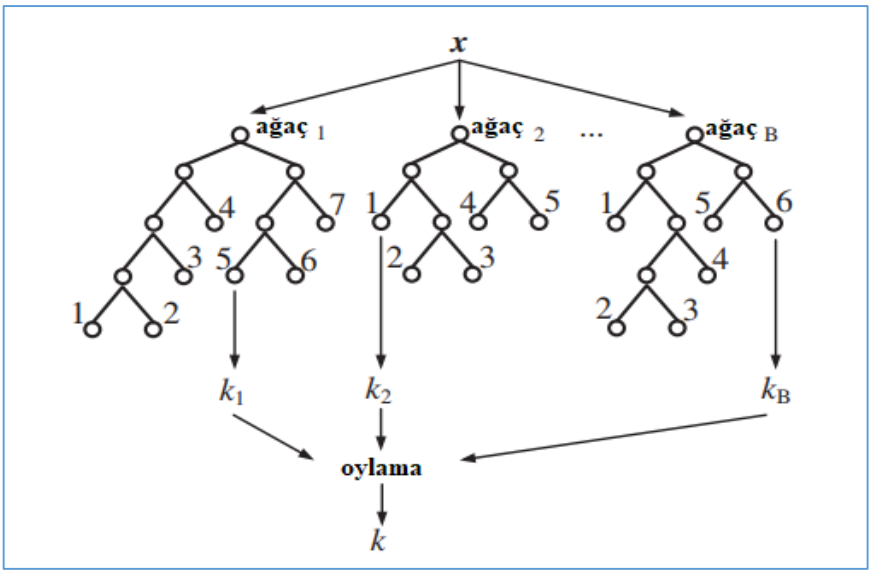
Breiman (2001), torbalama (bagging) algoritmasına ek bir rastlantısallık katmanı ekleyerek rastgele orman algoritmasını önermiştir. Rastgele Orman yöntemi verilerden, farklı bir bootstrap örneği kullanarak her bir ağacın oluşturulmasına ek olarak, sınıflandırma veya regresyon ağaçlarının oluşturulmasında da farklılıklar sağlamaktadır. Standart karar ağaçlarında, her düğüm, tüm değişkenler arasında en iyi bölünme kullanılarak bölünür. Rastgele orman yönteminde ise, her düğüm, o düğümde rastgele seçilen bir tahminci değişken (predictors) alt kümesi arasından en iyisi kullanılarak bölünür. Bu stratejinin uygulanması ile rastgele orman yöntemi, diskriminant analizi, destek vektörü makineleri ve sinir ağları dahil olmak üzere diğer pek çok sınıflayıcı ile kıyaslandığında çok iyi bir performans göstermektedir ve ezberlemeye (overfitting) karşı da oldukça güçlüdür. Buna ek olarak, rastgele orman algoritmasında kullanıcıdan yalnızca iki parametre istenmektedir. Bunlar; her düğümde kullanılan rasgele alt kümedeki değişken sayısı ve ormandaki ağaç sayısıdır. Ayrıca, rastgele orman yöntemi genellikle kullanıcıdan alınan bu değerlere çok duyarlı değildir (Liaw, 2002).

Birden fazla ağacın kullanılmasıyla sınıflamadaki hata oranının önümle ölçüde azaltılabileceği yapılan bazı çalışmalarda görülebilmektedir. Bu tür bir yaklaşımın ilk örneklerinden birisi "bagging" (Breiman, 1996) olup daha farklı yöntemlerle de benzer şekilde iyileştirmeler sağlanabilmektedir. Bu yaklaşımların ortak bileşeni ise *k*. ağaç için, önceki $\Theta_1, ..., \Theta_{k-1}$ rastgele vektörden bağımsız ancak aynı ortak dağılıma sahip bir rasgele Θ_k vektörünün üretilmesidir. Böylece eğitim seti ve Θ_k kullanılarak bir ağaç geliştirilerek bir $h(\mathbf{x}, \Theta_k)$ sınıflandırıcısı elde edilir. Burada $\boldsymbol{x}$ bir giriş vektörünü göstermektedir. Bu şekilde çok sayıda ağaç oluşturulduktan sonra en çok oyu alacak

sınıfı belirlemek için oylama işlemi yapılır. Sonuç olarak, Rastgele Orman, $\{h(\mathbf{x}, \Theta_k), k=1,...\}$ şeklinde ifade edilen ve ağaç yapısında sınıflandırıcıların bir araya gelmesi ile oluşturulan bir sınıflandırıcı olarak tanımlanmaktadır. Burada $\{\Theta_k\}$, bağımsız ve özdeş dağılıma sahip rastgele vektörlerdir. Rastgele orman algoritmasında her ağaç, ***x*** girdisinin en popüler sınıfı için birim oy kullanmaktadır (Breiman, 2001).

Rastgele orman algoritması (hem sınıflandırma hem de regresyon için) adımları şunlardır (Liaw, 2002; Akman vd., 2011):

1. Orijinal veri kümesinden n tane bootstrap (iadeli yeniden örnekleme) örneğin alınması. Rastgele orman algoritmasında, bootstrap yöntemi ile örnekleme yapılmaktadır. Bootstrap yöntemi non-parametrik koşullar söz konusu olduğunda kullanılan tekrarlı örnekleme yöntemlerinden biridir. Veri kümesi üzerinde rastgele seçilen örneklerin özellikleri kullanılarak büyük veri kümesi hakkında tahminlerde bulunulmaktadır. Çapraz doğrulama yönteminde, veriler her bir örnekte yalnızca bir kez görülebileceği ayrı bölümlere ayrılırken, bootstrap yönteminde yerine koyarak örnekleme yapıldığı için örneklerin birden çok kez seçilmesine izin verilmektedir (Lantz, 2013).
2. Bootstrap yöntemiyle oluşturulan her bir örneklemin 2/3'ünün eğitim (inBag), 1/3'ünün test (Out Of Bag-OOB) verisi olarak ayrılması.
3. Seçilen eğitim veri kümesi üzerinde aşağıdaki kurallara dayanarak budanmamış (unpruned) sınıflama veya regresyon ağacı oluşturulur. Bu adımlar;
   a. Her bir düğüm için, en iyi dallara ayıracak tahmin edici değişkenin, tüm tahmin edici değişkenler arasından değil, rastgele seçilecek olan ***m*** sayıda tahmin edici değişkenin arasından seçilmesi
   b. En iyi dallanma kriteri, seçilen tahmin değişkeni için hesaplanır. Bu yöntemle bulunan değere göre veri kümesi her bir düğümde iki alt dala ayrılır.
   c. Bu işlemler yaprak düğüm elde edilene kadar her düğüm için tekrar edilerek yapılır. Şekil 1.6.'da rastgele orman algoritmasındaki ağaç yapısı gösterilmektedir. B rastgele ormandaki ağaç sayısını, k_1, k_2, k_3 ve k sınıf etiketlerini göstermektedir.



Şekil 2.1. Rastgele orman algoritması ağaç yapısı (Englund ve Verikas, 2012).

4. Oluşturulan ağaçların tahminlerinin birleştirilerek yeni tahmininin oluşturulması. Yeni tahmin oluşturulurken, sınıflama ağaçları için en çok oyu alan sınıf seçilerek, regresyon ağaçları için ise oyların ortalaması alınarak yeni tahmin oluşturulur.

Eğitim verilerine dayalı hata oranı tahmini aşağıdaki gibi yapılır:

a. Her bir bootstrap döngüsünde, bootstrap örneklemede kullanılmayan verileri (Breiman'ın "çanta dışı" veya OOB olarak adlandırdığı) kullanarak oluşturulan ağaç test edilir.

b. Her bir ağaç için yapılan OOB tahmini toplanarak hata oranı tahmini yapılır.

Rastgele orman yönteminde ek olarak iki farklı bilgi daha üretilmektedir. Bunlar; tahmin değişkeninin önem derecesi ve verinin iç yapısının bir ölçüsü olarak farklı veri noktalarının birbirine yakınlığıdır.

Değişkenin Önemi:

Genel olarak, bir değişkenin önemi, diğer değişkenlerle olan (muhtemelen karmaşık) etkileşimler sonucu oluştuğu için, bu önemin hesaplanabilmesi de oldukça zordur. Rastgele orman algoritmasında, bir değişkenin önemi, incelenmek istenen değişken dışındaki tüm değişkenler aynı bırakılırken, test verisinde incelenmek istenen değişkenin değerleri kendi içerisinde değiştirilerek hesaplanır. Bu değişim ormandaki tüm ağaçlar üzerinde uygulanarak hata tahmininde oluşan değişim değerlendirilir. İncelenen değişken için hata tahmininde oluşan farkların ortalaması alınarak önem derecesi hesaplanmış olur (Liaw, 2002). Bu yöntem standart yöntem olarak da isimlendirilmektedir.

Değişkenin önem derecesinin hesaplanmasında kullanılan diğer yöntem ise Gini yöntemidir. Rastgele orman algoritmasında, belirli bir ***m*** değişkeninden dallara bölünme olmadan önce ve sonra veriler için Gini değerleri hesaplanır. Hesaplanan bu değerler arasındaki fark ormanda yer alan her bir ağaç için bulunarak toplanır. Elde edilen bu değer ***m*** değişkeni için Gini önem derecesini vermektedir (Akman vd., 2011). Eşitlik 2.4'de $GI(t)$ Gini indeksini, $p(k|t)$, k sınıfının t düğümünde doğru bir biçimde ayrılabilme oranını göstermektedir. Eşitlik 2.5'te ise, $\Delta GI(t)$ Gini farkını, $P_L GI(t_L)$ nodun sol tarafındaki Gini indeksini, $P_R\, GI(t_R)$ nodun sağ tarafındaki Gini indeksini, P_t bölünmeden önceki örnek sayısını, P_L bölünmeden sonraki soldaki örnek sayısını, R_R bölünmeden sonraki sağdaki örnek sayısını göstermektedir (Kawakubo ve Yoshida, 2012).

$$GI(t) = 1 - \sum_k p(k|t)2 \tag{2.4}$$

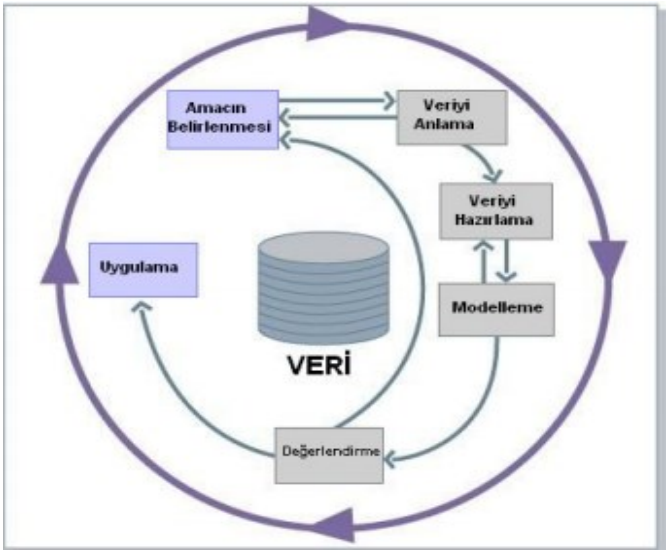
$$\Delta GI(t) = P_t\, GI(t) - P_L GI(t_L) - P_R\, GI(t_R) \tag{2.5}$$

Veri Yakınlık (Proximity) Matrisi:

Rastgele orman algoritması sınıflama ve regresyon modeline ek olarak veri yakınlık matrisi de sağlamaktadır. Veri yakınlık matrisi oldukça önemli bir bilgi kaynağı oluşturarak veri kümeleme, çok boyutlu verilerin görselleştirilmesi, aykırı (outlier) değerlerin tespit edilmesi, eksik değerlerin yerine konması, yanlış etiketlenen verilerin bulunması, destek vektör makinelerinde kernel matrisinin oluşturulması gibi veri madenciliği görevlerinde etkin bir biçimde kullanılmaktadır (Englund ve Verikas, 2012). Yakınlık matrisini elde etmek için, oluşturulan ağaçta veriler yukarıdan aşağı doğru yerleştirilirler. Bu yerleştirme işleminden sonra, x_i ve x_j gözlemleri ağacın aynı terminal düğümünü işgal ediyorsa prox (i,j) değeri bir artırılır. Rastgele orman oluşturulduğunda, yakınlık matsisi değerleri ormandaki ağaç sayısına bölünerek matrisin son hali elde edilir (Breiman ve Cutler, 2004).

2.4. Makine Öğrenmesi Süreci

Makine öğrenmesi yaklaşımının uygulanmasında sistematik bir yaklaşıma gereksinim duyulmaktadır. Literatürde, bilgi keşfi sürecine paralelel olarak bu süreçte geliştirilen algoritmaların uygulanmasını kolaylaştırmak amacıyla geliştirilen araçlar ve modeller bulunmaktadır. Bu modeller içerisinde veri madenciliği sürecinde en çok kullanılan model Veri Madenciliği için Çarpraz Endüstri Standard Süreç Modelidir (CRoss-Industry Standard Process for Data Mining - CRISP). Bu model DaimlerChrysler AG, SPSS, NCR ve OHRA gibi önde gelen veri madenciliği kullanıcıları ve tedarikçileri konsorsiyumu tarafından geliştirilmiştir (Wirth ve Hipp, 2000; Marban vd., 2009). CRISP-DM modeli, veri madenciliği ve bilgi keşfi çalışmaları için hem endüstriden hem de kullanılan teknolojiden bağımsız bir süreç modeli tanımlar. CRISP-DM süreç modeli 6 aşamadan oluşmaktadır. Bu aşamalar Şekil 1.7.'de gösterilmiştir.



Şekil 2.2. CRISP-DM süreç modeli aşamaları (Çınar ve Arslan, 2008).

Aşağıda CRISP-DM süreç modelinin aşamaları kısaca özetlenmiştir:

Amacın Belirlenmesi: Başlangıç aşamasında çalışmanın hedefleri ve gereksinimleri, problem alanı özelinde belirlenir ve bu aşamada edinilen bilgiler ışığında problem tanımı yapılır.

Veriyi Anlama: Bu aşama, veri toplama ile başlayarak, veri kalite problemlerinin tanımlanması, verilerin ilk değerlendirmesi, hipotezlerin oluşturulmasında kullanılmak üzere farklı alt kümelerin tanımlanması gibi veriyi anlama aktiviteleri ile devam eder.

Veriyi Hazırlama: Veri hazırlama aşaması, başlangıç veri kümesinden nihai veri kümesini oluşturmak için gereken tüm etkinlikleri kapsar. Veri hazırlama görevlerini gerçekleştirmek için öncesinde belirlenmiş bir sıra bulunmamaktadır.

Modelleme: Bu aşamada, çeşitli modelleme teknikleri seçilir ve uygulanır. Modelin uygulanması aşamasında optimum sonuç elde etmek için parametreler en uygun değerler olacak şekilde belirlenir. Modelin uygulanması genellikle veri formunda düzenlemeler gerektirdiği için, veri hazırlama aşamasına geri dönmek gerekebilmektedir.

Değerlendirme: Bu aşamada, modeli daha ayrıntılı bir şekilde değerlendirmek ve başlangıçta belirlenen hedefleri doğru bir şekilde gerçekleştirdiğinden emin olmak için atılan adımları gözden geçirmek önemlidir. Bu aşamanın sonunda, sonuçlarının nasıl kullanılacağına dair bir karara varılması gerekmektedir.

Uygulama: Bu aşamada, modelin uygulanması sonucunda ede edilen bilginin problem alanında kullanılabilecek şekilde organize edilmesi ve sunulması yer almaktadır.

2.5. Modelin Değerlendirilmesi

Sınıflama modelinin performansının değerlendirilmesinde Karışıklık Matrisi (Confusion Matrix) ve Alıcı İşlem Karakteristikleri (Receiver Operating Characteristic-ROC) eğrisi kullanılmıştır.
Karışıklık Matrisi, uygulanan sınıflama performansının değerlendirilmesi açısından önemli bir araçtır. Kalp hastalıkları veri kümesi için oluşturulan Karışıklık Matrisi bileşenleri Çizelge 2.5.'de verilmiştir (Sharan ve Sathees, 2016).

Çizelge 2.5. Karışıklık Matrisi Yapısı

	Tahmin Edilen Değerler		
Gerçek Değerler		Hastalık yok	Hastalık var
	Hastalık yok	TN	FP
	Hastalık var	FN	TP

True Positive (TP): Kalp hastalığı olan bireyler, kardiyovasküler hastalığı var şeklinde doğru bir biçimde sınıflandırılmıştır.
False Posivite (FP): Sağlıklı bireyler yanlış bir biçimde kardiyovasküler hastalığı var şeklinde sınıflandırılmıştır.
True Negative (TN): Sağlıklı bireyler doğru bir biçimde sağlıklı şeklinde sınıflandırılmıştır.

False Negative (FN): Kalp hastalığı olan bireyler yanlış bir biçimde sağlıklı şeklinde sınıflandırılmıştır.

ROC eğrisi duyarlılık (sensitivity) ve seçicilik (specifity) arasındaki ilişkinin gösterilmesi için grafiksel bir araç olarak kullanılmaktadır. ROC eğrisinde x ekseninde (1-seçicilik), y ekseninde ise duyarlılık oranları bulunmaktadır. Testin performansını değerlendirmede ROC eğrisi altında kalan alan (area under the curve-auc) önemli bilgi vermektedir. Bu alan 1 değerine yaklaştıkça model performansı da mükemmele yaklaşmış olmaktadır. Karışıklık matrisinden hesaplanabilecek parametreler aşağıdaki eşitliklerde verilmektedir (Kılıç, 2013).

$$\textbf{\textit{Doğruluk}} = \frac{TP+TN}{N} \qquad \textit{N: toplam örnek sayısı} \tag{2.4}$$

$$\textbf{\textit{Duyarlılık}} = \frac{TP}{TP+FN} \quad (\textit{True Positive Rate-Doğru Pozitif Oranı}) \tag{2.5}$$

$$\textbf{\textit{Seçicilik}} = \frac{TN}{FP+TN} \tag{2.6}$$

$$\textbf{\textit{False Positive Rate-Yanlış Pozitif Oranı}} = \textit{1-seçicilik} \tag{2.7}$$

3. BULGULAR ve TARTIŞMA

3.1. Eksik Veriler

Kalp hastalıkları veri kümesi üzerinde sınıflama modeli kurulmadan önce veri kalitesinin artırılması amacıyla veri kümelerindeki eksik verilerin oranları R programı aracılığıyla hesaplanmış ve sonuçlar Çizelge 3.1.'de verilmiştir. Veri kümelerindeki eksik veri oranları incelendiğinde Cleveland veri kümesinin yaklaşık %2 oranında eksik veriye sahip olduğu görülmektedir. Diğer 3 veri kümesinde ise %90'ın üzerinde kayıp veriye sahip değişkenlerin olduğu saptanmıştır. Veri kümeleri üzerinde yapılan analizlerin ve kurulacak modelin daha sağlıklı olabilmesi açısından %60 ve ya daha fazla oranda eksik veri içeren değişkenler veri kümelerinden çıkarılmıştır. Kalan değişkenlere uygulanacak eksik veri tamamlama yöntemine karar verebilmek için literatürde medikal veriler üzerinde yaygın olarak kullanılan klasik yöntem (ortanca-mod), Rastgele Orman ve k-En Yakın Komşuluk algoritmaları uygulanarak hata parametrelerine göre performansları karşılaştırılmıştır.

Çizelge 3.1. Veri kümelerinin eksik veri oranlarının dağılımı

Değişkenler	Cleveland Veri Kümesi	Macaristan Veri Kümesi	İsviçre Veri Kümesi	Va LB Veri Kümesi
Age				
Sex				
Cp				
Trestbps		% 0,3	% 1,6	% 28
Chol		% 7,8	**% 100**	% 28
Fbs		% 2,7	**% 60,9**	% 3,5
Restecg		% 0,3	% 0,8	
Thalach		% 0,3	% 0,8	% 26,5
Exang		% 0,3	% 0,8	% 26,5
Oldpeak			% 4,8	% 28
Slope		**% 64,6**	% 13,8	% 51
Ca	% 1,3	**% 98,9**	**% 95,9**	**% 99**
Thal	% 0,6	**% 90,4**	% 42,2	**% 83**

Cleveland veri kümesinde yer alan sayısal değişkenler için eksik veri tamamlama yöntemlerinin hesaplanan hata parametreleri sonuçları Çizelge 3.2.'de, kategorik değişkenler için sonuçlar ise Çizelge 3.3.'de verilmiştir. Sayısal değişkenler açısından yöntem performanslarına bakıldığında, klasik yöntem olan eksik verilerin sayısal veriler için ortanca ile tamamlanması yaklaşımından elde edilen hata parametrelerinin özellikle %40'ın altında ki kayıp veri oranlarında iyi sonuçlar verdiği görülmektedir. Ayrıca, klasik yöntem %40'ın üzerinde kayıp veriye sahip veri kümeleri üzerinde de Rastgele Orman gibi daha karmaşık modellere yaklaşık sonuçlar vermektedir. Kategorik değişkenler açısından yöntem performansları karşılaştırıldığında ise, mod değeri ile kayıp verileri tamamlama yönteminin, tüm kayıp oranlarında, hata parametreleri açısından en iyi sonuçlara sahip olduğu görülmektedir. Bu nedenle veri kümelerinin eksik verilerini tamamlamak amacıyla karmaşık modeller kullanmak yerine sayısal veriler için ortanca, kategorik veriler için mod değerinin kullanılmasına karar verilmiştir.

Çizelge 3.2. Sayısal değişkenler için yöntem performansları

Eksik Veri Tamamlama Yöntemi	**Sayısal Değişkenler İçin**								
	% 10 Eksik Veri			**%20 Eksik Veri**			**% 40 Eksik Veri**		
	NRMSE	RMSE	MAE	NRMSE	RMSE	MAE	NRMSE	RMSE	MAE
Medyan	3,00 5,30	13,05 6,95	**3,24** 1,79	**4,60** 7,90	**20,35** 10,40	**7,26** 3,76	12,90 14,00	36,40 15,67	**16,76** 7,62
Rastgele Orman	**2,90** 4,20	**12,69** 5,55	3,35 1,36	5,30 **6,00**	23,05 **7,86**	7,89 **2,82**	**12,80** **11,40**	**36,32** **12,78**	17,08 **6,12**
K-En yakın komşuluk	3,00 **4,10**	12,95 **5,41**	3,41 **1,35**	6,60 7,60	28,74 9,98	9,98 3,41	14,70 13,00	41,56 14,60	18,65 7,14

Çizelge 3.3. Kategorik değişkenler için yöntem performansları

Eksik Veri Tamamlama Yöntemi	Kategorik Değişkenler İçin		
	% 10 Eksik Veri	% 20 Eksik Veri	% 40 Eksik Veri
	PFC	PFC	PFC
Mod	**1,65** **5,61**	**2,97** **7,26**	**6,27** **19,47**
Rastgele Orman	**1,65** **5,61**	**2,97** 7,59	6,93 20,13
K-En yakın komşuluk	1,98 6,60	4,62 11,55	7,92 22,77

Kayıp verilerin belirlenmesi ve tamamlanması, değişkenlerin etiketlenmesi ve kategorize edilmesi gibi veri hazırlığı işlemlerinden sonra veri kümelerinin analizleri yapılmıştır. Yapılan ön analizler sonucunda, diğer veri kümeleri ile karşılaştırıldığında az sayıda kayıp veri içermeleri ve hasta ve sağlıklı bireylerin oranı açısından daha dengeli bir dağılım göstermeleri nedeniyle, detaylı analizlerin yapılması ve sonuçların sunulması amacıyla Cleveland ve Macaristan veri kümeleri seçilmiştir.

3.2. Cleveland Veri Kümesinin Analizi

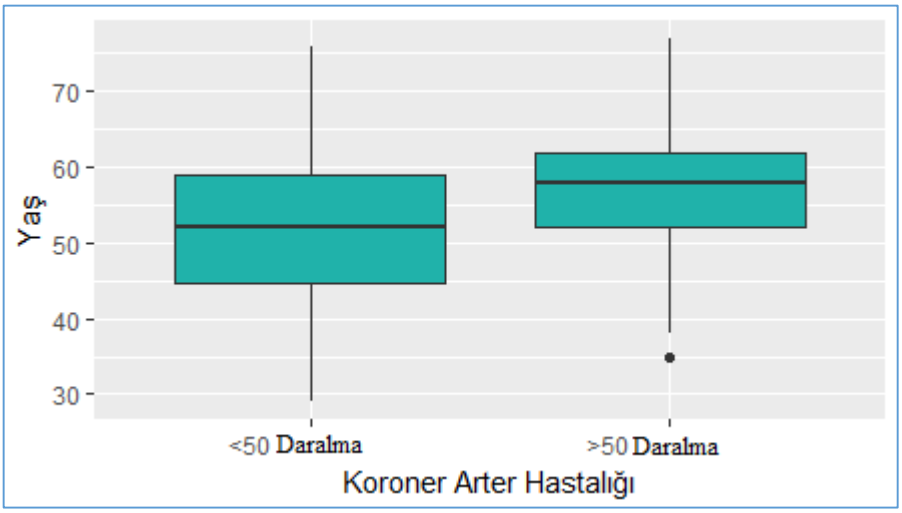
Veri kümelerinin analizine sayısal değişkenler için; tanımlayıcı istatistikler, kutu grafikleri, histogram dağılımı, saçılım garfikleri ve korelasyon analizi, kategorik değişkenler için ise hedef değişkene göre oluşturulmuş çubuk grafikleri kullanılmıştır. Cleveland veri kümesinin analizinin ilk aşamasında veri kümesinin tanımlayıcı istatistikleri hesaplanmıştır. R programı ile elde edilen tanımlayıcı istatistikler Çizelge 3.4.'de verilmiştir.

Çizelge 3.4. Sayısal değişkenler için tanımlayıcı istatistikler

Değişken	Minimum	1.Çeyrek	Ortanca	Ortalama	3.Çeyrek	Maksimum
Yaş	29	48	56	54	61	77
Trestbps	94	120	130	131,7	140,0	200
Chol	126	211	241	246,7	275	564
Thalach	71	133,5	153	149,6	166	202
Oldpeak	0	0	0,8	1,04	1,6	6,2

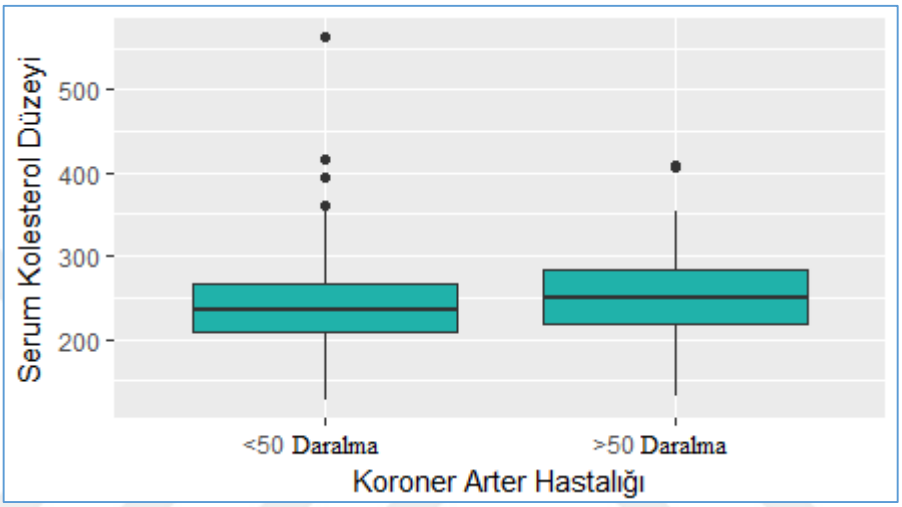
3.2.1. Sayısal Değişkenler için Kutu Grafikleri:

Şekil 3.1.'de verilen yaş değişkeninin kutu grafiği incelendiğinde, önemli koroner arter daralmasının olmadığı grupta ortanca yaş 52-53 yaş civarındayken ciddi koroner arter daralmasının olduğu grupta ortanca değeri 55 yaşın üzerine çıkmaktadır. Ayrıca, bu grupta 55-60 yaş arasında yığılma olduğu görülmektedir. Buna ek olarak, ciddi koroner arter daralması olan grupta 1 uç değer (outlier) bulunmaktadır.



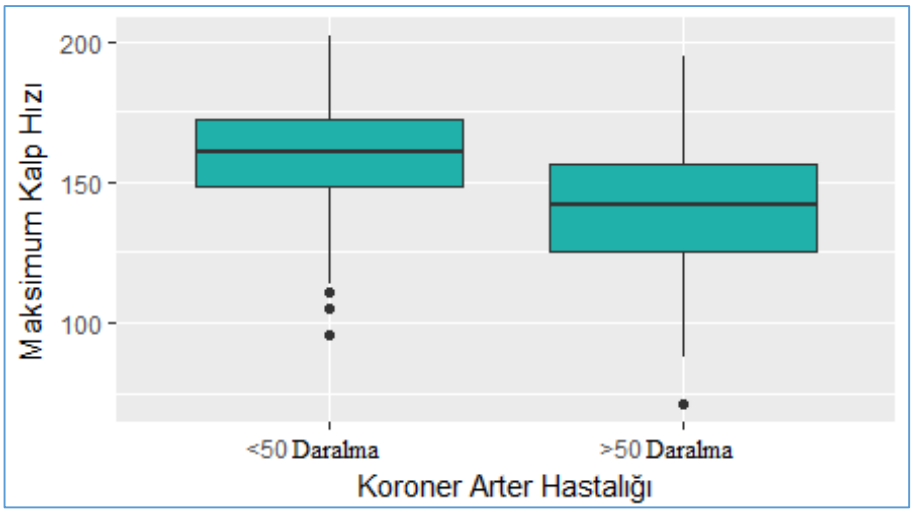
Şekil 3.1. Yaş değişkeni kutu grafiği

Şekil 3.2.'de verilen serum kolesterol düzeyinin kutu grafiği incelendiğinde, koroner daralmanın olduğu her iki grupta da ortanca kolesterol düzeyi ve kutuların kapsadığı bölgeler benzerlik göstermektedir. Bu nedenle, kolesterol değişkeninin sınıflama modelinde belirleyici bir değişken olamayacağı düşünülmektedir. Bununla birlikte, ciddi daralmanın olmadığı grupta 3 uç değer ve 1 aşırı (extreme) değer, diğer grupta ise 1 uç değer saptanmıştır.



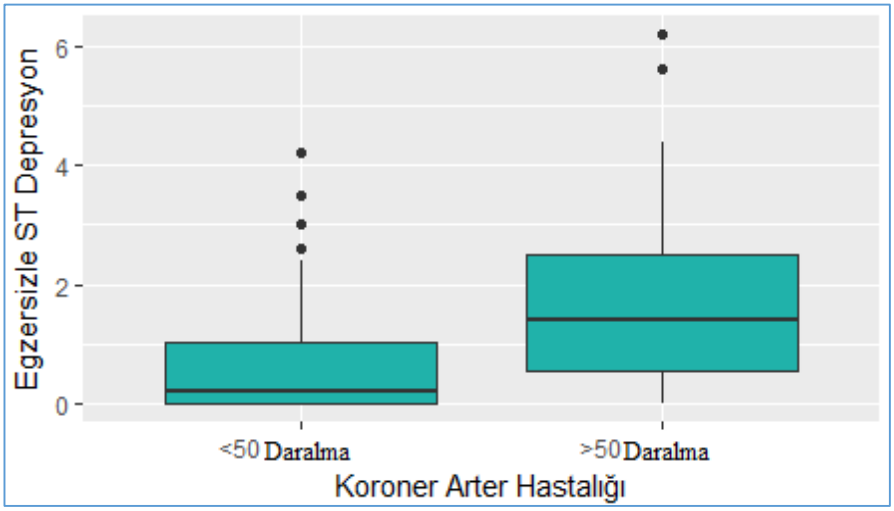
Şekil 3.2. Serum kolesterol düzeyi değişkeni kutu grafiği

Şekil 3.3.'de verilen maksimum kalp hızı değişkeninin kutu grafiği incelendiğinde, ciddi daralmanın olmadığı grubun ortanca kalp hızı değerinin diğer gruba göre yüksek olduğu görülmektedir. Ayrıca, kutuların kapsadığı bölgeler de birbirinden önemli ölçüde farklıdır. Bu nedenle, maksimum kalp hızı değişkeninin sınıflama modelinde önemli bir değişken olması beklenmektedir. Bu grupta ayrıca 3 uç değer bulunmaktadır. Ciddi daralmanın olduğu grupta ise 1 uç değer bulunmaktadır.



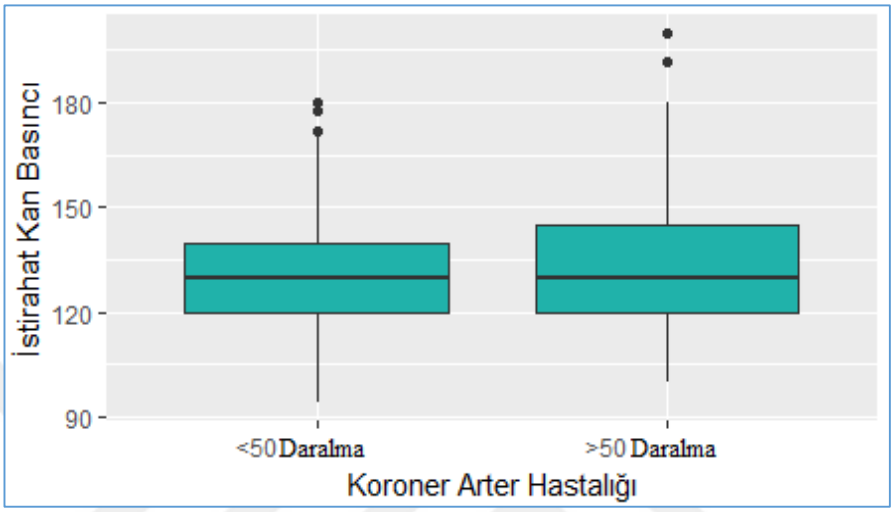
Şekil 3.3. Maksimum hızı kalp değişkeni kutu grafiği

Şekil 3.4.'de verilen egzersizle ST depresyon değişkeninin kutu grafiği incelendiğinde, koroner arter hastalığının olmadığı grupta ST depresyon oranı 0 değeri civarında yığılma göstermektedir. Ayrıca, iki grubun ortalamaları ve kutuların kapsadığı bölgelerin farklılığı ST depresyonu değişkeninin model için önemli bir değişken olduğunu göstermektedir. Ciddi koroner daralmanın olmadığı grupta 4 uç değer, diğer grupta ise 2 uç değer vardır.



Şekil 3.4. Egzersizle ST depresyon değişkeni kutu grafiği

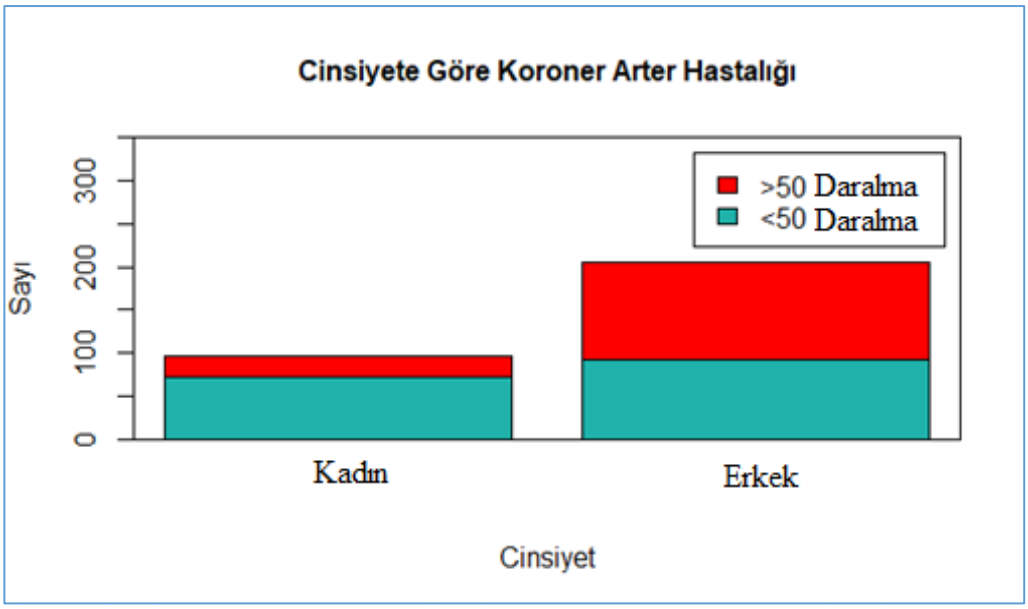
Şekil 3.5.'de verilen istirahat kan basıncı değişkeninin kutu grafiği incelendiğinde, koroner daralmanın olduğu her iki grupta da ortanca değerin ve kutuların kapsadığı bölgelerin birbirine oldukça yakın olduğu görülmektedir. İlk grupta 3 uç değer, ikinci grupta ise 2 uç değer bulunmaktadır.



Şekil 3.5. İstirahat kan basıncı değişkeni kutu grafiği

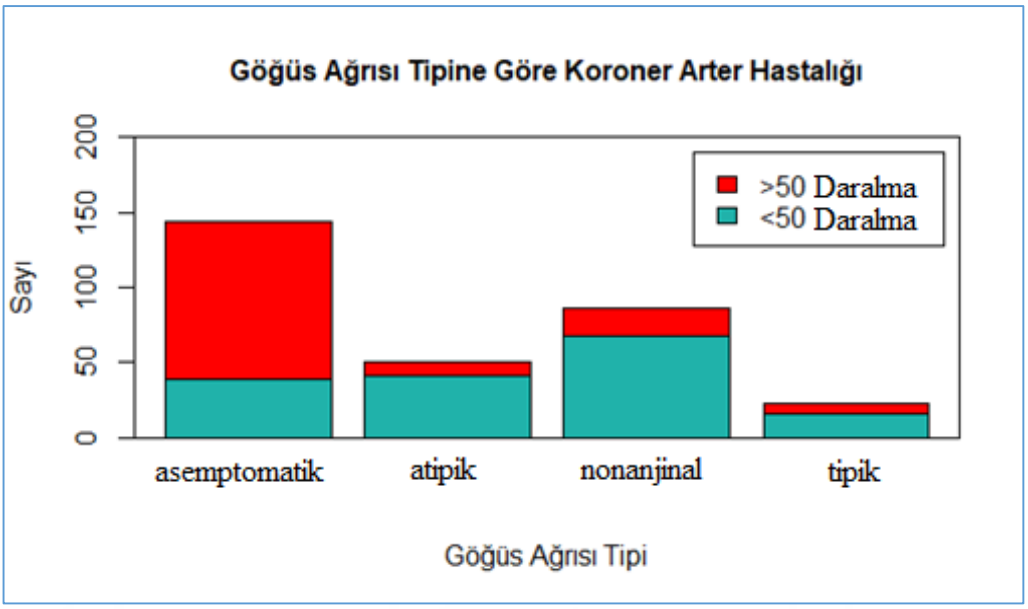
3.2.2. Kategorik Değişkenler için Çubuk Grafikleri:

Şekil 3.6.'da verilen, cinsiyet ile koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde, erkeklerde kadınlara göre koroner arter hastalığının daha fazla olduğu görülmektedir. Kadınlarda menopoz öncesi dönemde koroner kalp hastalığı erkeklere göre daha az görülmektedir. Menopoz öncesinde koroner arter hastalığının daha az görülmesi östrojenin lipid göstergeleri üzerindeki olumlu etkileri ile açıklanabilmektedir. Menopoz sonrası dönemde ise koroner kalp hastalığı prevalansı her iki cinsiyet için eşitlenmektedir.



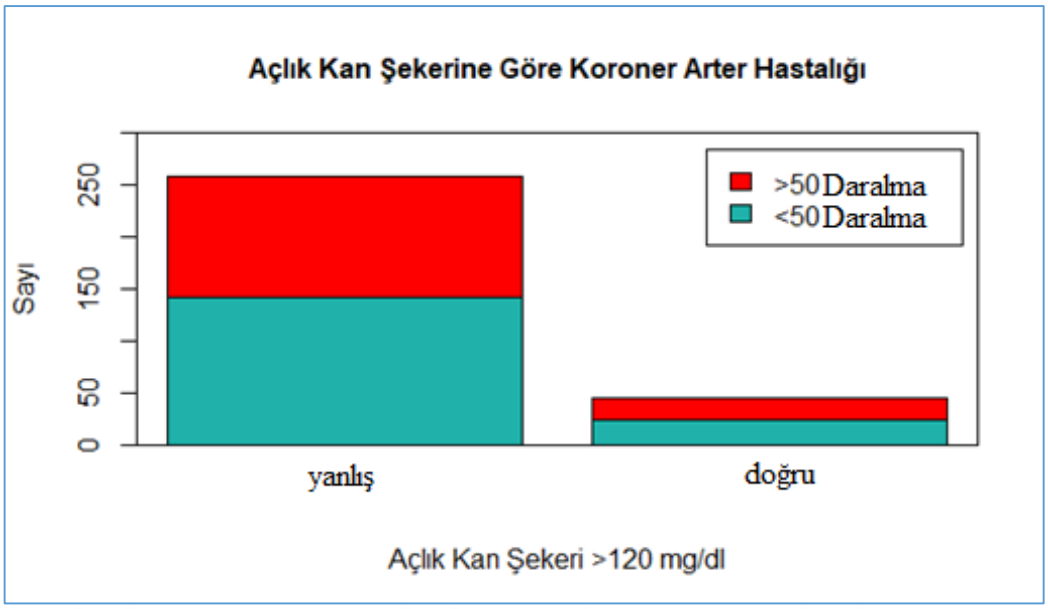
Şekil 3.6 Cinsiyet ve koroner arterlerde daralma

Şekil 3.7.'de verilen göğüs ağrısı tipi ile koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde, göğüs ağrısı olmayan kişilerde koroner arterlerdeki ciddi daralma oranının önemli ölçüde yüksek olduğu görülmektedir. Bu durum, göğüs ağrısı olmayan hastalar için beklen bir sonuç değildir. Bu duruma neden olabilecek klinik hasta özellikleri incelenerek ileri analizler ile bu grupta yer alan hastaların profili çıkarılmalıdır. Göğüs ağrısı olan gruplar incelendiğinde, tipik anjinası olan grupta koroner arterlerdeki ciddi daralma oranının daha yüksek olduğu ve nonanjinal grupta ciddi daralma oranının beklendiği gibi diğer iki gruba göre daha az olduğu saptanmıştır. Göğüs ağrısı yerleşimi substernal (göğüs kemiği altı) bölgede ise, eforla ve duygusal stresle ortaya çıkıyorsa ve 5 ila 20 dakika dinlenince düzeliyorsa tipik anjina olarak adlandırılmaktadır. Bu klinik özelliklerden 2 veya daha azı bulunuyorsa atipik anjina, eğer bunlardan hiçbiri bulunmuyorsa bu durum nonanjinal göğüs ağrısı olarak tanımlanmaktadır. Asemptomatik grupta ise göğüs ağrısı bulunmamaktadır.



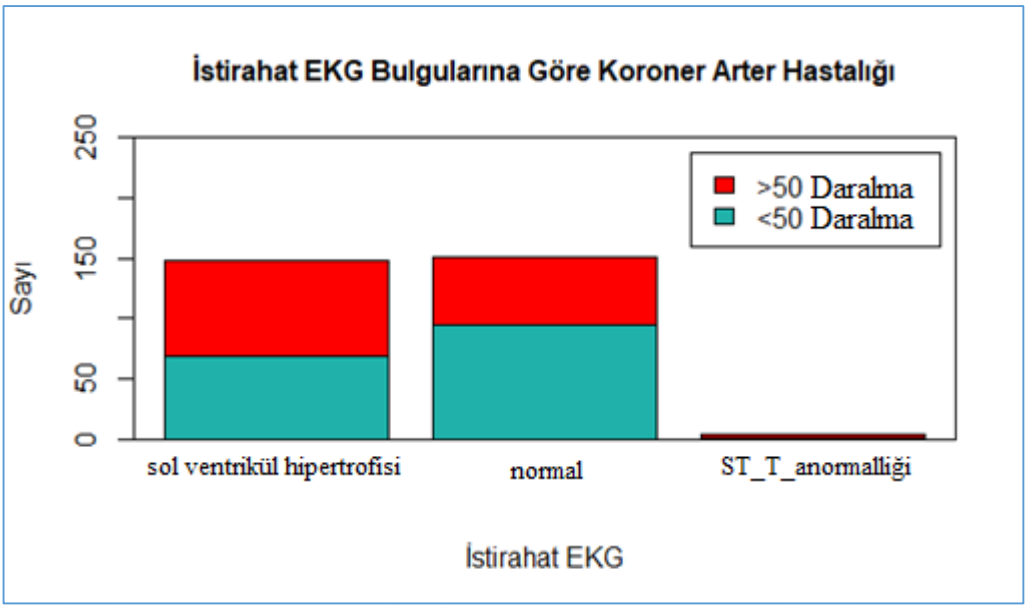
Şekil 3.7. Göğüs ağrısı tipi ve koroner arterlerde daralma

Şekil 3.8.'de verilen, açlık kan şekeri ve koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde, açlık kan şekeri 120 mg/dl' den fazla olan grupta koroner arterlerde ciddi daralma oranının daha yüksek olduğu görülmektedir. Diyabet, koroner kalp hastalığı için en önemli bir risk faktörlerinden biridir. Diyabetik hastalarda koroner arter hastalığı riski, 2 ila 4 kat arasında artmaktadır. Diyabet tanısı almamış ancak, diyabetin gelişim sürecinde yer alan bireylerde bozulmuş açlık glikozu ve bozulmuş glikoz toleransı bulunmaktadır. Bu kişilerde de kardiyovasküler hastalık riskinin artış gösterdiği bilinmektedir.



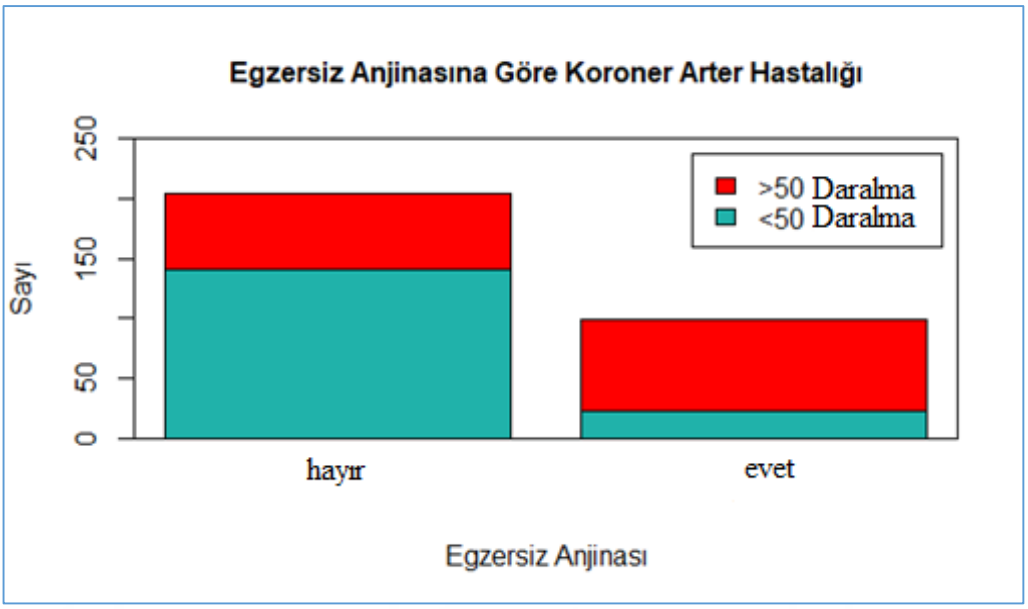
Şekil 3.8. Açlık kan şekeri ve koroner arterlerde daralma

Şekil 3.9.'da verilen istirahat elektrokardiyografisi ile koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde, istirahat EKG'sinde ST ve T dalga anormaliliği olan kişi sayısının oldukça az olduğu ve bunların çoğunda da ciddi koroner arter daralması olduğu görülmektedir. Bunun yanında, Elektrokardiyografi ile sol ventrikül hipertrofisi saptanan hastalarda sol ventrikül hipertrofisi olmayanlara göre koroner arter hastalığı oranının daha yüksek olduğu görülmektedir. İstirahat EKG'si koroner arter hastalığının başlangıç taramasında oldukça faydalıdır. Ancak göğüs ağrısı olan hastaların yaklaşık %60'ında EKG'nin normal olduğu bilinmektedir. EKG'de sol ventrikül hipertrofisi saptanması, koroner arter hastalığı için majör bir risk faktörü olan hipertansiyonda sık görülen bir bulgudur.



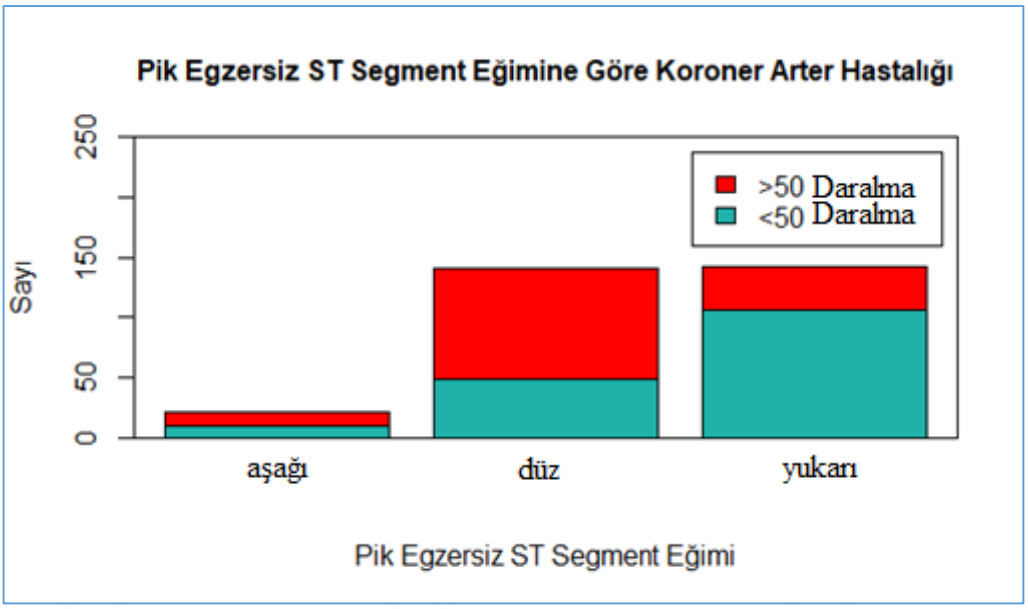
Şekil 3.9. İstirahat EKG ve koroner arterlerde daralma

Egzersizle tetiklenen anjina ve koroner kalp hastalığı arasındaki ilişkiyi gösteren çubuk grafiğe bakıldığında (Şekil 3.10.) egzersiz anjinası olan hastalarda egzersiz anjinası olmayan hasta grubuna göre ciddi koroner arter hastalığı oranı belirgin bir şekilde yüksektir. Anjina pektoris, göğüs ağrısı anlamında kullanılır ve koroner arter hastalığının en önemli belirtisidir. Egzersiz, kalbin oksijen ihtiyacını artıran bir durumdur. Koroner arter hastalığı varlığında, egzersiz sırasında kalp kasının oksijen ihtiyacı ve sunumu arasında ortaya çıkan dengesizlik nedeniyle kalp kası hücrelerine eforla gereksinim duyulan miktarda oksijen sağlanamadığı için miyokard iskemisi gelişmekte ve göğüs ağrısı ortaya çıkmaktadır.



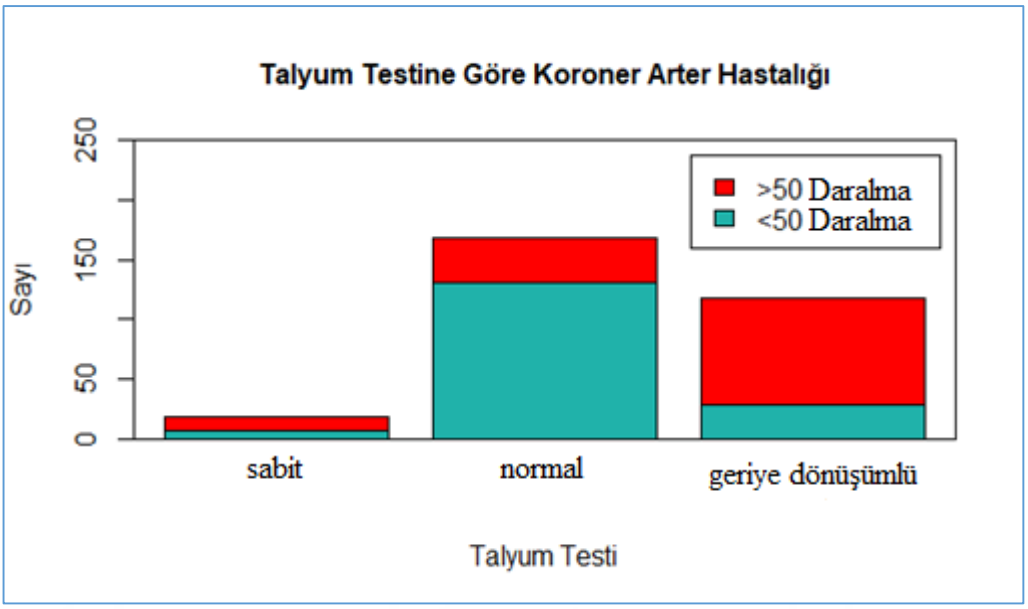
Şekil 3.10. Egzersizle tetiklenen anjina ve koroner arterlerde daralma

ST segment depresyonunun eğimi ve koroner kalp hastalığı arasındaki ilişkiyi gösteren çubuk grafik incelendiğinde (Şekil 3.11.), down sloping (aşağı eğimli) ST depresyonu olan kişi sayısı diğerlerine göre az olmakla birlikte bu grupta ciddi koroner arter hastalığı oranı en yüksektir. Flat (düz) ST depresyonu olanlarda up sloping (yukarı eğimli) ST depresyonu olanlara göre koroner arter hastalığı oranı daha yüksektir. Efor testi egzersiz esnasındaki ST segment değişiklikleri ile koroner arter hastalığının tanısında ve fonksiyonel kapasitenin değerlendirilmesinde kullanılabilir. Efor testinde ST depresyonu eğimi, koroner arter hastalığını olasılığını belirlemede yol göstericidir. Aşağı eğimli ST depresyonunda olasılık en yüksekken yukarı eğimli ST depresyonda en düşüktür.



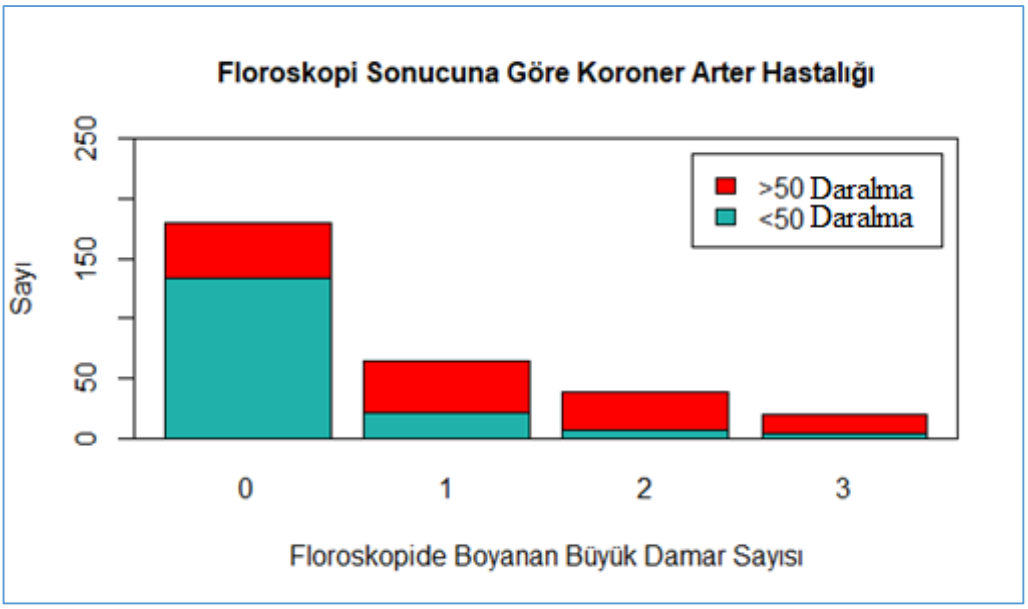
Şekil 3.11. Pik egzersiz ST segment eğimi ve koroner arterlerde daralma

Talyum testi ve koroner kalp hastalığı arasındaki ilişkiyi gösteren çubuk grafik incelendiğinde (Şekil 3.12.), koroner arter hastalığı oranının reversible defect (geriye dönüşümlü) tespit edilen grupta en yüksek olduğu saptanmıştır. Talyum testinde, fixed defect (sabit defekt) tespit edilen kişi sayısı diğer iki gruba göre oldukça az olmakla birlikte bu grupta koroner arter hastalığı oranı yarıdan fazladır. Talyum testi sonucu normal olan grupta ise koroner arter hastalığı oranı beklendiği gibi düşüktür. Nükleer görüntüleme, girişimsel olmayan bir biçimde koroner arter hastalığının tanısında ve kalp kası canlılığının değerlendirilmesinde sık olarak kullanılmaktadır. Bu teknikte, sabit defektler canlı iskemik (kanlanması azalmış) kalp dokusunu gösterdiği gibi ölü (skar) kalp dokusu alanlarını da gösterebilir. Reversible defektlerin varlığı ise iskemik canlı kalp kasının göstergesi olarak kabul edilmektedir.



Şekil 3.12. Talyum testi ve koroner arterlerde daralma

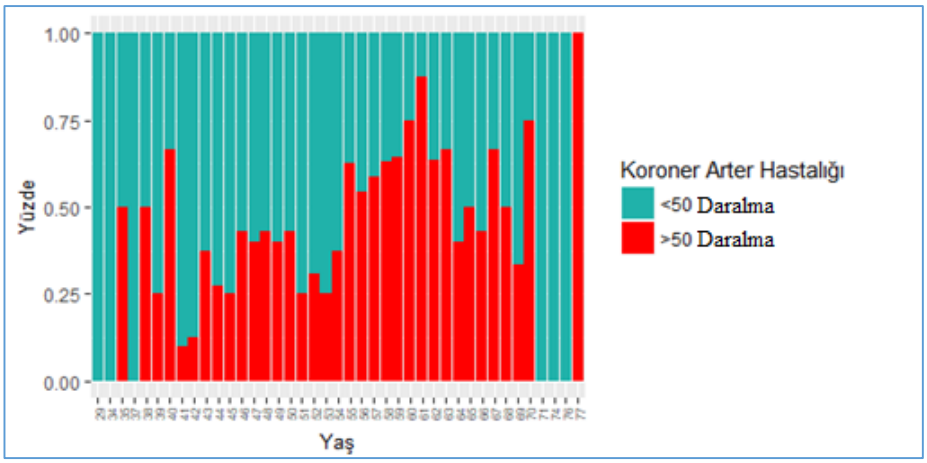
Fluoroskopi ile koroner kalsifikasyon tespit edilen büyük damar sayısı ve koroner kalp hastalığı arasındaki ilişkiyi gösteren çubuk grafik incelendiğinde (Şekil 3.13.), floroskopi ile kalsifikasyon görüntülenen damar sayısının artması ile birlikte koroner arterlerdeki ciddi daralma oranının da arttığı görülmektedir. Koroner kalsiyum skorlaması aterosklerotik plaklardaki kalsiyum miktarının ölçülerek koroner arter hastalığı riskinin belirlendiği bir testtir. Koroner arter kalsifikasyonu ile koroner arter hastalığı arasında doğrusal bir ilişki olduğu düşünülmektedir. Koroner arterde %50'den fazla darlık tespit edilenlerin %75'inde, önemli koroner arter hastalığı tespit edilmeyenlerin ise %10'unda kalsifikasyon görülmüştür (Adalet, 2013). Koroner kalsifikasyon miktarı yüksek olanlarda koroner arter hastalığının daha yaygın olduğu gösterilmiştir. Koroner arterlerde kalsifikasyonun tespit edilmemesi ciddi koroner darlığını dışlamada ve ilerde koroner arter hastalığı riskinin düşük olduğunu göstermede oldukça değerli bir öngörücüdür.



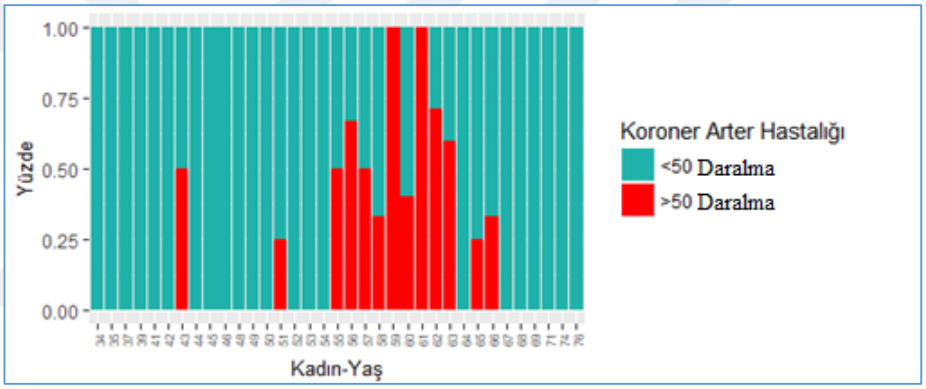
Şekil 3.13. Floroskopide boyanan damar sayısı ve koroner arterlerde daralma

3.2.3. Sayısal Değişkenler için Normalize Histogram Dağılımı

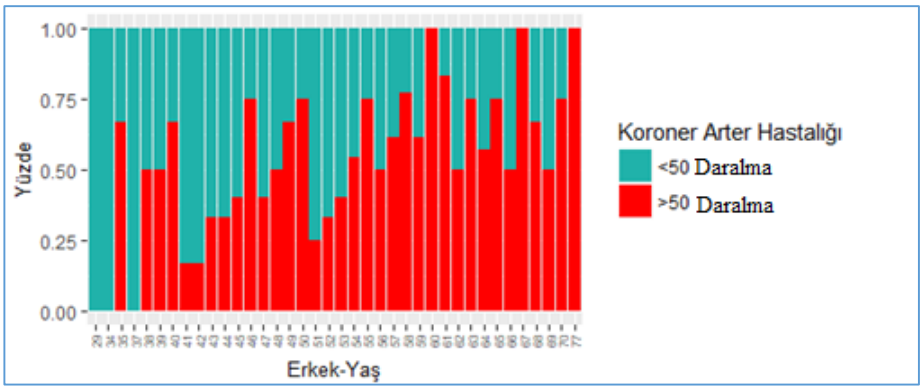
Şekil 3.14.'de verilen yaş değişkeninin histogram dağılımı incelendiğinde, 40 yaş ve sonrasında koroner arterlerdeki ciddi daralma oranının arttığı görülmektedir. Koroner arter hastalığı görülme sıklığı yaş ile artmaktadır. Yaş en önemli koroner arter hastalığı risk faktörü olarak düşünülmektedir. Bu durum histogram dağılımı ile uyumludur. Literatürde erkeklerde 45 yaşından, kadınlarda ise 55 yaşından büyük olmak koroner arter hastalığı için majör risk faktörlerinden biri olarak kabul edilmektedir. Veri kümesi kadın ve erkek olarak ayrıldıktan sonra oluşturulan histogram dağılımları Şekil 3.15 ve 3.16'da verilmiştir. Kadınlarda yaş değişkenin histogram dağılımının medikal literatürle oldukça uyumlu olduğu saptanmıştır.



Şekil 3.14. Yaş değişkeninin histogram dağılımı

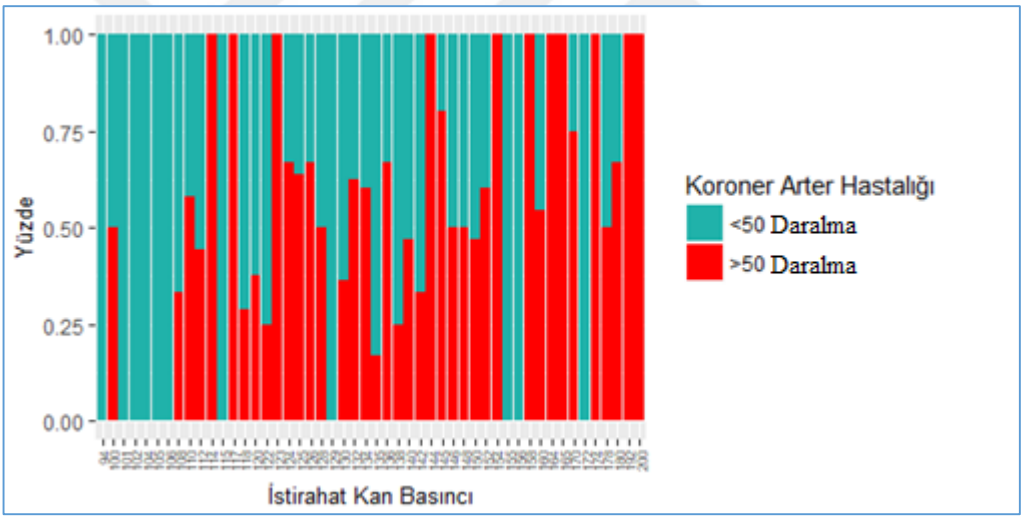


Şekil 3.15. Kadınlarda yaş değişkeninin histogram dağılımı

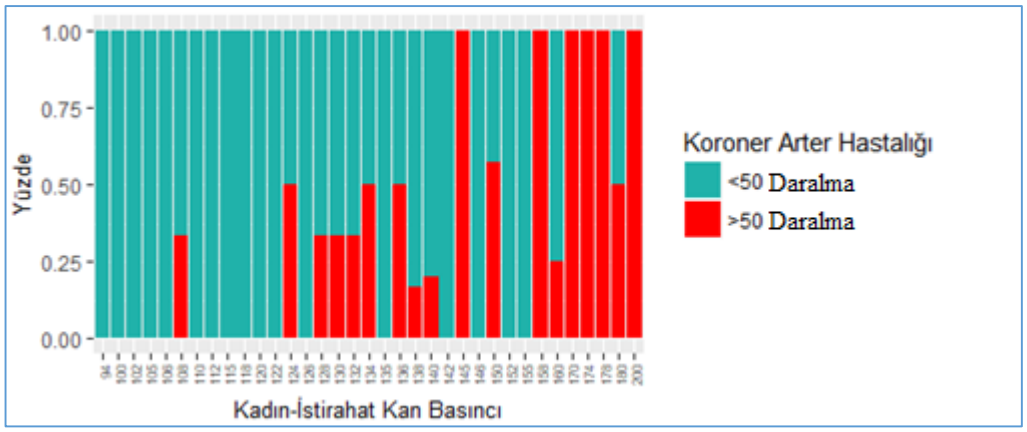


Şekil 3.16. Erkeklerde yaş değişkeninin histogram dağılımı

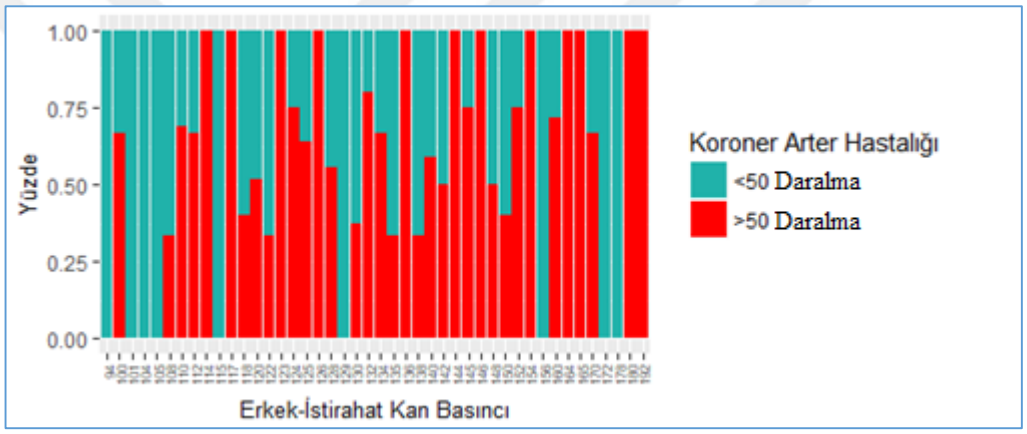
Şekil 3.17.'de verilen istirahat kan basıncını değişkeninin histogram dağılımı incelendiğinde, kan basıncının yüksek olduğu alanlarda, koroner arterlerdeki ciddi daralma oranının da arttığı görülmektedir. Literatürde hipertansiyon, koroner arter hastalığı için temel risk faktörü olarak kabul edilmektedir. Kan basıncındaki yükselmenin koroner arter hastalığı riskini arttırdığını gösteren çok sayıda çalışma bulunmaktadır. Koroner arter hastalığı hipertansiyonu olan hastalarda normal tansiyon değerlerine sahip bireylere göre 2-3 kat daha fazla olduğu gösterilmiştir. Veri kümesi kadın ve erkek olarak ayrıldıktan sonra oluşturulan istirahat kan basıncı değişkenine ait histogram dağılımları Şekil 3.18. ve 3.19.'da verilmiştir. Kadınlarda istirahat kan basıncı değişkenin histogram dağılımının medikal literatürle oldukça uyumlu olduğu görülürken erkeklerde bu ilişkiyi açık bir şekilde gösteren bir örüntü elde edilememiştir.



Şekil 3.17. İstirahat kan basıncı değişkeninin histogram dağılımı

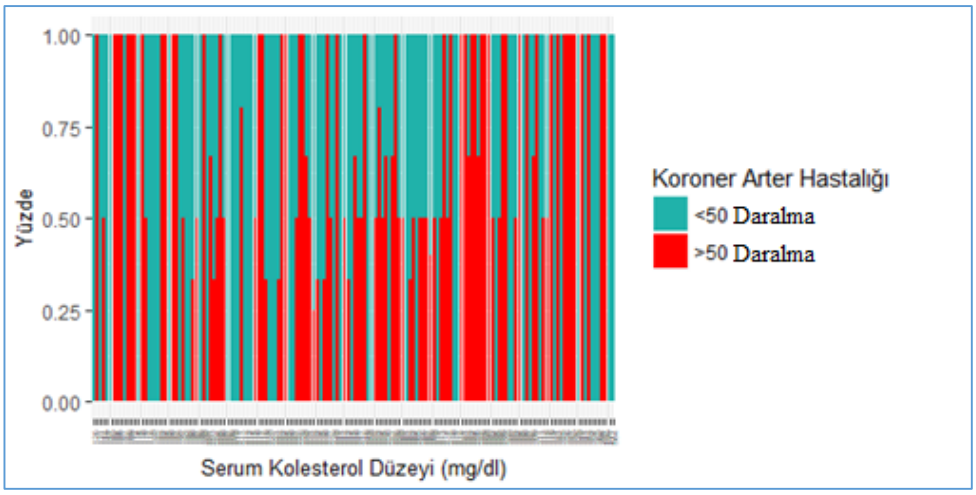


Şekil 3.18. Kadın istirahat kan basıncı değişkeninin histogram dağılımı

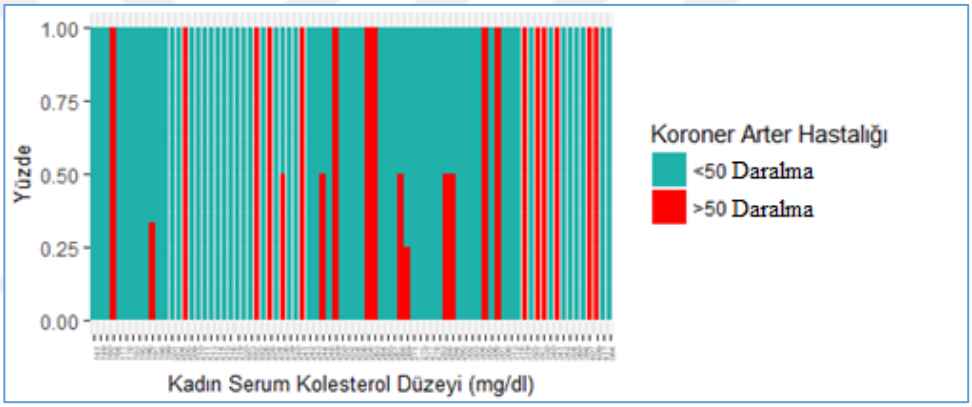


Şekil 3.19. Erkek istirahat kan basıncı değişkeninin histogram dağılımı

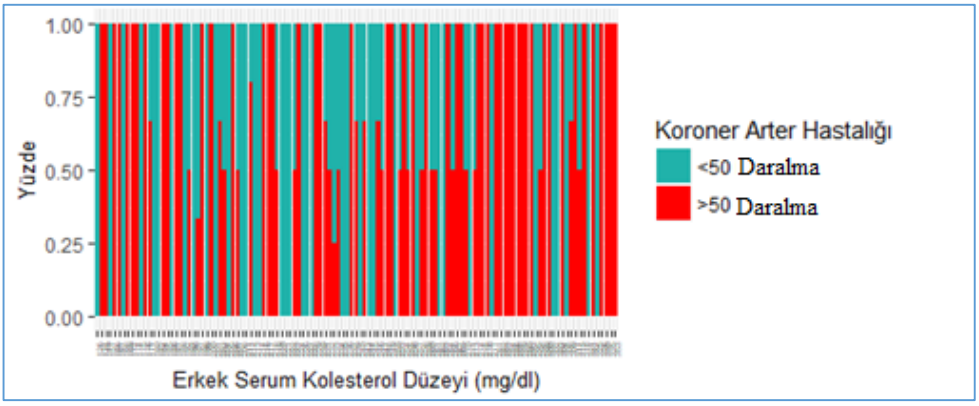
Şekil 3.20.'de verilen kolesterol değişkeninin histogram dağılımı incelendiğinde, kolesterol değerinin yüksek olduğu alanlarda koroner damarlardaki ciddi daralma oranı da hafif düzeyde arttığı görülmektedir. Ancak, literatürde koroner arter hastalığı açısından temel risk faktörü olarak kabul edilen kolesterol yüksekliğinin histogramda ciddi daralma oranında daha belirgin bir artış göstermesi beklenmektedir. Veri kümesi kadın ve erkek olarak ayrıldıktan sonra oluşturulan kolesterol değişkenine ait histogram dağılımları Şekil 3.21. ve 3.22.'de verilmiştir. Ancak her iki dağılımda da koroner arter hastalığı ve kolesterol seviyesi arasındaki ilişkiyi açık bir şekilde gösteren bir örüntü elde edilememiştir.



Şekil 3.20. Serum kolesterol düzeyi değişkeninin histogram dağılımı

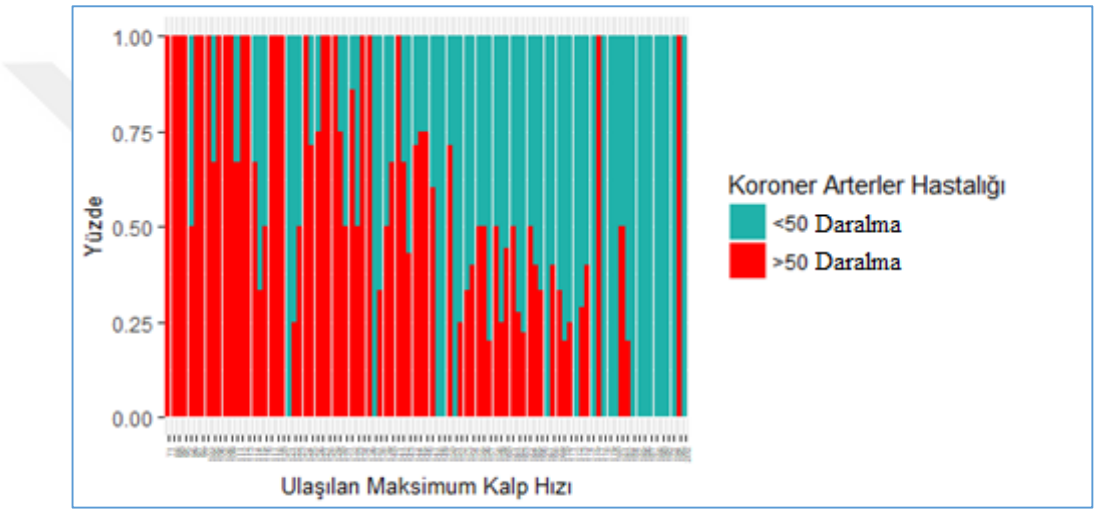


Şekil 3.21. Kadın serum kolesterol düzeyi değişkeninin histogram dağılımı



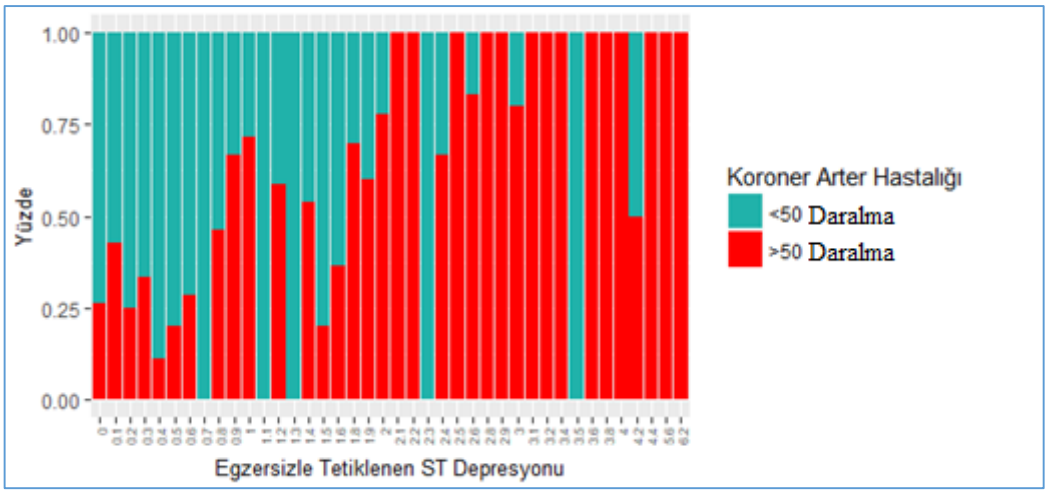
Şekil 3.22. Erkek serum kolesterol düzeyi değişkeninin histogram dağılımı

Şekil 3.23.'de verilen maksimum kalp hızı değişkeninin histogram dağılımı incelendiğinde efor sırasında ulaşılabilen maksimum kalp hızı azaldıkça, koroner arterlerdeki ciddi daralma oranının arttığı ve maksimum kalp hızına ulaşan, fonksiyonel kapasitesi iyi olan grupta ise azaldığı görülmektedir. Efor testi yapılacak hastalar için hasta yaşına göre öngörülen maksimum kalp hızı belirlenir ve efor sırasında bu kalp hızına ulaşması hedeflenir. Ancak, hastaların bir kısmı koroner arter hastalığı varlığı, kalp yetmezliği, kronik akciğer hastalığı, ortopedik problemler gibi çeşitli nedenlerle bu maksimum kalp hızına ulaşamamaktadır.



Şekil 3.23. Ulaşılan maksimum kalp hızı değişkeninin histogram dağılımı

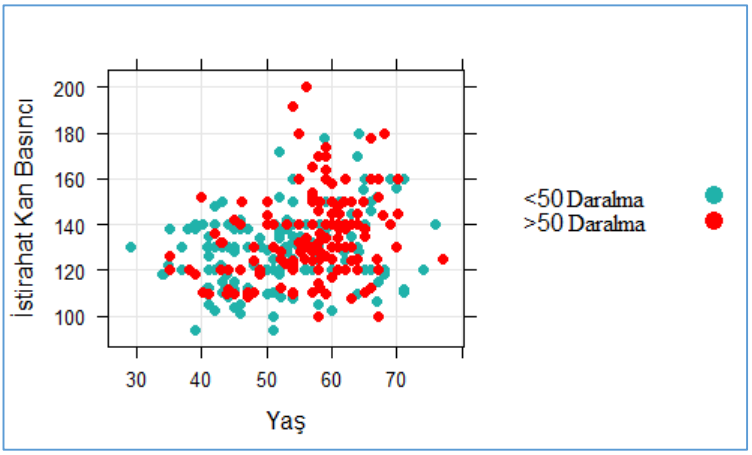
Şekil 3.24.'de verilen ST depresyon değişkeninin histogram dağılımı incelendiğinde, istirahat EKG'sine göre ST segmentindeki çökme miktarı arttıkça, koroner arterlerdeki ciddi daralma oranının da belirgin bir biçimde arttığı görülmektedir. Efor testinde koroner arter hastalığı varlığını düşündüren temel bulgu ST segmentinde çökmenin izlenmesidir. ST depresyonunun testin erken safhasında oluşması, normale dönmesi için gereken sürenin uzun olması ve ST depresyon miktarının fazla olması testin koroner damar daralmasını gösterme ihtimalini artırdığı kabul edilmektedir. Bu durum histogram verileri ile uyumludur.



Şekil 3.24. Egzersizle tetiklenen ST depresyonu değişkeninin histogram dağılımı

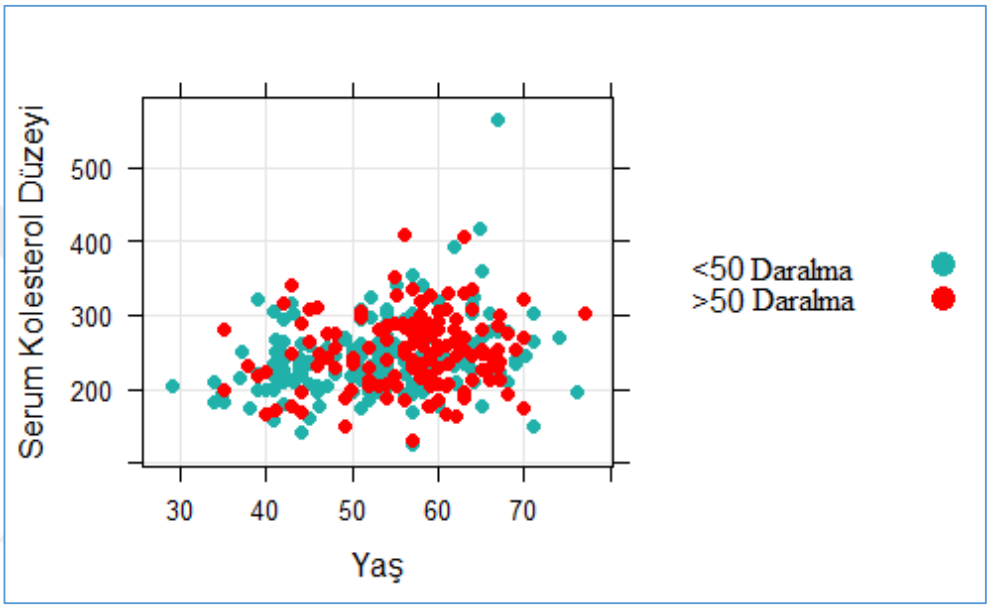
3.2.4. Sayısal Değişken Çiftleri İçin Saçılım Grafikleri

Şekil 3.25.'de verilen yaş ve istirahat halindeki kan basıncı arasındaki saçılım grafiği incelendiğinde, kan basıncının 120 mmHg' nın üzerinde olduğu 50 yaş üzerindeki hastalarda koroner arterlerde ciddi daralma oranının fazla olduğu görülmektedir. Hipertansiyon sıklığının yaşla birlikte arttığı bilinmektedir. Hem yaş hem de hipertansiyon koroner arter hastalığı açısından temel risk faktörleridir. Birlikte bulunmaları koroner arter hastalığı görülme sıklığını artırmaktadır.



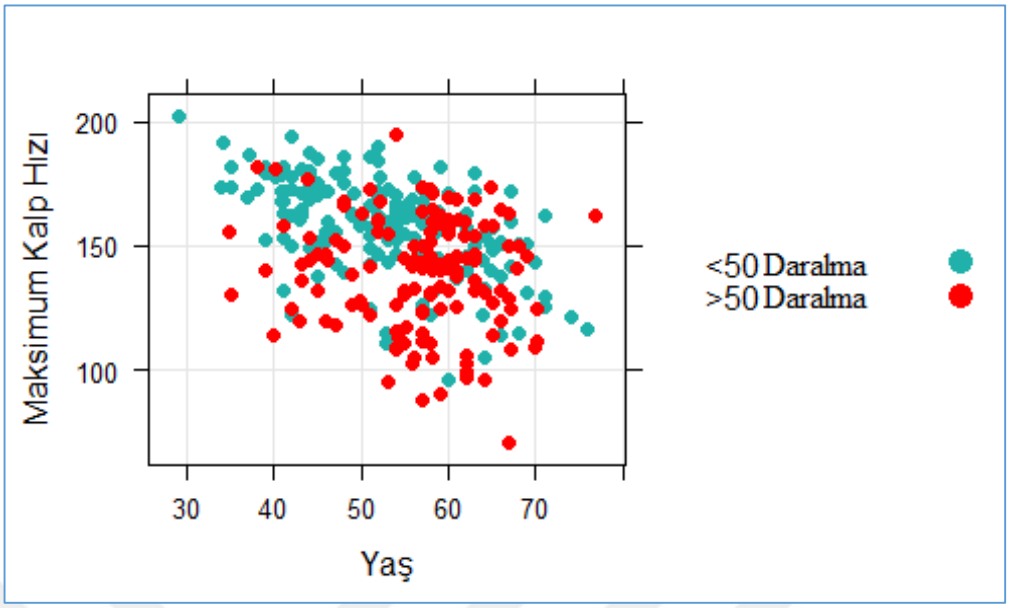
Şekil 3.25. Yaş ve istirahat kan basıncı değişkenleri saçılım grafiği

Şekil 3.26.'da verilen yaş ve serum kolesterol düzeyi arasındaki saçılım grafiği incelendiğinde, kolesterol düzeyi 200 mg'ın üzerinde ve yaşı 50'nin üzerinde olan grupta koroner arterlerde ciddi daralma oranının arttığı görülmektedir. Hiperlipidemi ve yaş koroner arter hastalığı açısından bağımsız iki temel risk faktörüdür. Bu iki temel risk faktörünün birlikteliği beklendiği gibi ciddi koroner arter hastalığı riskini artırmaktadır.



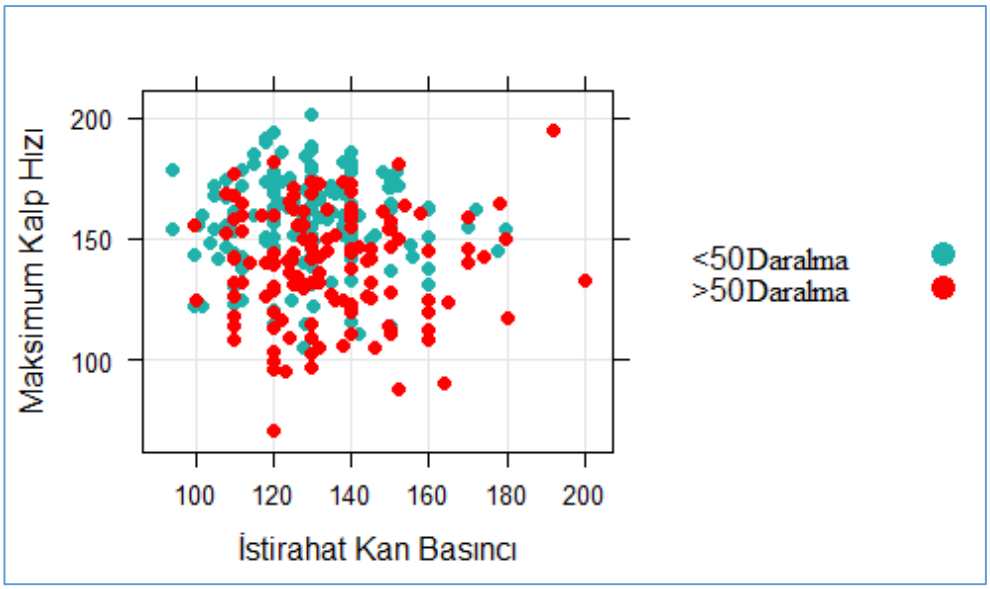
Şekil 3.26. Yaş ve serum kolesterol düzeyi değişkenleri saçılım grafiği

Şekil 3.27.'de verilen yaş ve ulaşılan maksimum kalp hızı arasındaki saçılım grafiği incelendiğinde, 55 yaş üzerinde ve maksimum kalp hızı 150'nin altında olan bireylerde ciddi koroner arter hastalığı oranının yüksek olduğu görülmektedir. Efor testinde ulaşılması hedeflenen maksimum kalp hızı yaşa göre belirlenmektedir. Koroner arter hastalığı varlığında hastanın ulaşabildiği maksimum kalp hızı hem fonksiyonel kapasitenin azalması hem de hedeflenen kalp hızına ulaşılmadan ST depresyonunun ortaya çıkması nedeniyle testin sonlandırılması nedeniyle düşük seviyelerde kalabilmektedir. Koroner arter hastalarında düşük maksimum kalp hızı aynı zamanda yaygın damar hastalığı ve kötü prognoz göstergesi olarak kabul edilmektedir.



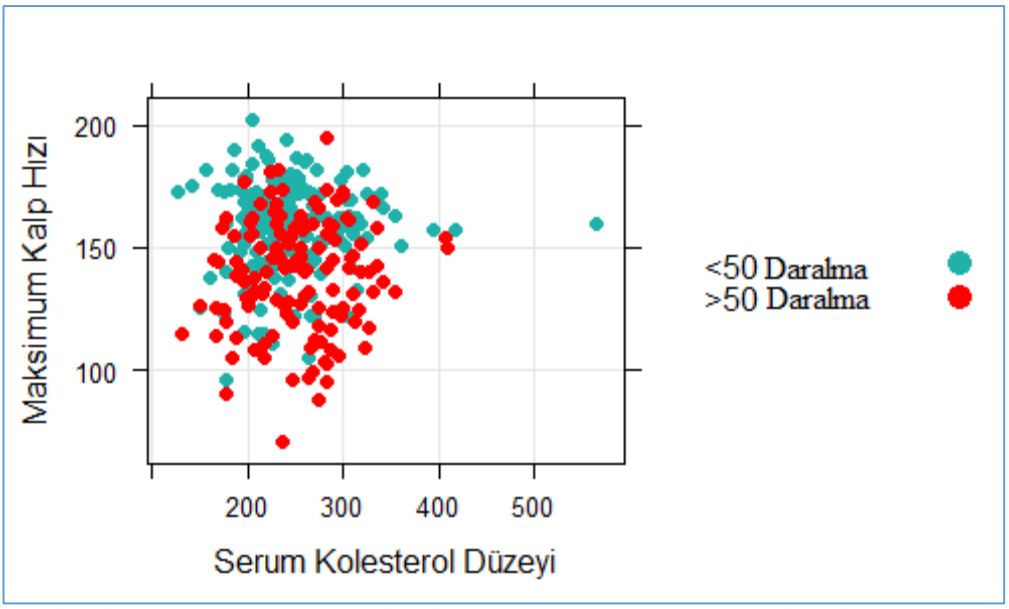
Şekil 3.27. Yaş ve maksimum kalp hızı değişkenleri saçılım grafiği

Şekil 3.28.'de verilen istirahat halindeki kan basıncı ve ulaşılan maksimum kalp hızı arasındaki saçılım grafiği incelendiğinde, kan basıncı 140 mmHg' nın altında ve ulaşılan maksimum kalp hızı yüksek olan hasta grubunda ciddi koroner arter hastalığı oranı düşük izlenmektedir. Koroner arter hastalığının temel risk faktörlerinden biri olan hipertansiyon yokluğunda, koroner kalp hastalığı oranının daha düşük olduğu bilinmektedir. Efor testinde ulaşılan maksimum kalp hızı arttıkça koroner arter hastalığı riski azalmaktadır. Saçılım grafiğinde bu iki durumun birlikteliği beklendiği şekilde ciddi koroner daralmanın az olduğu hasta grubunu oluşturmaktadır.



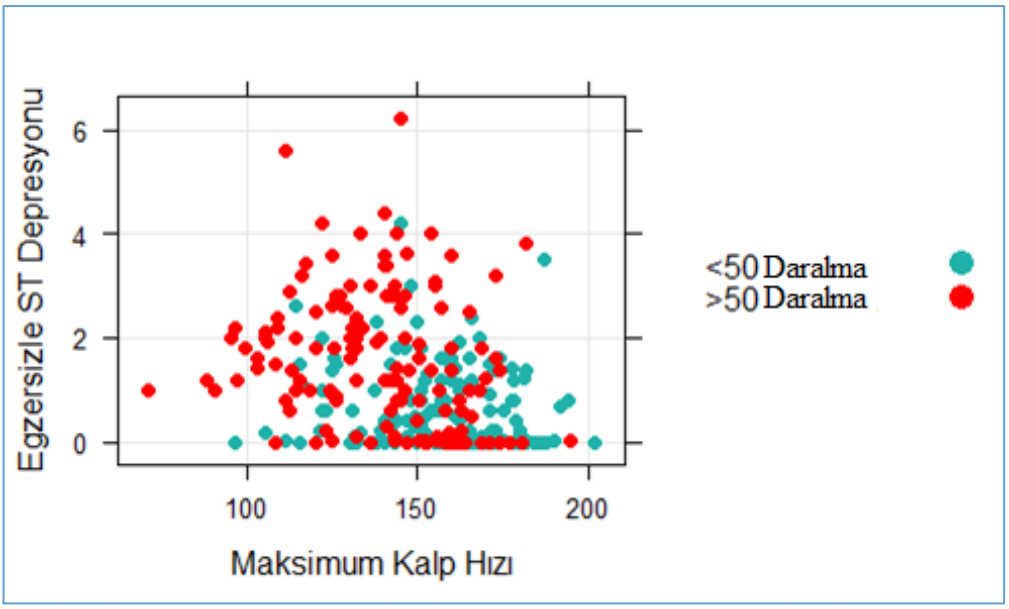
Şekil 3.28. Kan basıncı ve maksimum kalp hızı değişkenleri saçılım grafiği

Şekil 3.29.'da verilen kolesterol seviyesi ve ulaşılan maksimum kalp hızına ulaşma arasındaki saçılım grafiği incelendiğinde ulaşılan maksimum kalp hızı düşük ve kolesterol düzeyleri yüksek olan grupta ciddi koroner arter hastalığı oranının yüksek olduğu görülmektedir. Hiperlipidemi, koroner arter hastalığını artırdığı iyi bilinen temel bir risk faktörüdür. Efor testinde kötü prognoz ve yaygın koroner arter hastalığı varlığı ile ilişkili olan düşük maksimum kalp hızına sahip hastalarda hiperlipidemi varlığı beklendiği gibi ciddi koroner arter hastalığı ihtimalini arttırmaktadır.



Şekil 3.29. Kolesterol düzeyi ve maksimum kalp hızı değişkenleri saçılım grafiği

Şekil 3.30.'da verilen maksimum kalp hızı ve ST depresyon arasındaki saçılım grafiği incelendiğinde ulaşılan maksimum kalp hızı 150'nin altında ve ST depresyon miktarı 2 mm'nin üzerinde olan grupta ciddi koroner arter hastalığı oranının yüksek olduğu görülmektedir. Efor testinde koroner arter hastalığı olma ihtimalini artıran en önemli kriterlerden biri ST depresyon miktarının fazla olmasıdır. ST depresyonunun 2 mm'den fazla olduğu grupta yaygın koroner arter hastalığı ve kötü prognoz göstergesi olan düşük maksimum kalp hızı varlığı beklendiği gibi ciddi koroner arter darlığını göstermektedir.



Şekil 3.30. Maksimum kalp hızı ve kolesterol düzeyi değişkenleri saçılım grafiği

Çalışmada, saçılım grafikleri ile birlikte veri kümesinin sayısal değişken çiftleri arasındaki ilişkiyi incelemek amacıyla korelasyon analizi de yapılmıştır. Korelasyon analizi öncesinde değişkenlerin dağılım özellikleri Shapiro-Wilk normallik testi ile incelenmiştir. Shapiro-Wilk normallik testi sonuçları Çizelge 3.5.'de verilmiştir. Normallik test sonuçlarına bakıldığında p değerlerinin tamamı 0.05 değerinden küçük olduğu için değişkenler normal dağılıma sahip değildir. Değişkenlerin hiçbiri normal dağılım özelliği göstermediği için korelasyon analizinde Spearman Korelasyon uygulanmıştır. Spearman Korelasyon analizi sonuçları Çizelge 3.6.'da verilmiştir. Korelasyon analizi sonuçları incelendiğinde, efor sırasında ulaşılan maksimum kalp hızını gösteren *thalach* değişkeninin yaş ve egzersizle tetiklenen ST depresyonu gösteren *oldpeak* değişkenleri ile az ve orta düzeyde doğrusal ilişkiye sahip olduğu görülmektedir.

Çizelge 3.5. Sayısal değişkenler için Shapiro-Wilk normallik testi sonuçları

Değişkenler	w değeri	p değeri
age	0,98	0,01
trestbps	0,96	1,8e-06
chol	0,94	5,9e-09
thalach	0,97	6,9e-05
oldpeak	0,84	2,2e-16

Çizelge 3.6. Sayısal değişkenler için Spearman Korelasyon sonuçları

Değişkenler	age	trestbps	chol	thalach	oldpeak
age	1,00	0,29	0,19	**-0,39**	0,25
trestbps	-	1,00	0,13	-0,04	0,15
chol	-	-	1,00	-0,03	0,03
thalach	-	-	-	1,00	**-0,43**
oldpeak	-	-	-	-	1,00

3.3. Macaristan Veri Kümesi Analizi

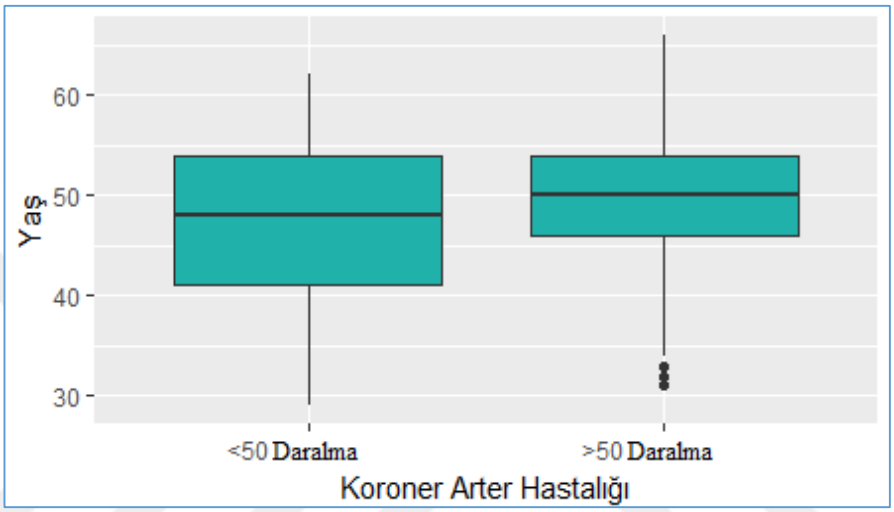
Macaristan veri kümesine ait analizler Cleveland veri kümesinden farklılıkları temel alınarak yapılmıştır. Macaristan veri kümesinin tanımlayıcı istatistikler Çizelge 3.7.'de verilmiştir.

Çizelge 3.7. Veri kümesi değişkenleri tanımlayıcı istatistikleri

Değişken	Minimum	1.Çeyrek	Ortanca	Ortalama	3.Çeyrek	Maksimum
Yaş	29	42	49	48	54	66
Trestbps	92	120	130	133	140	200
Chol	85	212	244	251	277	603
Thalach	82	122	140	139	155	190
Oldpeak	0	0	0	0.5	1	5

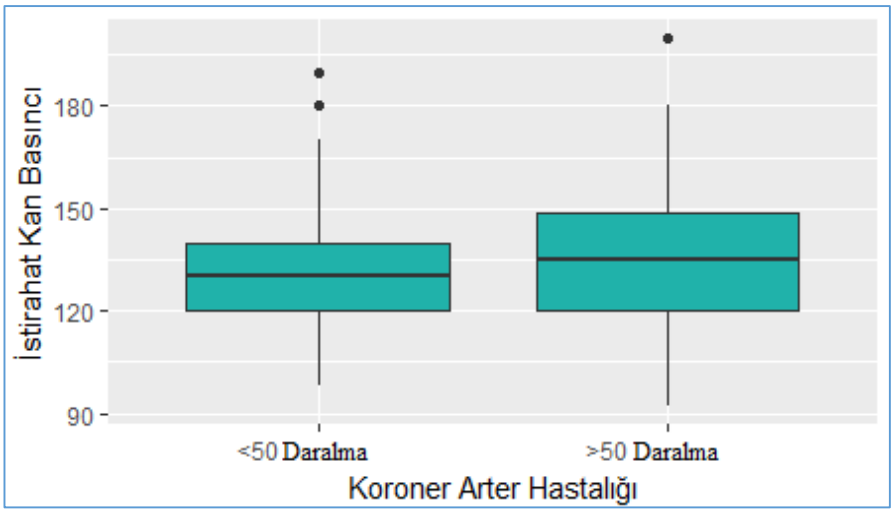
3.3.1. Sayısal Değişkenlerin Hedef Değişkene Göre Kutu Grafikleri:

Şekil 3.31.'de verilen Macaristan veri kümesi yaş değişkeninin kutu grafiği incelendiğinde her iki grupta da, ortanca yaşın, Cleveland veri kümesinden farklı olarak azaldığı ve 50 yaş civarında olduğu görülmektedir. Buna ek olarak, ciddi koroner arter daralması olan grupta 3 uç değer (outlier) bulunmaktadır.



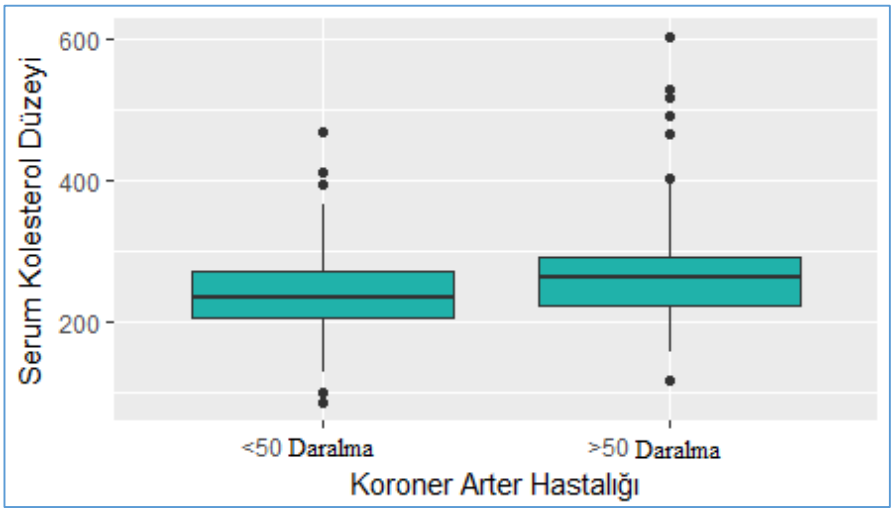
Şekil 3.31. Yaş değişkeni kutu grafiği

Şekil 3.32.'de verilen istirahat kan basıncı değişkeninin kutu grafiği incelendiğinde, Cleveland veri kümesine benzer bir biçimde koroner daralmanın olduğu her iki grupta da ortanca değerin birbirine oldukça yakın olduğu görülmektedir. İlk grupta 2 uç değer, ikinci grupta ise 1 uç değer bulunmaktadır.



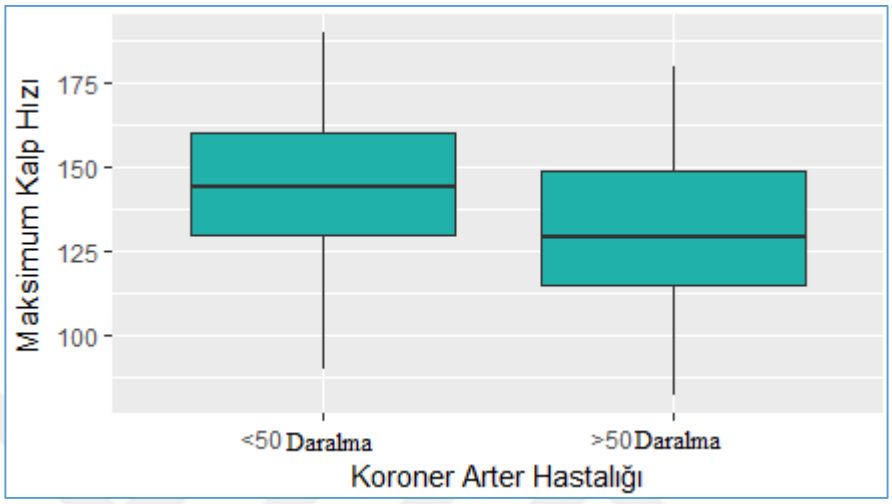
Şekil 3.32. İstirahat kan basıncı düzeyi değişkeni kutu grafiği

Şekil 3.33.'de verilen serum kolesterol düzeyinin kutu grafiği incelendiğinde Cleveland veri kümesine benzer biçimde koroner daralmanın olduğu her iki grupta da ortanca kolesterol düzeyi benzerlik göstermektedir. Ayrıca, ciddi daralmanın olmadığı grupta 5 uç değer saptanırken diğer gruptaki uç değerler Cleveland veri kümesine göre artmıştır.



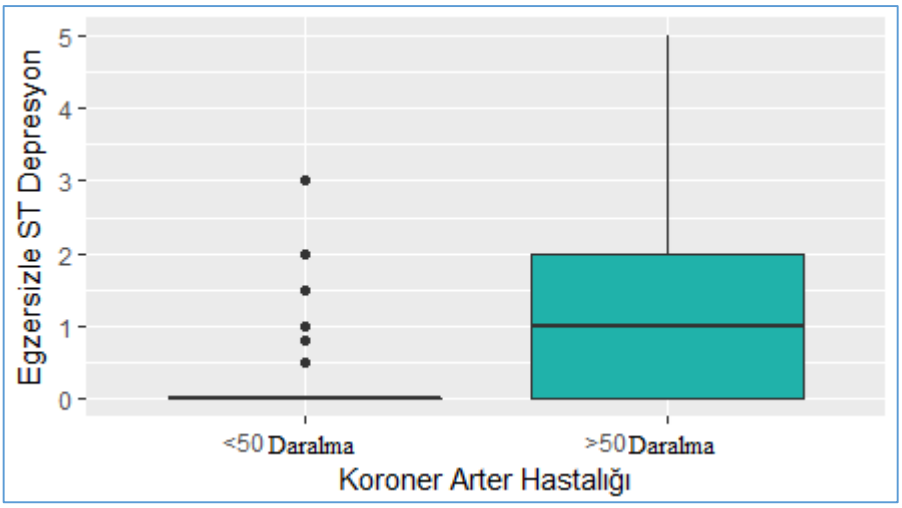
Şekil 3.33. Serum kolesterol düzeyi değişkeni kutu grafiği

Şekil 3.34.'de verilen maksimum kalp hızı değişkeninin kutu grafiği incelendiğinde Cleveland veri kümesine benzer bir biçimde ciddi koroner daralmanın olmadığı grubun ortanca kalp hızı değerinin diğer gruba göre yüksek olduğu görülmektedir.



Şekil 3.34. Maksimum hızı kalp değişkeni kutu grafiği

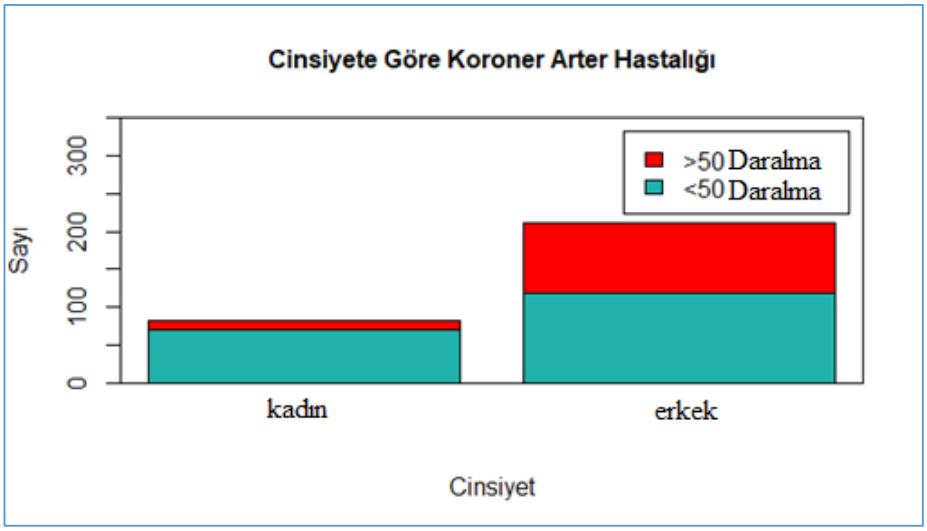
Şekil 3.35.'de verilen egzersizle ST depresyon değişkeninin kutu grafiği incelendiğinde ciddi daralmanın olmadığı grubun ortanca ST depresyonu değerinin, Cleveland veri kümesinden farklı olarak, 0 değerinde yığıldığı görülmektedir. İki grup arasında ortanca değerin ve kutuların kapsadığı alanların farklı olması egzersizle ST depresyon değişkeninin sınıflama modeli açısından önemli bir değişken olduğunu göstermektedir.



Şekil 3.35. Egzersizle ST depresyon değişkeni kutu grafiği

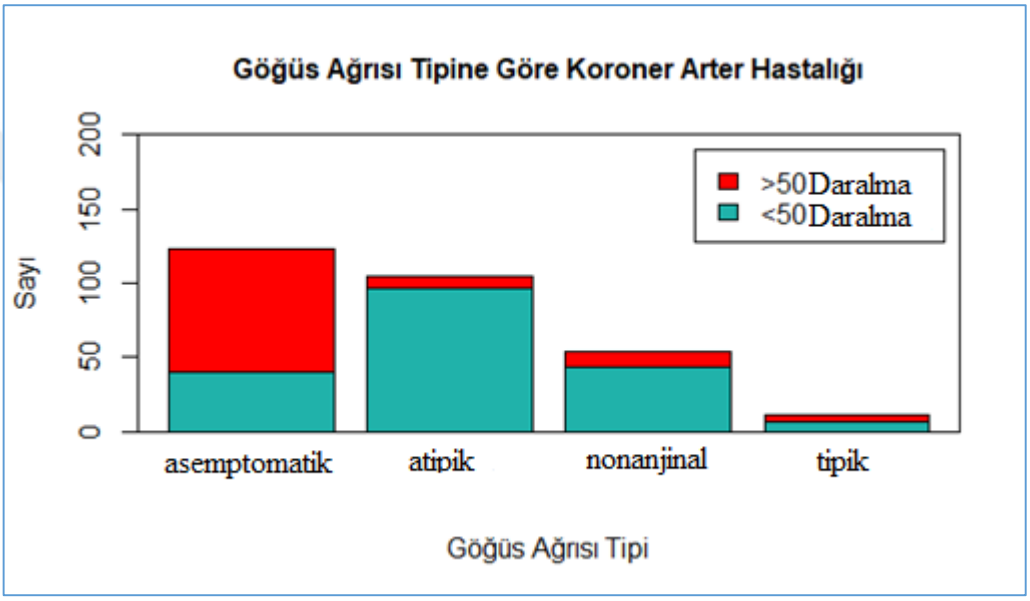
3.3.2. Kategorik Değişkenler için Çubuk Grafikleri

Şekil 3.36.'da verilen macaristan veri kümesinde cinsiyet ile koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde sonuçların Cleveland veri kümesi ile oldukça benzer olduğu görülmektedir. Her iki veri kümesinde de erkeklerde kadınlara oranla koroner arter hastalığının daha fazla olduğu saptanmıştır.



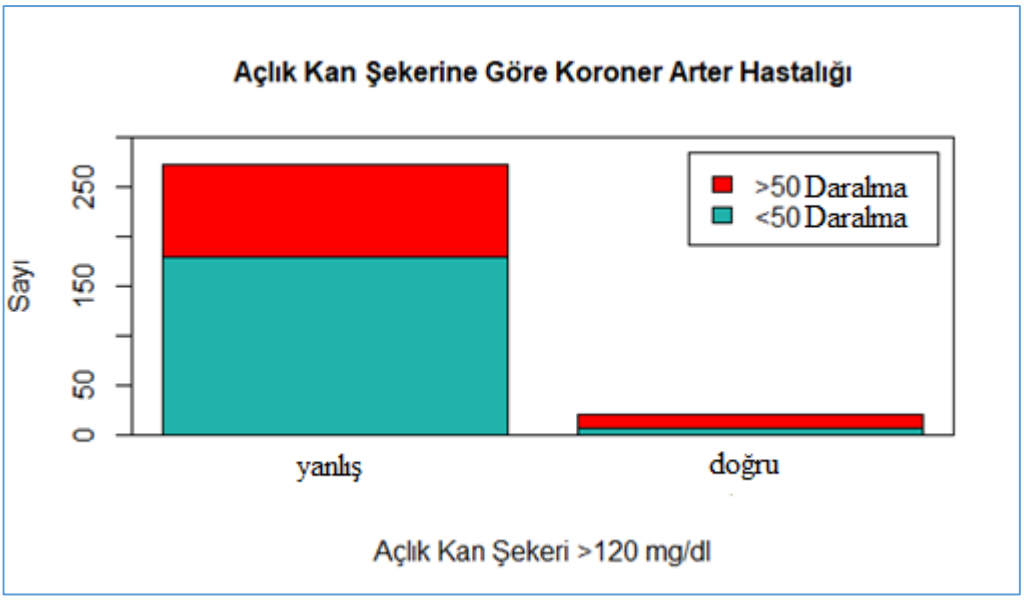
Şekil 3.36. Cinsiyet ve koroner arterlerde daralma

Şekil 3.37.'de verilen Macaristan veri kümesinde göğüs ağrısı tipi ile koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde her iki veri kümesinde de asemptomatik hasta sayısının daha fazla olduğu görülmektedir. Göğüs ağrısı olan grup içerisinde ise atipik göğüs ağrısı olan hasta sayısı Macaristan veri kümesinde daha fazla, tipik göğüs ağrısı olan hasta sayısının ise daha az olduğu görülmektedir. Ancak her iki veri kümesinde de ciddi koroner damar darlığı oranlarının benzer olduğu saptanmıştır.



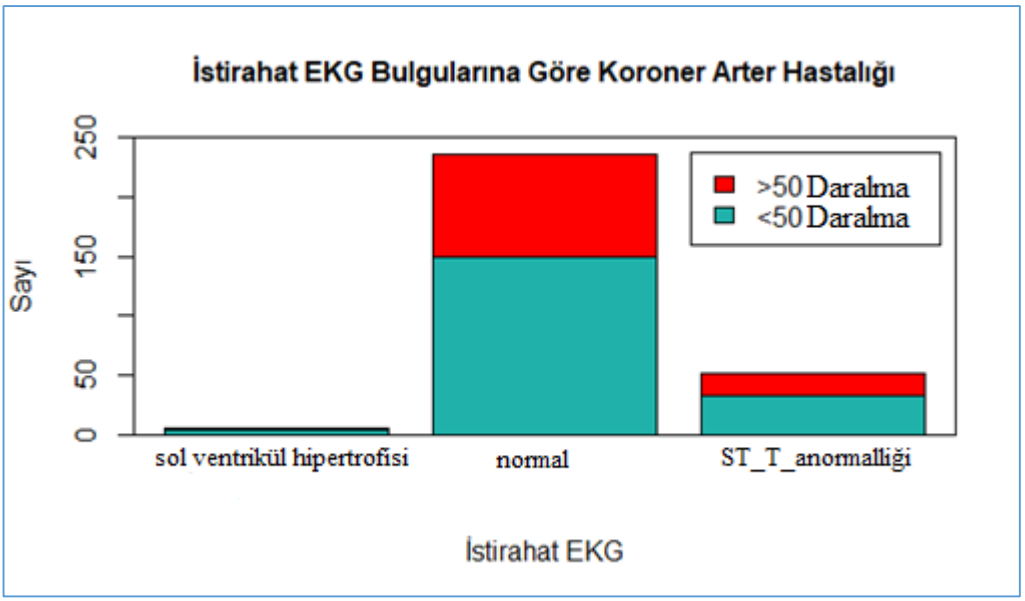
Şekil 3.37. Göğüs ağrısı tipi ve koroner arterlerde daralma

Şekil 3.38.'de verilen Macaristan veri kümesinde açlık kan şekeri ile koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde açlık kan şekeri düzeyinin 120 mg/dl' nin altında olduğu hastalarda, ciddi koroner damar darlığı oranının Cleveland veri kümesine göre daha az olduğu görülmektedir. Diğer grupta ise hasta sayısı daha az olmakla birlikte ciddi koroner kalp hastalığı oranı daha fazladır. Bu sonuçlara bakıldığında Macaristan veri kümesindeki oranların medikal literatürde belirtilen oranlara biraz daha yakın olduğu görülmektedir.



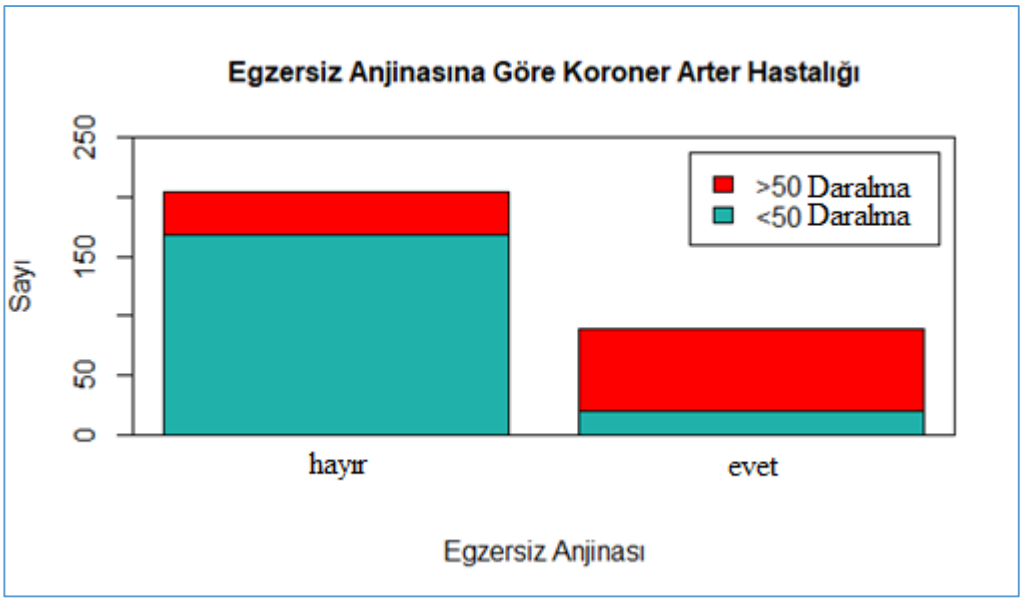
Şekil 3.38. Açlık kan şekeri ve koroner arterlerde daralma

Şekil 3.39.'da verilen Macaristan veri kümesinde istirahat elektrokardiyografisi ile koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde Cleveland veri kümesine göre sol ventrikül hipertrofisi ve ST anormaliliği olan bireylerin sayısının az iken normal istirahat EKG'sine sahip birey sayısının daha fazla olduğu görülmektedir. Sol ventrikül hipertrofisi olan bireylerin az olması veri kümesinde hipertansif hasta sayısının daha az olması ile açıklanabilir. Sol ventrikül hipertrofisi ve ST anormaliliği olan olan gruplarda Cleveland veri kümesine göre ciddi koroner kalp hastalığı daha az, normal istirahat EKG'sine sahip olan grupta ise benzer oranlarda bulunmuştur. Normal istirahat EKG' sine sahip grup dışarıda bırakıldığında diğer iki gruba ait verilerde ciddi koroner damar daralma oranı Cleveland veri kümesinde literatürle daha uyumlu olduğu görülmektedir.



Şekil 3.39. İstirahat EKG ve koroner arterlerde daralma

Şekil 3.40.'da verilen Macaristan veri kümesinde egzersizle tetiklenen anjina ile koroner arter hastalığı arasındaki ilişkiyi gösteren çubuk grafiği incelendiğinde Cleveland veri kümesine göre egzersiz anjinası olmayan grupta ciddi koroner damar darlığı oranı daha düşüktür. Egzersiz anjinası olan grupta ise oranların oldukça benzer olduğu görülmektedir. Egzersiz anjinası olmayan gruptaki oranlar Macaristan veri kümesinde medikal literatürdeki oranlara biraz daha yakındır.



Şekil 3.40. Egzersizle tetiklenen anjina ve koroner arterlerde daralma

Macaristan veri kümesinde yer alan sayısal değişkenler için oluşturulan histogram dağılımı ve sayısal değişken çiftleri için oluşturulan saçılım grafikleri incelendiğinde Cleveland veri kümesine göre önemli farklılıklar saptanmamıştır.

3.4. Rastgele Orman Algoritması

3.4.1. Cleveland Veri Kümesi

Veri kümesi üzerinde R programı aracılığıyla uygulanan Rastgele Orman algoritması sonuçları karışıklık matrisi, ROC eğrisi ve Gini indeksi olarak verilmiştir. Cleveland veri kümesi üzerinde, Cihan vd. (2018) çalışmalarında Karar Ağacı algoritmasından elde ettikleri 6 değişken temel alınarak uygulanan Rastgele Orman algoritması sonuçları Çizelge 3.8.'de karışıklık matrisi olarak verilmiştir. Karışıklık matrrisi bileşenlerinden modelin **doğruluk** oranı **%83,49**, **duyarlılığı** %**76,25** ve **seçiciliği** %**89,63** olarak hesaplanmıştır. Cihan vd. (2018) çalışmalarında elde edilen Karar Ağacı algoritması sonuçları ile karşılaştırıldığında % 78 olan doğruluk oranının arttığı

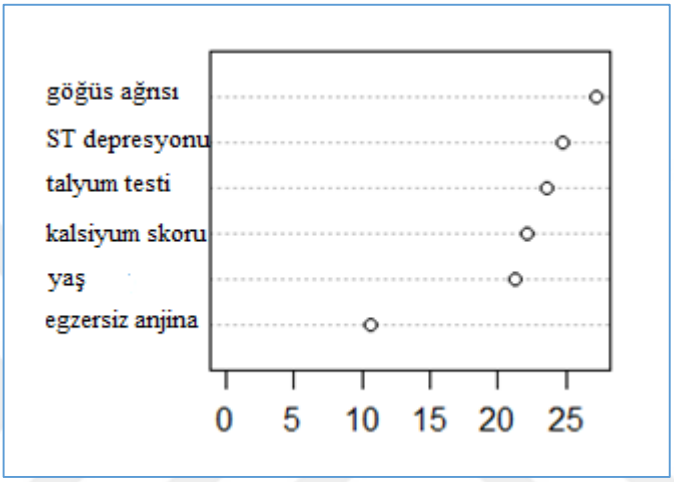
ve FP oranının azaldığı saptanmıştır. Ayrıca, FN grubundaki hasta sayısının da azaltıldığı belirlenmiştir.

Karar ağaçları tek bir örneklem üzerinde sınıflama yapmaktadır. Bu durum modelin güvenirliğini sınırlandırmaktadır. Rastgele Orman algoritmasında ise çok sayıda karar ağacının görüşleri sonucunda genel tek bir yargıya varıldığı için daha güvenilir tahminler yapılabilmektedir. Buna ek olarak, Rastgele Orman yönteminde tahmin sırasında tüm ağaçların tahminlerinin göz önünde bulundurulması veriler üzerinde daha iyi bir genelleme yapılabilmesini sağlamaktadır.

Medikal alanda yapılan hastalık sınıflama çalışmalarında FN oranı oldukça önemlidir. Bu grup hastalık olduğu halde yanlışlıkla hastalık yok şeklinde sınıflandırılan gruptur. Bu nedenle uygulanan sınıflama algoritması sonucunda bu sayının azaltılması hastalık yönetiminde kritik öneme sahiptir. Bununla birlikte, Şekil 3.41.'de Rastgele Orman algoritması Gini indeksine göre değişkenlerin önem sırası görülmektedir. Buna göre göğüs ağrısı tipini gösteren cp, egzersizle tetiklenen ST depresyonunu gösteren oldpeak ve talyum sintigrafi sonucunu gösteren thal değişkenleri sınıflandırmada en önemli 3 değişken olarak bulunmuştur. Anbarasi vd. (2010) yaptıkları çalışmada kalp hastalıkları veri kümesindeki değişken sayısını genetik algoritma ile 6 değişkene indirgemişlerdir. Bu değişkenler; göğüs ağrısı tipi, istirahat kan basıncı, egzersizle tetiklenen anjina, ST depresyon, floroskopide boyanan damar sayısı ve ulaşılan maksimum kalp hızıdır. Mukherjee vd. (2017) çalışmasında ulaşılan maksimum kalp hızı, floroskopide boyanan damar sayısı, ST segment eğimi, göğüs ağrı tipi ve talyum tarama testi sonuçlarının en önemli faktörler olduğu saptanmıştır. Ayrıca Ahmadi vd. (2017) yaptıkları çalışmada da duyarlılık analizi sonuçlarına göre, kalp hastalığının tahmin edilmesinde, en fazla katkıyı yapan değişkenlerin floroskopide boyanan damar sayısı ve talyum sintigrafi sonucu olduğu belirlenmiştir. Bu çalışmada, Rastgele Orman algoritması ile belirlenen değişken önem dereceleri literatürle benzerlik göstermektedir.

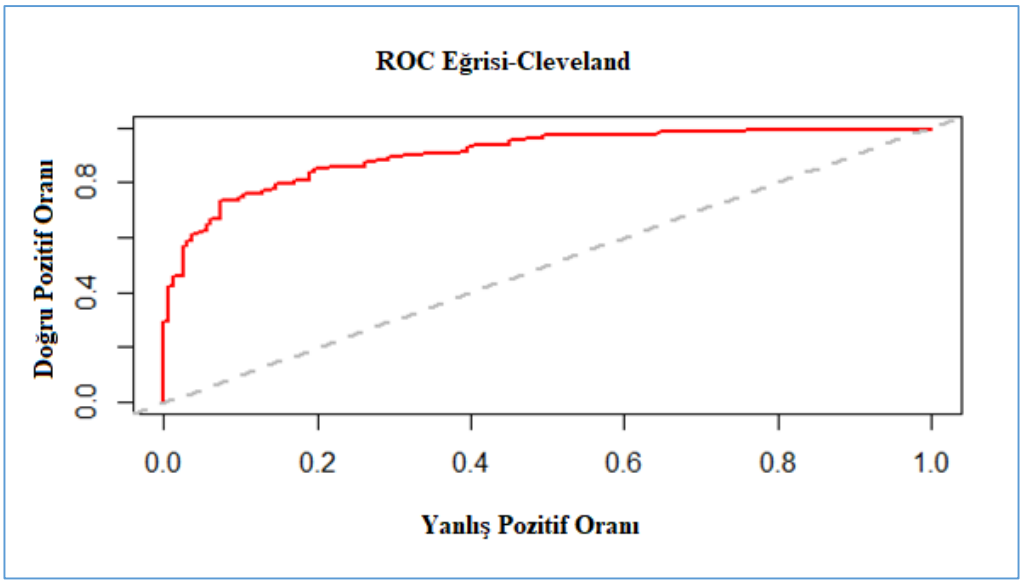
Çizelge 3.8. Cleveland veri kümesi karışıklık matrisi

		Tahmin Edilen Değerler	
Gerçek Değerler		Hastalık yok	Hastalık var
	Hastalık yok	147	17
	Hastalık var	33	106



Şekil 3.41. Cleveland veri kümesi için Gini indeksi

Şekil 3.42.' de Cleveland veri kümesi üzerinde uygulanan rastgele orman algoritması sonucunda elde edilen doğru pozitif ve yanlış pozitif oranları arasındaki ilişkiyi gösteren ROC eğrisi verilmiştir. ROC eğrisi analizinde eğri altında kalan alanın **%90,40** olduğu saptanmıştır. ROC eğrisi altında kalan alan, algoritmanın koroner kalp hastalığı olan bireylerle sağlıklı bireyleri ayırmadaki başarısını göstermektedir. Bu sonuç oluşturulan modelin çok iyi bir sınıflama performansına sahip olduğunu göstermektedir.

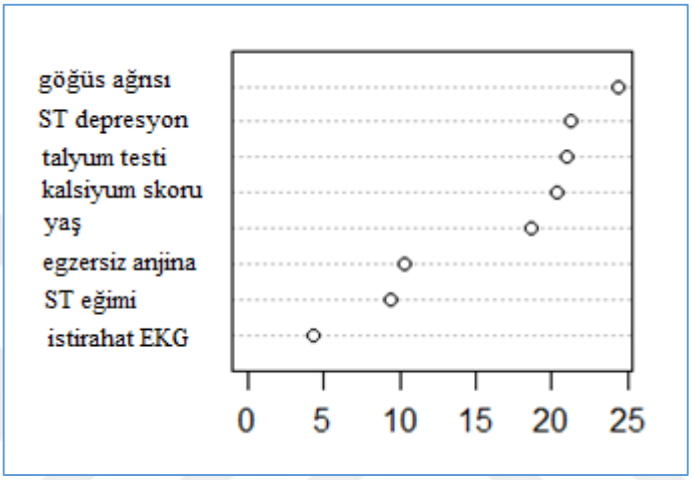


Şekil 3.42. Cleveland Veri Kümesi ROC Eğrisi

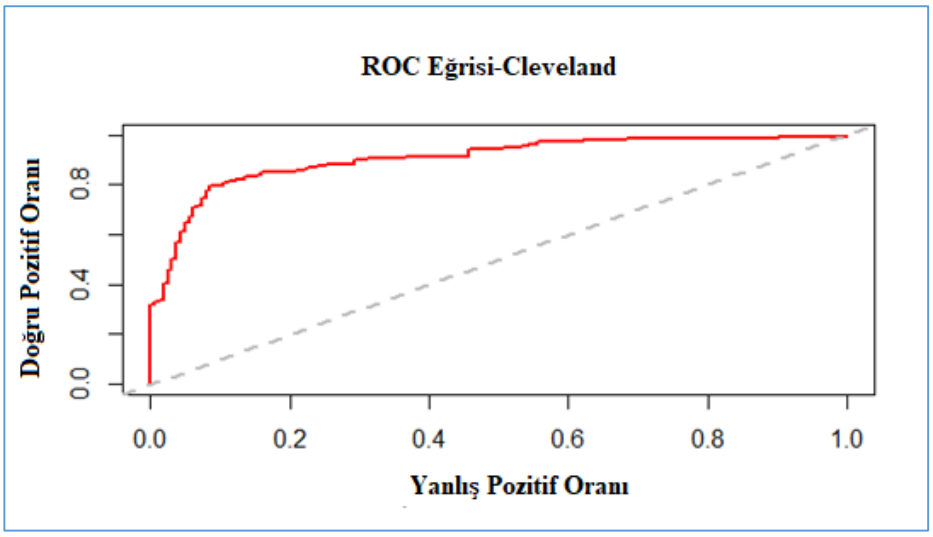
Cleveland veri kümesinde, ön analiz sonuçlarına dayanarak belirlenen değişkenler üzerinde uygulanan Rastgele Orman algoritması sonuçları Çizelge 3.9.'da karışıklık matrisi olarak verilmiştir. Karışıklık matrrisi bileşenlerinden modelin **doğruluk** oranı **%86,13**, **duyarlılığı** %**79,85** ve **seçiciliği** %**91,46** olarak hesaplanmıştır. Sınıflama modelinin oluşturulmasında grafiksel, istatistiksel analizler ve uzman görüşüne dayanan ön analiz sonuçlarının göz önünde bulundurulmasının karışıklık matrisi bileşenlerinde, dolayısıyla bu bileşenlerden hesaplanan tüm parametrelerde iyileşme sağladığı saptanmıştır. Bununla birlikte, Şekil 3.43.'de Rastgele Orman algoritması Gini indeksine göre değişkenlerin önem sırası görülmektedir. Her iki sınıflama modelinde de değişkenlerin önem sırasının değişmediği görülmektedir. Şekil 3.44.' de, ön analiz sonuçları dikkate alınarak oluşturulan sınıflama modelinin ROC eğrisi verilmiştir. ROC eğrisi analizinde eğri altında kalan alanın **%90,44** olduğu saptanmıştır.

Çizelge 3.9. Analiz sonuçlarına göre Cleveland veri kümesi karışıklık matrisi

		Tahmin Edilen Değerler	
Gerçek Değerler		Hastalık yok	Hastalık var
	Hastalık yok	150	14
	Hastalık var	28	111



Şekil 3.43. Analiz sonuçlarına göre Cleveland veri kümesi için Gini indeksi



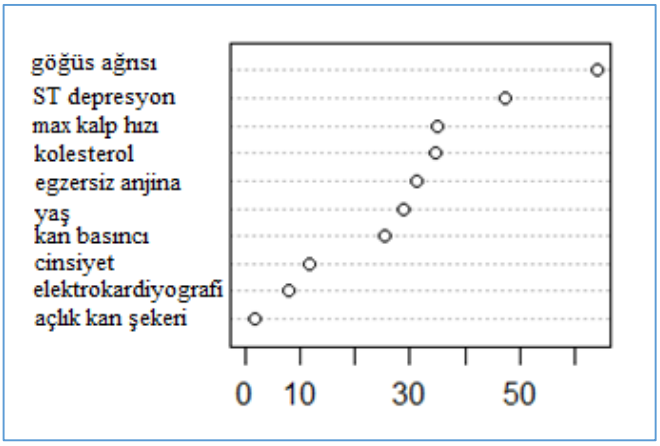
Şekil 3.44. Analiz sonuçlarına göre Cleveland Veri Kümesi ROC Eğrisi

3.4.2. Cleveland ve Macaristan Veri Kümeleri

Macaristan ve Cleveland veri kümelerinin birleştirilmesi ile oluşturulan, 596 hasta kaydı ve 11 değişken içeren veri kümesi üzerinde uygulanan Rastgele Orman algoritması sonuçları Çizelge 3.9.'da karışıklık matrisi olarak verilmiştir. Karışıklık matrisi bileşenlerinden modelin **doğruluk** oranı **%80,20**, **duyarlılığı %72,65** ve **seçiciliği %85,47** olarak hesaplanmıştır. İki veri kümesinin birleştirilmesi ile oluşturulan veri kümesinden elde edilen oranların üç parametre için de azaldığı görülmektedir. Şekil 3.43.'de Rastgele Orman algoritması Gini indeksine göre değişkenlerin önem sırası görülmektedir. Gini indeks sonuçları incelendiğinde ise göğüs ağrısı tipi ve egzersizle tetiklenen ST depresyonu değişkenlerinin diğer iki sınıflama modelinin sonuçlarına benzer biçimde en önemli iki değişen olduğu görülmektedir.

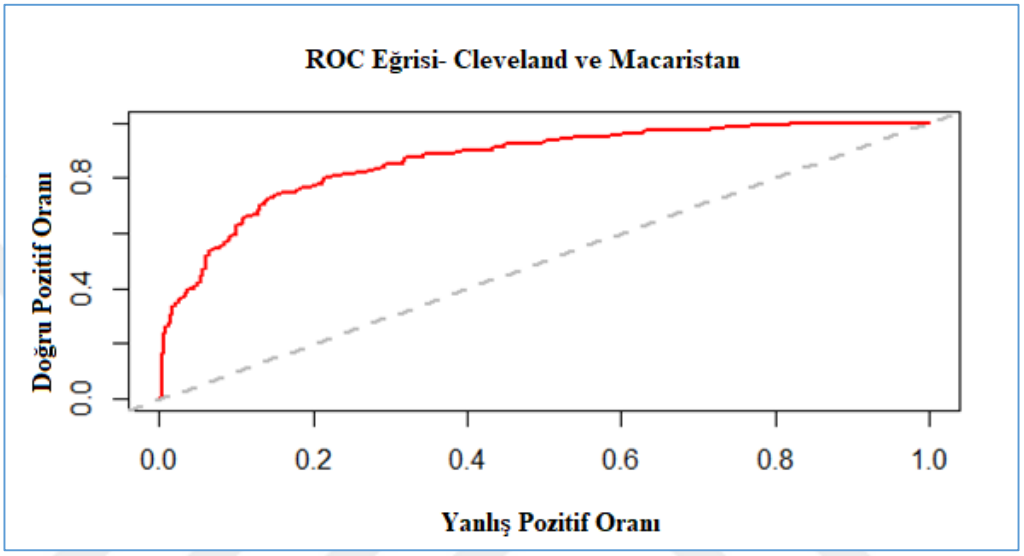
Çizelge 3.10. Cleveland ve Macaristan veri kümeleri karışıklık matrisi

		Tahmin Edilen Değerler	
Gerçek Değerler		Hastalık yok	Hastalık var
	Hastalık yok	300	51
	Hastalık var	67	178



Şekil 3.45. Cleveland ve Macaristan veri kümesi için Gini indeksi

Şekil 3.44.' de Macaristan ve Cleveland veri kümeslerinin birleştirilmesi ile oluşan veri kümesi üzerinde uygulanan Rastgele Orman algoritması sonucunda elde edilen ROC eğrisi verilmiştir. ROC eğrisi analizinde eğri altında kalan alanın **%88,26** olduğu saptanmıştır. Diğer parametrelere benzer biçimde birleştirilmiş veri kümesi için ROC eğrisi altında kalan alanın oranında da azalma saptanmıştır.



Şekil 3.46. Cleveland ve Macaristan veri kümesi ROC eğrisi

Cleveland ve Macaristan veri kümelerinin birleştirilmesi ile oluşturulan yeni veri kümesinde, floroskopide boyanan damar sayısını gösteren *ca* ve talyum sintigrafi sonucunu gösteren *thal* değişkenleri gibi sınıflama açısından kritik iki değişken, kayıp veriler içermeleri nedeniyle, bulunmamaktadır. Ancak, bu veri kümesinde, iki veri kümesi birleştirilerek, hasta sayısı artırılmıştır. Bu nedenle, veri kümesi çok sayıda veriden öğrenebilmektedir. İki kritik değişken olmamasına rağmen, algoritma performansında azalma oranının fazla olmaması bu durumla açıklanabilir. Bununla birlikte, sınıflama sonucunda elde edilen verilerin daha kolay yorumlanabilmesi, anlaşılabilmesi ve sonuçların güvenirliği açısından tüm değişkenler yerine, model açısından önemli olduğu ön analiz ya da yöntemlerle belirlenen az sayıda değişkenle çalışmak önem taşımaktadır.

4. SONUÇLAR ve ÖNERİLER

Bu çalışmada, makine öğrenmesi algoritmalarından Rastgele Orman algoritması kullanılarak koroner arter hastalığı riski analiz edilmiştir. Çalışmada, UCI veri kümesi koleksiyonundan alınan Cleveland, Macaristan, İsviçre ve VA Long Beach kalp hastalığı veri kümeleri kullanılmıştır. Veri kümelerinin makine öğrenmesi yaklaşımı ile analiz edilmesinde CRISP-DM süreç modelinin adımları izlenmiştir.

Çalışmada öncelikle, veri kümelerindeki eksik veri oranları çıkarılmış ve veri kümeleri üzerinde yapılacak analizlerin ve kurulacak modelin daha sağlıklı olabilmesi açısından %60 ve ya daha fazla oranda eksik veri içeren değişkenler veri kümelerinden çıkarılmıştır. Eksik verileri tamamlamada seçilecek yönteme karar verebilmek için Cleveland veri kümesi üzerinde yapay ve rastsal olarak %10, %20 ve %40 oranlarında kayıp veri oluşturulmuş ve bu veri kümeleri üzerinde klasik yöntem (ortanca-mod), Rastgele Orman ve k-En Yakın Komşuluk algoritmaları uygulanmıştır. Her bir yöntemle tamamlanan veri kümesi orijinal veri kümesi ile karşılaştırılarak hata parametreleri hesaplanmıştır. Hata parametre sonuçlarına göre sayısal veriler için ortanca, kategorik veriler için mod değerinin kullanılmasına karar verilmiştir.

Veri analizinde grafiksel ve istatistiksel yöntemler kullanılmıştır. Sayısal değişkenlerin grafiksel analizinde kutu grafiği, histogram dağılımı ve saçılım grafikleri kullanılmıştır. Kategorik değişkenlerin analizinde ise çubuk grafikleri kullanılmıştır. Grafiksel yöntemlerle birlikte Shapiro-Wilk normallik testi ve sayısal değişkenler arasındaki ilişkiyi incelemek amacıyla Spearman Korelasyon analizi gibi istatistiksel yöntemler de uygulanmıştır. Veriler, medikal literatür incelemesi ve kardiyoloji alanında uzman olan bir hekimin görüşleri doğrultusunda detaylı bir biçimde analiz edilmiştir.

Verilerin analiz edilmesinden sonra kayıp veri oranının en az olduğu ve dengeli bir dağılıma sahip olan Cleveland veri kümesi üzerinde sınıflama modeli kurulmuştur. Ayrıca, kayıp veri oranı İsviçre ve VA Long Beach veri kümelerine göre daha az olan ve dengeli bir dağılım gösteren Macaristan veri kümesi ile Cleveland veri kümesinin

birleştirilmesi ile elde edilen 596 hasta kaydı ve 11 değişkenden oluşan veri kümesi üzerinde de bir sınıflama modeli oluşturulmuştur. Her iki veri kümesinden elde edilen sınıflama performansları karşılaştırılmıştır.

Cleveland veri kümesi üzerinde uygulanan Rastgele Orman sınıflama algoritmasının doğruluk oranı %83,49, duyarlılığı %76,25, seçiciliği ise %89,63 olarak saptanmıştır. Ayrıca yapılan ROC eğrisi analizinde eğri altında kalan alanın %90,40 olduğu belirlenmiştir. Ayrıca, Cleveland veri kümesi üzerinde yapılan ön analiz sonuçlarının dikkate alınarak belirlendiği değişkenlerle uygulanan Rastgele Orman sınıflama algoritmasının doğruluk oranı %86,13, duyarlılığı %79,85, seçiciliği ise %91,46 olarak saptanmıştır. ROC eğrisi analizinde ise eğri altında kalan alanın %90,44 olduğu belirlenmiştir. Bu sonuçlar doğrultusunda, veri kümesi üzerinde grafiksel ve istatistiksel yöntemlerle yapılan ön analizlerin, oluşturulacak sınıflama modelinin performansını önemli ölçüde geliştirdiği saptanmıştır. Bu iki sınıflama modeline ek olarak, Macaristan ve Cleveland veri kümelerinin birleştirilmesi ile oluşan veri kümesi üzerinde uygulanan Rastgele Orman algoritmasının doğruluk oranının %80,20 duyarlılığının %72,65 ve seçiciliğinin ise %85,47 olduğu belirlenmiştir. ROC eğrisi analizinde ise eğri altında kalan alanın % 88,26 olduğu saptanmıştır. Ayrıca, her üç sınıflama modelinde de göğüs ağrısı tipi ve egzersizle tetiklenen ST depresyonu değişkenlerinin Gini indeksine göre en önemli iki değişen olduğu belirlenmiştir.

Koroner arter hastalığının kesin tanısında ve hastalık seyrinin izlenmesinde sıklıkla girişimsel bir yöntem olan anjiyografi işlemi altın standart olarak kullanılmaktadır. Ancak, anjiyografi işlemi girişimsel bir tanı işemi olması sebebiyle ciddi klinik komplikasyonlara yol açabilen, maliyeti yüksek ve ileri seviyede teknik uzmanlık isteyen bir işlemdir. Etkin hastalık yönetiminde temel amaç tanı, tedavi ve klinik izlemde yalnızca gerekli klinik işlemleri kullanarak hasta güvenliği artırmak ve sağlık bakım maliyetini düşürmektir. Bu düşünceden hareketle, bu çalışmanın, koroner kalp hastalığı açısından risk taşıyan hasta grubunun tespit edilerek, anjiografi gibi girişimsel işlem uygulanacak hasta grubuna karar vermede sağlık çalışanlarına rehberlik edeceği düşünülmektedir. Ayrıca, bu model, sağlık çalışanlarının manuel olarak yapmalarının oldukça güç olabileceği çok sayıdaki hasta verisinin semptom odaklı ve bilgisayar tabanlı bir sistem tarafından analiz ederek, sağlık çalışanları için

bir klinik karar destek sistemi oluşturacaktır. Bununla birlikte, makine öğrenmesi yaklaşımını kullanan modellerin klinik kullanımı ile birlikte girişimsel tanı işlemi uygulanması gereken hasta sayısının azaltılması sonucunda, medikal hatalar, klinik bakım maliyeti ve teknik insan gücü gereksinimi azaltılırken, hasta güvenliği ve klinik karar kalitesinin önemli ölçüde artırılacağı düşünülmektedir.

Gelecek çalışmalarda, makine öğrenmesi algoritmaları ile birlikte metin madenciliği gibi yöntemler kullanılarak, sağlık bakım sistemi veritabanları içerisinde oluşan çok büyük miktardaki verinin değerlendirilmesi, risk gruplarının saptanması ve bu bilgilerden elde edilecek bilgi ve örüntülerin hastalık yönetiminde kullanılması sağlanabilir. Bununla birlikte, gelecek çalışmalarda kümeleme algoritmaları ile benzer klinik özelliklere sahip hastaların profilinin çıkarılarak analiz ve sınıflamaların yapılmasına gereksinim duyulmaktadır. Ayrıca ileride yapılacak çalışmalarda sağlık alanında yapılan çalışmalar açısından çok büyük öneme sahip TP ve FN oranlarını birlikte optimize edecek algoritmalar geliştirilebilir.

KAYNAKLAR

Abdullah, A. S., A data mining model to predict and analyze the events related to coronary heart disease using decision trees with particle swarm optimization for feature selection. International Journal of Computer Applications. 55 (8), 2012.

Abdullah, A. S., Rajalaxmi, R., A data mining model for predicting the coronary heart disease using random forest classifier. In International Conference in Recent Trends in Computational Methods, Communication and Controls, April 2012, 22-25, 2012.

Adalet, K., Klinik Kardiyoloji Tanı ve Tedavi. İstanbul Medikal Yayıncılık, 1. Baskı, İstanbul, 2013.

Ahmadi, E., Weckman, G. R., Masel, D. T., Decision making model to predict presence of coronary artery disease using neural network and C5. 0 decision tree. Journal of Ambient Intelligence and Humanized Computing. 1-13, 2017.

Akman, M., Genç, Y., Ankaralı, H., Random forests yöntemi ve sağlık alanında bir uygulama. Turkiye Klinikleri Journal Of Biostatistics. 3 (1), 36-48, 2011.

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Sani, Z. A., A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine. 111 (1), 52-61, 2013.

Alizadehsani, R., Hosseini, M. J., Sani, Z. A., Ghandeharioun, A., Boghrati, R., Diagnosis of coronary artery disease using cost-sensitive algorithms. In Data Mining Workshops (ICDMW), IEEE 12th International Conference on, December 2012, Brussels-Belgium, s. 9-16, 2012.

Alpaydin, E., Introduction to machine learning. MIT press, Cambridge, Massachusetts London, England, 2010.

Anbarasi, M., Anupriya, E., Iyengar, N. C. S. N., Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology. 2 (10), 5370-5376, 2010.

Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., Yarifard, A. A., Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer Methods And Programs İn Biomedicine. 141, 19-26, 2017.

Avşar, A., Önder, Akçı., Beyter, M. E., Aterosklerozun patogenezi (aterogenez). Turkiye Klinikleri Journal of Cardiology Special Topics. 4 (2), 1-15, 2011.

Barter, P., The role of HDL-cholesterol in preventing atherosclerotic disease. European Heart Journal Supplements. 7 (suppl_F), F4-F8, 2005.

Bonow, R. O., Mann, D. L., Zipes, D. P., Libby, P., Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine. Elsevier Health Sciences, Philadelphia, 2015.

Breiman, L., Random forests. Machine learning. 45 (1), 5-32, 2001.

Breiman, L., Cutler, A., RFtools—for predicting and understanding data. In Interface'04 Workshop, 2004.

Burke, A. P., Farb, A., Malcom, G. T., Liang, Y. H., Smialek, J., Virmani, R., Coronary risk factors and plaque morphology in men with coronary disease who died suddenly. New England Journal of Medicine. 336 (18), 1276-1282, 1997.

Chaurasia, V., Early prediction of heart diseases using data mining. Caribbean Journal of Science and Technology. 1, 208–217, 2013.

Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., Lin, E. J., HDPS: Heart Disease Prediction System. In Computing in Cardiology, IEEE, 2011 September, s. 557-560, 2011.

Cihan, Ş., Karabulut, B., Arslan, G., Cihan, G., Koroner arter hastalığı riskinin veri madenciliği yöntemleri ile incelenmesi. Uluslararası Mühendislik Araştırma Ve Geliştirme Dergisi. 10 (1), 85-93, 2018.

Coutinho, M., Gerstein, H. C., Wang, Y., Yusuf, S., The relationship between glucose and incident cardiovascular events. A metaregression analysis of published data from 20 studies of 95,783 individuals followed for 12.4 years. Diabetes Care. 22 (2), 233-240, 1999.

Çınar, H. ve Arslan, G., Veri madenciliği ve CRISP-DM yaklaşımı. XVII. İstatistik Araştırma Sempozyumu, 2008 Ankara, s. 304-314, 2008.

Davies, M. J., Ho, S. Y., Atlas of coronary artery disease. Lippincott Williams & Wilkins, Philadelphia, 1998.

Deo, R. C. , Machine learning in medicine. Circulation, 132 (20), 1920-1930, 2015.

Du, K. L., Swamy, M. N., Neural networks and statistical learning. Springer Science & Business Media. 2013.

El-Bialy, R., Salamay, M. A., Karam, O. H., Khalifa, M. E., Feature analysis of coronary artery heart disease data sets. Procedia Computer Science. 65, 459-468, 2015.

Englund, C., Verikas, A., A novel approach to estimate proximity in a random forest: An exploratory study. Expert Systems With Applications. 39 (17), 13046-13050, 2012.

Erkuş, E., Kaya, Z., Yıldız, A., kardiyovasküler risk değerlendirmesi. Turkiye Klinikleri Journal Of Cardiology Special Topics. 6 (4), 1-8, 2013.

Foster, K. R., Koprowski, R., Skufca, J. D., Machine learning, medical diagnosis, and biomedical engineering research-commentary. Biomedical Engineering Online. 13 (1), 94, 2014.

Gui, C., Chan, V., Machine learning in medicine. University of Western Ontario Medical Journal. 86 (2), 76-78, 2017.

Han, J., Pei, J., Kamber, M. Data Mining: Concepts And Techniques. Elsevier, 2011.

Hansson, G.K., Nilsson, J., Pathogenesis of Atherosclerosis. 3-15. Ed: by Crawford, M.H., Di Marco, J.P., Paulus, W.J. Philadelphia. Mosby Elsevier, 2010.

Harrison D. G., Endothelial function and oxidant stress. Clin Cardiol. 20, 11-17, 1997.

Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N., Stokes, J., Factors of risk in the development of coronary heart disease six year follow up experience: the Framingham Study. Annals Of Internal Medicine. 55 (1), 33-50, 2017.

Kawakubo, H., Yoshida, H., Rapid feature selection based on random forests for high-dimensional data. In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing, January 2012, Las Vegas-USA, s.1, 2012.

Kılıç, S., Klinik karar vermede ROC analizi. Journal Of Mood Disorders. 3 (3), 135-40, 2013.

Kim, J., Lee, J., Lee, Y., Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Healthcare İnformatics Research. 21 (3), 167-174, 2015.

Kinge, D.,, Gaikwad, S. K., Survey on data mining techniques for disease prediction. International Research Journal of Engineering and Technology. 5 (1), 630-636, 2018.

Kononenko, I., Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence In Medicine. 23 (1), 89-109, 2001.

Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., Fettich, J., Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artificial Intelligence In Medicine. 16 (1), 25-50, 1999.

Lantz, B. Machine Learning With R. Packt Publishing Ltd, Birmingham, 2013.

Laslett, L. J., Alagona Jr, P., Clark III, B. A., Drozda Jr, J. P., Saldivar, F., Wilson, S. R., …., Hart, M. The worldwide environment of cardiovascular disease: prevalence, diagnosis, therapy, and policy issues: a report from the American College of Cardiology. Journal of the American College of Cardiology. 60 (25), s.1-49, 2012.

Liaw, A., Wiener, M., Classification and regression by random Forest. R News. 2 (3), 18-22, 2002.

Libby, P., Theroux, P., Pathophysiology of coronary artery disease. Circulation, 111 (25), 3481-3488, 2005.

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., Wang, Q., A hybrid classification system for heart disease diagnosis based on the rfrs method. Computational And Mathematical Methods In Medicine, 2017.

Lloyd-Jones, D., Adams, R. J., Brown, T. M., Carnethon, M., Dai, S., De Simone, G., .., Go, A., Heart disease and stroke statistics 2010 update. Circulation. 121 (7), 46-215, 2010.

Malinowski, P., Milewski, R., Ziniewicz, P., Milewska, A. J., Czerniecki, J., Więsak, T., ..., Wołczyński, S., Classification of patients treated for ınfertility using the IVF method. Studies In Logic, Grammar And Rhetoric. 43 (1), 49-59, 2015.

Marbán, Ó., Mariscal, G., Segovia, J., A Data Mining & Knowledge Discovery Process Model. I-Tech, Austria, 2009.

Marikani, T., Shyamala, K., Prediction of heart disease using supervised learning algorithms. International Journal of Computer Applications. 165 (5), 2017.

Masethe, H. D., Masethe, M. A., Prediction of heart disease using classification algorithms. In Proceedings Of The World Congress On Engineering And Computer Science, October 2014, San Francisco-USA, s. 22-24, 2014.

Mendis, S., Puska, P., Norrving, B., World Health Organization., Global atlas on cardiovascular disease prevention and control. Geneva: World Health Organization, 2011.

Michalski, R. S., Carbonell, J. G., Mitchell, T. M., Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

Mitchell, T. M., Machine Learning, McGraw-Hill Science/Engineering/Math, New York, 1997.

Mukherjee, S., Kapoor, S., Banerjee, P., Diagnosis and ıdentification of risk factors for heart disease patients using generalized additive model and data mining techniques. Journal of Cardiovascular Disease Research. 8 (4), 2017.

Muthukaruppan, S., Er, M. J., A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Systems with Applications. 39 (14), 11657-11665, 2012.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P., Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications. 40 (4), 1086-1093, 2013.

National Cholesterol Education Program (NCEP)., The third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III) final report. Circulation. 106 (25), 2002.

Nazari, S. , Fallah, M., Kazemipoor, H, Salehipour,A., A fuzzy inference- fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases. Expert Systems With Applications. 95, 261–271, 2018.

Netter FH. The Netter collection of medical illustrations. In Yonkman FF Volume 5 Heart New York, ICON learning systems, 16-17, 2001.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I., Ishii, S., A bayesian missing value estimation method for gene expression profile data. Bioinformatics. 19 (16), 2088-2096, 2003.

Onat, A., TEKHARF 2017: Tıp Dünyasının Kronik Hastalıklara Yaklaşımına Öncülük. Logos Yayıncılık, İstanbul, 2017.

Ökçün, B., Gürmen, T., Koroner anjiyografi komplikasyonları ve tedavisi. Turkiye Klinikleri Journal of Internal Medical Sciences. 3 (42), 48-72, 2007.

Pandey, A. K., Pandey, P., Jaiswal, K. L., A heart disease prediction model using decision tree. IUP Journal of Computer Sciences. 7 (3), 43, 2013.

Patil R Priya, Kinariwala A S., Automated diagnosis of heart disease using data mining techniques. International Journal of Advance Research, Ideas and Innovations in Technology. 3 (2), 560-567, 2017.

Portugal, I., Alencar, P., Cowan, D., The use of machine learning algorithms in recommender systems: a systematic review. Expert Systems with Applications. 2017.

Prakash, S., Sangeetha, K., Ramkumar, N., An optimal criterion feature selection method for prediction and effective analysis of heart disease. Cluster Computing. 1-7, 2018.

Schlemmer, A., Zwirnmann, H., Zabel, M., Parlitz, U., & Luther, S., Evaluation of machine learning methods for the long-term prediction of cardiac diseases. In Cardiovascular Oscillations (ESGCO), 2014 8th Conference of the European Study Group on, Mayıs 2014, Trento-Italy, s. 157-158, 2014.

Schmitt, P., Mandel, J., Guedj, M., A comparison of six methods for missing data imputation. Journal of Biometrics & Biostatistics. 6 (1), 1, 2015.

Scrutinio, D., Bellotto, F., Lagioia, R., Passantino, A., Physical activity for coronary heart disease: cardioprotective mechanisms and effects on prognosis. Monaldi Archives For Chest Disease. 64 (2), 2005.

Shafey, O., Eriksen, M., Ross, H., Mackay, J., The tobacco atlas. Atlanta: American Cancer Society. 3, 38-39, 2009.

Shafique, U., Majeed, F., Qaiser, H., Mustafa, I. U., Data Mining in Healthcare for Heart Diseases. International Journal of Innovation and Applied Studies. 10 (4), 1312, 2015.

Shamsollahi, M., Badiee A., Ghazanfari, M., Using combined descriptive and predictive methods of data mining for coronary artery disease prediction: a case study approach. Journal of AI and Data Mining. 2018.

Sharan, M. L., Sathees, K. B., Analysis of cardiovascular heart disease prediction using data mining techniques. International Journal of Modern Computer Science (IJMCS). 4 (1), 55-58, 2016.

Sharma, M., Khan F., Ravichandran, V., Comparing data mining techniques used for heart disease prediction. International Research Journal of Engineering and Technology. 4 (6), 56-72, 2017.

Smith Jr, S. C., Collins, A., Ferrari, R., Holmes Jr, D. R., Logstrup, S., ... ,Taubert, K., Our time: a call to save preventable death from cardiovascular disease. European Heart journal. 33 (23), 2910-2916, 2012.

Soni, J., Ansari, U., Sharma, D., Soni, S., Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications. 17 (8), 43-48, 2011.

Stamler, J., Wentworth, D., & Neaton, J. D., Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded?: findings in 356 222 primary screenees of the multiple risk factor intervention trial (mrfit). Jama. 256 (20), 2823-2828, 1986.

Stary, H. C., Chandler, A. B., Dinsmore, R. E., Fuster, V., Glagov, S., Insull, W., ...,Wissler, R. W., A definition of advanced types of atherosclerotic lesions and a histological classification of atherosclerosis: A report from the committee on vascular lesions of the council on arteriosclerosis, American Heart Association. Circulation. 92 (5), 1355-1374, 1995.

Stemme, S., Faber, B., Holm, J., Wiklund, O., Witztum, J. L., Hansson, G. K., T lymphocytes from human atherosclerotic plaques recognize oxidized low

density lipoprotein. Proceedings of the National Academy of Sciences. 92 (9), 3893-3897, 1995.

Takcı, H., Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences. 26 (1), 1-10, 2018.

Tantimongcolwat, T., Naenna, T., Isarankura-Na-Ayudhya, C., Embrechts, M. J., Prachayasittikul, V., Identification of ischemic heart disease via machine learning analysis on magnetocardiograms. Computers In Biology And Medicine. 38 (7), 817-825, 2008.

Tokgözoğlu, L., Dislipidemi, ateroskleroz ve hassas plaklar: Atorvastatinin ateroskleroz ve plak yapısına etkisi. Türk Kardiyol Dern Arş-Arch Turk Soc Cardiol. 37, 11-16, 2009.

Uyar, K., İlhan, A., Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia Computer Science. 120, 588-593, 2017.

Verma, L., Srivastava, S., Negi, P. C., A hybrid data mining model to predict coronary artery disease cases using non-ınvasive clinical data. Journal of Medical Systems. 40 (7), 1-7, 2016.

Williams, R. R., Hopkins, P. N., Wu, L. L., Schumacher, C., Hunt, S. C., Evaluating family history to prevent early coronary heart disease. Primer in preventive cardiology. Dallas: American Heart Association. 93, 1994.

Wirth, R., Hipp, J., CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, s.29-39, 2000.

Wong, N. D., Epidemiological studies of chd and the evolution of preventive cardiology. nature reviews. Cardiology. 11 (5), 276, 2014.

Yang, P., Hwa Yang, Y., B Zhou, B., Y Zomaya, A., A review of ensemble methods in bioinformatics. Current Bioinformatics. 5 (4), 296-308, 2010.

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., ..., Lisheng, L., Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. The Lancet. 364 (9438), 937-952, 2004.