DOKUZ EYLÜL ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DENETİMLİ ÖĞRENME ALGORİTMALARI KULLANILARAK KARDİYOVASKÜLER KALP HASTALIĞIYLA İLGİLİ DENGESİZ VERİLERİN SINIFLANDIRILMASI

Buse Nur BALTACIOĞLU

Ağustos, 2021

İZMİR

DENETİMLİ ÖĞRENME ALGORİTMALARI KULLANILARAK KARDİYOVASKÜLER KALP HASTALIĞIYLA İLGİLİ DENGESİZ VERİLERİN SINIFLANDIRILMASI

Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü

Tezsiz Yüksek Lisans Dönem Projesi

İstatistik Anabilim Dalı, Veri Bilimi Programı

Buse Nur BALTACIOĞLU

Ağustos, 2021

İZMİR

TEZSİZ YÜKSEK LİSANS DÖNEM PROJESİ SONUÇ FORMU

BUSE NUR BALTACIOĞLU, tarafından **DOÇ. DR. NESLİHAN DEMİREL** yönetiminde hazırlanan "**DENETİMLİ ÖĞRENME ALGORİTMALARI KULLANILARAK KARDİYOVASKÜLER KALP HASTALIĞIYLA İLGİLİ DENGESİZ VERİLERİN SINIFLANDIRILMASI**" başlıklı Dönem Projesi tarafımdan okunmuş, kapsamı ve niteliği açısından bir Tezsiz Yüksek Lisans Dönem Projesi olarak kabul edilmiştir.

Kabul edilen Tezsiz Yüksek Lisans Dönem Projesi;

- ☐ Kapsamlı bir derleme
- ☐ Eleştirel bir rapor
- ■ Uygulamaya dönük bir proje
- ☐ Deneysel bir çalışma

.....

Doç. Dr. Neslihan DEMİREL

Danışman

TEŞEKKÜR

Öncelikle değerli fikir ve tecrübeleri ile beni yönlendiren, proje konusu seçiminde ve projenin her aşamasında bana destek olan, farklı bakış açısı kazandıran; bana bu çalışmayı yapma fırsatı veren saygıdeğer danışmanım Doç. Dr. Neslihan DEMİREL'e teşekkürlerimi sunarım.

Ayrıca lisans ve yüksek lisans sürecinde her zaman bilgilerini, deneyimlerini bizimle paylaşan ve bu alanda gelişimime büyük katkısı olan İstatistik Anabilim Dalı akademisyenlerine teşekkürlerimi sunarım.

Son olarak da her durumda hevesim ve isteğimin olmasını sağlayan değerli annem Gülayşe DEMİRCİ ve ablam Hande BALTACIOĞLU'na sonsuz teşekkürlerimi sunarım.

Buse Nur BALTACIOĞLU

DENETİMLİ ÖĞRENME ALGORİTMALARI KULLANILARAK KARDİYOVASKÜLER KALP HASTALIĞIYLA İLGİLİ DENGESİZ VERİLERİN SINIFLANDIRILMASI

ÖZ

Kardiyovasküler hastalıklar, kalp ve kan damarlarının bir grup rahatsızlığıdır ve koroner kalp hastalığı, serebrovasküler hastalık, romatizmal kalp hastalığı ve diğer durumları içerir. Denetimli öğrenmenin sağlık sorunlarında umut verici çözümler sağladığı kanıtlanmıştır. Ayrıca tıbbi verilerin doğru yorumlanmasıyla hastalığın erken tahmin edilmesine yardımcı olurlar. Üstelik bu erken tahmin, hastalığın semptomlarının kontrol altına alınmasında ve hastalığın uygun tedavisinin yapılmasında yardımcı olabilir. Sınıflandırma modelleri geliştirilerek, kalp hastalıkları gibi kronik hastalıkların tahmininde denetimli öğrenme yaklaşımları kullanılabilir. Bu araştırmada, kardiyovasküler kalp hastalıklarını tahmin etmek için kapsamlı bir veri ön işleme yaklaşımı öneriyoruz. Yaklaşım özellikle kayıp değerlerin, gürültünün, standardizasyonun, dengesizliğin iyileştirilmesini ve bununla birlikte, sınıflandırma ve tahminleri içerir. Bu araştırma, Sınıflandırma Ağacı, Bagging, Rassal Ormanlar, Lojistik Regresyon, Destek Vektör Makineleri (Doğrusal - Radyal - Polinom), Xgboost, Naive Bayes ve K-En Yakın Komşu gibi denetimli öğrenme algoritmalarını kullanarak kardiyovasküler kalp hastalığı riskini tahmin etmeyi amaçlamaktadır. Ayrıca bu algoritmalar arasında tahmin doğruluğu bazında karşılaştırmalı bir çalışma yapılmıştır. Değerlendirme, R yazılımı ve ilgili denetimli öğrenme kitaplıkları kullanılarak gerçekleştirilmiştir.

Anahtar kelimeler: Denetimli öğrenme algoritmaları, veri ön işleme, dengesiz veri, kardiyovasküler kalp hastalığı

CLASSIFICATION OF IMBALANCED DATA ON CARDIOVASCULAR HEART DISEASE USING SUPERVISED LEARNING ALGORITHMS

ABSTRACT

Cardiovascular diseases are a group of disease of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Supervised learning algorithms has been proven to provide promising solutions in healthcare problems. This early prediction, moreover, can be helpful in controlling the symptoms of the disease as well as the proper treatment of disease. Supervised learning approaches can be used in the prediction of chronic diseases, such as heart diseases, by developing the classification models. In this research, we propose a data pre-processing extensive approach to predict cardiovascular heart diseases. The approach especially involves improving lost values, noise, standardization, imbalance and however classification and predictions. This research aims to predict the risk of cardiovascular heart diseases using supervised learning algorithms like Classification Tree, Bagging, Random Forest, Logistic Regression, Support Vector Machines (Linear - Radial - Polynomial), Xgboost, Naive Bayes, and K-Nearest Neighbours. Also, a comparative study among these algorithms on the basis of prediction accuracy is performed. The evaluation has been performed using R software and related supervised learning libraries.

Keywords: Supervised learning algorithms, data pre-processing, imbalance data, cardiovascular heart diseases

İÇİNDEKİLER

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

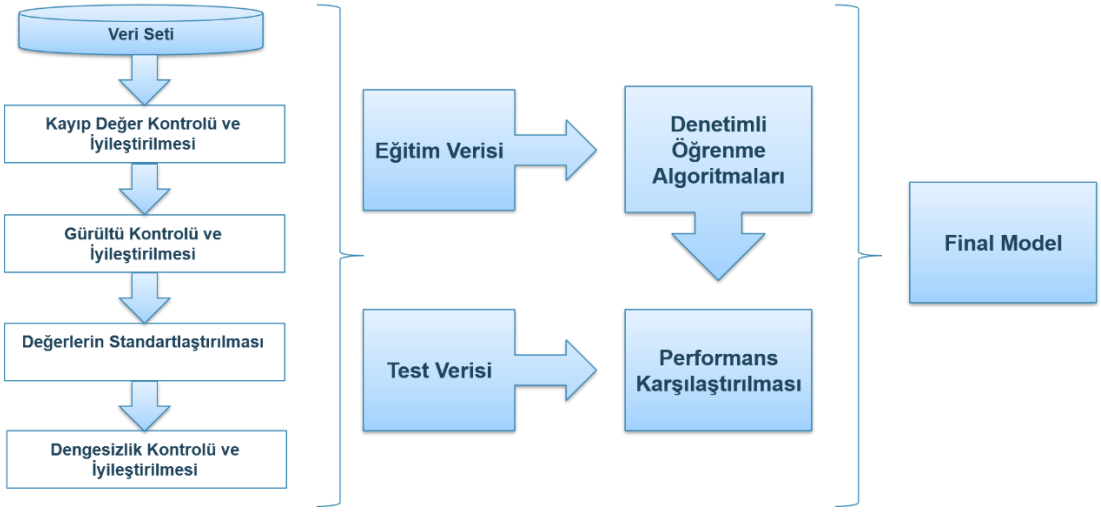
TABLOLAR LİSTESİ

Sayfa

BÖLÜM BİR
GİRİŞ

1.1 Amaç, Kapsam ve Hedefler

Dünya Sağlık Örgütü, kardiyovasküler hastalıkların dünya çapında bir numaralı ölüm nedeni olduğunu ve her yıl tahmini 17,9 milyon can aldığını bildirmiştir. Kardiyovasküler hastalıklar, kalp ve kan damarlarının bir grup rahatsızlığıdır ve koroner kalp hastalığı, serebrovasküler hastalık, romatizmal kalp hastalığı ve diğer durumları içerir. Kardiyovasküler ölümlerinin %80'i kalp krizi ve felçten kaynaklanmıştır. Bu ölümlerin üçte biri erken dönemde yani 70 yaşın altındaki insanlarda meydana gelmiştir. Kardiyovasküler hastalıkların erken prognozu, yüksek riskli hastalarda yaşam tarzı değişiklikleri konusunda karar vermede yardımcı olabilir ve dolayısıyla komplikasyonları azaltabilir (Dülek ve diğer., 2018). Bu araştırma, kardiyovasküler hastalığı tetikleyen faktörleri tespit etmeyi ve hastanın on yıl içinde koroner kalp hastası olup olmama riskini tahmin etmeyi amaçlamaktadır.



Şekil 1.1 Süreç tablosu

Araştırmada kullanılan veri seti için, Şekil 1.1'deki süreç uygulanmıştır. İlgili veri setinde listelenen kişilerin on yıl içinde kardiyovasküler kalp hastalığının varlığını tahmin etmek için sınıflandırma ağacı, bagging ile sınıflandırma ağacı, rassal ormanlar ile sınıflandırma ağacı, lojistik regresyon, destek vektör makineleri (doğrusal – radyal

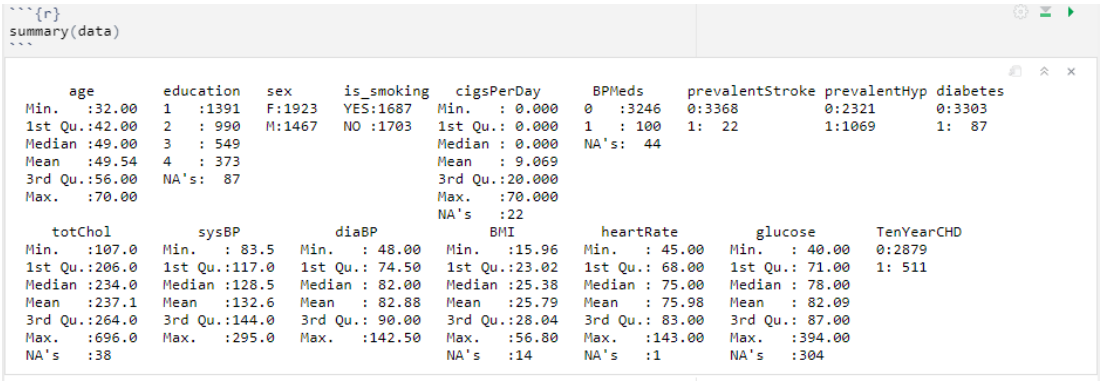
– polinom), xgboost, naive bayes, k – en yakın komşu algoritmaları ile modellenmiştir. Bu modellerin performansları ise F1 score, doğruluk (accuracy), AUC (area under curve), duyarlılık (sensitivity), özgüllük (specifity) ve ROC eğrisi gibi kriterler kullanılarak karşılaştırılmıştır.

1.2 Veri Seti Tanıtımı

Framingham kalp çalışması veri setinin bir alt kümesi olan bu veri seti üzerinde çalışılmıştır, çalışmada kullanılan veri setinin mevcut bölümü 3390 katılımcının kayıtlarını içermektedir. Veri seti, Framingham, Massachusetts'teki bir kitle üzerinde uzun süreli çalışma ile oluşturulmuştur. Çalışma, kardiyovasküler kalp hastalığına yol açan neden ve kökene dayanmaktadır (Krishnani ve diğer., 2019). Bu çalışma esas olarak, bir kişinin koroner kalp hastalığını algılamasında sağlığı üzerinde etkisi olan risk faktörlerini bulmaya odaklanmıştır. Veri setinde koroner kalp hastalığını etkileyen 16 değişken Tablo 1'de verilmektedir.

Attribute	Interpretation
Gender	0: Female ; 1: Male
Age	Age at the examination time
Education	1: high school ; 2: high school or GED ; 3: college or vocational school ; 4: college
CurrentSmoker	0: nonsmoker ; 1: smoker
Diabetes	0: No ; 1:Yes
TotChol	Total cholesterol inside patient's body (mg/dL)
SysBP	Systolic Blood Pressure (mmHg)
DiaBP	Diastolic Blood Pressure (mmHg)
CigsPerDay	Numer of cigarettes smoked per day (average)
BPMeds	Is the person on BP medicines
PrevalentStroke	If the person hadany prevalent stroke
PrevalentHyp	Any beneath prevalent
BMI	Body Mass Index: Weight (kg) / Height (meter-squared)
HeartRate	Beats / Min (Ventricular)
Glucose	Amount of glucose in mg/dL
TenYearCHD	Risk of developing CHD (0: No; 1:Yes)

Tablo 1 Veri seti değişken tanımları



```r
summary(data)
```

```
      age        education   sex      is_smoking   cigsPerDay      BPMeds     prevalentStroke prevalentHyp diabetes
 Min.   :32.00   1   :1391   F:1923   YES:1687   Min.   : 0.000   0   :3246   0:3368          0:2321       0:3303
 1st Qu.:42.00   2   : 990   M:1467   NO :1703   1st Qu.: 0.000   1   : 100   1:  22          1:1069       1:  87
 Median :49.00   3   : 549                      Median : 0.000   NA's:  44
 Mean   :49.54   4   : 373                      Mean   : 9.069
 3rd Qu.:56.00   NA's:  87                      3rd Qu.:20.000
 Max.   :70.00                                  Max.   :70.000
                                                NA's   :22
    totChol         sysBP          diaBP            BMI          heartRate         glucose       TenYearCHD
 Min.   :107.0   Min.   : 83.5   Min.   : 48.00   Min.   :15.96   Min.   : 45.00   Min.   : 40.00   0:2879
 1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 74.50   1st Qu.:23.02   1st Qu.: 68.00   1st Qu.: 71.00   1: 511
 Median :234.0   Median :128.5   Median : 82.00   Median :25.38   Median : 75.00   Median : 78.00
 Mean   :237.1   Mean   :132.6   Mean   : 82.88   Mean   :25.79   Mean   : 75.98   Mean   : 82.09
 3rd Qu.:264.0   3rd Qu.:144.0   3rd Qu.: 90.00   3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00
 Max.   :696.0   Max.   :295.0   Max.   :142.50   Max.   :56.80   Max.   :143.00   Max.   :394.00
 NA's   :38                                       NA's   :14      NA's   :1        NA's   :304
```

Şekil 1.2 Orijinal veri seti tanımlayıcı istatistikleri

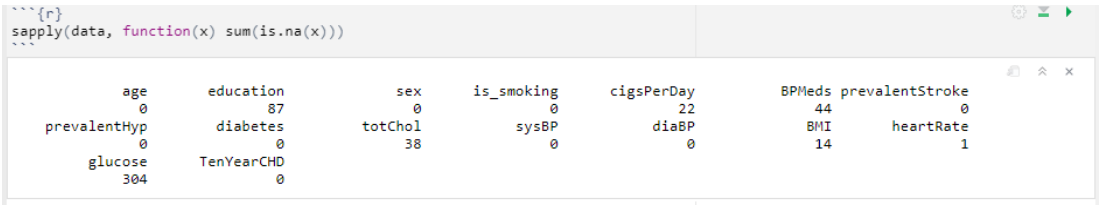
Bu değişkenlere ait tanımlayıcı istatistikler incelendiğinde kayıp gözlemlere sahip olduğu, değişkenlerin ölçeklerinin ve ortalamaların birbirinden faklı olduğu görülmektedir. Doğru analiz sonuçları elde edebilmek için model oluşturmadan önce veri setinde ön işlemeler yapılmalıdır.

BÖLÜM İKİ
VERİ ÖN İŞLEME

2.1 Kayıp Değer Kontrolü ve İyileştirilmesi

Büyük hacimli veri setlerinde kayıp değerlerin bulunması sıkça karşılaşılan bir durumdur. Veri toplama ya da veri kaydı sırasında veri setinde kayıp değerlere yol açan çeşitli nedenler olabilir. Veri setinde kayıp değerlerin tespit edilmesi, kayıp değerlere yol açan nedenlerin incelenmesi ve sorunun giderilmesi gerekir (Cebeci, 2020).

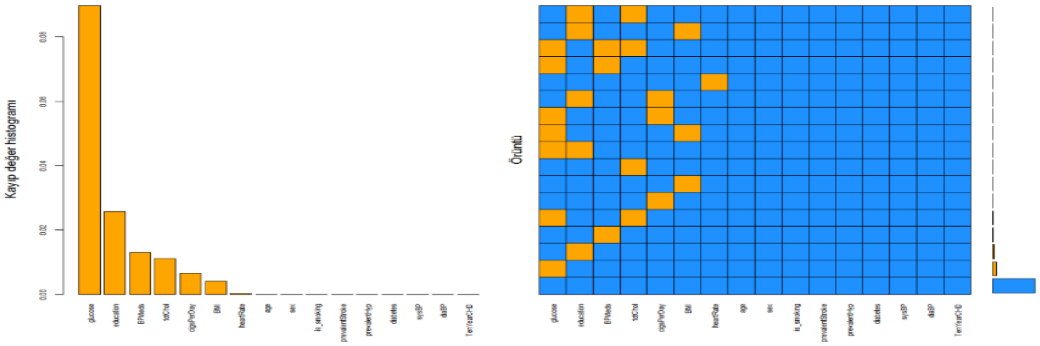
Kayıp değer saptama işlemleri için R yazılımının temel paketlerini kullanarak aşağıdaki fonksiyonla elde edilebilir.



Şekil 2.1 Orijinal veri setindeki kayıp gözlem sayıları

Yukarıdaki çıktı incelendiğinde en fazla kayıp değerin glucose değişkeninde olduğunu toplamda 7 değişkende kayıp değerler olduğunu görülmektedir.

Bu kayıp değerlerin oranlarını görselleştirmek ve daha sonra kayıp değerleri iyileştirmek için R yazılımının *VIM*, *mice*, *imputeTS* paketleri kullanılmıştır.



Şekil 2.2 Kayıp değer oranları

Şekil 2.2'in sol grafiği incelendiğinde, örneğin glucose değişkeninin yaklaşık %9'u, education değişkeninin yaklaşık %3'ü kayıp değer olarak görülmektedir. Sağ tarafta görülen grafikte ise değişkenler üzerinde çakışan kayıp değerlerin örüntüsü verilmektedir.

Kayıp değerleri tamamlamak için çeşitli yöntemler bulunmaktadır. Bunlar silme ve atama yöntemleridir. Bu çalışmada yapılan denemeler sonucunda, rassal ormanlar ile kayıp değer ataması uygulanmıştır.

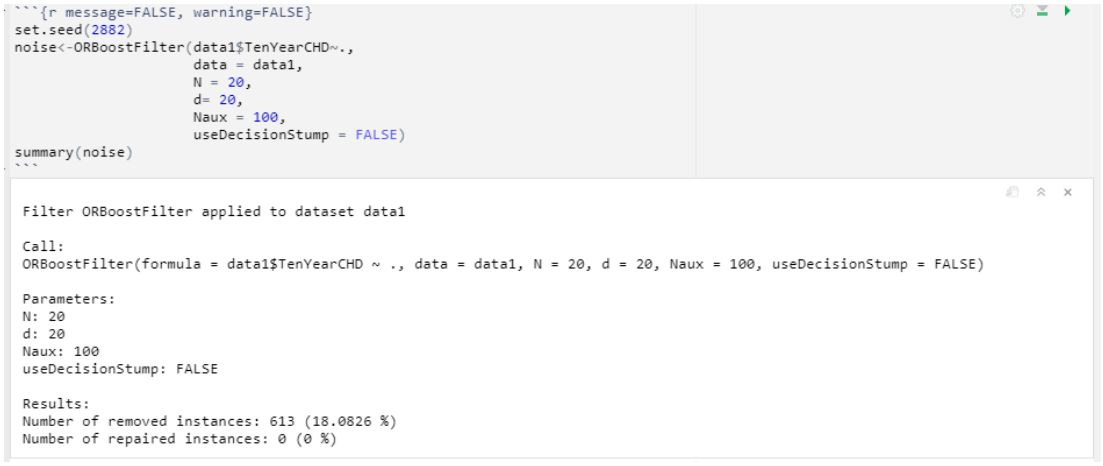
```r
set.seed(2882)
dfimp<-mice(data, m=5, meth="rf", maxit = 25)
data1<-complete(dfimp)
```

Şekil 2.3 Rassal ormanlar ile kayıp değer atama kodları

2.2 Gürültü Değer Kontrolü ve İyileştirilmesi

Gürültü, bir değişkenin değerleri arasında olmaması gerekirken istenmeyen bir şekilde veriye karışan ve değişkenin türü veya sınırlarına yabancı herhangi bir veriyi ifade eder. Aykırı değer kavramı ile sıkça karıştırılmaktadır. Aykırı değerler varlık olarak anlamlı iken gürültü değerleri anlamsızdır. Bu nedenle herhangi bir kullanışlılığı yoktur. Gereksiz yere veri depolama alanı işgal ettiğinden ve ayrıca veri madenciliği algoritmalarının yanlış sonuçlar üretmesine neden olduğundan tanınması ve veri setinden kaldırılması veya onarılması gerekir (Cebeci, 2020).

Bu çalışmada R yazılımı paketi olan *NoiseFiltersR* kullanılmıştır. Paket içinde çeşitli gürültü filtreleme algoritmaları bulunmakla birlikte, ORBoostFilter yöntemi uygulanmıştır. ORBoostFilter topluluk algoritmalarına dayanan gürültü değerlerini kaldırma yöntemidir.

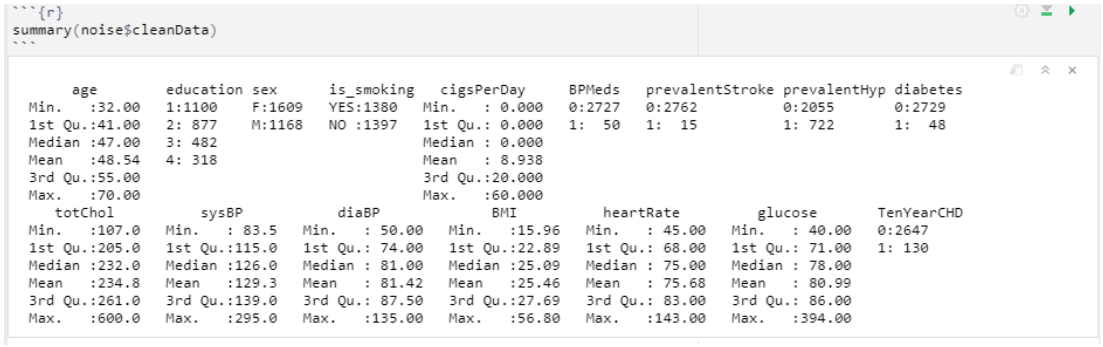


```{r message=FALSE, warning=FALSE}
set.seed(2882)
noise<-ORBoostFilter(data1$TenYearCHD~.,
                     data = data1,
                     N = 20,
                     d= 20,
                     Naux = 100,
                     useDecisionStump = FALSE)
summary(noise)
```

```
Filter ORBoostFilter applied to dataset data1

Call:
ORBoostFilter(formula = data1$TenYearCHD ~ ., data = data1, N = 20, d = 20, Naux = 100, useDecisionStump = FALSE)

Parameters:
N: 20
d: 20
Naux: 100
useDecisionStump: FALSE

Results:
Number of removed instances: 613 (18.0826 %)
Number of repaired instances: 0 (0 %)
```

Şekil 2.4 Gürültü değerlerinin saptanması

Çıktıdan da anlaşılacağı üzere ORBoostFilter algoritmasına göre 613 gözlem gürültü olarak değerlendirilmiş ve veri setinin %18,08'inin gürültü olduğu rapor edilmiştir.



```{r}
summary(noise$cleanData)
```

```
      age        education sex      is_smoking   cigsPerDay     BPMeds   prevalentStroke prevalentHyp diabetes
 Min.   :32.00   1:1100    F:1609   YES:1380   Min.   : 0.000   0:2727   0:2762          0:2055       0:2729
 1st Qu.:41.00   2: 877    M:1168   NO :1397   1st Qu.: 0.000   1:  50   1:  15          1: 722       1:  48
 Median :47.00   3: 482                        Median : 0.000
 Mean   :48.54   4: 318                        Mean   : 8.938
 3rd Qu.:55.00                                 3rd Qu.:20.000
 Max.   :70.00                                 Max.   :60.000
    totChol          sysBP           diaBP             BMI          heartRate         glucose       TenYearCHD
 Min.   :107.0   Min.   : 83.5   Min.   : 50.00   Min.   :15.96   Min.   : 45.00   Min.   : 40.00   0:2647
 1st Qu.:205.0   1st Qu.:115.0   1st Qu.: 74.00   1st Qu.:22.89   1st Qu.: 68.00   1st Qu.: 71.00   1: 130
 Median :232.0   Median :126.0   Median : 81.00   Median :25.09   Median : 75.00   Median : 78.00
 Mean   :234.8   Mean   :129.3   Mean   : 81.42   Mean   :25.46   Mean   : 75.68   Mean   : 80.99
 3rd Qu.:261.0   3rd Qu.:139.0   3rd Qu.: 87.50   3rd Qu.:27.69   3rd Qu.: 83.00   3rd Qu.: 86.00
 Max.   :600.0   Max.   :295.0   Max.   :135.00   Max.   :56.80   Max.   :143.00   Max.   :394.00
```

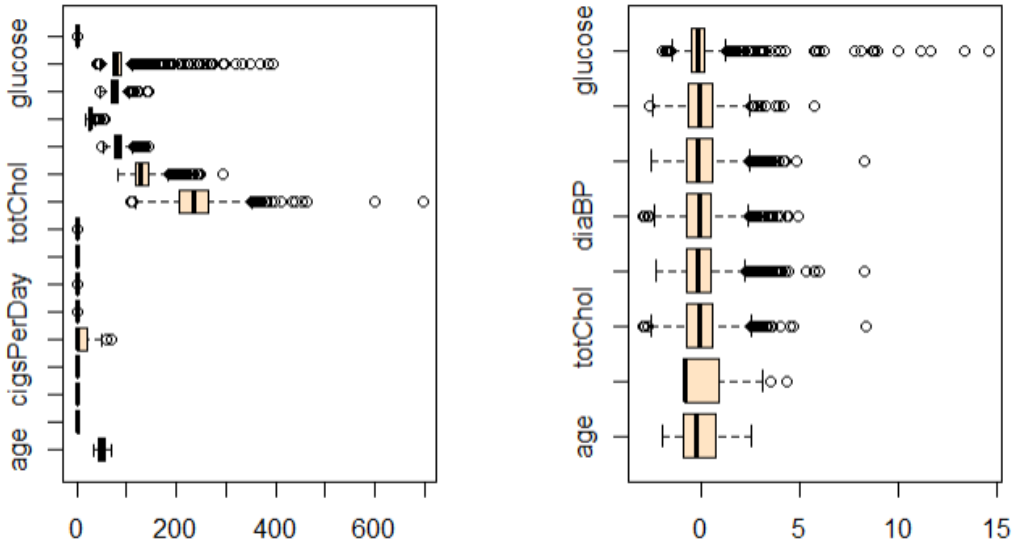
Şekil 2.5 Gürültüsüz veri setinin tanımlayıcı istatistikleri

Bu çalışmada gürültüsüz veri seti kullanılarak diğer işlemlere devam edilmiştir.

2.3 Değerlerin Standartlaştırılması

Veri seti, değişkenlerin özellikleri doğrultusunda farklı birimlerden oluşmaktadır. Bu durum nedeniyle değişkenler arasında karşılaştırma ve hesaplama yapılamamaktadır. Farklı birimlerde ölçülmüş değişkenleri 0 ortalamalı ve 1 varyanslı hale dönüştürerek değişkenleri birimsizleştirme işlemi bir standartlaştırma yöntemidir.

R yazılımının temel paketlerinden *base*'de bulunan *scale* fonksiyonu ile veri seti standartlaştırılmıştır.



Şekil 2.6 Orijinal veri seti ile standartlaştırılmış veri setinin kutu grafiği

2.4 Veri Seti Dengesizliği Kontrolü ve İyileştirilmesi

Sınıflandırma çalışmalarında dengesiz veri setleriyle karşılaşmak çok olağan bir durumdur. Bu durum sınıflandırma hatalarına, yanlış tahminlere ve dolayısıyla düşük doğruluk oranlarına sebep olmaktadır.

Dengesizlik, çarpık sınıf oranlarına sahip, genellikle iki kategorili hedef değişkenlerde görülmektedir. Hedef değişkenin büyük bir bölümünü oluşturan kategoriye çoğunluk sınıf, daha küçük bir oranını oluşturan kategoriye azınlık sınıf denmektedir (Google, 2020).

Azınlık / Çoğunluk Sınıfı Oranı	Dengesizlik Derecesi
$oran \leq \%1$	Aşırı
$\%1 < oran \leq \%20$	Orta
$\%20 < oran \leq \%40$	Hafif
$\%40 < oran$	Dengeli

Tablo 2 Dengesizlik derecesi

Bu çalışmada R yazılımında bulunan *imbalance* paketi kullanılarak hedef değişkenin dengesizlik derecesi Tablo 2'de incelenmiştir.

Şekil 2.7 Veri setinin dengesizlik oranı

Yukarıdaki çıktı incelendiğinde kalp hastası olma riski 130 kişide görülmekteyken olmama durumu 2647 kişide görülmektedir. Dengesizlik oranı yaklaşık 0,05 olarak bulunmuştur. Şekil 2.7'de bu değer veri setinin aşırı dengesiz olduğu göstermektedir. Bu sorun algoritmalarda hasta olmamayı daha iyi öğrenmelerine ve yanlılığa sebep olmaktadır bu yüzden hedef değişken dengelenmelidir.

Yetersiz örnekleme ve aşırı örnekleme dengesizlik sorunun çözümü için geliştirilen yöntemlerdir. Bu çalışmada aşırı örnekleme yöntemlerinden çoğunluk ağırlıklı azınlık aşırı örnekleme tekniği olan MWMOTE kullanılmıştır. Dengesizlik oranı %80 olarak ayarlanmıştır.

```r
set.seed(2882)
df_over<-oversample(df1, ratio = 0.8,
                    method = "MWMOTE",
                    filtering = FALSE,
                    classAttr = "TenYearCHD")
```

```r
table(df_over$TenYearCHD)
```

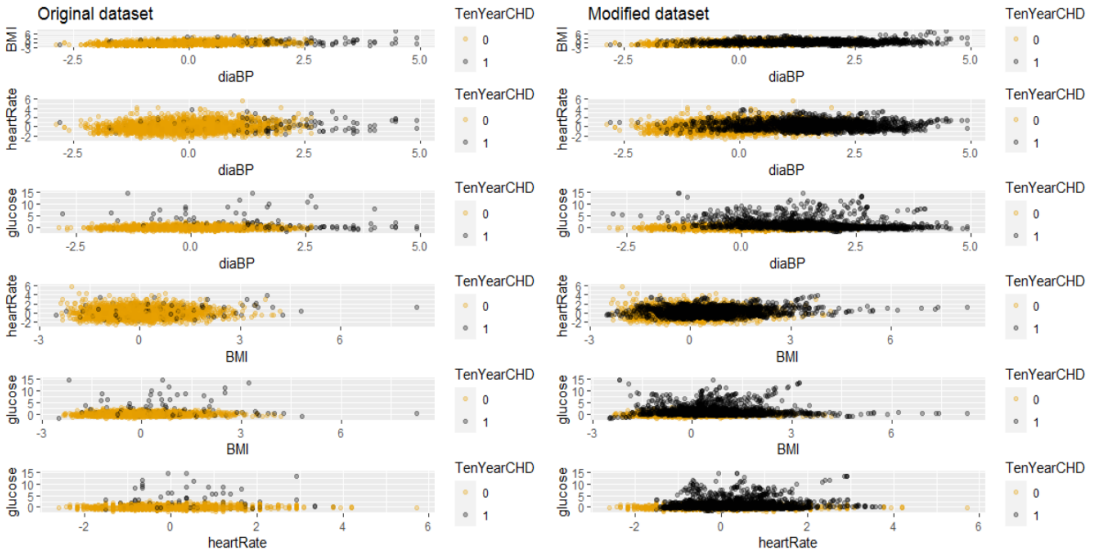
```
   0    1
2647 2118
```

```r
imbalanceRatio(df_over, classAttr = "TenYearCHD")
```

```
[1] 0.8001511
```

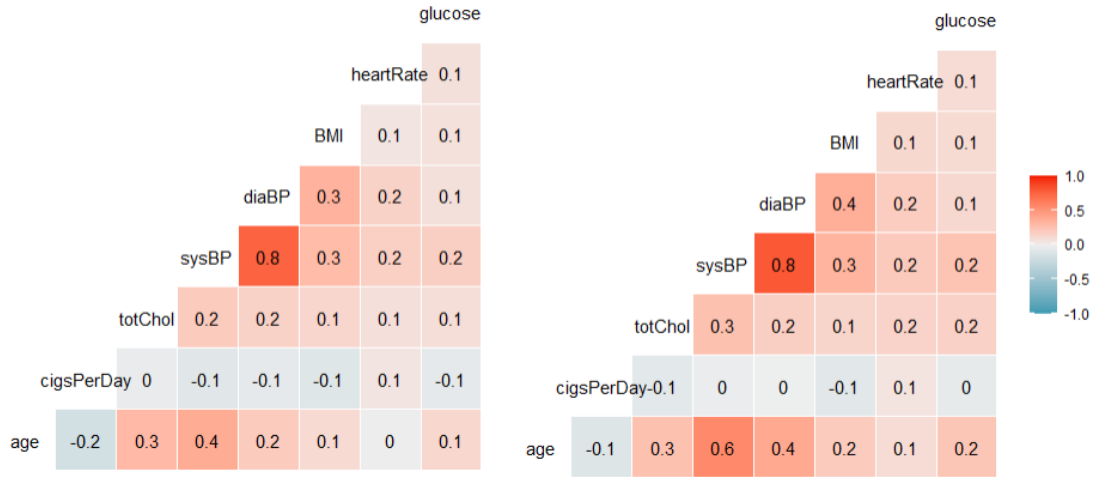
Şekil 2.8 Veri seti dengeleme kodları

Dengelenen hedef değişken 2647 kişinin kalp hastası olmama durumunu, 2118 kişinin de kalp hastası olma durumunu ifade eden yeni veri seti oluşturulmuştur.



Şekil 2.9 Orijinal veri seti ile dengeli veri seti grafiği

Şekil 2.9 incelendiğinde, orijinal veri setindeki azınlık sınıfı olan siyah değerler dengesizlik iyileştirilmesi yapıldıktan sonra çoğunluk sınıfıyla örtüşmüştür.



Şekil 2.10 Orijinal veri ile dengeli veriye ilişkin korelasyon grafikleri

Şekil 2.10'daki grafikler incelendiğinde dengelenen veri seti, orijinal verideki korelasyon yapısına uymaktadır. İki grafikte de sysBP ile diaBP arasında doğrusal pozitif yönlü yüksek bir ilişki bulunmaktadır.

2.5 Eğitim ve Test Verisine Ayırma

Veri ön işlemelerle elde edilen veri seti için eğitim ve test verisine, %70 ile %30'luk bir rassal seçimle ayrılmıştır.

```{r}
set.seed(2882)
train_df1<-sample(1:nrow(df2),(nrow(df2)*.7))
trainn<-df2[train_df1,]
testt<-df2[-train_df1,]
dim(trainn)
dim(testt)
```

```
[1] 3335   16
[1] 1430   16
```

Şekil 2.11 Eğitim ve test veri setlerine ayırma

Eğitim verisi 16 değişken ve 3335 gözlemden oluşmaktadır. Test verisi ise 16 değişken ve 1430 gözlemden oluşmaktadır.

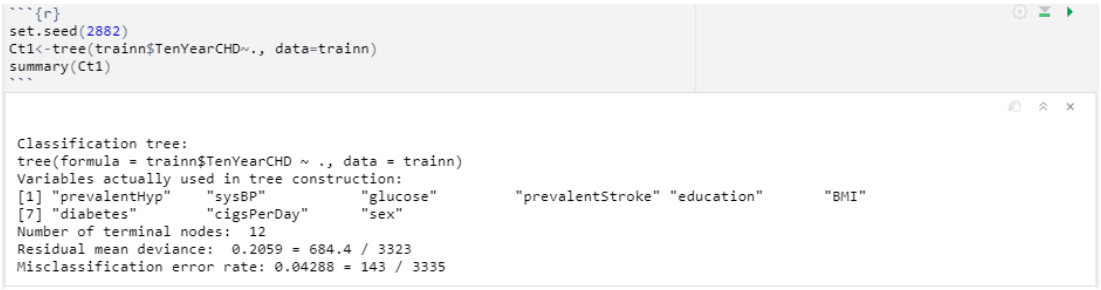
BÖLÜM ÜÇ
UYGULAMA

Bu çalışmada, on yılda kalp hastası olma durumunu Sınıflandırma Ağacı, Bagging, Rassal Ormanlar, Lojistik Regresyon, Destek Vektör Makineleri (Doğrusal - Radyal - Polinom), Xgboost, Naive Bayes ve K-En Yakın Komşu algoritmaları kullanılarak modellenmiştir.

3.1 Sınıflandırma Ağacı

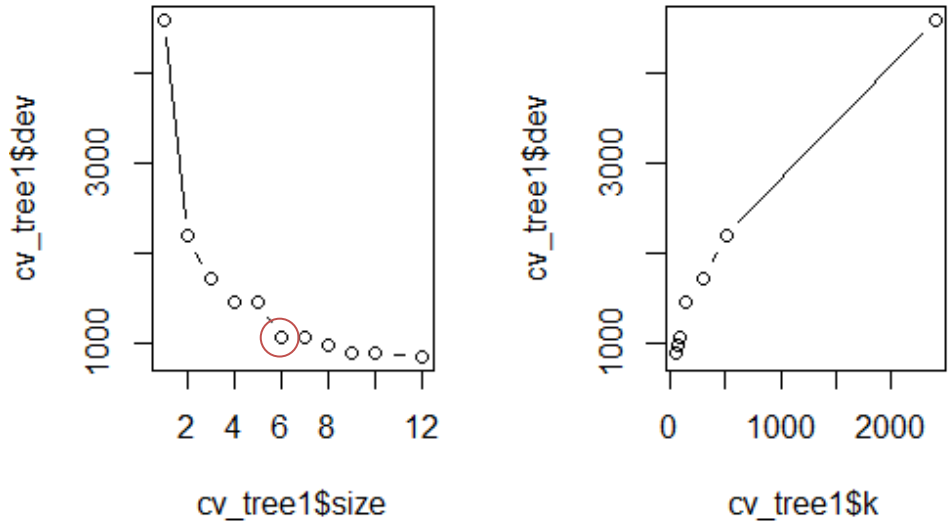
Sınıflandırma ağacı modeli temelde regresyon ağacı modeline çok benzemektedir. En büyük fark nicel bağımlı değişken yerine nitel bağımlı değişken kullanılmasıdır. Sınıflandırma ağacında her gözlem; ait olduğu bölgedeki eğitim gözlemlerine ait en yaygın olan sınıfa aittir (James ve diğer., 2013).

Bu çalışmada sınıflandırma ağacı modeli için R yazılımında bulunan *tree* paketi kullanılmıştır.



Şekil 3.1 Sınıflandırma ağacı model çıktısı

Sınıflandırma ağacı yöntemiyle elde edilen model budanmıştır. Terminal node sayısı aşağıdaki grafikteki dirsek noktası olarak belirlenerek en uygun model kurulmuştur.

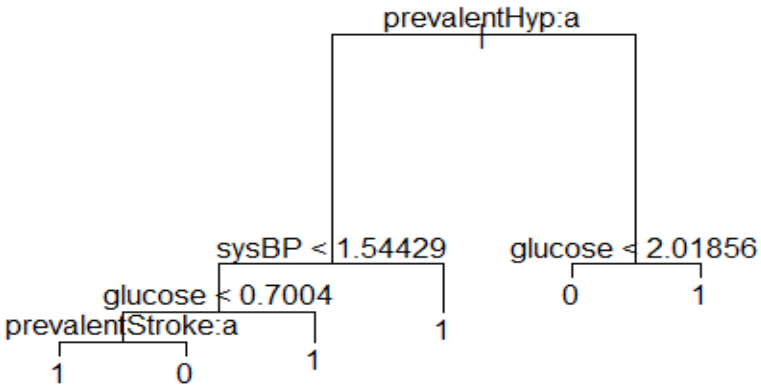


Şekil 3.2 Sınıflandırma ağacı terminal node – sapma grafiği

```r
prune_ctree1<-prune.tree(Ct1, best=6)
summary(prune_ctree1)
```

```
Classification tree:
snip.tree(tree = Ct1, nodes = c(17L, 9L, 5L))
Variables actually used in tree construction:
[1] "prevalentHyp"    "sysBP"           "glucose"         "prevalentStroke"
Number of terminal nodes:  6
Residual mean deviance:  0.3259 = 1085 / 3329
Misclassification error rate: 0.04858 = 162 / 3335
```

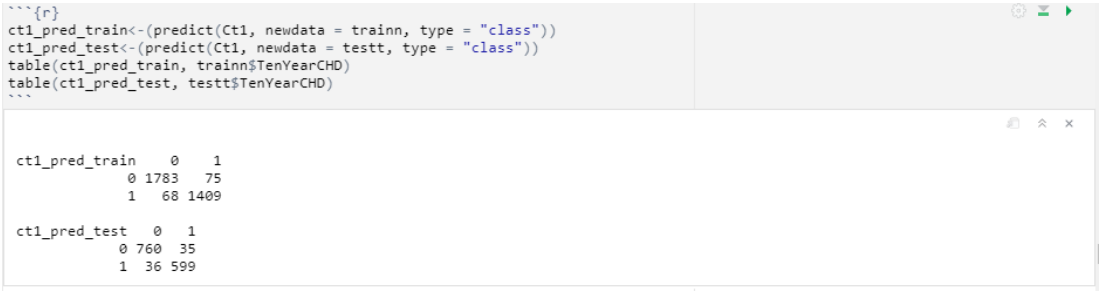
Şekil 3.3 Budanmış sınıflandırma ağacı model çıktısı



Şekil 3.4 Budanmış sınıflandırma ağacı

Budanmış sınıflandırma ağacı modeline göre on yılda kalp hastası olma riskini etkileyen faktörler prevalentHyp, sysBP, glucose, prevalentStroke olarak bulunmuştur.

Çapraz geçerlilik matrisinde eğitim veri seti için 3335 kişiden 3192 kişinin hastalık durumunu doğru tahminlenmiştir. Test veri seti için ise 1430 kişiden 1359 kişinin hastalık durumunu doğru tahminlenmiştir.



```{r}
ct1_pred_train<-(predict(Ct1, newdata = trainn, type = "class"))
ct1_pred_test<-(predict(Ct1, newdata = testt, type = "class"))
table(ct1_pred_train, trainn$TenYearCHD)
table(ct1_pred_test, testt$TenYearCHD)
```

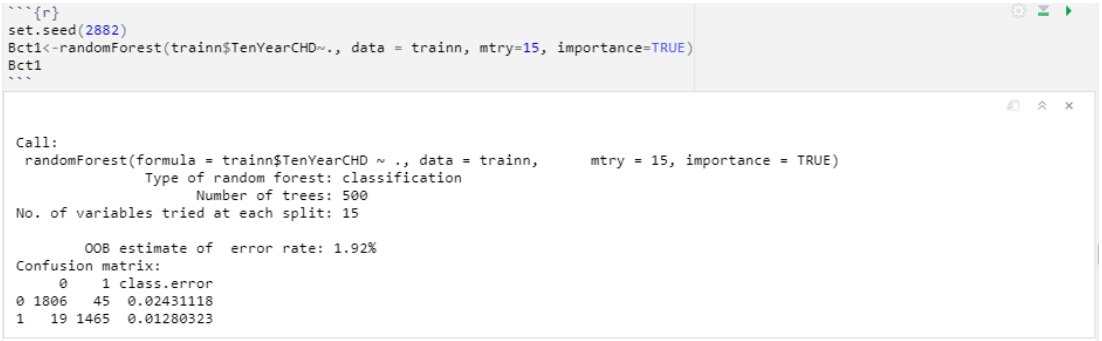
```
ct1_pred_train    0    1
             0 1783   75
             1   68 1409

ct1_pred_test   0   1
            0 760  35
            1  36 599
```

Şekil 3.5 Budanmış sınıflandırma ağacı çapraz geçerlilik matrisi

3.2 Bagging ile Sınıflandırma Ağacı

Bagging sınıflandırma ağacı yöntemi, orijinal veri setinden rastgele alt kümelere ayıran ve daha sonra nihai bir tahmin oluşturmak için bireysel tahminlerini toplayan bir topluluk meta tahmincisidir.

Bu çalışmada bagging modeli, R yazılımında bulunan *randomForest* paketiyle yapılmıştır.



```{r}
set.seed(2882)
Bct1<-randomForest(trainn$TenYearCHD~., data = trainn, mtry=15, importance=TRUE)
Bct1
```

```
Call:
 randomForest(formula = trainn$TenYearCHD ~ ., data = trainn,      mtry = 15, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 15

        OOB estimate of  error rate: 1.92%
Confusion matrix:
     0    1 class.error
0 1806   45  0.02431118
1   19 1465  0.01280323
```

Şekil 3.6 Bagging sınıflandırma ağacı model çıktısı

500 ağaç ve her ayrımda 15 değişken kullanılacak şekilde oluşturulan bagging modelinde hata oranı yani modelin hatalı karar verme oranını %1,92'dir.

Bagging ile sınıflama ağacı modeline göre on yılda kalp hastası olma riskini etkileyen faktörler sysBP, glucose, age ve prevalentStroke olarak bulunmuştur.

Çapraz geçerlilik matrisinde eğitim veri seti için 3335 kişinin hepsinin hastalık durumunu doğru tahminlenmiştir. Test veri seti için 1430 kişinin 1396'sının hastalık durumunu doğru tahminlenmiştir.

```{r}
bct1_pred_train<-predict(Bct1, newdata = trainn, type = "class")
bct1_pred_test<-predict(Bct1, newdata = testt, type = "class")
table(bct1_pred_train, trainn$TenYearCHD)
table(bct1_pred_test, testt$TenYearCHD)
```

```
bct1_pred_train    0    1
              0 1851    0
              1    0 1484

bct1_pred_test   0   1
             0 774  12
             1  22 622
```

Şekil 3.7 Bagging sınıflandırma ağacı çapraz geçerlilik matrisi

3.3 Rassal Ormanlar ile Sınıflandırma Ağacı

Rassal ormanlar algoritması, bağımsız bir karar ağaçları ormanı oluşturmak için hem bagging algoritmasını hem de değişkenlerin alt kümelerini oluşturarak modelleme yapar.

Bu çalışmada rassal ormanlar ile sınıflama modeli, R yazılımında bulunan *randomForest* paketiyle yapılmıştır. Alt küme sayısı, bağımsız değişken sayının karekökü olarak belirlenmiştir.

```{r}
set.seed(2882)
Rf1<-randomForest(trainn$TenYearCHD~., data = trainn, mtry = 4, importance = TRUE)
Rf1
```

```
Call:
 randomForest(formula = trainn$TenYearCHD ~ ., data = trainn,      mtry = 4, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 1.26%
Confusion matrix:
     0    1 class.error
0 1816   35 0.018908698
1    7 1477 0.004716981
```

Şekil 3.8 Rassal ormanlar ile sınıflandırma ağacı model çıktısı

500 ağaç ve her ayrımda 4 değişkenli alt kümeler kullanılacak şekilde oluşturulan modelde hata oranı %1,26'dır.

Rassal ormanlar ile sınıflama modeline göre on yılda kalp hastası olma durumu için risk faktörleri sysBP, glucose, age ve prevalentHyp olarak bulunmuştur.

Çapraz geçerlilik matrisine göre model, eğitim veri seti için 3335 kişinin hepsini doğru tahminlenmiştir. Test veri seti için 1430 kişinin 1404'ünü doğru tahminlenmiştir.

```{r}
rf1_pred_train<-predict(Rf1, newdata = trainn, type = "class")
rf1_pred_test<-predict(Rf1, newdata = testt, type = "class")
table(rf1_pred_train, trainn$TenYearCHD)
table(rf1_pred_test, testt$TenYearCHD)
```

```
rf1_pred_train    0    1
             0 1851    0
             1    0 1484

rf1_pred_test   0   1
            0 776   6
            1  20 628
```

Şekil 3.9 Rassal ormanlar ile sınıflandırma ağacı çapraz geçerlilik matrisi

3.4 Lojistik Regresyon

Lojistik regresyon, bir sınıflama algoritmasıdır. Bir dizi bağımsız değişkene dayalı ikili bir sonucu tahmin etmek için kullanılan bir yöntemdir.

Bu çalışmada lojistik regresyon modelini kurmak için *kknn*, değişken seçimi için *mass* paketleri kullanılmıştır.

Full modelle lojistik regresyon modeli kurulduktan sonra değişken seçimi yöntemi uygulanmıştır. AIC değeri en düşük model seçilip yeni model kurulmuştur.

```r
summary(lr1)
```

```
Call:
glm(formula = trainn$TenYearCHD ~ age + education + sex + is_smoking +
    cigsPerDay + BPMeds + prevalentStroke + prevalentHyp + diabetes +
    totChol + sysBP + diaBP + BMI + glucose, family = "binomial",
    data = trainn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9540  -0.0068  -0.0002   0.0156   4.2258

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       7.2052     1.7729   4.064 4.82e-05 ***
age               1.2446     0.2091   5.952 2.64e-09 ***
education2       -1.4892     0.4400  -3.385 0.000713 ***
education3       -0.3228     0.5467  -0.590 0.554928
education4        2.0274     0.3682   5.506 3.68e-08 ***
sex1             -1.7276     0.3375  -5.119 3.08e-07 ***
is_smoking1      -1.0441     0.4709  -2.217 0.026624 *
cigsPerDay        1.6597     0.2495   6.653 2.87e-11 ***
BPMeds1          -3.1423     0.5552  -5.660 1.51e-08 ***
prevalentStroke1 -6.2799     1.3844  -4.536 5.73e-06 ***
prevalentHyp1    -2.6946     0.6007  -4.486 7.26e-06 ***
diabetes1        -4.1906     1.0113  -4.144 3.42e-05 ***
totChol           0.4381     0.1337   3.276 0.001054 **
sysBP             2.5755     0.2495  10.324  < 2e-16 ***
diaBP             1.1718     0.2002   5.853 4.82e-09 ***
BMI               0.6804     0.1400   4.861 1.17e-06 ***
glucose           2.4947     0.2390  10.437  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4582.8  on 3334  degrees of freedom
Residual deviance:  345.8  on 3318  degrees of freedom
AIC: 379.8

Number of Fisher Scoring iterations: 10
```

Şekil 3.10 Lojistik regresyon model çıktısı

Kurulan model tahminlerinin kesim noktasını en yüksek doğruluk verecek döngü yazılmış ve 0,73 değerinin kesim noktası olmasına karar verilmiştir.

Çapraz geçerlilik matrisine göre model eğitim veri setinde 3335 kişiden 3280 kişiyi kalp hastası olma durumuna göre doğru tahminlenmiştir. Test veri seti için 1430 kişinin 1402’sini doğru tahminlenmiştir.

```r
lr1_pred_train<-ifelse(lr1_pred_train<=0.73,0,1)
lr1_pred_test<-ifelse(lr1_pred_test<=0.73,0,1)
table(lr1_pred_train, trainn$TenYearCHD)
table(lr1_pred_test, testt$TenYearCHD)
```

```
lr1_pred_train    0    1
             0 1843   47
             1    8 1437

lr1_pred_test   0   1
            0 785  17
            1  11 617
```

Şekil 3.11 Lojistik regresyon çapraz geçerlilik matrisi

3.5 Doğrusal Destek Vektör Makinesi

Destek vektör makinesi, eğitim verilerindeki herhangi bir noktadan en uzak olan iki sınıf arasında bir karar sınırı bulan vektör uzayı tabanlı makine öğrenme yöntemi olarak tanımlanır (Schölopf ve diğer. 2002).

Bu çalışmada destek vektör makinesi modelleri için, R yazılımında bulunan *e1071* ve *SparseM* paketleri kullanılmıştır.

Hiper parametre belirlenmesi için yapılan denemelerde en yüksek doğruluk verecek cost değeri 2 ve gamma değeri 0,3 olarak tanımlanmıştır.

```r
svm1_linear<-svm(formula = trainn$TenYearCHD~.,
                 data = trainn,
                 type = "C-classification",
                 kernel = "linear",
                 cost = 2,
                 gamma = 0.3,
                 scale = TRUE)

summary(svm1_linear)
```

```
Call:
svm(formula = trainn$TenYearCHD ~ ., data = trainn, type = "C-classification", kernel = "linear",
    cost = 2, gamma = 0.3, scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  2

Number of Support Vectors:  183

 ( 89 94 )


Number of Classes:  2

Levels:
 0 1
```

Şekil 3.12 Doğrusal destek vektör makinesi model çıktısı

Çapraz geçerlilik matrisine göre eğitim veri seti için 3335 kişiden 3281 kişinin kalp hastası olma durumu doğru tahmin edilmiştir. Test verisi için 1430 kişinin 1399'u doğru tahminlenmiştir.

```r
svm1_lin_pred_train<-predict(svm1_linear, trainn[,-16])
table(trainn$TenYearCHD, svm1_lin_pred_train)

svm1_lin_pred_test<-predict(svm1_linear, testt[,-16])
table(testt$TenYearCHD, svm1_lin_pred_test)
```

```
   svm1_lin_pred_train
      0    1
  0 1827   24
  1   30 1454
   svm1_lin_pred_test
      0   1
  0 777  19
  1  12 622
```

Şekil 3.13 Doğrusal destek vektör makinesi çapraz geçerlilik matrisi

3.6 Radyal Destek Vektör Makinesi

Her bir noktanın belirli bir noktaya ne kadar benzediğini normal dağılım ile hesaplar, ona göre sınıflandırır. Dağılımın genişliği gamma hiper parametresi ile kontrol edilir. Yapılan denemeler sonucu en uygun gamma parametresi 0,3 olarak bulunmuştur.

```r
set.seed(2882)
svm1_radial<-svm(trainn$TenYearCHD~.,
                 data = trainn,
                 kernel = "radial",
                 type ="C-classification",
                 gamma = 0.3,
                 cost =2)
summary(svm1_radial)
```

```
Call:
svm(formula = trainn$TenYearCHD ~ ., data = trainn, kernel = "radial", type = "C-classification",
    gamma = 0.3, cost = 2)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  2

Number of Support Vectors:  714

 ( 334 380 )


Number of Classes:  2

Levels:
 0 1
```

Şekil 3.14 Radyal destek vektör makinesi model çıktısı

Çapraz geçerlilik matrisine göre eğitim veri seti 3335 kişinin hepsini doğru tahminlenmiştir. Test veri seti için ise 1430 kişinin 1419'unu doğru tahminlenmiştir.

```r
svm1_rad_pred_train<-predict(svm1_radial, trainn[,-16])
table(trainn$TenYearCHD, svm1_rad_pred_train)

svm1_rad_pred_test<-predict(svm1_radial, testt[,-16])
table(testt$TenYearCHD, svm1_rad_pred_test)
```

```
   svm1_rad_pred_train
      0    1
  0 1851    0
  1    0 1484
   svm1_rad_pred_test
      0   1
  0 787   9
  1   2 632
```

Şekil 3.15 Radyal destek vektör makinesi çapraz geçerlilik matrisi

3.7 Polinom Destek Vektör Makinesi

Doğrusal olmayan destek vektör makine modellerinden bir diğeri de polinom modeldir.

Yapılan çalışmalar sonucunda en uygun hiper parametrelerle model kurulmuştur.

```r
set.seed(2882)
svm1_poly<-svm(trainn$TenYearCHD~.,
               data = trainn,
               kernel = "polynomial",
               type ="C-classification",
               gamma = 0.3,
               coef0 = 0.3,
               degree = 3,
               cost = 2)
summary(svm1_poly)
```

```
Call:
svm(formula = trainn$TenYearCHD ~ ., data = trainn, kernel = "polynomial", type = "C-classification",
    gamma = 0.3, coef0 = 0.3, degree = 3, cost = 2)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  2
     degree:  3
     coef.0:  0.3

Number of Support Vectors:  136

 ( 66 70 )


Number of Classes:  2

Levels:
 0 1
```

Şekil 3.16 Polinom destek vektör makinesi model çıktısı

Çapraz geçerlilik matrisine göre eğitim veri seti 3335 kişinin hepsini doğru tahminlenmiştir. Test veri seti 1430 kişiden 1418 kişiyi doğru tahminlenmiştir.

```r
svm1_poly_pred_train<-predict(svm1_poly, trainn[,-16])
table(trainn$TenYearCHD, svm1_poly_pred_train)

svm1_poly_pred_test<-predict(svm1_poly, testt[,-16])
table(testt$TenYearCHD, svm1_poly_pred_test)
```

```
   svm1_poly_pred_train
      0    1
  0 1851    0
  1    0 1484
   svm1_poly_pred_test
     0   1
  0 788   8
  1   4 630
```

Şekil 3.17 Polinom destek vektör makinesi çapraz geçerlilik matrisi

3.8 Xgboost

Xgboost, hız ve performans için tasarlanmış gradyan destekli karar ağaçlarından oluşan bir algoritmadır.

Yapılandırılmış veya tablo halindeki veriler için uygulanır. Bu yüzden eğitim ve test veri setleri one-hot-coding dönüşümü yapılmıştır.

Bu çalışmada xgboost modeli, *xgboost*, *caret* ve *mlr* paketleri kullanılmıştır.

```r
labels1<-trainn$TenYearCHD
ts_label1<-testt$TenYearCHD

new_tr1<-model.matrix(~.+0, data = trainn[,-16])

new_ts1<-model.matrix(~.+0, data = testt[,-16])

labels1<-as.numeric(labels1)-1

ts_label1<-as.numeric(ts_label1)-1
```

```r
dtrain1<-xgb.DMatrix(data = new_tr1, label = labels1)

dtest1<-xgb.DMatrix(data = new_ts1, label = ts_label1)
```

Şekil 3.18 One – hot – coding dönüşümü

Dönüşüm yapılan veri seti için uygun parametreleri belirlenip model kurulmuştur.

```r
set.seed(2882)
params<-list(booster = "gbtree",
             objective = "binary:logistic",
             eta = 0.3,
             gamma = 0,
             max_depth = 6,
             min_child_weight = 1,
             subsample = 1,
             colsample_bytree = 1)
```

```r
set.seed(2882)
xgbcv1<-xgb.cv(params = params, data = dtrain1,
               nrounds = 100, nfold = 5, showsd = T,
               stratified = T, print_every_n = 10,
               early_stopping_rounds = 20, maximize = F)
```

```r
xgbcv1$best_iteration
```

```
[1] 89
```

```r
xgb1<-xgb.train(params = params, data = dtrain1,
                nrounds = 89,
                watchlist = list(val = dtest1, train = dtrain1),
                print_every_n = 10, early_stopping_rounds = 10,
                maximize = F, eval_metric = "error")
```

Şekil 3.19 Xgboost model çıktısı

Kesim noktası yapılan döngüler sonucu en yüksek doğruluk veren noktanın 0,5 olduğu belirlenmiş ve tahminler buna göre yapılmıştır.

Xgboost modeline göre on yılda kalp hastası olma durumu için risk faktörleri sysBP, prevalentHyp, glucose ve diabetes olarak bulunmuştur.

Çapraz geçerlilik matrisine göre eğitim veri seti 3335 kişinin hepsini doğru tahminlenmiştir. Test veri seti ise 1430 kişiden 1406 kişiyi doğru tahminlenmiştir.

```r
xgb1_pred_train<-ifelse(xgb1_pred_train<=0.5,0,1)
xgb1_pred_test<-ifelse(xgb1_pred_test<=0.5,0,1)

table(xgb1_pred_train, trainn$TenYearCHD)
table(xgb1_pred_test, testt$TenYearCHD)
```

```
 xgb1_pred_train    0    1
               0 1851    0
               1    0 1484

 xgb1_pred_test   0   1
              0 782  10
              1  14 624
```

Şekil 3.20 Xgboost çapraz geçerlilik matrisi

3.9 Naïve Bayes

Naive Bayes modeli tahmincileri arasında bağımsızlık varsayımı ile Bayes Teoremi'ne dayanan bir sınıflandırma tekniğidir. Basit bir ifadeyle, Naive Bayes sınıflandırıcısı, bir sınıftaki belirli bir özelliğin varlığının başka herhangi bir özelliğin varlığıyla ilgisi olmadığını varsayar (Analyticsvidhya, 2017).

Bu çalışmada Naive Bayes modeli için *e1071* ve *klaR* paketleri kullanılmıştır.

Veri setinde sysBP ile diaBP değişkenleri arasında doğrusal pozitif yönlü bir ilişki bulunmaktadır. Naïve Bayes varsayımını sağlamak için diğer modellerde daha az risk faktörü olarak tanımlanan diaBP modele dahil edilmemiştir.

```
```{r message=FALSE, warning=FALSE}
set.seed(2882)
nb1<-train(trainn[,-c(12,16)], trainn[,16], "nb",
          trControl=trainControl(method = "cv", number = 10))
nb1
```
```

```
Naive Bayes

3335 samples
  14 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3002, 3002, 3002, 3002, 3001, 3002, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.9616154  0.9228167
   TRUE      0.9580199  0.9156030

Tuning parameter 'fL' was held constant at a value of 0
Tuning parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = FALSE and adjust = 1.
```

Şekil 3.21 Naïve Bayes model çıktısı

Çapraz geçerlilik matrisine göre eğitim veri seti 3335 kişiden 3209 kişiyi, test veri seti için 1430 kişinin 1380'ini doğru tahmin etmiştir.

```{r message=FALSE, warning=FALSE}
nb1_pred_train<-predict(nb1, newdata = trainn, type = "raw")
nb1_pred_test<-predict(nb1, newdata = testt, type = "raw")
table(nb1_pred_train, trainn$TenYearCHD)
table(nb1_pred_test, testt$TenYearCHD)
```

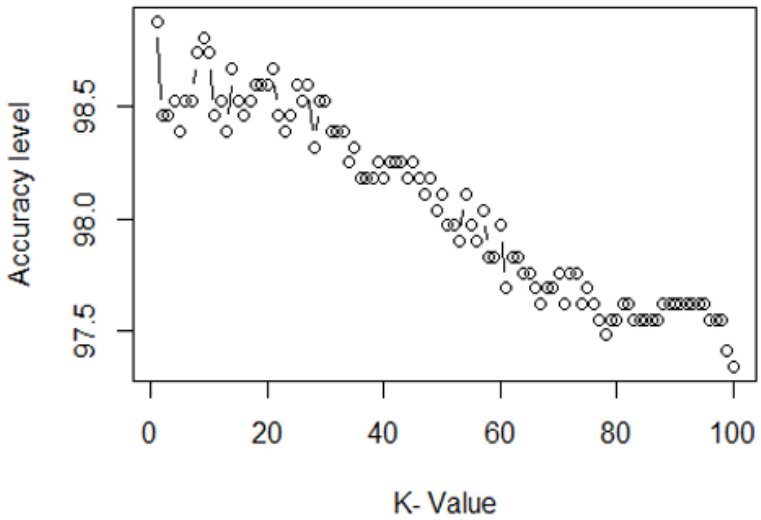
```
 nb1_pred_train    0    1
              0 1741   16
              1  110 1468

 nb1_pred_test   0   1
               0 750   4
               1  46 630
```

Şekil 3.22 Naïve Bayes çapraz geçerlilik matrisi

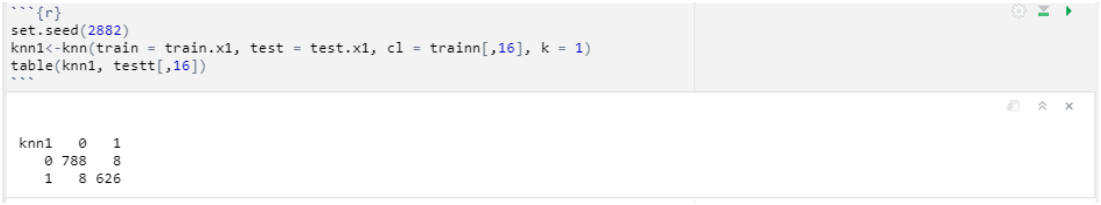
3.10 K En Yakın Komşu Algoritması

K en yakın komşu algoritması, mevcut tüm durumları saklayan ve yeni durumları bir benzerlik ölçüsüne göre sınıflandıran basit bir algoritmadır.

Bu çalışmada k en yakın komşu algoritması için *class* paketi kullanılmıştır. Doğruluk oranını en yüksek verecek k değeri, döngüler sayesinde elde edilmiş. Şekil 3.23 incelendiğinde 1 olarak belirlenmiştir.



Şekil 3.23 k belirleme grafiği



Şekil 3.24 K en yakın komşu model ve çapraz geçerlilik matrisi

Yukarıdaki çapraz geçerlilik matrisine göre model test verisi için 1430 kişiden 1414'ünü kalp hastalığı durumu için doğru tahminlenmiştir.

BÖLÜM DÖRT
SONUÇ

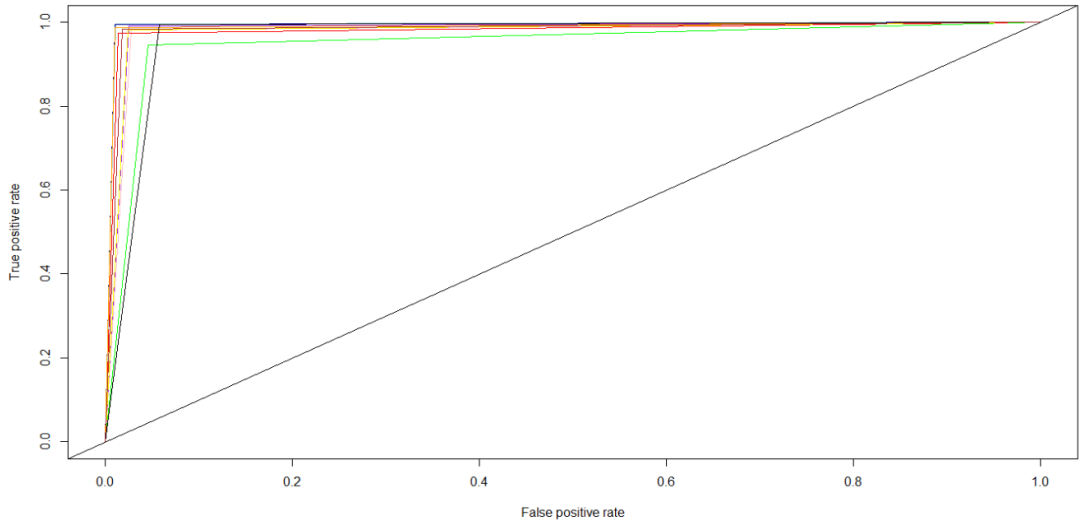
Uygulanan tüm yöntemlerin model metrikleri R yazılımında bulunan *ModelMetrics* ve *ROCR* paketleriyle hesaplanmıştır. Bu çalışmada F1-score, duyarlılık (sensitivity), özgüllük (specifity), doğruluk (accuracy) ve roc eğrisi altında kalan alan (area under curve) metrikleri ile karşılaştırma yapılmıştır.

Tablo 3'de test veri setinin performans değerleri karşılaştırılmıştır. Doğruluk değerlerinin oldukça yüksek olduğu görülmektedir. Destek vektör makinelerinden radyal ve polinom bununla birlikte k en yakın komşu algoritması yaklaşık %99 doğruluk oranına sahiptirler.

	Accurancy	F1 Score	Sensitivity	Specifity	AUC
Svm Radyal	0.9923	1.0000	0.9887	0.9968	0.9928
Svm Polinom	0.9916	0.0084	0.9899	0.9937	0.9918
KNN	0.9888	1.0000	0.9899	0.9874	0.9887
Xgboost	0.9832	0.9811	0.9824	0.9842	0.9833
Rassal Ormanlar	0.9818	1.0000	0.9749	0.9905	0.9827
Lojistik Regresyon	0.9804	0.0273	0.9862	0.9732	0.9797
Svm Doğrusal	0.9783	1.0000	0.9761	0.9811	0.9786
Bagging	0.9762	1.0000	0.9724	0.9811	0.9767
Naive Bayes	0.9650	1.0000	0.9422	0.9937	0.9679
Sınıflandırma Ağacı	0.9503	1.0000	0.9548	0.9448	0.9498

Tablo 3 Test verisi için model metrikleri

Duyarlılık, özgüllük, doğruluk ve eğri altında kalan alan metriklerinde hiçbir model %94'ün altına düşmemiştir.



Şekil 4.1 Test verisi roc eğrisi

Şekil 4.1 incelendiğinde tüm modeller oldukça güçlü tahmin doğruluğuna sahiptirler. Optimal modelin destek vektör makine algoritmalarından olan radyal modeli eğitim verisinde %100'lük, test verisinde %99,3'lük tahmin doğruluğuna sahiptir. En düşük tahmin oranına sahip olan sınıflandırma ağacı modeli bile %95'lik doğruluk oranına sahiptir.

Sonuçlar gösteriyor ki bu modeller kalp hastalığı riskini verimli bir şekilde tahmin edebilmektedir. Ayrıca ön işlemenin daha da iyileştirilmesine yapılan vurgu, doğru sonuçlar vermiştir.

Kardiyovasküler kalp hastası olma durumunun risk faktörleri, sysBP, prevalentHyp, glucose ve prevalentStroke olarak belirlenmiştir. Kalp hastalığı riskini azaltmak için önerilen yaşam tarzı; sağlıklı beslenmek, düzenli spor yapmak ve stresten uzak durmaktır.

Yapılan bu çalışma sayesinde erken teşhis, hastalığın semptomlarının kontrol altına alınması ve yaşam tarzı değişikliği kalp hastalığının önlenmesine yardımcı olabilir.

KAYNAKLAR

Analyticsvidhya. (2017). *6 easy steps to learn naive bayes algorithm with codes in python and R*. 26 Haziran 2021, https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Cebeci, Z., (2020). *Data Preprocessing with R*. Ankara: Nobel Akademik Yayıncılık

Dülek, H., Vural, Z. T., ve Gönenç, I. (2018). Kardiyovasküler hastalıklarda risk faktörleri. *The Journal of Turkish Family Physician, 9*(2), 53-58.

Google. (2020). *Imbalanced data*. 25 Mayıs 2021, https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

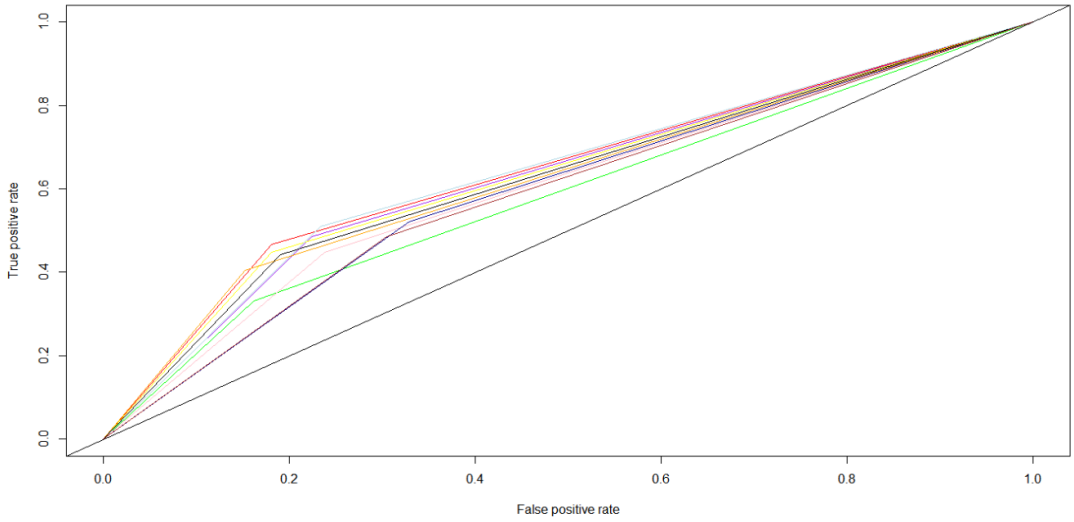
James, G., Witten, D., Hastie, T., ve Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Krishnani, D., Kumari, A., Dewangan, A., Singh, A., ve Naik, N. S. (2019). Prediction of coronary heart disease using supervised machine learning algorithms. *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* 367-372.

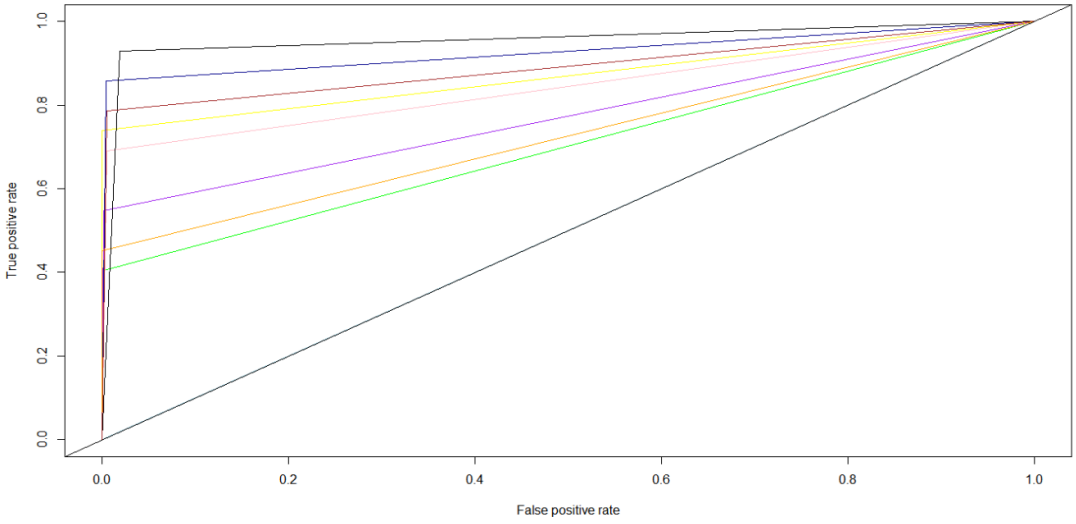
Schölkopf, B., Smola, A. J. ve Bach, F. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. London: MIT Press.

EKLER

Veri Ön İşleme Katkıları



Şekil A.1 Kayıp değer iyileştirilmesi yapılmış ve standartlaştırılmış modelin roc eğrisi



Şekil A.2 Kayıp değer ve gürültü iyileştirilmesi yapılmış, standartlaştırılmış modelin roc eğrisi