T.C.

BİLECİK ŞEYH EDEBALİ ÜNİVERSİTESİ

İKTİSADİ VE İDARİ BİLİMLER FAKÜTESİ

YÖNETİM BİLİŞİM SİSTEMLERİ BÖLÜMÜ

2021-2022 Öğretim Yılı Güz Dönemi

ÖĞRENCİNİN

AD : BURAK

SOYAD : ACAROĞLU

NO : 13238801400

DERS : VERİ MADENCİLİĞİ

Dr. Öğr. Üyesi Nur Kuban TORUN

Veri Madenciliği Nedir ?

Veri madenciliği, verinin ayıklaması ya da yığın bir veri içerisinden madencilik yapılarak anlamlı ve işe yarar bilginin çekilmesi aşamalarına verilen isimdir. Veri madenciliği aslında eski zamanlarda toprak içerisinden altın arayan kişilere yapılan bir göndermedir. İnsanlar ellerinde bulunan elekler ile büyük parça toprağı alarak, içerisinden altın aramaktadırlar. Veri madenciliği de aslında tam olarak budur. Anlamlı bir bilgi elde edebilmek amacıyla bir çok veri taranmakta ve analiz edilmektedir. Veri madenciliğinin bu açıdan bakıldığında ismi bilgi arayışında veri madenciliği ya da bilgi madenciliği olarak da kullanılabilmektedir. Ancak bir çok araştırmacı tarafından kullanılan ismi veri madenciliğidir (Jiawei ve Kamber, 2006).

Cabena (vd., 1998) veri madenciliğini, büyük veri kaynaklarından anlamlı bilgi elde etmek için algoritmaları, istatistiği ve görselliği kullanan disiplinlerarası alan olarak tanımlamıştır. Hand (vd., 2001) ise veri madenciliğini gözlenebilir veri setlerinin, anlaşılır ve yararlı bilgiye ihtiyaç duyanlar için analiz edilmesi ve elde edilen bilgilerin raporlanması olarak tanımlamaktadır. Wang ve Weigeng'e (2004) göre veri madenciliği yığın bir veri seti içerisinden gizli ve karmaşık ilişkilerin modern istatistik, akıllı bilgi sistemleri, makine öğrenmesi, örüntü tanıma, karar teorileri, veri mühendisliği ve veri bankası yönetimini birleştirerek çıkarılması olarak tanımlanmaktadır. Ayrıca veri madenciliği, otomatik veya yarı-otomatik biçimlerde verinin analiz edilerek gizli örüntülerin bulunması olarak tanımlamaktadır (Tang ve MacLennan, 2005; Witten ve Frank 2005). Keşfedilen örüntüler anlamlı olmalı ve ekonomik olarak araştırmacıya bir yararı dokunmalıdır. Tan, (vd., 2006), veri madenciliğini, büyük veri kaynaklarında yer alan yararlı bilgilerin otomatik olarak keşfedilmesi süreci olarak tanımlamaktadır. Gartner Group (2007), veri madenciliğini, veri ambarlarında depolanan büyük miktarlardaki verinin istatistiksel ve matematiksel tekniklerle birlikte örüntü tanıma teknolojilerinin de kullanılarak 4 incelenmesi yoluyla anlamlı, yeni ilişkiler, örüntüler ve eğilimler bulunması süreci olarak tanımlamaktadır.

Veri madenciliği günümüzde araştırmacılarca oldukça fazla kullanılan bir teknik olarak karşımıza çıkmaktadır. Özellikle toplumsal konularda ve endüstriye yönelik konularda veri madenciliğine başvuranların sayısı gitgide artmaktadır. Veri madenciliğine bu denli ilginin artmasının altında yatan ana sebep ise teknolojik gelişmelerdir. Teknolojik gelişmeler ile birlikte verilerin toplanması, depolanması ve çağrılması hususlarında kolaylık yaşanmaya başlanmıştır. Şekil 1'de görüldüğü üzere veri madenciliği belli aşamalardan geçerek günümüze gelmiştir (Jiawei ve Kamber, 2006).

Problemin Tanımlanması

Solunum Arresti süphesi ile kardiyoloji polikliniğe gelen hastaların oluşturdğu veri seti üzerinde, veri madenciliği yöntemleri kullanılarak bir kişinin Solunum Arresti olup olmadığını öngörebilmek bu çalışmanın ana amacıdır.

Veri Setini Anlama

Kullanılan veri seti İstanbul Okmeydanı ve Eğitim Araştırma Hastanesi'nden temin edilmiştir. Sağlık Bakanlığı'na bilimsel araştırma izni başvurusu yapılmış. Sağlık Bakanlığı'nın belirlediği gerekli koşullar sağlanmış, hastaneden onay alınmış ve süreç sonunda iznin çıkmasıyla çalışmaya başlanmıştır. Bu veri setinde hastalara ait kimlik bilgileri mevcut değildir. Değişkenler belirlenirken hastaların hikayeleri tek tek okunmuştur. Sonunda hastaarın nitelikleri (yaş, cinsiyet), laboratuvar sonuçları ve hastalık tanılarından oluşan 8 değişken belirlenmiştir.

Veri setinde 3 adet nümerik değer ile tanımlanmış sonuç bulunmaktadır. Bunlar Adm, Yıl, Ay'dır. Geri kalan değişkenler yaş değişkeni hariç kategorik değişkenlerdir. Cinsiyet değişkeninde erkek 1 ile, kadın 2 ile gösterilmiştir. Diğer değikenlerde ki 0 değeri o hastalığın yokluğunu, 1 değeri o hastalığın varlığını tanımlamaktadır.

Veri setinde bulunan niteliklere ait özellikler

	Değişken	**Veri Tipi**	**Veri Setinde Gösterimi**
1	Yaş	Nümerik	
2	Cinsiyet	Kategorik	1 = Erkek 2 = Kadın
3	Kardiyo-Solunum Arresti	Kategorik	0 = Yok 1 = Var
4	Hipoglisemik Aresti	Kategorik	0 = Yok 1 = Var
5	EKG (elektrokardiyogram)	Nümerik	
6	Kan Şekeri	Nümerik	
7	Hs-CRP	Nümerik	
	Hedef Nitelik		
8	Solunum Arresti	Kategorik	0 = Yok 1 = Var

Veriyi Hazırlama

Veri setinde bulunan uç noktalar incelenmiş, tekrar eden değerler çıkarılmıştır.

Veri seti temizlendikten sonra analiz için hazırlanmaktadır.

Analize Hazırlık

Bundan sonraki aşamalar RStudio'da yapılmıştır. Uygulama kodları eklerdedir.

Veri seti 200 gözlem ve 8 değişkenden oluşmaktadır.

Değişkenler sırasıyla: Yaş, Cinsiyet, Kardiyo-Solunum Arresti, Hipoglisemik Arresti, EKG, Kan Şekeri, Hs-CRP ve hedef nitelik olan Solunum Arresti'dir.

Öncelike veri setinin yapısı incelenmiş, nümerik ve faktör seklinde düzenlenmiştir. Nümerik değişkenler nümerik olarak, kategorik değişkenlerde faktör şeklinde tanımlanmıştır. Düzenlendikten sonra şu hale dönüşmüştür.

'data.frame': 200 obs. of 8 variables:

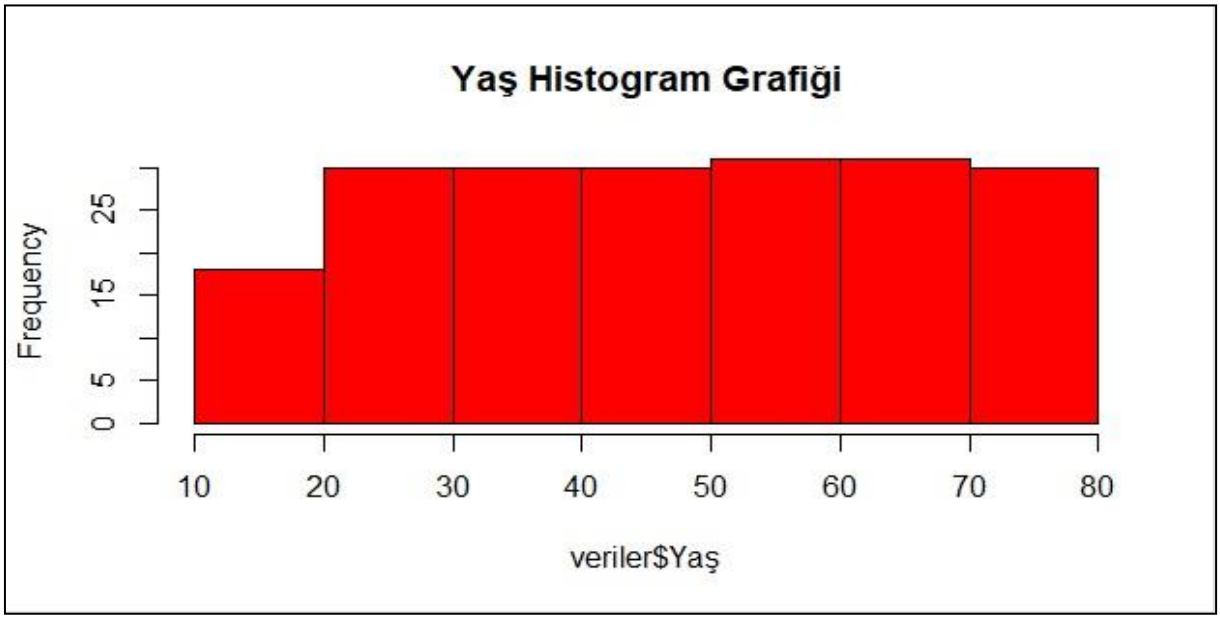
```
$ Yaş                    : num  25 19 25 19 24 18 25 18 23 20 ...
$ Cinsiyet               : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 ...
$ Kardiyo.Solunum.Arresti: Factor w/ 2 levels "0","1": 2 2 2 1 2 1 2 ...
$ Solunum.Arresti        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 ...
$ Hipoglisemik.Arresti   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 ...
$ EKG                    : num  52 80 62 56 80 86 96 90 62 76 ...
$ Kan.Şekeri             : num  56 112 76 110 64 110 96 102 106 78 ...
$ Hs.CRP                 : num  1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9...
```

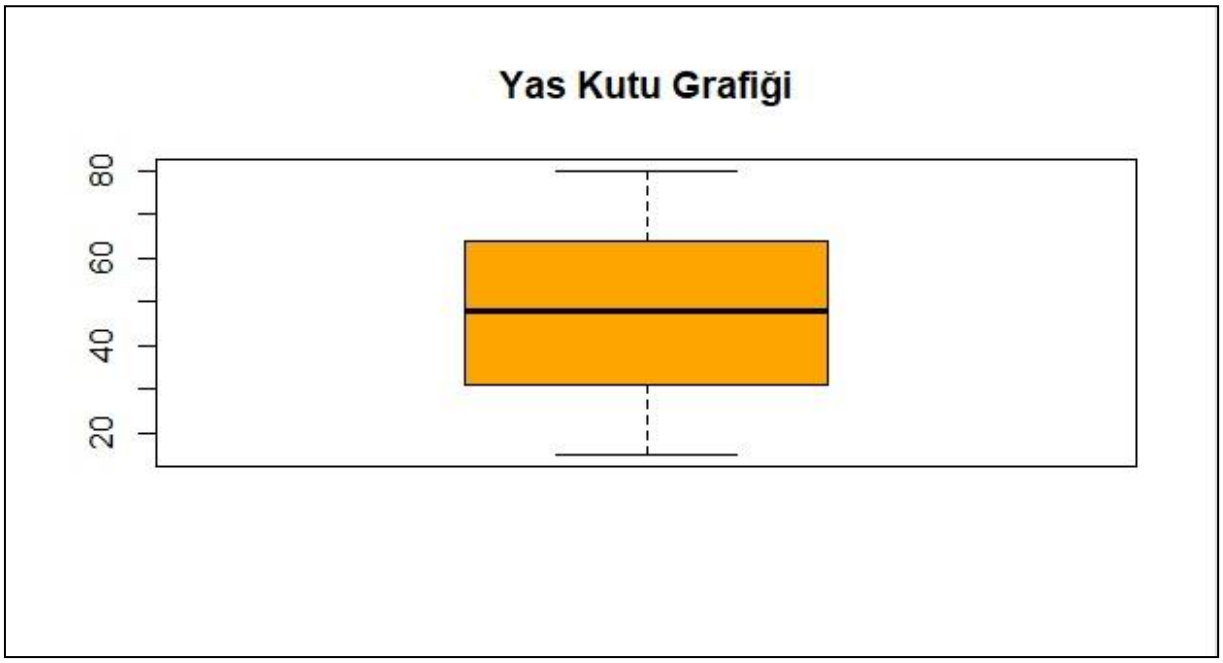
Bu tabloda kategorik değişkenler ve nümerik değişkenlerin minimum değerleri, 1. kartil, medyan, ortalama, 3. kartil ve maksimum değerleri görülür.

Yaş	Cinsiyet	Kardiyo Solunum Arresti	Solunum Arresti	Hipoglisemik Arresti	EKG	Kan Şekeri	Hs. CRP
Min. :15.00	Erkek : 99	0: 80	0:152	0:183	Min. : 50.0	Min. : 50.0	Min. :1.000
1st Qu.:31.00	Kadın :101	1:120	1: 48	1: 17	1st Qu.: 66.0	1st Qu.: 69.5	1st Qu.:1.700
Median :48.00					Median : 82.0	Median : 89.0	Median :2.400
Mean :47.63					Mean : 83.4	Mean : 89.1	Mean :2.441
3rd Qu.:64.00					3rd Qu.:100.5	3rd Qu.:108.5	3rd Qu.:3.200
Max. :80.00					Max. :120.0	Max. :130.0	Max. :4.000

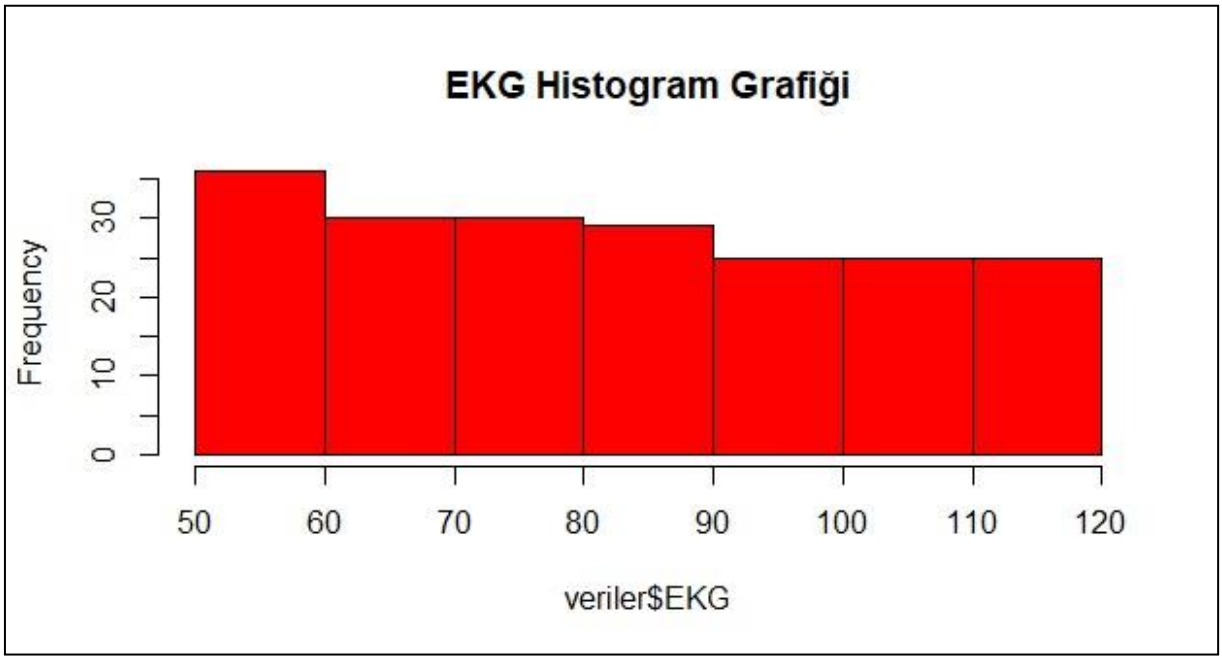
Veri setindeki değişkenler tek tek incelenmiştir. Bunun için her birine uygun grafikler çizilmiştir. Nümerik değişkenler için histogram grafikleri, kategorik değişkenler için çubuk grafikleri çizilmiştir. Ayrıca değişkenler kutu grafikleri ile de gösterilmiştir. Böylece dağılımları hakkında daha kolay bilgi edinilmiştir.

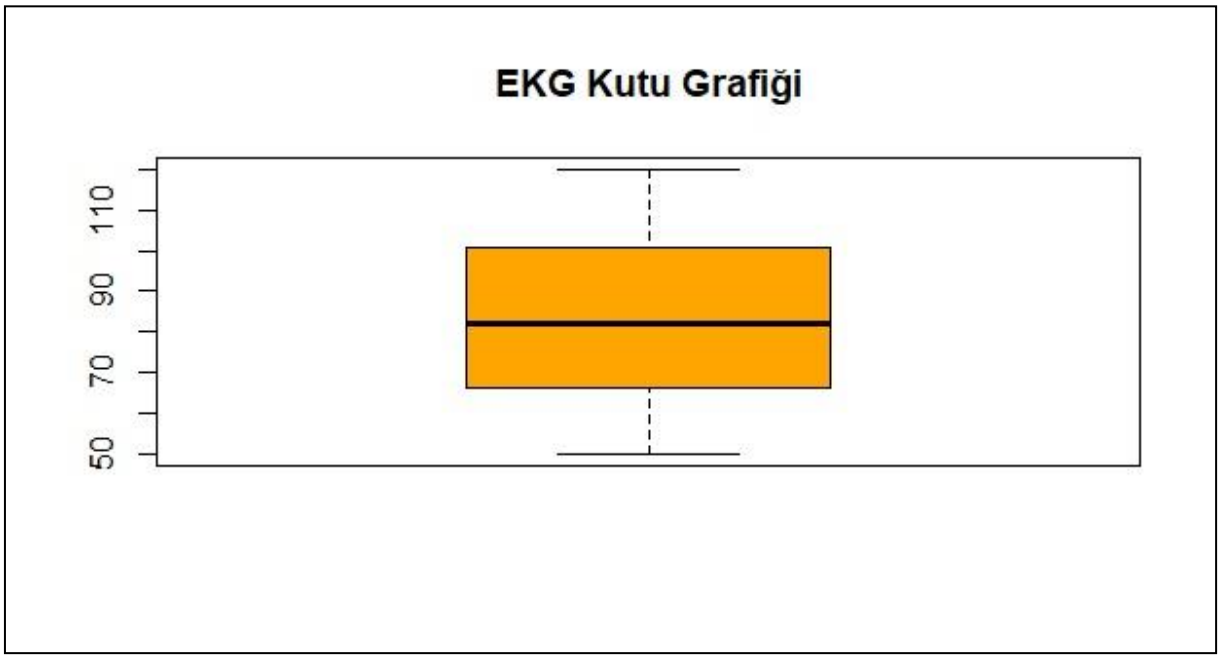
Veri setinde yaş değişim aralığı en küçük 18 yaşındaki hasta ile en büyük 25 yaşındaki hasta arasındadır. Frekanslarına bakıldığında 19-23 yaş civarı hasta sayısının çok olduğu görülmektedir.



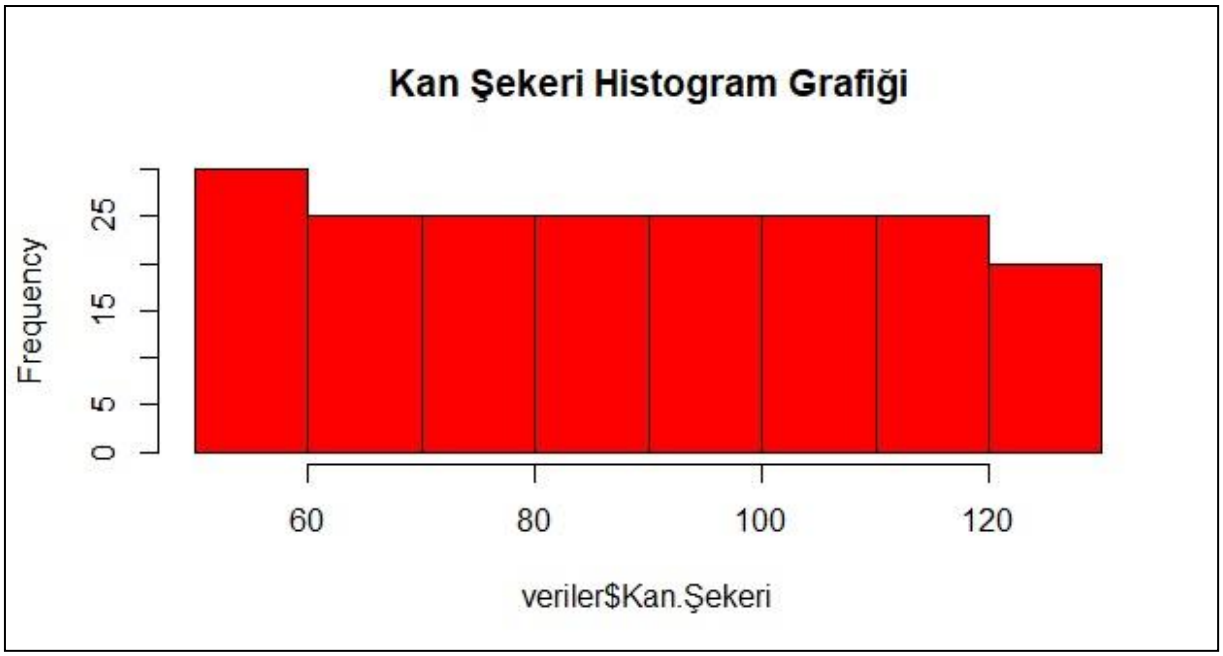


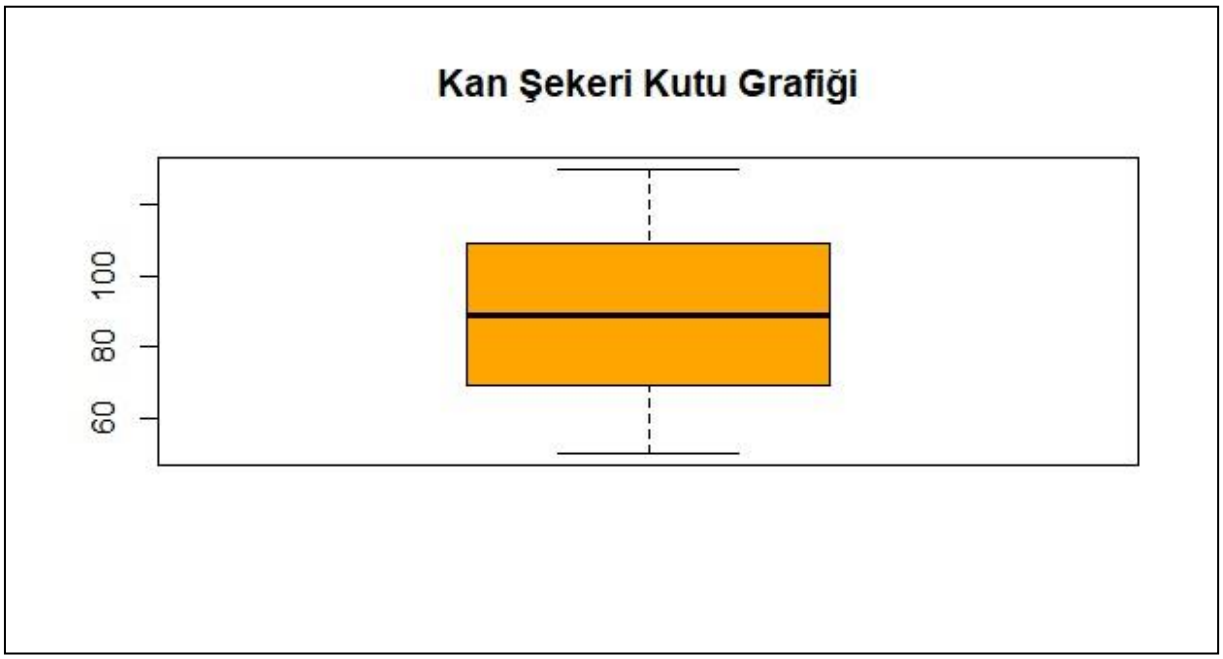
Veri setinde yaş değişim aralığı en küçük 15 yaşındaki hasta ile en büyük 80 yaşındaki hasta arasındadır. Frekanslarına bakıldığında 50 yaş civarı hasta sayısının çok olduğu görülmektedir.



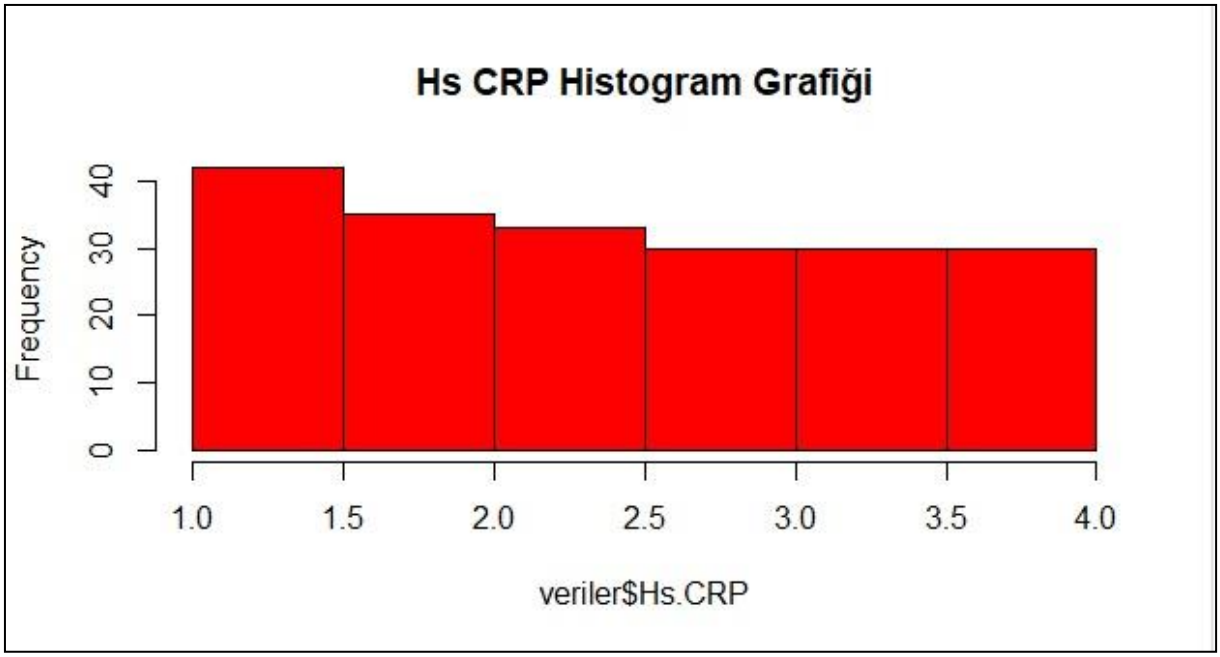


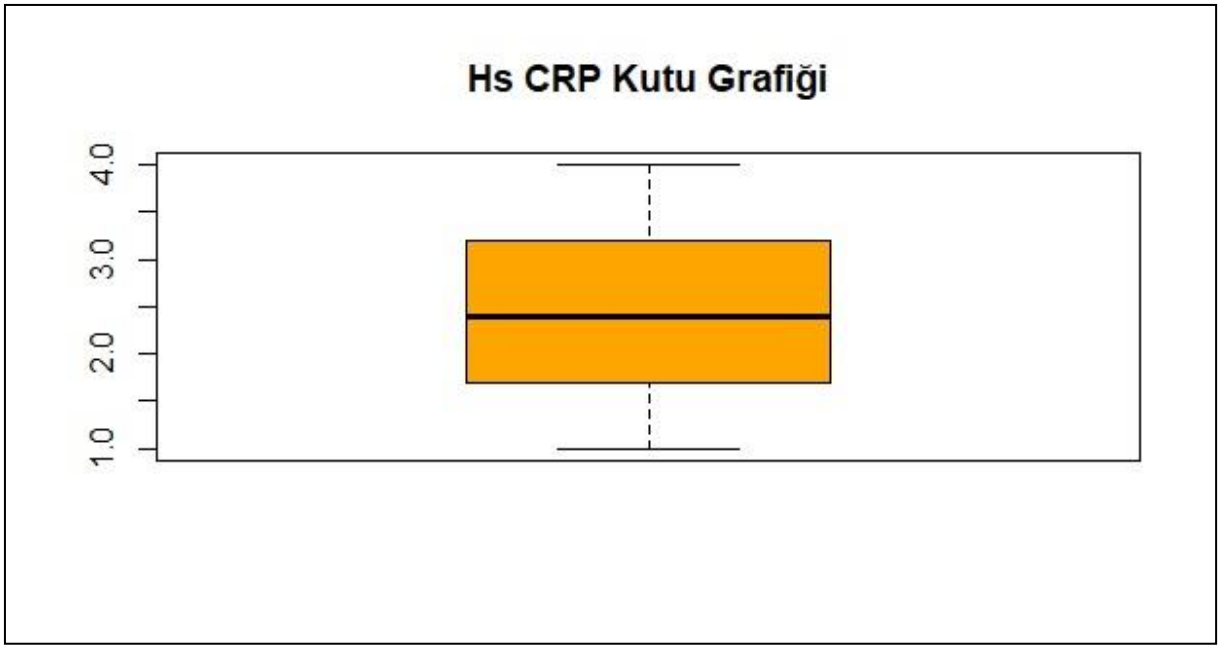
EKG değişkeni değerleri 50.0 ile 120.0 arasında değişmektedir. Ortalaması yaklaşık 83'dür.



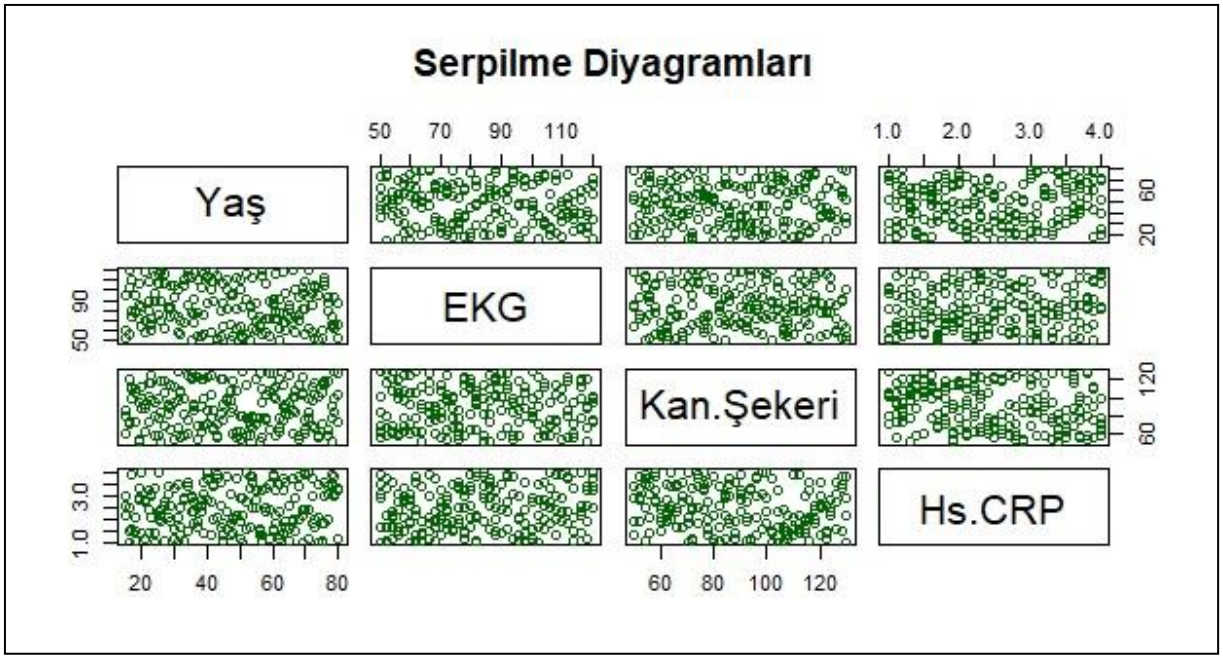


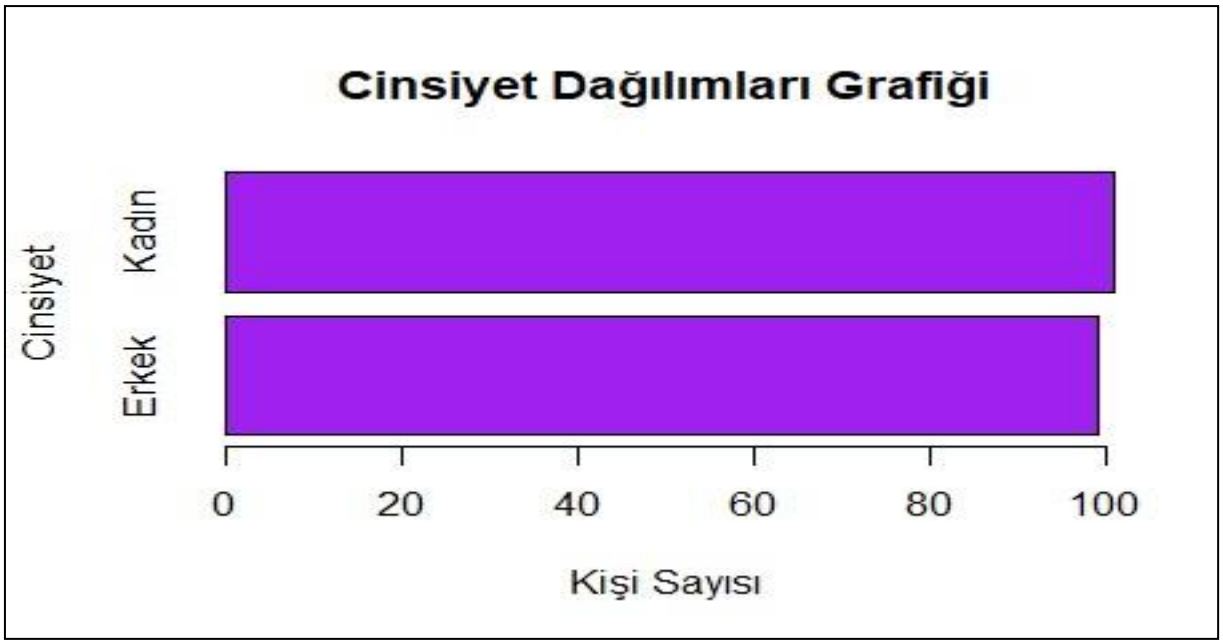
Kan şekeri değişkeni değerleri 50 ile 130 arasında değişmektedir. Ortalaması yaklaşık 90'dır



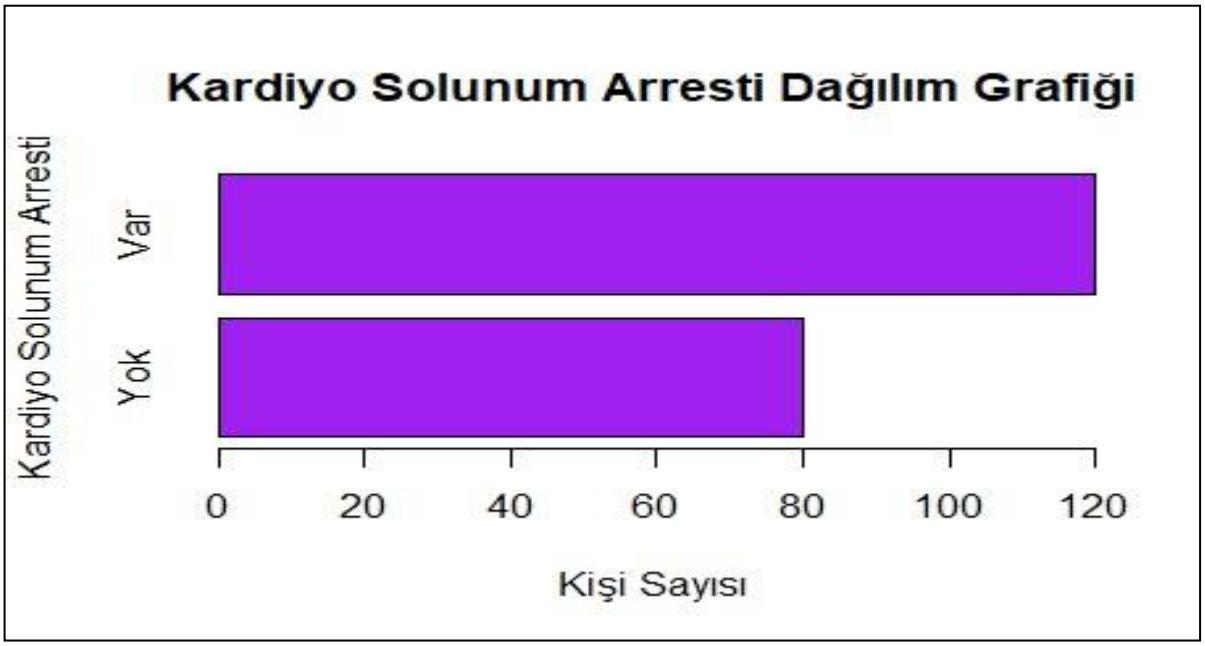


Hs CRP değişkeni değerleri 1.0 ile 4.0 arasında değişmektedir. Ortalaması 2.441 ‘dir.

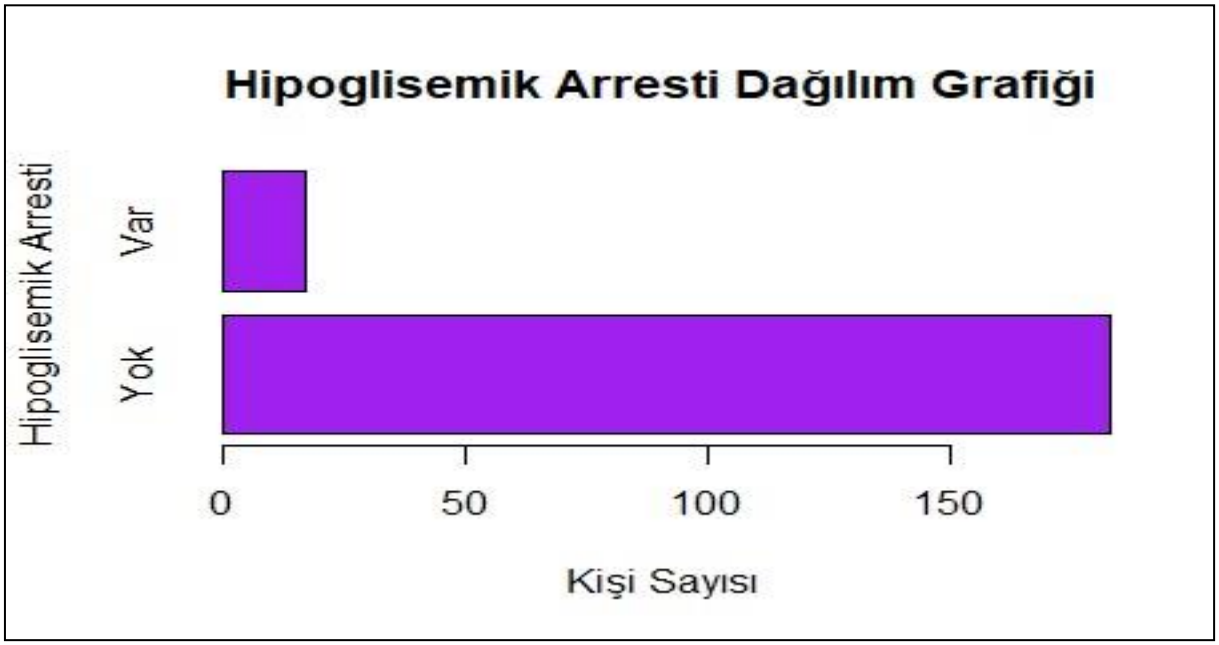




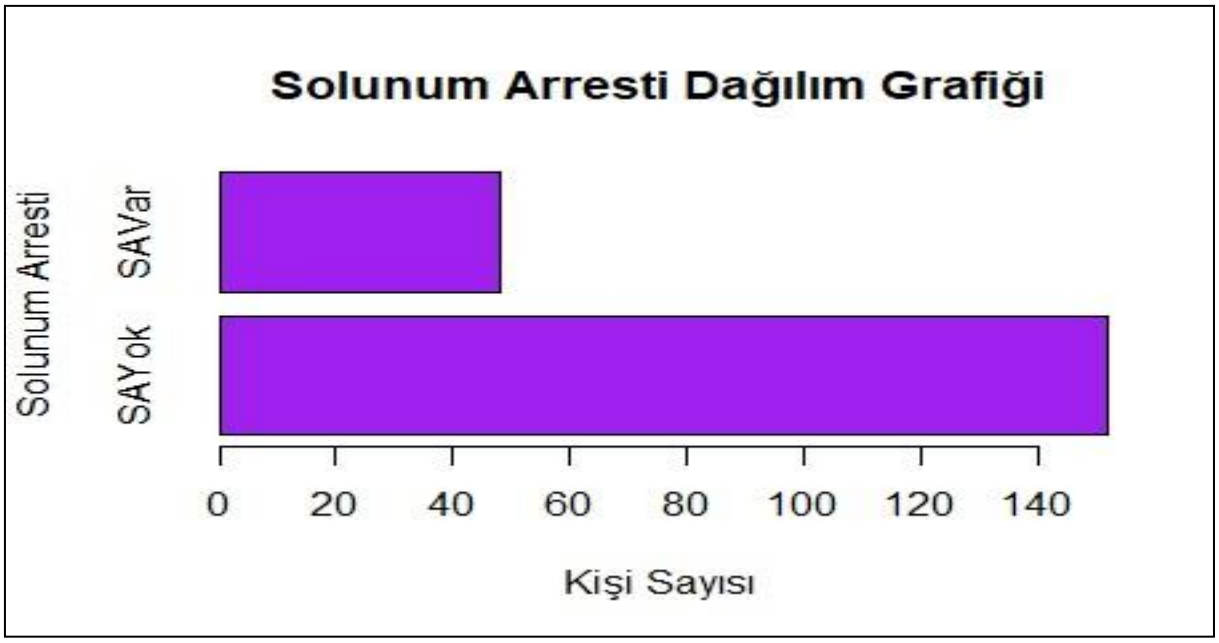
200 hastanın 99'u erkek, 101'i kadındır. Veri setinin % 49'u erkek, %51'i kadındır.



Kardiyo Solunum Arresti hastalığı olan 120 hasta, olmayan 80 hasta vardır. Veri setinin %60'ı Kardiyo Solunum Arresti tanısı konulan hastalardan, %40'ı Kardiyo Solunum Arresti tanısı konulmayan hastalardan oluşmaktadır.



Hipoglisemik Arresti hastalığı olan 17 hasta, olmayan 183 hasta vardır. Veri setinin %9'u Hipoglisemik Arresti tanısı konulan hastalardan, %91'i Hipoglisemik Arresti tanısı konulmayan hastalardan oluşmaktadır.



Solunum Arresti hastalığı olan 48 hasta, olmayan 152 hasta vardır. Veri setinin %24'ü Solunum Arresti tanısı konulan hastalardan, %76'sı Solunum Arresti tanısı konulmayan hastalardan oluşmaktadır.

Veri Dönüştürme

Veri setindeki nümerik değerler normalize edilmiştir. Bu işlem için minimum-maksimum yöntemi kullanılmıştır. En küçük değere 0 , en yüksek değere 1 atanarak, veri seti normalizasyon işlemi sonrası aşağıdaki şekle dönüşmüştür.

Veri seti artık analiz için hazırdır.

Yaş	EKG	Kan Şekeri	Hs.CRP
Min. :0.000	Min. :0.000	Min. :0.000	Min. :0.000
1st Qu. : 0.2462	1st Qu. : 0.2286	1st Qu. : 0.2437	1st Qu. : 0.2333
Median : 0.5077	Median : 0.4571	Median : 0.4875	Median : 0.4667
Mean : 0.5021	Mean : 0.4771	Mean : 0.4888	Mean : 0.4802
3rd Qu. : 0.7538	3rd Qu. : 0. 0.7214	3rd Qu. : 0.7312	3rd Qu. : 0.7333
Max. : 1.0000	Max. : 1.0000	Max. : 1.0000	Max. : 1.0000

KNN Algoritması

Sınıflandırma algoritmalarından olan K-En Yakın Komşu (KNN) algoritması danışmanlı öğrenen bir algoritmadır. Yani veri setinden öğrenim yapar. Veri seti, eğitim ve test veri seti olarak ikiye ayırılır. Eğitim veri setine algoritma öğretilir. Test veri setiyle de algoritma modeli test edilir. Gerçek veri ile öngörülen veri kıyaslanır. Kontenjans tablosu (confusion matrix) oluşturulur. Bu matrise göre modelin performans değerlendirme ölçütleri bulunur. Performans değerlendirme ölçütleri kurulan modelin ne kadar performans verdiği ölçer. Bunun için doğruluk oranı, hata oranı gibi ölçütler kullanılır.

KNN algoritması yeni bir veriyi sınıfa atarken, belirlenen bir k değeri kullanır. Bu k değerine göre mevcut örneklem içindeki verilere olan uzaklığı hesaplanır. Bu şekilde en yakın komşusu bulunarak o kümeye atanır.

Bu çalışmada Öklid uzaklığı ile uzaklıklar hesaplanarak K-en yakın komşu algoritması uygulanmıştır. Bunun için, veri setinden yalnızca nümerik değerler taşıyan ve hedef niteliğin de olduğu bir alt küme elde edilmiştir. Bu değişkenler, "Yaş", " EKG", "Kan Şekeri" ve "Hs.CRP" değişkenleridir, hedef nitelik de "Solunum Arresti" değişkenidir.

Formülü uyguladığımızda 200 veri ve 5 değişkenden oluşan yeni bir data.frame elde edilir.

Veri seti %60'ı eğitim ve %40'ı test veri seti olarak ayrılmıştır. Performans değerlendirme yöntemi olarak "Hold-out" yöntemi kullanılmıştır.

k değeri 1'den 5'e kadar denenmiştir.

Modelin performansının ölçülmesi için kontenjans tablosu kurulur.

k = 1 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri değerleri;

KNN	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	9	10
	SAyok	10	50

True positives(TP), doğru pozitif değeri 9'dur.

Gerçekte solunum arresti hastası olan 9 kişi, tahminde de solunum arresti hastasıdır diye tahmin edilmiştir. Kontenjans tablosu true positives yani doğru pozitif değeri bunu gösterir.

False positives(FP), yanlış pozitif değeri 10'dur

Var olan durum yani gerçekte solunum arresti hastalığı olmayan 10 kişi, solunum arresti hastalığı vardır şeklinde tahmin edilmiştir. Yani gerçekte olmayan bir durum tahminde var tahmin edilmişitir. Buna tip 1 hata denir. Kontenjans tablosunda yanlış pozitif sınıfı bu değeri gösterir.

False negatives(FN), yanlış negatif değeri 10'dur.

Gerçekte solunum arresti hastası olan 10 kişi, tahminde solunum arresti hastası değildir şeklinde tahmin edilmiştir. Gerçekte var olan(pozitif) bir durumun, tahminde yoktur(negatif) şeklinde bulunmasına tip 2 hata denir. Kontenjans tablosunda yanlış negatif sınıfı bu değeri gösterir.

True negatives(TN), doğru negatif değeri 50'dir.

Gerçekte solunum arresti olmayan hastalardan, modelin solunum arresti değildir biçiminde tahmin ettiği hasta sayısıdır. kontenjans tablosunda doğru negatif sınıfı bu değeri gösterir.

k=1 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=1 için
Doğruluk Oranı	% 0.74
Hata Oranı	% 0.25

k=2 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri değerleri;

KNN	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	7	12
	SAyok	12	48

k=2 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=2 için
Doğruluk Oranı	% 0.69
Hata Oranı	% 0.30

k=3 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri değerleri;

KNN	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	3	6
	SAyok	16	54

k=3 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=3 için
Doğruluk Oranı	% 0.72
Hata Oranı	% 0.27

k=4 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri değerleri;

KNN	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	3	7
	SAyok	16	53

k=4 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=4 için
Doğruluk Oranı	% 0.70
Hata Oranı	% 0.29

k=5 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri değerleri;

KNN	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	6	4
	SAyok	13	56

k=5 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=5 için
Doğruluk Oranı	% 0.78
Hata Oranı	% 0.21

Tüm k değerleri performans değerlendirme ölçütleri tablosu

Performans Değerlendirme Ölçütleri	**k=1**	**k=2**	**k=3**	**k=4**	**k=5**
Doğruluk Oranı	% 0.74	% 0.69	% 0.72	% 0.70	% 0.78
Hata Oranı	% 0.25	% 0.30	% 0.27	% 0.29	% 0.21

Tüm k değerleri kontenjans tablosu

k=1	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	9	10
	SAyok	10	50

k=2	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	7	12
	SAyok	12	48

k=3	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	3	6
	SAyok	16	54

k=4	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	3	7
	SAyok	16	53

k=5	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	6	4
	SAyok	13	56

k-NN algoritması performans değerlendirme ölçütleri oranları incelendiğinde doğruluk oranları için en yüksek değerler k=4 ve k=5'dir. Yanlış sınıflandırılmış örnek sayısının tüm veriye oranı olan hata oranları için en düşük değerler yine k=4 ve k=5 değeridir. Duyarlılık oranı yani doğru sınıflandırılmış pozitif değerlerin tüm pozitif değerlere oranı en yüksek k=1 değerinde çıkmıştır. Performans değerlendirme ölçütlerine göre en iyi değerleri veren model k=5 gibi gözükse de kontenjans tablosu incelenmelidir. k=1 kontenjans tablosuna bakıldığında gerçek pozitif değerin 9 çıktığı, yani gerçekte SA olan 9 hastanın tahminde de doğru tahmin edildiği görülmektedir. Oysa diğer matrislerde bu oran sürekli düşmektedir. k=2 matrisinde 7 kişi hem SA olup hemde doğru tahmin edilmiş, k=5 matrisinde 6 kişi hem SA olup hem de doğru tahmin edilmiştir. Bu matrisler hariç k değerlerinde, doğru pozitif değerleri 3 çıkmıştır. Bu çalışmada gerçek veriler ile hastalık tahmini yapılmaya çalışıldığından k=5 değeri sonuçları en yüksek doğruluk oranını ve en düşük hata oranını verse de, karışık matrisinde ortaya çıkan sonuçtan ötürü k=1 değeri tercih edilebilir.

k=1 değeri için kontenjans tablosu incelendiğinde, doğru pozitif(TP) değeri yani gerçekte solunum arresti olan hastalardan modelin solunum arresti olarak tahmin ettiği hastaların sayısı 9'dur.

Yanlış pozitif(FP) değeri yani gerçekte solunum arresti olmayan hastalardan modelin solunum arresti hastasıdır biçiminde tahmin ettiği kişi sayısı 10'dur. Tip 1 hatayı verir.

Yanlış negatif(FN) değeri yani gerçekte solunum arresti hastalığı olan kişilerin, modelin solunum arresti hastası değildir olarak tahmin ettiği kişi sayısı 10'dur. Tip 2 hatayı verir.

Doğru negatif(TN) değeri yani gerçekte solunum arresti olmayan hastalardan modelin solunum arresti hastası değildir biçiminde tahmin ettiği kişi sayısı 50'dir.

KNN algoritması için k=1 ve k=5 değeri ayrı ayrı incelenmiştir.

C4.5 Karar Ağacı Algoritması

C4.5 algoritması bir karar ağacı algoritmasıdır. Değişkenleri ağaç şeklinde dallanma yaparak sınıflandırır. C4.5 karar ağacı algoritması uygulanmadan önce veri setinin yapısı incelenmiştir. Değişkenler nümerik ve faktör şeklinde atanmıştır. Sınıflandırma algoritması olduğu için veri seti eğitim ve test veri seti olarak ayrılmıştır. Diğer algoritmalar ile bütünlük oluşturması açısından %60 eğitim veri seti, %40 test veri seti olarak ayrım yapılmıştır.

Uygulamanın yapılabilmesi için R programlamaya RWeka paketi yüklenmiş ve kütüphaneden çağırılmıştır. Paketin içindeki J48() fonksiyonu C4.5 karar ağacı algoritması çözümünde kullanılmıştır.

Eğitim veri setine uygulanan karar ağacı algoritması sonuçları verilmiştir.

```
=== Summary ===

Correctly Classified Instances          99               81.8182 %
Kappa statistic                          0.3469
Mean absolute error                      0.2911
Root mean squared error                  0.3815
Relative absolute error                 79.3935 %
Root relative squared error             89.3676 %
Total Number of Instances              121

=== Confusion Matrix ===

  a  b   <-- classified as
  8 21 |  a = SAvar
  1 91 |  b = SAyok
```

Burada correctly classified instances doğru yerleşen tahmin sayısıdır. Bunun toplam 121 kişi içinden 99 kişi olduğu gözükmektedir ve %81.8 doğruluk oranına sahiptir. Modelin oluşturduğu ağaç şu şekildedir:

```
J48 pruned tree
------------------

EKG <= 88: SAyok (120.0/26.0)
EKG > 88
|   EKG <= 114
|   |   Yaş <= 53
|   |   |   Kan.Şekeri <= 100
|   |   |   |   Kan.Şekeri <= 68: SAvar (10.0/3.0)
|   |   |   |   Kan.Şekeri > 68: SAyok (22.0/4.0)
|   |   |   Kan.Şekeri > 100: SAvar (9.0/2.0)
|   |   Yaş > 53: SAyok (24.0/3.0)
|   EKG > 114: SAyok (15.0/1.0)

Number of Leaves  :     6

Size of the tree :      11
```

Ağaç yapısı incelenir. Number of leaves yani yaprak sayısı 6 tanedir. Yapraktan sonra parantez içinde verilen değerler o kategoriye ait doğru ve yanlış sınıflandırmayı açıklar.

Örneğin;

EKG <= 88: SAyok (120.0/26.0)

Burada EKG 88'den düşük veya eşitse Solunum Arresti yoktur kuralı elde edilmiştir. Parantez ile ifade edilen, bu kategoride 120 örneğin doğru sınıflandırıldığı, 26 örneğin yanlış sınıflandırıldığıdır.

Karar ağacından elde edilen kurallar şu şekildedir.

KURAL1:

EKG değerleri 88'den küçük veya eşitse Solunum Arresti yoktur.

KURAL2:

EKG değerleri 88'den büyük ve EKG 114'den büyükse Solunum Arresti yoktur.

KURAL3:

EKG değerleri 114'den küçük veya eşitse VE yaş 53'den büyükse Solunum Arresti yoktur.

KURAL4:

Yaş 53'den küçük veya eşitse VE kan şekeri 100'den büyükse Solunum Arresti vardır.

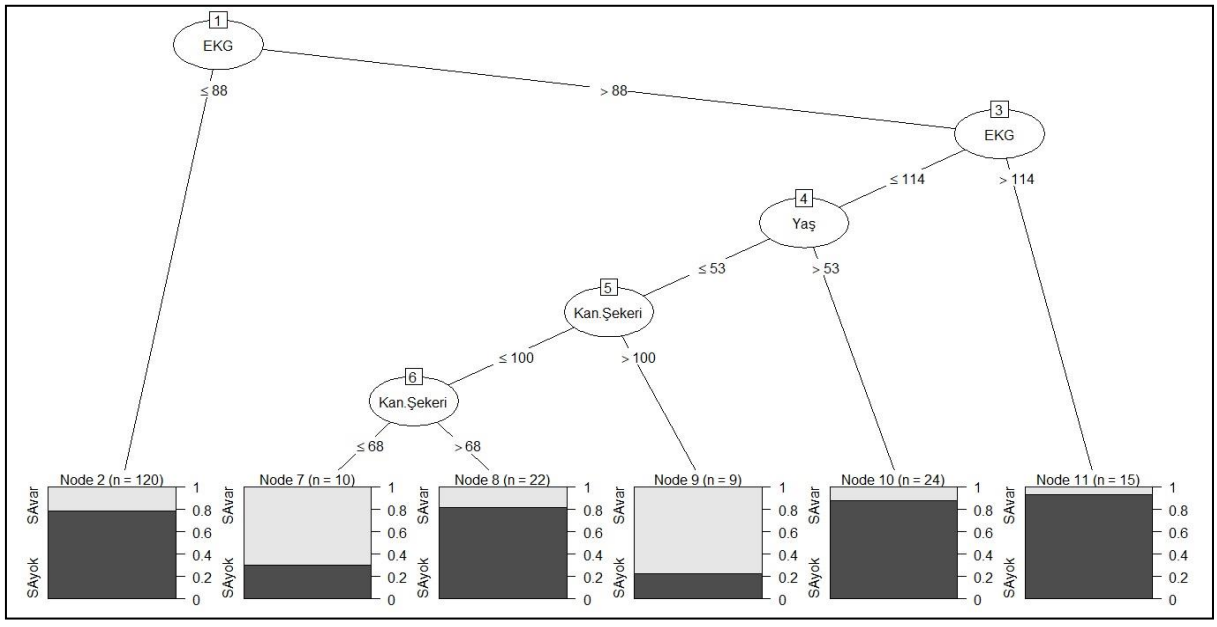
KURAL5:

Kan şekeri 100'den küçük veya eşitse VE 68'den büyükse Solunum Arresti yoktur.

KURAL6:

Kan şekeri 68'den küçük veya eşitse Solunum Arresti vardır.

C4.5 karar ağacı



Kontenjans Tablosu

C4.5	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	2	2
	SAyok	17	58

Performans Değerlendirme Ölçütleri Tablosu

Performans Değerlendirme Ölçütleri	**C4.5**
Doğruluk oranı	% 0.75
Hata oranı	% 0.24
TPR(Duyarlılık oranı)	% 0.10
SPC(Belirleyicilik oranı)	% 0.96
PPV(Pozitif öngörü oranı)	% 0.5
NPV(Negatif öngörü oranı)	% 0.77
FPR(Yanlış pozitif oranı)	% 0.02
FNR(Yanlış negatif oranı)	% 0.89
F-ölçütü	% 0.17

- C4.5 karar ağacı algoritması kontenjans tablosu sonuçları incelendiğinde, doğru pozitif değerinin 2 çıktığı görülmektedir. Yani gerçekte solunum arresti olan hastalardan, model tahminde solunum arresti hastasıdır diye 3 kişiyi doğru tahmin etmiştir.

- Gerçekte solunum arresti hastalığı olmayan kişileri model, solunum arresti hastasıdır şeklinde tahmin etmiştir. Yanlış pozitif değeri 2'dir.

- Gerçekte solunum arresti hastalığı olan kişileri model, solunum arresti hastalığı yoktur şeklinde tahmin etmiştir. Yanlış negatif değeri 17'dir.

- Gerçekte solunum arresti olmayan hastaları model solunum arresti değildir şeklinde, 58 kişide doğru tahmin etmiştir. Doğru negatif değeri 58'dir.

- Modelin doğruluk oranı 0.75 ve hata oranı 0.24 çıkmıştır. Kurallarda ortaya çıkan belirleyici değişkenler, EKG, Yaş ve Kan şekeri değerleridir.

Naive (Basit) Bayes Sınıflandırıcı Algoritması

Naive (Basit) Bayes ile bütün koşullu olasılık değerleri çarpılarak sınıflandırılır. Temeli Bayes teoremine dayanmaktadır. Bayes teoreminde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişki gösterilmektedir. Naive bayes yöntemi sınıflandırma algoritmaları içerisinde yer almaktadır.

Analiz öncesi değişkenler faktör ve nümerik olarak tanımlanmış, nümerik veriler normalize edilip analize uygun hale getirilmiştir. Diğer snıflandırma algoritmalarında olduğu gibi veri seti eğitim veri seti ve test veri seti olarak ayrılmıştır. Eğitim veri seti %60, test veri seti %40 olarak bölünmüştür. Eğitim ve test veri setine tahmininde kullanılacak nitelik ve hedef nitelik(diyabetik polinöropati) atanmıştır.

Naive Bayes algoritmasının kullanılması için R programına "e1071" paketi yüklenmeli ve kütüphaneden çağrılmalıdır. Bu paketteki naiveBayes() fonksiyonu kullanılmıştır. Model tahmin edilmiş ve aşağıdaki koşullu olasılık değerleri bulunmuştur.

Naive Bayes Classifier for Discrete Predictors

Call: naiveBayes.default(x = egitimNitelikleri, y = egitimHedefNitelik)

A-priori probabilities:

egitimHedefNitelik

SAvar	SAyok
0.2396694	0.7603306

Conditional probabilities:

Yaş

egitimHedefNitelik	[,1]	[,2]
SAvar	0.4647215	0.3048323
SAyok	0.5113712	0.3089608

Cinsiyet

egitimHedefNitelik	1	2
SAvar	0.4482759	0.5517241
SAyok	0.5000000	0.5000000

Kardiyo.Solunum.Arresti

egitimHedefNitelik	0	1
SAvar	0.5172414	0.4827586
SAyok	0.3804348	0.6195652

Hipoglisemik.Arresti

egitimHedefNitelik	0	1
SAvar	0.89655172	0.10344828
SAyok	0.91304348	0.08695652

EKG

egitimHedefNitelik	[,1]	[,2]
SAvar	0.5083744	0.3184675
SAyok	0.4649068	0.2922058

Kan.Şekeri

egitimHedefNitelik	[,1]	[,2]
SAvar	0.5206897	0.3289641
SAyok	0.4915761	0.2766131

Hs.CRP

egitimHedefNitelik	[,1]	[,2]
SAvar	0.4034483	0.3129918
SAyok	0.5043478	0.3066307

Tahmin edilen değerlerin ve gerçek değerlerin kıyaslanması için kontenjans tablosu elde edilmiştir.

Naive bayes kontenjans tablosu

Naive Bayes	Gerçek		
Tahmin		SAvar	SAyok
	SAvar	0	0
	SAyok	19	60

Naive bayes performans değerlendirme ölçütler

Performans Değerlendirme Ölçütleri	**Naive Bayes**
Doğruluk oranı	% 0.75
Hata oranı	% 0.24

- Naive Bayes algoritması kontenjans tablosu sonuçlarına göre gerçekte solunum arresti hastalığı olan 0 hasta, tahminde de solunum arresti hastası olarak tahmin edilmiştir. Doğru pozitif değeri 0'dır.
- Gerçekte solunum arresti hastalığı bulunmayan, ama tahminde solunum arresti hastasıdır çıkan 0 kişi vardır. Yanlış pozitif yani tip 1 hata değeri 0'dır.
- Gerçekte solunum arresti hastası olan, tahminde solunum arresti hastası değildir çıkan 19 kişi vardır. Yanlış negatif yani tip 2 hata değeri 19'dur.
- Gerçekte solunum arresti hastalığı olmayan, tahminde de solunum arresti hastalığı yoktur çıkan 60 kişi vardır. Doğru negatif değeri 60'dır.
- Modelin doğruluk oranı 0.75 ve hata oranı 0.24 çıkmıştır.

SONUÇ

Veri madenciliği sürecine sadık kalınarak, uygulama aşamaları anlatılmış ve uygulamada kullanılan tekniklere yer verilmiştir. K-nn, C4.5 ve Naive-Bayes analizi kullanılmıştır.

K-nn algoritması k değeri 1'den 5'e kadar değer verilerek tahminlenmeye çalışılmıştır. Bu algoritma için performans değerlendirme ölçütlerine bakıldığında k=4 ve k=5 algoritmalarının en iyi sonuçları verdiği gözlenmiştir. k=3 ve k=4 değerleri ile tahminlenen modeller birbirinin aynısı olduğu için k=5 algoritma sonuçlarının performans ölçüm değerlendirmesi bakımından en iyi sonuçları verdiği söylenebilir. **Bunu göre doğruluk oranı % 0.78 ve hata oranı %0.21'dir.**

K-nn algoritması kontenjans tablosu incelendiğinde doğru pozitif, yanlış pozitif, yanlış negatif ve doğru negatif tahmin değerleri elde edilmektedir. Buna göre matrisler incelendiğinde k=1 değeri ile tahminlenen modelde gerçekte solunum arresti hastası olan hastaların, modelin solunum arresti hastasıdır diye tahmin ettiği 9 hasta olduğu görülmektedir. Yanlış pozitif olarak tahmin ettiği yani gerçekte solunum arresti olmayan hastalardan modelin solunum arresti vardır olarak tahmin ettiği 10 hasta olduğu görülmektedir. Yanlış negatif değerleri yani gerçekte solunum arresti olan hastaları modelin solunum arresti hastası değildir olarak yanlış tahmin ettiği 10 hasta vardır. Doğru negatif değeri ise gerçekte solunum arresti olmayan hastaların modelde de solunum arresti hastası değildir şeklinde tahmin edilmesidir. Bu sayı da 50 kişidir.

K-nn algoritması k değerleri karışıklar matrisi için incelendiğinde k=2 için doğru pozitif tahmin sayısının 7'ye düştüğü görülmektedir. k=3 ve k=4 değeri için doğru pozitif değeri 3 olmuştur ve incelendiğinde k=5 değeri ile tahminlenen matrislerde doğru pozitif değerinin 6 olduğu görülmektedir.

Buna göre düşünüldüğünde model doğruluk oranı en yüksek ve hata oranını en düşük veren k=3 ve k=4 değeri mi, yoksa gerçek pozitif değeri 9 kişiyi doğru tahmin eden k=1 değeri mi alınmalıdır. Gerçek hasta verisi ile çalışıldığından ve sadece bir hastayı değil de daha çok hastayı doğru tahmin etmek istenildiğinden k=1 ile çözümlenen knn algoritması daha iyi performans göstermektedir şeklinde yorumlanabilir.

C4.5 karar ağacı algoritması karışık matrisi sonuçları incelendiğinde "doğru pozitif" 2 kişi, "yanlış pozitif" 2 kişi, "yanlış negatif" 17 kişi ve "doğru negatif" 58 kişi tahmin edilmiştir.

Modelin performans ölçümlerine bakıldığında **doğruluk oranı %0.75 ve hata oranı %0.24 çıkmıştır.**

Naive Bayes algoritması sonuçları incelendiğinde gerçekte solunum arresti hastası olan hastalardan modelin 0 kişiyi solunum arresti vardır şeklinde doğru tahmin ettiği görülmektedir. Doğru pozitif değeri 0'dır.

Gerçekte solunum arresti olmayan hastalardan, modelin solunum arresti vardır şeklinde, "yanlış pozitif " tahmin ettiği 0 kişi vardır.

Gerçekte solunum arresti olan hastalardan, modelin solunum arresti yoktur şeklinde, "yanlış negatif " tahmin ettiği 19 kişi vardır.

Gerçekte solunum arresti hastası olmayan hastaların, modelde de solunum arresti hastası değildir şeklinde tahminlendiği 60 kişi vardır. Doğru negatif değeri 60'dır.

Naive Bayes algoritması performans ölçümlerine bakıldığında **doğruluk oranı %0.75 ve hata oranı %0.24'tür.**

Tüm modeller birlikte değerlendirildiğinde performans ölçüm modelleri değerlendirme ölçütlerine göre en yüksek doğruluk ve en düşük hatayı veren **K-nn algoritması en uygun modeldir denilebilir.**

KAYNAKÇA

- UC Irvine Machine Repository
- Nur Kuban ders içi kaynakları