T.C BİLECİK ŞEYH EDEBALİ ÜNİVERSİTESİ

İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ

YÖNETİM BİLİŞİM SİSTEMLERİ

SAĞLIK HİZMETLERİNDE VERİ ANALATİĞİ

ÇİĞDEM DEDEOĞLU

ÖZ

SAĞLIK HİZMETLERİNDE VERİ ANALATİĞİ

ÇİĞDEM DEDEOĞLU

Bu raporun amacı kalp yetmezliği tanıları şüphesiyle oluşturulmuş veri seti üzerinde analiz yaparak hastaların kalp yetmezliği sonucunda ölüm olayı hedefi belirlenmiştir. Çalışmada kullanılan veri seti UCİ MACHİNE LEARNİNG sitesinden temin edilmiştir. Veri seti sınıflandırma algoritmaları ile analiz edilmiştir. K-en yakın komşu algoritması, Naive Bayes sınıflayıcı, C4.5 karar ağacı algoritması uygulanmış ve farklı modeller oluşturulmuştur. Modellerin performans doğruluk oranları ve hata oranları kıyaslanmıştır ve model performans doğruluk oranı yüksek olan model seçilmiştir. Veri analizleri R programla dili ile RStudio'da yapılmıştır.

Anahtar Kelimeler: Veri Madenciliği, R programlama, Sağlık Hizmetleri, Kalp Yetmezliği

ÖNSÖZ

Yapılan araştırmalara göre ülkemizde 2 milyonun üzerinde insan kalp yetmezliği yaşamaktadır. Kalp yetmezliği sonucu ölüm oranları başta bağırsak, meme ve prostat kanseri olmak üzere pek çok kanser hastalığına bağlı ölüm oranlarından daha yüksektir. Öte yandan, kalp yetmezliğinin pek çok tipi önlenebilir niteliktedir. Kalp yetmezliği tanısı konulmuş hastalarda bile, semptomlar hakkında daha çok bilgi sahibi olmaları ve bu doğrultuda tıbbi desteğe başvurmaları konusunda doğru yönlendirmeler yapılırsa, zamansız ölümlerin önlenmesi sağlanabilir. Bu sebeple, Sağlık hizmetleri veri analitiği çalışmasında konu olarak kalp yetmezliği seçilmiştir.

Modelleme aşamasında K-en yakın komşu algoritması, Naive-Bayes sınıflandırma, c4.5 karar ağacı algoritmaları kullanılmıştır. Elde edilen sonuçlar karşılaştırılarak en uygun model belirlenmiştir.

Bu Çalışma konusunun belirlenmesinde ve çalışmanın hazırlanma sürecinin her aşamasında bilgilerini, tecrübelerini ve değerli zamanlarını esirgemeyerek bana her fırsatta yardımcı olan değerli hocam Sayın Dr. Öğr. Üyesi Nur Kuban TORUN'a teşekkürü bir borç bilirim.

ÇİĞDEM DEDEOĞLU

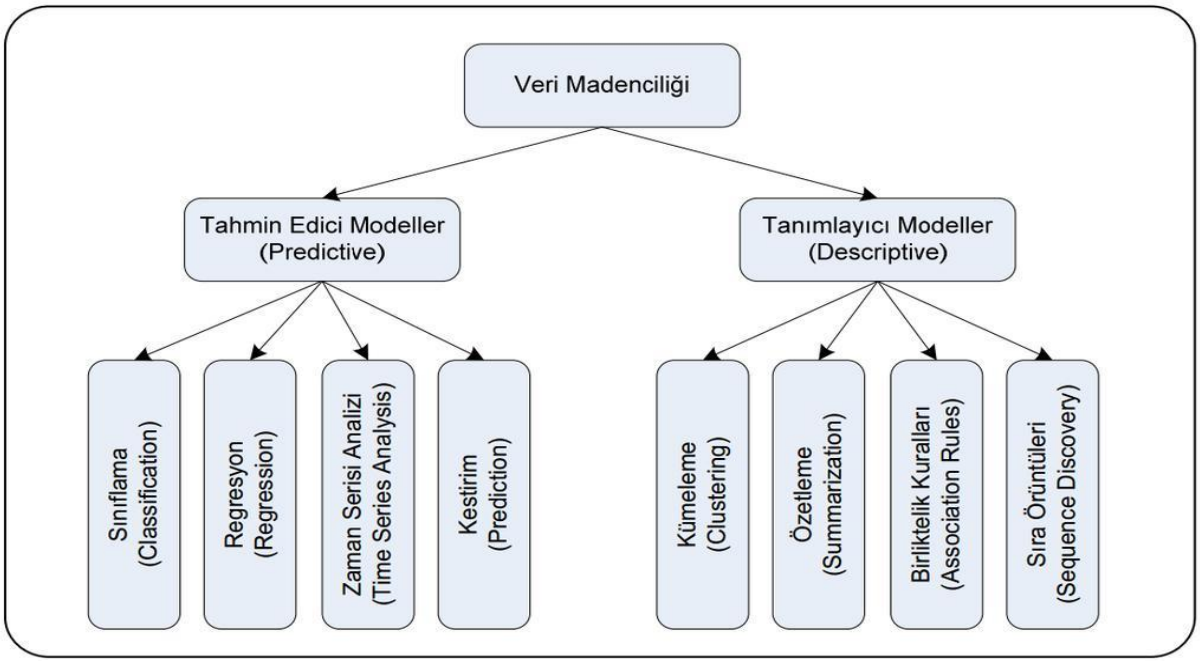
İÇİNDEKİLER

1.VERİ MADENCİLİĞİ NEDİR

"Veri tabanlarında bilgi keşfi" ve "veri madenciliği" literatürde birbirine yakın anlamlarda kullanılmaktadır. Usama M. Fayyad'a göre "veritabanlarında bilgi keşfi" veriden faydalı bilgiyi keşfetme süreci, "veri madenciliği" ise veriden örüntülerin çıkarılması için algoritmaların uygulanması olarak tanımlanır (Fayyad ve diğerleri, 1996). Veri madenciliği hedeflenen sonuçları elde edebilmek için, analiz edilmek üzere hazırlanmış verilere algoritmaların uygulandığı bilgi keşif sürecinin adımı olarak görülmektedir. Bununla beraber endüstride, medya ve veritabanı araştırmalarında "veri madenciliği" terimi "veritabanlarında bilgi keşfi" teriminden daha yaygın olarak kullanılmaktadır. Bu nedenle sürecin tamamı genellikle veri madenciliği olarak anılmaktadır. Veri madenciliği veritabanı sistemleri, istatistik, makine öğrenmesi, görselleştirme ve enformasyon bilimini içeren disiplinler arası bir alandır.

1.1 VERİ MADENCİLİĞİ GÖREVLERİ VE UYGULAMA ADIMLARI

Veri madenciliğinde farklı görevleri yerine getirmek için pek çok farklı algoritmalar kullanılır. Bu algoritmalar verilere uygun modeli bulmaya çalışır. Algoritmalar verileri inceler ve özelliklerine en uygun modeli seçer. Veri madenciliği görevleri "tahmin edici" ve "tanımlayıcı" modeller olmak üzere iki kategoriye ayrılır (Dunham,2003). Bu kategoriler ve modeller Şekil 1'de gösterilmiştir.



Tahmin edici modeller, sonuçları bilinen verileri kullanarak ilgili unsurlar için bir tahmin modeli oluşturur. Elde edilen bu model, sonuçları bilinmeyen unsurların tahmin edilmesinde kullanılır. Örneğin bir hastanede bir hastalığa ilişkin veri setini düşünelim. Veri madenciliği teknikleri uygulanarak hastalığa ilişkin geçmiş olaylardan elde edilmiş tıbbi veriler ve hasta durumu verilerinden bir tahmin modeli oluşturulabilir. Bu model sayesinde, hastaneye yeni gelmiş bir hastanın hastalığına ilişkin tahmin testler sonrası oluşan tıbbi veriler kullanılarak yapılabilir. Tahmin edici modeller sınıflama, regresyon, zaman serisi analizi ve kestirim olmak üzere dört grup olarak sınıflandırılabilir. Tanımlayıcı modeller verilerdeki örüntü veya ilişkileri tanımlarlar. Bu modeller tahmin edici modellerin aksine analiz edilen verilerin özelliklerini incelemek için kullanılan modellerdir. Örnek olarak sigorta poliçesini yenilememiş müşterilerin benzer özelliklerini belirleyecek bir kümeleme çalışması verilebilir. Kümeleme, özetleme, birliktelik kuralları, sıra örüntüleri keşfi modelleri tanımlayıcı modeller olarak nitelendirilir. Pek çok veri madenciliği sistem yazılımı geliştiren kuruluş, kullanıcılara yol göstermek amacıyla bir uygulama süreç modeli önerirler. Bu modeller genellikle ardışık adımların yürütülmesiyle kullanıcıları hedefe ulaştırmayı amaçlar.

1.2 VERİ MADENCİLİĞİ SÜRECİ

Veri madenciliği verilerden bilgi keşfi sürecinin içinde bir adımdır. Bu yaklaşıma göre verilerden bilgi keşfi süreci birbirini takip eden aşağıdaki adımlardan oluşur:

1. Verilerin temizlenmesi (tutarsız, aykırı, yanlış verilerin çıkarılması)

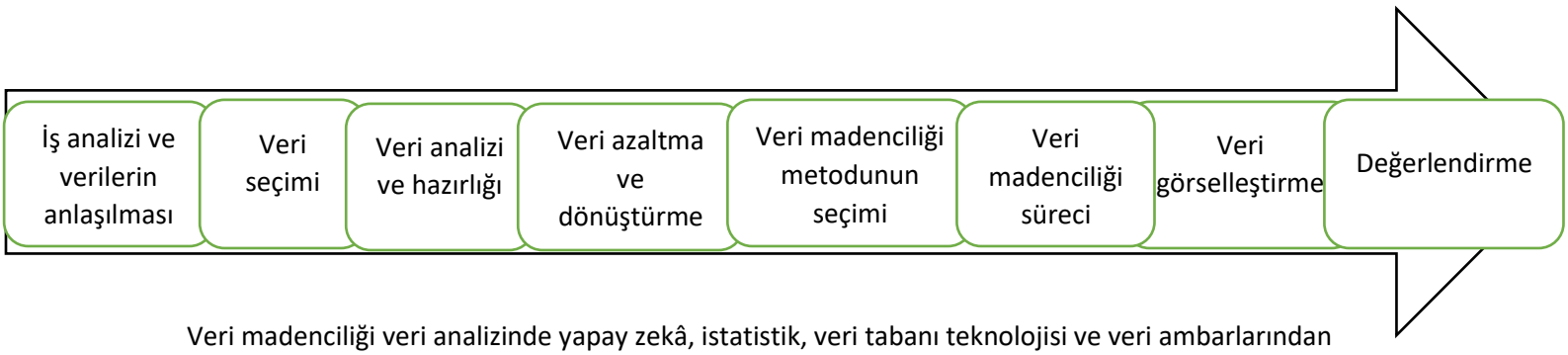
2. Verilerin bütünleştirilmesi (farklı kaynaklardan elde edilen verilerin bir araya getirilmesi)

3. Verilerin seçilmesi (veri tabanından analiz yapılacak verilerin seçilmesi örnek: değişken birey seçimi gibi)

4. Verilerin dönüştürülmesi (verilerin madencilik tekniklerinin gerektirdiği uygun yapıya getirilmesi)
5. Veri madenciliği deseni (veri deseni elde etmek için ilgili tekniklerin kullanıldığı temel işlem)

6. Desen değerlendirme (ilginçlik ölçütlerine göre gerçek ilginç desenlerin tanımlanması)

7. Bilgi sunumu (uygulama sonuçlarının görselleştirme ve betimlemeyle kullanıcıya sunumu)

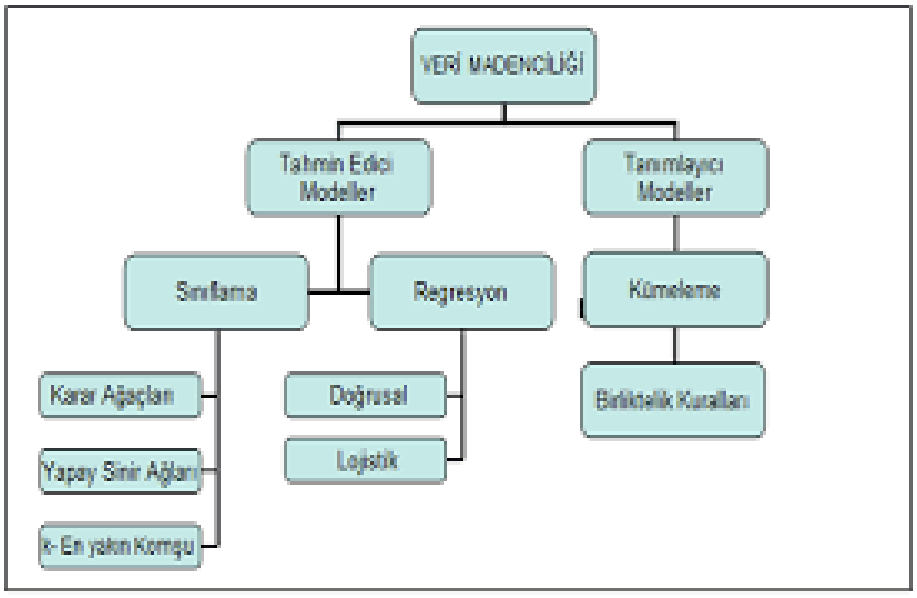


Veri madenciliği veri analizinde yapay zekâ, istatistik, veri tabanı teknolojisi ve veri ambarlarından önemli ölçüde yararlanmaktadır. Çünkü veri madenciliği yalnızca hazır verinin analizinden ibaret değildir. Veri madenciliği, veri analizinin yanı sıra araştırılacak problemle ilgili veri tabanının hazırlanması, verinin ilgili veri tabanlarından sorgulanması, verinin analize hazır hale getirilip analiz sonucunda elde edilen enformasyonun bilgiye dönüştürülmesi işlemlerini içeren uzun bir süreçtir.

Veri madenciliği sürecini karar probleminin belirlenmesi, veri önişleme, veri analizi ve sonuçların yorumlanması şeklinde ayırabiliriz (Vatansever, 2008).

1.3 VERİ MADENCİLİĞİ MODELLERİ

Veri madenciliğinde kullanılan modeller, tanımlayıcı ve tahmin edici olarak iki ana başlık altında incelenmektedir. Tanımlayıcı modellerde, karar vermeye yardımcı olarak kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. Tahmin edici modellerde, sonuçları bilinen verilerden hareketle bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Gerek tanımlayıcı gerekse tahmin edici modellerde yoğun olarak kullanılan belli başlı istatistiki yöntemler; sınıflama, regresyon, kümeleme, birliktelik kuralları, yapay sinir ağları olmak üzere beş ana başlık altında incelemek mümkündür. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları tanımlayıcı modellerdir.



1.4 VERİ MADENCİLİĞİ YÖNTEMLERİ

Veri Madenciliği yöntemlerini denetimli ve denetimsiz olmak üzere iki ana kategoriye ayırmak mümkündür. Veri Madenciliğinde iyi tanımlanmış veya kesin bir hedef olduğunda denetimli (supervised) ifadesi kullanılır. Elde edilmesi istenen sonuç için özel bir tanımlama yapılmamışsa veya belirsizlik söz konusu ise denetimsiz (unsupervised) ifadesi kullanılır [9].

Denetimli ve denetimsiz ifadeleri birbirinin tersine karşılık gelmektedir. Denetimli ve denetimsiz yöntemleri sürecin bütünü açısından değerlendirmek gerekirse;

• Denetimsiz yöntemler daha çok veriyi anlamaya, tanımaya, keşfetmeye yönelik olarak kullanılan ve sonraki uygulanacak yöntemler için fikir vermeyi amaçlamaktadır,

• Denetimli yöntemler ise veriden bilgi ve sonuç çıkarmaya yönelik kullanılmaktadır, denilebilir.

Bu nedenle denetimsiz bir yöntemle elde edilen bir bilgi veya sonucu, eğer mümkünse denetimli bir yöntemle teyit etmek, elde edilen bulguların doğruluğu ve geçerliliği açısından önem taşımaktadır. Denetimli (Supervised) Veri Madenciliği yöntemleri:

• En yakın k komşuluk (k-Nearest-Neighbor)

• K-ortalamalar kümeleme (K-means clustering)

• Regresyon modelleri (Regression models)

• Kural çıkarımı (Rule induction)

• Karar ağaçları (Decision trees)

• Sinir ağları (Neural networks) Denetimsiz (Unsupervised) Veri Madenciliği yöntemleri:

• Aşamalı kümeleme (Hierarchical clustering

• Kendi kendini düzenleyen haritalar (Self organized maps) olarak sınıflandırılabilir [9, 12].

Veri Madenciliği ile ilgili kullanılan pek çok yöntemin yanına hemen her geçen gün yeni yöntem ve algoritmalar eklenmektedir. Bunlardan bir kısmı onlarca yıldır kullanılan klasik teknikler diyebileceğimiz ağırlıklı olarak istatistiksel yöntemlerdir. Diğer yöntemler de genellikle istatistiği temel alan ama daha çok makine öğrenimi ve yapay zekâ destekli yeni nesil yöntemlerdir.

Veri Madenciliğinde kullanılan klasik yöntemlerin başlıcaları;

• Regresyon,

• K - En Yakın Komşuluk,

• Kümeleme, olarak sayılabilir.

Yeni nesil yöntemlerin başlıcaları ise;

• Karar Ağaçları,

• Birliktelik Kuralları,

• Sinir Ağları, olarak sıralanabilir

Ayrıca diğer Veri Madenciliği yöntemlerinin başlıcaları da;

• Temel Bileşenler Analizi,

• Diskriminant Analizi,

• Faktör Analizi,

• Kohonen Ağları,

• Bulanık Mantığa Dayalı Yöntemler,

• Genetik Algoritmalar,

• Bayesci Ağlar,

• Pürüzlü (Rough) Küme Teorisine Dayalı yöntemler, olarak sıralanabilir.

VERİ MADENCİLİĞİ TEKNİKLERİ

2 KARAR AĞACI ALGORİTMALARI

Veri madenciliği denildiğinde, sinir ağları ile birlikte ilk akla gelen yöntemlerden olan karar ağaçları, yeni jenerasyon veri madenciliği yöntemlerindendir. Bir ağaç diyagramı biçiminde, her bir dal ve yaprağı bir sınıflandırma sorgusu olacak biçimde dallanan yöntem; nitel, nicel, sürekli, kesikli tüm değişkenlere uygulabilen algoritmaları, ağaç diyagramı şeklindeki görsel desteği, SQL sorgusuna kolay dönüştürülebilir yapısıyla en popüler segmentasyon yöntemlerinden birisidir. C 4.5, C5.0, C&RT ve CHAID en popüler yöntemlerdir.

2.1 K-EN YAKIN KOMŞU ALGORİTMASI

2.2 YAPAY SİNİR AĞLARI

İnsan beyninin hesaplama mantığı baz alınarak oluşturulmuş (yapay) sinir ağları, karar ağaçları gibi yeni jenerasyon veri madenciliği yöntemlerindendir. Girdi ve çıktı arasında, küçük hesaplama birimlerinden elde edilen sonuçları birleştirerek sonuçlandıran bir modelleme yöntemidir. Karar ağaçları uygulama, anlama ve yorumlama açısından ne kadar kolaysa, sinir ağları da o derece zordur. Yalnızca model oluşturma, sonuçları yorumlama aşamasının ötesinde; doğru bir model kurabilmek için ağın eğitimindeki dengenin önemi oldukça büyüktür. Fazla eğitilmiş bir ağ, önceden gözlenmemiş bir gözleme yönelik tahmin kabiliyetini yitirirken; az eğitilmiş bir ağ ise yanlış tahmin verebilmektedir.

2.3 KARAR DESTEK MAKİNA SİSTEMLERİ

Karar destek makine ilk olarak 1995 yılında girdi bileşenli vektör olarak sınıflandırmada kullanılmak amacıyla ortaya çıkmıştır. Karar destek makine nesneleri, nesnelerin doğrusal ayrımlarına bakarak iki sınıf altında toplamaktadır. Bu doğrusal ayrıma karar vermede nesneler uzayında yer alan nesnelerin hiperdüzlemleri etki etmektedir. Karar destek makinede temel mantık, maksimum genelleme seviyesini sağlayan hiperdüzlemi bulmak ve aşırı uyumdan kaçınmaktır. Eğitim setlerinde minimal uzaklığa sahip maksimum marjinli hiperdüzleme destek vektörü denilmektedir. Maksimum marjinli hiperdüzlemin pozisyonunu bulmak ve destek vektörü saptamak amacıyla dual optimizasyon problemi kurgulanmaktadır.

2.4 NAİVE BAYES SINIFLANDIRICI

Bayes sınıflandırma tekniği, elde var olan, hali hazırda sınıflanmış verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine girme olasılığını hesaplayan bir yöntemdir. Belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar. Belirsizlik taşıyan durumlarda karar verme konusunda çok kullanışlıdır.

2.5 LOJİSTİK REGRESYON

Doğrusal regresyonda değişkenler arasındaki ilişkiler incelenirken, bu değişkenlerin kategorik olmasından çok, sürekli olması tercih edilmektedir. Kategorik verilerin olduğu durumlarda, doğrusal regresyona benzeyen bir yöntem olan lojistik regresyon ortaya çıkmaktadır. Lojistik regresyonda, kategorik veriler arasındaki ilişkiler ortaya koyulmaktadır (Larose, 2006). Lojistik regresyon yöntemi günümüzde özellikle tıp alanında yaygın olarak kullanılmaya başlanmıştır. Lojistik regresyon, diskriminant analizi ve crosstabs yöntemine alternatif olarak uygulanmaktadır. Doğrusal regresyonda

değişkenler arasındaki ilişki incelendiği gibi (Şıklar, 2000:5), lojistik regresyonda da ilişkiler incelenmektedir. Ancak lojistik regresyonu doğrusal regresyondan ayıran en önemli fark, bağımlı değişkenin kategorik olmasıdır. Lojistik regresyondaki bu fark hipotezlerde kendisini göstermektedir (Bircan, 2004).

2.6 BİRLİKTELİK KURALLARI

Gözlem değerleri arasındaki ilişkiyi, koşullu olasılık bazlı değerlendirmelerle özet olarak sunan ve uygulayıcı tarafından baştan tanımlanmış bir başarı oranının üzerindeki kuralları sıralayan bir yaklaşım izlenmektedir. Hesaplama mantığı nedeniyle hızlı sonuç vermesi ve çok büyük veri setlerine kolaylıkla uygulanabilmesi Birliktelik Kuralı Analizi'ni ticari veri tabanlarının madenciliğinde gittikçe popülerleşen bir araçtır haline getirmiştir.

KALP YETMEZLİĞİ HASTALIĞI

3 KALP YETMEZLİĞİ NEDİR?

Kalp yetmezliği (KY), kalp vücudun ihtiyaçlarını karşılamak için yeterli kanı pompalayamadığında ortaya çıkar. Özellikle kalp yetmezliği, kalp vücuda yeterince kan pompalayamadığında ortaya çıkar ve genellikle diyabet, yüksek tansiyon veya diğer kalp rahatsızlıkları veya hastalıklarından kaynaklanır.

3.1 KALP YETMEZLİĞİ TANILARI

Yapılan araştırmalar neticesinde yaş, anemi, yüksek kan basıncı, kreatinin fosfokinaz, şeker hastalığı, ejeksiyon fraksiyonu, cinsiyet, trombositler, serum kreatinin, serum sodyum, sıgara içmek, zaman ve hedef (ölüm olayıdır).

3.2 KALP YETMEZLİĞİ TEŞHİS

Hızlı ve doğru bir teşhis , uygun bir tedavinin belirlenmesi ve hastalık sonuçlarının iyileştirilmesi bakımından esas bir unsurdur. Ancak kalp yetmezliğinin teşhis edilmesi güçtür çünkü klinik semptomlar değişiklik göstermekte ve genellikle spesifik nitelik taşımaktadır. Teşhis, hasta öyküsünün, fiziksel muayenenin, görüntüleme ve laboratuar testlerinin bir kombinasyonuna dayalı olmalıdır. NT-proBNP ve Galectin-3 gibi biyo-göstergelere yönelik test, kalp yetmezliğinin **erken teşhisine ve prognozuna** olanak tanımaktadır[3,4]. Prokalsitonin (PCT) testi, akut kalp yetmezliğinden şikayetçi hastalarda **eşlik eden bakteriyel zatürreenin** teşhis edilmesi bakımından faydalıdır.[5]

3.3 ÖNLEME

Kalp yetmezliğini **önlemenin** en iyi yolu, yüksek kan basıncı ve koroner arter hastalığı gibi iyi bilinen risk faktörlerini azaltan **sağlıklı bir yaşam tarzı benimsemektir**.

Kalp yetmezliğine yakalanmamak için başvurabileceğiniz yöntemlerden bazıları aşağıda sıralanmıştır:

- Sigara içmemek
- Fiziksel olarak aktif olmak
- Şeker, tuz ve doymuş yağlar bakımından düşük ve dengeli bir beslenme düzenine sahip olmak
- Sağlıklı bir kiloda kalmak
- Stresi azaltmak ve kontrol altına almak

- Yüksek kan basıncı, yüksek kolesterol ve diyabet gibi belirli koşullara sahipseniz veya risk grubundaysanız bu koşulları kontrol altına almak ve gerekli görüldüğü takdirde ilaç tedavisine başlamak için doktorunuza danışınız

3.4 TEDAVİ

Kalp yetmezliği için önerilen **tedavi**, hastalığın evresine ve ciddiyetine bağlıdır. Terapötik yönetim, risk faktörlerini kontrol altına almakla birlikte temelde yatan nedenleri/kötüleştirici faktörleri ele almaktadır. Burada amaç, semptomlarda rahatlama sağlamak ve hastalığın ilerlemesini engellemektir.

Kalp yetmezliğinde rol oynayan çeşitli patofizyolojik mekanizmaları hedef alan ve kalp, böbrekler ve periferik dolaşımı içeren bazı **bulguya dayalı ilaç tedavileri** mevcuttur.

TABLO 1: VERİ KÜMESİNİN HER BİR ÖZELLİĞİNİN ANLAMLARI, ÖLÇÜ BİRİMLERİ VE ARALIKLARI

ÖZELLİK	AÇIKLAMA	ÖLÇÜM	MENZİL
YAŞ	Hastanın yaşı	Yıllar	[40,..., 95]
ANEMİ	Kırmızı kan hücrelerinin veya hemoglobinin azalması	Boole	0, 1
YÜKSEK KAN BASINCI	Hastanın hipertansiyonu varsa	Boole	0, 1
KREATİNİN FOSFOKİNAZ	Kandaki CPK enziminin seviyesi	mcg/L	[23,..., 7861]
(CPK)			
ŞEKER HASTALIĞI	Hastanın şeker hastalığı varsa	boole	0, 1
EJEKSİYON FRAKSİYONU	Kan bırakma yüzdesi	Yüzde	[14,, 80]
	Her kasılmada kalp		
CİNSİYET	Kadın veya erkek	ikili	0, 1
TROMBOSİTLER	Kandaki trombositler	kiloplatelet/mL	[25.01,..., 850.00]
SERUM KREATİNİN	Kandaki kreatinin seviyesi	Mg/dL	[0.50,..., 9.40]
SERUM SODYUM	Kandaki sodyum seviyesi	mEq/L	[114,..., 148]
SİGARA İÇMEK	Hasta sigara içiyorsa	Boole	0, 1
ZAMAN	Takip dönemi	Günler	[4,...,285]

ÖLÜM OLAYI	Hasta takip döneminde öldüyse	Boole	0, 1

mcg/L: litre başına mikrogram. ml: mikrolitre. mEq/L: litre başına milieşdeğer

- KREATİNİN FOSFOKİNAZ (CPK)

  Kandaki CPK enziminin seviyesini belirtir. Bir kas dokusu hasar gördüğünde, CPK kana akar. Bu nedenle, bir hastanın kanındaki yüksek CPK seviyeleri, kalp yetmezliği veya yaralanmayı gösterebilir.

- EJEKSİYON FRAKSİYONU

  Her kasılma ile sol ventrikülün ne kadar kan pompaladığının yüzdesini belirtir.

- SERUM KREATİNİN

Özellikle doktorlar böbrek fonksiyonunu kontrol etmek için kandaki serum kreatininine odaklanır. Bir hastada yüksek serum kreatinin seviyeleri varsa, böbrek fonksiyon bozukluğunu gösterebilir.

- SERUM SODYUM

Hastanın kanında normal sodyum düzeyleri olup olmadığını gösteren rutin bir kan muayenesidir. Kandaki anormal derecede düşük sodyum seviyesi kalp yetmezliğinden kaynaklanabilir.

- ÖLÜM OLAYI

Özelliği, hastanın ortalama 130 gün olan takip süresi sona ermeden öldü mü veya hayatta mı kaldığını belirtir.

Hayatta kalan hastalar (ölüm olayı = 0) 203 iken ölen hastalar (ölüm olayı = 1) 96'dır. İstatistiksel olarak, %32,11 pozitif ve %67,89 negatif vardır.

Bu veri setini 299 satır (hasta) ve 13 sütun (özellikler) içeren bir tablo olarak sunduk.

TABLO2: KATEGORİ ÖZELLİKLERİNİN İSTATİKSEL NİCEL ANLAŞILMASI

	TAM ÖRNEK		ÖLÜ HASTALAR		HAYATTA KALAN HASTALAR	
KATEGORİ ÖZELLİĞİ	#	%	#	%	#	%
• Anemi (0: anemi yok)	170	56.86	50	52.08	120	59.11
• Anemi (1: anemi var)	129	43.14	46	47.92	3	40.89
• Yüksek tansiyon (1: yok)	194	64.88	57	59.38	137	67.49
• Yüksek tansiyon (0: var)	105	35.12	39	40.62	66	32.51
• Diyabet (0: d.var)	174	58.19	56	58.33	118	58.13
• Diyabet (1: d.yok)	125	41.81	40	41.67	85	41.87
• Cinsiyet (1: kadın)	105	35.12	34	35.42	71	34.98
• Cinsiyet (0: erkek)	194	64.88	62	64.58	132	65.02
• Sigara (0: içiyor)	203	67.89	66	68.75	137	67.49
• Sigara (1: içmiyor)	96	32.11	30	31.25	66	32.51

#: Hasta sayısı

%: Hastaların yüzdesi

Tam Örnek: 299 kişi

Ölü Hastalar: 96 kişi

Hayatta kalan Hastalar: 203 kişi

4.UYGULAMA: KALP YETMEZLİĞİ BULUNAN HASTALARA İLİŞKİN VERİ ANALATİĞİ

4.1PROBLEMİN TANIMLANMASI

2015 yılında toplanan kalp yetmezliği olan 299 hastadan oluşan bir veri setini analiz ediyoruz. Hem hastaların sağ kalımını tahmin etmek hem de en önemli risk faktörlerine karşılık gelen özellikleri sıralamak için çeşitli makine öğrenimi sınıflandırıcıları uyguluyoruz.

4.2VERİ SETİNİ ANLAMA

Kullanılan veri seti UCI Machine Learning Repository sitesinden temin edilmiştir.Veri seti incelenmiştir ve toplam 12 değişken belirlenmiştir.

Veri setinde 6 adet nümerik değişken bulunmaktadır. Bunlar; yaş, kreatinin fosfokinaz, ejeksiyon fraksiyonu, trombositler, serum sodyum ve zaman'dır. Geri kalan değişkenler kategorik değişkenlerdir.

Anemi değişkeninde 1= anemi var 0= anemi yok,

Cinsiyet değişkeninde 1=kadın, 0= erkek,

Diyabet değişkeninde 1=diyabet yok 0=diyabet var,

Yüksek kan basıncı değişkeninde 0= var 1=yok,

Sigara değişkeninde 0=içiyor, 1= içmiyor,

Ölüm olayı değişkeninde 0= ölüm yok 1= ölüm var.

TABLO3: VERİ SETİNDE BULUNAN NİTELİKLERE AİT ÖZELLİKLER

	TAHMİN İÇİN KULLANILAN VERİNİN YAPISI		
	DEĞİŞKEN	VERİ TİPİ	VERİ SETİNDE GÖSTERİMİ
1	➢ Yaş	Nümerik	
2	➢ Anemi	Kategorik	0: Anemi yok 1: Anemi var
3	➢ Kreatinin fosfokinaz	Nümerik	
4	➢ Diyabet	Kategorik	0: var 1: yok
5	➢ Ejeksiyon fraksiyonu	Nümerik	
6	➢ Yüksek kan basıncı	Kategorik	0: yok 1: var
7	➢ Trombositler	Nümerik	
8	➢ Serum sodyum	Nümerik	
9	➢ Cinsiyet	Kategorik	0: Erkek 1: Kadın
10	➢ Sigara	Kategorik	0: İçiyor 1: İçmiyor
11	➢ Zaman	Nümerik	

12	➢ Ölüm olayı	Kategorik	0: Ölüm yok 1: Ölüm var

4.3ANALİZE HAZIRLIK

Bundan sonraki aşamalar R studio'da yapılmıştır. Uygulama kodları eklerdedir.

Veri seti toplam 299 gözlem ve 12 değişkenden oluşmaktadır. Değişkenler sırasıyla: yaş, anemi, kreatinin fosfokinaz, diyabet, ejeksiyon fraksiyonu, yüksek kan basıncı, trombositler, serum sodyum, cinsiyet, sigara, zaman, ölüm olayı.

Öncelikle veri setinin yapısı incelenmiş ve değişkenler nümerik ve faktör şeklinde düzenlenmiştir. Düzenlendikten sonra şu hale dönüşmüştür:

'data.frame' :299 obs. of 12 variables:

$ yas :num 75 55 65 50 65 90 75 60 65 80 ...

$ anemi :Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 1 2 ...

$ kreatinin_.fosfokinaz :num 582 7861 146 111 160 ...

$ diyabet : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...

$ ejeksiyon_fraksiyonu : num 20 38 20 20 20 40 15 60 65 35 ...

$ yüksek_kan_basıncı : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...

$ trombositler : num 265000 263358 162000 210000 327000 ...

$ serum_sodyum : num 130 136 129 137 116 132 137 131 138 133 ...

$ cinsiyet : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 1 2 ...

$ sigara : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...

$ zaman : num 4 6 7 7 8 8 10 10 10 10 ...

$ ölüm_olayı : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

Veri setinin özeti Tabloda verilmiştir. Bu tabloda kategorik değişkenler ve nümerik değişkenlerin minimum değerleri, 1. kartil, medyan, ortalama, 3. kartil ve maksimum değerleri görülür.

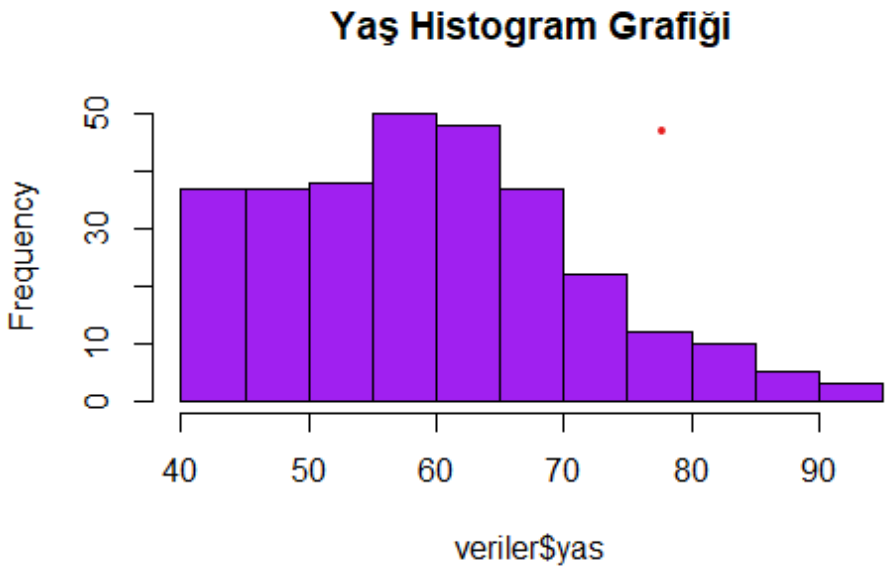
Veri setinin özeti aşağıdaki tabloda verilmiştir.

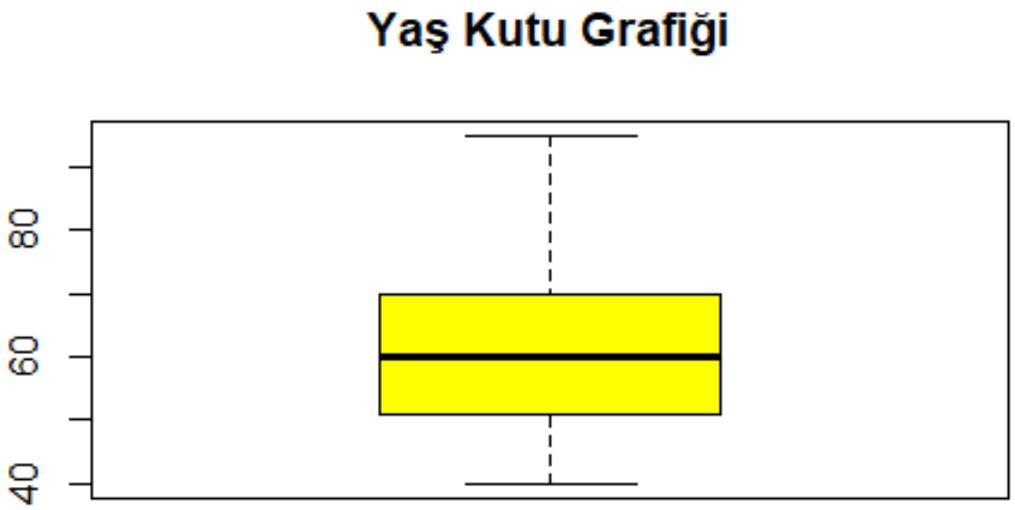
YAŞ	ANEMİ	KREATİNİN FOSFOKİNAZ	DİYABET	EJEKSİYON FRAKSİYONU	YÜKSEK KAN BASINCI	TROMBOSİTLER	SERUM SODYUM	CİNSİYET	SİGARA	ZAMAN	ÖLÜM OLAYI
Min. :40.00	0:170	Min. : 23.0	0:174	Min. :14.00	0:194	Min. : 25100	Min. :113.0	0:105	0:203	Min. : 4.0	0:203
1st Qu. :51.00	1:129	1st Qu. : 116.5	1:125	1st Qu.:30.00	1:105	1st Qu. :212500	1st Qu. :134.0	1:194	1: 96	1st Qu.: 73.0	1: 96
Median :60.00		Median : 250.0		Median :38.00		Median :262000	Median :137.0			Median :115.0	
Mean :60.83		Mean : 581.8		Mean :38.08		Mean :263358	Mean :136.6			Mean :130.3	
3rd. Qu:70.00		3rd Qu.: 582.0		3rd Qu.:45.00		3rd Qu.:303500	3rd Qu.:140.0			3rd Qu.:203.0	

Max. :95.00		Max. : 7861.0		Max. :80.00		Max. :850000	Max. :148.0			Max. :285.0	

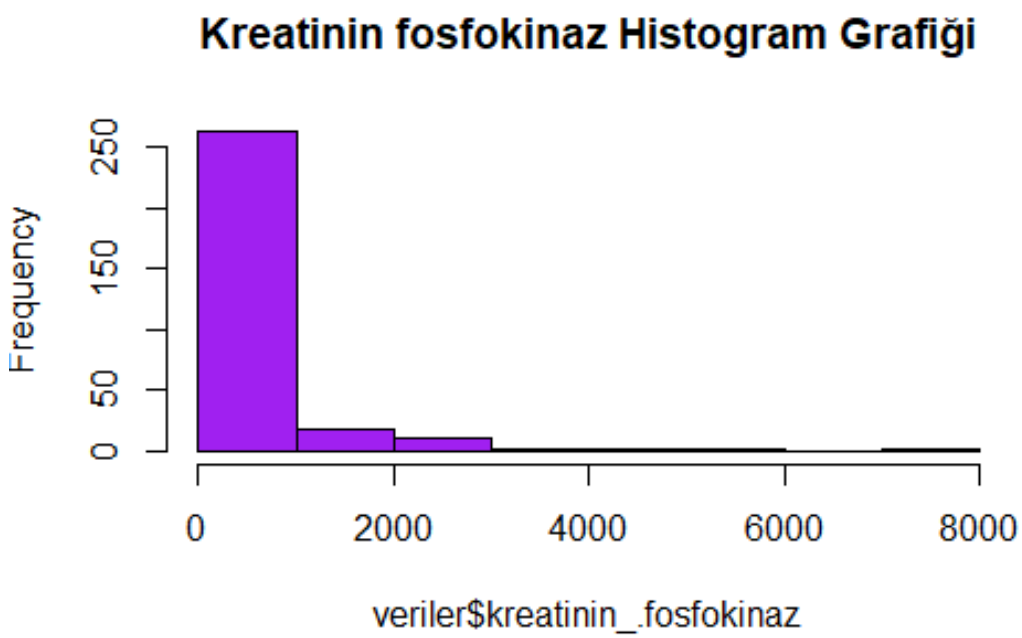
Veri setindeki değişkenler tek tek incelenmiştir. Değişkenlerin her birine uygun grafikler çizilmiştir.

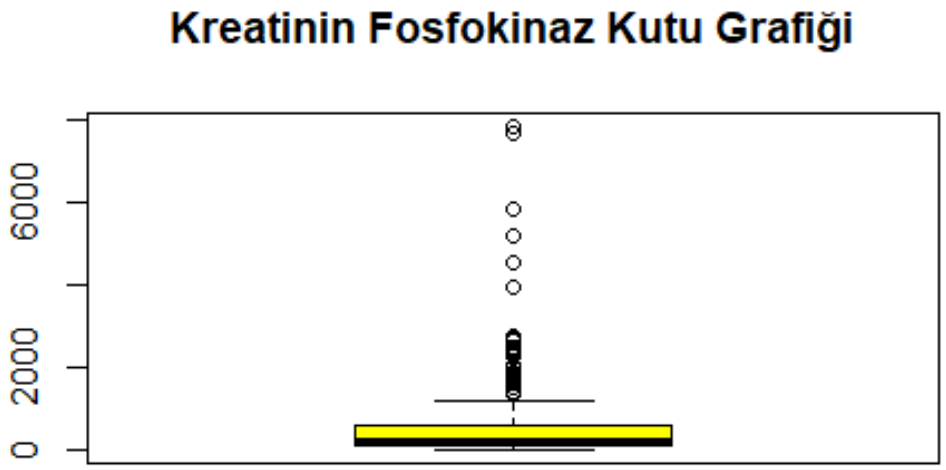
Nümerik değişkenler için Histogram ve Kutu grafikleri çizilmiştir.



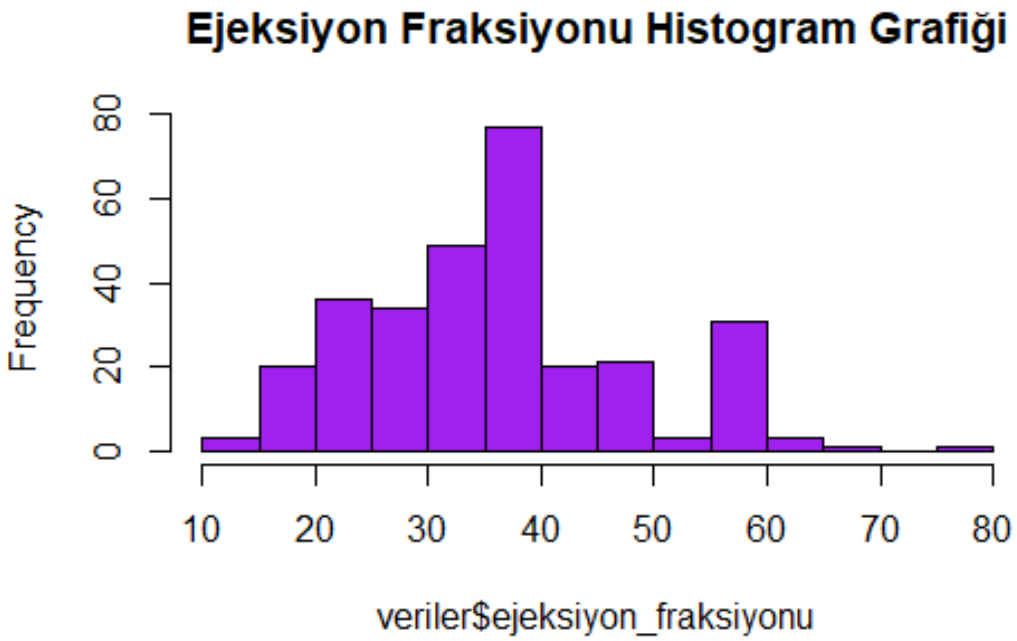


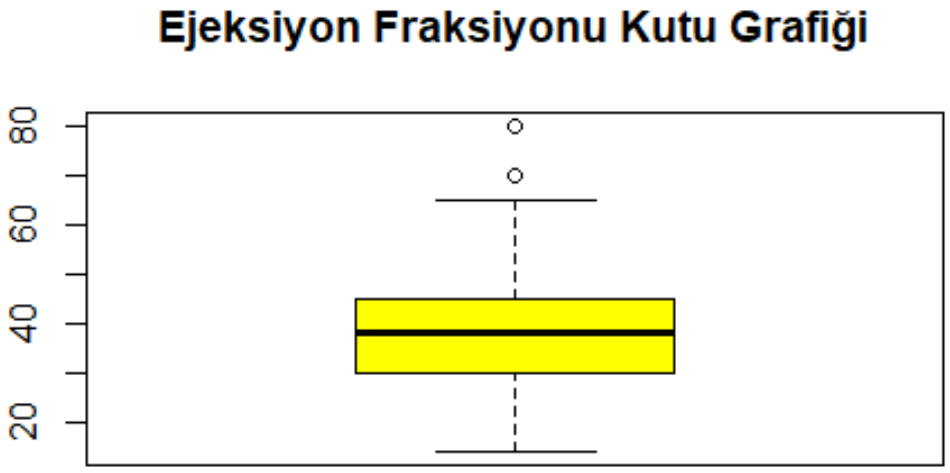
Veri setinde yaş değişim aralığı en küçük 40 yaşındaki hasta ile en büyük 95 yaşındaki hasta arasındadır. Frekanslarına bakıldığında 60 yaş civarı hasta sayısının çok olduğu görülmektedir.



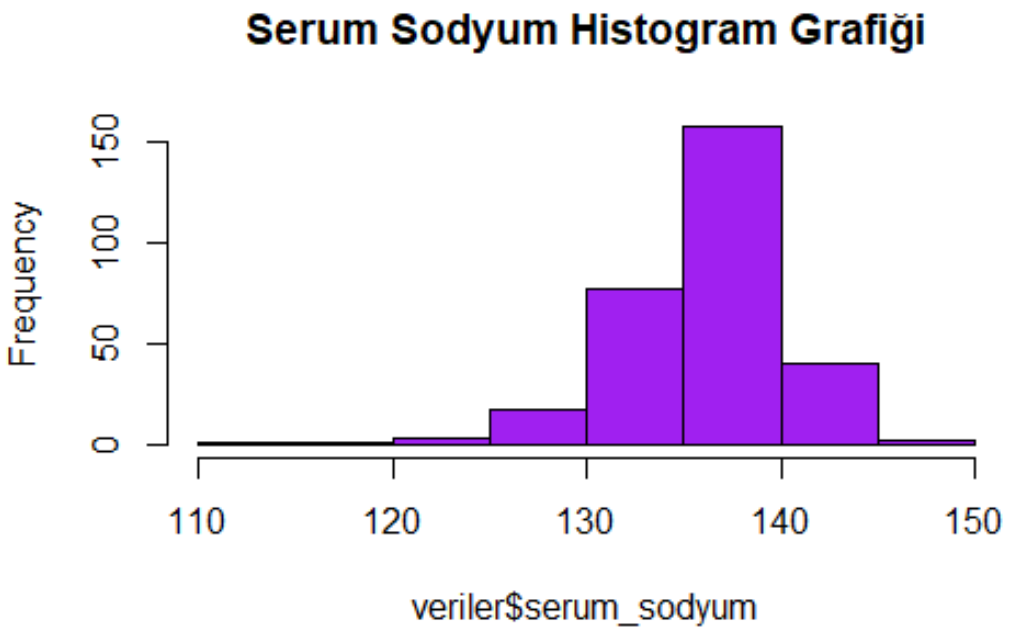


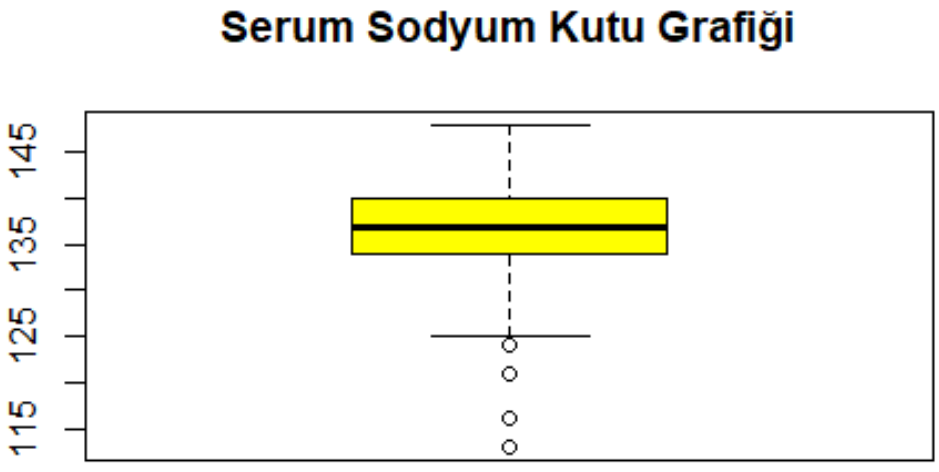
Frekanslara bakıldığında 250 mcg/L değerinin en çok tekrar eden değer olduğu görülmektedir.



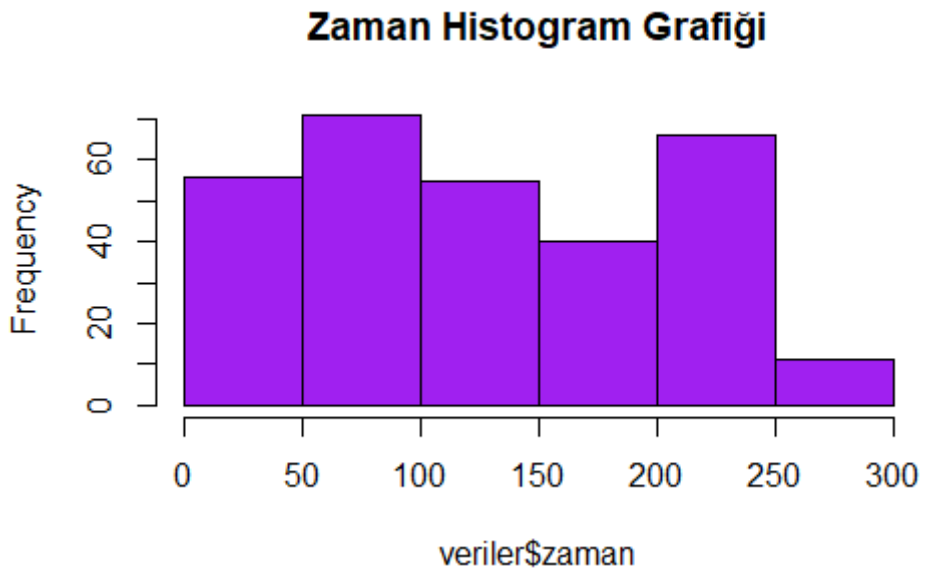


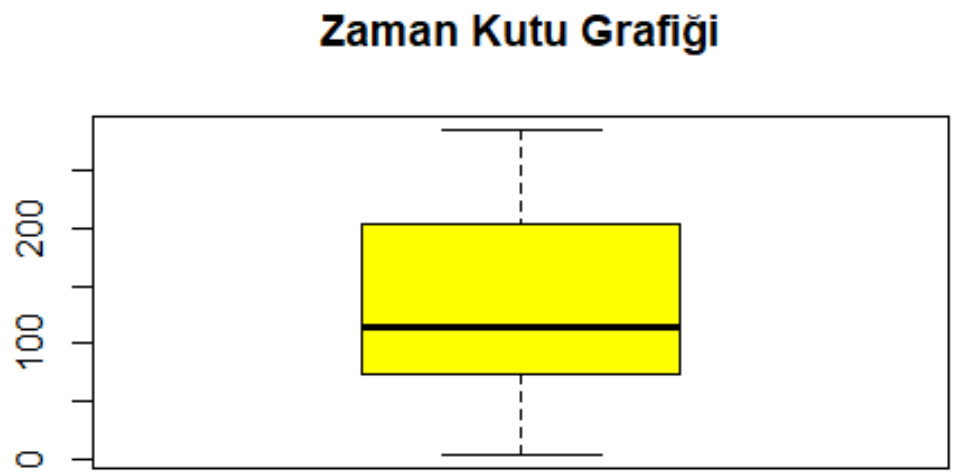
Frekanslara bakıldığında yüzde 40 değerinin en çok tekrar eden değer olduğu görülmektedir.



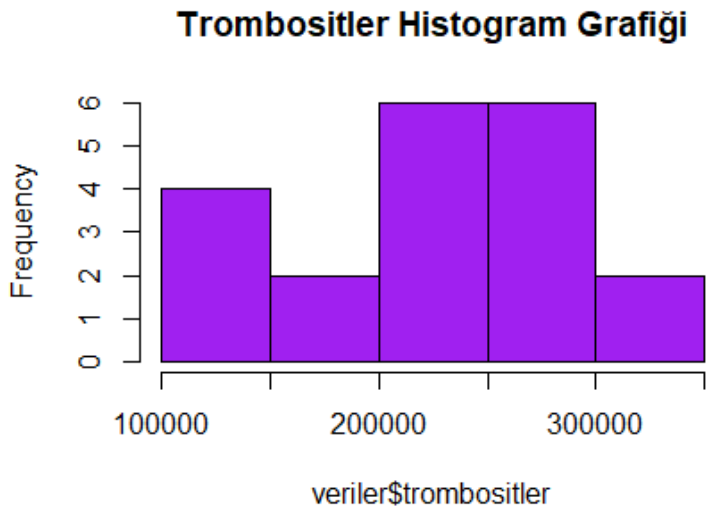


Frekanslara bakıldığında 135-140 mEg/L değerlerinin en çok tekrar eden değerler olduğu belirlenmiştir.

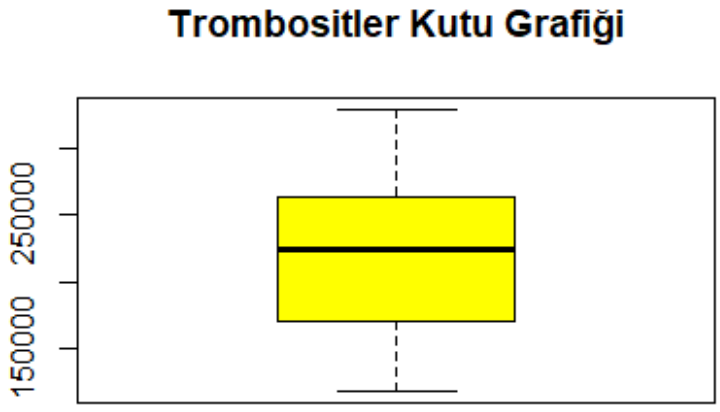


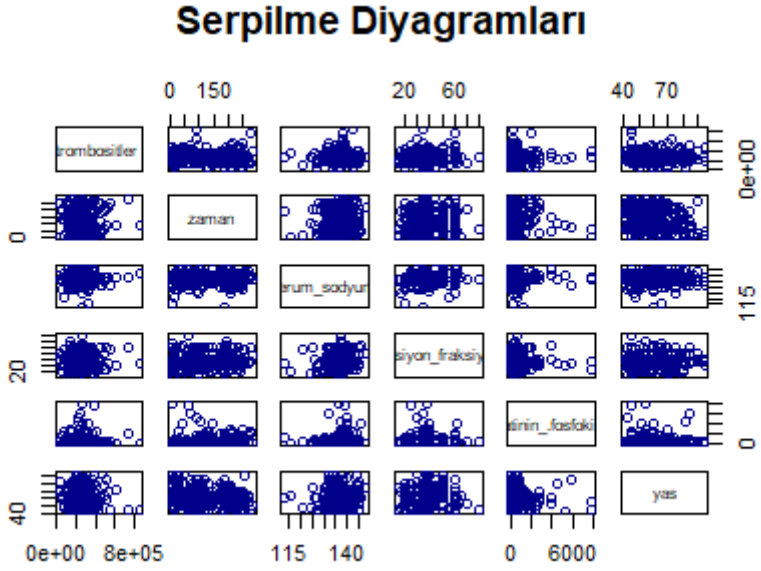


Frekans değerlerine bakıldığında 50-100 arasındaki değerlerin en çok tekrar eden değerler olduğu belirlenmiştir.

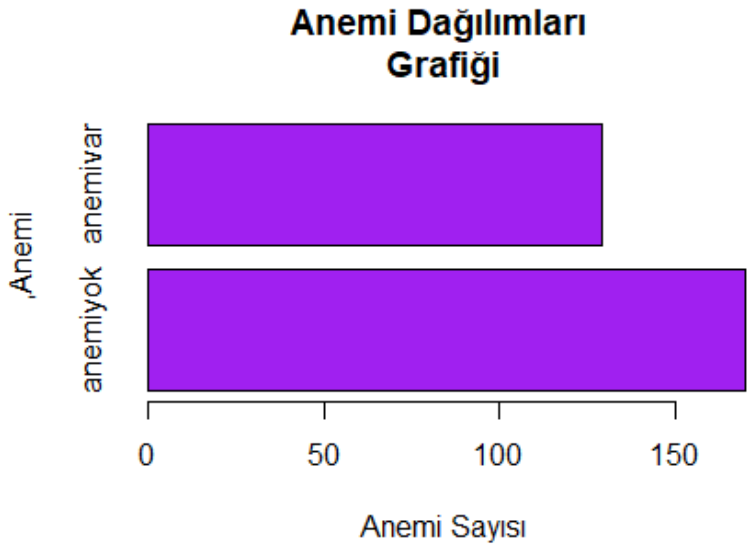


Frekans değerlerine bakıldığında 200000 ile 300000 arasındaki değerler en çok tekrar eden değerlerdir.

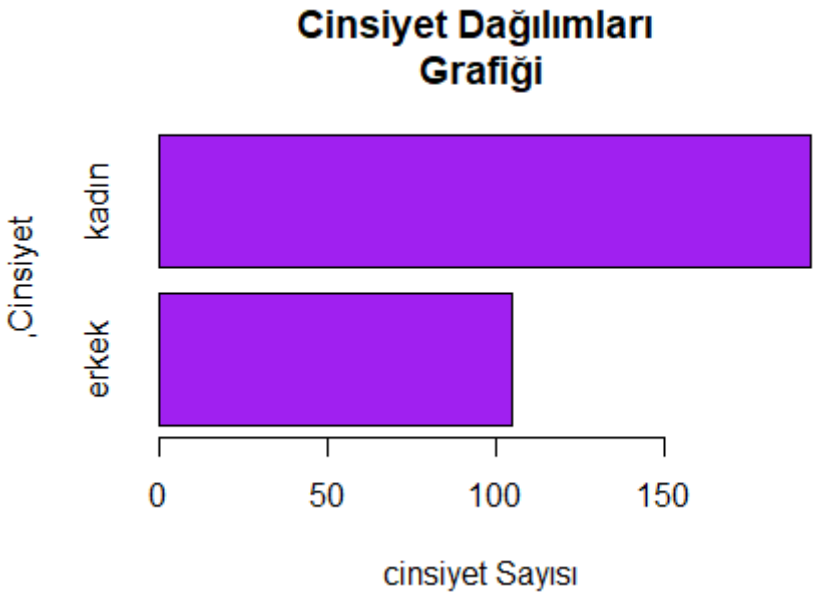




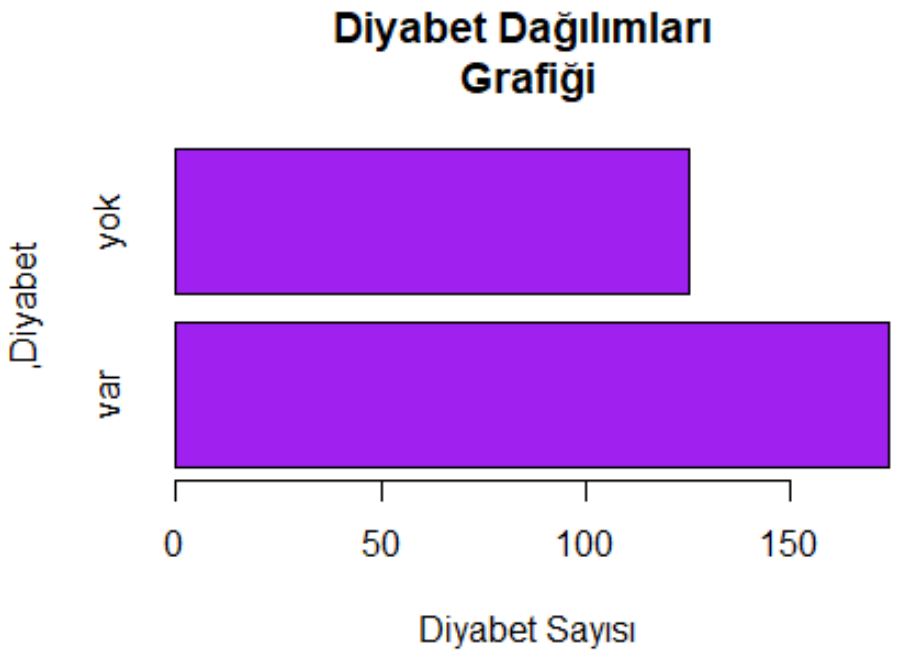
Kategorik değişkenler için Dağılım grafikleri çizilmiştir.



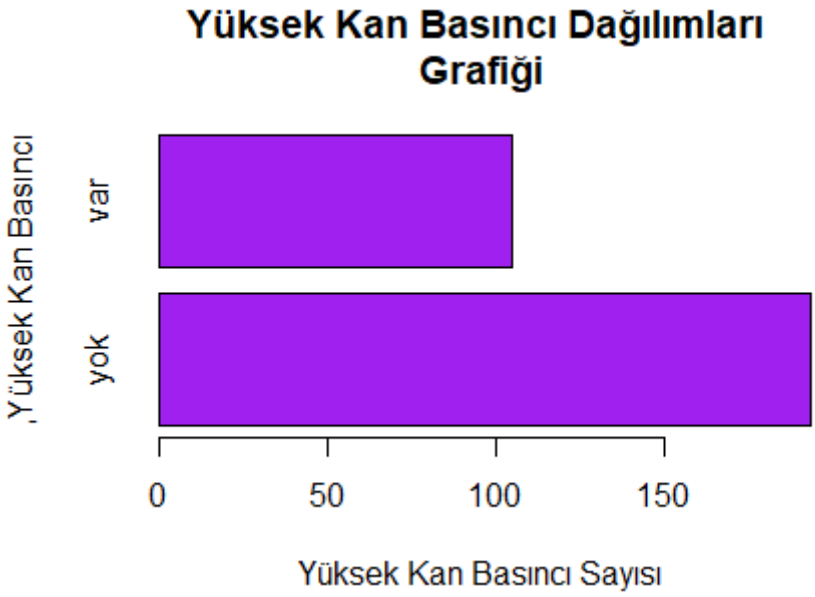
299 hastanın 170 tanesinde anemi yoktur 129 tanesinde anemi vardır. Veri setinin %56,86'da anemi tanısı yoktur, %43,14 'de anemi tanısı konulan hastalardan oluşmaktadır.



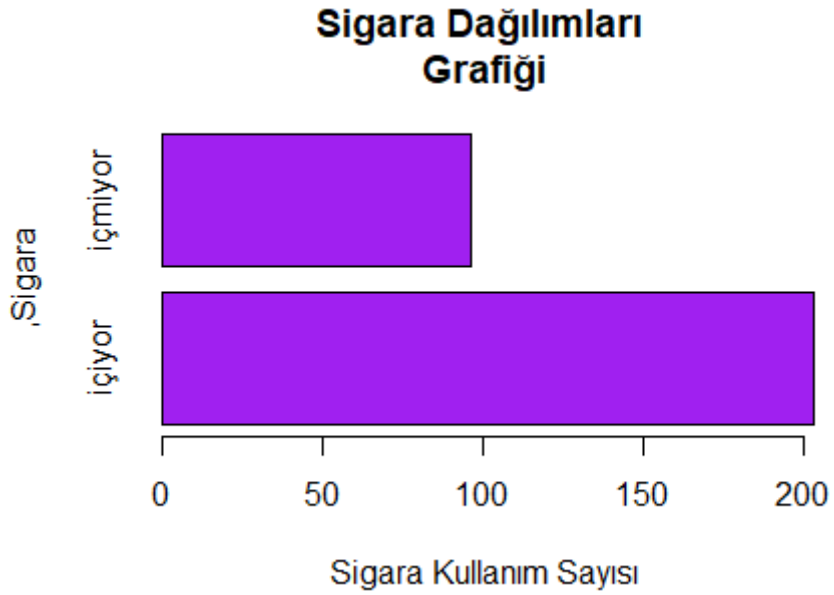
299 hastanın 194 tanesi kadın, 105 tanesi erkektir. %64,88 kadın hastadan, %35,12 erkek hastalardan oluşmaktadır.



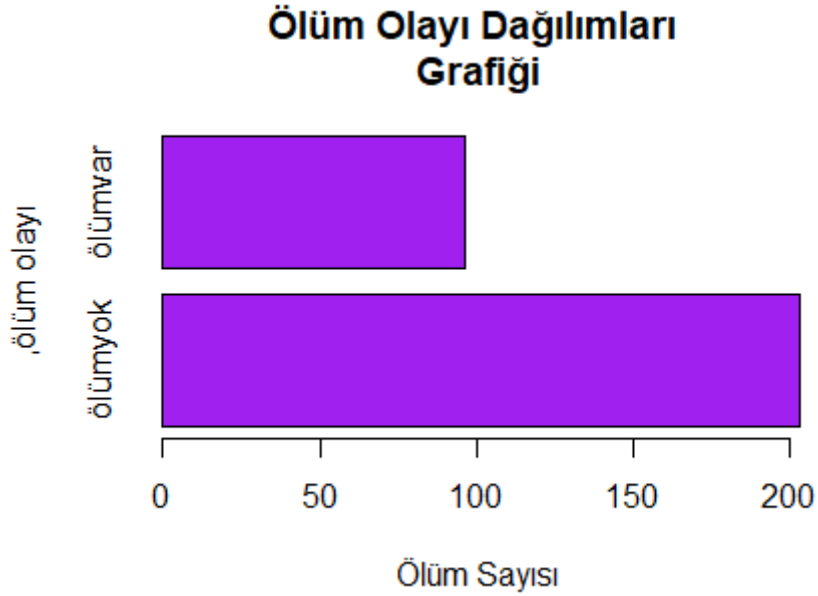
299 Hastanın 174 kişide diyabet var, 125 kişide diyabet yoktur. %58,19 diyabet tanısı konulan hastalardan oluşmaktadır. %41,81'de diyabet tanısı konulmayan hastalardan oluşmaktadır.



299 Hastanın 194 kişide yüksek kan basıncı yoktur, 105 kişide vardır. %64,88 yüksek kan basıncı tanısı olmayan hastalardır. %35,12 ise yüksek kan basıncı tanısı konulan hastalardan oluşmaktadır.



299 Hastadan 203 kişi sigara içiyor,96 kişi sigara içmiyor. %67,89 oranı sigara kullanan hastalardan oluşmaktadır. %32,11 oranı sigara kullanmayan hastalardan oluşuyor.



299 hastadan 203 kişi hayattadır,96 kişi vefat etmiştir.

4.4 C4.5 ALGORİTMASI

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.

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

Veri setine uygulanan karar ağacı algoritması sonuçları verilmiştir.

Burada correctly classified instances doğru yerleşen tahmin sayısıdır. Bunun toplam 299 kişi içinden 283 kişi olduğu gözükmektedir ve %94.6488 doğruluk oranına sahiptir. Modelin oluşturduğu ağaç şu şekildedir:

```
> summary(modelC11)

=== Summary ===

Correctly Classified Instances         283               94.6488 %
Kappa statistic                          0.8731
Mean absolute error                      0.0936
Root mean squared error                  0.2164
Relative absolute error                 21.452  %
Root relative squared error             46.3388 %
Total Number of Instances              299

=== Confusion Matrix ===

   a   b   <-- classified as
 201   2 |   a = ölümyok
  14  82 |   b = ölümvar
> |
```

Karar ağacından elde edilen kurallar bu şekildedir:

```
> print(modelC11)
J48 pruned tree
------------------

zaman <= 73
|   serum_sodyum <= 136
|   |   ejeksiyon_fraksiyonu <= 45: ölümvar (32.0)
|   |   ejeksiyon_fraksiyonu > 45
|   |   |   anemi = anemiyok: ölümyok (2.0)
|   |   |   anemi = anemivar: ölümvar (5.0)
|   serum_sodyum > 136
|   |   serum_sodyum <= 139
|   |   |   yüksek_kan_basıncı = yok
|   |   |   |   zaman <= 20: ölümvar (4.0)
|   |   |   |   zaman > 20: ölümyok (9.0/1.0)
|   |   |   yüksek_kan_basıncı = var
|   |   |   |   kreatinin_.fosfokinaz <= 84: ölümyok (2.0)
|   |   |   |   kreatinin_.fosfokinaz > 84: ölümvar (7.0)
|   |   serum_sodyum > 139: ölümvar (15.0/1.0)
zaman > 73
|   ejeksiyon_fraksiyonu <= 30
|   |   ejeksiyon_fraksiyonu <= 20
|   |   |   anemi = anemiyok: ölümvar (6.0/1.0)
|   |   |   anemi = anemivar: ölümyok (2.0)
|   |   ejeksiyon_fraksiyonu > 20
|   |   |   anemi = anemiyok: ölümyok (27.0/5.0)
|   |   |   anemi = anemivar
|   |   |   |   trombositler <= 194000: ölümyok (6.0)
|   |   |   |   trombositler > 194000
|   |   |   |   |   trombositler <= 228000: ölümvar (6.0)
|   |   |   |   |   trombositler > 228000
|   |   |   |   |   |   trombositler <= 336000: ölümyok (9.0/2.0)
|   |   |   |   |   |   trombositler > 336000: ölümvar (3.0)


|   ejeksiyon_fraksiyonu > 30
|   |   yas <= 70: ölümyok (140.0/5.0)
|   |   yas > 70
|   |   |   anemi = anemiyok
|   |   |   |   serum_sodyum <= 135
|   |   |   |   |   zaman <= 162: ölümvar (6.0)
|   |   |   |   |   zaman > 162: ölümyok (3.0)
|   |   |   |   serum_sodyum > 135: ölümyok (6.0)
|   |   |   anemi = anemivar: ölümyok (9.0/1.0)

Number of Leaves  :     20

Size of the tree :      39
```

Örneğin; anemi= varsa: anemivar: ölümyok(9.0/1.0) Burada anemi olan hastalarda ölüm olayı yoktur kuralı elde edilmiştir. Karar ağacından elde edilen kurallar şu şekildedir:

KURAL1:

zaman <= 73, serum_sodyum <= 136, ejeksiyon_fraksiyonu <= 45 ise ölüm vardır.

KURAL2:

zaman <= 73, serum_sodyum <= 136, ejeksiyon_fraksiyonu > 45 , anemi yok ise ölüm yoktur.

KURAL3:

zaman <= 73, serum_sodyum <= 136, ejeksiyon_fraksiyonu > 45 , anemi var ise ölüm vardır.

KURAL4:

serum_sodyum > 136, serum_sodyum <= 139, yüksek_kan_basıncı = yok, zaman <= 20 ise ölüm vardır.

KURAL5:

serum_sodyum > 136, serum_sodyum <= 139, yüksek_kan_basıncı = yok, zaman > 20 ise ölüm yoktur.

KURAL6:

serum_sodyum <= 139, yüksek_kan_basıncı = var, kreatinin_.fosfokinaz <= 84 ise ölüm yoktur.

KURAL7:

serum_sodyum <= 139, yüksek_kan_basıncı = var, kreatinin_.fosfokinaz > 84 ise ölüm vardır.

KURAL8:

serum_sodyum > 139 ise ölüm vardır.

KURAL9:

zaman > 73, ejeksiyon_fraksiyonu <= 30, anemiyok, ölüm vardır.

KURAL10:

zaman > 73, ejeksiyon_fraksiyonu <= 30, anemivar, ölüm yoktur.

KURAL11:

ejeksiyon_fraksiyonu > 20, anemiyok ise ölüm yoktur.

KURAL12:

trombositler <= 194000 ise ölüm yoktur.

KURAL13:

trombositler <= 228000 ise ölüm vardır.

KURAL14:

trombositler <= 336000 ise ölüm yoktur.

KURAL15:

trombositler > 336000: ölüm vardır

KURAL16:

ejeksiyon_fraksiyonu > 30, yas <= 70: ölüm yoktur.

KURAL17:

Anemi yok, serum_sodyum <= 135, zaman <= 162: ölüm vardır.

KURAL18:

Anemi yok, serum_sodyum <= 135, zaman> 162: ölüm yoktur.

KURAL19:

serum_sodyum > 135: ölüm yoktur.

KURAL20:

yas > 70, anemivar: ölüm yoktur.

C4.5 Kontenjans Tablosu

C4.5	Gerçek		
Tahmin		ölümyok	ölümvar
	ölümyok	201	2
	ölümvar	14	82

C4.5 karar ağacı algoritması kontenjans tablosu sonuçları incelendiğinde, doğru pozitif değerinin 201 çıktığı görülmektedir. Yani gerçekte 201 hastanın ölüm olayı yoktu, model tahmininde de ölüm olayı yoktur diye 201 kişiyi doğru tahmin etmiştir.

Gerçekte 14 hastanın ölüm olayı yoktur, model tahmininde ise 14 hastanın ölüm olayı vardır diyerek yanlış tahminde bulunmuştur. Yanlış pozitif değeri 14'dür.

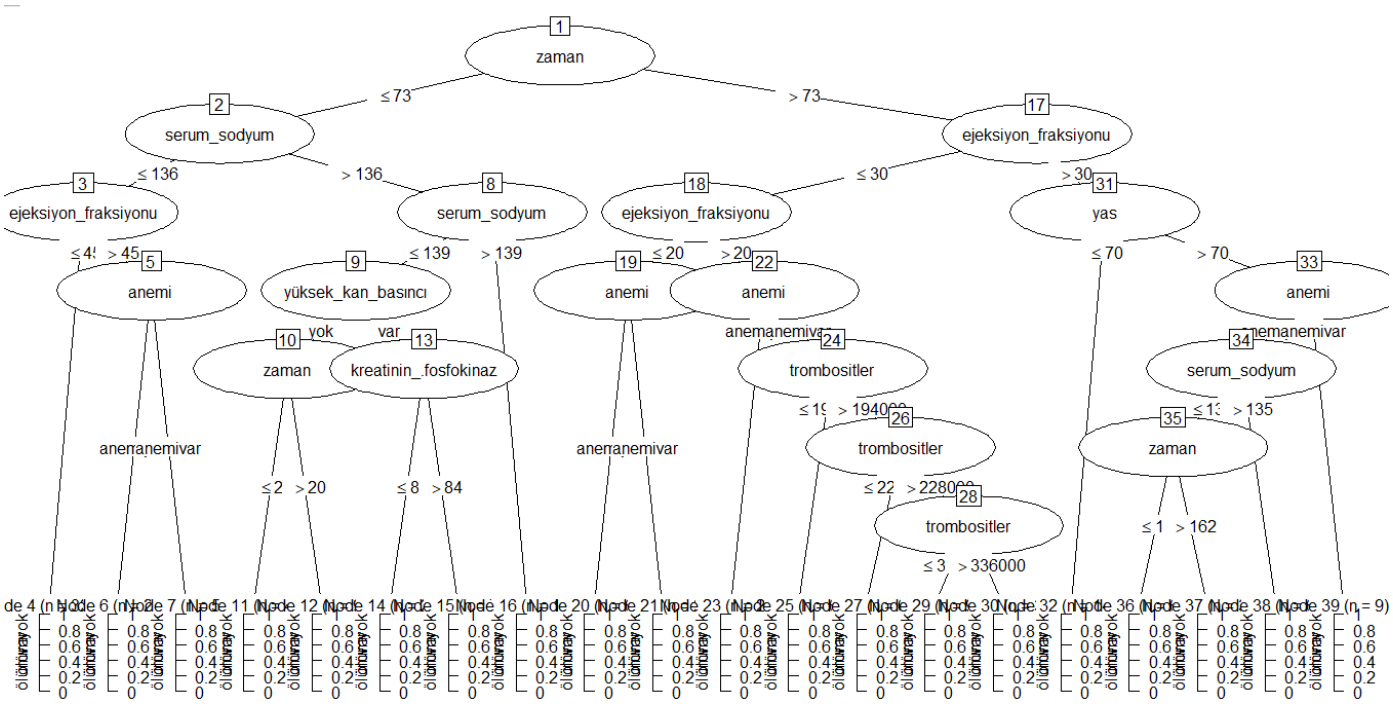
Gerçekte 82 hastanın ölüm olayı olayı var, model tahmininde de 82 hastada ölümolayı vardır diye tahmin etmiştir.

C4.5 PERFORMANS DEĞERLENDİRME ÖLÇÜTÜ

PERFORMANS DEĞERLENDİRME ÖLÇÜTÜ	C4.5
DOĞRULUK ORANI	%94.6488
HATA ORANI	%5.3512

Modelin doğruluk oranı %94.6488 ve hata oranı %5.3512 çıkmıştır. C4.5 karar ağacı algoritmasında ortaya çıkan belirleyici değişkenler; zaman,serum_sodyum, ejeksiyon_fraksiyonu, anemi, yüksek_kan_basıncı, kreatinin_fosfokinaz, trombositler değerleridir.

C4.5 KARAR AĞACI



4.5KNN ALGORİTMASI

K-NN *(K-Nearest Neighbor)* algoritması en basit ve en çok kullanılan sınıflandırma algoritmasından biridir. K-NN **non-parametric (parametrik** olmayan), **lazy** (tembel) bir öğrenme algoritmasıdır.

KNN algoritmasında sadece nümerik değere sahip olan değişkenler ve hedef nitelik olarak bir tane faktör değişken atanıp kullanılmıştır. Bunun için veri setinde nümerik değer taşıyan bir alt küme elde edilmiştir. Bu değişkenler; "yas", "kreatinin_. fosfokinaz", "ejeksiyon_fraksiyonu", "trombositler", "serum_sodyum", "zaman", "ölüm_olayı".

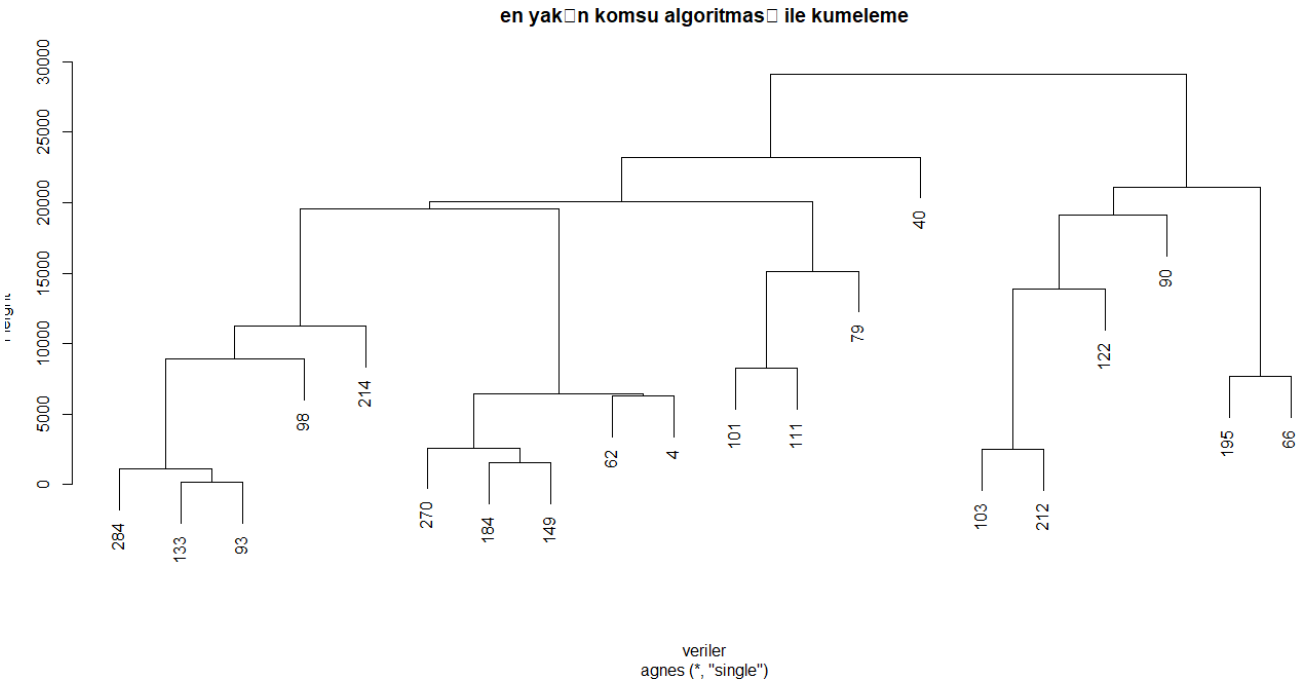
Formülü uyguladığımızda 299 veri ve 7 değişkenden oluşan yeni bir data.frame elde edilir.

En yakın komşu algoritması ve bannerplot grafiklerini görsel olarak bize vermesi için;

Sadece nümerik değerlerden oluşan ve bir hedef nitelik faktör alınan bir alt küme oluşturuldu. Oluşturulan bu alt küme ile 299 veri ve 7 değişken belirlendi.

Bu küme içerisinden set.seed(1234) fonksiyonu ile ind <- sample(1:299,20) 299 veri içerisinden rastgele 20 veri belirlenmiştir. Oluşan bu yeni alt küme ile rastgele seçilen 20 verinin ölüm_olayı belirlenmiştir.

EN YAKIN KOMŞU ALGORİTMASI İLE KÜMELEME

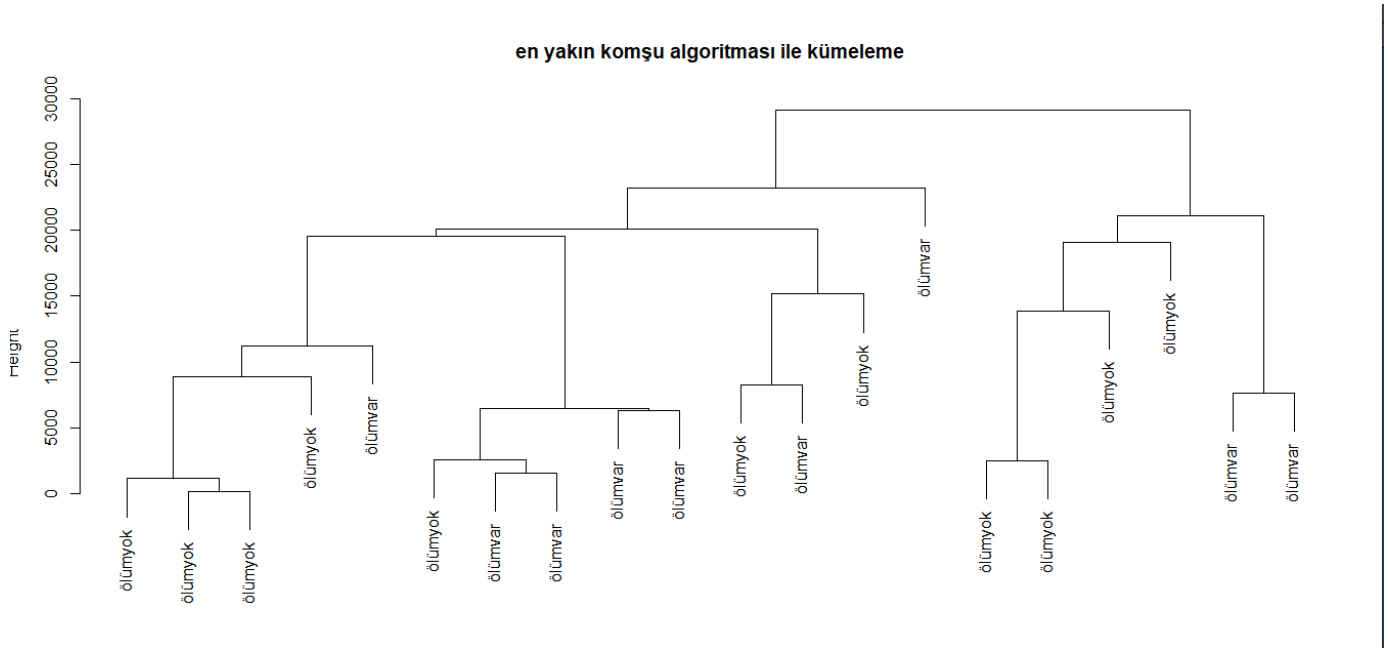


Rastgele seçilen 20 verinin ölüm olayı sonucu bu şekilde kümelenmiştir. Ölüm olayı faktörünün belirlenmesi için" revalue" kullanılarak ölüm_olayı dönüştürülmüştür.

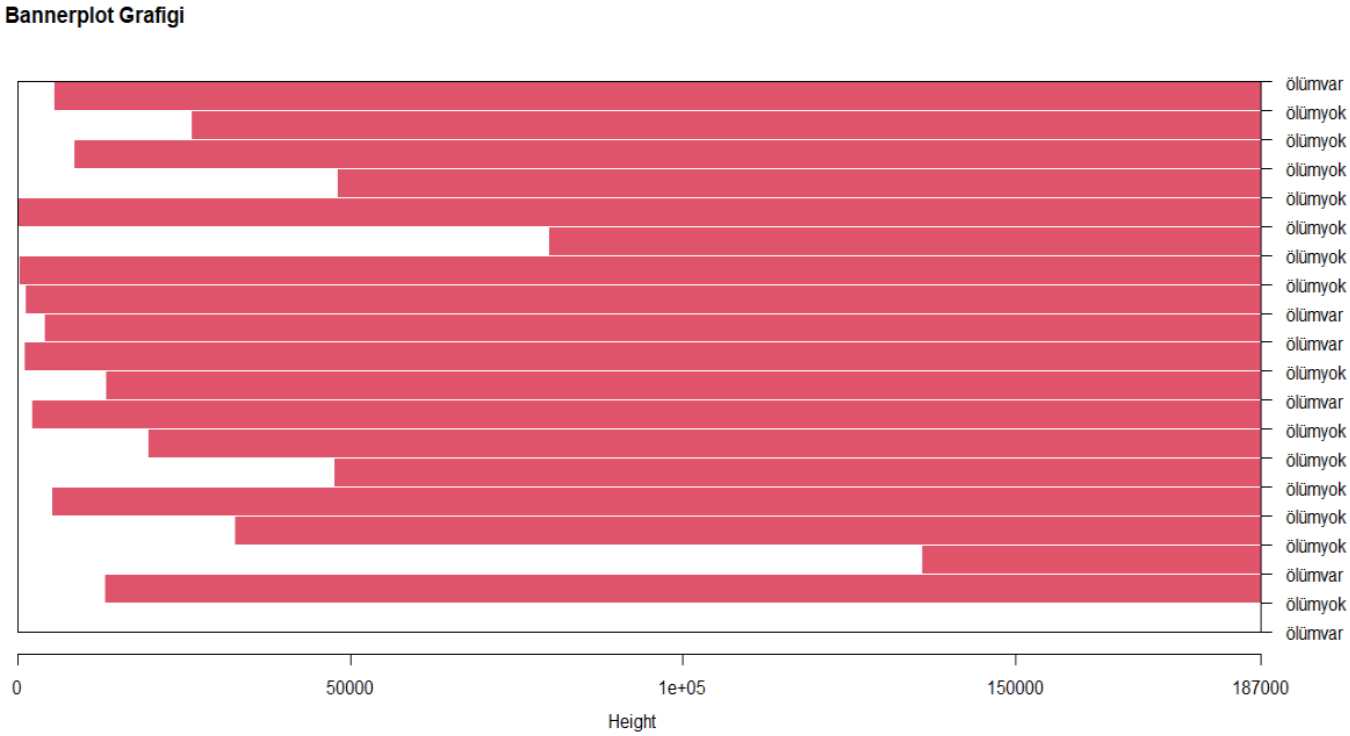
Hedef nitelik, ölüm_olayı değişkeninin değerleri 0= ölümyok, 1=ölümvar şekline dönüştürülür.

veriler$ölüm_olayı <- revalue(veriler$ölüm_olayı, c("0"="ölümyok","1"="ölümvar"))

EN YAKIN KOMŞU ALGORİTMASI İLE KÜMELEME



BANNERPLOT GRAFİĞİ



KNN DOĞRULUK ORANI İÇİN C4.5 KARAR AĞACI KNN'DE UYGULANMIŞTIR

KNN algoritmasının doğruluk ve hata oranını belirleyebilmek amacı ile KNN'de oluşturduğumuz sadece nümerik değerlerden ve hedef niteliği ölüm_olayı olarak belirlenen değişkenlerden oluşan bir alt küme belirlenir ve belirlenen bu alt küme ile C4.5 algoritmasında karar ağacı oluşturularak bu alt kümenin doğruluk ve hata oranı belirlenmiştir.

Sadece nümerik değişkenlerden oluşan ve bir hedef nitelik nitelik belirlenerek bu değişkenlerden oluşan bir alt küme oluşturulmuştur.

Oluşturulan bu alt kümedeki değişkenler; "yas", "kreatinin_. fosfokinaz", "ejeksiyon_fraksiyonu", "trombositler", "serum_sodyum", "zaman", "ölüm_olayı".

Formülü uyguladığımızda 299 veri ve 7 değişkenden oluşan yeni bir data.frame elde edilir.

Hedef nitelik ölüm olayı değişkeninin değeri 0: ölümyok, 1:ölümvar şekline dönüştürülür.

```
veriler$ölüm_olayı <- revalue(veriler$ölüm_olayı, c("0"="ölümyok","1"="ölümvar"))
```

=== Summary ===

Correctly Classified Instances	269	89.9666 %
Kappa statistic	0.7565	
Mean absolute error	0.1672	
Root mean squared error	0.2891	
Relative absolute error	38.3109 %	
Root relative squared error	61.9258 %	
Total Number of Instances	299	

Burada correctly classified instances doğru yerleşen tahmin sayısıdır. Bunun toplam 299 kişi içinden 269 kişi olduğu gözükmektedir ve %89.9666 doğruluk oranına sahiptir.

PERFORMANS DEĞERLENDİRME ÖLÇÜTÜ

PERFORMANS DEĞERLENDİRME ÖLÇÜTÜ	C4.5
DOĞRULUK ORANI	%89.9666
HATA ORANI	%10.0334

4.6NAİVE (BASİT) BAYES SINIFLANDIRICI ALGORİTMASI

Bayes teoremi, olasılık kuramı içinde incelenen önemli bir konudur. Bu **teorem** bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir.

Naive Bayes sınıflandırıcısının temeli Bayes teoremine dayanır. **lazy** (tembel) bir öğrenme algoritmasıdır aynı zamanda dengesiz veri kümelerinde de çalışabilir. Algoritmanın çalışma şekli bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır.

Analiz öncesi değişkenler faktör ve nümerik olarak tanımlanıp analize uygun hale getirilmiştir.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
  ölümyok   ölümvar
0.6777778 0.3222222

Conditional probabilities:
                    yas
egitimHedefNitelik     [,1]     [,2]
           ölümyok 58.44809 10.48833
           ölümvar 63.58621 14.02133

                    anemi
egitimHedefNitelik  anemiyok  anemivar
           ölümyok 0.6393443 0.3606557
           ölümvar 0.5172414 0.4827586

                    kreatinin_.fosfokinaz
egitimHedefNitelik     [,1]      [,2]
           ölümyok 532.0082  674.0966
           ölümvar 788.1724 1591.7726

                    diyabet
egitimHedefNitelik       var       yok
           ölümyok 0.5245902 0.4754098
           ölümvar 0.5862069 0.4137931

                    ejeksiyon_fraksiyonu
egitimHedefNitelik     [,1]     [,2]
           ölümyok 39.84426 10.90835
           ölümvar 33.10345 12.28562
```

```
                    yüksek_kan_basıncı
egitimHedefNitelik       yok       var
          ölümyok 0.6393443 0.3606557
          ölümvar 0.6206897 0.3793103

                    trombositler
egitimHedefNitelik     [,1]     [,2]
          ölümyok 271937.1 107371.2
          ölümvar 258369.5 107497.6

                    serum_sodyum
egitimHedefNitelik     [,1]     [,2]
          ölümyok 137.3689 4.346939
          ölümvar 135.7586 4.878562

                    cinsiyet
egitimHedefNitelik     erkek     kadın
          ölümyok 0.3688525 0.6311475
          ölümvar 0.3793103 0.6206897

                    sigara
egitimHedefNitelik    içiyor   içmiyor
          ölümyok 0.7049180 0.2950820
          ölümvar 0.7068966 0.2931034

                    zaman
egitimHedefNitelik      [,1]     [,2]
          ölümyok 165.98361 65.98760
          ölümvar  70.32759 61.59585
```

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

NAİVE BAYES KONTENJANS TABLOSU

NAİVE BAYES	GERÇEK SINIFLAR		
TAHMİNİ SINIFLAR		ölümyok	ölümvar
	ölümyok	72	14
	ölümvar	9	24

Naive bayes performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	Naive Bayes
Doğruluk oranı	%80,6
Hata oranı	%19,3

Naive bayes algoritması kontenjans tablosu sonuçlarına göre gerçekte ölüm_olayı= ölümyok olan 72 hasta tahminde de 72 olarak doğru tahmin edilmiştir.Doğru pozitif değeri 72'dir.

Gerçekte ölümvar olan, ama tahminde ölümolayı=ölümyok olan 14 kişi vardır. Yanlış pozitif yani tip 1 hata değeri 14'dür.

Gerçekte ölüm_olayı=ölümyok olan 9 kişi vardır, tahminde ölüm_olayı=ölümvar olan 9 kişi vardır.Yanlış negatif yani tip 2 hata değeri 9'dur.

Gerçekte ölüm_olayı=ölümvar olan 24 kişi vardır, tahminde ölüm_olayı=ölümvar olan 24kişi vardır.Doğru pozitif değeri 24'dür.

Modelin doğruluk oranı 0.8067'dir

Modelin hata oranı 0.193277'dir

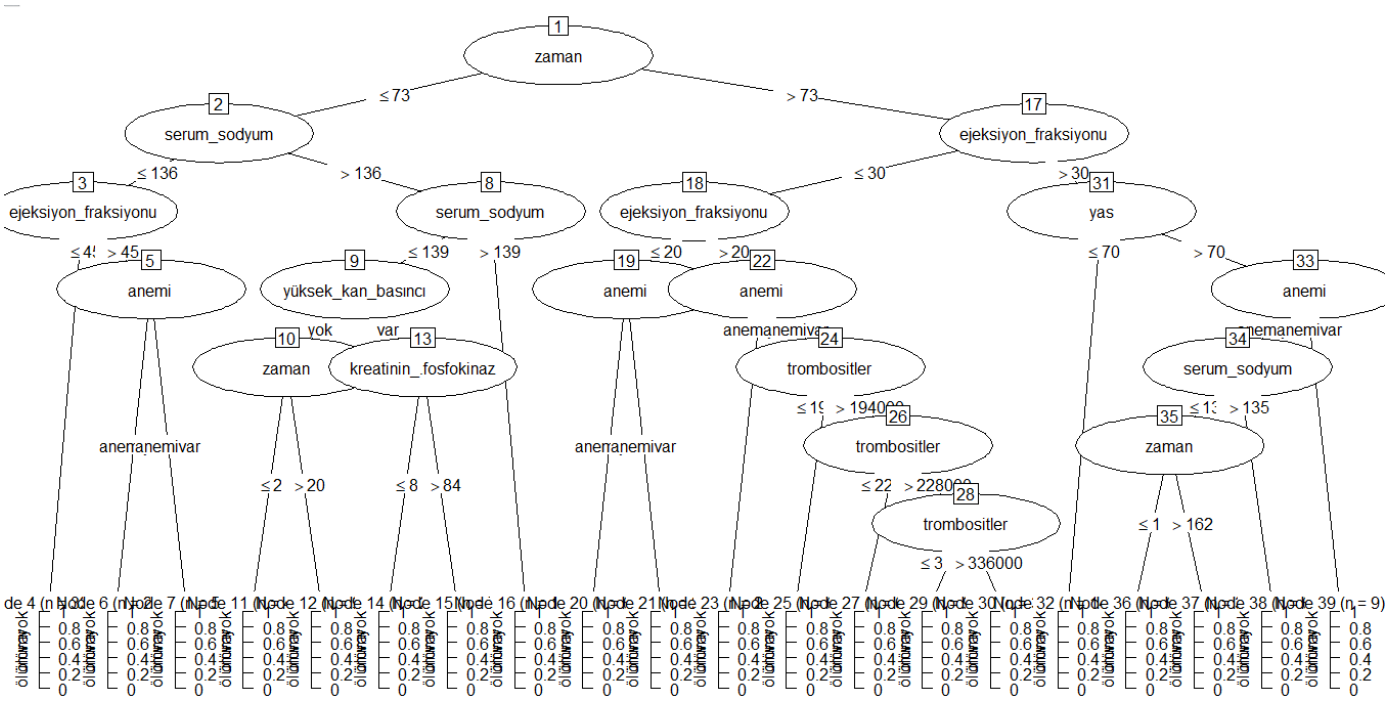
4.7GENEL DEĞERLENDİRME VE MODEL SEÇİMİ

	Doğruluk Oranı	Hata Oranı
KNN ALGORİTMASI	%89.9666	%10.0334
Naive Bayes Algoritması	%80,6	%19,3
C4.5 Karar Ağacı Algoritması	%94.6488	%5.3512

Belirlenen performans değerlendirme oranlarından doğruluk ve hata oranlarına bakılarak en iyi performans veren algoritma c4.5 karar ağacı algoritmasıdır.

C4.5 karar ağacı algoritması modeli oluşturulurken modelin kullandığı belirleyici değişkenler; zaman,serum_sodyum, ejeksiyon_fraksiyonu, anemi, yüksek_kan_basıncı, kreatinin_fosfokinaz, trombositler değerleridir.

C4.5 KARAR AĞACI ALGORİTMASI



SONUÇ

Kalp yetmezliği genel anlamda **kalbin** çeşitli nedenlere bağlı olarak zarar gördüğü veya zayıfladığı durumlarda ortaya çıkan bir hastalıktır. **Kalbin** kan pompalama odacıkları olan ventriküllerin sertleşmesi, **kalbin** iki atışı arasında tam olarak kan ile doldurulamamasına neden olarak **kalp yetmezliğine** yol açabilir.

Kalp Yetmezliği Dünya çağında 23 milyon kişiyi etkilemektedir ve hali hazırda artış göstermeye devam etmektedir. Gelişmiş dünyada 65 yaşın üzerindeki hastaların hastaneye yatırılmasındaki öncü nedenlerdendir. 5 kişiden biri , bir yıl içerisinde ve bu kişilerin %50 si 5 yıl içerisinde kalp yetmezliği teşhisi nedeniyle hayatını kaybetmektedir.

Konunun önemi sebebiyle bu veri çalışmasında kalp yetmezliğini etkileyen faktörler belirlenmiştir.

Çalışmanın birinci bölümünde Veri madenciliği nedir? kavramı, veri madenciliği görevleri ve uygulama adımları, veri madenciliği süreci, veri madenciliği modelleri ve veri madenciliği yöntemleri ele alınmıştır.

Çalışmanın ikinci bölümünde veri madenciliği teknikleri ele alınarak; karar ağacı algoritmaları, k- en yakın komşu algortiması, yapay sinir ağları, karar destek makine sistemleri, naive bayes sınıflandırması, lojsitik regresyon ve birliktelik kuralları anlatılmıştır.

Çalışmanın üçüncü bölümünde kalp yetmezliği hastalığı, kalp yetmezliği nedir? kalp yetmezliği tanıları, kalp yetmezliği teşhisi, önleme ve tedavi başlıkları altında anlatılmıştır.

Çalışmanın dördüncü bölümünde kalp yetmezliği bulunan hastalara ilişkin veri analitiği başlığı altında,problemin tanımlanması, veri setini anlama, analize hazırlık, C4.5 algoritması, knn algortiması, naive bayes sınıflandırıcı algortima ve genel değerlendirme ve model seçimi ele alınmıştır.

Knn algoritmasında sadece nümerik değerlerden oluşan bir alt küme oluşturulmuştur ve bu küme içerisinde rastgele seçilen 20 kişinin ölüm olayı en yakın komşu algortiması ile kümelenerek ve bannerplot grafiği ile sunulmuştur. Daha sonra oluşturduğum bu alt kümenin doğruluk oranını hesaplayabilmek için C4.5 algoritması bu küme içerisinde uygulanmıştır ve modelin doğruluk oranı %89.9666'dır hata oranı ise %10.0334'dür.

Naive Bayes algoritması performans ölçümlerine bakıldığında doğruluk oranı %80.6 ve hata oranı %19.3'dür.

C4.5 karar ağacı algoritmasının performans ölümlerine bakıldığında doğruluk oranı %94.6488 , hata oranı %5.3512'dir. C4.5 karar ağacı algoritmasında ortaya çıkan belirleyici değişkenler; zaman,serum_sodyum, ejeksiyon_fraksiyonu, anemi, yüksek_kan_basıncı, kreatinin_fosfokinaz, trombositler değerleridir.C4.5 karar ağacı algoritması ile ortaya çıkan 20 tane kural vardır.

Tüm modeller birlikte değerlendirildiğinde performans ölçüm modelleri değerlendirme ölçütlerine göre en yüksek doğruluk ve en düşük hatayı veren C4.5 algoritması en uygun modeldir denilebilir.

Bu çalışmada kalp yetmezliği tanıları belirlenip bu tanılar sonucunda 299 hastanın ölüm olayı belirlenmiştir ve ölüm olayını etkileyen belirleyici değişkenler belirlenmiştir. Gelecek çalışmalarda değişken sayısı arttırılıp azaltılırarak veya farklılaştırılarak daha farklı bulgular elde edilebilir. Bu araştırmada gerçek veri seti ile bir hastalık teşhisini tahminlemeye çalışılmıştır. Gelecek çalışmalar için değişken sayısı arttırılarak bu değerlerin düşürülmesi sağlanabilir.

KAYNAKÇA

file:///C:/Users/90536/Downloads/05.sinan_aydin.pdf

file:///C:/Users/90536/Downloads/yokAcikBilim_10085746%20(1).pdf

file:///C:/Users/90536/Downloads/Veri%20Madencili__i_%20T__p%20ve%20Sa__l__k%20Hizmetlerinde%20Kullan__m__%20ve%20Uygulamalar__ [%2387893] -75259.pdf

https://medium.com/@ekrem.hatipoglu/machine-learning-classification-naive-bayes-part-11-4a10cd3452b4

https://archive.ics.uci.edu/ml/datasets.php

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5/tables/1

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5#Tab1

https://www.drmustafaguler.com/kalp-yetmezligi

https://www.biomerieux.com.tr/kaynaklar/saglik-bilgileri/kalp-yetmezligi

Nur Kuban Torun Doktora Tez (1).pdf

EKLER

Ek 1: Veri Önişleme İçin Kullanılan R kodları

#Kullanılan veri seti dosyadan seçilir.

```
> veriler =read.table (file.choose(),header=T,sep=";")
```

#Veri yapısı incelenir.

```
> str(veriler)
```

```
'data.frame':   299 obs. of  12 variables:
 $ yas                   : num  75 55 65 50 65 90 75 60 65 80 ...
 $ anemi                 : int  0 0 0 1 1 1 1 1 0 1 ...
 $ kreatinin_.fosfokinaz: int  582 7861 146 111 160 47 246 315 157 123 ...
 $ diyabet               : int  0 0 0 0 1 0 0 1 0 0 ...
 $ ejeksiyon_fraksiyonu : int  20 38 20 20 20 40 15 60 65 35 ...
 $ yüksek_kan_basıncı   : int  1 0 0 0 0 1 0 0 0 1 ...
 $ trombositler          : num  265000 263358 162000 210000 327000 ...
 $ serum_sodyum          : int  130 136 129 137 116 132 137 131 138 133 ...
 $ cinsiyet              : int  1 1 1 1 0 1 1 1 0 1 ...
 $ sigara                : int  0 0 1 0 0 1 0 1 0 1 ...
 $ zaman                 : int  4 6 7 7 8 8 10 10 10 10 ...
 $ ölüm_olayı            : int  1 1 1 1 1 1 1 1 1 1 ...
```

#Veri yapısı nümerik ve faktör olarak tanımlanır.

```
veriler$yas <- as.numeric(veriler$yas)

veriler$anemi <- as.factor(veriler$anemi)

veriler$kreatinin_.fosfokinaz <- as.numeric(veriler$kreatinin_.fosfokinaz)

veriler$diyabet <- as.factor(veriler$diyabet)

veriler$ejeksiyon_fraksiyonu <- as.numeric(veriler$ejeksiyon_fraksiyonu)

veriler$yüksek_kan_basıncı <- as.factor(veriler$yüksek_kan_basıncı)

veriler$trombositler <- as.numeric(veriler$trombositler)

veriler$serum_sodyum <- as.numeric(veriler$serum_sodyum)

veriler$cinsiyet <- as.factor(veriler$cinsiyet)

veriler$sigara <- as.factor(veriler$sigara)

veriler$zaman <- as.numeric(veriler$zaman)

veriler$ölüm_olayı <- as.factor(veriler$ölüm_olayı)
```

Veri yapısı tekrar incelenir

> str(veriler)

```
'data.frame':   299 obs. of  12 variables:
 $ yas                   : num  75 55 65 50 65 90 75 60 65 80 ...
 $ anemi                 : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 1 2 ...
 $ kreatinin_.fosfokinaz: num  582 7861 146 111 160 ...
 $ diyabet               : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
 $ ejeksiyon_fraksiyonu : num  20 38 20 20 20 40 15 60 65 35 ...
 $ yüksek_kan_basıncı    : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...
 $ trombositler          : num  265000 263358 162000 210000 327000 ...
 $ serum_sodyum          : num  130 136 129 137 116 132 137 131 138 133 ...
 $ cinsiyet              : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 1 2 ...
 $ sigara                : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...
 $ zaman                 : num  4 6 7 7 8 8 10 10 10 10 ...
 $ ölüm_olayı            : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
>
```

> install.packages("plyr")

> library(plyr)

veriler$anemi <- revalue(veriler$anemi, c("1"="anemivar","0"="anemiyok"))

veriler$cinsiyet <- revalue(veriler$cinsiyet, c("0"="erkek","1"="kadın"))

veriler$diyabet <- revalue(veriler$diyabet, c("1"="yok","0"="var"))

veriler$yüksek_kan_basıncı <- revalue(veriler$yüksek_kan_basıncı, c("0"="yok","1"="var"))

veriler$sigara <- revalue(veriler$sigara, c("0"="içiyor","1"="içmiyor"))

veriler$ölüm_olayı <- revalue(veriler$ölüm_olayı, c("0"="ölümyok","1"="ölümvar"))

#Veri setinin özetine bakılır

> summary(veriler)

#Nümerik değişkenlerin grafikleri çizilir.

> hist(veriler$yas, col="red", main = "Yaş Histogram Grafiği")

> hist(veriler$ kreatinin_.fosfokinaz, col="red", main = "Kreatinin Fosfokinaz Histogram Grafiği")

> hist(veriler$ ejeksiyon_fraksiyonu, col="red", main = "Ejeksiyon Fraksiyonu Histogram Grafiği")

> hist(veriler$ trombositler, col="red", main = "Trombositler Histogram Grafiği")

> hist(veriler$ serum_sodyum, col="red", main = "Serum Sodyum Histogram Grafiği")

> hist(veriler$ zaman, col="red", main = "Zaman Histogram Grafiği")

#Kategorik değişkenlerin grafikleri çizilir.

#öncesinde anemi için; 1=anemivar, 0=anemiyok

Cinsiyet için; 0= erkek, 1= kadın

Diyabet için; 1=yok, 0=var

Yüksek_kan_basıncı için; 0=yok, 1=var

Sigara için; 0= içiyor, 1=içmiyor

Ölüm_olayı için; 0=ölümyok, 1=ölümvar

#Kategorik değişkenlerin grafikleri çizilir

#Anemi için grafik çizimi>

frekansanemi <- table(veriler$anemi)

> barplot(frekansanemi, col="purple" , main="Anemi Dağılımları Grafiği",xlab="Anemi Sayısı",ylab = "Anemi" , horiz = TRUE)

#Cinsiyet için grafik çizimi>

frekanscinsiyet <- table(veriler$cinsiyet)

> barplot(frekanscinsiyet, col="purple" , main="Cinsiyet Dağılımları Grafiği",xlab="Cinsiyet Sayısı",ylab = "Anemi" , horiz = TRUE)

#Diyabet için grafik çizimi>

frekansdiyabet <- table(veriler$diyabet)

> barplot(frekansdiyabet, col="purple" , main="Diyabet Dağılımları Grafiği",xlab="DiyabetSayısı",ylab = "Diyabet" , horiz = TRUE)

#Yüksek_kan_basıncı için grafik çizimi>

Frekansyüksek_kan_basıncı <- table(veriler$yüksek_kan_basıncı)

> barplot(frekansyüksek-kan_basıncı, col="purple" , main="Yüksek Kan Basıncı Dağılımları Grafiği",xlab="Yüksek Kan Basıncı Sayısı",ylab = "Yüksek Kan Basıncı" , horiz = TRUE)

#Sıgara için grafik çizimi>

frekanssigara <- table(veriler$sigara)

> barplot(frekanssigara, col="purple" , main="sigara Dağılımları Grafiği",xlab="Sıgara Kullanım Sayısı",ylab = "Sigara" , horiz = TRUE)

#Ölüm Olayı için grafik çizimi>

frekansÖlüm_olayı <- table(veriler$ölüm_olayı)

> barplot(frekansölüm_olayı, col="purple" , main="Ölüm Olayı Dağılımları Grafiği",xlab="Ölüm Sayısı",ylab = "Ölüm Olayı" , horiz = TRUE)

#kutu grafikleri çizimi

```
> boxplot(veriler$yas, col=" yellow ", main="Yas Kutu Grafiği")

> boxplot(veriler$ kreatinin_.fosfokinaz, col=" yellow ", main="kreatinin Fosfokinaz Kutu Grafiği")

 > boxplot(veriler$ejeksiyon_fraksiyonu, col=" yellow ", main="Ejeksiyon Fraksiyonu Kutu Grafiği")

> boxplot(veriler$trombositler, col=" yellow ", main="Trombositler Kutu Grafiği")

 > boxplot(veriler$serum_sodyum, col=" yellow ", main="Serum Sodyum Kutu Grafiği")

> boxplot(veriler$zaman, col="yellow", main="Zaman Kutu Grafiği")


 #serpilme diyagramı çizimi

> pairs( ~ yas+ zaman+ serum_sodyum + trombositler + ejeksiyon_fraksiyonu + kreatinin_fosfokinaz , data= veriler, col=" dark blue", main= "Serpilme Diyagramları")
```

KNN Algortiması Uygulaması Kodları

```
#önce veri seti çağırılır

veriler= read.table(file.choose(), header = T, sep = ";")

# veri seti incelenir, nümerik ve kategorik veriler tanımlanır

> str(veriler)

veriler$yas <- as.numeric(veriler$yas)

veriler$anemi <- as.factor(veriler$anemi)

veriler$kreatinin_.fosfokinaz <- as.numeric(veriler$kreatinin_.fosfokinaz)

veriler$diyabet <- as.factor(veriler$diyabet)

veriler$ejeksiyon_fraksiyonu <- as.numeric(veriler$ejeksiyon_fraksiyonu)

veriler$yüksek_kan_basıncı <- as.factor(veriler$yüksek_kan_basıncı)

veriler$trombositler <- as.numeric(veriler$trombositler)

veriler$serum_sodyum <- as.numeric(veriler$serum_sodyum)

veriler$cinsiyet <- as.factor(veriler$cinsiyet)

veriler$sigara <- as.factor(veriler$sigara)

veriler$zaman <- as.numeric(veriler$zaman)

veriler$ölüm_olayı <- as.factor(veriler$ölüm_olayı)


#library(plyr)

veriler$ölüm_olayı <- revalue(veriler$ölüm_olayı, c("0"="ölümyok","1"="ölümvar"))
```

```
#sadece küm. Değerlerden oluşan alt küme oluşturuldu
n_veriler <- veriler [,c(1,3,5,7,8,11,12)]

set.seed(1234)

ind <- sample(1:299,20)
veriler <- n_veriler[ind,]
#oluşturduğumuz verileri görselleştirmek için
pltree(model, main="en yakın komsu algoritması ile kumeleme")
pltree(model, main="en yakın komşu algoritması ile kümeleme", labels=veriler$ölüm_olayı)
bannerplot(agnes(veriler),main="Bannerplot Grafigi", labels = veriler$ölüm_olayı)
#KNN kümesinin doğruluk oranlarını bulabilmek için bu küme de c4.5 karar ağacı uygulaması uygulanmıştır.
veriler= read.table(file.choose(), header = T, sep = ";")
install.packages("caret")
library(caret)
install.packages("cluster")
library(cluster)
View(veriler)
summary(veriler)
str(veriler)
attributes(veriler)
#VERİLER nümerik ve faktör olarak tanımlanır
veriler$yas <- as.numeric(veriler$yas)
veriler$anemi <- as.factor(veriler$anemi)
veriler$kreatinin_.fosfokinaz <- as.numeric(veriler$kreatinin_.fosfokinaz)
veriler$diyabet <- as.factor(veriler$diyabet)
veriler$ejeksiyon_fraksiyonu <- as.numeric(veriler$ejeksiyon_fraksiyonu)
veriler$yüksek_kan_basıncı <- as.factor(veriler$yüksek_kan_basıncı)
veriler$trombositler <- as.numeric(veriler$trombositler)
veriler$serum_sodyum <- as.numeric(veriler$serum_sodyum)
```

```
veriler$cinsiyet <- as.factor(veriler$cinsiyet)
veriler$sigara <- as.factor(veriler$sigara)
veriler$zaman <- as.numeric(veriler$zaman)
veriler$ölüm_olayı <- as.factor(veriler$ölüm_olayı)
veriler$ölüm_olayı <- revalue(veriler$ölüm_olayı, c("0"="ölümyok","1"="ölümvar"))
#sadece nümerik değerlerden oluşan alt küme oluşturuldu ve nümerik değerlere karşılık gelen
#verilerin değerleri sayısal olarak girildi.
n_veriler <- veriler [,c(1,3,5,7,8,11,12)]
#rastgele veri seçimi için set.seed kullanılır.
set.seed(1234)
ind <- sample(1:299,299)
veriler <- n_veriler[ind,]
modelC11 <- J48(ölüm_olayı~.,data = veriler)

#kurallari gorelim
```

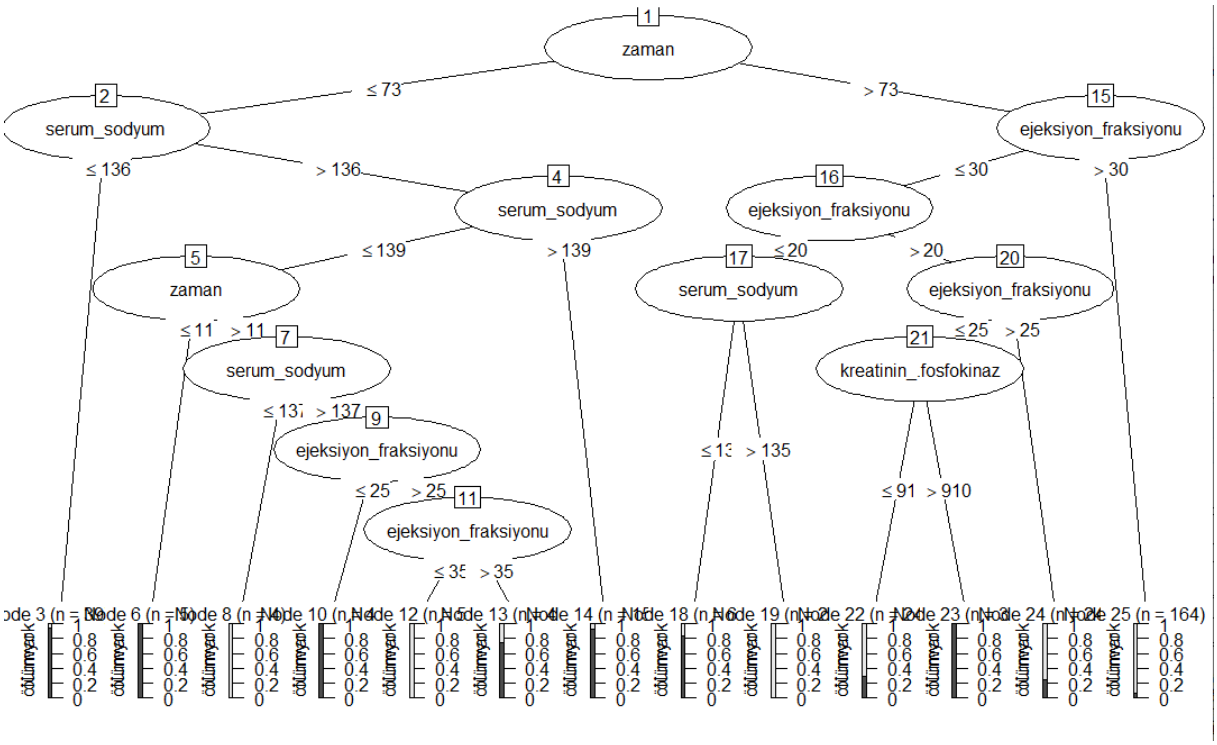
```
> print(modelC11)
J48 pruned tree
------------------

zaman <= 73
|   serum_sodyum <= 136: ölümvar (39.0/2.0)
|   serum_sodyum > 136
|   |   serum_sodyum <= 139
|   |   |   zaman <= 11: ölümvar (5.0)
|   |   |   zaman > 11
|   |   |   |   serum_sodyum <= 137: ölümyok (4.0)
|   |   |   |   serum_sodyum > 137
|   |   |   |   |   ejeksiyon_fraksiyonu <= 25: ölümvar (4.0)
|   |   |   |   |   ejeksiyon_fraksiyonu > 25
|   |   |   |   |   |   ejeksiyon_fraksiyonu <= 35: ölümyok (5.0)
|   |   |   |   |   |   ejeksiyon_fraksiyonu > 35: ölümvar (4.0/1.0)
|   |   serum_sodyum > 139: ölümvar (15.0/1.0)
zaman > 73
|   ejeksiyon_fraksiyonu <= 30
|   |   ejeksiyon_fraksiyonu <= 20
|   |   |   serum_sodyum <= 135: ölümvar (6.0/1.0)
|   |   |   serum_sodyum > 135: ölümyok (2.0)
|   |   ejeksiyon_fraksiyonu > 20
|   |   |   ejeksiyon_fraksiyonu <= 25
|   |   |   |   kreatinin_.fosfokinaz <= 910: ölümyok (24.0/7.0)
|   |   |   |   kreatinin_.fosfokinaz > 910: ölümvar (3.0)
|   |   |   ejeksiyon_fraksiyonu > 25: ölümyok (24.0/6.0)
|   ejeksiyon_fraksiyonu > 30: ölümyok (164.0/12.0)

Number of Leaves  :     13

Size of the tree :      25
```

summary(modelC11)

```
=== Summary ===

Correctly Classified Instances         269               89.9666 %
Kappa statistic                          0.7565
Mean absolute error                      0.1672
Root mean squared error                  0.2891
Relative absolute error                 38.3109 %
Root relative squared error             61.9258 %
Total Number of Instances              299

=== Confusion Matrix ===

   a   b   <-- classified as
 198   5 |   a = ölümyok
  25  71 |   b = ölümvar
> |
```

#grafigini cizelim

plot(modelC11)



C4.5 Karar Ağacı Algoritması Kodları

```
#veri seti çağırılır

veriler= read.table(file.choose(), header = T, sep = ";")

#veriler nümerik ve faktör olarak tanımlanır.

veriler$yas <- as.numeric(veriler$yas)

veriler$anemi <- as.factor(veriler$anemi)

veriler$kreatinin_.fosfokinaz <- as.numeric(veriler$kreatinin_.fosfokinaz)

veriler$diyabet <- as.factor(veriler$diyabet)

veriler$ejeksiyon_fraksiyonu <- as.numeric(veriler$ejeksiyon_fraksiyonu)

veriler$yüksek_kan_basıncı <- as.factor(veriler$yüksek_kan_basıncı)

veriler$trombositler <- as.numeric(veriler$trombositler)

veriler$serum_sodyum <- as.numeric(veriler$serum_sodyum)

veriler$cinsiyet <- as.factor(veriler$cinsiyet)

veriler$sigara <- as.factor(veriler$sigara)

veriler$zaman <- as.numeric(veriler$zaman)

veriler$ölüm_olayı <- as.factor(veriler$ölüm_olayı)

print(veriler)
```

```
View(veriler)
summary(veriler)
str(veriler)
attributes(veriler)

install.packages("plyr")
library(plyr)

#anemi değişkeninde 1=anemivar,0=anemiyok şeklinde tanımlanır.
veriler$anemi <- revalue(veriler$anemi, c("1"="anemivar","0"="anemiyok"))

#cinsiyet değişkeni 0=erkek 1=kadın olarak tanımlanır.
veriler$cinsiyet <- revalue(veriler$cinsiyet, c("0"="erkek","1"="kadın"))

#diyabet değişkeni 1=yok 0= var şeklinde tanımlanır.
veriler$diyabet <- revalue(veriler$diyabet, c("1"="yok","0"="var"))

#yüksek kan basıncı değişkeni 0=yok 1=var
veriler$yüksek_kan_basıncı <- revalue(veriler$yüksek_kan_basıncı, c("0"="yok","1"="var"))

#sigara değişkeni 0=içiyor 1=içmiyor şeklinde tanımlanır.
veriler$sigara <- revalue(veriler$sigara, c("0"="içiyor","1"="içmiyor"))

#ölüm olayı değişkeni 0=ölümyok 1=ölümvar şeklinde tanımlanmıştır.
veriler$ölüm_olayı <- revalue(veriler$ölüm_olayı, c("0"="ölümyok","1"="ölümvar"))

#Rweka peketi icinde C4.5 algoritmasinin J48() isimli bir uyarlamasi yer almaktadir.

modelC11 <- J48(ölüm_olayı~.,data = veriler)
#kurallari gorelim
print(modelC11)
```

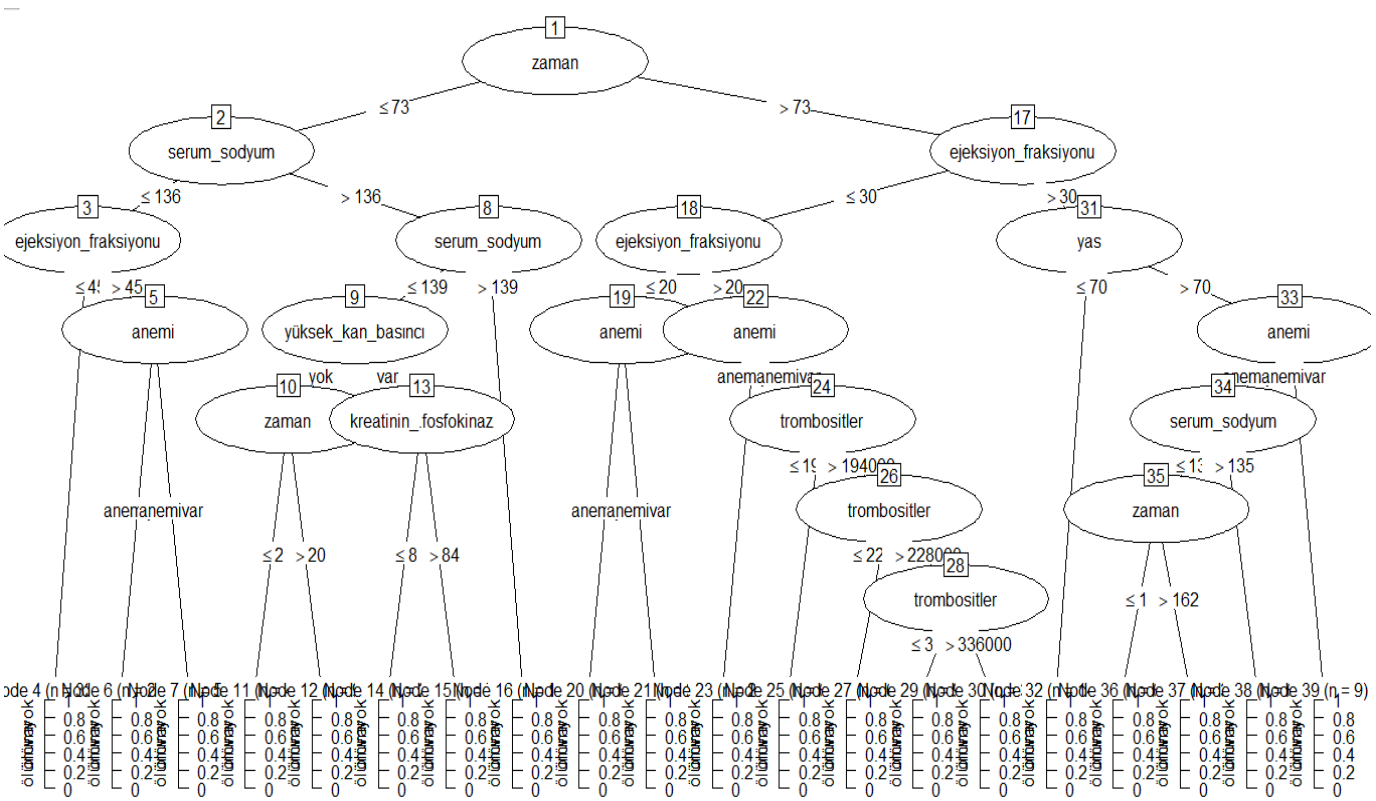
```
> print(modelC11)
J48 pruned tree
------------------

zaman <= 73
|   serum_sodyum <= 136
|   |   ejeksiyon_fraksiyonu <= 45: ölümvar (32.0)
|   |   ejeksiyon_fraksiyonu > 45
|   |   |   anemi = anemiyok: ölümyok (2.0)
|   |   |   anemi = anemivar: ölümvar (5.0)
|   serum_sodyum > 136
|   |   serum_sodyum <= 139
|   |   |   yüksek_kan_basıncı = yok
|   |   |   |   zaman <= 20: ölümvar (4.0)
|   |   |   |   zaman > 20: ölümyok (9.0/1.0)
|   |   |   yüksek_kan_basıncı = var
|   |   |   |   kreatinin_.fosfokinaz <= 84: ölümyok (2.0)
|   |   |   |   kreatinin_.fosfokinaz > 84: ölümvar (7.0)
|   |   serum_sodyum > 139: ölümvar (15.0/1.0)
zaman > 73
|   ejeksiyon_fraksiyonu <= 30
|   |   ejeksiyon_fraksiyonu <= 20
|   |   |   anemi = anemiyok: ölümvar (6.0/1.0)
|   |   |   anemi = anemivar: ölümyok (2.0)
|   |   ejeksiyon_fraksiyonu > 20
|   |   |   anemi = anemiyok: ölümyok (27.0/5.0)
|   |   |   anemi = anemivar
|   |   |   |   trombositler <= 194000: ölümyok (6.0)
|   |   |   |   trombositler > 194000
|   |   |   |   |   trombositler <= 228000: ölümvar (6.0)
|   |   |   |   |   trombositler > 228000
|   |   |   |   |   |   trombositler <= 336000: ölümyok (9.0/2.0)
|   |   |   |   |   |   trombositler > 336000: ölümvar (3.0)

|   ejeksiyon_fraksiyonu > 30
|   |   yas <= 70: ölümyok (140.0/5.0)
|   |   yas > 70
|   |   |   anemi = anemiyok
|   |   |   |   serum_sodyum <= 135
|   |   |   |   |   zaman <= 162: ölümvar (6.0)
|   |   |   |   |   zaman > 162: ölümyok (3.0)
|   |   |   |   serum_sodyum > 135: ölümyok (6.0)
|   |   |   anemi = anemivar: ölümyok (9.0/1.0)

Number of Leaves  :     20

Size of the tree :      39
```

summary(modelC11)

```
> summary(modelC11)

=== Summary ===

Correctly Classified Instances         283               94.6488 %
Kappa statistic                          0.8731
Mean absolute error                      0.0936
Root mean squared error                  0.2164
Relative absolute error                 21.452  %
Root relative squared error             46.3388 %
Total Number of Instances              299

=== Confusion Matrix ===

   a   b   <-- classified as
 201   2 |   a = ölümyok
  14  82 |   b = ölümvar
> |
```

#c4.5 algortimasının karar ağacı

plot(modelC11)



Naive – Bayes Algortiması İçin Kodlar

```
#veri seti çağırılır
veriler = read.table(file.choose(),header = T,sep = ";")
```

```
#veri seti çağırılır
veriler = read.table(file.choose(),header = T,sep = ";")
```

```
#modelin çalışması için aşağıdaki veriler kurulur.
install.packages("caret")
library(caret)
install.packages("e1071")
library(e1071)
data(veriler)
View(veriler)
str(veriler)
summary(veriler)
attributes(veriler)
```

```
#veriler nümerik ve faktör olarak tanımlanır.
veriler$yas <- as.numeric(veriler$yas)
veriler$anemi <- as.factor(veriler$anemi)
veriler$kreatinin_.fosfokinaz <- as.numeric(veriler$kreatinin_.fosfokinaz)
veriler$diyabet <- as.factor(veriler$diyabet)
veriler$ejeksiyon_fraksiyonu <- as.numeric(veriler$ejeksiyon_fraksiyonu)
veriler$yüksek_kan_basıncı <- as.factor(veriler$yüksek_kan_basıncı)
veriler$trombositler <- as.numeric(veriler$trombositler)
veriler$serum_sodyum <- as.numeric(veriler$serum_sodyum)
veriler$cinsiyet <- as.factor(veriler$cinsiyet)
veriler$sigara <- as.factor(veriler$sigara)
veriler$zaman <- as.numeric(veriler$zaman)
veriler$ölüm_olayı <- as.factor(veriler$ölüm_olayı)
```

```
#revalue komutunun çalışması için plyr kullanılır.
install.packages("plyr")
library(plyr)

#anemi değişkeninde 1=anemivar,0=anemiyok şeklinde tanımlanır.
veriler$anemi <- revalue(veriler$anemi, c("1"="anemivar","0"="anemiyok"))

#cinsiyet değişkeni 0=erkek 1=kadın olarak tanımlanır.
veriler$cinsiyet <- revalue(veriler$cinsiyet, c("0"="erkek","1"="kadın"))

#diyabet değişkeni 1=yok 0= var şeklinde tanımlanır.
veriler$diyabet <- revalue(veriler$diyabet, c("1"="yok","0"="var"))
#yüksek kan basıncı değişkeni 0=yok 1=var
veriler$yüksek_kan_basıncı <- revalue(veriler$yüksek_kan_basıncı, c("0"="yok","1"="var"))

#sigara değişkeni 0=içiyor 1=içmiyor şeklinde tanımlanır.
veriler$sigara <- revalue(veriler$sigara, c("0"="içiyor","1"="içmiyor"))

#ölüm olayı değişkeni 0=ölümyok 1=ölümvar şeklinde tanımlanmıştır.
veriler$ölüm_olayı <- revalue(veriler$ölüm_olayı, c("0"="ölümyok","1"="ölümvar"))

set.seed(1)
verisetibolme <- createDataPartition(y=veriler$ölüm_olayı, p=0.6, list=FALSE)

#veri setini egitim ve test olarak rastgele ikiye ayiracagiz
egitim <- veriler[verisetibolme,]
test <- veriler[-verisetibolme,]

# Eğitim ve test veri setine tahmininde kullanılacak nitelik ve hedef nitelik(ölüm_olayı) atanır.
#Ölüm_olayı 12. Sütunda olduğu için 12 kullanıldı.
```

testNitelikleri <- test[,-12]

testHedefNitelik <- test[[12]]

egitimNitelikleri <- egitim [,-12]

egitimHedefNitelik <- egitim [[12]]

 # Naive bayes için e1071 paketi çağrıldı. Bu paketteki naiveBayes() fonksiyonu kullanıldı.

library(e1071)

naiveBayes_modeli_kuruldu <- naiveBayes(egitimNitelikleri, egitimHedefNitelik)

naiveBayes_modeli_kuruldu

#modelin tahminleri bulunur.

(tahminiSiniflar <- predict(naiveBayes_modeli_kuruldu, testNitelikleri))

#gerçek sınıflar ile tahmini sınıflar kıyaslanır.

(karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini Siniflar", "Gercek Siniflar")))

(TP <- karisiklikmatrisi [1])

[1] 72

(FP <- karisiklikmatrisi [3])

[3] 14

(FN <- karisiklikmatrisi [2])

[2] 9

(TN <- karisiklikmatrisi [4])

[4] 24

#performans değerlendirme ölçütleri hesaplanır.

paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))

Modelin Doğruluk Oranı: 0.8067’dir

paste0("Hata = ",(Hata <- 1-Dogruluk))

Modelin Hata Oranı: 0.193277