T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

# VERİ MADENCİLİĞİ İLE ÇOCUKLUK ÇAĞINDAKİ AKUT ROMATİZMAL ATEŞİN KALP HASTALIĞINA ETKİLERİNİN ANALİZİ

**İlkim Ecem EMRE**

**Enformatik Anabilim Dalı**

**Enformatik Programı**

**DANIŞMAN**
**Yrd. Doç. Dr. Çiğdem EROL**

**II. DANIŞMAN**
**Yrd. Doç. Dr. Yalçın ÖZKAN**

**Haziran, 2017**

**İSTANBUL**

Bu çalışma 15.06.2017 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı Enformatik Programı'nda Yüksek Lisans tezi olarak kabul edilmiştir.

**Tez Jürisi**

Yrd. Doç. Dr. Çiğdem EROL (Danışman)
İstanbul Üniversitesi
Enformatik Bölümü

Prof. Dr. Sevinç GÜLSEÇEN
İstanbul Üniversitesi
Enformatik Bölümü

Prof. Dr. Yücel YILMAZ
Marmara Üniversitesi
İşletme Fakültesi

Doç. Dr. Çağatay NUHOĞLU
Haydarpaşa Numune Eğitim ve
Araştırma Hastanesi
Çocuk Kliniği

Doç. Dr.Seda TOLUN
İstanbul Üniversitesi
İşletme Fakültesi

20.04.2016 tarihli resmi gazetede yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

Bu tez, İstanbul Üniversitesi Bilimsel Araştırma Projeleri Yürütücü Sekreterliğinin 23585 numaralı projesi ile desteklenmiştir.

# ÖNSÖZ

Her zaman güler yüzü, sabrı ve titizliğiyle çalışmama destek olan, paniğe kapıldığım anlarda beni toparlayan, yol gösteren sevgili danışmanım Yrd. Doç. Dr. Çiğdem EROL'a çok teşekkür ederim.

Veri madenciliği alanındaki deneyimiyle tezimin analiz aşamasında her türlü desteği veren ve yol gösteren sevgili danışmanım Yrd. Doç. Dr. Yalçın ÖZKAN'a çok teşekkür ederim.

Tez çalışmamda yararlandığım veri setini kullanmamı sağlayan, İstanbul Medeniyet Üniversitesi Göztepe Eğitim ve Araştırma Hastanesi, Çocuk Kardiyoloji Bölümü'nden Uzm. Dr. Yusuf İzzet AYHAN ve tez sürecimde çalışmamın en başından beri hiç bir desteğini esirgemeyen, sabrı ve deneyimiyle çalışmalarıma eşlik eden Uzm. Dr. Nurdan EROL'a çok teşekkür ederim.

Tez sürecimde sıkıştığım her an gösterdikleri tüm destek ve emekleri için Arş. Gör. Dr. Elif KARTAL ve Arş. Gör. Dr. Zeki ÖZEN'e çok teşekkür ederim.

Her zaman her koşulda yanıbaşımda olan Utku SÖZERİ'ye en içten teşekkürlerimi sunarım.

Her zaman her koşulda beni destekleyen ve bugüne gelmemi sağlayan sevgili annem ve babama en içten teşekkürlerimi sunarım.

Haziran 2017 İlkim Ecem EMRE

# İÇİNDEKİLER

**Sayfa No**

# ŞEKİL LİSTESİ

**Sayfa No**

# TABLO LİSTESİ

**Sayfa No**

# SİMGE VE KISALTMA LİSTESİ

| Kısaltmalar | Açıklama |
|---|---|
| **AHA** | : Amerikan Kalp Derneği (*American Heart Association*) |
| **ARA** | : Akut Romatizmal Ateş |
| **ASO** | : Antistreptolizin O Antikor |
| **EÇH** | : Eritrosit Çökme Hızı |
| **EKG** | : Elektrokardiyogram |
| **CART** | : Sınıflandırma ve Regresyon Ağaçları (*Classification and Regression Trees*) |
| **CRP** | : C-reaktif Protein |
| **CRISP-DM** | : Veri Madenciliği için Çapraz Endüstri Standardı Süreci (*Cross Industry Standard Process for Data Mining*) |
| **KDD** | : Veri Tabanlarında Bilgi Keşfi (*Knowledge Discovery in Databases*) |
| **SEMMA** | : Örnekle, İncele, Düzelt, Modelle, Değerle (*Sample, Explore, Modify, Model, Assess*) |
| **WHO** | : Dünya Sağlık Örgütü (*World Health Organisation*) |
| **YÖK** | : Yükseköğretim Kurulu |

# ÖZET

## YÜKSEK LİSANS TEZİ

### VERİ MADENCİLİĞİ İLE ÇOCUKLUK ÇAĞINDAKİ AKUT ROMATİZMAL ATEŞİN KALP HASTALIĞINA ETKİLERİNİN ANALİZİ

**İlkim Ecem EMRE**

**İstanbul Üniversitesi**

**Fen Bilimleri Enstitüsü**

**Enformatik Anabilim Dalı**

**Danışman : Yrd. Doç. Dr. Çiğdem EROL**

**II. Danışman : Yrd. Doç. Dr. Yalçın ÖZKAN**

Günümüzde içinde yaşadığımız dönem bilgi toplumu ya da bilgi çağı olarak adlandırılmaktadır. Kuşkusuz ki bilginin ve bilgi teknolojilerinin hayatın her alanına nüfuz ettiği gözlemlenmektedir. Bu teknolojilerin hızlı gelişimi bilgi kavramının önemi de arttırmış, farklı kaynaklar tarafından üretilen çok çeşitli ve büyük hacimli veri ile karşılaşmamıza neden olmuştur. Veri miktarındaki artış ise elde edilen büyük boyutlu veriden bilgi elde etme sürecini zorlaştırmaya başlamıştır. İstatistik bilimi veri analizinde yüzyıllardan beri kullanılmaktadır; ancak veri miktarındaki artış, temeli istatistiğe dayanan veri madenciliği kavramını ortaya çıkarmıştır. Geçmişi istatistik kadar eskilere dayanmasa da veri madenciliği birçok farklı alandan araştırmacının ilgisini çekmektedir. Bu kapsamda veri madenciliği, veri analizinde gün geçtikçe popülerliğini ve işlevini arttıran bir alan olarak karşımıza çıkmakta ve birçok farklı araştırma alanında kullanılmaktadır. Sağlık çalışmalarında ise hastalardan elde edilen veri kümeleri istatistiksel yöntemlerle analiz edilebiliyor olsa da veri madenciliği yöntemlerinin kullanımı da sağlık verisinin analizinde kullanılabilmektedir.

Bu tez çalışmasının amacı veri madenciliği yöntemlerini kullanarak çocuk yaşta görülen akut romatizmal ateşin kalp üzerindeki etkilerinin analiz edilmesidir. Bu etkilerin belirlenmesi çocukluk yaşlarında görülen romatizmanın kalp kapağına vereceği zararların

en aza indirilmesi açısından önemlidir. Bu tespitin yapılabilmesi için veri madenciliğindeki sınıflandırma yöntemlerinden yararlanılmış ve farklı beş algoritma denenmiştir. Sade Bayes sınıflandırıcı, karar ağaçları (CART, C4.5, C5.0, C5.0 boosted) ve rastgele orman algoritmaları ile modeller kurularak akut romatizmal ateş tanısı konmuş hasta kayıtları analiz edilmiş ve algoritmaların performansları karşılaştırılmıştır. Model performans değerlendirme yöntemlerinden holdout, çapraz geçerleme (cross validation) ve bootstrap yöntemleri farklı şekillerde veri setine uygulanarak algoritmalar denenmiştir. Çalışma kapsamında, İstanbul Medeniyet Üniversitesi Göztepe Eğitim ve Araştırma Hastanesi Çocuk Kliniği ile ortak çalışarak, 297 hastaya ait kayıtlardan oluşan veri seti kullanılmış, ön işleme süreci sonunda kalan 201 hasta verisi ile analiz yapılmıştır. Farklı algoritmalardan elde edilen sonuçlar model performans değerlendirme ölçütlerine göre karşılaştırılmıştır. En iyi sonucu CART modeli vermiştir.

Bu tez çalışması, tıp ve sağlık alanındaki verinin analizinde veri madenciliği metotlarının kullanımının yaygınlaşmasına katkı sağlaması açısından önemli bir çalışma olarak görülmektedir.

Haziran 2017, 120 sayfa.

**Anahtar kelimeler:** Veri madenciliği, sınıflandırma yöntemleri, akut romatizmal ateş.

# SUMMARY

## M.Sc. THESIS

### ANALYSIS OF EFFECTS OF ACUTE RHEUMATIC FEVER IN CHILDHOOD ON HEART DISEASE WITH DATA MINING

**İlkim Ecem EMRE**

**İstanbul University**

**Institute of Graduate Studies in Science and Engineering**

**Department of Informatics**

**Supervisor : Asst. Prof. Dr. Çiğdem EROL**

**Co-Supervisor : Asst. Prof. Dr. Yalçın ÖZKAN**

Today the era we live in is called knowledge/information society or information age. No doubt, it is observed that information and information technologies penetrate every aspect of life. The rapid development of these technologies has also increased the importance of the concept of information and caused us to encounter a wide variety of large volumes of data produced by different sources. The increase in the amount of data has begun to complicate the process of obtaining the large-sized dataview obtained. Statistics has been used for centuries in data analysis but the increase in the amount of data reveals the concept of data mining which is based on statistics. Data mining attracts many researchers in many different fields, although it does not have a history as old as statistics. In this context, data mining appears as a field which increases in popularity and function day by day in data analysis and used in many different research fields. While data sets of health studies can be analyzed using statistical methods, the use of data mining methods can also be used in the analysis of health data.

The purpose of this thesis is to analyze the cardiac effects of acute rheumatic fever in childhood using the data mining methods. It is important to determine these effects in order to minimize the damage on cardiac valve the that the disease can cause.

Different classification algorithms are applied and five different algorithms have been tested to determine these effects. Naïve Bayes classifier, decision trees (CART, C4.5, C5.0, boosted by C5.0) and random forest algorithms are used to analyze the patient records with acute rheumatic fever diagnoses and to compare the performances of the algorithms. Algorithms are tried by applying holdout, cross validation and bootstrap methods to data set. In the scope of the study, a data set consisting of records of 297 patients is used in collaboration with Istanbul Medeniyet University Göztepe Education and Research Hospital Children's Clinic. 201 patient records could be included after preprocessing phase. The results obtained from different algorithms were compared according to the model performance evaluation criteria. CART model has given the best result.

This thesis is seen as an important study in terms of contributing to the widespread use of data mining methods in the analysis of data in medicine and health fields.

June 2017, 120 pages.

**Keywords:** Data mining, classification methods, acute rheumatic fever.

# 1. GİRİŞ

Günümüzde içinde yaşadığımız dönem bilgi toplumu ya da bilgi çağı olarak adlandırılmaktadır. Kuşkusuz ki bilginin ve bilgi teknolojilerinin hayatın her alanına nüfus ettiği gözlemlenmektedir. Bu teknolojilerin hızlı gelişimi bilgi kavramının önemi arttırmış, farklı kaynaklar tarafından üretilen çok çeşitli ve büyük hacimli veri ile karşılaşmamıza neden olmuştur. Veri miktarındaki artış ise büyük boysağlıkutlu veriden bilgi elde etme sürecini zorlaştırmaya başlamıştır. İstatistik bilimi veri analizinde yüzyıllardan beri kullanılmaktadır; ancak veri miktarındaki artış, temeli istatistiğe dayanan ve geçmişi o kadar eski olmayan veri madenciliği kavramını öne çıkarmıştır. Veri madenciliği, veri analizinde gün geçtikçe popülerliğini ve işlevini arttıran bir alan olarak karşımıza çıkmakta ve farklı araştırma alanlarında kullanım bulmaktadır.

Ortak bir görüşe varılmış kesin bir tanımı olmamakla beraber veri madenciliği için birbirine paralel tanımlamalar yapılmaktadır; en genel anlamıyla büyük hacimli veri kümeleri arasından anlamlı ve kullanışlı bilgi elde etme süreci olarak tanımlanabilir. Değerli madenlerin toprak altında aranması gibi veri madenciliği çalışmalarında da değerli ve anlamlı olabilecek olan bilginin gün yüzüne çıkarılması amaçlanır ve algoritmalar bu amaç doğrultusunda uygulanır.

Sağlık çalışmalarında, hastalardan elde edilen veri kümeleri uzun yıllardan beri istatistiksel yöntemlerle analiz edilebilmektedir. Bu kapsamda tıp çalışanlarının istatistiksel yöntemlere aşina olduğu bilinmektedir. Veri madenciliği yöntemlerinin sağlık verisi üzerinde uygulamaları olsa da bu çalışmaların alan dışından enformatik, mühendislik, yönetim bilişim sistemleri vb. alanlarda çalışan araştırmacılar tarafından gerçekleştirildiği görülmektedir. Veri madenciliği yöntemleri sağlık çalışanlarına/araştırmacılarına henüz istatistik bilimi kadar yakınlaşamamıştır. Oysa veri madenciliği analizlerinin uygulama alanı bulabileceği alanlardan biri de sağlık alanıdır ve bu alan veri miktarı açısından hem çok zengindir hem de varolan veri miktarı mikro ve makro ölçekte sürekli olarak artmaktadır. Veri madenciliği yöntemleri ile yapılabilecek olan analizler hastalıkların tespiti, hastalık tahmini ile erken teşhis koyma,

hastalıkların gruplandırılması, teşhis koymada doktorlara destek sağlayacak karar destek sistemlerinin geliştirilmesi, ender görülen hastalıkların/anormalliklerin tespit edilmesi gibi farklı şekillerde sağlık alanında katkı sağlayabilir.

Bu tez çalışmasının amacı, veri madenciliği yöntemlerini kullanarak çocuk yaşta görülen akut romatizmal ateşin kalp üzerindeki etkilerinin analiz edilmesidir. Bu etkilerin belirlenmesi hastalığın seyrinin gözlemlenmesi açısından önemlidir. Çalışma hem Türkiye'de sıklıkla görülen bir hastalığın ele alınması hem de sağlık alanında veri madenciliği yöntemlerinin kullanılabileceğini göstermesi açısından önemli görülmektedir. Akut romatizmal ateş ile ilgili yapılan literatür çalışmasında tıp çalışanları tarafından yapılan, istatistiksel yöntemlerin kullanıldığı birçok çalışmaya rastlanmıştır. Kalp ile ilgili veri setleri ile yapılmış çalışmalar olsa da bu hastalığı konu alan ve veri madenciliği yöntemlerinin kullanıldığı bir çalışmaya rastlanmamıştır. Bu sebeble, kullanılan veri seti ve analiz yöntemlerinin sentezlenmesi açılarından bu çalışmanın özgün bir çalışma olduğu düşünülmektedir. Veri madenciliği yönünden bakıldığında, sağlık alanındaki verinin analizinde veri madenciliği metotlarının kullanımının yaygınlaşmasına katkı sağlaması açısından da önemli bir araştırma olarak görülmektedir.

Bu tez çalışmasının GENEL KISIMLAR bölümünde öncelikle, çalışmanın iki ana başlığı olan akut romatizmal ateş hastalığı ve veri madenciliği konuları açıklanmıştır. İlk ana başlık olarak akut romatizmal ateş ile ilgili olarak tanım, tanı koymada kullanılan Jones Kriterleri, bulgular, tedavi ile ilgili genel bilgiler verilmiştir. Aynı zamanda Türkiye'de yapılan akut romatizmal ateş ile ilgili çalışmalara yer verilen literatür çalışması bölümün sonunda yer almaktadır. Aynı bölümde ikinci ana başlık olarak veri madenciliği kavramı ele alınmıştır. Veri madenciliğinin tanımı, farklı araştırmacılar tarafından oluşturulan veri madenciliği süreç tanımları, analizlerde büyük önem taşıyan ön işleme süreci, kullanılan farklı yazılımlar ve uygulama alanlarından bahsedilmiştir. Ardından veri madenciliği yöntemleri genel olarak verilmiş; bu çalışma kapsamında kullanılan sınıflandırma algoritmaları, model performans değerlendirme yöntemleri, model performans değerlendirme ölçütleri ile ilgili bilgiler yer almıştır. Kalp hastalıkları ile ilişkili veri madenciliği çalışmalarını içeren, hem yabancı hem de Türkçe kaynaklarda yapılan literatür çalışması başlığının sonunda verilmiştir. MALZEME VE YÖNTEM bölümü, bir önceki bölümde bahsedilmiş olan Fayyad ve diğ. (1996b) kaynağından alınan veri

madenciliği sürecine paralel olarak oluşturulmuş, bu süreçteki aşamalar takip edilerek analizler gerçekleştirilmiştir. Seçilen sürece paralel olarak beş başlık altında yapılan işlemlerden bahsedilmiştir. Seçim, ön işleme, dönüştürme, veri madenciliği ve yorumlama başlıklarından oluşan bu bölümde veri seti ile ilgili bilgiler, veri setinin orijinal hali, ön işleme aşamasında yapılan değişiklikler ve analizlerde kullanılan algoritmalar, performans değerlendirme yöntemleri ve ölçütleri ile ilgili açıklamalara yer verilmiştir. BULGULAR bölümünde, R programlama dili ve RStudio ile sınıflandırma algoritmaları kullanılarak yapılan analizlerin performans değerlendirmeleri karşılaştırmalı olarak verilmiştir. TARTIŞMA VE SONUÇ bölümünde ise bulgular tartışılmış ve gelecek araştırmalara yönelik öneriler ortaya konulmuştur.

# 2. GENEL KISIMLAR

## 2.1. AKUT ROMATİZMAL ATEŞ

Akut romatizmal ateş (ARA), A grubu beta hemolitik streptokokların neden olduğu, kalbi, eklemleri, deri ve derialtı dokusunu ve merkezi sinir sistemini tutan, üst solunum yolu enfeksiyonu sonrasında ortaya çıkan sistemik inflamatuar bir hastalıktır (Semizel ve diğ., 2005; Saltık, 2007; Düzgün, 2014). Bu hastalık farklı şekillerde kendini göstermektedir. ARA; kalp, deri, beyin ve eklemler üzerinde etkili olsa da kalıcı etkiyi sadece kalp üzerinde bırakmaktadır (Akalın, 2010). Bu durumu, 1884 yılında Ernst-Charles Lasègue "ARA eklemleri yalar, kalbi ısırır." şeklinde ifade etmiştir (Güler Eroğlu, 2016). 1500'lü yıllardan beri bilinmekte olan akut romatizmal ateş (Seckeler ve Hoke, 2011; Güler Eroğlu, 2016), hastalığın mekanizması, komplikasyonları ve dünya üzerinde önemli sağlık problemleri arasında olması nedeniyle araştırmacılar tarafından halen üzerinde çalışılan bir hastalıktır. Streptokok enfeksiyonlarının yoğun olarak görülmesinden dolayı ARA'nın kış ve ilkbahar aylarında daha sık görüldüğü belirtilmektedir (Köksal ve diğ., 2016)

Akut romatizmal ateş; "*Rheumatism*", "*Rheuma*", "*Bouillaud's Disease*", "*Poliarthritis Subacuta Rheumatismus*", "*Poliarthritis Acuta*", "*Poliarthritis Rheumatica Acuta*", "*Rheumatismus Infectiosus*", "*Rheumatismus Cerus*", "*Morbus Rheumaticus Spesificus*" gibi birçok farklı isimle anılmıştır (Saltık, 2007).

### 2.1.1. Görülme Sıklığı (İnsidans)

ARA, az gelişmiş veya gelişmekte olan ülkelerde ve zengin ülkelerdeki yerel alt gruplarda daha sık görülmektedir (Carapetis ve diğ., 2005; Carapetis ve Zühlke, 2011). Kötü yaşam şartları, kötü beslenme, yetersiz tıbbi bakım gibi olumsuz standarttaki yaşam şartlarının ve genetik yatkınlıkların hastalığın ortaya çıkışında birlikte rol oynadığı düşünülmektedir (Düzgün, 2014). ARA, sanayileşmiş, yaşam standartlarının iyi olduğu ülkelerde nadir görülmektedir. Hastalığın gelişmiş ülkelerde 0,5-3/100.000 civarında olduğu tahmin edilmektedir (Örün ve diğ., 2012). Hastalığın sık görüldüğü bölgeler Afrika ülkeleri, Brezilya ve Orta Güney Asya'dır. Ayrıca bazı yerli topluluklarda da ARA daha sık

görülmektedir (Akalın, 2010). Gelişmekte olan ülkelerdeki görülme sıklığının, Carapetis ve diğ. (2005) tarafından 50/100.000 olduğu belirtilmiştir.

Hastalığın Türkiye'de görülme sıklığıyla ilgili genel kapsamlı bir çalışma olmadığı belirtilmektedir. Bu sebeple ya bölgesel çalışmalara ya da genel durumu temsil eden birbirine yakın görülme oranlarına rastlanmaktadır. Saraçlar ve diğ. (1978), tarafından yapılan çalışmada 1972-1976 yılları arasında hastalığın Türkiye'deki görülme sıklığının 20/100.000 olduğunu Özer ve diğ. (2005) ve Örün ve diğ. (2012) aktarmışlardır. Beyazova ve diğ. (1987)'ne ait çalışmada ARA görülme sıklığının 56,5/100,000 olduğu, bu oranın 15 yıl sonraki bir başka çalışmalarında 36,7/100.000'e düştüğü Çağatay ve diğ. (2010) tarafından aktarılmıştır.

Saltık (2007) Türkiye'deki görülme sıklığının bölgesel çalışmalarda 50-100/100.000 arasında bulunduğunu belirtmiştir. Akalın (2010) ise Türkiye'deki görülme sıklığının Orta Doğu ve Akdeniz ülkelerindeki ile benzer sıklıkta, 25-100/100.000 arasında olduğunun düşünüldüğünü belirtmiştir. 1998-2011 yılları arasındaki 624 hastaya ait retrospektif veriden oluşan bir çalışmada Kayseri bölgesindeki ARA görülme sıklığı 7,4/100.000 olarak ortaya konmuştur (Narin ve diğ., 2015). 1980-2009 yılları arasındaki 1.115 hastaya ait veri ile Ankara'da retrospektif bir çalışma yapılmıştır. Bu bölgesel çalışmanın sonuçlarına göre ise 1980-1959 yılları arasında ARA görülme sıklığı 37,6/100.000, 1990-1999 arasında 60/100.000, 2000-2009 arasında ise 21/100.000 olarak ortaya konmuştur (Örün ve diğ., 2012).

### 2.1.2. Ortaya Çıkış Süresi

ARA en sık olarak 5-15 yaş arasındaki bireylerde görülür. Beta hemolitik streptokoklara bağlı boğaz enfeksiyonundan yaklaşık 3 hafta sonra (1-5 hafta), ARA'nın ortaya çıktığı belirtilmiştir (Semizel ve diğ., 2005; Saltık, 2007; Düzgün, 2014).

### 2.1.3. Tanı

ARA tanısı bazı kriterlere ve destekleyici bulgulara göre konmaktadır. Tanı koyarken yararlanılan kriterler Jones Kriterleri olarak adlandırılmaktadır. Bu kriterler 1944 yılında T. Duckett Jones tarafından oluşturulmuştur (Jones, 1944). Kriterler; 1965, 1984, 1992 (Amerikan Kalp Derneği – AHA tarafından), 2002/2003 (Dünya Sağlık Örgütü – WHO tarafından) yıllarında revize edilmiştir (Akalın, 2010; Seckeler ve Hoke, 2011; Güler

Eroğlu, 2016; Köksal ve diğ., 2016). Son güncelleme ise 2015 yılında gerçekleşmiştir (Gewitz ve diğ., 2015) (Tablo 2.1).

Köksal ve diğ. (2016) kriterlerin güncellemesindeki amacın "ilk atak ARA tanısında klinisyenlere yardımcı olmak ve yanlış tanı olasılığını en aza indirmek" olduğunu belirtmiştir. Carapetis ve diğ. (2005) her revizyonun kriterlerin spesifikliğini arttırırken hassaslığını azalttığını belirtmiştir. Yazarlara göre Jones Kriterleri teşhis kriterleri olmalarına rağmen, ARA riskinin yüksek olduğu toplumlar gibi bazı koşullara adapte edilerek hassaslığın arttırılması gerekmektedir.

**Tablo 2.1:** 2015 yılında revize edilmiş Jones Kriterleri (Gewitz ve diğ., 2015; Güler Eroğlu, 2016).

| **Güncellenen Jones Kriterleri** | |
|---|---|
| **A.** Tüm hastalarda geçirilmiş A grubu streptokok enfeksiyonu kanıtı olmalı (kore dışında) | |
| Tanı: ilk atak ARA | İki majör veya bir majör, iki minor bulgu |
| Tanı: tekrarlayan atak ARA | İki majör veya bir major, iki minör veya üç minör |
| | |
| **B.** Majör bulgular | |
| Düşük riskli topluluklara | Orta ve yüksek riskli topluluklar |
| Kardit (Klinik ve/veya subklinik[b]) | Kardit (Klinik ve/veya subklinik[b]) |
| Artrit (Sadece poliartrit) | Artrit (Monoartrit veya poliartrit veya poliartralji[c)] |
| Kore | Kore |
| Eritema marginatum | Eritema marginatum |
| Deri altı nodülleri | Deri altı nodülleri |
| | |
| **C.** Minör bulgular | |
| Düşük riskli topluluklar[a] | Yüksek riskli topluluklar |
| Poliartralji | Monoartralji |
| Ateş (≥38,5°C) | Ateş (≥38°C) |
| EÇH ≥60 mm/sa ve/veya CRP ≥3 mg/dL[d] | EÇH ≥30 mm/sa ve/veya CRP ≥3 mg/dL[d] |
| EKG'de uzamış PR (yaşa göre) | EKG'de uzamış PR (yaşa göre) |
| (kardit major bulgu değilse) | (kardit major bulgu değilse) |

[a]Düşük riskli topluluklarda okul çağı çocuklarında ARA sıklığı yılda ≤2/100 000 veya tüm yaşlarda romatizmal kalp hastalığı ≤1/1 000.
[b]Subklinik kardit patolojik ekokardiyografik valvulittir.
[c]Poliartralji yüksek risklilerde diğer nedenler dışlanırsa major bulgu kabul edilir. Eritema marginatum ve deri altı nodülleri eskisi gibi nadiren tek başına majör bulgu
kabul edilir. Eklem bulguları aynı hastada ya majör ya da minör bulgu kabul edilir.
[d]CRP laboratuvarın üst sınırının üstünde olmalıdır, en yüksek çıkan EÇH değeri kullanılır.

Tanı konabilmesi için ilk atak için iki majör kriterin veya bir majör iki minör kriterin bulunması ve A grubu beta streptokok enfeksiyonunun geçirilmiş olduğuna ait bulgularla desteklenmesi gerekmektedir. Tekrarlayan ataklarda, geçirilmiş streptokok kanıtının

varlığında son kriterlere göre, iki majör veya bir majör, iki minör veya üç minör bulgu ile tanı koyulabilmektedir (Güler Eroğlu, 2016).

Ancak istisnai durumlarda kriterlerin sağlanmasına gerek olmadan ARA tanısı konabilmektedir. Aşağıdaki üç durum için iki majör veya bir majör iki minör kriterin varlığı aranmadan ARA tanısı konabilmektedir (Behrman ve diğ., 2000; Park, 2008):

- Sydenham Koresi varlığında A grubu beta streptokok bulgusu olması gerekmez.
- Sinsi kardit varsa A grubu beta streptokok bulgusunun olması gerekmez.
- Tekrarlayan ataklarda bir majör veya birkaç minör kriterle beraber önceden geçirilmiş A grubu beta streptokok enfeksiyonu kanıtına ihtiyaç duyulmaktadır.

Tanı koyarken, klinik açıdan yüksek şüphenin olduğu durumlarda, hem ilk atak hem tekrarlayan atak için, bütün kriterleri sağlamasa bile hastalara *"şüpheli tanı"* konabilmektedir (Sika-Paotonu ve diğ., 2017). Bu gibi durumlarda doktorun tecrübesine bağlı olarak karar vermesi büyük önem taşımaktadır (Güler Eroğlu, 2016).

### 2.1.4. Majör Bulgular

Majör bulgular; kardit, artrit, Sydenham koresi, eritema marginatum ve cilt altı nodüllerdir.

#### *2.1.4.1. Kardit*

ARA'nın kalpte yaptığı inflamatuar değişiklikler kardit olarak adlandırılmaktadır; bu değişikler farklı ağırlıkta olabilir. Saltık (2007) karditin, kalpte kalıcı hasar bırakması sebebiyle ARA'nın en ciddi majör bulgusu olduğunu belirtmiştir Semizel ve diğ. (2005) ve Akalın (2010) karditin, hastaların %45-50'sinde görüldüğünü aktarmıştır. Kalpteki tutulum kalbin farklı tabakalarını kapsayabilir ve farklı derecelerde olabilir (Saltık, 2007; Düzgün, 2014). Kalbin en iç tabakası endokard, ortadaki tabaka miyokard, en dıştaki tabaka ise perikard olarak adlandırılmaktadır. Kalpteki tutulum da bu sırayı izleyerek içeriden dışarıya doğru bir şekilde gerçekleşir; en dıştaki tabakanın iltihaplanması için en içerideki tabakanın iltihaplanmış olması gerekir (Akalın, 2010; Köksal ve diğ., 2016).

Akalın (2010), kalp tabakalarının tutulumu ile ilgili olarak aşağıdaki bilgileri aktarmıştır:

- **Endokard tutulumu**

Endokard tabakasındaki tutulum kalpteki kapaklarla ilgili yetersizlikleri kapsamaktadır. Kalpte yer alan dört farklı kapak vardır; mitral kapak, aort kapağı, triküspit kapak ve pulmoner kapak. Bunlar arasında en sık mitral kapak ikinci olarak aort kapağı tutulur. Diğer iki kapak daha nadir olarak tutulur.

- **Miyokard tutulumu**

Miyokard tutulumu, kalbin kas yapısındaki/orta tabakasındaki tutulumdur.

- **Perikard tutulumu**

Perikard tutulumu, kalp zarının iltihaplanmasıdır. Bazı durumlarda ise kardit sessiz bir şekilde farkettirmeden, herhangi bir bulgu ortaya çıkarmadan seyredebilir. Bu duruma ise sessiz kardit denir (Akalın, 2010; Düzgün, 2014).

#### *2.1.4.2. Artrit*

Artrit, ARA'da en sık görülen bulgudur (Semizel ve diğ., 2005; Saltık, 2007). Düzgün (2014) hastaların %75'inde artrit görüldüğünü bildirmiştir. Akalın (2010)'a göre hastaların %75-80'inde artrit görülmektedir. Semizel ve diğ. (2005)'ne göre ise bu oran %60-80 aralığındadır.

Artritte genellikle diz, ayak bileği, el bileği dirsek, omuz gibi büyük eklemler tutulur ve tutulma gezici bir şekilde gerçekleşir (Semizel ve diğ., 2005; Saltık, 2007; Düzgün, 2014). Yani tutulan eklem bölgeleri zaman içerisinde değişiklik gösterir; bir kısmı iyileşirken diğer eklemler bulgular ortaya çıkar. Kardit kalpte kalıcı hasar bırakabilirken artrit vücutta kalıcı hasar bırakmaz (Akalın, 2010).

#### *2.1.4.3. Kore*

Kore, Sydenham Koresi olarak adlandırılmaktadır. Koredeki tutulum beyinde gerçekleşir. Koresi olan bir kişi; istemsiz, amaçsız, hızlı, düzensiz hareketler gerçekleştirir (Semizel ve diğ., 2005; Saltık, 2007; Düzgün, 2014).

Semizel ve diğ. (2005) ve Saltık (2007) kore görülme sıklığının hastalarda %10-15 aralığında olduğunu belirtmiştir. Düzgün (2014) hastaların %10-30'unda kore geliştiğini söylemektedir. Akalın (2010) ise hastaların %15-20 aralığında kore görüldüğünü belirtmiştir.

#### *2.1.4.4. Eritema Marginatum*

Nadir görülen belirtilerden biri olan eritema marginatum deride oluşan kaşıntısız, ortası soluk, koyu pembe renkte döküntülerdir (Semizel ve diğ., 2005; Saltık, 2007; Akalın, 2010). Eritema marginatum da artrit gibi gezici bir karaktere sahiptir; günden güne vücudun farklı yerlerinde kendini gösterebilir (Saltık, 2007).

Semizel ve diğ. (2005) ve Düzgün (2014) eritema marginatumun hastaların sadece %5'inde görüldüğünü belirtmiştir. Akalın (2010) hastalarda %5-10 aralığında eritema marginatum görüldüğünü belirtmiştir.

#### *2.1.4.5. Cilt Altı Nodülleri / Subkütan Nodüller*

Nadir görülen cilt altı nodülleri, adından da anlaşılacağı üzere, deri altında oluşan ağrısız nodüllerdir (Semizel ve diğ., 2005; Saltık, 2007).

Semizel ve diğ. (2005) ise daha küçük bir oran vererek görülme sıklığının %1'den az olduğunu belirtmiştir. Akalın (2010) hastaların %5'inde subkütan nodüllerin görüldüğünü belirtmiştir. Düzgün (2014) cilt altı nodüllerinin görülme sıklığının %10-20 arasında olduğunu söylemiştir.

### 2.1.5. Minör Bulgular

Minör bulgular; ateş, atralji, PR aralığında uzama ve akut faz reaktanlarındaki yükselmedir.

#### *2.1.5.1. Ateş*

Ateş, ARA'nın minör kriterlerinden biridir. Saltık (2007)'a göre ateş 37,8-40 derece arasında değişmektedir. Akalın (2010), 38 derece ve üzerindeki ateşin minör bulgu olarak kabul edildiğini belirtmiştir. Düzgün (2014) ise, ateşin 38-40 derece arasında olduğunu belirtmektedir.

#### *2.1.5.2. Atralji*

Atralji, şişlik veya kızarıklık olmadan görülen eklem ağrısıdır (Seckeler ve Hoke, 2011). Büyük eklemlerde görülür. Artrit varsa atralji minör bulgu olarak kabul edilmemektedir (Saltık, 2007; Akalın, 2010; Seckeler ve Hoke, 2011).

#### *2.1.5.3. PR Aralığında Uzama*

EKG sonucuna göre PR aralığında uzama olması ARA'nın minör kriterlerinden sayılır (Akalın, 2010; Saltık, 2007). Karditli hastalarda PR uzaması minör kriter sayılmamaktadır (Köksal ve diğ., 2016).

#### *2.1.5.4. Akut Faz Reaktanlarında Artışı*

Kandaki lökosit sayısı, eritrosit çökme (sedimantasyon) hızının (EÇH) artması ve C-reaktif proteindeki (CRP) yükselme akut faz reaktanlarındaki artış olarak ele alınır ve vücuttaki inflamasyonu gösterir (Semizel ve diğ., 2005; Düzgün, 2014; Köksal ve diğ., 2016).

### 2.1.6. Destekleyici Bulgular

Majör ve minör kriterlere ek olarak, tanı konmasında destekleyici olması amacıyla, geçirilmiş A grubu beta streptokok enfeksiyonunun varlığına dair kanıt gösterilmesi gerekmektedir. Çünkü bu bulgular başka hastalıklarda da görülebilmektedir. Enfeksiyon geçirilip geçirilmediğine, aşağıdaki bulgulara göre karar verilir (Seckeler ve Hoke, 2011; Gewitz ve diğ., 2015):

- Boğaz kültürü sonuçlarının A grubu beta hemolitik streptokokları için pozitif olması
- Hızlı antijen testinin pozitif olması
- Antistreptolizin antikor (ASO) titresinde artış

Bunların dışında Dünya Sağlık Örgütü tarafından yapılan düzenleme ile geçirilmiş kızıl öyküsü de enfeksiyon geçirildiğine dair kanıtlardan biri olarak sayılmaktadır (Akalın, 2010).

### 2.1.7. Tedavi ve Korunma

ARA tedavisi hastalığın akut döneminde bulguların tedavisine dayanmaktadır. Akut tedavi dışında hastalığın oluşmasının engellenmesi için yapılan koruyucu tedavi de

yapılması gerekmektedir. Hastalığın hiç oluşmadan streptekok enfeksiyonunun önlenmesi primer proflaksi (birincil önlem), hastalık oluştuktan sonra tekrarlama ve nükslerin önlenmesi için yapılan korumaya sekonder proflaksi (ikincil önlem) denilmektedir. Ağır vakalarda ise cerrahi müdahale gerekebilmektedir.

#### *2.1.7.1. Akut Tedavi*

Akut tedavi, antiinflamatuar ve korkikosteroid ilaç kullanımını ve yatak istirahatini içerir. Sıklıkla kullanılan antiinflamatuar ilaçlar aspirin (salisilatlar) ve steroidlerdir (kortizon) (Akalın, 2010; Seckeler ve Hoke, 2011; Köksal ve diğ., 2016). Bu ilaçlar; eklem, kalp ve diğer dokulardaki iltihaplanmanın baskılanması ve semptomların giderilmesi amacıyla kullanılır (Düzgün, 2014).

#### *2.1.7.2. Primer Proflaksi*

Primer proflaksi, A grubu beta hemolitik streptokoklara bağlı olarak gelişen üst solunum yolu enfeksiyonlarının antibiyotiklerle yapılan erken tedavisini içerir (Saltık, 2007; Akalın, 2010). Dolayısıyla birincil önlem ARA'nın ilk atağının önlenmesine yönelik korunmayı ifade eder (Semizel ve diğ., 2005). Penisilin, benzatin penisilin, makrolid, eritromisin tedavide kullanılan antibiyotiklerdendir (Akalın, 2010; Seckeler ve Hoke, 2011). Antibiyotikler oral yoldan (ağızdan) veya enjeksiyon yoluyla verilmektedir.

#### *2.1.7.3. Sekonder Proflaksi*

Sekonder proflaksi, A grubu beta hemolitik streptokoklara bağlı olarak gelişebilecek olan üst solunum yolu enfeksiyonlarının oluşmasını engelleyerek ARA'nın tekrarlamasının önlenmeye çalışılmasıdır (Semizel ve diğ., 2005; Saltık, 2007). Penisilin tedavisi ile hastalığın tekrarlamasının önüne geçilmeye çalışılır, bazı hastalarda ömür boyu penisilin ile proflaksi önerilir (Akalın, 2010).

Cerrahi tedavi, akut tedavide tercih edilmemekle beraber, ağır kapak yetersizliği olan hastalarda kapak replasmanı için yapılır (Akalın, 2010; Köksal ve diğ., 2016).

### 2.1.8. Literatür Taraması

Türkiye'de ARA ile ilgili birçok farklı çalışma yapılmaktadır. Bu bölümdeki literatür taramasında ARA ile ilgili olarak yapılmış olan çalışmalar bir araya getirilerek Tablo 2.2'de sunulmuştur. Bu kapsamda bulunan çalışmaların büyük bir kısmının retrospektif çalışmalardan oluştuğu görülmektedir.

**Tablo 2.2:** Türkiye'de ARA ile ilgili yapılan çalışmalar.

| Çalışma Başlığı | Kişi Sayısı | Araştırmanın Kapsadığı Tarihler | Araştırmanın Yapıldığı Kurum | Kaynak |
|---|---|---|---|---|
| Rheumatic Heart Disease Prevalence Among Schoolchildren in Ankara, Turkey | 4086 | Mart 1995-Haziran 1995 | Gazi Üniversitesi, Tıp Fakültesi, Pediatrik Kardiyoloji Bölümü | (Olguntürk ve diğ., 1999) |
| Acute Rheumatic Fever in Konya, Turkey | 274 | Haziran 1993-Haziran 1998 | Selçuk Üniversitesi, Tıp Fakültesi, Pediatrik Kardiyoloji Bölümü | (Karaaslan ve diğ., 2000) |
| Bursa İlindeki Çocuklarda Akut Romatizmal Ateşin Değerlendirilmesi | 207 | Ocak 1994-Temmuz 2000 | Uludağ Üniversitesi Tıp Fakültesi, Çocuk Sağlığı ve Hastalıkları Anabilim Dalı, Çocuk Kardiyoloji Bilim Dalı | (Bostan ve Çil, 2001) |
| Çocukluk Çağında Görülen Akut Romatizmal Ateş Olgularımızın Retrospektif Değerlendirilmesi | 121 | Ocak 1993 - Ocak 1999 | Ankara Eğitim ve Araştırma Hastanesi Çocuk Sağlığı ve Hastalıkları Kliniği | (Dallar ve diğ., 2002) |
| Akut Romatizmal Ateş: 60 Olgunun Retrospektif Değerlendirilmesi | 60 | Ocak 1997-Ocak 2000 | Haydarpaşa Numune Eğitim ve Araştırma Hastanesi, Çocuk Kliniği | (Erol ve diğ., 2002) |
| Review of 609 Patients with Rheumatic Fever in Terms of Revised and Updated Jones Criteria | 609 | Ocak 1982-Ocak 1991 | Gazi Üniversitesi, Tıp Fakültesi, Pediatrik Kardiyoloji Bölümü | (Olgunturk ve diğ., 2006) |
| Childhood Acute Rheumatic Fever in Ankara, Turkey | 129 | Ocak 1999-Temmuz 2002 | Hacattepe Üniversitesi, Tıp Fakültesi, Pediatrik Kardiyoloji Bölümü | (Özer ve diğ., 2005) |
| Long Term Follow-up Results of 139 Turkish Children and Adolescents with Rheumatic Heart Disease | 139 | Mart 1989-Haziran 2003 | İstanbul Tıp Fakültesi, Çocuk Sağlığı Ve Hastalıkları Anabilim Dalı, Pediatrik Kardiyoloji Bilim Dalı | (Yavuz ve diğ., 2008) |
| Akut Romatizmal Ateş: Klinik Bir Değerlendirme | 45 | Ocak 2000-Aralık 2008 | Zeynep Kamil Kadın ve Çocuk Hastalıkları Eğitim ve Araştırma Hastanesi, Çocuk Kliniği | (Çağatay ve diğ., 2010) |
| Akut Romatizmal Ateşte Sessiz Düşman: Subklinik Kardit | 80 | Mayıs 2007-Mayıs 2010 | Keçiören Eğitim ve Araştırma Hastanesi Çocuk Sağlığı ve Hastalıkları Kliniği | (Osman Özdemir ve diğ., 2011) |

**Tablo 2.2 (devam):** Türkiye'de ARA ile ilgili yapılan çalışmalar.

| Çalışma Başlığı | Kişi Sayısı | Araştırmanın Kapsadığı Tarihler | Araştırmanın Yapıldığı Kurum | Kaynak |
|---|---|---|---|---|
| Acute Rheumatic Fever in the Central Anatolia Region of Turkey: a 30-year Experience in a Single Center | 1115 | Ocak 1980-Aralık 2009 | Dr. Sami Ulus Kadın Doğum Çocuk Sağlığı ve Hastalıkları Eğitim ve Araştırma Hastanesi | (Örün ve diğ., 2012) |
| Akut Romatizmal Ateşli Çocuklarda Klinik ve Laboratuar Bulguların Geriye Dönük Olarak İncelenmesi: Reaktivasyon ve Koruyucu Tedaviye Uyumun Araştırılması | 255 | Ocak 2004-Ocak 2008 | Dicle Üniversitesi Tıp Fakültesi, Çocuk Sağlığı ve Hastalıkları Anabilim Dalı, Kardiyoloji Bölümü | (Gözü Pirinççioğlu ve diğ., 2012) |
| Incidence and Clinical Features of Acute Rheumatic Fever in Kayseri, Central Anatolia, 1998–2011 | 624 | Ocak 1998-Aralık 2011 | Erciyes Üniversitesi Tıp Fakütesi, Çocuk Kardiyolojisi Bölümü | (Narin ve diğ., 2015) |
| Akut Romatizmal Ateş Tanısı Konulan Hastaların Klinik Özellikleri ve Ekokardiyografik Bulguları | 65 | Ocak 2010-Mayıs 2014 | Selçuk Üniversitesi, Tıp Fakültesi, Çocuk Kardiyolojisi Bölümü | (Yılmaz ve diğ., 2015) |

## 2.2. VERİ MADENCİLİĞİ

Veri madenciliği, günümüzde veri analizinde sıklıkla karşılaşılan bir kavram olup birçok farklı disiplinle bağlantısı olan bir araştırma alanıdır. Genel geçer bir tanımını yapmak zor olsa da en genel tanım, büyük veri kümelerinden/setlerinden çeşitli algoritmalar ve veri analiz araçları/yazılımları kullanılarak, anlamlı, işe yarar bilgilerin keşfedilmesi olacaktır.

Veri madenciliği çalışmalarında, Usama Fayyad, Gregory Piatetsky-Shapiro ve Padhraic Smyth tarafından 1996 yılında yazılmış, *"The KDD Process for Extracting Useful Knowledge from Volumes of Data"* ve *"From Data Mining to Knowledge Discovery in Databases"* makaleleri temel kaynaklardır. İki makale birbirine paralel olarak, veri madenciliği ve veri tabanlarında bilgi keşfi kavramlarını ortaya koymakta ve sürecin adımlarını incelemektedir. KDD, İngilizce'de *"knowledge discovery in databases"* kavramının kısaltması yerine kullanılmaktadır ve bu kavram Türkçe'ye "veri tabanlarında bilgi keşfi" olarak çevrilmektedir.

Fayyad ve diğ. (1996b)'ne göre orijinal metindeki tanım şu şekildedir*: "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."*

Akpınar (2014); Fayyad ve diğ. (1996b) tarafından yapılan veri madenciliği tanımının çevirisini şu şekilde aktarmaktadır: *"Veri madenciliği veya veri tabanlarında bilgi keşfi, veri dizilerinden geçerli, yeni, mümkünse faydalı ve anlaşılır örüntülerin ortaya çıkartılabilmesi için gerçekleştirilen apaçık olmayan bir süreçtir."*

Farklı araştırmacılar tarafından birbirine benzer veri madenciliği tanımları yapılmıştır:

*"Veri madenciliği, umulmadık ilişkileri bulmak ve veriyi, veri sahibi tarafından hem anlaşılabilir hem de yararlı olması amacıyla özgün yollarla özetlemek için (genellikle büyük) gözlemsel veri kümelerinin analizidir." (Hand ve diğ., 2001)*

*"Veri madenciliği, veri içerisindeki kullanışlı örüntülerin bulunması sürecidir." (Roiger ve Geatz, 2003)*

*"Veri madenciliği, veri içerisindeki örüntüleri keşfetme sürecidir. Süreç otomatik veya (genellikle) yarı otomatiktir. Keşfedilen örüntüler bazı avantajlara, yani genellikle ekonomik avantaja yol gösterdiğinden anlamlı olmalıdır."(Witten ve diğ., 2011)*

*"Veri madenciliği, büyük miktarlardaki veriden ilginç örüntüleri keşfetme sürecidir." (Han ve diğ., 2012)*

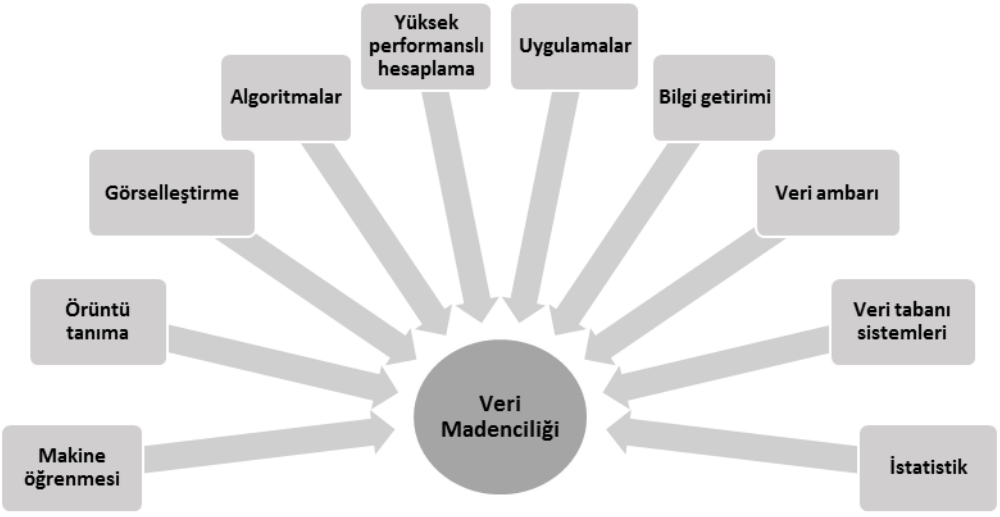
*"Veri madenciliği, büyük ölçekli veriler arasından "değeri olan" bir bilgiyi elde etme işidir." (Özkan, 2013)*

Kavramsal olarak bakıldığında; veri madenciliği çalışmaları "değerli" olan bilgiyi ortaya çıkarma amacıyla yapıldığından, analizler, değerli madenlerin elde edilmesi için yapılan gerçek madencilik çalışmalarına benzetilmekte ve veri madenciliği, ismini buradan almaktadır (Zaïane, 1999). Han ve diğ. (2012) kavramla ilgili olarak bir noktaya dikkat çekmiştir. Örneğin altın madenciliği kavramı, aranan değerli madenin "altın" olduğunu işaret ederken, veri madenciliği kavramının odak noktası "bilgi"yi aramaktır. Yazarlar bununla ilgili olarak, aranan şey olan bilgiyi vurgulamak amacıyla, *"knowledge mining*

*from data"(veriden bilgi madenciliği)* kavramının daha uygun olabileceğini belirtmiştir. Ancak bu durumun büyük miktardaki "veri"nin önemini yansıtmayabileceğini de aktarmışlardır. Günümüzde kavram veri madenciliği olarak kullanılmaktadır.

Han ve diğ. (2012)'ne göre veri miktarındaki fazlalık güçlü veri analiz araçlarına olan ihtiyacı doğurmaktadır. Bu durumu ise yazarlar *"a data rich but information poor situtation"* olarak ifade etmektedirler. Yani günümüzde elimizde çok büyük boyutlarda veri bulunmakta ancak bunlardan elde edilen enformasyon miktarı veri boyutlarının yanında küçük kalmaktadır. Bu nedenlerle veri madenciliği, veri ve enformasyon arasındaki boşluğu da doldurarak bilgiye giden yolu aydınlatacaktır. Ayrıca Hand ve diğ. (2001) veri madenciliğinin tek seferlik bir uygulama olarak görülmemesi gerektiğini belirtmiştir. Büyük veri setleri veri madenciliği yöntemleri ile birden fazla kez ve sınırsız farklı şekilde analiz edilebilir.

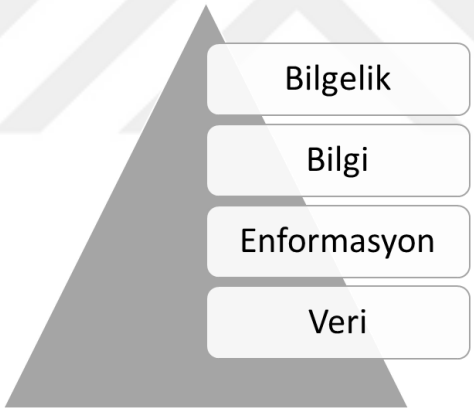
Veri madenciliği, disiplinler arası bir yapıya sahiptir (Han ve diğ., 2012). Birçok alandaki çalışmadan beslenen veri madenciliğinin çok disiplinli yapısı Şekil 2.1'de görülmektedir. Veri madenciliği hem bu alanlardaki çalışmalardan etkilenmekte hem de yöntemlerden yararlanmaktadır. Dolayısıyla alanlar arasında keskin bir ayrım yaparak alanları birbirinden ayırmak mümkün değildir (Hand ve diğ., 2001).




**Şekil 2.1:** Veri madenciliğinin ilgili olduğu alanlar (Han ve diğ., 2012).

Veri madenciliğini geleneksel metotlardan ayıran özelliğinin farklı türdeki veriden oluşan çok büyük veri kümelerine ölçeklenebilir *(scalability)* olmasıdır (Maimon ve Rokach, 2010). Ölçeklenebilirlik, Maimon ve Rokach (2010) tarafından çok sayıda kayıt, çok boyutluluk, çok sayıda sınıf ya da heterojenliğin varlığı olarak tanımlanmaktadır.

Veri kümeleri arasından bilginin keşfi bilgi piramidinde de görülebilmektedir. Piramit veri madenciliği süreçlerinin temel amacı olan veri kümelerinden değerli olan bilginin keşfedilmesini temsil etmektedir. Veriden bilgiye ve en sonunda bilgelik aşamasına giden yol *(DIKW – data, information, knowledge, wisdom)* Şekil 2.2'deki piramitte görülmektedir. "Bilgi hiyerarşisi" *(knowledge hierarchy)*, "Enformasyon hiyerarşisi" *(information hierarchy)*, "Bilgi piramidi" *(knowledge pyramid)* gibi isimlerle de anılan bu piramit, literatürde bu dört kavram arasındaki ilişkiyi gösteren temel bir modeldir (Rowley, 2007). Chaffey and Wood (2005), veriden bilgiye doğru çıkıldıkça "anlam" ve "değer"in arttığını aktarmıştır (Rowley (2007).




**Şekil 2.2:** Veri, enformasyon, bilgi, bilgelik (DIKW) piramidi.

Ackoff (1989), *"From Data to Wisdom"* başlıklı çalışmasında bu kavramlardan bahsetmiş ve aralarındaki bağlantıları ele almıştır. Araştırmacılar tarafından hala bu kavramlar tartışılmaktadır.

- Veri - Objelerin ve olayların özellikleridir (Ackoff, 1989). Hamdır (Ahsan ve Shah, 2006). Gözlemlerin sonucudur, enformasyona dönüştürülünceye kadar bir değeri yoktur (Bernstein, 2009).
- Enformasyon - Tanımlamalardır. Kim, ne, ne zaman, nerede, kaç tane gibi soruların cevaplarını içerir (Ackoff, 1989). Bağlantılar yoluyla verinin bir anlam

kazanmasıdır (Ahsan ve Shah, 2006). Soruların cevaplarını barındırır (Bernstein, 2009).

- Bilgi - Talimatları içerir, nasıl yapılır sorusunun cevaplarını içerir (Ackoff, 1989). İşe yarar enformasyonların bütünüdür (Ahsan ve Shah, 2006).
- Bilgelik - Değerlerle ilgilenir; sorgulamayı barındırır (Ackoff, 1989). Eylemlerin, uzun vadeli sonuçlarını görme ve bunları kontrol etme fikrine göre değerlendirme yeteneği anlamına gelir (Bernstein, 2009).

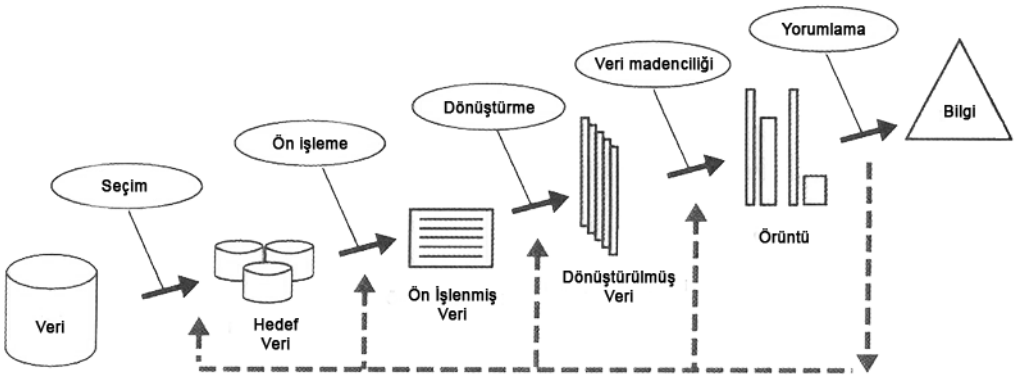
Büyük veri setleri veya kümeleri, günümüzde veri kapasitlerinin artık "terabayt", "petabayt" gibi birimlerle ifade edildiği veri depolarını tanımlamaktadır (Maimon ve Rokach, 2010). Gelişen bilgi teknolojileri ile birçok farklı alanda her gün terabaytlarca veri üretilmekte ve saklanmaktadır. Verinin büyüklüğünün yanında çeşitliliği de dikkat çekmektedir. Web verisi, sosyal medyada üretilen veri, biyolojik veri ya da multimedya verisi (ses, video, görüntü verisi), zaman serisi verisi gibi birçok farklı kaynaktan çok çeşitli veri üretilmektedir (Zhang ve Zhang, 2010; Han ve diğ., 2012). Zhang ve Zhang (2010) günümüzde bilgi keşfinin, sadece geleneksel veri tabanlarındaki yapılandırılmış veriye odaklanmasının yeterli olamayacağını ve veri tabanlarından veri ambarlarına, yapılandırılmış veriden yapılandırılmamış veriye doğru bir değişimin gözlendiğini belirtmiştir. Yapılandırılmış, yarı yapılandırılmış veya yapılandırılmamış veri olarak adlandırılan farklı kaynaklardan türeyen farklı tipteki verinin içerisinden bilgi keşfini gerçekleştirmek veri madenciliği alanının önüne zorluklar çıkarmaktadır (Han ve diğ., 2012). Ancak günümüzde veri madenciliği ile yapılan çalışmaların fazlalığı ve bu alana gösterilen yoğun ilgi farklı şekillerde bu zorlukların aşılmaya çalışıldığını göstermektedir.

Mevcut verinin türü aynı olmamakla beraber veri setlerinin çok boyutluluğu da veri madenciliği araştırmalarındaki zorluklardan birini oluşturmaktadır. Veri kümelerinin çok fazla sayıda nitelik yani sütundan oluşması dikkat çekmekte ve veri kümeleri ile analiz yapılmasını zorlaştırmaktadır. Fayyad ve diğ. (1996a), 21 yıl önce bu durumun analizlerde zorluk yaratabildiğinden bahsetmiştir. Bu durum literatürde "çok boyutluluğun laneti" *(curse of dimensionality)* olarak adlandırılmaktadır (Maimon ve Rokach, 2010). Veri setlerinin boyutlarındaki artış analizler esnasında daha fazla bellek ihtiyacı ve düşük işlem/hesaplama hızı olarak geri dönmektedir (Gorunescu, 2011).

### 2.2.1. Veri Madenciliği Süreci

Farklı kaynaklara bakıldığında veri madenciliği ile ilgili süreçlerin birkaç farklı şekilde ele alındığı görülmektedir. KDD *(Knowledge Discovery in Databases),* SEMMA *(Sample, Explore, Modify, Model, Assess)* ve CRISP-DM *(Cross-Industry Standard Process for Data Mining)* isimli üç farklı süreçten söz edilmektedir (Fayyad ve diğ., 1996a, 1996b; Olson ve Delen, 2008). Akpınar (2014) ise bunlara ek olarak yeni bir süreç önerisinde bulunmaktadır. Roiger ve Geatz (2003), farklı araştırmacılar tarafından veri tabanlarında bilgi keşfi süreci 4 ila 12 adımda gösterilebildiğini ancak adımların sayısı farketse de içeriklerin birbirine paralel şekilde oluşturulduğunu belirtmiştir. Bu tez çalışması kapsamında ise alandaki temel çalışmalardan biri olduğu için Fayyad ve diğ. (1996b)'ne ait olan süreçler takip edilmiştir.

Fayyad ve diğ. (1996b) tarafından veri tabanlarında bilgi keşfi süreci Şekil 2.3'teki gibi ortaya konmuştur. Buna göre veriden bilgiye giden süreç beş adımdan oluşmaktadır ve veri madenciliğinin bu sürecin adımlarından biri olduğu görülmektedir.




**Şekil 2.3:** Veri madenciliği süreci (Fayyad ve diğ., 1996a; Akpınar, 2014).

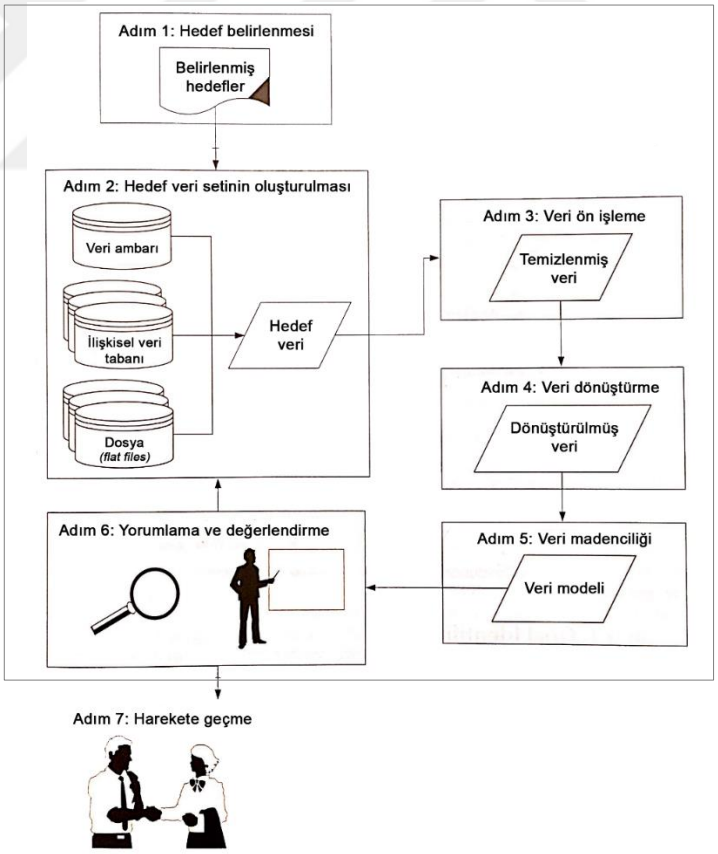
Adımlar aşağıdaki gibi açıklanmaktadır (Fayyad ve diğ., 1996a, 1996b; Akpınar, 2014) :

1- Seçim *(Selection)*: Verinin seçilmesini temsil eden aşamadır.
2- Ön işleme *(Preprocessing)*: Verinin analize hazır hale getirilmesi için işlendiği aşamadır.
3- Dönüştürme *(Transformation)*: Analiz için verinin uygun formata dönüştürüldüğü aşamadır.
4- Veri madenciliği *(Data mining)*: Çalışmanın amacına uygun olan veri madenciliği yöntemlerinin veriye uygulandığı aşamadır.

5- Yorumlama *(Interpretation)*: Veri madenciliği ile elde edilen sonuçların yorumlandığı aşamadır.

Fayyad ve diğ. (1996a) veri tabanlarında bilgi keşfi sürecinin, verinin elde edilmesi ile ilgili sürecin tamamını kapsayan bir kavram olduğunu, veri madenciliğinin ise bu süreç içerisindeki adımlardan biri olduğunu ve veri içerisindeki örüntüleri keşfedebilmek amacıyla çeşitli algoritmaların veriye uygulandığı adım olduğunu belirtmiştir. Ancak kimi araştırmacılar tarafından veri madenciliği; veri tabanlarında bilgi keşfi kavramının yerine kullanılabilirken, kimi araştırmacılar tarafından veri tabanlarında bilgi keşfinin bir adımı olarak kabul edilmektedir (Han ve diğ., 2012). Akpınar (2014)'ın aktarımında veri madenciliği ve veri tabanlarında bilgi keşfi kavramları birbirine eş kavramlar olarak kullanıldığı görülmektedir.

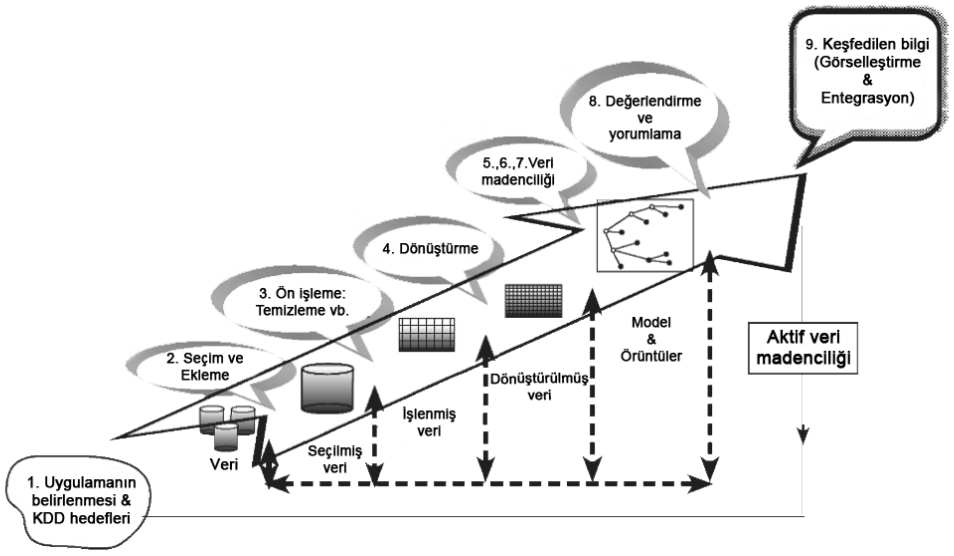
Roiger ve Geatz (2003) ise yedi adımda KDD sürecini tanımlamaktadır (Şekil 2.4):




**Şekil 2.4:** Yedi adımda veri tabanlarında bilgi keşfi süreç modeli (Roiger ve Geatz, 2003).

1- Hedef belirlenmesi *(Goal identification)*: Amacın belirlendiği, bilgi keşfi ile neye ulaşılmak istendiğinin/neyin başarılmak istendiğinin belirlendiği aşamadır.
2- Hedef veri setinin oluşturulması *(Creating a target data set)*: Analiz edilecek veri setinin uzman desteği veya bilgi keşfi araçları kullanılarak bulunduğu aşamasıdır.
3- Veri ön işleme *(Data preprocessing)*: Eksik verinin ne yapılacağına karar verildiği, verinin temizlendiği aşamadır.
4- Veri dönüştürme *(Data transformation)*: Verinin ihtiyaca uygun olarak farklı formatlara dönüştürüldüğü aşamadır.
5- Veri madenciliği *(Data mining)*: Bir ya da birden fazla veri madenciliği algoritmasının veri setine uygulandığı aşamadır.
6- Yorumlama ve değerlendirme *(Interpretation and evaluation)*: Veri madenciliği algoritması ile elde edilen sonuçların kullanışlı ve ilginç olup olmadığının yorumlandığı ve önceki aşamaların tekrarlanmasına gerek olup olmadığına karar verildiği aşamadır.
7- Harekete geçme *(Taking action)*: Keşfedilen bilgi kullanışlı bilgi ise bunun uygun problemlerin çözümünde kullanıldığı aşamadır.

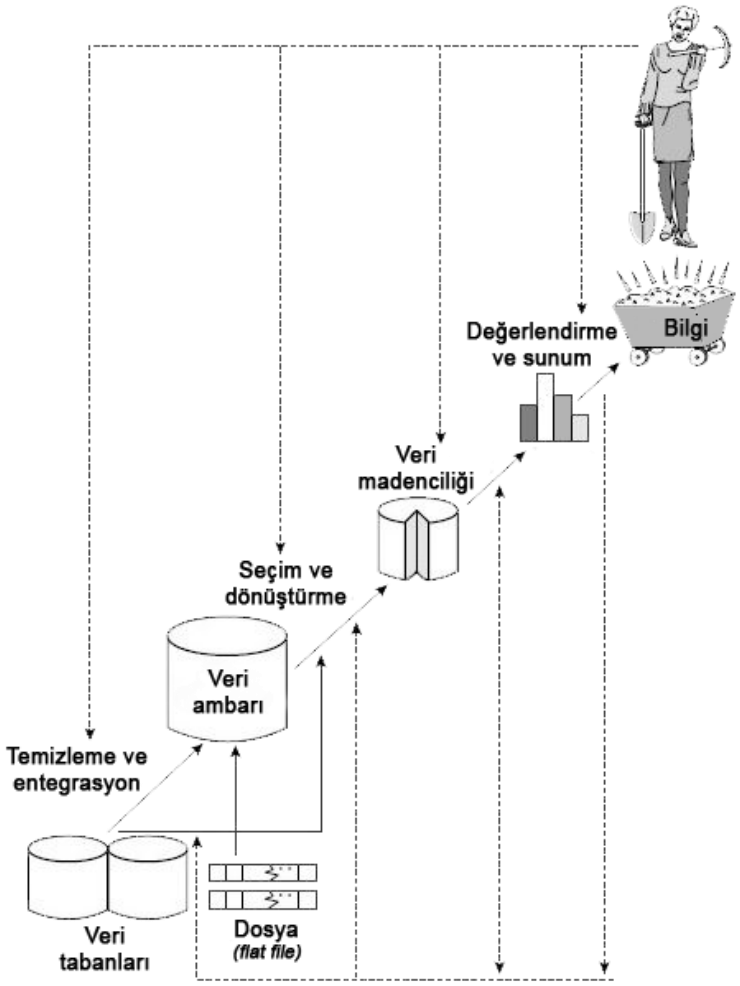
Maimon ve Rokach (2010) ise yine öncekilere paralel olarak süreci şu şekilde yapılandırmıştır (Şekil 2.5):




**Şekil 2.5:** Veri tabanlarında bilgi keşfi süreci (Maimon ve Rokach, 2010).

1- Uygulamanın belirlenmesi *(Domain understanding and KDD goals*): Ne yapılacağına karar verildiği, hedeflerin belirlendiği hazırlık aşamasıdır.
2- Seçim ve veri setinin oluşturulması *(Selection and addition)*: Belirlenen hedefe göre bilgi keşfinin yapılacağı veri setindeki mevcut verinin bulunması, gerekiyorsa yeni verinin eklenmesi, hangi niteliklerin değerlendirileceği gibi işlemlerin yapıldığı aşamadır.
3- Ön işleme ve temizleme *(Preprocessing and cleansing)*: Verinin temizlendiği yani eksik verinin ele alındığı, kirli ve uç noktalardaki verinin çıkarıldığı aşamadır.
4- Dönüştürme *(Transformation)*: Veri madenciliği için verinin hazırlandığı aşamadır. Nitelik seçimi, çıkarılması, ayrıklaştırma gibi işlemler bu aşamada yapılır.
5- Veri madenciliği *(Data mining):*
   5.1 Uygun veri madenciliği yönteminin seçimi *(Choosing the appropriate data mining task)*: Sınıflandırma kümeleme gibi veri madenciliği yöntemlerinden hangisinin kullanılacağına karar verilen aşamadır.
   5.2 Veri madenciliği algoritmasının seçilmesi *(Choosing data mining algorithm)*: Taktiklerin belirlendiği aşama olarak tanımlanmıştır. Hangi koşullar altında hangi veri madenciliği algoritmasının en uygun seçim olacağı belirlenmeye çalışılmaktadır.
   5.3 Veri madenciliği algoritmasının uygulanması *(Employing the data mining algorithm)*: Veri madenciliği algoritmasının veri setine uygulandığı aşamadır. Bu aşamada farklı parametreler kullanılarak algoritmalar birden fazla kez denenebilmektedir.
6- Değerlendirme ve Yorumlama *(Evaluation and interpratation)*: Elde edilen örüntülerin değerlendirildiği aşamadır. Modelin anlaşılırlığı ve kullanışlılığı üzerine odaklanılır.
7- Keşfedilen bilginin kullanılması *(Discovered knowledge)*: Elde edilen bilginin artık bir başka sisteme dâhil edilmek üzere hazır halde bulunduğu aşamadır.

Veri madenciliği alanındaki önemli kaynaklardan birinin yazarları olan Han ve diğ. (2012) tarafından bilgi keşfinin adımları ise aşağıdaki gibi yedi adımda yorumlamaktadır (Han ve diğ., 2012; Özkan, 2013; Kartal, 2015) (Şekil 2.6):




**Şekil 2.6:** Veri tabanlarında bilgi keşfinin bir adımı olarak veri madenciliği (Han ve diğ., 2012).

1- Veri temizleme *(Data cleaning):* Tutarsız ve kirli verinin çıkarıldığı aşamadır.
2- Veri bütünleştirme *(Data integration):* Farklı veri kaynaklarının bütünleştirilmesi, birbirine bağlandığı aşamadır.
3- Veri seçimi *(Data selection):* Veri tabanından analizle alakalı olan verinin seçildiği süreçtir.
4- Veri dönüştürme *(Data transformation):* Analize uygun hale gelmesi için verinin standartlaştırılması ve normalleştirilmesi yani uygun hale getirildiği aşamadır.
5- Veri madenciliği *(Data mining):* Örüntüleri keşfetmek için veriye akıllı yöntemlerin uygulandığı aşamadır.

6- Örüntü değerlendirme *(Pattern evaluation):* Bulunan örüntüleri çekiciliklerine göre tanımlayabilmeyi ifade eden aşamadır.
7- Bilgi gösterimi *(Knowledge presentation):* Kullanıcılara elde edilen bilgiyi sunmak amacıyla görselleştirme ve bilgi gösterim tekniklerinin uygulandığı aşamadır.

Farklı araştırmacılar tarafından veri madenciliği süreçleri birbirine paralel şekillerde ifade edilmiştir. Ele alınan adımlar birbirine benzer şekilde veriden bilgiye doğru keşfi temsil etmektedir.

Roiger ve Geatz (2003) bilgi keşfi sürecinin adımları ile bilimsel yöntemin adımlarını birlikte ele alarak KDD sürecinin adım adım veriden bilgiye doğru nasıl geliştiğini ortaya koymaktadır (Tablo 2.3). KDD'nin, bilimsel yöntemin, veri madenciliği üzerine uygulanması olarak tanımlandığını aktarmaktadır.

**Tablo 2.3:** Veri madenciliğine bilimsel yöntemin uygulanması (Roiger ve Geatz, 2003).

| **Bilimsel Yöntem** | **Veri Tabanlarında Bilgi Keşfi Süreci** |
|---|---|
| Problemin tanımlanması | Hedefin tanımlanması |
| Hipotezin kurulması | |
| | Hedef veri setinin oluşturulması |
| | Veri ön işleme |
| Deney yapma | Veri dönüştürme |
| | Veri madenciliği |
| Sonuç çıkarma | Yorumlama/Değerlendirme |
| Sonuçların doğrulanması | Harekete geçme |

Görüldüğü gibi farklı araştırmacılar tarafından ortaya konan veri madenciliği süreç tanımları mevcuttur. Elbette hangi sürecin izleneceği veri madenciliği araştırması yapacak olan araştırmacının tercihine/ihtiyacına bağlı olarak değişebilmektedir. Adımların sayısı veya isimleri değişmekte olsa da temelde, tanımlanan süreçlerin hepsi birbirini destekler biçimde benzer adımlardan oluşmaktadır.

### 2.2.2. Veri Madenciliğinde Ön İşleme Süreci

Veri madenciliği süreçleri araştırmacılar tarafından nasıl tanımlanırsa tanımlansın her süreçte en kritik olan aşamalardan biri ön işlemedir. Ön işleme süreci, eldeki verinin analizlere uygun bir şekilde temizlenmesini ve hazırlanmasını ifade eder. Eldeki verinin durumuna göre birden fazla işlem veri setine uygulanabilir.

Veri setleri; farklı sebeplerden dolayı kirli/parazit *(noisy)*, eksik *(missing)* ve tutarsız *(inconsistent)* olabilmektedir (Han ve diğ., 2012). Veri setlerindeki eksik, tekrar eden, aykırı/uç değer olan veri modelin doğruluğunu olumsuz şekilde etkileyebilmektedir (Zaïane, 1999). Çoğu zaman eldeki veri setleri doğrudan algoritmaların uygulanmasına elverişli olmamakta veya algoritmalar eksik veri ile çalışamamaktadır. Ön işleme aşaması, kurulan modellerin hem düzgün çalışmasına hem de doğruluğa etki etmesi sebepleriyle önemlidir. Han ve diğ. (2012), düşük kaliteli verinin düşük kaliteli veri madenciliği sonuçlarına sebep olacağını belirtmiştir. Bu tespit ön işleme aşamasının kritikliğine dikkat çekmektedir. Araştırmacılar tarafından bilinen bir kavram olan *"çöp içeri, çöp dışarı - garbage in, garbage out"* kavramı da basit bir şekilde ön işlemenin kritikliğini belirtmek amacıyla kullanılmaktadır. Veri ön işleme aşamasında aşağıdaki başlıklar gerçekleştirilmektedir.

#### *2.2.2.1. Veri Temizleme (Data Cleaning)*

Veri temizleme; eksik verinin tamamlanması, kirli verinin temizlenmesi, uç değerlerin tespiti ve tutarsızlıkların giderilmesini içerir (Han ve diğ., 2012). Akpınar (2014), veri temizlemenin önemine "*verinin temiz olması veri analizine güvenin en temel anahtarlarından biridir"* sözüyle kitabında belirtmektedir.

Gorunescu (2011), veri kalitesinin temel özelliklerinin doğruluk *(accuracy)*, güvenilirlik *(reliability)*, geçerlilik *(validity)* / hareket *(movement)*, eksiksizlik *(completeness)* ve uygunluk/ilgililik *(relevance)* olduğunu belirtmiştir. Buna bağlı olarak da veri setlerinde karşılaşılan sorunları kirli veri, aykırı değerler, eksik veri ve tekrarlayan veri olarak sıralamıştır.

- **Eksik veri**

Eksik veri birçok farklı sebepten dolayı veri setinde bulunabilir. Bunlar araştırmanın tasarımında yapılan hatalardan, cevaplayıcının kişisel özelliklerinden, bilgi birikimi

eksikliğinden veya cevapları abartma, özensiz davranma gibi yaklaşımlarından, ölçüm araçlarının özelliklerinden, veri toplama ortamındaki bozukluklardan, veri yönetimindeki eksikliklerden kaynaklanabilir (Akpınar, 2014).

Eksik verinin tamamlanması için birden fazla yol tercih edilebilir (Han ve diğ., 2012; Özkan, 2013):

- ✓ Eksik verinin bulunduğu kayıt veri setinden çıkartılabilir.
- ✓ Eksik verinin olduğu alanlar manuel olarak doldurulabilir.
- ✓ Kayıp değerlerin yerine genel sabit *(global constant)* kullanılabilir. Bütün kayıp değerler "bilinmeyen" ya da "NA" gibi bir sabitle doldurulabilir.
- ✓ Kayıp değerler, nitelik alanındaki tüm değerlere ait ortanca değer *(median)* veya ortalama değeri *(mean)* ile doldurulabilir.
- ✓ Kayıp değerler, ilgili nitelik alanındaki kayıtların ait oldukları sınıfların ortalamalarına veya ortanca değerlerine göre doldurulabilir.
- ✓ Kayıp değer yerine gelebilecek en muhtemel değer, regresyon, bayes tabanlı araçlar veya karar ağacı kullanılarak doldurulabilir.
- ✓ Nitelik değerlerinin kategorik olduğu durumlarda kayıp veri nitelik alanındaki en sık tekrar eden *(mode)* ile tamamlanabilir (Grzymala-Busse ve Grzymala-Busse, 2010; Kartal, 2015).

- **Tutarsızlıkların saptanması**

Veri girişindeki insan hataları, veri toplamada kullanlan cihazlardan kaynaklı hatalar, kasıtlı olarak doğru cevaplanmayan sorular, güncel olmayan adres verisi gibi sebeplerden dolayı veri setlerinde tutarsızlıklar meydana gelebilmektedir (Akpınar, 2014). Tekrarlayan verinin varlığı da bu başlık altında incelenmektedir. Tekrarlayan kayıtlar veri setinden çıkarılmalıdır (Gorunescu, 2011).

- **Uç/aykırı değerlerin saptanması**

Veri setlerinde, genelin yapısına aykırı davranış gösteren veri bulunabilmektedir (Han ve diğ., 2012). Bu veri bazen bir girdi hatasından kaynaklanırken (Akpınar, 2014) bazen ise dolandırıcılık tespiti gibi durumlarda anormal bir durumu işaret edebilmektedir (Han ve diğ., 2012). Aykırı değerlerin tespiti, sonuçları etkileyebileceğinden önem teşkil etmektedir (Gorunescu, 2011).

#### *2.2.2.2. Veri Entegrasyonu (Data Integration)*

Veri entegrasyonu ya da veri bütünleştirme farklı kaynaklardan elde edilen verinin bir araya getirilmesini ifade eder (Özkan, 2013).

#### *2.2.2.3. Veri Dönüştürme (Data Transformation)*

Literatürdeki bazı kaynaklarda, veri dönüştürme aşaması da ön işleme sürecinin alt başlığı olarak verilmiştir (Larose, 2005; Maimon ve Rokach, 2010; Han ve diğ., 2012; Özkan, 2013; Akpınar, 2014). Bu tez çalışması kapsamında izlenen Fayyad ve diğ., (1996a)'ne ait veri madenciliği sürecinde ön işleme süreci ve veri dönüştürme süreci ayrı birer aşama olarak ele alınmış; ancak dönüştürme başlığı ön işleme başlığı altında verilmiştir. Veri setlerinde dönüştürme adına ayrıklaştırma ve normalizasyon işlemleri yapılmaktadır.

Veri setindeki nümerik verinin kategorik veriye dönüştürülmesi işlemine ayrıklaştırma adı verilmektedir (Olson ve Delen, 2008). Bu dönüştürme sonucunda 1-5, 6-10, 11-15 gibi aralık etiketleri *(interval label)* veya düşük, normal, yüksek gibi kavramsal etiketler *(conceptual label)* oluşturulmaktadır (Han ve diğ., 2012). Veri setlerine analizin amacına uygun olarak farklı ayrıklaştırma yöntemleri uygulanabilmektedir (Koçoğlu, 2012).

Veri setindeki değişkenlerin aldığı değer aralıklarının birbirinden farklı olduğu durumlarda normalizasyon yöntemlerinden yararlanılır. Örneğin bir değişkendeki değerler 100-50000 arasında başka bir değişkeninki ise 0,25-0,90 arasında değişiyor olabilir. Bu tip durumlarda daha geniş aralıklı olan değişken sonuçları etkilemektedir ve bu sebeple veri madenciliğinde her bir değişkenin sonuçlar üzerindeki etkisini standartlaştırmak için nümerik veri normalizasyon yöntemleri ile normalize edilir (Larose, 2005).

Normalizasyon için aşağıdaki yöntemlerden yararlanılabilir (Larose, 2005):

- **Min-Max Normalizasyonu**

Bu yöntemde veri setindeki nümerik değerler 0-1 aralığına çekilir. $X$ gözlem değerleri olmak üzere, ilgili nitelik alanındaki en küçük ve en büyük değerler hesaplamada kullanılır. Bunun için (2.1) formülünden yararlanılır.

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (2.1)$$

- **Z-Score Normalizasyonu**

$X$ gözlem değerleri olmak üzere, ilgili nitelik alanındaki ortalama ve standart sapma değerleri hesaplamada kullanılır. Bunun için (2.2) formülünden yararlanılır.

$$X^* = \frac{X - \text{ortalama}(X)}{Standart\ Sapma\ (X)} \quad (2.2)$$

#### *2.2.2.4. Veri İndirgeme (Data Reduction)*

Veri madenciliği analizlerinde kullanılan veri setleri genellikle hem kayıt sayısı hem de nitelik sayısı açısından büyük veri setleri olmaktadır. Bu durum yani veri setlerinin çok boyutluluğu hem enine *(nitelik sayısı)* hem de boyuna *(kayıt/satır sayısı)* olmak üzere çok boyutlu büyüklüğünü ifade etmektedir (Akpınar, 2014). Her geçen gün boyut sayısı artmakta ve bu durum veri madenciliği analizlerinde birtakım zorluklara sebep olabilmektedir. Büyük veri setleri ile çalışmak veri madenciliği analizlerinin hesaplama maliyetinin yüksek olmasına sebep olmaktadır (Chizi ve Maimon, 2010).

- **Nitelik sayısının azaltılması**

Nitelik sayısının azaltılması, boyut sayısının azaltılması *(dimensionality reduction)* olarak tanımlanmaktadır; tekrarlayan veya ilgisiz olan nitelik alanlarının veri setinden çıkarılmasını ifade eder (Han ve diğ., 2012; Akpınar, 2014). Ayrıca nitelik alanları azaltılırken, nitelik seçimi *(feature selection, attribute selection, variable selection)* de yapılabilir ki bu durum modelin kurulması için en etkili rolü oynayacak niteliklerin seçimini ifade eder (Akpınar, 2014).

Bazı durumlarda ise tam tersine veri setindeki nitelik alanlarından yararlanılarak/ nitelik alanları birleştirilerek, yeni nitelik alanları oluşturulabilmektedir; bu durum *attribute construction* ya da *feature construction* olarak literatürde belirtilmektedir (Han ve diğ., 2012).

- **Örnek/Gözlem sayısının azaltılması**

Örnek/gözlem sayısının azaltılması, veri setini temsil edecek şekilde örnek sayısının azaltılması *(numerosity reduction)* olarak tanımlanmaktadır. Örnek sayısının azaltılması için veri setinden farklı yöntemlere göre bir örneklem seçilerek, kümeleme yapılarak veya başka yöntemlerle (parametik veya parametrik olmayan) veri setini temsil eden daha az sayıda örnek ile çalışılabilir (Han ve diğ., 2012).

Parametrik olmayan yöntemlerden örnekleme, büyük bir veri setinin daha küçük bir veri seti ile temsil edecek için daha küçük bir veri seti oluşturulması olarak tanımlamaktadır (Han ve diğ., 2012). Örnekleme yapmak için tekrarlı örnekleme *(sampling with replacement)*, tekrarsız örnekleme (*sampling without replacement)*, tabakalı örnekleme *(stratified sampling)*, küme örneklemesi *(cluster sampling)* yöntemlerinden yararlanılabilmektedir (Witten ve diğ., 2011; Han ve diğ., 2012).

### 2.2.3. Veri Madenciliği Araçları

Veri madenciliği analizleri farklı araçları kullanılarak yapılabilmektedir (Olson ve Delen, 2008; Mikut ve Reischl, 2011). Araçların bazıları ticari yazılımlarken bazıları ücretsiz ve açık kaynak kodlu yazılımlardır. Tablo 2.4'te bu yazılımlardan bazıları verilmiştir. Bu tez kapsamında ücretsiz edinilebilen ve açık kaynak kodlu olan RStudio (RStudio, 2017) tercih edilmiştir. Özkan ve Özkan (2017), istatistik ve veri madenciliği ile ilgili çalışan araştırmacıların yararlanabileceği birçok paket kütüphanesinin sunulması sebebiyle R dilinin avantaj sağladığını ve araştırmacıların ilgisini çektiğini belirtmiştir.

**Tablo 2.4:** Veri madenciliği araçları/yazılımları (Olson ve Delen, 2008; Mikut ve Reischl, 2011).

| Ücretsiz ve Açık Kaynak Kodlu Olan Yazılımlara Örnekler | Ticari Yazılımlara Örnekler |
|---|---|
| | SPSS Clementine |
| KNIME | SAS Enterprise Miner |
| Orange | IBM Intelligent Miner |
| **RStudio** | Oracle Data Mining |
| Rapid Miner | Megaputer PolyAnalyst |
| Tanagra | SAP Netweaver Business Warehouse |
| WEKA | SQL Server Analysis Services |
| | Statistica |

Hangi yazılımın kullanılacağı elbette araştırmacının tercihine ve yazılımların niteliklerine göre değişmektedir. Farklı araştırmacılar tarafından farklı yazılımlarla ilgili karşılaştırma çalışmaları yapılmıştır (Tablo 2.5).

**Tablo 2.5:** Veri madenciliği yazılımlarının karşılaştırıldığı yayınlar.

| Çalışmanın Başlığı | Kaynak |
|---|---|
| A Comparison Study Between Data Mining Tools Over Some Classification Methods | (Wahbeh ve diğ., 2011) |
| Comparison of Data Mining Techniques and Tools for Data Classification | (Borges ve diğ., 2013) |
| Comparison of Various Tools for Data Mining | (Rana ve Kaur, 2014) |
| Comparative Study of Data Mining Tools | (Rangra ve Bansal, 2014) |
| An Analytical Review of Data Mining Tools | (Peace, 2015) |
| A Comparison of Open Source Tools for Data Science | (Wimmer ve Powell, 2016) |

### 2.2.4. Veri Madenciliği Uygulama Alanları

Veri madenciliğinin uygulama alanlarına bakıldığında birçok farklı araştırma alanı ve sektörde kullanılabildiği görülmektedir. Gorunescu (2011) gerçek hayattaki problemlerin çözümünde veri madenciliğinin önemli bir yeri olduğuna dikkat çekmiştir. Ona göre ekonomi, sağlık, astronomi, meteoroloji, biyoloji, dil bilimi gibi büyük miktarlarda verinin olduğu alanlarda veri madenciliği analizlerine ihtiyaç duyulmaktadır. Özellikle veri miktarının fazla olduğu sağlık alanında veri madenciliği analizlerinin uygulanabileceğini gösteren yeni çalışmalar (Selçukcan Erol, 2016; Özkan ve Selçukcan Erol, 2017) yapılmaktadır.

Türkiye'de YÖK veri tabanındaki tez bazında veri madenciliği çalışmalarına bakıldığında farklı alanlardaki araştırmacılar tarafından veri madenciliği yöntemlerinin kullanıldığı görülmektedir. Tarama bu tez çalışmasının konusu olan sınıflandırma yöntemleri üzerinden yapılmıştır. Özellikle istatistikte kullanılan bazı yöntemlerin veri madenciliğinde de kullanılabildiği görülmektedir. Dolayısıyla alanların iç içe geçtiği gözlemlenmektedir. Çalışmalara bakıldığında mühendislik, işletme, sağlık bilimleri, sosyal bilimler gibi farklı alanlarda veri madenciliği ile ilgili çalışmalar yapıldığı görülmektedir (Tablo 2.6).

**Tablo 2.6:** YÖK veri tabanındaki çalışmalar.

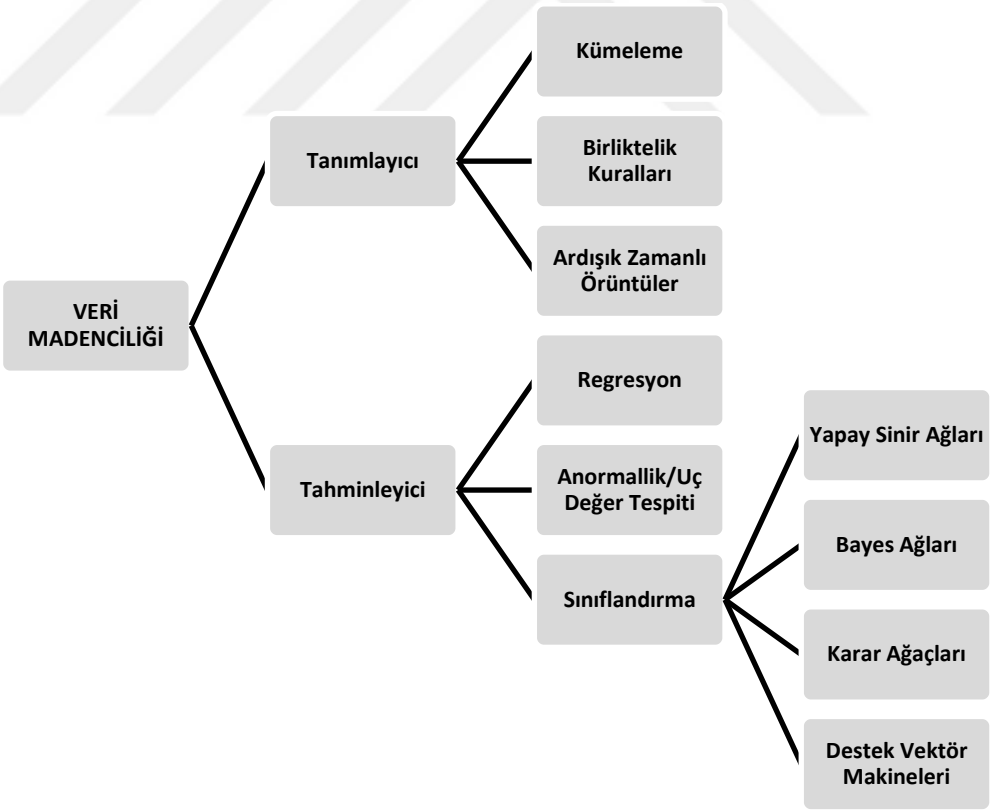
| Tez Adı | Tez Türü | Çalışmanın Yapıldığı Kurum | Sınıflandırma Yöntemi | Kaynak |
|---|---|---|---|---|
| k-En Yakın Komşuluk, Yapay Sinir Ağları ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması | Doktora | Bülent Ecevit Üniversitesi, Biyoistatistik Anabilim Dalı | KNN, yapay sinir ağları, karar ağaçları | (Köktürk, 2012) |
| Trafik Kazalarının Sınıflandırılmasında Karar Ağacı Kullanımı: Bodrum İlçesi Örneği | Yüksek Lisans | Gazi Üniversitesi, Endüstri Mühendisliği Anabilim Dalı | CART, CHAID, QUEST | (Parıldar, 2014) |
| Tanker Şamandıra Bağlama Sistemlerinin Yapay Sinir Ağları Tekniğiyle Optimizasyonu | Yüksek Lisans | İstanbul Teknik Üniversitesi, Gemi ve Deniz Teknoloji Mühendisliği Anabilim Dalı | yapay sinir ağları | (Yetkin, 2014) |
| Özellik Çıkarma Ve DVM Tabanlı Adaboost Algoritması ile Biyomedikal Veri Sınıflandırma | Yüksek Lisans | Selçuk Üniversitesi, Elektrik-Elektronik Mühendisliği | adaboost | (Mücahit, 2014) |
| Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama | Yüksek Lisans | Eskişehir Osmangazi Üniversitesi, Biyoistatistik Anabilim Dalı | rastgele orman, KNN | (Yılmaz, 2014) |
| Enformasyon Sistemlerinde Saklı Markov Modeli ve Bayes Tabanlı Sınıflandırıcılar ile Bilgi Modellerinin Geliştirilmesi | Doktora | Fırat Üniversitesi, İstatistik Anabilim Dalı | saklı Markov modeli, Bayes tabanlı sınıflandırıcılar | (Doğaner, 2015) |
| Veri Madenciliği Yöntemleri ile Ankilozan Spondilit Hastalığında Radyografik Progresyona Etkili Faktörlerin Analizi | Yüksek Lisans | İstanbul Üniversitesi, Enformatik Anabilim Dalı | C4.5, Gini, CART,rastgele orman | (Atasoy, 2015) |
| Kimlik Doğrulaması için Tuş Vuruş Dinamiklerine Dayalı Bir Güvenlik Sisteminin Yapay Sinir Ağları ile Geliştirilmesi | Doktora | İstanbul Üniversitesi, Enformatik Anabilim Dalı | yapay sinir ağları | (Özen, 2016) |
| Destek Vektör Makineleri Yardımıyla Tüketici Kredilerinin Sınıflandırılması | Yüksek Lisans | İstanbul Teknik Üniversitesi, İşletme Mühendisliği Anabilim Dalı | destek vektör makineleri | (Kaya, 2016) |
| Lojistik Regresyon Modeli İle Elde Edilen Tahminlerin Roc Eğrisi Yardımıyla Değerlendirilmesi: Türkiye'de Hanehalkı Yoksulluğu Üzerine Bir Araştırma | Yüksek Lisans | Süleyman Demirel Üniversitesi, Ekonometri Anabilim Dalı | lojistik regresyon | (Karcı, 2017) |

### 2.2.5. Veri Madenciliği Yöntemleri

Özkan (2013) ve Akpınar (2014) veri madenciliği yöntemlerini genel olarak üç ana başlık altında toplamaktadır:

1- Sınıflandırma *(classification),*
2- Kümeleme *(clustering),*
3- Birliktelik kuralları *(association rules)* ve ardışık zamanlı örüntüler *(sequential patterns).*

Maimon ve Rokach (2010), Gorunescu (2011) ve Han ve diğ. (2012) bunun dışında daha genel bir kapsamda tanımlayıcı (*descriptive*) ve tahminleyici/kestirimci *(predictive)* yöntemler olmak üzere iki ana kategori altında yöntemlerin toplanabileceğini aktarmıştır (Şekil 2.7). Tanımlayıcı yöntemler, bir veri setindeki özelliklerin ortaya konmasını ve anlaşılmasını sağlayan yöntemlerdir. Tahminleyici/ yöntemler ise eldeki veri ile tahmin yapılması için kullanılan yöntemlerdir.




**Şekil 2.7:** Tanımlayıcı ve tahminleyici veri madenciliği yöntemleri (Maimon ve Rokach, 2010; Gorunescu, 2011; Han ve diğ., 2012).

Larose (2005) altı farklı veri madenciliği yöntemi başlığını aktarmıştır:

- Tanımlayıcı *(descriptive)* - Veri içerisindeki örüntülerin veya trendlerin ortaya çıkarılmasıdır.
- Tahmin *(estimation)* - Sınıflandırma yöntemlerine benzer ancak sınıf değerleri kategorik değil nümerik değerlerden oluşur. Modelin değişkenler ve hedef sınıf ile öğrenmesi ve yeni bir değerin sınıfının bulunmasına dayanır. İstatistiksel yöntemlerden lineer regresyon, korelasyon, çoklu regresyon analizleri ve veri madenciliği yöntemlerinden yapay sinir ağları bu yöntemler arasına girer.
- Öngörü *(prediction)* - Sınıflandırma yöntemlerine benzer ancak gelecekteki bir şeyin değeri bulunmaya çalışılır. İstatistiksel yöntemlerden lineer regresyon, korelasyon, çoklu regresyon analizleri ve veri madenciliği yöntemlerinden yapay sinir ağları, karar ağaçları ve k-en yakın komşu bu yöntemler arasında girer. Sınıflandırma ve tahmin yöntemlerinden ayırması zordur (Roiger ve Geatz, 2003).
- Sınıflandırma *(classification)* - Modelin değişkenler ve hedef sınıf ile öğrenmesi ve yeni bir değerin sınıfının bulunmasına dayanır. Sınıf değerleri kategoriktir.
- Kümeleme *(clustering)* - Birbiriyle benzer özellikteki gözlemlerin/örneklerin gruplandırılmasıdır. Sınıflandırma yöntemlerinden hedef niteliğin yani sınıf değerlerinin olmaması ile ayrılır.
- Birliktelik *(association)* - İki veya ikiden fazla nitelik arasındaki ilişkiyi belirlemeye çalışan yöntemlerdir. En bilineni pazar sepet analizidir.

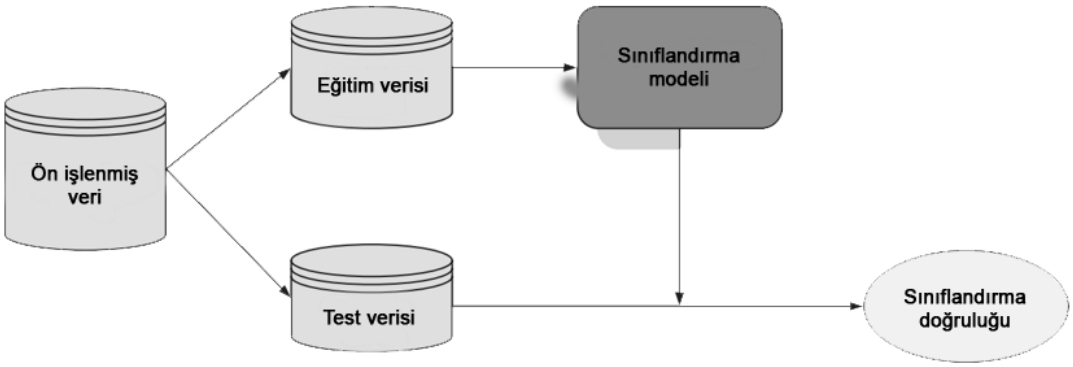
Veri madenciliği yöntemlerinden bahsederken, denetimli/danışmanlı öğrenme *(supervised learning)* ve denetimsiz/danışmansız öğrenme *(unsupervised learning)* kavramlarından da bahsetmek gerekmektedir. Makine öğrenmesinde *(machine learning)* kullanılan bu iki kavram veri madenciliği yöntemlerindeki yaklaşımları ifade etmek için de kullanılmaktadır. Danışmanlı öğrenme, sınıf değerleri/etiketleri belli olan veri setinin eğitilerek tahmin etmede kullanılmasıdır (Gorunescu, 2011; Flach, 2012; Harrington, 2012). Yani danışmanlı öğrenmede her bir kaydın hangi sınıf değerine ait olduğu önceden bellidir. Kurulan model buna göre öğrenerek yeni bir kayıt veri setine eklendiğinde veya veri setinin bir kısmı test etme için kullanıldığında, örneklerin hangi sınıflara ait olduğunu bulmaya çalışır. Sınıflandırma yöntemleri, danışmanlı öğrenme yaklaşımını temel alır. Danışmansız öğrenmede ise, danışmanlının tam tersine, çıktılar/etiketler/sınıf değerleri

belli değildir ve veri setindeki değerlere göre birbirine benzer kayıtlar aynı kümede birleştirilmeye çalışılır (Gorunescu, 2011; Flach, 2012; Harrington, 2012). Kümeleme yöntemleri ve birliktelik analizi danışmansız öğrenme yaklaşımını temel alır. Özetle danışmanlı öğrenmede sınıf değerleri önceden belli iken, danışmansız öğrenmede sınıf değerleri belli değildir.

Tez çalışması kapsamında sınıflandırma yöntemleri kullanıldığı için sınıflandırma yöntemleri detaylı olarak ele alınmıştır. Sınıflandırma, çeşitli sınıflandırma yöntemlerinin kullanılarak verinin hangi sınıfa ait olduğunun bulunmasını sağlar (Akpınar, 2014). Bu sınıflar önceden belirlenmiş sınıflardır. Bu sebeple sınıflandırma yöntemleri, danışmanlı öğrenme başlığı altında değerlendirilir. Hangi algoritma/yöntem kullanılırsa kullanılsın bu yöntemlerin temel mantığı aynıdır. Verinin bir kısmı modelin/algoritmanın eğitiminde, geri kalan kısmı ise test edilmesinde kullanılır. Girdi değerlerinin, hangi çıktı değerine karşılık geldiği başka bir deyişle hangi sınıfa ait olduğu bellidir (Kartal, 2015). Model eğitilerek sınıflandırma kuralları ortaya çıkarılır ve verinin hangi sınıfa ait olacağı bu kurallar çerçevesinde tahmin edilir (Özkan, 2013). Dolayısıyla yeni bir kayıt bir veri setine eklendiğinde veya hangi sınıfa ait olduğu bilinmeyen bir kayıt varsa modelin oluşturduğu kurallara göre sınıfın tahmin edilmesi sağlanır. Sınıflandırma yöntemlerinin çalışma şekli Şekil 2.8'de görülmektedir. Sınıflandırma yöntemleri *"resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama"* gibi farklı alanlarda farklı amaçlarla kullanılmaktadır (Silahtaroğlu, 2008). Sınıflandırma yöntemlerindeki temel kavramlar şu şekilde özetlenebilir (Gorunescu, 2011):

- Sınıf *(class)* - Sınıflandırmada kullanılan kategorik tipteki bağımlı değişkenler, sınıflandırmadan sonra kayıtların aldığı etiket/sınıf, sınıflandırmadaki çıktı, hedef sınıf, hedef nitelik
- Nitelik *(predictor/attributes)* - Her bir sütundaki veri seti elemanı, sınıflandırmada kullanılan bağımsız değişkenler
- Örnek/kayıt *(records/tuples/vectors/instances/objects/samples)* - Her bir satırdaki veri seti elemanı
- Model/sınıflandırıcı *(model/classifier)* - Matematiksel olarak bakıldığında değeri ilgili sınıf, değişkenleri de nitelikler olan fonksiyon

- Eğitim veri seti *(training dataset)* - Nitelikleri ve sınıfı barındıran, modelin eğitilmesinde kullanılan veri seti, sınıflandırmadaki girdi
- Test veri seti *(testing dataset)* - Modelin/sınıflandırıcının performansının test edilmesinde kullanılan veri seti




**Şekil 2.8:** Sınıflandırma yöntemlerinin iş akışı (Olson ve Delen, 2008).

Sınıflandırma yöntemleri birçok farklı alana uyarlanıp kullanılabilir (Bramer, 2007):

- Bir hastalıkla ilgili olarak yüksek, orta, düşük riskli hastaların belirlenmesinde
- Öğrencilerin başarılı veya başarısız olma durumlarını tespit edilmesinde
- Havanın bir sonraki gün yağmurlu olup olmayacağının tahmin edilmesinde
- Kişilerin oy verecekleri farklı partilerin tahmin edilmesinde
- Kişilerin suç işlemeye meyilli olup olmadığının tahmin edilmesinde
- Belli bir süre içerisinde evlerin değerinin artıp artmayacağının belirlenmesinde
- Belirli bir ürünü alabilecek ya da almayacak müşterilerin belirlenmesinde

Sıklıkla karşılaşılan sınıflandırma yöntemlerinden bazıları şunlardır (Gorunescu, 2011):

- k-en yakın komşu *(k-nearest neighbour)*
- Karar ağaçları *(decision trees)*
- Yapay sinir ağları *(artificial neural networks)*
- Sade Bayes sınıflandırıcı *(naive Bayes classifier)*
- Lojistik regresyon *(logistic regression)*
- Genetik algoritmalar *(genetic algorithms)*
- Destek vektör makineleri *(support vector machines)*

Bu tez çalışması kapsamında sade Bayes sınıflandırıcı, karar ağaçları (CART C4.5, C5.0, C5.0 boosted) ve rastgele orman algoritmaları kullanılmıştır. Algoritmalara ait detaylı bilgi sırasıyla 2.2.5.1, 2.2.5.2, 2.2.5.3 bölümlerinde yer almaktadır.

#### *2.2.5.1. Sade Bayes Sınıflandırıcı (Naive Bayes Classifier)*

Bu sınıflandırma yöntemi istatistikteki Bayes teoremini temel alır (Bayes ve Price, 1763). Hangi sınıfa ait olduğu bilinmeyen bir $X$ verisinin, herhangi bir sınıfa ait olma olasılığının hesaplanarak sınıf değerinin bulunmasıdır. Bu yöntemde sonrasal olasılık *(posterior probability)* yaklaşımı izlenir. Farklı nitelik değerleri (sütun) olan $X$ verisinin $C$ sınıfına ait olma olasılığının ($P(X|H)$ hesaplanması söz konusudur.

Bayes bağıntısına bakılacak olursa (2.3):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2.3}$$

$P(H|X)$: $H$'nin $X$ üzerinde koşullandırılması ile oluşan sonrasal olasılıktır. $X$'in $C$ sınıfına ait olduğu iddia edilmesi $H$ hipotezi olarak tanımlanır. Sınıfı bilinmeyen bir $X$ verisinin hangi sınıfa ait olduğu bu bağıntıya göre hesaplanır.

$P(X|H)$: $X$'in $H$ üzerinde koşullandırılması ile oluşan sonrasal olasılıktır. $X$'in bilinen niteliklerine göre sınıflardan birine ait olma olasılığının hesaplanmasıdır. Her bir nitelik değeri için ayrı ayrı koşullu olasılıkların hesaplanması gerekmektedir.

Örneğin iki sınıf değeri ($C_1$: pozitif, $C_2$: negatif) ve $X$'e ait üç nitelik ($X_1$, $X_2$, $X_3$) olduğu varsayılırsa her bir niteliğe göre $X$'in "pozitif" ve "negatif" sınıflarına ait olma olasılıklarının hesaplanması gerekmektedir. Bu durumda aşağıdaki olasılıkların hepsi hesaplanır ((2.4), (2.5)).

$$P(X_1|sınıf = pozitif), P(X_2|sınıf = pozitif), P(X_3|sınıf = pozitif) \tag{2.4}$$

$$P(X_1|sınıf = negatif), P(X_2|sınıf = negatif), P(X_3|sınıf = negatif) \tag{2.5}$$

$P(X|sınıf = pozitif)$ ve $P(X|sınıf = negatif)$ yani $X$ değerinin "pozitif" ve "negatif" sınıflarına ait olma olasılıkları, her bir nitelik için yukarıdaki gibi hesaplanan değerlerin çarpımına eşittir.

$P(H)$: $H$'ye ait önsel olasılıktır. $X$'in herhangi bir sınıfa ait olma olasılığıdır ((2.6), (2.7)). Bu durumda en baştan olasılığın ne olduğu bilinmektedir.

$$P(sınıf = pozitif) \tag{2.6}$$

$$P(sınıf = negatif) \tag{2.7}$$

$P(X)$: $X$'e ait önsel olasılıktır. Tüm sınıflar için sabittir.

Olasılıklar hesaplamak için (2.8) formülünden yararlanılır.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{2.8}$$

(2.8) formülü yerine hesaplamada basitleştirme yapmak adına $X$'e ait nitelik değerlerinin birbirinden bağımsız olduğu *(class-conditional independent)* kabul edilerek (2.9) formülü kullanılmaktadır.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i)P(x_2|Ci) \dots P(x_n|C_i) \tag{2.9}$$

$X$'e ait sınıf değerini hesaplamak için yapılan işlemlerde amaç $P(X|C_i)$ değerinin maksimize edilmesidir. Bu durum en büyük sonrasal sınıflandırma yöntemi *(maximum posteriori hypothesis)* adı verilir. Yani $X$, $P(X|C_i)P(C_i)$ ifadesinin maksimum olduğu sınıfa aittir. Bu (2.10) ile gösterilen formül ile elde edilir.

$$argmax\{P(X|C_i)P(C_i)\} \tag{2.10}$$

Başka bir ifade ile eğer $X$'in $C_i$ sınıfına ait olma olasılığı $C_j$ sınıfına ait olma olasılığında büyükse $X$, $C_i$ sınıfına aittir ((2.11)).

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \qquad 1 \leq j \leq m, j \neq i \tag{2.11}$$

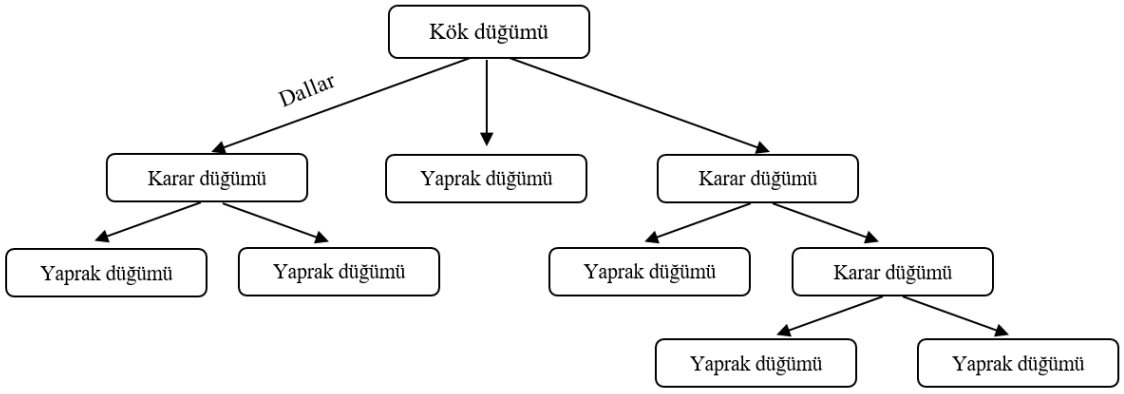
Yukarıda bahsedilen hesaplamalar nitelik değerlerinin kategorik olduğu durumlar için geçerlidir. Nümerik veri ile çalışırken olasılıkların hesaplanması için aşağıdaki bağıntıdan yararlanılmaktadır. Nümerik değerlerin Gauss dağılımı *(Gaussian distribution)* gösterdiği varsayılır, μ nümerik değerin gösterdiği sınıfa ait ortalama değer, $\sigma$ ise nümerik

değerin gösterdiği sınıfa ait standart sapma olarak ifade edilerek (2.12) formülü ile işlem yapılır.

$$P(X_k|C_i) = f\left(x_i, \mu_{C_i}, \sigma_{C_i}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.12)$$

### *2.2.5.2. Karar Ağaçları (Decision Trees)*

Karar ağaçları, ağaca benzer hiyerarşik bir yapı oluşturarak sınıflandırma yapılmasını sağlayan yöntemlerdir. Karar ağaçları altında farklı algoritmalar bulunmaktadır ancak algoritmaların temeldeki mantığı benzerdir. Karar ağaçları, böl ve fethet *(divide and conquer)* mantığı ile çalışırlar (Quinlan, 1996). Kolay anlaşılabilmesi ve gerçek hayattaki problemlere uyarlanabilmesi, hem nümerik hem de kategorik veri ile çalışma imkânı sağlamaları açısından karar ağaçları avantaj sağlamaktadır (Roiger ve Geatz, 2003). Karar ağacı üzerindeki temel elemanlar karar düğümleri *(decision nodes),* dallar *(branches),* yaprak düğümleri *(leaf nodes),* kök düğüm *(root node)* yani en tepedeki düğümdür (Larose, 2005; Balaban ve Kartal, 2015) (Şekil 2.9).




**Şekil 2.9:** Örnek karar ağacı ve elemanlarının görünümü.

Ağaç üzerindeki hiyerarşik bölünme, bölme kriterine *(split criterion)* göre yapılır (Aggarwal, 2015). Bölme kriteri; "*splitting criterion", "splitting attribute", "split point", "splitting subset*" olarak da kullanılmaktadır (Han ve diğ., 2012). Bölme kriteri birtakım ölçülere göre belirlenir. Bunlar nitelik seçim ölçüleri *(attribute selection measures)* ya da bölünme kuralları *(splitting rules)* olarak tanımlanmaktadır ve bunlardan bazıları bilgi kazancı *(information gain),* kazanç oranı *(gain ratio) ve* Gini katsayısıdır *(Gini index)* (Maimon ve Rokach, 2010; Han ve diğ., 2012).

ID3, C4.5 ve C5.0 algoritmaları, entropi temelli bölünme yapılan algoritmalardır (Gorunescu, 2011). Bölme kriterlerinden, karar ağaçlarında kullanılanlar aşağıda açıklanmıştır:

- **Bilgi kazancı *(information gain)***

ID3 algoritması tarafından kullanılır. Bilgi kazancı, Shannon (1948) tarafından ortaya konan entropi ***("Shannon entropy")*** kullanılarak hesaplanır. (2.13) formülü entropiyi verir.

$$Entropi(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{2.13}$$

Bulunan $Entropi(D)$, belirsizliğin ölçütüdür. Entropi, bölünme için optimal değerin bulunmasında kullanılmaktadır (Gorunescu, 2011). Bunun için (2.14) formülünden yararlanılır.

$$Entropi_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Entropi(D_j) \tag{2.14}$$

İki formülden yola çıkarak, (2.15) formülü ile bilgi kazancı elde edilir.

$$Bilgi\ Kazancı(A) = Entropi(D) - Entropi_A(D) \tag{2.15}$$

Bilgi kazancı $A$'dan yapılacak bir dallanmada ne kadar bilgi elde edileceğini göstermektedir. En yüksek bilgi kazancı elde edilmesini sağlayan nitelik dallanma için seçilir.

***Sınıflandırma ve Regresyon Ağaçları***

CART *(classification and regression trees)* (Breiman ve diğ., 1984), sürekli olarak ikili şekilde bölünerek oluşan bir ağaç yapısı oluşturur ve önemli bir karar ağacı algoritmasıdır (Akpınar, 2014). Ağaç her zaman ikili *(binary)* dallanmalardan ve hiyerarşik bir yapıdan oluşur. CART, bölünme için en uygun değişkeni seçer ve her seferinde tek bir değişken kullanarak bölünmeyi gerçekleştirir (Hand ve diğ., 2001). Hem nümerik hem de kategorik

veri ile çalışma imkânı sunan CART*, gini katsayısı, twoing, ordered twoing ve en küçük kareler sapması* yöntemlerini kullanarak bölünme yapılmasını sağlar (Akpınar, 2014).

- **Gini katsayısı *(Gini index)*** (Breiman ve diğ., 1984)

CART tarafından kullanılan Gini katsayısı her bir nitelik için ikili bölünme yapılması üzerine kuruludur ((2.16)).

$$Gini(D) = 1 - \sum_{i=1}^{m} {p_i}^2 \tag{2.16}$$

İkili dallanma yapılacağı için Gini hesaplanırken $D$ niteliğinin $D_1$ ve $D_2$ şeklinde bölünmesi gerekmektedir. Buna göre (2.17) formülü hesaplanır.

$$GiniA(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{2.17}$$

Her nitelik değeri için Gini katsayısı hesaplandıktan sonra en küçük Gini değerine sahip niteliğe göre bölünme gerçekleştirilir.

***C4.5 Algoritması***

C4.5 (Quinlan, 1993), karar ağacı algoritmalarından ID3'ün (Quinlan, 1986) devamı niteliğindeki algoritmadır. Bölme kriteri olarak entropi ile beraber kazanç oranını kullanır. C4.5, sınıflandırma ve regresyon ağaçlarında olduğu gibi ikili dallanmadan daha fazla dallanma gerçekleştirilmesine imkân sağlar (Roiger ve Geatz, 2003).

- **Kazanç oranı *(gain ratio)***

C4.5 ve C5.0 algoritmaları tarafından kullanılmaktadır. Önce (2.18) formülü ile bölme bilgisi *(split information)* sonrasında (2.19) formülü ile kazanç oranı hesaplanır.

$$B\ddot{o}lme\ bilgisi_A(D) = -\sum_{i=1}^{k} \frac{|D_i|}{|D|} \log_2\left(\frac{|D_i|}{|D|}\right) \tag{2.18}$$

$$Kazanç\ oranı(A) = \frac{Bilgi\ kazancı(A)}{B\ddot{o}lme\ bilgisi_A(D)} \tag{2.19}$$

Kazanç oranı her bir nitelik için ayrı ayrı hesaplanır. Buna göre kazanç oranı en yüksek olan nitelikten dallanma yapılmaya başlanır.

***C5.0 Algoritması***

C5.0, karar ağacı algoritmalarından C4.5'in (Quinlan, 1993) devamı niteliğindeki algoritmadır. Bölme kriteri olarak entropi ile beraber kazanç oranını kullanır. C5.0 ticari olarak temin edilebilmektedir (Witten ve diğ., 2011).

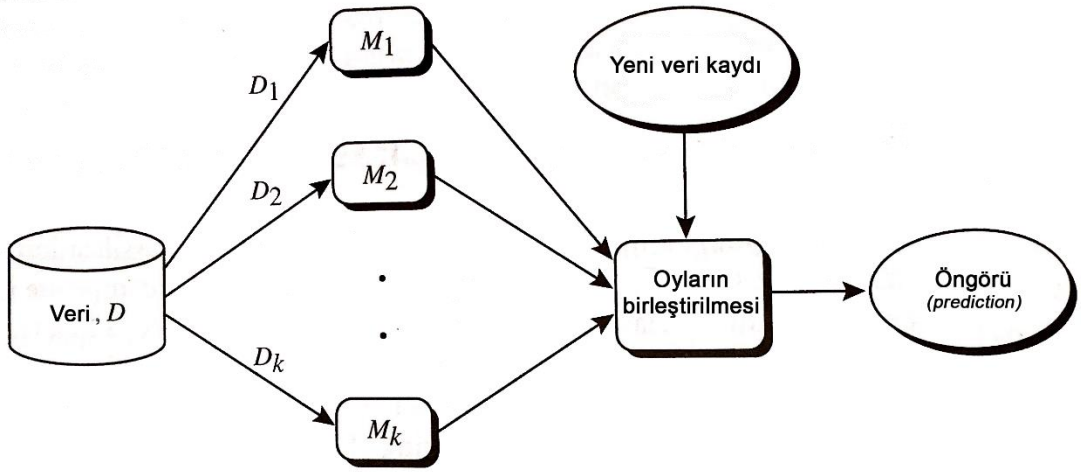
ID3, C4.5, C5.0 algoritmaları farklı araştırmacılar tarafından karşılaştırılmaktadır. Araştırmacıların belirttiklerine göre C5.0 daha etkili çalışması ve boosting özelliği sayesinde daha iyi doğruluk değerleri vermektedir (Hssina ve diğ., 2002; Patil ve diğ., 2012; Pandya ve Pandya, 2015).

C5.0'ın C4.5'ten ayrılmasını sağlayan en önemli özelliklerinden biri boosting özelliğinin olmasıdır. Doğruluğun arttırılması için boosting yönteminden yararlanılmaktadır (Han ve diğ., 2012; Harrington, 2012). Boosting yönteminde eğitim kümesindeki her bir veri kaydı için bir ağırlık atanır ve sırayla birden fazla sınıflandırma modeli oluşturulur. Her bir modelin oluşturulmasından sonra bu ağırlık değerleri güncellenir. Yeni sınıflandırma modeli yanlış sınıflandırılan kayıtları göz önünde bulundurarak oluşturulur. Yani her bir model kendinden önce oluşturulmuş olan modelin performansından etkilenir ve bunun sonucundan birbirinden farklı modeller oluşturulur.

### *2.2.5.3. Rastgele Orman Algoritması (Random Forest Algorithm)*

Rastgele orman algoritması, ensemble öğrenme *(ensemble learning)* yöntemlerinden biridir. Bagging, boosting, random forest gibi popüler ensemble yöntemlerinin temel amacı doğruluğun arttırılmasıdır (Han ve diğ., 2012). *Model averaging and combination, stacking, bucket of models* yöntemleri de diğer ensemble yöntemlerindendir (Aggarwal, 2015).

Aggarwal (2015), bu yöntemlerin kullanımındaki amacın birden fazla modele ait sonucun birleştirilerek "daha dayanıklı" *(more robust)* sonuçlar elde edilmesi olduğunu belirtmiştir. Ensemble yöntemine göre çalışan algoritmalar Şekil 2.10'daki gibi birden fazla sınıflandırma modeli oluşturur ve yeni bir örnek veri setine geldiğinde sınıfının belirlenmesi için oylama yöntemine başvurularak sınıf belirlenir.




**Şekil 2.10:** Sınıflandırıcı doğruluğunun arttırılması (Han ve diğ., 2012).

Şekil 2.10'da görülen çalışma mantığına göre rastgele orman algoritmasında oluşturulan her bir model karar ağacı yapısındadır ve rastgele orman, karar ağacı yapısında oluşturulan modelleri kapsayan yapıdan oluşur. Algoritma birden fazla karar ağacının oluşturulması ve en popüler olan sınıfın oylanması temeline dayanır (Breiman, 2001). Her bir karar ağacı oluşurken CART metodolojisi kullanılır; ağaçlar budanmaz; nitelik seçimi ise rastgele yapılır (Breiman, 2001). Algoritmanın ismi birden fazla karar ağacını bünyesinde barındırmasından, nitelik seçiminin rastgele yapılmasından gelmektedir.

### 2.2.6. Model Değerlendirme ve Seçimi

#### *2.2.6.1. Model Performans Değerlendirme Yöntemleri*

Analiz edilecek veri setinden örneklemenin nasıl yapılacağına dair farklı yöntemler mevcuttur ve bu yöntemler model performans değerlendirme yöntemleri olarak adlandırılırlar (Balaban ve Kartal, 2015). Yani veri setinde sınıflandırma yapmak amacıyla, eğitim ve test için veri setinin ne şekilde bölüneceğine bu yöntemler kullanarak karar verilmektedir.

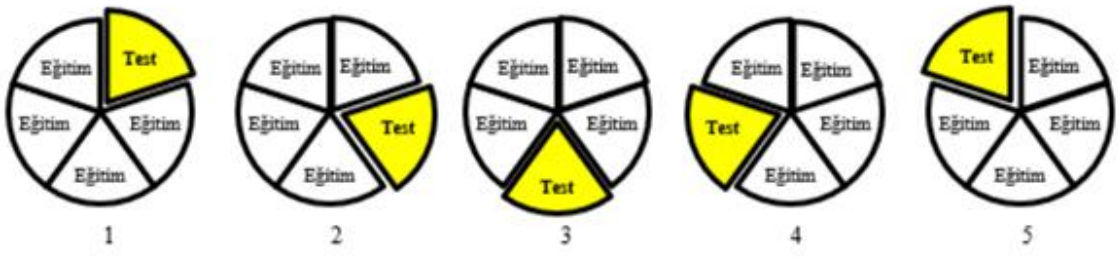
- **Holdout**

Holdout yönteminde veri setinin belirli bir kısmı test geri kalan kısmı ise eğitim verisi olarak bölünür ve genellikle veri setinin üçte biri test için üçte ikisi eğitim için ayrılır (Kohavi, 1995). Ancak bu yöntemin kullanımında birtakım dezavantajlar olabileceğinden bahsedilmiştir (Witten ve diğ., 2011; Balaban ve Kartal, 2015). Test veya eğitim veri setlerinin temsilci olmayabileceği araştırmacılar tarafından belirtilmektedir. Yani verinin

seçimi rastgele gerçekleştiğinden, farklı sınıflara ait örneklerin/verinin test ve eğitim setleri içerisindeki sınıf dağılımları eşit oranda olmayabilir. Bu yüzden eğitim ya da test veri setinde herhangi bir sınıftan çok fazla veri olması veya hiç veri olmaması gibi durumların önüne geçmek gerekmektedir. Bu durumu engellemek için rastgele seçimde, her sınıfın eğitim ve test veri setlerinde eşit oranda dağıldığından emin olmak amacıyla tabakalı örnekleme (*stratified sampling/stratified holdout*) tercih edilebilir. Holdout yönteminin farklı test ve eğitim setleri ile birden fazla kez tekrarlanarak gerçekleştirilmesi yöntemine ise tekrarlı holdout *(repeated holdout)* adı verilmektedir.

- **Çapraz Geçerleme *(Cross Validation)***

Çapraz geçerleme yönteminde öncelikle veri seti seçilen bir sayı *(k)* kadar parçaya bölünür. Bu yöntemin adı bu bölünme sebebiyle k-kat çapraz geçerleme ya da İngilizce'de *k-fold cross validation* olarak bilinmektedir. Stone (1974), farklı araştırmacılar tarafından çapraz geçerleme yönteminin ele alındığını; ancak açıkça yapılan ilk tanımın Mosteller ve Tukey (1968) tarafından yapıldığını belirtmiştir. Buna göre, veri seti *k* eşit parçaya bölündükten sonra her parça bir defasında test veri seti, geri kalan *k-1* sayıdaki parçalar ise eğitim veri seti olarak ayrılır. Her seferinde bir parça test verisi olacağından, seçilen *k* sayısı kadar bu işlem tekrarlanır. Yani *k* değerinin 5 seçilmesi durumunda eğitim ve test olarak veri seti 5 eşit parçaya bölünecek ve işlem her bir parçanın test verisi olarak kullanılmasına imkân sağlayacak şekilde 5 kez tekrarlanacaktır (Şekil 2.11). Çapraz geçerleme yönteminin kullanıldığı durumlarda, modelin doğruluğu hesaplanırken ise *k* kez elde edilen doğrulukların ortalaması alınır. *k* değerinin 10 olarak seçilmesinin en optimal seçim olduğu araştırmacılar tarafından belirtilmektedir (Olson ve Delen, 2008; Witten ve diğ., 2011).




**Şekil 2.11:** 5 kat çapraz geçerleme (Balaban ve Kartal, 2015).

- **Bootstrap**

Bootstrap yönteminde eğitim veri setinin oluşturulması için, örnekler rastgele seçilmekte ancak her bir örnek seçildikten sonra tekrardan veri kümesine dâhil edilmektedir (Efron ve Tibshirani, 1993). Bu sebeple bootstrap yöntemi *"sampling with replacement"* olarak tanımlanmaktadır (Witten ve diğ., 2011; Han ve diğ., 2012). Yani seçilen her bir örnek tekrardan veri kümesine dâhil edilmekte ve oluşturulan eğitim veri setindeki örnekler tekrar edebilmektedir. Geri kalan örnekler ise test veri seti olarak ele alınmaktadır.

#### *2.2.6.2. Model Performans Değerlendirme Ölçütleri*

Model kurulduktan ve analiz yapıldıktan sonra modelin ne kadar doğru sonuç verdiğinin değerlendirilmesi için yararlanılan ölçütlere model performans değerlendirme ölçütleri adı verilmektedir. Bu ölçütlere göre yapılan karşılaştırma, en iyi performansı veren modelin seçilmesini sağlamaktadır. Model performansını değerlendirmek için kullanılan değerler Tablo 2.8'de gösterilmiştir. Bu değerlerin hesaplanabilmesi için doğru pozitif, yanlış pozitif, doğru negatif, yanlış negatif değerlerinin kullanılması gerekmektedir (Han ve diğ., 2012; Balaban ve Kartal, 2015).

**Doğru pozitif *(true positive - TP):*** Gerçekte "pozitif" sınıfına ait olup, modelin "pozitif" sınıfına ait olarak tahmin ettiği örneklerin sayısını ifade eder. Tahmin edilen örneklerin gerçekteki sınıfı pozitif sınıftır; sınıflandırma doğru yapılmıştır.

**Doğru negatif *(true negative - TN):*** Gerçekte "negatif" sınıfına ait olup, modelin "negatif" sınıfına ait olarak tahmin ettiği örneklerin sayısını ifade eder. Tahmin edilen örneklerin gerçekteki sınıfı negatif sınıftır; sınıflandırma doğru yapılmıştır.

**Yanlış pozitif *(false positive - FP):*** Gerçekte "negatif" sınıfına ait olup, modelin "pozitif" sınıfına ait olarak tahmin ettiği örneklerin sayısını ifade eder. Tahmin edilen örneklerin gerçekteki sınıfı negatif sınıftır; sınıflandırma yanlış yapılmıştır.

**Yanlış negatif *(false negative - FN):*** Gerçekte "pozitif" sınıfına ait olup, modelin "negatif" sınıfına ait olarak tahmin ettiği örneklerin sayısını ifade eder. Tahmin edilen örneklerin gerçekteki sınıfı pozitif sınıftır; sınıflandırma yanlış yapılmıştır.

Buna göre doğru pozitif *(TP)* ve doğru negatif değerleri *(TN),* doğru sınıflandırmayı gösterirken, yanlış pozitif *(FP)* ve yanlış negatif *(FN)* değerleri yanlış yapılan

sınıflandırmayı gösterir (Witten ve diğ., 2011). Bu değerleri yani gerçek sınıfları ve modelin tahmin ettiği değerleri karmaşıklık matrisi *(confusion matrix, contingency matrix)* üzerinde görmek mümkündür (Kohavi ve Provost, 1998). Sınıflandırma modelinin performansı örneklerin doğru ya da yanlış sınıflarda tahmin edilmesine göre yapılır (Gorunescu, 2011). Karmaşıklık matrisi (Tablo 2.7) üzerindeki değerlere göre çeşitli hesaplamalar yapılarak modelin performansı değerlendirilir (Tablo 2.8). Karmaşıklık matrisi için, kontenjans tablosu, karışıklık matrisi ve konfüzyon matrisi gibi farklı isimler kullanılmaktadır.

**Tablo 2.7:** Karmaşıklık matrisi (Han ve diğ., 2012).

| **Karmaşıklık Matrisi** | | **Tahmin** | | |
|---|---|---|---|---|
| | | Pozitif | Negatif | Toplam |
| **Gerçek** | Pozitif | **TP**<br>**doğru pozitif** | **FN**<br>**yanlış negatif** | P |
| | Negatif | **FP**<br>**yanlış pozitif** | **TN**<br>**doğru negatif** | N |
| | Toplam | P' | N' | P+N |

**P:** Gerçekte pozitif sınıfa ait olan örneklerin sayısı
**N:** Gerçekte negatif sınıfa ait olan örneklerin sayısı
**P':** Pozitif sınıfta olduğu tahmin edilen örneklerin sayısı
**N':** Negatif sınıfta olduğu tahmin edilen örneklerin sayısı

**Tablo 2.8:** Değerlendirme ölçütleri (Gorunescu, 2011; Han ve diğ., 2012).

| **Ölçüt** | **Açıklama** | **Formül** |
|---|---|---|
| Doğruluk<br>*(accuracy, recognition rate)* | Doğru tahmin edilen örneklerin, tüm örneklerin sayısına oranıdır. | $\frac{TP+TN}{TP+FP+TN+FN} = \frac{TP+TN}{P+N}$ |
| Hata oranı<br>*(error rate, misclassification rate)* | Doğruluk değeri ile toplamı bire eşittir. | $\frac{FP+FN}{TP+FP+TN+FN} = \frac{FP+FN}{P+N}$<br>$1 - doğruluk$ |
| Duyarlılık<br>*(sensitivity, recall, true positive rate-TPR)* | Doğru tahmin edilen pozitif sınıftaki örneklerin, gerçekte pozitif sınıftaki örneklerin sayısına oranıdır. | $\frac{TP}{TP+FN} = \frac{TP}{P}$ |
| Belirleyicilik<br>*(specificity, true negative rate-TNR)* | Doğru tahmin edilen negatif sınıftaki örneklerin, gerçekte negatif sınıftaki örneklerin sayısına oranıdır. | $\frac{TN}{TN+FP} = \frac{TN}{N}$ |
| Pozitif öngörü değeri/Kesinlik<br>*(positive predictive value, precision)* | Doğru tahmin edilen pozitif sınıftaki örneklerin, pozitif sınıfta tahmin edilen örneklerin sayısına oranıdır. | $\frac{TP}{TP+FP} = \frac{TP}{P'}$ |
| Negatif öngörü değeri<br>*(negative predictive value)* | Doğru tahmin edilen negatif sınıftaki örneklerin, negatif sınıfta tahmin edilen örneklerin sayısına oranıdır. | $\frac{TN}{TN+FN} = \frac{TN}{N'}$ |

**Tablo 2.8 (devam):** Değerlendirme ölçütleri (Gorunescu, 2011; Han ve diğ., 2012).

| Ölçüt | Açıklama | Formül |
|---|---|---|
| Yanlış pozitif oranı *(false positive rate-FPR)* | Yanlış tahmin edilen negatif sınıftaki örneklerin, negatif sınıftaki örneklerin sayısına oranıdır. | $\frac{FP}{FP+FN} = \frac{FP}{N}$ $1 - belirleyicilik$ |
| Yanlış negatif oranı *(false negative rate-FNR)* | Yanlış tahmin edilen pozitif sınıftaki örneklerin, pozitif sınıftaki örneklerin sayısına oranıdır. | $\frac{FN}{FN+TP} = \frac{FN}{P}$ $1 - TPR$ |
| F-ölçütü *(F-score)* | Duyarlılık ve kesinliğin birlikte değerlendirildiği ölçüttür. | $\frac{2 \times kesinlik \times duyarlılık}{kesinlik \times duyarlılık}$ |

### 2.2.7. Literatür Taraması

Kalp hastalıkları ve veri madenciliği ile ilgili, veri madenciliği, sınıflandırma, Bayes sınıflandırıcı, karar ağacı, KNN, C4.5, C5.0 anahtar kelimeleri ile yapılan kaynak taramasında birçok çalışmanın yapıldığı görülmüştür (Tablo 2.9). Sağlık verisi ile yapılan bu çalışmalarda veri madenciliği yöntemleri kullanılmış olup çalışmaların büyük çoğunluğunda sınıflandırma algoritmalarının yardımıyla erken hastalık tespiti veya tahmini yapılmaya çalışılmıştır. Ancak çalışmaların yapıldığı alanlara bakıldığında tıp dışındaki farklı, genellikle bilişim teknolojileri ile alakalı alanlardaki araştırmacılar tarafından bu çalışmaların yapıldığı dikkat çekmektedir. Bu durum sağlık alanında, farklı pozisyonlardaki araştırmacıların; veri madenciliği, makine öğrenmesi veya yapay zekâ gibi, hastalıkların tespiti ve tahmininde destek niteliğinde kullanılabilecek araştırmalar uzak olduğunu düşündürmektedir. Tablo 2.9'da görülen çalışmalar mühendislik veya bilişim teknolojileri alanlarında çalışan araştırmacılar tarafından yapılmıştır. Rajathi ve Radhamani (2016) çalışmalarında bu tez çalışmasındakine benzer şekilde akut romatizmal ateşi konu almışlardır; ancak kullanılan veri seti nitelik bakımından farklıdır.

**Tablo 2.9:** Kalp hastalıkları ile ilgili yapılan veri madenciliği çalışmaları.

| Çalışma Başlığı | Kullanılan Yöntemler | Yayımlandığı Kaynak | Kullanılan Araçlar | Kaynak |
|---|---|---|---|---|
| Using Decision Tree for Diagnosing Heart Disease Patients | J48, Bagging | Proceedings of the 9th Australasian Data Mining Conference (AusDM'11) | WEKA | (Shouman ve diğ., 2011) |
| Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction | RIPPER, karar ağaçları, yapay sinir ağları, destek vektör makineleri | International Journal of Computer Science & Technology | WEKA | (Kumari ve Godara, 2011) |
| A Data Mining Approach for Diagnosis of Coronary Artery Disease | sade Bayes, SMO, bagging SMO, yapay sinir ağları | Computer Methods and Programs in Biomedicine | RapidMiner | (Alizadehsani ve diğ., 2013) |
| Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm | KNN, genetik algoritmalar | International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) | belirtilmemiş | (Jabbar ve diğ., 2013) |
| Early Prediction of Heart Diseases Using Data Mining Techniques | CART, ID3 | Caribbean Journal of Science and Technology | WEKA | (Chaurasia ve Pal, 2013) |
| Computational Intelligence for Heart Disease Diagnosis: A Medical Knowledge Driven Approach | sade Bayes, destek vektör makineleri, IBK, AdaBoostM1, J48, PART | Expert Systems with Applications | WEKA | (Nahar ve diğ., 2013) |
| Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability | destek vektör makineleri, lineer diskriminant analizi, C4.5, KNN, BLR, MLR, PLS-DA, k-ortalamalar, EMC, apriori | International Journal of Scientific and Research Publications | Tanagra | (Lakshmi ve diğ., 2013) |
| An Overview of Data Mining Classification Methods in Aortic Stenosis Prediction | karar ağaçları, destek vektör makineleri | International Journal of Engineering and Advanced Technology | WEKA | (Revathi ve Sumathi, 2014) |
| Early Heart Disease Prediction Using Data Mining Techniques | karar ağaçları, sade Bayes, yapay sinir ağları | Computer Science & Information Technology (CS & IT) | WEKA | (Methaila ve diğ., 2014) |

**Tablo 2.9 (devam):** Kalp hastalıkları ile ilgili yapılan veri madenciliği çalışmaları.

| Çalışma Başlığı | Kullanılan Yöntemler | Yayımlandığı Kaynak | Kullanılan Araçlar | Kaynak |
|---|---|---|---|---|
| Diagnosis of Heart Disease Using Data Mining Algorithm | KNN | International Journal of Computer Science and Information Technologies (IJCSIT) | WEKA, MATLAB | (Chandna, 2014) |
| Prediction of Heart Disease using Classification Algorithms | J48, sade Bayes, REPTREE, CART, Bayes ağları | Proceedings of the World Congress on Engineering and Computer Science 2014 | WEKA | (Masethe ve Masethe, 2014) |
| Analysis of Data Mining Techniques for Diagnosing Heart Disease | sade Bayes, bagging, ID3, J48 CART, lojistik regresyon, REPTree | International Journal of Advanced Research in Computer Science and Software Engineering | WEKA | (Rohilla ve Gulia, 2015) |
| Comparative Study on Heart Disease Prediction System Using Data Mining Techniques | yapay sinir ağları, sade Bayes, C4.5 | International Journal of Science and Research | belirtilmemiş | (Revathi ve Jeevitha, 2015) |
| Diagnosis of Rheumatoid Arthritis Using an Ensemble Learning Approach | C4.5, ID3, J48, CHAID, KNN, destek vektör makineleri, ADABOOST, CSBOOST | Computer Science & Information Technology (CS & IT) | MATLAB | (Shiezadeh ve diğ., 2015) |
| Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO | destek vektör makineleri, KNN, KNNACO | 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) | belirtilmemiş | (Rajathi ve Radhamani, 2016) |
| Human Heart Disease Prediction System Using Data Mining Techniques | ID3, KNN | 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT) | belirtilmemiş | (Thomas ve Princy, 2016) |
| Comparative Study of Data Mining Techniques on Heart Disease Prediction System: a case study for the “Republic of Chad” | sade Bayes, destek vektör makineleri | International Journal of Science and Research (IJSR) | ASP.NET, C#, Phyton | (Ngueilbaye ve diğ., 2016) |

Türkiye'de yapılmış kalp hastalıkları ile ilgili çalışmalara bakıldığında ise veri madenciliği, makine öğrenmesi veya sınıflandırma yöntemlerinin kullanıldığı çok fazla çalışmaya ulaşılamamıştır (Tablo 2.10). Bulunan çalışmalar arasında makale, konferans bildirisi, yüksek lisans ve doktora tezleri bulunmaktadır. Bu tez kapsamında ele alınan, akut romatizmal ateş hastalığı ile ilgili yapılan taramada, bu hastalık veya benzer veri seti ile ilgili yapılmış herhangi bir çalışmaya rastlanmamıştır. Ayrıca analizlerde kullanılan araçlar açısından bakıldığında da yalnızca ulaşılan bir doktora tez çalışmasında (Kartal, 2015) bu çalışmada olduğu gibi R Programlama Dili ve RStudio kullanılmıştır. Hem yabancı hem de Türkçe kaynaklar arasında yapılan taramada analiz aracı olarak genellikle WEKA'nın tercih edildiği göze çarpmaktadır.

**Tablo 2.10:** Türkiye'de kalp hastalıkları ile ilgili yapılan veri madenciliği çalışmaları.

| Çalışma Başlığı | Çalışma Türü | Kullanılan Yöntemler | Birim | Kullanılan Araçlar | Kaynak |
|---|---|---|---|---|---|
| Çoklu Sınıflandırıcı Sistemleri ile Konjestif Kalp Yetmezliği Teşhisi | Yüksek Lisans Tezi | KNN, lineer diskriminant analizi, yapay sinir ağları, destek vektör makineleri, radyal tabanlı fonksiyon | Bülent Ecevit Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik-Elektronik Mühendisliği Anabilim Dalı | WEKA | (Narin, 2013) |
| Twoing Algoritması ile Sınıflandırma: Kalp Hastalığı Uygulaması | Bildiri | twoing | XVI. Akademik Bilişim Konferansı Bildirileri | belirtilmemiş | (Uysal ve diğ., 2014) |
| Makine Öğrenmesi Algoritmaları Kullanılarak Kalp Hastalığı Tespiti | Bildiri | yapay sinir ağları, sade Bayes | International Conference On Education In Mathematics, Science & Technology Proceeding Book | MATLAB | (Boyraz ve diğ., 2014) |
| Kalp Krizi Riskinin Bir Veri Madenciliği Uygulaması ile Analizi | Yüksek Lisans Tezi | gri algoritması, apriori algoritması | Muğla Sıtkı Koçman Üniversitesi, Fen Bilimleri Enstitüsü, Elektronik ve Bilgisayar Eğitimi Anabilim Dalı | SPSS Clementine | (Elmaz, 2014) |
| Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama | Doktora Tezi | sade Bayes, KNN, Logistik Regresyon, ID3, C4.5 | İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Enformatik Anabilim Dalı | RStudio | (Kartal, 2015) |

**Tablo 2.10 (devam):** Türkiye'de kalp hastalıkları ile ilgili yapılan veri madenciliği çalışmaları.

| Çalışma Başlığı | Çalışma Türü | Kullanılan Yöntemler | Birim | Kullanılan Araçlar | Kaynak |
|---|---|---|---|---|---|
| AdaBoost ile Kalp Krizi Risk Tespiti | Makale | CART, AdaBoost | Celal Bayar Üniversitesi Fen Bilimleri Dergisi | MATLAB, C++ | (Bulut, 2016) |
| Predicting heart Diseases by Using Machine Learning Methods | Yüksek Lisans Tezi | Yapay sinir ağları, destek vektör makineleri, KNN | Atılım Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı | belirtilmemiş | (Benzreig, 2016) |

# 3. MALZEME VE YÖNTEM

Veri analizi Fayyad ve diğ. (1996b) tarafından ortaya konulan veri tabanlarında bilgi keşfi süreci takip edilerek yapılmıştır. Buna göre, uygulanan adımlarda ne yapıldığı ilgili başlıklarda ele alınmıştır.

## 3.1. VERİ

Bu tez çalışmasında kullanılmak üzere İstanbul Anadolu Kuzey Kamu Hastaneleri Birliğine bağlı Medeniyet Üniversitesi Göztepe Eğitim ve Araştırma Hastanesi, Çocuk Kliniği Çocuk Kardiyolojisi Polikliniği'nden, polikliniği ziyaret etmiş hastalara ait veri alınmıştır.

Modifiye Jones Kriterleri'ne göre ARA şüphesi veya reaktivasyonu tanısı ile 01.06.2003-01.03.2012 tarihleri arasında takibe alınan 297 hastaya ait veri kullanılmıştır. Hasta verisi retrospektif olarak toparlanmıştır. Bu nedenle veri setinde bulunan eksik verinin hasta kayıtlarına geri dönerek tamamlanması veya yanlış olduğundan şüphelenilen veriyi teyit etmek mümkün olmamış ve şüpheli kayıtlar veri setinden çıkarılmıştır.

Retrospektif olarak toplanan verinin analizlerde kullanılabilmesi için İstanbul Üniversitesi Tıp Fakültesi Klinik Araştırmalar Etik Kurulu'ndan rapor alınmıştır (EK.1).

### 3.1.1. Orijinal Veri Seti

Orijinal veri setinde 297 kayıt (satır), 35 adet nitelik (sütun) bulunmaktadır (Şekil 3.1). Nitelikler çoğunlukla kategorik değerlerden oluşmakta ancak veri setinde nümerik değerler de bulunmaktadır. 20 kategorik, 8 nümerik tipinde, 3 tarih ve 4 yazılı olarak ifade edilmiş nitelik mevcuttur; nitelikler Tablo 3.1'de verilmiştir.

**Tablo 3.1:** Orijinal veri setindeki nitelikler.

| | Nitelik İsimleri | Veri Tipi |
|---|---|---|
| 1. | Cinsiyet | Kategorik |
| 2. | Doğumtarihi | Tarih |
| 3. | Hastanesüre | Nümerik |
| 4. | İlkataktarih | Tarih |
| 5. | İlkatakyaş | Nümerik |
| 6. | Artrit | Kategorik |
| 7. | Kardit | Kategorik |
| 8. | Korea | Kategorik |
| 9. | Diğer | Kategorik |
| 10. | Lokosit | Nümerik |
| 11. | Hb | Nümerik |
| 12. | HTC | Nümerik |
| 13. | CRP | Nümerik |
| 14. | Sedim | Nümerik |
| 15. | ASO | Nümerik |
| 16. | Boğazkültürü | Kategorik |
| 17. | ARAöyküsü | Kategorik |
| 18. | Rekurrens | Kategorik |
| 19. | EKG | Kategorik |
| 20. | DLAB | Yazılı açıklama |
| 21. | EKOilkatak | Kategorik |
| 22. | EKOtakip | Kategorik |
| 23. | EKOdiğer | Yazılı açıklama |
| 24. | VAR000001 | Kategorik |
| 25. | VAR000003 | Kategorik |
| 26. | Penisilin | Kategorik |
| 27. | ASA | Kategorik |
| 28. | Kortizon | Kategorik |
| 29. | Diğerilaç | Yazılı açıklama |
| 30. | Açıklama | Yazılı açıklama |
| 31. | Özgeçmiş | Kategorik |
| 32. | Aile hikaye | Kategorik |
| 33. | Operasyon | Kategorik |
| 34. | Sonkontrol | Tarih |
| 35. | Takiptençıkma | Kategorik |

| Cins | DT | hastanesüre | İLKATAKT | İLKATAKYAŞ | Artrit | Kardit | Korea | Diğer | Lökosit | Hb | HTC | CRP | Sedim | ASO | bogazkültürü | ARAöyküsü | rekurrens | EKG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K | 01-mmm-1990 | | | | 1 | 0 | 0 | 0 | 8100 | 10 | | 12 | 120 | 400 | 2 | 1 | 1 | 1 |
| E | | | 05-mmm-1998 | | 0 | 1 | 0 | 0 | | 10 | 34 | 48 | 130 | | | 0 | 0 | |
| E | 06-mmm-1990 | | 06-mmm-1990 | | 1 | 1 | 0 | 0 | | | | | 115 | | | 0 | 0 | 0 |
| K | 06-mmm-2001 | | 02-mmm-2001 | | 1 | 1 | 0 | 0 | | | | | | | | | | |
| K | 04-mmm-1998 | 17 | 10-mmm-1998 | | 1 | 1 | 0 | 0 | 7300 | 11 | | 24 | 120 | 800 | 0 | 0 | 0 | 0 |
| K | 11-mmm-1998 | | 03-mmm-2004 | 0 | 0 | 0 | 0 | 0 | | | | | | | | 0 | 0 | 0 |
| E | 02-mmm-1992 | 10 | 02-mmm-2000 | 0 | 0 | 0 | 0 | 23 | 16600 | 11 | 36 | 46 | 55 | 400 | 0 | 0 | 0 | |
| E | 11-mmm-1988 | | 02-mmm-1999 | 0 | 0 | 0 | 0 | 0 | | | | 48 | | 400 | | 0 | 0 | |
| K | 01-mmm-1992 | | 11-mmm-2002 | 0 | 0 | 0 | 0 | 0 | | | | | | | | 0 | 0 | |
| E | 09-mmm-1994 | 2 | 10-mmm-2001 | 0 | 0 | 0 | 0 | 0 | 8870 | 11 | 31 | 0 | 20 | | | 0 | 0 | 0 |
| K | 07-mmm-1988 | 8 | 09-mmm-2001 | 0 | 1 | 0 | 0 | 0 | 6700 | 12 | 36 | 35 | 87 | 615 | 0 | 0 | 0 | 0 |
| K | 11-mmm-1984 | 0 | 11-mmm-1996 | 0 | 1 | 0 | 0 | 0 | | | | | | | | 0 | 0 | 0 |
| K | 11-mmm-1990 | 2 | 05-mmm-1994 | 0 | 1 | 0 | 0 | 0 | 14000 | 12 | | 48 | 60 | 100 | | 0 | 0 | |
| K | | | 11-mmm-1997 | 0 | 0 | 1 | 0 | 0 | | | | | | | | | | |
| K | 06-mmm-1997 | | 11-mmm-2001 | 4 | 0 | 1 | 0 | 0 | | | | | | | | 1 | 0 | |
| K | 07-mmm-1999 | | 01-mmm-2003 | 4 | 1 | 1 | 0 | 0 | | | | | | | | 0 | 0 | |
| E | 10-mmm-1997 | | 04-mmm-2002 | 5 | 0 | 0 | 0 | 0 | 9200 | 09 | | 24 | 80 | 600 | 0 | 1 | 0 | 0 |
| K | 02-mmm-1988 | 0 | 12-mmm-1996 | 5 | 1 | 0 | 0 | 0 | | | | | | | | 0 | 0 | 0 |
| K | 08-mmm-1992 | 11 | 12-mmm-1996 | 5 | 0 | 1 | 0 | 0 | 27300 | 10 | | | 40 | | 0 | 0 | 0 | 0 |
| E | 06-mmm-1992 | 11 | 08-mmm-1997 | 5 | 0 | 1 | 0 | 0 | 12400 | 09 | 33 | 0 | 90 | 800 | 0 | 0 | 0 | 0 |
| E | 01-mmm-1995 | 14 | 09-mmm-2000 | 5 | 1 | 1 | 0 | 0 | 10700 | 10 | 37 | 32 | 125 | | 0 | 0 | 0 | 0 |
| K | 06-mmm-1987 | 14 | 08-mmm-2002 | 5 | 1 | 1 | 0 | 0 | 11300 | 07 | 24 | 96 | 140 | 1440 | 0 | 1 | 1 | |
| E | 08-mmm-1989 | | 11-mmm-2001 | 5 | 1 | 1 | 0 | 0 | | | | | | | | 1 | | |
| K | 10-mmm-1998 | 24 | 05-mmm-2002 | 5 | 1 | 1 | 0 | 0 | 12300 | 10 | 30 | 19 | 120 | 1530 | 0 | 0 | 0 | |
| K | 04-mmm-1993 | 14 | 10-mmm-1999 | 6 | 1 | 0 | 0 | 123 | 17900 | 12 | 35 | 48 | 50 | | 2 | 0 | 0 | 0 |
| E | 09-mmm-1990 | 12 | 02-mmm-1996 | 6 | 1 | 0 | 0 | 0 | 7300 | 10 | 31 | 12 | 100 | 800 | 2 | 1 | 0 | 0 |
| K | 08-mmm-1990 | 7 | 08-mmm-1996 | 6 | 1 | 0 | 0 | 0 | | | | | | | | 0 | 0 | 0 |

**Şekil 3.1:** Orijinal veri setinin MS Excel'deki örnek görünümü.

## 3.2. ÖN İŞLEME

Veri setinin orijinal hali analiz için elden geçirilmiş; bazı nitelik alanları çıkarılmış, bazıları ise uzman görüşü alınarak birleştirilmiştir veya farklı sütunlar eklenerek gösterilmiştir. Şüphe bırakan ve tekrarlayan kayıtlar çıkarılmış, eksik veri tamamlanmış ve uç değerler tespit edilerek veri seti R programlama dili ile analize uygun hale getirilmiştir (EK.2).

- *Karar sınıflarını içeren sütun oluşturulmuştur.*

Öncelikle "karar" sütununun oluşturulmasında birkaç nitelik alanı kullanılmış, uzman görüşüne başvurularak hedef sınıflar belirlenmiştir. Böylece her bir kaydın ait olduğu sınıflar oluşturulmuştur.

Orijinal veri setinde, "EKOdiğer" niteliği ilgili doktorun not aldığı yazılı ifadelerden oluşmaktaydı. Nitelik, herhangi bir kategorik veya nümerik veriden değil standart bir şekilde yapılandırılmamış olan yazılı ifadelerden oluşuyordu. Ancak bu sütun hastanın baştaki durumundan takipteki durumuna kadar olan sürece dair veriyi içinde barındırdığından önem teşkil ediyordu.

"EKOdiğer"deki kalple ilgili yazılı ifadeler, "EKOilkatak" ve "EKOtakip" nitelikleri kullanılarak, aort kapağı ve mitral kapaktaki değişimi ifade eden, "ilk atak aort kapağı", "ilk atak mitral kapak", "takip aort kapağı", "takip mitral kapak" başlıklı 4 sütun veri setine eklenmiştir. "EKOdiğer"deki yazılı ifadeler, kapak durumlarının/tutulmalarının şiddetini ifade edecek şekilde "yok", "minimal", "hafif", "orta", "ağır/şiddetli" olarak 5

şekilde kategorize edilmiştir. Bu sütunların yanı sıra karar sütununun oluşturulmasında yardımcı olarak kullanılan bir başka sütun da "Açıklama" sütunu olmuştur. Yine yapılandırılmamış yazılı veriden oluşan "Açıklama", ilgili doktorun önemli gördüğü noktaları belirttiği veya kapaklardaki duruma dair not aldığı olan yazılı ifadelerden oluşmaktaydı. Bu sütun da karar sütununun oluşturulmasında rol oynamıştır. Uzman görüşü eşliğinde kayıtlar tek tek elden geçirilmiş; her bir kayıt için sınıflar belirlenmiştir. Her bir hastanın, "ilk atak aort kapağı", "ilk atak mitral kapak", "takip aort kapağı", "takip mitral kapak", "Açıklama" sütunları uzmanla birlikte incelenerek hastanın takipteki durumunun "ilerlemis", "iyilesmis", "gerilemis", "degismemis" ya da "takipsiz" sınıflarından hangisinde yer alması gerektiğine karar verilmiştir. Hastaların ait oldukları sınıf değerlerini ifade eden "karar" sütunu veri setine eklenmiştir.

- *Nitelik alanlarında düzenlemeler yapılmış; bazıları birleştirilmiş bazıları ise veri setinden çıkarılmıştır.*

"İlkataktarih" çıkarılmış, tarihe göre belirlenen "ay" ve "mevsim" sütunları veri setine eklenmiştir.

"Diğer" sütununda bir hücrede birden fazla seçenek bulunduğundan seçenek sayısı kadar nitelik alanı oluşturulmuştur. "ates", "atralji", "akutfazArtisi" sütunları eklenmiştir. Artritli hastalarda atralji minör kriter sayılmadığından, "atralji" "var" ise minör kriter "yok" olarak değiştirilmiştir.

Hem kategorik hem de nümerik veri ile analiz yapmak amacıyla "lokosit", "Hb", "HTC", "CRP", "Sedim" ve "ASO" değerleri kategorize edilmiştir/ayrıklaştırılmıştır. Hem orijinal veri setindeki gibi nümerik olarak analize dâhil edilmiş hem de uzman görüşü temel alınan kategorizasyona göre analize dâhil edilmişlerdir (Tablo 3.2). Kategorizasyona göre "lokosit_grup", "Hb_grup", "HTC_grup", "CRP_grup", "Sedim_grup" sütunları eklenmiştir. "lokosit", "CRP", "sedim" niteliklerinden biri yüksek ise "akutfazArtisi" "var" olarak işaretlenmiştir.

**Tablo 3.2:** Değerlerin uzman görüşünce kabul edilen kategorizasyonu.

| Nitelik Alanı | Veri Setindeki Aralıklar | Kategoriler | | | |
|---|---|---|---|---|---|
| | | 0=bilinmeyen | 1=düşmüş | 2=normal | 3=yükselmiş |
| Lökosit | 11-27300 | Veri yok. | <4500 | 4500-11000 cells/mm3 | >11000 |
| HB | 07-12,7 | Veri yok. | <11,5 | 11,5-13,5 g/dl | >13,5 |
| HTC | 11-54 | Veri yok. | <%35 | %35-45 | >%45 |
| CRP | 0-211 | Veri yok. | | <5 todd ünitesi | >5 todd ünitesi |
| Sedimantasyon | 1-169 | Veri yok. | <4 | 4-20 mm/saat | >20 |
| ASO | 0-6000 | Veri yok. | | 0-200 | >200 |

Yüksek "ASO" ve "boğazkültürü" streptokok enfeksiyonunun varlığını kanıtlayan nitelikler olduğundan "streptokokKaniti" sütununda birleştirilmiştir. "ASO", "ASO_grup" ve "boğazkültürü" sütunları çıkarılmıştır. Ayrıca uzman uyarısı ile orijinal veri setinde "kızıl" kelimesi taratılmıştır. Yazılı olarak kızılla ilgili öyküsü not alınan hastalar kontrol edilmiş, uzman görüşü doğrultusunda bu hastalar arasından "streptokokkaniti" "yok" olan 1 kişi için "var" olarak işaretlenmiştir. "Nüks", "relaps", "rekurrens", "reaktivasyon" kelimeleri orijinal veri setinde taratılmıştır. Bu kayıtlardan "rekurrens" sütunu işaretlenmeyen varsa, "rekurrens" sütunu "var" olarak işaretlenmiştir.

"EKG" ve orijinal veri setindeki "Diğer" sütunundaki değerlerden biri olan "PRuzamasi" birleştirilmiştir. "PRuzamasi" sütunu eklenmiş; "EKG" sütunu çıkarılmıştır. Karditli hastalarda PR uzaması minör kriter sayılmadığından, bu hastalarda "PRuzamasi" "var" olan kayıtlar "yok" olarak değiştirilmiştir.

"DLAB" sütunu yerine "teleNormal", "kardiyomegali", "enfeksiyon", "kansizlik" sütunları eklenmiştir. Uzman tarafından önemli kabul edilen "normal", "kardiyomegali", "enfeksiyon" değerleri alınmıştır. Bunlar dışında varolan ifadeler yeni veri setine dâhil edilmemiştir. Uzman görüşü alınarak, kardiyotorasik oran 0,50'nin altında ve sınırda olanlar normal kabul edilmiştir.

"EKOilkatak", "EKOtakip", "EKOdiğer" ve "Açıklama" sütunları veri setinden çıkarılmıştır. Bu nitelikler yukarıda açıklandığı üzere "karar" sütunun oluşturulmasında kullanılmıştır.

“VAR000001” ve “VAR000003” sütunlarının ne ile ilgili olduğu retrospektif veri içerisinden bulunamadığından veri setinden bu iki sütun çıkarılmıştır.

“Penisilin” sütunu yerine “benzatinPenisilin”, “prokainPenisilin” ve “eritromisin” sütunları eklenmiştir.

“Diğerilaç” sütunu elden geçirilerek kullanılan ilaçlar türlerine göre kategorize edilmiş, sonrasında her bir ilaç türü için 15 tane sütun veri setine eklenmiştir. Ancak ayrı birer sütun olarak “kortizon” ve “aspirin” sütunları veri setinde bulunduğundan “Diğerilaç” içerisindeki kortizon ve aspirin sütunları eklenmemiştir.

“Operasyon” sütunundaki veriden hareketle “aortRep”, “mitralRep”, “digerOp” olmak üzere 3 sütun oluşturulmuştur.

“Sonkontrol” sütunu çıkarılmış, yerine “Doğumtarihi”ne göre hesaplanan son kontrole gelme yaşı yazılmıştır. Sütun ismi “sonkontrolYas” olarak değiştirilmiştir.

“Takiptençıkma” sütunu veri setinden çıkarılmıştır.

- *Bazı kayıtlar veri setinden çıkarılmıştır.*

“karar” niteliğinde ait olduğu sınıf boş olan 44 kişi veri setinden çıkarılmıştır.

“karar” niteliğinde ait olduğu sınıf “takipsiz” olarak belirlenmiş olan 34 kişi veri setinden çıkarılmıştır.

Tanı koymaya yardımcı olan minör ve majör kriter sayısını karşılamayan 5 kişi veri setinden çıkarılmıştır.

Şüpheli olan ve/veya tekrarlayan 14 kişi veri setinden çıkarılmıştır.

- *Eksik veri tamamlanmıştır.*

Tüm nitelik alanları için boş olan hücrelerdeki veri hasta dosyalarında bulunmamakta, veri seti retrosopektif nitelikte olduğundan veri güncellenememekte veya teyit edilememektedir. Bu sebeplerden dolayı boş olan alanlar verinin olmadığı *(missing value)* alanlar olarak kabul edilmiştir.

Eksik verinin tamamlanması için RStudio üzerinde *"DMwR"* (Torgo, 2010) paketi içerisindeki *"knnImputation*" fonksiyonundan yararlanılmıştır. "knnImputation" fonksiyonu *NA* olan değerin *k* tane en yakın komşusunun değerlerine bakar ve bunların ağırlıklı ortalamasını alır. Boş olan değer bu şekilde doldurulur. Fonksiyon çalıştırılırken varsayılan değer olan *k=10* değerine göre işlem gerçekleştirilmiştir. Eksik veri tamamlanmadan önce veri setinin özeti Şekil 3.2'deki gibi görülmektedir. Örnek olarak kategorik veri setindeki görünüm konmuştur.

Eksik değer tamamlama, RStudio üzerinde farklı paket ve fonksiyonlar kullanılarak gerçekleştirilebilmektedir. *"knnImputation"* fonksiyonu dışında; *"mice"* paketindeki *"complete"* fonksiyonu (Buuren ve Groothuis-Oudshoorn, 2011), *"Hmisc"* paketindeki *"aregImpute"* fonksiyonu (Harrell, 2017) yararlanılabilecek diğer yöntemlere örnek olarak verilebilir.

```
> summary(veriseti)
    cins      hastaneSure          ay         mevsim     ilkatakYas    artrit    kardit    korea       ates
 erkek: 97   Min.   : 2.00   aralik :26   ilkbahar:53   Min.   : 4.00   var:141   var:186   var: 24   var : 27
 kadin:104   1st Qu.: 9.00   ocak   :22   kis      :66   1st Qu.: 9.00   yok: 60   yok: 15   yok:177   yok :172
             Median :13.00   nisan  :20   sonbahar:37   Median :11.00                                 NA's:  2
             Mean   :13.63   haziran:18   yaz      :44   Mean   :10.44
             3rd Qu.:16.00   mayis  :18   NA's     : 1   3rd Qu.:12.00
             Max.   :51.00   (Other):96                 Max.   :16.00
             NA's   :43      NA's   : 1                 NA's   :6
 atralji    akutfazArtisi    lokosit           Hb              HTC              CRP              sedim
 var : 18   var:162       Min.   : 4300   Min.   : 7.40   Min.   :11.00   Min.   :  0.00   Min.   :  4.00
 yok :181   yok: 39       1st Qu.: 8240   1st Qu.:10.00   1st Qu.:31.10   1st Qu.:  6.80   1st Qu.: 75.50
 NA's:  2                 Median :10100   Median :10.80   Median :33.40   Median : 22.00   Median :110.00
                          Mean   :10926   Mean   :10.93   Mean   :33.78   Mean   : 31.49   Mean   : 97.04
                          3rd Qu.:12800   3rd Qu.:11.93   3rd Qu.:36.10   3rd Qu.: 48.00   3rd Qu.:120.00
                          Max.   :26000   Max.   :17.70   Max.   :53.70   Max.   :211.00   Max.   :169.00
                          NA's   :35      NA's   :37      NA's   :46      NA's   :44       NA's   :34
 streptokokKaniti ARAoykusu          rekurrens   PRuzamasi teleNormal kardiyomegali enfeksiyon kansizlik
 var :110         var : 34   rekurrens     : 27   var :38   0   : 20   0   : 92      0   : 95   0   : 97
 yok : 19         yok :131   rekurrensdegil:133   yok :85   1   : 79   1   :  7      1   :  4   1   :  2
 NA's: 72         NA's: 36   NA's          : 41   NA's:78   NA's:102   NA's:102      NA's:102   NA's:102

  ilkatakAort  ilkatakMitral    takipAort    takipMitral benzatinPenisilin prokainPenisilin eritromisin
 hafif   :65   hafif   :87    hafif   :29   hafif   :61   0   :  2          0   : 68         0:199
 minimal :24   minimal :19    minimal :27   minimal :71   1   :198          1   :132         1:  2
 orta    :21   orta    :50    orta    : 6   orta    : 9   NA's:  1          NA's:  1
 siddetli: 5   siddetli:26    siddetli: 4   siddetli: 9
 yok     :11   yok     : 9    yok     :56   yok     :44
 NA's    :75   NA's    :10    NA's    :79   NA's    : 7

 aspirin    kortizon   alerji  antibiyotik atesdusurucu balgamsokucu idrarsokturucu kalpilaci kalpyetmezligi
 0   : 25   0   : 51   0:200   0:186       0:199        0:200        0:197          0:200     0:138
 1   :156   1   :126   1:  1   1: 15       1:  2        1:  1        1:  4          1:  1     1: 63
 NA's: 20   NA's: 24

 kanilaci kansulandirici mideilaci norolojikilac plazma  potasyumtakviyesi ozgecmis   ailehikaye aortRep
 0:200    0:200          0:200     0:187        0:200   0:199             var :  7   var : 10   0   :193
 1:  1    1:  1          1:  1     1: 14        1:  1   1:  2             yok :192   yok :189   1   :  7
                                                                        NA's:  2   NA's:  2   NA's:  1

 mitralRep  digerOp    sonkontrolYas                  takip          karar
 0   :199   0   :199   Min.   : 4.28   eriskinegonderilmis: 46   degismemis: 30
 1   :  1   1   :  1   1st Qu.:12.71   takipsiz           :153   gerilemis :104
 NA's:  1   NA's:  1   Median :15.78   NA's               :  2   ilerlemis : 21
                       Mean   :15.20                             iyilesmis : 46
                       3rd Qu.:18.02
                       Max.   :25.01
                       NA's   :5
```

**Şekil 3.2:** Eksik veri tamamlanmadan önce veri setinin özeti.

- *Aykırı (outlier) değer tespiti yapılmıştır.*

Veri setindeki eksik veri tamamlandıktan sonra uç değer tespiti yapılmıştır. Bu işlem sadece nümerik veri için yapılmıştır. RStudio üzerinde, "lokosit", "Hb", "HTC", "CRP", "Sedim" nitelikleri için *"DMwR"* (Torgo, 2010) paketindeki *"lofactor"* fonksiyonu kullanılarak uç değer olabilecek kayıtlar incelenmiştir. Breunig ve diğ. (2000)'nin çalışmasına dayanan fonksiyonda, her bir değer için *k* en yakın komşusuna olan uzaklıklara göre lokal yoğunluk *(local density)* hesaplanır. Her değerin *k* en yakın komşu ile yoğunluğu karşılaştırılır ve benzer yoğunlukta olmayan değerler uç değer olarak kabul edilir. Bir çeşit kümeleme işlemi *(density based clustering)* yapılarak uç değerler ortaya çıkarılmış olur (Breunig ve diğ., 2000). Fonksiyon bu işlemi yaparken aykırı değerlere ait skorlar oluşturur. Bu çalışmada, diğer skorlara yakın değerde olmasına rağmen, skoru 2,26 ve 2,13 olan iki kaydın aykırı değer olabileceğinden şüphelenilmiştir (Şekil 3.3). Bu nedenle, bu iki değer veri setinden çıkarılıp ve çıkarılmadan analizler yapılmış; sonuçlar arasında belirgin bir fark saptanamadığından tez kapsamında bu iki kaydın da dâhil olduğu veri seti *(n=201)* ile yapılan analiz sonuçları paylaşılmıştır.

Aykırı değerler tespiti, RStudio üzerinde farklı paket ve fonksiyonlar kullanılarak gerçekleştirilebilmektedir. *"lofactor"* fonksiyonu dışında; *"mvoutlier"* paketindeki *"pcout"* fonksiyonu (Filzmoser ve Gschwandtner, 2017), *"randomForest"* paketindeki *"outlier"* fonksiyonu (Liaw ve Wiener, 2002) yararlanılabilecek diğer yöntemlere örnek olarak verilebilir.

```
> outlier.scores
  [1] 1.1588162 1.0377618 1.0092754 2.2644087 0.9728406 1.0180163 0.9810113 1.2323139 1.0037543 1.3103050 1.0952712
 [12] 1.2146072 1.0448742 1.0556111 1.0177221 1.0938493 1.0558793 0.9716109 0.9620943 1.0098658 0.9979555 1.0657118
 [23] 1.0607949 1.0612628 0.9663283 0.9625222 1.1606189 1.0291315 1.0227085 0.9588681 0.9967235 0.9610857 1.2458935
 [34] 1.0874805 1.0124660 0.9778361 1.0668375 1.0935295 1.0884765 0.9812817 1.1750194 0.9823227 1.0377608 1.0736415
 [45] 1.2986882 1.6330329 1.4142520 1.1501300 0.9813119 0.9473273 0.9829335 2.1358465 1.0445853 1.1408001 1.1404553
 [56] 1.2226609 0.9895408 1.1612140 0.9887375 1.0682843 1.0747798 1.0092754 0.9736886 1.0576917 0.9935158 1.0924446
 [67] 0.9703638 0.9660903 0.9509947 1.2798999 1.1445750 1.0992954 1.0095341 0.9720098 1.1354244 1.0478634 1.0935295
 [78] 1.0537982 0.9609309 0.9791305 1.1344149 0.9574821 1.1187815 1.0083451 1.0217039 1.1686401 1.0015715 1.0139799
 [89] 0.9744538 1.0704752 1.0271600 1.0065948 1.0587475 1.0853300 1.0030993 1.0257294 1.0345070 0.9727410 1.0039464
[100] 1.1417524 0.9868243 1.0195992 1.0257294 0.9660903 0.9885078 1.1570588 0.9987018 0.9797533 1.1111957 0.9778902
[111] 1.2116108 1.3371816 1.0929237 1.1256305 1.0167871 1.1483897 1.0364383 0.9645983 1.2200617 0.9784284 0.9973320
[122] 0.9723268 1.0877495 1.0747321 1.2437503 1.1417422 0.9738665 1.0474185 1.0910046 1.0668048 1.0465280 0.9997949
[133] 1.0085289 1.0218969 0.9818624 0.9838304 1.0052783 1.0133386 1.1594797 1.2246494 1.2505234 0.9765066 1.0014779
[144] 1.3470990 1.0284029 1.0237904 1.2089682 0.9746684 1.0557378 0.9807028 1.0749796 1.0360369 0.9697906 1.0935295
[155] 0.9909092 0.9533806 1.0008600 1.0348561 0.9984410 1.1640495 1.1156287 0.9862859 1.1436052 1.0275513 1.1586166
[166] 0.9602938 0.9735531 1.0934240 0.9591052 1.1264726 1.0100302 1.3103675 1.1629111 1.1132409 0.9763480 1.2159420
[177] 1.0474562 0.9995843 0.9923037 0.9923037 1.1767975 1.0053624 1.0842026 1.0709337 1.1688072 1.0599258 0.9888824
[188] 1.0686217 0.9909318 1.1013111 0.9995843 0.9734559 1.1252557 1.0062182 1.0230668 1.1952147 1.1254632 0.9740550
[199] 1.0132098 1.0312554 0.9808819
```

**Şekil 3.3:** Aykırı değerlere ait skorlar.

### 3.2.1. İşlenmiş Veri Seti

Ön işlemenin ardından veri seti 201 kayıt (satır), 62 adet nitelikten (sütun) oluşan bir hale gelmiştir. Analizler için RStudio üzerinde veri seti tekrardan gözden geçirilmiş birkaç değişiklik daha yapılmıştır. "dogumtar" sütunu çıkarılmıştır. Silinen kayıtlar olması sebebiyle tek kategorisi (sadece "0") kalan "agrikesici" ve "parazitilaci" sütunları çıkarılmıştır. Kan ile ilgili değerleri ifade eden sütunlar hem nümerik ("lokosit", "Hb", "HTC", "CRP", "Sedim") hem de kategorik ("lokosit_grup", "Hb_grup", "HTC_grup", "CRP_grup", "Sedim_grup") olacak şekilde analize alınmış bu sebeple her seferinde hangi veri tipi ile çalışılacaksa diğer veri tipindeki 5 sütun veri setinden çıkarılmıştır. Her algoritma için analiz bir kez nümerik değerler bir kez de kategorik değerlerle olmak üzere iki kez tekrar edilmiştir. Sonuç olarak analizlere dâhil edilen veri setinde 201 kayıt, 54 nitelik alanı kullanılmıştır (Tablo 3.3).

**Tablo 3.3:** Analiz edilen veri setindeki nitelik açıklamaları.

| | **Nitelik İsimleri** | **Veri Tipi** |
|---|---|---|
| 1. | cins | Cinsiyet |
| 2. | hastaneSure | Hastanede yatış süresi |
| 3. | ay | İlk atağın geçirildiği ay |
| 4. | mevsim | İlk atağın geçirildiği mevsim |
| 5. | ilkatakyas | İlk atağın geçirildiği yaş |
| 6. | artrit | Majör kriterler |
| 7. | kardit | |
| 8. | korea | |
| 9. | ates | Minör kriterler |
| 10. | atralji | |
| 11. | akutfazartisi | |
| 12. | lokosit/lokosit_grup | Akyuvar hücreleri |
| 13. | hb/hb_grup | Hemoglobin |
| 14. | HTC/HTC_grup | Hematokrit |
| 15. | CRP/CRP_grup | C-reaktif protein |
| 16. | sedim/sedim_grup | Sedimantasyon, kanın çökme hızı |
| 17. | streptokokkaniti | Beta streptokok olup olmadığına dair kanıt, ASO (Antistreptolizin O) ve boğaz kültürü kanıtları |
| 18. | araoykusu | Önceden romatizma geçirilmiş mi? |
| 19. | rekurrens | Hastalık tekrarlamış mı? |
| 20. | pruzamasi | EKG çekiminde PR aralığında uzama var mı? |
| 21. | telenormal | Kalbin görüntüsü normal mi? |
| 22. | kardiyomegali | Kalpte büyüme var mı? |
| 23. | enfeksiyon | Enfeksiyon var mı? |
| 24. | kansizlik | Kansızlık var mı? |
| 25. | ilkatakaort | İlk atakta aort kapağındaki tutulumun derecesi |
| 26. | ilkatakmitral | İlk atakta mitral kapaktaki tutulumun derecesi |

**Tablo 3.3 (devam):** Analiz edilen veri setindeki nitelik açıklamaları.

| | Nitelik İsimleri | Veri Tipi |
|---|---|---|
| 27. | takipaort | Takipte aort kapağındaki tutulumun derecesi |
| 28. | takipmitral | Takipte mitral kapaktaki tutulumun derecesi |
| 29. | benzatinPenisilin | Kullanılan ilaçlar |
| 30. | prokainPenisilin | |
| 31. | eritromisin | |
| 32. | aspirin | |
| 33. | kortizon | |
| 34. | alerji | |
| 35. | antibiyotik | |
| 36. | atesdusurucu | |
| 37. | balgamsokucu | |
| 38. | idrarsokucu | |
| 39. | kalpilaci | |
| 40. | kalpyetmezligi | |
| 41. | kanilaci | |
| 42. | kansulandirici | |
| 43. | mideilaci | |
| 44. | norolojikilac | |
| 45. | plazma | |
| 46. | potasyumtakviyesi | |
| 47. | ozgecmis | Boğaz enfeksiyonuna sebebiyet verecek herhangi bir hastalık geçirildi mi? |
| 48. | ailehikaye | Ailede romatizma var mı yok mu? |
| 49. | aortRep | Kapak ameliyatı olmuş mu? |
| 50. | mitralRep | Kapak ameliyatı olmuş mu? |
| 51. | digerOp | Başka ameliyatı olmuş mu? |
| 52. | sonkontrolYas | Son kontorole gelme yaşı |
| 53. | takip | Hasta takipten çıkmış mı ya da erişkin kliniğine gönderilmiş mi? |
| 54. | karar | Hastaların takiplerinde kalp kapak tutulumunun değerlendirilmesi (“değişmemiş”, “gerilemiş”, “ilerlemiş”, “iyileşmiş”) |

Tezin geri kalan kısımlarında, veri setleri arasındaki ayrımı belirtmek için “nümerik veri seti” ve “kategorik veri seti” ifadeleri kullanılmıştır. Analiz edilen “nümerik veri seti”ndeki, nitelik alanları Şekil 3.4’te; “kategorik veri seti”ndeki nitelik alanları Şekil 3.5’te verilmiştir. Veri seti, orijinalinde olduğu gibi farklı tipteki veriden oluşmaktadır.

```
> colnames(tamveri)
 [1] "cins"              "hastaneSure"       "ay"                "mevsim"            "ilkatakYas"
 [6] "artrit"            "kardit"            "korea"             "ates"              "atralji"
[11] "akutfazArtisi"     "lokosit"           "Hb"                "HTC"               "CRP"
[16] "sedim"             "streptokokKaniti"  "ARAoykusu"         "rekurrens"         "PRuzamasi"
[21] "teleNormal"        "kardiyomegali"     "enfeksiyon"        "kansizlik"         "ilkatakAort"
[26] "ilkatakMitral"     "takipAort"         "takipMitral"       "benzatinPenisilin" "prokainPenisilin"
[31] "eritromisin"       "aspirin"           "kortizon"          "alerji"            "antibiyotik"
[36] "atesdusurucu"      "balgamsokucu"      "idrarsokturucu"    "kalpilaci"         "kalpyetmezligi"
[41] "kanilaci"          "kansulandirici"    "mideilaci"         "norolojikilac"     "plazma"
[46] "potasyumtakviyesi" "ozgecmis"          "ailehikaye"        "aortRep"           "mitralRep"
[51] "digerOp"           "sonkontrolYas"     "takip"             "karar"
```

**Şekil 3.4:** Analiz edilen nümerik veri setinin son halindeki nitelik alanları.

```
> colnames(tamveri)
 [1] "cins"              "hastaneSure"       "ay"                "mevsim"            "ilkatakYas"
 [6] "artrit"            "kardit"            "korea"             "ates"              "atralji"
[11] "akutfazArtisi"     "lokosit_grup"      "hb_grup"           "HTC_grup"          "CRP_grup"
[16] "sedim_grup"        "streptokokKaniti"  "ARAoykusu"         "rekurrens"         "PRuzamasi"
[21] "teleNormal"        "kardiyomegali"     "enfeksiyon"        "kansizlik"         "ilkatakAort"
[26] "ilkatakMitral"     "takipAort"         "takipMitral"       "benzatinPenisilin" "prokainPenisilin"
[31] "eritromisin"       "aspirin"           "kortizon"          "alerji"            "antibiyotik"
[36] "atesdusurucu"      "balgamsokucu"      "idrarsokturucu"    "kalpilaci"         "kalpyetmezligi"
[41] "kanilaci"          "kansulandirici"    "mideilaci"         "norolojikilac"     "plazma"
[46] "potasyumtakviyesi" "ozgecmis"          "ailehikaye"        "aortRep"           "mitralRep"
[51] "digerOp"           "sonkontrolYas"     "takip"             "karar"
```

**Şekil 3.5:** Analiz edilen kategorik veri setinin son halindeki nitelik alanları.

Analiz edilen nümerik veri setinin özeti ise Şekil 3.6'da görülmektedir.

```
> summary(tamveri)
    cins       hastaneSure          ay         mevsim      ilkatakYas    artrit     kardit     korea      ates
 erkek: 97   Min.   : 2.00   aralik :26   ilkbahar:53   Min.   : 4.00   var:141   var:186   var: 24   var: 27
 kadin:104   1st Qu.: 9.00   ocak   :22   kis      :67   1st Qu.: 9.00   yok: 60   yok: 15   yok:177   yok:174
             Median :14.00   nisan  :20   sonbahar:37   Median :11.00
             Mean   :14.49   haziran:19   yaz      :44   Mean   :10.46
             3rd Qu.:17.93   mayis  :18                 3rd Qu.:12.00
             Max.   :51.00   subat  :18                 Max.   :16.00
                             (Other):78
 atralji   akutfazArtisi    lokosit           Hb             HTC             CRP              sedim
 var: 18   var:162       Min.   : 4300   Min.   : 7.40   Min.   :11.00   Min.   :  0.00   Min.   :  4.00
 yok:183   yok: 39       1st Qu.: 8700   1st Qu.:10.06   1st Qu.:31.50   1st Qu.:  9.60   1st Qu.: 90.00
                         Median :10743   Median :10.69   Median :32.76   Median : 22.93   Median :110.00
                         Mean   :10980   Mean   :10.85   Mean   :33.41   Mean   : 29.89   Mean   : 98.86
                         3rd Qu.:12500   3rd Qu.:11.50   3rd Qu.:35.30   3rd Qu.: 32.99   3rd Qu.:120.00
                         Max.   :26000   Max.   :17.70   Max.   :53.70   Max.   :211.00   Max.   :169.00

 streptokokKaniti ARAoykusu          rekurrens    PRuzamasi teleNormal kardiyomegali enfeksiyon kansizlik
 var:182          var: 41   rekurrens     : 35   var: 62   0: 24      0:194         0:197      0:199
 yok: 19          yok:160   rekurrensdegil:166   yok:139   1:177      1:  7         1:  4      1:  2


   ilkatakAort   ilkatakMitral    takipAort    takipMitral benzatinPenisilin prokainPenisilin eritromisin aspirin
 hafif   :135   hafif   :96    hafif   :31   hafif   :61   0:  3             0: 68            0:199       0: 32
 minimal : 28   minimal :20    minimal :40   minimal :73   1:198             1:133            1:  2       1:169
 orta    : 22   orta    :50    orta    :49   orta    :12
 siddetli:  5   siddetli:26    siddetli:25   siddetli:11
 yok     : 11   yok     : 9    yok     :56   yok     :44


 kortizon alerji  antibiyotik atesdusurucu balgamsokucu idrarsokturucu kalpilaci kalpyetmezligi kanilaci
 0: 52    0:200   0:186       0:199        0:200        0:197          0:200     0:138          0:200
 1:149    1:  1   1: 15       1:  2        1:  1        1:  4          1:  1     1: 63          1:  1



 kansulandirici mideilaci norolojikilac plazma  potasyumtakviyesi ozgecmis  ailehikaye aortRep mitralRep digerOp
 0:200          0:200     0:187         0:200   0:199            var:  9   var: 11    0:194   0:200     0:200
 1:  1          1:  1     1: 14         1:  1   1:  2            yok:192   yok:190    1:  7   1:  1     1:  1



 sonkontrolYas                  takip           karar
 Min.   : 4.28   eriskinegonderilmis: 46   degismemis: 30
 1st Qu.:12.74   takipsiz           :155   gerilemis :104
 Median :15.81                             ilerlemis : 21
 Mean   :15.23                             iyilesmis : 46
 3rd Qu.:17.98
 Max.   :25.01
```

**Şekil 3.6:** Analiz edilen nümerik veri setinin özeti.

Analiz edilen kategorik veri setinin özeti ise Şekil 3.7'de görülmektedir.

```
> summary(tamveri)
    cins       hastaneSure          ay          mevsim      ilkatakYas     artrit     kardit     korea      ates
 erkek: 97   Min.   : 2.00   aralik :26   ilkbahar:53   Min.   : 4.00   var:141   var:186   var: 24   var: 27
 kadin:104   1st Qu.: 9.00   ocak   :22   kis      :67   1st Qu.: 9.00   yok: 60   yok: 15   yok:177   yok:174
             Median :14.00   nisan  :20   sonbahar:37   Median :11.00
             Mean   :14.60   haziran:18   yaz      :44   Mean   :10.46
             3rd Qu.:18.24   mayis  :18                 3rd Qu.:12.00
             Max.   :51.00   subat  :18                 Max.   :16.00
                             (Other):79
 atralji   akutfazArtisi lokosit_grup   hb_grup      HTC_grup     CRP_grup    sedim_grup  streptokokKaniti
 var: 19   var:162       dusuk :39    dusuk :146   dusuk :142   normal: 44   normal: 42   var:182
 yok:182   yok: 39       normal:95    normal: 47   normal: 58   yuksek:157   yuksek:159   yok: 19
                        yuksek:67    yuksek:  8   yuksek:  1


 ARAoykusu          rekurrens   PRuzamasi teleNormal kardiyomegali enfeksiyon kansizlik   ilkatakAort
 var: 42   rekurrens     : 35   var: 61   0: 24      0:194         0:197      0:199     hafif   :130
 yok:159   rekurrensdegil:166   yok:140   1:177      1:  7         1:  4      1:  2     minimal : 32
                                                                                       orta    : 23
                                                                                       siddetli:  5
                                                                                       yok     : 11

  ilkatakMitral    takipAort    takipMitral benzatinPenisilin prokainPenisilin eritromisin aspirin kortizon
 hafif   :97    hafif   :30   hafif   :61   0:  3            0: 68            0:199       0: 33   0: 52
 minimal :19    minimal :40   minimal :73   1:198            1:133            1:  2       1:168   1:149
 orta    :50    orta    :56   orta    :13
 siddetli:26    siddetli:19   siddetli:10
 yok     : 9    yok     :56   yok     :44


 alerji  antibiyotik atesdusurucu balgamsokucu idrarsokturucu kalpilaci kalpyetmezligi kanilaci kansulandirici
 0:200   0:186       0:199        0:200        0:197          0:200     0:138          0:200    0:200
 1:  1   1: 15       1:  2        1:  1        1:  4          1:  1     1: 63          1:  1    1:  1


 mideilaci norolojikilac plazma  potasyumtakviyesi ozgecmis  ailehikaye aortRep mitralRep digerOp sonkontrolYas
 0:200     0:187         0:200   0:199             var:  9   var: 11    0:194   0:200     0:200   Min.    : 4.28
 1:  1     1: 14         1:  1   1:  2             yok:192   yok:190    1:  7   1:  1     1:  1   1st Qu.:12.74
                                                                                                 Median :15.78
                                                                                                 Mean   :15.21
                                                                                                 3rd Qu.:17.98
                                                                                                 Max.   :25.01

                takip           karar
 eriskinegonderilmis: 46   degismemis: 30
 takipsiz           :155   gerilemis :104
                           ilerlemis : 21
                           iyilesmis : 46
```

**Şekil 3.7:** Analiz edilen kategorik veri setinin özeti.

## 3.3. DÖNÜŞTÜRME

Veri setinde, nümerik olan kan değerlerinin kategorik değerlere dönüştürülmesi yoluyla ayrıklaştırma yapılmıştır. Kategorizasyon yapılırken uzman görüşünden yararlanılmıştır.

## 3.4. VERİ MADENCİLİĞİ (MODELLEME)

### 3.4.1. Kullanılan Araçlar

Analizler, R programlama dilinde yazılan kodlar ile RStudio (Version 1.0.136) programı kullanılarak yapılmıştır (EK.3). Verinin hazırlanması için Microsoft Excel 2016 kullanılmıştır.

### 3.4.2. Algoritmalar

Analizler için sade Bayes sınıflandırıcı, karar ağaçları (CART, C4.5, C5.0, C5.0 boosted) ve rastgele orman algoritmaları kullanılmıştır. Analizlerde kullanılan paketler ve fonksiyonlara ait bilgiler Tablo 3.4'te verilmiştir.

**Tablo 3.4:** Analizlerde kullanılan paket ve fonksiyon bilgileri.

| Kullanılan Paket | Kullanılan Fonksiyonlar | Kullanım Amacı | Kaynak |
|---|---|---|---|
| readxl | readxl | Excel dosyasını okuma | (Wickham ve Bryan, 2017) |
| DMwR | knnImputation | Eksik veri tamamlama | (Torgo, 2010) |
| | lofactor | Uç değerlerin tespiti | |
| caret | sample | Bootstrap | (Kuhn, 2017) |
| | createDataPartition | Holdout | |
| | confusionMatrix | Karışıklık matrisi oluşturma | |
| stats | predict | Tahmin etme/Modelin kurulması | (R Core Team, 2017) |
| TunePareto | generateCVRuns | Çapraz geçerleme | (Müssel ve diğ., 2012) |
| e1071 | naiveBayes | Sade Bayes | (Meyer ve diğ., 2017) |
| rpart | rpart | CART | (Therneau ve diğ., 2017) |
| rpart.plot | rpart.plot | Karar ağacını çizdirme | (Milborrow, 2017) |
| RWeka | J48 | C4.5 | (Witten ve Frank, 2005; Hornik ve diğ., 2009) |
| C50 | C5.0 | C5.0 | (Kuhn ve diğ., 2015) |
| randomForest | randomForest | Rastgele orman | (Liaw ve Wiener, 2002) |

### 3.4.3. Model Performans Değerlendirme Yöntemleri

Model performans değerlendirme yöntemlerinde farklı yöntemler ile analizler gerçekleştirilmiş olup farklı algoritmalara ait performans sonuçları elde edilmiştir.

- Çapraz geçerleme yöntemi 5 kat, 10 kat, 15 kat olarak uygulanmıştır.
- Holdout yöntemi %80 eğitim %20 test, %70 eğitim %30 test, %60 eğitim %40 test olarak uygulanmıştır.
- Bootstrap yöntemi 50, 100, 200 örnek eğitim veri setinde kullanılmak üzere uygulanmıştır.

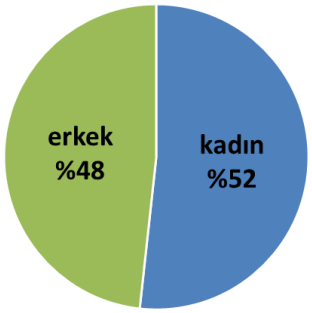
### 3.5. YORUMLAMA

Farklı model performans değerlendirme yöntemi ve farklı algoritmalara göre yapılan analizlere ait sonuçlar ve model performans değerlendirme ölçütlerine dayanarak yapılan yorumlara TARTIŞMA VE SONUÇ kısmında yer verilmiştir.

# 4. BULGULAR

Bu bölümde öncelikle veri setiyle ilgili temel birtakım istatistikler ardından veri madenciliği analizlerinin sonuçları verilmiştir.

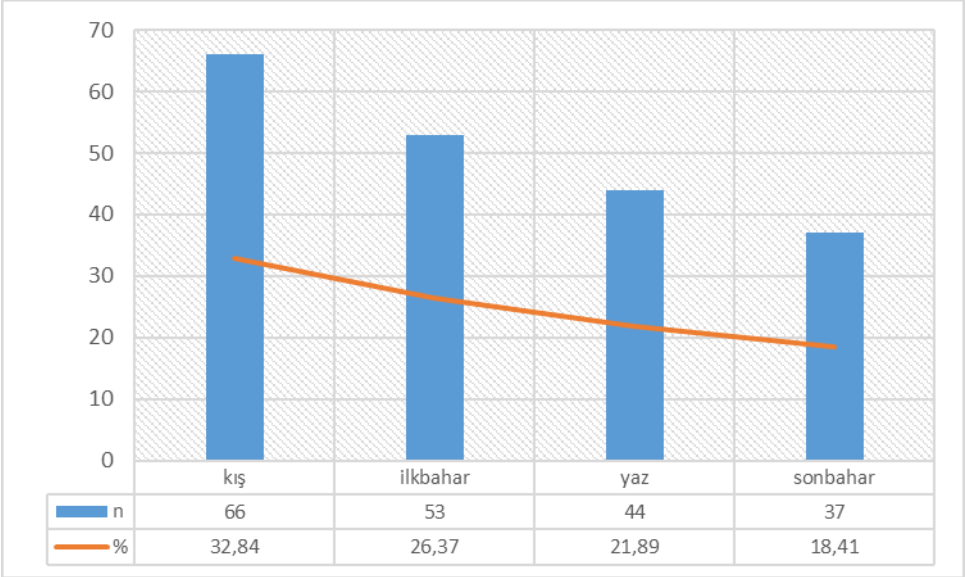
Veri setindeki cinsiyet dağılımlarına bakıldığında 104 hastanın kadın, 97 hastanın erkek olduğu görülmektedir. Hastalara ait cinsiyet dağılımı ve yüzde değerleri Şekil 4.1'de verilmiştir.




**Şekil 4.1:** Hastaların cinsiyet dağılımı.

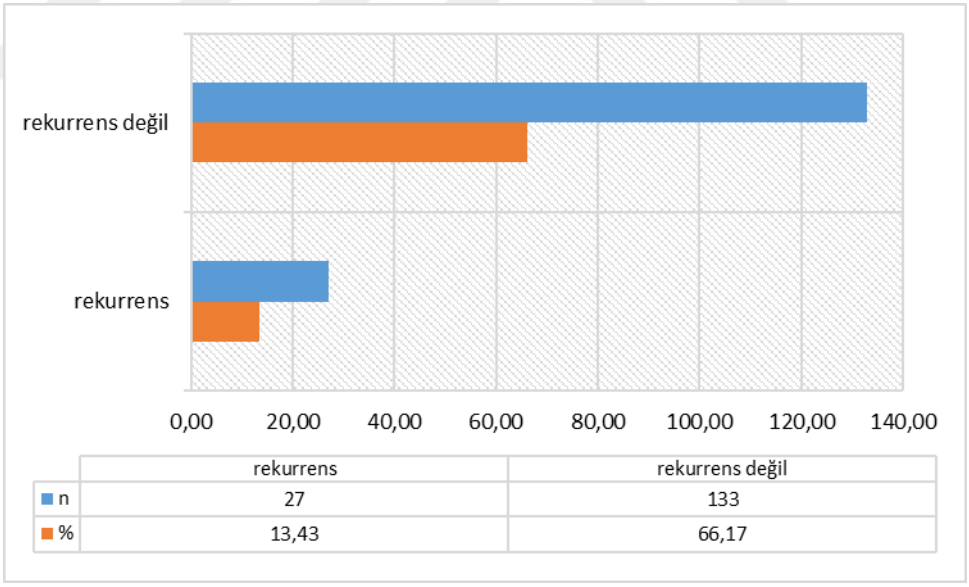
Hastaların ilk atak geçirme yaşlarının ortalaması 10,4'tür; yaş aralığı 4-16 arasında değişmektedir. Hastaların son kontrole gelme yaşlarının ortalaması ise 15,2'dir; yaş aralığı 4-25 arasında değişmektedir. Hastanede ortalama yatış süresi ise 14,4 gündür; hastanede yatış süresi 2-51 gün arasında değişmektedir.

Orijinal veri setindeki ilk atak geçirme tarihleri temel alınarak ilk atağın geçirildiği mevsimlere göre oluşturulan sütundan elde edilen, ilk atağın geçirildiği mevsimlerin dağılımı Şekil 4.2'de görülmektedir. Buna göre ilk ataklar en çok kış aylarında ikinci olarak da ilkbahar aylarında ortaya çıkmaktadır. Kişi sayılarına bakıldığında 66 kişi kış mevsiminde, 53 kişi ilkbaharda, 44 kişi yaz mevsiminde, 37 kişi ise sonbaharda ilk kez atak geçirmiştir. İlk atak geçirme tarihi belirtilmediği için bir kişiye ait mevsim verisi boş bırakılmış ve grafiğe eklenmemiştir.



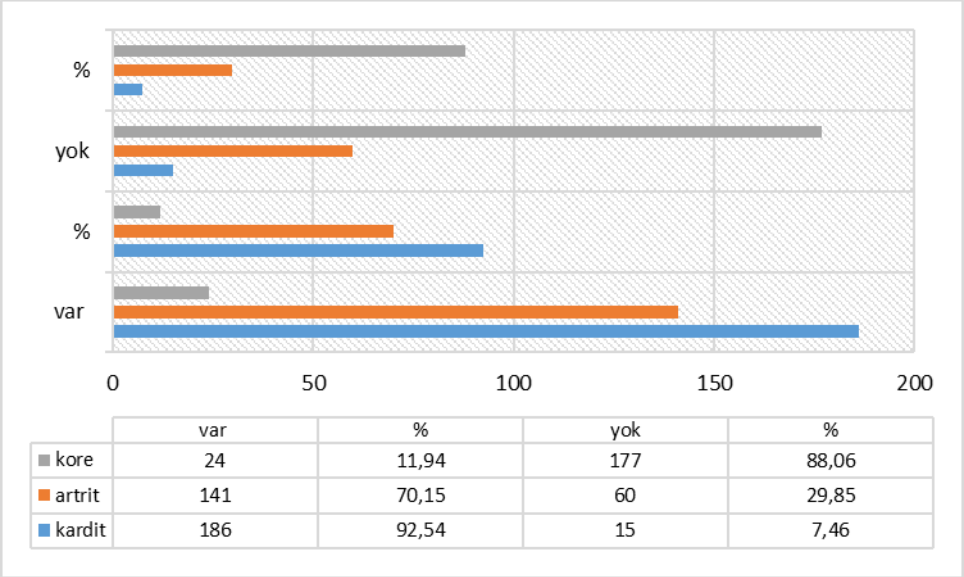

**Şekil 4.2:** İlk atakların mevsimsel dağılımı.

Veri setindeki hastaların % 13,43'ü hastalığı birden fazla kez geçirmiştir (Şekil 4.3). 41 kişiye ait veri, ilgili sütunda yer almadığından grafiğe dâhil edilmemiştir.




**Şekil 4.3:** Hastalığın tekrarlama durumu.

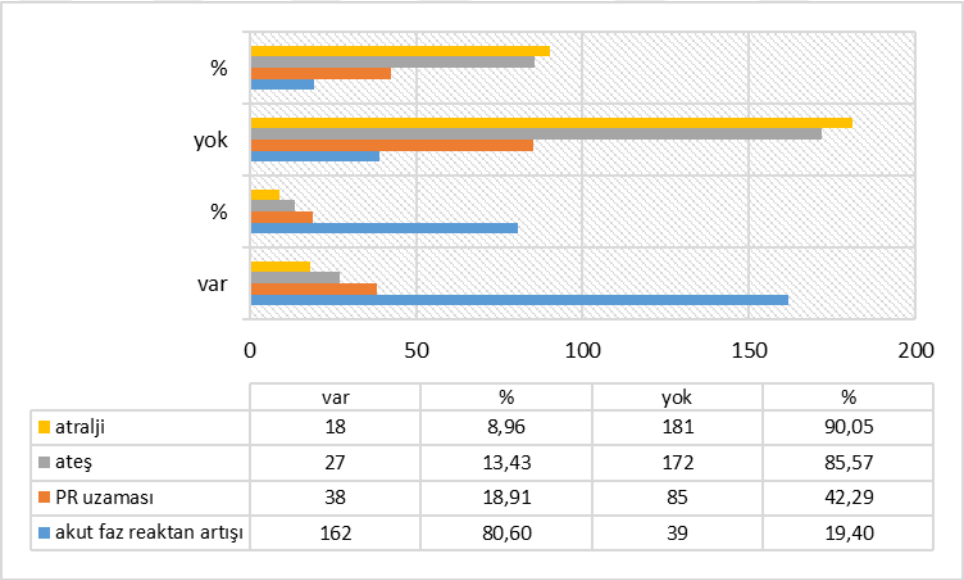
Hastalardaki majör kriterlerin görülme oranlarına bakıldığında (Şekil 4.4) hastalarda en yüksek oranla kardit ardından, artrit ve sonrasında ise kore olduğu görülmüştür. Majör Bulgular başlığı altında bahsedilen eritema marginatum ve cilt altı nodülleri olan hasta kaydı bulunmamaktadır.




| | var | % | yok | % |
|---|---|---|---|---|
| kore | 24 | 11,94 | 177 | 88,06 |
| artrit | 141 | 70,15 | 60 | 29,85 |
| kardit | 186 | 92,54 | 15 | 7,46 |

**Şekil 4.4:** Majör kriterlerin dağılımı.

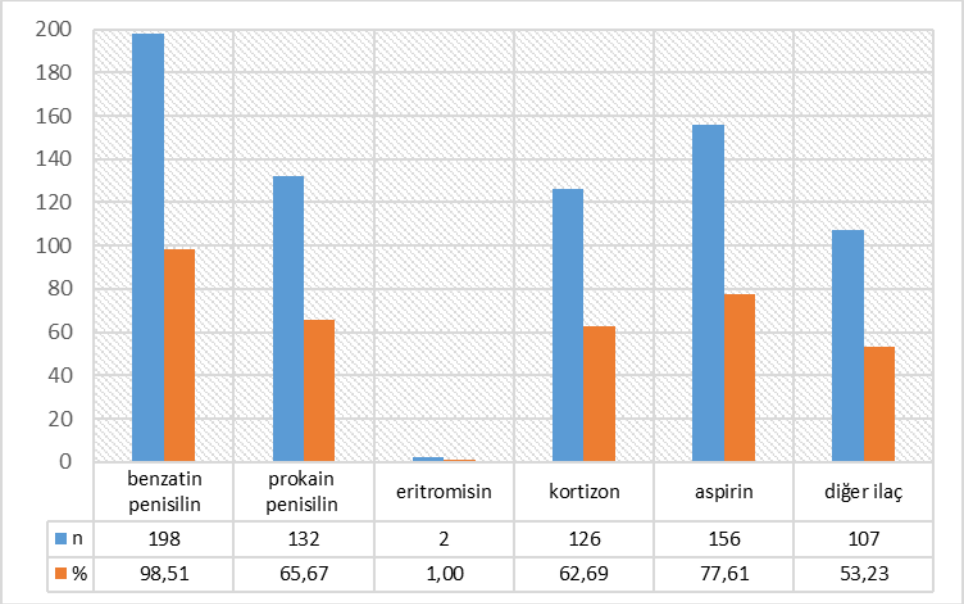
Minör kriterlere ait dağılım Şekil 4.5'te verilmiştir. Hastalardaki minör kriterlerin görülme oranlarına bakıldığında hasta sayısına göre azdan çoğa sırasıyla; atralji, ateş, PR uzaması ve akut faz reaktanlarında artışı bulguları görülmektedir.




| | var | % | yok | % |
|---|---|---|---|---|
| atralji | 18 | 8,96 | 181 | 90,05 |
| ateş | 27 | 13,43 | 172 | 85,57 |
| PR uzaması | 38 | 18,91 | 85 | 42,29 |
| akut faz reaktan artışı | 162 | 80,60 | 39 | 19,40 |

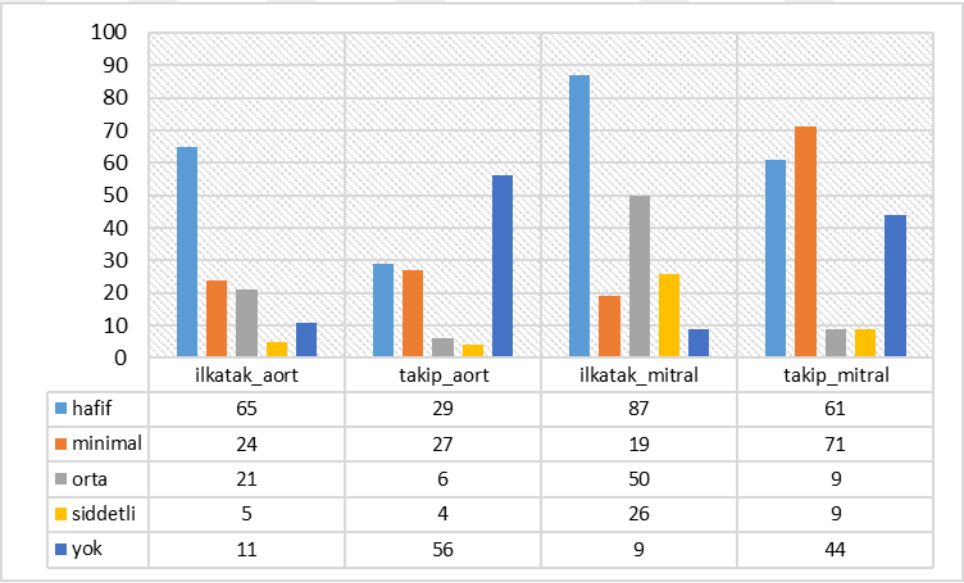
**Şekil 4.5:** Minör kriterlerin dağılımı.

Hastaların ilaç kullanımlarına ait dağılım ise Şekil 4.6'da verilmiştir. Hastaların %98,5'i benzatin penisilin, %77,61'i aspirin, %65,67'sinin prokain penisilin ve %62,68'inin kortizon kullanmıştır. Eritromisin ise sadece iki hasta (% 1) tarafından kullanılmıştır. Bunlar dışında kalan ağrı kesici, ateş düşürücü, alerji ilacı gibi diğer ilaçların kullanım oranı ise %53,23'tür.




**Şekil 4.6:** İlaç kullanımlarının dağılımı.

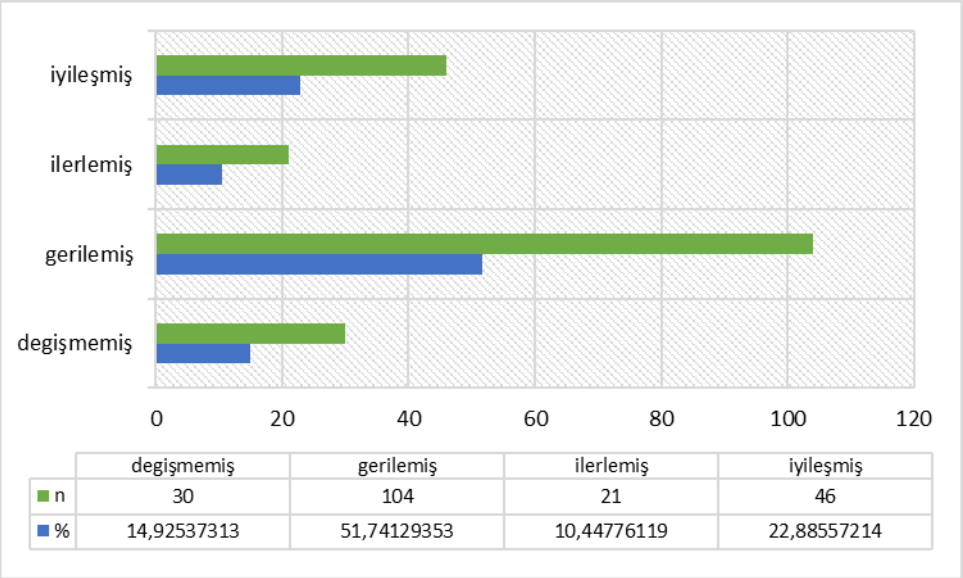
Hastalarla ilgili durum değerlendirmesinin yapılmasında yararlanılan, ilk atak ve takipteki mitral kapak ve aort kapağındaki tutulum derecelerini gösteren değerler Şekil 4.7'de verilmiştir.




**Şekil 4.7:** İlk atak ve takipte kapaklardaki tutulumların dereceleri.

Hastaların takiplerinde kalp kapak tutulumunun değerlendirilmesi (karar sütunundaki hasta durumlarının dağılımı) Şekil 4.8'de verilmiştir. 104 hasta ile en fazla hastanın "gerilemiş" sınıfında olduğu görülmektedir. Bu sınıfı çoktan aza doğru sırasıyla; 46 kişi

ile “iyileşmiş”, 30 kişi ile “değişmemiş” ve 21 kişi ile “ilerlemiş” sınıfları takip etmektedir.




| | degişmemiş | gerilemiş | ilerlemiş | iyileşmiş |
|---|---|---|---|---|
| n | 30 | 104 | 21 | 46 |
| % | 14,92537313 | 51,74129353 | 10,44776119 | 22,88557214 |

**Şekil 4.8:** Hastaların takiplerinde kalp kapak tutulumunun değerlendirilmesi (karar sütunundaki hasta durumlarının dağılımı).

Veri madenciliği analiz sonuçları ise hem nümerik (“lokosit”, “Hb”, “HTC”, “CRP”, “Sedim”) hem de kategorik (“lokosit_grup”, “Hb_grup”, “HTC_grup”, “CRP_grup”, “Sedim_grup”) değerleri temel alan “nümerik” ve “kategorik” veri setlerine göre Tablo 4.1’de verilmiştir. Tüm modellerden elde edilen sonuçlara göre en iyi performansı veren model (doğruluk değeri en yüksek olan), nümerik veri setinde holdout (%80 eğitim, %20 test) yöntemini kullanan CART modeli olmuştur (EK.4). Bu model ile doğruluk değeri 0,87 olarak elde edilmiştir .

Veri setleri temelinde bakıldığında her iki veri setinde de en yüksek doğruluk değerlerinin CART modeli ile elde edildiği görülmektedir. CART modeli, nümerik veri setinde ve tüm modeller bazında en yüksek olan 0,87 doğruluk değerini verirken; kategorik veri setinde ise 0,85 doğruluk değeri vermiştir.

**Tablo 4.1:** Modellerin doğruluk ve hata değerleri.

| NÜMERİK VERİ SETİ | | hold out 80/20 | hold out 70/30 | hold out 60/40 | CV 5 kat | CV 10 kat | CV 15 kat | bootstrap 50 | bootstrap 100 | bootstrap 200 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sade Bayes | Doğruluk | 0,74 | 0,64 | 0,68 | 0,65 | 0,68 | 0,69 | 0,56 | 0,53 | 0,61 |
| | Hata | 0,26 | 0,36 | 0,32 | 0,35 | 0,32 | 0,31 | 0,44 | 0,47 | 0,39 |
| CART | Doğruluk | 0,87 | 0,66 | 0,71 | 0,74 | 0,75 | 0,77 | 0,52 | 0,55 | 0,73 |
| | Hata | 0,13 | 0,34 | 0,29 | 0,26 | 0,25 | 0,23 | 0,48 | 0,45 | 0,27 |
| C4.5 | Doğruluk | 0,77 | 0,75 | 0,77 | 0,72 | 0,72 | 0,73 | 0,56 | 0,69 | 0,66 |
| | Hata | 0,23 | 0,25 | 0,23 | 0,28 | 0,28 | 0,27 | 0,44 | 0,31 | 0,34 |
| C5.0 | Doğruluk | 0,77 | 0,63 | 0,77 | 0,72 | 0,71 | 0,72 | 0,55 | 0,54 | 0,62 |
| | Hata | 0,23 | 0,37 | 0,23 | 0,28 | 0,29 | 0,28 | 0,45 | 0,46 | 0,38 |
| C5.0 boosted | Doğruluk | 0,85 | 0,66 | 0,82 | 0,75 | 0,73 | 0,76 | 0,59 | 0,61 | 0,69 |
| | Hata | 0,15 | 0,34 | 0,18 | 0,25 | 0,27 | 0,24 | 0,41 | 0,39 | 0,31 |
| Rastgele Orman | Doğruluk | 0,79 | 0,68 | 0,68 | 0,71 | 0,71 | 0,71 | 0,57 | 0,59 | 0,66 |
| | Hata | 0,21 | 0,32 | 0,32 | 0,29 | 0,29 | 0,29 | 0,43 | 0,41 | 0,34 |
| KATEGORİK VERİ SETİ | | hold out 80/20 | hold out 70/30 | hold out 60/40 | CV 5 kat | CV 10 kat | CV 15 kat | bootstrap 50 | bootstrap 100 | bootstrap 200 |
| Sade Bayes | Doğruluk | 0,69 | 0,64 | 0,66 | 0,65 | 0,67 | 0,65 | 0,55 | 0,56 | 0,64 |
| | Hata | 0,31 | 0,36 | 0,34 | 0,35 | 0,33 | 0,35 | 0,45 | 0,44 | 0,36 |
| CART | Doğruluk | 0,85 | 0,76 | 0,72 | 0,73 | 0,73 | 0,75 | 0,50 | 0,57 | 0,72 |
| | Hata | 0,15 | 0,24 | 0,28 | 0,27 | 0,27 | 0,25 | 0,50 | 0,43 | 0,28 |
| C4.5 | Doğruluk | 0,74 | 0,73 | 0,77 | 0,72 | 0,73 | 0,74 | 0,62 | 0,70 | 0,65 |
| | Hata | 0,26 | 0,27 | 0,23 | 0,28 | 0,27 | 0,26 | 0,38 | 0,30 | 0,35 |
| C5.0 | Doğruluk | 0,72 | 0,64 | 0,75 | 0,70 | 0,74 | 0,76 | 0,49 | 0,70 | 0,61 |
| | Hata | 0,28 | 0,36 | 0,25 | 0,30 | 0,26 | 0,24 | 0,51 | 0,30 | 0,39 |
| C5.0 boosted | Doğruluk | 0,77 | 0,68 | 0,73 | 0,76 | 0,74 | 0,75 | 0,61 | 0,60 | 0,73 |
| | Hata | 0,23 | 0,32 | 0,27 | 0,24 | 0,26 | 0,25 | 0,39 | 0,40 | 0,27 |
| Rastgele Orman | Doğruluk | 0,74 | 0,61 | 0,70 | 0,69 | 0,71 | 0,70 | 0,56 | 0,58 | 0,65 |
| | Hata | 0,26 | 0,39 | 0,30 | 0,31 | 0,29 | 0,30 | 0,44 | 0,42 | 0,35 |

En iyi performansı gösteren CART modeline ait karmaşıklık matrisi Şekil 4.9'da verilmiştir. Bu tabloya göre modelin yaptığı tahmini sınıflar ile gerçek sınıfların dağılımı görülmektedir.

```
> modelperformansi
Confusion Matrix and Statistics

                 gercek_siniflar
tahmini_siniflar degismemis gerilemis ilerlemis iyilesmis
      degismemis           5         0         0         0
      gerilemis            1        17         1         0
      ilerlemis            0         2         3         0
      iyilesmis            0         1         0         9
```

**Şekil 4.9:** CART modeline ait karmaşıklık matrisi.

Tablo 4.2'de karar ağacı oluşturulurken niteliklerin, kullanımlarına göre önem sırası verilmiştir. Bu doğrultuda "takipmitral", "takipaort", "ay", "ilkatakmitral" nitelikleri ilk sırada yer almaktadır. Karar verirken ilk bakılması gereken alanlar bu şekilde önceliklendirilmiştir.

**Tablo 4.2:** CART modeline göre niteliklerin önem sırası ve derecesi.

| Nitelik İsimleri* | Önem Derecesi |
|---|---|
| takipmitral | 35 |
| takipaort | 11 |
| ay | 9 |
| ilkatakmitral | 9 |
| lokosit | 5 |
| sedim | 4 |
| ilkatakaort | 4 |
| sonkontrolYas | 3 |
| mevsim | 3 |
| hastaneSure | 3 |
| HTC | 3 |
| kalpyetmezligi | 3 |
| atralji | 2 |
| aortRep | 2 |
| kansizlik | 1 |

*Nitelik isimlerinin açıklamaları Tablo 3.3'de verilmiştir.

CART modeline ait ağaç yapısı Şekil 4.10'da görülmektedir. Karar ağacının en alttaki yaprağında hastaların sınıflara ait olma yüzdeleri görülmektedir. Görülen dört adet yüzde değeri sırasıyla "değişmemiş", "gerilemiş", "ilerlemiş", "iyileşmiş" sınıflarına aittir. En yüksek değer hangisi ise kurallara göre o sınıfa ait olma olasılığı daha yüksektir. Modelden elde edilen 8 kural aşağıda listelenmiştir:

**Kural 1:** Eğer takipte mitral kapakta hafif, minimal, orta veya şiddetli derecede tutulum varsa ve takipte mitral kapakta hafif veya minimal derecede tutulum varsa ve ilk atakta mitral kapakta hafif veya minimal derecede tutulum varsa ve takipte mitral kapakta hafif derecede tutulum varsa hasta %79 olasılıkla "değişmemiş" sınıfına aittir.

**Kural 2:** Eğer takipte mitral kapakta hafif, minimal, orta veya şiddetli derecede tutulum varsa ve takipte mitral kapakta hafif veya minimal derecede tutulum varsa ve ilk atakta mitral kapakta hafif veya minimal derecede tutulum varsa ve takipte mitral kapakta

minimal derecede tutulum ve ilk atak ağustos, aralık, ekim, mayıs veya temmuz aylarında geçirilmişse hasta %87 olasılıkla “gerilemiş” sınıfına aittir.

**Kural 3:** Eğer takipte mitral kapakta hafif, minimal, orta veya şiddetli derecede tutulum varsa ve takipte mitral kapakta hafif veya minimal derecede tutulum varsa ve ilk atakta mitral kapakta hafif veya minimal derecede tutulum varsa ve takipte mitral kapakta minimal derecede tutulum varsa ve ilk atak eylül, haziran, kasım, mart, nisan, ocak veya şubat aylarında geçirilmişse ve takipte aort kapağında hafif, minimal veya orta derecede tutulum varsa hasta %85 olasılıkla “gerilemiş” sınıfına aittir.

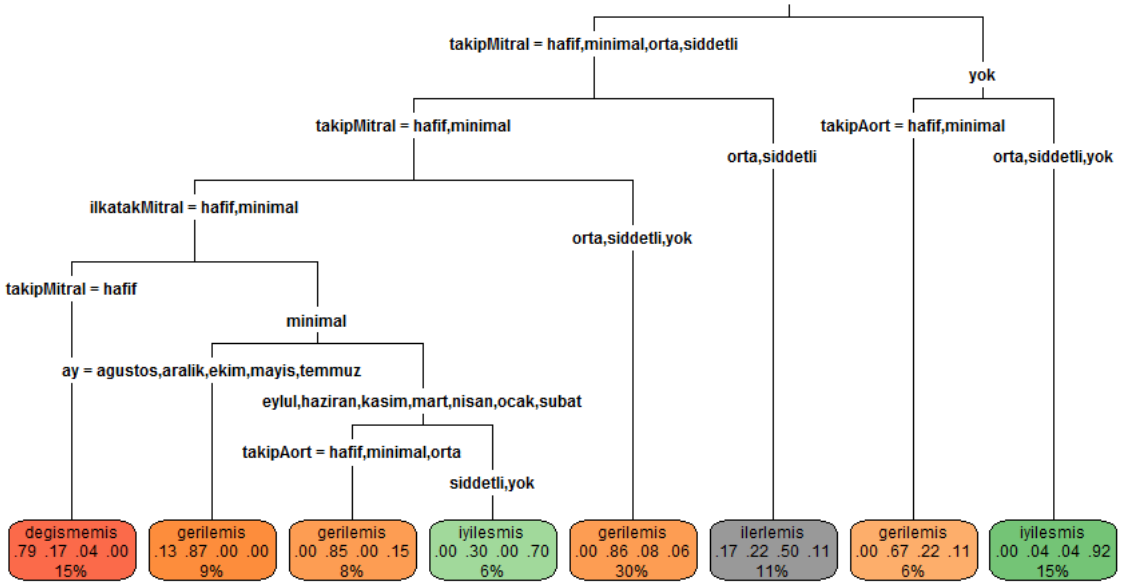
**Kural 4:** Eğer takipte mitral kapakta hafif, minimal, orta veya şiddetli derecede tutulum varsa ve takipte mitral kapakta hafif veya minimal derecede tutulum varsa ve ilk atakta mitral kapakta hafif veya minimal derecede tutulum varsa ve takipte mitral kapakta minimal derecede tutulum ve ilk atak eylül, haziran, kasım, mart, nisan, ocak veya şubat aylarında geçirilmişse ve takipte aort kapağında şiddetli derecede tutulum varsa veya tutulum yoksa hasta %70 olasılıkla “iyileşmiş” sınıfına aittir.

**Kural 5:** Eğer takipte mitral kapakta hafif, minimal, orta veya şiddetli derecede tutulum varsa ve takipte mitral kapakta hafif veya minimal derecede tutulum varsa ve ilk atakta mitral kapakta orta, şiddetli derecede tutulum varsa veya tutulum yoksa hasta %86 olasılıkla “gerilemiş” sınıfına aittir.

**Kural 6:** Eğer takipte mitral kapakta hafif, minimal, orta veya şiddetli derecede tutulum varsa ve takipte mitral kapakta orta veya şiddetli derecede tutulum varsa hasta %50 olasılıkla “ilerlemiş” sınıfına aittir.

**Kural 7:** Eğer takipte mitral kapakta tutulum yoksa ve takipte aort kapağında hafif veya minimal derecede tutulum varsa hasta %67 olasılıkla “gerilemiş” sınıfına aittir.

**Kural 8:** Eğer takipte mitral kapakta tutulum yoksa ve takipte aort kapağında orta, şiddetli derecede tutulum varsa veya tutulum yoksa hasta %92 olasılıkla “iyileşmiş” sınıfına aittir.




**Şekil 4.10:** CART modelinden elde edilen ağaç yapısı.

CART modelinin sonuçlarına göre sınıf bazında model performans değerlendirme ölçütleri incelendiğinde elde edilen model performans değerlendirme ölçütleri Tablo 4.3'te verilmiştir. Tüm sınıflar arasında, modelin en yüksek duyarlılık değerini (1,00) "iyileşmiş" sınıfı için; en düşük duyarlılık değerini ise (0,75) "ilerlemiş" sınıfı için verdiği görülmektedir. Belirleyicilik değerlerine bakıldığında ise model, en yüksek değeri (1,00) "değişmemiş" sınıfı için, en düşük değeri ise (0,89) "gerilemiş" sınıfı için vermiştir.

| | | **Model Performans Değerlendirme Ölçütleri** | | | | |
|---|---|---|---|---|---|---|
| | | **Duyarlılık** | **Belirleyicilik** | **Pozitif Öngörü Değeri** | **Negatif Öngörü Değeri** | **F-ölçütü** |
| **CART holdout80** | **sınıf: değişmemiş** | 0,83 | 1,00 | 1,00 | 0,97 | 0,91 |
| | **sınıf: gerilemiş** | 0,85 | 0,89 | 0,89 | 0,85 | 0,87 |
| | **sınıf: ilerlemiş** | 0,75 | 0,94 | 0,60 | 0,97 | 0,67 |
| | **sınıf: iyileşmiş** | 1,00 | 0,97 | 0,90 | 1,00 | 0,95 |

**Tablo 4.3:** CART modeli için sınıf bazında model performans değerlendirme ölçütleri.

# 5. TARTIŞMA VE SONUÇ

Ülkemizde akut romatizmal ateş (ARA) sıklıkla görülen bir hastalıktır. ARA ile ilgili literatür çalışmasında tıp çalışanları tarafından yapılan, istatistiksel yöntemlerin kullanıldığı birçok çalışmaya rastlanmıştır. Kalp ile ilgili veri setleri ile yapılmış çalışmalar olsa da bu hastalığı konu alan ve veri madenciliği yöntemlerinin kullanıldığı bir çalışmaya ise rastlanmamıştır. Bu tez çalışmasının amacı, veri madenciliği yöntemlerini kullanarak çocuk yaşta görülen akut romatizmal ateşin kalp üzerindeki etkilerinin analiz edilmesidir. Bu etkilerin belirlenmesi hastalığın seyrinin gözlemlenmesi açısından önemlidir. Çalışma hem Türkiye'de sıklıkla görülen bir hastalığın ele alınması hem de sağlık alanında veri madenciliği yöntemlerinin kullanılabileceğini göstermesi açısından önemli görülmektedir. Kullanılan veri seti ve analiz yöntemlerinin sentezlenmesi açılarından bu çalışmanın özgün bir çalışma olduğu düşünülmektedir.

Bu çalışma kapsamında, veri setinin yaş aralığı ve ortalaması literatür (Semizel ve diğ., 2005; Saltık, 2007; Düzgün, 2014) ile uyum göstermektedir.

ARA'nın kış ve ilkbahar aylarında daha sık görüldüğü literatürde belirtilmektedir (Köksal ve diğ., 2016). Tez kapsamında da ilk atakların mevsimsel olarak sırasıyla kış, ilkbahar, yaz ve sonbahar aylarında geçirildiği tespit edilmiştir. İlk atakların ARA ile ilişkisi açısından bu bulgunun literatür ile uyumlu olduğu düşünülmektedir.

ARA tanısı Jones Kriterleri (majör, minör ve destekleyici bulgular)'ne göre konmaktadır. Bu çalışmada majör kriterlerden, en çok kardit bulgusuna sahip hasta bulunduğu tespit edilmiştir (Şekil 4.4). Oysa Semizel ve diğ. (2005) ile Saltık (2007) tarafından en sık görülen bulgunun artrit olduğu belirtilmektedir. Bunun, veri setinin bir kardiyoloji merkezinden alınmış olmasından kaynaklandığı düşünülmektedir. Kardiyoloğu olmayan merkezler, şüpheli hastaları kardiyoloji merkezi olan hastanelere yönlendirdiğinden karditli hasta sayısı bu merkezde artritli hasta sayısının önüne geçmiş olabilir.

Hastaların ilk atak ve takipte kalp kapaklarındaki tutulumların derecelerine bakıldığında genel olarak takipteki hasta sayısının azaldığı görülmektedir (Şekil 4.7). İlk atakta aort ve mitral kapaklarda, orta ve şiddetli derecede tutulumu olan hasta sayısı takipte azalmış, minimal derecede tutulumu olan hasta sayısı ise takipte artmıştır. Bu artışın orta ve

şiddetli derecede tutulumu olan hastaların, takipte bu aşamaya gerilemesinden kaynaklanabileceği düşünülmektedir. Ayrıca tutulumu olmayan (yok kategorisi) hasta sayısının da takipte artış gösterdiği belirlenmiştir. Sonuç olarak takip edilen hastaların kalp kapaklarında iyileşme olduğu düşünülmektedir. Bu nedenle ARA'da takibin önemli olduğu ve hastaların düzenli olarak takip edilmesi gerektiği tarafımızdan vurgulanmaktadır.

Bu tez çalışması kapsamında, 5 algoritma (6 model) ve 9 performans değerlendirme yöntemi kullanılarak modeller oluşturulmuş ve bu modeller 2 veri setinde (nümerik ve kategorik) toplam 108 kere denenmiştir. Modellerin performansları hem model hem de sınıf bazında model performans değerlendirme ölçütlerine bakılarak incelenmiş ve ölçütlere göre karşılaştırma yapılarak en iyi performans gösteren algoritma veri setinin holdout yöntemi ile %80 eğitim %20 test oranında bölünmesi ile kurulan CART modeli (%87) olarak belirlenmiştir (Tablo 4.1). Nümerik ve kategorik veri setlerine göre performanslar karşılaştırıldığında her iki veri setinde de CART modelinin en yüksek doğruluk değerini verdiği görülmektedir.

Oluşturulan modellere göre en iyi performansı veren modelden elde edilen niteliklerin önem dereceleri göz önünde bulundurularak, önemli olduğu tespit edilen niteliklere (takipmitral, takipaort, ilkatakmitral, ay, lokosit, sedim, ilkatakaort, sonkontrolYas, mevsim, hastaneSure, HTC, kalpyetmezligi, atralji, aortRep, kansizlik) göre yeni bir veri seti oluşturulabilir veya sadece bu nitelikler ile analizler yapılabilir. Bu sayede veri setinde yer alıp karar ağacında kullanılmayan niteliklerin tetkik edilmesine gerek olmadığı düşünülmektedir (Tüm veri madenciliği çalışmalarında olduğu gibi alan uzmanı tarafından bu bilginin teyit edilmesi/denenmesi önerilmektedir). Örneğin akciğer filmi ile saptanan "telenormal" ve "kardiyomegali" nitelikleri ağaç yapısında bulunmamaktadır. Bu niteliklerin tespiti için yapılacak tetkik/çekim, ARA teşhis ve tedavisinde maddi kayba sebebiyet vermektedir. Ayrıca bu durum doktorlar, diğer sağlık çalışanları ve hasta açısından zaman kaybına da neden olabilmektedir. Bu tez çalışmasının sonuçlarına göre spesifik olarak ARA araştırılırken bu tetkikin yapılmaması özellikle çocukluk çağındaki hastaların film çekimi sebebiyle fazla ışına maruz kalmamasını sağlayacaktır. Bu gibi nedenlerle bundan sonra çocukluk çağındaki hastalarla yapılacak ARA ile ilgili çalışmalarda bu araştırmada belirlenen niteliklerin kullanılmasının maddi kayıpları ve iş

gücü kaybını azaltacağı, özellikle sağlık çalışanları için belirlenmiş olan muayene süresinin optimum şekilde kullanılmasına katkı sağlayacağı düşünülmektedir.

ARA'nın ilerleyişinin önüne geçilebilmesi yani çocukluk yaşta görülen ARA'nın kalp üzerindeki uzun dönemdeki olumsuz etkilerinin (ör. kalp yetmezliği, stenoz) önceden tahmin edilip azaltılabilmesi, mümkünse önlenebilmesi için çalışmada yer alan hastaların erişkin dönemde de takip edilmesi önem arz etmektedir. Tez çalışması kapsamında kişilere ait erişkin dönem verisine erişilemediği için bu analizler yapılamamıştır. Gelecekte bu konunun göz önünde bulundurularak veri toplanması önerilmektedir. Hastalığın uzun vadedeki takibi için veri toplanması ve bu verinin analiz edilmesi hastalığın hem ülke genelindeki indidansının belirlenmesi hem de etkilerinin azaltılması açısından büyük önem taşımaktadır.

ARA ile ilgili çalışmalarda, bu tez çalışmasında kullanılan veri madenciliği yöntemlerinin sağlık çalışanlarına/doktorlara bu alanın tıptaki veri ile ilgili fayda sağlayabileceğini gösterebilmesi umulmaktadır.

Veri madenciliği yöntemleri açısından bakıldığında, bu çalışmanın genişletilmesi için, kullanılmış olan veri setine yeni nitelik alanları eklenebileceği düşünülmektedir. Aynı veri seti ile destek vektör makineleri, genetik algoritmalar, k en yakın komşu, lojistik regresyon, yapay sinir ağları gibi farklı sınıflandırma algoritmaları denenebilir. Sınıflandırma algoritmalarının dışında kümeleme algoritmaları ile aynı veri seti analiz edilerek, örneklerin "karar" sınıfındaki dağılımları karşılaştırılabilir. Güncel hasta verisi toplanarak yeni verinin aynı model ile sınıflandırmada kullanılması ve yeni hastaların hangi sınıfa ait olabileceklerinin tahmin edilmesi sağlanabilir.

Çalışmanın geliştirilmesine ek olarak doktorlara destek sağlayabilecek bir web ara yüzünün oluşturulması planlanmaktadır. Böylece gerek deneyimli gerekse deneyimsiz doktorlar, daha önceki hasta verisine dayanarak oluşturulmuş bir karar destek sistemine danışabilecekler ve takipte hastanın durumunun ne olabileceğine dair öngörüde bulunan bir sistemden faydalanabileceklerdir. Benzer bir ara yüzün bir de Jones Kriterleri ile teşhis koymada yararlanmak için geliştirilmesi planlanmaktadır. Bu ara yüz sayesinde hastaya ait bulgulara göre, teşhis konmasına destek olacak bir sistem tasarlanacaktır. Bu sistem,

tek başına teşhis koymada kullanılmasa bile, doktorlara teşhisi sağlamlaştırma veya deneyimsiz doktorlara teşhisin güvenilirliğini arttırma açısından fayda sağlayabilir.

Bu çalışma kapsamında ele alınan veri seti ve analizlerin bu alanda çalışan araştırmacıların ilgisini çekmesi umulmaktadır. Bu tez çalışmasının farklı şekillerde genişletilebileceği düşünülmektedir. Çalışma hem veri madenciliği ile hem de hastalıkla/sağlık alanı ile ilgili olarak farklı yönlere doğru genişletilebilir görülmektedir. Veri madenciliği ile ilgili yapılabilecek geliştirmeler; hasta verisi ile yapılacak başka çalışmalarda farklı yöntemlerin denenmesini, çalışmaların teknik/analiz tarafının güçlendirilmesini sağlayacaktır. Sağlık çalışanları/doktorlar açısından bakıldığında ise çalışma; pratikteki uygulamalara katkı sağlayabilecek şekilde genişletilebilir ve veri madenciliği ile ilgili çalışmaların sağlık alanında yaygınlaşmasını sağlayabilir niteliktedir.

# KAYNAKLAR


Ackoff, R.L., 1989, From Data to Wisdom, *Journal of Applied Systems Analysis*, 16 (1), 3–9.

Aggarwal, C.C., 2015, An Introduction to Data Classification, *Data Classification: Algorithms and Applications*, CRC Press, Taylor & Francis Group, Boca Raton (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series), ISBN: 978-1-4665-8675-8, 2–36.

Ahsan, S. ve Shah, A., 2006, Data, Information, Knowledge, Wisdom: A Doubly Linked Chain, *The Proceedings of the 2006 International Conference on Information Knowledge Engineering*, 270–278.

Akalın, F., 2010, Akut Romatizmal Ateşte Yeni Klinik Görünümler, Yeni Görüşler, *Türkiye Klinikleri Journal of Pediatrical Sciences*, 6 (1), 9–19.

Akpınar, H., 2014, *Data: Veri Madenciliği Veri Analizi*, 1. baskı, Papatya Yayıncılık Eğitim, İstanbul, ISBN: 6054220816.

Alizadehsani, R., Habibi, J., Hosseini, M.J. ve diğ., 2013, A Data Mining Approach for Diagnosis of Coronary Artery, *Computer Methods and Programs in Biomedicine*, 111 (1), 52–61.

Atasoy, Y., 2015, *Veri Madenciliği Yöntemleri ile Ankilozan Spondilit Hastaliğinda Radyografik Progresyona Etkili Faktörlerin Analizi*, Yüksek Lisans Tezi, İstanbul Üniversitesi.

Balaban, M.E. ve Kartal, E., 2015, *Veri Madenciliği ve Makine Öğrenmesi*, 1. baskı, Çağlayan Kitapevi, İstanbul, ISBN: 978-975-436-089-9.

Bayes, M. ve Price, M., 1763, An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S., *Royal Society of London*, İngiltere.

Behrman, R.E., Kliegman, R. ve Jenson, H.B., 2000, *Nelson Textbook of Pediatrics*, 16. baskı, W.B. Saunders Company, ISBN: 978-0-7216-7767-5.

Benzreig, A.M.S., 2016, *Predicting Heart Diseases By Using Machine Learning Methods/Makine Öğrenme Yöntemleriyle Kalp Hastalıklarını Tahmin Etme*, Yüksek Lisans Tezi, Atılım Üniversitesi.

Bernstein, J.H., 2009, The Data-Information-Knowledge-Wisdom Hierarchy and It's Antithesis, *Proceedings from North American Symposium on Knowledge Organization*, 2 ,68-75.

Borges, L.C., Marques, V.M. ve Bernardino, J., 2013, Comparison of Data Mining Techniques and Tools for Data Classification, *Proceedings of the International Conference on Computer Science and Software Engineering*, ACM, 113–116.

Bostan, Ö.M. ve Çil, E., 2001, Bursa İlindeki Çocuklarda Akut Romatizmal Ateşin Değerlendirilmesi, *Türkiye Klinikleri Journal of Cardiology*, 14 (5), 276–281.

Boyraz, Ö.F., Seymen, V., Bozkurt, M.R. ve diğ., 2014, Makine Öğrenmesi Algoritmaları Kullanılarak Kalp Hastalığı Tespiti, *International Conference on Education in Mathematics, Science & Technology Proceeding Book*, Konya, ISBN: 978-605-61434-3-4, 1258–1262.

Bramer, M., 2007, *Principles of Data Mining*, 1. baskı, Springer Verlag, London (Undergraduate Topics in Computer Science), DOI: 10.1007/978-1-4471-4884-5, ISBN: 978-1-4471-4883-8.

Breiman, L., 2001, Random Forests, *Machine Learning*, 45 (1), 5–32.

Breiman, L., Friedman, J., Stone, C.J. ve diğ., 1984, *Classification and Regression Trees*, Taylor & Francis, ISBN: 978-0-412-04841-8.

Breunig, M.M., Kriegel, H.-P., Ng, R.T. ve diğ., 2000, LOF: identifying density-based local outliers, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ACM, New York, 93–104.

Bulut, F., 2016, *AdaBoost ile Kalp Krizi Risk Tespiti*, Celal Bayar Üniversitesi Fen Bilimleri Dergisi, 12 (3), DOI: 10.18466/cbayarfbe.280652, 459–472.

Buuren, S. van ve Groothuis-Oudshoorn, K., 2011, mice: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software*, 45 (3), 1–67.

Carapetis, J.R., McDonald, M. ve Wilson, N.J., 2005, Acute Rheumatic Fever, *The Lancet*, 366 (9480), DOI: 10.1016/S0140-6736(05)66874-2, 155–168.

Carapetis, J.R. ve Zühlke, L.J., 2011, Global Research Priorities in Rheumatic Fever and Rheumatic Heart Disease, *Annals of Pediatric Cardiology*, 4 (1), DOI: 10.4103/0974-2069.79616, 4–12.

Chandna, D., 2014, Diagnosis of Heart Disease Using Data Mining Algorithm, *International Journal of Computer Science and Information Technologies (IJCSIT)*, 5 (2), 1678–1680.

Chaurasia, V. ve Pal, S., 2013, Early Prediction of Heart Diseases Using Data Mining Techniques, *Caribbean Journal of Science and Technology*, 1, 208–217.

Chizi, B. ve Maimon, O., 2010, Dimension Reduction and Feature Selection, *Data Mining and Knowledge Discovery Handbook*, 2. baskı, Springer US, ABD, ISBN: 978-0-387-09822-7, 83–100.

Çağatay, D., Yıldız, F., Temel, Ö. ve diğ., 2010, Akut Romatizmal Ateş: Klinik Bir Değerlendirme, *Çocuk Dergisi*, 10 (4), DOI: 10.5222/j.child.2010.183, 183–189.

Dallar, Y., Şıklar, Z., Tanyer, G. ve diğ., 2002, Çocukluk Çağında Görülen Akut Romatizmal Ateş Olgularımızın Retrospektif Değerlendirilmesi Orijinal Araştırma, *Türk Pediatri Arşivi*, 37 (2).

Doğaner, A., 2015, *Enformasyon Sistemlerinde Saklı Markov Modeli ve Bayes Tabanlı Sınıflandırıcılar ile Bilgi Modellerinin Geliştirilmesi*, Doktora Tezi, Fırat Üniversitesi.

Düzgün, N., 2014, *Akut Romatizmal Ateş*, 314–321.

Efron, B. ve Tibshirani, R., 1993, *An Introduction to the Bootstrap*, Chapman & Hall, New York, ISBN: 978-0-412-04231-7.

Elmaz, F., 2014, *Kalp Krizi Riskinin Bir Veri Madenciliği Uygulaması ile Analizi*, Yüksek Lisans Tezi, Muğla Sıtkı Koçman Üniversitesi.

Erol, N., Türkmen, A., Özgüner, A. ve diğ., 2002, Akut Romatizmal Ateş: 60 Olgunun Retrospektif Değerlendirilmesi, *Kartal Eğitim ve Araştırma Hastanesi Tıp Dergisi*, 13 (3), 165–169.

Fayyad, U., Piatetsky-Shapiro, G. ve Smyth, P., 1996a, From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, 17 (3), 37–54.

Fayyad, U., Piatetsky-Shapiro, G. ve Smyth, P., 1996b, The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM*, 39 (11), 27–34.

Flach, P., 2012, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*, Cambridge University Press, ABD, ISBN: 978-1-107-09639-4.

Filzmoser, P. ve Gschwandtner, M., 2017, mvoutlier: Multivariate Outlier Detection Based on Robust Methods.

Gewitz, M.H., Baltimore, R.S., Tani, L.Y. ve diğ., 2015, Revision of the Jones Criteria for the Diagnosis of Acute Rheumatic Fever in the Era of Doppler Echocardiography, *Circulation*, DOI: 10.1161/CIR.0000000000000205, CIR.0000000000000205.

Gorunescu, F., 2011, *Data Mining - Concepts, Models and Techniques*, 1. baskı, Springer-Verlag Berlin Heidelberg, (Intelligent Systems Reference Library), ISBN: 978-3-642-19720-8.

Gözü Pirinççioğlu, A., Alyan, Ö., Kanğın, M. ve diğ., 2012, Akut Romatizmal Ateşli Çocuklarda Klinik ve Laboratuvar Bulguların Geriye Dönük Olarak İncelenmesi: Reaktivasyon ve Koruyucu Tedaviye Uyumunun Araştırılması, *Türk Kardiyoloji Derneği Arşivi*, 40 (5), DOI: 10.5505/tkda.2012.87405, 427–435.

Grzymala-Busse, J.W. ve Grzymala-Busse, W.J., 2010, Handling Missing Attribute Values, *Data Mining and Knowledge Discovery Handbook*, 2. baskı, Springer US, ABD, ISBN: 978-0-387-09822-7, 33–51.

Güler Eroğlu, A., 2016, Akut Romatizmal Ateş Tanısında Güncelleme: 2015 Jones Ölçütleri, *Türk Pediatri Arşivi*, DOI: 10.5152/TurkPediatriArs.2016.2397, 1–7.

Han, J., Kamber, M. ve Pei, J., 2012, *Data mining: concepts and techniques*, 3. baskı, Morgan Kaufman Publishers, ABD, ISBN: 978-0-12-381479-1.

Hand, D., Mannila, H. ve Smyth, P., 2001, *Principles of Data Mining*, MIT Press, ISBN: 0-262-08290-X.

Harrell, F.E., 2017, *Hmisc: Harrell Miscellaneous.*

Harrington, P., 2012, *Machine Learning in Action*, Manning Publications Co., ABD, ISBN: 978-1-61729-018-3.

Hornik, K., Buchta, C. ve Zeileis, A., 2009, Open-source Machine Learning: R meets Weka, *Computational Statistics*, 24 (2), DOI: 10.1007/s00180-008-0119-7, 225–232.

Hssina, B., Merbouha, A., Ezzikouri, H. ve diğ., 2002, A Comparative Study of Decision tree ID3 and C4.5, *International Journal of Advanced Computer Science and Applications* ,Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 13–19.

Jabbar, M.A., Deekshatulu, B.L. ve Chandra, P., 2013, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, *Procedia Technology*, 10, DOI: 10.1016/j.protcy.2013.12.340, 85–94.

Jones, T.D., 1944, The Diagnosis of Rheumatic Fever, *Journal of the American Medical Association*, 126 (8), DOI: 10.1001/jama.1944.02850430015005, 481–484.

Karaaslan, S., Oran, B., Reisli, I. ve diğ., 2000, Acute Rheumatic Fever in Konya, Turkey, *Pediatrics International*, 42 (1), 71–75.

Karcı, Z., 2017, *Lojistik Regresyon Modeli İle Elde Edilen Tahminlerin Roc Eğrisi Yardımıyla Değerlendirilmesi: Türkiye'de Hanehalkı Yoksulluğu Üzerine Bir Araştırma*, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi.

Kartal, E., 2015, *Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama*, Doktora Tezi, İstanbul Üniversitesi.

Kaya, K., 2016, *Destek Vektör Makineleri Yardımıyla Tüketici Kredilerinin Sınıflandırılması*, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi.

Koçoğlu, F.Ö., 2012, *Veri Madenciliği Sürecinde Veri Ayrıklaştırma Yöntemlerinin Karşılaştırılması ve Bir Uygulama*, Yüksek Lisans Tezi, İstanbul Üniveristesi.

Kohavi, R., 1995, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Stanford, CA, 1137–1145.

Kohavi, R. ve Provost, F., 1998, Glossary of Terms Journal of Machine Learning, *Machine Learning*, (30), 271–274.

Köksal, A.O., Gültekin Soylu, A. ve Özdemir, O., 2016, Akut Romatizmal Ateş, *Türkiye Çocuk Hastalıkları Dergisi*, 10 (4), DOI: 10.12956/tjpd.2015.182, 283–296.

Köktürk, F., 2012, k-*En Yakın Komşuluk, Yapay Sinir Ağları ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması*, Doktora Tezi, Bülent Ecevit Üniversitesi.

Kuhn, M., Weston, S., Coulter, N. ve diğ., 2015, *C50: C5.0 Decision Trees and Rule-Based Models*.

Kuhn, M., 2017, *caret: Classification and Regression Training.*

Kumari, M. ve Godara, S., 2011, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, *International Journal of Computer Science & Technology*, 2 (2), 304–308.

Lakshmi, K.R., Krishna, M.V. ve Kumar, S.P., 2013, Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability, *International Journal of Scientific and Research Publications*, 3 (6), 1–10.

Larose, D.T., 2005, Discovering *Knowledge in Data: An Introduction to Data Mining*, Wiley-Interscience, ABD, ISBN: 978-0-471-66657-8.

Liaw, A. ve Wiener, M., 2002, Classification and Regression by randomForest, *R News*, 2 (3), 18–22.

Maimon, O. ve Rokach, L., 2010, Introduction to Knowledge Discovery and Data Mining, *Data Mining and Knowledge Discovery Handbook,* 2. baskı, Springer US, ABD, ISBN: 978-0-387-09822-7, 1–15.

Masethe, H.D. ve Masethe, M.A., 2014, Prediction of Heart Disease using Classification Algorithms, *Proceedings of the World Congress on Engineering and Computer Science 2014*, San Francisco, ABD (World Congress on Engineering and Computer Science, WCECS 2014, San Francisco, USA, 22 - 24 October, 2014), ISBN: 978-988-19253-7-4.

Methaila, A., Kansal, P., Arya, H. ve diğ., 2014, Early Heart Disease Prediction Using Data Mining Techniques, *Academy & Industry Research Collaboration Center (AIRCC)*, DOI: 10.5121/csit.2014.4807, ISBN: 978-1-921987-10-6, 53–59.

Meyer, D., Dimitriadou, E., Hornik, K. ve diğ., 2017, e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

Mikut, R. ve Reischl, M., 2011, Data Mining Tools, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1 (5), DOI: 10.1002/widm.24, 431–443.

Milborrow, S., 2017, *rpart.plot: Plot "rpart" Models: An Enhanced Version of "plot.rpart"*.

Mosteller, F. ve Tukey, J., 1968, Data Analysis, Including Statistics, *Handbook of Social Psychology*, In: Lindzey, G. (ed.), Addison Wesley, 80–203.

Mücahit, B., 2014, *Özellik Çıkarma ve DVM Tabanlı Adaboost Algoritması ile Biyomedikal Veri Sınıflandırma*, Yüksek Lisans Tezi, Selçuk Üniversitesi.

Müssel, C., Lausser, L., Maucher, M. ve diğ., 2012, Multi-Objective Parameter Selection for Classifiers, *Journal of Statistical Software*, 46 (5), 1–27.

Nahar, J., Imam, T., Tickle, K.S. ve diğ., 2013, Computational Intelligence for Heart Disease Diagnosis: A Medical Knowledge Driven Approach, *Expert Systems with Applications*, 40 (1), DOI: 10.1016/j.eswa.2012.07.032, 96–104.

Narin, A., 2013, *Çoklu Sınıflandırıcı Sistemleri ile Konjestif Kalp Yetmezliği Teşhisi*, Yüksek Lisans Tezi, Bülent Ecevit Üniversitesi.

Narin, N., Mutlu, F., Argun, M. ve diğ., 2015, Incidence and Clinical Features of Acute Rheumatic Fever in Kayseri, Central Anatolia, 1998–2011, *Cardiology in the Young*, 25 (4), DOI: 10.1017/S1047951114000900, 745–751.

Ngueilbaye, A., Lei, L. ve Wang, H., 2016, Comparative Study of Data Mining Techniques on Heart Disease Prediction System: a case study for the "Republic of Chad", *International Journal of Science and Research*, 5 (5), 1564–1571.

Olguntürk, R., Canter, B., Tunaoğlu, F.S. ve diğ., 2006, Review of 609 Patients with Rheumatic Fever in Terms of Revised and Updated Jones Criteria, *International Journal of Cardiology*, 112 (1), DOI: 10.1016/j.ijcard.2005.11.007, 91–98.

Olguntürk, R., Aydın, G., Tunaoğlu, F.S. ve diğ., 1999, Rheumatic Heart Disease Prevalence Among Schoolchildren in Ankara, Turkey, *The Turkish Journal of Pediatrics*, 41 (2), 201–206.

Olson, D.L. ve Delen, D., 2008, *Advanced Data Mining Techniques*, Springer Verlag, Berlin Heidelberg, ISBN: 978-3-540-76917-0.

Osman Özdemir, D., Işık, Ş., Abacı, A. ve diğ., 2011, Akut Romatizmal Ateşte Sessiz Düşman: Subklinik Kardit, *Türk Kardiyoloji Derneği Arşivi*, 39 (1), 41–46.

Örün, U.A., Ceylan, Ö., Bilici, M. ve diğ., 2011, *Acute Rheumatic Fever in the Central Anatolia Region of Turkey: a 30-year experience in a single center*, European Journal of Pediatrics, 171 (2), DOI: 10.1007/s00431-011-1555-x, 361–368.

Özen, Z., 2016, *Kimlik Doğrulaması için Tuş Vuruş Dinamiklerine Dayalı Bir Güvenlik Sisteminin Yapay Sinir Ağları ile Geliştirilmesi*, Doktora Tezi, İstanbul Üniversitesi.

Özer, S., Hallioğlu, O., Özkutlu, S. ve diğ., 2005, Childhood Acute Rheumatic Fever in Ankara, Turkey, *The Turkish Journal of Pediatrics*, 47 (2), 120–124.

Özkan, B. ve Özkan, Y., 2017, *R ile Programlama*, 1. baskı, Papatya Yayıncılık Eğitim, İstanbul, ISBN: 978-605-9594-20-2.

Özkan, Y. ve Selçukcan Erol, Ç., 2017, *Biyoenformatik DNA Mikrodizi Veri Madenciliği*, Papatya Bilim, İstanbul, ISBN: 978-605-4220-89-2.

Özkan, Y., 2013, *Veri Madenciliği Yöntemleri*, 2. baskı, Papatya Yayıncılık Eğitim, ISBN: 0-201-74128-8.

Pandya, R. ve Pandya, J., 2015, C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning, *International Journal of Computer Applications*, 117 (16).

Parıldar, O., 2014, *Trafik Kazalarının Sınıflandırılmasında Karar Ağacı Kullanımı: Bodrum İlçesi Örneği*, Yüksek Lisans Tezi, Gazi Üniversitesi.

Park, K.M., 2008, *Pediatric Cardiology for Practitioners*, 5. baskı, Elsevier, ABD, ISBN: 978-0-323-04636-7.

Patil, P.N., Lathi, P.R. ve Chitre, P.V., 2012, Comparison of C5.0 & CART Classification Algorithms Using Pruning Technique, *International Journal of Engineering Research and Technology*, 1 (4).

Peace, I.C., 2015, An Analytical Review of Data Mining Tools, *International Journal of Engineering Research and Technology*, 4 (4), DOI: http://dx.doi.org/10.17577/IJERTV4IS040611, 387–389.

Quinlan, J.R., 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ISBN: 978-1-55860-238-0.

Quinlan, J.R., 1986, Induction of decision trees, *Machine Learning*, 1 (1), 81–106.

Quinlan, J.R., 1996, Learning Decision Tree Classifiers, *ACM Computing Surveys (CSUR)*, 28 (1), 71–72.

R Core Team, 2017, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Rajathi, S. ve Radhamani, G., 2016, Prediction and Analysis of Rheumatic Heart Disease Using kNN Classification with ACO, *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, DOI: 10.1109/SAPIENCE.2016.7684132, 68–73.

Rana, Q.P. ve Kaur, P., 2014, Comparison of Various Tools for Data Mining, *International Journal of Engineering Research and Technology*, 3 (10), 393–397.

Rangra, K. ve Bansal, K.L., 2014, Comparative Study of Data Mining Tools, *International Journal of Advanced Research in Computer Science and Software Engineering*, 4 (6).

Revathi, T. ve Jeevitha, S., 2015, Comparative Study on Heart Disease Prediction System Using Data Mining Techniques, Int*ernational Journal of Science and Research (IJSR)*, ISSN, 4 (7), 2120–2123.

Revathi, T. ve Sumathi, P., 2014, An Overview of Data Mining Classification Methods in Aortic Stenosis Prediction, *International Journal of Engineering and Advanced Technology (IJEAT)*, 3 (6), 173–175.

Rohilla, J. ve Gulia, P., 2015, *A*nalysis of Data Mining Techniques for Diagnosing Heart Disease, *International Journal of Advanced Research in Computer Science and Software Engineering*, 5 (7), 696–700.

Roiger, R. ve Geatz, M., 2003, *Data Mining: A Tutorial Based Primer*, Addison Wesley, ABD, ISBN: 0-201-74128-8.

Rowley, J., 2007, The Wisdom Hierarchy: Representations of the DIKW Hierarchy, *Journal of Information Science*, 33 (2), DOI: 10.1177/0165551506070706, 163–180.

RStudio, 2017, RStudio – Open source and enterprise-ready professional software for R, *https://www.rstudio.com/*, [Ziyaret Tarihi: 07.06.2017].

Saltık, İ.L., 2007, Akut Romatizmal Ateş, *Güncel Pediatri*, 5 (1), 85–93.

Seckeler, M.D. ve Hoke, T., 2011, The Worldwide Epidemiology of Acute Rheumatic Fever and Rheumatic Heart Disease, *Clinical Epidemiology*, DOI: 10.2147/CLEP.S12977, 67.

Selçukcan Erol, Ç., 2016, Sağlık Bilimlerinde R ile Veri Madenciliği, *R ile Veri Madenciliği Uygulamaları*, In: Balaban, M.E. ve Kartal, E. (ed.), 1. baskı, Çağlayan Kitapevi, İstanbul, ISBN: 978-975-436-093-6, 25–47.

Semizel, E., Bostan, Ö.M. ve Çil, E., 2005, Akut Romatizmal Ateş, *Güncel Pediatri,* 3, 57–61.

Shannon, C.E., 1948, A Mathematical Theory of Communication, *The Bell System Technical Journal, 27, 379–423.*

Shiezadeh, Z., Sajedi, H. ve Aflakie, E., 2015, Diagnosis of Rheumatoid Arthritis Using an Ensemble Learning Approach, *Computer Science & Information Technology (CS & IT)*, 5 (15), DOI: 10.5121/csit.2015.51512, 139–148.

Shouman, M., Turner, T. ve Stocker, R., 2011, Using Decision Tree for Diagnosing Heart Disease Patients, *Proceedings of the 9th Australasian Data Mining Conference*, Australian Computer Society, Avusturalya, 23–29.

Sika-Paotonu, D., Beaton, A., Raghu, A. ve diğ., 2017, Acute Rheumatic Fever and Rheumatic Heart Disease, Streptococcus pyogenes: Basic Biology to Clinical Manifestations, University of Oklahoma 2016, Oklahoma, 738.

Silahtaroğlu, G., 2008, *Veri Madenciliği*, 1. baskı, Papatya Yayıncılık Eğitim, İstanbul, ISBN: 978-975-6797-81-5.

Stone, M., 1974, Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society. Series B (Methodological)*, 36 (2), 111–147.

Therneau, T., Atkinson, B. ve Ripley, B., 2017, *rpart: Recursive Partitioning and Regression Trees*.

Thomas, J. ve Princy, R.T., 2016, Human Heart Disease Prediction System Using Data Mining Techniques, *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, IEEE, DOI: 10.1109/ICCPCT.2016.7530265, ISBN: 978-1-5090-1277-0, 1–5.

Torgo, L., 2010, *Data Mining With R, Learning With Case Studies*, Chapman and Hall/CRC.

Uysal, İ., Bilen, M. ve Ulukuş, S., 2014, Twoing Algoritması ile Sınıflandırma: Kalp Hastalığı Uygulaması, *XVI. Akademik Bilişim Konferansı Bildirileri*, Mersin.

Wahbeh, A.H., Al-Radaideh, Q.A., Al-Kabi, M.N. ve diğ., 2011, A Comparison Study Between Data Mining Tools Over Some Classification Methods, *(IJACSA) International Journal of Advanced Computer Science and Applications*, Special Issue on Artificial Intelligence, 18–26.

Wickham, H. ve Bryan, J., 2017, *readxl: Read Excel Files*.

Wimmer, H. ve Powell, L.M., 2016, A Comparison of Open Source Tools for Data Science, *Journal of Information Systems Applied Research*, 9 (2), 4–12.

Witten, I., Frank, E. ve Hall, M., 2011, Data *Mining: Practical Machine Learning Tools and Techniques*, 3. baskı, Morgan Kaufman Publishers, ABD, ISBN: 978-0-12-374856-0.

Witten, I.H. ve Frank, E., 2005, *Data mining: Practical Machine Learning Tools And Techniques*, 2nd ed., Morgan Kaufman, Amsterdam ; Boston, MA (Morgan Kaufmann series in data management systems), ISBN: 978-0-12-088407-0.

Yavuz, T., Nisli, K., Oner, N. ve diğ., 2008, Long Term Follow-up Results of 139 Turkish Children and Adolescents with Rheumatic Heart Disease, *European Journal of Pediatrics*, 167 (11), DOI: 10.1007/s00431-008-0799-6, 1321–1326.

Yetkin, M., 2014, *Tanker Şamandıra Bağlama Sistemlerinin Yapay Sinir Ağları Tekniğiyle Optimizasyonu*, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi.

Yılmaz, H., 2014, *Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama*, Yüksek Lisans Tezi, Eskişehir Osmangazi Üniversitesi.

Yılmaz, İ., Güvenç, O., Yılmaz, F.H. ve diğ., 2015, Akut Romatizmal Ateş Tanısı Konulan Hastaların Klinik Özellikleri ve Ekokardiyografik Bulguları, *Selçuk Tıp Dergisi*, 31 (2), 73–76.

Zaïane, O.R., 1999, *Introduction to Data Mining*, University of Alberta.

Zhang, Z. ve Zhang, R., 2010, Multimedia Data Mining, *Data Mining and Knowledge Discovery Handbook*, 2. baskı, Springer US, ABD, ISBN: 978-0-387-09822-7, 1081–1109.

# EKLER

**EK 1.** Etik kurul belgesi.

T.C.
İSTANBUL ÜNİVERSİTESİ
İSTANBUL TIP FAKÜLTESİ
KLİNİK ARAŞTIRMALAR ETİK KURULU

**Sayı : 1361** **Tarih : 28.11.2016**
**Konu : Yrd. Doç. Dr. Çiğdem EROL hk.**

**Sayın Yrd. Doç. Dr. Çiğdem EROL**
**Enformatik Bölümü**

**İlgi: Enformatik Bölümünün 11/11/2016 gün ve 297 sayı yazısı**

Sorumlu araştırıcılığını üstlendiğiniz ve Yüksek Lisans Öğrencisi İlkim Ecem EMRE' nin yürüteceği 2016/1342 dosya numaralı "Veri Madenciliği ile Çocukluk Çağındaki Akut Romatizmal Ateşin Kalp Hastalığına Etkilerinin Analizi" başlıklı çalışma kurulumuzun 25/11/2016 gün ve 20 sayılı toplantısında görüşülerek etik yönden uygun bulunmuş olup, tutanaklar ekte sunulmuştur.

Bilgilerinizi rica ederim.

**Prof. Dr. A.Yağız ÜRESİN**
**İstanbul Tıp Fakültesi Klinik Araştırmalar**
**Etik Kurul Başkanı**

**Eki:** İstanbul Tıp Fakültesi Klinik Araştırmaları Etik Kurulu Karar Formu

İSTANBUL TIP FAKÜLTESİ KLİNİK ARAŞTIRMALARI ETİK KURULU KARAR FORMU

| ETİK KURUL BİLGİLERİ | | |
|---|---|---|
| | ETİK KURULUN ADI | **İSTANBUL TIP FAKÜLTESİ KLİNİK ARAŞTIRMALARI ETİK KURULU** |
| | AÇIK ADRESİ: | **İ.Ü.İSTANBUL TIP FAKÜLTESİ HULUSİ BEHÇET KÜTÜPHANESİ KAT:3 FATİH/İSTANBUL** |
| | TELEFON | **0 (212) 414 21 53** |
| | FAKS | **0 (212) 414 21 53** |
| | E-POSTA | **itfetikkurul@istanbul.edu.tr.** |

| BAŞVURU BİLGİLERİ | | | | | |
|---|---|---|---|---|---|
| | ARAŞTIRMANIN AÇIK ADI | **"Veri Madenciliği ile Çocukluk Çağındaki Akut Romatizmal Ateşin Kalp Hastalığına Etkilerinin Analizi"** | | | |
| | ARAŞTIRMA PROTOKOL KODU | --- | | | |
| | KOORDİNATÖR/SORUMLU ARAŞTIRMACI UNVANI/ADI/SOYADI | **Yrd. Doç. Dr. Çiğdem EROL** | | | |
| | KOORDİNATÖR/SORUMLU ARAŞTIRMACININ UZMANLIK ALANI | **Biyoenformatik, Veri Madenciliği, Bilgi Teknolojileri** | | | |
| | KOORDİNATÖR/SORUMLU ARAŞTIRMACININ BULUNDUĞU MERKEZ | **İstanbul Üniversitesi Enformatik Bölümü** | | | |
| | DESTEKLEYİCİ | **İstanbul Üniversitesi Bilimsel Araştırmalar Birimi** | | | |
| | DESTEKLEYİCİNİN YASAL TEMSİLCİSİ | --- | | | |
| | ARAŞTIRMANIN FAZI | FAZ 1 | ☐ | | |
| | | FAZ 2 | ☐ | | |
| | | FAZ 3 | ☐ | | |
| | | FAZ 4 | ☐ | | |
| | ARAŞTIRMANIN TÜRÜ | Yeni Bir Endikasyon | ☐ | | |
| | | Yüksek Doz Araştırması | ☐ | | |
| | | Diğer ise belirtiniz : | | | |
| | ARAŞTIRMAYA KATILAN MERKEZLER | TEK MERKEZ ☐ | ÇOK MERKEZLİ ■ | ULUSAL ■ | ULUSLAR ARASI ☐ |

İSTANBUL TIP FAKÜLTESİ KLİNİK ARAŞTIRMALARI ETİK KURULU KARAR FORMU

| ARAŞTIRMANIN AÇIK ADI | "Veri Madenciliği ile Çocukluk Çağındaki Akut Romatizmal Ateşin Kalp Hastalığına Etkilerinin Analizi" |
|---|---|

| | Belge Adı | Tarihi | Versiyon Numarası | Dili |
|---|---|---|---|---|
| DEĞERLENDİRİLEN BELGELER | ARAŞTIRMA PROTOKOLÜ | 21/11/2016 | | Türkçe ■ İngilizce ☐ Diğer☐ |
| | BİLGİLENDİRİLMİŞ GÖNÜLLÜ OLUR FORMU | ☐ | | Türkçe ☐ İngilizce ☐ Diğer☐ |
| | OLGU RAPOR FORMU | ☐ | | Türkçe ☐ İngilizce ☐ Diğer☐ |
| | ARAŞTIRMA BROŞÜRÜ | ☐ | | Türkçe ☐ İngilizce ☐ Diğer☐ |

| | Belge Adı | | Açıklama |
|---|---|---|---|
| DEĞERLENDİRİLEN DİĞER BELGELER | TÜRKÇE ETİKET ÖRNEĞİ | ☐ | |
| | SİGORTA | ☐ | |
| | ARAŞTIRMA BÜTÇESİ | ■ | |
| | BİYOLOJİK MATERYEL TRANSFER FORMU | ☐ | |
| | HASTA KARTI/GÜNLÜKLERİ | ☐ | |
| | İLAN | ☐ | |
| | YILLIK BİLDİRİM | ☐ | |
| | SONUÇ RAPORU | ☐ | |
| | GÜVENLİLİK BİLDİRİMLERİ | ☐ | |
| | DİĞER: | ■ | Anabilim Dalı Başkanlığından Üst Yazı ve Akademik Kurul Kararı, Literatür Kaynağı, Sorumluluk Paylaşım Belgesi, Olgu Rapor Formu, İlgili Elemanların Bilgilendirildiğine Dair Belge, CV, CD |

| KARAR BİLGİLERİ | Karar No:20 | Tarih: 25/11/2016 |
|---|---|---|
| | İstanbul Üniversitesi Enformatik Bölümünde görevli **Yrd. Doç. Dr. Çiğdem EROL' un** sorumluluğunda ve Yüksek Lisans Öğrencisi İlkim Ecem EMRE' nin yürüteceği yukarıda bilgileri verilen araştırma başvuru dosyası ile ilgili belgeler araştırmanın gerekçe, amaç, yaklaşım ve yöntemleri dikkate alınarak incelenmiş, gerçekleştirilmesinde etik ve bilimsel sakınca bulunmadığına toplantıya katılan Etik Kurul üye tam sayısının salt çoğunluğu ile karar verilmiştir. | |

| İSTANBUL TIP FAKÜLTESİ KLİNİK ARAŞTIRMALARI ETİK KURULU | |
|---|---|
| ÇALIŞMA ESASI | 19.08.2011 tarihli, 28030 sayılı Resmi Gazetede yayınlanan Klinik Araştırmalar Hakkındaki Yönetmelik |
| BAŞKANIN UNVANI / ADI / SOYADI: | Prof. Dr. A. Yağız ÜRESİN |

| Unvanı/Adı/Soyadı | Uzmanlık Alanı | Kurumu | Cinsiyet | | Araştırma ile ilişki * | | Katılım ** | | İmza |
|---|---|---|---|---|---|---|---|---|---|
| Prof. Dr. A. Yağız ÜRESİN | Farmakoloji ve Klinik Farmakoloji | İstanbul Tıp Fakültesi (Etik Kurul Başkanı | E ■ | K ☐ | E ☐ | H ■ | E ■ | H ☐ | |
| Prof. Dr. Berrin UMMAN | Kardiyoloji | İstanbul Tıp Fakültesi (Etik Kurul Başkan Yardımcısı) | E ☐ | K ■ | E ☐ | H ■ | E ■ | H ☐ | |
| Prof. Dr. Ahmet GÜL | Romatoloji | İstanbul Tıp Fakültesi | E ■ | K ☐ | E ☐ | H ■ | E ■ | H ☐ | |
| Prof. Dr. Oğuzhan ÇOBAN | Nöroloji | İstanbul Tıp Fakültesi | E ■ | K ☐ | E ☐ | H ■ | E ■ | H ☐ | |
| Dr. Sevda ÖZEL YILDIZ | Biyoistatistik | İ.Ü. İstanbul Tıp Fakültesi Biyoistatistik | E ☐ | K ■ | E ☐ | H ■ | E ■ | H ☐ | |

* :Araştırma ile İlişki
** :Toplantıda Bulunma

**İ.Ü. İstanbul Tıp Fakültesi Klinik araştırmalar Etik kurulu 13.04.2013 tarih, 28617 sayılı Resmi Gazetede yayınlanan Klinik Araştırmalar Hakkında Yönetmelik çerçevesinde kurulmuş ve T.C. Sağlık Bakanlığı Türkiye İlaç ve Tıbbi Cihaz Kurumu tarafından onaylanmıştır. İlgili yönetmelik kapsamında kalan araştırmalar Sağlık Bakanlığından izin almak zorundadır. Yönetmelik kapsamı dışında kalan araştırmalar ise Etik Kurul bünyesinde oluşturulmuş 5 kişilik alt komisyon tarafından değerlendirilmekte olup Sağlık Bakanlığı iznine tabi değildir.**

**EK 2.** Veri ön işleme R kodları.

```
#VERİ ÖN İŞLEME_KATEGORİK VERİ SETİ İÇİN
#veri seti excel dosyası olarak yüklendi
setwd("D:/ilkim ecem emre/Documents/veriSeti")
library(readxl)
veriseti <- read_excel("ARAveriSeti.xlsx")

#veri setinin özeti görüntülendi
summary(veriseti)

#dogumTar sütunu çıkarıldı
veriseti <- subset( veriseti, select = -dogumTar )
#tek kategorisi olan agrikesici ve prazitilaci sütunları çıkarıldı
veriseti <- veriseti[,-c(39,51)]

#kategorik veri ile çalışmak için kan değerlerindeki nümerik sütunlar
çıkarıldı
veriseti <- veriseti[,-c(12,14,16,18,20)]

#hastanesure,ilkatakyas,sonkontrolyas sütunları nümerik olarak, diğerleri
kategorik düzenlendi
for(i in 1:ncol(veriseti)){
  if(i != 2 && i != 5 && i != 52)
    veriseti[[i]]<-as.factor(veriseti[[i]])}

#eksik veri knn yöntemi ile dolduruldu
install.packages("DMwR")
library(DMwR)
tamveri<-knnImputation(veriseti)
tamveri<-as.data.frame(tamveri)

#tekrar eden ve eksik veri kontrolü yapıldı
anyDuplicated(tamveri)
anyNA(tamveri)

#veri setinin özeti görüntülendi
summary(tamveri)

#veri setinin ön işlenmeden sonraki versiyonu kaydedildi
save(tamveri,file="tamVeri_kategorik.RData")

#veri setinin ön işlemeden sonraki versiyonu sonradan yüklemek için
setwd("D:/ilkim ecem emre/Documents/veriSeti")
load(file="tamVeri_kategorik.RData")

################################################################################
```

```
#VERİ ÖN İŞLEME_NÜMERİK VERİ SETİ İÇİN
#veri seti excel dosyası olarak yüklendi
setwd("D:/ilkim ecem emre/Documents/veriSeti")
library(readxl)
veriseti <- read_excel("ARAveriSeti.xlsx")

#veri setinin özeti görüntülendi
summary(veriseti)

#dogumTar sütunu çıkarıldı
veriseti <- subset( veriseti, select = -dogumTar )
#tek kategorisi olan agrikesici ve prazitilaci sütunları çıkarıldı
veriseti <- veriseti[,-c(39,51)]

#nümerik veri ile çalışmak için kan değerlerindeki kategorik sütunlar
çıkarıldı
veriseti <- veriseti[,-c(13,15,17,19,21)]

#hastanesure,ilkatakyas,sonkontrolyas ve kan değerleri ile ilgili olan
sütunlar nümerik olarak, diğerleri kategorik düzenlendi
for(i in 1:ncol(veriseti)){
  if(i != 2 && i != 5 && i != 12 && i != 13 && i != 14 && i != 15 && i != 16
&& i != 52)
    veriseti[[i]]<-as.factor(veriseti[[i]])}

#eksik veri knn yontemi ile dolduruldu
install.packages("DMwR")
library(DMwR)
tamveri<-knnImputation(veriseti)
tamveri<-as.data.frame(tamveri)

#tekrar eden ve eksik veri kontrolü yapıldı
anyDuplicated(tamveri)
anyNA(tamveri)

#uç değerler kontrol edildi
outlier.scores <- lofactor(tamveri[,12:16], k=10)
outlier.scores

#veri setinin özeti görüntülendi
summary(tamveri)

#veri setinin ön işlenmeden sonraki versiyonu kaydedildi
save(tamveri,file="tamVeri_numerik.RData")
#veri setinin ön işlemeden sonraki versiyonunu sonradan yüklemek için
setwd("D:/ilkim ecem emre/Documents/veriSeti")
load(file="tamVeri_numerik.RData")
```

**EK 3.** Modellere ait R kodları.

```
#MODEL KURMA _ SADE BAYES
################################################
#BAYES modelini oluşturmak için gerekli paket yüklendi
install.packages("e1071")
install.packages("caret")
install.packages("TunePareto")
library(e1071)
library(caret)
library(TunePareto)

################################################
#HOLD OUT yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.80,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.70,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.60,list=FALSE)

egitim<-tamveri[egitimindisleri,]
test<-tamveri[-egitimindisleri,]

egitim_nitelikleri<-egitim[,-54]
egitim_hedef<-egitim[[54]]

test_nitelikleri<- test[,-54]
test_hedef<-test[[54]]

#BAYES modeli oluşturuldu
naiveBayes_modeli<- naiveBayes(egitim_nitelikleri,egitim_hedef)

#test verisi ile deneme yapıldı
tahminisiniflar<-predict(naiveBayes_modeli,test_nitelikleri)

#karmaşıklık matrisi ile gerçek sınıflar ve tahmin edilen sınıflar
karşılaştırıldı
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))

#########################
#CROSSVALIDATION yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde katsayısı ve nrow aynı anda
5,10,15 olarak değiştirildi
tekrar<-1
katsayisi<-5
set.seed(8)

kat<-generateCVRuns(tamveri$karar,ntimes=tekrar,nfold=katsayisi,
leaveOneOut = FALSE,stratified = TRUE)
kat
accuracyFolds<-data.frame(matrix(ncol = 2, nrow = 5))
colnames(accuracyFolds) <- c("accuracy", "katsayısı")
```

```
i<-1
for(i in 1:katsayisi)

{test<-tamveri[kat$`Run  1`[[i]],]
egitim<-tamveri[-kat$`Run  1`[[i]],]

egitim_nitelikleri<-egitim[,-54]
egitim_hedef<-egitim[[54]]

test_nitelikleri<-test[,-54]
test_hedef<-test[[54]]

naiveBayes_modeli<- naiveBayes(egitim_nitelikleri,egitim_hedef)
tahminisiniflar<-predict(naiveBayes_modeli,test_nitelikleri)

modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test_hedef,
dnn=c("tahmini_siniflar","gercek_siniflar"))
print(modelperformansi$overall["Accuracy"])
accuracyFolds$accuracy[i] <- modelperformansi$overall["Accuracy"]
accuracyFolds$katsayısı[i] <- i
View(accuracyFolds)}

modelAcc<- mean(accuracyFolds$accuracy)

#########################
#BOOTSTRAP yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-sample(1:50,replace = TRUE)
egitimindisleri<-sample(1:100,replace = TRUE)
egitimindisleri<-sample(1:200,replace = TRUE)

egitim<-(tamveri[egitimindisleri,])
test<-(tamveri[-egitimindisleri,])

egitim<-tamveri[egitimindisleri,]
test<-tamveri[-egitimindisleri,]

egitim_nitelikleri<-egitim[,-54]
egitim_hedef<-egitim[[54]]

test_nitelikleri<- test[,-54]
test_hedef<-test[[54]]

naiveBayes_modeli<- naiveBayes(egitim_nitelikleri,egitim_hedef)
tahminisiniflar<-predict(naiveBayes_modeli,test_nitelikleri)

modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
```

```
#MODEL KURMA _ CART
################################################
#CART modelini oluşturmak için gerekli paket yüklendi
install.packages("rpart")
install.packages("caret")
install.packages("TunePareto")
library(rpart)
library(caret)
library(TunePareto)

#HOLD OUT yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.80,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.70,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.60,list=FALSE)

egitim<-(tamveri[egitimindisleri,])
test<-(tamveri[-egitimindisleri,])

#CART modeli modeli oluşturuldu
CART_modeli<- rpart(karar ~ . , method="class", data=egitim)

#test verisi ile deneme yapıldı
tahminisiniflar<-predict(CART_modeli,test[,-54],type = "class")

#karmaşıklık matrisi ile gerçek sınıflar ve tahmin edilen sınıflar
karşılaştırıldı
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))

#########################
#CROSSVALIDATION yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde katsayısı ve nrow aynı anda
5,10,15 olarak değiştirildi
tekrar<-1
katsayisi<-5
set.seed(8)

kat<-generateCVRuns(tamveri$karar,ntimes=tekrar,nfold=katsayisi,
leaveOneOut = FALSE,stratified = TRUE)
kat
accuracyFolds<-data.frame(matrix(ncol = 2, nrow = 5))
colnames(accuracyFolds) <- c("accuracy", "katsayısı")

i<-1
for(i in 1:katsayisi)

{test<-as.data.frame(tamveri[kat$`Run  1`[[i]],])
egitim<-as.data.frame(tamveri[-kat$`Run  1`[[i]],])

CART_modeli<- rpart(karar ~ . , method="class", data=egitim)
tahminisiniflar<-predict(CART_modeli,test[,-54],type = "class")
```

```
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
print(modelperformansi$overall["Accuracy"])
accuracyFolds$accuracy[i] <- modelperformansi$overall["Accuracy"]
accuracyFolds$katsayısı[i] <- i
View(accuracyFolds)}

modelAcc<- mean(accuracyFolds$accuracy)

#########################
#BOOTSTRAP yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-sample(1:50,replace = TRUE)
egitimindisleri<-sample(1:100,replace = TRUE)
egitimindisleri<-sample(1:200,replace = TRUE)

egitim<-(tamveri[egitimindisleri,])
test<-(tamveri[-egitimindisleri,])

CART_modeli<- rpart(karar ~ . , method="class", data=egitim)
tahminisiniflar<-predict(CART_modeli,test[,-54],type = "class")

modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
modelperformansi

#MODEL KURMA _ C4.5
################################################
#C4.5 modelini oluşturmak için gerekli paket yüklendi
install.packages("RWeka")
install.packages("caret")
install.packages("TunePareto")
library(RWeka)
library(caret)
library(TunePareto)

################################################
#HOLD OUT yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.80,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.70,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.60,list=FALSE)

egitim<-as.data.frame(tamveri[egitimindisleri,])
test<-as.data.frame(tamveri[-egitimindisleri,])

#C4.5 modeli oluşturuldu
C45_modeli<- J48(karar~.,egitim)

#test verisi ile deneme yapıldı
tahminisiniflar<-predict(C45_modeli,test[-54])
```

```
#karmaşıklık matrisi ile gerçek sınıflar ve tahmin edilen sınıflar
karşılaştırıldı
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))

#########################
#CROSSVALIDATION yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde katsayısı ve nrow aynı anda
5,10,15 olarak değiştirildi
tekrar<-1
katsayisi<-5
set.seed(8)

kat<-generateCVRuns(tamveri$karar,ntimes=tekrar,nfold=katsayisi,
leaveOneOut = FALSE,stratified = TRUE)
kat
accuracyFolds<-data.frame(matrix(ncol = 2, nrow = 5))
colnames(accuracyFolds) <- c("accuracy", "katsayısı")

i<-1
for(i in 1:katsayisi)

{test<-tamveri[kat$`Run  1`[[i]],]
egitim<-tamveri[-kat$`Run  1`[[i]],]

test<-as.data.frame(tamveri[kat$`Run  1`[[i]],])
egitim<-as.data.frame(tamveri[-kat$`Run  1`[[i]],])

C45_modeli<- J48(karar~.,egitim)
(tahminisiniflar<-predict(C45_modeli,test[,-54]))
print(summary(C45_modeli))

modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
print(modelperformansi$overall["Accuracy"])
accuracyFolds$accuracy[i] <- modelperformansi$overall["Accuracy"]
accuracyFolds$katsayısı[i] <- i
View(accuracyFolds)}

modelAcc<- mean(accuracyFolds$accuracy)

#########################
#BOOTSTRAP yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-sample(1:50,replace = TRUE)
egitimindisleri<-sample(1:100,replace = TRUE)
egitimindisleri<-sample(1:200,replace = TRUE)

egitim<-as.data.frame(tamveri[egitimindisleri,])
test<-as.data.frame(tamveri[-egitimindisleri,])

C45_modeli<- J48(karar~.,egitim)
tahminisiniflar<-predict(C45_modeli,test[-54])
```

```
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
```

```
#MODEL KURMA _ C5.0
################################################
#C5.0 modelini oluşturmak için gerekli paket yüklendi
install.packages("C50")
install.packages("caret")
install.packages("TunePareto")
library(C50)
library(caret)
library(TunePareto)

################################################
#HOLD OUT yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.80,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.70,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.60,list=FALSE)

egitim<-as.data.frame(tamveri[egitimindisleri,])
test<-as.data.frame(tamveri[-egitimindisleri,])

#C5.0 modeli oluşturuldu
C50_modeli<- C5.0(x = egitim[,-54],y = egitim$karar)

#boosting için aşağıdaki şekilde model oluşturuldu
#C50_modeli<- C5.0(x = egitim[,-54],y = egitim$karar, trials = 10)

#test verisi ile deneme yapıldı
tahminisiniflar<-predict(C50_modeli,test[-54])

#karmaşıklık matrisi ile gerçek sınıflar ve tahmin edilen sınıflar
karşılaştırıldı
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))

#########################
#CROSSVALIDATION yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde katsayısı ve nrow aynı anda
5,10,15 olarak değiştirildi
tekrar<-1
katsayisi<-5
set.seed(8)

kat<-generateCVRuns(tamveri$karar,ntimes=tekrar,nfold=katsayisi,
leaveOneOut = FALSE,stratified = TRUE)
kat
accuracyFolds<-data.frame(matrix(ncol = 2, nrow = 5))
colnames(accuracyFolds) <- c("accuracy", "katsayısı")

i<-1
for(i in 1:katsayisi)

{test<-as.data.frame(tamveri[kat$`Run  1`[[i]],])
egitim<-as.data.frame(tamveri[-kat$`Run  1`[[i]],])

C50_modeli<- C5.0(x = egitim[,-54],y = egitim$karar)
```

```
#boosting için aşağıdaki şekilde model oluşturuldu
#C50_modeli<- C5.0(x = egitim[,-54],y = egitim$karar, trials = 10)

tahminisiniflar<-predict(C50_modeli,test[-54])

modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
print(modelperformansi$overall["Accuracy"])
accuracyFolds$accuracy[i] <- modelperformansi$overall["Accuracy"]
accuracyFolds$katsayısı[i] <- i
View(accuracyFolds)}

modelAcc<- mean(accuracyFolds$accuracy)

#########################
#BOOTSTRAP yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-sample(1:50,replace = TRUE)
egitimindisleri<-sample(1:100,replace = TRUE)
egitimindisleri<-sample(1:200,replace = TRUE)

egitim<-as.data.frame(tamveri[egitimindisleri,])
test<-as.data.frame(tamveri[-egitimindisleri,])

C50_modeli<- C5.0(x = egitim[,-54],y = egitim$karar)

#boosting için aşağıdaki şekilde model oluşturuldu
#C50_modeli<- C5.0(x = egitim[,-54],y = egitim$karar, trials = 10)

tahminisiniflar<-predict(C50_modeli,test[-54])

modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
```

```
#MODEL KURMA _ RASTGELE ORMAN
################################################
#RASTGELE ORMAN modelini oluşturmak için gerekli paketler yüklendi
install.packages("randomForest")
install.packages("caret")
install.packages("TunePareto")
library(randomForest)
library(caret)
library(TunePareto)

################################################
#HOLD OUT yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.80,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.70,list=FALSE)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.60,list=FALSE)

egitim<-as.data.frame(tamveri[egitimindisleri,])
test<-as.data.frame(tamveri[-egitimindisleri,])

#RASTGELE ORMAN modeli oluşturuldu
rf_modeli<- (randomForest(karar~.,data = egitim))

#test verisi ile deneme yapıldı
tahminisiniflar<-predict(rf_modeli,test[-54])

#karmaşıklık matrisi ile gerçek sınıflar ve tahmin edilen sınıflar
karşılaştırıldı
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))

#########################
#CROSSVALIDATION yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde katsayısı ve nrow aynı anda
5,10,15 olarak değiştirildi
tekrar<-1
katsayisi<-5
set.seed(8)

kat<-generateCVRuns(tamveri$karar,ntimes=tekrar,nfold=katsayisi,
leaveOneOut = FALSE,stratified = TRUE)
kat
accuracyFolds<-data.frame(matrix(ncol = 2, nrow = 5))
colnames(accuracyFolds) <- c("accuracy", "katsayısı")

i<-1
for(i in 1:katsayisi)

{test<-as.data.frame(tamveri[kat$`Run  1`[[i]],])
egitim<-as.data.frame(tamveri[-kat$`Run  1`[[i]],])

rf_modeli<- (randomForest(karar~.,data = egitim))
tahminisiniflar<-predict(rf_modeli,test[-54])
```

```
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
print(modelperformansi$overall["Accuracy"])
accuracyFolds$accuracy[i] <- modelperformansi$overall["Accuracy"]
accuracyFolds$katsayısı[i] <- i
View(accuracyFolds)}

modelAcc<- mean(accuracyFolds$accuracy)

#########################
#BOOTSTRAP yöntemine göre eğitim ve test veri setleri oluşturuldu
#üç farklı şekilde deneme yapıldı, her seferinde bir satır çalıştırıldı
set.seed(8)
egitimindisleri<-sample(1:50,replace = TRUE)
egitimindisleri<-sample(1:100,replace = TRUE)
egitimindisleri<-sample(1:200,replace = TRUE)

egitim<-as.data.frame(tamveri[egitimindisleri,])
test<-as.data.frame(tamveri[-egitimindisleri,])

rf_modeli<- (randomForest(karar~.,data = egitim))
tahminisiniflar<-predict(rf_modeli,test[-54])

modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))
```

**EK 4.** En iyi performans veren modele (CART) ait R kodları.

```
#EN İYİ MODELE AİT KODLAR
#CART modelini oluşturmak için gerekli paketler yüklendi
install.packages("rpart")
library(rpart)
library(caret)

#HOLD OUT yöntemine göre eğitim ve test veri setleri oluşturuldu
set.seed(8)
egitimindisleri<-createDataPartition(y=tamveri$karar,p=.80,list=FALSE)
egitim<-(tamveri[egitimindisleri,])
test<-(tamveri[-egitimindisleri,])

#CART modeli oluşturuldu
CART_modeli<- rpart(karar ~ . , method="class", data=egitim)

#test verisi ile deneme yapıldı
tahminisiniflar<-predict(CART_modeli,test[,-54],type = "class")

#karmaşıklık matrisi ile gerçek sınıflar ve tahmin edilen sınıflar
karşılaştırıldı
modelperformansi<-
confusionMatrix(data=tahminisiniflar,reference=test$karar,
dnn=c("tahmini_siniflar","gercek_siniflar"))

#sınıf bazında sonuçlar görüntülendi
sonuc<-as.data.frame(modelperformansi$byClass)
View(sonuc)

############################
#en iyi modele ait karar ağacı oluşturuldu
install.packages("rpart.plot")
library(rpart.plot)
rpart.plot(CART_modeli,type = 3)
```

# ÖZGEÇMİŞ

| Kişisel Bilgiler | |
|---|---|
| Adı Soyadı | İlkim Ecem EMRE |
| Doğum Yeri | İstanbul |
| Doğum Tarihi | 06.02.1991 |
| Uyruğu | ☑ T.C. ❒ Diğer: |
| Telefon | |
| E-Posta Adresi | ecem.emre@marmara.edu.tr |
| Web Adresi | |

| Eğitim Bilgileri | |
|---|---|
| **Lisans** | |
| Üniversite | Marmara Üniversitesi |
| Fakülte | İktisadi İdari Bilimler Fakültesi |
| Bölümü | Almanca İşletme Enformatiği |
| Mezuniyet Yılı | 2015 |

| Yüksek Lisans | |
|---|---|
| Üniversite | İstanbul Üniversitesi |
| Enstitü Adı | Fen Bilimleri Enstitüsü |
| Anabilim Dalı | Enformatik Anabilim Dalı |
| Programı | Enformatik |
| Mezuniyet Tarihi | 2017 |

**Makale ve Bildiriler**

Özen, Z., Kartal, E. ve Emre, İ.E., 2017, A Case Study on Improving E-Learning Services Using Google Analytics in Turkey, *International Journal of E-Adoption (IJEA)*, 9 (1), DOI: 10.4018/IJEA.2017010103, 26–37.

Emre, İ.E. ve Erol, Ç.S., 2017, Veri Analizinde İstatistik mi Veri Madenciliği mi?, *International Journal Of Informatics Technologies*, 10 (2), DOI: 10.17671/btd.63043, 161.

Koçoğlu, F.Ö., Emre, İ.E. ve Erol, Ç.S., 2017, Observation of Success Status of Employees in E-Learning Courses in Organizations with Data Mining, *International Journal of E-Adoption (IJEA)*, 9 (1), DOI: 10.4018/IJEA.2017010104, 38–49.

Bozkurt, P., Güngör, G., Özen, Z. ve diğ., 2016, Older Age Is Related with Higher Pocp in Adult Patient Population in Turkey- Preliminary Report, *Anesthesia & Analgesia*, 123, DOI: 10.1213/01.ane.0000492682.07025.46, 368.

Emre, İ.E., Kartal, E. ve Gülseçen, S., 2016, Fen Bilimleri ve Sosyal Bilimler Alanlarındaki Öğrencilerin Eğitimde Bilgi Teknolojilerine Bakış Açısı: İstanbul Üniversitesi İncelemesi, *Hasan Ali Yücel Eğitim Fakültesi Dergisi*, 13 (3), 1–15.

Koçoğlu, F.Ö., Emre, İ.E. ve Selçukcan Erol, Ç., 2016, Observation of Success in e-Learning with Data Mining Methods, *6th International Conference on 'Innovations in Learning for the Future' 2016: Next Generation Book of Abstracts*, İstanbul, ISBN: 978-605-07-0607-9, 75.

Özen, Z., Kartal, E. ve Emre, İ.E., 2016, A Case Study on Improving E-Learning Services Using Google Analytics in Turkey, *6th International Conference on 'Innovations in Learning for the Future' 2016: Next Generation Book of Abstracts*, İstanbul, ISBN: 978-605-07-0607-9, 73.