**BEYAN**

Bu bitirme tezinin hazırlanmasında bilimsel ahlak kurallarına uyulduğunu, başkalarının çalışmalarından yarar sağlanması durumunda bilimsel kurallara uygun atıfta bulunulduğunu, yararlanılan bilgilerden herhangi bir tahrifat yapılmadığını, bitirme tezinin herhangi bir bölümünün bu üniversitede yahut başka bir üniversitede bitirme tezi olarak sunulmadığını beyan ederim.

Selçuk AKARIN

26.05.2019

**ÖNSÖZ**

Modern dünyanın getirileri arasında olan devasa verisetleri, günümüz dünyasınında en popüler çalışma disiplinlerinden biri olan veri madenciliğini ortaya çıkartmıştır. Veri madenciliği en basit tabiriyle verilerden işe yarar çıkarımlar elde etmektir. Özellikle reklam ve tıp alanında hayati öneme sahip olan verilerden çıkarım yapma sanatı birçok devlet ve şirketin ilgi odağı olmuştur. Reklamcılık sektöründe şirketlerin satış stratejilerini belirleyen bu disiplin tıp dünyasında ise hastalıklara karşı mücadele stratejilerini belirlemektedir. Bu alanda yapılan çalışmalar belki de yakın gelecekte birçok hastalığın antibiyotiği olacaktır.

Bu bitirme tezinin yazımı aşamasında çalışmalarımı özenle inceleyen değerli hocam. Dr. Muhammet Baykara’ya, lisans hayatım süresince derslerinde bulunduğum bütün saygıdeğer hocalarıma ve projeyi gerçekleştirme aşamasında benden yardımlarını esirgemeyen herkese bir teşekkürü borç bilirim.

Selçuk AKARIN

26.05.2019

İÇİNDEKİLER

[ÖZET](#_Toc354949698) III

[GİRİŞ 1](#_Toc354949699)

[BÖLÜM 1: VERİ MADENCİLİĞİ TARİHÇESİ VE LİTERATÜR TARAMASI](#_Toc354949700)

[1.1. Veri Madenciliği Tarihçesi 2](#_Toc354949701)

[1.2. Literatür Taraması 3](#_Toc354949702)

[BÖLÜM 2. GEREKLİ KURULUM VE KONFİGÜRASYONLAR, VERİ SETİNE GENEL BAKIŞ VE AMAÇ](#_Toc354949730)

[2.1. Gerekli Kurulumlar ve Konfigürasyonlar](#_Toc354949731) 4

[2.1.1. İndirmeler ve Kurulumlar](#_Toc354949732) 4

[2.1.2. R Studio’ya Genel Bakış ve Konfigürasyonlar](#_Toc354949733) 6

[2.2. Veri Setinin incelenmesi](#_Toc354949737) 9

[2.2.1. Veri Setinin Genel Tanıtımı](#_Toc354949738) 9

[2.2.2. Veri Setinin R Studio’ya yüklenmesi](#_Toc354949747) 10

[2.2.3. Veri Setinin Modeller İçin Düzenlenmesi](#_Toc354949747) 10

[2.2.4. Veri Seti Üzerinde Keşifçi Veri Analizi](#_Toc354949747) 10

[2.3. Amaç 1](#_Toc354949755)6

[BÖLÜM 3. VERİ SETİ ÜZERİNE MODEL KURMA VE KURULAN MODELLERİN DEĞERLENDİRİLMESİ](#_Toc354949756)

[3.1. Lojistik Regresyon Modeli](#_Toc354949757) 19

[3.1.1. Veri Setinin Test ve Eğitim Kümelerine Ayrılması](#_Toc354949760) 19

[3.1.2. Lojistik Regresyon Modelinin Kurulması](#_Toc354949760) 19

[3.1.3. Lojistik Regresyon Modeli ile Tahmin Yapılması](#_Toc354949760) 20

[3.1.4. Lojistik Regresyon Modeli Tahminlerinin Değerlendirilmesi](#_Toc354949760) 21

[3.1.5. Lojistik Regresyon Modelindeki Hatayı Ölçme](#_Toc354949760) 25

[3.2. KNN Modeli](#_Toc354949758) 27

[3.2.1. Veri Setinin Test ve Eğitim Kümelerine Ayırılması](#_Toc354949760) 29

[3.2.2. KNN Modelinin Kurulması](#_Toc354949760) 29

[3.2.3. KNN Modelinin Optimize Edilmesi ve Tahmin](#_Toc354949760) 30

[3.3. Yapay Sinir Ağı Modeli 3](#_Toc354949759)3

[3.3.1 Yapay Sinir Ağı Modeli İçin Standartlaştırma İşlemi 3](#_Toc354949760)6

[3.3.2 Veri Setinin Test ve Eğitim Kümelerine Ayırılması](#_Toc354949761) 36

[3.3.3 Yapay Sinir Ağı Modelinin Oluşturulması 3](#_Toc354949762)6

[3.3.4 Yapay Sinir Ağı Modeli ile Tahmin](#_Toc354949763) 37

[SONUÇ](#_Toc354949788) 38

[KAYNAKÇA](#_Toc354949789) 39

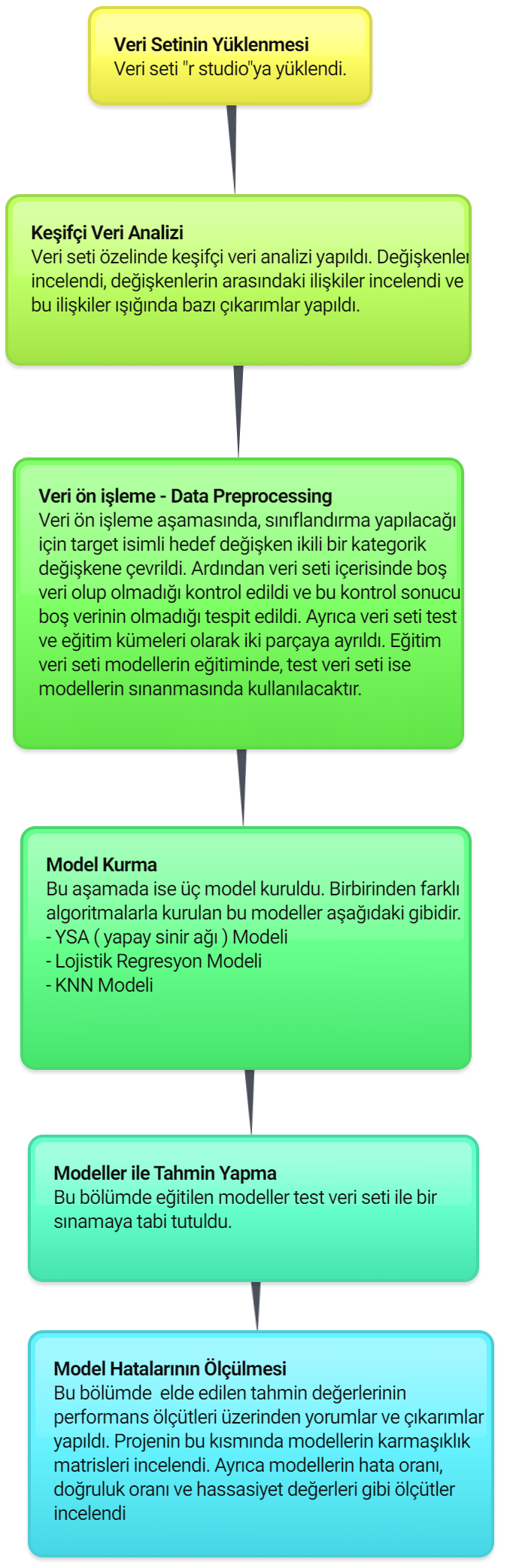
# 

# ÖZET

Önlenemez bir şekilde gelişen teknoloji dünyası sayesinde, her alanda olduğu gibi tıp alanında da teknoloji destekli devrimsel çalışmalar sürmektedir. Özellikle son dönemlerin yükselen yıldızlarından olan veri madenciliği konularında önemli çalışmalar yapılmaktadır. Günümüz dünyasında dijital verilerin toplanması ve işlenmesi için gereken veriler her geçen gün üstel bir şekilde katlanarak çoğalıyor. Hâl böyleyken, birçok devlet ve şirket bu verilerden mantıklı çıkarımlar elde etmek için büyük bir yarışa girmiştir. Bu çıkarımlara hastalıkların erken evrelerde teşhisleri, yapılan operasyonların etkilerinin incelenmesi veya ilaçların yan etkilerinin saptanması gibi çalışma konuları örnek olarak verilebilir.

Hastalıkların erken evrelerde teşhisi birçok hastalık için hayati derecede öneme sahiptir. Özellikle kanser, kalp hastalıkları, damar tıkanıklıkları gibi rahatsızlıkların erken evrelerde tespiti hastaya uygulanacak tedavilerin şeklini değitirip, belki de hayatına mâl olacak bu hastalıktan kurtulmasını dahi sağlayabilecektir. Bu çalışmada; kalp rahatsızlığına neden olabilecek verilerin toplanmasıyla oluşan *“heart\_disease”* adlı veri seti, bazı sınıflandırma algoritmaları ışığında incelenmiştir.

**Yapılan Çalışmanın Kavram Haritası**

****

# GİRİŞ

Günümüzde insanoğlunun etrafını saran büyük veri ağları sayesinde, dev veritabanları oldukça çok çeşitli bilgiyle dolmaktadır. Bu bilgilere birtakım örnekler verilecek ise; kredi veya banka kartı bilgileri, site kayıt bilgileri, alışveriş bilgileri, hava durumu bilgileri, güvenlik için kullanılan parmak izi ve retina bilgileri, kamera kayıtları bunlara örnek olarak sayılabilir. İnsanların her saniye etrafından akan bu verilerin kayıt edildiği yerler devasa veritabanlarıdır. Bu veritabanlarındaki verileri işleyen, sadeleştiren ve çıkarımlar yapan insanlara ise genel tabirle veri madencileri denilmektedir.

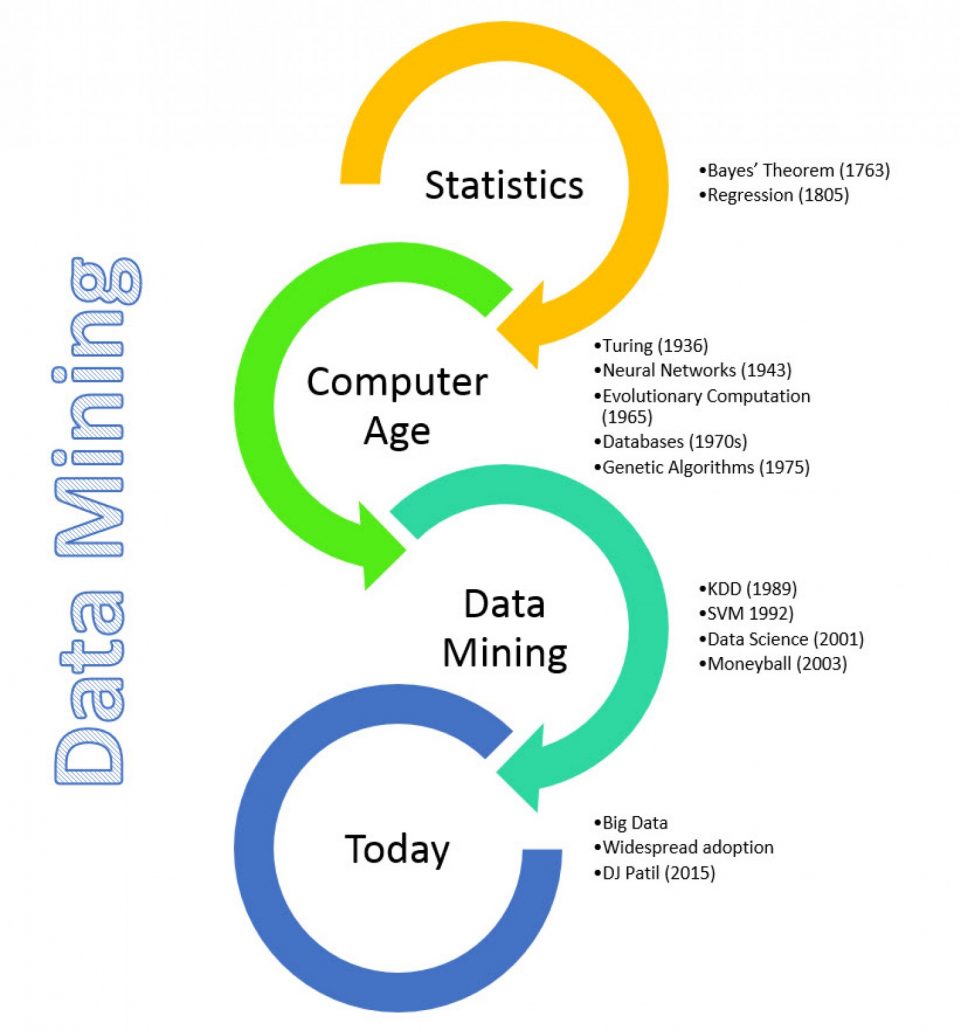
Çağımızın çok önemli istihbari bilgileri de bu verilerin işlenmesiyle ortaya çıkarılmakta, insan hayatının kurtulmasına ön ayak olacak bilgiler de bu veriler sayesinde öğrenilmektedir.. Bunun için verilerin en etkin şekillerde değerlendirilmesi gerekmektedir. Bu açıdan bakarsak veri madenciliği oldukça çok disiplindeki uygulamaları bünyesinde barındırır. Verilerin sağlıklı şekillerde işlenebilmesi için; istatistik, yapay zeka uygulamaları, veri tabanı teknolojileri, makine öğrenmesi ve görselleştirme gibi birbirinden çeşitli alt disiplin uygulamalarının veri madenciliği özelinde kullanmak gerekliliği doğmuştur. Veri madencileri genel olarak buradaki disiplinlere hakim olmalıdır. Hangi aracın, hangi teknik veya teknik kombinasyonun kullanılacağı uygulama özelinde elde edilmesi istenilen çıkarımlarla ilgilidir.

Tezin ilk bölümünde veri madenciliğinin tarihçesi ve literatür taraması yer almaktadır. İkinci bölümünde *“heart\_disease”*  veri setinin tanıtımı, gerekli kurulum ve konfigürasyonlar, bu çalışmanın asıl amacının yer aldığı bir kısım vardır. Üçüncü bölümünde ise *“heart\_disease”*  veri setine uygulanan modeller, bu modeller yardımıyla yapılan çıkarımlar ve sınıflandırmalar, bu sınıflandırmaların doğruluk oranları ve performans ölçütlerinin değerlendirmeleri, modeller ile ilgili düzenleme ve optimum modelin bulunmasına yönelik çalışmalar yer almaktadır. Dördüncü bölümde bu çalışmadan elde edilen sonuçlar ve değerlendirmeler mevcuttur. Son bölümde ise kaynaklar bulunmaktadır.

# BÖLÜM 1: VERİ MADENCİLİĞİ TARİHÇESİ, VERİ MADENCİLİĞİ UYGULAMALARI VE LİTERATÜR TARAMASI

## Veri Madenciliği Tarihçesi

Veri madenciliğinin son dönemlerde popüler bir dal olmasıyla çoğu kişi veri madenciliğinin yakın bir tarihte ortaya çıktığını sanmaktadır. Fakat hakim olan bu düşünce aslında yanlıştır. Veri madenciliği 1700’lerde Bayes Teoremi ve 1800’lerde Regresyon analizi gibi genelde veri içerisindeki anlamlı desenleri analiz eden erken veri madenciliği uygulamalarıyla başlayan oldukça köklü bir tarihe sahiptir. 1936’da ilk elektronik bilgisayar olarak kabul gören turing makinesi icat edilmiştir. Bunun devamında 1950’lerde sayımlar için bilgisayarlar kullanılmıştır. 1960’larda veritabanları ile verilerin depolanması kavramları teknoloji dünyasına kazandırılmıştır. 1960’ların sonlarına doğru basit öğrenmeli bilgisayarlar geliştirilebilmiştir. 1970’lerde ilişkisel veritabanı modelleri kullanılmaya başlanmıştır. 1980’lerde ise veritabanı yönetim sistemleri oldukça yaygın bir hale gelmiş ve bilimsel uygulamalarda ve mühendislik dallarında kullanılmaya başlanmıştır. 1990’larda içlerinde oldukça fazla veri biriken bu veritabanlarından önemli verilerin nasıl çıkarılabileceğiyle ilgili fikirler öne sürülmüştür. En nihayetinde 1992 yılında veri madenciliği ile ilgili ilk yazılım geliştirilmiştir. Bu noktadan sonra günümüze kadar veri madenciliği oldukça fazla alana girmiş ve yaygınlaşmıştır. (Savaş, Topaloğlu ve Yılmaz, 2012)



**Şekil 1. Veri madenciliği tarihine genel bakış**

## 1.2. Literatür Taraması

“Veri Madenciliği Yöntemlerine Genel Bakış” adlı bilimsel makalede veri madenciliğine genel bakış, veri madenciliği kullanım alanları ve veri madenciliği metotlarına kısaca değindikten sonra veri madenciliğinin tıptaki yeri hakkında bilgiler verilmiştir. (KÖKTÜRK, ANKARALI, SÜMBÜLOĞLU 2008)

“Veri Madenciliği Modelleri ve Uygulama Alanları” isimli makalede veri madenciliği içerisinde kullanılan modeller sınıflandırma, kümeleme, regresyon ve birliktelik kuralları çatılarında incelenmiştir. Bu çalışma veri madenciliğinde en önemli aşamalardan olan verinin analizi ve problemin anlaşılması konuları üzerinde durmuş ve öneminden bahsetmiştir. Bu unsurların, veri madenciliğinin başarısını ne derecede etkilediği gözler önüne serilmiştir. Ayrıca bu konular idrak edildikten sonra ilgilenen probleme en uygun teknik veya tekniklerin de seçiminde özenli davranılması konusunda önemli uyarılarda bulunulmuştur.

“Veri Madenciliği ve Türkiye’deki Uygulama Örnekleri” adlı çalışmada genel olarak Türkiye’deki veri madenciliği uygulamaları konu alınmış ve incelenmiştir. Yazarlar bunu yaparken konuyu önce belli başlıklara bölmüş ve buna göre incelemişlerdir. Bu başlıklar şöyle sıralanabilir;

* Eğitim
* Ticaret
* Mühendislik
* Bankacılık ve Borsa
* Tıp
* Telekomünikasyon

Yukarıdaki başlıklar altında Türkiye’de yapılan faaliyetlerden bahsedilmiş ve bu projelere atıflarda bulunulmuştur.Sonuç olarak ise veri madenciliğinin bu kullanım alanlarındaki amaçları ve yöntemlerine değinilmiş, veri madenciliğinin hangi cihetlere doğru yönlendiğinden bahsedilmiştir. (Savaş, Topaloğlu ve Yılmaz, 2012)

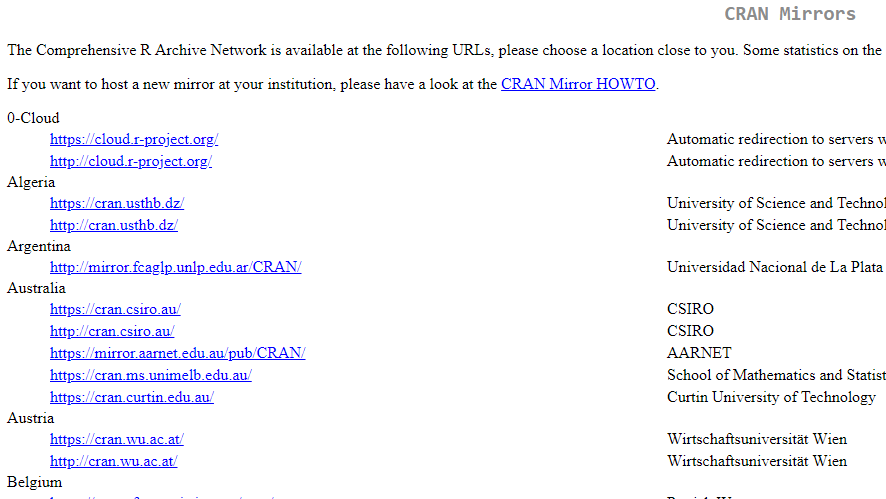
“Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları” isimli makalede ise sağlıkta veri madenciliğinin kullanımına yönelik başta sağlık çalışanları olmak üzere insanlara bir bakış açısı kazandırmak ve insanlara genel itibariyle bir veri madenciliği kavram çerçevesi oluşturmak hedefine yönelinmiştir. Bu çalışmada veri madenciliği ve veri madenciliği yöntemleri, veri ambaları gibi genel konular anlatıldıktan sonra tıp ve sağlık sektöründe veri madenciliği uygulamalarından bahsedilmiştir. Tüm bunlardan sonra makelenin asıl hedefi olan insanlara tıp sektöründe veri madenciliğinin nasıl ve ne şekillerde kullanılabileceğine yönelik konulara değinilmiştir. (KOYUNCUGİL ve ÖZGÜLBAŞ 2009)

# BÖLÜM 2. GEREKLİ KURULUM VE KONFİGÜRASYONLAR, VERİ SETİNE GENEL BAKIŞ VE AMAÇ

## 2.1.Gerekli Kurulumlar ve Konfigürasyonlar

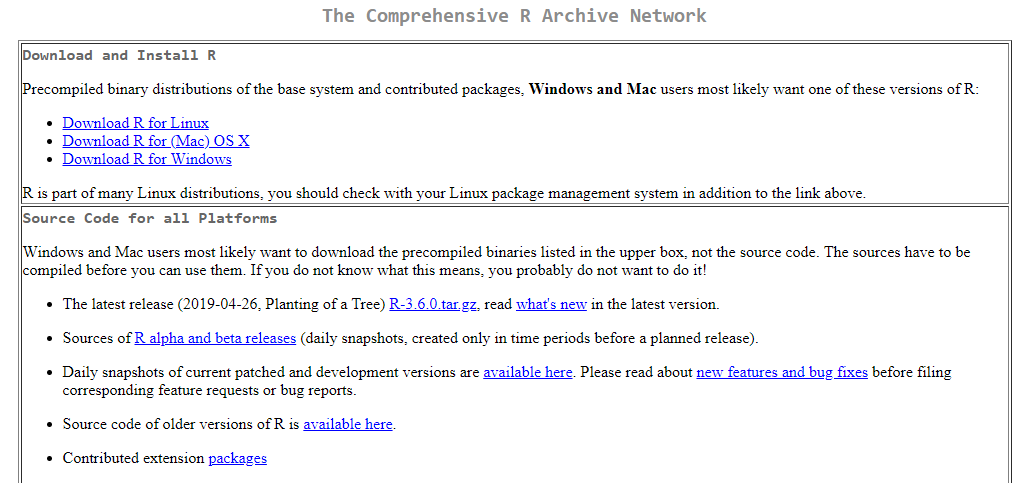
## 2.1.1.İndirmeler ve Kurulumlar

Tezin devamında yazılan kodları çalıştırmak için r project ve r studio kurulumlarının yapılması gereklidir. Öncelikle r project’i indirmek için *“*<https://cran.r-project.org/mirrors.html>*”* adresinden herhangi bir sunucu seçilip indirme işlemi başlatılabilir.

**

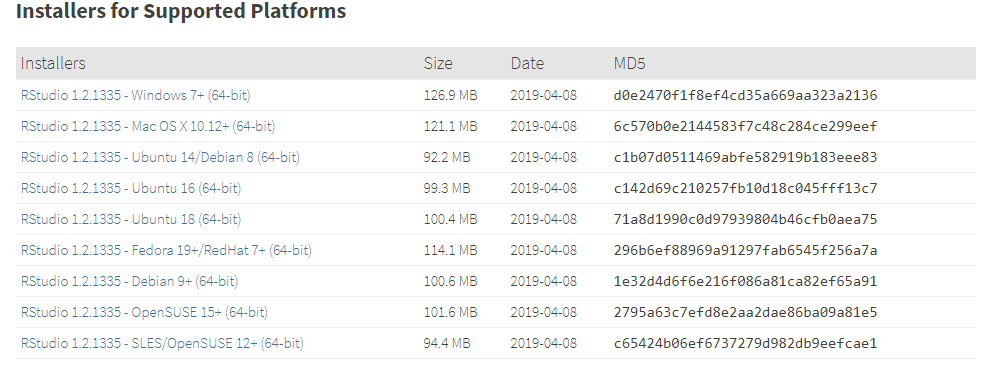
**Şekil 2. R project’in yansılarının tutulduğu sunucular**

Bir sunucu seçildikten sonra aşağıdaki ekranda işletim sistemine uygun bir sürüm seçilip indirilir.



Ş**ekil 3. R project’in yansıları**

Bu işlemin sonrasında r studio için bir indirme işlemi gerçekleştirilir. Bunun için *“*<https://www.rstudio.com/products/rstudio/download/#download>*”* adresine gidilip işletim sistemine uygun sürüm seçilerek indirme işlemi başlatılır.

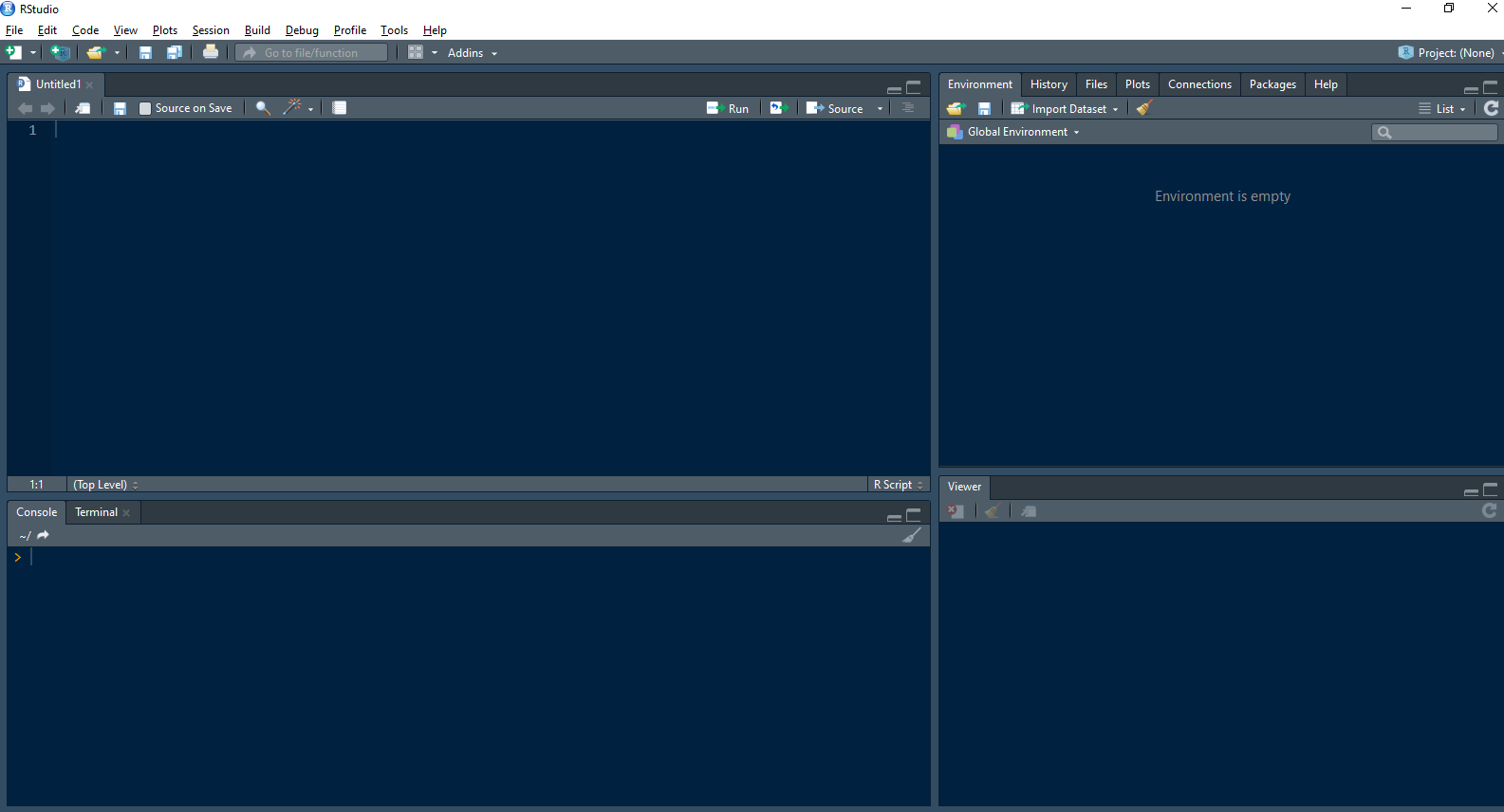


Ş**ekil 4. R studio dosyaları**

Ardından indirilmiş olan r project adlı kurulum dosyası “next,next,next” kolaylığında kurulacaktır. Bir diğer kurulum işlemi olan r studio kurulumu da ilk kurulumdan farksız olacaktır. Kurulum dosyasını çalıştırıp “ileri ileri” demek yeterli olacaktır.

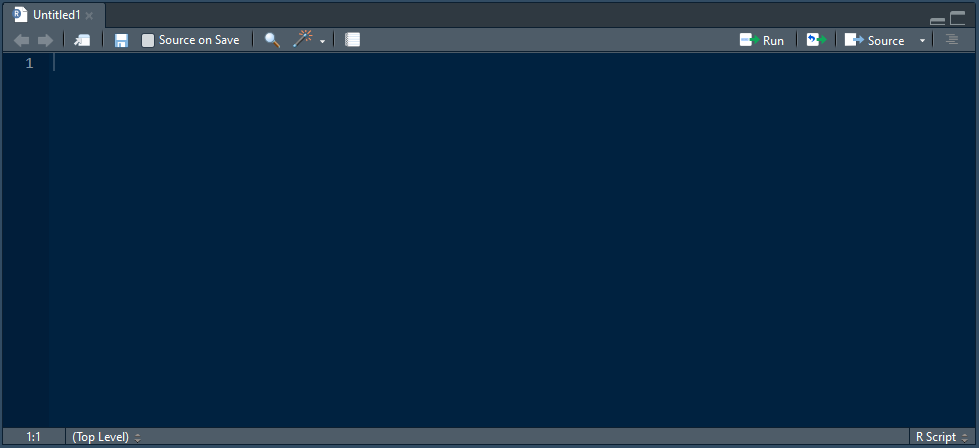
## 2.1.2. R Studio’ya Genel Bakış ve Konfigürasyonlar

R studio ilk yüklendiğinde kullanıcıyı aşağıdaki gibi bir ekran karşılar.

****

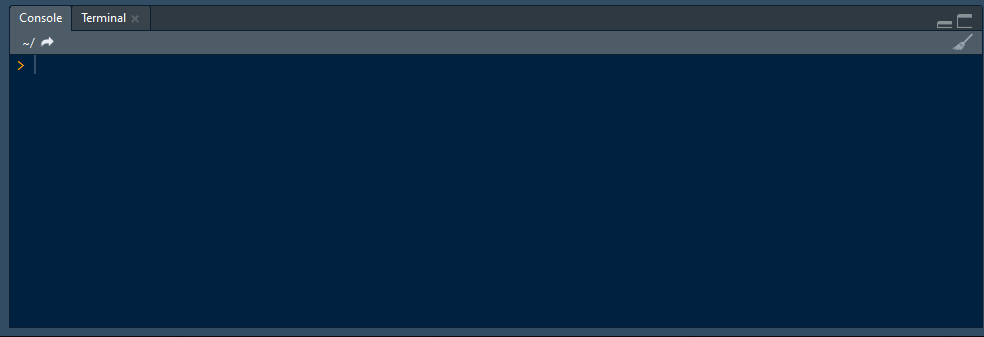
Ş**ekil 5. R studio**

Bu ekranlardan genel itibariyle söz etmek gerekirse; aşağıdaki sol üst kısımda yer alan ekran kodların yazıldığı kısımdır.

****

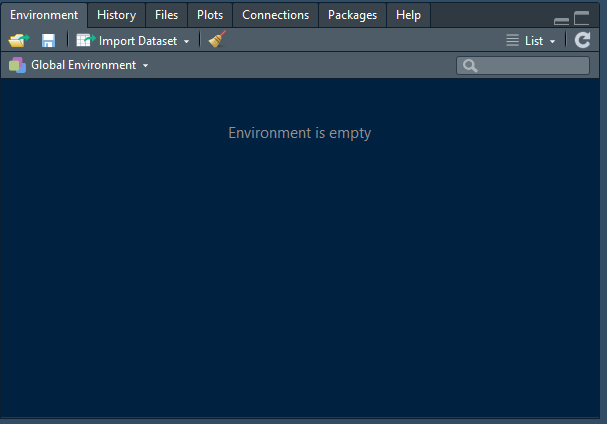
Ş**ekil 6. R studio’nun kod kısmı**

Sol alttaki bu konsol kısmı ise kodların çıktılarının görüntülendiği kısımdır.



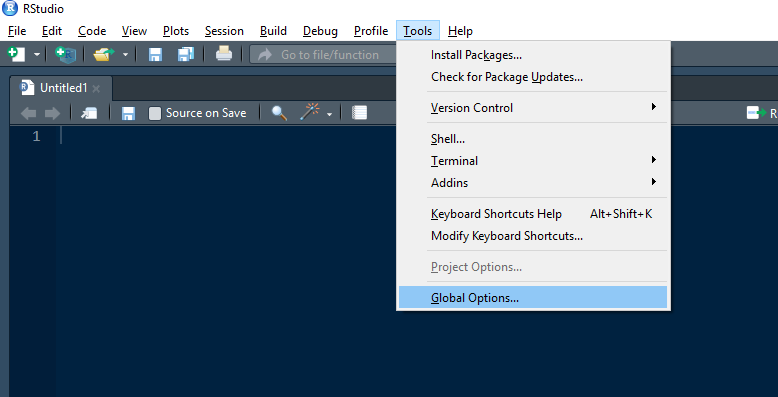
Ş**ekil 7. R studio’nun kod çıktılarının bulunduğu console kısmı**

Sağ üstte bulunan bu kısım ise birçok sekmeden oluşmaktadır. Bu sekmelerden environment sekmesi kod içinde tanımlanan değişkenlerin görüntülenmesini sağlayan kısımdır. History çalışan kodların geçmişini files ise projenin dizinini gösterir. Plots sekmesi herhangi bir şekil çizdirilirse bu şekli görüntüler. Connections r studio ile kurulmuş bağlantıları gösterir. Packages sekmesi ise içerisinde birçok kütüphane barındırır. Packages sayesinde bu kütüphaneleri görebilir ve projede kullanmak için projeye dahil edebiliriz.. Help sekmesinde ise herhangi bir r komutu veya kütüphanenin ne işe yaradığına dair bilgileri içerisinde barındırır.



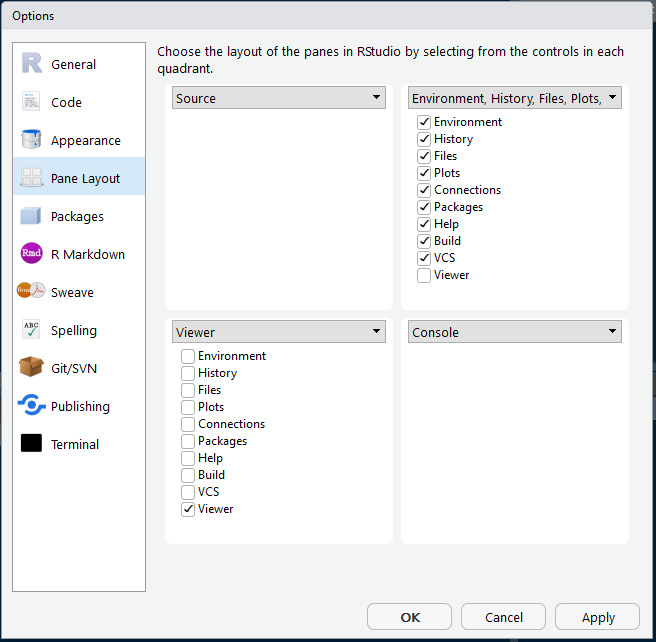
Ş**ekil 8. R studio’nun diğer sekmeleri**

Aşağıda gösterildiği gibi Tools --> Global Options diyerek r studio için özelleştirmeler yapılabilecek bir menüye gidilir.

**

Ş**ekil 9. R studio’nun genel ayarlarının bulunduğu sekmeye geçiş**

Aşağıdaki Pane Layout sekmesinde r studio’nun Pencere ayarları yapılabilir. Appearance kısmında ise r studio içerisinde kullanılan yazı tipi, yazı büyüklüğü, renklendirmeler gibi kısımlar ile kullanıcıya özel bir şablon oluşturulabilir..

**

Ş**ekil 10. R studio’nun genel ayarları**

## 2.2.Veri Setinin incelenmesi

## 2.2.1.Veri Setinin Genel Tanıtımı

Kullanılan *“heart\_disease”* veri seti 4 farklı hastaneden elde edilen kalp rahatsızlığına yönelik verilerle oluşturulmuştur. Bu veri setinin oluşturulmasında bahsi geçen hastaneler şöyledir; University Hospital - Zurich, Hungarian Instute of Cardiology – Budapest, University Hospital – Basel, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. Veri seti ham halindeyken verilerin hepsi double formatlı ve sürekli değişken olarak kabul edilmiştir. Fakat uygulamada veri seti kullanılırken hedef değişken kategorik değişkene çevirilip öylece sınıflandırma için kullanılacaktır. Bu değişkenlerden bahsedilecek olursa;

* Age : hastanın yaşı
* Sex: hastanın cinsiyeti
* Cp: göğüş ağrısı tipi (chest pain type)
* Trestbps: dinlenme kan basıncı (resting blood pressure)
* Chol: glukoz seviyesi mg/dl cinsinden
* Fbs : açlık kan şekeri (formülü: açlık kan şekeri>120 mg/dl) (1: doğru, 0: yanlış)
* Restecg: dinlenmiş haldeki hastanın elektrokardiyografi sonucu
* Thalach: elde edilen maksimum kalp atış hızı
* Exang: egzersize bağlı anjin (1: evet, 0: hayır)
* Oldpeak: Dinlenmeye göre egzersiz ile indüklenen ST depresyonu
* Slope: ST segmentinin eğimi
* Ca: floroskopi ile renklendirilmiş ana damarların sayısı (0-3)
* Thal: 3 = normal; 6 = sabitlenmiş kusur; 7 = tersinir kusur
* Target: 1 = kalp rahatsızlığı var 0 : kalp rahatsızlığı yok

Aşağıdaki kısımda, yukarıdaki değişkenlere sahip veri setinin üzerinde bir keşifçi veri analizi yapılmıştır.

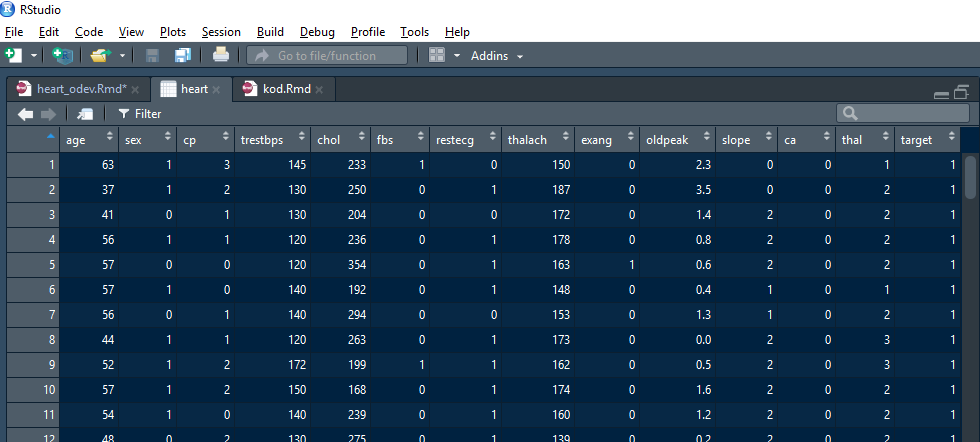
## 2.2.2. Veri Setinin R Studio’ya yüklenmesi

Veri setini r studio’ya yükleyebilmek için aşağıdaki kodlar kullanılır.



Veri setini r studio içerisinde görüntüleyebilmek için aşağıdaki kod kullanılır.





Ş**ekil 11. Heart\_disease adlı veri setinin r studio içerisinden görünümü**

Daha sonra aşağıdaki kodlar ile veri seti içerisindeki double tipindeki target isimli hedef değişken, yes – no değerlerini alacak bir kategorik değişkene çevirilip “dataframe\_heart” adlı bir dataframe’e atıldı.

## 2.2.3. Veri Setinin Modeller İçin Düzenlenmesi

## 2.2.4. Veri Seti Üzerinde Keşifçi Veri Analizi

Ardından veriye ilk bakış amaçlı aşağıdaki fonksiyonlar çalıştırıldı.

Kullanılan colnames fonksiyonu ile verisetindeki değişken isimleri görüntülendi.



Ş**ekil 12. Veri setine ait kolon isimleri**

Burada veri setine uygulanan nrow fonksiyonu ile 303 adet gözleme sahip olunduğu belirlendi.



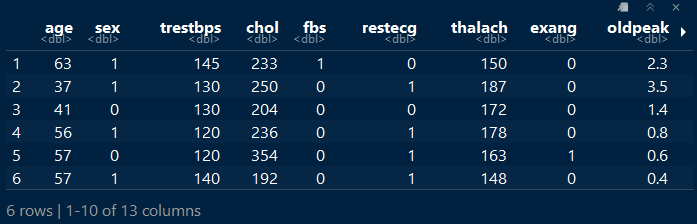
Ş**ekil 13. Veri setindeki gözlemlerin sayısı**

Aşağıda kullanılan ncol değişkeni ile veri setinde 13 adet değişkene sahip olunduğu görüldü.



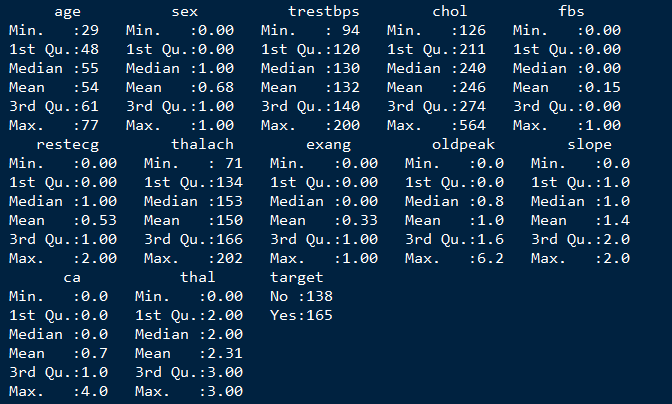
Ş**ekil 14. Veri setindeki kolon sayısı**

Head fonksiyonunun verdiği sonuçlar, veri setindeki ilk 6 gözleme ait sonuçlardır. Aşağıda değişkenlerin ne tarz değerler aldığı görüntülenmiştir.



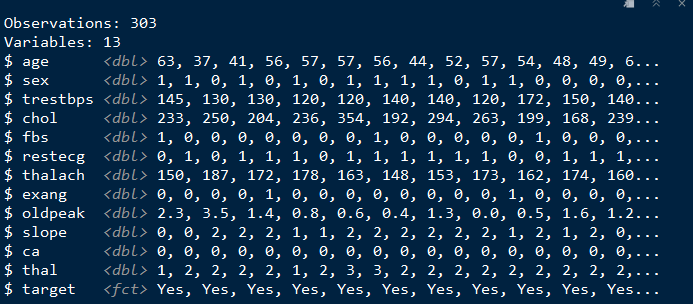
Ş**ekil 15. İlk altı verinin değerleri**

Summary fonksiyonu ise çok daha çeşitli bilgilerin bulunduğu bir istatistik özeti sunmaktadır. Kullanılan veri seti özelinde bakıldığında age değişkeninin minimum değerinin 29 amksimum değerinin 77 olduğu gözlenmektedir. İlk çeyreklikteki ortalamanın 48, gerçek ortalamanın 54 ve üçüncü çeyreklik ortalamasının da 61 olduğu tespit edilmiştir. Ayrıca medyan değerininde 55 olduğu saptanmıştır. Buradaki bilgilere bakıldığında hedef değişken olan target değişkeninin ye ve no değerlerini hangi frekansta aldığı görüldü.



Ş**ekil 16. Summary fonksiyonunun çıktısı – özet istatistikler**

Aşağıdaki glimpse fonksiyonu ise biraz daha farklı bir çıktı sunmuştur. Burada yine 303 gözlem ve 13 değişken olduğu ifade edilmiştir. Ayrıca bu değişkenlerin tipleri ve aldıkları değerlerden bir kestite sunulmuştur.



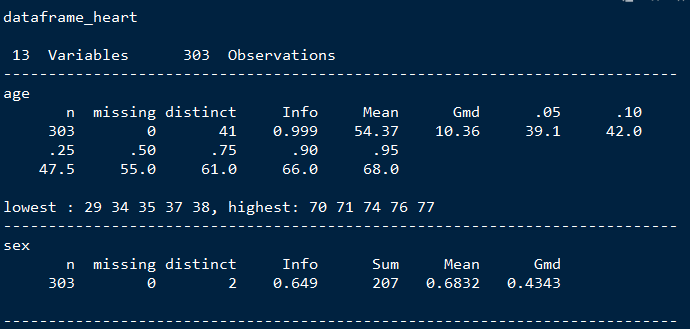
Ş**ekil 17. Glimpse fonksiyonunun çıktısı – özet istatistikler**

Aşağıdaki şekil funModelling kütüphanesinde yer alan plot\_num fonksiyonuyla oluşturulmuştur. Burada sürekli veya kategorik değişkenlerin dağılım grafiklerini bir arada görme imkanı sağlanmıştır.



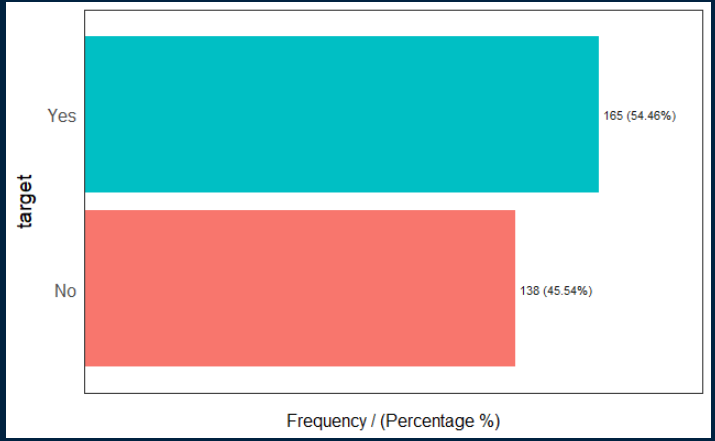
Ş**ekil 18. Bağımsız değişkenlerin görselleştirilmiş hali**

Yine kullanılan Hmisc::describe fonksiyonu ile gözlemler farklı bir formatta incelenmiştir. Burada diğerlerinden farklı olarak kayıp değer sayısı, benzersiz değer sayısı gibi bilgilerde verilmiştir. Örneğin age değişkeni için verisetinde 41 adet benzersiz değer varken, sex yani cinsiyet değişkeni için 2 adet benzersiz değer bulunmaktadır.



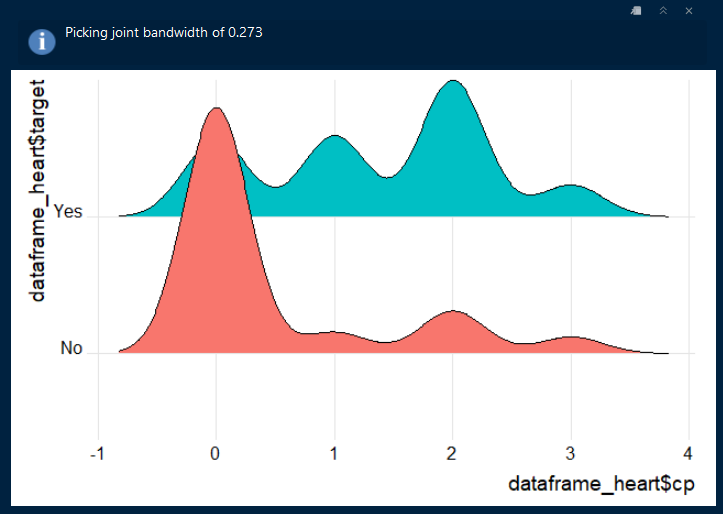
Ş**ekil 19. Hmisc::describe fonksiyonunun çıktısı**

Aşağıdaki grafiği elede etmek amaçlı kullanılan freq fonksiyonu ise kategorik değişkenler için bir dağılım grafiği oluşturmaktadır. Kullanılan veri setinde kategorik olarak sadece target olması sağlandığı için burada sadece target’a ait bilgiler yer almıştır. Buradaki grafikte target değişkeni 165 tane yes değeri ile yani kalp rahatsızlığı olan hastaları %54.46 oranıyla temsil ederken, 138 tane no değeriyle de kalp rahatsızlığı bulunmayan bireyleri %45.54 oranıyla temsil etmiştir.



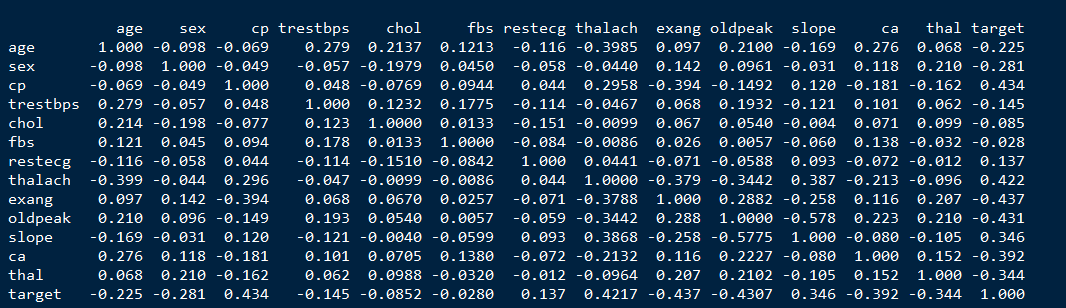
Ş**ekil 20. Bağımlı değişken olan targetin dağılımını gösteren bir grafik**

Aşağıda ise ggplot fonksiyonunun bir türevi kullanılmıştır. Buradaki grafikte bağımsız değişkenlerden olan cp değişkeninin bağımlı değişken olan target değişkeni üzerindeki dağılımı gözlemlenmiştir.



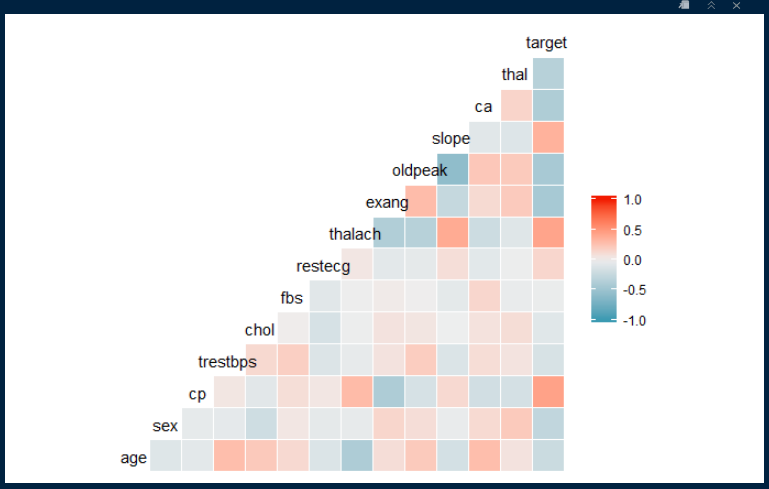
Ş**ekil 21. Cp bağımsız değişkeninin target bağımlı değişkeni üzerindeki dağılımı**

Aşağıda bulunan GGally kütüphanesinden cor fonksiyonu değişkenler arasındaki korelasyonu gösteren bir korelasyon matrisi oluşturmaktadır. Buradaki korelasyon matrisiyle hangi değişkenin hangi değişkenle ne kadar bağıntılı olduğu görülebilir. Burada öncelikle bakılması gereken hedef değişken olan target’ın diğer değişkenlerle olan ilişkisidir. Örneğin target değişkeninin en fazla ilişkide olduğu değişken cp değişkenidir. Target değişkeni cp değişkeniyle pozitif bir korelasyon içerisindedir. Bu da cp değişkeninin değeri artarken target değişkeninin de değerinin arttığını göstermektedir. Bu matris bu şekilde yorumlanabilir.



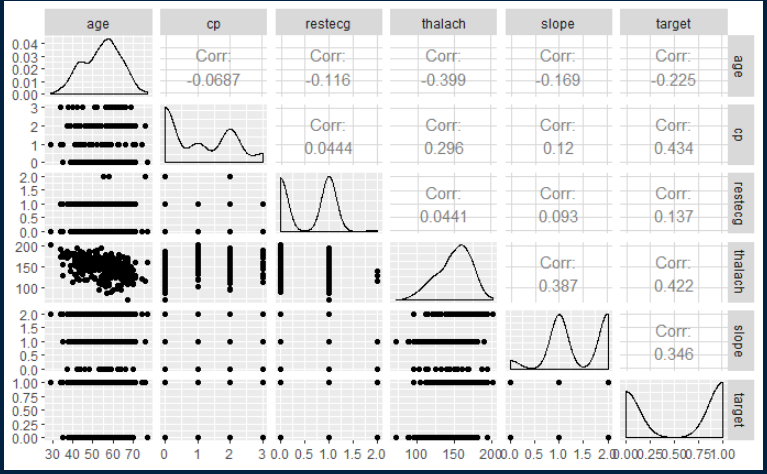
Ş**ekil 22. Değişkenlerin birbiri arasındaki ilişkiyi gösteren korelasyon matrisi**

Yine korelasyonlarla ilgili bir grafik olan aşağıdaki grafik ggcor fonksiyonu ile oluşturulmuştur. Buradaki grafik daha çok ısı haritası tekniğine benzer bir yöntemle oluşturulmuştur. Buradaki grafikte target kolonuna bakılır ise biraz önce incelenen korelasyon matrisinde target değişkeniyle daha çok ilişkide olan değişkenlerin renklerinin daha kızıl tonlarda olduğu görülecektir. Bu ilişki zayıfladıkça ise değişkenin rengi maviye doğru gitmiştir. Buradan da kalp rahatsızlığına etken olacak değişkenler bir nebze tespit edilebilir.



Ş**ekil 23. Korelasyonların ısı haritası üzerinde gösterimi**

Aşağıdaki ggpairs fonksiyonu ile bazı seçilmiş değişkenlerin yine korelasyonları ölçümlenmiştir. Buradaki grafiktede görüldüğü üzere kalp rahatsızlığındaki en önemli etmenin şu anki bilgilere göre göğüş ağrısı tipi olduğu söylenebilir.



Ş**ekil 24. Korelasyonların bir başka gösterimi**

## 2.3. Amaç

Buraya kadar yapılan çalışma ile veri içerisinde keşifçi bir veri analizi yapılmıştır. Değişkenlerin özellikleri, dağılımları ve aldıkları değerler gözlemlenmiş ve bu gözlemler görselleştirilmiştir. Burada veri setine hakim olmak için bir öğrenim gerçekleştirilmiştir. Kullanılan veri setinin 303 gözlem ve 13 değişkenden meydana geldiği ve hedef değişkenin yes ve no değerleri alan bir kategorik değişken olduğu tespit edilmiştir. Ardından kullanılan fonksiyonlar ile oluşturulan çıktılar üzerinden değişkenlerin birbiriyle ilişkileri hem korelasyon matrisi hem de ısı haritalarıyla gözlemlenmiştir. Özellikle bağımlı değişken ile bağımsız değişkenler arasındaki korelasyonlar dikkatlice incelenmiş ve bazı çıkarımlar yapılmıştır.

Buradan sonra yapılacak çalışmalar gözlemlenen ve bilgi sahibi olunan bu değişkenler ile bazı sınıflandırma algoritmaları kullanılarak tahmin modelleri oluşturmaktır. Bu modeller oluşturulduktan sonra algoritmaların bu problem özelinde ne gibi sonuçlar verdikleri ve doğruluk değerleri yorumlanacaktır. Ayrıca modellerin daha iyi hale gelmesini sağlayacak bazı yararlı düzeltmeler bulunup bulunamayacağı hakkında araştırmalar da yapılacaktır.

## 3. Veri Seti Üzerine Model Kurma ve Kurulan Modellerin Değerlendirmesi

## 3.1. Lojistik Regresyon Modeli

**Sözde Kod**

Veri seti test train olarak ayrılır. testing, training

formülü ile, sıfırla bir aralığında bir lojit dönüşüm yapılır ve varsayılan olarak hedef sınıf Y=Yes olma olasılığı hesaplanır.

formülü ile lojistik regresyonun odds’u hesaplanır.

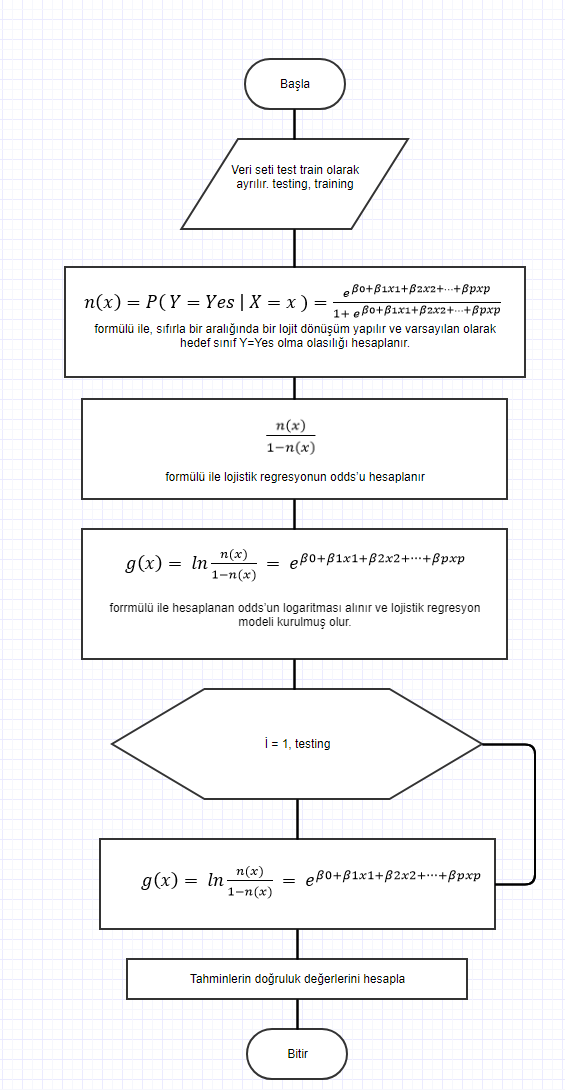
formülü ile hesaplanan odds’un logaritması alınır ve lojistik regresyon modeli kurulmuş olur.

For i=1 to testing do

End For

Yapılan tahminler için hata oranını hesapla.

**Akış diyagramı**



Lojistik regresyon bağımlı ve bağımsız değişkenler arasındaki bağlantıyı tanımlayan doğrusal bir model kurmak için kullanılır.

Lojistik regresyonun önemli özellikleri

* Bağımlı değişken kategoriktir.
* Bağımlı değişkenin değerinin 1 olduğu durumun gerçekleşme olasılığı hesaplanır.
* Lojistik fonksiyonu sayesinde üretilen değerler 0-1 arası olur.

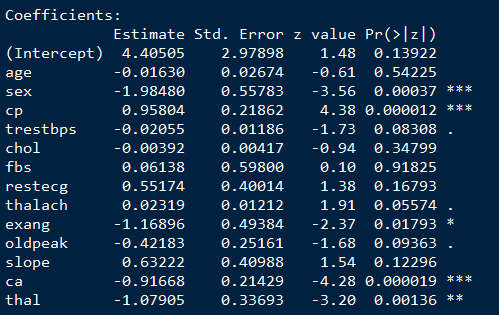
## 3.1.1 Veri Setinin Test ve Eğitim Kümelerine Ayrılması

Aşağıdaki kod yardımıyla veri setinin yüzde seksene yüzde yirmi, test ve eğitim kümeleri olarak ayrılması sağlandı. Veri setinin yüzde seksen değerini eğitim yüzde yirmi değerini ise test değeri içerisinde barındırmaktadır.



## 3.1.2. Lojistik Regresyon Modelinin Kurulması

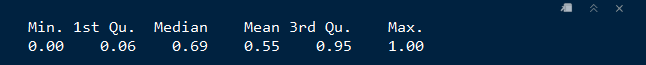
Burada ise lojistik regresyon modelinin oluşturulması sağlanmıştır. Glm fonksiyonunun bir lojistik regresyon modeli oluşturmasını sağlamak için fonksiyona binomial parametresi verilmiştir. Modelin öğrenme sürecinde hangi değişkenlerin daha etkili olduğu aşağıdaki çıktı ile belirlenebilir. Burada sex,cp ve ca değişkenleri çıktı da da görüldüğü gibi üç yıldız almışlardır. Bu sonuçta bu bağımsız değişkenlerin bağımlı değişken olan target üzerinde etkili olduklarını göstermektedir.



Ş**ekil 25. Lojistik regresyon modeli ile gözlemlenen değişkenlerin katsayıları gözlendi**

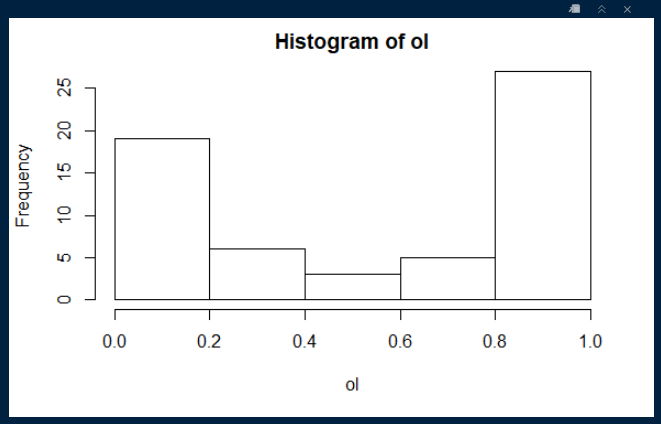
## 3.1.3. Lojistik Regresyon Modeli ile Tahmin Yapılması

Oluşturulan model ile bir tahmin yapmak istenirse aşağıdaki adımlar takip edilir. Burada kurulan model ile 0 ile 1 arasında bazı tahminler yapıldı. Aşağıda yine summary fonksiyonu ile bu tahmin değerlerine ait bilgiler incelendi. Burada tahmin değerlerinin ortalaması 0.55, medyan değeri ise 0.69 çıkmış yani nerdeyse yarı yarıya dağıldığı belirlenmiştir.



Ş**ekil 26. Lojistik regresyon ile yapılan tahminlerin özet istatistiği**

Bu tahmin değerleri bir histogram üzerinde incelenecek olur ise, tahminlerin neredeyse eşit bir şekilde dağıldığı gözlemlenebilir. Buradaki eşitlik 0.5 üzeri ve 0.5 aşağısı olarak düşünülebilir.



Ş**ekil 27. Lojistik regresyon ile yapılan tahminlerle oluşturulan histogram**

Aşağıda tahminler 0.5’ten fazla ve yahut az olarak sınıflandırılmıştır. Buradaki amaç 0’a yakın olan tahminlerin 0, 1’e yakın tahminlerin ise 1 olarak değerlendirilmesidir. Çünkü tahmin edilen target değişkeninin bu tarz bir ikili sınıflandırması söz konusudur. Sonuç olarak aşağıdaki table fonksiyonu yardımıyla oluşturulan çıktı incelenirse, buradan veri setinde 27 adet kalp rahatsızlığı bulunmayan bireyin, 33 adet ise kalp rahatsızlığı bulunan bireyin olduğu çıkarılabilir.



Ş**ekil 28. Lojistik regresyon ile tahmin edilen değerler**

## 3.1.4. Lojistik Regresyon Modeli Tahminlerinin Değerlendirilmesi

Burada çıkarılan sonuçların ne derece güvenli olduğu hususunda da aşağıdaki çalışma yapılmıştır. Burada yanlış sınıflandırma oranı ve doğruluk değeri (accuracy) değeri hesaplamaları yapılmıştır. 0.13 çıkan ilk sonuç yanlış sınıflandırma oranıdır.



Ş**ekil 29. Oluşturulan lojistik regresyon modelinin test hatası**

Bu yanlış sınıflandırma oranı 1 değerinden çıkarılır ise de doğruluk değeri (accuracy) bulunur. Buradaki modelin doğruluk değeri 0.87 olarak tespit edilmiştir.

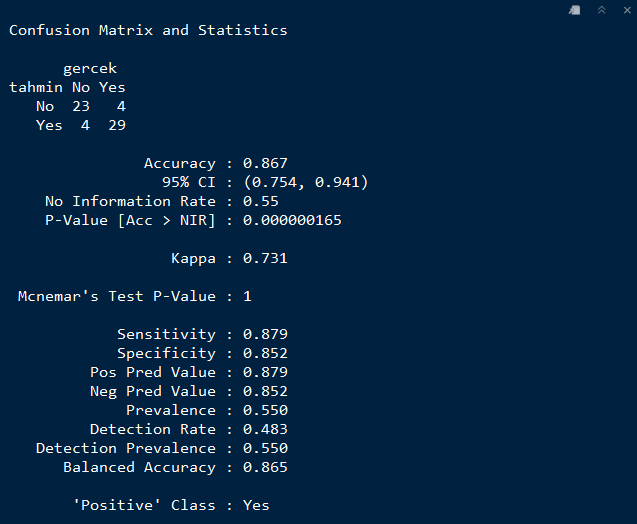


Ş**ekil 30. Lojistik regresyon modelinin doğruluk oranı**

Aşağıda tahmin değerleri ile bir karmaşıklık matrisi oluşturulmuştur. Bu karmaşıklık matrisine bakılır ise şöyle sonuçlar söz konusu olmuştur;

* 23 adet kalp rahatsızlığı bulunmayan bireye kalp rahatsızlığınız yok denilmiştir.
* 4 adet kalp rahatsızlığı bulunan bireye kalp rahatsızlığınız yok denilmiş ve hata yapılmıştır.
* 4 adet kalp rahatsızlığı bulunmayan bireye kalp rahatsızlığınız bulunmaktadır denilmiş ve hata yapılmıştır.
* 29 adet kalp rahatsızlığı bulunan bireye kalp rahatsızlığınız bulunmaktadır denilmiştir.

Bu sonuçlar %88’lik bir hassaslık ve %87’lik bir doğrulukta tahmin edilmiştir.



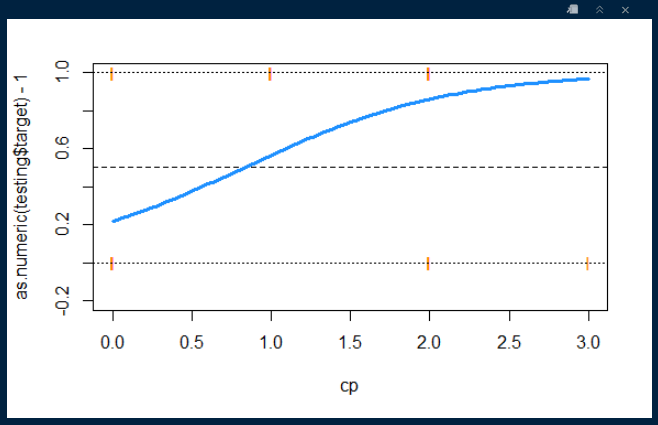
Ş**ekil 31. Lojistik regresyona ait karmaşıklık matrisi**

Aşağıdaki bir başka çıktıda da yine aynı hassaslık ve doğruluk oranları verilmiştir.



Ş**ekil 32. Lojistik regresyon modelinin doğruluk ve hassaslık değerleri**

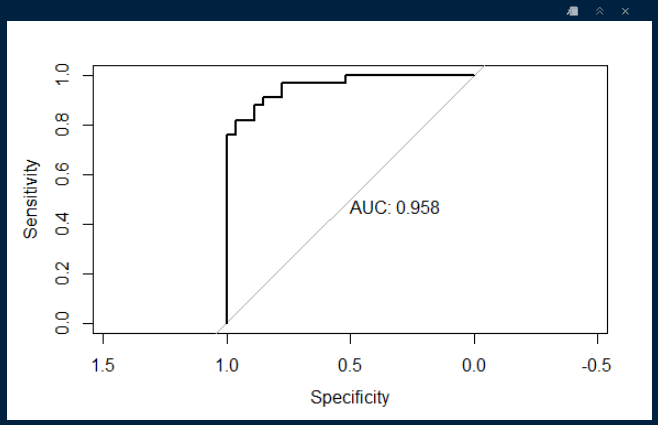
Aşağıdaki görsel de ise bağımlı değişkenin en fazla ilişkide olduğu tespit edilmiş olan bağımsız değişkenle olan ilişkisi görselleştirimiştir. Burada kalp ağrısı tipinin, bireyin kalp rahatsızlığının bulunmasındaki etkisi görselleştirilmiştir.



Ş**ekil 33. Cp değişkeni ile target değişkeni arasındaki bağıntının grafiği**

ROC eğrisini tanımlarken ikili sınıflama için yapılan ayrım eşik değerlerinin değişkenlik gösterdiği durumlarda, hassasiyet/kesinlik hesaplamasıyla oluşturulduğu söylenebilir. Daha basit bir anlatımla doğru-pozitif değerlerinin yanlış-pozitif değerlere bölünmesiyle elde edilen bir grafiktir. ROC eğrisindeki grafik sol üst köşeye yaklaştıkça testin genel doğruluğu artar. Ayrıca ROC eğrisinin altında kalan alana AUC değeri denilmektedir ve bu değer bir değişkendeki iki gurubun ne kadar iyi ayırt edilebildiğini göstermektedir.

Aşağıda çizilen ROC eğrisine ve AUC değerine bakılır ise kurulan lojistik regresyon modelinin çok kötü bir model olduğu söylenemez.



Ş**ekil 34. Lojistik regresyona ait ROC eğrisi**

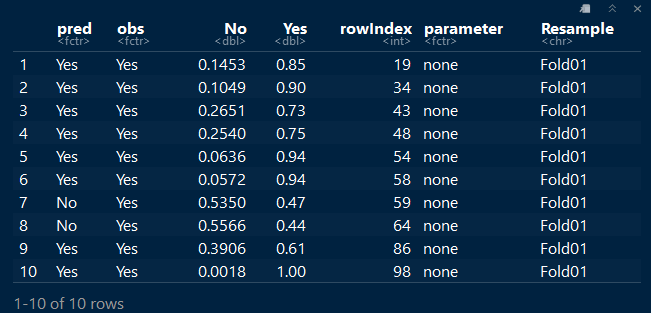
## 3.1.5. Lojistik Regresyon Modelindeki Hatayı Ölçme

Bu kısımda ise oluşturulan modelin ne kadar güvenilir bir model olduğuyla alakalı bir çalışma yapılmıştır. Kullanılan çapraz doğrulama (cross-validation) yöntemi şöyle bir fikir ile çalışmaktadır; eğitim kümesinde bulunan verilerden birbirinden farklı doğrulama kümeleri oluşturulur ve bu kümelerin üzerinde modelin performansı sınanır. Böylece yeni gözlemler öngrülmek istenilirse performası test edilen modelin nasıl bir performans göstereciğine dair bir fikir oluşmuş olur.

Aşağıda uygulanan k katlı çapraz doğrulama işleminde doğrulama kümeleri sayısı 10 olarak seçilmiştir ve modele uygulanmıştır.



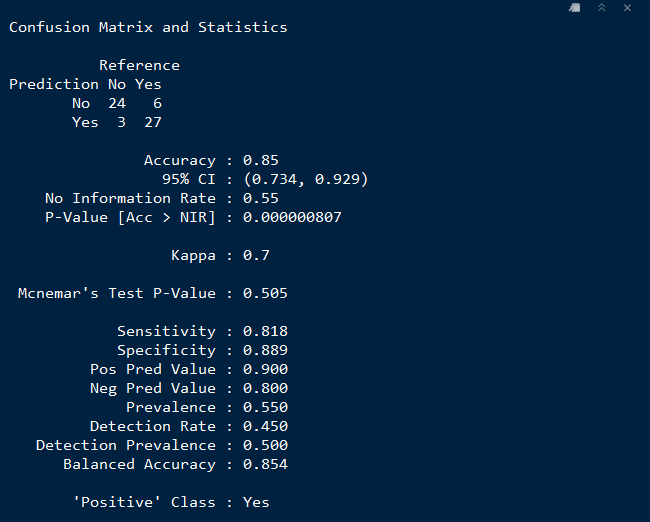
Bu uygulama sonucunda ilk 10 kaydın gerçek gözlemleri ve tahmin değerleri aşağıdaki çıktıda verilmiştir.



Ş**ekil 35. Cross-validation uygulanmış modelin ilk 10 tahmini**

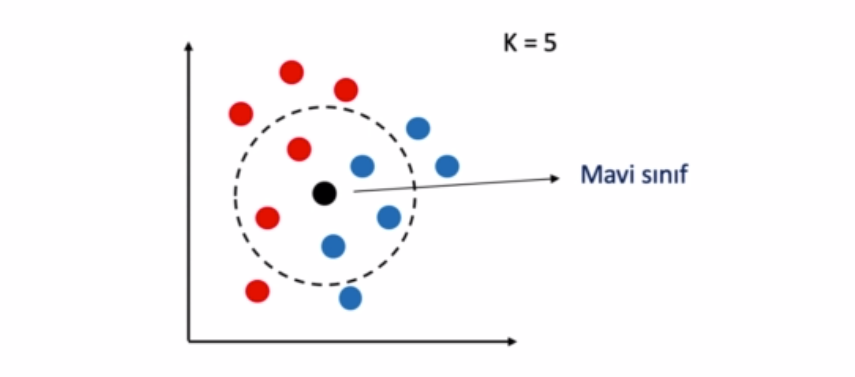
Sonuç olarak modele uygulanan çapraz doğrulama işlemi sonucunda elde edilen tahmin değerleri aşağıdaki karmaşıklık matrisinde verildiği gibi olmuştur. Çıktıdan da görüldüğü gibi eğitilen model %85’lik bir değerle doğru tahminde bulunmuş ve değişen doğrulama kümeleri karşısında sonuçta %2’lik az bir değişim olduğu gözlemlenmiştir.





Ş**ekil 36. Cross-validation uygulanmış modelin karmaşıklık matrisi**

## 3.2. KNN Modeli



Ş**ekil 37. KNN’e ait bir görsel**

KNN, en yakın komşuluk yaklaşımına dayanan bir modeldir. Bu modelde yukarıdaki siyah noktada olduğu gibi en yakın komşulara bakılıp henüz sınıflandırılmamış olan siyah noktayı sınıflandırmaya yönelik bir algoritmadır. Burada k ile belirtilmiş komşuluk değeri 5 olduğu için yeni gelen siyah noktanın etrafındaki en yakın 5 komşusuna bakılır. Bu komşuların sınıfları incelenir ve en fazla olan sınıf yeni gelen gözlemin sınıfı olur.

**Sözde Kod**

Sınıflandırma(X,Y,x) // X: eğitim verisi, Y: X’in sınıf etiketleri, x: bilinmeyen değer

For i=1 to m do

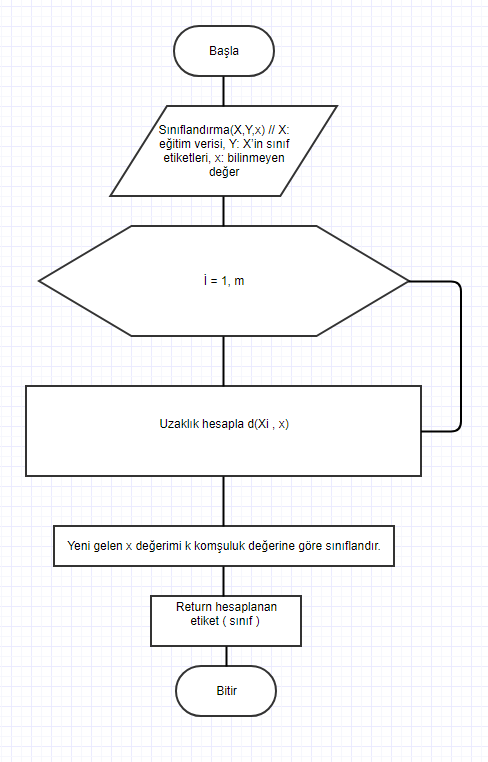
Uzaklık hesapla d(Xi , x)

End For

Yeni gelen x değerimi k komşuluk değerine göre sınıflandır.

Return hesaplanan etiket ( sınıf )

**Akış diyagramı**



## 3.2.1. Veri Setinin Test ve Eğitim Kümelerine Ayırılması

Aşağıdaki kodlar ile veri seti eğitim ve test verileri olarak yüzde seksen ve yüzde yirmlik oranlarla ayrılmıştır. Buradaki ayrım yüzde yirmi test, yüzde seksen eğitim verisi olarak seçilmiştir.



## 3.2.2. KNN Modelinin Kurulması

Buradaki knn fonksiyonu çok esnek bir fonksiyon değildir fakat yine de belli sonuçlar alınmıştır. Bu sonuçlara ulaşabilmek için öncelikle FNN kütüphanesinde bulunan knn fonksiyonu kullanılıp belli parametrelerle model kurulmuştur. Bu parametrelerden ilki eğitim verisi, ikincisi test verisi, üçüncüsü eğitim setindeki hedef değişken ve son olarak da komşuluk değeri olan k değeri fonksiyona verilmiştir.



Aşağıdaki kodlar yardımıyla, oluşturulan knn modelinin hata oranı bulundu. Bu hata oranı 0.3 olarak hesaplandı.

Ş**ekil 38. KNN hata oranı**

Bu elde edilen hata oranı değeri 1 değerinden çıkarılır ise doğruluk oranı (acccuracy) değerine ulaşılır. Bu knn modelinin doğruluk değeri yüzde yetmiş oranında çıkmıştır.



Ş**ekil 40. KNN doğruluk oranı**

Burada ise bir önceki modelden farklı olarak k komşuluk değerini 5 yapılarak bir model kuruldu. Bir önceki kurulan modelde k komşuluk değeri 3 olarak alınmıştı. Burada k komşuluk değeri 5 alınarak model kurma işlemi gerçekleştirilmiştir. Ardında hata ölçümü yapıldığında hata oranının biraz olsun düştüğü gözlemlenmiştir. Bu oran 0.27 olarak bulunmuştur.



Ş**ekil 41. K = 5 komşuluk değerinde KNN’in hata oranı**

Daha sonra k komşuluk değeri 10 alınarak bir model kurulmuş ve bu modelin de hata oranı biraz daha düşük çıkmıştır. Söz konusu hata oranı 0.25 olmuştur.

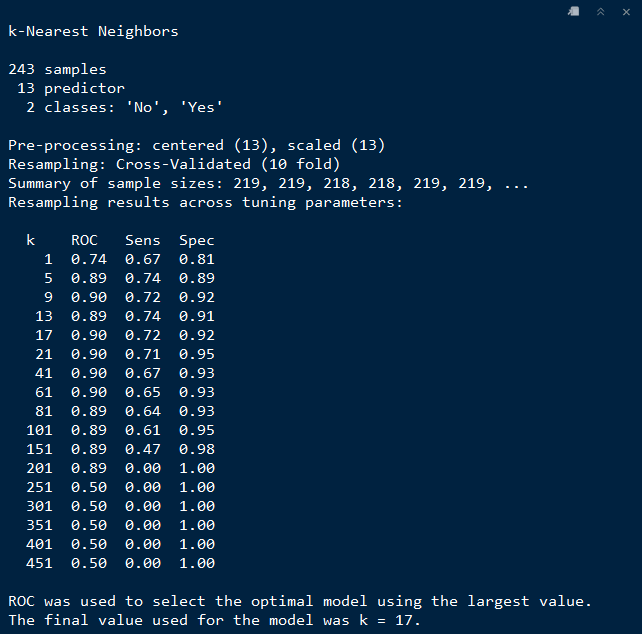


Ş**ekil 42. K = 10 komşuluk değerinde KNN’in hata oranı**

## 3.2.3. KNN Modelinin Optimize Edilmesi ve Tahmin

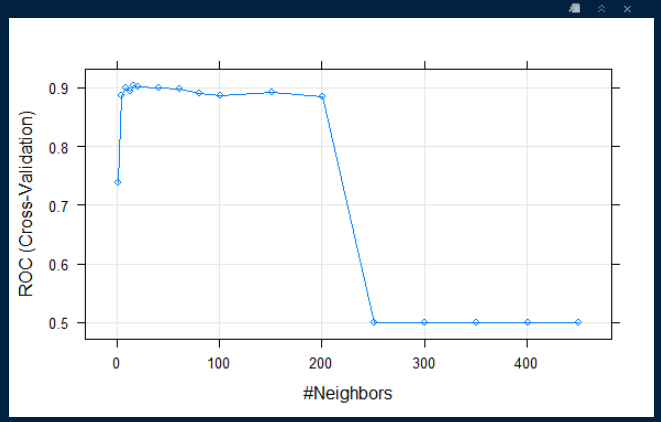
Kurulan bu modelin daha optimize bir model olması adına train adlı bir fonksiyon kullanıldı. Bu fonksiyonun çalışabilmesi açısından modele öncelikle bir çapraz doğrulama işlemi uygulandı. Ardından bir arama vektörü oluşturuldu. Bu arama matrisi aşağıda görülen çıktıdaki k değerlerinin ne olacağını belirleyen bir matristir. Daha sonra uygun değerlerle oluşturulan train fonksiyonunun çıktısı aşağıdaki şekildeki gibi olmuştur. Burada modelin en optimum sonuçlar üretebilmesi açısından k komşuluk değerinin 17 olarak seçilmesi tavsiye edilmiştir. Bu 17 sonucu arama vektörünün içerisindeki k değerlerinden, modeli en optimum sonuca götüren değer olarak belirlenmiştir.





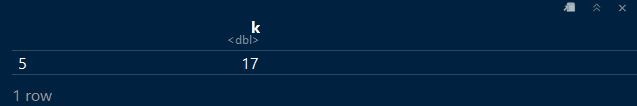
Ş**ekil 43. Train fonksiyonunun çıktısı**

Bu arama vektörünün içerisindeki değerlerin başarımları bir grafik şeklinde gösterilir ise aşağıdaki çıktı elde eldilir. Bu grafikteki y eksenindeki en üst nokta 17 değerini temsil etmektedir.



Ş**ekil 44. Arama vektöründeki komşuluk değerlerinin performansları**

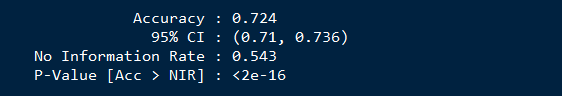
bestTune fonksiyonu ile de kısaca en iyi k değerine ulaşılabilir. Fakat unutulmamalıdır ki bu k değeri sadece arama vektörünün içerisindeki en iyi değerdir.



Ş**ekil 45. Arama vektöründeki en iyi komşuluk değeri**

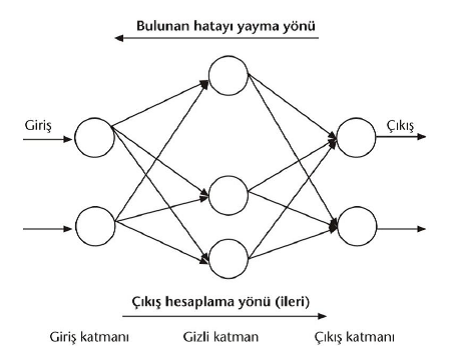
Arama vektöründen elde edilen en optimum sonuçla oluşturulan knn modelinin doğruluk değeri 0.72 olmuştur. Aşağıda bu uygulamada elde edilen en optimum değerler sıralanmıştır;

* k=3 için, doğruluk değeri = 0.7
* k=17 için, doğruluk değeri = 0.724
* k=5 için, doğruluk değeri = 0.73
* k=10 için, doğruluk değeri = 0.75



Ş**ekil 46. Arama vektörüne göre optimize edilmiş modelin çıktıları**

## 3.3. Yapay Sinir Ağı Modeli



Ş**ekil 47. Geri yayılımlı yapay sinir ağı**

Yapay sinir ağları genel olarak birbirleriyle ilişkili yapay sinir hücrelerinden meydana gelmektedir. Bu ağı oluşturan birimlere nöron adı verilir. Yapay sinir ağları genel olarak yukarıdaki yapıdadır. Yapay sinir ağlarında giriş katmanı, gizli katmanlar ve çıkış katmanları bulunur. Bu katmanlardaki nöronlar belli ağırlık değerlerine göre birbirleriyle ilişkidedirler. Yukarıdaki şekilde belirtilen aslında bir geri yayılımlı ağ örneğidir. Geriye yayılımlı ağların temelinde yatan fikir, kabul edilebilir bir hata miktarı elde edilinceye kadar ağ çıktısı ile gerçek çıktısı arasındaki farkları en aza indirgeyebilmek için girdi ağırlıklarının değiştirilmesidir. Girdi ağırlıkları başlangıçta genelde rastgele dağılır. Bu ağırlıkların güncellemesi olayı aslında ağın öğrenmesi olayıdır. Yapay sinir ağlarına, çok katmanlı algılayıcılar da denilmektedir.

**Sözde Kod**

Ağırlık için rastgele bir değer varsayımı yap. w = 0.25;

For i = 1 to ön tanımlı bir iterasyon sayısı do

Eğitim setini geçici bir değişkene at;

Δwi = 0;

While eğitim verisi boş değilse do

d = GüncelEgitimVerisiniAl();

r = MatristekiSatirIndeksiniBul(d);

c = MatristekiSutunIndeksiniBul(d);

od = AgCiktilariniHesapla();

tr = SatirBazindaMaksimumBenzerligiBul(r);

tc = SutunBazindaMaksimumBenzerligiBul(r);

Δwi = Δwi + η[(tr – od) + (tc – od)]sid;

GüncelEgitimVerisiniSil();

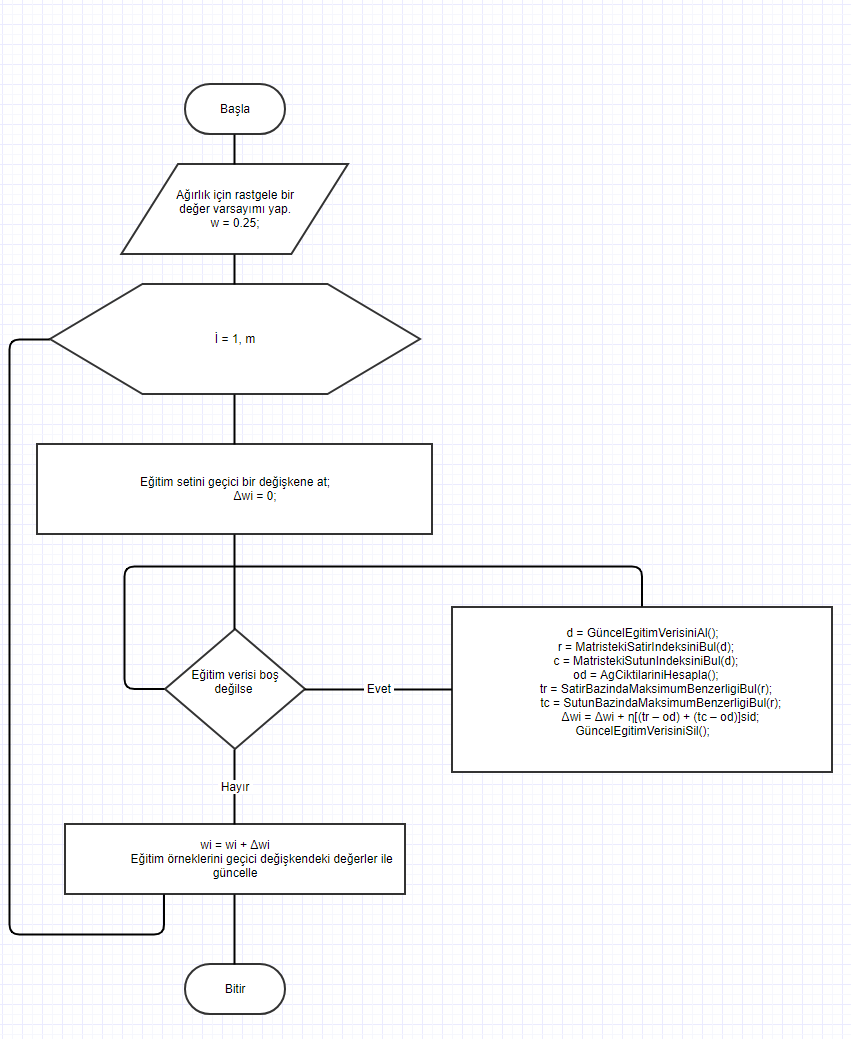
End While

wi = wi + Δwi

Eğitim örneklerini geçici değişkendeki değerler ile güncelle

End For

**Akış Diyagramı**



## 3.3.1 Yapay Sinir Ağı Modeli İçin Standartlaştırma İşlemi

Yapay sinir ağlarını sınıflandırma problemlerinde kullanabilmek için öncelikle veri setindeki bağımsız değişkenlerin bir standartlaştırma işlemine tabi tutulaması gerekmektedir. Bunun için aşağıdaki kodlar ile bağımsız değişkenler 0 ile 1 değerleri aralığına indirgenmiştir.



## 3.3.2 Veri Setinin Test ve Eğitim Kümelerine Ayırılması

Burada modelimizde kullanabilmek amacıyla verisetimizi uygun bir biçimde test ve eğitim verisi olarak iki kümeye bölüyoruz. Aşağıdaki örnekte yüzde seksen eğitim kümesi ve yüzde yirmi test kümesi olarak bir ayrım yapılmıştır.



## 3.3.3 Yapay Sinir Ağı Modelinin Oluşturulması

Aşağıdaki kodlar yardımıyla nnet kütüphanesinde bulunan yapay sinir ağı fonksiyonu projeye dahil edildi. Ardından ilgili parametrelerle veri seti için bir yapay sinir ağı modeli oluşturuldu. Buradaki parametrelerden ilki hedef değişken, ikincisi veri seti, üçüncüsü gizli katman sayısı ve sonuncu parametre de aşırı öğrenmenin önüne geçmek için kullanılan bir değer olan decay değeri olmuştur.

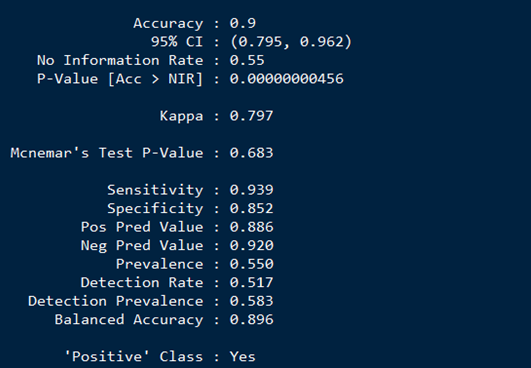


## 3.3.4 Yapay Sinir Ağı Modeli ile Tahmin

Model eğitildikten sonra, bu yapay sinir ağı modeli ile test verilerinin tahmini yapıldı.



Yapılan bu tahminlerin çıktıları aşağıda gözükmektedir. Burada oluşturulan yapay sinir ağı modelinin %90 oranında bir doğruluk değeri ile tahmin yaptığı saptanmıştır. Bu tahminlerin hassaslık oranları ise %90’dır.



Ş**ekil 48. Oluşturulan yapay sinir ağı modelinin karmaşıklık matrisi**

# SONUÇ

Bu tezin yazılması için yapılan uygulamada kullanılan algroritmaların veri seti üzerindeki performanları aşağıdaki şekilde olmuştur.

* Yapay sinir ağı (YSA) – doğruluk değeri (accuracy) = 0.9
* Lojistik regresyon – doğruluk değeri ( accuracy) = 0.87
* KNN
* k=10 için, doğruluk değeri = 0.75
* k=5 için, doğruluk değeri = 0.73
* k=17 için, doğruluk değeri = 0.724
* k=3 için, doğruluk değeri = 0.7

Yukarıdaki sonuçlardanda görüldüğü üzere çok güçlü iki algoritma ile kurulan yapay sinir ağı modeli ve lojistik regresyon modelleri KNN modeline karşı ezici bir üstünlük sağlamıştır. Bu durumun bu veri seti özelinde olduğu unutulmamalıdır.

Veri setinde target değişkenindeki dağılım %54.46 yes ve %45.54 no değerleridir. Yani bu veri setinde target değişkeninin hepsine yes denseydi %54.46’lık bir doğruluk değeri elde edilecekti. Fakat kurulan modeller ile elde edilen sonuçlar bundan çok daha iyidir. Özellikle YSA modeliyle %90’lık bir doğruluk oranına ulaşılmış ve bu sonuç oldukça başarılı kabul edilebilirdir.

Sonuç olarak bu veri seti özelinde kurulan modeller arasıdan en iyi tahmin değerlerini veren model yapay sinir ağıyla kurulan model olmuştur.

# KAYNAKÇA

SAVAŞ, Serkan, TOPALOĞLU, Nurettin ve YILMAZ, Mithat (2012), “Veri Madenciliği ve Türkiye’deki Uygulama Örnekleri”

KOYUNCUGİL, Ali Serhan ve ÖZGÜLBAŞ, Nermin (2009), “Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları”

ÖZEKES, Serhat (2003), “Veri Madenciliği ve Türkiye’deki Uygulama Örnekleri”

KÖKTÜRK, Firuzan, ANKARALI, Handan, SÜMBÜLOĞLU, Vildan (2008), “Veri Madenciliği Yöntemlerine Genel Bakış”