İnönü Üniversitesi Sosyal Bilimler Enstitüsü
Ekonometri Ana Bilim Dalı

VERİ MADENCİLİĞİ VE KANSER ERKEN TEŞHİSİNDE KULLANIMI

Muhammed Şamil ŞIK

Danışman: Yrd.Doç.Dr. Hasan SÖYLER

Yüksek Lisans Tezi

Malatya, 2014

VERİ MADENCİLİĞİ VE KANSER ERKEN TEŞHİSİNDE KULLANIMI

Muhammed Şamil ŞIK

İnönü Üniversitesi Sosyal Bilimler Enstitüsü
Ekonometri Ana Bilim Dalı

Danışman: Yrd.Doç.Dr. Hasan SÖYLER

Yüksek Lisans Tezi

Malatya, 2014

KABUL VE ONAY

Muhammed Şamil ŞIK tarafından hazırlanan "Veri Madenciliği Ve Kanser Erken Teşhisinde Kullanımı" başlıklı bu çalışma, 27.06.2014 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Prof. Dr. Mehmet GÜNGÖR (Başkan)

Yrd. Doç. Dr. Hasan SÖYLER (Danışman)

Yrd. Doç. Dr. Yunus BULUT (Üye)

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

T.C.
Prof. Dr. Mehmet KARAGÖZ
Enstitü Müdürü

Enstitü Müdürü

ONUR SÖZÜ

"Yrd.Doç.Dr. Hasan SÖYLER'in danışmanlığında yüksek lisans tezi olarak hazırladığım **VERİ MADENCİLİĞİ VE KANSER ERKEN TEŞHİSİNDE KULLANIMI** başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığını ve yararlandığım bütün yapıtların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım."

Muhammed Şamil ŞIK

TEŞEKKÜR

Veri Madenciliği konusunda araştırma olanağı sağlayan, bu çalışma sırasında ilgisini ve desteğini esirgemeyen tez danışmanım Sayın Yrd.Doç. Dr. Hasan SÖYLER'e çok teşekkür ederim. Ayrıca ellerindeki veri ve bilgileri benimle paylaşan İnönü Üniversitesi Medikal Onkoloji Bilim Dalı Doç.Dr. Hakan Harputluoğlu'na, Malatya Halk Sağlık Müdürlüğü Bulaşıcı Olmayan Hastalıklar Programlar Ve Kanser Şube Müdürü Dr. Selma Aydın Felek'e ve İstanbul Tuzla Devlet Hastanesi Baş Hekimi Dr. İrfan Fırat'a teşekkürü bir borç bilirim. Veri bulma aşamasındaki katkılarından dolayı Anadolu Tıbbi Onkoloji Derneği Sekreteri Ayhan Yıldız'a ve Malatya Halk Sağlık Müdürlüğü Kanser Kayıt Uzmanlarına çok teşekkür ederim. Ve her zaman yanımda olan aileme; düzenlemelerde benden yardımını esirgemeyen kardeşim Emre'ye teşekkürlerimi sunarım.

ÖZET

ŞIK, Muhammed Şamil. Veri Madenciliği Ve Kanser Erken Teşhisinde Kullanımı, Yüksek Lisans Tezi, Malatya, 2014.

İstatistiksel modelleme kullanarak veriden bir sonuca ulaşmanın iki yolu vardır. Bunlardan biri, veri modelinin nasıl çalıştığına dair bir bakış açısıyla işe başlayan ve modeli tanımlayan parametreleri açığa çıkarmaya çalışan ekonometri, diğeri de veri modeli hakkında kabullerde bulunmayan ve eldeki veriyi kullanarak en iyi modeli algoritmik olarak kurmaya çalışan veri madenciliğidir. Yapılan bu çalışmada veri madenciliği kavramı, süreci, teknikleri ve veri madenciliği açısından önemli kavramlar olan OLAP, veri ambarı ve CRISP-DM tanımlanmıştır. WEKA programı ile gerçekleştirilen bir veri madenciliği uygulamasıyla günümüzün yaygın bir sağlık problemi olan kanserde hayatta kalma ihtimalini arttıran erken teşhis sürecinde veri madenciliğinin kullanımı gösterilmiştir.

Anahtar Sözcükler

VTBK, Veri Madenciliği, Veri Tabanı, OLAP, Veri Ambarı, CRISP-DM, WEKA, Göğüs Kanseri.

ABSTRACT

ŞIK, Muhammed Şamil. Data Mining and Its Use In Cancer Early Diagnosis, Master Thesis, Malatya, 2014.

There are two ways of reaching conclusions from data by using statistical modeling. One of them is econometrics which starts with a perspective of how data model works and trying to reveal the parameters which identify the model, the other is data mining which doesn't make assumptions about data model and trying to find best model by algorithmically by using the available data. In this study data mining concept, its process, techniques and OLAP, data warehouse and CRISP-DM which are important concepts in terms of data mining were described. With a data mining application performed by WEKA program use of data mining shown in early detection process which increases the probability of survival in cancer which is one common health problem of todays.

Key Words

KDD, Data Mining, Database, OLAP; Data Warehouse, CRISP-DM, WEKA, Breast Cancer.

VERİ MADENCİLİĞİ VE KANSER ERKEN TEŞHİSİNDE KULLANIMI

Muhammed Şamil ŞIK

İÇİNDEKİLER

TABLOLAR

ŞEKİLLER

KISALTMALAR

ARFF: Attribute Relation File Format

CRISP-DM: Cross Industry Standard Process for Data Mining (Çapraz Endüstri Veri Madenciliği Standart Süreci)

ETL: Extraction, Transformation, and Loading (çıkarsama, dönüştürme ve yükleme)

FNA: Fine Needele Aspiration (İnce İğne Aspirasyonu)

OLAP: Online Analytical Processing (Çevrim İçi Analitik İşleme)

OLTP: Online Transaction Processing (Çevrim Hareket İşleme)

SQL: Structured Query Language (Yapılandırılmış Sorgu Dili)

VTBK: Veri Tabanlarında Bilgi keşfi

VTYS: Veri Tabanı Yönetim Sistemleri

WEKA: Waikato Environment for Knowledge Analysis

GİRİŞ

1990'ların başında dünyadaki bilgi miktarının her 20 ayda bir, iki katına çıktığı tahmin edilmekteydi ve veri tabanlarının boyutları ve sayıları muhtemelen bundan daha hızlı artmaktaydı. Çünkü telefon etmek, kredi kartı kullanmak veya tıbbi bir test gibi basit işlemler bile genellikle bir bilgisayara kaydedildiğinden iş faaliyetlerinin otomasyonu giderek artan bir veri akışı oluşturmaya başlamıştı. Bu nedenle bilimsel ve idari veri tabanları hâlâ hızla büyümekteydi. Örneğin NASA 1990'larda bile analiz edebileceğinden fazla veriye sahipti. O yıllarda dünya gözlem uydularının, önceki tüm görevlerin toplamından fazla olarak, her gün bir terabayt (10^{12} bayt) veri oluşturması bekleniyordu. Bu durumda bir kişinin, gözlem uydularının bir günde oluşturdukları resimlere bir saniyede bir resim oranı ile geceleri ve hafta sonları da dâhil olarak sadece bakması birkaç yıl sürerdi ve şüphe yok ki, böyle büyük bir veri yığını içindeki verilerin çok azı insan gözüyle görülecekti. Eğer eldeki bu verilerin hepsi anlaşılsaydı, bilgisayarda analiz edilebilirlerdi (Piatetsky-Shapiro ve diğ., 1992).

Her ne kadar, veri analizi için temel istatistiksel teknikler uzun süre önce geliştirilmiş olsa da "Bu kadar ham veri akışıyla ne yapmamız lazım?" sorusunu cevaplayabilmek için verilerin akıllıca analiz edilmesini, anlaşılmasını, öngörülmesini ve elde edilen bilginin sunulmasını sağlayan gelişmiş teknikler henüz olgunlaşmamıştır. Bunun sonucunda veri üretimi ve üretilen veriyi anlama arasında giderek büyüyen bir uçurum oluşmuştur (Piatetsky-Shapiro ve diğ., 1992). Bu uçurumu kapatmak için yapılan çalışmalar sonucunda veri tabanı, çevrim içi analitik işleme, veri ambarı ve veri madenciliği kavramları ortaya çıkmış ve gelişmeye devam etmişlerdir.

BÖLÜM 1: VERİ TABANI

1.1 VERİ NEDİR?

Veriler; ölçüm, sayım, deney, gözlem ya da araştırma yolu ile elde edilen değerlerdir. Ölçüm ya da sayım yolu ile toplanan ve sayısal bir değer bildiren veriler *nicel* veriler, sayısal bir değer bildirmeyen metin, resim, ses veya video gibi veriler de *nitel* veriler olarak adlandırılmaktadır.

Her ne kadar veri (*data*) ve bilgi (*information*) kelimelerinin aynı anlamı karşıladıkları düşünülse de aralarındaki fark verinin işlenmemiş bilgi olmasıdır. Anlamlı biçimde derlenen ve birleştirilen veriler kullanılarak öğrenilen ve gelecekte karar vermeye yardımcı olan yeni değerlere bilgi denir. Yani bilgi veriden araştırmacıya iletilen mesajın içeriğidir.

1.2 VERİ TABANI NEDİR?

Veri tabanı, bilgisayar kullanımında çözüme erişmek için işlenebilir duruma getirilmiş bilgi ortamıdır (Türk Dil Kurumu). Bu tanımı biraz daha açacak olursak veri tabanları; birbiriyle ilişkili olan, gereksiz yinelemelerden arınmış ve belirli bir amaca uygun olarak düzenlenmiş verilerin mantıksal ve fiziksel olarak tanımlarının olduğu tablolardan oluşan bilgi depolarıdır. Kısaca, veri tabanı kullanım amacına uygun veri depolayan bir yazılımdır. Veri tabanı yazılımlarının diğer veri depolayan yazılımlardan farkları; depoladıkları veriyi verimli ve hızlı bir şekilde yönetmeleri, değiştirebilmeleri ve bu veriye hızlı bir biçimde erişme imkânı sağlamalarıdır.

Hayatımız boyunca çeşitli kaynaklardan öğrendiğimiz/aldığımız, belirli konular hakkında toplanmış ve çeşitli başlıklar altında beynimizde tutulan verileri istediğimizde tek olarak veya diğer verilerle birleştirerek kullanabiliriz. Örneğin

renkler başlığı altında bildiğimiz, tanıdığımız, sevdiğimiz ve sevmediğimiz renkler yer alır; tatlar başlığı altında acı, tatlı, ekşi, sevdiğimiz ve sevmediğimiz tatlar yer alır; arkadaşlar başlığı altında ise arkadaşlarımızın isimleri, telefon numaraları, adresleri ve hatta yüzlerinin resimleri de yer alır. Veri tabanları ise günlük hayatta beynimizin gerçekleştirdiği bu fonksiyonların bir benzerini bilgisayar ortamı için sağlar. Veri tabanlarında toplanan verilerden istenildiğinde; toplanılan verilerin tümü veya istenilen özelliklere uyanları görüntülenebilir, yazdırılabilir ve hatta bu verilerden yeni veriler üretilerek bunlar çeşitli amaçlarla kullanılabilir (Şentürk, 2006: 4; Burma, 2005: 12).

Pek çok kurum ve kuruluşun düzenli ve verimli bir şekilde işleyebilmesi için hayati önem taşıyan veri tabanlarının kullanıldığı alanlara ve sakladıkları verilere örnek olarak şunlar verilebilir:

- Ticari bir şirket için; şirketin ürettiği malların, bu malları üretmek için şirketin kullanması gereken malzemelerin, bu malzemelerin alındığı diğer şirketlerin, şirketin; müşterilerinin, ödemelerinin, tahsilatlarının, borçlarının, bakiyesinin, çalışanlarının verileri, vb. kayıtlar,
- Bir okul için; öğretmen ve öğrenci verileri, boş ve dolu derslikler, sınav tarihleri, sınav sonuçları, eğitim planları, vb. kayıtlar,
- Bir hastane için; hasta verileri, yatakların doluluk ve boşluk verileri, teşhis-tedavi verileri, sağlık personeli nöbet çizelgeleri, doktor performans puanları, döner sermaye ve gider verileri, vb. kayıtlar.

Bu örneklerin yanı sıra düğün davetli listeleri, bir telefon rehberine kayıtlı kişiler ve adresleri, *Windows Başlat* menüsündeki tüm program kısa yolları, *Sık Kullanılanlar* klasöründe düzen verilmiş internet kısa yolları birer veri tabanıdır.

1.3 VERİ TABANI BİLEŞENLERİ

Bir veri tabanı; tablolar, formlar, raporlar, sorgular, makrolar ve modüller gibi bileşenlerden oluşur (Database Basics).

Tablolar: Bir veri tabanı tablosu şekil olarak satır ve sütunlarda verinin depolandığı bir çalışma tablosu gibidir.

Her bir satıra bir kayıt denk gelir. Kayıtlar bilgi parçacıklarının saklandığı yerlerdir. Her bir kayıt, bir veya daha fazla alandan oluşur. Alanlar tabloda sütunlara karşılık gelir. Örneğin "Hasta" isimli bir tabloda, her bir kayıt (satır) farklı bir hasta hakkında bilgi içerir ve her bir alan (sütun) ad, soyad, yaş, vb. gibi farklı türlerde bilgiler içerir.

Formlar: Veri giriş ekranları olarak ta adlandırılan formlar veri ile çalışmak için kullanılan arayüzlerdir[1] ve çoğu zaman çeşitli komut düğmelerini içerirler.

Formlar, veri ile çalışmada kolaylıklar sağlarlar ve ayrıca formlara komut düğmeleri gibi işlevsel öğeler de eklenebilir. Komut düğmeleri formda hangi verinin görüneceğini belirlemek için, diğer formları veya raporları açmak için veya diğer pek çok görev için programlanabilirler. Örneğin; hasta verileriyle birlikte kullanılan "Hasta Formu" diye adlandırılmış bir form, yeni bir hastalık tanısının kayda geçilmesi için bir hastalık tanı formunu açan bir düğmeye sahip olabilir.

Ayrıca, formlar diğer kullanıcıların veri tabanındaki verilerle nasıl etkileşime geçeceklerini de kontrol etmeye yararlar. Örneğin, veri tabanındaki verinin sadece belirli bir kısmının görüntülenmesine izin veren bir form oluşturulabilir. Bu, verinin

[1] Bilgisayar yazılımlarının kullanıcı tarafından çalıştırılmasını sağlayan, çeşitli resimlerin, grafiklerin, yazıların yer aldığı ön sayfa.

korunmasına yardım eder ve verinin uygun bir şekilde kayda geçmesini garanti altına alır.

Raporlar: Tablolardaki verileri özetlemek ve temsil etmek için kullanılırlar. Bir rapor genellikle, "Bu yıl içinde her bir hastanın tedavi masrafı ne kadar tuttu?" veya "Hastalar hangi semtlerde oturuyorlar?" gibi belirli bir soruyu cevaplar. Her bir rapor bilgiyi olabilecek en okunaklı şekilde temsil etmek üzere biçimlendirilir.

Sorgular: Veri tabanlarındaki yük atları olarak da ifade edilebilirler. En yaygın işlevleri tablolardaki veriyi geri çağırmaktır. Ulaşılmak istenen veri genellikle farklı tablolara yayılmış şekilde bulunur ve sorgular bu verilerin tek bir çalışma tablosunda görüntülenmesine izin verir. Sorgular, tüm kayıtları tek seferde görüntülemek istemeyen kullanıcının veriyi filtreleme ölçütleri eklemesine izin vererek sadece istenilen verilere ulaşılmasını sağlar.

Makrolar: Veri tabanına işlevsellik eklenmesinde kullanılabilen basitleştirilmiş bir programlama dili olarak düşünülebilirler. Örneğin; bir makro, bir form üzerindeki bir komut düğmesine iliştirilebilir böylece komut düğmesine her tıklandığında makro da kullanılmış olur. Makrolar bir raporu açmak, bir sorguyu çalıştırmak veya veri tabanını kapatmak gibi görevleri gerçekleştiren faaliyetleri içerirler. Veri tabanında el yordamı ile yürütülen çoğu operasyon makrolar sayesinde otomatikleştirilerek zamandan tasarruf sağlanır.

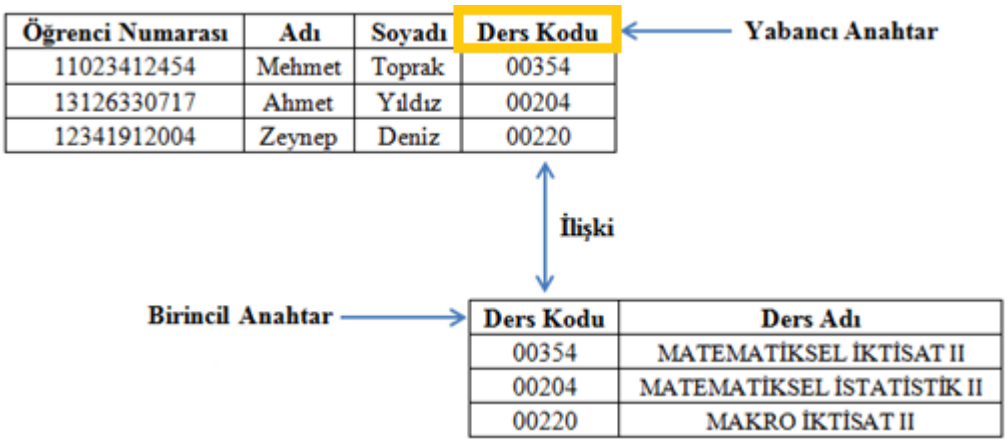
Modüller: Modüller, makrolar gibi, veri tabanına işlevsellik eklenmesinde kullanılabilen araçlardır. Makrolar, makro faaliyetleri listesinden seçilerek oluşturulurken; modüller, uygulamalar için Visual Basic programlama dili ile yazılırlar. Bir modül; tanımların, durumların ve tek bir ünite olarak bir arada depolanan süreçlerin bir koleksiyonudur.

1.4 İLİŞKİSEL VERİ TABANLARI

Bir ilişkisel veri tabanı, tümü ilişkisel modele göre tanımlanmış ve düzenlenmiş veri tablolarının bir kümesini içeren bir veri tabanıdır. Bir tek tablodaki veri bir ilişkiyi temsil eder. Çoğu zaman tablolar, ek olarak, birbiriyle tanımlanmış ilişkilere de sahip olabilirler.

Herhangi bir ilişkisel veri tabanı tablosunda birincil anahtar ve ikincil anahtar olmak üzere önemli iki tip sütun vardır. Birincil anahtar, herhangi bir belirli kaydı benzersiz bir şekilde tanımlayabilen (müşteri kimlik numarası, seri numarası, vb.) sütun veya sütunlardır. Yabancı anahtar ise bir başka tablonun birincil anahtarını işaret eden ve diğer tablolar arasında bağlantı kuran sütun veya sütunlardır.

Şekil 1.1 ilişkisel veri tabanlarındaki bir ilişki modelini göstermektedir.



Şekil 1.1: İlişki Modeli

İlişkisel veri tabanlarının adlandırılmasında kullanılan ilişkisel terimi sadece tablolar arasındaki ilişkilere karşılık gelmez. İlk olarak, tablonun kendi içindeki sütunlar arasındaki ilişkilere, yani tablonun kendisine karşılık gelir; ikinci olarak tablolar arasındaki bağlantılara karşılık gelir.

İlişkisel veri tabanları boyutlarından ve karmaşıklıklarından dolayı, genellikle, ilişkileri sürdürmek için alt programlara[2] ve seçilmiş veriyi geri çağırmak ve sunmak için sorgu gibi harici programlara erişim arayüzlerine ihtiyaç duyarlar.

1.5 VERİ TABANI YÖNETİM SİSTEMLERİ

Veri tabanı yönetim sistemleri (VTYS) bir sabit disk veya bir ağ gibi sistemlere yüklenen tüm veri tabanlarının düzenlenmesi, depolanması ve geri çağrılmaları gibi yönetim işlemlerine tahsis edilmiş bilgisayar programlarıdır.

Her veri tabanı yönetim sisteminde dört önemli eleman vardır.

1. **Modelleme Dili:** VTYS'de bulunan veri tabanlarının dilleridir. Hiyerarşik, ağ, ilişkisel ve nesne veri tabanı dilleri günümüzde yaygın olarak kullanılan bazı modelleme dilleridir.
2. **Veri Yapıları:** Bireysel kayıtlar, dosyalar, alanlar ve bunların tanımları ve görsel öğeler gibi verileri düzenlemeye yardım ederler.
3. **Sorgu Dili:** Bağlantı verisini, farklı kullanıcılar için erişim haklarını ve protokolleri görüntüleyerek veri tabanının güvenliğini sağlar. Örneğin; SQL (*Structured Query Language*; Yapılandırılmış Sorgu Dili) ilişkisel veri tabanı yönetim sistemlerinde kullanılan yaygın bir sorgu dilidir.
4. **Hareket Mekanizması:** Çeşitli hareketlerin aynı anda gerçekleştirilmelerini düzenleyen mekanizmadır. Bu mekanizma bir kaydın aynı anda birden fazla kullanıcı tarafından değiştirilmesine izin vermeyerek veri bütünlüğü korur.

[2] Sık tekrar edilen işlemlerde kullanılan kodları tekrar tekrar yazmak yerine bu işlemler alt programlar olarak tanımlanıp işlem kolaylığı sağlanabilir.

1.6 ÇEVRİM İÇİ ANALİTİK İŞLEME (*OLAP*)

İlişkisel veri tabanlarının kullanımı ve sonrasında ortaya çıkan veri ambarlarının büyüklüğü ile beraber, verilere daha hızlı şekilde erişme ve çok boyutlu analiz ihtiyaçları doğmuştur. SQL'in, verinin muazzam karmaşıklığı ve aşırı büyümesine bağlı olarak analiz için yeterli gelmemeye başlaması ile çevrimiçi analitik işleme ortaya çıkmıştır.

OLAP (*Online Analytical Processing*) terimi, çoğunlukla, bir veri tabanı veya veri ambarındaki veriyi analiz ederken girişilen çeşitli sorgu güdümlü analiz türlerini tanımlamada kullanılır (Berry ve Linoff, 2000). OLAP, boyutlar olarak tanımlanan, farklı bakış açılarından verinin çekilip çıkarılmasını ve görüntülenmesini sağlar (Fayyad, 2001). "OLAP'ın, bir veri ambarı içindeki özetlenmiş verinin çoklu ve dinamik görünümlerini sağlama kabiliyeti veri madenciliği için sağlam bir zemin oluşturur" (Han ve Kamber, 2001). Bu nedenle, veri madenciliği ve OLAP birbirini tamamlayan araçlar olarak kabul edilir.

OLAP, hareket işlemeden ziyade, sorgulama ve raporlama için optimize edilmiş, iş zekâsı sorgularını kolaylaştıran bir veri tabanı teknolojisidir. Ayrıca, OLAP verileri hiyerarşik olarak düzenlenir ve tablolar yerine küplerde depolanırlar.

OLAP, verileri analiz etmek için verilere hızlı erişim sağlayan çok boyutlu yapılar kullanan karmaşık bir teknolojidir. Çok boyutlu yapılar kullanmak, örneğin; iş zekâsında, özet tablo raporunun ve özet grafik raporunun tüm ülke veya bölgedeki toplam satışlar gibi yüksek düzeyde özetleri görüntülemesini ve ayrıca satışların özellikle yüksek veya düşük olduğu bölgeler için ayrıntıları görüntülemesini kolaylaştırır.

OLAP, çok boyutlu sorguları cevaplamayı sağlayan bir sistemler sınıfı olarak da ifade edilebilir. Genel olarak OLAP; pazarlama, planlama, tahminde bulunma ve

benzer uygulamalar için kullanılır. Bu nedenle OLAP için kullanılan veri tabanları hızlı bir şekilde, karmaşık ve amaca özel sorgular yapabilmek için tasarlanmışlardır. Bu sorgular sonucunda elde edilen OLAP çıktılarını göstermek için satırları ve sütunları sorgunun boyutları tarafından şekillendirilen bir matris kullanılır. Ve OLAP ile özetler elde edebilmek için, çoğunlukla, çoklu tablolar üzerinde birleştirme yöntemleri kullanılır. OLAP, örneğin; bu yılın satışları ile geçen yılın satışlarının kıyaslamasında, ilerideki mevsimin satışlarını tahmin etmede ve yüzdelik değişime bakarak eğilim hakkında ne söylenebileceğini belirlemede kullanılabilir.

OLAP, karar destek sistemi araçlarının bir parçasıdır. Geleneksel sorgu ve raporlama araçları veri tabanında ne olduğunu tanımlar ve OLAP bundan daha da öteye giderek bazı şeylerin neden doğru olduğunu cevaplar. Kullanıcı bir bağlantı hakkında bir dizi varsayıma dayalı örüntü ve hipotez üretir ve bu varsayımları doğrulamak veya reddetmek için veri tabanına dayalı sorgular kullanır.

1.6.1 OLAP Veri Tabanının Özellikleri

OLAP veri tabanının sahip olduğu özellikler şunlardır:

- Çok boyutlu inceleme özelliğine sahip olması,
- Şeffaflık,
- Erişilebilirlik,
- Her seviyede sorgulama için aynı performansı gösterebilme özelliği,
- İstemci-Sunucu yapısında olması,
- Sınırsız şekilde çapraz raporlama olanağının olması,
- En alt seviyedeki verilerin otomatik olarak ayarlanması,
- Her şarta uygun boyutlandırılabilir olması,
- Çoklu kullanıcı desteğinin olması,
- Her seviyede verilerin değiştirilebilir olması,
- Esnek raporlama özelliği,

- Boyut ve gruplamalarda sınır olmaması.

OLAP, işlevsel olarak, aşağıdaki analitik ve dolaşımsal faaliyetleri de içeren, son kullanıcıyı destekleyen birleştirilmiş verinin dinamik çok boyutlu analizi olarak tanımlanır:

- Boyutlar içinde, hiyerarşiler boyunca ve üyeler arasında uygulanan hesaplamalar ve modelleme,
- Zaman serileri üzerinde eğilim analizi,
- Görüntülemek için alt kümeleri dilimleme,
- Bütünleştirilmiş yapının daha derin seviyelerindeki detay veriye ulaşma ve
- Görüntülenen alanda yeni boyutsal kıyaslamalar için döndürme operasyonları gerçekleştirme.

Bir OLAP veri tabanı, örneğin; bölgelere, ürün türüne ve satış kanalına[3] göre toplanmış satış verisini içerebilir. Tipik bir OLAP sorgusu her bir bölgenin her bir ürün türüne göre tüm ürün satışlarını bulmak maksadıyla gigabaytlarca/çok sayıda yılın satış veri tabanına ulaşabilir. Sonuçları görüntüledikten sonra, analistin bölge/ürün sınıflandırması bünyesindeki her bir satış kanalı için satış hacimlerini bulmak adına ileride sorguyu iyileştirebilir. Son adım olarak, analist her bir satış kanalı içi yıldan yıla veya mevsimden mevsime kıyaslamalar yapmak isteyebilir. Tüm bu süreç hızlı yanıtlama zamanıyla çevrim içi olarak yürütülmelidir ki analiz süreci bozulmasın.

OLAP sorguları aşağıdaki çevrim içi işlemlerle karakterize edilebilirler:

- Çok büyük miktarda veriye erişme, örneğin, birkaç yılın verisi,
- Pek çok iş unsuru arasındaki bağlantıları analiz etme, örneğin; satış, ürünler, bölgeler, satış kanalları,

[3] Tüketicilerin satın alabilmeleri için ürünlerin veya hizmetlerin bir pazara getirilme yollarına satış kanalları denir.

- Birleştirilmiş veriyi isteme, örneğin; satış hacimleri, bütçedeki para ve harcanan para,
- Hiyerarşik zaman aralıkları üzerinde birleştirilmiş veriyi kıyaslama, örneğin; aylık, mevsimlik veya yıllık,
- Veriyi farklı bakış açılarıyla sunmak, örneğin; bölgeye göre satışlara karşı her bölgedeki ürüne göre kanallar yoluyla satış,
- Veri öğeleri arasında karmaşık hesaplamalar yapmak, örneğin; beklenen kârın, belirli bir bölgedeki her bir satış kanalı türü için satış gelirinin bir fonksiyonu olarak hesaplanması,
- Kullanıcının, sistem tarafından engellenmeden bir analitik düşünce sürecini sürdürebilmesi için kullanıcının isteklerine hızla yanıt verebilen çevrim içi işlemler.

Veri ambarları ve OLAP birbirini tamamlayan iki kavramdır. Veri ambarı verileri barındırmaya yarar. OLAP ise bu yığın halinde duran verileri anlamlı hale getirip; analizler yapmaya yarar.

OLAP veri tabanları ölçüler ve boyutlar olmak üzere iki temel veri türü içerir:

- **Ölçüler:** Bilinçli kararlar vermede kullanılan miktarlar ve ortalamalar gibi nümerik verilerdir.
- **Boyutlar:** Ölçüleri düzenlemede kullanılan sınıflardır.

OLAP veri tabanları, veriyi analiz ederken bilinen sınıfları kullanarak veriyi çok detaylı seviyelerde düzenlemeye yardımcı olurlar.

1.6.2 OLAP Kavramları

Aşağıda OLAP veri tabanları ile ilgili kavramlar tanımlanmıştır (OLAP And OLAP Server Definitions; OLAP Küpü Nedir?; Overview of Online Analytical Processing (OLAP)).

OLAP Sunucusu (*OLAP Server*)**:** Bir OLAP sunucusu, özel olarak, çok boyutlu veri yapılarını desteklemek ve bu veri yapıları üzerinde çalıştırılmak üzere tasarlanmış, yüksek kapasiteli, çoklu kullanıcıya sahip bir veri yönlendirme motorudur. Sunucunun tasarımı ve verinin yapısı, esnek hesaplama ve formülsel bağlantılar için olduğu gibi her yönde ve amaca özel bilgi geri çağırımı için optimize edilmiştir. OLAP sunucusu ya son kullanıcıya uygun ve hızlı yanıt zamanlarıyla işlenmiş çok boyutlu bilgiyi ulaştırır ya da veri yapılarını ilişkisel veya diğer veri tabanlarından gelen verilerle çabucak oluşturur ya da kullanıcıya her iki işlemi de içeren bir tercih sunar.

OLAP veri tabanları verinin geri çağırılmasını hızlandırmak için tasarlanmıştır. Çünkü OLAP sunucusu özetlenmiş değerleri hesaplar ve dolayısıyla bir rapor oluşturulurken veya değiştirilirken daha az veriye ihtiyaç duyulur. Bu yaklaşım, geleneksel veri tabanlarında düzenlenmiş verilerden daha büyük miktardaki kaynak veri ile çalışabilmeyi sağlar.

OLAP sunucusu yer kullanımını optimize ederek fiziksel depolama ihtiyaçlarını minimize edebilir, böylelikle son derece büyük miktarda verinin analiz edilmesini mümkün kılar (Sumathi ve Sivanandam, 2006: 88-94).

OLAP veri tabanı sunucuları; yuvarlamayı, detaya inmeyi ve dilimlemeyi ve küplere ayırmayı içeren, sık kullanılan, analitik operasyonları destekler.

Yuvarlama (*Roll-Up*)**:** Çok boyutlu veri tabanları, genellikle, her bir boyuttaki verinin hiyerarşilerine dayalı veya formüle dayalı bağlantılarına sahiptir. Ne var ki, analist bu hiyerarşilerin en alt seviyelerindeki tüm veriyi ya nadiren dikkate alır ya da hiçbir zaman dikkate almaz. Analist veriye daha az detaylı bir şekilde bakmak isteyebilir, örneğin; her bir ürünün tek tek satışları yerine ürün türüne göre satışları kullanmak isteyebilir. Analist bunu yaparak, ürün boyutu boyunca veriyi ürün türü seviyesine yuvarlamış olur. Başka bir deyişle bir yuvarlama işlemi, bir boyut boyunca veriyi özetleme ile alakalıdır. Özetleme kuralı, bir hiyerarşi boyunca toplamları hesaplama veya "$kâr = satışlar - maliyetler$" gibi bir takım formüller uygulama olabilir.

Detaya İnme (*Drill-Down*)**:** Yuvarlama operasyonunun tersi olan detaya inme, kullanıcının en çok özetlenmiş veri seviyeleri ve en detaylı veri seviyeleri arasında dolaşmasını sağlayan özel bir analitik tekniktir. Örneğin, Türkiye'nin kuzeyi için satışlar görüntülenirken, bölge boyutunda bir detaya inme operasyonu Marmara, Karadeniz ve Doğu Anadolu bölgelerini görüntüler. Karadeniz bölgesi üzerinde daha ileri seviyede bir detaya inme operasyonu gerçekleştirilirse Doğu, Orta ve Batı Karadeniz bölümleri görüntülenebilir.

Dilimleme ve Küplere Ayırma (*Slice and Dice*)**:** Dilimleme ve küplere ayırma kavramları veri tabanına farklı bakış açılarından bakma yeteneği ile ilgilidir. Dilimleme, boyutlardan birinden sadece bir değer seçip, küpün dikdörtgen şeklinde bir alt kümesini ayırmak suretiyle bir boyutu eksik yeni bir küp oluşturma işlemidir. Son kullanıcının bakış açısıyla bir dilim, küpten seçilen iki boyutlu bir tablodur. Küplere ayırma ise analistin çoklu boyutun belirli değerlerini seçmesine izin vererek bir alt küp oluşturmasını sağlayan bir operasyondur.

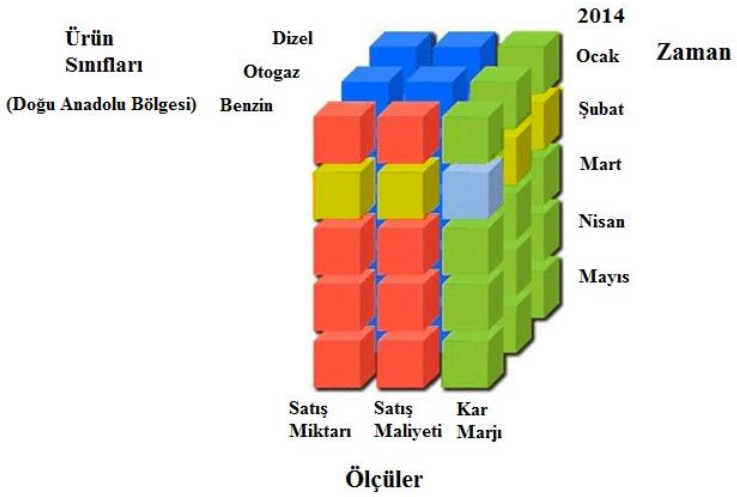
Dilimleme ve küplere ayırma, genellikle, eğilimleri analiz etmek ve örüntüler bulmak amacıyla zaman ekseni boyunca uygulanır.

OLAP İstemcisi (*OLAP Client*)**:** OLAP sunucularından dilimler isteyen ve görselleştirme veya dolaşım amaçları için iki boyutlu veya çok boyutlu gösterimler sağlayan son kullanıcı uygulamalarıdır.

OLAP, çoklu kullanıcı istemci/sunucu modunda uygulanır ve veri tabanı boyutu ve karmaşıklığı ne olursa olsun sorgulara çabuk yanıt verir.

Küp: Analiz edilmek istenen her bir boyutun seviyelerine ve hiyerarşilerine göre ölçüleri birleştiren bir veri yapısıdır. Küpler; satış veya stok rakamları gibi özetlenmiş veriler ile zaman, coğrafya ve ürün hatları gibi birkaç boyutu birleştirir. Küpler matematiksel anlamdaki "küpler" değillerdir, çünkü eşit kenarlara sahip olmak zorunda değillerdir. Yine de böyle karmaşık bir kavram için uygun bir benzetmedir.

Şekil 1.2 bir akaryakıt firmasının Doğu Anadolu Bölgesi'ne ait verilerini içeren bir OLAP küpünü göstermektedir.



Şekil 1.2: OLAP küpü.

Ölçü: Bir küpün olgu tablosunda bir sütunu temel alan ve genellikle nümerik değerler olan bir veriler kümesidir. Ölçüler; bir kübteki işlenmiş, birleştirilmiş ve analiz edilmiş merkezi değerlerdir. Yaygın örnekleri arasında satışlar, kâr, gelir ve maliyet sayılabilir.

Üye: Bir hiyerarşide, verinin bir veya daha fazla kez bulunmasını temsil eden bir elemandır. Bir üye benzersiz veya benzersiz olmayan olabilir. Örneğin, 2007 ve 2008 zaman boyutunun yıl seviyesindeki benzersiz üyeleri temsil ederken; Ocak, ay seviyesindeki benzersiz olmayan üyeleri temsil eder çünkü elde bir veya daha fazla yılın verisi varsa zaman boyutunda bir veya daha fazla Ocak ayı olacaktır.

Hesaplanan Üye: Değeri, yürütme esnasında hesaplanan bir boyut elemanıdır. Hesaplanan üye değeri, diğer üyelerin değerlerinden türetilebilir. Örneğin; kâr hesaplanan üyesi, gider üyesinin değerini satışlar üyesinin değerinden çıkararak belirlenebilir.

Hücre: Çok boyutlu yapıdaki her bir boyuttan bir eleman seçilmesiyle oluşturulmuş kesişim kümesinde bulunan yalnız bir veri değeridir. Örneğin, boyutlar; ölçüler, zaman, ürün ve coğrafya ise Satışlar, Ocak 2014, Ekmek ve Türkiye boyut elemanları, tüm boyutlar boyunca kesişen ve 2014 Ocak ayında Türkiye'deki ekmek satışlarının değerini içeren yalnız bir veri hücresini benzersiz bir şekilde tanımlar.

Boyut: Bir boyut, kullanıcının veri algısına göre hepsi benzer türden elemanların bir listesi olan, küpün yapısal bir özelliğidir. Kullanıcının kolaylıkla anladığı ve veri analizi için temel olarak kullandığı, bir küpün içindeki seviyelerin bir veya daha fazla düzenlenmiş hiyerarşisidir. Örneğin; tüm aylar, mevsimler, yıllar, vs. bir zaman boyutu oluşturur ve benzer şekilde tüm şehirler, bölgeler, ülkeler, vs. bir coğrafya boyutu oluşturur. Bir boyut, çok boyutlu yapı içerisindeki değerleri tanımlamak için kullanılan bir dizin olarak hareket eder. Eğer boyutun bir elemanı seçilmişse, kalan elemanlar bir alt küp tanımlar. Eğer iki boyutun seçilmiş yalnız bir elemanı varsa, geriye kalan iki boyut bir tablo tanımlar. Eğer tüm boyutların seçilmiş yalnız bir elemanı varsa, bu yalnızca bir hücre tanımlar. Boyutlar; geri çağırma, keşfetme ve analiz etme için veriyi düzenlemenin ve seçmenin çok kısa ve sezgisel bir yolunu sunar.

Hiyerarşi: Her bir üyenin bir üst üyesi olacak ve sıfır veya daha fazla alt üyesi olacak şekilde bir boyutun üyelerini düzenleyen mantıksal bir ağaç yapısıdır. Bir alt üye, hiyerarşide doğrudan mevcut üye ile bağlantılı bir sonraki düşük seviyedeki bir üyedir. Örneğin; mevsim, ay ve gün içeren bir zaman hiyerarşisinde Ocak ayı mevsimin bir alt üyesidir. Bir üst üye, hiyerarşide doğrudan mevcut üye ile bağlantılı bir sonraki üst seviyedeki bir üyedir. Örneğin; mevsim, ay ve gün içeren bir zaman hiyerarşisinde mevsim Ocak ayının üst üyesidir.

Seviye: Bir hiyerarşide veri; bir zaman hiyerarşinde yıl, mevsim, ay ve gün seviyelerinde olduğu gibi düşük veya yüksek detay seviyeleri olarak düzenlenebilir.

Döndürme (*Pivot*)**:** Döndürme terimi, verinin alternatif temsillerini görüntülemek amacıyla veri eksenlerini döndüren bir görselleştirme operasyonuna karşılık gelir. Döndürme bir satır boyutunu sütun boyutuna taşıma, yani satırlar ile sütunların yerini değiştirme işlemleri ile gerçekleştirilebilir.

Dolaşım (*Navigation*)**:** Dolaşım; kullanıcının, genellikle, bir OLAP sunucusuna bağlı grafiksel bir OLAP istemcisi kullanarak bir küpü keşfetmek için kullandığı detaya inme, döndürme ve görüntüleme süreçlerini tanımlamak için kullanılan bir terimdir.

1.6.3 Çevrim İçi Hareket İşleme (*OLTP*)

Çevrim içi hareket işleme uygulamaları, günlük veri tabanı işlemlerinin gereksinimlerini ve kullanıcıların operasyonel veri[4] ihtiyaçlarını karşılamak için geliştirilmiştir. OLTP (*Online Transaction Processing*), çok sayıda kısa çevrim içi işlem tarafından (ekleme, güncelleme, silme) karakterize edilir. OLTP sistemleri için; çok hızlı sorgu işleme, çoklu erişim ortamlarında veri bütünlüğünü koruma ve

[4] Veriyi genel olarak, enformasyonel veri ve operasyonel veri olarak ikiye ayırmak mümkündür. Enformasyonel veri, kişiye yönelik, bütünleşik, zaman içinde oluşan ve birleştirilmiş veriler olarak tanımlanabilir. Operasyonel veri ise, uygulamaya yönelik, dağınık, kısa zamanda oluşan ve tekrarlayabilen veriler olarak tanımlanmaktadır (Özmen, 2001).

saniyedeki işlem sayısı ile ölçülen etkinlik önemlidir. Bir OLTP veri tabanında detaylı, güncel veriler ve işlemsel veri tabanlarını depolamada kullanılan varlık modeli olan şema vardır.

Genellikle, veri ambarlarında depolanmış olan OLTP veri tabanları OLAP'ın veri kaynağıdır. OLAP verisi bu tarihsel (geçmiş) veriden türer ve karmaşık analizlere müsaade eden yapılar halinde derlenir.

Tablo 1.1 OLTP ve OLAP sistem tasarımları arasındaki temel farkları göstermektedir (Sumathi ve Sivanandam, 2006: 70-72; Gürsoy, 2009: 21; Özkan, 2013, 32-34; OLTP vs. OLAP).

	OLTP Sistemi Çevrim İçi Hareket İşleme (Operasyonel Sistem)	**OLAP Sistemi Çevrim İçi Analitik İşleme (Veri Ambarı)**
Verinin Kaynağı	Operasyonel Veri: OLTP'ler verinin esas kaynağıdır.	Birleştirilmiş Veri: OLAP verisi çeşitli OLTP veri tabanlarından gelir.
Verinin Amacı	Temel iş görevlerini kontrol etmek ve yürütmektir.	Planlamaya, problem çözmeye ve karar desteğe yardımcı olmaktır.
Veri Nedir?	Devam etmekte olan iş sürecinin anlık bir görüntüsünü ortaya koyar.	Çeşitli iş aktivitelerinin çok boyutlu görüntüleridir.
Fonksiyon	Günlük operasyonlar.	Uzun vadeli bilgi gereksinimleri, karar destek.
Eklemeler ve Güncellemeler	Son kullanıcılar tarafından gerçekleştirilen kısa ve hızlı eklemeler ve güncellemelerdir.	Periyodik olarak uzun süren toplu veri tazelemelerdir.
Sorgular	Nispeten standartlaştırılmış ve basit sorgulardır.	Çoğu zaman birleştirmeler içeren karmaşık sorgulardır.
İşleme Hızı	Genellikle çok hızlıdır.	İşleme giren veri miktarına bağlı; toplu veri tazelemeler ve karmaşık sorgular saatlerce sürer; sorgu hızı dizinler oluşturularak arttırılabilir.
Erişilen Kayıt Miktarı	Onlarca	Milyonlarca
Kullanıcı sayısı	Binlerce	Yüzlerce

Tablo 1.1: OLTP ve OLAP arasındaki farklar

Alan İhtiyacı	Eğer tarihsel bilgi arşivlenmişse nispeten küçük olabilir.	Birleştirme yapılarının ve tarihsel verinin bulunmasına bağlı olarak daha büyük; OLTP'den daha çok indekse ihtiyaç duyar.
Veri Tabanı Tasarımı	Pek çok tabloyla son derece normalize edilmiştir. Varlık–ilişki veri tabanı, uygulama odaklı.	Genel olarak daha az tabloyla denormalize[5] edilmiştir; yıldız veya kar tanesi şemalarını kullanır. Konu odaklıdır.
Yedekleme ve Kurtarma	Düzenli olarak yedekler; operasyonel veri işi yürütmek açısından hayati öneme sahiptir, veri kaybı muhtemelen önemli miktarda mali kayba ve borçlanmaya neden olur.	Düzenli yedeklemeler doğrultusunda, bazı ortamlar kurtarma yöntemi olarak OLTP verilerini geri yükleyebilir.

Tablo 1.1: OLTP ve OLAP arasındaki farklar

1.7 VERİ AMBARI

1.7.1 Veri Ambarı Kavramları

İş Zekâsı (*Business Intelligence*)**:** İş zekâsı, bir anlamda, anlamlı bir şekilde (çoğunlukla farklı veri görselleştirme teknikleri kullanılarak) tarihsel verinin temsil edilmesi sanatı ve bilimidir.

Veri tabanlarında depolanan ham veri, iş zekâsı sürecinin uygulanması süreci boyunca değerli bilgiye dönüşür.

[5] Programlamada denormalizasyon, verileri gruplayarak bir veri tabanını okuma hızını optimize etmeye çalışma sürecidir.

"İş zekası teknolojilerinin yaygın fonksiyonları; raporlama, çevrim içi analitik işleme, veri madenciliği, süreç madenciliği, karmaşık olay işleme, iş performansı yönetimi, kıyaslama (*benchmarking*), metin madenciliği ve öngörüsel analizdir".

Karar Destek Sistemleri (*Decision Support Systems*)**:** Karar destek sistemleri, veri hakkındaki ayrıntılı bilgiye ulaşabilen ve karar verme sürecinde kullanıcıların sistemle karşılıklı olarak etkileşimde bulunduğu, bilgisayar tabanlı bilişim sistemleridir. Verinin nereden geldiği, nasıl olduğu, ne olması gerektiği ve gelecekte ne olabileceği hakkındaki sorulara cevap arar.

Operasyonel Veri Tabanı (*Operational Database*)**:** Operasyonel veri tabanı, veri ambarının veri kaynağıdır. Günlük iş operasyonlarını gerçekleştirmede kullanılan detaylı veriyi içerir. Veri, güncellemeler yapıldıkça sürekli değişir ve son işlem mevcut değerini yansıtır.

Bir operasyonel veri tabanı, bir kuruluşa ait güncel ve değiştirilebilir verileri içerir. Bir kuruluşa ait veri yönetim sisteminde bir operasyonel veri tabanının istatistiksel analiz amaçlı olarak çıkarılmış ve değiştirilemez verileri içeren bir karar destek veri tabanının zıttı olduğu söylenebilir. Örneğin; karar destek veri tabanları pek çok farklı işçinin ortalama gelirini tespit etmeye yarayan veriyi sağlarken, operasyonel veri tabanları aynı veriyi verilen bir zaman aralığında işçilerin çalıştıkları gün sayısına dayanarak işçilere ne kadar ödeme yapılacağının hesaplanmasında kullanır.

Tanımlayıcı Veri (*Meta Data*)**:** Tanımlayıcı veri; bir bilgi kaynağını tanımlayan, açıklayan, yerini belirleyen veya kullanmayı ve yönetmeyi kolaylaştıran yapısal bilgidir. Tanımlayıcı veriye çoğu zaman veri hakkında veri veya bilgi hakkında bilgi denir. Örneğin her satırına bir hasta gelecek şeklinde veri içeren bir tabloda hangi alanlar var, her bir alanın boyu ve tipi nedir gibi özellikler tanımlayıcı veridir ve tablonun kendisinde yer almaz.

Veri Marketi (*Data Mart*)**:** İş hakkında stratejik kararlar almalarında, yöneticilere yardımcı olan bir veri tabanı veya veri tabanı topluluğudur. Veri ambarı bir kuruluşun tüm veri tabanlarını bir araya getirirken veri marketi sadece belirli bir iş fonksiyonuna odaklanır. Örneğin son üç ay içindeki satış veya muhasebe kayıtları gibi. Bu yönüyle veri marketi veri ambarının bir alt kümesidir.

1.7.2 Veri Ambarı Nedir?

Veri ambarı (*Data Warehouse*) Bill Inmon tarafından şöyle tanımlanmıştır: ”Bir veri ambarı yönetim kararlarını desteklemek için konuya yönelik, kalıcı, entegre edilmiş, zamana bağlı veri topluluğudur”.

Veri ambarları; ilişkisel verilerin bulunduğu, büyük hacimli, anlık bilgiden tarihsel derinliği olan veriye ulaşabilmeyi, konu-mekân-zaman temelli raporlar alabilmek için altyapı oluşturmayı, veri güvenliğini ve geleceğe yönelik analizler yapabilmeyi sağlayan yapılardır. Veri ambarı sistemleri, kullanıcıya çeşitli operasyonel sistemlerden gelen verilerin temizlenmesi, dönüştürmesi ve saklanması işlevlerinde ve karar destek sistemlerinde ve yardımcı olmaktadır.

Bir veri ambarı, hareket işlemeden[6] ziyade sorgulama ve analiz için tasarlanmış; genellikle işlemsel veriden gelen tarihsel veriyi içeren, fakat diğer kaynaklardan gelen verileri de bulundurabilen bir ilişkisel veri tabanıdır. Analiz iş yükünü, hareket iş yükünden ayırır ve bir işletmenin çeşitli kaynaklardan veri toplamasını mümkün kılar.

[6]Bir işletmede kullanılan kaynakların, işletme içinde ve dışındaki çıkar ve ilgi gruplarının her biri açısından anlamlı olan ve zamanla meydana gelen her bir değişimine *hareket* (ya da işlem) denir. **Hareket işleme,** bir işletmede meydana gelen yapılandırılmış ve sürekli yinelenen olguları kaydetme, izleme, saklama, işleme ve yayımlama işlemleridir. Bu olgulara örnek olarak sipariş almak, fatura ve irsaliye hazırlamak, mal ve hizmet teslim almak ya da etmek, bordro hazırlamak gösterilebilir.

Bir veri ambarı, ilişkisel veri tabanı uygulamalarına ek olarak; çıkarsama, dönüştürme ve yükleme (*Extraction, Transformation, and Loading; ETL*), çevrim içi analitik işleme ve veri toplayıp kullanıcıya ulaştırma süreçlerini sağlayan diğer uygulamaları da içerir (Data Warehouse).

Veri ambarlarının çoğu entegre edilmiş bir ambar şekillendirmek için veriyi çok sayıda kaynaktan toplar. Bu verilerin ambara yüklenmeden önce ETL tarafından yapılan özel bir müdahaleye ihtiyacı vardır. ETL veriye iki tür müdahaleden sorumludur:

1. **Veri Entegrasyonu:** Farklı sistemlerden gelen veriler arasında bazı bağlantıların kurulabilmesi için veriyi entegre eder.
2. **Nitel müdahale:** Veri ambarına yüklenmeden önce verinin doğruluğu ve kalitesi kontrol edilebilir, eğer gerekirse düzeltilebilir.

Veri ambarcılığını tanımlamanın bir yolu da William Inmon tarafından ileri sürülen aşağıdaki veri ambarı özelliklerini açıklamaktır (Oracle9i Data Warehousing Guide Release 2 (9.2)):

- Konuya yönelik (*Subject oriented*)
- Entegre edilmiş (*Integrated*)
- Kalıcı (*Nonvolatile*)
- Zaman bağlı (*Time variant*)

Veri Ambarı Konuya Yöneliktir: Veriyi analiz etmeye yardımcı olmaları için tasarlanmış veri ambarının tanımlanmış bir faaliyet alanı vardır ve sadece bu faaliyet alanına dâhil olan verileri depolarlar. Örneğin, bir firmanın satış ekibi firmanın satış verileri hakkında daha fazla öğrenmek için bir veri ambarı oluşturuyorsa, tanımı gereğince bu veri ambarının üretim yönetimi ile ilgili verileri değil de satışlarla ilgili veri içermesi gerekir. Bu veri ambarını kullanarak " Geçen sene belirli bir ürünü en çok alan müşteri kimdir?" gibi sorulara cevap verilebilir. Bir veri ambarının çalışma

alanına göre tanımlanabilmesi veri ambarını konuya yönelik kılar. Bu sayede veri ambarı hem belirli bir alandaki sorulara cevap almamızı kolaylaştırır hem de bizi gereksiz veri tekrarlarından kurtarmış olur.

Veri Entegre Edilmiştir: Entegrasyon, konuya yönelik olmayla yakından ilişkilidir. Veri ambarları, farklı kaynaklardan gelen verileri tutarlı bir biçime bir araya getirmelidir. İsim çakışmaları ve ölçü birimleri arasındaki uyuşmazlıklar gibi problemleri çözmelidir. Bunu başardıkları zaman, veri ambarlarının entegre edilmiş oldukları söylenir.

Veri Kalıcıdır: Veri bir kez veri ambarında depolandı mı veri ambarından kaldırılmaz veya silinmez ve her ne olursa olsun her zaman veri ambarında kalır. Bu mantıklıdır, çünkü veri ambarının amacı ne gerçekleşmişse onu analiz edebilmeyi mümkün kılmaktır.

Veri Zamana Bağlıdır: Eğilimleri keşfetmek için analistin çok büyük miktarda veriye ihtiyacı vardır. Bu, performans ihtiyaçları nedeniyle tarihsel verinin bir arşive taşınmasını talep eden çevrim içi hareket işlemeye çok zıttır. Bir veri ambarının zamana bağlılığıyla kast edilen, veri ambarının zamanla değişime odaklanmasıdır. Bu, veri ambarına yeni veri yüklendikçe veri ambarı da boyutça büyür anlamına gelir.

Veri madenciliği; veri ambarını, yapay zekâ ve istatistikle bağlantılı yöntemlerin bir karışımı olan bilgi keşfi sistemleri vasıtası ile birliktelikleri bulmak, sınıflandırmalar ve kümelemeler yapmak ve tahminlerde bulunmak için bir bilgi kaynağı olarak kullanır (Gray ve Watson, 1998).

Veri ambarlarındaki veri; temizlenme, diğer bilgilerle birlikte özetlenme veya arşivlenme aşamalarının uygulanabileceği yaşa gelene kadar veri ambarında bulunmaya devam eder.

Bir veri ambarı genel anlamda (Ponniah, 2010; Guerra,2013):

- Tüm verileri operasyonel veri tabanlarından alır.
- Gerektiğinde, dışarıdan konu ile ilgili veri dâhil eder.
- Çeşitli kaynaklardan veri toplar.
- Tutarsızlıkları giderir ve veriyi dönüştürür.
- Karar vermek için kullanılacak veriyi kolay ulaşılabilir uygun formatlarda saklar.
- Veri ambarı fonksiyonuna tahsis edilmiş bilgisayarlarda bulunur.
- Oracle, Microsoft veya IBM gibi veri tabanı yönetim sistemleri üzerinde devam ettirilir.
- Veriyi uzun süre muhafaza eder.
- Pek çok kaynaktan gelen veriyi birleştirir.
- Üretilen veriyi yüksek hızlı bir veri girişi tasarımından yüksek hızlı geri çağırmayı destekleyen bir veri girişi tasarımına dönüştüren bir veri modeli etrafında özenle inşa edilir.

Bir veri ambarı, operasyonel sistemlerden[7] çıkarılan ve amaca özel sorgular ve çizelgelenmiş raporlama için tarihsel anlık görüntüler şeklinde kullanıma sunulan bir veri kavramıdır. Veri ambarında bulunan veriyi operasyon ortamında bulunan veriden ayıran özellikler şunlardır:

- Uygun veriler, kolay ulaşılabilmeleri için birlikte kümelenmiş şekilde bulunur.
- Değişik zamanlarda elde edilen verinin birkaç kopyası bir arada tutulur.
- Veri, veri ambarına bir kez yerleştirildikten sonra güncellenmez. Bunun yerine, veri ambarında saklanan tarihsel anlık görüntüler olarak operasyonel veri tabanlarından gelen veriler ile periyodik olarak yenilenir.

[7] Veri ambarcılığında operasyonel sistemler bir kuruluşun günlük hareketlerinin verimliliğini ve hareket verilerinin bütünlüğünü koruyacak şekilde tasarlanmış, günlük hareketlerinin işlendiği bir sistemdir.

1.7.3 Veri Ambarcılığının Sağladığı Çözümler Nelerdir?

Operasyonel veri (günlük işleri yürüten veri) içeren sistemler kullanıcılar için faydalı bilgiler içerir. Örneğin analist; aykırı durumları araştırmak veya gelecek satışları projelendirmek için hangi ürünlerin, hangi bölgelerde, yılın hangi döneminde satıldığı bilgisini kullanabilir.

Bir veri ambarının işaret ettiği esas problem operasyonel veriye doğrudan ulaşan son kullanıcıların amaca özel veya diğer özel sorgulara ve raporlara ulaşmakta zorluk çekmeleridir. Bu durum birkaç faktöre bağlıdır:

- Verilerin çoğu kullanıcı tarafından ulaşılması zor olan uyarlanabilir veri tabanı yönetim sisteminde[8] saklanır.
- Veri depoları hareket işleme için tasarlanmıştır, amaca özel raporlama için değil.
- Bir veriyi veya raporu elde etmek için, genellikle, raporu oluşturması veya özleştirilmiş bir indirme programı sağlaması için bir programcıya ihtiyaç duyulur.
- Tüm veriler, aynı zamana ait olsalar bile, tutarlı olmayabilirler.
- Operasyonel sistemlerde tarihsel raporlama için saklanan verinin yeterli kopyası olmayabilir.
- Son kullanıcılar mevcut depolarda neyin saklandığı bilgisine sahip olmayabilir.
- Kullanıcı, operasyonel veri tabanını sorgulayacak uzmanlığa sahip olmayabilir. Örneğin, IMS[9] veri tabanları özel bir tür veri yönlendirme dili kullanan bir uygulama programı gerektirir.
- Performans, bankalar için olan veri tabanlarında olduğu gibi, çoğu operasyonel veri tabanı için çok önemlidir. Sistem, kullanıcıların amaca özel sorgular yapmasının üstesinden gelemeyebilir.

[8] Uyarlanabilir veri tabanı yönetim sistemi, hem verinin bütünlüğünü koruyan hem de yüksek hareket işleme hızına sahip bir veri tabanı yönetim sistemidir. Özellikle aynı anda binlerce kullanıcının ihtiyacına saniyenin altında sürelerde karşılık verebilecek şekilde tasarlanmıştır.

[9] IBM Bilgi yönetim sistemi (Information Management System; IMS) bir bileşik hiyerarşik veri tabanıdır ve geniş hareket işleme kapasitesine sahip bir bilgi yönetim sistemidir.

- Operasyonel veri, genellikle, kullanıma en uygun şekilde değildir. Örneğin; ürüne, bölgeye ve sezonuna göre özetlenmiş satış verileri analist için ham veriden daha kullanışlıdır.

Veri ambarcılığı bu problemleri çözer ve operasyonel veriden çıkarılmış, karar verme için dönüştürülmüş bilgi içeren verinin depolarını oluşturur. Örneğin, bir veri ambarcılığı aracı tüm satış verilerini operasyonel veri tabanından kopyalayabilir, veriyi temizler, veriyi özetleyecek hesaplamalar gerçekleştirir ve özetlenmiş veriyi operasyonel veri tabanından bir ayrık veri tabanındaki (veri ambarındaki) bir hedefe yazar. Bu sayede kullanıcılar, ayrık veri tabanını operasyonel veri tabanlarına temas etmeden sorgulayabilirler.

Veri ambarcılığı kavramını anlamak zor değildir. Veri ambarcılığının ana fikir raporlama, analiz ve diğer iş zekâsı fonksiyonlarını desteklemek için ihtiyaç duyulan veriye kalıcı bir depolama alanı yaratmaktır. İlk bakışta, veriyi birden fazla yerde saklamak gereksiz görünebilir. Ne var ki faydaları bunu yapmanın emeğini ve maliyetini fazlasıyla karşılar.

Veri ambarı şu faktörleri işaret eder ve son kullanıcılara pek çok fayda sağlar:

- Son kullanıcının çok çeşitli veriye iyileştirilmiş erişimi,
- Arttırılmış veri tutarlılığı,
- Verinin ek belgelenme işlemi,
- Potansiyel olarak düşük hesaplama maliyetleri ve artan üretkenlik,
- Farklı kaynaklardan gelen ve birbiriyle alakalı verileri aynı yerde toplamayı sağlamak,
- Bilgisayar sistemlerindeki değişimleri destekleyen bir programlama altyapısının oluşturulması,
- Operasyonel sistemin performansını etkilemeden son kullanıcılara her seviyede özel amaçlı sorgulama veya raporlama yetkisi verme.

1.7.4 Veri Ambarında Hangi Veriler Bulunur?

İşlemsel veri tabanı sistemlerinde yüksek hacimli detaylı veriler bulunur. Bu verinin bir çekirdek alt kümesi, ilgilenilen konuya göre öncelikli olarak, veri ambarına aktarılır.

Veri ambarının temel bir aksiyomu veri ambarına aktarılan verinin hem sadece okunabilir hem de kalıcı olmasıdır. Veri ambarındaki verinin boyutu arttıkça, kullanıcının veriyi daha uzun vadeli analiz etmesini sağlayan, değeri de artar.

Operasyonel veri, genellikle, gerçek zamanlı veya gerçek zamanlıya yakın iken veri ambarındaki veri tarihseldir. Veri aktarım süreci belirli aralıklarla, muhtemelen günde bir kez ve gece yarısında, gözlenir. Veri ambarı, öncelikli olarak, nispeten büyük hacimli tarihsel verinin gelecekte ne yapılacağına karar vermek amacıyla raporlanması ve analiz edilmesi için kullanıldığından böyle bir aktarım çizelgesi yeterlidir.

1.7.5 Veri Ambarı Mimarisi

Veri temizleme ve veri bütünleştirmeyi içeren veri ambarının inşası, veri madenciliği için önemli bir ön işleme basamağı olarak görülebilir. Pek çok kaynaktan veri toplayan bir veri ambarı inşa etmek veri bütünlüğü problemini çözerek bazen yıllar süren ve milyonlarca dolara mal olan veriyi bir veri tabanına yükleme işlemini gerçekleştirir (Gray ve Watson, 1998). Ne var ki, veri madenciliği uygulamak için bir veri ambarı şart değildir. Eğer bir veri ambarı müsait değilse madencilik uygulanacak veri bir veya daha fazla operasyonel veya işlemsel veri tabanlarından veya veri marketlerinden alınabilir. Alternatif olarak, veri madenciliği için kullanılacak veri tabanı bir veri ambarının mantıksal veya fiziksel bir alt kümesi olabilir.

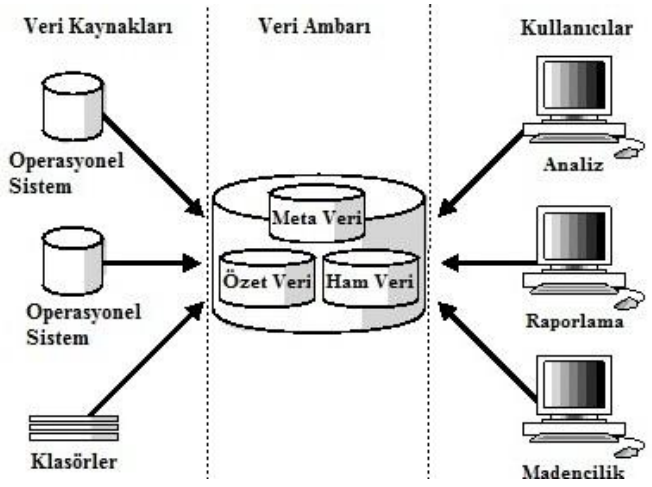
Veri ambarı mimarisi genellikle üç bileşenden oluşur:

1. **Veri Toplama Yazılımları:** Eski sistemlerden ve harici kaynaklardan verileri alıp, birleştirip, özetleyip bir veri ambarına yükleyen yazılımlardır.
2. **Veri Tabanı Yazılımı:** Veri ambarının sakladığı verinin ilişkilendirildiği, genellikle hedef veri tabanı olarak adlandırılan yazılımdır.
3. **İstemci Yazılım:** Kullanıcıların ambardaki veriye ulaşmalarına ve bu veriyi analiz etmelerine izin veren yazılımdır.

Veri ambarları ve onların mimarileri veri ambarını kullanan kuruluşların özelliklerine göre çeşitlilik gösterir. Yaygın olan üç mimari şunlardır:

- Temel veri ambarı mimarisi,
- Hazırlanma alanı içeren veri ambarı mimarisi,
- Hazırlanma alanı ve veri marketi içeren veri ambarı mimarisi.

Temel Veri Ambarı Mimarisi: Bu mimaride, son kullanıcılar çeşitli kaynak sistemlerden veri ambarına gelen veriye doğrudan ulaşırlar. Bu veri ambarında rapor almak, analiz yapmak ve geleceğe yönelik veri madenciliği çalışmaları yapmak çok zordur. Şekil 1.3 bir veri ambarı için temel mimariyi göstermektedir.

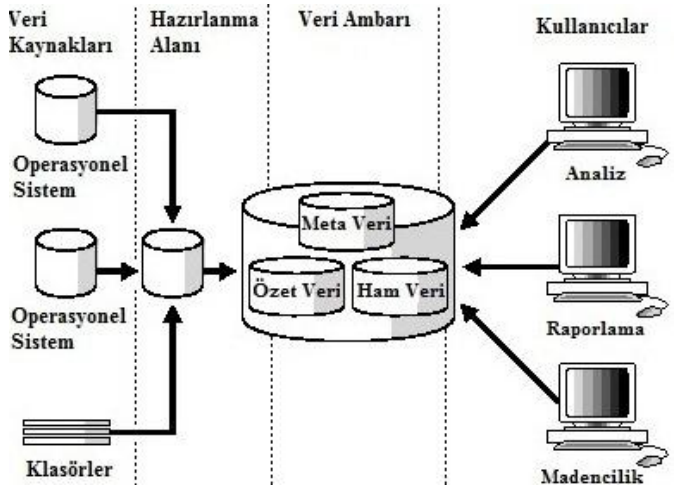


Şekil 1.3: Temel veri ambarı mimarisi (Oracle9i Data Warehousing Guide Release 2 (9.2)).

Şekil 1.3'de özet veriye ek olarak OLTP sistemlerinin geleneksel verileri olan meta (tanımlayıcı) veri ve ham veri de bulunmaktadır. Uzun işlemlerin hesaplanma süresini kısaltmalarından dolayı özet veriler çok önemlidirler.

Hazırlanma Alanı İçeren Veri Ambarı Mimarisi: Şekil 1.3'de görüldüğü üzere operasyonel verinin, veri ambarına konmadan önce temizlenmesi ve işlenmesi gerekir. Çoğu veri ambarı bu amaçla; verilerin geçici olarak saklandığı, veri ambarına gitmeden önce temizlendiği, kaliteli hale getirildiği, sınıflandırılabildiği ve sıralanabildiği yerler olan hazırlanma alanlarını kullanır.

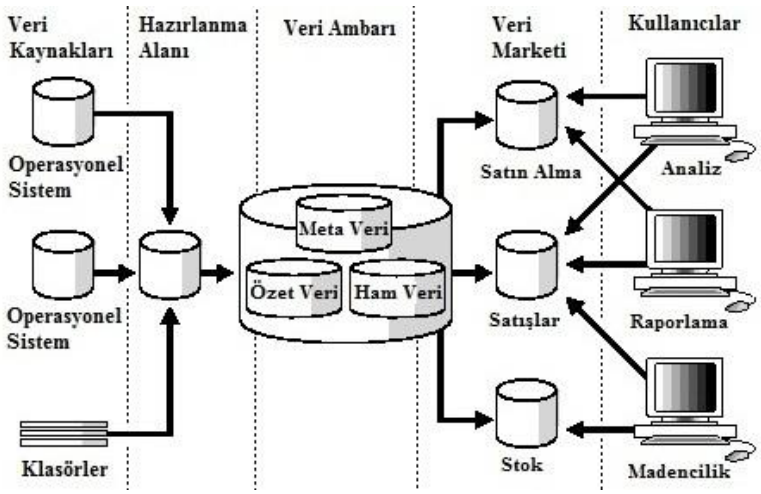
Hazırlanma alanı, veri özetlemeyi ve genel veri ambarı yönetimini kolaylaştırır. Şekil 1.4 bu tip mimariyi temsil etmektedir.



Şekil 1.4: Hazırlanma alanı içeren veri ambarı mimarisi (Oracle9i Data Warehousing Guide Release 2 (9.2)).

Hazırlanma Alanı ve Veri Marketi İçeren Veri Ambarı Mimarisi: Şekil 1.4'teki veri ambarı mimarisi yaygın olsa da kuruluşlar içindeki farklı gruplar için veri ambarı mimarisi özelleştirilmek istenebilir. Bu, işin belirli bir hattı için tasarlanmış sistemler olan veri marketleri eklenerek yapılabilir. Veri marketleri sayesinde iş yükü azaltılır, performans arttırılır, raporlama daha kolay ve raporlar daha ulaşılabilir hale gelir. Şekil 1.5 satın alma, satışlar ve stokların ayrıldığı bir örneği temsil etmektedir.

Bu örnekte bir finansal analist, satın alma ve satışlar için tarihsel veriyi analiz edebilir.



Şekil 1.5: Hazırlanma alanı ve veri marketi içeren veri ambarı mimarisi (Oracle9i Data Warehousing Guide Release 2 (9.2)).

BÖLÜM 2: VERİ MADENCİLİĞİ

2.1 VERİ TABANLARINDA BİLGİ KEŞFİ

Bilgi keşfi; veriden gelen, örtülü, daha önceden bilinmeyen ve potansiyel olarak faydalı olan bilgilerin önemsiz olamayan çıktılarıdır. Durumların (verilerin) bir kümesi F, bir dil L ve kesinliğin bir ölçüsü C olarak verilsin. Biz bir örüntüyü[10], F'nin bir F_S alt kümesi içindeki bağıntıları bir C kesinliğiyle açıklayan, L'deki bir S raporu olarak tanımlarız; öyle ki S, F_S'deki tüm durumların sayımından daha basittir. Kullanıcının ilgi ölçüsüne göre dikkate değer ve kullanıcının ölçütlerine göre yeterince kesin olan bir örüntüye bilgi denir. Bir veri tabanındaki durumların kümesini gözlemleyen ve örüntüler üreten bir programın çıktısı bu anlamda keşfedilmiş bilgidir (Piatetsky-Shapiro ve diğ., 1992). Veriden anlamlı örüntüler çıkarma sürecine literatürde, veri işleme sürecinde bilginin son ürün olduğunu vurgulamak için **Veri Tabanlarında Bilgi Keşfi - VTBK** *(Knowledge Discovery in Databases)* tanımlaması yapılmıştır. Geleneksel sorgu veya raporlama araçlarının, hacmi sürekli büyüyen bu veri yığınları karşısında yetersiz kalması üzerine "Bu verilerden nasıl faydalanılabilir?" sorusuna cevap arayan VTBK süreci ortaya çıkmıştır.

Veri tabanlarında bilgi keşfinin dört temel özelliği vardır:

1. **Üst Düzey Dil:** Keşfedilmiş bilgi üst düzey bir dille tasvir edilir. Bu dilin doğrudan insanlar tarafından kullanılmasına gerek yoktur, fakat söylemleri insanlar tarafından anlaşılabilmelidir.
2. **Kesinlik:** Keşifler kesin olarak veri tabanının içeriğini tasvir etmelidir. Bu tasvirin ne derece kusurlu olduğu kesinlik ölçüleriyle ifade edilir.

[10] Örüntü, olay ve nesnelerin düzenli bir biçimde birbirini takip ederek gelişmesidir.

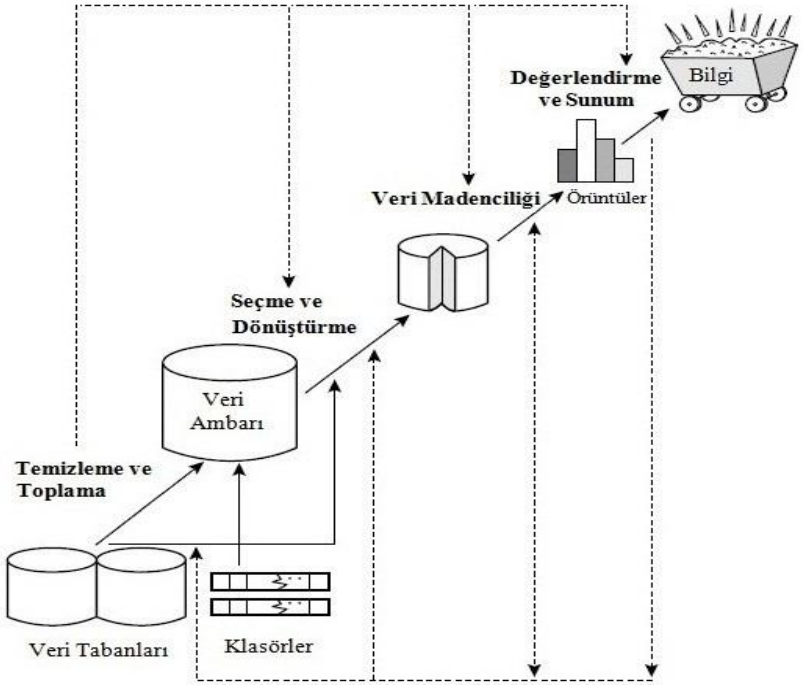
3. **Kayda Değer Sonuçlar:** Keşfedilen bilgi, kullanıcı tanımlı hükümlere göre kayda değerdir. Kayda değer olmakla kast edilen, örüntülerin yeni, potansiyel olarak kullanışlı oldukları ve keşif sürecinin önemsiz olmadığıdır.
4. **Etkinlik:** Keşif süreci etkindir. Büyük boyutlu veri tabanları için işlem süreleri tahmin edilebilir ve makuldür.

Kısaca VTBK süreci, veri analizi çabaları ve veri tabanlarındaki veri setlerinin içindeki örüntü ve düzenleri bulma teknikleriyle alakalıdır. VTBK'yı uygun bir şekilde yapmak, önceden bilinmeyen veya fark edilemeyecek kadar net olmayan örüntülerin açığa çıkarılması ihtimalini arttırır.

Büyük veri setlerindeki düşük seviyedeki veriden yüksek seviyede bilgi çıkarımını sağlamayı amaçlayan VTBK sürecinin en önemli basamağı modelin kurulması ve değerlendirilmesi aşamalarından oluşan veri madenciliğidir. Bu önem birçok araştırmacının, birbirine paralel süreçlere sahip olan VTBK ve veri madenciliğini eş anlamlı olarak da kullanmasına neden olmaktadır (Akpınar, 2000).

2.1.1 VTBK'nın Basamakları

VTBK, verinin nasıl saklanması ve algoritmaların büyük veri setlerine nasıl uygulanması gerektiği, sonuçların nasıl yorumlanacağı sorularının cevabını arama aşamalarıdır (Altıntop, 2006). Veri madenciliği ise, VTBK süreci ile elde edilen bilgi ve örüntüleri seçip çıkarmak için algoritmaların kullanılmasıdır. VTBK süreci basamakları Şekil 2.1 de verilmiştir:



Şekil 2.1: VTBK sürecinde bir basamak olarak veri madenciliği (Dunham, 2003).

VTBK sürecini oluşturan basamaklar kısaca şunlardır (Han ve Kamber, 2006):

1. **Veri Temizleme** (*Data Cleaning*)**:** Kirli, gürültülü ve tutarsız veri ayıklanır.
2. **Veri Bütünleştirme** (*Data Integration*)**:** Birden fazla veri kaynağı birleştirilebilir[11].
3. **Veri Seçme** (*Data Selection*)**:** Analize uygun olan ve konuyla ilgili veri, veri tabanından alınır.
4. **Veri Dönüştürme** (*Data Transformation*)**:** Veriler, özetleme veya toplama yoluyla madencilik için uygun biçimlere dönüştürülür veya birleştirilirler[12].
5. **Veri Madenciliği** (*Data Mining*)**:** Akıllı yöntemlerin, veri örüntülerini ortaya çıkarmak amacıyla uygulandığı önemli bir aşamadır.

[11] Enformasyon endüstrisinde; sonuçta elde edilen veriler veri ambarlarında toplanırken, veri temizlemenin ve veri bütünleştirmenin sürecin bir basamağı olarak uygulanması yaygın bir akımdır.
[12] Bazen, veri ambarı oluşturma ve kullanma işlemi anlamına gelen veri ambarcılığında, veri dönüştürme veya birleştirme işlemleri veri seçme sürecinden önce uygulanır. Ayrıca veri indirgeme de esas verinin doğruluğundan ödün vermeden daha küçük bir temsil elde etmek için uygulanabilir.

6. **Örüntü Değerlendirme** (*Pattern Evaluation*)**:** Keşfedilen bilgiyi temsil eden aranan örüntülerin bazı ilgi ölçülerine dayanılarak tanımlanmasıdır.
7. **Bilgi Sunumu** (*Knowledge Presentation*)**:** Veri madenciliği ile elde edilen bilginin kullanıcıya aktarımı için görselleştirme ve sunum tekniklerinin kullanılmasıdır.

2.1.2 Veri Madenciliği Tanımlar ve Temel Kavramlar

Verinin bol olduğu ve güçlü veri analiz araçlarının kıt olduğu duruma *veri zengini fakat bilgi yoksulu* durumu denir. Bu kavramı biraz daha açacak olursak pek çok sayıda büyük veri havuzlarında toplanmış ve depolanmış, hızla büyüyen, muazzam miktardaki veri güçlü araçlar olmadan insanların anlamasının mümkün olmadığı hale gelmiştir. Karar vericilerin büyük miktardaki veri içinde gömülü olan değerli bilgiyi açığa çıkaracak donanıma sahip olmamaları önemli kararların veri havuzlarında depolanmış bilgi yönünden zengin veriye bağlı olmaktan ziyade karar vericinin önsezisine bağlı olmasına neden olmaktadır. Bunun sonucunda, verilerin toplandığı büyük veri havuzları nadiren ziyaret edilen "veri mezarları" haline gelmişlerdir. Buna ek olarak uzman sistem[13] teknolojilerinin, kullanıcının veya ilgili alanın uzmanlarının bilgi tabanlarına bilgiyi elle girmelerine dayandığını düşünün. Ne yazık ki bu durumda karar alma süreci önyargılara ve hatalara meyilli ve son derece masraflı ve zaman alıcıdır. Bu ihtiyaçlar ve veri ve bilgi arasındaki giderek büyüyen uçurum veriyi analiz eden ve iş stratejilerine, bilimsel ve tıbbi araştırmalara katkıda bulunan önemli veri örüntülerinin ortaya çıkarılmasını sağlayan veri madenciliği kavramını ortaya çıkarmıştır.

Madencilik, gerçek anlamda, yer altındaki madenlerin araştırılması, çıkarılması ve işletilmesiyle ilgili teknik ve yöntemlerin uygulanma süreçlerinin bütününü tanımlayan bir terimdir. Veri madenciliği ise farklı çevreler tarafından çeşitli şekillerde tanımlanmıştır.

[13] Kullanıcılarına, uzmanların bilgi ve muhakeme yeteneklerine ulaşma ve bu yeteneklerden faydalanma olanağı veren bir bilgisayar paketidir

Gartner Group'a[14] göre veri madenciliği "Depolama ortamlarında saklanmış büyük boyutlu verilerin örüntü tanıma teknolojilerinin yanı sıra istatistiksel ve matematiksel teknikler kullanılarak elenmesiyle anlamlı yeni korelasyonları, örüntüleri ve eğilimleri keşfetme sürecidir". Diğer tanımlar ise şunlardır:

"Veri üzerinde, veri örüntülerin (veya modellerin) özel bir sıralamasını ortaya çıkaran veri analizi ve keşif algoritmalarından oluşan VTBK sürecinin bir adımıdır" (Fayyad ve diğ., 1996).

"Veri madenciliği, önceden bilinmeyen, geçerli ve etkin bilginin büyük veri tabanlarından çekilmesi ve daha sonra bu bilginin son iş kararlarını almak için kullanılmasını kapsayan bir süreçtir" (Cabenave diğ., 1998).

"Veri madenciliği, büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanılarak aranmasıdır" (Gürsoy, 2009).

"Veri madenciliği, büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işlemidir. Bu sayede veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik kestirimlerde bulunmak mümkün görülmektedir" (Özkan, 2013).

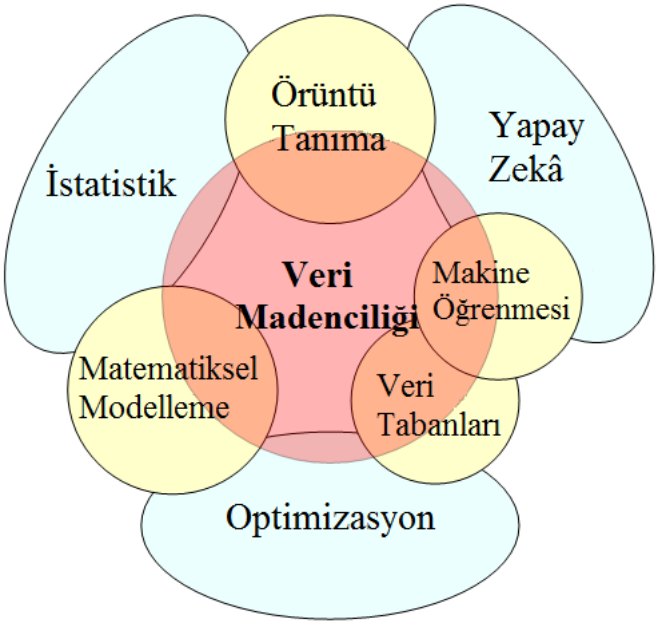
"Veri madenciliği (genellikle büyük) gözlemsel veri[15] setlerinin, önceden akla gelmeyen ilişkileri bulmak ve kullanıcıya göre hem anlaşılabilir hem de kullanışlı yeni yollarla özetlemek için analiz edilmesidir" (Hand ve diğ., 2001).

"Bir hipotezi ispatlamak (yukarıdan aşağıya veri madenciliği) veya kesin istatistiksel korelasyonlar üzerine kurulmuş yeni hipotezler üretmek amacıyla çok büyük veri tabanlarını sorgulama işlemidir" (Sullivan, 2012).

"Veri madenciliği, bilgisayar teknolojilerinin sağlamış olduğu çok hızlı veri işleme ve yüksek hacimde veri depolama imkanları yardımıyla ve farklı disiplinlerin (yapay zeka, makine öğrenmesi, uzman sistemler, veri tabanı teknolojileri, paralel bilgi işleme, dağıtık veri işleme, görselleştirme, optimizasyon, veri ambarcılığı, istatistik,…) katkısıyla sağlanan araçlarla, sahip olunan çok büyük hacimlerdeki veriden, karar vericinin etkin ve daha fazla bilgiye dayalı karar vermesinde kullanılabilmesi amacıyla önceden bilinmeyen, gizli, örtük, klasik metotlarla ortaya çıkarılması güç, faydalı, ilginç, anlaşılabilir ilişki, örüntü, bağıntı veya trendlerin

[14] Gartner Group, Inc, merkezi Birleşmiş Milletler 'in Connecticut Eyaleti'nin Stamford Şehri'nde olan bir bilgi teknolojileri araştırma ve danışma firmasıdır.
[15] "Gözlemsel veri" kavramı deneysel verinin zıttı anlamında kullanılmıştır.

otomatik veya yarı otomatik bir şekilde ortaya çıkarılması işlemidir" (Şentürk, 2006: 3).



Şekil 2.2: Veri madenciliği disiplinler arası bir çalışma alanıdır.

Esas itibarıyla modeller oluşturmak ve örüntüler bulmak amacıyla olmak üzere iki tip veri madenciliği yaklaşımı vardır. Model oluşturma ile ilgili olan yaklaşım, veri setlerinin büyük boyutlarından kaynaklanan problemler dışında, geleneksel istatistiksel araştırma yöntemlerine benzer. Örneğin, ileriki bölümlerde göreceğimiz, bir veri setinin ayrılması için kümeleme analizi; öngörüde bulunmak için regresyon analizi, vb. yöntemler. Model oluşturma yaklaşımında işlemsel ve mekanik modeller olmak üzere de bir ayrım yapılır. İşlemsel (deneysel) modeller model ilişkilerini, ilişkileri herhangi bir teoriye dayandırmadan araştırır. Mekanik (görüngüsel[16]) modeller ise temel veri üretme sürecinde bir teoriye veya mekanizmaya dayanır. Veri madenciliği, tanımı gereği, daha çok işlemsel modellerle ilgilenir (Jackson, 2002).

[16] Olayları, iç yüzünü ve temelindeki nedenleri düşünmeksizin dış görünümleri ile incelemeye ilişkin.

Örüntü tespiti ile ilgili olan ikinci tip veri madenciliği yaklaşımı; sahtekârlık tespiti için bir kredi kartı kullanımındaki olağan dışı harcamalar, EEG[17] işaretlerindeki düzensiz dalgalanmalar ve diğerlerinden farklı örüntü özelliklerine sahip nitelikler gibi alışılmadık örüntü davranışlarını tespit etmek için standart örüntüdeki küçük ama muhtemelen önemli sapmaları belirlemek ister. Bu tip yaklaşımlar veri madenciliğinin, veri yığınları içinden değerli bilgiyi araştırması kavramının ortaya çıkmasına yol açmıştır (Fayyad ve diğ., 1996).

Örüntü tespiti için kullanılan pek çok veri madenciliği algoritması esas örüntüyü tarayarak anlamsız özelliklerin etkilerini ayırt edebilirken, veri tabanlarının içerdikleri verilerin süreksiz, gürültülü, muğlak ve eksik olması gibi anormalliklerin sayısı arttığında madencilik algoritmalarının öngörü gücü azalabilir. Bu anlamda ticari veri tabanları karmaşıklıklarından dolayı örüntü çıkarımında benzersiz bir problem arz eder.

2.1.3 Denetimli Ve Denetimsiz Öğrenme

Yazılımların, algoritmalar kullanarak bazı sonuçlar ve kurallar çıkarmak için eldeki verileri incelemesi işlemine *öğrenme* denir. Veri madenciliği ise hiçbir ön kabulde bulunmadan verilerden öğrenmek demektir.

Veri madenciliğinde veriden öğrenme, denetimli ve denetimsiz öğrenme olmak üzere iki şekilde gerçekleşir. Yönlendirilmiş veri madenciliği olarak da adlandırılan denetimli öğrenmede hedef, açıklayıcı (bağımsız) değişkenler ve açıklanan (bağımlı) değişken(ler) arasındaki ilişkileri tanımlamaktır. Veri madenciliğinde denetimli öğrenmede, eldeki giriş verileri ile sonuç verileri arasında bir ilişki kurabilmek için veri setinin yeterince büyük bir kısmının bağımlı değişken değerleri ve o verilerden çıkan sonuçlar bilinmelidir.

[17] EEG ya da Elektroansefalografi, beyin dalgaları aktivitesinin elektriksel yöntemle izlenmesini ölçen yöntemdir.

Yönlendirilmiş veri madenciliği uygulanırken eldeki veri; eğitim, test ve doğrulama verileri olmak üzere birbiriyle ortak veri içermeyen üç gruba ayrılır. Eğitim verisi bir model kurmak ve gerekli parametreleri tahmin etmek için kullanılır. Test verisi seçilmiş bir örneklemden elde edilmiş modelin diğer verilere genelleştirilip genelleştirilemeyeceğini görmeye yardım eder. Test verisini kullanmak, özellikle, aşırı öğrenme[18] durumundan kaçınmayı sağlar. Doğrulama verisi ise parametreleri iyileştirmek, çeşitli yöntemleri değerlendirmek ve uzun vadede en iyi işi yapan modeli seçmek için kullanılır (Gentle ve diğ., 2004: 801).

Denetimsiz öğrenme veya bir öğretmen olmadan öğrenme, veri madenciliği tanımına daha uygundur. Denetimsiz öğrenmede değişkenler bağımlı ve bağımsız değişkenler olarak ayrılmaz ve bağımlı değişkenin değeri ya bilinmiyordur ya da çok az durum için kaydedilmiştir. Denetimsiz öğrenme süresince çıkış değerlerinin girilmesine gerek yoktur, sadece örnek giriş değerleri verilir ve sistemin örnekler arasındaki ilişkileri kullanarak kendi kendisine öğrenmesi sağlanır (Valpola, 2000; Gentle ve diğ., 2004: 791).

2.1.4 Veri Madenciliği Uygulama Alanları

Bu bölümde, günlük hayatta pek çok kullanım alanı olan veri madenciliğinin yaygın olarak kullanıldığı bazı alanlar ve bu alanlardaki uygulamalara yer verilmiştir (Akpınar, 2000; Özçınar, 2006; Baykal 2006).

Pazarlama

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
- Posta kampanyalarında cevap verme oranının artırılması,

[18] İstatistikte ve makine öğrenmesinde, bir istatistiksel model veriler arasındaki temel ilişki yerine bir rassal hatayı veya gürültüyü tanımladığında aşırı öğrenme (*overfitting*) gözlenir. Aşırı öğrenme, modelin genel eğilimlerden öğrenmek yerine eğitim verisini ezberlemeye başlamasıdır.

- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi *(Market Basket Analysis),*
- Müşteri ilişkileri yönetimi *(Customer Relationship Management),*
- Müşteri değerlendirme *(Customer Value Analysis),*
- Satış tahmini *(Sales Forecasting).*

Bankacılık

- Usulsüzlük tespiti,
- Risk analizleri,
- Risk yönetimi,
- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi.

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri tipinin belirlenmesi.

Perakendecilik

- Satış noktası veri analizleri,
- Alış-veriş sepeti analizleri,
- Tedarik ve mağaza yerleşim optimizasyonu.

Borsa

- Hisse senedi fiyat tahmini,
- Genel piyasa analizleri,
- Alım-satım stratejilerinin optimizasyonu.

Telekomünikasyon

- Kalite ve iyileştirme analizleri,
- Hisse tespitlerinde,
- Hatların yoğunluk tahminleri.

Biyoloji, Tıp, Genetik ve İlaç

- Test sonuçlarının tahmini,
- Ürün geliştirme,
- Tıbbi teşhis,
- Tedavi sürecinin belirlenmesi.

Endüstri

- Kalite kontrol analizlerinde,
- Lojistik,
- Üretim süreçlerinin optimizasyonunda.

Bilim ve Mühendislik

- Deneysel veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesi.

Eğitim

- Öğrenci davranışlarının öngörülmesi,
- Öğrencilerin ders seçme eğilimlerinin belirlenmesi.

2.1.5 Türkiye'de Veri Madenciliği Örnekleri

Dünyada olduğu gibi ülkemizde de yaygınlaşan veri madenciliği çalışmalarına dair bazı örneklere bu bölümde yer verilmiştir.

1. Abdullah Baykal tarafından 2007 yılında yapılan çalışma ile Dicle Üniversitesi web/mail sitesi günlükleri üzerinde, veri madenciliği uygulamalarından biri olan web içerik madenciliği teknikleri kullanılarak saldırı amaçlı yapılan erişimler tespit edilmiştir (Baykal, 2007).

2. Cebrail Çiftlikli ve çalışma arkadaşları tarafından 2007 yılında yapılan çalışmada bulanık mantık kullanılarak elde edilen, cep telefonu kullanıcılarına ait verilerin karar ağaçları ile analiz edilmesi ile pazarlama araştırmasında kullanılabilecek, gerçeğe daha yakın ve daha doğru sonuçlar elde edilmiştir (Çiftlikli ve diğ., 2007).

3. Ali Serhan Koyuncugil ve Nermin Özgülbaş tarafından 2008 yılında yapılan çalışmada İstanbul Menkul Kıymetler Borsası'nda işlem gören KOBİ'lere ait veriler üzerinde karar ağaçları kullanılarak KOBİ'lerin güçlü ve zayıf yönleri tespit edilmiştir (Koyuncugil ve Özgülbaş, 2008).

4. Tuğçe Üner tarafından 2010 yılında yapılan çalışmada otomotiv sektörü müşterilerine ait bilgiler veri madenciliği ile analiz edilerek Müşteri İlişkileri Yönetimi (MİY) ve Elektronik Müşteri İlişkileri Yönetimi (e-MİY)'in önemi ortaya çıkarılmıştır (Üner, 2010).

5. Abdulkadir Özdemir ve çalışma arkadaşları tarafından 2010 yılında yapılan çalışmada Gümüşhane Devlet Hastanesi veri tabanından alınan verilere veri madenciliği uygulanmış ve oluşan örüntülerden hastanenin hasta sayısı, hasta ve hastalık portföyü, poliklinik hasta ilişkileri tespit edilmiştir (Özdemir ve diğ., 2010).

6. Mehpare Timor ve çalışma arkadaşları tarafından 2011 yılında yapılan çalışmada hazır giyim perakende sektöründe faaliyet gösteren bir firmanın müşterilerinin verileri kullanılarak birliktelik kuralları analizi ile müşterilerin alışveriş

alışkanlıkları belirlenmeye çalışılmış ve kümeleme analizi ile müşteriler, demografik özellikleri dikkate alınarak bölümlendirilmiştir (Timor ve diğ., 2011).

7. Özcan Asilkan tarafından 2011 yılında yapılan çalışmada veri madenciliği tekniklerinden regresyon analizi ve yapay sinir ağları kullanılarak ikinci el otomobillerin pazardaki güncel fiyatları modellenmiş ve veri madenciliğinin, ikinci el otomobillerin güncel pazar fiyatlarını modellemede başarılı olarak kullanılabileceğini ortaya konmuştur (Asilkan, 2011).

8. Mehmet Ali Alan tarafından 2012 yılında yapılan çalışmada Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü lisansüstü öğrencilerine ait verileri kullanan bir sınıflandırma algoritması ile öğrencilerin notlarını etkileyen ve etkilemeyen faktörler tespit edilmiştir (Alan, 2012).

9. Özgün Çöllüoğlu Gülen ve Selçuk Özdemir tarafından 2013 yılında yapılan çalışmada Yasemin Karakaya Bilim ve Sanat merkezinde okuyan 12 yaş ve üzeri özel yetenekli öğrencilere ait veriler kullanılarak özel yetenekli öğrencilerin ilgi alanları öngörülmüş ve bu alanlar arasındaki ilişkiler keşfedilmiştir. Bu sayede bilim ve sanat merkezlerine, bireysel ihtiyaçları karşılayarak farklılaştırılmış öğretim verme ve ders programlarını daha etkili bir şekilde düzenleme gibi pek çok fayda sağlanmıştır (Gülen ve Özdemir, 2013).

2.2 VERİ MADENCİLİĞİNİN İŞLEV VE GÖREVLERİ

İstatistik ve makine öğrenmesi ait pek çok tekniği kullanan veri madenciliğinin işlev ve görevleri; özetleme, tanımlayıcı (*descriptive*) modelleme ve tahmin edici (*predictive*) modelleme olmak üzere üç ana başlık altında toplanabilir.

2.2.1 Özetleme

Veri madenciliğinin amacı bazen veriyi üreten unsurlar (insanlar, ürünler, süreçler, vs.) hakkındaki bilgimizi arttırmak kadar basit olabilir. Veriyi tanımak; veri madenciliğinin veri hakkında ne söyleyebileceği, hangi alanlarda kısıtlamaların olacağı ve ileride hangi analiz basamaklarının kullanılmasının uygun olacağı hakkında fikir sahibi olmak için önemlidir.

Her veri analizi uygulamasının başlangıcında verinin genel eğilimlerini ve uç değerlerini görmek için veriyi gözden geçirmek gerekir. Bu amaçla, verinin özetlenmesi gerekir. Örneğin bir GSM firması müşterilerinin telefon görüşmeleri toplam dakika, toplam harcama, toplam arama sayısı, vs. şeklinde özetlenerek gözden geçirilebilir.

Özet tabloları, basit tanımlayıcı istatistikler ve basit grafikler kullanarak veriyi gözden geçirme analiste verinin, veri kalite problemleri ve gerekli alt yapı bilgisi gibi bazı belli başlı özelliklerini verecektir.

Verinin tabiatını anlama ve veri içinde saklı bilgiyi bulmak için kullanılacak potansiyel hipotezleri bulma konusunda analiste yardımcı olan ve her ne kadar doğru olmasa da basit görevler olarak tanımlanan veri özetleme ve veri tanımlama gibi görevler için araştırmacılar çok fazla zaman harcamak istemezler. Bu nedenle özetleme ve tanımlama basamaklarında veri işleme hızı çok önemlidir. Bir frekans tablosu veya bir serpilme diyagramı saniyeler içinde görülebilmelidir ki bu sadece bazı bilgisayar programları kullanılarak başarılabilir. Diğer bir önemli nokta da tüm değişkenlerin hızla taranmasıdır. Eğer bir program, oluşturulan tablonun veya grafiğin açık ve detaylı bir tanımına ihtiyaç duyuyorsa, kullanıcı birkaç defadan sonra bu yorucu uğraşı sonlandırır. Konuya özel fonksiyonlar ve değişken türüne bağımlı cevaplar bu tür problemler için uygun bir çözüm sağlar (Gentle ve diğ., 2004: 792-793).

2.2.2 Tanımlayıcı Modelleme

Tanımlayıcı modellemenin amacı; var olan verileri yorumlayarak davranış biçimleri ile ilgili tespitler yapmak ve bu davranış biçimini gösteren alt veri setlerinin özelliklerini tanımlamaktır, belirli bir hedefi tahmin etmek değildir. Bunun bir sonucu olarak, tanımlayıcı modelleme denetimsiz öğrenme kapsamında değerlendirilir. Tanımlayıcı modelleme ile veri setinde yer alan veriler arasındaki ilişkiler, bağlantılar ve davranışlar bulunur. Tanımı bilmek, tekrarlanan bir faaliyette veya tanımı bilinen yeni bir verinin yapıya katılmasında ne şekilde hareket edileceği konusunda karar almaya destek olur (Arguden ve Erşahin, 2002: 39).

Karakteristik tanımlayıcı modelleme yöntemleri şunlardır:

- Kümeleme *(Clustering),*
- Birliktelik Kuralları *(Association Rules)* ve
- Ardışık Örüntüler *(Sequential Patterns).*

2.2.2.1 Kümeleme

Nesneleri kümelemek insanın, insanların ve eşyanın belli başlı özelliklerini tasvir etmek ve onları bir sınıfla tanımlamak ihtiyacı kadar eskidir. Bu nedenle kümeleme matematik ve istatistikten, biyoloji ve genetiğe kadar pek çok bilimsel öğretiyi kucaklar. Biyolojide "taksonomi "den tıpta "sendrom" a ve genetikte "genotip" ten imalatta "grup teknolojisi" ne kadar tüm alanlarda üzerinde çalışılan problem, birimlerin sınıflarını oluşturmak ve birimleri uygun sınıflara atamaktır (Maimon ve Rokach, 2010: 270-271).

Kümeleme, bir veri kümesindeki veriler içinde çeşitli özellikler itibariyle birbirine benzeyenlerin aynı sınıflarda toplanması ile veri kümesinin alt kümelere ayrılması işlemidir. Kümelemedeki amaç alt kümeler içi benzerlikleri maksimize

(küme içi uzaklıkları minimize) etmek ve alt kümeler arası benzerlikleri minimize (kümeler arası uzaklıkları maksimize) etmektir. Kümelemede tahmin edici modelleme yöntemlerinden biri olan sınıflandırmadan farklı olarak; başlangıç aşamasında veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı bilinmemekte olup, konunun uzmanı olan bir kişi tarafından kümelerin neler olacağı tahmin edilmektedir.

Kümeleme matematiksel olarak şöyle gösterilebilir: S kümesinin alt kümelerinin bir ailesi $C = \{C_1, \dots, C_k\}$'dır, öyle ki $i \neq j$ için $S = \cup_{i=1}^{k} C_i$ ve $C_i \cap C_j = \emptyset$'dir. Sonuç olarak klasik küme teorisinde S'deki her bir eleman yalnız ve yalnız bir kümeye aittir (Maimon ve Rokach, 2010:269).

Kümelemeye gerçek dünyadan bir örnek verecek olursak; bir banka müşterilerini yaşlarının, gelirlerinin, ev sahibi olup olmama, vs. durumlarının benzerliklerine göre gruplandırabilir ve bir gruptaki müşterilerin genel özelliklerini o gruptaki müşterileri tanımlamak için kullanabilir. Kümeleme bankaya, müşterilerini daha iyi anlamasında yardım edebilir ve böylece banka daha uygun ve kişiye özel hizmetler sunabilir. İş ve araştırma sahalarında kümelemenin görevlerine dair diğer örnekler olarak şunlar verilebilir (Larose, 2005: 17):

- Küçük bütçeli ve düşük sermayeli bir işletme için belli bir müşteri grubunun tercih ettiği bir ürünün hedef kitle pazarlamasında,
- Muhasebe denetim amaçlı; mali davranışları tehlikesiz ve şüpheli sınıflarına ayırmak için,
- Veri kümelerinin yüzlerce özelliği olduğu zaman bir boyut indirgeme aracı olarak,
- Çok sayıda gen benzer davranış sergilediğinde, genlerin dışavurumlarının kümelenmesi için kullanılabilir.

Kümeleme, çoğu zaman, veri madenciliği sürecinde sonuç olarak çıkan kümelerin ileriki basamaklarda yeni girdiler olarak kullanıldığı bir ön hazırlık basamağı olarak kullanılır.

2.2.2.2 Birliktelik Kuralları

Birliktelik, verilerin birlikteliğinin veya bağıntısının keşfedilmesidir. Bu tür bir birliktelik veya bağlantı, bir birliktelik kuralı olarak adlandırılır. Bir birliktelik kuralı, bir veri kümesindeki veriler arasında yüksek sıklıkta birlikte görülen özelliklere ait ilişkisel kuralları açığa çıkarır. Bu tanımdan faydalanarak bir veri tabanındaki, bir veri kümesinin ortaya çıkmasının diğer veri kümelerinin ortaya çıkmasıyla yakından ilgili olduğunu söyleyebiliriz.

Veri madenciliğinde birlikteliğin görevi hangi özelliklerin birbirine uyduğunu bulmaktır. İş dünyasında yaygın olarak birliktelik analizi veya pazar sepeti analizi olarak bilinen birliktelik, iki veya daha fazla özellik arasındaki ilişkiyi sayısallaştırmaya yarayan kuralları açığa çıkarmaya çalışır. Sonuç olarak bir birliktelik kuralı, *destek* ve *güven* kriterleri ile birlikte, "eğer *önce gelen,* sonra *takip eden*" şeklinde ifade edilir.

Destek kriteri, ürünler arasındaki bağlantının ne kadar sık olduğunu belirtir ve güven kriteri de tanımlanan kuralın kabul edilebilirliğini gösterir. Kullanıcı tarafından belirlenen minimum destek eşik değerini ve minimum güven değerini aşan birliktelik kuralları dikkate alınır.

Büyük veri tabanlarında birliktelik kurallarının bulunması iki aşamalı bir süreçtir. İlk aşamada her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanan öğeler bulunur. İkinci aşamada sık tekrarlanan öğeler arasında çok büyük bir destek ve yüksek seviyeli bir güven kriterine sahip güçlü birliktelik kuralları oluşturulur.

Birliktelik kuralları madenciliğinin ilk örnek uygulaması olan "Pazar Sepeti Analizi"nin hedefi, müşterinin pazar sepetinin içeriğini analiz ederek satış oranlarını yükseltmek ve kârı maksimize etmektir. Birliktelik kuralları kullanılarak birlikte

alınan belli ürünlerin hangilerinde indirim yapılabileceği bulunabilir veya birlikte satılan ürünleri birbirlerine yakın yerlere yerleştirerek mağazaya A ürününü almaya gelen bir müşteriye ayrıca B ürününe de ihtiyacı olabileceğini hatırlatılabilir. Örneğin bir süper market Perşembe gecesi alış veriş yapan 1000 müşterisinden 200 tanesinin bebek bezi aldığını ve bebek bezi alan bu 200 müşteriden 50 tanesinin de sigara aldığını bulabilir. Böylece birliktelik kuralı $^{200}/_{1000} = \%20$ destek ve $^{50}/_{200} = \%25$ güven oranıyla "eğer bebek bezi alırsa sigara da alır" olacaktır. Bu verilere sahip olan marketler, birlikte satılan ürünleri yakın raflara koyarak, katalogda birlikte satılan ürünlerin birlikte görülmesini sağlayarak veya müşteriler için cazip ürün paketleri oluşturarak satışları artırabilirler.

İş ve araştırma sahalarında birliktelik kurallarının görevlerine dair diğer örnekler olarak şunlar verilebilir (Larose, 2005: 36):

- Bir cep telefonu operatörünün paket yükseltme teklifine olumlu yanıt veren abonelerin oranlarının incelenmesi,
- Ebeveynleri kendilerine kitap okuyan ve ebeveynlerinin kendileri de iyi okuyucular olan çocukların oranı,
- Telekomünikasyon ağlarındaki verim kaybının öngörülmesi,
- Yeni bir ilacın tehlikeli yan etkiler sergileyebileceği durumların oranının tespiti.

2.2.2.3 Ardışık Örüntüler

Birliktelik kurallarından farklı olarak, işlem hareketlerini zaman ve mekân faktörlerini de dikkate alarak gösteren örüntülere *ardışık örüntüler* denir. Bu anlamda bir müşterinin birinci gün A ürününü, onu izleyen gün veya günlerin birinde B ürününü ve daha sonraki bir gün de C ürününü alması zaman içinde ardışık bir örüntü oluşturur.

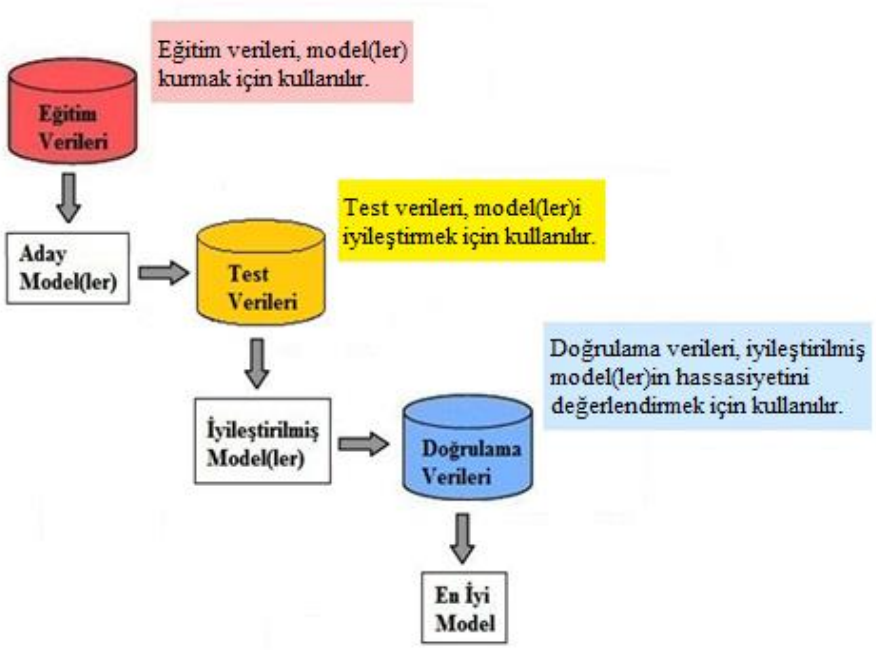
Verilen bir ardışık örüntüler kümesindeki sık gözlenen tüm ardışık alt örüntülerin bulunması işlemi olan ardışık örüntü madenciliği:

- Minimum destek kriterini sağlayan tüm örüntüleri bulabilir,
- Az sayıda veri tabanının tarandığı durumlarda son derece etkilidir,
- Kullanıcı tanımlı çeşitli kısıtlar eklenmesine uyumludur.

2.2.3 Tahmin Edici Modelleme

Tahmin edici modellemenin amacı geçmiş verilerden yararlanarak, yeni veya henüz tam anlaşılmamış verileri önceden tanımlanmış bir sınıfa atamaktır. Tahmin edici modelleme için girilen veriler, verinin temel özelliklerini tanımlayan açıklayıcı değişkenler ve değeri tahmin edilecek olan açıklanan değişken(ler) olmak üzere ikiye ayrılırlar. Bunun bir sonucu olarak, tahmin edici modelleme denetimli öğrenme kapsamında değerlendirilir.

Tahmin edici modelleme sürecinin basamakları Şekil 2.3'te verilmiştir.



Şekil 2.3: Tahmin edici modelleme süreci (Jackson, 2002).

Tahmin edici modelin türünü, açıklanan değişkenin tabiatı belirler. Eğer açıklanan değişken kesikli ise tahmin için sınıflandırma modeli kullanılır, sürekli ise regresyon modeli kullanılır.

Karakteristik tahmin edici modelleme yöntemleri şunlardır:

- Sınıflandırma *(Classification*) ve
- Regresyon *(Regression)*.

2.2.3.1 Sınıflandırma

Sınıflandırma, bir verinin özelliklerine dayanarak o verinin sınıfını belirleyen bir fonksiyonun veya modelin türetilmesidir. Modeli kurabilmek için sonuçları önceden bilinen durumlar ve bu durumlarda ilgili faktörlerin aldığı değerlerin bir veriler kümesi, eğitim kümesi olarak verilir. Eğitim verileri dışında kalan verilerin bir kümesi de modelin doğruluğunu belirlemek için kullanılmak üzere test kümesi olarak belirlenir. Bir sınıflandırma fonksiyonu veya modeli, özellikler ve eğitim kümesindeki verilerin sınıfları arasındaki ilişkiler analiz edilerek kurulur. Böyle bir sınıflandırma fonksiyonu veya modeli kümeye eklenen yeni verileri sınıflandırmak ve veri tabanındaki verilerin sınıflarının daha iyi bir anlayışını geliştirmek için kullanılabilir. Örneğin, eğitim kümesi vazifesi gören teşhis konulmuş hastalar kümesinden bir hastanın hastalığının ne olduğuna, onun teşhis verisinden ulaşabilecek bir sınıflandırma modeli kurulabilir. Bu sınıflandırma modeli, yeni bir hastanın hastalığını da yaş, cinsiyet, ağırlık, vücut ısısı, kan basıncı, vs. gibi teşhis verilerine dayanarak teşhis edebilir (Yongjian, 1997).

Sınıflandırma matematiksel olarak şöyle tanımlanabilir: $x = (x_1, \dots, x_n)$ veri ile ilgili değişkenler kümesi ve $y = (y_1, \dots, y_m)$ değerleri bilinen kesikli değişkenlerin sınıflarının sonlu bir kümesi olmak üzere, sınıflandırma $f : x \rightarrow y$ fonksiyonun öğrenilmesi işlemidir.

2.2.3.2 Regresyon

Sınıflandırmadan farklı olarak, sürekli bağımlı değişkenlerin değerlerini tahmin etmek için kullanılan bir yöntemdir. Regresyon, sürekli hedef (bağımlı) değişkenin (y) değerinin bir veya daha fazla tahminci (bağımsız) değişkenin ($x_1, \dots, x_n$) bir fonksiyonu (F) olarak hesaplanması sürecidir. Örneğin; emlak fiyatlarının tahmin edildiği bir regresyon modeli belli bir zaman diliminde gözlenmiş emlak satışlarından toplanmış; konutların yaşı, metrekaresi, oda sayısı, vergisi, tutulu satış (*mortgage*) oranları, okullara veya alış veriş merkezlerine yakınlığı, vs. gibi verilere dayanarak geliştirilebilir.

Regresyon matematiksel olarak şöyle tanımlanabilir: Bağımsız değişken(ler)in katsayıları olan parametreler ($\theta_1, \dots, \theta_n$), bağımsız değişken(ler)in bağımlı değişken üzerindeki etkisinin bir ölçüsünü göstermek ve bağımlı değişkenin beklenen ile tahmin edilen değeri arasındaki fark hata terimi (e) olmak üzere, $y = F(x, \theta) + e$ eşitliğini sağlayan en iyi F fonksiyonu için θ değerlerinin araştırılması sürecidir.

2.2.4 Veri Madenciliği Ve İstatistik

İstatistik ve veri madenciliği bilim dallarının her ikisi de veri içindeki yapıyı keşfetmeyi, veriden öğrenmeyi ve verinin bilgiye dönüştürülmesini amaçlamaktadır. Amaçları o denli örtüşür ki kimileri veri madenciliğini istatistiğin bir alt kümesi olarak tanımlar. Fakat veri madenciliği, özellikle istatistikçilerin ağırlıklı olarak ilgilenmedikleri, veri tabanı teknolojisi ve makine öğrenmesi gibi diğer bilim dallarının araç ve yöntemlerini kullandığından bu düşünce gerçekçi değildir (Hand, 1998). Buna rağmen istatistiksel süreçler veri madenciliğinde, özellikle model kurma ve geliştirmede, büyük rol oynarlar. Öğrenme algoritmalarının çoğu, kuralları veya karar ağaçlarını inşa ederken ve ayrıca aşırı öğrenen modelleri düzeltirken istatistiksel testleri kullanırlar. Veriden öğrenmenin veya veriyi bilgiye

dönüştürmenin bu iki yaklaşımının birbirini tamamlayıcı özellikleri şöyle özetlenebilir:

- Veri madenciliği ile elde edilmiş, önemli ilişkileri ve eğilimleri tanımlayan, bilgi bu keşiflerin neden kullanışlı olduklarını ve ne ölçüde geçerli olduklarını açıklayamaz. İstatistiksel analizin doğrulayıcı araçları keşfedilen bilgileri doğrulamak ve bu keşifler üzerine kurulu kararların kalitesini hesaplamak için kullanılabilir.
- İstatistiksel bir analiz uygulayarak veriler arasında gözlemlenmiş davranış için olası açıklamalar tasarlanır ve bu hipotezlerin analiz edilecek veriyi dikte etmelerine izin verilir. Sonra, veri madenciliği uygulanarak verinin test edilecek yeni hipotezler sunması sağlanır.

Veri madenciliğinin istatistik ile benzer yönlerinin yanında birçok farklı yönü de mevcuttur. Bu farkları şöyle sıralayabiliriz (Ohri, 2011; Moss, 2009; Brusilovsky, 2011):

1. İstatistiksel analiz problemleri iyi yapılandırılmış[19], veri madenciliği problemleri ise yapılandırılmamıştır[20].
2. İstatistik matematiksel formüllerin bir kümesi iken veri madenciliği algoritmaların bir kümesidir.
3. İstatistikte ana kütle küçük ($n \leq 10^4$) ve aynı cinsten (homojen) birimlerin meydana getirdiği topluluktur, veri madenciliğinde ise ana kütle büyük ($n \geq 10^4$) ve heterojendir.
4. İstatistiksel analizde sinyal gürültü oranı[21] $S/N > 3$ iken veri madenciliğinde bu oran $0 < S/N \leq 3$'tür.

[19] İyi yapılandırılmış problemler; yüksek oranda bütünlükle tanımlanabilir, yüksek oranda kesinlikle çözülebilir, genellikle uzmanların en iyi yöntem ve en iyi çözüm üzerinde uzlaştıkları, kolaylıkla sayısal karşılığına tercüme edilebilir, amacın en iyi sonucu bulmak olduğu, çok basitten karmaşığa doğru giden problemlerdir.
[20] İyi yapılandırılmamış problemler; yüksek oranda bütünlükle tanımlanamayan, yüksek oranda kesinlikle çözülemeyen, uzmanların en iyi yöntem ve en iyi çözüm üzerinde uzlaşamadıkları, kolaylıkla sayısal karşılığına tercüme edilemeyen, amacın mantıklı çözümü bulmak olduğu, karmaşıktan çok karmaşığa doğru giden problemlerdir.

5. İstatistiksel analiz sadece nümerik (kesikli veya sürekli) verileri kullanırken veri madenciliği araçları sadece nümerik veriyi değil diğer veri türlerini de (metin, döngüsel, sıralı, mantıksal, vs.) kullanabilirler.
6. İstatistiğin önsel bilgi sağlanmasına ihtiyacı vardır. Veri madenciliği ise önsel bilgi üretir.
7. Veri madenciliği başlangıçta hedefe ihtiyacı olmayan bir inceleme aracıdır. İstatistiğin ise bir hedef fonksiyonuna ihtiyacı vardır.
8. Eğer gerekli şartlar sağlanırsa istatistik bir optimum üretir, veri madenciliği üretmez.
9. Veri madenciliği karmaşıklıkla, eksik veriyle, dinamiklerle ve bilinmeyen ana kütle karışımlarıyla başa çıkmayı bilir. İstatistik bunlar hakkında çok daha sınırlıdır.
10. İstatistiksel analiz önsel bilgiye dayalı bir hipotez ile başlar ve hipotezin onaylanması veya ret edilmesi üzerine kanıt arar, analiz tipi doğrulayıcıdır. Veri madenciliğinin görevi daha sonra istatistik tarafından çalıştırılmak üzere hipotezleri tanımlamaktır, analiz tipi arama amaçlıdır.

Veri madenciliğinde kullanılan bazı yaygın istatistiksel analiz teknikleri aşağıda verilmiştir (Jackson, 2002).

Tanımlayıcı Ve Betimleyici Yöntemler: Ortalamalar; varyans ölçüleri, dağılımlar, olasılık tabloları ve basit korelasyonlar gibi verinin yapısını anlamak için kullanışlı, temel tanımlayıcı istatistiksel yöntemlerdir. Betimleme de çok miktarda veriyi yorumlamak için histogramlar, kutu grafikleri ve dağılım grafikleri gibi betimleme araçlarını kullanan temel bir keşif yöntemidir.

Kümeleme Analizi: Nispeten homojen gruplar veya kümeler oluşturabilmek için değişkenlerle ilgili bilgileri düzenlemeye çalışır. Oluşturulan kümeler kendi içlerinde

[21] Sinyal gürültü oranı, değişim katsayısının çarpmaya göre tersidir. Yani bir ölçümün ortalamasının standart sapmasına oranıdır.

homojendirler, yani elemanlar birbirine benzer ve dışlarında heterojendirler, yani bir kümenin elemanı diğer kümelerin elemanlarına benzemez (Jackson, 2002).

Korelasyon Analizi: İki değişken arasındaki ilişkinin gücünü ölçer. Bulunan korelasyon katsayısı bir değişkende gözlenen değişimin başka bir değişkende sebep olacağı değişimi gösterir. İki değişken arasındaki korelasyonun kıyaslanmasında amaç, bağımsız değişkendeki değişimin bağımlı değişkende değişime neden olup olmayacağını görmektir. Bu bilgi bağımsız değişkenin öngörü yeteneklerini anlamamıza yardım eder. Korelasyon bulguları, regresyon bulgularında olduğu gibi, nedensel ilişkileri analiz etmekte kullanılabilirler, fakat kendi başlarına nedensel örüntüler oluşturamazlar.

Ayırma Analizi: Bir takım tahmincilerden gelen iki veya daha fazla ayrık grup içindeki üyeliklerin, doğal bir sıralama yokken, öngörülmesi için kullanılır. Örneğin oy kullanacak kişiler içinden *A* veya *B* partisine oy verecek kişileri sosyal statülerine, yaşlarına, vb. bilgilere bakarak öngörebiliriz.

Ayrıma analizi, birden fazla grup içeren bir bağımsız değişkenin birden fazla bağımlı değişken üzerindeki etkisini inceleyen tek yönlü çok değişkenli varyans analizinin tersi olarak görülebilir. Çok değişkenli varyans analizinin bağımsız değişkenleri ayırma analizinin bağımlı değişken grupları olurken, çok değişkenli varyans analizinin bağımlı değişkenleri ayırma analizinin tahmincileri olurlar (Frenandez, 2002: 263).

Faktör Analizi: Bir grup değişken arasındaki korelasyonlara neden olan etkenlerin anlaşılması için kullanılır. Temel faktör analizi yöntemleri, değişken sayılarını indirgemek ve değişkenleri sınıflandırmak için değişkenler arasındaki ilişkilerin yapısını tespit etmektir. Bu bakımdan faktör analizi veri indirgeme veya yapı tespiti yöntemi olarak uygulanabilir. Doğrulama amaçlı yapılan faktör analizinde faktör yapısının önceden bilindiği kabul edilir ve amaç kabul edilen bu faktör yapısının

doğru olduğunu deneysel olarak ispatlamaktır. Araştırma amaçlı bir faktör analizinde ise amaç bir faktör yapısı keşfetmektir.

Regresyon Analizi: İki veya daha fazla sayısal değişken arasındaki ilişkiyi kullanarak değişkenlerden birinin (bağımlı değişkenin) diğerleriyle (bağımsız değişkenlerle) tahmin edilmesini sağlayan istatistiksel bir araçtır. Regresyon modeli, değişkenler arasındaki ilişki ne kadar güçlü olursa olsun, hiçbir sebep-sonuç ilişkisi belirtmez.

Lojistik Regresyon: Bağımlı değişken, iki elemanlı (*binary*) veya nitel bir değer olduğunda kullanılır. Lojistik regresyon, doğrusal regresyonda olduğu gibi, bir en uygun eşitlik bulur. Fakat bağımsız değişkenlerin dağılımları hakkında hiçbir varsayımda bulunmadığından bu en uygun eşitliği bulmak için en küçük kareler sapma kriteri yerine, verilen uygun regresyon katsayılarının gözlemlenmiş sonuçlarının elde edilme olasılığını maksimize eden maksimum olabilirlik yöntemini kullanır.

2.2.5 Veri Madenciliği Ve Makine Öğrenmesi

Makine öğrenmesi tecrübeden gelen bilgi kazanımının mekanikleştirilmesi ile performansı iyileştirmek için hesaplama yöntemlerinin kullanılmasıdır (Langley ve Simon, 1995). Makine öğrenmesinin amacı bilgi mühendisliği[22] sürecinde çok fazla zaman tüketen insan faaliyeti yerine eğitim verisindeki düzenleri keşfedip; doğruluğu ve etkinliği iyileştiren otomatik yöntemler kullanarak otomasyon seviyelerini yükseltmeyi sağlamaktır.

[22] Bilgi mühendisi, uzman sistemler konusunda eğitilmiş bilgisayar sistem uzmanıdır. Bilgi mühendisleri, alan uzmanlarının kendilerine sundukları bilgiyi yorumlar ve bilgisayar programcılarına/kodlayıcılara iletirler. Kodlayıcılar bilgiyi son kullanıcılar tarafından erişilecek şekilde sistem veri tabanına kodlar.

Veri madenciliğinin merkezini oluşturan makine öğrenmesi, bilimsel yöntem olarak; bir olayın gözlenmesi, bir hipotezin kurulması, öngörülerde bulunmak ve daha ileri gözlemler yapmadan önce hipotezin iteratif olarak yeniden gözden geçirilmesi süreci olarak tanımlanabilir. Bu süreç pek çok görevin yerine getirilmesi gereken karar verme basamaklarında da kullanılabilir. Makine öğrenmesinin esas hedefi bu tip süreçleri otomatikleştirmektir. Böylece bilgisayar, veri kümesine yeni veriler eklendikçe daha doğru öngörülerde bulunmak için verimli bir şekilde öğrenir. Bu sayede otomatik olarak bilgisayarın sağladığı öngörüler, bir işlemin sadece insana bağlı karar verme yetilerinin kullanılarak yapılmasından daha etkin ve daha doğru veya daha uygun maliyetli olarak yapılmasını sağlayabilir.

Makine öğrenmesine şunlar örnek verilebilir:

- Bir oda için tercih edilen aydınlatma seçimi,
- Nesneleri sınıflandırmak,
- Görüntüler içindeki belirli örüntüleri tanıma,
- Optik karakter tanıma: El yazısı bir metinde harfleri tanıyarak kelimeleri sınıflandırma,
- Konuşulan dili anlama,
- Sensör verilerine dayanan sistemleri kontrol etme,
- Güvenlik açısından kritik sitemler için riskleri öngörme,
- Bir ağdaki hataları tespit etme,
- Bir sistemdeki anormal durumları teşhis etme,
- Yüz tanıma: Görüntüler içindeki yüzleri bulma,
- İstenmeyen Posta Filtreleme: Elektronik postaları istenmeyen veya istenmeyen olmayan olarak tanımlama,
- Konu Tanıma: Haber makalelerini politika, spor, eğlence, vs. gibi sınıflandırma,
- Tıbbi Teşhis: Bir hastanın bir hastalığı taşıyıp taşımadığının teşhisinde,
- Müşteri Segmentasyonu: Hangi müşterilerin belirli bir satış teşvikine tepki vereceklerini öngörme gibi,

- Sahtekârlık Tespiti: Kredi kartı işlemlerinin hangilerinin dolandırıcılık olabileceğini tanımlama,
- Hava durumu Tahmini: Yarın yağmurun yağıp yağmama olasılığının hesaplanması gibi.

Veri madenciliğinde kullanılan bazı yaygın makine öğrenmesi teknikleri aşağıda verilmiştir (Goebel ve Gruenwald, 1999; Jackson, 2002):

Yapay Sinir Ağları: "Bir sinir ağı paralel, tek yönlü bağlantı kanalları vasıtası ile birbirine bağlı (yerel bir hafızası olan ve yerel bilgi işleme operasyonlarını yürütebilen) bağlantı denilen işleme elemanlarından oluşan dağıtılmış bilgi işleme yapısıdır. Her bir işleme elemanı, arzu edildiği kadar yan bağlantıya (her biri aynı işleme elemanı çıktı sinyalini taşıyan) dallanan (yayılan) sadece bir çıktı bağlantısına sahiptir. İşleme elemanı çıktı sinyali arzu edilen herhangi bir matematiksel türde olabilir. Her bir işleme elemanının sürdürdüğü bilgi işleme görevi tamamen yerel olmalıdır, yani sadece bağlantılar vasıtasıyla işleme elemanına ulaşan girdi sinyalinin mevcut değerlerine ve işleme elemanının yerel hafızasında depolanmış değerlere dayanmalıdır." (Nielsen, 1989).

Yapay sinir ağları insan beyninden modellenmiş bir sistemdir. İnsan beyninin sinapslar sayesinde birbiriyle bağlantılı milyonlarca sinir hücresinden oluşması gibi, yapay sinir ağları da beyindeki sinir hücrelerinin birbiriyle bağlantılı olmasına benzer şekilde çok sayıda temsili sinir hücresinden oluşmaktadır. Tıpkı insan beyninde olduğu gibi sinir hücreleri arasındaki bağlantının gücü, uyarana veya ağın öğrenmesini sağlayan elde edilen çıktıya göre değişebilir veya öğrenme algoritması tarafından değiştirilebilir.

Yapay sinir ağlarının başlangıç modelini kurmak çok fazla zaman alıcıdır, çünkü veri girme işlemi genellikle ham verinin dönüştürülmesi anlamına gelir ve değişken (özellik) seçimi için analistin çok fazla zamana ve yeteneğe ihtiyacı vardır. Ayrıca, teknik arka planı olmayan bir kullanıcı için yapay sinir ağlarının nasıl çalıştığını anlamak zordur.

Durum Tabanlı Çıkarsama: En yeni yapay zekâ tekniklerinden biri olan durum tabanlı çıkarsama, verilen bir problemi doğrudan geçmiş tecrübeleri ve çözümleri kullanarak çözmeye çalışan bir teknolojidir. Bir durum, genellikle, daha önceden karşılaşılmış ve çözülmüş belirli bir problemdir. Verilen yeni bir problemde, durum tabanlı çıkarsama depolanmış durumlar kümesini inceler ve bu yeni probleme benzeyenleri anlamsal ağları kullanarak bulur. Eğer benzer durumlar mevcutsa, onların çözümleri yeni probleme uygulanır ve yeni problem ileride kullanmak üzere durum tabanına eklenir.

Durum tabanlı çıkarsama yaklaşımı özellikle biçimsel temsil veya parametre tahmini için yeterli bilgi olmadığında kullanışlıdır. Durum tabanlı çıkarsama istatistiksel kabullere dayanmaz, insanların problem çözerken sıklıkla kullandığı benzeşim tabanlı çıkarsama ilkelerine dayanır ve bu nedenle gerekçeleri insanlarca anlaşılabilirdir (Arshadi ve Jurisica, 2005).

Durum tabanlı çıkarsamada, durum veri tabanında bulunan çözümler yeni problemlere uygulanırken benzer şartlar altında ne yapılması gerektiği ile değil de geçmişte yapılanlarla sınırlı olduklarından optimal olmayabilirler. Bu nedenle, durum veri tabanında bulunan çözümlerdeki hatalar yeni problemlerin çözümünde de devam ettirebilir.

Genetik Algoritmalar: Genetik algoritmalar öngörü ve sınıflandırma problemlerinin iyi çözümlerini araştırmak için doğal seleksiyonun, üremenin ve mutasyonun evrimsel biyolojik süreçlerinden modellenmiş bir algoritmik optimizasyon stratejisidir. Birbirleriyle rekabet halinde olan potansiyel çözümlerin kümesindeki en iyi çözümleri seçer ve birleştirir. Bunu yaparken çözüm kümesinin toplam iyiliğinin, canlı popülasyonlarının evrimine benzer bir şekilde, gittikçe daha da iyi olması beklenir. Veri madenciliğinde genetik algoritmalar, değişkenler arasındaki bağımlılıklar ile ilgili hipotezlerin birliktelik kuralları veya bazı diğer veriler arası bağlantı formlarında formülüze edilmesinde kullanılır.

Genetik algoritmaların sağladıkları çözümleri açıklamak zordur ve neden bu çözümlere ulaştıklarının anlaşılmasını sağlayan açıklayıcı istatistiksel ölçümler sağlamazlar.

Karar Ağaçları: Bir karar ağacı, uçta olmayan her bir düğümün (dairenin) eldeki veri ile ilgili, karar vericinin kontrolü dışında olan, bir testi veya kararı temsil ettiği ve ilk olay ile son sonuçlara ulaşma aşamasına kadar ortaya çıkan tüm olay ve eylemlerin düzenlenmesiyle oluşan bir ağaçtır. Düğümlerden çıkan her bir dal bir olayı simgeler ve düğümdeki testin sonucuna göre belli bir dal seçilir. Belirli bir veriyi sınıflandırmak için kök düğümden başlanır ve sonuçlara göre uçtaki bir düğüme veya yaprağa varılana kadar ilerlenir. Bir uç düğüme ulaşıldığında bir karar verilmiş olur ve bu karar bir kare ile gösterilir. Karar ağaçları hiyerarşik olarak düzenlenmiş kurallar ile nitelenen bir kurallar kümesinin bir hali olarak da yorumlanabilir.

Bazı veri madenciliği uygulamalarında önemli olan tek şey öngörünün doğruluğudur, modelin nasıl çalıştığı önemli değildir. Bazılarında ise, bir kararın nedeninin açıklanması çok önemli olabilir. Koşullu olasılıklara dayanarak kurallar üreten karar ağacı algoritması bu tür uygulamalar için uygundur.

Karar ağaçları eğitim sürecinde veriyi çok hızlı kullanır ve asla küçük veri kümeleri ile birlikte kullanılamaz. Ayrıca verideki gürültüye karşı son derece hassas ve aşırı öğrenmeye yatkındır.

Birliktelik Kuralları: Birliktelik kuralları; bilinen bir grup verinin ve bu verilerle aynı grupta olmayan, fakat gruptaki verilerle aynı özellikleri taşıyan diğer verilerin durumu hakkında öngörüde bulunmayı sağlayan bir veya daha fazla yönü ve özellikleri arasındaki bağlantıların ifadesidir. Daha genel haliyle, bir birliktelik kuralı verideki belli özelliklerin ortaya çıkmalarının veya bir veri kümesindeki belli veriler arasındaki istatistiksel korelasyonun ifadesidir.

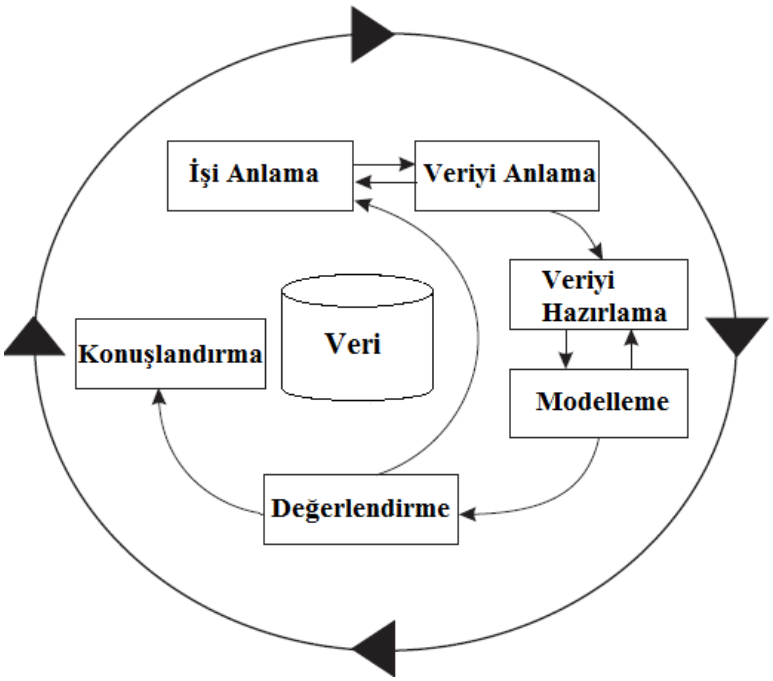
Bir veri setindeki bazı özelliklerin bir kümesi $x = \{x_1, \dots, x_n\}$ olmak üzere bu özellikler kullanılarak bir y kuralının c güven ve s doğruluk seviyesinde öngörülmesi kısaca $\{x_1, \dots, x_n\} \Rightarrow y[c, s]$ şeklinde gösterilir.

BÖLÜM 3: CRISP-DM

1996'nın sonlarına kadar veri madenciliği projelerini yürütmek için standart bir taslağın kullanılmaması, veri madenciliği projesinin başarısının veya başarısızlığının yüksek oranda projeyi gerçekleştiren kişilere bağlı olmasına neden olmaktaydı ve bir proje için başarılı bir uygulamanın başka projeler için tekrarlanması ile başarılı sonuçlar elde edileceği de kesin değildi. Veri madenciliğinin, iş problemlerini veri madenciliği problemlerine dönüştürecek, uygun veri dönüşümlerini ve veri madenciliği tekniklerini belirtecek ve sonuçların geçerliliğini değerlendirecek araçlara ve edinilen tecrübeleri kaydedecek standart bir yaklaşıma ihtiyacı vardı.

Veri madenciliğindeki bu sorunları ve ihtiyaçları azaltmak için veri madenciliği ile uğraşan bir takım kuruluşlar (NCR, ISL(SPSS), Daimler-Chrysler ve OHRA) Çapraz Endüstri Veri Madenciliği Standart Süreci (*Cross Industry Standard Process for Data Mining; CRISP-DM*) olarak adlandırılan bir sistem geliştirmişlerdir. CRISP-DM hem veri madenciliği çalışmasının yapıldığı endüstrilerden (alanlardan) hem de veri madenciliği için kullanılan yazılımlardan bağımsız olarak veri madenciliği projelerinin yürütülmesi için bir taslak sağlayan ve veri madenciliği projelerini daha az maliyetli, daha güvenilir, daha kolay tekrarlanabilir, daha kullanışlı ve daha hızlı yapmayı hedefleyen bir süreç modelidir.

Veri madenciliği projesinin yaşam döngüsü Şekil 3.1'de gösterilen altı aşamaya ayrılmıştır. Bu aşamaların sıralaması sabit değildir. Oklar aşamalar arasındaki en önemli ve en sık tekrarlayan bağlantıları göstermektedir. Çoğu zaman farklı aşamalar arasında ileri ve geri hareket etmek gerekir. Bir aşamadan sonra hangi aşamanın veya bir aşamanın hangi basamağının gerçekleştirileceği o aşamanın sonucuna bağlıdır.



Şekil 3.1: Çapraz Endüstri Veri Madenciliği Süreç Modeli (Chapman ve diğ., 2000).

Aşağıda CRISP-DM süreci aşamaları açıklanmıştır (Chapman ve diğ., 2000; Sumathi ve Sivanandam, 2006: 662-663; Jackson, 2002; Wirth ve Hipp, 2000; Küçüksille, 2009; Arguden ve Erşahin, 2002: 20-25; IBM Knowledge Center, The data mining process).

3.1 İŞİ ANLAMA

CRISP-DM'nin bu ilk aşaması veri madenciliği çalışmasının hedeflerini ve gereksinimlerini anlamaya odaklanır. Bu aşamada elde edilen bilgiler bir veri madenciliği problemine dönüştürülür ve hedeflere ulaşmak için bir hazırlık planı tasarlanır.

3.1.1 İş Hedeflerini Belirleme

Veri madenciliği uzmanının ilk hedefi, çalışmanın yapıldığı alanın bakış açısıyla, karar vericinin ulaşmak istediği hedefi tüm yönleriyle anlamaktır. Bu

basamakta amaç; karar vericinin, uygun bir şekilde dengelenmesi gereken, çoğu zaman birbiriyle rekabet eden hedef ve kısıtları içinde projenin sonucunu etkileyebilecek önemli etkenleri projenin başlangıcında açığa çıkarmaktır.

Proje sonuçlarını, çalışmanın yapıldığı alan açısından, değerlendirme ölçütlerinin ve öncelikli hedeflerin belirlenmesi görevleri de bu basamakta yürütülür.

Projenin püf noktalarına hızlıca ulaşmak amacı ile yürütülen bu basamağın ihmal edilmesinin olası bir sonucu olarak yanlış sorulara doğru cevaplar üretmek için çok fazla emek harcanır.

3.1.2 Durumu Değerlendirme

Proje sonucunda verilecek yanlış kararların zararlarını ve doğru kararların faydalarını değerlendirmek önemli bir ihtiyaçtır. Bu basamak; tüm kaynaklar, kısıtlar, varsayımlar ve veri analizinin amacını ve projenin planını belirlemeye dâhil edilen diğer etkenler hakkında daha detaylı bilgi toplamayı, değerlendirmeyi ve bunları elde edilecek faydanın boyutu ile karşılaştırmayı içerir.

3.1.3 Veri Madenciliği Hedeflerini Belirleme

Bu basamakta projenin, iş hedeflerini belirleme basamağında çalışmanın yapıldığı alanın terminolojisi ile belirtilen, amaçları veri madenciliği terminolojisi ile ifade edilir.

Elde edilen bir sonucun, veri madenciliği açısından, proje için başarılı bir sonuç olup olmadığını değerlendirme kriteri de bu basamakta tanımlanır**.**

3.1.4 Proje Planının Oluşturulması

Proje planının oluşturulması basamağında veri madenciliği hedeflerine ve dolayısıyla iş hedeflerine ulaşmak için, kullanılacak yazılım ve tekniklerin seçildiği ilk adımlarda dâhil, proje boyunca uygulanması öngörülen adımlar belirlenir.

Proje planının oluşturulması, projede gerçekleştirilecek aşamaların gerçekleştirilme süreleri, hangi kaynaklarla gerçekleştirileceklerinin listelenmesi, zaman çizelgesi ve riskler arasındaki bağımlılıkların analiz edilmesi bakımından önemlidir.

3.2 VERİYİ ANLAMA

Veriyi anlama aşaması; başlangıç verisini toplama, veri kalitesi problemlerini tanımlama, veri içindeki saklı bilgiyi açığa çıkartmak için hipotezler oluşturarak ilgili alt kümeleri tespit etme gibi veriyi tanımak ve veri hakkında ilk izlenimleri edinmek amacıyla yapılan faaliyetleri içerir.

3.2.1 Başlangıç Verisini Toplama

Projede kullanılacak verinin elde edildiği basamaktır. Başlangıç verisinin toplanması basamağında, ileride aynı projenin tekrarlanmasında veya benzer projelerin gerçekleştirilmesinde kullanılmak üzere; verinin toplandığı kaynakların, veriyi elde etmek için kullanılan yöntemlerin, karşılaşılan problemlerin ve erişilen her çözümün kaydı tutulur. Bu basamak, veriyi anlamak için özel bir yazılım kullanılıyorsa, verinin toplandığı kaynaklardan analizin yapılacağı yazılıma yüklenmesini de içerir.

Eğer veri pek çok farklı kaynaktan toplanmışsa bu basamakta veya daha sonra veriyi hazırlama aşamasında verinin birleştirilmesi ile ilgili sorunların çözülmesi gerekir.

3.2.2 Veriyi Tanımlama

Toplanan verilerin formatının, elde ne kadar sayıda kayıt olduğunun, kayıtların kaç alandan (özellikten) oluştuğunun ve diğer yüzeysel özelliklerin incelenmesi ve değerlendirilmesi verinin, projenin ihtiyaçlarını karşılayıp karşılamadığını anlamak bakımından önemlidir.

3.2.3 Veriyi Keşfetme

Veriyi keşfetme basamağı; sorgulama, görselleştirme ve raporlama yöntemleri kullanılarak ele alınabilen, sonuca yönelik olmaktan daha çok projenin gerçekleştirilebilmesi için veri ile ilgili eksiliklerin tespitine dair veri madenciliği soruları ile uğraşır. Verinin önemli özelliklerinin dağılımını, özellikler arasındaki bağlantıları ve basit istatistiksel analizleri içeren bu basamakta, veri hakkında edinilen ilk izlenimler veya başlangıç hipotezleri ve bunların projenin geri kalanı üzerindeki etkileri üzerine analizler yapılır. Bu analizler doğrudan veri madenciliği hedeflerine hitap edebilir. Örneğin; veri tanımını ve kalite raporlarını iyileştirebilir veya bunlara katkıda bulunabilir, veri dönüştürme ve veri hazırlama aşamasında ihtiyaç duyulan daha ileri basamaklardaki diğer analizlere katkıda bulunabilir.

3.2.4 Verinin Kalitesinin Belirlenmesi

Aşağıdaki gibi hem veriye hem de çalışmanın yapıldığı alanın bilgisine dayanan sorularla veri kalitesi incelenebilir:

- Veri gerekli olan tüm durumları kapsıyor mu?
- Veri doğru mu veya hatalar içeriyor mu?

- Eğer veri hatalar içeriyorsa bu hatalar ne kadar yaygın?
- Veride kayıp değerler var mı?
- Eğer veride kayıp değerler varsa hangi verilerin değerleri kayıp, bunlar nasıl temsil ediliyor ve kayıp değerler ne kadar yaygın?

3.3 VERİYİ HAZIRLAMA

Veriyi hazırlama aşaması, başlangıç aşamasındaki ham veriden modellemede kullanılacak veri setinin inşa edilmesine kadar yapılan tüm faaliyetleri kapsar. Modelleme için verinin dönüştürülmesini ve temizlenmesini içerdiği gibi tablo oluşturma, kayıt ve özellik seçimini de içeren veri hazırlama görevleri belirli bir sıralama gözetilmeden pek çok kez tekrarlanabilir.

3.3.1 Veriyi Seçme

Bu basamakta analizde kullanılacak veriye karar verilir. Veri seçme kriteri; verinin veri madenciliği hedeflerine uygunluğunu, kalitesini ve verinin hacmi veya türü gibi teknik kısıtları içerir. Veriyi seçme basamağında, modeli olumsuz etkileyebilecek bazı özellikler ve kayıtlar çıkarılabilir.

3.3.2 Veriyi Temizleme

Veriyi anlama aşamasının veri kalitesini belirleme basamağı boyunca belirlenen veri kalitesi problemlerini ele alan veri temizleme basamağı, veri kalitesinin modelde kullanılabilecek seviyeye yükseltildiği basamaktır. Kayıp veriler, boş değerler, var olmayan değerler ve tam olmayan veri içeren kirli veri ile çalışmak modelin oluşmasına engel olabilecek veya modeli bozabilecek problemler ortaya çıkabilir. Veri temizleme; verinin temiz alt kümelerinin seçimini, eksik değerler yerine uygun değerlerin atanmasını veya eksik verileri bir model aracılığıyla tahmin

ederek elde edilen değerlerle değiştirmeyi veya kirli veri içeren elemanlar için ayrı modeller inşa edilmesini içerebilir.

3.3.3 Verinin İnşası

Bu basamak eldeki verileri kullanarak özellik türetme gibi yapısal veri hazırlama işlemlerini içerir. Örneğin; en ve boy verileri kullanılarak alan özelliği $en \times boy = alan$ formülü ile türetilebilir.

3.3.4 Veriyi Birleştirme

Farklı formatlarda verilerin ve farklı ölçü birimlerinin kullanılması, aynı değerler için farklı kodlamaların yapılması ve verilerin farklı zamanlara ait olmaları veri uyumsuzluklarına neden olur. Farklı kaynaklardan toplanmış verileri birleştirmek ve bu uyumsuzlukları ortadan kaldırmak için tablo kaynaştırma ve toplanmış değerler üretme yöntemleri kullanılır. Tablo kaynaştırma yönteminde aynı nesneler hakkında farklı bilgiler içeren iki veya daha fazla tablonun birleştirilmesi ve toplanmış değerler üretme yönteminde ise çok sayıda kayıttan, tablodan veya diğer veri kaynaklarından gelen verinin özetlenmesi ile yeni değerler hesaplanması yardımıyla veriler birleştirilir.

3.3.5 Veriyi Biçimlendirme

Veriyi, modelleme aşamasını gerçekleştirecek yazılımın kullanabileceği şekilde verinin anlamını değiştirmeden (örneğin veri setindeki kayıtların sıralamasını değiştirmek) düzenleme basamağıdır.

3.4 MODELLEME

Bu aşamada, çeşitli modelleme yöntemleri seçilir ve uygulanır. Genellikle, benzer veri madenciliği problemleri için farklı çözüm yöntemleri uygulanabilir. Bazı modelleme yöntemleri sadece belirli veri biçimlerini kullanabildiğinden, çoğu zaman, veri hazırlama aşamasına geri dönülmesi gerekir.

3.4.1 Modelleme Yönteminin Seçilmesi

Genellikle, işi anlama aşamasında veri madenciliği çalışmasında kullanılacak yazılımın seçilmesi ile birlikte modelleme sırasında kullanılacak yöntemler hakkında da bir takım öngörüler kazanılmış olur. Bu basamakta ise birden fazla yöntem arasından modelleme sırasında kullanılacak esas yönteme karar verilir.

Modelleme yöntemini seçerken çoğu yöntemin; tüm verilerin düzgün dağılıma sahip olması, hiçbir kayıp değerin dikkate alınmaması, sınıf değişkeninin nominal olması, vb. kabullerde bulunduğuna dikkat etmek gerekir.

3.4.2 Model Test Tasarımını Oluşturma

Gerçekten bir model oluşturmadan önce modelin kalitesini ve doğruluğunu test etmek için bir süreç veya mekanizma oluşturulması gerekir. Örneğin, veri madenciliği modelleri için kalite ölçüsü olarak hata oranlarını kullanan sınıflandırma gibi denetimli veri madenciliği görevlerinde, genellikle, veri seti eğitim ve test diye ikiye ayırılır. Sonra model eğitim seti üzerine kurulur, modelin kalitesi eğitim setinden farklı test seti üzerinde değerlendirilir ve yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı hesaplanır. Bu anlamda öncelikli olarak eldeki verinin nasıl eğitim ve test veri setlerine ayırılacağının belirlenmesi gerekir.

3.4.3 Modeli Kurma

Modeller, öngörüler edinmek ve bu öngörülere dayanan bilgiyle desteklenen kararlar almak amacıyla kurulur. Bir modeli kurarken ulaşılmak istenen en önemli hedef, modelin henüz karşılaşılmamış verilere uygulanması halinde doğru öngörülerde bulunması anlamına gelen güvenilirliktir.

Kullanılan veri madenciliği yöntemi ne olursa olsun tahmin edici modeller kurmak için takip edilen adımlar şunlardır:

1. Eğitim modeli, modelin gelecek hakkında öngörülerde bulunması amacıyla geçmişe ait verilerden oluşturulur. Bu süreçte veri madenciliği algoritmaları öngörüsel değeri olan örüntüleri bulur.
2. Modelin eğitim setini ezberlemesini engellemek için modelin test seti kullanılarak iyileştirilmesi gerekir. Test setinin kullanılması aşaması modelin daha güvenilir olmasını ve karşılaşılmamış veri üzerinde de iyi sonuçlar verecek şekilde uygulanmasını sağlar.
3. Tamamen eğitim ve test setlerinden ayrı bir veri seti olan doğrulama seti kullanılarak modelin, model seti dışındaki verilere uygulanması durumundaki performansı değerlendirilir.

3.4.4 Modeli Değerlendirme

Modeli değerlendirme basamağı, modelleme aşamasının sonuçlarına dayanan tamamen teknik bir basamaktır. Bu basamakta veri madenciliği uzmanı kendi alan bilgisine, yani veri madenciliği başarı ölçütüne ve arzulanan test tasarımına göre modeli yorumlar. Veri madenciliği uzmanı modeli uygulamanın başarısını veri madenciliği açısından değerlendirir, sonra veri madenciliği sonuçlarını çalışmanın yapıldığı alanın bakış açısıyla değerlendirmek için alan uzmanları ile görüşür.

Bu basamakta veri madenciliği uzmanı, üretilen modellerin doğruluk açısından iyi özelliklerini listeler ve modellerin başarılarını kıyaslayarak modelleri derecelendirmeye çalışır. Modelleri veri madenciliği ölçütüne göre değerlendirmeye çalışırken mümkün olduğu kadar iş hedeflerini ve iş başarı kriterlerini de hesaba katar. Genellikle veri madenciliği uzmanı ya yalnız bir yöntemi birden fazla kez uygulayarak ya da çeşitli farklı yöntemlerle veri madenciliği sonuçları üretir ve veri madenciliği değerlendirme ölçütüne göre tüm sonuçları karşılaştırır.

Veri madenciliği uzmanı en iyi modeli (veya modelleri) bulana kadar model kurma ve model değerlendirme aşamalarını tekrar eder.

3.5 DEĞERLENDİRME

Konuşlandırma aşamasından önce, modelin uygun bir şekilde iş hedeflerine ulaştığından emin olmak için, modelin baştan aşağı değerlendirilmesi ve modeli kurmak için gerçekleştirilen basamakların gözden geçirilmesi önemlidir.

Modeli değerlendirme basamağı sadece modelle ilgili doğruluk ve modelin güvenilirliği gibi sonuçları dikkate alırken, değerlendirme aşaması proje doğrultusunda üretilen tüm sonuçları dikkate alır. Bu aşamanın sonunda, ulaşılan veri madenciliği sonuçlarının kullanılıp kullanılamayacağına karar verilir.

3.5.1 Sonuçları Değerlendirme

Bu basamakta modelin iş hedeflerini ne kadar karşıladığı değerlendirilir, yeterli bir şekilde dikkate alınmamış önemli iş meselelerinin olup olmadığı ve modelin yetersiz olduğu bir nokta varsa bu yetersizliğe dair iş sebepleri belirlenmeye çalışılır. Eğer yeterli zaman ve bütçe varsa sonuçları değerlendirmek için model (veya modeller) uygulanarak çıkan sonuçlar değerlendirilebilir.

3.5.2 Süreci Gözden Geçirme

Modelin (veya modellerin) kabul edilebilir ve iş hedeflerini karşılar nitelikte olduğuna karar verdikten sonra bir şekilde gözden kaçmış herhangi bir önemli etken veya basmak var mı diye kalite daha derinlemesine bir inceleme yapılması gerekir. Veri madenciliğinin değerlendirme aşamasında süreci gözden geçirme kalite kontrolü amacıyla yapılır.

3.5.3 Sonraki Adımları Belirleme

Veri madenciliği ve alan uzmanları değerlendirme ve gözden geçirme basamaklarının sonuçlarına göre kararları etkileyebilecek, kalan kaynaklar ve bütçe kısıtları altında, gelecekteki olası faaliyetleri ve her bir faaliyetin lehine ve aleyhine olan nedenleri göz önüne alarak projenin nasıl devam edeceğine bu basamakta karar verirler. Bu basamakta alınan kararlar sonucunda ya proje bitirilip konuşlandırma aşamasına (son aşamaya) geçilir ya süreci iyileştirmek için ek çalışmalar yapılır ya da süreç tekrar edilir.

3.6 KONUŞLANDIRMA

Genellikle, modelin oluşturulması projenin bitmesi anlamına gelmez. Modelin amacı veriden kazanılan bilgiyi arttırmak olsa da açığa çıkarılan bilginin düzenlenmesi ve karar vericinin kullanabileceği bir şekilde sunulması önemlidir. Konuşlandırma aşaması, karar vericinin ihtiyaçlarına göre, bir rapor oluşturmak kadar basit veya yinelenen bir veri madenciliği sürecini uygulamak kadar karmaşık olabilir.

3.6.1 Konuşlandırmayı Planlama

Veri madenciliği sonuçlarını çalışmanın yapıldığı alanda kullanabilmek için değerlendirme sonuçları alınır ve konuşlandırma için bir strateji geliştirilir.

3.6.2 Planı Takip Etme ve Devam Ettirme

Veri madenciliği sonuçları günlük hayatta kullanılmaya devam edildikleri sürece konuşlandırma planını takip etme ve devam ettirme önemli meselelerdir. Zaman içerisinde verilerde ve çalışmanın yapıldığı alanda ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak takip edilmesini ve kullanılmaya devam edilebilmeleri için yeniden düzenlenmelerini gerektirir. Bu sayede, uzun vadede, veri madenciliği sonuçlarının yanlış kullanımı engellenmiş olur.

3.6.3 Son Raporun Hazırlanması

Projenin sonunda proje ekibi son bir rapor hazırlar. Bu rapor, konuşlandırma planına bağlı olarak, projenin ileride tekrar uygulanabilmesini sağlamak için edinilen tecrübelerin bir özeti veya veri madenciliği sonuçlarının son ve kapsamlı bir sunumu olabilir.

3.6.4 Projeyi Gözden Geçirme

Proje gözden geçirilerek proje boyunca kazanılan önemli tecrübeler özetlenir; nelerin doğru nelerin yanlış gittiği, nelerin yeterince iyi bir şekilde yapıldığı ve nelerin daha iyi yapılması gerektiği değerlendirilir. Örneğin, tespit edilmiş gizli tuzakları ve yanlış yola sevk eden yaklaşımları veya benzer projelerde en uygun veri madenciliği yöntemlerini seçmek için kullanılabilecek ipuçlarını açıklamak bu basamağın bir parçası olabilir.

BÖLÜM 4: VERİ MADENCİLİĞİNİN GÖĞÜS KANSERİ ERKEN TEŞHİSİNDE KULLANIMI

Bu bölümde, üçüncü bölümde anlatılan CRISP-DM veri madenciliği sürecinin gerçekleştirilen uygulamada nasıl hayata geçirildiği hakkında bilgi verilecektir.

4.1 İŞİ ANLAMA

Vücudumuzun farklı yerlerinde iyi huylu ve kötü huylu tümörler bulunmaktadır. İyi huylu tümörler komşu bölgelere yayılmazlar ve çıkarıldıktan sonra tekrarlamazlar. Kötü huylu tümörler ise komşu bölgelere yayılırlar ve zarar verirler. Bu tümörleri kanser olarak adlandırırız. Kanserden korunmanın en önemli yolları kanser yapan etkenlerden kaçınma ve erken teşhistir.

Günümüzde kanser vakalarının çoğu hastalığın ilerlemiş safhalarında teşhis edilmektedir. Bunun sonucunda uygulanan tedaviler çoğu zaman işe yaramamakta ve hasta hayatını kaybetmektedir. Öncelikle insan hayatının önemi göz önüne alınıp sonrasında da kansere yakalanmış hastaların teşhis ve tedavi maliyetlerinin incelendiğinde kanserin varlığının veya riskinin başlangıç safhasında tespit edilmesinin önemi anlaşılmaktadır. İleride daha büyük tıbbi müdahale gerektirecek operasyonların önceden alınan kararlarla en aza indirgenmesi, tıbbi kaynak planlaması açısından faydalı olmakta ve hastalık kontrol altına alınarak, kişinin iç huzurunda kalıcı hasarlar oluşması engellenmektedir.

Kanser hastalarına ait tüm laboratuvar sonuçları, çekilen film ve röntgen görüntüleri, hastanın hikâyesi dahi veri tabanlarında tutulmaktadır. Fakat sağlık kurumlarında biriken bu veriler zamana ve bölgelere göre değişmektedir. Ayrıca, bu veriler içerisinde tutarsız, eksik, aykırı ve uç veriler bulunmaktadır. Bu nedenle veri tabanlarından geleneksel sorgulama metotlarıyla bilgiyi süzmek ve bu bilgileri

raporlar halinde sunmak bilgiler içerisinde saklı bulunan gizli-önemli kuralların ortaya çıkmasını sağlamaz. Ancak kanserin teşhis ve tedavi sürecinde alınacak kararlara yön verebilecek cevapların bulunması, büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanılarak aranmasını sağlayan veri madenciliği teknikleri ile mümkündür.

Kanser erken teşhisinin en önemli hedefleri; teşhis amacıyla yürütülen testlerin hastaya mümkün olduğunca zarar vermemesi, teşhisin çabuk ve güvenilir olmasıdır.

Röntgen, biyopsi ve diğer tahlil sonuçlarının yanlış yorumlanması sonucunda kanser hastası olmayan bir hastaya kanser teşhisi konması hastaya hem ruhsal hem de maddi açıdan zarar verecektir. Ve gerçekte kanser hastası olan bir hastaya sağlam teşhisi konması, sonradan yanlış teşhis konulduğu fark edilse bile, kanser tedavisi için geç kalınmasına neden olacaktır.

Doğru erken teşhis yardımıyla hasta en az kanser hastalığının kendisi kadar ölümcül olan gereksiz tedavi sürecinin fizyolojik, psikolojik ve mali açıdan vereceği zararlardan da korunmuş olur.

Ülkemizde kanser hasta kayıtlarının çoğunun dosya-işlem sistemi ile tutulması, kayıtların standart bilgiler içermemesi, kanser kayıtçılığının henüz yaygınlaşmamış olması ve yasaların kanser hastalarına ait verilere ulaşılmasını hasta haklarının ihlali kapsamında değerlendirmesi kanser erken teşhisine yönelik güvenilir ve hızlı karar verme tekniklerinin geliştirilmesini zorlaştırmaktadır.

4.1.1 Kullanılacak Yazılım

Yapılacak veri madenciliği çalışması için kullanılacak yazılım olarak belirlenen WEKA (*Waikato Environment for Knowledge Analysis*), Waikato Üniversitesi (Yeni Zelanda) tarafından Java programlama dilinde yazılmış açık kaynak kodlu bir veri madenciliği programıdır. Veri madenciliği çalışmasında kullanılmak üzere WEKA programı şu nedenlerle seçilmiştir:

- GNU, genel kamu lisanslı (GPL), bir yazılım olmasından dolayı düşük bilgi işlem maliyetine sahiptir.
- Java programlama dilinde yazıldığı için günümüzdeki her bilgisayarda çalıştırılabilir.
- Kapsamlı veri ön işleme, sınıflandırma, kümeleme, birliktelik analizi, özellik seçimi ve görselleştirme tekniklerini içermektedir.
- Uygulamaların komut girilerek gerçekleştirilmesine imkân tanır.

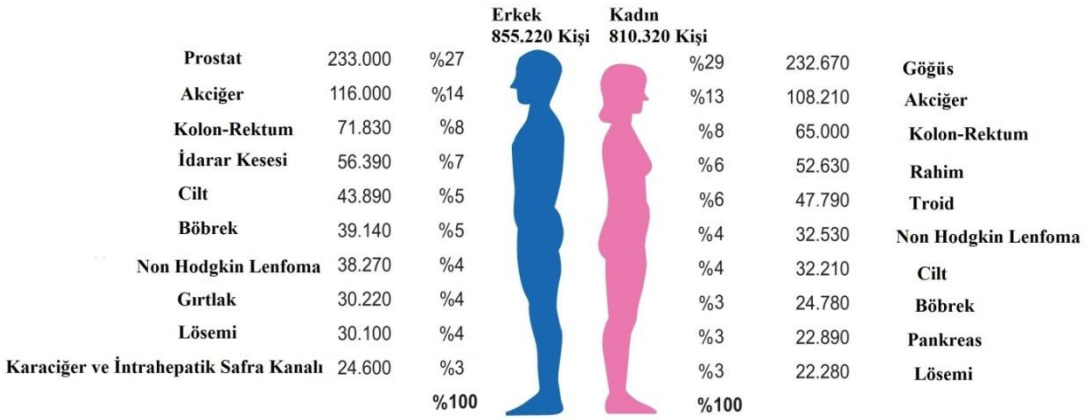
Halen yeni sürümleri geliştirilmeye devam eden WEKA'nın bugüne kadar geliştirilmiş tüm sürümleri ücretsiz olarak http://www.cs.waikato.ac.nz/ml/weka/ adresinden edinilebilmektedir.

Şekil 4.1: WEKA kullanıcı arayüzü.

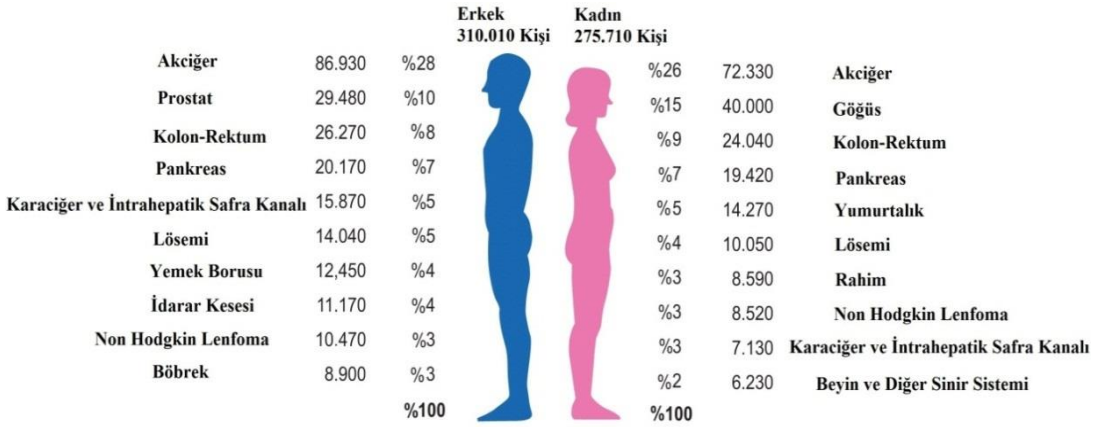
4.2 VERİYİ ANLAMA

Yapılacak olan çalışma "veri madenciliği ve kanser erken teşhisinde kullanımı" olduğundan kullanılacak veri seti kanser erken teşhisine yönelik kayıtları içermelidir. Bu amaçla hastanelerin kanser veri tabanlarının kullanılmasına karar verilmiş ve Turgut Özal Tıp Merkezi Medikal Onkoloji Polikliniğine başvurulmuştur. Fakat hastanenin hasta kayıtlarını dosya-işlem sistemi ile tutması ve teşhis amaçlı yapılan test çeşitlerinin hastadan hastaya değişmesinden dolayı hali hazırda bir veri tabanına ulaşılamamıştır. Bu nedenle İl Halk Sağlık Müdürlüğü Kanser Kayıt Merkezi'ne başvurulmuştur. Her ne kadar, hem Turgut Özal Tıp Merkezi Medikal Onkoloji Polikliniği doktor ve çalışanları hem de Malatya Kanser Kayıt Merkezi uzman ve çalışanları yapılacak olan çalışmada ellerindeki veriyi paylaşma konusunda son derece yardımcı olmuşlarsa da Kanser Daire Başkanlığı'nın uyarıları doğrultusunda kanser verilerinin paylaşılmasının yasalara aykırı olduğu tespit edilmiştir. Dolayısıyla Türkiye'ye ait gerçek kanser verileri ile çalışmak yerine uluslararası yasaların bilimsel çalışmalarda kullanılmasına izin verdiği Wisconsin Göğüs Kanseri Teşhis Verisi kullanılmıştır.

Şekil 4.2 ve 4.3, sırasıyla, 2014 yılında Amerika'da görülen başlıca kanser türleri için cinsiyete göre tahmini kanser tanısı ve kanserden ölüm oranlarını göstermektedir.



Şekil 4.2: Tahmini kanser tanısı konan hasta sayısı ve oranları (Siegel ve diğ., 2014).



Şekil 4.3: Tahmini kanserden ölüm oranları (Siegel ve diğ., 2014).

Kanser bir hücrenin kontrol edilemeyen bölünmesi ile başlar ve tümör adı verilen gözle görülebilir bir kitle ile sonuçlanır. Tümör iyi huylu veya kötü huylu olabilir. Kötü huylu tümörler hızla büyürler ve etraflarını çevreleyen dokulara zarar vererek yayılırlar. Göğüs kanseri göğüste büyümeye başlayan kötü huylu bir dokudur. Göğüste yabancı bir kitlenin bulunması, göğüsün şeklinde ve boyutunda değişiklik olması, göğüs derisinin renginde değişiklik olması, göğüs ağrıları vs. gibi anormallikler göğüs kanseri belirtileridir. Diğer kanserlerde olduğu gibi göğüs kanserinin erken teşhisi hayat kurtarıcı olabilir.

Çoğu göğüs kanseri, hasta tarafından göğüste bir kitle olarak tespit edilir. Göğüsteki kitlelerin çoğu iyi huyludur bu yüzden iyi huylu tümörleri kötü huylu tümörlerden ayırmak doktorların görevleridir. Göğüs kanserini teşhis etmek için üç yol vardır: Mamografi, görsel yorumlama ile ince iğne aspirasyonu (*Fine Needele Aspiration; FNA*) ve cerrahi biyopsi. Bu yöntemlerin hastalığı doğru teşhis etme oranları aşağıdaki gibidir (Mangasarian ve diğ., 1994):

- Mamografide doğruluk oranı $\%68 - \%79$ arasındadır ve sonuçları hassasiyetten uzaktır.
- Görsel yorumlama ile FNA'da doğruluk oranı $\%65 - \%98$ arasındadır ve doğruluğu teşhisi koyan doktorun tecrübesine dayanır.

– Cerrahi biyopside ise doğruluk oranı %100'e yakındır. Ne var ki cerrahi müdahale pahalıdır, hastaya zarar verir, zaman alır ve acı vericidir.

Wisconsin Göğüs Kanseri Teşhis Verisi, FNA tekniği ile alınan örneklerin Xcyt görüntü analiz programı ile incelenmesi ile elde edilen verilerden oluşmaktadır (Mangasarian ve diğ., 1994). UCI makine öğrenmesi veri havuzundan alınan bu veri seti 357'sine iyi huylu tümör, 212'sine kötü huylu tümör tanısı konmuş 569 örnekten oluşmaktadır. Kötü huylu tümörlerin olduğu vakalar cerrahi biyopsi yapılarak doğrulanırken, iyi huylu tümörlerin olduğu vakalar ya cerrahi biyopsi ile ya da birbirini takip eden periyodik tıbbi incelemeler sonucunda doğrulanmıştır.

FNA ve Xcyt programı kullanılarak gerçekleştirilen teşhis süreci aşağıdaki gibidir:

- Göğüsteki kitleden FNA ile bir örnek alınır. Alınan örnek içindeki çekirdekler net görünmeleri için boyanır ve mikroskop altında incelenir. Preparatın, hücrelerin birbirinden farklı olduğu, bir bölümü dijital bir kamera kullanılarak taranır.
- Sonra kullanıcı her bir çekirdeği Xcyt programını kullanarak izole eder. Bu için kullanıcı fare imlecini kullanarak her bir çekirdeğin yaklaşık çeperini çizer ve sonra "aktif kontur modeller"[23] (yılanlar) olarak bilinen bir bilgisayar görüntü yaklaşımını kullanarak yaptığı bu yaklaşık çeper çizimlerinin gerçek çekirdek çeperlerine yakınsamasını sağlar.
- Normal bir görüntü 10 ile 40 arasında değişen sayıda çekirdekten oluşur. Bir kez çekirdeklerin tümünün (veya çoğunun) çeperi çizilip birbirlerinden ayrıldıktan sonra program her bir çekirdek için aşağıdaki özellikleri hesaplar:

[23] Literatürde, ilk olarak aktif kontur (yılan) adı verilen biçim değiştirebilir model Kass ve diğerleri (Kass ve diğ., 1988) tarafından sunulmuştur. En bilinen yarı otomatik detay çizim yaklaşımlarından bir tanesidir. Bu optimizasyon tekniği elastodinamik modelleri ve bu modellerin iç ve dış kuvvetlerin etkisi altındaki davranışlarını kullanan eğrinin kontrastlık ve pürüzsüzlük modellerinin optimizasyonunu içermektedir. Aktif kontur, imge üzerinde şeklini değiştirip, nesnenin sınırlarına geldiği zaman duran kapalı bir elastik şerit gibi düşünülebilir. Aktif kontur, şekil değiştirmeyi enerji minimizasyonu ile gerçekleştirir. Aktif konturun enerjisi, iç ve dış güç olmak üzere ikiye ayrılmıştır. Dış güç, imge özelliklerinden gelen imge gradyan şiddeti gibi değerlere bağlıyken, iç güç ise aktif konturun şeklinin düz olmasını sağlamaktadır.

- **Yarıçap** (*Radius*)**:** Bir tek çekirdeğin yarıçapı, aktif kontur modelinin merkezinden çıkan yarıçap doğru parçalarının ortalamasının alınması ile bulunur.
- **Doku** (*Texture*)**:** Hücre çekirdeğinin dokusu, piksel bileşenleri içindeki gri ölçeğin[24] değerlerinin standart sapmasının ölçülmesi ile bulunur.
- **Çevre** (*Perimeter*)**:** Birbirini takip eden aktif kontur noktaları arasındaki toplam uzaklık çekirdek çevresini verir.
- **Alan** (*Area*)**:** Alan, aktif konturun iç kısmındaki piksellerin sayısına çevre üzerindeki piksellerin yarısının eklenmesi ile hesaplanır.
- **Düzgünlük** (*Smoothness*)**:** Çekirdeğin bir yarıçapı ile diğer yarıçaplarının ortalama uzunluğu arasındaki farkın ölçülmesi ile sayısallaştırılır.
- **Kompaktlık** (*Compactness*)**:** $\left(\frac{Çevre^2}{Alan} - 1\right)$ formülü yardımı ile hesaplanır.
- **Konkavlık** (*Concavity*)**:** Konkavlık, hücre çekirdeklerinin çeperlerindeki girintilerin boyu ölçülerek hesaplanır. Bitişik olmayan aktif kontur noktaları arasına kirişler çizilir ve her bir kirişin içinde kalan gerçek çekirdek sınırı ölçülür.
- **Konkav Noktalar** (*Concave Points*)**:** Bu özelliğin konkavlıktan farkı sadece çekirdek çeperinin konkav bölgelerindeki sınır noktalarının sayısını dikkate almasıdır.
- **Simetri** (*Symmetry*)**:** Simetriyi ölçmek için, esas eksen veya merkezden geçen en uzun kiriş bulunur. Sonra esas eksene, hücre çeperine kadar, dik olan doğruların uzunlukları arasındaki fark esas eksenin böldüğü her iki parça için ölçülür.
- **Oransal Kırılma Boyutu** (*Fractal Dimension*)**:** Bir çekirdek çeperinin oransal kırılma boyutu Mandelbrot tarafından tanımlanan "kıyı şeridi yaklaşımı" kullanılarak tahmin edilir.

- Bu özelliklerin her birinin ortalaması, standart hatası ve en büyük üç değerinin ortalaması olan en büyük değerin hesaplanması sonucunda her bir örnek için Tablo 4.1'de gösterilen 30 adet çekirdek özelliği elde edilmiştir.

[24] Fotoğrafçılık ve programlamacılıkta gri ölçek dijital imgesi her bir pikselin değerinin bir örneklem olduğu, yani sadece yoğunluk bilgisi içeren bir imgedir.

– **Örneklem Sayısı =** n

– **Ortalama:** $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

– **Standart Sapma:** $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

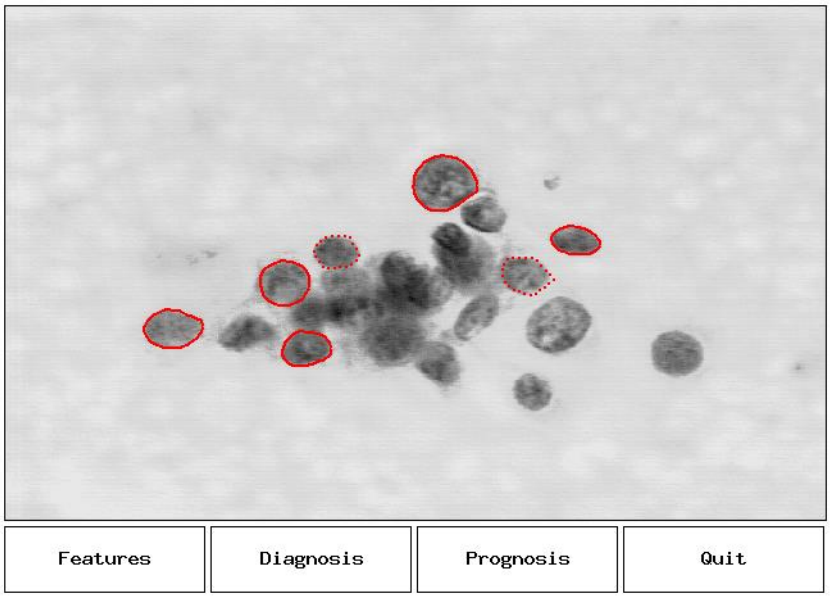
– **Standart Hata =** $\frac{s}{\sqrt{n}}$

1- yarıçap_ort.	11- yarıçap_std.h.	21- yarıçap_en_büyük
2- doku_ort.	12- doku_std.h.	22- doku_ en_büyük
3- çevre_ort.	13- çevre_std.h.	23- çevre_ en_büyük
4- alan_ort.	14- alan_std.h.	24- alan_ en_büyük
5- düzgünlük_ort.	15- düzgünlük_std.h.	25- düzgünlük_ en_büyük
6- kompaktlık_ort.	16- kompaktlık_std.h.	26- kompaktlık_ en_büyük
7- konkavlık_ort.	17- konkavlık_std.h.	27- konkavlık_ en_büyük
8- konkav_noktalar_ort.	18- konkav_noktalar_std.h.	28- konkav_noktalar_ en_büyük
9- simetri_ort.	19- simetri_std.h.	29- simetri_ en_büyük
10- oransal_kırılma_ boyutu_ort.	20- oransal_kırılma_ boyutu_std.h.	30- oransal_kırılma_ boyutu_ en_büyük

Tablo 4.1: FNA yöntemi ve Xcyt programı kullanılarak elde edilen 30 özellik.

Yüksek değerli şekil özellikleri taşıyan çekirdekler daha az düzgün bir şekle ve daha yüksek kötü huylu tümör olma ihtimaline sahiptir.

– Bu etkileşimli süreç preparat başına iki ile beş dakika arasında değişir.



Şekil 4.4: Xcyt programının kullanımını gösteren bir görüntü.

4.2.1 Çekirdek Özelliklerine Ait Bazı İstatistiksel Analizler

Tablo 4.2'de "Farklı Değer" ifadesi ile kastedilen, seçilen özellik için gözlenen birbirinden farklı değer sayısıdır. Örneğin yarıçap ortalama özelliğinin $[6.98 - 28.11]$ kapalı aralığında değişen 456 farklı değeri olduğu için $Farklı\ Değer = 456$'dır. "Bir Değerli" ifadesi ile kastedilen ise seçilen özelliğin herhangi bir değerinin sadece bir örnekleme ait olanlarının sayısıdır. Örneğin yarıçap ortalama özelliği için 1'den 358'e kadar olan tüm örneklemler $[6.98 - 28.10]$ kapalı aralığında sadece bir değere sahip olsun ve 359'dan 569'a kadar olan tüm örneklemler için yarıçap ortalama değeri 28.11 olsun. Bu durumda $Bir\ Değerli = 359$'dur.

Özellik	Farklı Değer	Bir Değerli (%)	Minimum	Maksimum	Ortalama	Standart Sapma
Yarıçap Ort.	456	359 (%63)	6.981	28.11	14.127	3.524
Doku Ort.	479	398 (%70)	9.71	39.28	19.29	4.301
Çevre Ort.	522	478 (%84)	43.79	188.5	91.969	24.299
Alan Ort.	539	510 (%90)	143.5	2501	654.889	351.914
Düzgünlük Ort.	474	404 (%71)	0.053	0.163	0.096	0.014
Kompaktlık Ort.	537	507 (%89)	0.019	0.345	0.104	0.053
Konkavlık Ort.	537	517 (%91)	0	0.427	0.089	0.08
Konkav Noktalar Ort.	542	527 (%93)	0	0.201	0.049	0.039
Simetri Ort.	432	327 (%57)	0.106	0.304	0.181	0.027
Oransal Kırılma Boyutu Ort.	499	434 (%76)	0.05	0.097	0.063	0.007
Yarıçap Std.H.	540	513 (%90)	0.112	2.873	0.405	0.277
Doku Std.H.	519	473 (%83)	0.36	4.885	2.217	0.552
Çevre Std.H.	533	499 (%88)	0.757	21.98	2.866	2.022
Alan Std.H.	528	491 (%86)	6.082	542.2	40.337	45.391
Düzgünlük Std.H.	547	525 (%92)	0.002	0.031	0.007	0.003
Kompaktlık Std.H.	541	516 (%91)	0.002	0.135	0.025	0.018
Konkavlık Std.H.	533	508 (%89)	0	0.396	0.032	0.03
Konkav Noktalar Std.H.	507	459 (%81)	0	0.053	0.012	0.006
Simetri Std.H.	498	437 (%77)	0.008	0.079	0.021	0.008
Oransal Kırılma Boyutu Std.H.	545	521 (%92)	0.001	0.03	0.004	0.003
Yarıçap En Büyük	457	372 (%65)	7.93	36.04	16.269	4.833

Tablo 4.2: Çekirdek özelliklerine ait bazı istatistiksel analizler

Özellik	Farklı Değer	Bir Değerli (%)	Minimum	Maksimum	Ortalama	Standart Sapma
Doku En Büyük	511	455 (%80)	12.02	49.54	25.677	6.146
Çevre En Büyük	514	462 (%81)	50.41	251.2	107.261	33.603
Alan En Büyük	544	519 (%91)	185.2	4254	880.583	569.357
Düzgünlük En Büyük	411	289 (%51)	0.071	0.223	0.223	0.023
Kompaktlık En Büyük	529	491 (%86)	0.027	1.058	0.254	0.157
Konkavlık En Büyük	539	522 (%92)	0	1.252	0.272	0.209
Konkav Noktalar En Büyük	492	435 (%76)	0	0.291	0.115	0.066
Simetri En Büyük	500	437 (%77)	0.157	0.664	0.29	0.062
Oransal Kırılma Boyutu En Büyük	535	502 (%88)	0.055	0.208	0.084	0.018

Tablo 4.2: Çekirdek özelliklerine ait bazı istatistiksel analizler.

4.3 VERİYİ HAZIRLAMA

Bu aşamada yapılacak olan veriyi seçme, birleştirme ve verinin inşası işlemleri elimizdeki Wisconsin Göğüs Kanseri Teşhis Verisi için hali hazırda yapılmıştır.

```
842302,M,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.
6,2019,0.1622,0.6656,0.7119,0.2654,0.4601,0.1189
842517,M,20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,
8.8,1956,0.1238,0.1866,0.2416,0.186,0.275,0.08902
84300903,M,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.
,1709,0.1444,0.4245,0.4504,0.243,0.3613,0.08758
84348301,M,11.42,20.38,77.58,386.1,0.1425,0.2839,0.2414
87,567.7,0.2098,0.8663,0.6869,0.2575,0.6638,0.173
84358402,M,20.29,14.34,135.1,1297,0.1003,0.1328,0.198,0
.2,1575,0.1374,0.205,0.4,0.1625,0.2364,0.07678
843786,M,12.45,15.7,82.57,477.1,0.1278,0.17,0.1578,0.08
,741.6,0.1791,0.5249,0.5355,0.1741,0.3985,0.1244
844359,M,18.25,19.98,119.6,1040,0.09463,0.109,0.1127,0.
,1606,0.1442,0.2576,0.3784,0.1932,0.3063,0.08368
84458202,M,13.71,20.83,90.2,577.9,0.1189,0.1645,0.09366
110.6,897,0.1654,0.3682,0.2678,0.1556,0.3196,0.1151
844981,M,13,21.82,87.5,519.8,0.1273,0.1932,0.1859,0.093
39.3,0.1703,0.5401,0.539,0.206,0.4378,0.1072
84501001,M,12.46,24.04,83.97,475.9,0.1186,0.2396,0.2273
```

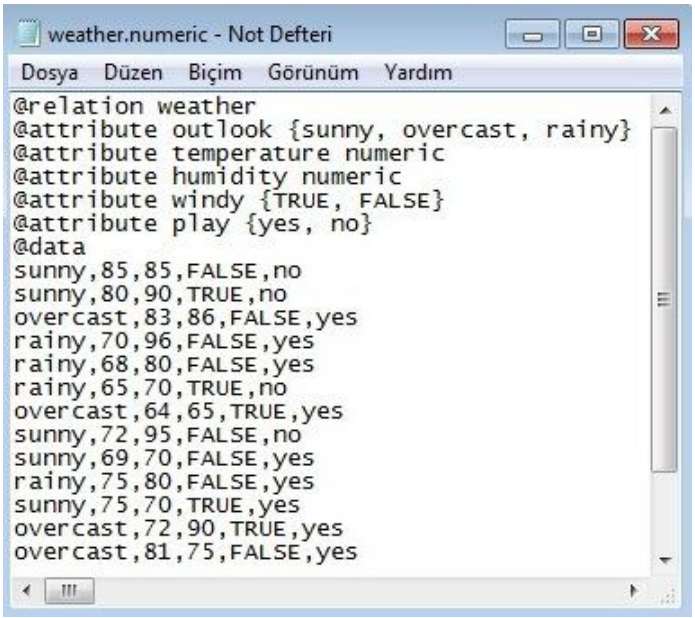
Şekil 4.5: Düzenlenmemiş Wisconsin Göğüs Kanseri Teşhis Verisi (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29).

Veriyi temizleme basamağında, örneklemin alındığı bireyi temsil eden ve her kaydın ilk özelliği olan "ID Number" bilgi keşfi sürecinde anlamlı olmadığı için veri setinden çıkarılmış ve sonuçların daha anlaşılır olması için özelliklerin Türkçeleri kullanılmıştır.

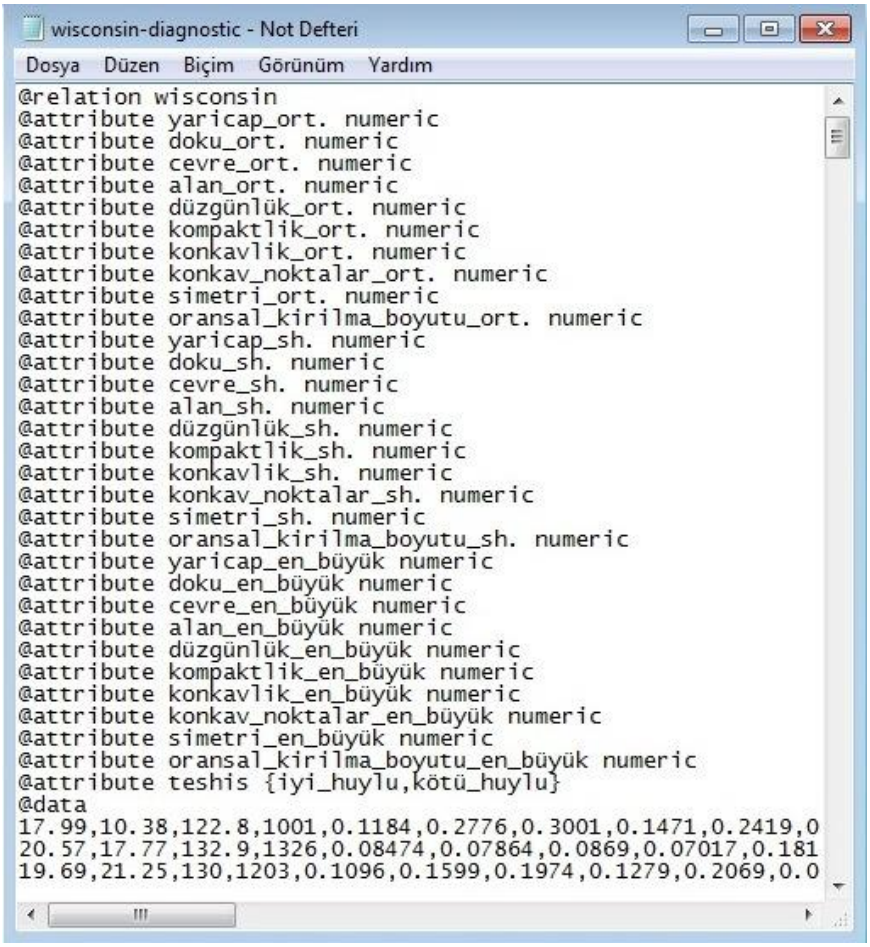
```
842302 ,M,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871
6.2019,0.1622,0.6656,0.7119,0.2654,0.4601,0.1189
842517 ,M,20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05
8.8,1956,0.1238,0.1866,0.2416,0.186,0.275,0.08902
84300903 ,M,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.1279,0.2069,0.05999
,1709,0.1444,0.4245,0.4504,0.243,0.3613,0.08758
```

Şekil 4.6: Veri içindeki ID Number özellikleri.

Veriyi biçimlendirme basamağında veri seti, WEKA'nın kullandığı "ARFF" (*Attribute Relation File Format*) biçimine dönüştürülmüştür. Bunun için WEKA/data klasöründeki arff dosyalarından weather.numeric taslak olarak kullanılmıştır.



Şekil 4.7: weather.numeric arff. dosyası.



Şekil 4.8: Arff. dosya formatına dönüştürülmüş Wisconsin Göğüs Kanseri Teşhis Verisi.

4.4 MODELLEME VE DEĞERLENDİRME

4.4.1 Özellik Seçimi

Yaygın olarak istatistik; örüntü tanıma ve medikal alanlarında kullanılan ve sınıflandırmada önemli bir rol oynayan özellik seçimi, veri madenciliğindeki ön işleme yöntemlerinden biridir. Veri setinin, tahmin sürecine uygun olmayan pek çok veri içermesi durumunda sınıflandırma performansının düşmesi nedeni ile özellik seçimi gibi yöntemler kullanarak çok miktardaki veriden daha güvenilir veri üretme ihtiyacı doğmuştur. Özellik seçimi, en uygun nitelikleri belirleme ve gereksiz nitelikleri mümkün olduğu kadar veri setinden çıkarma süreci olarak tanımlanabilir (Lavanya ve Rani 2011; Hall ve Holmes 2003; Bani Ahmad, 2013: 119).

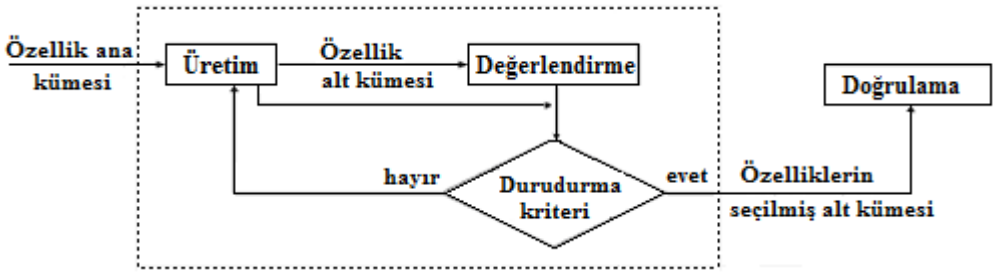
Özellik seçme süreci aşağıdaki basamaklardan oluşur:

1. Araştırma teknikleri kullanılarak özelliklerin ana kümesinden özelliklerin alt kümeleri üretilir.
2. Her bir alt kümenin sınıflandırmaya uygun olup olmadığı mesafe, bağımlılık, tutarlılık, sınıflandırma hata oranı gibi ölçüler kullanılarak değerlendirilir.
3. Optimal özellik alt kümesini saptamak için durdurma kriteri belirlenir.
4. Seçilen alt kümenin optimal olup olmadığı kontrol edilir.

Özellik seçiminin amaçları şunlardır (Bani Ahmad, 2013: 65; Guyon, 2003):

- Öngörünün doğruluk yüzdesini arttırmak,
- Daha hızlı çalışan bir sınıflandırma yöntemi seçmek,
- Daha az önemli ve gereksiz özellikleri görmezden gelmek,
- Veri kalitesini arttırmak,
- Aşırı öğrenmeden kaçınmak ve
- Büyük miktarda kullanılabilir veri problemini çözmeye ve bu veriden nasıl etkili bir şekilde faydalanılacağını bulmaya yardımcı olmaktır.

Şekil 4.9'da özellik seçme süreci gösterilmiştir.



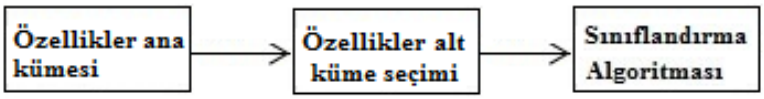
Şekil 4.9: Özellik seçme süreci (Hall ve diğ., 1997).

Özellik seçimi yöntemleri filtreleme ve sarmalama olarak ikiye ayrılır. Filtreleme yaklaşımı sınıflandırma yapılmadan önce veriye uygulanır. Bu yöntemde verinin genel niteliklerine dayanan sezgisel bir yöntemle özellikler değerlendirilir.

Sarmalama yaklaşımında ise özellikler sınıflandırma algoritmaları kullanılarak değerlendirilir (Lavanya ve Rani, 2011).

4.4.1.1 Filtreleme Yöntemi

Filtreleme yöntemleri özelliklerin yeterliliklerini değerlendirirken bir özellikler kümesi ile sınıf özelliği arasındaki istatistiksel korelasyonu kullanır. Çoğu durumda özellikler sıralanır ve alt sıralardaki özellikler öğrenme süreci boyunca değerlendirilmeye alınmaz. Daha sonra üst sıralardaki özelliklerin alt kümesi sınıflandırma algoritması için eğitim seti olarak kullanılır (Bani Ahmad, 2013: 67; Saeys ve diğ., 2007). Şekil 4.10'da görüldüğü üzere filtreleme yönteminin sarmalama yönteminden esas farkı filtreleme yönteminin, özellik alt kümelerini araştırma süreci boyunca, sınıflandırma algoritmalarını dikkate almamasıdır.



Şekil 4.10: Özellik seçiminde filtreleme yaklaşımı (Bani Ahmad, 2013: 69; Meesad ve. Yen, 2003).

Filtreleme yöntemleri ve sınıflandırma algoritmalarının birbirinden bağımsız olmalarının esas amacı özellik seçiminin sadece bir kez yapılması ve sonra farklı sınıflandırma algoritmaları ile seçilen özellik alt kümelerini değerlendirmektir. Ne var ki filtreleme yöntemleri ve sınıflandırma algoritmalarının birbirinden bağımsız olmaları sınıflandırma doğruluk yüzdesinin düşük seviyede olmasına neden olabilir (Bani Ahmad, 2013: 67; Saeys ve diğ., 2007).

Tablo 4.3 özellik seçiminde filtreleme yöntemini kullanmanın avantaj ve dezavantajlarını göstermektedir.

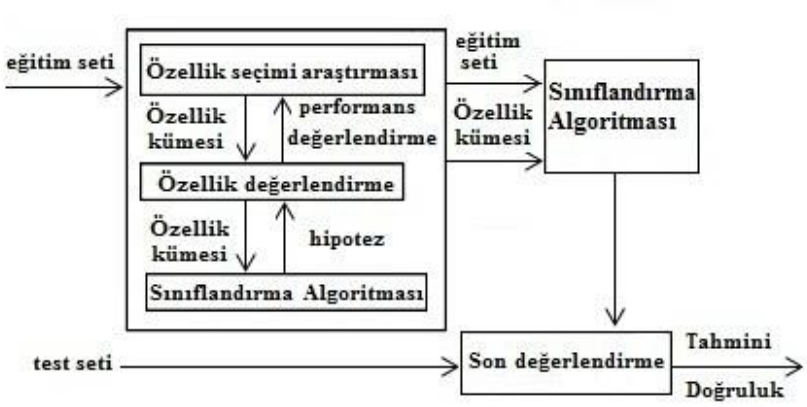
Avantajlar	Dezavantajlar
• Nispeten hızlıdır. • Çok fazla durum ve nitelik içeren çok büyük veri tabanlarına uygulanabilir. • Sınıflandırma algoritmasından bağımsızdır.	• Özelliklerin birbirleriyle olan bağımlılıklarını dikkate almaz. • Sınıflandırma algoritması ile etkileşimi dikkate almaz.

Tablo 4.3: Filtreleme özellik seçimi yönteminin avantaj ve dezavantajları (Bani Ahmad, 2013: 69; Saeys ve diğ., 2007).

4.4.1.2 Sarmalama Yöntemi

Sarmalama yöntemlerinde özellik seçme algoritması sınıflandırma (öğrenme) algoritmasının etrafına bir sargı gibi sarılır. Süreç, sınıflandırma algoritmasını kullanarak özelliklerin uygun bir alt kümesinin araştırılması ile başlar.

Şekil 4.11 sarmalama yaklaşımının değerlendirme sürecinde veri setine nasıl uygulandığını göstermektedir. Önce sınıflandırma algoritması araştırma yöntemi ile elde edilen özellik alt kümelerini değerlendirir ve özellik alt kümelerinin kaliteleri ve uygunlukları hakkında hipotezler elde eder. Sonra en yüksek tahmini değere sahip özellik alt kümesi sınıflandırma algoritmasının kullanacağı veri seti olarak seçilir. Son olarak eğitim ve test süreçlerinin birbirinden bağımsız olduğundan emin olmak için model, araştırma yönteminin kullandığı veri setinden başka bir veri seti üzerinde değerlendirilir. Sonuç olarak en uygun özellik alt kümesinin sınıflandırma algoritması tarafından kullanılması ile tahmini bir doğruluk yüzdesi elde edilir (Bani Ahmad, 2013: 65; Kohavi ve John, 1997).



Şekil 4.11: Özellik seçiminde sarmalama yaklaşımı (Bani Ahmad, 2013: 66; Kohavi ve John, 1997).

Sarmalama yöntemlerinde sınıflandırma algoritmaları, sınıflandırmanın içyapısına müdahale etmeden özellik alt kümelerinin kalitelerinin ölçümünde kullanılır.

Tablo 4.4 özellik seçiminde sarmalama yöntemini kullanmanın avantaj ve dezavantajlarını göstermektedir.

Avantajlar	Dezavantajlar
• Kullanması ve uygulaması kolaydır. • Sınıflandırma algoritması ile etkileşimlidir. • Özellikler arasındaki bağımlılıkları modeller.	• Aşırı öğrenme riski vardır. • Hesaplaması uzun sürer.

Tablo 4.4: Sarmalama özellik seçimi yönteminin avantaj ve dezavantajları (Bani Ahmad, 2013: 69; Saeys ve diğ, 2007).

Sarmalama yöntemi, sınıflandırmada kullanılacak algoritma ile hesaplanan doğruluk yüzdesine dayanarak özellikleri değerlendirir ve seçer. Sarmalama yöntemi, belirli bir sınıflandırma algoritması kullanarak özellik uzayını bazı özellikleri ihmal ederek araştırır ve ihmal edilen özelliklerin tahmin sonuçlarına nasıl etki ettiğini analiz eder. Öğrenme sürecinde bir özelliğin belirgin bir fark ortaya çıkarması o özelliğin önemli olduğu ve öğrenme açısından kaliteli bir özellik olduğu anlamına gelir.

4.4.2 Sınıflandırma

Kanser erken teşhisi için yapılacak çalışmanın hedefi sonuçları önceden bilinen hem kanser teşhisi konmuş hem de sağlıklı teşhisi konmuş hastaların durumları ve bu durumlarda ilgili faktörlerin aldığı değerler arasındaki ilişkileri tanımlamaktır. Bu amaçla denetimli öğrenme kapsamında değerlendirilen sınıflandırma yöntemi kullanılacaktır.

4.4.2.1 Eğitim ve Test Veri Setlerini Oluşturma ve Doğrulama

Eldeki verinin, rassal olarak, eğitim ve test setleri olmak üzere birbirinden bağımsız iki veri setine ayrılması yöntemine tutma (*holdout*) yöntemi denir. Genellikle verinin üçte ikisi eğitim setine ve geriye kalan üçte biri de test setine ayrılır. Eğitim seti modeli kurmak için kullanılır ve sonra modelin doğruluğu test seti kullanılarak değerlendirilir. Bu yöntemde modeli kurarken başlangıç verisinin sadece bir kısmı kullanıldığı için değerlendirme güvenilir sonuçlar vermez.

Bu çalışmada hem eğitim ve test için elimizdeki veri miktarı sınırlı olduğundan hem de küçük boyutlu veriler için daha iyi sonuçlar verdiğinden k-kat çapraz doğrulama (*k-fold cross-validation*) yöntemi kullanılmıştır.

k-kat çapraz doğrulama yönteminde başlangıç veri seti, rassal olarak, her biri yaklaşık olarak eşit büyüklükte olan k tane ayrık alt kümeye; $A_1, A_2, \dots, A_k$'ya ayrılır. Eğitim ve test süreçleri k defa tekrarlanır. i-inci iterasyonda A_i alt kümesi test seti olarak tutulur ve geriye kalan alt kümeler modeli eğitmek için kullanılır. Yani ilk iterasyonda $A_2, \dots, A_k$ alt kümeleri bir model elde etmek için eğitim seti olarak kullanılır ve model A_1 üzerinde test edilir, ikinci iterasyonda $A_1, A_3, \dots, A_k$ alt kümeleri eğitim seti olarak kullanılır ve model A_2 üzerinde test edilir ve işlem böyle devam eder. Tutma yönteminden farklı olarak burada her bir örneklem aynı sayıda tekrarla eğitim için ve bir defa da test için kullanılır.

WEKA, çapraz doğrulama işlemini otomatik olarak gerçekleştirir. Eldeki veriyi k tane ayrık alt kümeye ayırır ve öğrenme algoritmasını çapraz doğrulamanın her bir alt kümesi için bir kez ve son olarak tüm veri seti üzerinde olmak üzere toplam $k + 1$ defa çalıştırır. Her ne kadar k arttıkça tahmini doğruluk yüzesinin varyansı azalsa da WEKA programı, k-kat çapraz doğrulama yönteminde genellikle daha iyi sonuçlar verdiği gözlendiği için, varsayılan seçenek olarak $k = 10$ alır. Bu nedenle bu çalışmada 10-kat çapraz doğrulama yöntemi kullanılmıştır.

Yalnız bir 10-kat çapraz doğrulama yapmak güvenilir bir sonuç almak için yeterli olmayabilir. Daha güvenilir bir sonuç elde etmek için çapraz doğrulama süreci 10 defa tekrarlanır, yani 10 defa 10-kat çapraz doğrulama yapılır ve sonuçların ortalaması alınır. Bu sayede öğrenme algoritması elimizdeki verinin tamamının onda dokuzunu 100 defa kullanmış olur ve tahmini doğruluk yüzdesinin varyansı azalır (Witten ve diğ., 2011:152-154).

4.4.2.2 Sınıflandırmada Kullanılan Değerlendirme Kriterleri

1. **Hata Matrisi** (*Confusion Matrix*)**:** Hata matrisi, genellikle, sınıflandırma modellerinin doğruluğunu temsil eden bir görselleştirme aracıdır (Han ve Kamber, 2006: 365). Sonuçlar ve tahmin edilen sınıflar arasındaki ilişkileri gösterir.

Sınıflandırma modellinin etkinliği, her birinin olası değeri hata matrisinde verilmiş olan doğru ve yanlış sınıflandırılmış değişkenlerin sayısı ile hesaplanır (Cabena, ve diğ., 1998).

Bu çalışmada kullanılan hata matrisindeki değerler ve anlamları aşağıda verilmiştir:

– ***TP*** (*True Positive*)**:** İyi huylu tümörün doğru sınıflandırıldığı örneklem sayısıdır.

- ***TN*** (*True Negative*)**:** Kötü huylu tümörün doğru sınıflandırıldığı örneklem sayısıdır.
- ***FP*** (*False Positive*)**:** İyi huylu tümörün kötü huylu tümör olarak yanlış sınıflandırıldığı örneklem sayısıdır.
- ***FN*** (*False Negative*)**:** Kötü huylu tümörün iyi huylu tümör olarak yanlış sınıflandırıldığı örneklem sayısıdır.

		Öngörülen Sınıf	
		İyi Huylu Tümör	Kötü Huylu Tümör
Gerçek Sınıf	İyi Huylu Tümör	***TP***	***FP***
	Kötü Huylu Tümör	***FN***	***TN***

Tablo 4.5: Hata Matrisi

2. **Kappa İstatistiği** (*Kappa Statistics*)**:** Kappa istatistiği sınıflandırılmış verinin iki kümesi arasındaki uyum seviyesinin ölçüsü olarak tanımlanabilir. Örneğin, bir hastanın iki farklı uzman tarafından değerlendirilen FNA sonuçlarının birbirleriyle ne seviyede uyumlu oldukları Kappa değeri ile gösterilir. Kappa sonuçları $[0,1]$ aralığında değişir. Eğer Kappa$= 0$ ise hiç uyum yoktur. Kappa$=$ $[0.40,0.59]$ ise orta seviyede bir uyum vardır, Kappa$=$ $[0.60,0.79]$ ise iyi seviyede bir uyum vardır ve Kappa $=$ $[0.80,0.99]$ ise çok iyi seviyede bir uyum vardır. Eğer Kappa$= 1$ ise mükemmel bir uyum vardır (Landis ve Koch, 1977).

3. **Doğruluk** (*Accuracy*)**:** Doğru sınıflandırılmış örneklem sayısının toplam örneklem sayısına oranıdır.

$$Doğruluk = \frac{TN + TP}{FN + TN + TP + FP}$$

4. **Kesinlik** (*Precision*)**:** Doğru sınıflandırılmış pozitif örneklem sayısının sınıfı pozitif olarak öngörülmüş toplam örneklem sayısına oranıdır. $[0,1]$ aralığında bir değer alır. Yüksek bir kesinlik değeri, algoritmanın uygun olmayanlardan çok

uygun sonuçlar bulması anlamına gelir. Bir sınıflandırma işleminde bir C sınıfı için 1.0 kesinlik değeri C sınıfına ait olduğu tespit edilen tüm kayıtların gerçekten C sınıfında olduğu anlamına gelir, fakat C sınıfına ait olduğu halde C sınıfına dâhil edilmemiş kayıtların sayısı hakkında bir bilgi vermez.

$$Kesinlik = \frac{TP}{TP + FN}$$

5. **Duyarlılık** (*Recall*)**:** Doğru sınıflandırılmış pozitif örneklem sayısının gerçek sınıfı pozitif olan tüm örneklemlerin sayısına oranıdır. Duyarlılık, gerçek pozitiflik oranı (*True Positive Rate*) olarak da adlandırılır. $[0,1]$ aralığında bir değer alır. Yüksek bir duyarlılık değeri, algoritmanın uygun sonuçların çoğunu bulması anlamına gelir. Bir sınıflandırma işleminde 1.0 duyarlılık değeri gerçekte C sınıfına ait olan tüm kayıtların doğru sınıflandırıldığı anlamına gelir, fakat C sınıfına ait olmadığı halde C sınıfına dâhil edilen kayıtların sayısı hakkında bir bilgi vermez.

$$Duyarlılık = \frac{TP}{TP + FP}$$

6. **F-Ölçütü** (*F-Measure*)**:** Çoğunlukla kesinlik ve duyarlılık arasında zıt bir ilişki vardır ki birinin değerini arttırmak diğerinin değerini düşürebilir. Beyin tümörü ameliyatı bu duruma iyi bir örnektir. Farz edelim ki bir beyin cerrahı bir hastanın beynindeki kötü huylu tümörleri çıkarmak istesin. Cerrahın tüm tümörleri çıkarması gerekir aksi takdirde kalan kanserli hücreler yeniden tümör oluşturacaklardır. Buna karşılık cerrahın sağlıklı beyin hücrelerine dokunmaması gerekir ki aksi takdirde hastanın beyin fonksiyonları bozulabilir. Cerrah tüm kanserli hücreleri temizlediğinden emin olmak için beyin üzerinde daha serbest çalışarak hücrelerin çoğunu çıkarabilir. Cerrahın bu kararı duyarlılığı arttırır, fakat kesinliği azaltır. Öte yandan cerrah daha ihtiyatlı davranarak sadece kanserli olduğu tespit edilen hücreleri de temizleyebilir. Bu karar ise kesinliği artırırken duyarlılığı azaltır. Başka bir deyişle, yüksek duyarlılık sağlıklı hücreleri çıkarma ihtimalini arttırırken (olumsuz sonuç) tüm kanserli hücreleri çıkarma

ihtimalini de arttırır (olumlu sonuç) ve yüksek kesinlik sağlıklı hücreleri çıkarma ihtimalini azaltırken (olumlu sonuç) tüm kanserli hücreleri çıkarma ihtimalini de azaltır (olumsuz sonuç). Bu nedenle daha kesin ve duyarlı sonuçlar elde etmek için her iki ölçütün harmonik ortalaması olan F-ölçütü kullanılır.

$$F - \text{Ölçütü} = \frac{2 \times Duyarlılık \times Kesinlik}{Duyarlılık + Kesinlik}$$

7. **Yanlış Pozitiflik Oranı** (*False Positive Rate*)**:** Yanlış sınıflandırılmış pozitif örneklem sayısının gerçek sınıfı pozitif olan tüm örneklemlerin sayısına oranıdır.

$$Yanlış\ Pozitiflik\ Oranı\ = \frac{FP}{TP + FP}$$

8. **ROC Alanı** (*ROC Area*)**:** İkili bir sınıflandırma problemi için ROC eğrisi duyarlılığın, yanlış pozitiflik oranının bir fonksiyonu olarak gösterilmesi ile elde edilir. Bu eğrinin altındaki ROC alanı sınıflandırmada öngörülen değerler ve gerçek değerler arasında kıyaslama yapmayı sağlar. Bu özelliğinden dolayı doğruluk yüzdesi yüksek modelleri kullanmaktansa ROC alanı değeri yüksek modeller tercih edilir. ROC alanının değeri [0,1] aralığında değişir. Eğer algoritmanın bulduğu sonuçlar %100 duyarlılığa ve %0 yanlış pozitiflik oranına sahipse ROC eğrisi (0,1) noktasında kesişen ve birbirine dik iki doğru parçasından oluşur ve bu durumda ROC alanı değeri 1.00 olur.

4.4.2.3 Kullanılan Özellik Değerlendiriciler

1. **Sınıflandırıcı Özellik Değerlendirici** (*Classifier Attribute Eval*)**:** Bir özelliğin değerini, kullanıcı tarafından belirlenmiş bir sınıflandırma algoritması yardımıyla hesaplar.

2. **Ki-Kare Özellik Değerlendirici** (*Chi-Squared Attribute Eval*)**:** Bir özelliğin değerini, özelliğin sınıfa göre ki-kare istatistiğinin değerini hesaplayarak bulur.

3. **Korelasyon Özellik Değerlendirici** (*Correlation Attribute Eval*)**:** Bir özelliğin değerini, özellik ve sınıf arasındaki korelasyonu ölçerek hesaplar.

4. **Filtrelenmiş Özellik Değerlendirici** (*Filtered Attribute Eval*)**:** Keyfi olarak seçilen ve özelliklerin sırasını değiştirmeyen bir filtreleme yöntemi ile düzenlenmiş veri için keyfi olarak seçilen bir özellik değerlendirici kullanır.

5. **Kazanım Oranı Özellik Değerlendirici** (*Gain-Ratio Attribute Eval*)**:** Bir özelliğin değerini, özelliğin sınıfa göre kazanım oranını ölçerek hesaplar.

   H entropi[25] olmak üzere, kazanım oranı aşağıdaki formülle hesaplanır.

$$GainR(Class, Attribute) = \frac{H(Class) - H(Class \mid Attribute)}{H(Attribute)}$$

6. **Bilgi Kazanım Özellik Değerlendirici** (*Info-Gain Attribute Eval*)**:** Bir özelliğin değerini sınıfa göre bilgi kazanım oranını ölçerek hesaplar.

   H entropi olmak üzere, bilgi kazanım oranı aşağıdaki formülle hesaplanır.

$$InfoGain(Class, Attribute) = H(Class) - H(Class \mid Attribute)$$

7. **OneR Özellik Değerlendirici** (*OneR Attribute Eval*)**:** Bir özelliğin değerini OneR sınıflandırma algoritmasını kullanarak hesaplar. OneR sınıflandırma algoritması eğitim verisi üzerinde ve hata oranına göre özellikleri sıralar. Tüm nümerik değerli özellikleri sürekli kabul eder ve onları birkaç ayrık aralığa böler (Tripoliti ve Manis, 2012).

[25] Entropi, birden fazla olası sınıflandırma olmasına bağlı olarak, bir eğitim setindeki belirsizliğin teorik bilgi ölçüsüdür.

8. **ReliefF Özellik Değerlendirici** (*ReliefF Attribute Eval*)**:** Bir özelliğin değerini, bir örneklemi defalarca örnekleyerek hesaplar ve aynı ve farklı sınıfların en yakın örneği için verilen özellik değerini dikkate alır. Hem kesikli hem de sürekli veri üzerinde uygulanabilir.

9. **Önem Özellik Değerlendirici** (*Significance Attribute Eval*)**:** Olasılıksal önemi, özellik-sınıflar ve sınıflar-özellik ilişkisi şeklinde iki yönlü bir fonksiyon olarak hesaplayarak bir özelliğin değerini bulur.

10. **SVM Özellik Değerlendirici** (*SVM Attribute Eval*)**:** Bir özelliğin değerini, destek vektör makinesi (*Support Vector Machine; SVM*) sınıflandırma algoritmasını kullanarak bulur. Özellikler, SVM tarafından atanan ağırlığın karesi alınarak sıraya konulur.

11. **Simetrik Belirsizlik Özellik Değerlendirici** (*Symmetrical Uncert Attribute Eval*)**:** Bir özelliğin değerini özelliğin sınıfa göre simetrik belirsizliğini ölçerek hesaplar. Bu ölçüye göre eğer bir özellik bir sınıf ile yüksek korelasyona sahip fakat veri setindeki diğer özellikler ile düşük korelasyona sahip ise iyi bir özelliktir (Tripoliti ve Manis, 2012).

H entropi olmak üzere, simetrik belirsizlik aşağıdaki formülle hesaplanır.

$$SymmU(Class, Attribute) = 2 \times \frac{H(Class) - H(Class \mid Attribute)}{H(Class) + H(Attribute)}$$

4.4.2.4 Kullanılan Sınıflandırma Algoritmaları

1. **Bayes Ağı** (*Bayes Net*)**:** Tüm özelliklerin nominal olduğu ve hiçbir kayıp değer olmadığı varsayımları altında Bayes ağlarını[26] öğrenirler (Witten ve diğ., 2011: 453).

2. **Sade Bayes** (*Navie Bayes*)**:** Bayes ağları içinde en basit yaklaşımdır ve bir veri setinin tüm özelliklerinin sınıf değişkeninin değerinden bağımsız olduğu kabulünü dikkate alır. Bundan dolayı Naive Bayes sınıflandırma algoritması, sınıfın bir ataya sahip olmadığı ve sınıfın her bir özellik için bir tek ata olduğu bir Bayes ağıdır (Kumar ve Sahoo, 2012

3. **Çok Katmanlı Algılayıcı** (*Multilayer Perceptron*)**:** Bir yapay sinir ağı; girdi değerlerini tanımlayan bir girdi katmanı, matematiksel fonksiyonu tanımlayan bir veya daha fazla saklı katman ve sonuçları tanımlayan bir çıktı katmanı olmak üzere üç katmandan oluşur. Her bir katman birbiriyle bağlantılı çok sayıda nörondan oluşur ve her bir katman bir önceki katmandan gelen girdiyi alıp bir sonraki katman için sonuç üreten bir aktivasyon fonksiyonuna (matematiksel fonksiyona) sahiptir. Dolayısıyla yapay sinir ağlarında öngörü aktivasyon fonksiyonu sayesinde yapılır. Çok katmanlı algılayıcı veya ileri beslemeli ağ ise saklı katmanda bir non-lineer (doğrusal olmayan) aktivasyon fonksiyonu içeren bir tür yapay sinir ağıdır. Girdi ve çıktı vektörleri arasında bir non-lineer bağlantı sağlar (Kumar ve Sahoo, 2012).

4. **Basit Lojistik** (*Simple Logistic*)**:** Lineer lojistik regresyon modellerini inşa etmek için kullanılan bir sınıflandırma algoritmasıdır.

[26] Değişkenlere ait koşullu olasılık dağılımlarını ortaya koymak ve değişkenlere ait alt kümeler arasındaki koşullu bağımsızlıkları tanımlamak üzere kullanılır.

5. **Olasılıksal Gradyan İniş**: Çeşitli lineer modelleri (ikili sınıf içeren destek vektör makineleri, lojistik regresyon ve lineer regresyon) öğrenmek için olasılıksal gradyan iniş (*Stochastic Gradient Descent; SGD*) uygular. Genelde tüm kayıp değerler için bir değer atar ve nominal özellikleri ikili özelliklere dönüştürür. Ayrıca tüm özellikleri normalleştirir, böylece çıktıdaki tüm katsayılar normalleştirilmiş veriye dayanır.

6. **Sıralı Minimal Optimizasyon**: Bir destek vektörü sınıflandırma algoritmasını eğitmek için sıralı minimal optimizasyon (*Sequential Minimal Optimization; SMO*) algoritmasını kullanır. Bu uygulama genelde tüm kayıp değerlere bir değer atar, nominal özellikleri ikili özelliklere dönüştürür ve tüm özellikleri normalleştirir.

7. **IB1:** Verilen test örneklemine en yakın Öklid uzaklığındaki eğitim örneklemini bulan ve bu test örnekleminin sınıfını eğitim örnekleminin sınıfı ile aynı sınıf olarak öngören bir temel örnek tabanlı sınıflandırma algoritmasıdır. Eğer birden fazla en yakın örneklem varsa bulunan ilk örneklem kullanılır (Witten ve diğ., 2011: 472).

8. **KYıldız** (*KStar; K**)**:** Bir test örnekleminin ait olduğu sınıfın eğitim örneklemlerine bağlı olduğu ve eğitim örneklemlerinin bu test örneklemine benzerliğinin bir benzerlik fonksiyonu tarafından belirlendiği bir temel örnek tabanlı sınıflandırma algoritmasıdır. Entropi tabanlı bir uzaklık fonksiyonu kullanması ile diğer temel örnek tabanlı sınıflandırma algoritmalarından ayrılır.

9. **PART:** Kısmi karar ağaçlarını[27] kullanarak kurallar elde eder. Kural oluştururken böl ve yönet stratejisini kullanır, kural oluşturmak için kullandığı örneklemleri

[27] Bir kısmi karar ağacı, tanımlanmamış alt ağaçlara giden dallar içeren bir karar ağacıdır. Böyle bir ağaç oluşturmak için inşa etme ve budama işlemleri daha fazla basitleştirilemeyen bir sabit alt ağaç bulunana kadar uygulanır. Bir kez sabit bir alt ağaç bulunduğunda inşa işlemi durur ve tek bir kural elde edilir.

kaldırır ve hiçbir örneklem kalmayana kadar kalan örneklemlerden tekrar tekrar kurallar oluşturur. Bu algoritma her bir kuralın oluşturuluş şekli ile standart karar ağacı yaklaşımdan farklıdır. Sadece bir kural oluşturmak için eldeki örneklemlerden bir budanmış karar ağacı oluşturulur, en büyük alanı kaplayan yaprak bir kural olarak kabul edilir ve ağacın geri kalanı atılır (Witten ve diğ., 2011: 208,459).

10. **Lojistik Model Ağaçları** (*Logistic Model Trees; LMT*)**:** Yapraklarında lojistik regresyon fonksiyonları olan sınıflandırma ağaçlarıdır. Bu algoritma ikili ve çoklu sınıf hedef değişkenlerinin, nümerik ve nominal özelliklerin ve kayıp değerlerin varlığı halinde kullanılabilir.

11. **Rassal Ormanlar** (*Random Forests*)**:** Her bir ağacın bağımsız olarak örneklenmiş bir rassal vektörün değerlerine bağlı olduğu ve her bir ağacın aynı dağılıma sahip olduğu bir ağaç kombinasyonudur. Ağaç sayısı arttıkça orman için genelleme hatası bir limite yakınsar. Bir ormanın genelleme hatası ormandaki her bir ağacın gücüne ve bu ağaçlar arasındaki korelasyona bağlıdır (Breiman, 2001).

4.4.2.5 Kullanılan Araştırma Yöntemleri

1. **En İyi İlk** (*Best First*)**:** Özellik uzayını bir geri iz sürme aracı ile iyileştirilmiş açgözlü[28] tepe tırmanma algoritmasını[29] kullanarak araştırır. Peş peşe gelen iyileşmeyen düğümlerin sayısını ayarlayarak geri iz sürmenin seviyelerini kontrol etmeyi sağlar. Boş bir özellikler kümesi ile başlar ve ileriye doğru araştırır veya tüm özellikler kümesi ile başlar ve geriye doğru araştırır veya herhangi bir noktadan başlar ve her iki yöne doğru araştırır.

[28] Bir genel optimum nokta bulmak umudu ile her safhada yerel optimal tercih yapar. Bu algoritmalara açgözlü denmesinin sebebi küçük örneklemlerin her biri için optimal çözümler bir çıktı sağlarken, algoritma daha büyük problemleri bir bütün olarak ele almaz.

[29] Arama işleminin yapıldığı grafikteki tepelerden ismini alır. Basitçe bir grafikte bulunan en düşük noktanın aranması sırasında grafikte yapılan hareketin aslında tepe tırmanmaya benzemesinden ismini almaktadır.

2. **Evrimsel Araştırma** (*Evolutionary Search*)**:** Düzgün rassal başlangıç, ikili turnuva seçimi, tek nokta çaprazlama, mutasyon ve seçkincilik ile kuşak değiştirme (yeni en iyi birey her zaman elde tutulur) gibi işleçler ile bir evrimsel algoritma[30] kullanarak özellik uzayını araştırır.

3. **Genetik Araştırma** (*Genetic Search*)**:** Parametreleri içinde ana kütle büyüklüğü, nesillerin sayısı, çaprazlama ve mutasyon olasılıkları olan basit bir genetik algoritma kullanır (Witten ve diğ., 2011: 493).

4. **Açgözlü Adımsal Araştırma** (*Greedy Stepwise*)**:** Özellik alt kümeleri boyunca ileriye veya geriye doğru bir açgözlü araştırma gerçekleştirir. Boş bir özellik kümesi ile veya tüm özellikler kümesi ile veya özellik uzayındaki keyfi bir noktadan başlayabilir. Kalan özelliklerin eklenmesi veya silinmesi değerlendirmede bir düşüşe neden olana kadar devam eder. Ayrıca özellik uzayını bir uçtan diğer uca tarayarak özelliklerin bir sıralamasını oluşturabilir ve seçilen özelliklerin sırasını kaydeder.

5. **Lineer İleri Doğru Seçim** (*Linear Forward Selection*)**:** En iyi ilk araştırma yönteminin uzantısıdır, en iyi ve sınırlı sayıda k tane özelliği dikkate alır. Bu araştırma yöntemi tüm özellikleri sıralar ve sabit k tane en iyi özelliği seçer.

6. **PSO Araştırma:** Bu araştırma yöntemi özellik uzayını, sürekli değişkenlere sahip ve kesikli değişkenlere sahip optimizasyon problemlerini çözmede kullanılan popülasyon tabanlı stokastik bir yaklaşım olan parçacık sürü optimizasyonu[31] (*Particle Swarm Optimization; PSO*) yöntemini kullanarak tarar.

---

[30] Biyolojik evrimden esinlenerek üreme, mutasyon, rekombinasyon ve doğal seçilime benzer mekanizmalar kullanır. Optimizasyon problemlerinin aday çözümleri bir popülasyondaki bireyleri temsil eder ve seçilim değeri fonksiyonları çözümlerin içinde yaşadığı çevreyi belirler.

[31] Optimizasyon probleminin araştırma uzayında hareket eden bir parçacığın konumu, optimizasyon probleminin eldeki bir aday çözümünü temsil eder. Her bir parçacık esas olarak kuş sürülerinin davranış modellerinden ilham almış kurallara göre hızını değiştirerek araştırma uzayında kendisine daha iyi bir konum arar.

7. **Yarış Araştırma** (*Race Search*)**:** Bu araştırma yöntemi çapraz doğrulamada karşılaştırılan özellik alt kümelerinin çapraz doğrulama hataları üzerinde eşleştirilmiş t- testi[32] (*paired t-test*) ve eşleştirilmemiş t-testini[33] (*unpaired t-test*) uygular. Karşılaştırılan iki özellik alt kümesinin hata ortalamaları arasında belirgin bir fark gözlendiğinde hata ortalaması yüksek olan özellik alt kümesi yarıştan elenir. Eğer karşılaştırılan iki özellik alt kümesinin hata ortalamaları arasında belirgin bir fark yoksa iki kümeden biri yarıştan elenir.

8. **Sıralama Araştırma** (*Rank Search*)**:** Tüm özellikleri sıralamak için bir özellik değerlendirici kullanır. Eğer bir özellik alt kümesi değerlendirici belirlenmişse bu değerlendirici, özellik alt kümelerinin sıralanmış bir alt kümesini oluşturmak için ileri doğru seçim araştırma yöntemini kullanır. Sıralanmış özellikler listesinden faydalanılarak oluşturulan artan boyuttaki özellik alt kümeleri değerlendirir, yani en iyi özelliği, en iyi özellik ve en iyi özellikten bir sonraki en iyi özelliği içeren ve bu şekilde devam eden kümeleri değerlendirir.

9. **Dağınık Arama** (*Scatter Search*)**:** Özellik alt kümeleri uzayı boyunca bir dağınık arama[34] gerçekleştirir.

10. **Alt Küme Boyutu İleri Doğru Seçim** (*Subset Size Forward Selection*)**:** Lineer ileri doğru seçim araştırma yönteminin bir uzantısıdır. Başlangıç değeri (*seed*) ve kat sayısı belirlenmiş bir çapraz doğrulama gerçekleştirir. Verilen bir alt küme

[32] Eşleştirilmiş t-testi, bir örneklemdeki gözlem sonuçlarının diğer bir örneklemin gözlem sonuçları ile karşılaştırılabildiği durumlarda ana kütle ortalamalarının karşılaştırılmasında kullanılır. Analiz edilen verilerin normal dağılıma sahip olduğunu kabul eder ve genellikle bir grup örneklemin iki farklı test koşulu altında verdiği sonuçların karşılaştırılmasında kullanılır.

[33] Eşleştirilmemiş t-testi, aynı veya farklı büyüklükte bağımsız iki grubun ortalamalarının karşılaştırılmasında kullanılır. Örneklemlerin alındığı ana kütlelerin ortalamalarının eşit olduğu sıfır hipotezini test eder. Sürekli değerler alan veriler için uygulanan eşleştirilmemiş t-testi verinin normal dağılıma sahip olduğunu ve her iki ana kütlenin standart sapmalarının aynı olduğunu kabul eder.

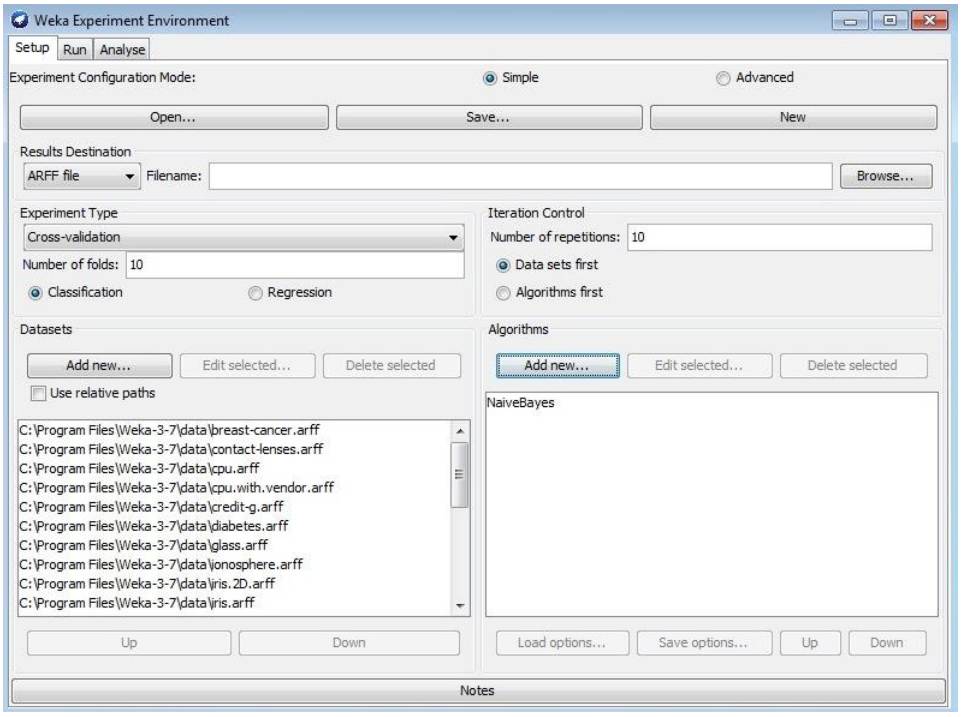
[34] Dağınık arama metodu, yeni çözümler elde edebilmek için referans küme adı verilen bir kümedeki çözümleri birleştirir. Dağınık arama metodunun diğer evrimsel algoritma metotlarından temel farkı, çözümü seçme yoludur. Dağınık Arama metodunda sistematik olarak referans kümesinden iki ya da daha fazla çözüm seçilerek yeni sonuçlar üretilir.

boyutu değerlendiriciyi kullanarak optimal alt küme boyutunu tespit etmek için her bir kat üzerinde lineer ileri doğru seçim gerçekleştirilir.

11. Sıralayıcı (*Ranker*)**:** Her bir özelliği ayrı ayrı değerlendirerek bir sıralama yapar. Özellik ReliefF, Gain Ratio gibi özellik değerlendiriciler ile birlikte kullanılır.

4.4.3 Sarmalama Ve Filtreleme Yaklaşımları İle Özellik Seçimi Ve Sınıflandırma

Bu çalışmada özellik seçimi ve sınıflandırma işlemleri hem 10 defa 10-kat çapraz doğrulama işlemini otomatik olarak yapması hem de elde edilen sonuçların ortalamalarını ve bu ortalamaların eşleştirilmiş t-testi ile karşılaştırmalarını vermesi sebebi ile WEKA'nın "Experimenter" ortamı kullanılarak yapılmıştır.



Şekil 4.12: WEKA Experimenter ortamı.

Bu çalışmada uygulanan filtreleme yöntemi ile özellik seçimi ve elde edilen özellik alt kümeleri ile oluşturulan modellerin karşılaştırmaları Ek-2'de ve sarmalama yöntemi ile özellik seçimi ve sınıflandırma sonuçlarının karşılaştırmaları Ek-3'te verilmiştir. Bu amaçla hem aynı yaklaşımla elde edilen sonuçları karşılaştırmak hem de her iki yaklaşımın en iyi sonuçlarını karşılaştırmak için eşleştirilmiş t-testi uygulanmıştır. Experimenter ortamından faydalanılarak uygulanan eşleştirilmiş t-testinin sonuç vermediği durumlarda ise ilk olarak modelin gerçeğe yakınlığının bir ölçüsü olan "ROC Alanı" değeri dikkate alınmıştır, eğer ROC alanı değerleri aynı olan iki model var ise ikinci olarak doğru sınıflandırmanın bir ölçüsü olan "Doğruluk" yüzdesine bakılmıştır. Hem ROC alanı değerlerinin hem de doğruluk yüzdelerinin eşit olması halinde ise kesinlik ve duyarlılığın ortak bir ölçüsü olan F-ölçütüne göre model seçilmiştir.

4.4.3.1 Filtreleme Yöntemi İle Özellik Seçimi Ve Elde Edilen Sınıflandırma Sonuçları

Filtreleme yöntemi ile özellik seçiminde, önce bir özellik değerlendirici ve bu özellik değerlendirici ile birlikte kullanılabilecek bir araştırma yöntemi seçilir. Sonra belirlenen uygun özellik alt kümeleri sınıflandırma işleminde kullanılır.

Tablo 4.6 en iyi sonuç veren 11 farklı sınıflandırma algoritması için filtreleme yaklaşımı ile yapılan özellik seçimi işlemi ile sıralanan 30 özellik içinden ilk 15 özellik kullanılarak elde edilen ve en yüksek doğruluk yüzdesini, F-ölçütünü ve ROC alanı değerini veren özellik alt kümelerini göstermektedir.

Özellik Değerlendirici: SVM Attribute Eval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	15	0.89	94.91	0.96	0.96	0.96	0.99	3.19
NaiveBayes	15	0.87	94.08	0.95	0.96	0.95	0.99	3.26
MultilayerPerceptron	15	0.92	96.49	0.97	0.98	0.97	0.99	2.42
SDG	15	0.94	97.33	0.97	0.99	0.98	0.97	2.12
SimpleLogistic	**15**	**0.94**	**97.38**	**0.97**	**0.99**	**0.98**	**1.00**	**2.18**
SMO	15	0.95	97.88	0.97	0.99	0.98	0.97	1.84
IB1	15	0.91	95.94	0.97	0.97	0.97	0.96	2.40
KStar	15	0.90	95.36	0.95	0.98	0.96	0.99	2.42
PART	15	0.89	94.89	0.96	0.96	0.96	0.95	2.79
LMT	15	0.94	97.31	0.97	0.99	0.98	1.00	2.32
RandomForest	15	0.92	96.21	0.96	0.98	0.97	0.99	2.64

Tablo 4.6: Filtreleme özellik seçimi yaklaşımında en iyi sonuçlar.

SVM Attribute Eval özellik değerlendirici ve Ranker araştırma yöntemi ile elde edilen özellikler Tablo 4.7'de sıralanmıştır.

1. yarıçap_en_büyük	2. konkav_noktalar_ en_büyük	3. çevre_ en_büyük
4. doku_ en_büyük	5. konkav_noktalar_ort.	6. alan_ en_büyük
7. simetri_ en_büyük	8. yarıçap_ort.	9. düzgünlük_ en_büyük
10. alan_ort.	11. yarıçap_std.h.	12.doku_ort.
13.çevre_ort.	14. konkavlık_ort.	15. kompaktlık_std.h.
16. çevre_std.h.	17. oransal_kırılma_ boyutu_ort.	18. konkavlık_ en_büyük
19. alan_std.h.	20. simetri_ort.	21. kompaktlık_ort.
22. düzgünlük_ort.	23. düzgünlük_std.h.	24.oransal_kırılma_ boyutu_std.h.
25. oransal_kırılma_ boyutu_ en_büyük	26. doku_std.h	27. konkavlık_std.h.
28. konkav_noktalar_std.h.	29. simetri_std.h.	30. kompaktlık_ en_büyük

Tablo 4.7: Filtreleme özellik seçimi yaklaşımı ile sıralanan özellikler ana kümesi.

4.4.3.2 Sarmalama Yöntemi İle Özellik Seçimi Ve Elde Edilen Sınıflandırma Sonuçları

Sarmalama yöntemi sınıflandırma algoritmasına bağımlı bir özellik seçimi gerçekleştirir. Hem en uygun özellik alt kümesinin seçiminde hem de sonrasında sınıflandırmada aynı sınıflandırma algoritmasını kullanır.

Tablo 4.8 sarmalama yaklaşımı ile yapılan özellik seçimi ve sınıflandırma işlemlerinde; 11 farklı özellik araştırma yöntemi için en yüksek doğruluk yüzdesini, F-ölçütünü ve ROC alanı değerini veren özellik alt kümelerini göstermektedir.

Özellik Değerlendirici: Classifier Subset Eval
Sınıflandırıcı: Simple Logistic

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	8	0.95	97.89	0.98	0.99	0.98	0.99	1.90
Evolutionary	10	0.96	97.93	0.98	0.99	0.98	0.99	2.03
Genetic	15	0.94	97.36	0.97	0.99	0.98	1.00	2.27
Greedy Stepwise	5	0.94	97.44	0.97	0.99	0.98	0.99	2.14
Linear Forward Selection	8	0.95	97.61	0.98	0.99	0.98	0.99	2.09
PSO	**15**	**0.94**	**97.43**	**0.97**	**0.99**	**0.98**	**1.00**	**2.00**
Race	6	0.92	96.08	0.96	0.97	0.97	0.99	2.48
Rank	20	0.94	97.45	0.97	0.99	0.98	0.99	2.12
Scatter	5	0.95	97.49	0.97	0.99	0.98	0.99	2.28
Subset Size Forward Selection	5	0.95	97.75	0.97	0.99	0.98	0.99	2.07

Tablo4.8: Sarmalama özellik seçimi yaklaşımında en iyi sonuçlar.

Tablo 4.8'de de görüldüğü üzere filtreleme ve sarmalama yöntemleri uygulandıktan sonra oluşturulan tüm modeller içinde en iyi sonuç Classifier Subset Eval özellik değerlendirici altında Simple Logistic sınıflandırma algoritması ve PSO özellik araştırma yöntemi uygulanarak elde edilmiştir. Buna göre Tablo 4.9'daki özellikleri içeren ve teşhis sonuçları belli olmayan hastalara ait bir veri seti üzerinde 10 defa 10-kat çapraz doğrulama kullanılarak Simple Logistic sınıflandırma algoritması ile gerçekleştirilecek bir sınıflandırma işleminde %100'e yakın doğrulukta ve güvenirlikte sonuçlar elde edilmesi beklenmektedir.

Classifier Subset Eval özellik değerlendirici altında Simple Logistic sınıflandırma algoritması ve PSO özellik araştırma yöntemi ile elde edilen en iyi özellikler Tablo 4.9'da verilmiştir.

1- yarıçap_ort.	4- konkav_ noktalar _ort.	7- alan_std.h.	10- simetri _std. h.	13- alan_en_büyük
2- doku_ort.	5- doku_std.h.	8- düzgünlük _std.h.	11- yarıçap _en _büyük	14- konkav_ noktalar_en_büyük
3- konkavlık _ort.	6- çevre_std.h.	9- kompaktlık _std.h.	12- doku _en _büyük	15- simetri_en _büyük

Tablo 4.9: Sarmalama özellik seçimi yaklaşımı ile edilen en uygun 15 özellik.

BÖLÜM 5: SONUÇ VE ÖNERİLER

Yapılan bu çalışma sonucunda, büyük veri yığınları içindeki altın değerindeki bilgiyi keşfetmeyi sağlayan veri madenciliği hakkında aşağıdaki sonuçlara ulaşılmıştır.

1. Veri madenciliği uygulamasının yapıldığı alana ait bilgi sadece veri madenciliği sürecinin başlangıcında hedefleri tanımlarken ve sonunda sonuçların konuşlandırılmasında değil, sürecin her aşamasında kullanılır. İşi anlama aşamasında elde edilen bilgileri bir veri madenciliği problemine dönüştürmede, veriyi anlama aşamasında probleme uygun verilerin seçiminde, veriyi hazırlama aşamasında veriye soru sorup cevap alarak veriden bilgi edinmek için veriyi şekillendirmede çalışma alanı bilgisi kullanılır ve konuşlandırma aşamasında, elde edilen veri madenciliği sonuçları çalışma alanında kullanılır. Ayrıca modelleme, öngörüsel modeller oluşturmak ve bu modelleri çalışma alanının açısıyla yorumlamak için veri madenciliği algoritmalarını kullanmak ve değerlendirme de bulunan uygun modelleri kullanmanın çalışma alanı üzerindeki etkisini anlamak anlamına gelir.

Kısaca, verinin taşıdığı bilgi ile bu bilginin gerçek hayattaki yeri arasında bir boşluk vardır ve bu boşluk veriden elde edilen bilginin çalışma alanı bilgisi ile yorumlanması ile doldurulur.

2. Veriyi hazırlama aşamasında yapılan temizleme, dönüştürme, vb. işlemlerin her biri analiz edilecek problem uzayının değişmesi anlamına gelir. Bu işlemleri gerçekleştirirken amaç veriyi, veri madenciliği sorularının sorulabileceği ve veri madenciliği algoritmaları gibi analitik tekniklerin bu soruları cevaplamasını kolaylaştıracak bir hale getirmektir. Bu nedenle veriyi hazırlama, veri madenciliği sürecinde önemli bir role sahiptir. Örneğin, bu çalışmada kullanılan Wisconsin Göğüs Kanseri Teşhis Verisi daha önceki çalışmalarda veri madenciliği uzmanlarının, algoritmaların bir çözüm bulmasını kolaylaştırmak

için problem uzayını özenle yeniden şekillendirmelerinden dolayı özellik seçimi işlemi gerçekleştirilmeden bile yüksek doğruluk ve güvenilirlikte sonuçlar vermiştir.

Özellik Değerlendirici	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	30	0.88	94.40	0.96	0.96	0.96	0.99	3.29
NaiveBayes	30	0.86	93.31	0.94	0.96	0.95	0.98	3.30
MultilayerPerceptron	30	0.93	96.70	0.97	0.98	0.97	0.99	2.37
SDG	30	0.95	97.80	0.97	0.99	0.98	0.97	1.80
SimpleLogistic	**30**	**0.94**	**97.40**	**0.97**	**0.99**	**0.98**	**1.00**	**2.36**
SMO	30	0.95	97.54	0.97	0.99	0.98	0.97	1.85
IB1	30	0.91	95.64	0.97	0.97	0.97	0.95	2.32
KStar	30	0.88	94.66	0.94	0.97	0.96	0.99	2.76
PART	30	0.88	94.24	0.96	0.95	0.95	0.95	2.74
LMT	30	0.94	97.36	0.97	0.99	0.98	1.00	2.36
RandomForest	30	0.91	95.73	0.96	0.98	0.97	0.99	2.52

Tablo 5.1: Özellikler ana kümesi için sınıflandırma sonuçları.

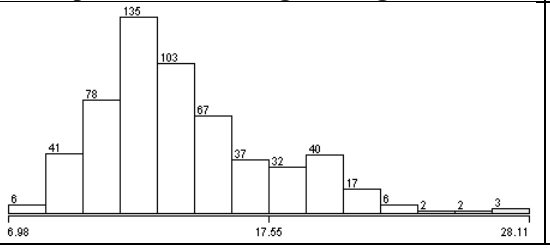
3. Veri madenciliği; uygulamanın yapıldığı alan bilgisine sahip bir kullanıcı tarafından, bir veri madenciliği problemine dönüştürülmüş ve iyi tanımlanmış iş hedeflerine ulaşmak için, eksiksiz ve uygun veriler kullanılarak yapılsa bile en uygun özelliklerin seçimi ve en iyi modelin kurulması için eldeki tüm ihtimallerin değerlendirilmesine ihtiyaç duyar. Yani bir uygulamada kullanılacak en uygun model sadece denenerek bulunabilir. Örneğin bu çalışmada yapılan özellik seçimi ve sınıflandırma basamaklarında, en uygun özellikler alt kümesini ve en iyi sınıflandırma modelini bulmak için Ek-2 ve Ek-3'teki tablolar oluşturulmuş ve karşılaştırılmıştır.

4. Veri madenciliği, veriden bilgi keşfine uzanan süreçte otomatik olarak çalışan bir teknoloji değildir. Her aşamasında insanla etkileşimde olan ve merkezinde bir veya daha fazla iş hedefi olan bir süreçtir. Öyle ki veri madenciliği, üzerinde çalışılan problem uzayındaki saklı örüntü ve düzenleri açığa çıkarsa bile bu örüntü ve düzenlerin nedenlerini açıklamak insanlara kalmıştır.

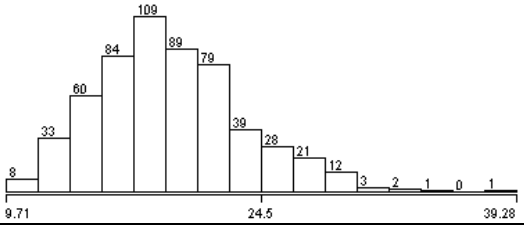
5. Kanser erken teşhis sürecinin başarı ile uygulanabilmesi için insan kaynaklı hataların en aza indirgenmesi gerekmektedir. Her ne kadar veri madenciliği bu anlamda hastalar için bir umut ışığı olsa da veri madenciliğinin kendisi de insan kaynaklı hatalar ve gerçek hayat problemlerinin her zaman bilgisayar mantığına uyarlanamaması nedeniyle yanılmaya açıktır. Yine de elde edilen yüksek doğrulukta ve gerçeğe son derece uygun sonuçlar göz önüne alındığında hastaya zarar vermeyecek, kanser erken teşhisinde verilen kararları destekleyecek ve veri içindeki gizli örüntüleri keşfetmeyi sağlayacak bir yöntem olarak veri madenciliğine önem vermenin gerekliliği görülmektedir.

Ek-1: ÖZELLİKLERE AİT İSTATİSTİKSEL DAĞILIMLAR

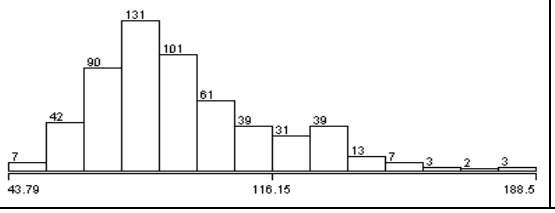
Yarıçap Ortalama Özelliğinin Dağılımı



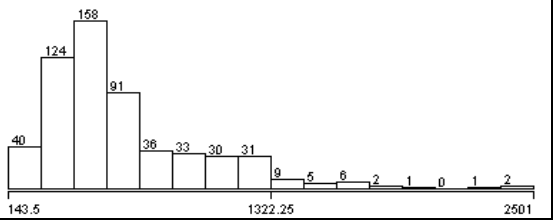
Doku Ortalama Özelliğinin Dağılımı



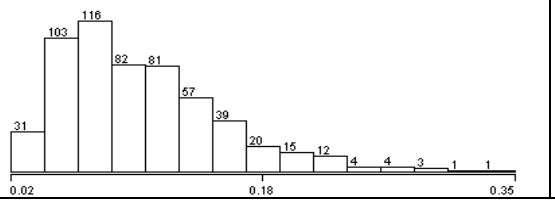
Çevre Ortalama Özelliğinin Dağılımı



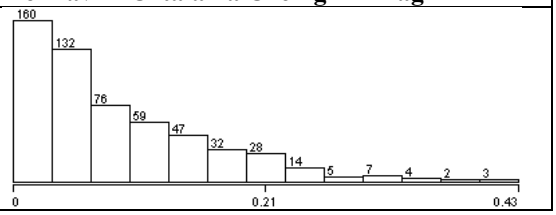
Alan Ortalama özelliğinin dağılımı



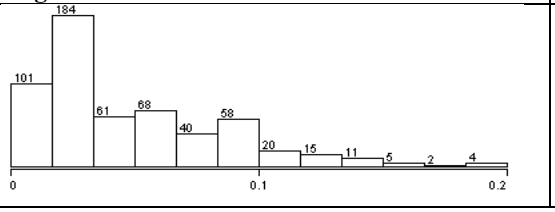
Kompaktlık Ortalama Özelliğinin Dağılımı



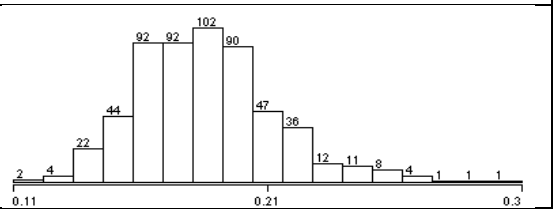
Konkavlık Ortalama Özelliğinin Dağılımı



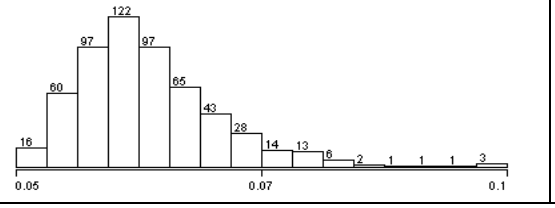
Konkav Noktalar Ortalama Özelliğinin Dağılımı



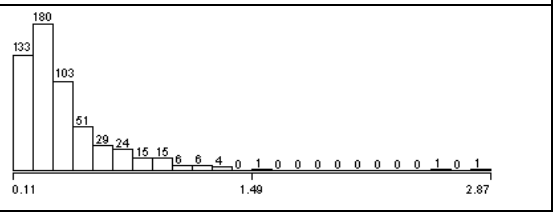
Simetri Ortalama Özelliğinin Dağılımı



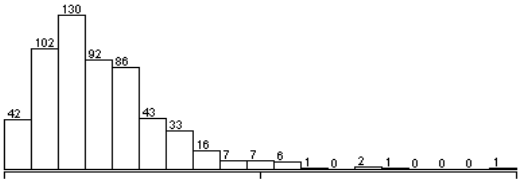
Oransal Kırılma Boyutu Ortalama Özelliğinin Dağılımı



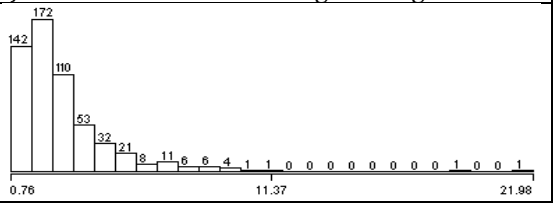
Yarıçap Standart Hata Özelliğinin Dağılımı



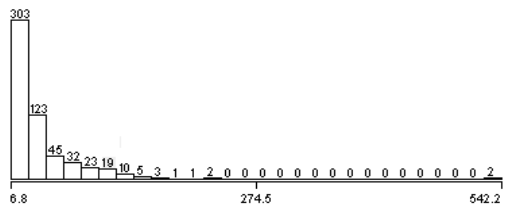
Doku Standart Hata Özelliğinin Dağılımı



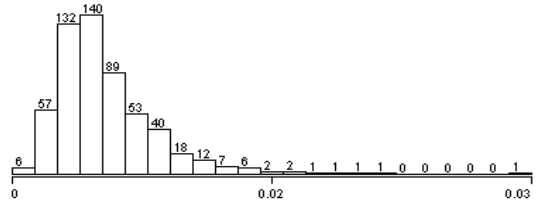
Çevre Standart Hata Özelliğinin Dağılımı



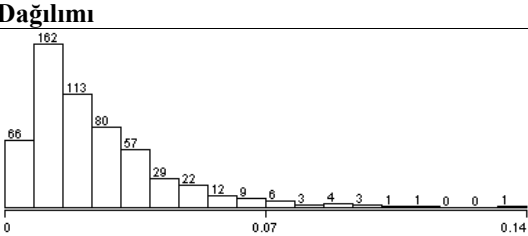
Alan Standart Hata Özelliğinin Dağılımı



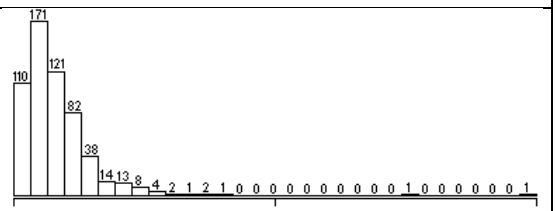
Düzgünlük Standart Hata Özelliğinin Dağılımı



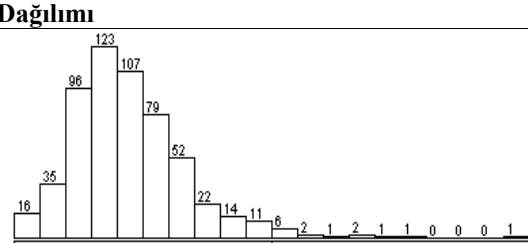
Kompaktlık Standart Hata Özelliğinin Dağılımı



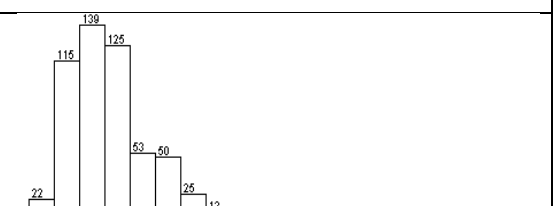
Konkavlık Standart Hata Özelliğinin Dağılımı



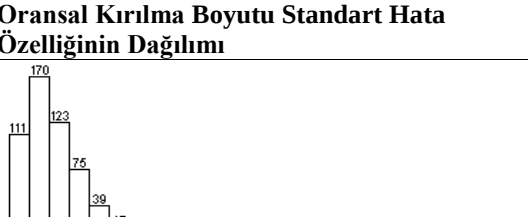
Konkav Noktalar Standart Hata Özelliğinin Dağılımı



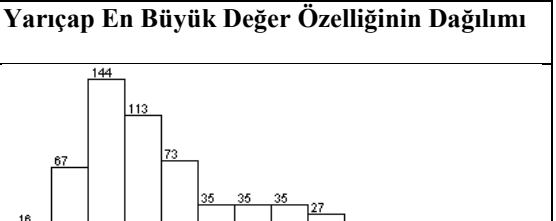
Simetri Standart Hata Özelliğinin Dağılımı



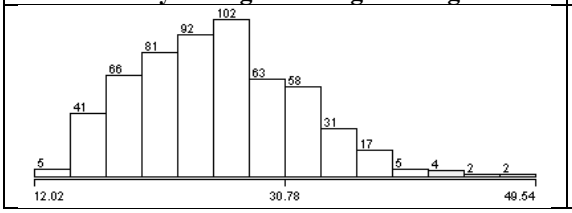
Oransal Kırılma Boyutu Standart Hata Özelliğinin Dağılımı



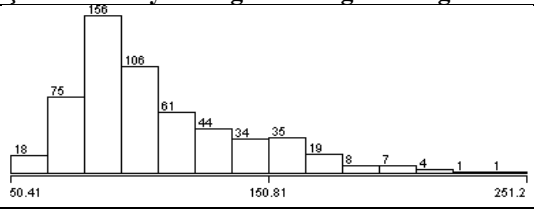
Yarıçap En Büyük Değer Özelliğinin Dağılımı



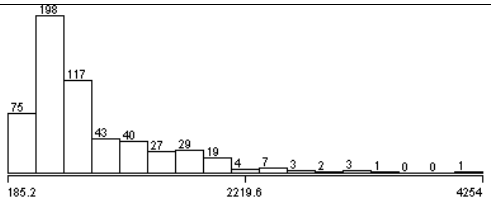
Doku En Büyük Değer Özelliğinin Dağılımı



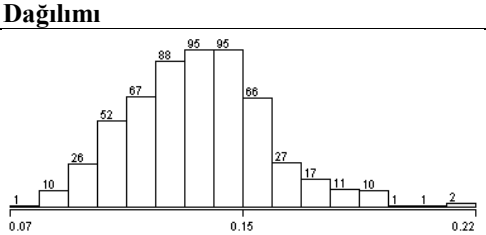
Çevre En Büyük Değer Özelliğinin Dağılımı



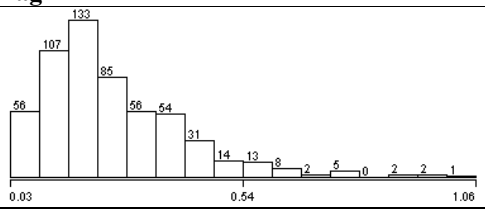
Alan En Büyük Değer Özelliğinin Dağılımı



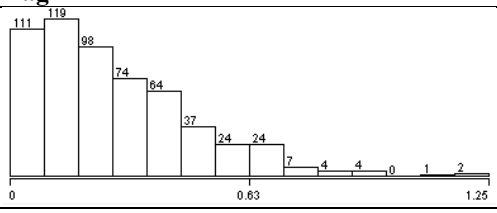
Düzgünlük En Büyük Değer Özelliğinin Dağılımı



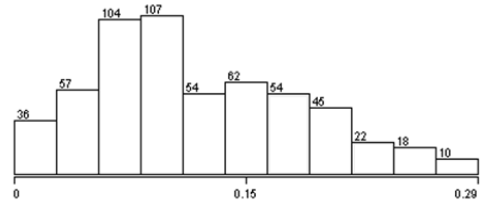
Kompaktlık En Büyük Değer Özelliğinin Dağılımı



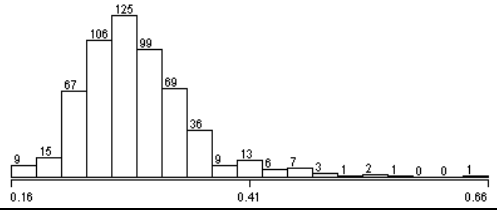
Konkavlık En Büyük Değer Özelliğinin Dağılımı



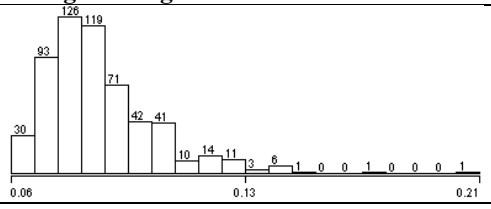
Konkav Noktalar En Büyük Değer Özelliğinin Dağılımı



Simetri En Büyük Değer Özelliğinin Dağılımı



Oransal Kırılma Boyutu En Büyük Değer Özelliğinin Dağılımı



Ek-2: FİLTRELEME ÖZELLİK SEÇİM YÖNTEMİ SONUÇLARI

Özellik Değerlendiriciler:
InfoGainAttributeEval – ChiSquaredAttributeEval – CorrelationAttributeEval – FilteredAttributeEval – GainRatioAttributeEval – OneRAttributeEval – SignificanceAttributeEval – SymmetricalUncertAttributeEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	15	0.87	93.85	0.96	0.95	0.95	0.98	3.48
NaiveBayes	15	0.85	92.96	0.94	0.95	0.94	0.98	3.24
MultilayerPerceptron	15	0.90	95.19	0.96	0.97	0.96	0.99	3.37
SDG	15	0.89	94.71	0.95	0.97	0.96	0.94	2.93
SimpleLogistic	**15**	**0.90**	**95.40**	**0.96**	**0.97**	**0.96**	**0.99**	**2.92**
SMO	15	0.86	93.73	0.94	0.96	0.95	0.93	3.10
IB1	15	0.85	92.80	0.94	0.94	0.94	0.92	3.24
KStar	15	0.84	92.73	0.94	0.95	0.94	0.98	3.23
PART	15	0.85	92.85	0.94	0.95	0.94	0.94	3.09
LMT	15	0.90	95.40	0.96	0.97	0.96	0.99	2.97
RandomForest	15	0.87	94.18	0.95	0.96	0.95	0.98	3.13

Özellik Değerlendiriciler: ReliefFAttributeEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	15	0.89	94.80	0.96	0.95	0.96	0.99	3.12
NaiveBayes	15	0.89	94.66	0.96	0.96	0.96	0.99	3.16
MultilayerPerceptron	15	0.92	96.24	0.97	0.97	0.97	0.99	2.31
SDG	15	0.94	97.22	0.97	0.99	0.98	0.97	2.16
SimpleLogistic	**15**	**0.93**	**96.82**	**0.97**	**0.98**	**0.97**	**0.99**	**2.24**
SMO	15	0.93	97.02	0.96	0.99	0.98	0.96	2.12
IB1	15	0.92	96.21	0.97	0.97	0.97	0.96	2.13
KStar	15	0.91	96.05	0.96	0.98	0.97	0.99	2.16
PART	15	0.89	94.68	0.96	0.96	0.96	0.95	2.89
LMT	15	0.93	96.71	0.97	0.98	0.97	0.99	2.28
RandomForest	15	0.91	95.91	0.96	0.98	0.97	0.99	2.16

Özellik Değerlendiriciler:
InfoGainAttributeEval – ChiSquaredAttributeEval – FilteredAttributeEval – OneRAttributeEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.87	94.03	0.95	0.95	0.95	0.98	3.32
NaiveBayes	10	0.85	92.36	0.93	0.96	0.94	0.98	3.35
MultilayerPerceptron	**10**	**0.90**	**95.33**	**0.96**	**0.97**	**0.96**	**0.99**	**2.91**
SDG	10	0.89	94.77	0.95	0.97	0.96	0.94	3.09
SimpleLogistic	10	0.90	95.33	0.96	0.97	0.96	0.99	2.98
SMO	10	0.87	94.04	0.94	0.97	0.95	093	3.04
IB1	10	0.87	93.94	0.95	0.96	0.95	0.93	3.07
KStar	10	0.88	94.31	0.94	0.97	0.96	0.98	2.87
PART	10	0.86	93.39	0.94	0.96	0.95	0.95	3.03
LMT	10	0.90	95.28	0.96	0.97	0.96	0.99	3.01
RandomForest	10	0.88	94.22	0.95	0.96	0.95	0.98	3.15

Özellik Değerlendiriciler:
CorrelationAttributeEval – GainRatioAttributeEval – SymmetricalUncertAttributeEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.87	93.68	0.96	0.94	0.95	0.98	3.30
NaiveBayes	10	0.88	94.18	0.95	0.96	0.95	0.98	3.02
MultilayerPerceptron	10	0.90	95.57	0.96	0.97	0.97	0.99	2.84
SDG	10	0.89	94.71	0.95	0.97	0.96	0.94	3.04
SimpleLogistic	**10**	**0.92**	**96.22**	**0.97**	**0.98**	**0.97**	**0.99**	**2.58**
SMO	10	0.86	93.75	0.94	0.96	0.95	0.93	3.09
IB1	10	0.87	93.97	0.95	0.96	0.95	0.93	2.79
KStar	10	0.88	94.43	0.95	0.97	0.96	0.98	2.74
PART	10	0.86	93.34	0.93	0.96	0.95	0.95	3.25
LMT	10	0.92	96.19	0.97	0.98	0.97	0.99	2.64
RandomForest	10	0.88	94.52	0.95	0.97	0.96	0.98	3.10

Özellik Değerlendirici: PrincipalComponents
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.82	91.83	0.93	0.94	0.94	0.98	3.50
NaiveBayes	10	0.81	91.25	0.92	0.94	0.93	0.97	3.72
MultilayerPerceptron	**10**	**0.88**	**94.36**	**0.95**	**0.96**	**0.96**	**0.99**	**3.16**
SDG	10	0.86	93.80	0.93	0.97	0.95	0.93	2.87
SimpleLogistic	10	0.86	93.46	0.94	0.96	0.95	0.99	3.10
SMO	10	0.86	93.53	0.92	0.98	0.95	0.92	3.13
IB1	10	0.83	92.13	0.95	0.93	0.94	0.92	2.91
KStar	10	0.84	92.58	0.94	0.95	0.94	0.97	2.87
PART	10	0.84	92.41	0.94	0.94	0.94	0.94	3.21
LMT	10	0.86	93.44	0.94	0.96	0.95	0.98	3.14
RandomForest	10	0.86	93.66	0.94	0.96	0.95	0.98	3.18

Özellik Değerlendirici: ReliefFAttributeEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.88	94.40	0.96	0.96	0.96	0.98	3.25
NaiveBayes	10	0.88	94.34	0.95	0.96	0.96	0.98	2.97
MultilayerPerceptron	10	0.93	96.70	0.97	0.98	0.97	0.99	2.51
SDG	10	0.94	97.38	0.97	0.99	0.98	0.97	2.12
SimpleLogistic	**10**	**0.94**	**97.28**	**0.97**	**0.98**	**0.98**	**0.99**	**2.09**
SMO	10	0.93	96.82	0.96	0.99	0.98	0.96	2.23
IB1	10	0.92	96.49	0.97	0.98	0.97	0.96	1.98
KStar	10	0.92	96.54	0.96	0.99	0.97	0.98	2.18
PART	10	0.90	95.22	0.96	0.97	0.96	0.96	2.93
LMT	**10**	**0.94**	**97.28**	**0.97**	**0.98**	**0.98**	**0.99**	**2.09**
RandomForest	10	0.91	95.87	0.96	0.98	0.97	0.98	2.56

Özellik Değerlendirici: SignificanceAttributeEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.87	93.83	0.95	0.95	0.95	0.98	3.55
NaiveBayes	10	0.87	93.87	0.95	0.96	0.95	0.98	3.10
MultilayerPerceptron	**10**	**0.90**	**95.52**	**0.96**	**0.97**	**0.96**	**0.99**	**2.98**
SDG	10	0.89	94.99	0.95	0.97	0.96	0.94	2.88
SimpleLogistic	10	0.90	95.26	0.96	0.97	0.96	0.99	2.95
SMO	10	0.87	93.83	0.94	0.97	0.95	0.93	3.01
IB1	10	0.86	93.39	0.94	0.95	0.95	0.93	2.92
KStar	10	0.88	94.31	0.94	0.98	0.96	0.98	3.10
PART	10	0.85	93.15	0.94	0.96	0.95	0.95	3.37
LMT	10	0.90	95.20	0.96	0.97	0.96	0.99	2.93
RandomForest	10	0.88	94.50	0.95	0.96	0.96	0.98	2.85

Özellik Değerlendirici: SVMAttributeEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.89	95.05	0.96	0.96	0.96	0.99	3.00
NaiveBayes	10	0.91	95.61	0.96	0.97	0.97	0.99	2.70
MultilayerPerceptron	10	0.94	97.10	0.97	0.98	0.98	0.99	2.23
SDG	10	0.94	97.21	0.97	0.98	0.98	0.97	2.13
SimpleLogistic	**10**	**0.95**	**97.59**	**0.98**	**0.99**	**0.98**	**0.99**	**2.01**
SMO	10	0.94	97.07	0.96	0.99	0.98	0.96	2.09
IB1	10	0.90	95.15	0.96	0.97	0.96	0.95	2.69
KStar	10	0.90	95.61	0.95	0.99	0.97	0.98	2.62
PART	10	0.89	94.76	0.96	0.96	0.96	0.95	2.79
LMT	**10**	**0.95**	**97.59**	**0.98**	**0.99**	**0.98**	**0.99**	**2.01**
RandomForest	10	0.91	95.75	0.96	0.98	0.97	0.99	2.59

Özellik Değerlendirici: CfsSubsetEval
Araştırma Yöntemleri: BestFirst – GreedyStepwise – LinearForwardSelection – PSO – Scatter – SubsetSizeForwardSelection – Tabu

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	11	0.90	95.36	0.97	0.96	0.96	0.99	3.01
NaiveBayes	11	0.88	94.55	0.95	0.97	0.96	0.99	3.03
MultilayerPerceptron	11	0.93	96.72	0.97	0.98	0.97	0.99	2.17
SDG	11	0.94	97.10	0.96	0.99	0.98	0.96	2.06
SimpleLogistic	**11**	**0.94**	**97.31**	**0.97**	**0.99**	**0.98**	**0.99**	**2.21**
SMO	11	0.92	96.51	0.95	1.00	0.97	0.95	2.26
IB1	11	0.91	95.71	0.96	0.97	0.97	0.95	2.44
KStar	11	0.91	95.77	0.96	0.98	0.97	0.99	2.48
PART	11	0.88	94.24	0.96	0.95	0.95	0.95	3.11
LMT	11	0.94	97.28	0.97	0.99	0.98	0.99	2.22
RandomForest	11	0.91	95.70	0.96	0.98	0.97	0.99	2.70

Özellik Değerlendirici: CfsSubsetEval
Araştırma Yöntemi: Evolutionary

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	13	0.90	95.24	0.97	0.96	0.96	0.99	3.10
NaiveBayes	13	0.87	93.97	0.95	0.96	0.95	0.99	3.30
MultilayerPerceptron	13	0.93	96.73	0.97	0.98	0.97	0.99	2.11
SDG	13	0.94	97.35	0.97	0.99	0.98	0.97	2.06
SimpleLogistic	**13**	**0.94**	**97.35**	**0.97**	**0.99**	**0.98**	**0.99**	**2.08**
SMO	13	0.93	96.68	0.95	1.00	0.97	0.96	2.15
IB1	13	0.93	96.87	0.97	0.98	0.98	0.96	2.07
KStar	13	0.91	95.98	0.96	0.98	0.97	0.99	2.41
PART	13	0.87	94.01	0.95	0.95	0.95	0.95	3.11
LMT	13	0.94	97.33	0.97	0.99	0.98	0.99	2.08
RandomForest	13	0.91	95.80	0.96	0.98	0.97	0.99	2.73

Özellik Değerlendirici: CfsSubsetEval
Araştırma Yöntemi: Genetic

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	17	0.89	94.96	0.96	0.96	0.96	0.99	3.04
NaiveBayes	17	0.88	94.54	0.95	0.97	0.96	0.99	3.07
MultilayerPerceptron	17	0.92	96.49	0.97	0.98	0.97	0.99	2.65
SDG	17	0.94	97.33	0.97	0.99	0.98	0.97	2.14
SimpleLogistic	**17**	**0.93**	**96.98**	**0.97**	**0.98**	**0.98**	**0.99**	**2.19**
SMO	17	0.94	97.19	0.96	1.00	0.98	0.96	2.17
IB1	17	0.92	96.15	0.97	0.97	0.97	0.96	2.40
KStar	17	0.90	95.50	0.95	0.98	0.96	0.99	2.34
PART	17	0.87	94.13	0.96	0.95	0.95	0.95	2.71
LMT	17	0.93	96.94	0.97	0.98	0.98	0.99	2.14
RandomForest	17	0.91	96.01	0.96	0.98	0.97	0.99	2.36

Özellik Değerlendirici: CfsSubsetEval
Araştırma Yöntemi: Rank

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.90	95.45	0.97	0.96	0.96	0.99	3.01
NaiveBayes	10	0.90	95.12	0.96	0.97	0.96	0.98	2.87
MultilayerPerceptron	10	0.93	96.82	0.97	0.98	0.97	0.99	2.05
SDG	10	0.94	97.03	0.96	0.99	0.98	0.96	1.98
SimpleLogistic	**10**	**0.95**	**97.59**	**0.97**	**0.99**	**0.98**	**0.99**	**2.05**
SMO	10	0.93	96.61	0.95	1.00	0.97	0.96	2.21
IB1	10	0.91	95.78	0.96	0.97	0.97	0.95	2.38
KStar	10	0.91	95.85	0.96	0.97	0.97	0.98	2.29
PART	10	0.88	94.62	0.96	0.96	0.96	0.95	3.00
LMT	10	0.95	97.52	0.97	0.99	0.98	0.99	2.09
RandomForest	10	0.91	95.70	0.96	0.98	0.97	0.98	2.57

Özellik Değerlendirici: ConsistencySubsetEval
Araştırma Yöntemi: BestFirst

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	8	0.90	95.34	0.96	0.96	0.96	0.99	2.89
NaiveBayes	8	0.88	94.27	0.94	0.97	0.95	0.99	3.00
MultilayerPerceptron	8	0.92	96.08	0.97	0.97	0.97	0.99	2.27
SDG	8	0.94	97.07	0.97	0.99	0.98	0.97	2.11
SimpleLogistic	8	0.92	96.22	0.96	0.98	0.97	0.99	2.37
SMO	8	0.92	96.38	0.96	0.99	0.97	0.96	2.15
IB1	8	0.88	94.38	0.95	0.96	0.96	0.94	2.84
KStar	8	0.87	94.25	0.94	0.98	0.96	0.99	2.86
PART	8	0.89	94.66	0.96	0.96	0.96	0.96	2.81
LMT	**8**	**0.92**	**96.31**	**0.96**	**0.98**	**0.97**	**0.99**	**2.48**
RandomForest	8	0.91	95.85	0.96	0.98	0.97	0.99	2.47

Özellik Değerlendirici: ConsistencySubsetEval
Araştırma Yöntemi: Evolutionary

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	9	0.83	91.99	0.93	0.94	0.94	0.97	3.46
NaiveBayes	9	0.85	92.94	0.93	0.96	0.94	0.98	3.47
MultilayerPerceptron	9	0.90	95.45	0.96	0.97	0.96	0.99	2.65
SDG	9	0.90	95.45	0.95	0.98	0.96	0.95	2.56
SimpleLogistic	**9**	**0.93**	**96.93**	**0.97**	**0.98**	**0.98**	**0.99**	**2.33**
SMO	9	0.89	94.92	0.94	0.98	0.96	0.94	2.67
IB1	9	0.87	93.92	0.95	0.96	0.95	0.93	2.97
KStar	9	0.87	93.67	0.94	0.96	0.95	0.99	3.30
PART	9	0.86	93.71	0.95	0.95	0.95	0.94	3.39
LMT	**9**	**0.93**	**96.33**	**0.97**	**0.98**	**0.98**	**0.99**	**2.33**
RandomForest	9	0.88	94.38	0.94	0.97	0.96	0.98	3.00

Özellik Değerlendirici: ConsistencySubsetEval
Araştırma Yöntemi: Genetic

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	12	0.88	94.24	0.95	0.96	0.95	0.98	3.12
NaiveBayes	12	0.84	92.44	0.92	0.96	0.94	0.98	3.25
MultilayerPerceptron	12	0.92	96.30	0.97	0.98	0.97	0.99	2.54
SDG	12	0.94	97.03	0.97	0.98	0.98	0.97	2.23
SimpleLogistic	**12**	**0.93**	**96.89**	**0.97**	**0.98**	**0.98**	**1.00**	**2.27**
SMO	12	0.92	96.45	0.95	0.99	0.97	0.96	2.26
IB1	12	0.88	94.59	0.95	0.96	0.96	0.94	2.76
KStar	12	0.89	94.87	0.94	0.99	0.96	0.99	2.64
PART	12	0.88	94.64	0.95	0.96	0.96	0.95	3.15
LMT	12	0.93	96.82	0.97	0.98	0.97	0.99	2.28
RandomForest	12	0.90	95.36	0.95	0.98	0.96	0.99	2.93

Özellik Değerlendirici: ConsistencySubseteval
Araştırma Yöntemi: GreedyStepwise

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	7	0.87	94.01	0.96	0.95	0.95	0.98	3.09
NaiveBayes	7	0.88	94.27	0.94	0.97	0.96	0.98	2.84
MultilayerPerceptron	7	0.93	96.59	0.97	0.98	0.97	0.99	2.19
SDG	7	0.94	97.08	0.97	0.98	0.98	0.97	2.13
SimpleLogistic	**7**	**0.94**	**97.00**	**0.97**	**0.98**	**0.98**	**0.99**	**2.33**
SMO	7	0.91	95.71	0.95	0.98	0.97	0.95	2.31
IB1	7	0.89	95.06	0.95	0.97	0.96	0.94	2.50
KStar	7	0.86	93.60	0.93	0.98	0.95	0.99	3.11
PART	7	0.89	94.80	0.96	0.96	0.96	0.95	2.99
LMT	7	0.93	96.98	0.97	0.98	0.98	0.99	2.34
RandomForest	7	0.90	95.55	0.96	0.97	0.96	0.99	2.76

Özellik Değerlendirici: ConsistencySubsetEval
Araştırma Yöntemleri: LinearForwardSelection – SubsetSizeForwardSelection

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	7	0.89	95.12	0.95	0.97	0.96	0.99	2.80
NaiveBayes	7	0.89	94.89	0.95	0.97	0.96	0.99	2.73
MultilayerPerceptron	7	0.92	96.40	0.97	0.98	0.97	0.99	2.41
SDG	7	0.94	97.23	0.97	0.98	0.98	0.97	2.18
SimpleLogistic	7	0.92	96.45	0.97	0.98	0.97	0.99	2.17
SMO	7	0.93	96.65	0.96	0.99	0.97	0.96	2.23
IB1	7	0.92	96.45	0.97	0.98	0.97	0.96	2.10
KStar	7	0.90	95.34	0.95	0.98	0.96	0.99	2.65
PART	7	0.89	95.03	0.96	0.97	0.96	0.95	3.00
LMT	**7**	**0.92**	**96.49**	**0.97**	**0.98**	**0.97**	**1.00**	**2.19**
RandomForest	7	0.91	96.01	0.96	0.98	0.97	0.99	2.39

Özellik Değerlendirici: ConsistencySubsetEval
Araştırma Yöntemi: PSO

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	9	0.90	95.42	0.95	0.97	0.96	0.99	2.72
NaiveBayes	9	0.85	93.20	0.93	0.96	0.95	0.99	3.17
MultilayerPerceptron	9	0.92	96.45	0.97	0.98	0.97	0.99	2.31
SDG	9	0.93	96.82	0.97	0.98	0.98	0.96	2.17
SimpleLogistic	**9**	**0.94**	**97.00**	**0.97**	**0.98**	**0.98**	**0.99**	**2.25**
SMO	9	0.90	95.59	0.95	0.99	0.97	0.94	2.55
IB1	9	0.89	94.83	0.95	0.97	0.96	0.94	2.63
KStar	9	0.87	94.20	0.94	0.97	0.96	0.99	3.15
PART	9	0.89	94.80	0.95	0.97	0.96	0.95	2.80
LMT	9	0.93	96.93	0.97	0.98	0.98	0.99	2.19
RandomForest	9	0.91	96.01	0.96	0.98	0.97	0.99	2.61

Özellik Değerlendirici: ConsistencySubsetEval
Araştırma Yöntemi: Rank

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	21	0.88	94.43	0.96	0.96	0.96	0.99	3.26
NaiveBayes	21	0.87	93.85	0.94	0.96	0.95	0.99	3.40
MultilayerPerceptron	21	0.93	96.98	0.97	0.98	0.97	0.99	2.43
SDG	21	0.94	97.19	0.97	0.99	0.98	0.97	2.20
SimpleLogistic	**21**	**0.93**	**96.89**	**0.97**	**0.98**	**0.98**	**0.99**	**2.30**
SMO	21	0.94	97.17	0.96	0.99	0.98	0.96	2.07
IB1	21	0.90	95.15	0.96	0.96	0.96	0.95	2.48
KStar	21	0.89	94.97	0.95	0.97	0.96	0.99	2.54
PART	21	0.88	94.22	0.96	0.95	0.95	0.95	2.76
LMT	**21**	**0.93**	**96.86**	**0.97**	**0.98**	**0.98**	**0.99**	**2.30**
RandomForest	21	0.92	96.26	0.96	0.98	0.97	0.99	2.55

Özellik Değerlendirici: ConsistencySubsetEval
Araştırma Yöntemi: Scatter

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	8	0.84	92.69	0.94	0.95	0.94	0.98	3.51
NaiveBayes	8	0.82	91.90	0.93	0.95	0.94	0.98	3.28
MultilayerPerceptron	8	0.91	95.85	0.96	0.98	0.97	0.99	2.45
SDG	8	0.91	96.08	0.95	0.99	0.97	0.95	2.46
SimpleLogistic	**8**	**0.91**	**95.99**	**0.96**	**0.98**	**0.97**	**0.99**	**2.57**
SMO	8	0.89	95.12	0.94	0.99	0.96	0.94	2.76
IB1	8	0.89	94.85	0.95	0.97	0.96	0.94	2.64
KStar	8	0.84	92.71	0.93	0.96	0.94	0.98	3.43
PART	8	0.97	93.99	0.95	0.96	0.95	0.95	2.87
LMT	8	0.91	95.91	0.96	0.98	0.97	0.99	2.65
RandomForest	8	0.89	94.82	0.95	0.97	0.96	0.98	2.76

Özellik Değerlendirici: FilteredSubsetEval
Araştırma Yöntemleri: BestFirst – GreedyStepwise – LinearForwardSelection – Scatter – SubsetSizeForwardSelection

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	9	0.89	95.07	096	0.96	0.96	0.99	3.19
NaiveBayes	9	0.88	94.31	0.95	0.96	0.96	0.99	3.23
MultilayerPerceptron	**9**	**0.92**	**96.12**	**0.97**	**0.97**	**0.97**	**0.99**	**2.33**
SDG	9	0.92	96.22	0.96	0.98	0.97	0.96	2.42
SimpleLogistic	9	0.91	96.03	0.96	0.97	0.97	0.99	2.59
SMO	9	0.93	96.65	0.96	0.99	0.97	0.96	2.16
IB1	9	0.90	95.26	0.96	0.97	0.96	0.95	2.83
KStar	9	0.90	95.13	0.96	0.97	0.96	0.99	2.49
PART	9	0.88	94.43	0.95	0.96	0.96	0.95	2.89
LMT	9	0.91	95.98	0.96	0.97	0.97	0.99	2.64
RandomForest	9	0.90	95.56	0.96	0.98	0.97	0.99	2.51

Özellik Değerlendirici: FiteredSubsetEval
Araştırma Yöntemi: Evolutionary

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	12	0.90	95.24	0.97	0.96	0.96	0.99	3.10
NaiveBayes	12	0.87	94.04	0.95	0.96	0.95	0.99	3.26
MultilayerPerceptron	12	0.93	96.84	0.97	0.98	0.97	0.99	2.07
SDG	12	0.94	97.44	0.97	0.99	0.98	0.97	2.03
SimpleLogistic	12	0.94	97.44	0.97	0.99	0.98	0.99	2.03
SMO	12	0.93	96.75	0.95	1.00	0.97	0.96	2.09
IB1	12	0.92	96.50	0.97	0.97	0.97	0.96	2.33
KStar	12	0.92	96.47	0.97	0.98	0.97	0.99	2.24
PART	12	0.88	94.31	0.95	0.96	0.95	0.95	3.14
LMT	**12**	**0.94**	**97.45**	**0.97**	**0.99**	**0.98**	**0.99**	**2.02**
RandomForest	12	0.91	95.75	0.96	0.98	0.97	0.99	2.51

Özellik Değerlendirici: FilteredSubsetEval
Araştırma Yöntemi: Genetic

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	16	0.89	94.96	0.96	0.96	0.96	0.99	3.04
NaiveBayes	16	0.88	94.55	0.95	0.97	0.96	0.99	3.05
MultilayerPerceptron	16	0.92	96.31	0.97	0.98	0.97	0.99	2.62
SDG	16	0.94	97.14	0.97	0.99	0.98	0.97	2.29
SimpleLogistic	**16**	**0.93**	**96.91**	**0.97**	**0.98**	**0.98**	**0.99**	**2.25**
SMO	16	0.94	97.07	0.96	1.00	0.98	0.96	2.19
IB1	16	0.91	95.79	0.97	0.97	0.97	0.95	2.43
KStar	16	0.92	96.12	0.96	0.98	0.97	0.99	2.26
PART	16	0.88	94.31	0.97	0.95	0.95	0.95	2.72
LMT	16	0.93	96.87	0.97	0.98	0.98	0.99	2.25
RandomForest	16	0.92	96.21	0.96	0.98	0.97	0.99	2.66

Özellik Değerlendirici: FilteredSubsetEval
Araştırma Yöntemi: PSO

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.90	95.96	0.97	0.96	0.96	0.99	3.01
NaiveBayes	10	0.89	94.68	0.95	0.97	0.96	0.99	2.93
MultilayerPerceptron	10	0.92	96.45	0.97	0.97	0.97	0.99	2.47
SDG	10	0.94	97.12	0.96	0.99	0.98	0.96	1.99
SimpleLogistic	**10**	**0.94**	**97.40**	**0.97**	**0.99**	**0.98**	**0.99**	**2.08**
SMO	10	0.93	96.68	0.95	1.00	0.97	0.96	2.21
IB1	10	0.91	95.94	0.97	0.96	0.97	0.96	2.54
KStar	10	0.92	96.47	0.97	0.98	0.97	0.99	2.24
PART	10	0.88	94.50	0.96	0.96	0.96	0.95	3.18
LMT	10	0.94	97.36	0.97	0.99	0.98	0.99	2.08
RandomForest	10	0.91	95.66	0.96	0.98	0.97	0.99	2.70

Özellik Değerlendirici: FilteredSubsetEval
Araştırma Yöntemi: Rank

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	7	0.88	94.20	0.97	0.94	0.95	0.99	3.45
NaiveBayes	7	0.88	94.38	0.95	0.96	0.96	0.98	3.08
MultilayerPerceptron	**7**	**0.90**	**95.47**	**0.96**	**0.97**	**0.96**	**0.99**	**2.71**
SDG	7	0.88	94.59	0.95	0.97	0.96	0.94	2.77
SimpleLogistic	7	0.89	94.87	0.96	0.96	0.96	0.99	2.91
SMO	7	0.88	94.38	0.94	0.97	0.96	0.93	3.04
IB1	7	0.86	93.61	0.95	0.95	0.95	0.93	2.87
KStar	7	0.89	95.01	0.95	0.97	0.96	0.98	2.79
PART	7	0.85	93.04	0.94	0.95	0.94	0.96	3.02
LMT	7	0.89	94.78	0.96	0.96	0.96	0.99	2.97
RandomForest	7	0.87	94.18	0.95	0.96	0.95	0.98	3.25

Ek-3: SARMALAMA ÖZELLİK SEÇİM YÖNTEMİ SONUÇLARI

Özellik Değerlendirici: ClassifierSubsetEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	15	0.87	93.85	0.96	0.95	0.95	0.98	3.48
NaiveBayes	15	0.85	92.96	0.94	0.95	0.94	0.98	3.24
MultilayerPerceptron	15	0.90	95.19	0.96	0.97	0.96	0.99	3.37
SDG	15	0.89	94.71	0.95	0.97	0.96	0.94	2.93
SimpleLogistic	**15**	**0.90**	**95.40**	**0.96**	**0.97**	**0.96**	**0.99**	**2.92**
SMO	15	0.86	93.73	0.94	0.96	0.95	0.93	3.10
IB1	15	0.85	92.80	0.94	0.94	0.94	0.92	3.24
KStar	15	0.84	92.73	0.94	0.95	0.94	0.98	3.23
PART	15	0.85	92.85	0.94	0.95	0.94	0.94	3.09
LMT	15	0.90	95.40	0.96	0.97	0.96	0.99	2.97
RandomForest	15	0.87	94.18	0.95	0.96	0.95	0.98	3.13

Özellik Değerlendirici: ClassifierSubsetEval
Araştırma Yöntemi: Ranker

Sınıflandırıcı	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BayesNet	10	0.87	94.03	0.95	0.95	0.95	0.98	3.32
NaiveBayes	10	0.85	92.96	0.93	0.96	0.94	0.98	3.35
MultilayerPerceptron	**10**	**0.90**	**95.33**	**0.96**	**0.97**	**0.96**	**0.99**	**2.91**
SDG	10	0.89	94.71	0.95	0.97	0.96	0.94	3.04
SimpleLogistic	10	0.90	95.33	0.96	0.97	0.96	0.99	2.98
SMO	10	0.86	93.75	0.94	0.96	0.95	0.93	3.09
IB1	10	0.87	93.97	0.95	0.96	0.95	0.93	2.79
KStar	10	0.88	94.31	0.94	0.97	0.96	0.98	2.87
PART	10	0.86	93.39	0.94	0.96	0.95	0.95	3.03
LMT	10	0.90	95.28	0.96	0.97	0.96	0.99	3.01
RandomForest	10	0.88	94.22	0.95	0.97	0.95	0.98	3.15

Özellik Değerlendirici: ClassifierSubsetEval
Sınıflandırıcı: MultilayerPerceptron

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	11	0.91	95.87	0.96	0.97	0.97	0.99	2.77
Evolutionary	15	0.92	96.15	0.96	0.98	0.97	0.99	2.40
Genetic	21	0.93	96.54	0.97	0.98	0.97	0.99	2.49
Greedy Stepwise	6	0.93	96.68	0.97	0.98	0.97	0.99	2.20
Linear Forward Selection	**9**	**0.95**	**97.47**	**0.98**	**0.98**	**0.98**	**0.99**	**2.19**
PSO	19	0.93	96.75	0.97	0.98	0.97	0.99	2.69
Race	5	0.92	96.15	0.96	0.98	0.97	0.99	2.31
Rank	17	0.92	96.26	0.97	0.98	0.97	0.99	2.56
Scatter	8	0.92	96.36	0.97	0.97	0.97	0.99	2.59
Subset Size Forward Selection	7	0.93	96.21	0.97	0.98	0.98	0.99	2.18

Özellik Değerlendirici: ClassifierSubsetEval
Sınıflandırıcı: SGD

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	9	0.95	97.82	0.98	0.99	0.98	0.97	1.98
Evolutionary	11	0.95	97.79	0.97	0.99	0.98	0.97	1.94
Genetic	**13**	**0.96**	**97.98**	**0.98**	**0.99**	**0.98**	**0.98**	**1.79**
Greedy Stepwise	6	0.94	97.22	0.96	0.99	0.98	0.96	2.18
Linear Forward Selection	13	0.95	97.72	0.97	0.99	0.98	0.97	2.11
PSO	15	0.95	97.91	0.98	0.99	0.98	0.97	1.93
Race	3	0.89	94.85	0.94	0.98	0.96	0.94	2.85
Rank	29	0.95	97.54	0.97	0.99	0.98	0.97	1.91
Scatter	5	0.94	97.40	0.97	0.99	0.98	0.97	2.19
Subset Size Forward Selection	6	0.94	97.30	0.97	0.99	0.98	0.97	2.22

Özellik Değerlendirici: ClassifierSubsetEval
Sınıflandırıcı: SMO

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	9	0.94	97.28	0.97	0.99	0.98	0.97	2.03
Evolutionary	11	0.94	97.40	0.97	0.99	0.98	0.97	1.97
Genetic	15	0.95	97.70	0.97	1.00	0.98	0.97	1.90
Greedy Stepwise	7	0.94	97.40	0.97	0.99	0.98	0.97	2.11
Linear Forward Selection	7	0.94	97.10	0.96	1.00	0.98	0.96	2.17
PSO	**14**	**0.95**	**97.77**	**0.97**	**0.99**	**0.98**	**0.97**	**1.93**
Race	3	0.88	94.66	0.93	0.99	0.96	0.93	2.88
Rank	22	0.95	97.65	0.97	1.00	0.98	0.97	1.87
Scatter	8	0.94	97.12	0.96	0.99	0.98	0.96	2.14
Subset Size Forward Selection	6	0.93	96.87	0.96	0.99	0.98	0.96	2.17

Özellik Değerlendirici: ClassifierSubsetEval
Sınıflandırıcı: IB1

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	2	0.77	89.19	0.91	0.92	0.91	0.88	3.65
Evolutionary	2	0.65	83.53	0.87	0.88	0.87	0.82	4.55
Genetic	5	0.86	93.50	0.95	0.95	0.95	0.93	2.99
Greedy Stepwise	2	0.77	89.19	0.91	0.92	0.91	0.88	3.65
Linear Forward Selection	2	0.82	91.60	0.93	0.94	0.93	0.91	3.82
PSO	2	0.61	81.46	0.86	0.85	0.85	0.80	4.43
Race	**9**	**0.88**	**94.55**	**0.95**	**0.96**	**0.96**	**0.94**	**3.17**
Rank	2	0.74	88.03	0.90	0.91	0.91	0.87	4.06
Scatter	2	0.74	87.68	0.91	0.90	0.90	0.87	4.49
Subset Size Forward Selection	2	0.82	91.60	0.93	0.94	0.93	0.91	3.82

Özellik Değerlendirici: ClassifierSubsetEval
Sınıflandırıcı: KStar

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	3	0.88	94.12	0.94	0.98	0.96	0.98	2.87
Evolutionary	4	0.74	88.00	0.88	0.94	0.91	0.96	3.98
Genetic	5	0.84	92.76	0.93	0.96	0.94	0.98	2.81
Greedy Stepwise	3	0.88	94.61	0.94	0.98	0.96	0.98	2.87
Linear Forward Selection	3	0.88	94.61	0.94	0.98	0.96	0.98	2.87
PSO	3	0.82	91.72	0.91	0.97	0.94	0.96	3.39
Race	**5**	**0.93**	**96.66**	**0.97**	**0.98**	**0.97**	**0.99**	**2.37**
Rank	5	0.86	93.64	0.94	0.96	0.95	0.98	3.05
Scatter	3	0.89	95.10	0.95	0.98	0.96	0.99	2.55
Subset Size Forward Selection	3	0.88	94.61	0.94	0.98	0.96	0.98	2.87

Özellik Değerlendirici: ClassifierSubsetEval
Sınıflandırıcı: PART

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	**8**	**0.90**	**95.10**	**0.97**	**0.96**	**0.96**	**0.96**	**2.78**
Evolutionary	11	0.89	94.75	0.96	0.96	0.96	0.95	2.73
Genetic	15	0.86	93.61	0.95	0.95	0.95	0.95	3.23
Greedy Stepwise	**8**	**0.90**	**95.10**	**0.97**	**0.96**	**0.96**	**0.96**	**2.78**
Linear Forward Selection	5	0.88	94.54	0.96	0.96	0.96	0.95	3.06
PSO	11	0.88	94.43	0.95	0.96	0.96	0.95	3.20
Race	4	0.89	94.97	0.96	0.97	0.96	0.95	3.00
Rank	22	0.87	94.13	0.96	0.95	0.95	0.95	2.83
Scatter	5	0.88	94.54	0.96	0.96	0.96	0.95	3.06
Subset Size Forward Selection	5	0.88	94.54	0.96	0.96	0.96	0.95	3.06

Özellik Değerlendirici: ClassifierSubsetEval
Sınıflandırıcı: LMT

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	11	0.93	96.78	0.97	0.98	0.97	0.99	2.43
Evolutionary	14	0.94	97.03	0.97	0.99	0.98	0.99	2.18
Genetic	13	0.93	96.89	0.97	0.98	0.98	0.99	2.42
Greedy Stepwise	4	0.93	96.66	0.97	0.98	0.97	0.99	2.31
Linear Forward Selection	13	0.93	96.78	0.97	0.99	0.97	0.99	2.10
PSO	13	0.93	96.84	0.97	0.98	0.97	0.99	2.46
Race	7	0.93	96.89	0.97	0.98	0.98	0.99	2.47
Rank	**20**	**0.94**	**97.33**	**0.97**	**0.99**	**0.98**	**0.99**	**2.41**
Scatter	5	0.90	95.59	0.96	0.97	0.97	0.99	2.54
Subset Size Forward Selection	4	0.93	96.66	0.97	0.98	0.97	0.99	2.31

Özellik Değerlendirici: ClassifierSubsetEval
Sınıflandırıcı: Randomforest

Araştırma Yöntemi	Özellik Sayısı	Kappa İstatistiği	Doğruluk (%)	Kesinlik	Duyarlılık	F-Ölçütü	ROC Alanı	Doğruluk Std.S.
BestFirst	3	0.84	92.66	0.93	0.96	0.94	0.96	3.74
Evolutionary	**10**	**0.91**	**95.89**	**0.95**	**0.98**	**0.97**	**0.99**	**2.66**
Genetic	9	0.90	95.27	0.95	0.97	0.96	0.99	2.75
GreedyStepwise	2	0.83	92.04	0.92	0.95	0.94	0.96	3.44
LinearForward Selection	3	0.87	93.83	0.94	0.96	0.95	0.97	3.16
PSO	19	0.91	95.94	0.96	0.98	0.97	0.99	2.34
Race	5	0.91	95.91	0.96	0.98	0.97	0.98	2.20
Rank	29	0.91	95.66	0.96	0.98	0.97	0.99	2.50
Scatter	4	0.86	93.29	0.94	0.95	0.95	0.97	2.76
Subset Size Forward Selection	3	0.87	93.83	0.94	0.96	0.95	0.97	3.16

KAYNAKÇA

Akpınar, Haldun (2000), "Veri Tabanlarında Bilgi Keşfi ve Veri madenciliği", *İ.Ü. İşletme Fakültesi Dergisi*, C:29, S: 1, s. 1-22.

Alan, Mehmet Ali (2012), "Veri Madenciliği ve Lisansüstü Öğrenci Verileri Üzerine Bir Uygulama", *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, S. 33, s. 165-174.

Altıntop, Ümmühan (2006), **"**İnternet Tabanlı Öğretimde Veri Madenciliği Tekniklerinin Uygulanması", *Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi*.

Arguden, Yılmaz, B. Erşahin (2002), Veri Madenciliği: Veriden Bilgiye, Masraftan Değere.

Arshadi, Niloofar, I. Jurisica (2005), **"**Data Mining for Case-Based Reasoning in High-Dimensional Biological Domains," *IEEE Transactions On Knowledge and Data Engineering*, C. 17, S. 8, s. 1127-1137.

Asilkan, Özcan (2011), "İkinci El Otomobillerin Güncel Pazar Fiyatlarının Veri Madenciliği Yöntemleriyle Modellenmesi", *Akademik Bakış Dergisi,* 2011, S. 24, s.1-19.

Baykal, Abdullah (2007), "Web Madenciliği ile Log Analizi", *D. Ü. Ziya Gökalp Eğitim Fakültesi Dergisi*, S. 8, s. 129-145.

Breiman, Leo (2001), "Random Forests", *Machine Learning*, C. 45, s. 5-32.

Brusilovsky, Pavel, Data Mining vs. Statistics.
http://www.bisolutions.us/Data-Mining-vs-Statistics.php (Erişim tarihi: 12.12.2013)

Cabena, Peter, P. Hadjinian, R. Stadler, J. Verhees, M. Kamber (1998), Discovering Data Mining: From Concept to Implementation, New Jersey: Prentice Hall.

Chapman, Peter, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth (2000), CRISP-DM 1.0: Step-by-Step Data Mining Guide.

CRISP
http://blog.visilabs.com/crisp/ (Erişim tarihi: 25.03.2014)

Çiftlikli, Cebrail, M. Nakip, E. Özyirmidokuz (2007), "Bilgi Keşfi Sürecinde Bulanık Gözatım Uygulaması" *Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, S. 23, s. 75-93.

Database Basics
http://office.microsoft.com/en-us/access-help/database-basics-HA010064450.aspx

Data Warehouse
http://datawarehouse4u.info/index_en.html (Erişim tarihi: 17.02.2014)

Data Warehousing and Business Intelligence Introduction
http://infogoal.com/datawarehousing/overview.htm (Erişim tarihi: 30.11.2013)

Data Warehouse vs Data Mart
http://www.datamartist.com/data-warehouse-vs-data-mart (Erişim tarihi: 07.05.2014)

Difference Between Data Mining and OLAP
http://www.differencebetween.com/difference-between-data-mining-and-vs-olap/
(Erişim tarihi: 22.02.2014)

Dunham, Margaret H., (2006), Data Mining Introductory and Advanced Topics, New Jersey: Pearson Education Inc.

Fayyad, Usama M., G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (1996), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge.

Fayyad, Usama (2001), "The Digital Physics of Data Mining", *Communications of the ACM*, C. 44, S. 3, s. 62-65

Fernandez, George (2002), Data Mining Using SAS Applications.

Gentle, James E. , W. Härdle, Y. Mori (2004), Handbook of Computational Statistics: Concepts and Methods.

Goals of a data warehouse
http://www.rpi.edu/datawarehouse/dw-goals.html (Erişim tarihi: 10.01.2014)

Goebel, Michael, L. Gruenwald (1999), **"**A Survey of Data Mining and Knowledge Discovery Software Tools", *SIGKDD Explorations*, C. 1, s. 20-33.

Gray Paul, H.J. Watson (1997), Decision Support in the Data Warehouse.

Gray, Paul, H. J. Watson (1998), "Professional Briefings: Present and Future Directions in Data Warehousing", *Database for Advances in Information Systems*, C. 29, S. 3, s. 83-90.

Guerra, Joseph (2013), Why You Need a Data Warehouse
http://datalyticstechnologies.com/ (Erişim tarihi: 30.12.2013)

Gülen, Özgün Çöllüoğlu, S. Özdemir (2013), "Analysis of Gifted Students' Interest Areas Using Data Mining Techniques", *Üstün Yetenekli Eğitimi Araştırmaları Dergisi*, C. 1, S. 3, s. 213-226.

Han, Jiawei, M. Kamber (2006), Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann Publishers.

Hand, David J., (1998), "Data Mining: Statistics and More?", *The American Statistician*, C.52, S.2, s. 112-118.

Hand, David J., H. Mannila, P. Smyth (2001), Principles of Data Mining**,** The MIT Press.

How Data Integration Works
http://computer.howstuffworks.com/data-integration3.htm (Erişim tarihi: 08.02.2014)

Jackson, Joyce (2002), "Data Mining: A Conceptual Overview", *Communications of the Association for Information Systems,* C. 8, s. 267-296.

Kohavi, Ron, G.H. John (1997), "Wrappers for Feature Subset Selection", *Artificial Intelligence*, C. 97, s.273-324.

Kounen, Diego (2004), Data Mining and Statistics: What is the Connection?
http://www.tdan.com/view-articles/5226/ (Erişim tarihi: 10.12.2013)

Koyuncugil, Ali Serhan, N. Özgülbaş (2008), "İMKB'de İşlem Gören KOBİ'lerin Güçlü ve Zayıf Yönleri: CHAID Karar Ağacı Uygulaması", *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* C. 23, S. 1, s. 1-21.

Kumar, Yugal, G. Sahoo (2012), "Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA", *I.J. Information Technology and Computer Science,* C. 4, S. 7, s. 43-49.

Küçüksille, Engin (2009), "Veri Madenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi ve İMKB Hisse Senetleri Piyasasında Bir Uygulama", *T.C. Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Doktora Tezi.*

Landis, J.Richard, G.G. Koch (1977). "The Measurement Of Observer Agreement For Categorical Data", *Biometrics*, C. 33, S. 1, s.59–174.

Langley, Pat, H. A. Simon (1995), "Applications of Machine Learning and Rule Induction", *Communications of the ACM*, S. 38, s. 55-64.

Larose, Daniel T. (2005), Discovering Knowledge In Data-An Introduction to Data Mining: Wiley Publishing

Lavanya, D., K.U. Rani (2011), "Analysis of Feature Selection With Classficatıon: Breast Cancer Datasets", *D.Lavanya et al./ Indian Journal of Computer Science and Engineering (IJCSE)*, S.2, s.756-763.

Leeuwen, Jan van, **"**Approaches In Machine Learning", *Institute of Information and Computing Sciences, Utrecht University, Padualaan 14, 3584 CH Utrecht, the Netherlands*.

Machine Learning Interaction Research in the Computer Science Department at Carnegie Mellon.
http://www.csd.cs.cmu.edu/research/areas/machinelearning/ (Erişim tarihi: 18.12.2013)

Maimon, Oded, L. Rokach (2010), Data Mining and Knowledge Discovery Handbook Second Edition: Springer.

Mangasarian, Olvi L., W.H. Wolberg, Machine Learning for Cancer Diagnosis and Prognosis.
http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html (Erişim tarihi: 15.04.2013)

Mariscal, Gonzalo, O. Marban, C. Fernandez (2010), A Survey Of Data Mining And Knowledge Discovery Process Models And Methodologies, *The Knowledge Engineering Review*, S 25: 2, s. 137–166.

Moss, Larissa (2009), "Data Mining", *EIMI Archives*, C. 3. http://www.eiminstitute.org/library/eimi-archives/volume-3-issue-1-january-2009-edition/data-mining (Erişim tarihi: 07.01.2014)

Nielsen, Robert H.(1989), Theory of the backpropagation neural network, *International Joint Conference on Neural Networks*.

Ohri, Edith (2011), What is the difference between statistical computing and data mining? http://www.analyticbridge.com/forum/topics/2004291:Topic:8561 (Erişim tarihi: 19.01.2014)

OLAP And OLAP Server Definitions http://altaplana.com/olap/glossary.html (Erişim tarihi: 13.02.2014)

OLAP Küpü Nedir? http://datawarehouse.gen.tr/olap-kupu-nedir-2/ (Erişim tarihi: 19.02.2014)

OLTP – OLAP http://fatma.molu.net/oltp-olap/ (Erişim tarihi: 03.02.2014)

OLTP vs. OLAP http://datawarehouse4u.info/OLTP-vs-OLAP.html (Erişim tarihi: 18.02.2014)

Oracle Data Mining Concepts. http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/toc.htm (Erişim tarihi: 28.11.2013)

Oracle9i Data Warehousing Guide Release 2 (9.2) http://docs.oracle.com/cd/B10501_01/server.920/a96520/toc.htm (Erişim tarihi: 05.01.2014)

Overview of Online Analytical Processing (OLAP) http://office.microsoft.com/en-us/excel-help/overview-of-online-analytical-processing-olap-HP010177437.aspx (Erişim tarihi: 18.02.2014)

Özçınar, Hüseyin (2006), **"**KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi", *Pamukkale Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi*.

Özdemir, Abdulkadir, F. Aslay, H. Çam (2010), "Veri Tabanında Bilgi Keşfi Süreci: Gümüşhane Devlet Hastanesi Uygulaması", *Sosyal ve Ekonomik Araştırmalar Dergisi - Selçuk Üniversitesi*, C. 14, S. 20, s. 347-365.

Özkan, Yalçın (2013), Veri Madenciliği Yöntemleri, Papatya Yayınevi

Özmen, Şule. (2001) "İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor", *V. Ulusal Ekonometri ve İstatistik Sempozyumu*, Çukurova Üniversitesi

Piatetsky-Shapiro, Gregory, W. Frawley, C. J. Matheus (1992), "Knowledge Discovery in Databases", *AI Magazine* C. 13, S. 3, s. 57-70.

Ponniah, Paulraj (2010), Data Warehousing Fundamentals for It Professionals, Wiley Publications..

Salama, Gouda I., M. B. Abdelhalim, M.A. Zeid (2012), "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", *International Journal of Computer and Information Technology (2277–0764)*, C. 01, S. 1, s.36-43.

Schapire, Rob (2013), Theoretical Machine Learning. http://www.cs.princeton.edu/courses/archive/spring13/cos511/scribe_notes/0205.pdf (Erişim tarihi: 02.02.2014)

Siegel, Rebecca, M. Jiemin, Z. Zou, A. Jemal (2014), "Cancer Statistics 2014", *CA: A Cancer Journal for Clinicians*, S.64, s. 9–29.

Statoo Consulting-Statistical Consulting + Data Analysis + Data Mining Services Switzerland.
http://www.statoo.com/en/datamining/ (Erişim tarihi: 25.12.2013)

Sullivan, Rob (2012), Introduction to Data Mining for the Life Sciences, Cincinnati, OH, USA: Springer.

Sumathi, S., Sivanandam, S.N. (2006), Introduction to Data Mining and its Applications: Springer.

Şentürk, Aysan (2006), Veri Madenciliği-Kavram ve Teknikler, Ekin Basım Yayın.

Şimşek Gürsoy, U. Tuğba (2009), Veri Madenciliği ve Bilgi Keşfi, Pegem Akademi Yayıncılık.

Timor, Mehpare, A. Ezerce, U.T. Gürsoy (2011), "Müşteri Profili ve Alışveriş Davranışlarını Belirlemede Kümeleme ve Birliktelik Kuralları Analizi: Perakende Sektöründe Bir Uygulama", *İstanbul Üniversitesi İşletme Fakültesi İşletme İktisadı Enstitüsü Dergisi-Yönetim*, C. 22, S. 68, s. 128-147.

IBM Knowledge Center, The data mining process
http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=%2Fcom.ibm.im.easy.doc%2Fc_dm_process.html (Erişim tarihi: 04.03.2014)

Tripoliti, Evanthia, G. Manis (2012), "Feature Selection in HRV Analysis of Young and Elderly Subjects", *5th European Conference of the International Federation for Medical and Biological Engineering IFMBE Proceedings*, C. 37, s. 516-519.

Türk Dil Kurumu
http://www.tdk.gov.tr/index.php?option=com_gts&arama=gts&guid=TDK.GTS.53afed1e92d0b8.21139613 (Erişim tarihi: 01.10.2013)

Üner, Tuğçe (2010), "Pazarlamada Müşteri İlişkileri Yönetimi (MİY) ve E-MİY Analizlerinin Değerlendirilmesi", *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, C. 12, S. 3, s. 87-104.

Valpola, Harri (2000), Bayesian Ensemble Learning for Nonlinear Factor Analysis, *Helsinki Teknoloji Üniversitesi Doktora Tezi*.

Veri Ambarı – ETL – Veri Madenciliği
http://emraharslanbm.wordpress.com/tag/data-warehouse/ (Erişim tarihi: 20.01.2014)

Veri Ambarı ve İş Zekâsı Yapısal Teorisi
http://www.cevizbilgi.com.tr/veri-ambari-ve-zekasi-yapisal-teorisi/ (Erişim tarihi: 05.05.2014)

Whats a Data Mart?
http://www.yellowfinbi.com/YFForum-Whats-a-Data-Mart-?thread=92098 (Erişim tarihi: 16.01.2014)

What is a data warehouse - A 101 guide to modern data warehousing
http://www.dwbiconcepts.com/data-warehousing/18-dwbi-basic-concepts/87-what-is-a-data-warehouse-guide.html (Erişim tarihi: 12.01.2014)

Wirth, Rüdiger, J. Hipp (2000), "CRISP-DM: Towards a Standard Process Model for Data Mining", *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*.

Witten, Ian H., F. Eibe, M.A. Hall (2011), Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed): Morgan Kaufmann Publishers, Elsevier.

Wolberg, William H., W.N. Street, O.L. Mangasarian (1994), "Machine Learning Techniques To Diagnose Breast Cancer From Fine-Needle Aspirates", *Cancer Letters*, C. 77, s. 163-171.

Yongjian, Fu (1997), **"**Data Mining: Tasks, Techniques and Applications", *IEEE Potentials*, C. 16, S. 4, s. 18-20.

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 (Erişim tarihi: 14.04.2014)

http://datawarehouse.gen.tr/ (Erişim tarihi: 17.02.2014)

http://www.gartner.com/it-glossary/data-mining (Erişim tarihi: 21.12.2013)