**CUMHURİYET ÜNİVERSİTESİ**
**Sosyal Bilimler Enstitüsü**
**İşletme Ana Bilim Dalı**
**Sayısal Yöntemler Bilim Dalı**

# VERİ MADENCİLİĞİNDE KULLANILAN KESTİRİM YÖNTEMLERİNİN PERFORMANSLARININ KARŞILAŞTIRILMASI

**Doktora Tezi**

**Esra GÜLTÜRK**

**Sivas**
**Temmuz 2016**

**CUMHURİYET ÜNİVERSİTESİ**
**Sosyal Bilimler Enstitüsü**
**İşletme Ana Bilim Dalı**
**Sayısal Yöntemler Bilim Dalı**

# VERİ MADENCİLİĞİNDE KULLANILAN KESTİRİM YÖNTEMLERİNİN PERFORMANSLARININ KARŞILAŞTIRILMASI

**Doktora Tezi**

**Esra GÜLTÜRK**

**Tez Danışmanı**
**Doç. Dr. Hüdaverdi BİRCAN**

**Sivas**
**Temmuz 2016**

## KABUL VE ONAY

| | |
|---|---|
| **Üniversite:** | **: Cumhuriyet Üniversitesi** |
| **Enstitü** | **: Sosyal Bilimler Enstitüsü** |
| **Ana Bilim Dalı** | **: İşletme Ana Bilim Dalı** |
| **Bilim Dalı** | **: Sayısal Yöntemler** |
| **Tezin Başlığı** | **: Veri Madenciliğinde Kullanılan Kestirim Yöntemlerinin Performanslarının Karşılaştırılması** |
| **Savunma Tarihi** | **: 02/06/2016** |
| **Danışmanı** | **: Doç. Dr. Hüdaverdi Bircan** |

| | **Unvanı - Adı Soyadı** | **İmza** |
|---|---|---|
| **Jüri Başkanı** | **: Prof. Dr. Mahmut KARTAL** | |
| **Üye** | **: Doç. Dr. Erdem KARABULUT** | |
| **Üye** | **: Doç. Dr. Serdal Kenan KÖSE** | |
| **Üye** | **: Doç. Dr. Hüdaverdi BİRCAN** | |
| **Üye** | **: Doç. Dr. Mehmet Ali ALAN** | |

**Oy Birliği** ☑

**Oy Çokluğu** ☐

**Esra GÜLTÜRK tarafından hazırlanan "Veri Madenciliğinde Kullanılan Kestirim Yöntemlerinin Performanslarının Karşılaştırılması" başlıklı tez, kabul edilmiştir. …./…./…….**

**Prof. Dr. Metin BOZKUŞ**
**Enstitü Müdürü**

## ETİK İLKELERE UYGUNLUK BEYANI

Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü bünyesinde hazırladığım bu Doktora tezinin bizzat tarafımdan ve kendi sözcüklerimle yazılmış orijinal bir çalışma olduğunu ve bu tezde;

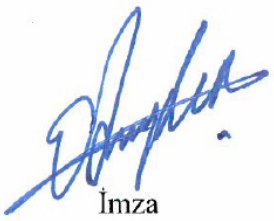
1- Çeşitli yazarların çalışmalarından faydalandığımda bu çalışmaların ilgili bölümlerini doğru ve net biçimde göstererek yazarlara açık biçimde atıfta bulunduğumu;

2- Yazdığım metinlerin tamamı ya da sadece bir kısmı, daha önce herhangi bir yerde yayımlanmışsa bunu da açıkça ifade ederek gösterdiğimi;

3- Başkalarına ait alıntılanan tüm verileri (tablo, grafik, şekil vb. de dahil olmak üzere) atıflarla belirttiğimi;

4- Başka yazarların kendi kelimeleriyle alıntıladığım metinlerini, tırnak içerisinde veya farklı dizerek verdiğim yine başka yazarlara ait olup fakat kendi sözcüklerimle ifade ettiğim hususları da istisnasız olarak kaynak göstererek belirttiğimi,

beyan ve bu etik ilkeleri ihlal etmiş olmam halinde bütün sonuçlarına katlanacağımı kabul ederim.



İmza

Esra GÜLTÜRK

# ÖN SÖZ

Yıllardır benden emeğini, bilgisini ve tecrübesini esirgemeyen bölüm hocam sayın Yrd. Doç. Ziynet ÇINAR'a, tez çalışmamda yardımlarından dolayı Doç Dr.Erdem KARABULUT ve tez danışmanım Doç. Dr. Hüdaverdi BİRCAN hocalarıma, tezin bu noktaya gelmesinde emeği olan tez izleme komitesindeki ve değerlendirme komitesindeki değerli hocalarıma içtenlikle teşekkür ederim.

Ayrıca manevi desteğini ve yardımlarını esirgemeyen Yusuf AKGÜL arkadaşıma ve bu doktora sürecinde her konuda yanımda olan sevgili ve değerli eşim Aziz'e, canım çocuklarım Yiğit ve Aysema'ya teşekkür ederim.

# İÇİNDEKİLER

# KISALTMALAR

| | |
|---|---|
| DVR: | Destek Vektör Regresyon |
| DVM: | Destek Vektör Makinası |
| HKO: | Hata Kareler Ortalaması- Mean Standart Error |
| IBM: | International Business Machines |
| KKKA: | Kırım Kongo Kanamalı Ateş |
| KNIME: | Konstanz Information Miner |
| OLAP: | Online Analytic Processing |
| $R^2$: | Açıklayıcılık katsayısı |
| OLTP: | Online Transaction Processing |
| ROLAP: | Realitional Online Analytic Processing |
| RF: | Random Forest- Rassal Orman |
| RA: | Regresyon Ağacı-Regression Tree |
| SAS: | Statistical Analysis Software |
| SPSS: | Statistical Package for the Social Sciences |
| SQL: | Yapısal sorgu dili- Structured query language |
| WEKA: | Waikato Environment for Knowledge Analysis |

# TABLOLAR LİSTESİ

# ŞEKİLLER LİSTESİ

# ÖZET

GÜLTÜRK, Esra. “Veri Madenciliğinde Kullanılan Kestirim Yöntemlerinin Performanslarının Karşılaştırlması”, Doktora Tezi, Sivas, Haziran 2016.

Bu çalışmada, literatürde eksikliği fark edilen "Destek Vektör Regresyon, Random Forest ve Regresyon Ağacı" yöntemlerinin kestirim performanslarının kıyaslanması amaçlanmıştır. Bağımlı değişkeni kategorik ve sürekli değişken olarak alıp hem sınıflama, hemde regresyon yöntemlerinin kestirim performansları incelenmiştir. Bu amaçla, Cumhuriyet Üniversitesi Tıp Fakültesi Enfeksiyon Hastalıkları ve Çocuk Sağlığı Hastalıkları servisinde yatan kırım kongo kanamalı ateş tanısı ile tedavi gören 2009-2011 yılları arası tüm hasta bireylerin verileri servis kayıtlarından alınmıştır. Bu üç yıl içerisindeki toplam 245 hastaya ait 6125 veri girişi yapılmıştır. Çalışmada yetişkin, çocuk ve tüm hasta olmak üzere toplam üç grup hasta verisi kullanılmıştır. Regresyon modellerinin performanslarını karşılaştırmak için hata kareler ortalaması ve açıklayıcı yüzdesine bakılmıştır. Sınıflamada modellerin karşılaştırılmalarına bakmak için duyarlılık, kesinlik, doğruluk oranı ve F ölçütüne bakılmıştır. Gerçek veri seti için regresyon yöntemlerinden, her üç gruptada destek vektör regresyon açıklayıcılık yüzdesi en fazla, hata kareler ortalaması en az olan regresyon modeli olarak bulunmuştur. Simülasyon çalışmasında, her bir senaryo 1000 kez tekrar edilmiş, her bir tekrarda sözü edilen regresyon yöntemleri uygulanmıştır. Senaryo yapılarına göre en iyi regresyon yöntemi destek vektör regresyon olarak bulunmuştur.

**Anahtar kelimeler**: Veri madenciliği, destek vektör regresyon, rassal orman, regresyon ağacı.

# ABSTRACT

GÜLTÜRK, Esra., “Performance Comparison of Estimation Methods Used In Data Mining ”, PhD Thesis, Sivas, June 2016.

In this study, performance comparison of estimation methods as "Support Vector Regression, Random Forest and Regression Tree" were aimed. By taking categorical and continuous variables as dependent variable, performances of classification and regression estimation methods were examined. For this purpose, data of all patients, who were hospitalized with the diagnosis of crimean-congo haemorrhagic fever between 2009 and 2011 years in Cumhuriyet University Faculty of Medicine, Infectious Diseases and Children's Health ward, were obtained from the service records. 6125 data entry of 245 patient’s were made within three years. In this study, three sets of data including adults, children and all patients were used. To compare the performances of regression models, mean square error and explanatory percentage were examined. Sensitivity, precision, accuracy and F measure were examined to look into comparison of models in classification. For real data set in all of three groups, explanatory percentage of support vector regression was maximum, mean square error of support vector regression was minimum. In the simulation study, each scenario was repeated 1000 times, relevant regression methods were applied in each repetition. According to the scenario structures, support vector regression was the best regression method.

**Key words:** Data mining, support rector regression, random forest, regression tree.

# GİRİŞ

Veri madenciliği, büyük kapasiteli veri tabanları içerisinden geleceğin tahmin edilmesine ve bilinmeyen verilerin çekilmesini sağlayan bir analiz yöntemidir (Savaş vd. 2012: 1). Veri madenciliği, veri arama bilimi ve teknolojisi olarak veri tabanlarında bilgi keşfetmeye yönelik genel sürecin bir parçasıdır. Günümüzün bilgisayar odaklı dünyasında, bu veri tabanları büyük çapta bilgiyi içerirler. Bu bilgilerin erişilebilirliği ve bolluğu veri madenciliğini çok önemli ve gerekli hale getirmektedir (Silahtaroğlu 2008:10; Savaş ve diğerleri 2012: 2).

Veri madenciliğinin gelişim adımları incelendiğinde veri toplama ilk 1960'lı yıllarda başlamıştır. Veri madenciliğinin ilk kullanımı 1970'lerde ve daha sonraki yıllarda geliştirilen uzman sistemlerle olmuştur. Uzman sistemlerin tıp alanında güçlü araçlar sunmasına rağmen, bu alandaki verilerin hızlı değişmesi ve uzmanlar arasındaki görüş farklılıkları nedeniyle çok yaygınlaşmamıştırlar. Daha sonraki yıllarda özellikle 1990'lı yıllarda hastaların gelecekteki sağlık durumları ve maliyet tahminleri gibi konuları araştırmak için sinir ağları kullanılmaya başlanmıştır. Sağlık sektörü 1990'lı yıllardan önce geleneksel istatistik teknikleri kullananırken, 1990'lı yıllardan itibaren veri madenciliği tekniklerini kullanmaya başlamıştır. Yönetimsel alanda, hastalık tespiti ve teşhisi, tedavi planlaması, risk tahminleri, davranış modellemeleri gibi klinik uygulamalarında yaygın olarak kullanılan veri madenciliği teknikleri, biyoistatistik alanına büyük katkılar sağlamıştır (Yıldırım ve diğerleri 2008:429; Koyuncugil, Özgülbaş 2009: 21-29).

Günümüzde artan rekabet koşullarıyla birlikte işletmelerin varlıklarını sürdürebilmeleri ve başarılı olabilmelerin de etkin ve doğru karar verebilmeleri önemlidir. Bu nedenle de büyük veritabanlarından istenilen anlamlı ve kullanılabilir verilere ulaşmak için çeşitli istatiksel metodlar geliştirilmiştir. Teknolojinin gelişmesiyle birlikte donanımın ucuzlaması verilerin uzun süre depolanmasına sebep olmuştur, dolayısıyla da büyük kapasiteli veri tabanları oluşmuştur (Emel ve diğerleri 2005: 31; Kaya, Köymen 2008: 159).

Geleneksel istatistiksel yöntemlerde bir hipotez kurulur ve bu hipotez kabul yada reddine yönelik istatistiksel testler yapılıp karar verilir, veri madenciliği ise

varlığı bilinen fakat kesin olmayan örüntü yapılarını araştırır. (Koyuncugil, Özgülbaş 2009: 25).

Veri madenciliği; veri tabanı sistemleri, veri görselliği, makina öğrenme, istatistik, yapay sinir ağları, vb. diğer disiplinler ile ilişkilidir (Koyuncugil, Özgülbaş 2009: 24). Veri madenciliği yöntemleri tahmin edici ve tanımlayı olarak iki kısımda incelenmiştir. Tahmin edicide; sınıflama ve regresyon modelleri kullanılırken, tanımyacıda daha çok kümeleme, birliktelik ve ardışık zaman örüntüleri yaygın şekilde kullanılmaktadır (Akpınar 2000: 5).

Veri madenciliği günümüzde giderek yaygınlaşmakta ve birçok alanda etkili bir şekilde kullanılmaktadır. Bunlardan bazıları; eğitim, sağlık, işletme, borsa, telekominasyon, mühendislik, bankacılıktır (Savaş ve diğerleri 2012: 2). Sağlık sektörü konusunun insan olması nedeniyle kendine has özellikleri olan bir hizmet sektörüdür ve uzmanların doğru ve güvenilir verilerle hızlı kararlar verebilmesi önemlidir. Sağlık sistemi politikalarının ve kararlarının amaçlarına uygun ve etkin olabilmesi; güvenilir, güncel ve doğru verilere ve bu verilerin karar destek sistemleri aracılığıyla kullanılmasına bağlıdır. Sağlık verileri hastaneler, diğer sağlık kurumları, sigorta şirketleri ve ilgili kamu kurumları başta olmak üzere birçok kuruluş tarafından toplanmaktadır (Koyuncugil, Özgülbaş 2009: 28).

Sağlık alanında veri madenciliği farklı yöntemler ve konu başlıkları altında son yıllarda yaygın bir şekilde kullanılmaktadır. Literatür taraması yapıldığında en çok sınıflama yöntemlerinin performans karşılaştırmaları ve kümeleme yöntemleri yaygın şekilde kullanılması dikkat çekmiştir. Veri madenciliğin kullanılan regresyon modellerinin sağlık alanında kullanımı ile ilgili çalışmalarda tek regresyon modelinin incelenmesi yada iki tanesinin performanslarının kıyaslanması kullanılmıştır. Örneğin Faridi A v.d. 2012 yılında yaptıkları bir çalışmada destek vektör regresyon ve sinir ağları modelleri ile performanslarını kıyaslamıştır. $R^2$ ve MSE (mean square error) göre en iyi performansı destek vektör regresyon yöntemi ile sağlamıştır.

Chayama K v.d. 2011 yılında yaptıkları çalışmada kronik hepatid C hastalarına bir terapi uygulayarak sonuçları çoklu lojistik regresyon, sınıflama ve regresyon ağacı ile duyarlılıklarına bakılmıştır. Veri madenciliğinde kullanılan yaklaşımların bu terapi sonuçlarının faydasını doğru bir şekilde ortaya koyduğu tesbit edilmiştir.

Leidy NK v.d. 2016 yılında KOAH hastalarına ait veriler kullanarak yapmış oldukları çalışmada; Random forest yöntemini kullanmıştır.

Basak D. v.d. 2007 yılında yaptıkları çalışmada destek vektör regresyon yöntemini teori, yöntem ve son yıllardaki gelişim hakkında detaylı bir şekilde ele almıştır.

Breiman 2001 yılında yayınladığı kitabında random forest yöntemini 13 bölüm halinde detaylı bir şekilde anlatılmıştır.

Veri madenciliğinde kullanılan regresyon modellerinin ve simülasyon yönteminin 1000 tekrarlı, farklı senaryolardaki sonuçları için performanslarının nasıl etkilendiğine dair literatürde çalışmalara rastlanamamıştır. Bu konuda çalışmalarda ise tek bir regresyon modeli ele alınmış ve sonuçlar genellenmiştir.

Bu tez çalışmasında veri madenciliğinin tanımı, temel kavramlarına, veri madenciliği yöntemlerine ve veri madenciliğinde kullanılan regresyon modellerine değinilmiştir. Esas olarak literatürde eksikliği fark edilen regresyon yöntemlerinden olan "Destek Vektör Regresyon", "Regresyon Ağaçları", "Random Forest" özelliklerinin açıklanması ve üç yöntemin kestirim performanslarının kıyaslanması ve gerçek veri ile herbir senaryonun 1000 kez tekrarlı simülasyon sonuçlarına göre regresyon modellerinin kestirim performanslarına bakılması, korelasyon yapısının düşük, orta, yüksek ilişkili durumuna göre farklı gözlem sayıları ile veri türetilip modellerin performanslarının değişip değişmeyeceğinin belirlenmesi amaçlanmıştır.

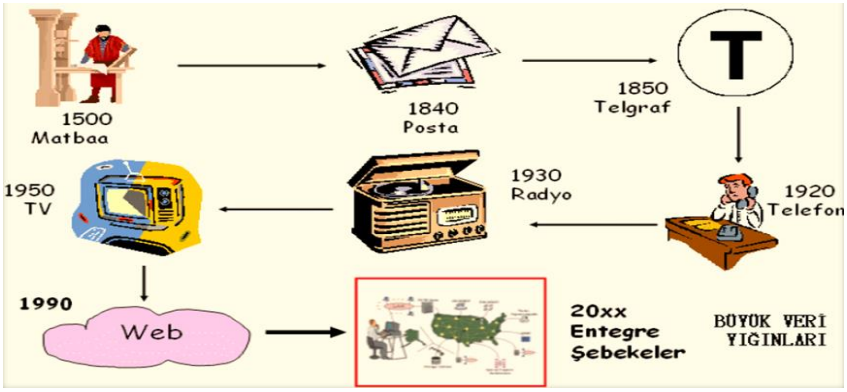
# 1. BÖLÜM

# VERİ MADENCİLİĞİ

Bu bölümde; veri keşif sürecinin en önemli kısımlarından olan veri, bilgi, özbilgi ve bilgelik kavramlarından yola çıkılarak veri madenciliğinin tanımı, veri madenciliğiyle ilgili disiplinler, veri madenciliğinde yaşanan sorunlar, veri madenciliği süreçleri, veri madenciliği modelleri ve son olarakda veri madenciliğinde kullanılan kestirim yöntemleri üzerinde durulmaktadır.

## 1.1. Veri Keşfi Süreci ve Temel Kavramlar

Günümüzde güçler dengesinin bilgi üzerinde yoğunlaştığı bilinen bir gerçektir. Farklı yöntemlerle çeşitli kaynaklardan elde edilen bilgilerin, yine belirli bir disiplin ve sistem içerisinde analiz edilmesiyle ortaya çıkan sonuçlar; teknolojik, ekonomik, siyasi, toplumsal ve askeri alanlar gibi çeşitli uzmanlık dallarında aktif olarak kullanılmaktadır. Elde edilen bilgiyi tam zamanında, rasyonel ve doğru olarak kullanan işletmeler hedefledikleri sonuca kısa yoldan ve hızlı bir şekilde erişmektedirler (Güllüoğlu 2011: 1). Bu erişimde veri madenciliği çok önemli bir rol oynamakla birlikte, öncelikle veri keşif süreci elemanlarından olan; veri, bilgi, özbilgi ve bilgelik kavramlarının tanımlanması konunun temelini teşkil etmektedir.



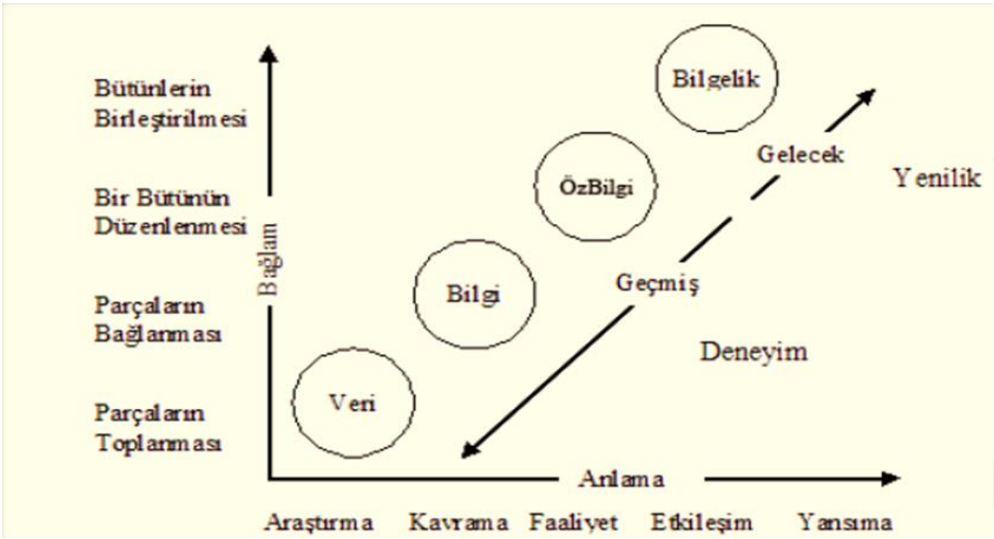

**Şekil 1.1.** Gelişen Bilgi Teknolojisi ve Veri Oluşumu

**Kaynak**: (Forboudi, 2009: 9)

Veri; bilgisayara giriş yapılabilen herhangi bir metin veya rakamsal ifade olarak tanımlanabilmektedir. Verilerin düzenlenmiş, yani işlenmiş haline ise bilgi adı verilmektedir. Başka bir ifadeyle ham verilerin işlenmesi sonucu bilgiye ulaşılmaktadır. Örneğin işletmelerin satış faaliyetleriyle ilgili veriler, hangi tür ürün veya hizmetlerin satıldığı ve bunların ne zaman satıldığı gibi daha birçok şey hakkında bilgi sağlayabilmektedir. Elde edilen bu bilgilerin daha öz bir forma kavuşturulması ile elde edilen bilgi ise özbilgi olarak ifade edilmektedir. İşletmelerin gelecek projeksiyonlar ve trendler hakkında öngörüde bulunabilmek için bilgiyi, özbilgi haline dönüştürebilmektedirler. Örneğin işletmeler satışlarıyla ilgili olarak özet bilgileri sayesinde müşterilerinin tüketim profillerini ve satın alma davranışlarını analiz edebilmekte ve bu özet bilgiler yardımıyla promosyona veya reklama ihtiyaç duyup duymadıklarını tespit edebilmektedirler. Kısaca öz bilgi; bilginin hacim itibariyle oldukça küçültülmüş, fakat kullanım değerinin ise oldukça arttırılmış bir türüdür (Tüzüntürk 2010: 65-66). Veri bir yöntem kullanılarak, tutarlı bir bağıntıya konulduğunda anlam çıkarmaya elverişli "ham olgu" dur. Ham olgu olarak ifade edilen verinin anlamlı sonuca dönüştürülebilmesi için ise bir dönüşüm sürecinden geçirilmesi gerekmektedir. Verinin bir dönüşüm sürecinden geçirilerek anlamlı sonuçlara dönüştürülmüş şekli ise "bilgi" olarak ifade edilir (Karakaya 1994:14).

Veri ve bilgi terimleri günlük yaşamda genellikle birbiri yerine kullanılabilmektedir. Fakat eş anlamlı olarak kullanılan bu iki terim aslında birbirinden çok farklı kavramları ifade etmektedir. Bu bağlamda veri; bir amaç doğrultusunda bir araya getirilmiş sistematik gerçekler bütünü, her türlü işaret, harf, rakam ve semboller kümesidir. Dolayısıyla bilgiye erişim için kullanılan ham gerçekler ve malzemeler, veri olarak ifade edilmektedir. Bilgi ise; karar verme sürecinde anlamlı bir biçimde işlenmiş veridir. Bilginin içeriği veri olmasına rağmen, tüm veriler bilginin oluşumunda kullanılmamaktadır. Verilerin düzgün, organize ve sistematik bir şekilde oluşturulması bilginin oluşumunda son derece önemlidir. Bir kişi için faydalı olan bilgi, diğer bir kişi için önemsiz olabilmektedir. Toplanan verinin amaca uygun bir biçimde faydalı bir bilgiye evrilmesi için konuyla ilgili yeterince detayı barındırması, doğrulanabilir olması, güncel ve zamanında elde edilebiliyor olması, tam olması, diğer veri kaynaklarıyla uyumlu olması ve kolay işlenebilme gibi özelliklere sahip olması gerekmektedir (Kelleci 2011: 3).

Öztürk'e göre " Hayatın her alanında kayıt altına alınan tüm olgu ve durumlara veri denir. Verinin en büyük özelliği ham olmasıdır. Veri tek başına bir anlam ifade etmez. Verilerin; ilişkilendirilmiş, düzenlenmiş, anlamlandırılmış ve işlenmiş haline bilgi denir. Bilgi, insanoğlunun aklının erdiği olgulara verilen addır. Bilgi düzeyinin en üst basamağında bilgelik yer alır. Kişilerin yetenek ve deneyimleri ile bilgelik yakından ilişkilidir" (2014: 11).




**Şekil 1.2.** Veri, Bilgi, Özbilgi ve Bilgelik Hiyerarşisi

**Kaynak**: (Terlemez 2008: 7)

## 1.2. Veri tabanı ve Veri Ambarı Kavramları

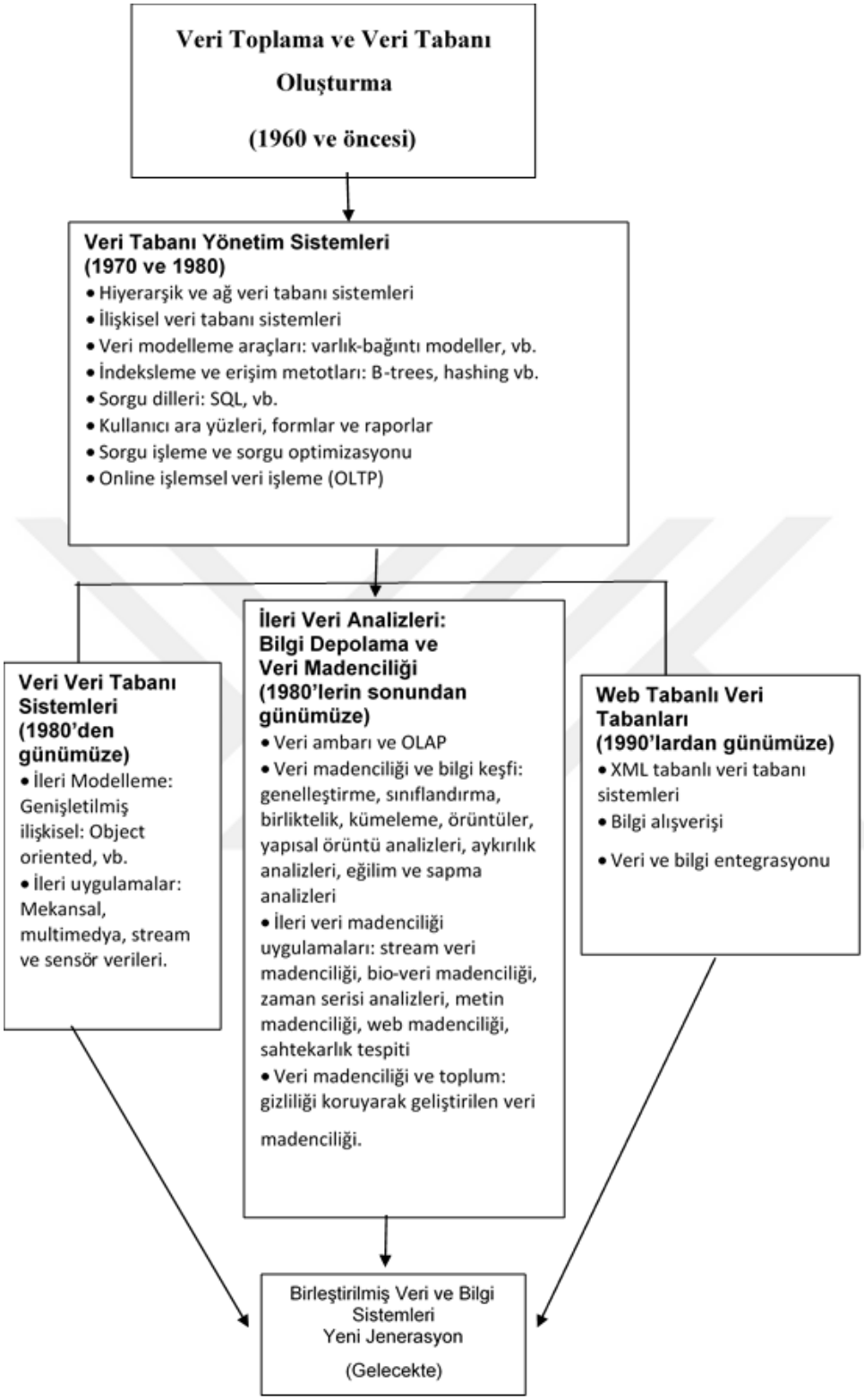
Bilgisayar teknolojisiyle ortaya çıkan veri tabanı, büyük miktardaki verileri dijital ortama güvenli, yapısal ve oldukça hızlı bir şekilde depolamaya, sorgulamaya ve kullanmaya yarayan bir yazılım sistemidir. Burada verilerin yapısal özellikleri ve birbirleriyle ilişkileri de saklanmaktadır. Günümüzde veri ağırlıklı bir biçimde kullanılan yazılımların büyük bir çoğunluğu, kullanmakta olduğu verileri veritabanlarında muhafaza etmektedir. Veri tabanlarında SQL (yapısal sorgu dili) bilmek gerekirken, günümüzde veri tabanları yönetim sistemlerini kullanarak SQL bilme zorunluluğu olmadan da birçok işlem grafik ortamında kolaylıkla yapmaya olanak tanımaktadır. Veri tabanları; ilişkisel, ilişkisel olmayan, nesne, nesne-ilişkisel ve masaüstü olmak üzere genel olarak beş grupta incelenmektedir (Asilkan 2008: 4-

5). Veri tabanı sistemleri yazılımları bilgisayar sistemlerinin önemli bir bileşeni olarak değerlendirilir. Veri tabanı yönetim sistemleri, birbiriyle ilişkili veri ve programlar topluluğundan oluşmaktadır. Veri topluluğu bir "veri tabanı" olarak değerlendirilir. Veri tabanı bir kuruluşa ilişkin bilgilerin yer aldığı ortamdır (Özkan 2013: 14).

Veri tabanı kavramı için sistematik erişim imkânı olan, yönetilebilir, güncellenebilir, taşınabilir, birbirleri arasında tanımlı ilişkiler bulunabilen bilgiler kümesidir. Veri tabanı, birbiri ile ilişkili veriler topluluğudur. Veri tabanı sadece veriler yığınını değil, bunlar arasındaki ilişkileri de saklar. Veri tabanları ve veri tabanı yönetim sistemleri bilgi kaynaklarının saklanmaları, korunmaları, güvenliklerinin sağlanmaları ve örgüt için gerekli zamanda gerekli bilginin kolaylıkla edinimleri konusunda büyük avantajlar sağlamaktadırlar (Şengül 2010: 70).

Veri tabanları standart özellikleri içerir. Bu özelliklerden bazıları; veriler depolayabilmelidir, Son kullanıcılar verilere istedikleri zaman erişip, kullanabilmelidirler, Veri tabanı güvenliği ile veriler korunmalıdırlar. Bir veri tabanı veri güncelleme, yeni veri ekleme, silme ve veriyi istenen şekilde kullanabilmek için çeşitli yöntemler içeren nispeten kolay bir veri yönetimine sahip olmalıdır, Bir veri tabanında tutulan veriler doğru olmalıdır. Aynı veri gerekli olmadıkça tekrar etmemelidir. Veri tabanları giderek artan verilere daha kolay erişim sağlamalıdır (Stephens, Plew 2003: 73).

Veri tabanı kullanımı, geleneksel dosya kullanımına göre bir çok yönden üstünlük sağlamaktadır. Veri tabanları, veri madenciliği sorgularına temel oluşturacak girdilerin sağlanması amacıyla oluşturulan yapılardır. Veri tabanındaki sorgu cümlecikleri, veri madenciliğinin istediği örneklem kümesini elde etmek için kullanılmaktadır. Dolayısyla veri madenciliği, veri tabanından farklıdır; çünkü veri tabanındaki örüntüler için sorgular çalıştırılırken, veri madenciliğindeki sorgular genellikle keşfe dayalı ve ortada olmayan örüntüleri saptamaya yöneliktir (Gürbüz 2009: 33). Bir veri tabanı; tablolar, formlar, raporlar, sorgular, makrolar ve modüller gibi toplam altı farklı elementten meydana gelmektedir (Şık 2014: 4).




**Şekil 1.3.** Veri tabanı Teknolojisinin Evrimi

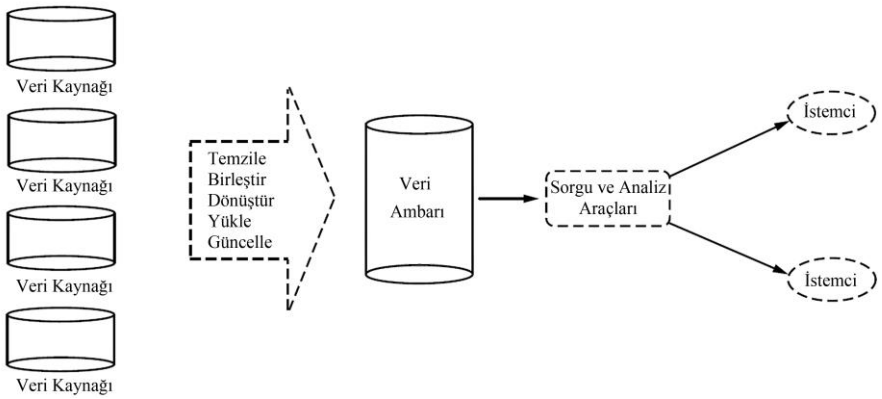
**Kaynak:** (Dolgun 2014: 9)

Veri ambarı; bir işletmede karar alıcı pozisyonunda olan personelin veya yöneticinin istediği bilgilere ulaşmasını sağlamak gayesiyle düzenlenen, özellikle sorgu işleme sistemlerinin ihtiyaçları için geliştirilmiş bir ya da bir grup ilişkisel veri tabanı yönetim sistemidir. Bu bağlamda veri ambarı; sorgulama ve analiz için uygun ve entegre bir bilgi havuzuna benzetilebilmektedir (Kolay 2006: 22).

Veri ambarı, kaynak sistemlerinden boyutlandırılmış veya standartlaştırılmış bir veri deposuna verileri periyodik olarak getiren ve birleştiren sistemdir. Veri ambarları, veri madenciliği ile eşzamanlı olarak düşünülen ve bu sürecin gerçekleştirildiği veriyi tedarik eden, özelleştirilmiş bir yapıya sahip olan veri tabanıdır. Özetle veri ambarı; farklı olarak birçok kaynaktan ve özellikle de farklı formlarda verilerin depolandığı ve tümünün de benzer bütünleşik bir çatı altında kullanımının arzu edildiği yapılardır (Koyuncugil, Özgülbaş 2009: 23).

Veri ambarlarına dair ilk araştırmaların Inmon ve Kimball gibi araştırmacılar tarafından yapıldığı literatürde yer almaktadır. Veri ambarının ne olup olmadığıyla ilgili ilk tanımlamalar Inmon tarafından yapılmıştır. Bu bağlamda veri ambarı; bütünleşik, zaman dilimli, özne tabanlı ve yöneticilerin karar verme eylemlerinde onlara faydalı olacak şekilde bir araya getirilmiş olan ve değişmeyen veriler kümesidir (Bozkır 2009: 14).

Veri ambarları, veri temizleme, veri birleştirme, veri dönüştürme, veri yükleme ve periyodik veri güncelleme işlemleri ile oluşturulur. Şekil 1.4'te veri ambarının oluşturulması ve kullanılması şematik olarak gösterilmiştir




**Şekil 1.4** Veri Ambarının Temel Yapısı

**Kaynak**: (Han, Kamber 2006: 12)

Veri ambarı, kaynak sistemlerinden boyutlandırılmış veya standartlaştırılmış bir veri deposuna verileri periyodik olarak getiren ve birleştiren sistemdir. Veri ambarlarının tasarlanması komplike bir süreç olup, aşağıda yer alan basamakların takip edilmesini gerekli kılmaktadır (Düzgünoğlu ve diğerleri 2006: 112).

1- Mimarinin belirlenmesi, kapasitenin planlanması ve veri tabanıyla ilgili araçların seçilmesi,
2- Sunucuların, saklama birimlerinin ve istemcilerin entegrasyonunun sağlanması,
3- Veri ambarına ait grafik ve görüntülerin tasarlanması,
4- Fiziksel ambar organizasyonu süreçlerinin tanımlanması,
5- Veri kaynaklarına bağlanılması,
6- Verilerin çekilmesi, temizlenmesi, taşınması ve güncelleştirilmesi,
7- Nihai kullanıcı uygulamalarının tasarlanması,
8- Ambar ve ambar uygulamalarının birleştirilmesi.

## 1.3. Veri tabanı Sistemleri

Literatür ekseninde tarama yapıldığında farklı veri tabanı sistemleri olduğu bilinmekle birlikte, veri tabanı sistemlerinin genel olarak; OLTP ( Çevrimiçi Hareket İşleme), OLAP (Çevrimiçi Analitik İşleme) ve ROLAP (İlişkisel Çevrimiçi Analitik İşleme) olmak üzere üç farklı segmentte ele alındığı görülmektedir.

### 1.3.1. OLTP Sistemler (Çevrimiçi Hareket İşleme)

OLTP'ler 1980 lerden önce kullanılmaya başlanmıştır. Bir işletmenin günlük verilerinin işlendiği ortamlara OLTP sistemler denilmektedir. Bu sistemler, operasyonel süreçlerdeki verilerin kaydedilip depolanması amacıyla kullanılmaktadır. Bu sistemlere kayıtlı olan veriler temel alınarak çeşitli sayı ve türde belge ve raporlar meydana getirilmektedir (Gökmen 2015: 28).

Organizasyonda satın alma, kaydetme, muhasebe ve bankacılık gibi günlük işlemlerin yapıldığı işlemsel veri tabanı sistemleridir. Detaylı bilgi içerir ve ayrıntılı görüntüye sahiptirler. Veriye erişim sağlanabilmekte ve üzerinde oynama yapılmasına izin vermektedir. Saklanan kayıt sayısı sınırlıdır (Cevahir 2011: 10).

### 1.3.2. OLAP Sistemler (Çevrimiçi Analitik İşleme)

Veritabanları üzerinde çeşitli stratejik kararlar almaya yardımcı olacak analiz ve sorgu işlemleridir. Geleneksel sorgu ve raporlama araçları, veri tabanında "ne?" sorusuna cevap ararken, OLAP bir kademe daha ilerisine yönelir ve "niçin?" sorusunu ispatlamak için kullanılmaktadır. Örneğin; bir analist, kredi borcu ödeme güçlüğüne neden olan risk faktörlerini belirlemek istiyor. Öncelikle düşük gelirli kişilerin kredi riskinin yüksek olacağı şeklinde bir hipotez ileri sürebilir ve veri tabanını bunun doğruluğunu göstermek için analiz edebilir. Eğer doğruluğunu ispatlayamazsa hipotezini değiştirir. Yüksek borç sahibi olmanın risk faktörü olduğunu düşünerek bunu doğrulamaya çalışır. Eğer bunu da doğrulayamazsa her iki faktörün birlikte, kredi riskinde etkili olduğu tezini araştırabilir. Yani; analist örüntü ve ilişkilerle ilgili bir seri hipotez üretir ve bunların doğruluk veya yanlışlığını ispat etmeye çalışır. Bu yüzden OLAP, tümdengelimsel bir işlemdir; ancak incelenmesi gerekli değişken ve parametre sayısı düzinelerce, yüzlerce olduğu zaman etkili hipotezler ileri sürmek ve bunları OLAP ile doğrulamak çok daha zorlaşır (Kocabaş 2010: 11).

OLAP, ilişkisel raporlama ve veri madenciliği konularını kapsayan iş zekasının konusudur. Çok boyutlu çevrede veri analizini destekleyen ve sorgu temelli bir yöntem olan bu metot, ham veriden dönüştürülmüş bilgiye yönelik incelemelerin hızlı ve interaktif bir biçimde yapılmasını sağlamaktadır. Birçok iş alanlarında sıkça kullanılan OLAP; verilere stratejik anlam katan bir dönüşüm aracı olarak tanımlanabilmektedir (Esen 2009: 56).

OLTP uygulamalarının genel amacı iş dünyasındaki güncel verilerin toplanmasıdır ve veri tabanına işlenmesidir. Bu veri tabanınadaki kayıtlar güncellenebilir, okunabilir, silinebilir ve yeni kayıtlar eklenebilir ama veri ambarındaki veriler okunabilir, yeni veri eklenmez ve silinmezler. Bütün şirketler için temel olan bilgi karlılıktır. Müşteri ile ilgili bilgilerin, üretimle ilgili bilgilerin, bu üretim için harcamalarla ilgili bilgilerin tutulduğu temel amacı veri toplamak olan uygulamalara OLTP uygulamaları denilebilmektedir. Örneğin; ERPL (Enterprime Resource Planning-Kurumsal Kaynak Planlaması) sistemleri ayrı ayrı satıcılardan almak yerine konsolide edilmiş ve şirketin yukarıdaki süreçlerin çoğunu sağlamayı

hedefleyen en geniş kapsamlı OLTP uygulamalarıdır (Arslan, Yılmaz 2010: 76; Imhoff ve diğerleri 2003: 299).

Tablo 1.1. e göre OLTP ve OLAP sistemleri arasındaki temel farklar verinin işleyişine göre özetlenmiştir.

**Tablo 1.1.** OLTP ve OLAP Sistemleri Arasındaki Farklar

| | OLTP (Çevrimiçi İşlem Süreci ) | OLAP (Çevrimiçi Analitik Süreç) |
|---|---|---|
| Verinin Kaynağı | OL TP'ler verinin esas kaynağıdır | OLAP verisi çeşitli OLTP veri tabanlarından gelir |
| Verinin Amacı | Temel iş görevlerini kontrol etmek ve yürütmek | Planlamaya, problem çözmeye ve karar desteğe yardımcı olmak |
| Veri Nedir? | Devam etmekte olan iş sürecinin anlık bir görüntüsünü ortaya koymak | Çeşitli iş aktivitelerinin çok boyutlu görüntüleri |
| Fonksiyon | Günlük operasyonlar | Uzun vadeli bilgi gereksinimleri, karar destek |
| Eklemeler ve Güncellemeler | Son kullanıcılar tarafından gerçekleştirilen kısa ve hızlı eklemeler ve güncellemeler | Periyodik olarak uzun süren toplu veri tazelemeleri |
| Sorgular | Nispeten standartlaştırılmış ve basit sorgular | Çoğu zaman birleştirmeler içeren karmaşık sorgular |
| İşleme Hızı | Genellikle çok hızlı | İşleme giren veri miktarına bağlı; toplu veri tazelemeler ve karmaşık sorgular saatlerce sürer |
| Erişilen Kayıt Miktarı | Onlarca | Milyonlarca |
| Kullanıcı Sayısı | Binlerce | Yüzlerce |
| Alan İhtiyacı | Eğer tarihsel veri arşivlenmişse nispeten küçük olabilir | Birleştirme yapılarının ve tarihsel verinin bulunmasına bağlı olarak daha büyük olabilir |
| Veri Tabanı Tasarımı | Pek çok tablo ile son derece normalize edilmiştir | Genel olarak daha az tablo ile denormalize edilmiştir |

| | | |
|---|---|---|
| Yedekleme ve Kurtarma | Düzenli olarak yedekler, operasyonel veri hayati öneme sahiptir ve veri kaybı mıhtemelen önemli miktarda mali kayba ve borçlanmaya neden olur | Düzenli yedeklemeler doğrultusunda, bazı ortamlar kurtarma yöntemi olarak OLTP verilerini geri yükleyebilir |

**Kaynak**: (Şık 2014: 17)

### 1.3.3. ROLAP Sistemler ( İlişkisel OLAP)

Dinamik çok boyutlu verilerin analizinde tercih edilen ilişkisel OLAP olarak da adlandırılan ROLAP; yüksek miktardaki veri, veri ambarlarında kullanılabilmektedir (Esen 2009: 60).

## 1.4. Veri Madenciliği Kavramına Genel Bakış

Veri madenciliği terimini net olarak anlayabilmek için öncelikle kelimelerin yalın olarak ne ifade ettiğini anlamak gerekmektedir. Madencilik; yeryüzündeki saklı kalmış ve değerli kaynakların günyüzüne çıkartılması süreci olup, bu sözcüğün veri kelimesi ile ilişkilendirilmesi ise veri yığınları arasında ilk etapta farkına varılamayan değerli bilgilerin saptanması ve ortaya çıkartılması düşüncesine dayanmaktadır (Irmak ve diğerleri 2012: 102).

Veri madenciliği, büyük verileri analiz ederek, verileri anlamlı hale getiren bilgisayar destekli işlemlerdir. Veri madenciliği terimi büyük veri tabanları içerisindeki değerli bilgiyi arama ve değerli bir maden cevherine sahip dağdaki madeni kazıp çıkartma arasındaki benzerlikten türemiştir. Her ikisi de ya uçsuz bucaksız miktardaki materyali eleme veya değerli olanı bulmak için derinlemesine araştırma gereksinimi duyan işlemlerdir (Akçay 2014: 34-35).

Teknolojik gelişmelere bağlı olarak işletmelerin depoladığı veri sayısı da günden güne artmaktadır. Veri madenciliği'de bu gelişmelere paralel olarak farklı yöntemlerin kullanılması yolu ile literatürde yerini almaktadır. İş hayatıyla ilgili her ortamda; bilhassa satış ve pazarlama alanlarında her geçen gün yeni bir uygulama faaliyete geçmektedir. Müşterilerin satın alma örüntülerinin belirlenmesi, müşteri ilişkileri yönetimi, müşteri terk ve pazar sepeti analizi gibi uygulamalar, aktif olarak

kullanılan uygulamalardan bazılarıdır (Çınar, Silahtaroğlu 2012: 390). En basit biçimde veri madenciliği büyük miktarlardaki veriden bilgi çıkartma olarak tanımlanabilmektedir. Veri madenciliği, pek çok analiz aracı kullanımıyla veri içerisinde örüntü ve ilişkileri keşfederek, bunları geçerli tahminler yapmak için kullanan bir süreçtir. Veri madenciliği bir veya daha fazla makine öğrenme tekniğinin uygulanarak otomatik olarak bir veri tabanı içinde bulunan verilerden bilgi çıkartılması, verilerin analiz edilmesi işlemidir (Akyol ve diğerleri 2012: 313).

Veri madenciliği, işlenmemiş verinin tek başına sunamadığı bilginin gün yüzüne çıkarılması ve veri analizi sürecine dahil edilmesidir (Jacobs 1999: 8). Bilinmeyeni tahmin etmede oldukça başarılı olan temel değişkenlerin, milyonlarca diğer değişkenlerden arındırılmasını sağlama becerisidir (Kitler, Wang 1998: 45). Gelecekle ilgili tahminlerde bulunabilmeye olanak sağlayabilecek potansiyele sahip bağlantıların, büyük çaplı ve kapasiteli veri yığınları arasından, bilgisayar programları vasıtasıyla aranması faaliyetidir (Doğan, Türkoğlu 2008: 86).

Literatürde veri madenciliğine ilişkin çeşitli tanımlar yapılmıştır.

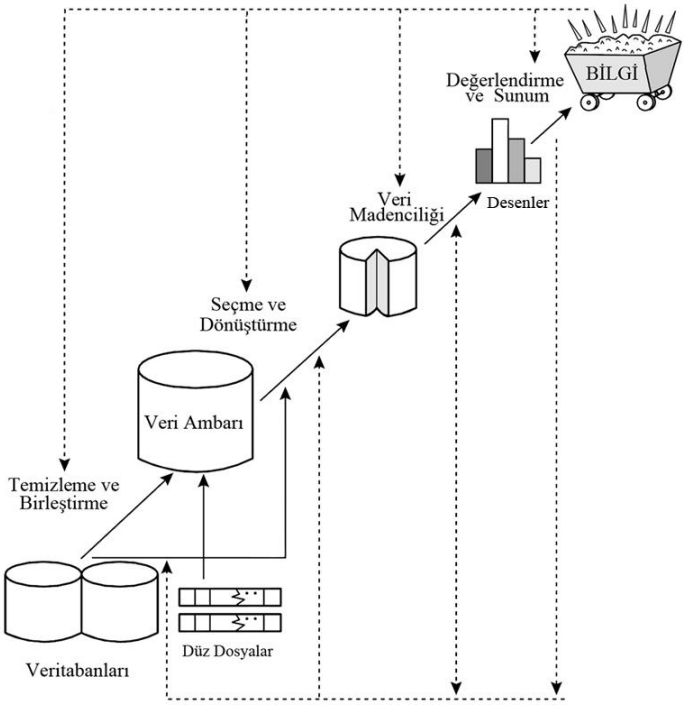
Cabena ve arkadaşlarına göre “Veri madenciliği, geniş veritabanlarından bilgi çıkarımını hedeflemek için makine öğrenimi, örüntü tanıma, istatistik, veri tabanı ve görselleştirme tekniklerini bir araya getiren disiplinler arası bir alandır” (1998: 517)

Witten ve Frank ‘a göre “Veri madenciliği veride var olan örüntüleri keşfetme sürecidir. Süreç otomatik veya (daha çok) yarı otomatiktir. Keşfedilen örüntüler anlamlı olmalıdır ve genellikle ekonomik avantaj olmak üzere fayda sağlamalıdır” (2005: 9).

Gartner Grubuna göre "Veri madenciliği örüntü tanıma teknolojilerinin yanı sıra istatistiki ve matematiksel teknikleri kullanarak depolarda saklanan büyük çaptaki veriler yoluyla eleme yapmak suretiyle anlamlı yeni korelasyonlar, örüntüler/kalıplar, ve trendler keşfetme sürecidir" (Silahtaroğlu 2008: 10).

Veri madenciliği ve veri tabanlarında bilgi keşfi (Knowledge Discovery in Databases - KDD) süreci kavramları birçok kaynakta birbirinin yerine kullanılmaktadır. Veri madenciliği, veri tabanlarında bilgi keşfi sürecinde önemli bir adım olmasına rağmen birçok çalışmada tüm süreci anlatmak için kullanılmaktadır.

Veri madenciliği ile büyük veri yığınlarından veri tabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veri tabanı teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır (Akpınar 2000: 2).




**Şekil 1.5** Veri Tabanlarında Bilgi Keşfi Sürecinin adımları

**Kaynak: (**Han, Kamber 2006: 6)

Veri madenciliği istatistiksel bir yöntem olarak ele alınsa da bazı noktalarda geleneksel istatistik yönteminden ayrılmaktadır. Veri madenciliği, mantıksal kurallara ve görsel sunumlar haline dönüştürülebilen nitel modellerin oluşturulabilmesi amacı taşımakta, bilgisayar ve insan arayüzünün entegrasyonu olarak tanımlanabilmektedir (Ekim 2011: 8). İstatistikçiler veri tabanlarını inceleyip, istatistiksel açıdan önemli ilişkiler aramaktadır. Veri madenciliği, bu süreci otomatik olarak gerçekleştirmektedir (Koyuncugil, Özgülbaş 2009: 24).

Veri madenciliği yöntemleri ilk olarak marketlerden yapılan alışverişlerin analizinde kullanılmış, marketler bu yöntemleri kullanarak yapılan alışverişlerde birlikte alınan ürünleri incelemiş ve bu ürünleri market raflarında birbirlerine yakın olacak şekilde yerleştirmişlerdir. Promosyon yapıldığı dönemlerde bu ürünler beraber

kullanılarak müşterilerin dikkatini çekmek ve tüketim faaliyetlerini artırmak için kullanılmıştır. Ayrıca bu analizler optimum stok ve envanter yönetimi için gerekli olan bilgiyi de sağlamışlardır (Akçetin, Çelik 2015: 7).

Bilişim teknolojilerinin yeterli olarak gelişmediği ve dolayısıyla işletmeler için kullanılmadığı zamanlarda, işletmelerde yürütülen faaliyetleri kayıt ve kontrol işlemleri manuel olarak yapılmaktaydı. Günümüz şartlarında bu denli büyük çaplı verilerin kayıt işlemleri, özellikle bilgisayar alanındaki gelişmeler sayesinde gigabayt ve terabayt genişliğindeki hard disklere kaydedilebilen bu verilerin manuel olarak kontrolü zaman kısıtlarından dolayı mümkün olmamaktadır. Her ne kadar excel ve access gibi ofis programları kullanılsa da bunlarda yetersiz kalmaktadır (Terzi 2012: 52).

Veri madenciliği; yeni ortaya çıkmış ve yeni kimlik kazanmış bir alan olmasına rağmen, bu yaklaşımın kökenleri neredeyse otuz yıllık bir geçmişe sahip olan araştırma ve uygulama geleneğine dayanmaktadır. Bu dönem süresince; SAS (Statistical Analysis Software), SPSS (Statistical Package for the Social Sciences) ve IBM (International Business Machines) gibi firmalar istatistiki analizin liderleri olarak var olmuştur. Günümüzde de bu işletmeler veri madenciliği alanında son derece aktiftirler ve yıllara dayanan sektör tecrübeleri sayesinde geliştirilmiş yüksek düzeyde kabul gören ürünleri mevcuttur (Irmak 2009: 6).

Veri madenciliği; işletmenin sahip olduğu veri, enformasyon kaynaklarında yöneticilerin veya analistlerin sormayı akıl edemediği sorulara, işletme hakkındaki yanıtların aranması işlemidir. Veri madenciliği'de temel amaç; büyük çaplı verilerdeki örüntüleri tespit etmeye yarayan algoritmaların kullanılarak, insanoğlunun kesinlikle hayal dahi edemeyeceği yeni trendlerin keşfedilmesini sağlamaktır. Veri madenciliği sadece bir bilim değil, aynı zamanda da bir sanattır. Çünkü veri madenciliği hipotezleri keşfeder, fakat bulguları entegre etmek için insanoğlunun becerilerine ihtiyaç duymaktadır (Çetin 2009: 4).

Veri madenciliği; bir veri kümesine uygulanan korelasyon, istatistik veya bir sınıflama raporu değildir, geleneksel metot analizleri ile ortaya çıkarılamayan ve dolayısıyla değerlendirilemeyen verilerden yeni örüntü ve eğilimlerin ortaya çıkarılması işlemidir. Veri madenciliği, kişinin sezgilerine dayalı, işlemleri tekrarlama

konusundaki sabrını zorlayan ve esnek düşünme yeteneği gerektiren bir tekniktir. Buna mukabil bazı durumlarda metodoloji dikkatli bir şekilde uygulanırsa, hemen örüntü almakta mümkündür (Özdemir 2004: 45). Veri madenciliğinde verinin analizi ve yazılım tekniklerinin kullanılarak veriler arasındaki ilişkiyi, kuralları ve özellikleri, daha önceden fark edilmemiş veri desenlerini tespit edebilmek için uygun bilgisayar yazılımlarının kullanılması gerekmektedir (Akgöbek, Kaya 2011:238).

Günümüzün ekonomik şartları ve hızla değişim gösteren iş ortamlarında, yöneticiler; gün içerisinde onlarca karar almak durumunda kalmaktadırlar. İş deneyimi ve sezgilere dayalı olarak alınan kararlar ise yöneticileri rasyonel olmayan karar verme eylemlerine sevk etmektedir. Dolayısıyla bu tür durumlarda kişilerin yanlış karar alma oranları da oldukça yüksektir. Bu bağlamda riski azaltmanın tek yolu ise bilgiye dayalı yönetimi öngören karar destek sistemlerinin kullanılmasıdır. Her ne kadar veri madenciliği ile ilgili teknikler gerçek anlamda bir karar destek sistemi teşkil etse de, bilgi teknolojileri olmadan tek başına bir varlık gösterememektedir. Dolayısıyla veri madenciliği bilgi teknolojilerinden sonuna kadar faydalanmak gerekir. (Savaş ve diğerleri 2012: 7).

Erşahin ve Argüden' e göre "veri madenciliği, özellikle kar ve pazar payı elde edebilmek için yoğun rekabetin yaşandığı pazarlama alanında ortaya çıkmaktadır. Hangi müşteri, hangi ürünü, ne zaman satın alabilir, kimler tedarikçilerinden vazgeçmekte ve bu tür müşterileri vazgeçirmek için neler yapılabilir, ürünün değerini yitirmesine hangi değişkenler neden olmaktadır, vb. soruların cevapları veri yığınlarının altındadır ve cevapları bulabilmek için veri madenciliği çözümleri gereklidir" (2008: 16-17).

Veri madenciliği; karar destek araçlarının niteliklerini yücelten, veri tabanlarına gizlenmiş bilgileri bulan ve iş uzmanları için kavrayış dağıtıcı bir sistem ve süreç olarak ifade edilmiştir (Albayrak, Yılmaz 2009: 1). Bunların dışında; veri tabanı analizi ve karar verme desteği; kalite kontrol, rekabet analizi, öngörü, kurumsal kaynakların optimum seviyede kullanılması, müşterilerin kredi risk incelemesi ve haber kümeleri gibi çeşitli uygulama örnekleri de literatürde yer almaktadır (Baykal 2006: 96).

Veri madenciliği, bilgi teknolojilerindeki doğal gelişim sürecinin bir yansıması, hatta sonucu olarak görülebilmektedir. Oldukça büyük ölçekli veriler, farklı alanlardaki büyük ölçekli veri tabanları içinde kıymetli verileri barındıran bir veri madeni gibi ele alınabilmektedir. Veri madenciliği'de; bilgisayar, makine öğrenmesi, veri ambarı yönetimi, algoritmalar ve analiz teknikleri kullanılarak bu süreç gerçekleştirilmektedir (Albayrak, Yılmaz 2009:2).

Veri madenciliği; veri toplamak, mevcut verilerden sorgulamalar yapmak veya ileri istatistiksel metotların kullanımının ötesinde bir noktadır. Örneğin; bir restoran zincirinde, hangi şubenin ne kadar ciro ettiği, hangi ürünlerin hangi şubelerde daha çok satıldığı, hangi günlerde veya saatlerde müşteri yoğunluklarının yaşandığı, hangi özellikteki müşterilerin devamlılık gösterdiği ve hangi bölgelerde performans düşüklüğü yaşadıklarının analiz edilmesi veri madenciliğinin konusu değildir. Veri madenciliği, yüzlerce değişkenin olduğu, bu değişkenler arasında sadece rakamsal değişkenlerin değil, kategorik değişkenlerin de olduğu milyonlarca veriden ancak doğru algoritmalar ve güçlü bir bilgisayar ile sonuca erişmenin olanaklı olduğu modelleri kurmaktır (Erşahin, Argüden 2008: 17).

Serek ve Ata'ya göre "veri madenciliği yalnızca bir takım araç ve tekniklerden ibaret olmayıp, veri toplama, veri temizleme, model oluşturma, model testi ve uygulama gibi birçok aşamaları içeren bir süreci ifade etmektedir. Ayrıca bu aşamaların tümünde, bilgisayarlar tarafından otomatik olarak gerçekleştirilemeyeceği insanın yorum ve katkısının çok önemli olduğu unutulmaması gereken önemli bir husustur" (2010: 71).

## 1.5. Veri Madenciliğinin Tarihsel Gelişimi

İlk bilgisayarların sayım amacıyla kullanılması 1950'li yıllara denk gelmektedir. 1960'larda ise veri tabanı ve verilerin depolanması kavramı teknolojik hayatta yerini almaya başlamıştır. Basit düzeyde öğrenmeli bilgisayarların geliştirilmesi ise 1960'lı yılların sonlarına doğru gerçekleştirilmiştir (Savaş ve diğerleri 2012: 4).

Veri madenciliği, kavram olarak ilk defa 1960'lı yıllarda ortaya çıkmıştır. Bu yıllarda veri madenciliği genellikle veri taranması ve veri yakalanması gibi isimler

almıştır. İlişkisel veri tabanı yönetim sistemlerinin kullanımı 1970'lerde ortaya çıkmıştır. Bu dönemlerde bilgisayar mühendisleri tarafından basit sistemleri olan uzman sistemler geliştirilmiş ve basit düzeyde makine öğrenimi sağlanmıştır. 1980'lerde veri tabanı yönetim sistemleri yaygın hale gelmiş ve çeşitli alanlarda kullanılmaya başlanmıştır. Bu dönemlerde işletmeler rakipleri, müşterileri ve ürünleriyle ilgili büyük çaplı verilerden oluşan veri tabanlarını meydana getirmişlerdir. Veri madenciliği isminin ilk kullanımı 1990' lara dayanmaktadır. Bu dönemlerde bilgisayar mühendisleri geleneksel istatistiki metotlar yerine algoritmik modülleri kullanarak veri analizini başlatmışlardır. Veri tabanlarındaki veri miktarı bu dönemlerde oldukça artış gösterdiği için büyük çaplı veriler içerisinden faydalı bilgilerin nasıl elde edileceği üzerinde kafa yorulmaya başlanmıştır. Bu amaç doğrultusunda da veri madenciliğinde kullanılmak üzere ilk bilgisayar yazılım programı 1992 yılında ortaya çıkarılmıştır. Günümüze kadar veri madenciliği sürekli olarak gelişimini sürdürmüş ve oldukça geniş bir alanda faaliyetlerini sürdürmüştür (Çalış 2013: 6).

## 1.6. Veri madenciliğinin Önemi

Veri miktarı ve çeşitliliğinin artması analizlerin eyleme dönük ve rasyonel sonuçlar içerecek bir biçimde yapılmasını zorunlu kılmaktadır. İşletmeler ve iş dünyası arasında rekabetin yoğunlaşması değişim ve uyum sürecinin zorunlu kıldığı bu hızı yakalayabilmenin, müşteri odaklı olmanın ve verimliliğin önemini her zamankinden daha fazla arttırmıştır. Dolayısıyla bu süreçte birçok interdisipliner bilimlerden faydalanmak kaçınılmaz hale gelmiştir (Ata ve diğerleri 2008: 34).

Gelişen teknolojiyle birlikte insanoğlunun verdiği kararlar, yaptığı alışveriş işlemleri, tüm mali hareketleri, ziyaret ettiği internet sitelerinin yanı sıra görüntü ve konuşmaları bile çeşitli nedenlerden dolayı kayıt altına alınmaktadır. Kaydedilen bu veriler; ticari finansal, bilimsel veya toplumsal ve benzeri şekillerde her formatta olabilir. Bu verilerin otomatik bir şekilde analiz edilmesi, anlamlı bir hale getirilmesi ve nihai olarak ihtiyacı karşılayacak bir bilgiye dönüştürülmesi gereksinimi ortaya çıkmıştır ki, bu da veri madenciliği'ye olan ihtiyacı işaret etmektedir (Sever 2010: 2).

Günümüzde kurumlar büyük miktarlarda veri üretmekte fakat bu veriler içinde anlamlı ve yararlı bilgiyi ortaya çıkarmakta zorluk çekmektedirler. Geleneksel istatistik yöntemlerle büyük boyuttaki veriyi çözümlemek kolay olmadığı için verileri işlemek ve çözümlemek için özel yöntemlere ihtiyaç duyulmuştur. Bu noktada, veri madenciliği bu gereksinimi karşılamak üzere ortaya çıkmıştır (Alan 2012: 165).

## 1.7. Veri madenciliği Sürecinde Yaşanan Zorluklar

Veri madenciliği; veri ambarları, yapay zeka, matematik ve istatistik gibi çeşitli multidisipliner bilimleri bir araya getirip, onlardan faydalanma gereksinimi ortaya çıkarır. Dolayısıyla yüksek performans özelliklerine sahip bilgisayarlar ve uzman kullanıcılar tarafından verilerin analiz süreci gerçekleştirilir. Tüm bu işlemlerde her bir faaliyet için en doğru olan yöntemin kullanılması gerekir. Şayet ilk etapta elde edilen veriler temizlenmez, doğru depolanmaz ve iyi bir analiz sürecine tabi tutulmaz ise ortaya beklenmeyen, yanlış sonuçlar çıkabilir. Tüm bunlara ek olarak veri kaynaklarının da genellikle büyük çaplı olması da sistemsel, donanımsal ve zaman açısından çeşitli problemlere yol açabilmektedir. Sonuç itibariyle veri ambarlarının gün geçtikçe büyüyen bir yapıda olduğu dikkate alındığında veri ambarlarının bu büyümeyi tolere edebilecek esneklikte ve yapıda tasarlanması gerekmektedir (Tosun 2006: 10).

Veri madenciliği işletmeler açısından son derece önemli yararlar sağlamakla birlikte, birtakım problemleri de beraberinde getirmektedir. Veri halinde kalmış girdiler işlenemedikleri sürece, bireylere ve işletmelere fayda veya zarar sağlayamazlar. Fakat herhangi bir işletme, veriden yola çıkarak bilgiye ulaştığı andan itibaren bu bilgilerin hukuki ve ahlaki kuralları gözeterek kullandığı zaman fayda sağlamalıdır. Bu noktadan sonra asıl sorun; bireylerden veya işletmelerden sağlanan verilerin işlenerek bilgi elde edilmesi aşamasında, uygun kullanılıp kullanılmaması aşamasında ortaya çıkmaktadır. Oldukça güçlü bir rekabet unsuru olarak kullanılabilen veri madenciliği, aynı zamanda da kişilerin veya firmaların haklarını ihlal eden, bir silah olarak da kullanılabilmektedir. Bu noktada elde edilen bilgilerin üçüncü taraflara sunulmaması ve mesleki sırlarını ifşa etmemesi gibi unsurların dikkate alınması

gerekmektedir. Aksi takdirde kanunların gerektiği şekilde düzenlenmesi ve uygun yaptırımların gerçekleşme zorunluluğu ortaya çıkmalıdır (Meriç 2004: 48-49).

## 1.8. Veri Madenciliğinin Kullanım Alanları

Veri madenciliği tekniklerini işletmeler özellikle pazarlama alanında kullanmaktadır. Müşterilerin sosyoekonomik ve demografik özelliklerinin belirlenmesi başta olmak üzere, bununla birlikte satın alma davranışlarının belirlenmesi, mevcut müşterilerin elde tutulması ve sadık müşteri oluşturma faaliyetleri, yeni müşterilerin kazanılması ve satış tahminleri gibi uygulamalar ile çeşitli müşteri profilleri arasındaki ilişkiler incelenmektedir (Ayanoğlu ve diğerleri 2004: 1).

Veri madenciliğinin kullanım ve uygulama alanları; sigortacılık, bankacılık, elektronik ticaret, pazarlama, biyoloji, tıp, genetik, kimya, uzay bilimleri ve teknolojisi, görüntü tanıma ve robot görüş sistemleri, yüzey analizleri ve coğrafi bilgi sistemleri, meteoroloji ve atmosfer bilimleri, sosyal bilimler ve davranış bilimleri, bilimsel, mühendislik ve sağlık hizmetleri bakım verileri ile metin madenciliği olarak belirtilmiştir (Atak 2014: 6-9).

Veri madenciliği uygulama alanları sektörel, bilimsel ve mühendislik verileri, şehircilik ve planlama, sağlık verileri, iş ve işletme verileri, perakendecilik-marketçilik verileri, bankacılık,meteoroloji verileri, finans ve borsa verileri, eğitim sektörü verileri, internet (web) verileri, doküman verileri, taşımacılık ve ulaşım verileri, turizm ve otelcilik verileri, belediyeler ve telekommünikasyon olarak sıralanmıştır (Ertuğrul ve diğerleri 2013: 98-99).

Veri madenciliği uygulamaları günümüzde başta ABD (Amerika Birleşik Devletleri) ve AB (Avrupa Birliği) ülkeleri olmak üzere pek çok ülkede; üniversite başvurularının değerlendirilmesi ve öğrencilere burs bağlanmasına, işsizlik sigortaları değerlendirmelerine, vergi iadelerindeki ususlsüzlük tespitinden sosyal sigortalardaki hileli kullanımların tespitine, vergi sistemindeki değişikliklerin bütçeye olan etkisini öngörmekten bütçe likitidesinin yönetimine, vergi usulsüzlüklerinin tespitinden güvenlik ve savunma alanlarındaki uygulamalara kadar daha bir çok sahada yaygın olarak kullanılmaktadır (Alkan, Falay 2007: 8).

Bu bağlamda; konuyla ilgili olarak yapılmış olan yüksek lisans ve doktora tezlerinin de, akademik anlamda oldukça fazla olduğu ve istatistiki açıdan da fazlaca tercih edilen bir konu ve yöntem olduğu Tablo 1.2. ve Tablo 1.3. 'de net bir şekilde görülmektedir.

**Tablo 1.2.** 2010-2015 Yılları Arasında Veri madenciliğiyle İlgili Yayımlanmış Tez Sayıları

| Yıllar | Yüksek Lisans Tezi | Doktora Tezi |
|---|---|---|
| 2015 | 9 | 3 |
| 2014 | 70 | 17 |
| 2013 | 73 | 26 |
| 2012 | 59 | 16 |
| 2011 | 61 | 12 |
| 2010 | 68 | 5 |
| Toplam | 340 | 79 |

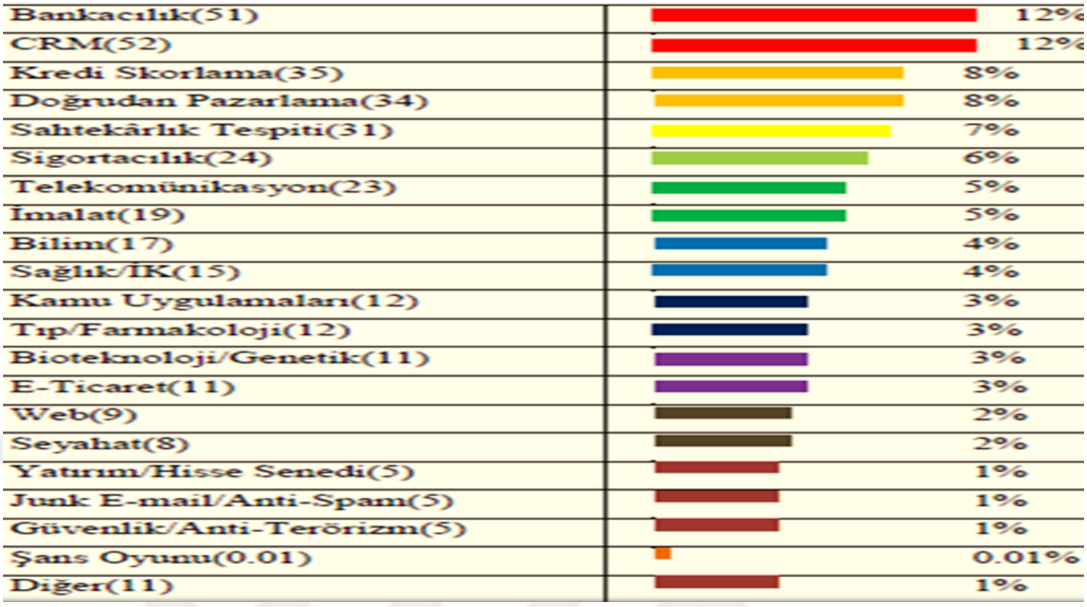
**Kaynak:** https://tez.yok.gov.tr (Erişim Tarihi: 21.07.2015).

**Tablo 1.3.** 2010-2015 Yılları Arasında Veri madenciliğiyle İlgili Yayımlanmış Tezlerin Alanları

| Alan | Yüksek Lisans Tezi | Doktora Tezi |
|---|---|---|
| Fen Bilimleri | 279 | 58 |
| Sosyal Bilimler | 54 | 16 |
| Tıp Bilimleri | 7 | 5 |
| Toplam | 340 | 79 |

**Kaynak:** https://tez.yok.gov.tr (Erişim Tarihi: 21.07.2015).

**Tablo 1.4**. Veri madenciliğinin Sektörler Bazında Kullanımı




**Kaynak: (**Saygılı 2013: 21)

## 1.9. Veri Madenciliğiyle İlişkili Disiplinler

Veri madenciliği aracılığıyla, büyük veri kümelerini içerisinde barındıran veri tabanı sistemlerinde gizli kalmış bilgilerin çekilmesi sağlanmaktadır. Bu işlem, istatistik, matematik, modelleme teknikleri, veri teknolojisi tabanı gibi birçok disiplinin bir araya gelmesi ile yapılmaktadır (Saygılı 2013: 18).

Veri madenciliği; veri tabanı teknolojisi, istatistik, bilgisayar bilimleri, makine öğrenimi, örüntü tanıma ve görselleştirme gibi pek çok teknik alan arasında köprü görevi gören çok disiplinli bir alandır (Özyirmidokuz, Kayseri: 16). Veri madenciliği her ne kadar multidisipliner bir alan olsa da, özellikle makine öğrenmesi ve istatistik üzerine yoğunlaşan bir alan olarak görülmektedir.


**Şekil 1.6.** Veri Madenciliğinin Diğer Disiplinlerle Olan İlişkisi

**Kaynak:** (Alagöz ve diğerleri 2014: 5)

### 1.9.1. Bilgi Bilimi

Bilgi bilimi veri madenciliği içinde çok sayıdaki köklü metotlardan biridir (Aggarwal, Wang 2010: 218). Bilgi bilimi uygulamalarında en çok kullanılan ölçümler geri çağırma ve kesinliktir. Arama motoru uygulamasında, geri çağırma; geri çağırılan ilgili sayfaların oranını ölçer. Kesinlik ise, ilgili sayfaların geri çağırılma oranını ölçer (Bramer 2007: 176).

Bilgi bilimi, belgeler veya belgelerin içindeki bilgiyi araştıran bilimdir. Belgeler yazı, multimedia ve web üzerinde bile olabilmektedir. Geleneksel bilgi yeniden çağırma (retrieval) ile veri tabanı arasındaki farklar şunlardır. Bilgi yeniden çağırmada; (1) araştırma verisi yapılandırılmamıştır, (2) sorgulamalar temelde kompleks yapısı olmayan anahtar cümleler ile şekillenebilmektedir (Han ve diğerleri 2012: 26).

Bilgi yeniden çağırmada tipik yaklaşımlar olasılıklı modelleri adapte etmektedir. Örneğin, yazı belgesi kelimelerin havuzu gibidir, bu belgede görülen kelimelerin çoklu setidir. Belgenin dil modeli belgedeki kelimelerin havuzunu oluşturan olasılık yoğunluk fonksiyonudur. İki belgedeki benzerlik karşılık gelen dil modelleri arasındaki benzerlik ile ölçülebilir (Han ve diğerleri 2012: 26).

### 1.9.2. Görselleştirme

Görselleştirme, işletme veri kümelerinden trendleri ve yeni örnekleri keşfetmek için veri analistleri ve iş yardımcısı için anahtardır. Görselleştirme karar vericiler için bu keşifler ile bağlantı kuran kanıtlanmış metottur (Soukup, Davidson 2002: 5).

Görselleştirme teknikleri kullanarak verinin analizi ve keşfi, diğer veri madenciliği tekniklerinin muaf tutarak yeterli ve yeni bilgi getirir. Görselleştirme fikirleri taşıma açısından ve inşaların algısında önemli rol oynayan görsellik için güçlü bir araçtır (Karahoca 2012: 154). Veri görselleştirmenin amacı gözlenen veriyi anlamayı yüksek seviyede elde etmektir (Shahbaba 2012: 5).

Veri görselleştirmede amaç, grafiksel gösterim olarak veri ile açık ve etkin bir şekilde bağlantı kurmayı amaçlar. Veri görselleştirme yoğun olarak bir çok uygulamada kullanılır. Örneğin çalışırken raporlamada, işletme operasyonlarını yönetirken, görev sürecini takip ederken daha popüler olarak görselleştirme tekniklerinin avantajlarını kullanarak veri ilişkilerini keşfedebiliriz aksi takdirde ham veriye bakarak kolayca gözleyemeyiz (Han ve diğerleri 2012: 56).

### 1.9.3. Veri madenciliği, makine öğrenmesi ve istatistik

Makine öğrenmesi, istatistik ve veri madenciliği arasında sıkı bir ilişki vardır. Bu üç disiplinde veri içindeki örüntüleri tespit etmeyi amaçlamaktadır. Makine öğrenmesinde yer alan yöntemler, veri madenciliği algoritmalarında kullanılan yöntemlerin çekirdeğini teşkil etmektedir (Saygılı 2013: 18).

Yapay zeka konusunun bir alt dalı olan makine öğrenmesi; veri eğitimi ile model oluşturulması ve analiz edilmesiyle ilgili bir alandır. Bilgisayar oyunları ve yapay zeka alanında öncü kişilerden olan Arthur Lee Samuel 1959'da makine öğrenmesi konusunu bilgisayarların yeniden programlanmaya ihtiyaç duymadan görev yapmasını sağlayan bir bilim olarak tanımlamıştır (Samuel 2000: 210).

Makine öğrenme, bilgisayarların "öğrenmesi" kolay algoritma ve teknikleri dizayn etmeye, geliştirmeye odaklanmaktadır. Belirli bir veri seti üzerinde çalışmayı hedeflememektedir. Geliştirdiği algoritmalar her sorunu çözmeyi hedeflemektedir. Veri madenciliği ise makine öğrenme yöntemlerini kullanarak "gerçek" veriler

üzerinde tanımlama, sınıflama, tahmin ya da kümeleme amaçlı çalışmaktadır. Bu doğrultuda danışmanlı ve danışmansız öğrenme algoritmaları kullanılmaktadır. Makine öğrenmesi, istatiksel analizinde herhangi bir modelleme yapılmamış genetik araştırmalarda kullanılabilmektedir. Bu sayede sadece bir veri setine özgü sonuçlar yerine belirli hastalıklara özgü modeller ortaya konabilmektedir. Bu ayrımın farkında olarak bilgi teknolojilerinin getirdiği yenilikleri istatistiksel analizlerde kullanmak gereklidir (Coşgun, Karaağaoğlu 2010: 162).

Makine öğrenmesi; denetimli ve denetimsiz öğrenme yöntemlerinden oluşmaktadır. Denetimli öğrenme; önceden gözlemlenmiş ve sonuçları bilinen verileri kullanarak, bu verileri ve sonuçlarını kapsayan bir fonksiyon oluşturmayı amaçlayan makine öğrenimi yöntemidir. Denetimsiz öğrenme ise; etiketlenmemiş verideki gizli yapıyı bulma işlemidir. Yani veriler arasında var olan; ama gözle görülmeyen bağıntının açığa çıkarılması işlemidir (Nizam, Akın 2015).

Veri üzerinde örüntü araştırma, bilgi keşfi için önemli bir aşamadır. Makine öğrenmesi; bilginin keşfedilmesi esnasında tümevarımsal algoritmaların uygulanması prosesini tamamlamak adına oldukça yaygın bir şekilde kullanılan akademik bir çalışma sahasıdır (Timur ve diğerleri 2011: 75).

Günümüzde makine öğrenmesi temeline dayanan teknikler kullanılarak, özellikle sağlık alanındaki sorunların çözümü için hekimlerin kararlarına destek sağlanabilmektedir (Karakoyun, Hacıbeyoğlu 2014: 31).

Makine öğrenimi, tecrübelerden elde edilen bilgileri makineleştirerek hesaplama yöntemlerinde performansı artırmak için kullanılan bir çalışmadır. Makine öğrenimi, bilgi mühendisliği sürecinde otomasyon düzeyini artırmayı, eğitim verilerindeki örüntülerin keşfedilmesi sürecinde etkinliği artıran otomatik tekniklerin, çok fazla zaman kaybına neden olan insan gücünün yerine geçmesini amaçlamaktadır (Jackson 2002: 272).

Bilgisayar yardımı ile bir problem çözümlemek istenirse, probleme uygun algoritmalar geliştirmek gereklidir. Günümüzdeki teknolojik gelişmeler sayesinde, veri tabanlarında milyarlarca veri kaydedilmekte ve bu verilerden çıkarsamalar yapılmaktadır. Bu verilerdeki örüntüleri ve düzenlilikleri araştırmak için birçok algoritma geliştirilmiştir. Bu algoritmalar programlanarak makine öğreniminin bir

parçasını oluşturmaktadırlar. Makine öğrenimi; bilgisayarların, algılayıcı verisi ya da veri tabanı gibi veri türlerine dayalı öğrenimini olanaklı kılan algoritmaların tasarım ve geliştirme süreçlerini konu edinen bir bilim dalıdır. Ancak makine öğrenimi sadece veri tabanı problemi değil, aynı zamanda yapay zekanın bir parçasıdır. Aynı zamanda makine öğrenimi, robot teknolojilerinde, görüntü ve ses tanıma sistemlerinde birçok probleme çözüm üretmektedirler (Alpaydın 2010: 3).

Veri analizinde istatistik bilimi de önemli rol oynamaktadır. İstatistiğin amacı istatistiksel analiz teknikleriyle veriler hakkında anlamlı bilgiler üretmek ve yorum yapılmasına olanak sağlamaktır. Bu durumda veri madenciliğinin istatistikten farkının ne olduğu sorusu akla gelmektedir. İstatistiğin doğuşu, bilgisayar icadından önceye dayanmaktadır. İstatistiksel yöntemler elle de uygulanabilmektedir. Bilgisayar teknolojisinin doğuşu ve gelişimi özellikle büyük verilerin kullanımında istatistiksel analizlerde kolaylık sağlamış olsa da verinin içindeki gizli örüntülerin bulunması, çözümlenmesi ve yorumlanmasında yeterli olmamıştır. Bu noktada bazı modellere, algoritmalara gereksinim duyulmuştur. Bu da veri madenciliği kavramının ortaya atılmasına neden olmuştur (Akçay 2014: 37).

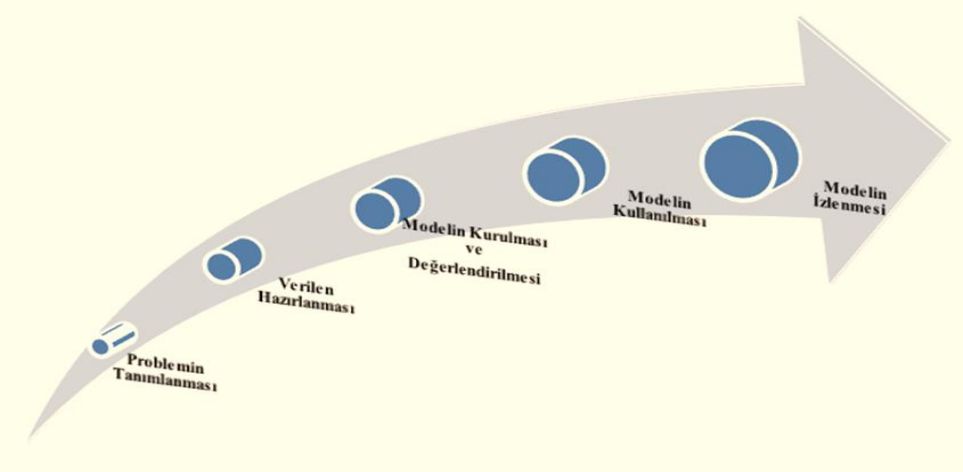
Veri madenciliği çalışması esas olarak bir istatistik uygulamasıdır. Verilen bir örnek kümesine bir kestirici oturtmayı amaçlar. İstatistik literatüründe son elli yılda bu amaç için değişik teknikler önerilmiştir. Bu teknikler istatistik literatüründe çok değişkenli analiz başlığı altında toplanır ve genelde verinin parametrik olduğu varsayımına göre bu varsayım altında sınıflama, faktör analizi, regresyon, kümeleme, varyans analizi, bağıntı kurma teknikleri istatistikte uzun yıllardır kullanılmaktadır (Çoban 2006: 67).

Veri madenciliği; İnce ve Alan'a göre "veri analiz etmeye dayalı bilgi keşfetmek ve reklam, bioenformatik, veri tabanı pazarlaması, sahteciliğin tespiti, e-ticaret, sağlık, güvenlik, web, finansal tahmin vs. dahil olmak üzere çeşitli uygulamalara tatbik edilebilir faydalı bilgiler üretmek" olarak bilinir (2014:67).

Veri madenciliğinin cazibesi anlaşılmaya başlandıkça, bu alanla ilgili bilgisayar programları da hızlı bir şekilde artmaktadır. Buna rağmen yine de bu alanda yapılan araştırmalarda istatistik temel bir görev üstlenmektedir (Oğuzlar 2003: 67).

## 1.10. Veri Madenciliği Süreçleri

Günümüz bilgisayar teknojilerinin gelişmesiyle veri madenciliği uygulamasının işletmelerde kullanımı giderek yaygınlaşmıştır. Ancak elde edilen verilerin doğruluğu ve bu verilerden elde edilen bilgilerin güvenirlilği tartışılır boyuttadır. Dolayısıyla üzerinde araştırma yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda ne kadar etkin olursa olsun hiç bir veri madenciliği algoritmasının fayda sağlaması mümkün değildir. Bu yüzden, ham veriler belirli işlemlerle hatalardan kurtarılması için ön işleme tabitutularak analize uygun hale getirilmesi gerekmektedir (Çam 2014: 21). Veri madenciliği sürecinde izlenen adımlar genellikle problemin tanımlanması, verilerin hazırlanması, modelin kurulması ve değerlendirilmesi, modelin kullanılması ve modelin izlenmesi şeklindedir. (Savaş ve diğerleri 2012: Çalış ve diğerleri 2014: 3).




**Şekil 1.7.** Veri Madenciliği Süreçleri

**Kaynak: (**Savaş ve diğerleri 2012)

### 1.10.1. Problemin Tanımlanması

Bu aşama veri madenciliği sürecinin en temel ve en önemli aşamasını teşkil etmektedir. Bu aşamada; araştırmanın amacı, mevcut durumun değerlendirilmesi, veri madenciliğinin amaçları ve proje planlama sürecinin belirlenmesi gibi alt başlıklar yer almaktadır (Albayrak, Yılmaz 2009: 36).

Veri madenciliği çalışmalarında başarılı olmanın en önemli şartı, uygulamanın hangi amaç için yapılacağının ve elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceğinin tanımlanmasıdır. Bu nedenle, veri madenciliği çalışmalarında öncelikli olarak amaç açık bir şekilde ortaya konulmalı ve durum değerlendirmesi yapılmalıdır (Çalış 2013: 14).

### 1.10.2. Verilerin Hazırlanması

Veri madenciliği çalışmalarında karşılaşılan en önemli sorunlar verilerdeki eksiklikler, araştırmaya uygun verilerin seçilmemesi, seçilen veriler arasında yüksek korelasyon bulunması, sapan değerler gibi sorunlardır. Bunun için veri madenciliği adımlarından olan veri önişleme başka bir deyişle veri hazırlama bütün veri madenciliği sürecinin en önemli adımıdır (Gemici 2012: 20).

Veri madenciliğinin en önemli ikinci aşamasını teşkil etmektedir. Bu süreçte işletmelerin var olan bilgi sistemleri sayesinde ürettikleri sayısal bilgilerin iyi analiz edilmesi, veriler ile var olan iş sorunları arasında ilişki olması gerektiği göz önünde bulundurulmalıdır. İlgili proje bünyesinde faydalanılacak olan sayısal verilerin ne tür iş süreçleri ile meydana getirildiği de bu sayısal veriler kullanılmadan önce analize tabi tutulmalıdır. Ancak bu yolla verilerin kalitesiyle ilgili bir yorum yapılabilir. Verilerin hazırlanması aşaması ise kendi arasında; toplama, birleştirme, temizleme ve dönüştürme gibi alt basamaklarından meydana gelmektedir (Tümen 2013: 11).

#### 1.10.2.1. Veri Temizleme

Tutarsız ve hatalı veriler, veri tabanı üzerinde yapılacak analizlerde yanlış sonuç verebileceğinden bu hatalı veriler veri ambarına aktarılmadan önce silinir. Örneğin; dersanede eğitim gören bir öğrenci eğer eğitimi boyunca hiçbir deneme sınavına girmediyse yapılacak analizde bu öğrenciye yer verilmemelidir (Hatipoğlu 2013: 18).

#### 1.10.2.2. Veri Birleştirme

Bazı durumlarda birçok veri kaynağından yararlanarak veri kümemizi oluşturmamız gerekmektedir. Veri birleştirme denilen bu işlemde farklı kaynaklardan gelen veriler aynı veri kümesi altında birleştirilmektedir. Farklı kaynaklardan aynı nitelik için farklı

değerler, ölçü birimleri ya da derecelendirmeler kullanılmış olabilmektedir. Bu durumlarda nitelik değerlerini birleştirirken dönüşüm yapmak gerekmektedir. Farklı kaynaklarda aynı nitelikler farklı nitelikmiş gibi ele alınmış olabilmekte ya da birleştirme sonucunda gereksiz veriler oluşabilmektedir. Bu tip niteliklerin belirlenmesi, gereksiz verilerin ayıklanması gerekmektedir (Coşkun 2010: 14).

#### 1.10.2.3. Veri Dönüştürme

Daha sağlıklı sonuçların elde edilebilmesi veya verinin kullanılan algoritmalarla uyumlu olabilmesi için, verinin tanımlanan bir fonksiyona uygun olarak farklı değer veya ölçeklere dönüştürülmesi işlemine denir. Örneğin; bir uygulamada bir yapay sinir ağı algoritmasının kullanılması durumunda kategorik değişken değerlerinin evet/hayır olması, bir karar ağacı algoritmasının kullanılması durumunda ise, örneğin; gelir değişken değerlerinin yüksek/ortak/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır (Yılmaz 2006: 87; Akpınar 2014: 132).

#### 1.10.2.4. Verilerin Toplanması

Tanımlanan problem için gerekli olduğu düşünülen verilerin toplanacağı veri kaynaklarının belirlenmesi aşamasıdır. Verilerin toplanmasında kuruluşun kendi kaynaklarının dışında, nüfus sayımı, meteroloji, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilmektedir (Yılmaz 2006: 86).

#### 1.10.2.5. Verilerin indirgenmesi

Veri madenciliği uygulamalarında bazen çözümleme işlemi uzun zaman alabilir. Sonuçların değişmeyeceğine inanılıyorsa veri sayısı veya değişkenlerin sayısı azaltılabilir. Veri idirgeme; veriyi birleştirme, boyut indirgeme, veri sıkıştırma, örnekleme ve genelleme şeklinde yapılır. (Özkan 2013: 41)

### 1.10.3. Modelin Kurulması ve Değerlendirilmesi

Basit geçerlilik testi, kurulmuş olan bir modelin doğruluğunun test edilmesinde yaygın olarak kullanılan en kolay metottur. Bu metotta verilerin %5 ile %33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi tamamlandıktan sonra test işlemine geçilmektedir. Bir sınıflama modelinde; Doğruluk

Oranı + Hata Oranı = 1 olmak zorundadır. Az sayıda verinin olması halinde başvurulabilecek bir diğer metot ise çapraz geçerlilik testi olup, veri kümesi rastgele şekilde k eşit parçaya bölüp bunlardan birisi dışarda tutulup, diğer parçaların dışarda tutulan parçanın modelin performans üzerindeki etkisine bakılır. (Altıntaş 2006: 11).

Modelin yayılma aşamasından önce, işletmenin amaçlarını tam olarak gerçekleştirdiğinden emin olmak gerekir. Modelin eksiksiz bir şekilde değerlendirilmesi ve modeli gerçekleştirmek için oluşturulan adımların gözden geçirilmesi önemli bir aşamadır. Buradaki temel amaç, yeteri derecede dikkate alınmayan bir işletme sorununun olup olmadığını belirlemektir. Bu evrenin sonunda veri madenciliği sonuçlarının kullanımıyla ilgili bir karara ulaşılabilmektedir (Küçüksille 2009: 34).

Modelleme aşamasında, çeşitli modelleme teknikleri seçilmekte, uygulanmakta ve optimum değerlere ulaşabilmek için parametreleri ayarlanmaktadır. Bu aşamada, seçilen teknikler veri setleri üzerinde çalıştırılmakta ve çıkan matematiksel denklemler yorumlanmaktadır. Süreç tekrarlandıkça performans iyileşmekte, sonuçlar daha güvenilir hale gelmektedir. Bu aşamada seçilen veri madenciliği aracının özellikli olarak hangi algoritmaları, teknikleri kullanacağına ve hangilerinin modele en uygun olduğuna karar verilmeye çalışılmaktadır. Genellikle, aynı veri madenciliği problem tipi için birden fazla teknik bulunduğundan ve bu tekniklerden bazıları verinin özel bir formunu gerektirdiğinden bu aşamadan veri hazırlama aşamasına geri dönüşler çoğunlukla gerekli olmaktadır (Yakut 2012: 12).

### 1.10.4. Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model, doğrudan doğruya bir uygulama olabileceği gibi başka bir uygulamanın alt parçası olarak da işlem görebilmektedir. Kurulan modeller; risk analizi, kredi değerlendirme ve dolandırılık saptama gibi işletme uygulamalarında doğrudan kullanılabilmektedir. Bunun yanı sıra kurulan bu modeller promosyon planlanması gibi farklı simülasyonlara entegre edilebilmekte veya tahmin edilen envanter düzeyleri, sipariş miktarının altına düştüğünde otomatik olarak sipariş verilmesine olanak tanıyacak şekilde bir uygulamanın içine gömülebilmektedir (Ataseven 2008: 12; Elmas 2014: 13).

### 1.10.5. Modelin İzlenmesi

Zaman içerisinde sistemlerin özelliklerinden ve dolayısıyla ürettikleri verilerde meydana gelen değişikliklerden dolayı kurulan modellerin sürekli olarak izlenip takip edilmesi ve eğer gerekli ise yeniden düzenlenmesi yapılmalıdır. Tüm bu süreçlerden sonra farklı görselleştirme ve raporlaştırma elemetleri vasıtasıyla saptanmış olan sonuçlar ilgilelere sunulmaktadır (Farboudi 2009: 25). Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler, model sonuçlarının izlenmesinde kullanılan faydalı bir yöntemdir (Terzi ve diğerleri 2011: 34).

Zamanla bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modelin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini zorunlu kılacaktır. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler, model sonuçlarının izlenmesinde kullanılan faydalı bir yöntemdir (Çil 2010: 12).

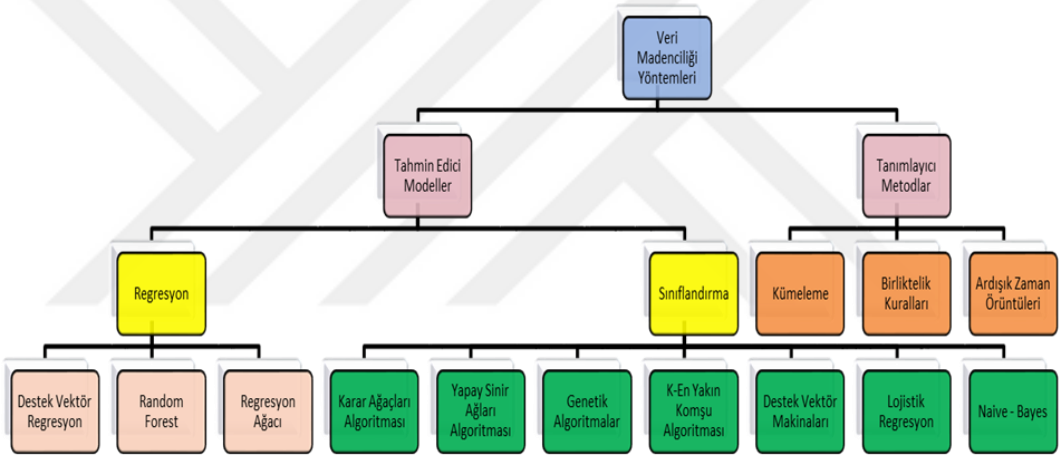
Uygun yazılım seçilerek toplanan ve hazırlanan veriler bu yazılım ile analiz edilmektedir. Bu aşamada kullanılacak olan yazılımlar marka ve açık kaynak kodlu yazılımlar olarak iki ayrı grupta toplanabilmektedir. Marka yazılımların arkasında bir teknik destek olması bu tür yazılımların en büyük avantajıdır. Özgür yazılım akımının gelişmesiyle birlikte açık kaynak kodlu yazılımlarında sayısı her geçen gün artmakla kalmamış, sadece güvenli sonuçlar üretebilen yazılımlar ortaya çıkmıştır. Bunlardan bazıları KNIME (Konstanz Information Miner), Tanogra, R, WEKA (Waikato Environment for Knowledge Analysis) isimli açık kaynak kodlu yazılımlar olup, akademik alanlarda olduğu kadar, değişik sektörlerde faaliyet gösteren her büyüklükteki firmanın veri madenciliği çözümüne katkıda bulunabilecek niteliktedirler (Çınar, Silahtaroğlu 2012: 312).

## 1.11. Veri Madenciliği Yöntemleri

Veri madenciliğinde kullanılan yöntemler; tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında incelenmektedir. Tahmin edici yöntemlerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Tanımlayıcı

modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır (Akpınar 2000: 5).

Veri madenciliği yöntemleri; tanımlayıcı ve tahmin edici modeller olmak üzere genel olarak iki ana başlıkta ele alınabilmektedir. Fakat bazı yöntemler hem tanımlayıcı hem de tahmin edici niteliklere sahip olabilmektedirler. Tahmin edici modellerde; sonuçları bilinen verilerden yola çıkarak bir model oluşturulması ve oluşturulan bu modelden sonuçları net olarak bilinmeyen verilerin tahmin edilmesi amaçlanmaktadır. Eldeki verilerin strateji geliştirme ve karar verme süreçlerinde kullanılması sonucunda yeni bilgiler elde edilmesi ise; tanımlayıcı modeller vasıtasıyla yapılmaktadır. Tanımlayıcı modeller genellikle veriler arasındaki gizli kalmış ilişkilerin gün yüzüne çıkartılmasında kullanılmaktadır (Akman ve diğeleri 2011: 37).




**Şekil 1.8.** Veri madenciliğinde Kullanılan Modeller

### 1.11.1. Tanımlayıcı Modeller

Tanımlayıcı modellerin amacı veri setinde yer alan veriler arasındaki ilişkileri, bağlantıları ve davranışları bulmaktır. Var olan verileri yorumlayarak davranış biçimleri ile ilgili tespitler yapmayı ve bu davranış biçimini gösteren alt veri setlerinin özelliklerini tanımlamayı hedeflemektedir. Tanımı bilmek; tekrarlanan bir faaliyete veya tanımı bilinen yeni bir verinin yapıya katılmasında ne şekilde hareket edileceği konusunda karar almaya destek olmaktadır (Erşahin, Argüden 2008: 39).

A/B aralığında geliri ve en az iki arabası olan çocuk sahibi olan aileler ile çocuk sahibi olmayan ve geliri A/B aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin saptanması, tanımlayıcı modellere örnek olarak verilebilmektedir (Sevindik ve diğerleri 2012: 188).

Farklı kaynaklarda farklı alt sınıflamalar olsa da tanımlayıcı modelleri kendi arasında; kümeleme, birliktelik kuralları ve ardışık zaman örüntüleri şeklinde toplam üç grupta ele almak mümkündür.

#### 1.11.1.1. Kümeleme

Kümeleme analizi (clustering) veri madenciliğinin en önemli alanlarından birisidir; amacı, nesneleri birbirlerine olan benzerliklerine göre gruplara ayırmaktır. Elde bulunan veriler incelenerek birbirlerine benzeyenler bir kümeye, benzemeyenler ise bir başka kümeye toplanmaktadırlar (Yeleğin 2012: 6).

Kümeleme; verilerin benzer özelliklerinin belirlenerek gruplandırılmasıdır. Gruplandırma işlemi, kavramsal kümeleme ilkelerinden sınıflar arası maksimum ve sınıf içi minimum benzerlik ilkelerine dayalı olarak gerçekleştirilmektedir (Koçoğlu 2012: 13).

Verilerin farklı gruplaşmalarını inceler ve eğer varsa bunları ortaya çıkarmaktadır. Birbirine benzeyen nesnelerin aynı grup altında toplanması işlemi genellikle kümeleme olarak tanımlanmaktadır. Kümeleme analizinde en önemli faktör; hangi kriterlere göre kümeleme yapılacağıdır. Bu kriterler konunun uzmanları tarafından tahmin yöntemiyle kestirilmeye çalışılmaktadır. Kümeleme analizi yapabilmek için verilerin normal dağılım şartına uyup uymadıkları göz önünde bulundurulmamaktadır. Bu bakımdan kümeleme analizi, diğer istatistiki tekniklerden farklılık arz etmektedir (Yurtay ve diğerleri 2013: 898-899).

Kümeleme analizinin ilk amacı; gözlem birimlerini benzer özelliklerine göre gruplandırmak olan çok değişkenli istatistiksel yöntemlerden birisidir. Bu analizde; elde edilen bir küme içerisindeki gözlem birimleri, önceden saptanmış bir nitelik açısından birbirine benzemektedir. Dolayısıyla elde edilen kümelerdeki gözlem birimleri homojen bir özellik göstermektedir. Bu analizin temel hedefi; dağınık bir

halde bulunan verileri benzerliklerini baz alarak bir araya getirmek ve sınıflandırarak işlenebilir bir forma kavuşturmaktır (Timor ve diğerleri 2011: 131).

Kümeleme sürecinde, küme kapsamına dahil olan elemanların benzerliği fazla, kümeler arasındaki benzerliğin ise az olması arzu edilmektedir. Bu şartı sağlayan bir kümeleme yönteminin kaliteli olduğu varsayılmaktadır. Bu yöntemin seçimi, kullanılacak verinin türüne ve uygulamanın amacına göre farklılık arz etmektedir (Bilgin, Çamurcu 2005: 139).

Uçan'a göre "kümeleme analizleri; bölünme merkezli kümeleme, hiyerarşik kümeleme, yoğunluk-merkezli kümeleme ve grid (ızgara) merkezli kümeleme olmak üzere dört gruba ayrılmakta ve incelenmektedir" (2010: 35). Yeşilbudak ve arkadaşları tarafından ise "hiyerarşik yöntemler, parçalı yöntemler, yoğunluk tabanlı yöntemler, şebeke tabanlı yöntemler ve sezgisel yöntemler olmak üzere toplam beş grupta ele alınmıştır" (2011: 29).

#### 1.11.1.2. Birliktelik Kuralları

Tanımlayıcı modellerin alt başlıklarından birisi olan birliktelik kuralları analizi; müşterilerin beraber satın aldıkları ürünlerin analizini yapmaktadır. Bu analizde temel amaç; satın alma davranışında bulunanlar arasındaki ilişkilerin tanımlanmasıdır. Bu türden ilşkilerin belirlenmesi, işletmelerin kar marjlarını arttırmada önemli bir rol üstlenmektedir. Örneğin bir A malını satın alanların B malını da çok yüksek ihtimalle satın aldıkları biliniyorsa ve eğer bir müşteri A malını alıyor fakat B malını almıyorsa, o müşteri potansiyel bir B müşterisidir. Şayet elde edilen veride, ürünlerle ilgili olarak sadece satın alındı veya alınmadı şeklinde bir bilgi varsa bu analizde ürünler arasındaki bağıntı, destek ve güven faktörleri ile hesaplanmaktadır (Şimşek 2006: 52).

Birliktelik kurallarının en önemli ve yaygın kullanım alanları; pazar sepet analizleri, çapraz pazarlama, promosyon analizleri, katalog ve yerleşim düzeni tasarımlarıdır (Erpolat 2012: 138).

Müşterilerin satın alma alışkanlıklarının belirlenmesinde, ürünlerin seçilmesinde ve ürünlerin teşhir edileceği raf lokasyonlarının saptanması gibi kritik raf alanı yönetimi konularında işletmelere, özellikle de binlerce çeşit ürünün yer aldığı

süpermarket ve benzeri işletmelere yol gösterici bir pozisyonda yer almaktadır (Özkan 2011: 32).

Birliktelik kuralları algoritması ile ardışık zaman örüntüleri algoritmaları arasındaki farklara değinilecek olursa; birliktelik kuralları eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılmaktadır. Örneğin; müşteriler bira satın aldıklarında % 75 olasılıkla patates cipsi de satın almaktadırlar. Ardışık zaman örüntüleri, birbirleriyle ilişkisi olan fakat birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılmaktadır (Çankırı ve diğerleri 2009: 159).

Birliktelik kuralları yaygın olarak kullanılan veri madenciliği yöntemlerinden birisidir. Veriler arasındaki birlikteliklerin, ilişkilerin ve bağıntıların kurallar halinde bulunması işlemidir. Veri nesneleri arasındaki ilginç ilişkiler ve eş zamanlı gerçekleşen durumlar araştırılmaktadır. Bu birliktelik kuralına örnek olarak bez ve mama ürününü satın alan müşterilerin %80 olasılıkla ıslak mendil ürününü de satın alması verilebilmektedir. Bu tür birliktelik kuralları, söz konusu nesneler ile ilgili durumun sıklıkla tekrarlanması durumunda anlamlıdır. Birliktelik kuralları analizi; ticaret, finans, mühendislik, fen ve sağlık sektörlerinin bir çok alanlarında kullanılmaktadır. Örneğin; pazar sepet analizlerinde sıklıkla birlikte satılan ürünleri tespit etmek, web sayfalarında ziyaretçilerin hangi sayfaları birlikte tıkladığını araştırmak, bağıntılı olarak geçirilen hastalıkları belirlemek için kullanılabilmektedir (Birant ve diğerleri 2010: 215).

#### 1.11.1.3. Ardışık Zaman Örüntüleri

Örüntü (pattern) sözcüğü; herhangi bir çizim, ses, resim veya parmak izi gibi bir şekil olarak tasvir edilebilmektedir. Ayrıca bir kimsenin yaptığı işler de örüntü olarak ele alınabilmektedir. Örnek vermek gerekirse; bir müşterinin marketten ekmek, peynir ve süt alması bir örüntüdür. Dolayısıyla bir müşterinin satın aldığı ilk ürünün A, onu izleyen gün veya günlerden birinde B ürününü ve daha sonraki bir günde ise C ürününü alması ise yine bir örüntüyü teşkil etmektedir. Fakat bu sefer zaman içerisinde birbirini izleyen bir örüntü meydana getirecektir (Yeleğin 2012: 5).

Birliktelik kurallarının çalışmasından sonra gelen doğal bir adımdır. Bir önceki adımdaki gibi bu teknik işlemsel veri tabanlarına uygulanmaktadır. Burada işlemin bir

sahibi ve bir zaman göstergesi olmalıdır. Örneğin; bir banka veri tabanında her bir işlem belirli bir müşteriye uygulanır ve belirli bir zamanda gerçekleşmiştir (Tuğ 2005: 28).

Ardışık zamanlı örüntüler; aşağıda sunulan örneklerde görüldüğü üzere; birbirleriyle ilişkisi olan; ancak birbirini izleyen dönemlerde gerçekleştirilen ilişkilerin tanımlanmasında kullanılmaktadır: X ameliyatı yapıldığında, 15 gün içerisinde Y enfeksiyonu oluşacaktır. İMKB endeksi düşerken A hisse senedinin değeri % 15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri % 60 olasılıkla artacaktır. Çekiç satın alan bir müşteri ilk üç ay içerisinde % 15, bu dönemi izleyen üç ay içerisinde % 10 ihtimalle çivi satın alacaktır (Akpınar 2000: 7).

### 1.11.2. Tahmin Edici Modeller

Tahmin edici modellemede temel amaç; veri tabanındaki bazı alanların diğer alanlara bağlı olarak tahmin edilmesidir. Tahmin edilecek alan eğer sayısal (sürekli) bir değişken ise tahmin problemi bir regresyon problemidir. Eğer tahmin edilecek alan kategorik bir değişken ise sınıflama problemidir. Sınıflama ve regresyon için kullanılan çok fazla sayıda değişken bulunmaktadır. Tahmin edici modellerde problem; diğer alanlardaki (girdiler), her gözlem için hedef değişken değerinin verilmiş olduğu veri seti ve problem hakkında önceden sahip olunan bilgileri yansıtan varsayımların kümesinin verilmesi durumunda tahmin edilecek değişkenin alabileceği muhtemel değerin belirlenmesi şeklinde özetlenebilmektedir. Girdilerin doğrusal olmayan dönüşümüyle birleştirilen doğrusal regresyon çok geniş alandaki problemlerin çözümünde kullanılabilmektedir. Girdi uzayının dönüşümü; problem hakkında bilgi ihtiyacını gerektiren genellikle zor bir sorundur (Kiremitçi 2005: 47).

Kestirim modelleri bir sınıflama modeli gibidir ancak bu modeli, tahmin ve sınıflama modellerinden ayıran özellik, gelecekteki verilerin tahmin etmesidir. Veri madenciliği uygulamalarında bu modele örnek olarak; sel baskınlarının tahmin edilmesi, konuşma sesinden sözcüklerin tahmin edilmesi ve örüntü tanımlama problemleri verilebilmektedir. Gelecek değerlerin zaman serisi analizi veya regresyon modelleri kullanılarak tahminlenmesine karşın, farklı metodolojiler de kullanılabilmektedir. Örneğin su baskınlarının tahminlenmesinde, ırmağın farklı bölgelerine konumlandırılan alıcılar, ırmağın su düzeyini, yağmur miktarı, nem ve

zaman gibi çeşitli verileri toplayarak su baskınlarına dair tahminsel modeller meydana getirebilmektedir (Aydın 2007: 12).

Bir çok kaynakta tahmin edici modelleri kendi arasında; sınıflama ve regresyon olarak ikiye ayrılır. Sınıflamada bağımlı değişken kategorik olurken, regresyon yönteminde bağımlı değişken sürekli'dir (Tüzüntürk 2010: 76).

#### 1.11.2.1. Sınıflama

Veri madenciliği tekniklerinden biri olan sınıflama; sınıfı tanımlanmış mevcut verilerden faydalanarak sınıfı belli olmayan verilerin, sınıfını tahmin etmek için kullanılanılan bir yöntemdir. Sınıflama yöntemi iki aşamadan oluşur. Birinci aşamada tahmin için kullanılan bir model oluşturulur, ikinci aşamada ise oluşturulan bu modele sınıfı belli olmayan yeni veriler uygulanarak, sınıflar tahmin edilmeye çalışılmaktadır. Sınıflama tahminleyici bir modeldir ve/veya makina öğrenimi yöntemleri kullanılarak sınıflara atanması işlemidir. Sınıflama işlemine örnek verilecek olursa havanın bir sonraki gün nasıl olacağı yada bir kutuda ne kadar mavi top olacağı tahmin edilebilir (Silahtaroğlu 2008: 45).

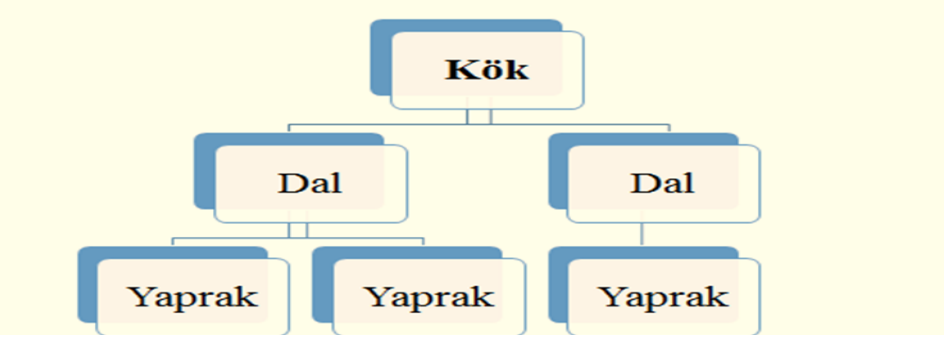
Sınıflama ile regresyon arasında önemli farklılıklar bulunmaktadır. Sınıflamada hedef değişken kategorik veri tipinde iken, regresyonda sürekli sayısal veri tipindedir ve hedef değişkenin sayısal değeri kestirilmeye çalışılmaktadır. Başlıca sınıflama teknikleri, Yapay Sinir Ağları (Artificial Neural Networks), Genetik Algoritmalar (Genetic Algorithms), K- En Yakın Komsu (K–Nearest Neighbour), Bellek Temelli Nedenleme (Memory Based Reasoning), Naive – Bayes, Lojistik Regresyon (Logistic Regression) ve Karar Ağaçlarıdır (Decision Trees), Destek vektör Makinaları, Random forest (Albayrak, Yılmaz: 2009).

Veri madenciliğinin en bilinen işi sınıflamadır. Girdilerin çeşitli niteliklere göre bir sınıflayıcı (model) tarafından sınıflara atanması sürecidir. Eldeki nesnelerin bir sınıfta atanıp atanmayacağının ya da sınıflardan hangisine atanacağının belirlenmesidir. Başka bir ifade ile nesneler veya durumlar için uygun sınıf tahmin edilmesidir. Sınıflama girdileri, her biri bir sınıf etiketi ile etiketlenecek gözlem veya örneklerden oluşan bir eğitim kümesidir. Çıktı ise modelin her bir gözleme niteliklerine dayalı olarak atadığı sınıf etiketidir (Emel, Taşkın 2005: 224).

Veriler genellikle özniteliklerindeki değerlere göre sınıflara ayrılmaktadır. Genel olarak veri sınıflaması, bir veri tabanında yer alan nesne kümeleri arasından genel özelliklerini bulur ve sınıflama modeline göre farklı sınıflara ayırmaktadır (Özgülbaş, Koyuncugil 2010: 497). Daha önce yapılmış olan çalışmalar ele alındığında genel olarak sınıflamanın; karar ağaçları algoritması, yapay sinir ağları algoritması, genetik algoritmalar, K-en yakın komşu algoritması, destek vektör makineleri algoritması, bulanık mantık, naive-bayes algoritması ve lojistik regresyon olmak üzere toplam sekiz farklı klasmanda ele alındığı görülmektedir.

#### 1.11.2.1.1. Karar Ağaçları Algoritması

Veri madenciliği yöntemlerinden sınıflama modelinin alt grubunda yer alan ve veri madenciliği uygulamasında sıklıkla kullanılan yöntemlerden birisi de karar ağaçları algoritmasıdır. Bu algoritma türünde her bir düğüm ayrı bir özelliği yansıtmaktadır. En üste yer alan öğe; kök, en altta bulunan öğe; yaprak ve bu ikisi arasında konumlanan öğe ise; dal olarak literatürdeki yerini almaktadır. Bir olayın sonuçlandırılmasında sorunun cevabına göre bir eylem planı çizen karar ağacı algoritmalarının yapısal şekli aşağıda yer almaktadır Karar ağaçlarının uygulandığı ID3, C4.5, C5.0, J48, CART algoritmaları en bilinen uygulamalardır (Daş, Türkoğlu 2014: 382).




**Şekil 1.9.** Karar Ağacı Algoritması Yapısı

**Kaynak: (**Daş, Türkoğlu 2014: 382)

Karar ağaçları, kullanıldıkları farklı algoritmalara göre; otomatik Ki-Kare etkileşim belirleme, sınıflama ve regresyon agacı ve hızlı-sapmasız-etkili istatistik ağacı olarak toplam üç farklı grupta ele alınabilmekte ve incelenebilmektedir (Kayri 2008: 106).

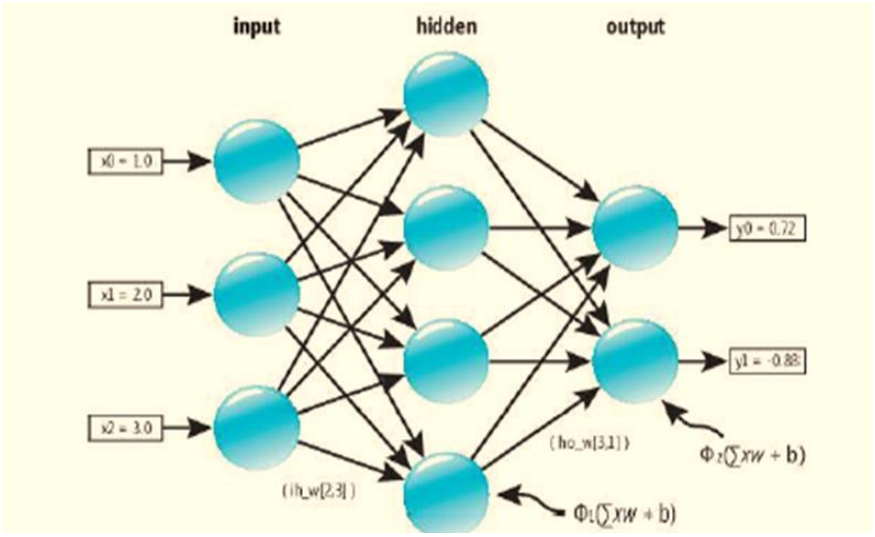
Karar ağaçları veya sınıflama ağaçları, istatistiksel öğrenme ve veri madenciliğinde çok bilinen sınıflama metodudur. Veri madenciliğinde çok kullanıldığından dolayı, veri madenciliğinin yük beygiri olarak da adlandırılmıştır (Özgür, Erdem 2012: 43).

Karar ağaçları yorumlarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir. Karar ağaçları, bir ağaç görünümünde, tahmin edici bir tekniktir (Erol 2013: 28).

#### 1.11.2.1.2. Yapay Sinir Ağları Algoritması

İnsanoğlunun beyin fonksiyonlarının çalışmasını taklit ederek, yeni sistemler oluşturmaya çalışan bir yaklaşımdır. Sınıflama yöntemleri içerisinde en yaygın olarak kullanılan metotlardan biridir. Yapay sinir ağları yapısı, insan beynindeki biyolojik sinir hücrelerinin yapısı temel alınarak oluşturulmaktadır. İnsan beyninde olduğu gibi yapay sinir ağlarında da öğrenme ve öğrenilen bilgiler ışığında karar verme mekanizmaları yer almaktadır (Dandıl 2013: 2).

Yapay sinir ağları 1980'lerden sonra yaygınlaşmıştır. Bu algoritmadaki amaç fonksiyon birbirine bağlı basit işlemci ünitelerinden oluşan bir ağ üzerine dağıtılmıştır. Yapay sinir ağlarında kullanılan öğrenme algoritmaları, veriden üniteler arasındaki bağlantı ağırlıklarını hesaplar. Uygulama alanı daha geniştir ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez (Çankırı ve diğerleri 2009: 154-155).




**Şekil 1.10.** Yapay Sinir Ağı Yapısı

**Kaynak: (**Hatipoğlu 2013: 26)

#### 1.11.2.1.3. Genetik Algotirmalar

Doğadaki gözlemlenen evrimsel sürece benzeyen bir arama ve eniyileme yöntemidir. Bu algoritma temel olarak en iyinin hayatta kalması ilkesine göre hareket eder ve en iyi entegre çözümü aramaktadır. Çeşitli türdeki sorunlara sadece bir çözüm üretmek yerine farklı çözüm sınıflarından oluşan bir küme meydana getirilmektedir. Genetik algoritmalar, sorunlara çözüm bulabilmek için evrimsel bir süreci bilgisayar ortamında taklit ederek iş yapmaktadır. Sorunun mümkün olan birçok çözümünü temsil eden bu kümeye, genetik algoritma terminolojisinde nüfus adı verilmektedir. Nüfuslar ise; vektör veya birey olarak tanımlanan sayı dizgelerinden meydana gelmektedir. Birey içerisindeki herbir eleman da gen adını almaktadır. Sonuç itibariyle nüfustaki bireyler, evrimsel bir süreç içerisinde genetik algoritma işlemcileri tarafından tayin edilmektedir (Ayık ve diğerleri 2007: 446).

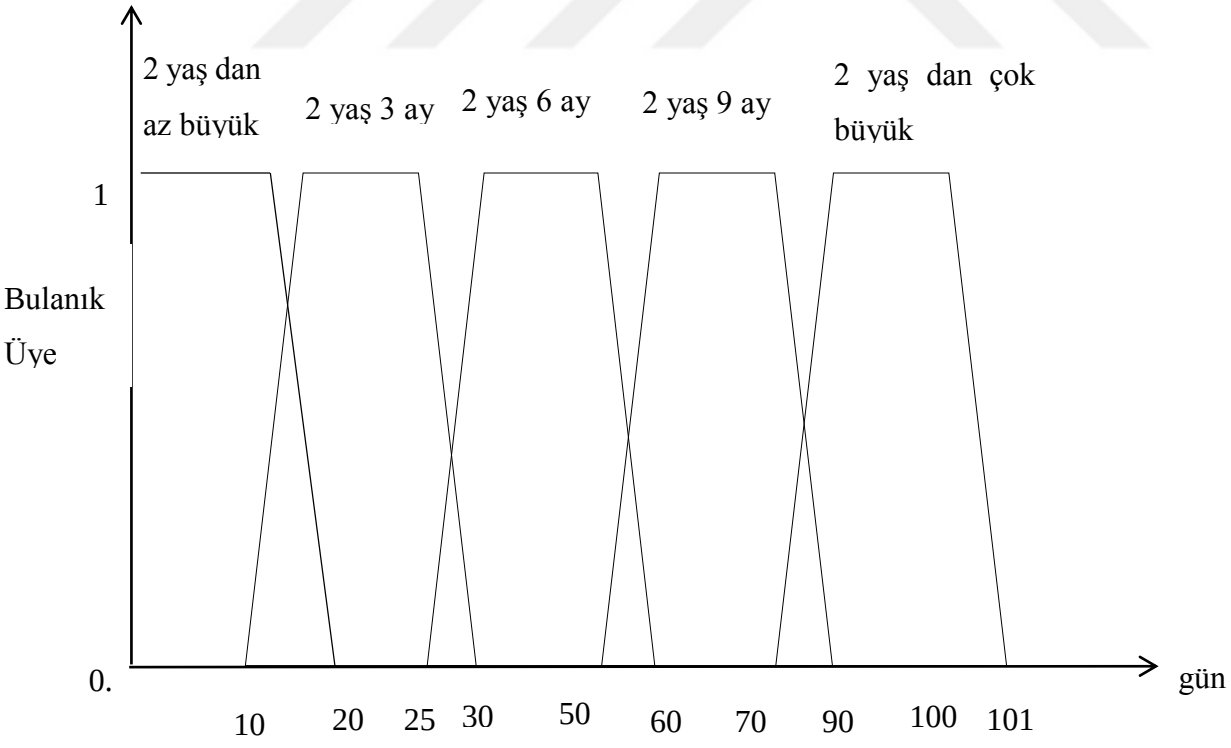
#### 1.11.2.1.4. Bulanık Mantık

Sınıflama için kullanılan kurala-dayalı sistemler, süreklilik özelliği gösteren durumlar için kesin sınırlı özellikler gerektirdiğinde dezavantaja sahiptirler. Örneğin çocuk hastalar için kullanılan bir A ilacının uygulama şekli; 2 yaşından büyük ve B tipi bir hastalığı geçirmemiş hasta olması şartı olsun. Bu durumu kurala dayalı sistem olarak belirtirsek;

$$eğer(yaş \geq 2) ve\ (hasatalık \neq B) o\ zaman\ ilaç = uygulanabilir.$$

Bu durumda ilaç 2 yaşından büyük ve B hastalığını geçirmeyen çocuk hastalara kullanılabilir olur. Ayrıca her farklı kategori için farklı dozlarda kullanılması gerekir. Peki eğer hastanın yaşı 2 den büyükse ve aylık olarak kategorilere ayrılırsa bu durumda her bir kategori için bulanık eşik değeri ve sınır belirlememiz gerekir. Yaş değişkenini; 2 yaş'dan az büyük, 2 yaş ve yaklaşık 3 aylık, 2 yaş ve yaklaşık 6 aylık, 2 yaş ve yaklaşık 9 aylık ve 2' den çok büyük olarak kategorilere ayırırsak, 2 yaş ve üstü için kesin sınırlar belirleyemeyiz. Yani gelen herhangi bir hastanın 2 yaş 15 günlük yada 2 yaş 75 günlük olduğunu düşünürsek ay olarak kesin bir değer atayamayız. Bu durumda bulanık mantık kuralları devreye girer.

Bulanık mantık kesin değerler yerine üyelik dereceleri ile gösterilen bulanık kümeleri kullanarak değişkenler için daha iyi tahminler yapmayı sağlar. Üyelik derecesi bir elemanın bulanık kümeye ait olma derecesini gösterir ve [0, 1] arasında değer alır. Olasılık (possibility) teorisi olarak da bilinen bulanık küme teorisi Lotfi Zadeh tarafından geliştirilmiştir (Han ve diğerleri 2012: 428).

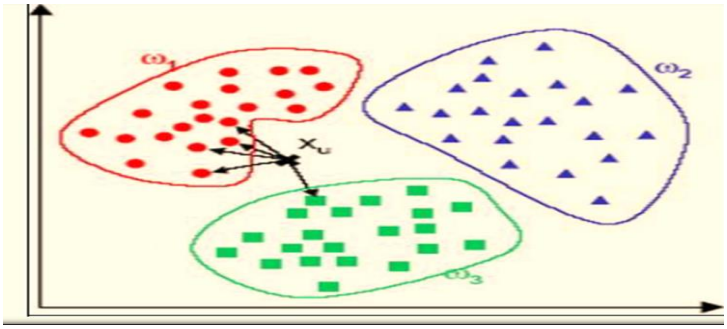



**Şekil 1.11.** Bulanık Aralık

Yukarıdaki şekil üzerinden ilacın veriliş şeklini yorumlarsak şu şekilde olur. Kliniğe 1 ay içerisinde 100 çocuk hasta tedavi için uğramış olsun. A ilacının bu hastalara verildiğini düşünelim. Örneğin bir hastanın B hastalığını geçirmediği ve yaşının ise nüfus kağıdına göre 2 yaş 55 gün olduğunu düşünürsek bu durumda hasta hem 2 yaş 6 ay ve 2 yaş 9 ay grubuna girer. Bu durumda uygulanacak ilaç dozajına karar vermek için 2 yaş 55 gün için en yüksek üyelik derecesini veren aralık çözüm olacaktır. En yüksek bulanık üyelik derecesini veren çözüm aralığına göre ilaç o dozajda verilebilir. Bulanık kümeler ile sınıflama, birçok alanda kullanılan ve araştırmacıya esnek çözüm yöntemleri sunan sınıflama tekniğidir.

#### 1.11.2.1.5. K-En Yakın Komşu Algoritması

Bellek tabanlı bir sınıflama yöntemidir. Bu sınıflama yöntemi, öğrenim kümesindeki hatayı ve saklanan alt kümenin büyüklüğü şeklinde ölçülen karmaşıklığı birlikte azaltan bir algoritma olarak tanımlanabilmektedir. Bu yöntemin tek amacı; örneğe eklenecek yeni gözlemin hangi sınıfa ait olduğunu saptamaktır. Örnekler; n boyutlu bir uzay kümesinde bir nokta olarak alınmakta ve verilen noktalara en yakın komşuların sayısı olan k parametresi belirlenmektedir. Uzaklık hesaplama amacında olan bu yöntem, verilen noktaya diğer tüm noktaların uzaklıkları tek tek hesaplanarak elde edilmektedir. Bu işlemi yapabilmek için öklid bağıntısı devreye girmektedir. Hesap edilen uzaklık düzeyleri hesaba katılarak satırlar sıralanmakta ve en küçük k tanesi seçilmektedir (Razbonyalı, Özkaya: 2-3).




**Şekil 1.12.** K-En Yakın Komşu Yöntemi

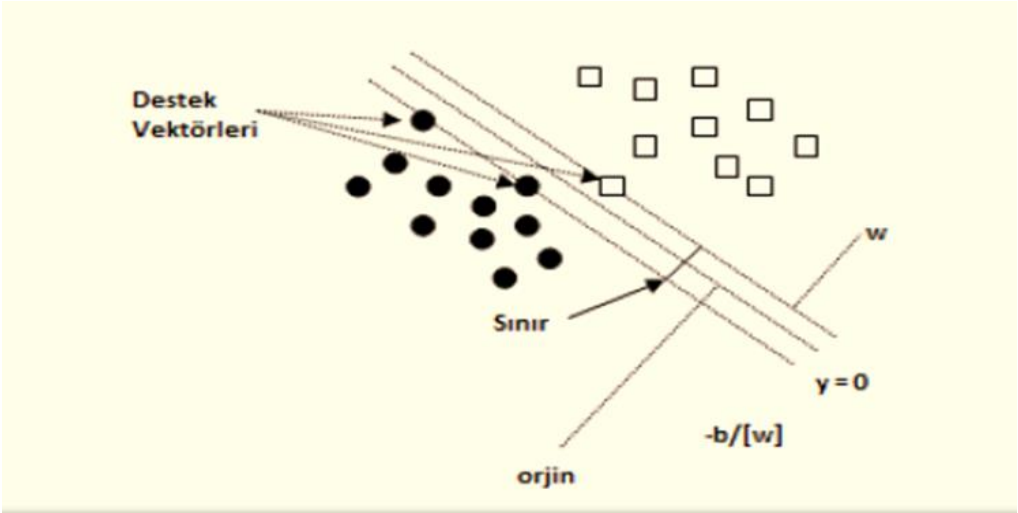
**Kaynak:** (Ceyhan, Kırlıoğlu 2014: 19)

#### 1.11.2.1.6. Destek Vektör Makineleri Algoritması

Bu sınıflama türü; istatistiki öğrenme teorisine dayanan kontrollü bir sınıflama algoritmasıdır. Bu algoritmalar ilk etapta iki sınıflı doğrusal verilerin sınıflandırılması sorunu için tasarlanmış ve daha sonraları ise çok sınıflı doğrusal olmayan verilerin sınıflandırılması amacıyla genelleştirilmiştir. Bu sınıflamanın basit bir ilkesi olup, iki sınıfı birbirinden ayırabilen en uygun karar fonksiyonunun tahmin edilmesi yöntemine dayanmaktadır (Kavzoğlu, Çölkesen 2010: 75).

Yapısal riskin minimum düzeye indirgenmesi amacıyla ortaya atılmış bir öğrenme türüdür. Destek vektör makinelerinin dayandığı teori; Vapnik Chervonenkis ve diğerleri tarafından 1960'larda başlatılmış ve 1970li yıllara doğru gelişen başarılı bir çalışmanın ürünü haline gelmiştir. İlk başarılı uygulamaları ise 1990'lı yıllara rastlamaktadır. Bu yöntem daha sonra yapay zeka uzmanları ve matematikçiler tarafından oldukça ilgi görmüş ve sıklıkla kullanılır hale gelmiştir (Hacıefendioğlu 2012: 22-23).

Destek vektör makinelerinin eğitim verileri çok az olduğu durumlarda dahi genelleme yetenekleri oldukça yüksektir. Buna ek olarak hiçbir yerel minimum içermezler. Kuadratik programlama problemi şeklinde formüle edildikleri için, problem kuadratik programlama teknikleriyle çözülebilmektedir (Tayyar, Tekin 2013: 196).




**Şekil 1.13.** Destek Vektör Makineleri Sınıflandırıcı

**Kaynak: (**Yakut 2012: 42)

#### 1.11.2.1.7. Naive-Bayes Algoritması

En etkili ve verimli makine öğrenmesi ve veri madenciliği algoritmalarından birisidir. Bu yöntem temel olarak bayes kuralına bağlı olasılıksal çıkarıma dayanmaktadır. Eldeki mevcut verilere dayalı olarak kurulan hipotezlerin doğru olma ihtimallerine göre faaliyet göstermekte ve elde edilen verilere göre maksimum olasılığa sahip olan hipotezi seçmede karar vermeye yardımcı olmaktadır (Değirmenci 2014: 17).

Sınıflama algoritmalarının kıyaslandığı araştırmalarda naive-bayes algoritmalarının, karar ağaçları ve yapay sinir ağları algoritmalarından daha yüksek doğruluk oranı ve hız parametleri bakımından daha iyi ve başarılı bir performans gösterdikleri tespit edilmiştir (Pala 2013: 71).

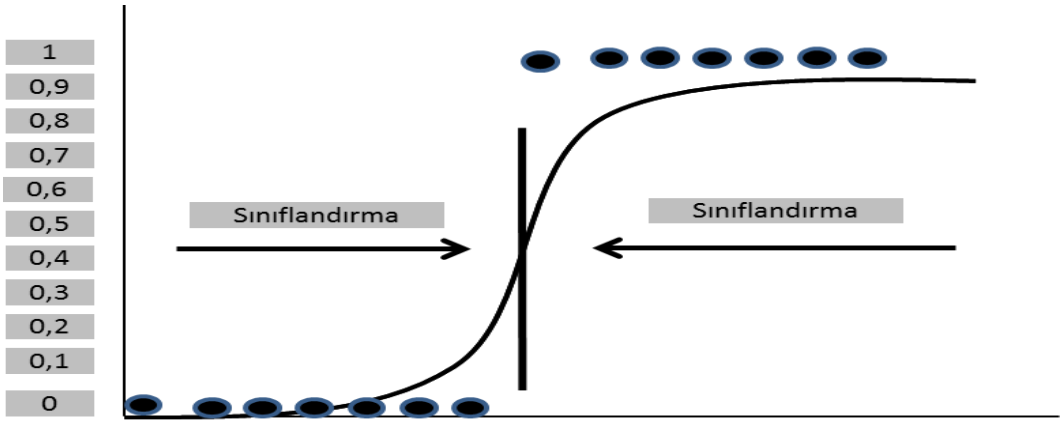
#### 1.11.2.1.8. Lojistik Regresyon

Esnekliği ve kolay yorumlanabilirliği açısından başta tıp, biyoloji, ekonomi gibi alanlarda sıkça kullanılan lojistik regresyon yöntemi; bağımlı değişkenin ikili, üçlü ve çoklu kategorilerde açıklayıcı değişkenlerle olan sebep – sonuç ilişkisini inceleyen bir veri madencliği yöntemidir. Bağımlı değişken katagorik iki değer aldığında model çeşitli dağılımlara dayalı olarak doğrusal regresyon modelinden farklı biçimde tanımlanmaktadır. Lojistik regresyonu (LR), bir veya daha fazla kategorik çıktı için ve açıklayıcı faktör kümeleri arasındaki ilişkiyi açıklamak için kullanırız. Sıradan regresyon modellerinde ortalamanın anahtar miktar olarak modellendiği gibi, kategorik çıktı değişken durumlarında log oranlı parametreler belirgin rol oynar (Fleiss ve diğerleri 2003: 284). Lojistik regresyon, veriyi lojistik eğriye uydurarak bir olayın olma ihtimalini tahmin etmek için kullanılır (Zhao 2013: 46). Lojistik regresyon, bağımlı değişkenin kesikli olduğu durumlarda, çoklu regresyonun fikrinin genişletilmesidir. Lojistik regresyonda bağımsız değişkenlerin dağılımları için herhangi bir varsayım yoktur (Liao, Triantaphyllou 2007: 116).

Bağımlı değişkenin iki ya da çok sınıflı kesikli değişken olması durumunda kullanılabilecek modeller çok çeşitlidir. Bu modellerden doğrusal olasılık modeli, lojit ve probit modeller arasında en fazla tercih edilen yöntem lojistik regresyondur.

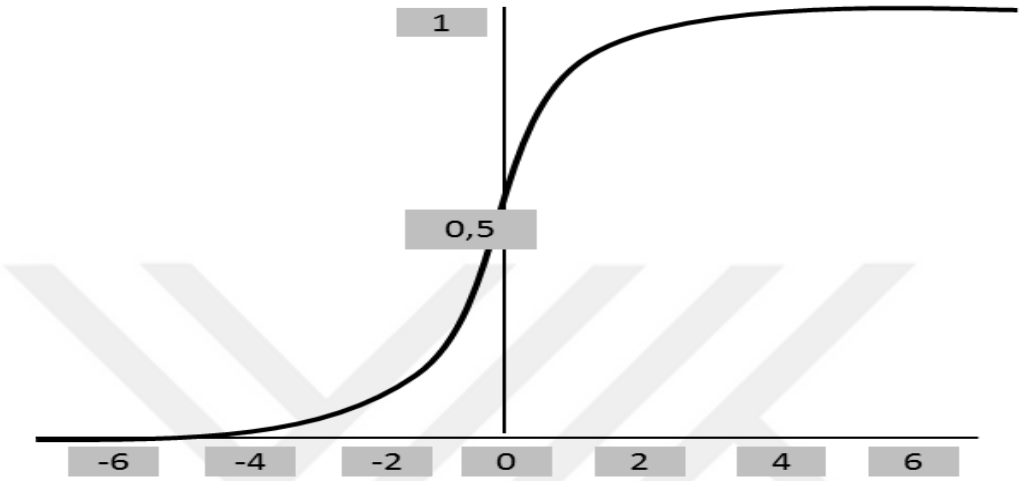
Lojistik regresyon analizini, doğrusal regresyon analizinden ayıran en belirgin özellik de lojistik regresyon analizinde bağımlı değişkenin iki ya da çok sınıflı olmasıdır. Doğrusal modelleri kullanan regresyon analizi numerik veriyi tahmin etmede kullanılan çok genel tekniktir ve lojistik regresyon ise genellikle doğrusal sınıflama uygulamak için kullanılır (Sullivan 2012: 282). Niteliksel bağımlı değişkenin kategori sayısına ve kategorilerin sırasız yada sıranabilir olmasına göre farklı lojistik regresyon yöntemleri vardır. (Hosmer, Lemeshow 1989: 5-50). Bağımlı değişken ikiden çok kategorili sırasız niteliksel değişken tipinde olduğunda çok kategorili lojistik regresyon yöntemleri kullanılırken, bağımlı değişken ikiden çok kategorili sıralanabilir niteliksel değişken tipinde olduğunda sıralı lojistik regresyon yöntemleri kullanılır. Bağımlı değişkenin iki kategorili niteliksel değişken tipinde olması durumunda iki kategorili (binary) lojistik regresyon yöntemi kullanılmaktadır (Rıffenburgh 2012: 193).

Lojistik regresyon değerlerdeki değişimler arasındaki korelasyonu hesaplar veya nominal değişkenlerin durumunu ve kardinal bağımsız boyutunu hesaplar. Temel lojistik regresyon bir tane ikili değişken durumunu varsayar. Diğer lojistik regresyon modelleri ise çok durumlu ve çok değişkenli durumları çalışır (Thomsen 2002: 559). Bağımlı değişken kodlanırken riskin olmadığı durum için 0 ve riskin olduğu durum için 1 kodu kullanılır. Bağımsız değişkenlerin tipi ile ilgili herhangi bir kısıtlama yoktur. Bağımsız değişkenler sürekli sayısal, kesikli sayısal, sırasız yada sıralanabilir niteliksel değişken tiplerinde olabilir.




**Şekil 1.14.** Lojıstık regresyon eğrisi

Lojistik regresyon, veriyi lojistik eğriye uydurarak bir olayın olma ihtimalini tahmin etmek için kullanılır (Zhao 2013: 46). Bu eğri 1844 veya 1845 yıllarında Pierre François Verhults tarafından populasyon büyümelerinde S şekli modeli olarak kullanılmıştır.




**Şekil 1.15.** Lojistik Fonksiyon

Şekil 1.15 'ndeki lojistik fonksiyonunu tanımlayan lojistik regresyon modeli eşitlik 1.1 ile verilir. Buna göre X=x olduğunda Y=1 olma olasılığı π'dir denir.

$$\mathbf{\Pi(X) = P(Y=1/X=x)} = \frac{\mathbf{e^{(\beta_0+\beta_1 x)}}}{\mathbf{1+e^{(\beta_0+\beta_1 x)}}} = \frac{\mathbf{1}}{\mathbf{1+e^{-(\beta_0+\beta_1 x)}}} \qquad 1.1$$

İki kategorili sonuç değişkenlerinin analizi için lojistik regresyonun tercih edilmesinin sebebi matematiksel açıdan çok esnek olması, kullanım kolaylığı ve sonuçların klinik açıdan anlamlı bir şekilde yorumlanabilmesidir. Eşitlik (1.1) ile verilen ve doğrusal olmayan lojistik regresyon fonksiyonu logit dönüşüm uygulandığında doğrusallaştırılabilmektedir. Lojit dönüşüm bir olayın odds'unun doğal logaritması alınarak yapılır. Bir olayın odds'u p/ (1-p) yada $\pi/(1-\pi)$ ile verilir ve bu oran 0 ile sonsuz (∞) arasında değer alabilir. Odds'ların doğal logaritması alındığında ise lojit dönüşüm yapılmış olur ve lojitler -∞ ve +∞ sonsuz arasında değer alabilir.

$$\text{Lojit } \pi(x) = g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \qquad 1.2$$

Lojistik regresyon modeli, yanıt değişkeninin odds'u türünden aşağıdaki gibi belirtilebilir.

$$\frac{\pi(x)}{1-\pi(x)} = e^{(\beta_0+\beta_1 x)} \qquad 1.3$$

Odds'un doğal logaritması alındığında model doğrusal modele dönüşür.

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln e^{(\beta_0+\beta_1 x)} = (\beta_0 + \beta_1 x) \qquad 1.4$$

Araştırmalarda önemli bir konu olan etken ile etkenlerle hastalık arasındaki lişkinin risk yönünden incelenmesidir. Lojistik regresyon Analizi değişkenlerin risk katsayılarını bularak bir risk analizi gibi kullanılır. Bağımlı değişken sonuç ve bağımsız değişkenlere de, bazen risk faktörleri denir. Bağımlı değişken dikkate alınarak, her bağımsız değişkenin ne kadar risk taşıdığı bulunmaktadir. Bu riskler odds katsayıları ile verilir.

Odds ve Odds oranı önemli kavramlardır. Odds gerçeleşen olay sayısının gerçekleşmeyen olay sayısına oranı olarak tanımlanır. Odds oranı olma olasılığının olmama olasılığına oranı olarak tanımlanır. Odds oranı güven aralığı (%95 CI): b ± 1, 96×SE şeklindedir. b katsayısının exp değerinin alınmış sonucudur (Çelik 2011: 392).

Lojistik regresyon analizinde katsayılarının kestirimi genellikle en çok olabilirlik yöntemi kullanılarak bulunur. En çok olabilirlik yönteminde, gözlenen veri setine elde etme olasılığını en büyük yapacak biçimde bilinmeyen parametreler için değerler üretilir. Bu yöntemi uygulayabilmek için olabilirilik fonsiyonu olarak adlandırılan bir fonksiyonun oluşturulması gerekir. Bu fonksiyon, gözlenen verinin olasılığını bilinmeyen parametrelerin bir fonksiyonu olarak belirtir. Bu fonksiyonu en büyük yapan değerler, bilinmeyen parametrelerin en çok olabilirlik kestiricileridir. Yani en çok olabilirlik yönteminde, bir olayın olması olasılığı en çok yapılmaya çalışılır.

Modeldeki değişkenlerin önemliliği olabilirlik oranı wald yada score testlerinden biriyle incelenebilir. Bağımsız değişkenin veya değişkenlerin önemliliği olabilirlik oranı G- istatistiği ile incelenir.

$$LR = G = -2LN\left(\frac{L(\text{değişken modelde olmadığında})}{L(\text{değişken modelde olduğunda})}\right) \quad 1.5$$

$$LR = G = -2(lnL\text{değişken modelde olma8dığında} - lnL(\text{değişken modelde olduğunda}))$$

LR asimtotik olarak ki-kare dağılır. Serbestlik derecesi, iki modelde kestirilen parametre sayısı arasındaki farka eşittir. Bu test olabilirlik oranı testi ya da sapma testi olarak adlandırılır. Bu test değerinin küçük olması, modele eklenen değişkenlerin lojit'in kestiriminde önemli bir katkı sağlamadığını ve değişkenlerin modelde bulunmasına gerek olmadığını gösterir. Bu test işlemi modeli uydurmak için gözlem sayısı (n) yeterince büyük olduğunda geçerlidir.

Wald testi değişkenlerin önemliliğini test etmek için kullanılan bir testtir. Wald testinde de olabilirlik oran testinde olduğu gibi beta katsayılarının en çok olabilirlik kestirimlerinden yararlanıılır. Wald testi, eğim parametresi $\beta_1$'in en çok olabilirlik kestiriminin ($\hat{\beta}_j$) standart hatasına S ($\hat{\beta}_j$)) bölünmesi ile elde edilir (Eşitlik (1.6)).

Modeldeki katsayıların test edlmesi için Wald $X^2$ istatistiği kullanılır (Eşitlik (1.7)).

$$W=\frac{\hat{\beta} j}{S(\hat{\beta} j)} \quad 1.6$$

$$W=\frac{\hat{\beta} j^2}{S(\hat{\beta} j)^2} \sim X^2 \quad 1.7$$

Bu değer serbestlik derecesi 1 olan $X^2$ dağılışı ile karşılaştırılarak karar verilir. Büyük örneklemler için olabilirlik oran testi ve wald testinin asimtotik olarak benzer sonuçlar verdiği belirtilmektedir. Küçük örneklemler için hangi testin daha iyi sonuç verdiği konusunda kuramsal bilgiler yetersiz olmakla birlikte wald testi yerine olabilirlik oranı testi kullanılması önerilmiştir (Alpar 2011: 624-627).

### 1.11.2.2. Regresyon

Regresyonun genel tanımı bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkiyi matematiksel modellerle açıklayarak bağıntı(lar) bulmak şeklinde verilir. Bulunan bağıntı değişik amaçlarla kullanılabilir. En önemli amaç kestirim yapmaktır. Bir bağımlı, bir bağımsız değişkenin olduğu doğrusal regresyon çözümlemesine "basit doğrusal regresyon çözümlemesi", bir bağımlı, birden çok bağımsız değişkenin olduğu doğrusal regresyon çözümlemesine "çoklu doğrusal regresyon çözümlemesi" denir (Alpar 2016: 459).

Bir bağımlı, bir bağımsız değişken arasında doğrusal ilişkinin var olduğu regresyon çözümlemesine basit doğrusal regresyon çözümlemesi denir. Basit doğrusal regresyonda, değişkenler arasındaki yapı doğrusal ise bu iki değişken için regresyon denklemi bulunabilir. $x_1$'e açıklayıcı değişken ya da etkileyen değişken de denir. y değişkeni ise, $x_1$ değişkenine bağlı olarak değiştiği düşünüldüğü için bağımlı değişken, açıklanan değişken ya da etkilenen değişken denir. $x_1$ değişkeni raslantıya bağlı değişken olabilir yada araştırıcı tarafından $x_1$ değişkeni denetlebilir olduğu durumlar vardır. Bu durum regresyon analizini, korelasyon analizinden ayıran önemli bir varsayımdır. y altkümelerinin oluşturduğu dağılımlara ilişkin ortalamalar bir doğru üzerindedir, bu varsayıma doğrusallık varsayımı denir (Alpar 2011: 413-415).

y bağımlı değişkeni çoğunlukla ölçümle elde edilen sürekli bir değişkendir. Sayımla elde edildiği durumlarda mevcuttur. Bu tür değişkenler, değişik dönüşümler yardımıyla sürekli değişken durumuna getirilebilir. Basit doğrusal regresyon modeli şu şekilde ifade edilir (Alpar 2011: 412-413).

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \qquad (i = 1,2,3,\ldots\ldots,n) \qquad 1.8$$

$\beta_0$ ve $\beta_1$ , regresyon modelinin bilinmeyenleridir. Bu bilinmeyenlere parametre yada regresyon katsayıları denir.

$B_0$; Regresyon doğrunun y eksenini kestiği noktayı gösterirken kesim nokttası veya sabit (değişmez) gibi adlar alır.

$B_1$; Regresyon katsayısı olarak adlandırılır ve bağımsız değişkende bir birimlik değişme (artma yada azalma) olduğunda, bağımlı değişkende meydana gelecek ortalama değişiklik miktarını verir.

$\epsilon_i$ ; $y_i = \beta_0 + \beta_1 x_{i1}$ doğrusal regresyon denlemi, y ile $x_1$ arasındaki gerçek ilişkiyi kabul edilebilir bir yaklaşım sağlar, ve y, $x_1$'in doğrusal bir fonksiyonu iken $\epsilon_i$ ; bu yaklaşımdan sapmaları ifade ettiğinden hata terimi olarak adlandırılır ve y'deki değişimin regresyon modeli ($\beta_0 + \beta_1 x_{i1}$ ) ile açıklanamayan bölümü tanımlar. ($\epsilon_i \sim N(0, \sigma^2)$) (Alpar 2011: 414).

Basit doğrusal regresyon çözümlemesinin önemli bir amacı bilinmeyenlerin ($\beta_0$ ve $\beta_1$) kestirilmesidir. Parametrelerin tahmini en küçük kareler yöntemi ile yapılır. Bunun için modelde yer alan artıklardan (residuals) yararlanılır. $\beta_0$, $\beta_1$ ve $\sigma^2$ parametrelerin tahminleri sırasıyla $b_0$ , $b_1$ve $\widehat{\sigma^2}$ ile gösterilecek olursa artıklar şu şekilde ifade edilir (Rawling ve diğerleri 1998: 3).

$e_i = y_i - (b_0 + b_1 x_{i1})$ $\quad i = 1,2,3,\ldots\ldots,n$ 1.9

Burada $e_i$ 'ye artık denir.$e_i$ 'ler, hataların ($\epsilon_i$) kestirimi olarak düşünülebilir.

Regresyon denleminin katsayıları şu şekilde bulunur.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_i x_n)^2 \qquad 1.10$$

Sağ taraf sıfıra eşitlendiğinde $b_0$ ve $b_1$'e gore türev alınırsa artık kareler toplamı olan $\sum_{i=1}^{n} e_i^2$ 'yi en küçük yapan $b_0$ ve $b_1$ katsayılarının formülleri bulunur. Eşitlik 1.11 ve 1.12'de verilmiştir.

$$b_0 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{i1} - \overline{x_1})}{\sum_{i=i}^{n}(x_{i1} - \overline{x_1})^2} = \frac{\sum_{i=1}^{n} x_{i1} y_i - \frac{\sum_{i=1}^{n} x_{i1} \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_{i1}^2 - \frac{\left(\sum_{i=1}^{n} x_{i1}\right)^2}{n}} = \frac{ÇT_{XY}}{KT_X} \qquad 1.11$$

$$b_0 = \bar{y} - b_1 \overline{x_1} \qquad 1.12$$

Yukarıdaki eşitlikte;

$KT_x$: $x_i$ lerin ortalamaya göre düzeltilmiş kareler toplamı yada X Ortalamadan Ayrılış Kareler Toplamı da (XOAKT) denir.

$ÇT_{xy}$: düzeltilmiş çarpımlar toplamı denir.

$\bar{y} = \sum_{i=1}^{n} yi/n$ ve $\bar{x}_{i1} = \sum_{i=1}^{n} x_{i1}/n$ 1.13

$\hat{y}_I = b_0 + b_1 x_{i1}$ 1.14

Bu denklemde, her bir $x_{i1}$ değerinin yerine konması ile elde edilen $\hat{y}_I$ değerleri regresyon doğrusu üzerinde olacaktır (Alpar 2011: 417).

Çoklu doğrusal regresyon denklemi;

$$\hat{y}= \beta_o + \beta_{1x_1} + \beta_{2x_2} + \beta_{3x_3} + \cdots . + \beta_{nx_n} + \epsilon \qquad 1.15$$

Basit doğrusal regresyon ve çoklu doğrusal regresyonda, bir takım varsayımların doğrusallık, normallik, varyansların homojenliği gibi yerine getirilmesinden sonra uygulanabilen bir test istatistiğidir. Varsayımların yerine gelmemesi durumunda bilinen bir takım dönüştürme işlemleri yapılarak veri kümesi, uygun hale getirilmeye çalışılmaktadır. Bu da veri setindeki orijinal değerlerin ya logaritmik dönüşümleri, ya da karekök gibi dönüşüm yöntemleri ile yapılabilmektedir. Parametrik yöntemlerde her ne kadar dönüştürme metotları kullanılarak ön koşul varsayımlar yerine getirilmeye çalışılıyorsa da, yapılan analiz dahilinde veri setine ilişkin yanlı sonuçlar elde edilmesi söz konusu olabilmektedir. Bu nedenle alternatifin olmadığı durumlarda verileri dönüştürme yoluna gidilmesi istatistiki açıdan daha doğrudur. Bahsedilen avantajlı yönleriyle çoklu regresyon analizine alternatif sayılabilecek ve çoklu regresyon analizinin gerektirdiği bir takım varsayımları taşımayan veri madenciliğinde kullanılan destek vektör regresyon, random forest ve regresyon ağacı yöntemleri kulanılmaktadır (Kayrı, Bostan 2008: 168).

#### 1.11.2.2.1. Destek Vektör Regresyonu

Destek vektör makinaları; istatistiki öğrenme teorisine dayanan kontrollü bir sınıflama algoritmasıdır. Bu algoritmalar ilk etapta iki sınıflı doğrusal verilerin sınıflandırılması sorunu için tasarlanmış ve daha sonraları ise çok sınıflı doğrusal olmayan verilerin sınıflandırılması amacıyla genelleştirilmiştir. Bu sınıflamanın basit bir ilkesi olup, iki sınıfı birbirinden ayırabilen en uygun karar fonksiyonunun tahmin edilmesi yöntemine dayanmaktadır (Kavzoğlu, Çölkesen 2010: 75).

Yapısal riskin minimum düzeye indirgenmesi amacıyla ortaya atılmış bir öğrenme türüdür. Destek vektör makinelerinin eğitim verileri çok az olduğu durumlarda dahi genelleme yetenekleri oldukça yüksektir. Buna ek olarak hiçbir yerel minimum içermezler. Kuadratik programlama problemi şeklinde formüle edildikleri

için, problem kuadratik programlama teknikleriyle çözülebilmektedir (Tayyar, Tekin 2013: 196)

Destek vektör (SV) algoritması, 1960'lı yıllarda Rusya'da geliştirilen genelleştirilmiş Portre algoritmasının doğrusal olmayan bir genellemesidir. 1970 li yıllara doğru gelişen başarılı bir çalışmanın ürünü haline gelmiştir. Regresyon için SVM'nin bir sürümü Vapnik, Steven Golowich ve Alex Smola tarafından 1997'de tasarlanmıştır. Bu yöntem daha sonra yapay zeka uzmanları ve matematikçiler tarafından oldukça ilgi görmüş ve sıklıkla kullanılır hale gelmiştir (Hacıefendioğlu 2012: 22-23).

Destek Vektörler sınflandırma ve regresyon için umut verici ve gelecek vaat eden çözüm yöntemidir (Chang, Lin 2005: 1). DVR regresyon teknikleri arasında en çok kullanılan yöntemdir (Hoa, Lin 2012: 3323).

Vapnik Chervonenkis ve diğerleri tarafından geliştirilen destek vektör regresyonu'nun temel özelliklerinden biri gözlenen çalışma hatasını en aza indirmek yerine, genelleştirilmiş performansa ulaşmak için genelleme hatası sınırını en aza indirmeyi denemesidir. Destek vektör regresyonu aynı zamanda zaman serisi tahminlerinde de kullanılmaktadır (Wu ve diğerleri 2009: 4276).

Klasik Regresyon modeli tanımlayan formül ile başlandığında örnek model şöyle olsun:

$$f(x) = (\omega, x) + b \qquad ve\ \omega \in X, b\ \in R \qquad 1.16$$

Yukarıdaki denklemde düzlemsellik durumu gözlendiğinde amaç en minimum $\omega$ yi aramaktır. Bunu yapabilmek için amaç fonksiyonu minimum olan optimizasyon problemi kurulur (Smola, Schölkopf 2004: 200). Örnek veri seti ($X_1$,$Y_1$),($X_2$, $Y_2$),…….($X_i$,$Y_i$) olsun ve $x_i$ girdi vektörü $y_i$ çıktı vektörü $x_i$ değerleri ile ilişki olsun. Bu durumda kurulan destek vektör regresyonu probleminin çözümü optimizasyon modelinin çözümü ile olacaktır. Optimizasyon problemi:

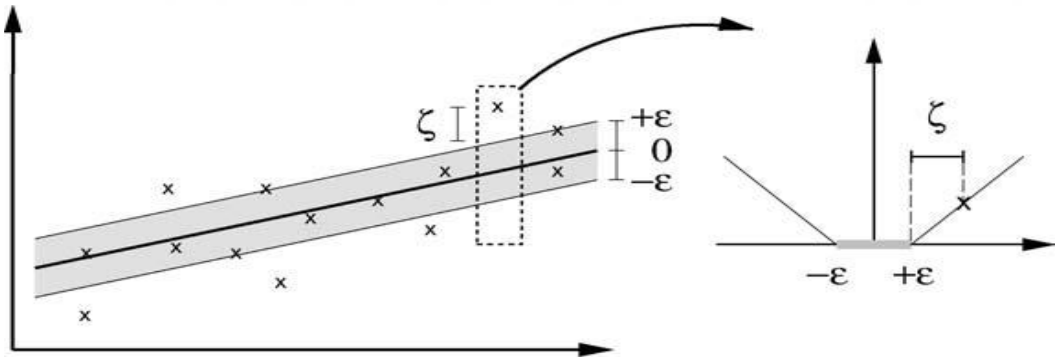
$$\min \qquad \tau\left(\omega, \xi^{(*)}, \varepsilon\right) = \frac{1}{2}\|\omega\|^2 + C\left(V\xi + \frac{1}{l}\sum(\xi_i + \xi_i^*)\right) \qquad 1.17$$

$$y_i - (\omega^T \phi(x_i) + b)\ \leq \varepsilon + \xi_i \qquad 1.18$$

$$(\omega^T \phi(x_i) + b) - y_i \leq \varepsilon + \xi_i^* \quad 1.19$$

$$\xi_i, \xi_i^* \geq 0 \quad \text{i=1,……, } I \quad V\epsilon\ (0,1)$$

$I$ = örneklem sayısı, i-örneğinin vektörü $x_i$ çekirdek fonksiyonu $\phi$ tarafından yüksek boyutlu uzayda eşleştirilmiş vektör, $\xi_i$ = üst sınır sapma hatası, $\xi_i^*$ = alt sınır sapma hatası, $\varepsilon$ = yoğunluk ve $y - (\omega^T \phi(x) + b) \leq \varepsilon$ tüp. Destek vektör regresyonunun tutarlılığı şu parametreler ile kontrol edilir. Bunlar hatanın maliyeti, Tüp'ün genişiliği ve çekirdek fonksiyon olarak adlandırılan ilişkilendirme fonksiyonudur. Destek vektör regresyonun modellenmesinde amaç veri seti $x_i$ 'nin yüksek boyutlu uzayda doğrusal olmayan ilişkilendirme üzerinden çözümünü yapmaktır. Çekirdek fonksiyonu $\phi$ girdi uzayı ile özellik uzayı arasında doğrusal olmayan bir ilişki kurar. $a_i^*, a_i$ lagrange çarpımlarıdır.




**Şekil 1.16.** Alt ve Üst Limitler

Destek vektör regresyonunda kullanılan çekirdek fonksiyonlar şunlardır:

$$Gaussan\ Radial\ Çekirdek\ Fonksiyonu: k(x_i, x_j) = \exp(-\gamma \|x - xi\|^2) \quad 1.20$$

$$Polinom\ Çekirdek\ Fonksiyonu: k(x_i, x_j) = (1 + x_i . x_j)^d \quad 1.21$$

$$Doğrusal\ Çekirdek\ Fonksiyuonu: k(x_i, x_j) = x_i^T x_j \quad 1.22$$

$x_i$ ve $x_j$ girdi vektör uzayıdır (Wu, Kumar: 2009: 45)

Bunun sonucunda v-SVR nin optmizasyon problem v≥0, c> 0

$$mak\ W(a^{(*)}) = \sum_{i=1}^{I}(a_i^* - a_i)\, y_i - \frac{1}{2}\sum_{i,j=1}^{I}(a_i^* - a_i)(a_j^* - a_j)k(x_i, x_j) \quad 1.23$$

$$st\ \ \sum_{i=1}^{I}(a_i - a_i^*) = 0 \quad 1.24$$

$$0 \le a_i^{(*)} \le \frac{C}{I} \quad 1.25$$

$$\sum_{i=1}^{I}(a_i + a_i^*) \le C.v \quad 1.26$$

En son SVR bu form da gösterilebilir

$$f(x) = \sum_{i=1}^{I}(a_i^* - a_i)k(x_i, x) + b \quad 1.27$$

#### 1.11.2.2.2. Regresyon Ağacı

Regresyon ağacı (CART) metodu 1984 yılında Berkley ve Stanford araştırmacıları Leo Breiman, Jerome Friedman, Richard Olshen, ve Charles Stone tarafından tanıtılmıştır. Regresyon modelleri bir veya birden çok tahminleyiciden sürekli değişken değerini tahmin etme olarak tanımlanır. Tahminleyiciler sürekli ya da kesikli değişkenler olabilir. Örneğin herhangi bir hastanın kandaki şeker seviyesini bağımsız değişkenler olan vücut kitle endeksi yada kandaki kolesterol seviyesi kullanılarak tahmin edilebilir. Sınıflama problemleri ise kategorik değişkenlere benzer özellikleri olan problemlerdir. Örneğin tedavi sürecinde uygulanan tedavi yöntemlerine rağmen hastaların tedaviden neden vazgeçtikleri yada durduklarını yaşadıkları coğrafyanın genel özelliklerini yada etnik kökenine bağlı değişkenlerin özelliklerini inceleyerek açıklayabiliriz (Sullivan 2012: 130).

Temel anlamda sınıflama ve regresyon ağaçları (CART), ağaçlara dayanan algoritmalar üreterek vakaları tahmin etmeye ve sınıflamaya yarayan, mantıksal şartları gözlememize yardımcı olur. CART süreci araştırmacının sırasıyla basit sorular kullanılarak oluşturulur.

CART algoritması ağacın yapı şeklini kullanan sınıflama ve regresyon modelleri oluşturmak için kullanılan çözüm yöntemidir. Tek değişkenli olarak kurulan modeller ikili olan karar konseptini içerir.

CART algoritması modelde kullanılan eldeki verileri alt kümelere ayırdığı için, mevcut alt kümenin içindeki durum bir önceki alt kümeden daha homojendir.

Kategorik veri içeren bağımlı değişkenlerin var olduğu türden karar ağaçlarına sınıflama ağaçları, sürekli değerler içeren bağımlı değişkenlerin kullanıldığı karar ağaçlarına ise regresyon ağaçları denmektedir.

#### 1.11.2.2.2.1. Dallandırma İşlemi

CART ağaçlarını oluşturmanın süreci şu adımlardan oluşur.

- Tahminlerin tutarlı olması için hedef kriterin çok iyi belirlenmesi ve tutarlı olması gerekir.
- Ağaçların oluşturulması için dalların tanımlanması gerekir.
- Ağaçların oluşturulmasından sonra gerektiğinde dalların iptal edilerek ağacın budanması.
- Tüm işlemlerden sonra doğru olan yada tam ölçekli ağacın seçilmesi (Sullivan 2012: 130).

#### 1.11.2.2.2.2. Tahminlerin Doğruluğu

CART algoritması teoride ve pratikte en uygun, doğru tahmini yapmak için kullanılır. Diğer manada ise en az oranda yanlış sınıflama yapmaya çalışmaktır. Bu algoritmanın bir diğer amacı da minimum düzeyde maliyet ile en doğru tahmini yapmaktır. Bu maliyetler genelde verilerin kullanılması, analiz edilmesi ve modele uygulanmasından kaynaklanan zamandan oluşur. Aynı zamanda en az yanlış modeli de elde etmektir. Çünkü yanlış modelin zamandan başka uygulandığı problem içinde başka türlü maliyetleri de vardır. Örneğin herhangi bir sınıflama problemlerinde seçilen kategorik değişkenlere uygun ön olasılıklar verilmez ise sonuç istenildiği gibi olmayacağı için yanlış çıkarımlara neden olacaktır (Sullivan 2012: 134).

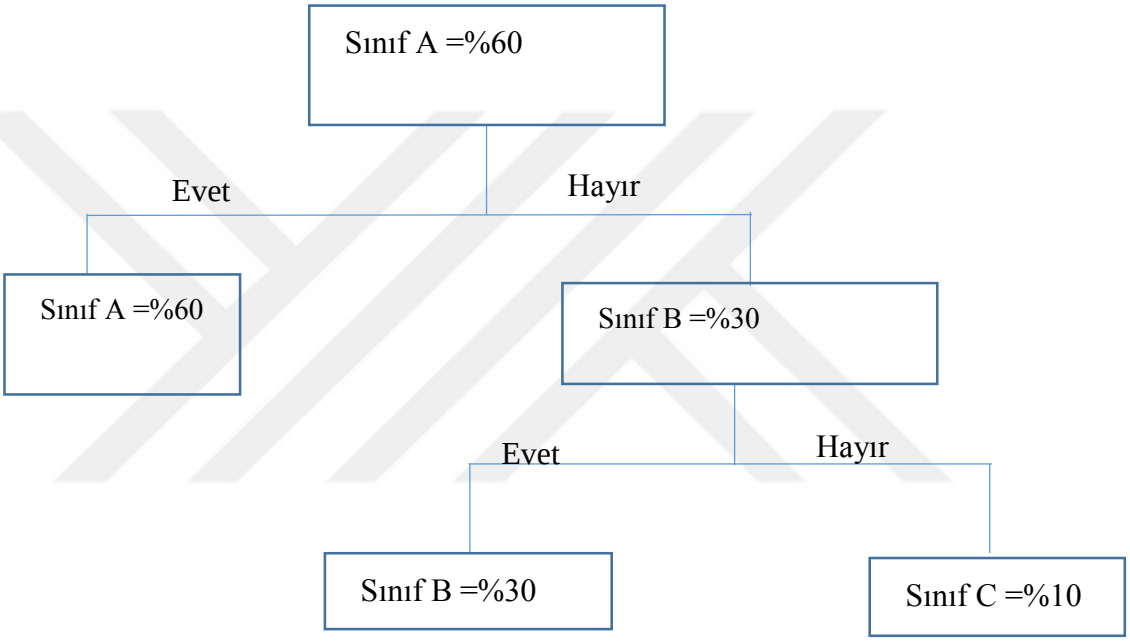
#### 1.11.2.2.2.3. Dalların Tanımlanması ve Dallandırma Tekniği

Doğru bir sınıflama süreci için veri seti içinde tanımlanan her vaka da dalların doğru oluşturulması gerekir. Çünkü çoğu durumda verideki kompleks yapıdan dolayı oluşturulan ağaç olması gerekenden daha karmaşık olur ve bu durumdan dolayı dallandırma işlemi aşırı şekilde gereğinden fazla olur. Aşırı uyum sağlama dolayısı ile tutarsız tahminlere neden olur. Bu yüzden dallandırma tekniği için farklı yöntemler

kullanılmıştır. Bunların en bilineni bilgi kazanımı, gini dallandırma tekniği ve twoing tekniği, $X^2$ ve entropidir (Sullivan 2012: 134).

#### 1.11.2.2.2.3.1. Gini Dallandırma Tekniği

Gini tekniği tek bir düğüm içindeki sınıfa düşen diğer tüm verileri dışlayan en geniş ölçekli veri sınıfına bakar. Örneğin 3 sınıftan oluşan veri seti için, A=%60, B=%30, C=%10 dir. Gini tekniğine göre dallandırma aşağıdaki gibi olur.




**Şekil 1.17.** Gini Ağacı

Gini tekniğine göre diğerlerine göre büyük olan sınıf sayısı ayrı yeni bir düğüme konurken, diğerleri birlikte diğer yeni düğüme konulur. Bir sonraki işlemler aynı şekilde yapılır ve en küçük sınıf sayıları elde edilene kadar devam eder. Fakat gerçek hayatta bu kadar düzgün ağaç elde edilmesi mümkün değildir.

$Gini(n) = 1 - \sum_{i=1}^{m}[p(i|n)]^2$ Her ne kadar en büyük sayıya sahip olan kendi düğüm noktasında olsa bile, büyük sınıftan bazı miktar yada oran diğer düğüm noktalarına kayabilir. Gini metodu her defasında tek ve en geniş veri setine odaklanır. Herhangi bir "n" noktasının Gini indeksi; $p(i|n)$ sınıf i'nin düğüm noktası "n" deki göreli frekansını gösterir. Maksimum değer $\sum_{i=1}^{m}[p(i|n)]^2$ nin en küçük değerinde hesaplanır. Tersinde ise minimum değer $\sum_{i=1}^{m}[p(i|n)]^2$nin maksimum olduğu

değerinde hesaplanır. Gini tekniği kullanıldığında en küçük Gini indeksi elde edilmeye çalışılır (Sullivan 2012: 135).

#### 1.11.2.2.2.2.3.2. Twoing Kuralı

Gini tekniğinin farklı bir yaklaşımı olan bu teknikte, sınıf gruplarını bulmak için hepsini beraber kullanarak %50 yakın olanları beraber alarak bir düğüme yerleştirir. Mesela 4 sınıftan oluşan bir grup; A=%40, B=%10, C=%20, D=%30 olsun. Bu durumda A ve B bir grup ve C ve D de başka bir grup olacaktır. Fakat bir çok sınıflamada rakamlar arasında bu tip yakınlık olmayabilir. Ama teknik yine aynı şekilde devam eder (Sullivan 2012: 137).

#### 1.11.2.2.2.2.3.3. Entropi

Entropi yöntemine göre hesaplama sonucu geçerli bilgi düğümün ayrılması ile elde etme yöntemine dayanır. Twoing yöntemine benzeyen yanları vardır. Gözlenen değişkene bakılarak, dallandırmayı incelemek için bu değişken kullanılarak en yüksek bilgi elde edilemeye çalışılır. Kullanılan bu değişken bölümlendirme işleminde, sonuç bölümlendirmelerindeki veriyi sınıflama için gerekli olan bilgiyi minimuma indirir ve bilgi kirliliğini en aza indirir. Bu yaklaşımı kullanma, verilen örneği sınıflama için kullanılan testlerin sayısını minimuma indirir. Bu sayede ayrıca basit ağaçlar oluşturmamızı sağlar. Her ne kadar basit ağaçlarda oluştursa sonuç çözümünün ağacı basit olmayabilir. Yöntemin matematiksel açıklaması şöyledir. Mevcut veri seti "D" nin bir kısmını kullanmak üzere, "m" tane farklı sınıflama $C_i$ i=1,…..m, ihtiyaç duyulan gerekli bilgi kazanımı şu formül ile verilir.

$$I_D = -\sum_{i=1}^{m} p(i|t)\log_2 p(i|t) \qquad 1.28$$

$p(i|t)$ , $C_i$ sınıfına ait D veri kümesine ait herhangi bir örneğin göreli frekansı şu şekilde tahmin edilir. $I_D$ aynı zamanda D veri kümesinin Entropisidir.

$$p(i|t) = \frac{|C_{i,D}|}{|D|} * I_D \qquad 1.29$$

$I_D$ aynı zamanda verilen herhangi bir düğümün homojenliğini de ölçer. Gini metot da olduğu gibi maksimum değer, örnekler minimum değerli, az bilgiye sahip sınıflar arasına eşit olarak dağılır. Bu aynı zamanda en çok bilgiye sahip sınıfa ait örnekler olduğunda gerçekleşir (Sullivan 2012: 138).

#### 1.11.2.2.2.2.3.4. Bilgi Kazanımı

Bir daldaki bilgi kazanımı şu şekilde ifade edilir.

$$\text{Kazanım} = I_D(t) - \left(\sum_{i=1}^{k} \frac{n_i}{n} I_D(i)\right) \quad 1.30$$

$I_D$ entropi ölçüsü olmak üzere "t" ana düğüm noktası, "k" daldaki bölümlendirme sayısı, $n_i$ i bölümdeki örnek sayısı, "n" ise "t" düğümündeki örneklerin toplam sayısını ifade eder.

Bilgi kazanımı dallandırmadan kaynaklanan entropi'yi azaltmayı amaçlar. Bilgi kazanımının en fazla olduğu düğüm aynı zamanda entropinin en az olduğu düğümdür.

Aynı zamanda kazanım oranını hesaplamak da mümkündür. Kazanım oranı;

$$\text{Kazanım}_{\text{oranı}} = \frac{I_D - \left(\sum_{i=1}^{k} \frac{n_i}{n} I_D(i)\right)}{I_D(t)} \quad 1.31$$

$$I_D = -\sum_{i=1}^{m} p(i|t) \log_2 p(i|t) \quad 1.32$$

Eğer $I_D$ yüksek değer alırsa geniş sayıda küçük bölmeler elde edilir ve bu kazanım oranı değerini düşürür (Sullivan 2012: 139).

#### 1.11.2.2.2.2.3.5. Yinelemeli Bölünme

En iyi ayırım bulunduğunda, CART aşağıda kalan her düğüm için araştırma sürecine devam eder. Bir sonraki ayrım ya bir kriter ile durdurulur yada ayırmak artık imkansızdır. Bu durumda genelde ayırım şu şartlarda durur (Han ve Kamber 2006: 296).

1-En az sayıda vakaya ulaşılmıştır.

2- Vakaların toplam sayısının belirli bir bölümü düğümün içindedir.

3- Ayırma işleminde en üst seviyeye ulaşılmıştır.

4- En fazla düğüm sayısına ulaşılmıştır.

Eğer düğüm içinde sadece bir vaka kaldıysa, diğer tüm vakalar birbirinin çifti ise ve düğümdeki tüm hedef değerler aynı ise daha fazla ayırım yapmak mümkün olur.

#### 1.11.2.2.2.2.3.6. Ağaçları Budama

Budamayı durdurmaya odaklanmaktan ziyade, CART işlevi, CART ağaçları ihtiyaç olduğundan fazla büyüdüğü zaman en iyi ağacı bulmak için budamak gerekir. Bu teknik bir anlamda sınıflama ağacını budama işlevi ile benzer özellikler taşır (Izenman 2008: 295). CART süreci V-katlı çapraz doğrulama sürecini kullanarak yada test verisi kullanırken iyi ağacı bulmayı ve test etmeyi amaçlar.

Modeli deneme için model oluşturulurken kullanılmayan veri seti kullanılarak ağaçlar puanlanarak test sürecinden geçirilir. Çapraz doğrulama yeniden örnekleme şekline benzer. Yani toplam dağılımdan örnek sayılar çeker ve modelleri tüm örnekler üzerinde dener.

V-katlı çapraz doğrulama şu şekilde yapılır.

1- Tüm veri seti V-katlı parçalara bölümlenir.
2- V-1 katlı parçanın farklı kombinasyonları üzerinde V tane model denenir, her defasında V'inci kat kullanılarak hata tahmin edilir.
3- Yeni veri üzerindeki ağaçların doğruluğunu tahmin etmek için V hata ölçümünün ortalaması kullanılır.
4- Üçüncü adımdaki hataları minimize eden dizayn parametreleri (karmaşıklık cezası) seçilir.
5- Adım dörtteki parametreler kullanılarak, tüm verileri kullanarak ağaç yeniden tamir edilir.

### 1.11.2.2.3. Rassal Ormanlar (Random Forest)

Rassal ormanlar algoritması biçimsel olarak Brieman tarafından çalışılmıştır. Rassal ormanlar aynı zaman toplu öğrenme metotlarından ve algoritmalarından biridir. (Torgo 2014: 254). Kullanılan bu algoritma rassal alt uzaylar ve paketleme kavramını bir araya getirir. Yani Rassal ormanlar paketleme tekniği kullanılarak oluşturulabilir.

Topluluk içindeki her bir sınıflayıcı bir karar ağacıdır ve bu karar ağaçları bir araya geldiğinde rassal ormanları oluştururlar. Tekil karar ağaçları, ayrımları gözlemleyebilmek için her düğümdeki davranışların rassal seçimi kullanılarak oluşturulur (Han ve diğerleri 2012: 383).

Rassal orman algoritması her değişken aralığından yapılan rassal seçimleri içeren ve her bir alt kümeye eklenen verilerin, farklı alt kümeleri üzerindeki ağaçlara rakamların yazıldığı algoritmadır. Bu tip gruplar genelde topluluk olarak da adlandırılır. Topluluk içindeki her bir karar ağacı her girdi durumunun sınıflaması için oy kullanır (Nısbet ve diğerleri 2009: 851).

Rassal ormanlarda ağacın büyümesinin adımları şöyledir.

1. Vaka sayılarından rassal bir örnek alınır. Diğer ağaçlar için olan sonraki örnekler ile değiştirme yapılır. Yani daha önce gelen ağacın yapısında kalanlar olsa bile hiçbir olay dışarıda bırakılmaz.
2. Değişkenlerin sayısından daha az olmak üzere değişkenlerin bir alt kümesi seçilir. Ve değişkenlerin alt kümeleri üzerindeki en iyi ayırım seçilir. Değişken alt küme sayısı "M" olmak üzere M'in değeri arttıkça ağaçlar arasındaki korelasyon artar aynı zamanda ağaçların tahmin gücü de artar. Bu sebebden dolayı kullanılacak ağaç sayısı ne fazla ne de az olmalıdır. M sayısı aynı zamanda modelin duyarlılığında ve algoritmanın içinde de kullanılan önemli bir faktördür.
3. Kullanılan vakaların üçte biri örnek test veri seti oluşturmak için kullanılır. Bu test veri seti hata oranını tahmin etmek için kullanılır. Ortalama hata oluşturulan tüm ağaçlardan hesaplanır.
4. Değişken önemi ağacın altına doğru örneklem dışı veri kullanılarak hesaplanır. Ve tahmin edilen sınıflar için kullanılan kararlar sayılır. Her değişkenin içindeki değerler rassal değişir ve ağacın aşağısına doğru bağımsız şekilde devam eder. Etkinin ölçüsünü elde etmek için, değiştirilen değişkenli ağaç için karar sayısı değişmeyen ağaçlardaki karar sayısından çıkarılarak elde edilir. Değişkenin önem değerini elde etmek için tüm ağaçlar arası ortalama etki hesaplanır (Nısbet ve diğerleri 2009: 852).

Rassal ormanlar Breiman tarafından öne sürülen Rassal orman algoritması kısaca şu şekilde yazılır. Rassal ormanda var olan K tane ağaç olsun ve $\{T_1,\ldots\ldots T_K\}$, ve k th ağaçtan rassal vektör $Q_k$ rassal vektörler oluşturulsun ve k=1,….K. Oluşturulan bu vektörler $\{Q_k\}$ ağaç modelleme için kullanılacak olan benzer bağımsız dağılımlı vektörlerdir. Bu vektörler ağaç yapısı şeklinde ifade edilirler. Rassal seçimde bu

vektörler N tane ayırım noktası olan ve {1,…..N} aralığından rassal seçilen tamsayıların karışımından oluşur. "k"th ağaç sınıflandırıcı $f(x,Q_k)$ olursa oluşturulacak rassal orman $\{f(x,Q_k)\}$, 1,…..K şeklinde ifade edilir (Dua, Du 2011: 37).

Rassaol ormanın birçok avantajlı yöntemleri vardır.Aşağıdaki gibi sıralamak mümkündür;

1- Sınıflama için kullanılan algoritmalar arasında yüksek doğruluğa en yakın algoritmadır.
2- Çok geniş veri setleri ile çalışmak mümkündür.
3- Değişkenlerin önemini tahmin etmeye yardımcı olan algoritmadır.
4- Kayıp veri durumlarında kayıp veriyi işleyen kuvvetli bir yöntemdir. Düğümlerde yer alan tüm vakaların arasındaki değişkenler için en sık rastlanan değer kayıp verilerin yerine konarak algoritma kullanılır.
5- Yardımcı yazılımlar sayesinde oldukça hızlı olan bu algoritma sayesinde binlerce ağaç, vaka ve değişken birkaç dakika içinde kullanılarak hızlı sonuçlar alınabilir.

Rassal ormanlar sınıflama işlemi için kullanılabilirlik olarak özellikleri bir sıralama halinde elde etmek için kullanılabilir (Torgo 2014: 246).

#### 1.12.3.1. Random Forest Regresyonu

Rassal ormanlar regresyonu sınıf etiketlerinin aksine sayısal değer alan ağaç temizleyicisine h(x,Q) benzer, rassal vektörlere Q dayanan büyüyen ağaçlar tarafından şekillenir, çıktı değerleri sayısaldır ve deneme verisi rassal vektör Y,X dağılımından bağımsızca çekilmiştir. Herhangi bir tahminleyici h(x) için otalama kareli genelleştirilmiş hata;

$$E_{X,Y}(Y-h(X))^2 \qquad 1.33$$

Rassal orman tahminleyici k tane ağacın $\{h(x,Q_k)\}$ ortalaması alınarak şekillenir. Benzer şekilde sınıflama için şu adımlar uygulanır.

1- Eğer ormandaki ağaçların sayısı sonsuza giderse;

$$E_{X,Y}(Y-av_k h(X,Q_k))^2 \longrightarrow E_{X,Y}(Y-E_Q h(X,Q'))^2 \qquad 1.34$$

Olur.

2- Eğer tüm Q, $EY=E_X h(X,Q)$ ise;

$PE^*$ (orman) ormanının genelleştirilmiş hatası olmak üzere;

$PE^*(\text{orman}) \leq \bar{p} PE^*(\text{ağaç})$ ve $\bar{p}$= Y-h(X,Q) ve Y-h(X,Q') arasındaki ağırlıklandırılmış korelasyondur Q' ise bağımsızdır (Breiman 2001: 25-26).

# 2. BÖLÜM

# MATERYAL ve YÖNTEM

## 2.1. Materyal

Çalışmamıza 2009-2010-2011 yıllarına ait Cumhuriyet Üniversitesi Tıp Fakültesi Enfeksiyon Hastalıkları ve Çocuk Sağlığı Hastalıkları servisinde yatan kırım kongo kanamalı ateş (KKKA) tanısı ile tedavi gören tüm hasta bireylerin verileri servis kayıtlarından alınmıştır. KKKA, ülkemizde son yıllarda sık görülmeye başlayan ve mortalitesi yüksek olan viral bir enfeksiyon hastalığıdır ve bu virüse sahip kenenin insanı ısırmasıyla bulaşır. Özellikle iç anadolu bölgesinde Tokat, Sivas ve Yozgat illerinde çok görülmüştür. Bu üç yıl içerisindeki toplam 245 hastaya ait 6125 veri girişi yapılmıştır. Bunun 113'ü yetişkin, 132' si çocuk hastasıdır. Bu iki grup hastalar birleştirilerek üçüncü grup oluşturulmuştur. Çalışmamızda yetişkin, çocuk ve tüm hasta olmak üzere toplam üç grup hasta verisi kullanılmıştır. Bağımlı değişkenimiz hastanede yatış süresi olup diğer bağımsız değişkenler yaş, cinsiyet, yaşadığı yer, hayvancılık yapıp yapmama durumu, konjonktivit, sarılık, akciğer tutulumu, hepatomegali, splenomegali, bilinç değişikliği, baş ağrısı, kas ağrısı, boğaz ağrısı, bulantı, kusma, ishal, öksürük, ateş, kırıklık, kanama, döküntü, leukopenia, kan ürünü ihtiyacı ve semptom gün sayısı olmak üzere toplam 25 tane değişkenimiz vardır.

## 2.2. Yöntem

### 2.2.1. Regresyon Yöntemi

Veri madenciliğinde kullanılan regresyon yöntemlerinden destek vektör regresyon (DVR), random forest (RF) ve regresyon ağaçlarının (RA) kestirim performanslarına bakılmıştır. Destek vektör regresyonda radial çekirdek fonksiyonu kullanılmıştır. Regresyon modellerinin performanslarını karşılaştırmak için hata kareler ortalaması (HKO) ve açıklayıcılık katsayısı ($R^2$) ölçü alınmıştır.

### 2.2.2. Sınıflama Yöntemi

Hastanede yatış süresi kategorileştirilerek veri madenciliğinde kullanılan sınıflama algoritmalarından destek vektör makinası, random forest, sınıflama ve regresyon ağaçlarının (CART) duyarlılık, kesinlik, doğruluk oranı ve F ölçütüne bakılmıştır

## 2.3. İstatistiksel Analiz

Analizler, simülasyon çalışması ve grafikler için R 3.2.0 programı kullanılmıştır. R programı "http://cran.r-project.org/" sitesinden indirilen ücretsiz bir yazılım olup ismini programı geliştiren Yeni Zelanda Auckland Üniversitesinde Ross Ilhaka ve Robert Gentleman' den almıştır. R bircok istatistiksel analizler ve grafikler için kullanılan programlama dilidir. Programın içindeki uygun paketler yüklenerek, kodlar elle yazılır. Bu çalışmada kullanılan tüm kodlar EK:1, EK:2, EK:3, EK:4 ve EK:5 de verilmiştir.

### 2.3.1. Simülasyon çalışması

Çalışmada herhangi bir ülkenin serveri seçilip, e1071, MASS, randomForest, rpart, corpcor paketleri yüklenerek analizler gerçekleşmiştir. Simülasyon çalışması için veriler, mvrnorm fonksiyonu ile çok değişkenli standart normal dağılımdan türetilmiştir. Ortalama vektörü "0" olarak alınmıştır. Korelasyon matrisi orjinal veri setinden hesaplanmıştır. Nitel değişkenler orjinal veri setindeki frekanslar dikkate alınarak belirlenen kesim noktalarına göre orjinal veri sayısı kadar türetilmiştir. Simülasyon çalışmasında tüm senaryolarda 1000 tekrar yapılarak modelin performansları karşılaştırılmıştır.

İkinci simülasyon çalışmasında gözlem sayısı n=100, n=250, n= 1000 olmak üzere değişkenler arası korelasyon miktarı düşük ilişki (r= 0,00-0,20), orta ilişki (r= 0,21-0,40), yüksek ilişki (r= 0,41-0,60) olmak üzere veri türetilmiştir ve buna göre regresyon modellerinin performanslarının karşılaştırılmasına bakılmıştır. Korelasyon matrisi uniform (tekdüze) dağılımından türetilmiştir. Toplam değişken sayısı 25, nitel değişken sayısı 22, sürekli değişken sayısı 3'tür. Düşük korelasyon için illişki matrisi

(r=0,00-0,20) EK:6'da , orta korelasyon için ilişki matrisi (r=0,21-0,40) EK:7'de ve yüksek korelasyon için ilişki matrisi (r=0,41-0,60) EK: 8 de verilmiştir.

### 2.3.2. Sınıflamada Kullanılan Ölçütler

Sınıflamada kullanılan algoritmaların model başarısını değerlendirirken kullanılan temel kavramlar doğruluk, kesinlik, duyarlılık ve F-ölçütüdür. Modelin başarısı, doğru sınıfa atanan gözlem sayısı ve yanlış sınıfa atılan gözlem sayısı nicelikleriyle alakalıdır.

Test sonucunda ulaşılan sonuçların başarım bilgileri karışıklık matrisi ile ifade edilebilir. Karışıklık matrisinde satırlar test kümesindeki gözlemlere ait gerçek sayıları, kolonlar ise modelin tahminlemesini ifade eder.

| Öngörülen sınıf | | Kesinlik Sınıfı | |
|---|---|---|---|
| | | Sınıf= 1 | Sınıf= 0 |
| | Sınıf= 1 | A = DP | B=YP |
| | Sınıf= 0 | C= YN | D=DN |

A= DP (doğru pozitif) B=YP (yanlış pozitif)

C= YN (yanlış negatif) D=DN (doğru negatif)

#### 2.3.2.1. Doğruluk – Hata oranı

Diğer adı geçerlilik katsayısı olan doğruluk oranı model başarısının ölçülmesinde kullanılan en popüler ve basit yöntemdir. Doğru sınıflandırılmış gözlem sayısının (DP +DN), toplam gözlem sayısına (DP+DN+YP+YN) oranıdır. Hata oranı ise bu değerin 1'e tamlayanıdır. Diğer bir ifadeyle yanlış sınıflandırılmış gözlem sayısının (YP+YN), toplam gözlem sayısına (DP+DN+YP+YN) oranıdır.

$$\text{Doğruluk} = \frac{DP+DN}{DP+YP+YN+DN} \qquad 2.1$$

$$\text{Hata Oranı} = \frac{YP+YN}{DP+YP+YN+DN} \qquad 2.2$$

#### 2.3.2.2. Kesinlik

Kesinlik, sınıfı 1 olarak tahminlenmiş doğru pozitif gözlem sayısının, sınıfı 1 olarak tahminlenmiş tüm gözlem sayısına oranıdır. Pozitif kestirim değeri adını alır. Geliştirilen yeni test sonucunda hasta olanların gerçekten hasta olma olasılığını verir.

$$\text{Kesinlik}=\text{DP}/(\text{DP}+\text{YP}) \quad (2.3)$$

#### 2.3.2.3. Duyarlılık

Gerçek hastalar içinden hastaları ayırma yeteneğidir. Diğer bir ifadeyle doğru sınıflandırılmış pozitif gözlem sayısının toplam pozitif gözlem sayısına oranıdır.

$$\text{Duyarlılık}= \text{DP}/(\text{DP}+\text{YN}) \quad (2.4)$$

#### 2.3.2.4. F-Ölçütü

Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu vermek açısından yeterli değildir. Bunun için F-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır.

$$\text{F- Ölçütü}= ((2 \times \text{Duyarlılık} \times \text{Kesinlik}))/(\text{Duyarlılık}+\text{Kesinlik}) \quad (2.5)$$

#### 2.3.2.5. Hata Kareler Ortalaması

HKO $\hat{\theta}$' nın gerçek anakütle parametresinden ortalamada ne kadar uzakta olduğunu ölçer. HKO varyans ve yanlılığa (bias) bağlı olduğundan yanlı tahmin edicilerin karşılaştırılmasında kullanılabilir.

$$\text{HKO}=\frac{1}{n}\sum_{i=1}^{n}\left(\theta_i-\hat{\theta_i}\right)^2 \quad (2.6)$$

#### 2.3.2.6. Açıklayıcılık Katsayısı

İncelenen değişkenler bağımlı-bağımsız değişken olarak tanımlanabiliyorsa, açıklayıcılık katsayısı bağımlı değişkendeki değişimin yüzde kaçının bağımsız değişken tarafından açıklanabildiğini belirtir ve $R^2$ ile gösterilir.

# 3. BÖLÜM

## BULGULAR

### 3.1. Demografik Özelliklerin Dağılımı

KKKA verisine ait değişkenlerin demografik özelliklerinin frekans dağılımları Tablo 3.1., Tablo 3.2. ve Tablo 3.3.' de verilmiştir.

**Tablo 3.1.** Çalışmaya Alınan Bireylerin Cinsiyetine Göre Frekans Dağılımı

| Grup | Cinsiyet | Sayı | % |
|---|---|---|---|
| Çocuk | Erkek | 94 | 71,2 |
| | Kadın | 38 | 28,8 |
| | Toplam | 132 | 100,0 |
| Yetişkin | Erkek | 49 | 43,4 |
| | Kadın | 64 | 56,6 |
| | Toplam | 113 | 100,0 |
| Toplam | Erkek | 143 | 55,4 |
| | Kadın | 102 | 41,6 |
| | Toplam | 245 | 100,0 |

Tablo 3.1.'e göre çalışmaya alınan bireylerin cinsiyetine göre frekans dağılımı incelendiğinde; çocuk hastaların %71,2'si erkek, % 28,8'i kadın, yetişkin hastaların % 43,4'ü erkek, %56,6'si kadın, toplamda ise %55,4'ü erkek, % 41,6'sı kadındır.

**Tablo 3.2.** Hastaların Yaşadıkları Yere Göre Dağılımı

| Grup | Yaşadıkları yer | Sayı | % |
|---|---|---|---|
| Çocuk | Sivas | 124 | 93,9 |
| | Diğer il | 8 | 6,1 |
| | Toplam | 132 | 100,0 |
| Yetişkin | Sivas | 35 | 31,0 |
| | Diğer il | 78 | 69,0 |
| | Toplam | 113 | 100,0 |
| Toplam | Sivas | 159 | 64,9 |
| | Diğer il | 86 | 35,1 |
| | Toplam | 245 | 100,0 |

Tablo 3.2.' ye göre çocuk hastaların Sivas ilinde ikametgah etme oranı % 93,9' dur. Diğer iller ise bu oran % 6,1'dir. Yetişkin hasta grubunda Sivas'da ikametgah etme durumları %31'dir. Diğer illerden gelme oranı % 69, 0'dır. Geneline bakıldığında en çok hasta Sivas ilinden olup % 64,9'dur. Diğer ilden gelme oranı %35,1'dir.

**Tablo 3.3.** Hastaların Hayvancılık Yapma Durumunun Dağılımı

| Grup | Hayvancılık yapma durumu | sayı | % |
|---|---|---|---|
| Çocuk | Evet | 27 | 20,5 |
| | Hayır | 105 | 79,5 |
| | Toplam | 132 | 100,0 |
| Yetişkin | Evet | 97 | 85,8 |
| | Hayır | 16 | 14,2 |
| | Toplam | 113 | 100,0 |
| Toplam | Evet | 124 | 50,6 |
| | Hayır | 121 | 49,4 |
| | Toplam | 245 | 100,0 |

Tablo 3.3 İncelendiğinde çocuk hastaların hayvancılık yapma oranı %20,5 iken yetişkin grupta bu oran %85,8'dir. Toplam grupta bu oran % 50,6'dır.

## 3.2. Nicel değişkenlerin dağılımı

Hasta gruplarının nicel değişkenlerine göre tanımlayıcı istatistikleri Tablo 3.4.'de gösterilmiştir.

**Tablo 3.4.** Bireylerin Nicel Değişkenlere Göre Durumu

| Grup | Nicel Değişkenler | n | Min. | Max | Ortalama | Standart Sapma |
|---|---|---|---|---|---|---|
| Çocuk | Hastanede kalış süresi | 132 | 4 | 30 | 9,83 | 3,05 |
| | Semptomlu gün sayısı | 132 | 0 | 18 | 4,44 | 3,57 |
| | Yaş | 132 | 0 | 17 | 11,59 | 3,95 |
| Yetişkin | Hastanede kalış süresi | 113 | 0 | 17 | 8,62 | 2,90 |
| | Semptom gün sayısı | 113 | 0 | 11 | 5,22 | 2,32 |
| | Yaş | 113 | 18 | 79 | 46,65 | 17,62 |
| Toplam | Hastanede kalış süresi | 245 | 0 | 30 | 9,27 | 3,04 |
| | Semptom gün sayısı | 245 | 0 | 18 | 4,80 | 3,07 |
| | Yaş | 245 | 0 | 79 | 27,76 | 21,39 |

Tablo 3.4. incelendiğinde çocuk hastaların hastanede kalış süreleri ortalama 9,83 ± 3,05, yetişkin grupta 8,62 ±2 ,90 ve tüm hasta grubunda ise 9,27 ± 3,04'tür. Semptomlu gün sayısı incelendiğinde ise çocuk hastalarında 4,44±3,57, yetişkin grupta 5,22±2,32 ve tüm hasta grubunda ise 4,80±3,07'dir. Yaşların ortalaması incelendiğinde çocuk yaş grubunun ortalaması 11,59± 3,95, yetişkin yaş grubunun yaş ortalaması 46,65± 17,62 ve tüm hasta grubuna bakıldığında yaş ortalaması 27,76 ± 21,39'dur.

## 3.3. Bağımsız Kategorik Değişkenlerin Dağılımı

Hasta gruplarının bağımsız ve kategorik değişkenlere göre frekans dağılımı Tablo 3.5.'te verilmiştir.

**Tablo 3.5.** Hasta Gruplarının Kategorik Değişkenlere Göre Dağılımı

| Semptomlar | Gruplar | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Çocuk | | | | Yetişkin | | | | Toplam | | | |
| | Var | | Yok | | Var | | Yok | | Var | | Yok | |
| | n | % | n | % | n | % | n | % | n | % | n | % |
| **Konjonktivit** | 55 | 58,3 | 77 | 41,7 | 59 | 47,8 | 54 | 52,2 | 114 | 46,5 | 131 | 53,5 |
| **Sarılık** | 0 | 0 | 132 | 100 | 8 | 7,1 | 105 | 92,9 | 8 | 3,3 | 237 | 96,7 |
| **Akciğer Tutulumu** | 95 | 72 | 37 | 28 | 6 | 5,3 | 107 | 94,7 | 101 | 41,2 | 144 | 58,8 |
| **Hepatomegali** | 3 | 2,3 | 129 | 97,7 | 25 | 22,1 | 88 | 77,9 | 28 | 11,4 | 217 | 88,6 |
| **Splenomegali** | 2 | 1,5 | 130 | 98,5 | 8 | 7,1 | 105 | 92,9 | 10 | 4,1 | 235 | 95,6 |
| **Bilinç Değişikliği** | 3 | 2,3 | 129 | 97,7 | 4 | 3,5 | 109 | 96,5 | 7 | 2,9 | 238 | 97,1 |
| **Başağrısı** | 42 | 31,8 | 90 | 68,2 | 92 | 81,4 | 21 | 18,6 | 134 | 54,7 | 111 | 45,3 |
| **Kas Ağrısı** | 20 | 15,2 | 112 | 84,8 | 102 | 90,3 | 11 | 9,7 | 122 | 50,2 | 123 | 49,2 |
| **Boğaz Ağrısı** | 6 | 4,5 | 126 | 95,5 | 17 | 15 | 96 | 85 | 23 | 9,4 | 222 | 90,6 |
| **Bulantı** | 58 | 43,9 | 74 | 56,1 | 92 | 81,4 | 21 | 18,6 | 150 | 61,2 | 95 | 38,8 |
| **Kusma** | 72 | 54,5 | 60 | 45,5 | 81 | 71,7 | 32 | 28,3 | 153 | 62,4 | 92 | 37,6 |
| **İshal** | 23 | 17,4 | 109 | 82,6 | 39 | 34,5 | 74 | 65,5 | 62 | 25,3 | 183 | 74,7 |
| **Öksürük** | 96 | 72,7 | 36 | 27,3 | 25 | 22,1 | 88 | 77,9 | 124 | 50,6 | 121 | 49,4 |
| **Ateş** | 127 | 96,2 | 5 | 3,8 | 104 | 92 | 9 | 8 | 231 | 94,3 | 14 | 5,7 |
| **Kırıklık** | 53 | 40,2 | 79 | 59,8 | 106 | 93,8 | 7 | 6,2 | 159 | 64,9 | 63 | 25,7 |
| **Kanama** | 23 | 17,4 | 109 | 82,6 | 40 | 35,4 | 73 | 64,6 | 63 | 25,7 | 182 | 74,3 |
| **Döküntü** | 16 | 12,1 | 116 | 87,9 | 24 | 21,2 | 89 | 78,8 | 40 | 16,3 | 205 | 83,7 |
| **Leukopenia** | 16 | 12,1 | 116 | 87,9 | 32 | 28,3 | 81 | 71,7 | 48 | 19,6 | 197 | 80,4 |
| **Kan ürünü ihtiyacı** | 7 | 5,3 | 125 | 94,7 | 73 | 64,6 | 40 | 35,4 | 80 | 32,7 | 165 | 67,3 |

Tablo 3.5.'e göre konjonktivit görülme oranı çocuk grubunda % 58,3' tür, yetişkinde ise bu oran % 47,8 dir, genelinde ise durum % 46,5 'tir. Sarılık değişkenine

bakıldığında çocuk grubunda hiç yoktur, yetişkin grupta bu oran %7, 1'dir. Toplam grupta bu oran %3,3 'tür. Akciğer tutulumu çocuk grubun %72'sinde mevcuttur ama yetişkin grupta durum tam tersi çok az olup %5, 3'tür. Hepatomegalinin var olma durumu çocuk grupta %2,3, yetişkin grupta %22,1 ve Toplam grupta %11,4'tür. Splenomegal ise çocukta %1,5, yetişkinde %7,1 ve Toplam grupta %4,1'dir.Çocuk grubunda bilinç değişikliği %2,3, yetişkin grupta %3,5 ve toplam grupta ise bu oran %2,9'dir. Boğaz ağrısı şikayeti çocuk grubunda %4,5'tir, yetişkin grupta %15'tir, Toplam grupta ise bu oran % 9,4'tür. Başağrısı şikayeti ile gelenlerin en fazla yetişkin grupta olup oranı %81,4'tür, çocuk grupta %31,8, genelinde ise %54,7'dir. Bulantı ile gelenlerin %81,4' ü yine yetişkin gruptur, genelinde ise bu durum %61,2'dir. Çocuk grupta bu durum %43,9' dur. Kusma şikayeti ile gelenlerin %71,7' si yetişkin gruptur, çocuk grubunda %54,5, toplam grupta bu oran %62,4'tür. Öksürük ile gelen hastaların çoğu çocuk gruptur, bu oran %72,7' dir, yetişkin grupta ise bu oran azdır %27,3'tür. Genelinde ise bu oran %49,4' tür. Ateş belirleyici bir değişken olup toplam gruplarda en yüksektir. Çocuk grupta %96,2 , yetişkin grupta %92 ve toplam grupta ise bu oran %94,3'tür. Kırıklık şikayeti ile gelen hastaların %93,8 olup yetişkin gruptadır. Çocuk grupta ise %40,2 genelinde ise bu oran %64,9'dur. Kas ağrısı şikayeti yetişkin grupta daha fazladır %90,3' tür .Çocuk grupta ise bu oran %15,2'dir.Toplam grupta ise bu oran %50,2'dir. İshal, kanama, döküntü, leukopenia görülme oranı genelde azdır. Çocuk grubunda ishal görülme durumu %17,4 yetişkin grupta %34,5 genelinde ise bu durum %25,3'tür. Kanama durumu ise çocuk grubunda %17,4, yetişkin grupta %35,4 ve genel toplam grupta bu oran %25,7'dir.döküntü ve leukopenia durumu ise çocuk grupta aynı olup %12,1'dir. Yetişkin grupta ise döküntü görülme durumu %21,2, leukopenia görülme durumu ise %28,3'tür. Genelinde ise bu durum döküntü görülme durumu için %16,3, leukopenia için %19,6'dır. Kan ürünü ihtiyacı için çocuk grupta bu oran az olup %5,3' tür yetişkin grupta bu oran daha fazla olup %64,6'tır. Toplam grupta bu oran % 32,7' dir.

## 3.4. Sınıflama Ölçütlerine Göre Algoritmaların Karşılaştırılması

Gerçek veri setindeki bağımlı değişken olan hastanede yatış süresini kesim noktası 10 olarak alındığında, hastanede yatış süresi (-1, 9] ve (9, 30] olarak iki kategoriye ayrılmıştır. Destek vektör makinası, random forest ve regresyon ağaçlarının algoritmalarının sınıflama performansların karşılaştırmak için kullanılan doğruluk, kesinlik, duyarlılık ve F-ölçütü değerleri Tablo 3.6 da verilmiştir. Tablo 3.6'daki çözüm sonuçlarını elde etmek için kullanılan R program kodları ve karışıklık matrisleri EK: 4'de verilmiştir. DVM için örnek çözüm'de eşitlik 2.1, 2.3, 2.4 ve 2.5' deki formüller kullanılmıştır.

$$\text{Doğruluk}=\frac{87+94}{87+25+39+94}=0,74 \qquad \text{Kesinlik}=\frac{87}{87+25}=0,78$$

$$\text{Duyarlılık}=\frac{87}{87+39}=0,69 \qquad \text{F ölçütü}=\frac{2x(0,78x0,69)}{0,78+0,69}=0,73$$

**Tablo 3.6.** Hasta Gruplarına Göre Sınıflama Performanslarının Karşılaştırılması

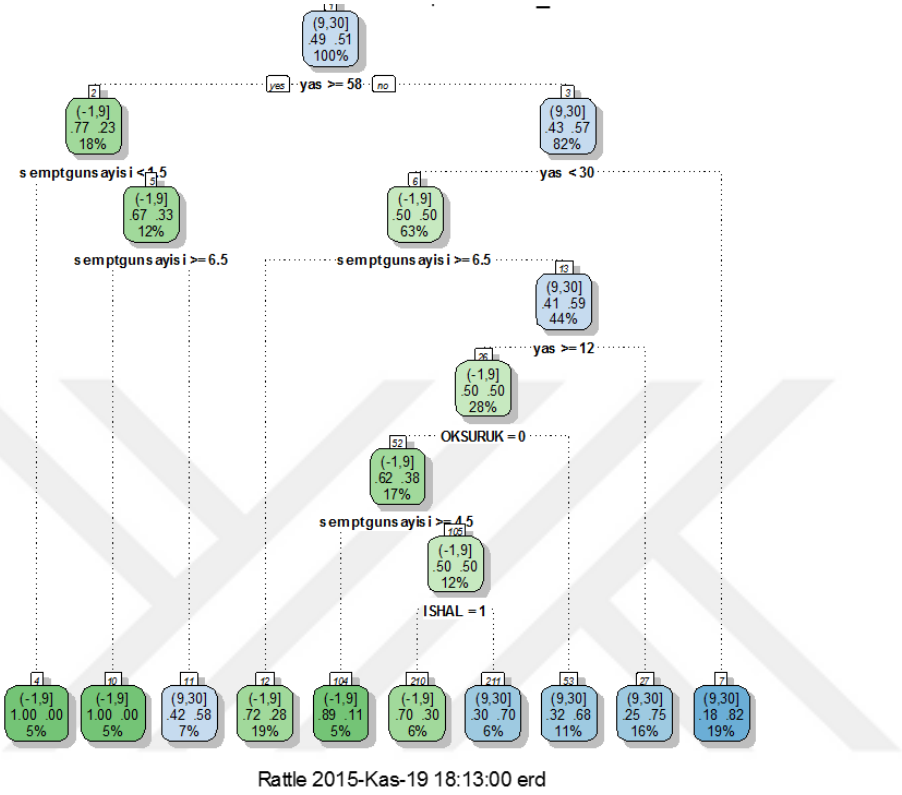
| Algoritma | Doğruluk % | Kesinlik% | Duyarlılık% | F-Ölçütü% |
|---|---|---|---|---|
| **DVM** | 74 | 78 | 69 | 73 |
| **RF** | 62 | 63 | 63 | 63 |
| **RA** | 76 | 72 | 70 | 71 |

Doğruluk kriterine göre en iyi sonucu 0,76 ile RA; 0,74 ile DVM; 0,62 ile RF'dir. Kesinlik ölçütüne göre 0,78 ile DVM; 0,72 ile RA; ve 0,63 RF' dir. Duyarlılık kriterine göre 0, 70 ile RA; 0,69 ile DVM; 0,63 ile RF 0,58 gelir. F ölçütüne göre 0,73 ile DVM; 0,71 ile RA ve 0,63 ile RF gelir.

## 3.5. Sınıflama Ağacı

Bağımlı değişkenin kategorik olmasına göre kullanılan sınıflama ağaçlarında, bağımlı değişken olan hastanede yatış süresi 10 günden az ve 10 günden fazla olarak iki gruba ayrılmıştır.Tüm değişkenler modele alındığında 10 tane sınıf oluşmuştur..

Regresyon ve sınıflama ağaçlarının R programında çözümü ve kodları EK: 4'de verilmiştir, sonuıçlar Şekil 3.1'deki gibidir.




**Şekil 3.1.** Yatış süresine göre sınıflama ağacı

Sınıflama ağaçları aşağıdaki gibi yorumlanır;

1. Yaşı ≥58 ve semptom gün sayısı 1,5< ise hastanede yatış süresi 10 günden azdır.
2. Yaş ≥58 ve 1,5 ≤ semptom gün sayısı < 6,5 ise hastanede yatış süresi 10 günden azdır.
3. Yaş ≥58 ve semptom gün sayısı ≥ 6,5 ise hastanede yatış süresi 10 günden fazladır.
4. Yaş< 30 ise semptom gün sayısı 6,5 ≥ ise hastanede yatış süresi 10 günden azdır.
5. 12≤ Yaş <30 ise 4,5≤ semptom gün sayısı < 6,5 arası öksürük yoksa hastanede yatış süresi 10 günden azdır.

6. 12≤ Yaş <30 ise semptom gün sayısı ≥ 4,5 , öksürük yoksa, ishal durumu varsa hastanede yatış süresi 10 günden azdır.
7. 12≤ Yaş <30 ise semptom gün sayısı≥ 4,5 , öksürük yoksa, ishal durumu yoksa hastanede yatış süresi 10 günden fazladır.
8. 12≤ Yaş <30 ise semptom gün sayısı≤6,5 öksürük yok ise hastanede yatış süresi 10 günden fazladır.
9. 12≤ Yaş <30 ise semptom gün sayısı≤6,5 hastanede yatış süresi 10 günden fazladır.
10. 30< Yaş≤ 58 ise hastanede kalma süresi 10 günden fazladır.

## 3.6. Regresyon Yöntemlerinin Karşılaştırılması

KKKA verisine ait bağımlı değişken hastanede yatış süresi ve tüm bağımsız değişkenlerin modele alınması ile oluşan regresyon yöntemlerinden destek vektör regresyon, random forest ve regresyon ağacının performanslarının karşılaştırılmasında HKO (Hata kareler ortalaması) en küçük, $R^2$ (açıklayıcılık yüzdesi) en büyük olan değer ile modelin performansıkarşılaştırılır..

### 3.6.1. Gerçek Veri Setine Göre Regresyon Yöntemlerinin Karşılaştırılması

KKKA gerçek hasta verilerine göre regresyon modellerinden; DVR, RF ve RA' larının hasta gruplarına göre HKO ve $R^2$'si, Tablo 3.7.'de verilmiştir. Çözüm için kullanılan R program kodları EK: 1'de verilmiştir.

**Tablo 3.7.** Hasta Gruplarına Göre Regresyon Modellerinin Karşılaştırılması

| **Hasta Grupları** | **Çocuk** | | **Yetişkin** | | **Toplam grup** | |
|---|---|---|---|---|---|---|
| **Yöntemler** | **HKO** | **$R^2$** | **HKO** | **$R^2$** | **HKO** | **R** |
| **DVR** | 2,00 | 0,84 | 0,30 | 0,98 | 1,35 | 0,89 |
| **RF** | 3,14 | 0,83 | 2,42 | 0,84 | 2,82 | 0,85 |
| **RA** | 6,88 | 0,26 | 5,90 | 0,29 | 6,23 | 0,32 |
| **n** | 132 | | 113 | | 245 | |

Tablo 3.7.' ye göre DVR için çocuk grupta $R^2$= 0,84, HKO= 2,00 iken yetişkin grupta $R^2$=0,98, HKO= 0,30'dur. Toplam hasta grubunda ise $R^2$ = 0,89, HKO=1,35'dir. RF için çocuk grupta $R^2$= 0,83, HKO= 3,14'tür. Yetişkin grupta

$R^2$ =0,84, HKO= 2,42, Toplam hasta grubunda $R^2$= 0,85, HKO= 2,82'dir.RA için çocuk grupta $R^2$ = 0,26 , HKO= 6,88 yetişkin grupta $R^2$= 0,29, HKO= 5,90 toplam hasta grubunda $R^2$= 0,32, HKO= 6,23 'tür.

Tüm gruplarda HKO'sı en az ve $R^2$'si en yüksek olan regresyon modeli DVR'dir. İkinci olarak en iyi model RF'dir. RA diğerlerine göre daha düşüktür.

#### 3.6.1.1. KKKA verisine göre değişkenlerin önem sırası

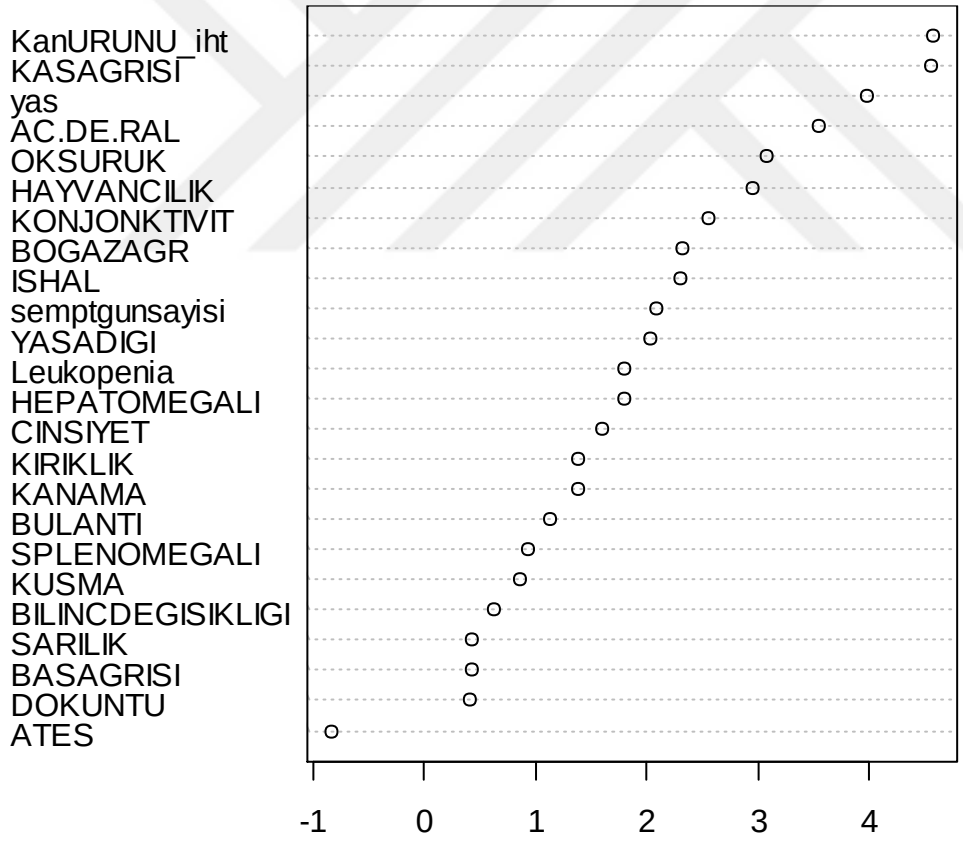
KKKA gerçek veri setine ait tüm değişkenler modelde önemli olmayabilir. Önemli olan değişkenlerin sıralamasını random forest regresyon yöntemine göre belirlenmiştir. Bu yöntemde, k adet değişkenle model kurulur, her seferinde bir değişken dışarıda bırakılarak oluşacak hata kareler ortalamasındaki artışa göre değişkenin modele olan katkısı bulunur. Buna göre toplam grup, yetişkin ve çocuk verisine ait değişkenlerin random forest yöntemine göre önem sırası Tablo 3.8., Tablo 3.9. ve Tablo 3.10.' da gösterilmiştir. Diğer yöntemlerde değişken seçim işlemi yapılamamaktadır.

**Tablo 3.8.** Toplam Grup Verisine Göre Tüm Veride Değişkenlerin Önemliliğe Göre Sıralaması

| Değişkenler | % Artan HKO |
|---|---|
| Kanurürünü ihtiyaç olma durumu | 4,5731 |
| Kas ağrısı | 4,5549 |
| Yaş | 3,9794 |
| Akcger tutulumu | 3,5421 |
| Öksürük | 3,0831 |
| Hayvancılık | 2,9521 |
| Konjonktıvıt | 2,5541 |
| Bogaz ağrısı | 2,3247 |
| Ishal | 2,2918 |
| Semptom gün sayısı | 2,0858 |
| Yasadığı yer | 2,0231 |
| Leukopenia | 1,8049 |
| Hepatomegalı | 1,7989 |

| | |
|---|---|
| Cinsiyet | 1,5921 |
| Kırıklık | 1,3793 |
| Kanama | 1,3763 |
| Bulantı | 1,1370 |
| Splenomegalı | 0,9324 |
| Kusma | 0,8605 |
| Bilinç değişikliği | 0,6211 |
| Sarılık | 0,4295 |
| Basagrısı | 0,4254 |
| Dokuntu | 0,4101 |
| Ates | -0,8258 |

Tablo 3.8.'e göre toplam gruptaki değişkenlerin önem sırasına göre sıraladığımızda kan ürünü ihtiyaç olma durumunun veri setinden çıkarılması modelde HKO'sı % 4,571'lık bir artış yaratırken, Ateş değişkeninin modelden çıkarılması HKO'sı % -0,82 lik bir artış sağlamıştır. Şekil 3.2.' de grafiksel gösterimi verilmiştir.



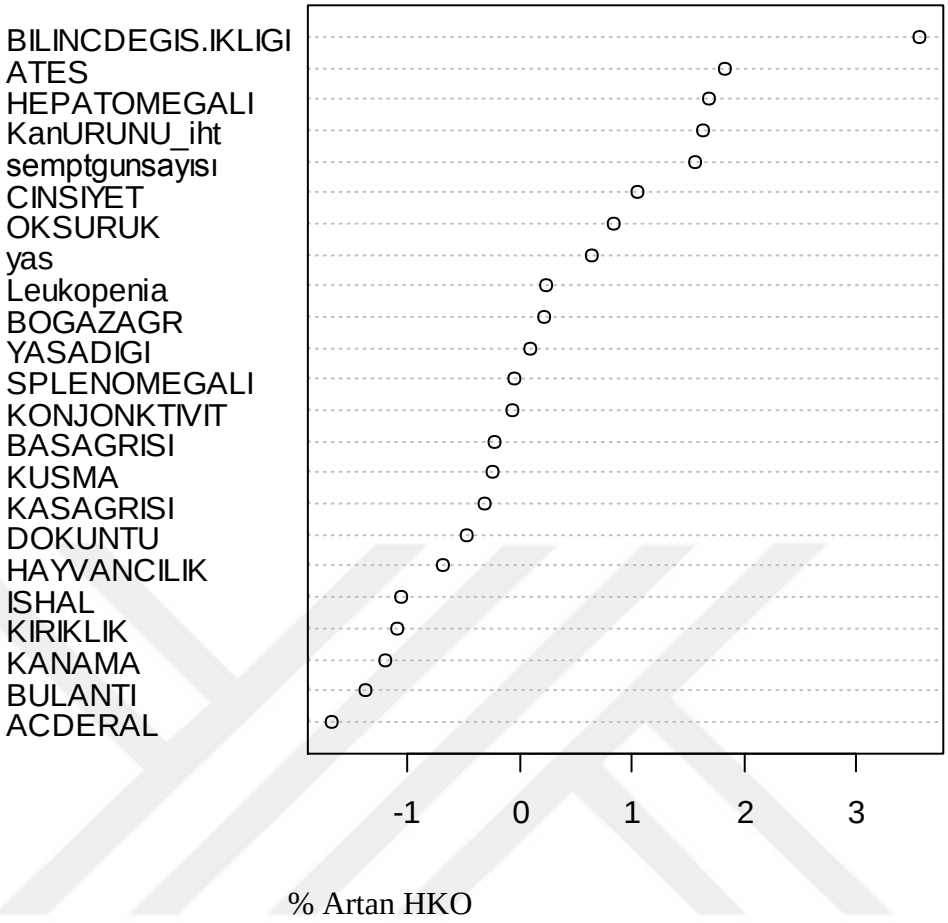

**Şekil 3.2.** Toplam gruptaki değişkenlerin modeldeki önem sırasının gösterimi

Şekil 3.2. ‘deki sonuçlar Tablo 3.8.’i doğrular nitelikte olup modeldeki en önemli değişkenler kan ürünü ihtiyaç olma durumu, kas ağrısı, yaş olarak sıralanırken en önemsiz değişken ates olup, olmama durumudur.

**Tablo 3.9.** Yetişkin Verisine Göre Tüm Veride Değişkenlerin Önemliliğe Göre Sıralaması

| Değişkenler | % Artan HKO |
|---|---|
| Bilinç değişikliği | 3,5736 |
| Ateş | 1,8307 |
| Hepatomegalı | 1,6869 |
| Kan ürünü ihtiyaç olma durumu | 1,6401 |
| Semptom gün sayısı | 1,5718 |
| Cinsiyet | 1,0450 |
| Öksürük | 0,8359 |
| Yaş | 0,6505 |
| Leukopenia | 0,2386 |
| Boğaz ağrısı | 0,2091 |
| Yaşadığı yer | 0,0956 |
| Splenomegalı | -0,0538 |
| Konjonktivit | -0,0740 |
| Baş ağrısı | -0,2249 |
| Kusma | -0,2507 |
| Kas ağrısı | -0,3195 |
| Döküntü | -0,4665 |
| Hayvancılık | -0,6926 |
| ishal | -1,0634 |
| Kırıklık | -1,0950 |
| Kanama | -1,1988 |
| Bulantı | -1,3770 |
| Akciğer tutulumu | -1,6757 |

Tablo 3.9.’a göre yetişkin verisindeki değişkenlerin önem sırasına göre sıraladığımızda bilinç değişikliği olma durumunun veri setinden çıkarılması modelde HKO’ sı % 3,5736’lık bir artış yaratırken, Akciğer tutulumu değişkeninin modelden çıkarılması HKO’ sı % -1,6757’ lik bir artış sağlamıştır. Şekil 3.2’ de grafiksel olarak gösterilmiştir.




**Şekil 3.3.** Yetişkin verisine göre değişkenlerin modeldeki önem sırasının gösterimi

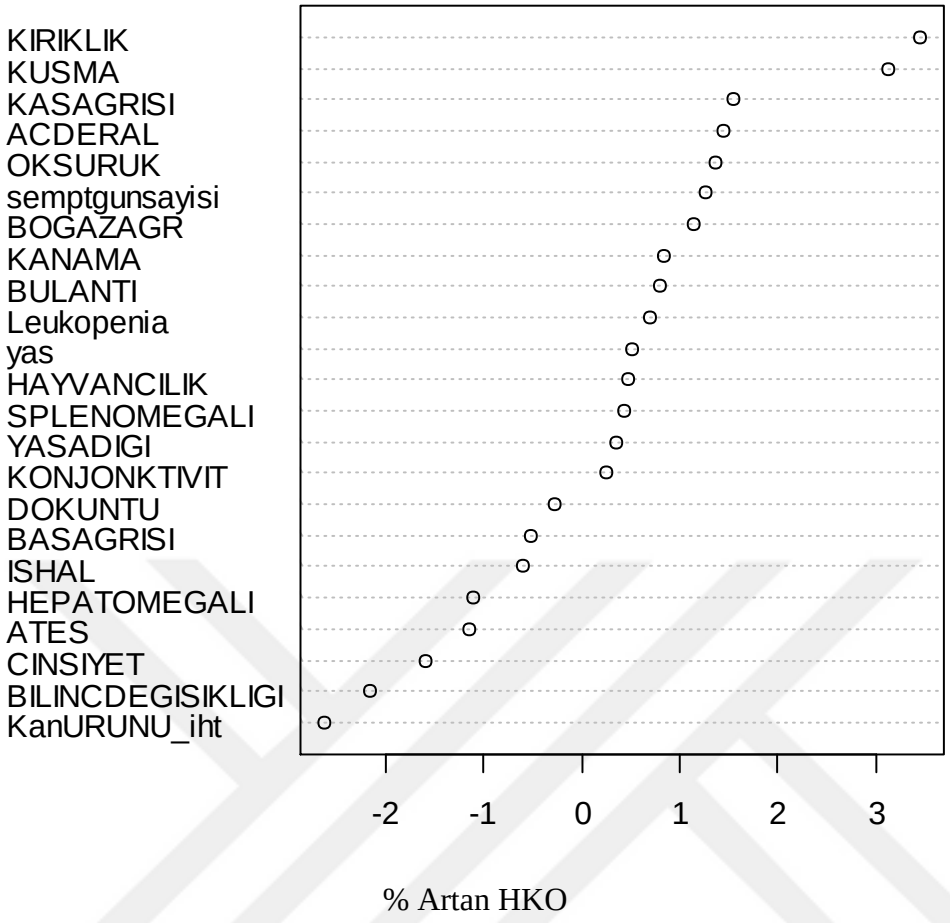
Şekil 3.3' deki sonuçlar Tablo 3.10'u doğrular şekilde olup modeldeki en önemli değişkenler bilinç değişikliği durumu, ateş, hepatomegali olarak sıralanırken en önemsiz değişken akciğer tutulumu ve bulantı olma durumudur.

**Tablo 3.10.** Çocuk Verisine Göre Değişkenlerin Önemliliğe Göre Sıralaması

| **Değişkenler** | **%IncMSE** |
|---|---|
| Kırıklık | 3,4580 |
| Kusma | 3,1374 |
| Kas ağrısı | 1,5375 |
| Akciğer tutulumu | 1,4400 |
| Öksürük | 1,3565 |
| Semptom gün sayısı | 1,2573 |
| Boğaz ağrısı | 1,1471 |
| Kanama | 0,8271 |
| Bulantı | 0,7945 |
| Leukopenia | 0,6926 |

| | |
|---|---|
| Yaş | 0,5143 |
| Hayvancılık | 0,4708 |
| Splenomegali | 0,4314 |
| Yasadığı yer | 0,3543 |
| Konjonktivit | 0,2457 |
| Döküntü | -0,2711 |
| Baş ağrısı | -0,5241 |
| İshal | -0,6129 |
| Hepatomegali | -1,1124 |
| Ateş | -1,1529 |
| Cinsiyet | -1,6022 |
| Bılınç değişikliği | -2,1710 |
| Kan ürünü ihtiyaç olma durumu | -2,6240 |

Tablo 3.10.'a göre çocuk verisindeki değişkenlerin önem sırasına göre sıraladığımızda kırıklık olma durumunun veri setinden çıkarılması modeled HKO'sı % 3,4580'lık bir artış yaratırken, kan ürününe ihtiyaç olma durumu değişkeninin modelden çıkarılması HKO'sı % -2,6240' lik bir artış sağlamıştır.




**Şekil 3.4.** Çocuk verisine göre değişkenlerin modeldeki önem sırasının gösterimi

Şekil 3.4.'de sonuçlar Tablo 3.10'u doğrular şekilde olup modeldeki en önemli değişkenler kırıklık olma durumu, kusma, kas ağrısı ve akciğer tutulumu olarak sıralanırken en önemsiz değişken bilinç değişikliği ve kan ürünü ihtiyaç olma durumudur

### 3.6.2. Gerçek Veri Setine Ait Simülasyon Sonuçları İçin Regresyon Modellerinin Karşılaştırılması

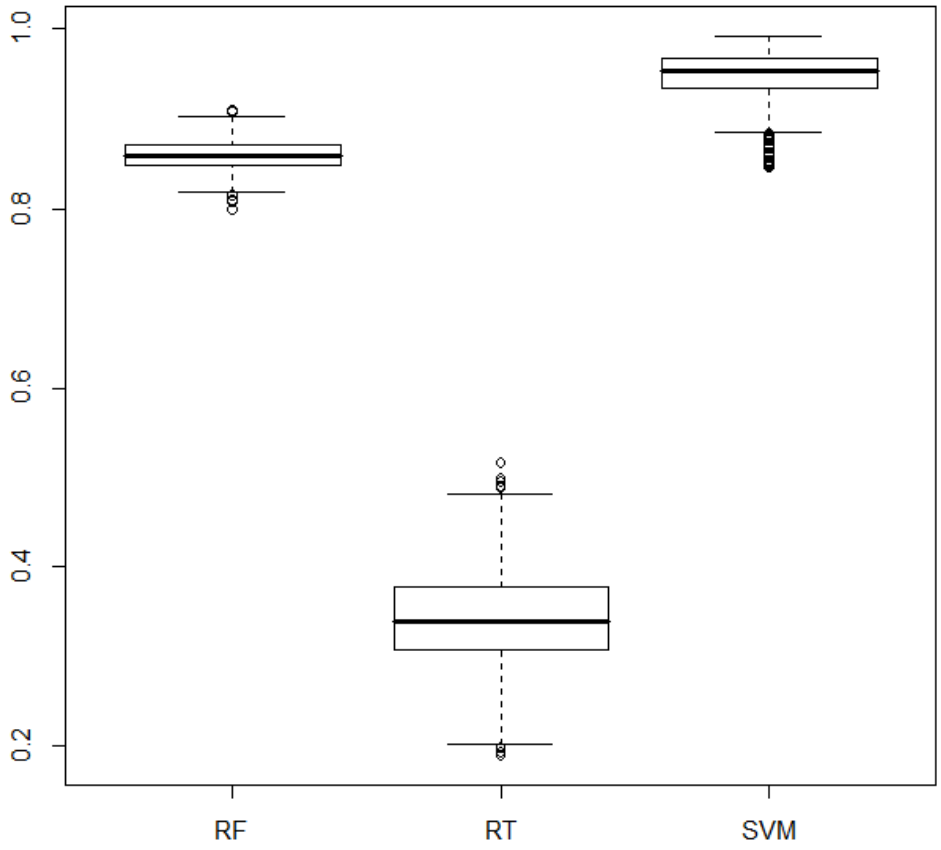
Simülasyon çalışmasına göre gerçek veri setinden elde edilen, gerçek veri sayısı kadar 1000 tekrarlı simülasyon sonuçlarına göre toplam hasta, çocuk ve yetişkin grup için regresyon modellerinin performanslarının karşılaştırılması Tablo 3.11'de gösterilmiştir. Sonuçların çözümüne ait R program kodları EK: 2'de verilmiştir.

**Tablo 3.11.** Hasta Gruplarının 1000 Tekrarlı Simülasyon Sonuçlarına Göre Regresyon Modellerinin Karşılaştırılması

| Yöntemler | Çocuk | | Yetişkin | | Toplam grup | |
|---|---|---|---|---|---|---|
| | HKO | $R^2$ | HKO | $R^2$ | HKO | $R^2$ |
| DVR | 0,069 | 0,95 | 0,035 | 0,98 | 0,054 | 0,96 |
| RF | 0,29 | 0,86 | 0,255 | 0,88 | 0,292 | 0,87 |
| RA | 0,649 | 0,34 | 0,621 | 0,37 | 0,670 | 0,32 |
| n | 132 | | 113 | | 245 | |

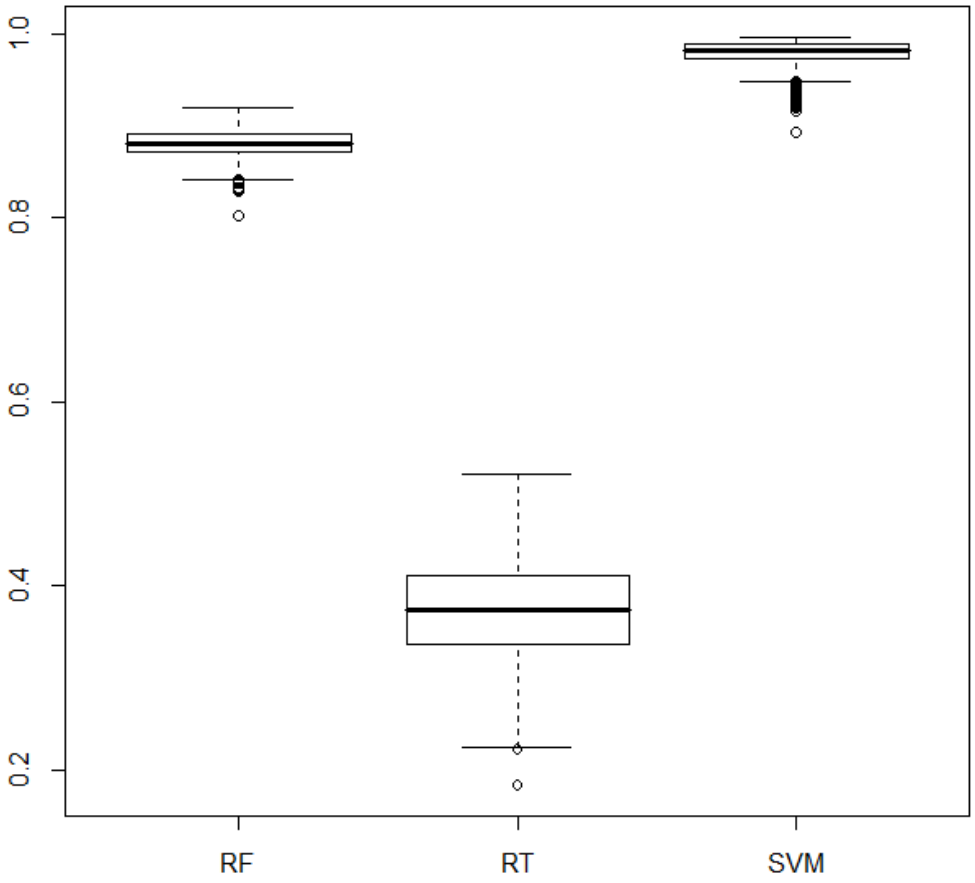
Şekil 3.5., Şekil 3.6. ve Şekil 3.7. görsellerde verilmiş ve Tablo 3.11.' ü doğrular niteliktedir. DVR için çocuk grupta $R^2$ =0,95, HKO =0,069, yetişkin grupta $R^2$= 0,98 HKO=0,035'dur. Toplam grupta ise $R^2$ =0,96, HKO= 0,054 bulunmuştur. RF için çocuk grubunda $R^2$=0,86 HKO=0,29, yetişkin grubunda $R^2$ =0, 88, HKO= 0,255'dir. Toplam grupta ise $R^2$ =0,87, HKO=0,292 bulunmuştur. RA için çocuk grupta $R^2$ =0,34, HKO=0,649, yetişkin grupta $R^2$ =0,37, HKO= 0,621'dir. Toplam grupta ise $R^2$ =0,32, HKO= 0,670 bulunmuştur.

Tüm gruplarda; açıklayıcılık yüzdesi en fazla, hata kareler ortalaması en düşük olan regresyon yöntemi DVR'dir bunu RF regresyon yöntemi izler. RA regresyon yöntemi diğerlerine açıklayıcılık yüzdesi en az ve hata kareler ortalaması en yüksek olandır. Tablo 3.11.'deki $R^2$ değerlerine ait box plot grafiği Şekil 3.5.'de verilmiştir.



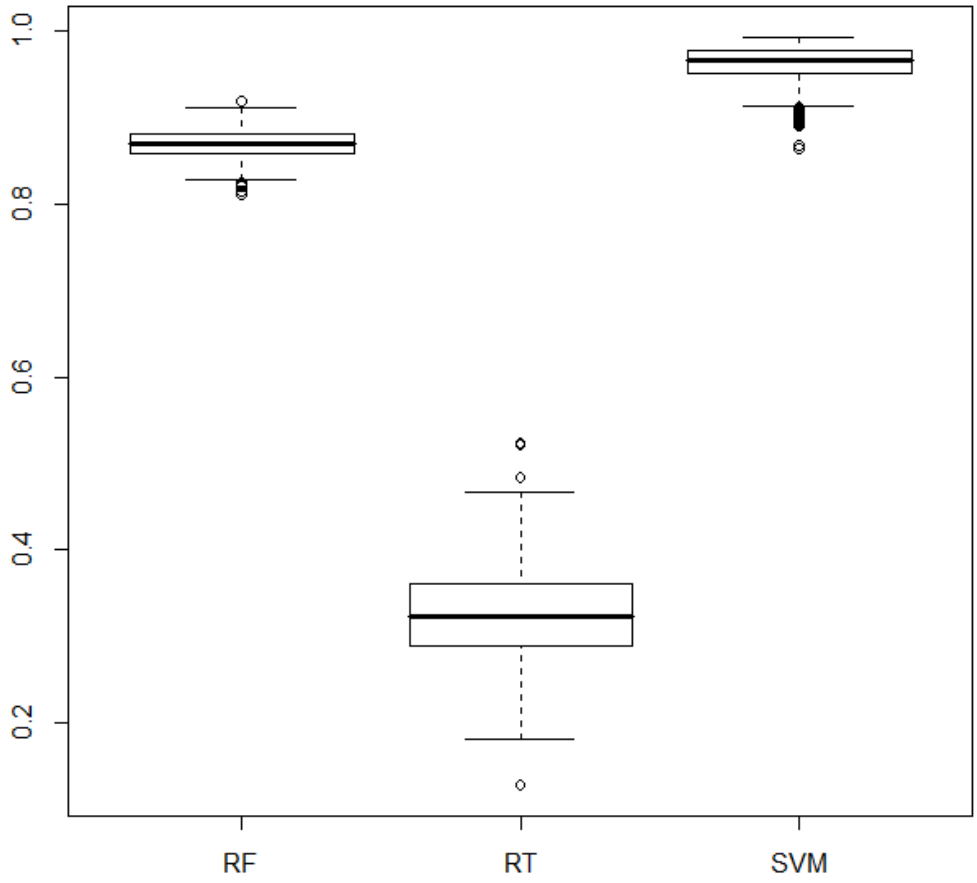

**Şekil 3.5.** Çocuk Grubuna İçin 1000 Kez Tekrarlı Simülasyona Göre Regresyon Yöntemlerinin Durumu

Şekil 3.5.' de ki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA(RT) olup Tablo 3.11.'i doğrular niteliktedir.




**Şekil 3.6.** Yetişkin Grubunun 1000 Kez Tekrarlı Simülasyon Sonuçlarına Göre Regresyon Yöntemlerinin Durumu

Şekil 3.6.' da ki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA(RT) olup Tablo 3.11.'i doğrular niteliktedir.




**Şekil 3.7.** Toplam Hastaya Ait 1000 Kez Tekrarlı Simülasyon Sonuçlarına Göre Regresyon Modellerinin Performansları

Şekil 3.7.' deki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA(RT) olup Tablo 3.11.'i doğrular niteliktedir.

### 3.6.3. Simülasyon Çalışması İle Farklı Senaryolara Ait Verilerin Regresyon Modellerinin Karşılaştırılması

Bağımlı ve bağımsız değişkenlerin korelasyon matrsi için aralarında düşük (r=0,00-0,20), orta (r=0,21-0,40) ve yüksek (r=0,41-0,60) korelasyon ilişkisine göre uniform dağılımdan, simülasyon çalışması ile gözlem sayıları 100, 250 ve 1000 tane

veri elde edilmiştir. Bu verilere göre destek vektör regresyon, random forest ve regresyon ağaçlarının performanslarının karşılaştırılması sonuçları sırasıyla Tablo 3.12. , Tablo 3.13. ve Tablo 3.14.'de verilmiştir. Simülasyon çalışmasıyla farklı senaryolara ait regresyon modellerinin performanslarının karşılaştırılmasında kullanılan, R programına ait kodlar EK: 3'tedir. Ayrıca simülasyon çalışmasında kullanılan düşük, orta ve yüksek yapılı korelasyon matrisleri sırasıyla EK:6, EK: 7 ve EK:8 verilmiştir.

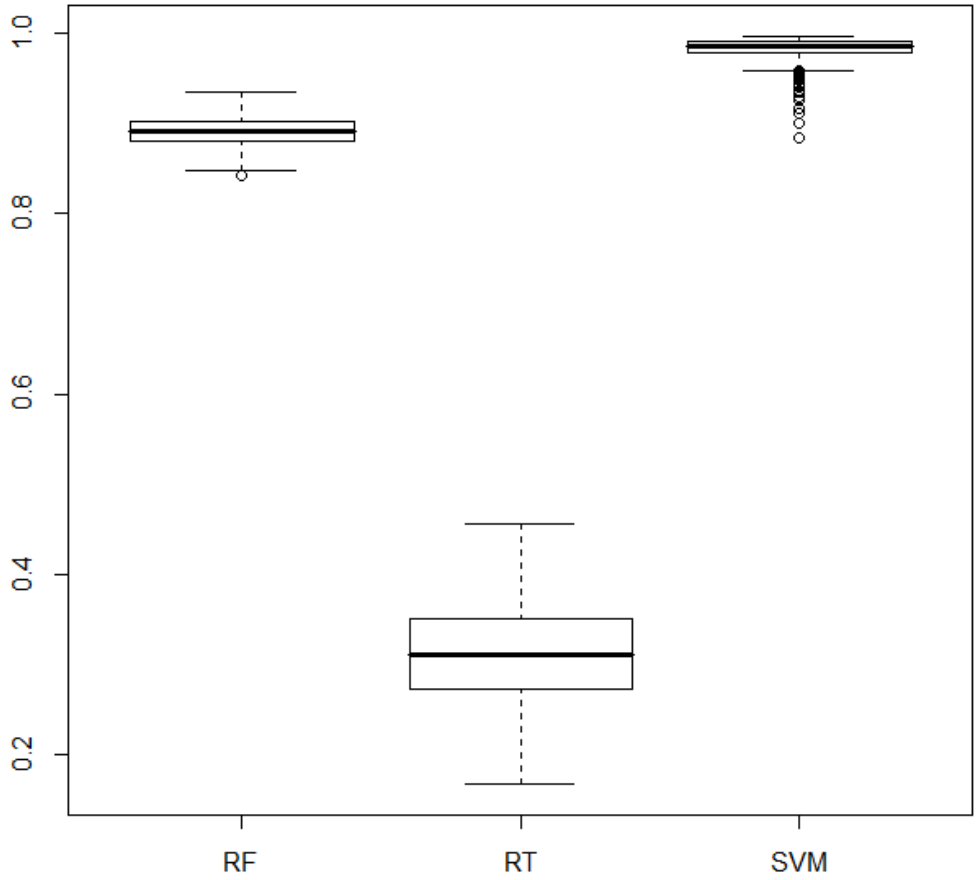
**Tablo 3.12.** Değişkenler Arası Korelasyon Yapısına Göre, n=100 İçin Regresyon Modellerinin Sonuçları

| Yöntemler | Düşük (r=0,00-0,20) | | Orta (r=0,21-0,40) | | Yüksek (r=0,41-0,60) | |
|---|---|---|---|---|---|---|
| | HKO | $R^2$ | HKO | $R^2$ | HKO | $R^2$ |
| DVR | 0,032 | 0,98 | 0,026 | 0,98 | 0,026 | 0,98 |
| RF | 0,283 | 0,89 | 0,227 | 0,86 | 0,180 | 0,86 |
| RA | 0,682 | 0,31 | 0,550 | 0,44 | 0,437 | 0,56 |

HKO ve $R^2$ değerleri için 1000 tekrardaki degişkenler arası ilişki yapısının yöntemler üzerine etkisi Tablo 3.12.'de gösterilmiştir, sonuçlar Şekil 3.8., Şekil 3.9. ve Şekil 3.10.'da görsel hale getirilmiştir.

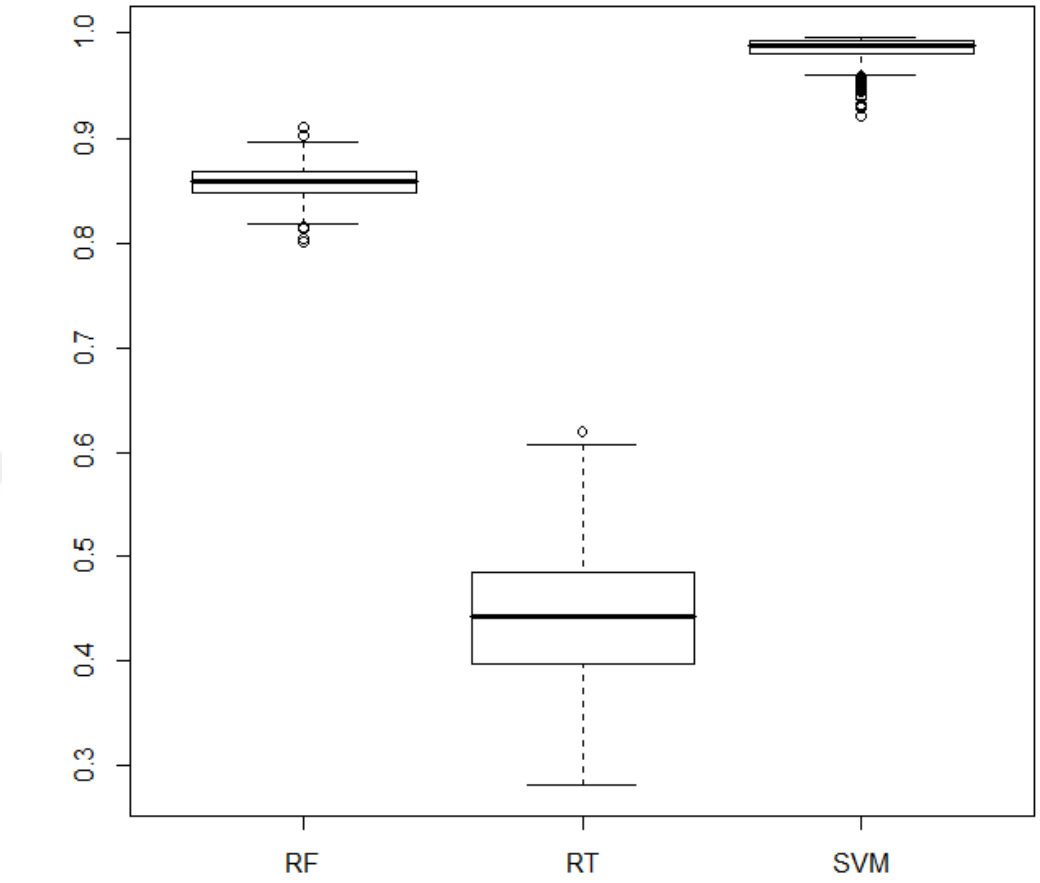
Düşük (r=0,00-0,20) düzey korelasyon yapısına göre DVR için $R^2$ =0,98, HKO= 0,032; RF için $R^2$ =0,89, HKO= 0,283; RA için $R^2$ =0,31, HKO= 0,682'dir. Orta (r=0,21-0,40) düzey korelasyon yapısına göre DVR için $R^2$ =0,98, HKO =0,026; RF için $R^2$ =0,86, HKO= 0,22; RA için $R^2$ =0,44, HKO= 0,550'dir. Yüksek (r=0,41-0,60) düzey korelasyon yapısına göre DVR için $R^2$=0,98, HKO=0,026; RF için $R^2$ =0,86, HKO= 0,180; RA için $R^2$= 0,56 HKO= 0,437'dir.

Buna göre her üç korelasyon yapısı için $R^2$ si en yuksek, HKO'sı en düşük olan regresyon yöntemi DVR'dir. İkinci en iyi yöntem RF'dir. RA'nın sonuçları diğerlerine göre daha düşüktür.



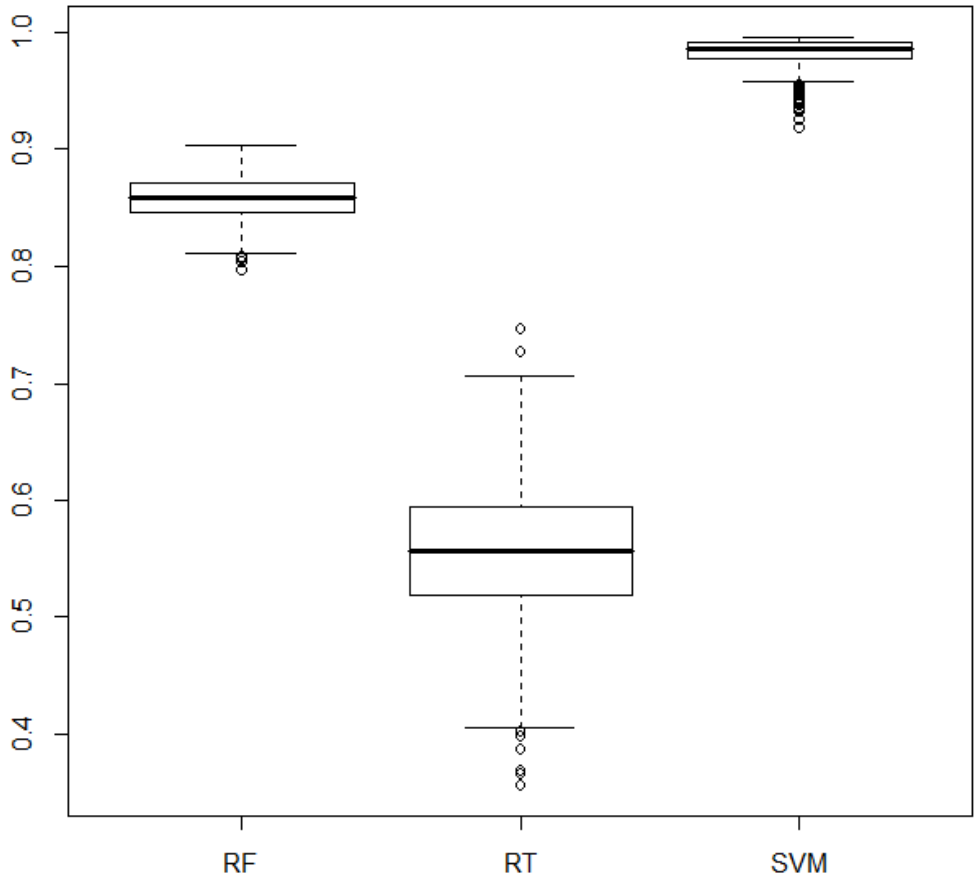

**Şekil 3.8.** (n=100) Düşük Düzey Korelasyon Yapısına Göre Regresyon Yöntemlerinin Performanları

Şekil 3.8.'de Tablo 3.12'yi doğrular nitelikte olup en yüksek performansı DVR (SVM) göstermiş olup arkasıdan onu RF takip etmiştir. En düşük performansı RA (RT) göstermiştir.




**Şekil 3.9.** (n=100) Orta Düzey Korelasyon Yapısına Göre Regresyon Yöntemlerinin Performanları

Şekil 3.9.'da Tablo 3.12.'yi doğrular nitelikte olup en yüksek performansı DVR göstermiş olup arkasıdan onu RF takip etmiştir. En düşük performansı RA (RT) göstermiştir.




**Şekil 3.10.** (n=100) Yüksek Düzey Korelasyon Yapısına Göre Regresyon Yöntemlerinin Performansları

Şekil 3.10.'de Tablo 3.12.'yi doğrular nitelikte olup en yüksek performansı DVR göstermiş olup arkasıdan onu RF takip etmiştir. En düşük performansı RA göstermiştir.

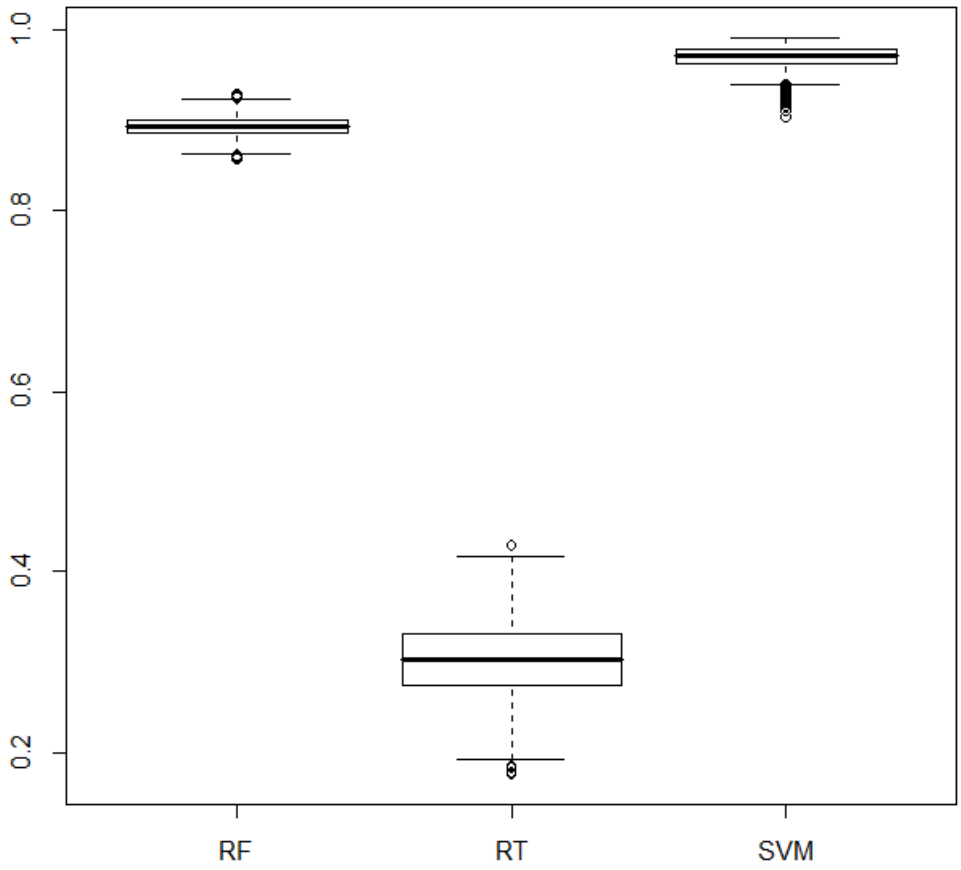
**Tablo 3.13.** Değişkenler Arası Korelasyon Yapısına Göre, n =250 İçin Regresyon Modellerinin Sonuçları

| yöntemler | Düşük (r=0,00-0,20) | | Orta (r=0,21-0,40) | | Yüksek (r=0,41-0,60) | |
|---|---|---|---|---|---|---|
| | HKO | $R^2$ | HKO | $R^2$ | HKO | $R^2$ |
| DVR | 0,047 | 0,97 | 0,041 | 0,97 | 0,044 | 0,96 |
| RF | 0,283 | 0,89 | 0,226 | 0,85 | 0,181 | 0,85 |
| RA | 0,696 | 0,30 | 0,570 | 0,43 | 0,456 | 0,54 |

Tablo 3.13.'ün sonuçları ile Şekil 3.11., Şekil 3.12. ve Şekil 3.13.' deki herbir korelasyon yapısına göre görseller birbirini destekler şekildedir.

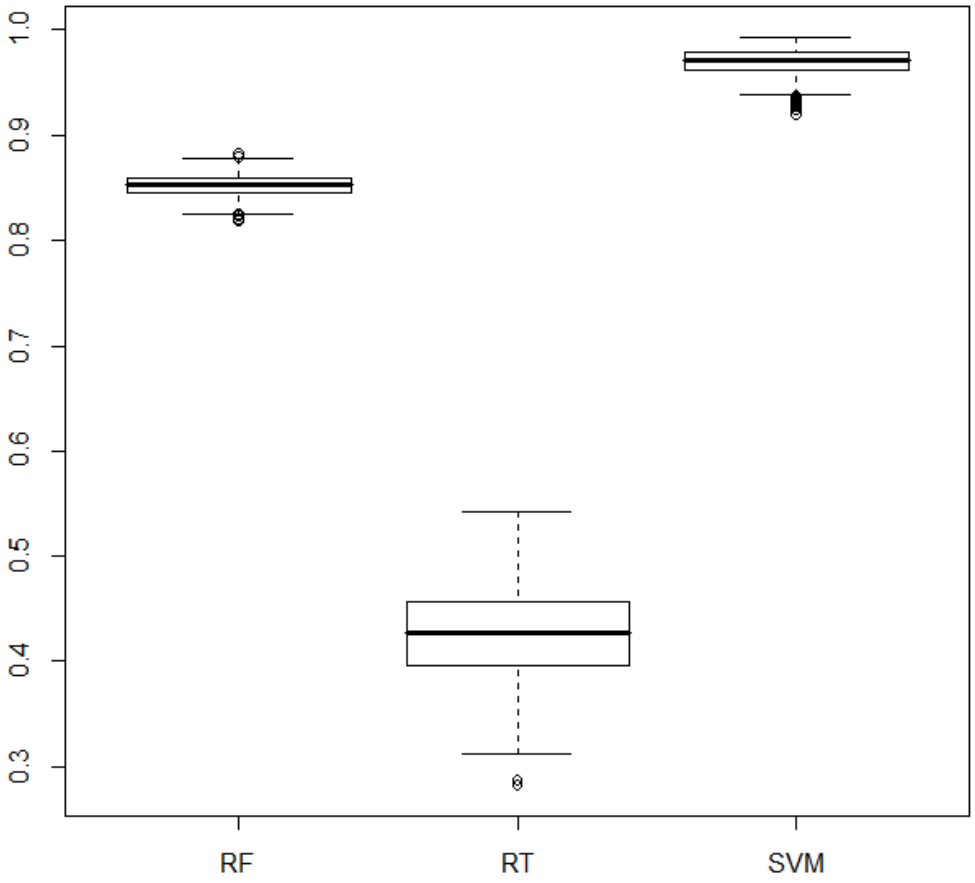
Düşük (r=0,00-0,20) düzey korelasyon yapısına göre DVR için $R^2$ =0,97, HKO= 0,047; RF için $R^2$= 0,89, HKO= 0,283; RA için $R^2$ = 0,30, HKO= 0,696'dir. Orta (r=0,21-0,40) düzey korelasyon yapısına göre DVR için $R^2$ =0,97, HKO =0,041; RF için $R^2$ 0,85, HKO =0,226; RA için $R^2$ = 0,43 HKO= 0,570'dir. Yüksek (r=0,41-0,60) düzey korelasyon yapısına göre DVR için $R^2$=0,96, HKO=0,044; RF için $R^2$ 0,85, HKO=0,181; RA için $R^2$ =0,54, HKO= 0,456'dir.

İlişki yapılarına göre her üç yöntemden en iyi performansı DVR verir. Arkasından RF gelir, RA dğerlerine göre daha düşük performans göstermiştir.



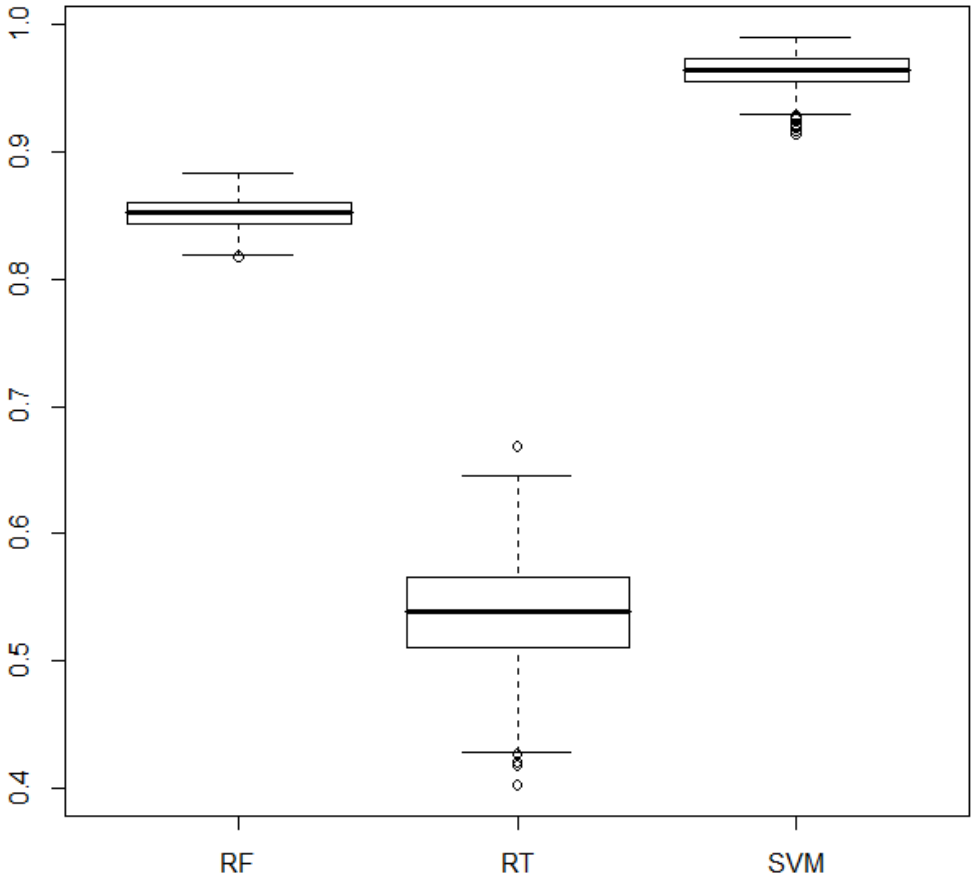

**Şekil 3.11.** (n=250) Düşük Düzey Korelasyon Yapısına Göre Regresyon Yöntemlerinin Performansları

Şekil 3.11.' deki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA (RT) olup Tablo 3.13.'ü doğrular niteliktedir.




**Şekil 3.12.** (n=250) Orta düzey korelasyon yapısına göre Regresyon yöntemlerinin performansları

Şekil 3.12.' deki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA (RT) olup Tablo 3.13.'ü doğrular niteliktedir.




**Şekil 3.13.** (n=250) Yüksek Düzey Korelasyon Yapısına Göre Regresyon Yöntemlerinin Performansları

Şekil 3.13.' deki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA (RT) olup Tablo 3.13.'ü doğrular niteliktedir.

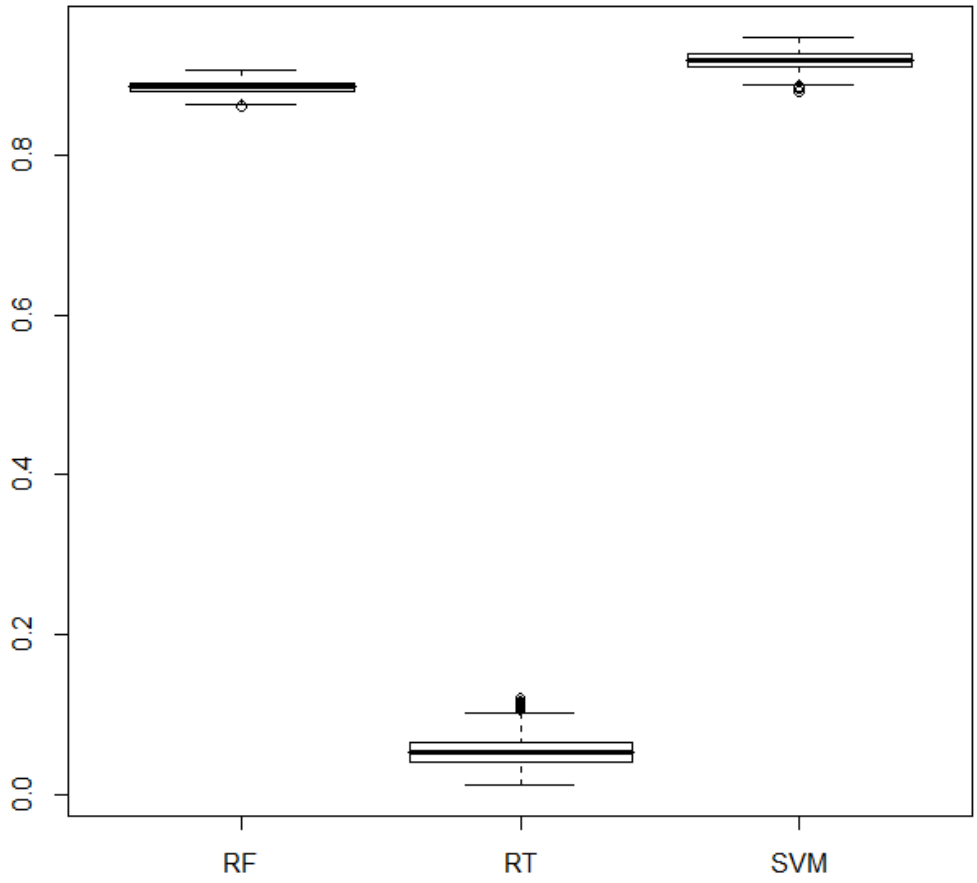
**Tablo 3.14.** Değişkenler Arası Korelasyon Yapısına Göre, n =1000 İçin Regresyon Modellerinin Sonuçları

| Yöntemler | Düşük (r=0,00-0,20) | | Orta (r=0,21-0,40) | | Yüksek (r=0,41-0,60) | |
|---|---|---|---|---|---|---|
| | HKO | $R^2$ | HKO | $R^2$ | HKO | $R^2$ |
| DVR | 0,103 | 0,92 | 0,095 | 0,92 | 0,103 | 0,90 |
| RF | 0,295 | 0,89 | 0,234 | 0,84 | 0,197 | 0,83 |
| RA | 0,944 | 0,13 | 0,740 | 0,26 | 0,563 | 0,43 |

Tablo 3.14.'ün sonuçları ile Şekil 3.14., Şekil 3.15. ve Şekil 3.16.' daki herbir korelasyon yapısına göre grafikler birbirini destekler şekildedir.

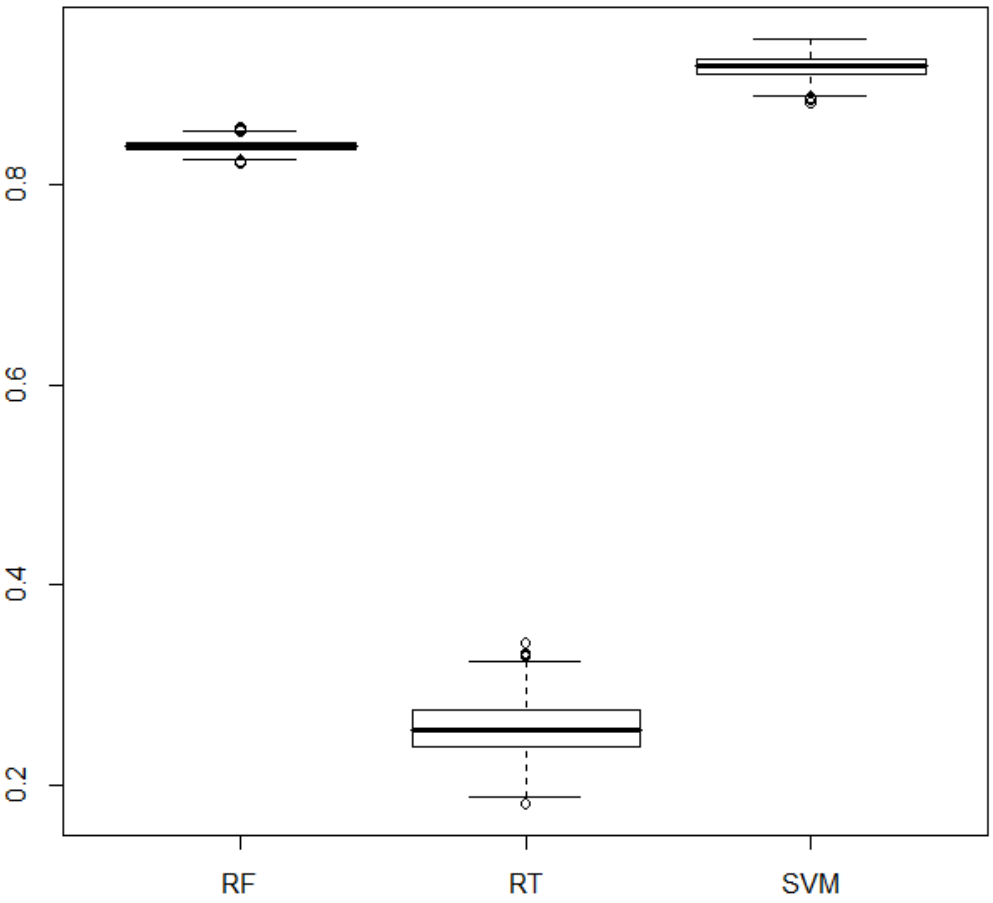
Düşük (r=0,00-0,20) düzey korelasyon yapısına göre DVR için $R^2$ =0,92, HKO= 0,103; RF için $R^2$=0,89, HKO= 0,295; RA için $R^2$ =0,13 HKO= 0,944'dir. Orta (r=0,21-0,40) düzey korelasyon yapısına göre DVR için $R^2$ =0,92, HKO= 0,095; RF için $R^2$ =0,84, HKO= 0,234; RA için $R^2$ = 0,26 HKO= 0,740'dir. Yüksek (r=0,41-0,60) düzey korelasyon yapısına göre DVR için $R^2$ =0,90, HKO= 0,103; RF için $R^2$=0,83, HKO= 0,197; RA için $R^2$ = 0,43 HKO= 0,563'dir.

İlişki yapılarına göre her üç yöntemden en iyi performansı DVR verir. Arkasından RF gelir, RA dğerlerine göre daha düşük performans göstermiştir.



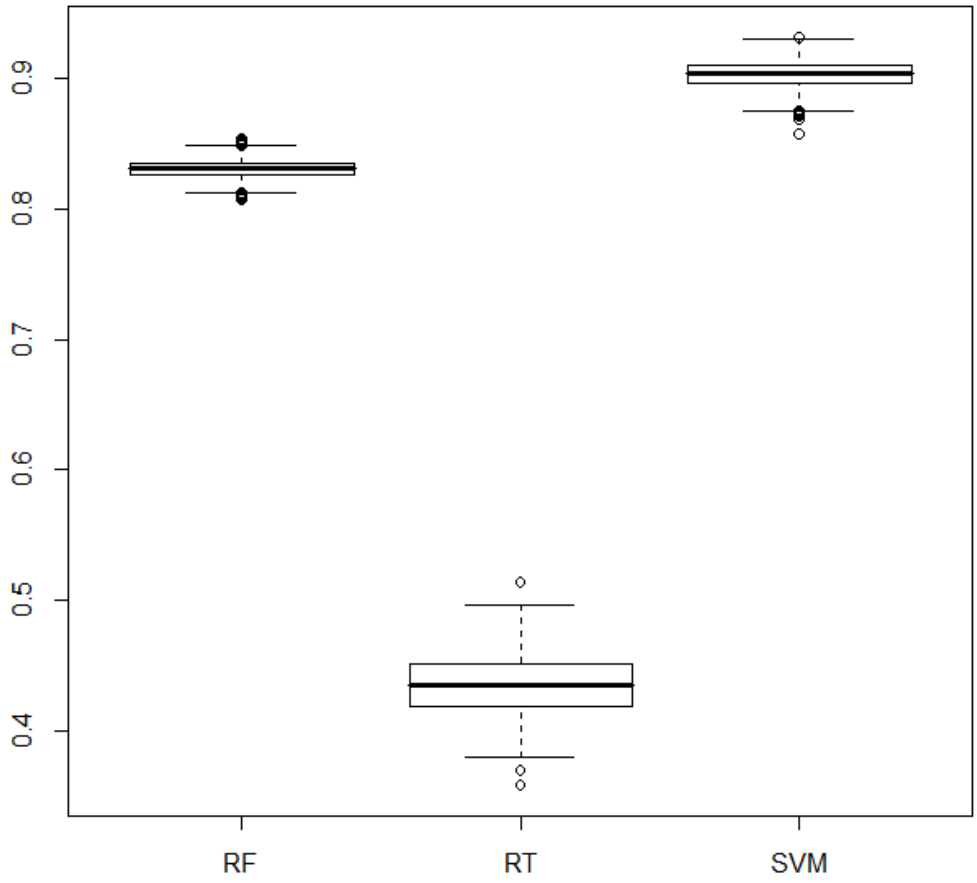

**Şekil 3.14** (n=1000) Düşük Düzey Korelasyon Yapısına Göre Regresyon Yöntemlerinin Performansları

Şekil 3.14.' deki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA (RT) olup Tablo 3.14.'ü doğrular niteliktedir.




**Şekil 3.15** (n=1000) Orta Düzey Korelasyon Yapısına Göre Regresyon Yöntemlerinin Performansları

Şekil 3.15.' deki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA (RT) olup Tablo 3.14.'i doğrular niteliktedir.




**Şekil 3.16** (n=1000) Yüksek Düzey Korelasyon Yapısına Göre Regresyon Yöntemlerinin Performansları

Şekil 3.16.' daki grafik incelendiğinde DVR (SVR) yönteminin performansı diğerlerinden daha iyidir. Arkasından RF takip etmiştir. En düşük performans RA (RT) olup Tablo 3.14.'i doğrular niteliktedir.

# SONUÇLAR ve ÖNERİLER

Bu çalışmada 2009-2010-2011 yıllarında Cumhuriyet Üniversitesi Tıp Fakültesi İntaniye ve Çocuk Sağlığı ve Hastalıkları servislerinde KKKA tanısıyla yatan 245 hastaya ait 25 değişken ile 6125 veri girişi yapılmıştır. Veri madenciliğinde kullanılan kestirim yöntemlerinin performanslarının karşılaştırılması amaçlandırılmıştır. Bu çalışma destek vektör regresyon, random forest ve regresyon ağacı yöntemlerinin performanslarını karşılaştırması bakımından yapılan ilk çalışma olması açısından önem taşımaktadır. Ayrıca bu çalışmada veri madenciliği yöntemleri kullanılmıştır, veri madenciliği taşıdığı özellikler ile diğer yöntemlerden avantajlar sağlamakta ve önem teşkil etmektedir. Bu özellikler (Koyuncugil ve Özgülbaş 2009: 26)

- Veri madenciliği veride var olan örüntüleri keşfetme sürecidir ve bu süreç otomatik veya yarı otomatiktir. Keşfedilen örüntüler anlamlı olmalıdır ve genellikle ekonomik avantaj olmak üzere fayda sağlamalıdır.
- Geleneksel istatistiksel yöntemlerin aksine çok büyük veri içerisinden yararlı bilginin çıkartılmasını sağladığı için yoğun veriden bilgi çıkartırken kullanılacak yegâne çözüm veri madenciliğidir.
- İstatistikçiler genellikle bir hipotez ile başlarlar, Veri madenciliği hipoteze gerek duymaz.
- İstastiksel analizler hipotezlerini eşleştirmek için kendi eşitliklerini geliştirmek zorundayken, veri madenciliği algoritmaları eşitlikleri otomatik olarak geliştirir.
- İstatistiksel analizler sadece sayısal verileri kullanırken, veri madenciliği metin, ses gibi farklı tiplerde veri kullanır.

Bu araştırma ile ilgili ortaya çıkan sonuçlar şu şekilde sıralanmıştır.

- Çalışmamızda genel olarak KKKA tanısı alan bireylerden erkek hasta sayısı daha fazladır. Ancak yetişkin grupta kadın hasta sayısı az da olsa daha fazladır.
- Bireylerin hastanede kalış sürelerine göre çocuk grunbunun hastanede kalma süresi daha uzundur. En az ise yetişkin gruptur. Semptomlu geçirilen gün sayısı incelendiğinde ortalama olarak yetişkin grup diğer gruplardan daha fazladır.
- Kırım kongo kanamalı ateş tanısı alan çocukların çoğu Sivas ili ve çevresi iken yetişkinlerde tanı alan hastaların büyük çoğunluğu diğer illerden gelmektedirler toplamda hastaların büyük çoğunluğu Sivas ili ve çevresindendir.
- Çocuk grubun yaş aralığı 0-17 yaş arası olduğundan hayvancılıkla ilgilenme durumu yetişkin ve Toplam gruba göre daha düşüktür.
- KKKA hastaların hastaneye geldiklerindeki bulgular incelendiğinde çocuklarda görülen bulgular en çok; konjoktivit, akciğer tutulumu, öksürük ve ateştir. Yetişkinlerde ise en çok görülen bulgular; sarılık, hepatomegali, splenomegali, bilinç değişikliği, kas ağrısı, boğaz ağrısı, bulantı, kusma, ishal, kırıklık, kanama, döküntü, leukopenia ve kan ürünü ihtiyacıdır. Ateş ise her iki grupta yüksektir.
- Gerçek veri setindeki bağımlı değişken olan hastanede yatış süresini kesim noktası 10 aldığımıızda (-1, 9] ve (9,30] olarak iki kategoriye ayrılmıştır. Sınıflama performanslarının karşılaştırılmasında algoritmalar incelendiğinde; model testine ait doğruluk ölçütü ile en iyi sonucu RA algoritması verir. DVM en iyi performans bakımından RA'ya yakındır. Arkasından RF gelir. Kesinlik ölçütüne göre en iyi algoritma DVM, arkasından RA ve RF gelir. Duyarlılık kriterine göre en iyi algoritma RA ile DVM'dir. Arkasından RF gelir. F ölçütüne göre en iyi sonuç üreten algoritma DVM'dir. Arkasından RA ve RF gelir.
- Toplam grup üzerinden tüm değişkenler random forest regresyon yöntemine göre, değişkenlerin önem sırası oluşturulmuştur buna göre en önemli değişkenler kan ürünü ihtiyaç olma durumu, kas ağrısı, yaş, akciğer

tutulumudur. Daha az önemli olanları splenogami, kusma, bilinç değişikliği, sarılık, başağrısı, döküntü ve ateştir.

- Yetişkin grup için random forest yöntemine ile tüm değişkenlerin önem sırası oluşturulmuştur. Buna göre en önemli değişken bilinç değişikliğidir. En az önemli olan değişkenleri ishal, kırıklık, kanama, bulantı ve akciğer tutulumu şeklinde sıralayabiliriz.
- Çocuk grubuna göre random forest yöntemi ile belirlenen tüm değişkenlerin sıralamasında en önemlileri kırıklık ve kusmadır. En önemsizleri cinsiyet, bilinç değişikliği, kan ürünü ihtiyac olma durumudur.
- Gerçek veri seti için her üç gruptada DVR açıklayıcılık yüzdesi en fazla, HKO en düşük olan,regresyon modelidir. Açıklayıcı yüzdesi olarak en fazla, HKO en düşük olan grup yetişkin gruptur.
- Hasta grupları 1000 kez tekrarlı simülasyon sonuçları icelendiğinde; Toplam gruplarda hata kareler ortalaması en düşük ve model uyumu bakımından en yüksek olan regresyon modeli DVR'dir. Arkasından RF yöntemi gelmektedir. RA yöntemi diğerlerine göre düşük performans göstermiştir.
- Korelasyon yapılarına göre gözlem sayısı 100 olduğunda ve yöntem performansları karşılaştırıldığında DVR, her üç korelasyon yapısı için HKO değerleri acısından en düşük $R^2$ değeri en büyüktür. Düşük düzey korelasyon yapısına göre regresyon yöntemlerinin performanslarına bakıldığında en iyi sonucu DVR göstermiş olup arkasından RF gelmektedir. En düşük performans RA göstermektedir. Orta düzey korelasyon yapısına göre regresyon yöntemlerinin performanslarına bakıldığında en iyi sonucu DVR göstermiş olup arkasından RF gelmektedir.En düşük performans RA göstermektedir. Yüksek düzey korelasyon yapısına göre regresyon yöntemlerinin performanslarına bakıldığında en iyi sonucu DVR göstermiş olup onu RF takip etmektedir. En düşük performans RA göstermektedir.
- Korelasyon yapılarına göre gözlem sayısı 250 olduğunda ve yöntem performansları karşılaştırıldığında en iyi regresyon yöntemi DVR'dir. Arkasından RF yöntemi gelmektedir. En düşük performans RA ya aittir. RA

regresyon modelinde korelasyon yapısındaki ilişki arttıkça HKO'de azalma, açıklama yüzdesinde artış olmuştur.

- Korelasyon yapılarına göre gözlem sayısı 1000 olduğunda ve yöntem performansları karşılaştırıldığında en iyi regresyon yöntemi DVR'dir. DVR en yüksek açıklama yüzdesine sahip ve en küçük HKO değerine sahiptir. Bunu RF takip etmektedir. En düşük değerler RA elde edilmektedir, ancak bu yöntemin korelasyon yapılarındaki ilişki arttıkça açıklama yüzdesi artmakta, HKO değeri oldukça azalmaktadır.

## ÖNERİLER

Çalışmadan elde edilen veriler doğrultusunda aşağıdaki öneriler geliştirilmiştir;

- Literatürde yapılan çalışmalarda veri madenciliğinde kullanılan regresyon yöntemerinin performansları genelde tekli veya ikili olarak incelenmiştir. Bu çalışmada üçlü olarak karşılaştırma yapılmıştır. ileriye yönelik çalışmalarda çoklu karşılaştırmalar yapılması önerilmektedir.
- Bu çalışma R programında yapılması bakımından bir ilk olup önem teşkil etmektedir, yapılacak çalışmalarında bu programda yapılması önerilmektedir.

# KAYNAKLAR


Aggarwal Charu C. and Wang Haixun (2010). *Managing and Mining Graph Data*. London: Springer.

Akçay Ahmet (2014). *Bilgi ve Belge Yönetiminde Veri Medenciliğ*i. İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi Anabilim Dalı.Yüksek Lisans Tezi.

Akçetin Eyüp ve Çelik Ufuk (2015). *Karınca Kolonisi Optimizasyonu Sınıflama Algoritması Yöntemi ile Telefon Bankacılığında Doğrudan Pazarlama Kampanyası Üzerine Bir Sınıflama Analizi.* İnternet Uygulamaları ve Yönetimi. 7.

Akgöbek Ömer ve Kaya Serkan (2011). "Veri Madenciliği Teknikleri İle Veri Kümelerinden Bilgi Keşfi: Medikal Veri Madenciliği Uygulaması*"."E-Journal of New World Sciences Academy*, 238.

Akman Muhammet ve Genç Yasemin ve Ankaralı Handan (2011). "Random Forest Yöntemi ve Sağlık Alanında Bir Uygulama". *Türkiye Klinikleri Journal of Biostatistics*, 3(1), 36-48.

Akpınar Haldun (2000). "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği*"*. *İstanbul Üniversitesi İşletme Fakültesi Dergisi, 5.*

Akpınar Haldun (2014). *Data Veri Madenciliği Veri Analizi*. Papatya: İstanbul

Akyol Kemal, Şen Baha ve Çalık Elif (2012). *Biyokimya Ve Hemogram Laboratuar Test Sonuçlarının Lojistik Regresyon Yöntemiyle Analizi*. Uşak: Uşak Üniversitesi. XIV. Akademik Bilişim Konferansı. 313.

Alagöz Ali, Öge Serdar ve Ortakarpuz Metehan (2014). "Bir Kurumsal Zeka Teknolojisi Olarak Veri Madenciliği ile Muhasebe Bilgi Sistemi Ilişkisi" *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Özel Sayı,* 1-21.

Alan Mehmet Ali (2012). "Veri Madenciliği ve Lisansüstü Öğrenci Verileri Üzerine Bir Uygulama*"*. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi,* 165.

Albayrak Ali Sait ve Yılmaz Koltan Şebnem (2009). "Veri Madenciliği: Karar Algoritmaları ve Imkb Verileri üzerine Bir Uygulama". *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, C: 14, S: 1, 38.

Alkan Ali ve Falay Emre (2007, Eylül-Ekim). "Kamu Uygulamalarında Çözüm Veri Madenciliği". *Strateji Geliştirme Başkanlığı Strateji Bülteni Sayı:* 5, S. 1-31.

Alpar Reha (2011). *Çok Değişkenli İstatiksel Yöntemler*. Ankara: Detay Yayıncılık.

Alpar Reha (2016). *Spor, Sağlık Ve Eğitim Bilimlerinde Örneklerle Uygulamalı İstatistik ve Geçerlik-Güvenirlik-Spss'de Çözümleme Adımları ile Birlikte*. Ankara: Detay Yayıncılık.

Alpaydın Ethem (2010). *Introduction To Machine Learning*. London, England: The Mit Press Cambridge.

Altıntaş Tamer (2006). *Veri Madenciliği Metotlarından Olan Kümeleme Algoritmalarının Uygulamalı Etkinlik Analizi*. Sakarya: Sakarya Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi.

Arslan Volkan ve Yılmaz Güray (2010). "*Karar Destek Sistemlerinin Kullanımı İçin Uygun Bir Model Geliştirilmesi*". Journal of Aeronautics & Space Technologies/Havacılık ve Uzay Teknolojileri Dergisi. 4(4):75-82.

Asilkan Özcan (2008). *Veri Madenciliği Kullanılarak Ikinci El Otomobil Pazarında Fiyat Tahmini.* Antalya: Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü Doktora Tezi.

Ata Nihal, Özkök Erengül ve Karabey Uğur (2008). Mühendislik "*Veri Madenciliğinde Yaşam Çözümlemesi: Kredi Kartı Sahipleriyle Ilgili Bir Uygulama"*.Fen Bilimleri Dergisi. 26(1): 33-42.

Atak Fatih (2014). *Gerçek Ağ Verisi Üzerinde Veri Madenciliği Uygulamalarının Karşılaştırılması.* Ankara: Gazi Üniversitesi Bilişim Enstitüsü .Yüksek Lisans Tezi.

Ataseven Sinan (2008). *Üniversitelerin Adaylar Tarafından Tercih Edilme Desenlerini Veri Madenciliği Yöntemleri ile Belirleyen Bir Model Önerisi.*

İstanbul: İstanbul Kültür Üniversitesi Fen Bilimleri Enstitüsü .Yüksek Lisans Tezi.

Ayanoğlu Murat, Mert Kazım ve Giray Emel (2004). "*Perakende Sektöründe Veri Madenciliği Vazgeçilmez Mi? Alternatifi Crm Mi?"* 15-18 Haziran.Yöneylem Araştırması/Endüstri Mühendisliği 24. Ulusal Kongresi. Gaziantep-Adana: 1-3.

Aydın Sinan (2007). *Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama.* Eskişehir: Anadolu Üniversitesi Sosyal Bilimler Enstitüsü Doktora Tezi.

Ayık Yusuf Ziya, Özdemir Abdulkadir ve Yavuz Uğur (2007). "*Lise Türü Ve Lise Mezuniyet Başarısının Kazanılan Fakülte Ile Ilişkisinin Veri Madenciliği Tekniği Ile Analizi*." Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi. 10(2): 441-454.

Basak Debasish, Pal Srimanta and Dipak Chandra Patranabis (2007). "*Support Vector Regression"*. Neural Information Processing. 11:203-224.

Baykal Abdullah. (2006). "*Veri Madenciliği Uygulama Alanları"*. Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi.7: 95-107.

Bilgin Turgay Tugay ve Çamurcu Yılmaz (2005). "*Dbscan, Optics Ve K-Means Kümeleme Algoritmalarını Uygulamalı Karşılaştırılması*". Politeknik Dergisi. 8(2): 139-145.

Birant Derya, Kut Alp, Ventura Medi, Altınok Hakan, Altınok Benal, Altınok Elvan, ve Ihlamur Murat (10-12 Şubat 2010). *İş Zekası Çözümleri İçin Çok Boyutlu Birliktelik Kuralları Analizi*. Muğla: XII. Akademik Bilişim Konferansı Bildirileri. 215-222

Bozkır Ahmet Selman (2009). *Olap Veri Madenciliği Teknolojilerinden Yararlanılarak Web Tabanlı Bir Karar Destek Sisteminin Gerçekleştirilmesi*. Ankara: Hacettepe Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Bramer Max (2007). *Principles of Data Mining*: London. Springer

Breiman Leo (2001). *Machine Learning*. Kluwer Academic Publishers 45:5-32

Cabena Peter, Hadjinian Pablo, Stadler Rolf, Verhees Jaap and Zanasi Alessandro (1998). *Discovering Data Mining: From Concept To Implementation*. USA: Prentice- Hall .Upper Saddle River. Nj.

Cevahir Fahrettin (2011). *Bir Perakende Firmasına Ait Veriler Üzerinden Veri Madenciliği Uygulamaları*. İstanbul: Mimar Sinan Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Ceyhan İsmail Fatih (2014). *Bağımsız Denetim Kalitesini Artırıcı Bir Yöntem Olarak Veri Madenciliği: Borsa İstanbul Uygulaması*. Sakarya Üniversitesi. Sosyal Bilimler Enstitüsü. Doktora Tezi.

Chang Ming-Wei ve Lin Chih-Jen (2005). *Leave-One-Out Bounds For Support Vector Regression Model Selection.* Taiwan: Deparment of Computer Science and Informantion Engineering National University Taipei.

Chayama K, Hayes Cn, Yoshioka K, Moriwaki H, Okanoue T, Sakisaka S, Takehara T, Oketani M, Toyota J, Izumi N, Hiasa Y, Matsumoto A, Nomura H, Seike M, Ueno Y, Yotsuyanagi H and Kumada H (2011). *"Factors Predictive Of Sustained Virological Response Following 72 Weeks of Combination Therapy For Genotype 1b Hepatitis C"*. Journal of Gastroenterology. 46(4): 545-55

Coşgun Erdal ve Karaağaoğlu Ergun (14-17 Ekim 2010). *Genetik Araştırmalarda Machine Learning ve Veri Madenciliği*. Magosa KKTC: VII. Ulusal Tıp Bilişimi Kongreleri. 162.

Coşkun Cengiz (2010). *Veri Madenciliği Algoritmaları Karşılaştırılması*. Diyarbakır: Dicle Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi.

Çalış Aslı (2013). *Veri Madenciliği Yaklaşımı İle Bireysel Müşterilerin Kredi Ödeme Performanslarının Değerlendirilmesi.* Kocaeli: Kocaeli Üniversitesi. Fen Bilimleri Enstitüsü. Endüstri Mühendisliği Anabilim Dalı. Yüksek Lisans Tezi.

Çalış Aslı, Kayapınar Sema ve Çetinyokuş Tahsin (2014). *Veri Madenciliğnde Karar Ağacı Algoritmaları Ile Bilgisayar Ve İnternet Güvenliği Üzerine Bir Uygulama*. Endüstri Mühendisliği Dergisi. 25 (3-4): 2-19.

Çankırı Süreyya, Kartal Elif, Yıldırım Kemal ve Gülseçen Sevinç (1-2 Ekim 2009). "*Organizasyonlarda Bilgi Yönetimi Sürecinde Veri Madenciliği Yaklaşımı"*

Bilgi Çağında Varoluş: Fırsatlar ve Tehditler Sempozyumu.ÜNAK. İstanbul: Yeditepe Üniversitesi. 148-167.

Çelik Yusuf (2011). *Biyoistatistik Bilimsel Araştırma Spss Nasıl?* Yazarın kendi yayını. 557.

Çetin Muhammed (2009). *Bir Üretim İşletmesinde Veri Madenciliği Uygulaması*. Sakarya: Sakarya Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Çınar Ayşe ve Silahtaroğlu Gökhan (2012). "Veri Madeniliği Teknikleri Ile Müşteri Memnuniyetine Etki Eden Gizli Nedenlerin Keşfi*"Marmara Üniversitesi İİBF Dergisi*. 33(2): 309-330.

Çil Fatih (2010). *Banka Yatırım Fonu Müşteri Hareketlerinin Belirlenmesine Yönelik Bir Veri Madenciliği Uygulaması.* Ankara: Gazi Üniversitesi. Fen Bilimleri Enstitüsü. Yükseklisans Tezi.

Çoban Aslan (2006). *İmalat Sanayinde Veri Madenciliği Destekli Tedarikçi Seçimi Uygulaması.* Sakarya: Sakarya Üniversitesi. Fen Bilimleri Enstitüsü. Makine Eğitimi Anabilim Dalı. Doktora Tezi.

Dandıl Emre (2013). "Karaciğerde Oluşan Hastalıkların Tespitinde Makine Öğrenmesi Yöntemlerinin Kullanılması". Akademik Bilişim Konferansı. 23-25 Ocak Antalya: Akdeniz Üniversitesi.1-3.

Daş Bihter ve Türkoğlu İbrahim (2014). "Dna Dizilimlerinin Sınıflandırılmasında Karar Ağacı Algoritmalarının Karşılaştırılması". Eleco 2014 Elektrik-Elektronik-Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu. 28-29 Kasım. Bursa: 381-383.

Değirmenci Tuğba (2014). *Resmi İstatistiklerde Veri Madenciliği Yaklaşımı*. Kayseri: Erciyes Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Doğan Şengul ve Türkoğlu İbrahim (2008). "*Iron-Deficiency Anemia Detection From Hemotology Parameters By Using Decision Trees"*. International Journal of Science and Technology. 85-92.

Dolgun Muhsin Özgür (2015). *Veri madencilği sınıflama yöntemlerinin başarılarının; bağımlı değişken prevalansı, örneklem büyüklüğü ve bağımsız değişkenler*

*ilişki yapısına göre karşılaştırılması*. Ankara: Hacettepe Üniversitesi. Sağlık Bilimler Ensitüsü. Doktora tezi.

Dua Sumet and Du Xian (2011). *Data Mining and Machine Learning In Cybersecurity. 6000 Broken Sound Parkway Nw, Suite 300 Boca Raton*. Taylor & Francis Group Crc Press.

Düzgünoğlu Selda, Yazıcı Adnan ve Yarımağan Ünal (2006) "Tıp Bilişiminde Veri Ambarı ve Veri Madenciliği Uygulaması". 16-19 Kasım. Antalya: III. Ulusal Tıp Bilişimi Kongresi.110-115.

Ekim Ufuk (2011). *Veri Madenciliği Algoritmalarını Kullanarak Öğrenci Verilerinden Birliktelik Kurallarının Çıkarılması*. Konya: Selçuk Üniversitesi. Fen Bilimleri Enstitüsü.Yüksek Lisans Tezi.

Elmas Fatih (2014). *Kalp Krizi Riskinin Bir Veri Madenciliği Uygulamasıyla Analizi*. Muğla: Muğla Sıtkı Koçman Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Emel Gül Gökay ve Taşkın Çağatan (2005). "Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması". *Osmangazi Üniversitesi. Sosyal Bilimler Dergisi*. 6 (2): 221-239.

Erol Bahar (2013). *Müşteri İlişkileri Yönetimi İçin Veri Madenciliği Kullanılması ve Sigortacılık Sektörü Üzerine Bir Uygulama*. İstanbul: Marmara Üniversitesi. Fen Bilimleri Enstitüsü.Yüksek Lisans Tezi.

Erpolat Semra (2012). "Otomobil Yetkili Servislerinde Birliktelik Kurallarının Belirlenmesinde Apriori ve Fp-Growth Algoritmalarının Karşılaştırılması*"*. *Anadolu Üniversitesi. Sosyal Bilimler Dergisi.* 12(2): 151-166.

Erşahin Burak ve Argüden Yılmaz (2008). *Veri Madenciliği Veriden Bilgiye, Masraftan Değere*. İstanbul: Arge Danışmanlık A. Ş.

Ertuğrul İrfan, Orfan Arzu ve Şavlı Ayşegül (2013). "Veri Madenciliği Uygulamasına İlişkin Pamukkale Üniversitesi Hastanesinde Hasta Profilinin Belirlenmesi". *Pamukkale Üniversitesi. Mühendislik Bilimleri Dergisi.* 19(2): 97-103.

Esen Fevzi M (2009). *Veri tabanlarından Bilgi Keşfi: Veri Madenciliği ve Bir Sağlık Uygulaması*. İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Yüksek Lisans Tezi.

Farboudi Sara (2009). *Tıp Bilişiminde İstatistiksel Veri Madenciliği.* Ankara: Hacettepe Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Faridi A, Sakomura Nk, Golian A and Marcato Sm (2012).'' Predicting Body And Carcass Characteristics Of 2 Broiler Chicken Strains Using Support Vector Regression and Neural Network Models". *Poultry Science* 91(12): 3286-94.

Fleiss Josebs L., Levin Bruce and Paik Cho Myunghee (2003). *Statistical Methods For Rates And Proportions.* New Jersey:John Wiley & Sons. Inc

Liao T.Warren and Triantaphyllou Evangelos (2007). *Recent Advances In Data Mining Of Enterprise Data: Algorithms And Applications.* Singapore: World Scientific Publishing Co. Pte. Ltd

Gemici Burhan (2012). *Veri Madenciliği ve Bir Uygulaması*. İzmir: Dokuz Eylül Üniversitesi. Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi.

Gökmen Şenol (2015). *Müşteri İlişkileri Yönetiminde Bir Araç Olarak Veri Madenciliği Ve Perakende Sektöründe Bir Uygulama*. İstanbul: Yıldız Teknik Üniversitesi Sosyal Bilimler Enstitüsü Yüksek Lisans Tezi.

Güllüoğlu Sabri Serkan (2011). "Tıp ve Sağlık Hizmetlerinde Veri Madenciliği Çalışmaları: Kanser Teşhisine Yönelik Bir Ön Çalışma". *Online Academic Journal Of Information Technology*. 2(5): 1-7.

Gürbüz Feyza (2009). *Hayvancılık Sektöründe Veri Madebciliği Ile Farklı Sınıflama Tekniklerinin Karşılaştırmalı Olarak Uygulanması*. Kayseri: Erciyes Üniversitesi. Fen Bilimleri Enstitüsü. Doktora Tezi.

Hacıefendioğlu Şerife (2012). *Makine Öğrenmesi Yöntemleri Ile Glokom Hastalığının Teşhisi.* Konya: Selçuk Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Han Jiawei and Kamber Micheline (2006). *Data mining: Concepts and Techiniques*.Second Edition. Morgan Kaufmann.

Han Jiawei, Kamber Micheline and Pei Jian (2012). *Data Mining Concepts and Techniques,* Third Edition. Usa: Elsevier

Hatipoğlu Bünyamin (2013). *Dershane Eğitiminin Üniversiteye Yerleşmedeki Etkisinin Veri Madenciliği Ile İrdelenmesi.* İstanbul: İstanbul Aydın Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Ho Chia-Hua and Lin Chih-Jen (2012).”Large Scale Linear Support Vector Regression*”. Taiwan: Journal of Machine Learning Research* 13: 3323-3348.

Hosmer David W. and Lemeshow Stanley (1989). *Applied Logistic Regression*. New York: John Wiley&Sons

Irmak Sezgin (2009). *Veri Madenciliği Yöntemleri Ile Sağlık Sektörü Veritabanlarında Bilgi Keşfi: Tanımlayıcı ve Kestirimci Model Uygulamaları*. Antalya: Akdeniz Üniversitesi. Sosyal Bilimler Enstitüsü. Doktora Tezi.

Irmak Sezgin, Köksal, Can Deniz ve Asilkan Özcan (2012). “Hastanelerin Gelecekteki Hasta Yoğunluklarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi*”. Uluslararası Alanya İşletme Fakültesi Dergisi*. 4(1): 101-114.

Izenman Alan Julian (2008). *Modern Multivariate Statistical Techniques*. Usa: Springer

İnce Ali Rıza ve Alan Mehmet Ali (2014). “Ürün Portföy Planlamasinda Veri Madenciliğindenn Yararlanilmasi Üzerene Bir Çalisma”. *Eul Journal Of Social Sciences (V:II) Laü Sosyal Bilimler Dergisi*. 65-67

İmhoff Claudia, Galemmo Nicholas and Geiger Jonathan (2003). *Mastering Data Warehouse Design Relation and Dimensional Techniques.* İndianapolis. Usa: Wiley Publishing.

Jackson Joyce (2002). *Data Mining: A Conceptual Overview*. Communication Of The Association For İnformation System Magazine. 8(1): 267-296.

Jacobs, P. (1999). Data Mining: What General Manegers Need To Know. Harward Management Update, C: 4, S: 10, 8.

Karahoca Adem (2012). A*dvances İn Data Mining Knowledge Discovery and Applications*. Croatia. Intech

Karakaya Mevlüt (1994). *Muhasebe Bilgi Sistemi ve Bilgi Teknolojisi*.Ankara

Karakoyun Murat ve Hacıbeyoğlu Mehmet (2014). "Biyomedikal Veri Kümeleri İle Makine Öğrenmesi Sınıflama Algoritmalarının Istatistiksel Olarak Karşılaştırılması." *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Mühendislik Bilimleri Dergisi.* 16(48): 30-42.

Kayrı Murat ve Boysan Murat (2008). **"**Bilişsel Yatkınlık İle Depresyon Düzeyleri İlişkisinin Sınıflandırma ve Regresyon Ağacı Analizi İle İncelenmesi*". Ankara: Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*. 34: 168-177.

Kavzoğlu Taşkın ve Çölkesen İsmail (2010). "Destek Vektör Makineleri İle Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi*". Harita Dergisi.* 144: 73-82.

Kaya Halil ve Köymen Kemal (2008). "Veri Madenciliği Kavramı ve Uygulama Alanları*" Doğu Anadolu Bölgesi Araştırma ve Uygulama Dergisi*. 2:159-164.

Kayri Murat (2008). "Elektronik Portfolyo Değerlendirmeleri İçin Veri Madenciliği Yaklaşımı." *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*. 5(1): 98-110.

Kelleci Ender, Ergen Sergen ve Uyuçgil Hakan (2011). *"*Konumsal Veri Modeli Içinde, Konumsal Veri Tabanı". *Eskişehir: Anadolu Üniversitesi. Aöf Yayını.*1338: 2-23.

Kiremitçi Barış (2005). *Veri Ambarlarında Veri Madenciliği Ve Ulaştırma-Lojistik Sektöründe Bir Uygulama*. İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü. İşletme Anabilim Dalı. Sayısal Yöntemler Bilim Dalı Yayımlanmamış. Yüksek Lisans Tezi.

Kitler Richard and Wang Weidong (1998). "*The Emerging Role Of Data Mining*". Solid State Technology. 42: 11- 45.

Kocabaş Fatma (2010). *Veri Madenciliği Süreci ve Gerçek Bir Veri Seti Üzerinde Uygulaması*. Ankara: Hacettepe Üniversite. Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi.

Koçoğlu Önay Fatma (2012). *Veri Madenciliği Sürecinde Veri Ayrıklaştırma Yöntemlerinin Karşılaştırması ve Bir Uygulama*. İstanbul: İstanbul Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Kolay Gökçe (2006). *İşletmelerde Bilgi Sistemleri Verimliliğini Arttırmada Veri Madenciliği Yöntemi: Bir Simülasyon Çalışması.* Zonguldak: Karaelmas Üniversitesi. Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi.

Koyuncugil Ali Serhan ve Özgülbaş Nermin (2009). "Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları". *Bilişim Teknolojileri Dergisi* 2(2): 21-32.

Küçüksille Engin (2009). *Veri Madenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi ve Imkb: Hisse Senetleri Piyasasında Bir Uygulama.* Isparta: Süleyman Demirel Üniversitesi. Sosyal Bilimler Enstitüsü. İşletme Anabilim Dalı. Doktora Tezi.

Leidy Nk, Malley Kg, Steenrod Aw, Mannino Dm, Make B, Bowler Rp, Thomashow Bm, Barr Rg, Rennard Si, Houfek Jf, Yawn Bp, Han Mk, Meldrum Ca, Bacci Ed, Walsh Jw and Martinez F (2016). "Insight Into Best Variables For Copd Case Identification: A Random Forests Analysis". *Chronic Obstructive Pulmonary Diseases* 3(1):406-418.

Meriç Osman Arda (2004). *Veri Madenciliği Aracı Olarak Genetik Algoritmalar ile Yapay Sinir Ağları ve Genetik Algoritma-Yapay Sinir Ağı Melez Modelinin Müşteri Değerlendirilmesinde Kullanılması*. İstanbul: İstanbul Üniversitesi. Sosyal Bilimler Enstitüsü. Doktora Tezi.

Nısbet Robert, Elder Jhon and Mıner Gary (2009). *Handbook Of Statistical Analysis and Data Mining Applications*. Usa: Elsevier Inc.

Nizam Hatice ve Akın Saliha Sıla (2015). "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması". erişim tarihi: 15 Temmuz. Inetr.Org.Tr/Inetconf19/Bildiri/10.Pdf

Oğuzlar Ayşe (2003). "Veri Önişleme". *Erciyes Üniversitesi İİBF Dergisi*. 21: 67-76.

Özdemir Abdulkadir (2004). *Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği-Sağlık Sektöründe Uygulama*. Erzurum: Atatürk Üniversitesi Sosyal Bilimler Üniversitesi. İşletme Anabilim Dalı. Doktora Tezi.

Özgülbaş Nermin ve Koyuncugil Ali Serhan 2010). "Sağlık Bakanlığı Hastanelerinin Finansal Risklere Göre Sınıflandırılması: Veri Madenciliği Modeli". 28 Nisan-01 Mayıs. III. Uluslararası Sağlıkta Performans ve Kalite Kongresi. 1-623.

Özgür Atilla ve Erdem Hamit (2012). "Saldırı Tespit Sistemlerinde Kullanılan Kolay Erişilen Makine Öğrenme Algoritmalarının Karşılaştırılması". *Bilişim Teknolojileri Dergisi*. 5 (2): 41-48.

Özcan Tuncay (2011). *Perakende Endüstrisinde Raf Alanı Yönetimine Veri Madenciliği Esaslı Analitik Bir Yaklaşım.* İstanbul: İstanbul Üniversitesi. Fen Bilimleri Enstitüsü. Doktora Tezi.

Özkan Yalçın (2013). *Veri Madenciliği Yöntemleri. Bilgisayar Bilimler ve Mühendisliği.*2.Basım. İstanbul: Papaya Yayıncılık Eğitim.

Öztürk Ümit (2014). *Lojistikte Fiyatlandırmayı Iyileştirme Amaçlı Olarak Veri Madenciliği Teknikleri ile Bir Öneri.* İstanbul: Beykent Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Özyirmidokuz Kahya Esra (2009). *Veri Madenciliği Tekniklerini Kullanarak İmalat Verilerinin Modellenmesi ve Analizi*. Kayseri: Erciyes Üniversitesi. Sosyal Bilimler Enstitüsü. Doktora Tezi.

Pala Tuba (2013). *Tıbbi Karar Destek Sisteminin Veri Madenciliği Yöntemleriyle Gerçekleştirilmesi*. İstanbul: Marmara Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi.

Rawlings Jhon.O, Pentula Sastry.G ve Dickey David A. (1998). A*pplied Regression Analysis: A Research Tool*.USA: Springer.

Razbonyalı Can ve Özkaya Aslı Uyar. (2014). "Makine Öğrenmesi ile Ürün Sınıflama İncelemesi". 5-7 Şubat. Mersin:14. Akademik Bilişim Konferansı. 2-3.

Riffenburgh Robert H (2012). *Statistics in Medicine*. Usa: Elsevier.

Samuel Arthur (2000). "Some Studies İn Machine Learning Using The Game Of Checkers". *Ibm Journal Of Research And Development*. 44(1): 206-226.

Savaş Serkan, Topaloğlu Nurettin ve Yılmaz Mithat (2012). "Veri Madenciliği ve Türkiye'deki Uygulama Alanları". *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*.7.

Saygılı Ahmet (2013). *Veri Madenciliği ile Mühendislik Fakültesi Öğrencilerinin Okul Başarılarının Analizi.* İstanbul: Yıldız Teknik Üniversitesi. Fen Bilimleri Enstitüsü. Bilgisayar Mühendisliği Anabilim Dalı. Bilgisayar Mühendisliği Programı. Yüksek Lisans Tezi.

Seyrek İbrahim Halil ve Ata H. Ali (2010). "Veri Zarflama Analizi ve Veri Madenciliği ile Mevduat Bankalarında Etkinlik Ölçümü". *Bddk Bankacılık ve Finansal Piyasalar Dergisi.* 4(2): 67-84.

Sever Süleyman Zafer (2010). *Yoğunluk Tabanlı Kümeleme Metotları Kullanılarak Paralel Veri Madenciliği Gerçekleştirilmesi*. İstanbul: Maltepe Üniversitesi Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Sevindik Tuncay, Kayışlı Korhan ve Ünlükahraman Orhan (2012). "Web Tabanlı Eğitimde Veri Madenciliği". *Turkish Journal Of Computer And Mathematics Education.* 3(3): 183-193.

Shahbaba Babak (2012). *Biostatistics with R*. London: Springer

Silahtaroğlu Gökhan (2008). *Kavram Ve Algoritmalarıyla Temel Veri Madenciliği*. İstanbul: Türkiye Papatya Yayıncılık.

Smola Alex J and Schölkopf Bernhard (2004). A *Tutorial On Suppor Vector Regression.* Germany: Statistics and Computing 14: 199–222.

Soukup Tom ve Davidson Ian (2002). *Visual Data Mining.* Canada: John Wiley & Sons.

Stephens Ryan ve Plew Ron (2003). *24 Saatte Veri Tabanları*. (Çev. Nalan Güven Küçüker). İstanbul: Alfa Yayınları.

Sullivan Rob (2012). *Introduction The Data Mining For Life Scince*. Usa: Human Press Springer.

Şengül Ayşe Yasemin (2010). *Dağitik Veri Tabanı Sistemlerinde Optimizasyon Süreçleri ve Uygulamalari*. İstanbul: Marmara Üniversitesi. Sosyal Bilimler Enstitüsü. İletişim Bilimleri Anabilim Dalı. Bilişim Bilim Dalı. Doktora Tezi.

Şık Muhammet Şamil. (2014). *Veri Madenciliği ve Kanser Erken Teşhisinde Kullanımı*. Malatya: İnönü Üniversitesi Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi.

Şimşek Umman Tuğba (2006). *Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (Crm) Bir Uygulama*. İstanbul: İstanbul Üniversitesi. Sosyal Bilimler Enstitüsü. Doktora Tezi.

Tayyar Nezih ve Tekin Selin. (2013). "İmkb-100 Endeksinin Destek Vektör Makineleri İle Günlük, Haftalık ve Aylık Verileri Kullanılarak Tahmin Edilmesi".*Bolu: Abant İzzet Baysal Üniversitesi. Sosyal Bilimler Enstitüsü Dergisi.* 13(13): 189-217.

Terlemez Levent (2008). *Eş İşlem Stratejisi Yöntemiyle İmkb'de Portföy Oluşturmada Veri Madenciliği Uygulaması*. Anadolu Üniversitesi. Fen Bilimleri Enstitüsü. İstatistik Anabilim Dalı. Doktora Tezi

Terzi Özlem, Küçüksille Ecir Uğur, Ergin Gülşah ve İlker Ahmet (2011). "Veri Madenciliği Süreci Kullanılarak Güneş Işınımının Tahmini". Sdu International Technologic Science. 3(2): 29-37.

Terzi Serkan (2012). "Hile ve Usulsüzlüklerin Tespitinde Veri Madenciliğinin Kullanımı". Muhasebe ve Finansman Dergisi. 54: 51-64.

Thomsen Eric (2002). *Olap Solutions*. *Building Multidimensional İnformation Systems*. Canada: John Wiley & Sons. Inc

Timor Mehpare, Ezerçe Ayşegül ve Gürsoy U. Tuğba (2011). "Müşteri Profili ve Alışveriş Davranışlarını Belirlemede Kümeleme ve Birliktelik Kuralları Analizi". *Perakende Sektöründe Bir Uygulama*. *Yönetim Dergisi*. 22(68): 128-147.

Timur Mustafa, Aydın Fatih ve Akıncı T.Çetin (2011). "İstanbul Göztepe Bölgesinin Makine Öğrenmesi Yöntemi İle Rüzgar Hızının Tahmin Edilmesi". *Makine Teknolojileri Elektronik Dergisi.* 8(4): 75-80.

Torgo Luis (2014). *Data Mining With R, Learning With Case Studies*. 6000 Broken Sound Parkway Nw. Suite 300 Boca Raton: Chapman&Hall/Crc.

Tosun Tuğba (2006). *Veri Madenciliği Teknikleri ile Kredi Kartlarında Müşteri Kaybetme Analizi.* İstanbul: İstanbul Teknik Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Tuğ Emine (2005). *Genel Algoritmalar ile Tıbbi Veri Madenciliği.* Konya: Selçuk Üniversitesi. Fen Bilimleri Enstitüsü. Bilgisayar Mühendisliği Ana Bilim Dalı. Yüksek Lisans Tezi.

Tümen Vedat (2013). *Veri Madenciliği Yöntemleri ile Güç Kalitesi Verilerinin İncelenmesi*. Tunceli: Tunceli Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Tüzüntürk Selim (2010). "Veri Madenciliği ve İstatistik". *Uludağ Üniversitesi IİBF Dergisi.* 29(1): 65-90.

Uçan Ömer (2010). *Dijital Kütüphanelerde Veri Madenciliği Uygulamaları. Akdeniz Üniversitesi Merkez Kütüphanesi Örneği.* Antalya: Akdeniz Üniversitesi. Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi.

Yakut Emre (2012). *Veri Madenciliği Tekniklerinden C 5.0 Algoritması Ve Destek Vektör Makineleri ile Yapay Sinir Ağlarının Sınıflandırma Başarılarının Karşılaştırılması: İmalat Sektöründe Bir Uygulama*. Erzurum: Atatürk Üniversitesi. Sosyal Bilimler Enstitüsü. İşletme Anabilim Dalı. Doktora Tezi.

Yeleğin Ali (2012). *Mesleki Eğitimde Öğrenci Altyapısının Öğrenci Eğitim Başarısına Etkisinin Veri Madenciliği Yöntemleriyle Ortaya Çıkartılması.* İstanbul: Beykent Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

Yeşilbudak Mehmet, Kahraman Hamdi Tolga ve Karacan Hacer (2011). "Veri Madenciliğinde Nesne Yönelimli Birleştirici Hiyerarşik Kümeleme Modeli". *Gazi Üniversitesi. Mühendislik ve Mimarlık Fakültesi Dergisi*. 26(1): 27-39.

Yıldırım Pınar, Uludağ Mahmut ve Görür Abdülkadir (2008). "Hastane Bilgi Sistemlerinde Veri Madenciliği". 30 Ocak-1 Şubat. Çanakkale: Onsekiz Mart Üniversitesi Akademik Bilişim.429-434

Yılmaz Emrah (2006). *Kütahya İlinde Sosyal Sınıfların Belirlenmesi ve Veri Madenciliği ile Tüketici Profilinin Çıkarılmasına Yönelik Bir Uygulama*. Kütahya: Dumlupınar Üniversitesi. Sosyal Bilimler Enstitüsü. İşletme Anabilim Dalı. Yüksek Lisans Tezi.

Yurtay Yüksel, Salman Yavuz, Salman Mehmet Emin ve Gençali Fatih (2013). "Kansızlık Tanısına İlişkin Bir Veri Madenciliği Uygulaması". 07-09 Haziran. Sakarya Üniversitesi. Isites 2013 International Symposium On Innovate Technologies İn Enginering Science. 898-899

Wu Chih-Hung, Tzeng Gwo-Hshiung ve Lin Rong-Ho (2009). *A Novel Hybrid Genetic Algorithm For Kernel Function And Parameter Optimization In Support Vector Regression*: Expert Systems. Usa: Elsevier.

Witten Ian H. and Frank Eibe (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Second Edition. SanFrancisco.Ca. Usa: Morgan Kaufmann Publishers.

Wu Xindong and Kumar Vipin (2009). *The Top Ten Algorithms In Data Mining*: Usa: By Taylor & Francis Group, Llc.

Zhao Yanchang (2013). *R and Data Mining*. Usa: Elsevier.

# EKLER

**EK:1** Kırım Kongo Kanamalı Ateş Verisine Göre Destek Vektör Regresyon,Random Forest Ve Regresyon Ağaçlarının Performanslarının Karşılaştırılmasında Kullanılan R Programındaki Kodlar

```
library(e1071)
library(rpart)
library(randomForest)
library(corpcor)
library(MASS)

kkka.veri=read.csv2(file.choose(), header=TRUE)

sonuc.svm = svm(YATCIK_T~.,data=kkka.veri, type="eps-
regression",kernel="radial", gamma=0.5, cost=2)
pred.svm=predict(sonuc.svm,kkka.veri)
mse.svm=sum((kkka.veri$YATCIK_T - pred.svm)^2)/nrow(kkka.veri)
r2.svm = (cor(kkka.veri$YATCIK_T , pred.svm))^2

sonuc.rf <- randomForest(YATCIK_T ~ ., data=kkka.veri,
mtry=5, ntree=100, importance=TRUE,na.action=na.omit)
pred.rf=predict(sonuc.rf, kkka.veri,predict.all=TRUE)
mse.rf=sum((kkka.veri$YATCIK_T -
pred.rf$aggregate)^2)/nrow(kkka.veri)
r2.rf = (cor(kkka.veri$YATCIK_T , pred.rf$aggregate))^2


sonuc.rt=rpart(YATCIK_T ~ ., data=kkka.veri)
pred.rt=predict(sonuc.rt, kkka.veri)
mse.rt=sum((kkka.veri$YATCIK_T - pred.rt)^2)/nrow(kkka.veri)
r2.rt = (cor(kkka.veri$YATCIK_T , pred.rt))^2
```

## EK: 2 Gercek Veri Setinin Korelasyon Yapısı ile Simulasyon Kodları

```
çocuk verisi için

c.kor=read.csv2(file.choose(),header=T)
c.kor=make.positive.definite(c.kor)

d=24
mean=rep(0,d)

kesim=c(-0.17,-0.50, 0, 0, 0,1.07,-
0.06,1.47,1.62,0.77,1.47,1.81,-0.89,-1.30,1.04,-0.89,-
0.57,0.40,0.77,-1.41,-1.54,0.37,0.80,0.57)

tez.fonksiyon=function(n, nbin, mu, Sigma, num.sim)
{
# bağımlı değişkenin veride en son sütunda yer alması
gerekiyor
# n: gözlem sayısı
# nbin: nitel değişken sayısı
# mu: ortalama vektörü
# Sigma: korelasyon matrisi
# num.sim: simulasyon sayısı

pred.svm=numeric(n)
mse.svm=numeric(num.sim)
r2.svm=numeric(num.sim)

pred.rf=numeric(n)
mse.rf=numeric(num.sim)
r2.rf=numeric(num.sim)


pred.rt=numeric(n)
mse.rt=numeric(num.sim)
r2.rt=numeric(num.sim)


for (i in 1:num.sim)
{

mydata=data.frame(mvrnorm(n,mu,Sigma))


# nitel değişkenleri tanımlama
for (j in 1:nbin) {
     mydata[,j]=cut(mydata[,j], breaks=c(min(mydata[,j])-1,
kesim[j], max(mydata[,j])+1))
     }

#SVM ile model oluşturma, radial kernel kullanılarak
```

```
sonuc.svm = svm(X24~.,data=mydata, type="eps-
regression",kernel="radial",gamma=0.5,cost=2)

# svm modeli kullanılarak kestirim değerleri hesaplatma
pred.svm=predict(sonuc.svm,mydata)

# hata kareler ortalamasını hesaplatma
mse.svm[i]=sum((mydata$X24- pred.svm)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma

r2.svm[i] = (cor(mydata$X24 , pred.svm))^2


# Random Forest ile model oluşturma

sonuc.rf <- randomForest(X24~.,data=mydata,  mtry=5,
ntree=100, importance=TRUE,na.action=na.omit)

# RF modeli kullanılarak kestirim değerleri hesaplatma
pred.rf=predict(sonuc.rf, mydata,predict.all=TRUE)

# hata kareler ortalamasını hesaplatma
mse.rf[i]=sum((mydata$X24 - pred.rf$aggregate)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
r2.rf[i] = (cor(mydata$X24 , pred.rf$aggregate))^2


# Regresyon ağacı ile model oluşturma

sonuc.rt=rpart(X24~.,data=mydata)

# Regresyon ağacı modeli kullanılarak kestirim değerleri
hesaplatma
pred.rt=predict(sonuc.rt, mydata)

# hata kareler ortalamasını hesaplatma
mse.rt[i]=sum((mydata$X24 - pred.rt)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma

r2.rt[i] = (cor(mydata$X24 , pred.rt))^2

}

mse.svm.ort=mean(mse.svm)
r2.svm.ort=mean(r2.svm)

mse.rf.ort=mean(mse.rf)
r2.rf.ort=mean(r2.rf)

mse.rt.ort=mean(mse.rt)
```

```
r2.rt.ort=mean(r2.rt)

sonuc=list(MSE.SVM=mse.svm, MSE.SVM.ortalama=mse.svm.ort,
R2.SVM=r2.svm, R2.SVM.ortalama=r2.svm.ort,
 MSE.RF=mse.rf,MSE.RF.ortalama=mse.rf.ort, R2.RF=r2.rf,
R2.RF.ortalama=r2.rf.ort,
 MSE.RT=mse.rt, MSE.RT.ortalama=mse.rt.ort, R2.RT=r2.rt,
R2.RT.ortalama=r2.rt.ort)

return(sonuc)
}


cocuk=tez.fonksiyon(132, nbin=21, mu=mean, Sigma=y.kor,
num.sim=1000)
boxplot(c(cocuk$R2.RT,cocuk$R2.RF,
cocuk$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))



#yetiskin verisi

y.kor=read.csv2(file.choose(),header=T)
y.kor=make.positive.definite(y.kor)


d=25
mean=rep(0,d)


kesim=c(-0.17,-0.50, 0, 0, 0,1.07,-
0.06,1.47,1.62,0.77,1.47,1.81,-0.89,-1.30,1.04,-0.89,-
0.57,0.40,0.77,-1.41,-1.54,0.37,0.80,0.57,-0.37)


tez.fonksiyon=function(n, nbin, mu, Sigma, num.sim)
{
# bağımlı değişkenin veride en son sütunda yer alması
gerekiyor
# n: gözlem sayısı
# nbin: nitel değişken sayısı
# mu: ortalama vektörü
# Sigma: korelasyon matrisi
# num.sim: simulasyon sayısı

pred.svm=numeric(n)
mse.svm=numeric(num.sim)
r2.svm=numeric(num.sim)

pred.rf=numeric(n)
mse.rf=numeric(num.sim)
r2.rf=numeric(num.sim)
```

```
pred.rt=numeric(n)
mse.rt=numeric(num.sim)
r2.rt=numeric(num.sim)


for (i in 1:num.sim)
{

mydata=data.frame(mvrnorm(n,mu,Sigma))


# nitel değişkenleri tanımlama
for (j in 1:nbin) {
     mydata[,j]=cut(mydata[,j], breaks=c(min(mydata[,j])-1,
kesim[j], max(mydata[,j])+1))
     }

#SVM ile model oluşturma, radial kernel kullanılarak
sonuc.svm = svm(X25~.,data=mydata, type="eps-
regression",kernel="radial",gamma=0.5,cost=2)

# svm modeli kullanılarak kestirim değerleri hesaplatma
pred.svm=predict(sonuc.svm,mydata)

# hata kareler ortalamasını hesaplatma
mse.svm[i]=sum((mydata$X25- pred.svm)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
r2.svm[i] = (cor(mydata$X25 , pred.svm))^2


# Random Forest ile model oluşturma

sonuc.rf <- randomForest(X25~.,data=mydata,  mtry=5,
ntree=100, importance=TRUE,na.action=na.omit)

# RF modeli kullanılarak kestirim değerleri hesaplatma
pred.rf=predict(sonuc.rf, mydata,predict.all=TRUE)

# hata kareler ortalamasını hesaplatma
mse.rf[i]=sum((mydata$X25 - pred.rf$aggregate)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
,
r2.rf[i] = (cor(mydata$X25 , pred.rf$aggregate))^2


# Regresyon ağacı ile model oluşturma

sonuc.rt=rpart(X25~.,data=mydata)
```

```
# Regresyon ağacı modeli kullanılarak kestirim değerleri
hesaplatma
pred.rt=predict(sonuc.rt, mydata)

# hata kareler ortalamasını hesaplatma
mse.rt[i]=sum((mydata$X25 - pred.rt)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
r2.rt[i] = (cor(mydata$X25 , pred.rt))^2

}

mse.svm.ort=mean(mse.svm)
r2.svm.ort=mean(r2.svm)

mse.rf.ort=mean(mse.rf)
r2.rf.ort=mean(r2.rf)

mse.rt.ort=mean(mse.rt)
r2.rt.ort=mean(r2.rt)

sonuc=list(MSE.SVM=mse.svm, MSE.SVM.ortalama=mse.svm.ort,
R2.SVM=r2.svm, R2.SVM.ortalama=r2.svm.ort,
 MSE.RF=mse.rf,MSE.RF.ortalama=mse.rf.ort, R2.RF=r2.rf,
R2.RF.ortalama=r2.rf.ort,
 MSE.RT=mse.rt, MSE.RT.ortalama=mse.rt.ort, R2.RT=r2.rt,
R2.RT.ortalama=r2.rt.ort)

return(sonuc)
}


yetiskin=tez.fonksiyon(113, nbin=22, mu=mean, Sigma=y.kor,
num.sim=1000)
boxplot(c(yetiskin$R2.RT,yetiskin$R2.RF,
yetiskin$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))



#tüm veri

t.kor=read.csv2(file.choose(),header=T)
t.kor=make.positive.definite(t.kor)


d=25
mean=rep(0,d)


kesim=c(-0.17,-0.50, 0, 0, 0,1.07,-
0.06,1.47,1.62,0.77,1.47,1.81,-0.89,-1.30,1.04,-0.89,-
0.57,0.40,0.77,-1.41,-1.54,0.37,0.80,0.57,-0.37)
```

```
tez.fonksiyon=function(n, nbin, mu, Sigma, num.sim)
{
# bağımlı değişkenin veride en son sütunda yer alması
gerekiyor
# n: gözlem sayısı
# nbin: nitel değişken sayısı
# mu: ortalama vektörü
# Sigma: korelasyon matrisi
# num.sim: simulasyon sayısı

pred.svm=numeric(n)
mse.svm=numeric(num.sim)
r2.svm=numeric(num.sim)

pred.rf=numeric(n)
mse.rf=numeric(num.sim)
r2.rf=numeric(num.sim)


pred.rt=numeric(n)
mse.rt=numeric(num.sim)
r2.rt=numeric(num.sim)


for (i in 1:num.sim)
{

mydata=data.frame(mvrnorm(n,mu,Sigma))


# nitel değişkenleri tanımlama
for (j in 1:nbin) {
     mydata[,j]=cut(mydata[,j], breaks=c(min(mydata[,j])-1,
kesim[j], max(mydata[,j])+1))
     }

#SVM ile model oluşturma, radial kernel kullanılarak
sonuc.svm = svm(X25~.,data=mydata, type="eps-
regression",kernel="radial",gamma=0.5,cost=2)

# svm modeli kullanılarak kestirim değerleri hesaplatma
pred.svm=predict(sonuc.svm,mydata)

# hata kareler ortalamasını hesaplatma
mse.svm[i]=sum((mydata$X25- pred.svm)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
r2.svm[i] = (cor(mydata$X25 , pred.svm))^2


# Random Forest ile model oluşturma
```

```
sonuc.rf <- randomForest(X25~.,data=mydata,  mtry=5,
ntree=100, importance=TRUE,na.action=na.omit)

# RF modeli kullanılarak kestirim değerleri hesaplatma
pred.rf=predict(sonuc.rf, mydata,predict.all=TRUE)

# hata kareler ortalamasını hesaplatma
mse.rf[i]=sum((mydata$X25 - pred.rf$aggregate)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
,
r2.rf[i] = (cor(mydata$X25 , pred.rf$aggregate))^2


# Regresyon ağacı ile model oluşturma

sonuc.rt=rpart(X25~.,data=mydata)

# Regresyon ağacı modeli kullanılarak kestirim değerleri
hesaplatma
pred.rt=predict(sonuc.rt, mydata)

# hata kareler ortalamasını hesaplatma
mse.rt[i]=sum((mydata$X25 - pred.rt)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
r2.rt[i] = (cor(mydata$X25 , pred.rt))^2

}

mse.svm.ort=mean(mse.svm)
r2.svm.ort=mean(r2.svm)

mse.rf.ort=mean(mse.rf)
r2.rf.ort=mean(r2.rf)

mse.rt.ort=mean(mse.rt)
r2.rt.ort=mean(r2.rt)

sonuc=list(MSE.SVM=mse.svm, MSE.SVM.ortalama=mse.svm.ort,
R2.SVM=r2.svm, R2.SVM.ortalama=r2.svm.ort,
 MSE.RF=mse.rf,MSE.RF.ortalama=mse.rf.ort, R2.RF=r2.rf,
R2.RF.ortalama=r2.rf.ort,
 MSE.RT=mse.rt, MSE.RT.ortalama=mse.rt.ort, R2.RT=r2.rt,
R2.RT.ortalama=r2.rt.ort)

return(sonuc)
}


tum=tez.fonksiyon(245, nbin=22, mu=mean, Sigma=t.kor,
num.sim=1000)
```

```
boxplot(c(tum$R2.RT,tum$R2.RF,
tum$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))
```

pred.rf=numeric(n)

```
mse.rf=numeric(num.sim)
r2.rf=numeric(num.sim)
pred.rt=numeric(n)
mse.rt=numeric(num.sim)
r2.rt=numeric(num.sim)
for (i in 1:num.sim)
{

mydata=data.frame(mvrnorm(n,mu=mean,Sigma=y.kor))


for (j in 1:22) {
     mydata[,j]=cut(mydata[,j], breaks=c(min(mydata[,j])-
1,kesim[j],max(mydata[,j])+1))
     }
sonuc.svm = svm(mydata[,25]~.,data=mydata, type="eps-
regression",kernel="radial")
pred.svm=predict(sonuc.svm,mydata)
mse.svm[i]=sum((mydata[,25] - pred.svm)^2)/n
r2.svm[i] = (cor(mydata[,25] , pred.svm))^2

sonuc.rf <- randomForest(mydata[,25]~.,data=mydata,  mtry=5,
ntree=100, importance=TRUE,na.action=na.omit)
pred.rf=predict(sonuc.rf, mydata,predict.all=TRUE)

mse.rf[i]=sum((mydata[,25] - pred.rf$aggregate)^2)/n
r2.rf[i] = (cor(mydata[,25] , pred.rf$aggregate))^2

sonuc.rt=rpart(mydata[,25]~.,data=mydata)
pred.rt=predict(sonuc.rt, mydata)

mse.rt[i]=sum((mydata[,25] - pred.rt)^2)/n r2.rt[i] =
(cor(mydata[,25] , pred.rt))^2

mse.svm.ort=mean(mse.svm)
r2.svm.ort=mean(r2.svm)

mse.rf.ort=mean(mse.rf)
r2.rf.ort=mean(r2.rf)

mse.rt.ort=mean(mse.rt)
r2.rt.ort=mean(r2.rt)

sonuc=list(MSE.SVM=mse.svm, MSE.SVM.ortalama=mse.svm.ort,
R2.SVM=r2.svm, R2.SVM.ortalama=r2.svm.ort,
```

```
MSE.RF=mse.rf,MSE.RF.ortalama=mse.rf.ort, R2.RF=r2.rf,
R2.RF.ortalama=r2.rf.ort,
MSE.RT=mse.rt, MSE.RT.ortalama=mse.rt.ort, R2.RT=r2.rt,
R2.RT.ortalama=r2.rt.ort)
}
return(sonuc)
}

aa=tez.fonksiyon(n, mu, Sigma, num.sim=1000)

boxplot(c(aa$R2.RT,aa$R2.RF,
aa$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))
```

## EK: 3 Korelasyon Yapılarına ve Gözlem Sayılarına Göre Destek Vektör Regresyon, Random Forest ve Regresyon Ağacının Performanslarının Karşılaştırılmasında, R Programında Kullanılan Kodlar.

```
library(corpcor)
library(MASS)
library(e1071)
library(randomForest)
library(rpart)

kesim=c(-0.17,-0.50,1.07,-0.06,1.47,1.62,0.77,1.47,1.81,-
0.89,-1.30,1.04,-0.89,-0.57,0.40,0.77,-1.41,-
1.54,0.37,0.80,0.57,-0.37)

d=25
mean=rep(0,d)

# düşük korelasyon
# r -> (0 - 0.20)

k=runif(d*(d-1)/2, 0, 0.20)

d.kor=diag(d)
d.kor[lower.tri(d.kor)]=k
d.kor[upper.tri(d.kor)]=t(d.kor)[upper.tri(d.kor)]

d.kor=make.positive.definite(d.kor)


# orta korelasyon
# r -> (0.201 - 0.40)

k=runif(d*(d-1)/2, 0.201, 0.40)

o.kor=diag(d)
o.kor[lower.tri(o.kor)]=k
o.kor[upper.tri(o.kor)]=t(o.kor)[upper.tri(o.kor)]

o.kor=make.positive.definite(o.kor)


# yüksek korelasyon
# r -> (0.401 - 0.60)

k=runif(d*(d-1)/2, 0.401, 0.60)

y.kor=diag(d)
y.kor[lower.tri(y.kor)]=k
```

```
y.kor[upper.tri(y.kor)]=t(y.kor)[upper.tri(y.kor)]

y.kor=make.positive.definite(y.kor)


# yukarıda tanımlanan 3 farklı korelasyon yapısı daha once
olusturduğumuz tez.fonksiyon adlı
# fonksiyonda çalıştırıldı. bu fonksiyonda n değeri sırasıyla
100, 2500 ve 1000 olarak alındı.


tez.fonksiyon=function(n, d, nbin, mu, Sigma, num.sim)
{
# bağımlı değişkenin veride en son sütunda yer alması
gerekiyor
# n: gözlem sayısı
# değişken sayısı
# nbin: nitel değişken sayısı
# mu: ortalama vektörü
# Sigma: korelasyon matrisi
# num.sim: simulasyon sayısı

pred.svm=numeric(n)
mse.svm=numeric(num.sim)
r2.svm=numeric(num.sim)

pred.rf=numeric(n)
mse.rf=numeric(num.sim)
r2.rf=numeric(num.sim)

pred.rt=numeric(n)
mse.rt=numeric(num.sim)
r2.rt=numeric(num.sim)


for (i in 1:num.sim)
{

mydata=data.frame(mvrnorm(n,mu,Sigma))


# nitel değişkenleri tanımlama
for (j in 1:nbin) {
     mydata[,j]=cut(mydata[,j], breaks=c(min(mydata[,j])-1,
kesim[j], max(mydata[,j])+1))
     }


#SVM ile model oluşturma, radial kernel kullanılarak

sonuc.svm = svm(X25~.,data=mydata, type="eps-
regression",kernel="radial",gamma=0.5,cost=2)
```

```
# svm modeli kullanılarak kestirim değerleri hesaplatma
pred.svm=predict(sonuc.svm,mydata)

# hata kareler ortalamasını hesaplatma
mse.svm[i]=sum((mydata$X25- pred.svm)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma

r2.svm[i] = (cor(mydata$X25 , pred.svm))^2


# Random Forest ile model oluşturma

sonuc.rf <- randomForest(X25~.,data=mydata,  mtry=5,
ntree=100, importance=TRUE,na.action=na.omit)

# RF modeli kullanılarak kestirim değerleri hesaplatma
pred.rf=predict(sonuc.rf, mydata,predict.all=TRUE)

# hata kareler ortalamasını hesaplatma
mse.rf[i]=sum((mydata$X25 - pred.rf$aggregate)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
r2.rf[i] = (cor(mydata$X25 , pred.rf$aggregate))^2


# Regresyon ağacı ile model oluşturma

sonuc.rt=rpart(X25~.,data=mydata)

# Regresyon ağacı modeli kullanılarak kestirim değerleri
hesaplatma
pred.rt=predict(sonuc.rt, mydata)

# hata kareler ortalamasını hesaplatma
mse.rt[i]=sum((mydata$X25 - pred.rt)^2)/n

# model uyumunu değerlendirmek için R-kare değerini hesaplatma
r2.rt[i] = (cor(mydata$X25 , pred.rt))^2


}

mse.svm.ort=mean(mse.svm)
r2.svm.ort=mean(r2.svm)

mse.rf.ort=mean(mse.rf)
r2.rf.ort=mean(r2.rf)

mse.rt.ort=mean(mse.rt)
r2.rt.ort=mean(r2.rt)
```

```
sonuc=list(MSE.SVM=mse.svm, MSE.SVM.ortalama=mse.svm.ort,
R2.SVM=r2.svm, R2.SVM.ortalama=r2.svm.ort,
 MSE.RF=mse.rf,MSE.RF.ortalama=mse.rf.ort, R2.RF=r2.rf,
R2.RF.ortalama=r2.rf.ort,
 MSE.RT=mse.rt, MSE.RT.ortalama=mse.rt.ort, R2.RT=r2.rt,
R2.RT.ortalama=r2.rt.ort)

return(sonuc)
}


# Düşük, orta ve yüksek korelasyon düzeyi ile n=100,250 ve
1000 gözlem sayısı için simülasyon analizi işlemleri

# düşük korelasyon, n=100


aa100d=tez.fonksiyon(n=100, nbin=22, mu=mean, Sigma=d.kor,
num.sim=1000)
boxplot(c(aa100d$R2.RT,aa100d$R2.RF,
aa100d$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa100d$MSE.RT.ortalama
aa100d$MSE.RF.ortalama
aa100d$MSE.SVM.ortalama

aa100d$R2.RT.ortalama
aa100d$R2.RF.ortalama
aa100d$R2.SVM.ortalama

fivenum(aa100d$MSE.RT)
fivenum(aa100d$MSE.RF)
fivenum(aa100d$MSE.SVM)
fivenum(aa100d$R2.RT)
fivenum(aa100d$R2.RF)
fivenum(aa100d$R2.SVM)


# n=250
aa250d=tez.fonksiyon(n=250, nbin=22, mu=mean, Sigma=d.kor,
num.sim=1000)
boxplot(c(aa250d$R2.RT,aa250d$R2.RF,
aa250d$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa250d$MSE.RT.ortalama
aa250d$MSE.RF.ortalama
aa250d$MSE.SVM.ortalama

aa250d$R2.RT.ortalama
aa250d$R2.RF.ortalama
aa250d$R2.SVM.ortalama

fivenum(aa250d$MSE.RT)
```

```
fivenum(aa250d$MSE.RF)
fivenum(aa250d$MSE.SVM)
fivenum(aa250d$R2.RT)
fivenum(aa250d$R2.RF)
fivenum(aa250d$R2.SVM)


# n=1000
aa1000d=tez.fonksiyon(n=1000, nbin=22, mu=mean, Sigma=d.kor,
num.sim=1000)
boxplot(c(aa1000d$R2.RT,aa1000d$R2.RF,
aa1000d$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa1000d$MSE.RT.ortalama
aa1000d$MSE.RF.ortalama
aa1000d$MSE.SVM.ortalama

aa1000d$R2.RT.ortalama
aa1000d$R2.RF.ortalama
aa1000d$R2.SVM.ortalama

fivenum(aa1000d$MSE.RT)
fivenum(aa1000d$MSE.RF)
fivenum(aa1000d$MSE.SVM)
fivenum(aa1000d$R2.RT)
fivenum(aa1000d$R2.RF)
fivenum(aa1000d$R2.SVM)



# orta korelasyon, n=10

aa100o=tez.fonksiyon(n=100, nbin=22, mu=mean, Sigma=y.kor,
num.sim=1000)
boxplot(c(aa100o$R2.RT,aa100o$R2.RF,
aa100o$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa100o$MSE.RT.ortalama
aa100o$MSE.RF.ortalama
aa100o$MSE.SVM.ortalama

aa100o$R2.RT.ortalama
aa100o$R2.RF.ortalama
aa100o$R2.SVM.ortalama

fivenum(aa100o$MSE.RT)
fivenum(aa100o$MSE.RF)
fivenum(aa100o$MSE.SVM)
fivenum(aa100o$R2.RT)
fivenum(aa100o$R2.RF)
fivenum(aa100o$R2.SVM)
```

```
# n=250
aa250o=tez.fonksiyon(n=250, nbin=22, mu=mean, Sigma=y.kor,
num.sim=1000)
boxplot(c(aa250o$R2.RT,aa250o$R2.RF,
aa250o$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa250o$MSE.RT.ortalama
aa250o$MSE.RF.ortalama
aa250o$MSE.SVM.ortalama

aa250o$R2.RT.ortalama
aa250o$R2.RF.ortalama
aa250o$R2.SVM.ortalama

fivenum(aa250o$MSE.RT)
fivenum(aa250o$MSE.RF)
fivenum(aa250o$MSE.SVM)
fivenum(aa250o$R2.RT)
fivenum(aa250o$R2.RF)
fivenum(aa250o$R2.SVM)


# n=1000
aa1000o=tez.fonksiyon(n=1000, nbin=22, mu=mean, Sigma=y.kor,
num.sim=1000)
boxplot(c(aa1000o$R2.RT,aa1000o$R2.RF,
aa1000o$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa1000o$MSE.RT.ortalama
aa1000o$MSE.RF.ortalama
aa1000o$MSE.SVM.ortalama

aa1000o$R2.RT.ortalama
aa1000o$R2.RF.ortalama
aa1000o$R2.SVM.ortalama

fivenum(aa1000o$MSE.RT)
fivenum(aa1000o$MSE.RF)
fivenum(aa1000o$MSE.SVM)
fivenum(aa1000o$R2.RT)
fivenum(aa1000o$R2.RF)
fivenum(aa1000o$R2.SVM)



# yüksek korelasyon, n=100


aa100y=tez.fonksiyon(n=100, nbin=22, mu=mean, Sigma=y.kor,
num.sim=1000)
```

```
boxplot(c(aa100y$R2.RT,aa100y$R2.RF,
aa100y$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa100y$MSE.RT.ortalama
aa100y$MSE.RF.ortalama
aa100y$MSE.SVM.ortalama

aa100y$R2.RT.ortalama
aa100y$R2.RF.ortalama
aa100y$R2.SVM.ortalama

fivenum(aa100y$MSE.RT)
fivenum(aa100y$MSE.RF)
fivenum(aa100y$MSE.SVM)
fivenum(aa100y$R2.RT)
fivenum(aa100y$R2.RF)
fivenum(aa100y$R2.SVM)



# n=250
aa250y=tez.fonksiyon(n=250, nbin=22, mu=mean, Sigma=y.kor,
num.sim=1000)
boxplot(c(aa250y$R2.RT,aa250y$R2.RF,
aa250y$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa250y$MSE.RT.ortalama
aa250y$MSE.RF.ortalama
aa250y$MSE.SVM.ortalama

aa250y$R2.RT.ortalama
aa250y$R2.RF.ortalama
aa250y$R2.SVM.ortalama

fivenum(aa250y$MSE.RT)
fivenum(aa250y$MSE.RF)
fivenum(aa250y$MSE.SVM)
fivenum(aa250y$R2.RT)
fivenum(aa250y$R2.RF)
fivenum(aa250y$R2.SVM)



# n=1000

aa1000y=tez.fonksiyon(n=1000, nbin=22, mu=mean, Sigma=y.kor,
num.sim=1000)
boxplot(c(aa1000y$R2.RT,aa1000y$R2.RF,
aa1000y$R2.SVM)~rep(c("RT","RF","SVM"),each=1000))

aa1000y$MSE.RT.ortalama
aa1000y$MSE.RF.ortalama
```

```
aa1000y$MSE.SVM.ortalama

aa1000y$R2.RT.ortalama
aa1000y$R2.RF.ortalama
aa1000y$R2.SVM.ortalama

fivenum(aa1000y$MSE.RT)
fivenum(aa1000y$MSE.RF)
fivenum(aa1000y$MSE.SVM)
fivenum(aa1000y$R2.RT)
fivenum(aa1000y$R2.RF)
fivenum(aa1000y$R2.SVM)
```

## EK: 4 R Pogramında Sınıflamanın Yapılışı

```
library(e1071)
library(rpart)
library(randomForest)
for (i in 1:23) y.veri[,i]=factor(y.veri[,i])
veri$YATCIK_kod=cut(veri$YATCIK_T,breaks=c(-1,9,max(veri[,26])))
veri=veri[,c(-4,-26)]
sonuc.svm = svm(YATCIK_kod~.,data=veri, type="C-classification",kernel="radial")
pred.svm=predict(sonuc.svm,veri)
sonuc.rf <- randomForest(YATCIK_kod~.,data=veri, mtry=5, ntree=100, importance=TRUE,na.action=na.omit)
pred.rf=predict(sonuc.rf, veri[,-25],predict.all=TRUE)
sonuc.rt=rpart(YATCIK_kod~.,data=veri)
pred.rt=predict(sonuc.rt, veri)
pred.rt=cut(pred.rt[,2],breaks=c(-1,0.50,1),labels=c(0,1))
table(pred.svm,veri[,25])
table(pred.rf$aggregate,veri[,25])
table(pred.rt[,2],veri[,25])
> pred.rt=predict(sonuc.rt, veri)
> pred.rt=cut(pred.rt[,2],breaks=c(-1,0.50,1),labels=c(0,1))
> table(pred.rt,veri[,25])
pred.rt (-1,9] (9,30]
      0     75     15
      1     44    111
> sonuc.rt
n= 245
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 245 119 (9,30] (0.4857143 0.5142857)
   2) yas>=55.5 44  14 (-1,9] (0.6818182 0.3181818)
     4) semptgunsayisi< 1.5 12   0 (-1,9] (1.0000000 0.0000000) *
     5) semptgunsayisi>=1.5 32  14 (-1,9] (0.5625000 0.4375000)
```

10) semptgunsayisi>=6.5 13 2 (-1,9] (0.8461538 0.1538462) *

11) semptgunsayisi< 6.5 19 7 (9,30] (0.3684211 0.6315789) *

3) yas< 55.5 201 89 (9,30] (0.4427861 0.5572139)

6) yas< 16.5 128 57 (-1,9] (0.5546875 0.4453125)

12) yas>=12.5 65 21 (-1,9] (0.6769231 0.3230769)

24) semptgunsayisi>=4.5 39 7 (-1,9] (0.8205128 0.1794872) *

25) semptgunsayisi< 4.5 26 12 (9,30] (0.4615385 0.5384615)

50) ISHAL=1 12 4 (-1,9] (0.6666667 0.3333333) *

51) ISHAL=0 14 4 (9,30] (0.2857143 0.7142857) *

13) yas< 12.5 63 27 (9,30] (0.4285714 0.5714286)

26) KONJONKTIVIT=1 27 10 (-1,9] (0.6296296 0.3703704)

52) yas< 9.5 14 2 (-1,9] (0.8571429 0.1428571) *

53) yas>=9.5 13 5 (9,30] (0.3846154 0.6153846) *

27) KONJONKTIVIT=0 36 10 (9,30] (0.2777778 0.7222222) *

7) yas>=16.5 73 18 (9,30] (0.2465753 0.7534247) *

table(pred.svm,veri[,25])

pred.svm (-1,9] (9,30]

(-1,9] 94 39

(9,30] 25 87

> sonuc.svm

Call:

svm(formula = YATCIK_kod ~ ., data = veri, type = "C-classification", kernel = "radial")

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 1

gamma: 0.04

Number of Support Vectors: 222

> sonuc.rf

Call:

randomForest(formula = YATCIK_kod ~ ., data = veri, mtry = 5, ntree = 100, importance = TRUE, na.action = na.omit)

Type of random forest: classification

```
                Number of trees: 100
No. of variables tried at each split: 5
      OOB estimate of  error rate: 37.96%
Confusion matrix:
      (-1,9] (9,30] class.error
(-1,9]     73     46   0.3865546
(9,30]     47     79   0.3730159
> pred.rf=predict(sonuc.rf, veri[,-25],predict.all=TRUE)
> table(pred.rf$aggregate,veri[,25])
```

## EK: 5 Tüm Veri ile Değişken Seçiminde Göre Kullanılan Kodlar

```
tum.data=read.csv2(file.choose(),header=TRUE)

sonuc.rft <- randomForest(YATCIK_T ~ ., data=tum.data,
mtry=5, ntree=100, importance=TRUE,na.action=na.omit)

# değişken önmelilikleri ile ilgili sonuçların elde edilmesi
importance(sonuc.rft,type=1)
# değişken önmeliliklerinin grafiksel gösterimi
varImpPlot(sonuc.rft,type=1)


# Yetişkin verisi üzerinde değişken seçimi

yetiskin.data=read.csv2(file.choose(),header=TRUE)

sonuc.rfy <- randomForest(YATCIK_T ~ ., data=yetiskin.data,
mtry=5, ntree=100, importance=TRUE,na.action=na.omit)

# değişken önmelilikleri ile ilgili sonuçların elde edilmesi
importance(sonuc.rfy,type=1)
# değişken önmeliliklerinin grafiksel gösterimi
varImpPlot(sonuc.rfy,type=1)



# Çocuk verisi üzerinde değişken seçimi

cocuk.data=read.csv2(file.choose(),header=TRUE)

sonuc.rfc <- randomForest(YATCIK_T ~ ., data=cocuk.data,
mtry=5, ntree=100, importance=TRUE,na.action=na.omit)

# değişken önmelilikleri ile ilgili sonuçların elde edilmesi
importance(sonuc.rfc,type=1)
# değişken önmeliliklerinin grafiksel gösterimi
varImpPlot(sonuc.rfc,type=1)
```

## EK:6 (r=0,00-0,20) Düşük Düzey Korelasyon Matrisi

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1,00 | 0,03 | 0,05 | 0,16 | 0,14 | 0,05 | 0,10 | 0,20 | 0,12 | 0,18 | 0,00 | 0,01 | 0,03 | 0,14 | 0,19 | 0,05 | 0,08 | 0,11 | 0,19 | 0,13 | 0,15 | 0,09 | 0,16 | 0,11 | 0,12 |
| 0,03 | 1,00 | 0,05 | 0,19 | 0,09 | 0,02 | 0,15 | 0,18 | 0,15 | 0,10 | 0,05 | 0,18 | 0,08 | 0,19 | 0,00 | 0,14 | 0,06 | 0,18 | 0,13 | 0,07 | 0,15 | 0,17 | 0,03 | 0,15 | 0,10 |
| 0,05 | 0,05 | 1,00 | 0,05 | 0,01 | 0,04 | 0,14 | 0,00 | 0,19 | 0,07 | 0,13 | 0,11 | 0,15 | 0,03 | 0,05 | 0,09 | 0,15 | 0,11 | 0,14 | 0,18 | 0,02 | 0,19 | 0,17 | 0,14 | 0,03 |
| 0,16 | 0,19 | 0,05 | 1,00 | 0,18 | 0,10 | 0,00 | 0,01 | 0,01 | 0,14 | 0,00 | 0,06 | 0,13 | 0,18 | 0,14 | 0,12 | 0,10 | 0,06 | 0,17 | 0,15 | 0,18 | 0,05 | 0,18 | 0,05 | 0,09 |
| 0,14 | 0,09 | 0,01 | 0,18 | 1,00 | 0,13 | 0,09 | 0,08 | 0,02 | 0,13 | 0,06 | 0,15 | 0,11 | 0,15 | 0,13 | 0,13 | 0,01 | 0,18 | 0,15 | 0,04 | 0,14 | 0,17 | 0,17 | 0,15 | 0,05 |
| 0,05 | 0,02 | 0,04 | 0,10 | 0,13 | 1,00 | 0,04 | 0,15 | 0,00 | 0,15 | 0,08 | 0,14 | 0,02 | 0,10 | 0,19 | 0,07 | 0,20 | 0,12 | 0,17 | 0,06 | 0,04 | 0,17 | 0,06 | 0,19 | 0,13 |
| 0,10 | 0,15 | 0,14 | 0,00 | 0,09 | 0,04 | 1,00 | 0,18 | 0,09 | 0,00 | 0,18 | 0,03 | 0,08 | 0,09 | 0,10 | 0,15 | 0,02 | 0,20 | 0,19 | 0,03 | 0,18 | 0,18 | 0,12 | 0,04 | 0,13 |
| 0,20 | 0,18 | 0,00 | 0,01 | 0,08 | 0,15 | 0,18 | 1,00 | 0,04 | 0,02 | 0,17 | 0,05 | 0,13 | 0,13 | 0,15 | 0,18 | 0,07 | 0,10 | 0,18 | 0,16 | 0,02 | 0,20 | 0,18 | 0,07 | 0,10 |
| 0,12 | 0,15 | 0,19 | 0,01 | 0,02 | 0,00 | 0,09 | 0,04 | 1,00 | 0,19 | 0,13 | 0,05 | 0,12 | 0,12 | 0,09 | 0,02 | 0,17 | 0,17 | 0,04 | 0,03 | 0,20 | 0,17 | 0,12 | 0,09 | 0,15 |
| 0,18 | 0,10 | 0,07 | 0,14 | 0,13 | 0,15 | 0,00 | 0,02 | 0,19 | 1,00 | 0,10 | 0,17 | 0,19 | 0,19 | 0,14 | 0,10 | 0,18 | 0,04 | 0,02 | 0,14 | 0,04 | 0,05 | 0,12 | 0,11 | 0,16 |
| 0,00 | 0,05 | 0,13 | 0,00 | 0,06 | 0,08 | 0,18 | 0,17 | 0,13 | 0,10 | 1,00 | 0,04 | 0,02 | 0,05 | 0,18 | 0,04 | 0,18 | 0,09 | 0,02 | 0,01 | 0,18 | 0,02 | 0,11 | 0,05 | 0,15 |
| 0,01 | 0,18 | 0,11 | 0,06 | 0,15 | 0,14 | 0,03 | 0,05 | 0,05 | 0,17 | 0,04 | 1,00 | 0,18 | 0,01 | 0,05 | 0,11 | 0,17 | 0,11 | 0,07 | 0,14 | 0,01 | 0,06 | 0,15 | 0,14 | 0,16 |
| 0,03 | 0,08 | 0,15 | 0,13 | 0,11 | 0,02 | 0,08 | 0,13 | 0,12 | 0,19 | 0,02 | 0,18 | 1,00 | 0,07 | 0,13 | 0,14 | 0,18 | 0,01 | 0,13 | 0,06 | 0,19 | 0,14 | 0,19 | 0,11 | 0,19 |
| 0,14 | 0,19 | 0,03 | 0,18 | 0,15 | 0,10 | 0,09 | 0,13 | 0,12 | 0,19 | 0,05 | 0,01 | 0,07 | 1,00 | 0,07 | 0,09 | 0,03 | 0,17 | 0,13 | 0,20 | 0,02 | 0,09 | 0,14 | 0,11 | 0,01 |
| 0,19 | 0,00 | 0,05 | 0,14 | 0,13 | 0,19 | 0,10 | 0,15 | 0,09 | 0,14 | 0,18 | 0,05 | 0,13 | 0,07 | 1,00 | 0,18 | 0,09 | 0,06 | 0,12 | 0,13 | 0,11 | 0,11 | 0,18 | 0,06 | 0,12 |
| 0,05 | 0,14 | 0,09 | 0,12 | 0,13 | 0,07 | 0,15 | 0,18 | 0,02 | 0,10 | 0,04 | 0,11 | 0,14 | 0,09 | 0,18 | 1,00 | 0,19 | 0,10 | 0,13 | 0,14 | 0,14 | 0,08 | 0,13 | 0,02 | 0,13 |
| 0,08 | 0,06 | 0,15 | 0,10 | 0,01 | 0,20 | 0,02 | 0,07 | 0,17 | 0,18 | 0,18 | 0,17 | 0,18 | 0,03 | 0,09 | 0,19 | 1,00 | 0,14 | 0,04 | 0,12 | 0,05 | 0,05 | 0,15 | 0,14 | 0,04 |
| 0,11 | 0,18 | 0,11 | 0,06 | 0,18 | 0,12 | 0,20 | 0,10 | 0,17 | 0,04 | 0,09 | 0,11 | 0,01 | 0,17 | 0,06 | 0,10 | 0,14 | 1,00 | 0,07 | 0,19 | 0,05 | 0,05 | 0,16 | 0,11 | 0,06 |
| 0,19 | 0,13 | 0,14 | 0,17 | 0,15 | 0,17 | 0,19 | 0,18 | 0,04 | 0,02 | 0,02 | 0,07 | 0,13 | 0,13 | 0,12 | 0,13 | 0,04 | 0,07 | 1,00 | 0,08 | 0,06 | 0,04 | 0,14 | 0,13 | 0,13 |
| 0,13 | 0,07 | 0,18 | 0,15 | 0,04 | 0,06 | 0,03 | 0,16 | 0,03 | 0,14 | 0,01 | 0,14 | 0,06 | 0,20 | 0,13 | 0,14 | 0,12 | 0,19 | 0,08 | 1,00 | 0,09 | 0,12 | 0,11 | 0,12 | 0,01 |
| 0,15 | 0,15 | 0,02 | 0,18 | 0,14 | 0,04 | 0,18 | 0,02 | 0,20 | 0,04 | 0,18 | 0,01 | 0,19 | 0,02 | 0,11 | 0,14 | 0,05 | 0,05 | 0,06 | 0,09 | 1,00 | 0,20 | 0,16 | 0,15 | 0,09 |
| 0,09 | 0,17 | 0,19 | 0,05 | 0,17 | 0,17 | 0,18 | 0,20 | 0,17 | 0,05 | 0,02 | 0,06 | 0,14 | 0,09 | 0,11 | 0,08 | 0,05 | 0,05 | 0,04 | 0,12 | 0,20 | 1,00 | 0,00 | 0,16 | 0,11 |
| 0,16 | 0,03 | 0,17 | 0,18 | 0,17 | 0,06 | 0,12 | 0,18 | 0,12 | 0,12 | 0,11 | 0,15 | 0,19 | 0,14 | 0,18 | 0,13 | 0,15 | 0,16 | 0,14 | 0,11 | 0,16 | 0,00 | 1,01 | 0,18 | 0,01 |
| 0,11 | 0,15 | 0,14 | 0,05 | 0,15 | 0,19 | 0,04 | 0,07 | 0,09 | 0,11 | 0,05 | 0,14 | 0,11 | 0,11 | 0,06 | 0,02 | 0,14 | 0,11 | 0,13 | 0,12 | 0,15 | 0,16 | 0,18 | 1,00 | 0,04 |
| 0,12 | 0,10 | 0,03 | 0,09 | 0,05 | 0,13 | 0,13 | 0,10 | 0,15 | 0,16 | 0,15 | 0,16 | 0,19 | 0,01 | 0,12 | 0,13 | 0,04 | 0,06 | 0,13 | 0,01 | 0,09 | 0,11 | 0,01 | 0,04 | 1,00 |

## EK:7 (r=0,21-0,40) OrtaDüzey Korelasyon Matrisi

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1,00 | 0,22 | 0,38 | 0,23 | 0,30 | 0,27 | 0,32 | 0,29 | 0,33 | 0,26 | 0,27 | 0,29 | 0,36 | 0,28 | 0,26 | 0,32 | 0,39 | 0,32 | 0,39 | 0,27 | 0,30 | 0,35 | 0,24 | 0,36 | 0,30 |
| 0,22 | 1,00 | 0,25 | 0,38 | 0,33 | 0,28 | 0,37 | 0,37 | 0,35 | 0,35 | 0,36 | 0,39 | 0,31 | 0,38 | 0,31 | 0,40 | 0,32 | 0,26 | 0,33 | 0,36 | 0,24 | 0,34 | 0,38 | 0,37 | 0,36 |
| 0,38 | 0,25 | 1,00 | 0,26 | 0,23 | 0,22 | 0,39 | 0,35 | 0,30 | 0,39 | 0,26 | 0,38 | 0,24 | 0,37 | 0,30 | 0,30 | 0,34 | 0,24 | 0,21 | 0,33 | 0,33 | 0,36 | 0,38 | 0,21 | 0,30 |
| 0,23 | 0,38 | 0,26 | 1,00 | 0,32 | 0,28 | 0,37 | 0,26 | 0,23 | 0,28 | 0,31 | 0,35 | 0,21 | 0,28 | 0,35 | 0,32 | 0,24 | 0,34 | 0,23 | 0,33 | 0,33 | 0,30 | 0,35 | 0,31 | 0,24 |
| 0,30 | 0,33 | 0,23 | 0,32 | 1,00 | 0,21 | 0,24 | 0,39 | 0,31 | 0,38 | 0,38 | 0,33 | 0,34 | 0,30 | 0,34 | 0,27 | 0,25 | 0,39 | 0,30 | 0,38 | 0,38 | 0,33 | 0,31 | 0,36 | 0,24 |
| 0,27 | 0,28 | 0,22 | 0,28 | 0,21 | 1,00 | 0,36 | 0,35 | 0,38 | 0,31 | 0,36 | 0,32 | 0,37 | 0,29 | 0,33 | 0,26 | 0,25 | 0,29 | 0,33 | 0,25 | 0,39 | 0,34 | 0,26 | 0,25 | 0,36 |
| 0,32 | 0,37 | 0,39 | 0,37 | 0,24 | 0,36 | 1,00 | 0,37 | 0,28 | 0,26 | 0,22 | 0,35 | 0,31 | 0,37 | 0,38 | 0,38 | 0,38 | 0,35 | 0,25 | 0,30 | 0,37 | 0,32 | 0,36 | 0,34 | 0,30 |
| 0,29 | 0,37 | 0,35 | 0,26 | 0,39 | 0,35 | 0,37 | 1,00 | 0,38 | 0,28 | 0,29 | 0,29 | 0,26 | 0,29 | 0,22 | 0,27 | 0,33 | 0,22 | 0,22 | 0,34 | 0,39 | 0,22 | 0,33 | 0,31 | 0,35 |
| 0,33 | 0,35 | 0,30 | 0,23 | 0,31 | 0,38 | 0,28 | 0,38 | 1,00 | 0,25 | 0,21 | 0,24 | 0,34 | 0,33 | 0,30 | 0,40 | 0,25 | 0,35 | 0,21 | 0,23 | 0,29 | 0,33 | 0,39 | 0,37 | 0,32 |
| 0,26 | 0,35 | 0,39 | 0,28 | 0,38 | 0,31 | 0,26 | 0,28 | 0,25 | 1,00 | 0,21 | 0,33 | 0,34 | 0,27 | 0,37 | 0,24 | 0,39 | 0,24 | 0,36 | 0,32 | 0,36 | 0,33 | 0,38 | 0,25 | 0,22 |
| 0,27 | 0,36 | 0,26 | 0,31 | 0,38 | 0,36 | 0,22 | 0,29 | 0,21 | 0,21 | 1,00 | 0,34 | 0,25 | 0,23 | 0,22 | 0,23 | 0,31 | 0,34 | 0,34 | 0,35 | 0,27 | 0,22 | 0,32 | 0,39 | 0,25 |
| 0,29 | 0,39 | 0,38 | 0,35 | 0,33 | 0,32 | 0,35 | 0,29 | 0,24 | 0,33 | 0,34 | 1,00 | 0,25 | 0,28 | 0,24 | 0,30 | 0,23 | 0,32 | 0,21 | 0,24 | 0,30 | 0,37 | 0,28 | 0,21 | 0,34 |
| 0,36 | 0,31 | 0,24 | 0,21 | 0,34 | 0,37 | 0,31 | 0,26 | 0,34 | 0,34 | 0,25 | 0,25 | 1,00 | 0,40 | 0,30 | 0,32 | 0,22 | 0,32 | 0,32 | 0,23 | 0,34 | 0,38 | 0,32 | 0,22 | 0,39 |
| 0,28 | 0,38 | 0,37 | 0,28 | 0,30 | 0,29 | 0,37 | 0,29 | 0,33 | 0,27 | 0,23 | 0,28 | 0,40 | 1,00 | 0,27 | 0,39 | 0,36 | 0,22 | 0,40 | 0,39 | 0,38 | 0,38 | 0,22 | 0,23 | 0,32 |
| 0,26 | 0,31 | 0,30 | 0,35 | 0,34 | 0,33 | 0,38 | 0,22 | 0,30 | 0,37 | 0,22 | 0,24 | 0,30 | 0,27 | 1,00 | 0,39 | 0,34 | 0,32 | 0,24 | 0,29 | 0,28 | 0,32 | 0,22 | 0,38 | 0,35 |
| 0,32 | 0,40 | 0,30 | 0,32 | 0,27 | 0,26 | 0,38 | 0,27 | 0,40 | 0,24 | 0,23 | 0,30 | 0,32 | 0,39 | 0,39 | 1,00 | 0,24 | 0,33 | 0,37 | 0,23 | 0,33 | 0,37 | 0,24 | 0,28 | 0,22 |
| 0,39 | 0,32 | 0,34 | 0,24 | 0,25 | 0,25 | 0,38 | 0,33 | 0,25 | 0,39 | 0,31 | 0,23 | 0,22 | 0,36 | 0,34 | 0,24 | 1,00 | 0,22 | 0,40 | 0,33 | 0,26 | 0,39 | 0,28 | 0,37 | 0,35 |
| 0,32 | 0,26 | 0,24 | 0,34 | 0,39 | 0,29 | 0,35 | 0,22 | 0,35 | 0,24 | 0,34 | 0,32 | 0,32 | 0,22 | 0,32 | 0,33 | 0,22 | 1,00 | 0,24 | 0,24 | 0,23 | 0,37 | 0,22 | 0,37 | 0,27 |
| 0,39 | 0,33 | 0,21 | 0,23 | 0,30 | 0,33 | 0,25 | 0,22 | 0,21 | 0,36 | 0,34 | 0,21 | 0,32 | 0,40 | 0,24 | 0,37 | 0,40 | 0,24 | 1,00 | 0,36 | 0,24 | 0,32 | 0,32 | 0,30 | 0,37 |
| 0,27 | 0,36 | 0,33 | 0,33 | 0,38 | 0,25 | 0,30 | 0,34 | 0,23 | 0,32 | 0,35 | 0,24 | 0,23 | 0,39 | 0,29 | 0,23 | 0,33 | 0,24 | 0,36 | 1,00 | 0,39 | 0,34 | 0,32 | 0,35 | 0,33 |
| 0,30 | 0,24 | 0,33 | 0,33 | 0,38 | 0,39 | 0,37 | 0,39 | 0,29 | 0,36 | 0,27 | 0,30 | 0,34 | 0,38 | 0,28 | 0,33 | 0,26 | 0,23 | 0,24 | 0,39 | 1,00 | 0,31 | 0,29 | 0,38 | 0,30 |
| 0,35 | 0,34 | 0,36 | 0,30 | 0,33 | 0,34 | 0,32 | 0,22 | 0,33 | 0,33 | 0,22 | 0,37 | 0,38 | 0,38 | 0,32 | 0,37 | 0,39 | 0,37 | 0,32 | 0,34 | 0,31 | 1,00 | 0,28 | 0,28 | 0,25 |
| 0,24 | 0,38 | 0,38 | 0,35 | 0,31 | 0,26 | 0,36 | 0,33 | 0,39 | 0,38 | 0,32 | 0,28 | 0,32 | 0,22 | 0,22 | 0,24 | 0,28 | 0,22 | 0,32 | 0,32 | 0,29 | 0,28 | 1,00 | 0,31 | 0,34 |
| 0,36 | 0,37 | 0,21 | 0,31 | 0,36 | 0,25 | 0,34 | 0,31 | 0,37 | 0,25 | 0,39 | 0,21 | 0,22 | 0,23 | 0,38 | 0,28 | 0,37 | 0,37 | 0,30 | 0,35 | 0,38 | 0,28 | 0,31 | 1,00 | 0,36 |
| 0,30 | 0,36 | 0,30 | 0,24 | 0,24 | 0,36 | 0,30 | 0,35 | 0,32 | 0,22 | 0,25 | 0,34 | 0,39 | 0,32 | 0,35 | 0,22 | 0,35 | 0,27 | 0,37 | 0,33 | 0,30 | 0,25 | 0,34 | 0,36 | 1,00 |

## EK:8 (r=0,41-0,60) Yüksek Düzey Korelasyon Matrisi

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1,00 | 0,54 | 0,50 | 0,53 | 0,60 | 0,52 | 0,44 | 0,57 | 0,48 | 0,51 | 0,55 | 0,45 | 0,52 | 0,57 | 0,54 | 0,57 | 0,44 | 0,58 | 0,43 | 0,57 | 0,55 | 0,59 | 0,58 | 0,50 | 0,55 |
| 0,54 | 1,00 | 0,57 | 0,42 | 0,54 | 0,59 | 0,59 | 0,54 | 0,49 | 0,50 | 0,45 | 0,41 | 0,48 | 0,58 | 0,48 | 0,47 | 0,47 | 0,55 | 0,50 | 0,47 | 0,54 | 0,51 | 0,46 | 0,50 | 0,58 |
| 0,50 | 0,57 | 1,00 | 0,49 | 0,57 | 0,41 | 0,54 | 0,46 | 0,57 | 0,53 | 0,51 | 0,44 | 0,53 | 0,51 | 0,57 | 0,56 | 0,49 | 0,43 | 0,50 | 0,44 | 0,52 | 0,58 | 0,43 | 0,43 | 0,48 |
| 0,53 | 0,42 | 0,49 | 1,00 | 0,53 | 0,51 | 0,57 | 0,45 | 0,50 | 0,52 | 0,54 | 0,55 | 0,58 | 0,47 | 0,51 | 0,58 | 0,47 | 0,59 | 0,42 | 0,51 | 0,44 | 0,46 | 0,46 | 0,53 | 0,51 |
| 0,60 | 0,54 | 0,57 | 0,53 | 1,00 | 0,51 | 0,50 | 0,41 | 0,57 | 0,48 | 0,57 | 0,54 | 0,44 | 0,46 | 0,55 | 0,59 | 0,51 | 0,52 | 0,46 | 0,58 | 0,51 | 0,41 | 0,51 | 0,57 | 0,48 |
| 0,52 | 0,59 | 0,41 | 0,51 | 0,51 | 1,00 | 0,54 | 0,58 | 0,44 | 0,58 | 0,51 | 0,58 | 0,59 | 0,58 | 0,47 | 0,57 | 0,52 | 0,57 | 0,57 | 0,44 | 0,41 | 0,59 | 0,49 | 0,41 | 0,58 |
| 0,44 | 0,59 | 0,54 | 0,57 | 0,50 | 0,54 | 1,00 | 0,54 | 0,48 | 0,49 | 0,42 | 0,43 | 0,44 | 0,43 | 0,43 | 0,43 | 0,48 | 0,43 | 0,53 | 0,54 | 0,57 | 0,58 | 0,55 | 0,48 | 0,44 |
| 0,57 | 0,54 | 0,46 | 0,45 | 0,41 | 0,58 | 0,54 | 1,00 | 0,59 | 0,50 | 0,47 | 0,57 | 0,54 | 0,54 | 0,59 | 0,50 | 0,47 | 0,43 | 0,57 | 0,51 | 0,50 | 0,44 | 0,50 | 0,57 | 0,55 |
| 0,48 | 0,49 | 0,57 | 0,50 | 0,57 | 0,44 | 0,48 | 0,59 | 1,00 | 0,52 | 0,52 | 0,54 | 0,59 | 0,53 | 0,54 | 0,56 | 0,53 | 0,42 | 0,51 | 0,52 | 0,43 | 0,57 | 0,45 | 0,53 | 0,58 |
| 0,51 | 0,50 | 0,53 | 0,52 | 0,48 | 0,58 | 0,49 | 0,50 | 0,52 | 1,00 | 0,60 | 0,48 | 0,55 | 0,43 | 0,42 | 0,41 | 0,54 | 0,58 | 0,57 | 0,48 | 0,48 | 0,50 | 0,50 | 0,57 | 0,56 |
| 0,55 | 0,45 | 0,51 | 0,54 | 0,57 | 0,51 | 0,42 | 0,47 | 0,52 | 0,60 | 1,00 | 0,45 | 0,50 | 0,53 | 0,56 | 0,58 | 0,42 | 0,54 | 0,51 | 0,56 | 0,59 | 0,44 | 0,48 | 0,44 | 0,50 |
| 0,45 | 0,41 | 0,44 | 0,55 | 0,54 | 0,58 | 0,43 | 0,57 | 0,54 | 0,48 | 0,45 | 1,00 | 0,51 | 0,57 | 0,49 | 0,53 | 0,47 | 0,46 | 0,43 | 0,49 | 0,57 | 0,43 | 0,48 | 0,58 | 0,44 |
| 0,52 | 0,48 | 0,53 | 0,58 | 0,44 | 0,59 | 0,44 | 0,54 | 0,59 | 0,55 | 0,50 | 0,51 | 1,00 | 0,47 | 0,46 | 0,53 | 0,51 | 0,42 | 0,48 | 0,51 | 0,42 | 0,46 | 0,55 | 0,49 | 0,45 |
| 0,57 | 0,58 | 0,51 | 0,47 | 0,46 | 0,58 | 0,43 | 0,54 | 0,53 | 0,43 | 0,53 | 0,57 | 0,47 | 1,00 | 0,42 | 0,46 | 0,51 | 0,58 | 0,47 | 0,49 | 0,52 | 0,59 | 0,50 | 0,59 | 0,50 |
| 0,54 | 0,48 | 0,57 | 0,51 | 0,55 | 0,47 | 0,43 | 0,59 | 0,54 | 0,42 | 0,56 | 0,49 | 0,46 | 0,42 | 1,00 | 0,48 | 0,50 | 0,56 | 0,51 | 0,56 | 0,48 | 0,48 | 0,56 | 0,50 | 0,42 |
| 0,57 | 0,47 | 0,56 | 0,58 | 0,59 | 0,57 | 0,43 | 0,50 | 0,56 | 0,41 | 0,58 | 0,53 | 0,53 | 0,46 | 0,48 | 1,00 | 0,49 | 0,56 | 0,47 | 0,42 | 0,44 | 0,42 | 0,52 | 0,46 | 0,55 |
| 0,44 | 0,47 | 0,49 | 0,47 | 0,51 | 0,52 | 0,48 | 0,47 | 0,53 | 0,54 | 0,42 | 0,47 | 0,51 | 0,51 | 0,50 | 0,49 | 1,00 | 0,43 | 0,49 | 0,51 | 0,44 | 0,57 | 0,54 | 0,55 | 0,54 |
| 0,58 | 0,55 | 0,43 | 0,59 | 0,52 | 0,57 | 0,43 | 0,43 | 0,42 | 0,58 | 0,54 | 0,46 | 0,42 | 0,58 | 0,56 | 0,56 | 0,43 | 1,00 | 0,58 | 0,54 | 0,46 | 0,48 | 0,56 | 0,45 | 0,54 |
| 0,43 | 0,50 | 0,50 | 0,42 | 0,46 | 0,57 | 0,53 | 0,57 | 0,51 | 0,57 | 0,51 | 0,43 | 0,48 | 0,47 | 0,51 | 0,47 | 0,49 | 0,58 | 1,00 | 0,54 | 0,41 | 0,50 | 0,51 | 0,57 | 0,46 |
| 0,57 | 0,47 | 0,44 | 0,51 | 0,58 | 0,44 | 0,54 | 0,51 | 0,52 | 0,48 | 0,56 | 0,49 | 0,51 | 0,49 | 0,56 | 0,42 | 0,51 | 0,54 | 0,54 | 1,00 | 0,41 | 0,46 | 0,52 | 0,59 | 0,52 |
| 0,55 | 0,54 | 0,52 | 0,44 | 0,51 | 0,41 | 0,57 | 0,50 | 0,43 | 0,48 | 0,59 | 0,57 | 0,42 | 0,52 | 0,48 | 0,44 | 0,44 | 0,46 | 0,41 | 0,41 | 1,00 | 0,43 | 0,52 | 0,59 | 0,59 |
| 0,59 | 0,51 | 0,58 | 0,46 | 0,41 | 0,59 | 0,58 | 0,44 | 0,57 | 0,50 | 0,44 | 0,43 | 0,46 | 0,59 | 0,48 | 0,42 | 0,57 | 0,48 | 0,50 | 0,46 | 0,43 | 1,00 | 0,51 | 0,43 | 0,56 |
| 0,58 | 0,46 | 0,43 | 0,46 | 0,51 | 0,49 | 0,55 | 0,50 | 0,45 | 0,50 | 0,48 | 0,48 | 0,55 | 0,50 | 0,56 | 0,52 | 0,54 | 0,56 | 0,51 | 0,52 | 0,52 | 0,51 | 1,00 | 0,41 | 0,48 |
| 0,50 | 0,50 | 0,43 | 0,53 | 0,57 | 0,41 | 0,48 | 0,57 | 0,53 | 0,57 | 0,44 | 0,58 | 0,49 | 0,59 | 0,50 | 0,46 | 0,55 | 0,45 | 0,57 | 0,59 | 0,59 | 0,43 | 0,41 | 1,00 | 0,43 |
| 0,55 | 0,58 | 0,48 | 0,51 | 0,48 | 0,58 | 0,44 | 0,55 | 0,58 | 0,56 | 0,50 | 0,44 | 0,45 | 0,50 | 0,42 | 0,55 | 0,54 | 0,54 | 0,46 | 0,52 | 0,59 | 0,56 | 0,48 | 0,43 | 1,00 |

# Öz Geçmiş

## KİŞİSEL BİLGİLER


Adı Soyadı : Esra Gültürk

Uyruğu : T.C.

Doğum Tarihi ve Yeri : 1979- SİVAS

e-posta : esra0709.eg@gmail.com


## EĞİTİM

| Derece | Kurum | Mezuniyet Yılı |
|---|---|---|
| Lisans | Ondokuz Mayıs Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü | 2005 |
| Yüksek Lisans | Cumhuriyet Üniversitesi Tıp Fakültesi Biyoistatistik ABD | 2009 |

## İŞ TECRÜBESİ

| Tarih | Kurum | Görev |
|---|---|---|
| 2006-2009 | Cumhuriyet Üniversitesi Tıp Fakültesi Biyoistatistik ABD | Araştırma Görevlisi |
| 2009- | Cumhuriyet Üniversitesi Tıp Fakültesi Biyoistatistik ABD | Araştırma Görevlisi |

## YABANCI DİL BİLGİSİ

| Yabancı Dilin Adı | KPDS (-) | ÜDS (58) | TOEFL (-) | EILTS (-) |
|---|---|---|---|---|