**T.C.**
**TRAKYA ÜNİVERİSTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# TIP 'DA VERİ MADENCİLİĞİ UYGULAMALARI: MEME KANSERİ VERİ SETİ ANALİZİ

**Oğuz POYRAZ**

**Yüksek Lisans Tezi**

**Bilgisayar Mühendisliği Anabilim Dalı**
**Danışman: Doç. Dr. Erdem UÇAR**

**EDİRNE-2012**

**T.C.**
**TRAKYA ÜNİVERİSTESİ**

**FEN BİLİMLERİ ENSTİTÜSÜ**

# TIP 'DA VERİ MADENCİLİĞİ UYGULAMALARI: MEME KANSERİ VERİ SETİ ANALİZİ

**Oğuz POYRAZ**

**Yüksek Lisans Tezi**
**Bilgisayar Mühendisliği Anabilim Dalı**

**Danışman: Doç. Dr. Erdem UÇAR**

**2012**
**EDİRNE**

TRAKYA ÜNİVERİSTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

# TIP 'DA VERİ MADENCİLİĞİ UYGULAMALARI: MEME KANSERİ VERİ SETİ ANALİZİ

Oğuz POYRAZ

YÜKSEK LİSANS TEZİ

Bu tez 27/07/2012 tarihinde aşağıdaki jüri tarafından kabul edilmiştir.

| | | |
|---|---|---|
| Doç. Dr. Erdem UÇAR | Doç. Dr. Necdet SÜT | Yrd. Doç. Dr. İlhan UMUT |
| (Danışman) | (Üye) | (Üye) |

**TEŞEKKÜR**

Bu çalışmanın hazırlanması esnasında bana yardımcı olan, bu alanda çalışmam için beni teşvik eden, yardımlarını ve desteklerini benden esirgemeyen değerli hocam Doç. Dr. Erdem UÇAR ‘a teşekkür ederim.

Yüksek Lisans tez çalışmalarım sırasında veri madenciliği konusunda değerli tecrübelerinden yararlandığım Doç. Dr. Mehmet KAYA’ ya teşekkür ederim.

Çalışmalarım sırasında değerli katkılarıyla bana yardım eden ve ortak çalışmalar yaptığımız Arş.Gör. Ümit Can KUMDERELİ’ye, ayrıca çalışmalarım da her türlü katkıyı sağlayan diğer arkadaşlarıma çok teşekkür ederim.

Beni lisansüstü çalışmaya teşvik eden akademik tecrübelerinden her daim yararlandığım sevgili babam Prof.Dr. Mustafa POYRAZ ‘a sonsuz sevgi ve saygılarımı sunarım.

## ÖZET

Veri madenciliği, günümüz bilgi çağında en güncel makine öğrenmesi yöntemlerden birisidir. Bilgisayar sistemlerinin her geçen gün hem daha ucuzluyor olması, hem de güç ve kapasitelerinin artıyor olması, bilgisayarlarda daha büyük miktarlarda verinin saklanabilmesine imkan vermektedir.

Veri madenciliği, dünya üzerinde artan veri miktarının etkili bir biçimde kullanılmasının neredeyse tek çözümü olarak görülmektedir. Bu yüzden, büyük miktardaki verileri isleyebilen teknikleri kullanabilmek, günümüzde büyük önem kazanmaktadır. Veri madenciliği bu gibi durumlarda kullanılan, büyük miktardaki veri setlerinde saklı durumda bulunan örüntü ve eğilimleri keşfetme işlemidir.

Veri ambarlarında toplanan veriler tek başlarına değersizdirler. Bu veriler ancak belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmektedir. Veriyi amacına uygun bilgiye dönüştürme isi veri madenciliği ile yapılabilmektedir.

Akıllı veri işleme metodu olan veri madenciliği, dünya üzerinde artan veri miktarının etkili bir biçimde kullanılmasının neredeyse tek çözümü olarak görünmektedir. Bu gelişme diğer alanlarda olduğu gibi tıp alanında da ilgi odağı haline gelmiştir. Özellikle tıp alanındaki verinin büyüklüğü ve hayati önem taşıması bu alandaki uygulamaları daha da önemli kılmaktadır.

Bu tezde sağlık verilerinden veri madenciliği uygulaması üzerine hazırlanmıştır. Veri madenciliğin tanımı ve veri madenciliği teknikleri ile kanser çeşitlerinden biri olan ve bayanlar arasında en sık görülen Meme Kanseri hakkında kısa bilgi verildikten meme kanseri üzerine weka'da yapılan uygulama anlatılacaktır.

Uygulamada Wisconsin veri seti kaynağından meme kanseri verileri üzerinden bir karar ağacı algoritması olan ve temeli ID3 ve C4.5 algoritmalarına dayanan J48, Bayes sınıflandırma algoritmalarından Naive-Bayes, regresyon tabanlı algoritmalardan lojistik regresyon ve örnek tabanlı sınıflandırma algoritmalarından Kstar algoritmaları kullanılarak modeller oluşturulmuş ve oluşturulan modellerin başarım dereceleri karşılaştırılmıştır.

**Anahtar Kelimeler:** Veri Madenciliği, Tıp Bilişimi, Meme Kanseri .

**ABSTRACT**

Data mining, machine learning methods in today's information age is one of the most up to date. Day by day computer systems are being cheaper and also their capacities are increasing, so it enables computers to store more data.

Data mining is seen as a unique solution in all over the world for using data capacity in an efficient way. Therefore using techniques that can process huge data gain more importance today. Data mining is exploring hidden patterns and trends that are use in these kinds of data sets.

Data that are collected in data warehouse solitarily is invaluable. These can be valuable if they will be processed for an aim. Replacing data into information can be performed by data mining.

Data mining having smart data process methods is seen as a unique solution in all over the world for using data capacity in an efficient way. These developments become the center of attention in all other areas but also in medicine. Especially the size of the data in medicine area and the content of the data make applications in this area more important.

This thesis is prepared upon data mining based on health data. After giving a short brief on breast cancer, which is seen most frequent between females and which is one of the types of cancer, the application that is performed on “weka” will be described.

These application models are constituted by using J48, which is based on ID3 and C4.5 and which is a decision tree algorithm on breast cancer, Naïve-Bayes algorithm, which is one of the classification algorithms of Bayes, Logistic regression, which is based on regression and sample based Kstar algoritms and the success degrees of these methods are compared.

**Keywords:** Data Mining, Medical Informatics, Breast Cancer .

## İÇİNDEKİLER

## TABLOLAR DİZİNİ

## ŞEKİLLER DİZİNİ

# BÖLÜM 1.

# GİRİŞ

Veri madenciliği alanı, bilgisayar teknolojisinin gelişmeye başladığı yıllardan sonra günümüzde en güncel teknolojilerinden biri olma yönünde en büyük eğilimi göstermiştir, elde edilen verilerin sürekli ve büyük bir hızla artmasıyla ortaya çıkan veri analizi ihtiyacına bağlı olarak büyük bir hızla gelişmeye başlamıştır. Bu gelişmeler bu kısa zaman diliminde pek çok akademik araştırma ve geliştirmeyi peşi sıra getirmiştir.

Veri madenciliği teknolojisi ile büyük miktardaki verileri işleyebilme teknikleri kullanılarak, gizli kalmış bilgileri keşfetmek, geleceğe dönük kararlar almamızda, karar destek oluşumuna katkı sağlamak gibi işlevleri üstlenmiştir. Günümüz de birçok kurumsal uygulamada veriler üzerinden karar verme sürecin de anlamlı bilgiyi üretmede etkin rolü oynamaktadır.

Günümüzde verinin, bilginin yönetiminde daha çok öz bilgi ile ilgilenilmektedir. Bilgi teknolojilerinin gelişimi ve gündelik hayatın her aşamasında kullanılabilir hale gelmesiyle beraber, her alanda oldukça büyük miktarda veri birikmeye başlamıştır. Böylece, banka, üniversite, okul, seyahat şirketi, hastane, devlet dairesi benzeri kuruluşların çalışıp işleyebilmesi için kayıt altında tutmak durumunda olduğu çeşitli veriler veritabanlarında depolanmıştır. Verilerin hafızadaki durumları, veritabanı yaratmak ve yönetmek, kullanıcıların erişimleri, verilerin yönetilmesi, yedeklerin alınması gibi işlemleri düzenleyen sistemlerin (veritabanı yönetim sistemlerin) artan kullanımı ve hacimlerindeki olağanüstü artış, kuruluşların elde toplanan bu verilerden nasıl faydalanabileceği problemi ile karşı karşıya bırakmıştır.

Geleneksel sorgu ve raporlama araçlarının veri yığınları karşısında yetersiz kalması Veri Tabanlarında Öz Bilgi Keşfi (Knowledge Discovery in Databases) adı altında, sürekli ve yeni arayışlara yöneltmiştir. Veri tabanlarında öz bilgi keşfi süreci

içerisinde, model kurulması ve değerlendirilmesi aşamalarından meydana gelen Veri Madenciliği (Data Mining) en önemli kesimi oluşturmaktadır. Veri madenciliği Veri Tabanlarında Öz bilgi Keşfi'nin özünü oluşturan keşif kısmının gerçekleştiği adım olarak alınabileceği gibi bağımsız süreç olarak da değerlendirilmektedir.

Veri madenciliğinin amacı, geçmiş faaliyetlerin analizini temel alarak gelecekteki davranışların tahminine yönelik karar-verme modelleri yaratmaktır.

Veri madenciliği ilk olarak müşteri ilişkilerinde başlamıştır. Veri madenciliği organizasyonel hedeflerin başarılmasında çok geniş kullanım alanına sahiptir. Ayrıca, bankacılıkta finansal göstergelere ilişkin gizli ilişkilerin bulunmasında, pazarlamada müşterilerin satın alma örüntülerinin belirlenmesinde ve sigortacılıkta ise riskli müşterilerin örüntülerinin belirlenmesinde veri madenciliği uygulamalarına çok sık rastlanılmaktadır.

Günümüzde, veri madenciliği firmalar tarafından öncelikle müşteri odaklı olarak (finansal, iletişimsel ve pazarlama) kullanılmaktadır. Veri madenciliği firmalara fiyat, üretim planlaması, personel becerileri gibi iç faktörleri belirlemelerine olanak tanımaktadır. Ayrıca, ekonomik göstergeler, rekabet ve pazarın yapısı gibi dış faktörleri belirlemelerine olanak tanımaktadır. Böylece, firmaların satışları, müşterilerinin tatmini ve şirket karları üzerindeki olumlu ya da olumsuz etkiler belirlenebilmektedir.

Sonuçta, öz bilgiyi elde etme ve veriler içindeki detayları görebilme olanağı sağlanmaktadır. Veri madenciliği birçok alanda uygulanabilmektedir.

Veri madenciliği her geçen gün yeni ve farklı alanlarda kullanılmaya başlamakla birlikte günümüzde yaygın olarak kullanıldığı başlıca alanlar şöyle özetlenebilir.

### 1.1 Sağlık ve Biyoloji

- Yeni virüs türlerinin keşfi ve sınıflandırılması
- Hastalıkların özelliklerinin belirlenerek teşhislerin kolaylaştırılması
- Tıbbi kayıtlar

- Birlikte kullanılan ilaçların yan etkilerinin araştırılması
- Test sonuçlarının tahmini
- Ürün geliştirme
- Tedavi sürecinin belirlenmesi

Tıp en çok verilerin tutulduğu sağlıklı verilerin bulunduğu alanların başında bulunmaktadır. Özellikle son yıllarda genetiğin inanılmaz hızda ilerlemesi sonucun oluşan gen haritaları ile hastalıklar sınıflandırılmaya, hangi genlere sahip bireylerin hangi hastalıklara yakalanma olasılığı olduğuna dair çalışmaları bol miktarda televizyonlardan ve gazetelerden öğreniyoruz.

Aynı şekilde virüslerin yapısı incelenerek onlar sınıflandırılıyor. İlaçların üretimi, kullanımı yan etkilerinin araştırılması konusunda da benzer çalışmalar veri madenciliği ile yapılabilmektedir.[Kocabaş,2006]

Sağlık alanında bilginin kullanım şeklinde meydana gelen değişiklikler, sağlık bakım hizmetini verenleri etkilemiştir; sağlık bakım hizmetinin verilmesinde bilgisayar kullanımı, bilginin paylaşım-ekip yaklaşımını, veri ve bilgi temelli uygulama gibi kavramlar yaygınlaşmaya başlamıştır. Bilgisayarlar hasta bakım hizmetlerinin destekleme, sağlık bakım hizmetlerinin kalitesinin değerlendirilmesi gibi doğrudan sağlık bakım hizmetlerinin sunulmasında kullanılmasının yanı sıra, karar verme, yönetim, planlama ve tıbbi araştırmalar gibi yönetsel ve akademik fonksiyonların yerine getirilmesinde daha fazla kullanılmaya başlanılmıştır.

Sağlık alanında bulunan mevcut veri oldukça fazla ve hayati öneme sahiptir. Hastane bilgi sistemleri sayesinde bu veriler düzenli olarak tutulmaktadır. Hayati öneme sahip olan bu verilerden daha fazla yaralanmak mümkündür. Hastane bilgi sistemlerinden veya diğer tıbbi veri toplayan sistemlerden alınan veriler üzerinde yapılan veri madenciliği çalışmaları hem uzmanlar için hem hastane yönetimi için hem de hastaların daha kaliteli bir hizmet almalarında etkin rol alabilir.

### 1.2 Telekomünikasyon

- İletişim Hatları yoğunluk tahmini
- Web sitesi ziyaretçilerinin profil analizi
- Kalite ve iyileştirme analizleri

Telekomünikasyon sektörünün inanılmaz boyutlara ulaştığı ve çok daha büyük bir ivme ile artacağı çağda yaşadığımızı düşünürsek kişilerin kullanım sıklıkları, amaçları ve hat yoğunluk tahminleri yapılarak firmalar altyapı güncellemelerine gidebilir, müşteriye ilişkin müşteriye özel kampanyalar düzenleyebilirler.

Aynı şekilde web siteleri her şeyin kayıt altına tutulabildiği uygulamalardır. Bu uygulamalar yardımıyla bırakın kullanıcı profili çıkarmayı (Google Analytics bile ihtiyaçlarımızı fazlasıyla karşılıyor) click mining diye tabir edilen kavram ile kullanıcıların web uygulamasında yaptıkları işlemler bile analiz edilebiliyor hatta bir butonun yeri bile değiştirebiliyor, uygulama da gerekli optimizasyon çalışmaları yapılabiliyor.

### 1.3 Finans (Bankacılık, Borsa)

Farklı finansal göstergeler arasında korelasyon tespiti, kredi kartı dolandırıcılıklarının tespiti, kredi taleplerinin değerlendirilmesi, kredi kartı harcamalarına göre müşteri profili belirlenmesinde, sigorta dolandırıcılıklarının tespitinde, yeni poliçe talep edecek müşterilerin tahmininde yoğun olarak kullanılmaktadır.

Farklı finansal göstergeler arasında gizli korelasyonların bulunması,

- Kredi kartı dolandırıcılıklarının tespiti,

- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,

- Kredi taleplerinin değerlendirilmesi.[Baykal,2]

### 1.4 Pazarlama

Müşterilerin satın alma örüntüleri, demografik bilgileri, kampanya ürünleri belirleme, mevcut müşterileri kaybetmeden yeni müşteriler kazanma, pazar sepeti analizi (Market Basket Analysis), müşteri ilişkileri yönetimi (CRM – Customer Relations Management), Müşteri değerlendirme (Customer Value Analysis), Satış tahmini (Sales Forecasting). alanları en yaygın veri madenciliği uygulama alanlarıdır.

### 1.5 Sigortacılık

Yeni poliçe talep edecek müşterilerin tahmin edilmesi, Sigorta dolandırıcılıklarının tespiti riskli müşteri örüntülerinin belirlenmesi. Fiyat belirlemesinde ülkenin coğrafi koşulları, kültürel yapısı, bölgelere göre bireylerin gelirleri ve daha bir çok değişken alınarak yeni oranların hesaplanması kritik önemdedir.

### 1.6 Astronomi

Gezegen yüzey şekillerinin ve gezegen yerleşimleri, yeni galaksiler keş fi, yıldızların konumlarına göre gruplandırılmasında kullanılmaktadır.

### 1.7 Biyoloji, Tıp ve Genetik

Bitki türleri ıslahı, gen haritasının analizi ve genetik hastalıkların tespiti, kanserli hücrelerin tespiti, yeni virüs türlerinin keşfi ve sınıflandırılması, fizyolojik parametrelerin analizi ve değerlendirilmesinde kullanılmaktadır.

Gen bilimi ile de ortaklaşa çalışma sayesinde hangi bireylerin suç işlemeye eğilimli olduklarına dair tahminleme çalışmaları yapılarak bu olayların daha ortaya

çıkmadan önlenmesinden tutun, kullanıcıların yazım karakterlerinden yola çıkarak birçok olasılığın hesaplanmasından çıkın veri madenciliğinin en çok kullanıldığı alanlardan birisi kriminolojidir.

### 1.8 Kimya

Yeni kimyasal moleküllerin keşfi ve sınıflandırılması, yeni ilaç türlerinin keşfinde kullanılmaktadır.

### 1.9 Yüzey Analizi ve Coğrafi Bilgi Sistemleri

Bölgelerin coğrafi özelliklerine göre sınıflandırılması, kentlerde yerleşim yerleri belirleme, kentlerde suç oranı, zenginlik-yoksulluk, köken belirleme, kentlere yerleştirilecek posta kutusu, otomatik para makinaları, otobüs durakları gibi hizmetlerin konumlarının tespitinde kullanılmaktadır

### 1.10 Görüntü Tanıma ve Robot Görüş Sistemleri

Çeşitli algılayıcılar aracılığı ile tespit edilen görüntülerden yola çıkarak engel tanıma, yol tanıma, yüz tanıma, parmak izi tanıma gibi tekniklerde kullanılmaktadır.

### 1.11 Uzay Bilimleri ve Teknolojisi

### 1.12 Meteoroloji ve Atmosfer Bilimleri

Bölgesel iklim, yağış haritaları oluşturma, hava tahminleri, ozon tabası deliklerinin tespiti, çeşitli okyanus hareketlerinin belirlenmesinde kullanılmaktadır.

### 1.13 Sosyal Bilimler ve Davranış Bilimleri

Kamuoyu yoklamaları inceleme, genel eğilim belirleme, seçim öngörüleri oluşturmada kullanılmaktadır.

### 1.14 Metin Madenciligi (Text Mining)

Çok büyük ve anlamsız metin yığınları arasından anlamlı ilişkiler elde etmekte kullanılmaktadır.

### 1.15 Internet Madenciliği (Web Mining)

İnternet üzerindeki veriler hem hacim hem de karmaşıklık olarak hızla artmaktadır.

Sadece düz metin ve resimden başka akan (streaming) ve sayısal veriler de web verileri arasında yer almaktadır. 5nternetin belirli kategorilere ayrılarak veriye ulaşım süresinin azaltılması web madenciliğinin temel hedefidir.

# BÖLÜM 2.

# VERİ MADENCİLİĞİ

Veri madenciliği özellikle sağlık verilerinin kullanımda yaygın bir yöntem haline gelmiştir. Bu çalışmada veri madenciliği modelleri işlevlerine göre Sınıflama, Regresyon, Kümeleme ve Birliktelik Kuralları baslıkları altında incelenmekte ve uygulama alanları açıklanmaktadır.

Veri madenciliği araç ve metotlarının gelişmesiyle iş dünyasından kaynaklanan, konuya yönelik taleplerden ötürü, algoritmaların ve yazılım araçlarının geliştirilmesine yönelik, hem iş çevresinde hem de akademik çevrede konuya yoğun bir ilgi oluşmuş, verilerin sürekli büyümesi ve algoritmaların kompleksliğinden dolayı daha iyi sonuçlar almanın yolları araştırılmıştır. Yapılan araştırmalarda ortaya çıkan pek çok yöntemden hangisinin daha iyi olduğu gibi sorular ortaya çıkmıştır. Uygulanan teknoloji ve algoritmaların verimliliği her ne kadar karşılaşılan problem alanına bağımlı olsa da akademik anlamda karşılaştırma yapılması ihtiyacı doğmuştur.

Veri analizinin istatistik bilimine bağlı olması, ayrıca yapay zeka ve makine öğrenme gibi temelde istatistik ve matematik bilimine bağlı farklı akademik disiplinlerin oluşturduğu veri madenciliği yöntemlerinin değerlendirilmesi doğal olarak yine istatistik biliminin temel kuramları üzerinden yapıla gelmiştir.

Veri madenciliği çalışmalarında çok çeşitli yöntemler kullanılmaktadır. Farklı alanlarda çok geniş bir uygulama alanına sahip olduğu için var olan yöntemler üzerinde iyileştirmeler yapılmakta ve yeni yöntemler geliştirilmektedir.

Aynı zamanda, matematik, istatistik, enformatik ve bilgisayar bilimlerindeki gelişmeler de bu alana yansımaktadır. Bu sebeple, veri madenciliği, geniş bir uygulama alanına sahip olup, gelişmeye açık, sadece akademik değil aynı zamanda iş dünyasının da yoğun ilgisini çeken bir alandır.

Veri madenciliği uygulama alanının çok geniş olması bu konuya olan ilgiyi de arttırmaktadır. Kullanılan pek çok model ve bu modellere ait farklı algoritmalar vardır. Bu algoritmalardan hangisinin daha efektif sonuçlar ürettiği, hangi algoritmanın hangi alanda daha başarılı olduğu sorusuna verilen cevaplar uygulamaların başarımını arttıracak ve yapılan işin verimini arttıracaktır. Bu sebeple algoritmaların karşılaştırılarak değerlendirilmesi büyük önem arz etmektedir.

Bu tezde sağlık verilerinin veri madenciliği algoritmalarının karşılaştırılması amaçlanmıştır. Farklı sınıflandırma algoritmalarının nasıl karşılaştırılabileceği ve kullanılabilecek metrikler üzerinde durulmuş, sık kullanılan ve de bilinen dört farklı sınıflandırma algoritması karşılaştırılarak veri ön işlemeden başlamak üzere model oluşturulması ve modellerin karşılaştırılması konusunda bilgi verilmesi amaçlanmıştır.

## 2.1. Veri Madenciliği Tarihçesi

Veri madenciliği son yıllarda adını duyurmaya başlasa da literatürde 1980'lerden itibaren yer almıştır.

Veri madenciliğinin kavramı üç temel başlık olarak gelişmiştir. Bunlardan ilki ve en eskiye dayananı klasik istatistik bilimidir. Klasik istatistik bilimi regresyon analizi, standart dağılım, standart sapma, diskriminant analizi, güven aralıkları gibi verileri ve veriler arasındaki ilişkiyi inceleyen yöntem ve klasik istatistik çalışmalarından oluşur. Klasik istatistiksel yöntemler veri madenciliğinin araç ve metotlarının esasını oluşturur.

Veri madenciliğinin diğer başlığı yapay zeka (AI) dır. Yapay zeka, sezgisel - heuristic - yaklaşımı temel alarak, insan-benzeri-düşünebilme prensibiyle, istatistikten farklı metotlarla, istatistiksel problemlere yaklaşır. Bu yaklaşım uygulanabilirlik açısından yüksek kapasitede bilgisayar gücü gerektirdiği için, güçlü bilgisayar sistemlerinin kullanıcının hizmetine sunulmaya başlandığı 1980‟ li yıllara kadar pratik uygulamalarda yer edinememiştir. Hala pek çok uygulama, süper bilgisayarlar gibi

kişisel bilgisayarlardan daha güçlü makineler gerektirdiği için, bu uygulamaların pek çoğunun büyük şirket ya da devlet kurumları ile sınırlı kaldığı söylenebilir.

Veri madenciliğinin üçüncü başlığı istatistik ve yapay zekadan alan makine öğrenmesidir. Makine öğrenme, yapay zekanın sezgisel - heuristic - yöntemleri ileri düzey istatistiksel yöntemlerle harmanlayıp evrimleşerek geliştiği ileri düzey halidir denebilir. Makine öğrenme, uygulandığı bilgisayar sistemlerinde, istatistiksel ve yapay zeka algoritmaları kullanarak eldeki verinin değerlendirilmesine, bu verilerden sonuçlar çıkarılmasına ve bu sonuçlara bakılarak kararlar alınmasına olanak sağlar.

Temel olarak veri madenciliği, öğrenme yöntemlerinin iş ve bilimsel verilere uygulanarak anlamlı bilginin çıkarılmasıdır. Veri madenciliği, istatistik, yapay zeka ve makine öğrenme disiplinlerinin gelişmesiyle ortaya çıkan, eldeki veriden öğrenme yoluyla gizli bilgileri ve örüntüleri ortaya çıkararak ileriye dönük tahminler yapmayı amaçlayan yeni bir bilim dalıdır. İş ve bilim alanında, normalde çok yoğun veri kümelerinden çıkarılması imkansız bilgiyi çıkarmada gün geçtikçe daha çok kabul görmektedir.

## 2.2 Literatür Özeti

Veri madenciliğinde bilgiye erişmede farklı metotlar kullanılmaktadır. Bu metotlara ait pek çok algoritma vardır. Bu algoritmalardan hangisinin daha üstün olduğu üzerine pek çok çalışma yapılmış, yapılan bu çalışmalarda farklı sonuçlar elde edilmiştir. Bunun en önemli sebebi, işlem başarımının, kullanılan veri kaynağına, veri üzerinde yapılan önişleme, algoritma parametrelerinin seçimine bağlı olmasıdır. Farklı kişiler tarafından, farklı veri kaynakları üzerinde, farklı parametrelerle yapılan çalışmalarda farklı sonuçlar oluşması doğaldır. Ancak, yaptığım çalışma, "benzer veri kümelerinde belli yöntemlerin daha başarılı olduğu" şeklindeki çıkarıma [Michie,D.,1994] uygun olarak, diğer çalışmalarla [Delen, D. ; Walker, G. ; Kadam, A.,2004] [Bellaachia, A. ; Guven, E.2006] benzer sonuç vermiştir. Göğüs kanseri vakalarının farklı yıllarını içeren Wisconsin veri kümesi kaynağının kullanıldığı çalışmada, bir karar ağacı algoritması olan C4.5 algoritmasının diğer algoritmalardan

daha iyi sonuç ürettiği sonucuna ulaşıldığı belirtilmiştir [Delen, D. ; Walker, G. ; Kadam, A.,2004, Bellaachia, A. ; Guven, E.2006] Bu tez çalışmasında da, yapılan karşılaştırma sonucunda, C4.5 algoritmasının Weka implementasyonu olan J48 karar ağacı algoritması, benzer şekilde diğer algoritmalara göre daha başarılı bulunmuştur.

Literatürdeki diğer karşılaştırma çalışmalarında sonucun kullanıcının yatkın olduğu modele bağlı olduğu, bu yüzden farklı makalelerde farklı sonuçlara ulaşılacağı belirtilmiştir. Bunun dışında bazı çalışmalarda kompleks algoritmaların klasik algoritmalara karşı daha başarılı olduğu şeklindeki iddiaların da aslında illüzyondan ibaret olduğu ifade edilmektedir [Hand, D. J 2006.]

Deneysel çalışmalar üzerine yapılan bu eleştirilerin haklılık payı büyüktür. Dolayısı ile yapılan bir karşılaştırma işlemine dayanarak bir algoritmanın diğer bir algoritmaya kesin bir üstünlüğünden söz etmek doğru olmayacaktır. Ancak model başarımı karşılaştırmalarının, bir veri madenciliği çalışmasında önemli katkıları olacağı açıktır. Bir kullanıcının bir problem üzerinde yapacağı model oluşturma işleminde farklı algoritmaları karşılaştırarak en başarılıyı bulmasının ve modelini o algoritma ile kurmasının elbette sonuçlar üzerinde olumlu etkisi olacaktır. Ancak, burada dikkat edilmesi gereken nokta öğrenme kümesinin seçimidir. Çünkü farklı öğrenme kümeleriyle yapılan farklı karşılaştırmalar farklı sonuçlar verebilir [Hand, D. J,2006]. Ayrıca yeni geliştirilen bir algoritmanın bilimsel anlamda geçerliliğinin belirlenmesinde deneysel çalışmaların önemli bir yeri vardır.

2003 yılında yapılan bir çalışmada 1991yılında Meme Kanseri Wisconsin (Orijinal) Veri Seti veri kaynağındaki göğüs kanseri hasta kayıtları üzerinde yapılan çalışmada eğiticili ve eğiticisiz nöral algoritmalar göğüs kanseri teşhisi amacıyla karşılaştırılmıştır. RBF eğitme setindeki en iyi sınıflayıcı olmasına rağmen en önemli sonuç Test kümesi düşünüldüğünde SOM en iyi sınıflama oranını vermektedir. Genel olarak sonuçlara bakıldığında WBCD verisinin sınıflandırılmasında en uygun nöron ağı modeli RBF ve SOM olmuştur. Ayrıca sonuçlar, eğiticili ve eğiticisiz nöral algoritmaların göğüs kanseri teşhisinde büyük başarı elde ettiğini göstermiştir. [KIYAN ,YILDIRIM,2003]

## 2.3. Veri Madenciliği Hakkında Temel Bilgiler

Veri madenciliği büyük veri kaynaklarındaki gizli, önemli ve yararlı bilgilerin bilgisayar yardımıyla keşfedilmesidir. Veriler arasındaki benzerliklerin, örüntülerin ya da ilişkilerin çıkarılması amacıyla uygulana işlemler bütünü oluşturur. Veri madenciliğinin ekonomi alanında pazar araştırması, müşteri profilinin çıkarılması, sepet analizi; bankacılıkta risk analizi, sahtekarlıkların saptanması; bilişimde web verilerinin analizi, ağ güvenliği, belgelerin sınıflandırılması gibi uygulamaları mevcuttur. Bunların dışında meteorolojide, tıpta, temel bilimlerde, ilaç biliminde ve diğer alanlarda da uygulamaları mevcuttur. Her ne kadar veri madenciliği yeni bir alan olsa da, aslında daha önceleri ekonomistler, istatistikçiler, hava durumu tahminleyicileri, eldeki verileri kullanarak ileriye dönük tahminler yapmakla uğraşıyorlardı. Son on yıllarda veri miktarlarındaki hızlı büyüme, farklı tarzlardaki verilerin farklı algoritma ihtiyacı, bu disiplinin kendi ayakları üzerinde durma gereksinimine sebep olmuştur.

Gelişen teknoloji ile birlikte hayatımızdaki veriler gün be gün büyümekte, daha önceleri kilobaytlarla ifade edebildiğimiz kişisel bilgisayarlardaki veriler artık megabaytlar, gigabaytlar ile ifade edilebilmektedir. Daha önceleri çöpe atılabilir tarzdaki veriler bile, bilgi depolama aygıtlarının gelişmesiyle beraber depolanmaya başlanmıştır. Günlük hayatımızı kolaylaştıran bankacılık işlemleri, online sistemler, internetin yaygınlaşması, bilgiye kolay erişim ve bilgi aktarım gereksinimini arttırmıştır. Bu gibi gelişmeler veri miktarının hızlı bir şekilde artmasında bir faktör olmuştur. Doğrulanabilirliği mümkün olmasa da, bir tahmine göre dünyadaki toplam veri miktarı her 20 ayda bir ikiye katlanmaktadır. Büyük şirketlerin, okulların, hastanelerin, bankaların, alışveriş merkezlerinin, diğer özel ve kamu kurumlarının veri bankaları büyük veri yığınlarından oluşmaktadır. Bu veriler analiz edilerek ileriye dönük politika belirlemede, geleceği öngörmede ya da var olan sistem hakkında karar alıcı

mekanizmalarda önemli rol oynarlar. Veri madenciliği bu büyük miktardaki verilerin analiz edilmesi için uygulanması gereken basamakların bütününü kapsar.

### 2.3.1. Veri

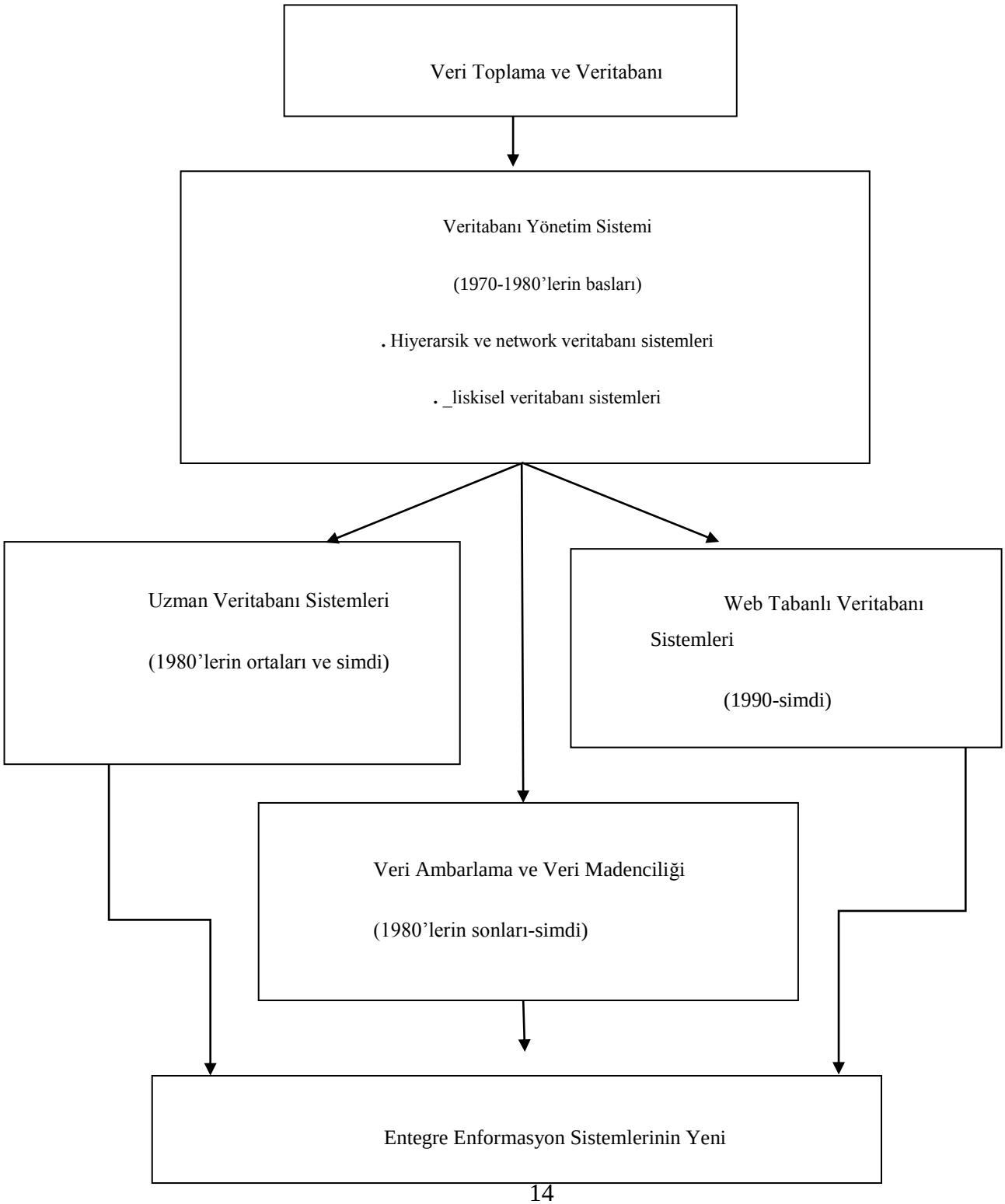
Veri, nesneleri, işlem ya da eylemleri niteliklerinin değerleriyle tanımlayan bilgi kümesidir. Nesneler ya da işlemler niteliklerden oluşur. Örneğin nesnemiz ‘otomobil ’ ise, ‘ rengi’, ‘yaşı’, ‘markası’, ‘modeli’ onun nitelikleridir. ‘otomobil’ nesnelerinin oluşturduğu küme ise veridir. Benzer şekilde bankadan para çekme işlemini ele alırsak ‘çekilen tutar’, ‘çekildiği hesap numarası’, ‘çekiliş saati’, ‘çekildiği yer’ bu işlemi tanımlayan niteliklerdir. İçerisinde birden çok para çekme işlemine ait bilgiyi barındıran küme ise veridir.

### 2.3.2. Veri Tabanı Teknolojisi

Veritabanı yönetim sistemleri sayesinde büyük ve karışık verilere ulamsak oldukça kolaydır. Çünkü verilerin depolandığı dosyalar, veri miktarı büyüdükçe istenilen veriye ulaşmayı zorlaştırmaktadır. Veritabanı teknolojileri ise bu dosyaları düzenleyerek verilere daha hızlı ve düzenli bir şekilde ulaşmayı sağlamaktadır. Ayrıca birden fazla kullanıcının aynı bilgileri birbirini etkilemeden kullanmasına da imkan vermektedir.

Veri madenciliği ile ilgili olarak yapılan tanımlardaki ortak nokta her bir tanımın ‘büyük ölçekli veriler’ den söz etmesidir. Bu büyük ölçekli verinin temeli ise veritabanlarına dayanmaktadır. Veri madenciliği disiplininin meydana gelmesinde veritabanı teknolojisindeki gelişmelerin önemi oldukça büyüktür. Veritabanı birbirleriyle ilişkili verilerin tekrara yer vermeden, çok amaçlı kullanımına olanak sağlayacak şekilde depolanması olarak tanımlanmaktadır. Kısaca, veritabanı bir veri kümesi olup, kullanıcıların ihtiyaçlarına göre sınıflandırılmalı, raporlanmalı ve analiz edilmesi gerekmektedir. Bu süreç ise veri analizi ile sağlanmaktadır.[Babadağ,2006]

Yukarıda da bahsedilen veritabanı gelişimi ve bu gelişme içerisinde veri madenciliğinin rolü sekil 1' de gösterilmiştir.[Han. J,Kamber. M,2001]



Sekil 2.1. Veritabanı Teknolojisinin Gelişimi ve Veri Madenciliği

Sekil 1 incelendiğinde de veri madenciliğinin veritabanı gelişim sürecinin bir ürünü olduğu görülmektedir. Son yıllarda veri ambarı, veri madenciliği ve OLAP (On-Line Analytical Processing) karar destek sistemleri teknolojilerindeki gelişmeler ile yapılan araştırmalar oldukça fazla önem kazanmıştır. Önceki yıllarda (1980'lerin sonlarında) veri erişimi ve veri değiştirme gibi işlemleri yapan OLTP (On-Line Transaction Processing) sistemlerini veri ambarları ve OLAP teknolojisinin gelişimi takip etmektedir. Bu nedenle aşağıda veri ambarı, OLTP ve OLAP teknolojilerinden kısaca bahsedilecektir.

### 2.3.3.Veri Ambarı

Veri ambarı; karar verme sürecinde kullanılan, konu tabanlı, birleştirilmiş, zamana bağımlı, verilerin sabit olduğu veri topluluğudur. Veri topluluklarının veri ambarı olarak adlandırılabilmesi aşağıdaki dört özelliği taşımalıdır.

1991 yılında ilk kez William H. Inmon tarafından ortaya atılan veri ambarı, birçok veritabanından alınarak birleştirilen verilerin toplandığı depolardır. Veri ambarlarının özelliği kullanıcılara farklı detay düzeyleri sağlayabilmesidir. Detayın en alt düzeyi arşivlenen kayıtların kendisi ile ilgili iken, daha üst düzeyler zaman gibi daha fazla bilginin toplanması ile ilgilidir. Veri ambarları ciddi yatırımlar gerektirmekte ve uygulanması bir yıl veya daha uzun zaman almaktadır

**a) Konu tabanlı:** Veri ambarları, satış verileri, müşteri bilgileri gibi belirli bir konuda veriler içerir.

**b) Birleştirilmiş (Integrated):** Veri ambarı birçok farklı kaynaktan gelen bilgilerin toplanması ile oluşur. Örneğin bir veri ambarı içinde ilişkisel veritabanları, düz metin dosyaları, işlemsel veritabanları bulunabilir.

**c) Zamana Bağımlı:** Veri ambarlarında veriler belirli periyodik aralıklarla eklenir. Veri ambarındaki her bir anahtar yapı tarihsel olarak dizilmiş olmalıdır. Örneğin günlere göre son beş yılın birinci basamak muayene verileri.

**d) Sabit:** Veri ambarında veriler işlevsel veritabanlarında olduğu gibi sürekli güncellenmez. Veri ambarına eklendiği andan itibaren sabit olarak kaydedilir.

Veritabanı ile veri ambarı arasındaki başlıca farklar aşağıdaki gibi açıklanabilir:

Veri ambarı bir işletmenin günlük kullanımda veri depoladığı işlevsel (operational) veritabanından ayrı tutulur. Bu yüzden veri ambarındaki bilgiler güncel değildir.

İşlevsel veritabanlarındaki bilgiler güncel önemlerini yitirdiklerinde veri ambarına gönderilirler.

Veritabanları okuma/yazma amaçlı, veri ambarları ise sadece okuma amaçlı kullanılırlar.

Veritabanları günlük giriş-çıkış işlemleri için kullanılırken veri ambarı uzun süreli veri analizi ve geleceğe yönelik öngörüler elde etme amaçlı kullanılır.

### 2.3.4. Veri Ambarlarının Kullanım Nedenleri

Veri ambarları bir karar verme mekanizması veya diğer adıyla karar destek sistemi olarak kullanılmaktadır. Veri Ambarı üzerinde veri madenciliği, çok boyutlu veri analizi (Online Analytical Processing - OLAP), müşteri ilişkileri yönetimi(CRM),istatistiksel analiz ve raporlama işlemleri gerçekleştirilir[Han, J.;Kamber, M.,2001]. Bu işlemlerin tamamına yakını, işlevsel veritabanları üzerinde de gerçekleştirilebilmesine rağmen, veri ambarı kurma ve kullanmanın temel nedeni her iki sistem için de yüksek performans elde etme isteğidir.

İşlevsel veritabanları, sıralama, arama ve hazır sorguları çalıştırma işlemleri için, veri ambarları ise özel veri organizasyonu, veri analizi ve çabuk erişim için optimize edilirler. Veri ambarı kurulmadığı durumlarda işlevsel veritabanı performansı önemli ölçüde düşerken, yapılan karar destek işlemleri doğruluktan uzaklaş maktadır. Ayrıca, karar verme işlemleri tarihsel veriler gerektirdiği için veri ambarı karar destek sistemleri için vazgeçilmez bir unsurdur.

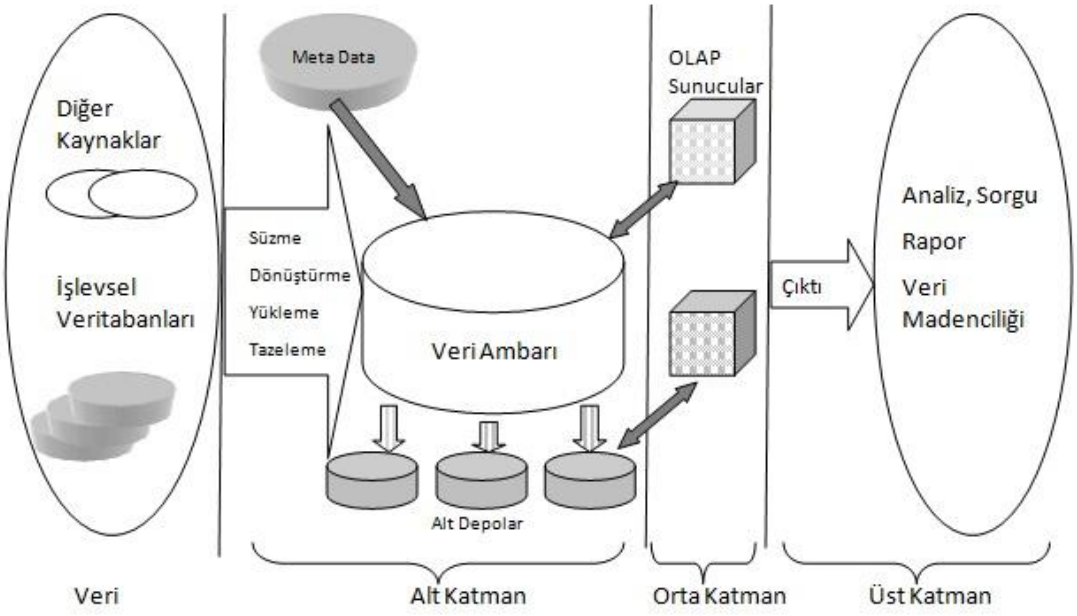
### 2.3.5. Veri Ambarı Mimarisi

Veri ambarı mimarisi J.Han'ın yaklaşımına göre 3 katmanlı bir yapıdan oluşmaktadır.

Sekil1'de görülen bu katmanlar şunlardır:

**a) Alt Katman**

Veri ambarı veritabanı sunucusudur ve genellikle ilişkisel bir veritabanı sisteminden oluşur. İşlevsel veritabanlarında veya dış kaynaklardan gelen veriler uygulama program ara yüzleri (geçit) tarafından seçilir. Geçit programları bir veritabanı yönetim sistemi ile desteklenir. Bu sayede istemci programların sunucu tarafına SQL kodu seklinde sorgu gönderebilmesine olanak sağlanır. Geçit programlarının en bilinenleri Microsoft firmasının ODBC (Open Database Connection) ve OLE-DB (Object Linking and Embedding for Databases) ve Sun Microsystems firmasının JDBC (Java Database Connection) adlı ürünleridir.




Şekil 2.2. Veri ambarını oluşturan katmanlar.[ Han, J.; Kamber, M.2001]

**b) Orta Katman**

Orta katman OLAP sunucudur. Bir alt katmandan gelen veriler OLAP sunucular tarafından analiz yöntemleri kullanılarak raporlama, analiz ve veri madenciliği işlemleri için anlamlı veriler haline getirilir.

**c) Üst Katman**

Üst katman istemciden oluşur, sorgulama ve raporlama araçları, analiz araçları ve veri madenciliği araçları içerir.

Veri ambarlarında yer alan bilgiler, bilgilerin kullanılacağı alanlara göre ayrı alt depolara(data-mart) dağıtılabilirler.

### 2.3.6. Veri Tabanlarında Bilgi Keşfi Aşamaları

Veri madenciliği, veritabanlarında bilgi keşfi (VTBK) (KDD–Knowledge Discover in Databases) işleminin temel bileşenlerinden biridir. Bununla beraber VTBK sadece veri madenciliğinden ibaret değildir.

**a. Veri Önişleme**

Nitelikler, sayısal,nominal ya da katarlar şeklinde değer alabilirler. Bu aşamada öncelikle veriler içindeki gürültüler, tutarsızlık ve düzensizlikler giderilir. Verilerin analize uygun bir yapıya getirilmesi işlemine veri önişleme denir.

Veri önişleme adımı bir veri madenciliği çalışmasının oldukça büyük bir kısmını kapsar ve analizin doğru sonuçlara ulaşmasında ve efektif bir şekilde uygulanmasında büyük öneme sahip olup problem alanında bilgi sahibi olmayı gerektirir.

Bu adım veri madenciliği aşamalarının ilk ve en uzun basamağını oluşturur. Veri temizleme, veri birleştirme, veri dönüşümü ve veri azaltma işlemlerini kapsar.

**b. Veri Temizleme**

Kullanıcı hataları, program hataları, bazı otomatize edilebilecek işlemleri kullanıcıya bırakma, veri girişinin önemsenmemesi gibi sebeplerle veri kümelerinde eksik ya da gürültülü veriler oluşabilir. Veri üzerindeki bazı nitelikler yanlış değer taşıyabilecekleri gibi, eksik, geçersiz veriler de olabilir. Veriler üzerinden faydalı ve doğru sonuçlar çıkarabilmek için bu tip bilgilerin düzeltilmesi ya da göz ardı edilmesi gerekir. Veri temizleme basamağında bu tip veriler tamamlanır, ayıklanır ya da tutarsız veri varsa bu tutarsızlıklar belirli mantıksal işlemlerle düzenlenir.

Eksik nitelik değeri taşıyan veriler, göz ardı edilebilir, rastgele değerlerle doldurulabilir.

Verilerimiz içerisinde tutarsız, ya da gerçeğe aykırı olan gürültülü veri diye adlandırılan hatalı veriler de bulunabilir. Gürültülü veriler, bölmeleme, demetleme ya da eğri uydurma gibi metotlarla düzeltilebilir.

Bölmeleme işlemi eldeki verinin sıralanarak eşit bölmelere ayrılması ve her bölmenin kendisine ait ortalama ya da uç değerlerle ifade edilmesidir. Böylece verilerdeki hata miktarlarının minimize edilmesi amaçlanır

Demetleme ile benzer veriler aynı demette olacak şekilde gruplanır ve dışarıda kalan veriler göz ardı edilir. Böylece muhtemel yanlış ölçümler ayıklanmış olur. Eğri uydurma metodu ile ise nitelik değeri diğer niteliklere bağlı olarak belli bir fonksiyona uydurulur. Bu fonksiyon kullanılarak nitelik değerindeki tutarsızlıklar giderilir.

**c. Veri Birleştirme**

Bazı durumlarda birçok veri kaynağından yararlanarak veri kümemizi oluşturmamız gerekir. Veri birleştirme denilen bu işlemde farklı kaynaklardan gelen veriler aynı veri kümesi altında birleştirilir. Farklı kaynaklarda aynı nitelik için farklı değerler, ölçü birimleri ya da derecelendirmeler kullanılmış olabilir. Bu durumlarda nitelik değerlerini birleştirirken dönüşüm yapmak gerekir. Farklı kaynaklarda aynı nitelikler farklı nitelikmiş gibi ele alınmış olabilir ya da birleştirme sonucunda gereksiz veriler oluşabilir. Bu tip niteliklerin belirlenmesi, gereksiz verilerin ayıklanması gerekir.

**d. Veri Seçme ve Dönüştürme**

Verilerde bazı nitelik tipleri uygulanacak olan algoritmaya uygun olmayabilir ya da veri nitelikleri belirleyici olmayabilir. Veri dönüşümü yapılarak nitelikler algoritmaya uygun hale getirilir ve nitelikler daha belirleyici olacak şekilde dönüştürülebilir. Bunun için normalizasyon ya da nitelik oluşturma işlemleri yapılabilir.

Bu aşamada, veri madenciliğinin sağlıklı yapılabilmesi için veriler üzerinde bir takım işlemler yapılır. Bu işlemler:

- Veri madenciliği konusu ile ilgili bilgi seçimi.
- Madencilik yapılacak veri türünün belirlenmesi.
- Veriler arasında hiyerarşik yapı ve genellemelerin belirlenmesi.
- Veri madenciliği sonunda bulunacak bilgi için yenilik ve ilginçlik ölçümü yöntemlerinin belirlenmesi.

- Veri madenciliği sonunda bulunacak veri için sunum ve görselleştirme araçlarının belirlenmesi.

**e. Veri Azaltma**

Analiz edilecek olan verinin aşırı büyük olması, uygulanacak olan algoritmanın daha uzun bir sürede tamamlanmasına ve aslında sonucu etkilemeyecek gereksiz işlemlere sebep olur; ayrıca bazı algoritmalar belirli tip veriler üzerinde çalışır, bu tipte olmayan verilerin göz ardı edilmesi ya da dönüştürülmesi gerekir. Bu sebeple veri önişleme aşamasında uygulanacak olan bir diğer işlem de sonucu etkilemeyecek bir şekilde gereksiz olan bilgilerin silinmesi, birleştirilmesi ya da diğer bazı yöntemlerle daha anlamlı ve algoritmaya uyumlu hale getirilmesidir. Nitelik birleştirme, nitelik azaltma, veri sıkıştırma, veri küçültme, veri ayrıştırma ve kavram oluşturma gibi yöntemlerle eldeki veri, sonucu değiştirmeyecek şekilde daha verimli bir hale getirilmektedir.

Nitelik seçme, problem alanına yönelik bilgiyi değerlendirerek yapılabileceği gibi istatistiksel yöntemlerle, karar ağaçlarıyla ya da bilgi kazancı değerleriyle tespit edilebilir. Veri sıkıştırma, büyük verinin sıkıştırma algoritmalarıyla boyutunu küçültmeyi, böylece veri saklamayı ve veri erişimini hızlandırmayı amaçlar. Bu yöntemin verimli olması için uygulanacak olan algoritmanın sıkıştırılmış veri üzerinde çalışabilmesi gerekir.

**f) Örüntü Değerlendirme (Pattern Evaluation):**

Bu aşamada belirlenen ilginçlik (interestingness) ölçüm yöntemleri kullanılarak veri madenciliği ile bulunan verilerin ne kadar ilginç ve yararlı olduğu tespit edilir.

**g) Bilgi Sunumu(Knowledge Presentation):**

Çeşitli görselleştirme ve raporlaştırma araçları kullanılarak bulunmuş olan veriler ilgili kullanıcılara sunulur.

VTBK süreci defalarca tekrar ve aşamalar arası atlamalar ve ileri geri hareketler içerebilmektedir. Günümüzde çoğunlukla veri madenciliği aşamasına odaklanılmakta,

fakat diğer tüm aşamalar VTBK işleminin bütünlüğü açısından en az veri madenciliği kadar önemlidir [Fayyad, U.M.; Piatesky-Shapiro, G.;1994].

# BÖLÜM 3.

# VERİ MADENCİLİĞİ TEKNİKLERİ

Veri madenciliği teknikleri eldeki veri türüne ve elde edilen sonuçların kullanım amacına göre farklılıklar gösterir. Temelde veri madenciliği iki kategoride incelenir

- Tanımlayıcı (Descriptive)
- Öngörüsel (Predictive)

Tanımlayıcı veri madenciliği, veritabanındaki verinin genel karakterini, mevcut durumu ortaya çıkarmaya yönelik yöntemleri ön plana çıkarır. Öngörüsel veri madenciliği ise verileri geleceğe yönelik tahminler yapma, sonuç çıkarma amaçlı işlemlerde kullanır.

Veri madenciliği teknikleri kullanıldıkları veri yapılarına ve keşfedebildikleri örüntü biçimlerine göre kategorilere ayrılır. Birçok kaynak veri madenciliği teknikleri için farklı gruplandırmalar yapmıştır. Bunlardan en yaygın kabul göreni J.Han'ın ortaya sürdüğü kategorilerdir. J.Han kategorilerini kullanan kaynaklar bile, hangi algoritmanın hangi kategoriye ait olduğu konusunda net görüş birliğine sahip değildir. Bu kategorileri aşağıdaki gibidir:

- Tanımlama ve Ayrımlama (Characterization and Discrimination)
- Birliktelik Analizi (Association Analysis)
- Sınıflandırma ve Öngörü (Classification and Prediction)
- Kümeleme Analizi (Cluster Analysis)
- Sıra dışılık (istisna) Analizi (Outlier Analysis)
- Evrimsel Analiz (Evolution Analysis)

### 3.1. Tanımlama ve Ayrımlama

Veriler gösterdikleri ortak özelliklere göre genelleştirilmiş sınıflara ayrılabilirler. Bir firma müşteri portföyünü alışveriş ortalaması belirli bir miktardan daha yüksek olan müşterileri "zengin", diğerlerini ise "orta halli" ya da "fakir" olarak tanımlayabilir. Bu tür genellemeler veri kümesinin elemanlarının ortak özellikleri ya da veri kümesinin diğer veri kümleri ile olan farklılıklarını yansıtacak şekilde yapılabilmektedir.

#### 3.1.1 Tanımlama (Characterization)

Bir veri kümesinin elemanlarının genel özelliklerini özetlemek amaçlı kullanılır. Örneğin bir alışveriş merkezinde bu yıl satısı oranı %25'in üzerinde artan mallar ifadesi bir Tanımlama işlemidir.

#### 3.1.2 Ayrımlama (Discrimination)

Bir veri kümesinin diğer bir veri kümesinden farklarını ortaya çıkarma işlemidir. Örneğin bu yıl satış oranı %10 artan mallar ile satış oranı %15 azalan malların karşılaştırılması Ayrımlama tabanlı veri madenciliğidir.

Her iki tür veri madenciliği yöntemi birbirine çok benzer yöntemler kullanırlar.

Ayrıca her iki yöntemle elde edilen sonuçlar pasta grafiği, sütun grafiği, eğriler ve çok boyutlu küpler ile sunulurlar.

### 3.2 Birliktelik Analizi

Birliktelik analizi bir veri kümesinde kendiliğinden, sıklıkla gerçeklesen, birlikte ya da aynı süre içinde alınma, yapılma, oluşma gibi etkileri keşfetme temeline dayanır.Bu yöntem bankacılık işlemlerinin analizinde ya da pazar sepeti analizi

yönteminde yaygın olarak kullanılır. Pazar sepeti analizi, bir alışveriş sırasında veya birbirini izleyen alış verişlerde müşterinin hangi mal veya hizmetleri satın alma eğiliminde olduğunun belirlenmesiyle müşteriye daha fazla ürün satılması yollarından biridir [Akpınar,H,2000].

Pazar sepeti analizi ile örneğin müşteriler bira satın aldığında %75 ihtimalle cips de alırlar seklinde bir ilişki ortaya çıkarılabilir. Bunun sonucunda bira ile cips yan yana raflara yerleştirilebilir veya bira alanlar cips aldığında cips fiyatında indirim yapılacak şekilde kampanyalar oluşturularak satışlar arttırılabilir.

Birliktelik analizi yalnızca mal ve hizmetlerin birlikte satın alınması için değil aynı zamanda hangi koşulları sağlayan müşterilerin hangi ürünleri alacağı hakkında da çözümler getirmektedir. Örneğin bir banka kredi kartı kayıtları incelendiğinde yasları 20 ile 29 arasında değişen müşterilerden, gelirleri 700 milyon ile 900 milyon TL arasında değişen müşterilerin bilgisayar satın aldıkları görülmüştür. Bu kural, birliktelik analizi yönteminde söyle ifade edilir:

Yaş(X , “20…29”) ^ Gelir(X , “700…900”) alır(X , “bilgisayar”)

### 3.3. Sınıflandırma ve Öngörü

Sınıflandırma işlemi insan düşünce yapısına en uygun veri madenciliği yöntemidir. İnsanoğlu çevresindeki nesneleri ve olayları daha iyi anlamak ve başkalarına anlatabilmek için hemen her şeyi sınıflandırma eğilimindedir. Örneğin, insanları davranışlarına göre, hayvanları türlerine göre, evleri görünüşlerine göre sınıflandırmaktadır.

Veri madenciliğinde sınıflandırma, eldeki mevcut verileri önceden belirlenen bir özelliğe göre sınıflara ayırmak ve yeni eklenecek verilerin hangi sınıfa dahil olacağını tayin etme işlemdir. Diğer bir deyişle, yeni karşılaşılan bir girdinin hangi sınıfa dahil olacağına karar verme işlemidir.

Sınıflandırma işlemine, bankaların kredi başvurularını düşük, orta ve yüksek riskli olarak sınıflandırması, bir okulda yeni gelen öğrencilerin hangi sınıfta eğitim görmesi gerektiğinin belirlenmesi örnek olarak verilebilir.

Öngörü işlemi sınıflandırma işlemine çok benzer. Ancak öngörü işleminde sınıflandırma, gelecek için tahmin edilen belirli bir davranışa ya da belirli bir değere göre yapılır. Öngörü işleminde yapılan sınıflandırmanın doğru olup olmadığını test etmenin tek yolu "bekle ve gör" prensibidir [Han, J.-Kamber, M., Morgan,2000].

Öngörü işlemine örnek olarak deprem tahmini, bir turizm şirketi müşterilerinden hangilerinin bu yaz yurt dışında tatil yapmak isteyeceğinin belirlenmesi verilebilir.

Sınıflandırma ve Öngörü işleminde Karar Ağaçları (Decision Tree), Yapay Sinir Ağları (Neural Networks), K-en yakın komsu (K-Nearest Neighbour), Genetik algoritmalar, Naive Bayesian sınıflama, Bellek Tabanlı Nedenleme (Memory Based Reasoning) yöntemleri kullanılır.

Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri kurabilen iki veri analiz yöntemidir .[Han. J,Kamber. M,2001]. Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır .[Han. J,Kamber. M,2001]. Örneğin, bir sınıflama modeli banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, regresyon modeli geliri ve mesleği verilen potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir.

Sınıflama ve regresyon modellerinde kullanılan baslıca teknikler şunlardır [Han, J.-Kamber, M., Morgan,2000]:

1 - Karar Agaçları (Decision Trees)

2- Yapay Sinir Agları (Artificial Neural Networks)

3- Genetik Algoritmalar (Genetic Algorithms)

4- K-En Yakın Komsu (K-Nearest Neighbor)

5- Bellek Temelli Nedenleme (Memory Based Reasoning)

6- Naive-Bayes

### 3.3.1 Karar Ağaçları (Decision Trees)

Karar ağaçları, veri madenciliğinde kuruluşlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir.

Karar ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde, tahmin edici bir tekniktir . Ağaç yapısı ile, kolay anlaşılabilen kurallar yaratabilen, bilgi teknolojileri işlemleri ile kolay entegre olabilen en popüler sınıflama tekniğidir.

Karar ağacı karar düğümleri, dallar ve yapraklardan oluşur .[Han. J,Kamber. M,2001]. Karar düğümü, gerçekleştirilecek testi belirtir. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur.

Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa, o daim sonucunda bir karar düğümü oluşur. Ancak daim sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden baslar ve yukarıdan aşağıya doğru yaprağa ulasana dek ardışık düğümleri takip ederek gerçekleşir.

Karar ağacı tekniğini kullanarak verinin sınıflanması iki basamaklı bir işlemdir [Han, J.-Kamber, M., Morgan,2000].

İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından analiz edilir.

Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır.

Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karşılaştırılır. Eğer modelin doğruluğu kabul edilebilir bir değer ise model, sınıfı bilinmeyen yeni verileri sınıflama amacıyla kullanılabilir.

Örneğin, bir eğitim verisi incelenerek kredi duruma sınıfını tahmin edecek bir model oluşturuluyor. Bu modeli oluşturan bir sınıflama kuralı

IF yas = "41...50" AND gelir = yüksek THEN kredi durumu = mükemmel

şeklindedir. Bu kural gereğince yası "41...50" kategorisinde olan (yası 41 ile 50 arasında olan) ve gelir düzeyi yüksek bir kişinin kredi durumunun mükemmel olduğu görülür.

Oluşturulan bu modelin doğruluğu, bir test verisi aracılığı ile onaylandıktan sonra model, sınıfı belli olmayan yeni bir veriye uygulanabilir ve sınıflama kuralı gereği yeni verinin sınıfı "mükemmel" olarak belirlenebilir.

Tekrarlamak gerekirse bir karar ağacı, bir alandaki testi belirten karar düğümlerinden, testteki değerleri belirten dallardan ve sınıfı belirten yapraklardan oluşan akış diyagramı seklindeki ağaç yapısıdır. Ağaç yapısında ki en üstteki düğüm kök düğümüdür.

Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi, çeşitli durumların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması, gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması, sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması, kategorilerin birleştirilmesi gibi alanlarda karar ağaçları kullanılmaktadır

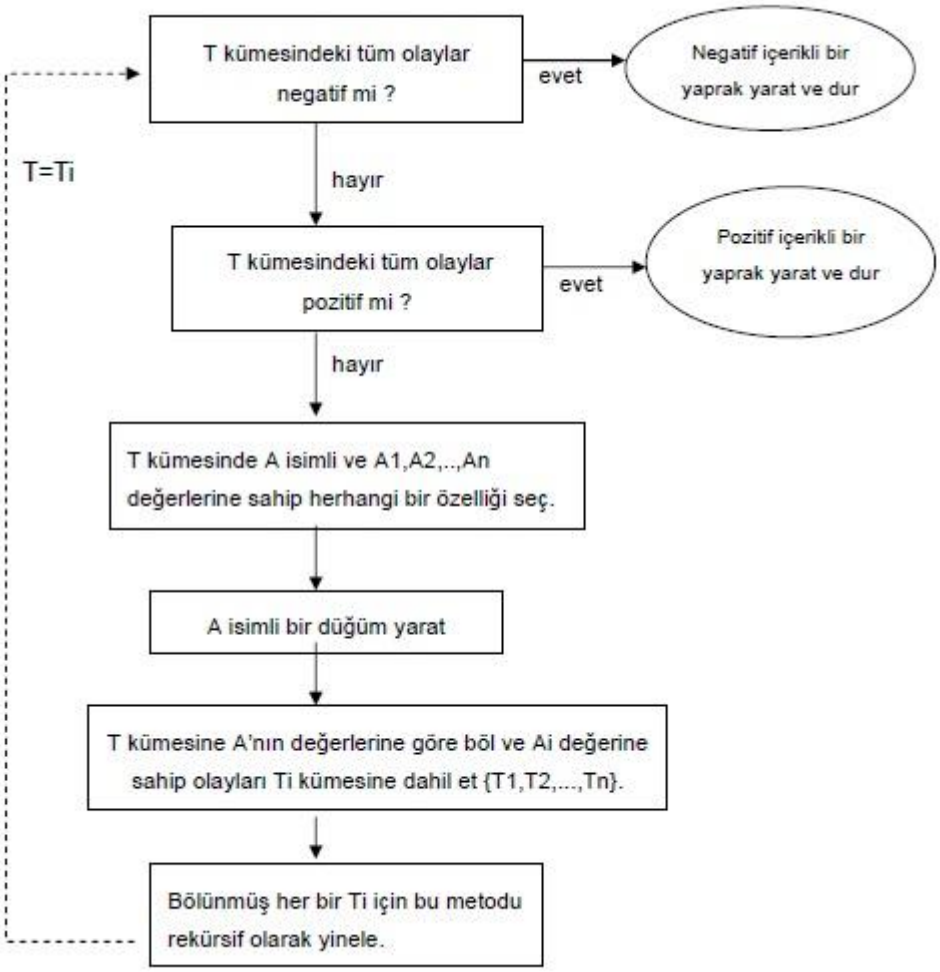
Karar ağaçları, hangi demografik grupların mektupla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip olduğunun belirlenmesi (Direct Mail), bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi (Credit Scoring), geçmişte isletmeye en faydalı olan bireylerin özelliklerini kullanarak ise alma süreçlerinin belirlenmesi, tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi, hangi değişkenlerin satışları etkilediğinin belirlenmesi, üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesi gibi uygulamalarda kullanılmaktadır [Akpınar,2000].

### 3.3.2.Karar Ağacı Oluşturma

Ağacın oluşturulmasına yönelik olarak çeşitli ağaç oluşturma metotları vardır. Ağacı oluşturmadaki en önemli kriter belli özelliklere göre toplanmış, güvenilir ve yeterli sayıda olay örneklerinin varlığıdır. Bu iki faktör ağaç oluşturmanın temelini oluşturur. Ağaç oluşturmadaki en önemli adım ise böl ve elde et aşamasıdır.

#### 3.3.2.1 Böl ve Elde Et (Divide and Conquer)

Bu metot Hunt'ın uyguladığı bir metottur. Bu metotta örnek uzay T ve sınıflar pozitif ve negatif olsun. Bu durumda bir KA oluşturma Şekil.3.1'deki gibi olacaktır.




Şekil 3.1. Hunt'ın ağaç oluşturma metodu

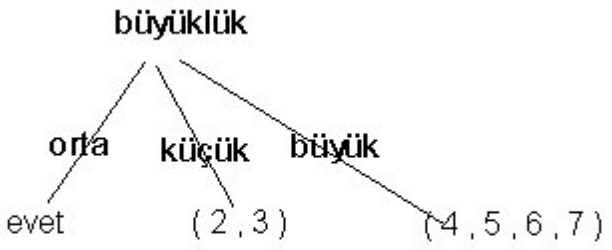
Bu algoritma en temel ağaç yaratma algoritmasıdır.Bu algoritmanın geliştirilmesine yönelik olarak çeşitli çalışmalar vardır. Bunlardan en önemlileri tek değişkenli karar ağaçları için Quinlan'ın 1983'te geliştirdiği ID3 algoritması ve ardından yine Quinlan'ın geliştirdiği C4.5 algoritmasıdır. Çok değişkenli karar ağaçları için ise Breiman'ın geliştirdiği CART algoritması vardır.

Örnek olarak bir maddenin bizim için uygun olup olmadığına bakan bir çalışma ele alınsın. Bu maddenin büyüklük, renk ve şekil gibi özellikleri olsun ve 7 adet örnek olay olsun. Bu örnekler evet-hayır olarak ikili sınıflandırılmış olarak Tablo3.1 'de gösterilmiştir.

Tablo 3.1 Örnek bir olay kümesi

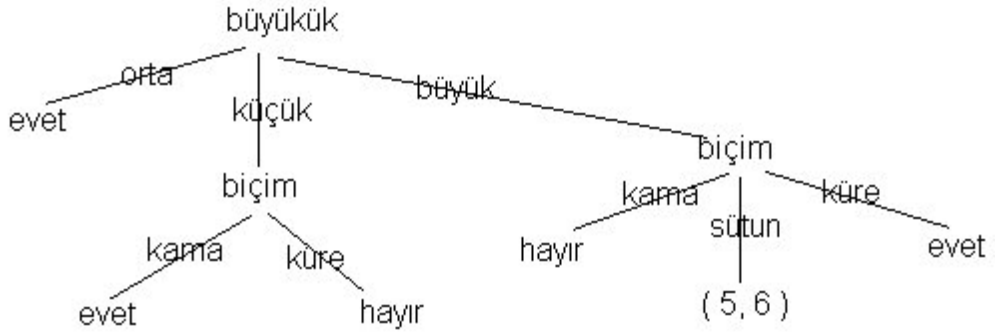
| | Büyüklük | Renk | Biçim | Sonuç |
|---|---|---|---|---|
| 1 | Orta | Mavi | Tuğla | Evet |
| 2 | Küçük | Kırmızı | Kama | Hayır |
| 3 | Küçük | Kırmızı | Küre | Evet |
| 4 | Geniş | Kırmızı | Kama | Hayır |
| 5 | Geniş | Yeşil | Sütun | Evet |
| 6 | Geniş | Kırmızı | Sütun | Hayır |
| 7 | Geniş | Yeşil | küre | Evet |

1 den 7'e kadar sıralanmış örnekler rasgele seçilen özelliği ile alt kümelere bölünsün. Şekil 3.2'de de gösterildiği gibi büyüklüğün muhtemel üç çeşit değeri olur ve üç tane dal oluşur.




Şekil 3.2. Tablo 1'in büyüklük sınıfına göre bölünmesi

Bu noktada büyüklük=küçük dalına ve büyüklük=büyük dalına yönelik olarak aynı işlem gerçekleştirilsin. Bölme işlemi yine rasgele seçilen biçim özelliğine göre yapılırsa Şekil 3.3 'teki ağaç oluşur.




Şekil 3.3. Şekil.3.2'deki ağacın bölünmüş kümeleri biçim özelliğine göre tekrar bölünmesi sonucu oluşan ağaç

### 3.3.2.2.ID3 Algoritması

HUNT'ın algoritmasındaki en önemli eksiklik özelliklerin rasgele seçilmesidir. Oysa ki bu seçim sırasında bilgi kazancı en yüksek olan özellik dikkate alınırsa oluşturulan ağaç o kadar sade ve anlaşılır olacaktır.

Buna yönelik olarak Quinlan entropy kurallarını içeren bilgi teorisini kullanmıştır. Shannon ve Weaver'ın Bilgi Teorisinde temel olarak kaynak, mesaj ve alıcı vardır. Bu sistemde bilgi, mesaja bakılarak değil de, alıcıya bakılarak elde edilir. Alıcı mümkün olan mesaj uzayı bilgisine ve bu mesajların olasılıklarına sahiptir. Ağaçlardaki bazı düğümler ve bu düğümlerdeki kararlar anlamsız ve gereksiz olabilmektedir. Ancak bu tip düğümler de negatif–pozitif olay balanslarına sahiplerdir. İşte bu şekilde sınıflandırma yapılabilmektedir. [Kocabaş, Ş.1991]

Örnek olarak X düğümünde 5 pozitif ve 3 negatif olay var. Bu noktada yapılacak bir sınıflandırmanın pozitif olasılığı 5/8'dir, negatif olasılığı 3/8'dir. İşte bu olasılıksal sınıflandırmayı türetme yeteneğinin anlamı şudur: Doğru olarak sınıflandırılmış bir örneğin söylediği mesajın bilgi içeriği artık hesaplanabilir.

Öyle ki bir tablonun sonuçları mesaj olsun ve mesajlar iki değere sahip olsunlar. Bu değerlerle birlikte p bilgisi pozitif olasılığını, q bilgisi negatif olasılığını gösterir. Bu

iki değerin toplamı zaten 1 (p+q) olmak zorundadır. Doğru sınıflandırma veren bir mesajın bilgi içeriği

I (p,n) = -p log2 p - q log2 q (2.1)

şeklinde hesaplanır.

Bu formül genel bilgi içerik formülünün özel bir durumudur. Çünkü özel olarak iki olasılık mevcuttur: pozitif ve negatif.

{A1,A2,...,An} değerlerine sahip A özelliği ağacın bölünmesi için kullandığında,T kümesi {T1,T2,...,Tn} şeklinde bölünecektir. Bu bölümleme de T kümesindeki A özelliğinin Ai olduğu bölgelere Ti densin. Bu kümedeki pozitif olayların sayısını pi temsil etsin, negatif olayların sayısını ni temsil etsin. Bu durumda Ti alt ağacı için beklenen bilgi gereksinimi ise I(pi,ni) olur. T ağacı için beklenen bilgi gereksinimi tüm Ti ağaçlarının beklenen bilgi gereksinmelerinin ağırlıklı ortalamalarının toplamı olur ve aşağıdaki gibi hesaplanır.

$$E(A) = \sum_{i-1}^{n} \frac{pi + ni}{p + n} I(pi, ni)$$

Dolayısı ile A özelliği üzerinden sağlanan bilgi kazancı

Bilgi kazancı(A)= I(p,n) – E(A) şeklinde ifade edilir.

Bilgi gereksinimi ve bilgi kazancı ID3 algoritmaları için iki önemli kavramdır. Belirleyici bir sınıflandırma için bilgi ihtiyacı aslında doğru sınıflandırmayı sağlayan mesajın bilgi içeriğinden başka bir şey değildir. Buna yönelik olarak yaratmak istenilen karar ağaçlarının amacı doğru soruları sormasıdır. Ve sonunda öyle bir noktaya ulaşılmalı ki, bu noktanın karar için bilgi gereksinimi 0 olsun. İşte bu noktada ID3 algoritmasının yaptığı şey, ağacı doğru kurmaktır. Kurulu karar ağacının her seviyesinde geriye kalan bilgi gereksinimi (remaining information required ) minimize edilir.

Bu bilgiler ışığında Tablo2.1'deki örnek ele alınsın. Bu olayların hepsi birden evet ya da hepsi birden hayır olamadıklarından bilgi kazancı en yüksek olan özellikten başlayarak bölümleme işlemi gerçekleştirilir. Örnek uzayda 4 adet pozitif olay olduğunda bir olayın pozitif gelme olasılığı 4/7=0.57'dir. Negatif gelme olasılığı 3/7=0.43'dür. Bundan dolayı doğru bir sınıflandırma için gerekecek bilgi kazancı

$$-(0.57 \times \log_2 0.57) - (0.43 \times \log_2 0.43) = 0.99$$

Şimdi her bir özellik için bilgi gereksinmeleri hesaplansın. Büyüklük özelliği için küçük, orta ve büyük olmak üzere üç tip değer vardır. Büyük değeri kümenin {4,5,6,7} elemanlarını kapsamaktadır. Bu küme içersinde 2 evet ve 2 hayır sınıfı bulunduğundan ve p,n = 2/4=0.5 olduğundan gereken bilgi kazancı

$$-(0.5 \times \log_2 0.5) - (0.5 \times \log_2 0.5) = 1$$

Aynı işlem küçük değeri {2,3} elemanlarını içermektedir. Bu kümede 1 evet ve 1 hayır sınıfı vardır. Bu durumda gereken bilgi kazancı

$$-(0.5 \times \log_2 0.5) - (0.5 \times \log_2 0.5) = 1$$

Orta değeri için hesap edildiğinde ilgili kümede sınıflardan sadece bir tanesi olduğundan sonuç 0 çıkar.

Şu anda büyüklük özelliği için beklenen bilgi gereksinimini hesap edilebilir. Bütün bilgi gereksinim sonuçları ilgili özellik değerlerinin orantısıyla çarpılarak toplanır ve

$$(1 \times 4/7) + (1 \times 2/7) + (0 \times 1/7) = 0.86$$

olur. Büyüklük özelliği için beklenen bilgi kazancı mevcut bilgi ihtiyacından beklenen bilgi ihtiyacı çıkarılarak hesaplanır;

$$0.99 - 0.86 = 0.13$$

Aynı işlemi renk ve biçim için yapılarak işlem tamamlanır. Bilgi kazancı renk için 0.52 ve biçim için 0.7 bulunur. Bu koşullar altında biçim özelliği en yüksek bilgi

kazancına sahip özellik olur. Buna göre ağaç tekrar oluşturulduğunda Şekil 3.4'deki gibi daha sade bir şekil alır.[YILDIRIM.2003]


Şekil 3.4.. ID3 ile oluşturulmuş KA

### 3.3.2.3.C4.5 Karar Ağacı Eğitme Algoritması

ID3 algoritmasında bazı eksiklikler ve sorunlar vardır. Bunlar aşağıda anlatılmaktadır. Bu sorunlar yine Quinlan'ın geliştirdiği C4.5 algoritmasıyla giderildi. C4.5 Algoritması ID3 algoritmasının bütün özelliklerini kendine miras alarak oluşturulmuş bir algoritmadır. Yukarıda bahsedilen tüm içeriğin üzerine yeni kavramlar eklenmiştir. Bölünme-Dağılma Bilgisi (Split-Info), özelliklerin kayıp değerleriyle baş edilmesi, sayısal özellik değerlerinin hesaplara katılması bu başlıklardan en önemlileridir. Bu başlıklardan biri olan Bölünme-dallanma bilgisi ile başlanabilir. .[YILDIRIM.2003]

#### 1. Bölünme-Dallanma Bilgisi (Split Information)

Bir kategorik özelliğin olası değer çeşitliliği ne kadar yüksek olursa o özelliğin bilgi kazancı gereksiz bir şekilde yüksek çıkar ve bu durum ağacın doğruluğunu kötü bir şekilde etkiler. Bu tip özellikler işe yaramadıkları gibi bilgi kazancı yüksek özelliklerin de önüne geçip veride gizlenmiş kuralların çıkarılmasına engel teşkil ederler. Yukarıda

ele alınan örnekte değer çeşitliliği en fazla dörttü. Şimdi bu veriye çeşitliliği çok yüksek bir özellik eklenirse nasıl bir sonuç çıkacaktır? Öyle ki Tablo1'deki 1'den 7'e kadar verilmiş olan etiket numaraları bir özelliğe karşılık gelsin. Bu aşamada bu özelliğin bilgi kazancı hesap edilsin; 1:evet, 2:hayır, 3:evet,4:hayır,5:evet,6:hayır ve 7:evet şeklinde her bir özellik değeri için bir tane sonuç elde edilecektir.

Bu durumda 1 değeri için gereken bilgi kazancı -(1 x log2 1/1 ) - (0 x log2 0/1) = 0 çıkacak ve aslında tüm değerleri için gereken bilgi kazancı 0 çıkacaktır. Etiket özelliği için ortaya çıkan bilgi kazancı ise

0 x (1/7) +0 x (1/7) +0 x (1/7) +0 x (1/7) +0 x (1/7) +0 x (1/7) +0 x (1/7) =0

olacaktır. Genel bilgi teorisi için bu sonucu mevcut bilgi gereksiniminden çıkarmak gerekecek. Bu durumda etiket özelliği için bilgi kazancı

0.99 – 0 = 0.99

olacaktır. Bu sonuç diğer sonuçlar arasındaki en yüksek sonuçtur ve buna göre bir ağaç oluşturulduğunda Şekil 3.5'deki ağaç yapısı oluşur.


Şekil 3.5. Tablo1'in etiket özelliğine göre bölümlenmesi

### 3.3.3 Sayısal Özellikler

Veri kümesinde iki tip veri vardır; Nominal (kategorik) ve sayısal. Nominal daha önceki bölümlerde kullanılan veri tipleridir. Örnek olarak renk özelliği mavi, kırmızı ve sarı ile, büyüklük özelliği büyük, orta ve küçük ile ifade edilir. ID3 algoritması daha önce sadece nominal değerlere sahip veri tipleri ile işlemler yapabilirdi. Ancak bu

algoritmanın varisi olan C4.5 algoritması ise bu tip veriler için de bir yöntem geliştirmiştir.

İlk bakışta sayısal özelliklerle uğraşmak ve onların bilgi kazancını hesaplamak oldukça zor gelebilir. Ancak bu iş o kadar da zor değildir. Yapılması gereken iş sadece bu özelliğin sayısal değerleri arasında uygun eşik değerini bulmaktır. Bu eşik değeri bulunduktan sonra ikili bir bölünme ile veri kümesi bölünebilir; Bu eşik değerinden büyük veriler ile bu eşik değerinden küçük veriler. Bu anlamda algoritma çok sade bir şekilde açıklanabilir. Öncelikle tüm sayısal değerler küçükten büyüğe sıralanır.Bu sıra {v1,v2,...,vm} ile ifade edilsin. Bu durumda seçilen eşik değeri vi ve vi+1 arasında olursa {v1, v2,...,vi} ile {vi+1, vi+2,...,vm} gibi iki grup ortaya çıkar. Buradan da görülüyor ki m-1 adet eşik değeri seçilebilir. Bu seçim işlemi için olası bütün eşit değerleri

$$\frac{v_i + v_{i+1}}{2}$$

formülü ile hesaplanır. Bu yapıyla sanki söz konusu özellik büyük-küçük değerleri olan nominal bir özelliktir. Bu anlayışla nominal değerlere uygulanan bilgi oranı formülü tüm eşik değerleri için uygulanır ve bilgi kazanımı en iyi olan eşik değeri söz konusu özelliğin eşiği olarak kabullenilir. Eğer en iyi eşik değeri e ise ve söz konusu özelliğin sayısal değerleri {v1, v2,...,vn} kümesi ile ifade ediliyorsa, bu kümedeki vi < e koşulunu sağlayan elemanlar küçük kategorisine ve vi> e koşulunu sağlayan elamanlar büyük kategorisine dahil edilir.

Böyle bir bölümleme işlemi bir çok veri üzerinde denendiğinde olumlu sonuçlar vermiştir. Ancak eksik olan yanı sadece ikili bir bölme işlemi gerçekleştirmesidir. Oysa ki bu tip bir bölümleme üçlü ya da daha fazla olursa veri yığını içersine gizlenmiş olan kuralları bulma olasılığı daha çok artar. .[YILDIRIM.2003]

### 3.3.4. Yapay Sinir Agları (Artificial Neural Networks)

Yapay sinir ağları ile sınıflandırmanın işleyiş yapısı, çıktı katmanına ulaşabilmek için ağırlıkların hesaplanmasına dayanır. Eğitim veri kümesi üzerinde hesaplanan ağırlıklar, test veri kümesi üzerinde kullanılarak öğrenmenin ne kadar gerçekleştiği belirlenir. Elde edilen ağırlıkların etkinliği doğrulanamazsa ağırlıklar üzerinde düzeltme ve yeniden hesaplama işlemleri gerçekleştirilir. Öğrenme süreci tamamlandığında ise ağırlıklar yardımıyla yeni bir verinin hangi sınıfa ait olduğu belirlenebilir. Yapay sinir ağlarında öğrenme süreci uzun sürse de oldukça duyarlı sınıflandırmalar yapabilmektedir [TAPKAN,ÖZBAKIR,BAYKASOĞLU,2011]

İnsan beyninin hesaplama mantığı baz alınarak oluşturulmuş (yapay) sinir ağları, karar ağaçları gibi yeni jenerasyon veri madenciliği yöntemlerindendir. Girdi ve çıktı arasında, küçük hesaplama birimlerinden elde edilen sonuçları birleştirerek sonuçlandıran bir modelleme yöntemidir. Karar ağaçları uygulama, anlama ve yorumlama açısından ne kadar kolaysa, sinir ağları da o derece zordur. Yalnızca model oluşturma, sonuçları yorumlama aşamasının ötesinde; doğru bir model kurabilmek için ağın eğitimindeki dengenin önemi oldukça büyüktür. Fazla eğitilmiş bir ağ, önceden gözlenmemiş bir gözleme yönelik tahmin kabiliyetini yitirirken; az eğitilmiş bir ağ ise yanlış tahmin verebilmektedir [KOYUNCUGİL,ÖZGÜLBAŞ,2009].

Genel anlamıyla yapay sinir ağları, beynin bir işlevini yerine getirme yöntemini modellemek için tasarlanan bir sistemdir. Yapay sinir ağı, yapay sinir hücrelerinin birbirleri ile çeşitli şekilde bağlanmasından oluşur ve genellikle katmanlar şeklinde düzenlenir . Tek tabaka ya da tek eleman içeren bazı başarılı ağlar oluşturulabilmesine rağmen çoğu uygulamalar en az üç tabaka içeren ağlara ihtiyaç duymaktadır [Şentürk, 2006]. Bunlar:

1. Girdi tabakası: Dışarıdan veri alan nöronları içermektedir.

2. Çıktı tabakası: Çıktıları dışarı ileten nöronları içeren tabakalardır.

3. Gizli tabaka: Girdi ve Çıktı tabakaları arasında birden fazla gizli

tabaka bulunabilir.

Bu gizli tabakalar çok sayıda nöron içermektedir ve bu tamamen nöronlar ağ içindeki diğer nöronlarla bağlantılıdırlar.

İleri beslemeli çok katmanlı modeller en sık kullanılan yapay sinir ağı modelleridir. Bundan başka geri beslemeli yapay sinir ağı modelleri de mevcuttur. İleri beslemeli sinir ağlarında, hücreler katmanlar şeklinde düzenlenir ve bir katmandaki hücrelerin çıktıları bir sonraki katmana ağırlıklar üzerinden giriş olarak verilir.

Geri beslemeli sistemde ise, en az bir hücrenin çıkışı kendisine ya da diğer hücrelere giriş olarak verilir ve genellikle geri besleme bir geciktirme elemanı üstünden yapılır. Yapay sinir ağlarının güçlü yönleri şunlardır:

• Çok sayıda gürültülü girdi verileri içeren veri kümelerinde iyi sonuçlar verir.

• Sayısal ve kategorik çıktıların ele alınıp tahmin edilmesine olanak tanır.

• Veri kümesinde zaman faktörünün gerekli olduğu uygulamalarda da kullanılır.

• Farklı alanlara iyi uyum gösterir.[TÜZÜNTÜRK,2010]

### 3.3.5. Genetik Algoritmalar

Genetik algoritmaların temel ilkeleri ilk kez Michigan Üniversitesi'nde John Holland tarafından ortaya atılmıştır. Holland 1975 yılında yaptığı çalışmaları "Adaptation in Natural and Artificial Systems" adlı kitabında bir araya getirmiştir. İlk olarak Holland evrim yasalarını genetik algoritmalar içinde eniyileme problemleri için kullanmıştır.

Genetik algoritmalar problemlere tek bir çözüm üretmek yerine farklı çözümlerden oluşan bir çözüm kümesi üretir. Böylelikle, arama uzayında aynı anda birçok nokta değerlendirilmekte ve sonuçta bütünsel çözüme ulaşma olasılığı yükselmektedir. Çözüm kümesindeki çözümler birbirinden tamamen bağımsızdır. Her biri çok boyutlu uzay üzerinde bir vektördür.

Genetik algoritmalar problemlerin çözümü için evrimsel süreci bilgisayar ortamında taklit ederler. Diğer eniyileme yöntemlerinde olduğu gibi çözüm için tek bir yapının geliştirilmesi yerine, böyle yapılardan meydana gelen bir küme oluştururlar

### 3.3.6. K-En Yakın Komşu (K-Nearest Neighbor)

K En Yakın Komşu yöntemi, sınıflandırma problemini çözen denetimli öğrenme yöntemleri arasında yer alır. Yöntemde; sınıflandırma yapılacak verilerin öğrenme kümesindeki normal davranış verilerine benzerlikleri hesaplanarak; en yakın olduğu düşünülen k verinin ortalamasıyla, belirlenen eşik değere göre sınıflara atamaları yapılır. Önemli olan, her bir sınıfın özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır. Yöntemin performansını k en yakın komşu sayısı, eşik değer, benzerlik ölçümü ve öğrenme kümesindeki normal davranışların yeterli sayıda olması kriterleri etkilemektedir.

### 3.3.7. Bellek Temelli Nedenleme (Memory Based Reasoning)

Bellek tabanlı yöntemler, denetimli öğrenmenin kullanıldığı VM tekniklerindendir. Bu tekniğin temel özelliği, daha önceki deneyimlerden faydalanarak mevcut problemlere benzer durumları tanımlayıp geçmiş benzer problemlere getirilen uygun çözümleri mevcut problemlere uygulamaya çalışmaktır.

Bellek tabanlı yönetim tekniğinin (BTY) performansını belirleyen iki fonksiyon vardır. Bunlar, uzaklık ve kombinasyon fonksiyonlarıdır. Uzaklık fonksiyonu iki kayıt arasındaki uzaklığın bulunmasına, kombinasyon fonksiyonu ise sonuçların anlamlı sonuç sunacak şekilde birleştirilmesine olanak sağlar. BTY tekniğinin sözü edilen fonksiyonları kullanmasının bir faydası her türlü veri tipi için geçerliliğinin olmasıdır. Bu tekniğin uygulamada çok avantajlı yanları olmasının (her türlü veriye uygulanabilirliği, adaptasyon kolaylığı) dışında, geçmiş tarihi verileri saklama maliyeti bu yöntemi oldukça pahalı bir teknik haline getirmektedir [Biçen, 2002].

### 3.3.8. Naive-Bayes

Naive Bayes algoritması sınıflandırıcı bir algoritmadır. Metin dökümanlarının sınıflandırılmasında yaygın olarak kullanılır. Uygulanabilirliği ve performansı ile ön plana çıkan bir algoritmadır. İstatistiksel yöntemler yardımı ile sınıflandırma yapar.

Naive Bayes algoritmasının uygulanmasında bir takım kabuller yapılır. Bunlardan en önemlisi niteliklerin birbirinden bağımsız olduğudur. Eğer nitelikler birbirini etkiliyorsa burada olasılık hesaplamak zordur. Niteliklerin hepsinin aynı derecede önemli olduğu kabul edilir.

Naive Bayes algoritması bit ağırlıklandırma yöntemi ile ve frekans ağırlıklandırma yöntemi ile kullanılabilir.

### 3.3.9.Lojistik Regresyon (Logistic Regression)

Lojistik Regresyon Analizinin kullanım amacı, istatistikte kullanılan diğer model yapılandırma teknikleri ile aynıdır. En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenler arasındaki ilişkiyi uygun olarak tanımlayabilen bir model kurmaktır. Lojistik regresyon modelleri, son yıllarda biyoloji, tıp, ekonomi, tarım ve veterinerlik ve tasıma sahalarında yaygın olarak kullanılmaktadır. [Bircan, 2004].

## 3.4. Kümeleme Analizi

### 3.4.1. Kümeleme Analizi Tanımı

Kümeleme analizi, bir veri kümesindeki bilgileri belirli yakınlık kriterlerine göre gruplara ayırma işlemidir. Bu grupların her birine “küme” adı verilir. Kümeleme

analizine kısaca “kümeleme” adı verilir. Kümeleme işleminde küme içindeki elemanların benzerliği fazla, kümeler arası benzerlik ise az olmalıdır.

Kümeleme işlemi sınıflandırma ve öngörü işleminin aksine, veri kümesini önceden sınıflara ayırmaz, bunun yerine veriler dağılımlarına göre irdelenerek doğal sınıflandırmalar oluşturur. Kümeleme işleminin sınıflandırma işleminden en önemli farkı önceden belirlenmiş sınıflar ya da sınıf tanımları (etiketleri) olmamasıdır. Bu yüzden kümeleme işlemi gözetimsiz (unsupervised) veri madenciliği yöntemidir.

Kümeleme, gözetimsiz sınıflama (unsupervised classification) yöntemidir [Han,J,Kamber,M,2001]. Gözetimli sınıflandırma işleminde veriler önceden sınıflandırılmış örüntülerdir.

Burada temel amaç, yeni gelecek ve henüz hangi sınıfta oldugu bilinmeyen verilerin var olan sınıflardan en uygun olanına yerleştirilmesidir. Gözetimsiz sınıflamada ise amaç, başlangıçta verilen ve henüz sınıflandırılmamış bir küme veriyi anlamlı alt kümeler oluşturacak şekilde öbeklemektir. Kümeleme işlemi tamamen gelen verinin özelliklerine göre yapılır.

Kümeleme işlemi sonunda elde edilen kümeler kullanılan yöntemin giriş parametrelerine bağımlı olsa da, giriş parametrelerinden bağımsız kümeleme teknikleri geliştirme çalışmaları sürmektedir [Berkhin, Pavel.2002].

Kümeleme işleminde temel prensip, sınıf içi benzerliği maksimum, sınıflar arası benzerliği minimum yapmaktır [Han,J,Kamber,M,2001]. Bir kümeleme yönteminin kalitesi bu prensibi sağlaması ile doğru orantılıdır.

Kümeleme analizi sadece veri madenciliğinde değil, örüntü tanıma, görüntü isleme, coğrafi bilgi sistemleri gibi birçok alanda yoğun olarak kullanılmaktadır.

Literatürde pek çok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, veri tipine ve amaca bağlıdır.

Genel olarak başlıca kümeleme yöntemleri su şekilde sınıflandınlabilir

1 - Bölümleme yöntemleri (Partitioning methods)

2- Hiyerarşik yöntemler (Hierarchical methods)

3- Yogunluk tabanlı yöntemler (Density-based methods)

4- Izgara tabanlı yöntemler (Grid-based methods)

5- Model tabanlı yöntemler (Model-based methods)

**1.Bölümleme Metodları**

Bölümleme metodları (partitioning methods), n adet nesneden oluşan veritabanını, giriş parametresi olarak belirlenen k adet bölüme (k <= n) ayırma temeline dayanır. Veritabanındaki her bir eleman bir farklılık fonksiyonuna (dissimilarity function) göre k adet bölümden birine dahil edilir. Bu bölümlerden her biri bir küme olarak adlandırılır.

Bölümleme metodları k sayısı doğru tahmin edilebilirse benzer şekilli dışbükey kümeleri bulmakta oldukça başarılı sonuçlar vermektedir. Eşer k sayısı hakkında önceden bir fikir belirlenemezse algoritmayı farklı k değerleri için tekrar tekrar uygulayarak en uygun k değeri bulunabilir.

Bölümleme metodlarının genel problemi k giriş parametresine bağımlı olmaları ve düzgün şekilli olmayan kümeleri bulamamalarıdır [24]. Bölümleme metodları kmeans, k-medoids ve CLARA-CLARANS olarak bilinen algoritmaları kullanır [ Han, J; Kamber, M.2001].

**2.Hiyerarsik Kümeleme Teknikleri**

Hiyerarşik modeller bir ağaç yapısı oluşturarak kümeleme işlemini gerçekleştirmektedir. Oluşturulan kümeleme ağacının bütün ağaç yapılarında olduğu gibi bir root düğümü ve çocuk düğümleri mevcutdur.

Hiyerarşik kümeleme tekniklerinde, başlangıçta küme sayısı belirtilmemektedir. Algoritma, x: veri seti s: uzaklıklar matrisi olmak üzere; ( x, s) girdi olarak tanımlanmaktadır.

Sonuçta çıktı olarak elde edilen kümeler hiyerarşiktir. Hiyerarşik kümeleme tekniklerinin bir çoğunda uygulanan süreç optimizasyon esaslı değildir. Bu tekniklerdeki amaç, birleşme tamamlanıncaya kadar bölmenin ilerlemesi için tekrarlamalar kullanarak bazı yaklaşımlar bulmaktır.

Hiyerarşik kümeleme teknikleri ağaca benzer (dendogram) bir grafik oluştururlar. Hiyerarşik kümeleme tekniklerini, hiyerarşik birleşmenin aşağıdan yukarıya (bottomup) ve yukarıdan aşağıya (top-down) yapılmasına bağlı olarak toplayıcı (agglomerative) ve ayırıcı (divise) hiyerarşik kümeleme teknikleri diye ikiye ayırmak mümkündür.

Aşağıdan yukarıya, toplaşım kümeleme algoritmaları ve yukardan aşağıya kümeleme algoritmaları olarak iki grupta toplanabilir.

Toplaşım kümeleme algoritmaları, başlangıçta veritabanındaki her bir noktayı bir küme olarak görür. Bu kümeleri birleştire birbirinden ayrı kümeler oluşturur. Bölünür kümeleme algoritmaları ise başlangıçta veritabanındaki tüm noktaları tek bir kümeymiş gibi görür. Veritabanını taradıkça, birbirine benzemeyen noktaları kümeden dışarı atarak önceden verilmiş, k kadar kümeye dağıtır. (Silahtaroğlu,2008)

Hiyerarşik Modelleri kullanan başlıca algoritmalar ise şunlardır.

### 2.1) SLINK Algoritması

SLINK algoritması, Tek Bağlantı tekniği ile anılmaktadır. SLINK algoritması yukarda anlatılmış olan toplaşımlı metoda göre çalışmaktadır. SLINK algoritması temelde2 küme grubunun en dışında olan ve birbirine yakın olan noktalar arasındaki mesafeye göre benzerlik teoremleri geliştirerek kümeleme işlemini gerçekleştirmektedir. Buradaki mesafenin ve benzerliği ölçümünde bilinen formullerden yararlanılmaktadır. Ayrıca burada en kısa yol algoritması (gezgin satıcı algoritması olarak da geçmektedir.) kullanıldığını da belirtmek gerekir.

### 2.2) CURE (Clustering Using Representatives ) Algoritması

Bu algoritma dağınık bir şekil gösteren küme yapılarında ki küme içine alınıp alınamayacağına karar verilemeyen nesnelerin değerlendirilmesine faydalı bir yaklaşım

önermektedir. Bu algoritma temelde bütün nesneleri birer küme oluşturabileceği yaklaşımına göre çalışmaktadır. Özetle, CURE algoritması temsilciler kullanarak kümeleme işlemini gerçekleştirmektedir

### 2.3) CHAMELEON Algoritması

CHAMELEON algoritması iki küme arasındaki benzerliği dinamik bir şekilde belirler ve çoğu algoritmanın(SLINK,CURE vb.) hatalı kümelemeler yaptığı durumlarda düzgün bir şekilde kümeleme işlemini gerçekleştirmektedir. Temelde bu algoritma kümelerin kendi iç benzerlikleri ile alt kümeleri arasında ki benzerliklere göre işlemlerini gerçekleştirmektedir. Bazı durumlarda kümenin dışına yakın nesneler komşu bir kümenin merkezine mesafe olarak yakın olabilir, bu gibi durumlarda k-means algoritması gibi algoritmalar hatalı işlemler yapmasına rağmen CHAMELEON için böyle bir şey söz konusu değildir. Çünkü bu algoritmada dinamik olarak küme içerisinde ki bağlantılar ve benzerliklerde dikkate alınmaktadır.

Aşağıdaki şekilde algoritmanın çalışma mantığını görebilirsiniz.

### 2.4) BIRCH Algoritması

BIRCH (Balanced Iterative Reducing and Clustering using Hierarcihes ) Algoritması çok büyük veritabanlarının kümelenmesi amacıyla oluşturulmuş, gürültülü verilerin kontrolünü de sağlayan ilk hiyerarşik kümeleme algoritmasıdır. BIRCH algoritması kümeleme işlemini bir ağaç yapısı oluşturarak gerçekleştirir ve sadece sayısal veriler üzerinde çalışır. Burada belirtilen ağaç yapısına CF ağacı denmektedir.

CF = (n,LS,SS).

Olarak 3 tane bilgiyi barındırır. Burada ki “n” kümedeki nokta sayısı, “LS” kümedeki noktaların değerlerinin toplamı, “SS” kümedeki noktaların değerlerinin karelerinin toplamıdır.

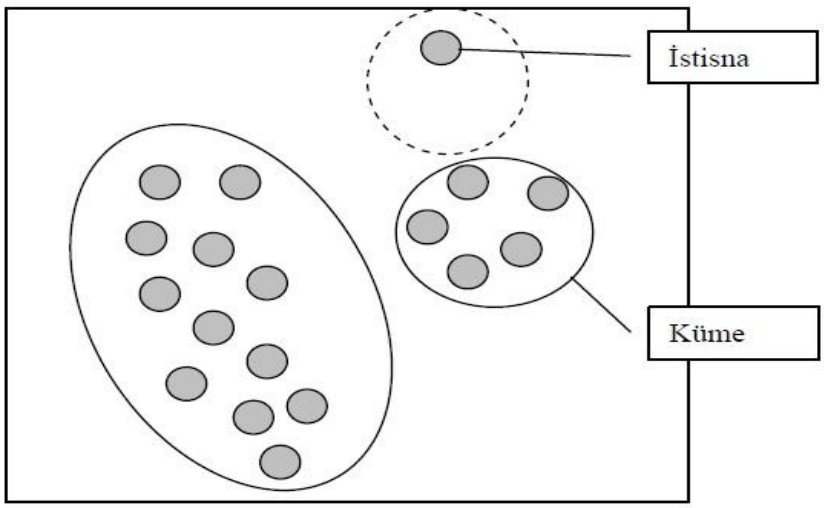
CF ağacı yukarıdan aşağıya doğru artış gösterir.(yani toplaşım algoritması değil, hiyerarşik ama bölünür bir kümeleme algoritmasıdır) CF ağacının dallarının artışı, daha

önceden belirlenmiş T (eşik değeri) ne kadar devam eder. T değerinin aşıldığı yerlerde bir aşağı düğüme geçilir. Aşağıdaki şekilde CF ağacının yapısı görülebilir.

### 3- Yoğunluk Tabanlı Yöntemler (Density-based methods)

Dağılmış verilere sahip veritabanlarının sadece uzaklığı temel alan bölümlemeli algoritmalar ile kümelenmesi oldukça güçtür. Çünkü hiç bir kümeye dâhil olmayan uç noktalar içeren bu dağılmış veritabanlarının bölümlemeli algoritmalar ile kümelenmesi neticesinde doğru kümeler ortaya çıkmayacaktır. Bu durumda birlikte bir yoğunluk oluşturan verilerin aynı kümeye alınmasına dayanan yoğunluğa dayalı algoritmalar kullanılmalıdır. Bu tür algoritmalara örnek olarak DBSCAN, OPTICS ve DENCLUE algoritmaları verilebilir.

Bu yöntemde her noktanın çevresindeki komsuları ile olan yakınlığı hesaplanır. Yakınlık hesaplamada genelde Öklit uzaklığı kullanılsa da veri türüne göre yakınlık hesaplama yöntemi farklılık gösterebilir. Bu yöntemin temel prensibi "yeterince komsusu olmayan noktaları" tespit etmektir. Bu durum Sekil 8'de görülmektedir.




Şekil.3.6 İstisna ve küme oluşumları

### 4- Izgara Tabanlı Yöntemler (Grid-based methods)

Büyük boyuttaki veritabanlarının kümelenmesinde numaralandırılmış çizgilerden oluşan hücresel yapıları kullanan algoritmalardır. Bu algoritmalara örnek olarak ise bölgenin dikdörtgen hücrelere bölünerek hiyerarşik bir yapının kullanıldığı STING algoritması, değişik şekillerde kümeler sunabilen ve hassas kümeleme kabiliyeti olan dalga kümeleme algoritması ve hem yoğunluğa hem de grid yapısına sahip CLIQUE algoritması verilebilir.

## 3.4.2. Kümeleme Analizinin Özellikleri

İyi bir kümeleme analizi yöntemi su özelliklere sahip olmalıdır [ Han, J; Kamber, M .2001]:

- Ölçeklenebilir olmalıdır. Birkaç yüz kayıttan oluşan veri kümesine de milyonlarca kayıt içeren kümeye de uygulanabilmelidir.
- Farklı veri türleri ile kullanılabilmelidir. Hem sayısal hem kategorik veriler içeren veritabanlarında kullanılabilmelidir.
- Düzgün şekilli olmayan kümeleri de bulabilmelidir.
- En az sayıda giriş değişkeni gerektirmelidir. Bir yöntem ne kadar az giriş değişkeni gerektiriyorsa o ölçüde kullanıcının kararlarından bağımsızdır.
- Gürültü içeren veriler ile de kullanılabilmelidir.
- Veri kümesindeki kayıtların sıralanmasından bağımsız olmalıdır. Kümenin hangi elemanından başlanırsa başlansın sonuç değişmemelidir.
- Çok boyutlu veritabanlarına uygulanabilmelidir.
- Veri kümesinin sahip olduğu sınırlıkları dikkate alabilmelidir.
- Kolay yorumlanabilir sonuçlar üretebilmeli ve işlevsel olmalıdır.

Bu özellikler ideal bir kümeleme algoritmasının niteliklerdir. Mevcut algoritmaların hiç biri bu özelliklerin tamamına sahip değildir. Kümeleme analizi gelişmekte olan bir araştırma konusudur ve ilerleyen yıllarda ideale yakın yöntemlerin geliştirileceği umulmaktadır.

### 3.4.3 Kümeleme Analizi Veri Türleri

Veri madenciliğinin birçok alanında olduğu gibi Kümeleme Analizinde de veri yapısı matris formundadır. Matris formu bilgisayar ortamında hesaplama yapabilmek için en uygun veri yapısı olarak kendini kanıtlamıştır.

Kümeleme işleminde kullanılan matrisler iki temel gruba ayrılır [ Han, J; Kamber, M.2001]

#### 3.4.3.1 Veri Matrisi (data matrix)

Bu matris n adet nesne için p adet özelliğin tanımlandığı satırların birleşmesinden oluşan (n x p) boyutundadır. Örneğin bir şehirdeki insanların yas, boy, ağırlık, cinsiyet, mahalle gibi özellikleri alt alta yazıldığında Denklem şekil.9'daki gibi bir matris oluşur. Burada her bir sütun bir niteliği, her bir satır ise niteliklerin değerlerini içermektedir.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Şekil.3.7 Veri Matrisi (data matrix)

### 3.4.3.2 Farklılık Matrisi (Dissimilarity matrix)

Nesnelerin diğer nesneler ile olan uzaklık bilgilerinin tutulduğu n x n boyutunda olan matristir. Bu matrisin genel ifadesi Denklem şekil3.6.'da görülmektedir

$$\begin{bmatrix} 0 & d(1,2) & d(1,3) & & d(1,n) \\ d(2,1) & 0 & d(2,3) & & d(2,n) \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Şekil 3.8 Farklılık matrisi (Dissimilarity matrix):

Nesneler arasındaki uzaklık fonksiyonu değişme özelliğine sahip olduğu için, diğer bir ifade ile :

d(i,j) = d(j,i)

olduğu için farklılık matrisinin asal köşegenin altında kalan değerler ile üstünde kalan değerler simetriktir. Bu yüzden farklılık matrisine tek yönlü (one-mode) matris denir ve yalnızca asal köşegen ve altında kalan elemanları içerir. Veri matrisinin böyle bir özelliği bulunmadığı için iki yönlü (two mode) matris denir.

Veri madenciliğinde çoğunlukla farklılık matrisi kullanılır. Farklılık matrisi elemanlarını bulabilmek için elemanlar arası farklar hesaplanabilmelidir.

## 3.5. Sıra Dışılık Analizi

Bir veri kümesinde verilerin genel davranışından veya veri dağılım modelinden farklılık gösteren nesnelere sıra dışı (Outlier) denir. Birçok veri madenciliği yöntemi istisnaları gürültü veya aşırı durumlar olarak görür, bu yüzden dikkate almaz. Fakat bazı

durumlarda istisna noktalar diğerlerine göre çok daha fazla bilgi içerir. Örneğin kredi kartı veya sigorta sahte karlıklarının tespitinde, tıp biliminde yeni bir hastalığın başlangıcını tespit etmede istisnalar analiz edilir. İstisna analizinde iki yöntem söz konusudur ayrılır [ Han, J; Kamber, M.2000]

### 3.5.1. İstatistik Tabanlı Yöntem

İstatistik yaklaşım kümeleme ve sınıflandırma yöntemlerinin her ikisini de kullanır.

İstatistik yaklaşım diğer tüm kümeleme modellerinde olduğu gibi sadece kümelenmeleri ortaya çıkarmakla kalmaz, bunun yanında kümelerin genel karakterleri ile ilgili bilgiler de verir. Bu isleme kavramsal kümeleme denir

## BÖLÜM 4.

## TIP VE HASTA BİLGİ SİSTEMLERİNDE VERİ MADENCİLİĞİ UYGULAMALARI

Sağlık sektörü bilginin içerik ve yapısal anlamda en hızlı değiştiği alanlardandır. Sağlık hizmetlerinin en hızlı, en doğru, en yüksek kalitede ve ihtiyaca cevap verecek şekilde sunulabilmesi için sağlık araştırmacı ve karar vericilerinin en doğru ve güncel bilgiye ulaşması gerekmektedir.

Günümüzde bilgi sistemleri ve iletişim teknolojilerindeki gelişmeler sayesinde tıp ve sağlık alanındaki birçok veri sayısal ortamda saklanabilmekte ve kolaylıkla erişilebilmektedirler.

Bu nedenle sağlık hizmetlerinin sunumu, her düzeydeki sağlık kurumlarının yönetimi ve sağlık politikalarının oluşturulmasında bir karar destek aracı olarak Veri Madenciliği'nin kullanılması sağlık araştırmacılarının en doğru kararları almasına yardımcı olacaktır.

Sağlık Bakanlığının sağlık bilgi sistemleri ve sağlık bilgi standartları üzerine çalışmaları her geçen gün yeni gereksinimlere cevap verecek nitelik de güncellenip yenilenmektedir, ulusal sağlık veri sözlüğünün 2.0 versiyonu ile veri toplama standartları daha da güncel hale gelmiştir.

### 4.1.Tıp da Veri Madenciliği Uygulamaları

Sağlık bilgi sistemlerindeki veri madenciliği tekniklerinin ilk kullanımı 1970'lerde ve daha sonraki yıllarda geliştirilen uzman sistemlerle olmuştur. Uzman sistemlerin tıp alanında güçlü araçlar sunmasına rağmen, sağlık alandaki verilerin hızlı değişmesi ve uzmanlar arasındaki görüş farklılıkları nedeniyle çok yaygınlaşmamıştırlar.

Daha sonraki yıllarda özellikle 1990'lı yıllarda hastaların gelecekteki sağlık durumları ve maliyet tahminleri gibi konuları araştırmak için sinir ağları kullanılmaya başlanmıştır.

Örneğin birçok insan, kalp hastalıkları, diabet ve astım hastalıkları gibi kronik hastalıklarla yaşamak zorundadır. Bu hastalıkların hem tıbbi açıdan hem de hastane kaynak ve maliyetleri açısından ele alınarak doğru yönetilmesi gerekmektedir. Veri madenciliği yöntemleri kullanılarak bu sistemlerdeki gizli ve önemli bilgiler keşfedilmelidir. Keşfedilen bu bilgiler hem tıbbi araştırmalar hem de yönetim planları için değerlendirilmelidir.

Sağlık alanında yapılan birçok veri madenciliği araştırmalarında hastaların elektronik tıbbi kayıtları ve idari işleri belgeleyen veriler kullanılmaktadır. Bu verilerden yararlanılarak farklı tahminler yapılabilir.

• Belirli bir hastalığa sahip kişilerin ortak özelliklerinin tahmin edilmesi

• Tıbbi tedaviden sonra hastaların durumlarının tahmin edilmesi

• Hastane maliyetlerinin tahmin edilmesi

• Ölüm oranları ve salgın hastalıkların tahmin edilmesi [ Kudyba, S.,2004].

Örneğin Tablo 4.1. de hastalığın olup olmaması durumuna göre hastaların yaş, tansiyon ve sigara kullanımı gibi bilgileri verilmiştir. Bir veri madenciliği algoritması bu verilerden yararlanarak hastalığın olup olmamasına dair kurallar çıkarabilir.

Tablo 4.1. Hastalık Sınıflandırma Veri Seti

| Örnek No | Yaş | Tansiyon | Sigara Kullanımı | Hastalık (Class) |
|---|---|---|---|---|
| 1 | 18 | Normal | Kullanmıyor | yok |
| 2 | 24 | Normal | Her zaman | var |
| 3 | 37 | Normal | Bazen | yok |
| 4 | 33 | Düşük | Her zaman | var |
| 5 | 52 | Yüksek | Bazen | var |
| 6 | 55 | Normal | Kullanmıyor | yok |
| 7 | 56 | Yüksek | Bazen | yok |

Hastalıkların yönetimi ile ilgili veri madenciliği çalışmaları hastalıkların ve durumlarının tanımlanmasını ve maliyetlerin modellenmesi gibi araştırmaları içerir. Bu çalışmalarda amaç pozitif sonuç elde etmektir. Örneğin Harleen Kaur ve arkadaşları hastaların yaş ve cinsiyet gibi verilerini karar ağacı yöntemleri ile analiz ederek göğüs kanseri olup olmadığını tahmin etmeye çalışmışlardır [Kaur, H.,Wasan., S,2006]. Hastane bilgi sistemlerindeki verilerle yapılmış diğer bir çalışmada, hastaların sık sık farklı doktorları ziyaret etmeleri araştırılmış ve hasta demografik bilgileri ve işlemsel veriler analiz edilmiştir. İlişkisel kural analizi (Association rules analysis) kullanılarak yapılan veri madenciliği çalışmasında, yaşın, cinsiyetin, hastanelerin özelliklerinin, kronik ve akıl hastalıklarının sürekli doktor ziyaret etme davranışında etkili oldukları ortaya çıkartılmıştır[Chen.Y., ve Wu., S,2003].

Hastanelerde maliyetleri etkileyen en önemli konulardan birisi de hastaların kalış süreleridir.

Kalış sürelerinin etkileyen faktörler de günümüzde veri madenciliği çalışmalarının araştırma konusudur ve birçok çalışma yapılmıştır.

Örneğin, yapılan bir çalışmada hastaların demografik ve çevresel bilgileri, sinir ağları ile analiz edilmiş ve bazı önemli bilgiler elde edilmiştir.

Bu bilgilere göre 40 yaşından büyük hastalar, şehirlerde yaşayan hastalar, alkol ve sigara bağımlılığı olan hastalar daha uzun süre hastanede kalmaktadırlar. Ayrıca özel hastanelerdeki kalış süreleri devlet hastanelerinden daha kısadır

Sağlık uygulamaları ve tedaviler büyük oranda maliyet gerektirirler. Yapılan tetkikler veya tedavilerden hile yapılarak çıkar sağlanmaya çalışılabilir. Özellikle Avrupa ve Amerika'da sağlık sigorta şirketleri bu konuları araştırmaktadırlar. Hile tespiti için veri madenciliği yöntemlerinden yararlanılır. Bu tür araştırmalarda hasta, tetkik ve doktor bilgileri analiz edilir ve anormal veriler incelenir. Örneğin ortalama maliyetin üzerindeki tetkikler veya tedaviler şüphe kaynağıdır. Bu çalışmalarda genellikle kümeleme(clustering) algoritmaları kullanılır.

Bu alanda ilaçlar da tıbbın önemli araştırma konularından birisidir. Amerika Birleşik Devletleri'nde yeni bir ilaç geliştirildiğinde, klinik denemelerden sonra FDA (Food and Drug Administrators) kurumu tarafından onaylanarak piyasaya sürülür. Onaylanmadan önce ilacın faydalarının risklerinden daha çok olması göz önünde bulundurulur. Bazı ilaçlar piyasaya sürüldükten sonra risklerinin çok fazla görülmesi nedeniyle kaldırılmışlardır. İlaçların önceden tanımlanmamış yan etkilerinin bulunabileceği olasılığı, web üzerinden tıbbi yayınlar analiz edilerek veri madenciliği çalışmaları da yapılmaktadır.[ Carino., C., Jia., Y., Lambert., B., West., P.,Yu., C,2005].,

Ülkemizde Sağlık Bakanlığı tarafından benzer çalışmalar yürütülmektedir.

### 4.2.Tıp ve Biyoinformatik Alanlarında Veri Madenciliği Çalışmaları

Tıp ve sağlık alanındaki verilerin birçoğu yapısal olmayan metinlerde saklanmaktadırlar. Örneğin, hastaların tıbbi durumları, tanı, tedavi bilgileri ve klinik dokumanlar metin olarak saklanmaktadırlar. Ayrıca, uygulanan işlemlere ait faturalar ve iş akışını belgeleyen raporlar da metin formatındadır. Tıp alanındaki bilimsel makaleler de sağlık alanında yapılan araştırmalar ve yenilikler için değerli bilgi kaynaklarıdır ve metinsel yapılarda saklanırlar. Bu yapılar üzerinde bilgi keşfi yapmak için metin madenciliği yöntemleri kullanılmaktadır.

Metin madenciliği Doğal Dil İşleme (Natural Language Processing), Tıp Bilişimi (Medical Informatics) ve İstatistik gibi alanlarla ortak çalışılan bir araştırma alanıdır. Tıp alanında yapılan metin madenciliği çalışmalarında özellikle Medline gibi tıbbi alanda yapılmış bilimsel yayınların saklandığı büyük veritabanları için bilgi keşfetme yöntemleri geliştirilir ve bu çalışmaların amacı metin yapısındaki verilerin analiz edilip bilgi keşfi yapmak ve bilgi yönetimi sağlamaktır. Tıp alanındaki makalelerden tedavi ve tanı ile ilgili yeni yaklaşımlar, kavramlar arasındaki gizli ilişkiler ortaya çıkartılabilir. Elde edilen önemli bilgiler hem araştırmalara büyük destek sağlar hem de sağlık kurumlarının başarısını artırır.

Sağlık ve tıp, günümüzün en çok bilgi ihtiyacı olan araştırma alanlarıdır. Son yıllarda özellikle sağlık veri modelleri, standartlar ve kodlama sistemlerindeki yenilikler sayesinde hastanelerde ve sağlık merkezlerinde kullanılan bilgi sistemlerinde önemli gelişmeler yaşanmıştır. Bu gelişmeler daha çok ve çeşitli verinin saklanabilmesini sağlamış ve beraberinde bilgi keşfi ihtiyacını ortaya çıkarmıştır. Veri Madenciliği, sağlık ve tıp alanındaki büyük veritabanlarından değerli bilgileri ortaya çıkartarak, hem tıp açısından hem de hizmet kalitesinin artırılması açısından büyük katkılar sağlar. Günümüzde uluslararası ortak projeler kapsamında geliştirilen ve biyoloji verilerinin saklandığı veritabanları, bu veritabanlarına erişim ve veri madenciliği sistemleri de klinik araştırmaların önemli bir parçası haline gelmişlerdir. Tıbbi gelişmelerde biyolojik araştırmaların da büyük katkısı vardır. Günümüzde biyolojik yapılar ve özellikle genlerle ilgili büyük uluslar arası veritabanları ve bunlara erişimi kolaylaştıran yazılım araçları kullanılmaktadır. Örneğin BIOMART, Avrupa Biyoinformatik Enstitüsü ve Cold Spring Harbor Laboratuarı (CSHL) tarafından geliştirilmiş, ilişkisel veriler için biyoinformatik dünyasında yaygın olarak kullanılan bir veri madenciliği sistemidir . Bu sistem karmaşık ilişkilere sahip biyolojik veriler OMIM (Online Mendelian Inheritance in Man) projesi ise National Center for Biotechnology Information (NCBI) tarafından geliştirilmiş genetik bozukluklarla ilgili bilinen hastalıkların saklandığı bir veritabanıdır. Bu veritabanı me tin bilgiler, resimler ve referans bilgilerinden oluşmuştur ve ayrıca Medline veritabanına da bağlantısı vardır. Büyük bir bilgi kaynağına sahip olan veritabanı genlerle ilgili araştırmalarda önemli katkılar sağlar. [Pınar YILDIRIM1, Mahmut ULUDAĞ2, Abdülkadir GÖRÜR1,2008]

### 4.3.Hastane Bilgi Sistemlerinde Veri Madenciliği Uygulamaları

Hastane bilgi sistemleri hastalara ait demografik bilgiler, hastalık ve tedavi durumları, yapılan tetkikler, faturalama ve idari işlere ait bilgileri içerir. Sağlık ve tıp, çağımızın en önemli bilimsel araştırma alanları olduğu için bu alandaki bilgi sistemleri de araştırmalar için en büyük veri kaynaklarıdır. Son otuz yılda dünyada sağlık bilgi sistemlerinde büyük gelişmeler yaşanmıştır. Sağlık Bilişiminin yeni bir alan olmasına

rağmen özellikle bilgi modelleme ve tanı araçlarında hızlı yenilikler yapılmıştır [Kudyba, S,2004].

Günümüz Hastane Bilgi Sistemleri hastalara ve onların tıbbi durumlarına ait birçok veri barındırmaktadır. Bu sağlık verileri, hastane ve klinik veritabanlarında gizlidir. Hastanelerin ve sağlık kuruluşlarının verimliliğini artırmak ve geleceğe dair planları yapabilmek için veri madenciliği teknikleri kullanılır tıbbi verilerden gizli kalmış önemli bilgileri ortaya çıkarır ve böylece bu teknikler hastaneler ve klinik araştırmalar için değerli bilgiler sağlarlar.

Ülkemizde, Sağlık Bakanlığı yaptığı değerlendirmeler ile sağlık alanında politika üretmek için hayati öneme sahip verilerin toplanmasında, saklanmasında ve analiz edilmesinde ulusal veya uluslararası standartların olmadığı, özellikle veri toplama konusunda ciddi bir karmaşanın mevcut olduğu tespitinde bulunmuş ve"Sağlıkta Dönüşüm Programı" kapsamında "Karar Sürecinde Etkili Bilgiye Erişim: Sağlık Bilgi Sistemi"başlığı ile çalışmalar başlatmıştır. Ulusal Sağlık Veri Sözlüğü, Minimum Veri Setleri, Sağlık Kodlama Referans Sunucusu ve sağlık verilerinin toplandığı Elektronik Sağlık Kaydı (ESK) veritabanı ve Karar Destek Sistemi bileşenleri bu çalışmaların kapsamını oluşturmaktadır [Sağlık Bakanlığı,.2009.].

# BÖLÜM 5.

# MEME KANSERİ

İnsan vücudundaki sağlıklı hücrelerin, kas ve sinir hücreleri hariç, bölünebilme yeteneği vardır. Hücreler, yenilenme ve yaralanan dokuların onarılması amacıyla bölünebilme yeteneklerini kullanırlar. Her hücrenin hayatı boyunca belli bir bölünebilme sayısı vardır. Sağlıklı bir hücre gerektiği yerde ve gerektiği kadar bölüneceğini bilir. Buna karşın kanser hücreleri, bu bilinci kaybedip kontrolsüz olarak bölünmeye başlayarak çoğalırlar. Bütün bu bilgilerden yola çıkarak kanser için aşağıdaki tanımlamaları yapabiliriz,

Kanser için en doğru tanımlamaları aşağıdaki gibi açıklayabiliriz.

Kanser, bir dokunun veya organın hücrelerinde sağlıksız bir değişme ortaya çıkıp bu hücrelerin denetimsiz çoğalmaya başlamasıyla meydana gelen hastalıkların tümü için kullanılan bir genel kavramdır.

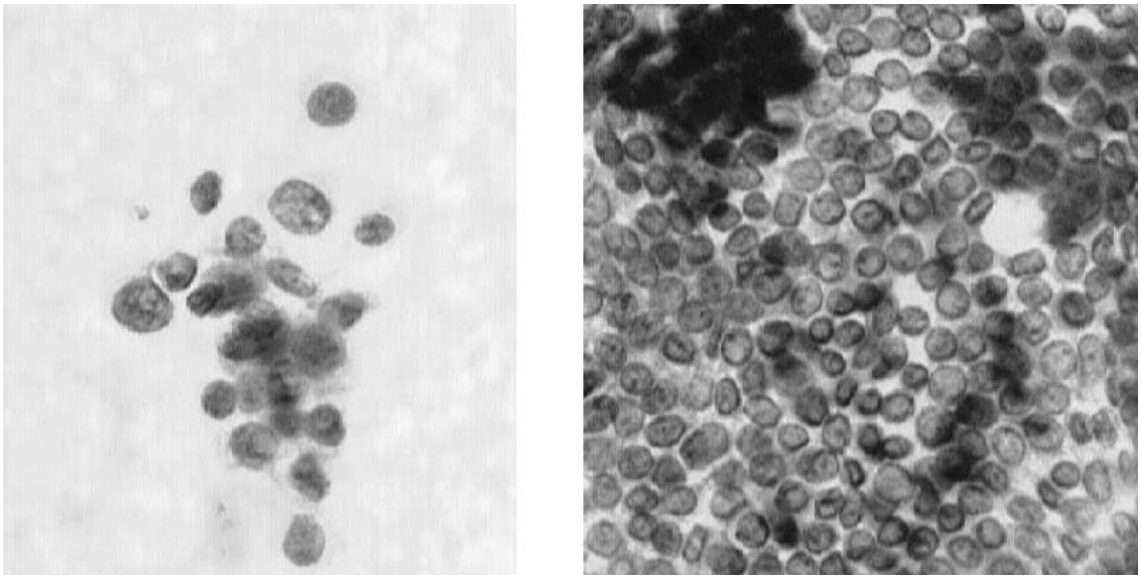
- Genellikle kontrolden çıkan hücrelerin hızlı ve sürekli çoğalmalarıdır.

Vücudumuzun her organı bir çok farklı hücre tiplerinden oluşmaktadır. Normal olarak bu hücreler vücut için gerekli olduğu şekilde belli bir düzen içinde büyüyerek bölünürler. Bu süreç bir düzen içinde yürür ve vücut sağlığının korunmasına yarar. Eğer, yeni hücrelere ihtiyaç olmadığı halde hücreler bölünmeye başlarsa, gereğinden fazla doku oluşmaya başlar. Bu fazlalık dokular tümör denen bir urun ortaya çıkmasına sebep olurlar. Böylece oluşan fazlalık doku iyi huylu veya kötü huylu olabilir.

- Kötü huylu (malignant) tümörlerdir; yani iyi huylu (benign) tümörlerin aksine başka dokulara sızma ve yayılma özelliği gösterirler.

İyi huylu tümörler :Kanser değildir. Bunlar normal olarak yok edilebilinir ve bir daha da meydana çıkmazlar.

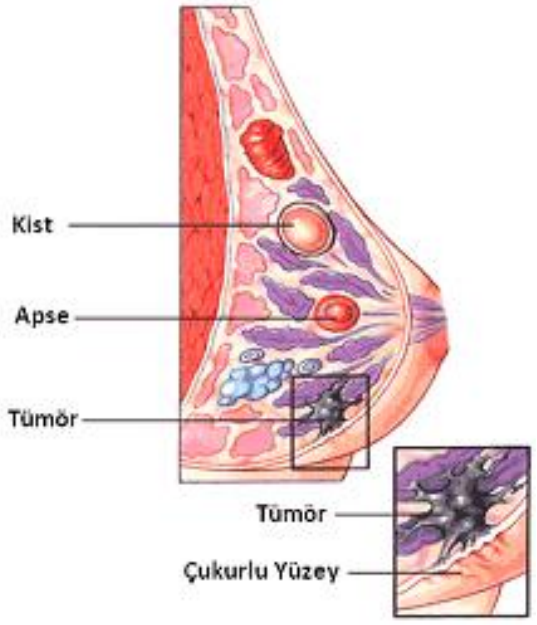
Kötü huylu tümörler: Kanser anlamına gelir. Kanser hücreleri denetimsiz büyüyüp bölünür. Bunlar yakınlarındaki sağlıklı dokuların içine girerek bunları bozabilirler. Kanser hücreleri Ayrıca ilk tümörden koparak kan dolaşımına veya lenf sistemine girebilirler. Göğüs kanseri de bu yoldan yayılarak vücudun diğer parçalarında yeni tümörler oluşturur. Kanserli dokunun yayılmasına metastaz oluşturma denir [Temiz,2007].



Şekil 4.1. Meme kanseri Hücreleri

Daha henüz kanserin erken evrelerinde bile kanser hücreleri kan damarları yolu ile kana karışır. Ancak kana karışan kanser hücrelerinin çoğu kanın içinde tahrip olur, bu yüzden kan yolu ile yayılım erken dönemlerde genellikle olmaz. Kanser, zamanla geliştikçe kana karışan kanser hücrelerinin sayısı da artar, bu artışa orantılı olarak kan yolu ile metastaz yapma olasılığı da artar. Kanser hücreleri kan yolu ile tüm vücuda yayılabildiği halde yerleşip metastaz yapabilmeleri için belli bir düzeyde oksijene ihtiyaç duyarlar.[Temiz,2007]

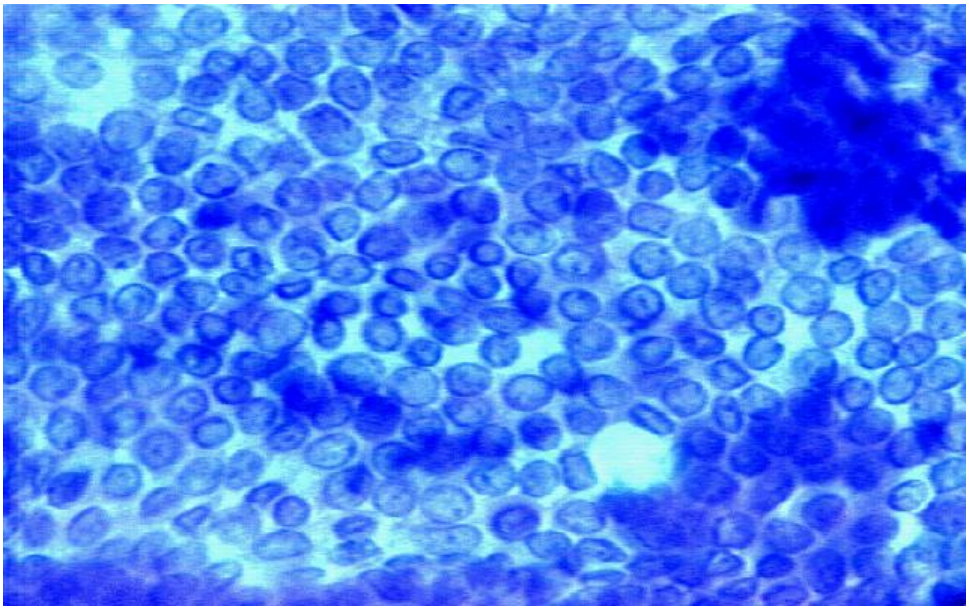
Şekil 12'de meme kanseri hücreleri görülmektedir. Kadınlarda meme, süt bezleri ve burada üretilen sütü meme ucuna taşıyan kanallardan oluşmaktadır. Bu süt bezleri ve kanalları döşeyen hücrelerin; erkeklerde ise, nadir olarak görülse de, meme hücrelerinin kontrol dışı olarak çoğalmasına meme kanseri denir. 2004 yılı Dünya Sağlık Örgütü kayıtlarına göre aynı yıl içerisinde gerçekleşen ölümlerin %13'ü kansere bağlıdır. 7.4 milyon insan kanser yüzünden hayatını kaybetmiştir ve meme kanseri yüzünden 519 bin insan hayatını kaybetmiştir [Mustafa DANACI, Mete ÇELİK, A. Erhan AKKAYA,2010]



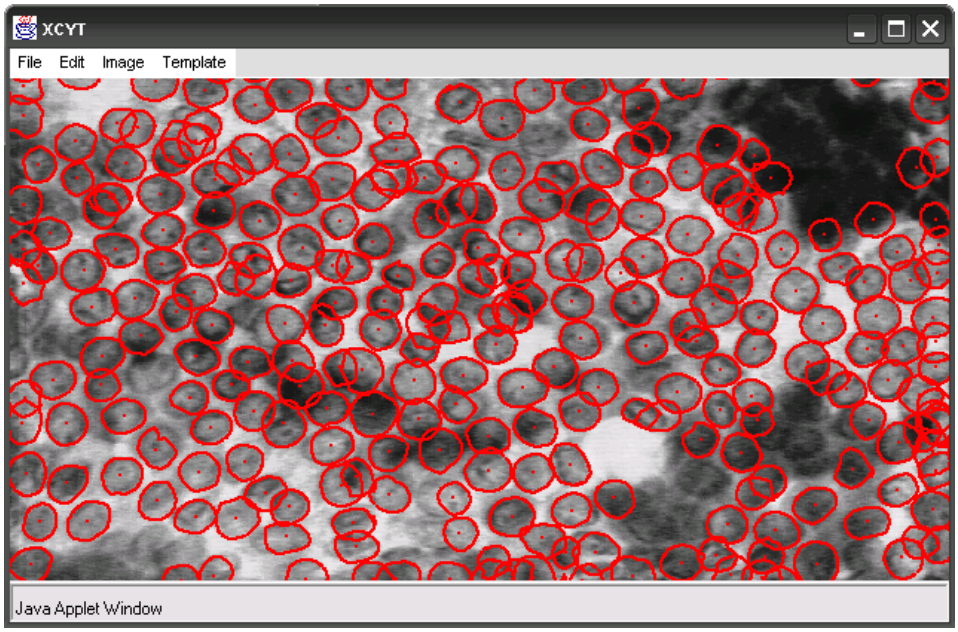

Şekil 4.2. Meme kanseri Hücreleri

## 5.1.ÖRNEK UYGULAMA

Xcyt [5], 1990 yıllarının başında Amerika Madison Eyaleti Wisconsin Üniversitesi'nde temelleri atılan bir yazılım projesinin adıdır. Yazılım, meme kanserindeki iki önemli safha olan tanı (diagnosis) ve tahmin (prognosis) safhalarını gerçekleştirmeyi hedefleyerek bir proje olarak ortaya çıkmıştır. Şekil 4.3'de kanser riski taşıyan bir hastanın göğsünden alınan doku örneğinin mikroskop altında 63 defa büyütülmesi ile elde edilen görüntü verilmiştir. Bu görüntü kullanılarak kenar çıkarma ve merkez saptama algoritmaları ile Xcyt programına hücrelerin özelliklerinden dokunun genel şablonu çıkarttırılmaktadır. Xcyt programı ile Şekil 4.4'de görüldüğü gibi önce verilen hücre topluluğu gri skalaya çevrilmiştir.



Şekil 4.3. Meme dokusu altından alınan 63 defa büyütülmüş hücre topluluğu




Şekil 4.4. Merkez saptama ve çevre çıkarma

Hücrelerin büyük bir kısmı program tarafından saptandıktan sonra bu hücrelere ait genel veriler elde edilmiştir.

- **Radius:** Tüm hücrelerin yarıçapları ortalaması, standart sapması ve en kötü değeri
- **Texture**: İç yüzeylerin gri skaladaki değişim oranlarının ortalaması, standart sapması ve en kötü değeri
- **Perimeter:** Hücrelerin çevre uzunlukları ortalaması, standart sapması ve en kötü değeri
- **Area**: Hücrelerin yüzey alanları ortalaması, standart sapması ve en kötü değeri
- **Smoothness:** Komşu hücrelerin yarıçap uzunluklarının ortalaması, standart sapması ve en kötü değeri
- **Compactness:** Çevre²/Alan = Yoğunluk ortalaması, standart sapması ve en kötü değeri
- **Concavity:** Hücre çevresindeki girinti ve çıkıntıların büyüklükleri ortalaması, standart sapması ve en kötü değeri
- **Concave Points**: Hücre çevresindeki girinti ve çıkıntı nokta sayısının ortalaması, standart sapması ve en kötü değeri

- **Symmetry**: Hücrelerin elips şekil değişikliği ortalaması, standart sapması ve en kötü değeri
- **Fractal Dimension**: İç içe geçmiş düzensiz hücrelerin tüm normal hücrelere oranının ortalaması, standart sapması ve en kötü değeridir.

# BÖLÜM 6.

# VERİ MADENCİLİĞİ PROGRAMLARI

Veri Madenciliği, veriden bilgi elde etme amaçlı kullanılan teknikler bütünüdür. İstatistiksel analiz tekniklerinin ve yapay zekâ algoritmalarının bir arada kullanılarak veri içerisindeki gizli bilgilerin açığa çıkarılması ve verinin nitelikli bilgiye dönüştürülmesi sürecidir. Veri Madenciliği uygulamalarını gerçekleştirmek için ticari ve açık kaynak olmak üzere birçok program mevcuttur.

Veri Madenciliği uygulamaları yapmak için bilgisayar programı kullanmak gereklidir. Bu kapsamda birçok yazılım geliştirilmiştir. Bu bölümde özellikle Açık Kaynak Kodlu Veri Madenciliği Programlarından olan WEKA'ya değinilmiştir.

## 6.1.Ticari Veri Madenciliği Programları

### 6.1.1.Spss

Merkezi Chicago‟ da bulunan SPSS (Statistical Package for Social Sciences) 1967 yılından bu yana verilerdeki gizli bilgileri keşfetme ve stratejik karar desteği sağlama yönünde ileri analitik çözümler sunmaktadır. SPSS‟ in veri madenciliği metadolojisi olarak kabul ettiği CRISP DM (CRoss Industry Standart Processing for Data Mining) %50‟ nin üzerinde bir kullanıma sahiptir. İnternet kayıtlarına ve elde edilen verilere gelişmiş veri madenciliği teknikleri uygulanarak, kullanıcılar ile birebir ilişki kurmayı sağlayacak öngörüler elde edilebilir. Bu aşamada SPSS çözümlerine, teknolojilerine ve danışmanlığına başvurarak, güvenilir sonuçlar elde etme yolunda bir adım atmış oluruz. SPSS veri madenciliği çalışmalarına kendi yeteneğini ve tecrübesini getirerek, öğrenme süresini azaltacak, çalışmalara en hızlı şekilde başlamamızı sağlayacaktır. [Farboudi,2009]

### 6.1.2. Clementine

Clementine SPSS firmasının veri madenciliği için geliştirmiş olduğu bir modüldür. SPSS istatistiksel bir araçtır. Clementine'nin SPSS içinde bir modül olarak kullanılması kullanıcıların SPSS'in istatistiksel fonksiyonlarından faydalanmasına imkan verir. Yapay sinir ağları ve kural tümevarım yöntemlerini kullanır. Clementine müşteri hizmetleri yönetimi, kimya sektöründe maddelerin aşındırıcılık tahmininde ve bankacılık alanında kredi kartı dolandırıcılıkları gibi konularda kendine uygulama alanı bulmuştur.[Doğan,2007]

### 6.1.3. Sas

SAS‟ ın (Statistical Analysis Software) dünya çapında 112 ülkede 44000‟ i aşkın kullanıcısı bulunmaktadır. Kullanımı SPSS programına göre biraz daha zordur. SAS programında komut yazmak gerekir. Veriler üzerinde gerekli istatistik tekniklerini kullanarak tahmini sonuçlar verir. SAS araştırma, kamu, perakende, sigorta, bankacılık, medya, eğitim ve telekomünikasyon sektörlerinde kullanılmaktadır. [Farboudi,2009]

### 6.1.4. Enterprise miner

SAS firmasının veri madenciliği aracıdır. SAS'ın Veri ambarı ve ÇAİ (çevrimiçi analitik işleme) araçlarıyla bütünleşik çalışabilmektedir. Enterprise Miner karar ağaçları, yapay sinir ağları, regresyon analizi, 2-aşama modelleri (two-stage models), kümeleme, zaman serileri, ilişkilendirme, vb. veri madenciliği sorgularını ele alabilmektedir. Grafiksel arayüzü sayesinde kullanım kolaylığı sağlar ve kullanıcılar uygulamanın karmaşıklığından habersiz bir şekilde sadece girdi ve çıktılara yoğunlaşabilirler. 2 katmanlı mimariyi kullanır. İstemci bilgisayardaki yazılım gereksinimi Windows 98, 2000 ve NT'dir. Sunucu bilgisayardaki yazılım gereksinimi Windows 98, 2000 ve NT ile Linux'dır .[Doğan,2007]

### 6.1.5. Kxen

KXEN (knowledge extraction engine), bir bilgi çıkarma motorudur. Veri madenciliği motorunu kullanmak için veri madenciliği araçlarını veri işleme akışına kolaylıkla ekleyebilen tek programdır .[Farboudi,2009]

### 6.1.6. Insightful miner

Basit veri madenciliği projelerini yapacak olan sıradan çözümleyiciler için mevcut olan en iyi programlardan biridir. İnsightful miner, S-Plus kullananlar için iyi bir veri madenciliği aracıdır. Çünkü S-Plus fonksiyonlarının tüm kütüphaneleri bu programla kullanılabilir. Bu sayede Statistica Data Miner kadar olmasada zengin istatistiksel çözümleme algoritmalarına sahiptir [Farboudi,2009]

### 6.1.7. Affinium model

Bu program piyasadaki yanıt (response) modelleme ürünleri içinde kullanımı en kolay programdır. İstatistiksel ve grafiksel temeli zayıf olan veri madencisi ve istatistikçi olmayan kişiler için en iyi programdır. Diğer programların kullanıcılar tarafından elle yapılan bir çok veri madenciliği işlemleri, algoritma seçimi dahil modelleme motoru tarafından otomatik olarak yapılır. Kullanıcı sadece hızlıdan kapsamlıya kadar çözümleme seviyesini seçmelidir. Program mevcut en iyi modeli koruyarak, modelleri az sayıdan çok sayıda algoritma ve parametre kümelerinden kurar. Dört değişken modelleme uygulaması vardır; yanıt modelleyici, çapraz satıcı, müşteri bölümleyici ve müşteri değerlendiriciler. Bu uygulamalar fonksiyon açısından birbirine çok benzer, sadece modeli oluştururken kullanılan terimler açısından farklılık gösterir . [Farboudi,2009]

### 6.1.8. Statistica Data Miner

STATISTICA Data Miner, KXEN gibi kendine has kategoride bir programdır. Veri madenciliği projesindeki tüm görevleri kolaylaştırmadaki başarısı ve bir çok işlemi başarıyla gerçekleştirmesi açsından eşsizdir. Diğer programların kullanımı daha kolay olabilir (Insightful Miner gibi) ya da daha otomatik olabilirler (Affinium Model ya da KXEN gibi) ancak hiçbir veri madenciliği programı STATISTICA Data Miner kadar fazla araç sunamaz. [Farboudi,2009]

### 6.1.9. Inlen

İlişkisel veri tabanından aldığı verileri makine öğrenimi teknikleriyle işledikten sonra ortaya çıkan sonuçları Veri tabanına yazmaktadır. Üretilen bilgi kesimi, basit ya da bileşik olabilmektedir(Doğan,2007). INLEN aracında dört işleç vardır:

1. Veri tabanı yönetim işleci: Veri tabanı sorgularını yazmak için geliştirilen bir işleçtir.

2. Bilgi yönetim işleci: Üretilen bilgiyi yönetmek için kullanılır.

3. Bilgi üretim işleci: Veri tabanından bilgi almak ve makine öğrenimi algoritmalarını çağırmak için kullanılır.

4. Makrolar: INLEN işleçlerini bir sırada tanımlamayı ve tek bir işleç gibi kullanabilmeyi sağlar.

### 6.1.10. DBMiner

Kanada Simon Fraser Üniversitesi tarafından geliştirilen bir sistemdir. DBMiner sınıflama, kümeleme, eşleştirme ve sıra örüntüleri sorgularını yapabilecek veri madenciliği algoritmalarını kullanır. DBMiner çevrimiçi analitik işleme özelliğiyle veri madenciliği algoritmalarının bütünleşik çalışabilme özelliği sayesinde ön plana çıkmaktadır. Bu özellik OLAM (Online Analytical Mining) olarak anılır. DBMiner OLAP ve veri madenciliği yöntemlerini dinamik bir şekilde seçebilme imkânına sahiptir. Kullanıcının kolay kullanabileceği bir ara yüze sahiptir. Bu ara yüz sayesinde elde edilen sonuçlar çok yönlü bir soyutlama kullanılarak gösterilebilmektedir. (Doğan,2007)

### 6.1.11. Darwin

Darwin Oracle firmasının veri madenciliği aracıdır. Darwin regresyon ağaçları, karar ağaçları, kümeleme, yapay sinir ağları, Bayesian öğrenme, k-yakınlığında komşuluk gibi birçok algoritmayı destekleyen bir veri madenciliği aracıdır. Paralel sunucular için geliştirilmiş bir veri madenciliği sistemidir. Darwin kullanımı kolay bir ara yüze sahiptir. Darwin veri madenciliği algoritmalarından CART, StarTree, StarNet ve StarMatch'i kullanır. [Doğan,2007]

## 6.2. WEKA

WEKA (Waikato Environment for Knowledge Analyses), Waikato Üniversitesi tarafından geliştirilerek 1996'da ilk resmi sürümü yayınlanmış olan bir makine öğrenme ve veri madenciliği yazılımıdır. Akademik araştırmalar, eğitim ve endüstriyel uygulama alanlarında kullanım yeri olan WEKA, veri analizi ve tahminleyici modelleme için geliştirilmiş algoritma ve araçların görsel bir birleşimini içerir. Geliştirilen yazılımın temel avantajları geniş veri önişleme ve modelleme tekniklerine sahip olması, grafiksel kullanıcı ara yüzü sayesinde kullanımının kolay olması ve Java programlama dili ile uygulandığından herhangi bir platformda kullanılabilmesi yani taşınabilir olmasıdır.

WEKA bir proje olarak başlayıp bugün dünya üzerinde birçok insan tarafından kullanılmaya başlanan bir Veri Madenciliği uygulaması geliştirme programıdır. WEKA java platformu üzerinde geliştirilmiş açık kodlu bir programdır. WEKA çalıştırıldıktan sonra Şekil 6.1'de görüldüğü gibi, Application menüsünde çalışılabilecek modlar listelenmektedir. Bunlar komut modunda çalışmayı sağlayan Simple CLI, projeyi adım adım görsel ortamda gerçekleştirmeyi sağlayan Explorer ve projeyi sürükle bırak yöntemiyle gerçekleştirmeyi sağlayan KnowledgeFlow seçenekleridir. Explorer seçeneği seçildikten sonra üzerinde çalışılacak verilerin seçilmesi, bu veriler üzerinde temizleme ve dönüştürme işlemlerinin gerçekleştirilebilmesini sağlayan ekran ile karşılaşılmaktadır.

Arff, Csv, C4.5 formatında bulunan dosyalar WEKA'da import edilebilir. Herhangi bir text soyadaki verileri WEKA ile işlemek olanaksızdır.

Ayrıca Jdbc kullanılarak veritabanına bağlanıp burada da işlemler yapılabilir. WEKA'nın içerisinde Veri İşleme, Veri Sınıflandırma, Veri

Kümeleme, Veri İlişkilendirme özellikleri mevcuttur. Bu adımdan sonra yapılacak olan projenin amacına göre açılan sayfadaki uygun tabdaki (Sınıflandırma, Kümeleme, İlişkilendirme) uygun algoritma veya algoritmalar seçilerek veriler üzerine uygulanmakta ve en doğru sonucu veren algoritma seçilebilmektedir.


Şekil 6.1 Weka menüsü

WEKA paket programı yardımıyla elde edilmiştir. WEKA paket programında veri kümesi için sırasıyla Naive Bayes, Kstar, RBFNetwork, J.48, JRIP, Ridor algoritmaları seçilerek program çalıştırılmış ve elde edilen sonuçlar hazırlanmıştır.

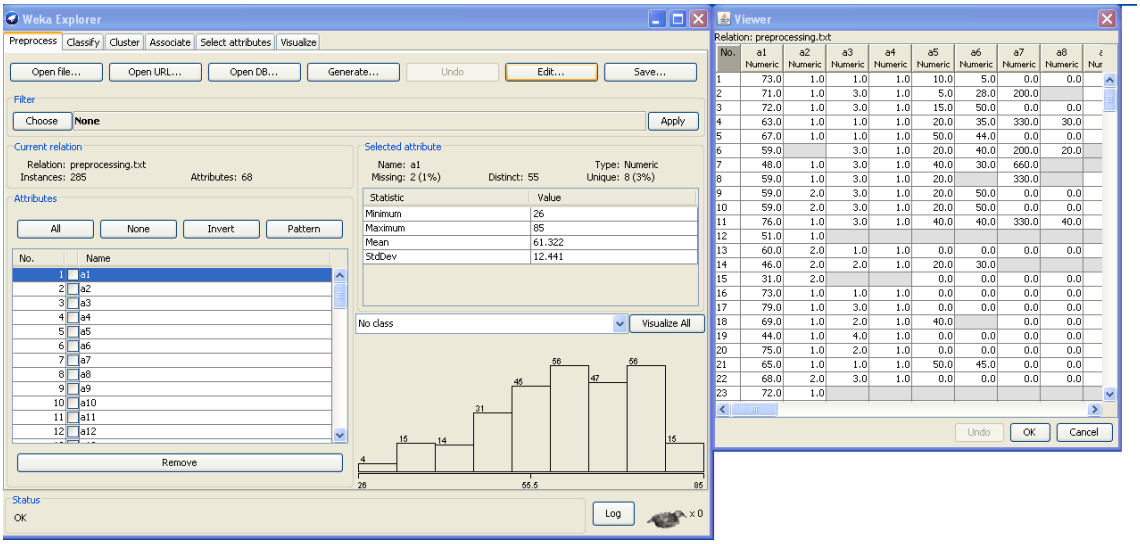
Ayrıca HyperPipes, VFI gibi birçok algoritma denenmiştir.

### 6.2.1. Veri Önişleme

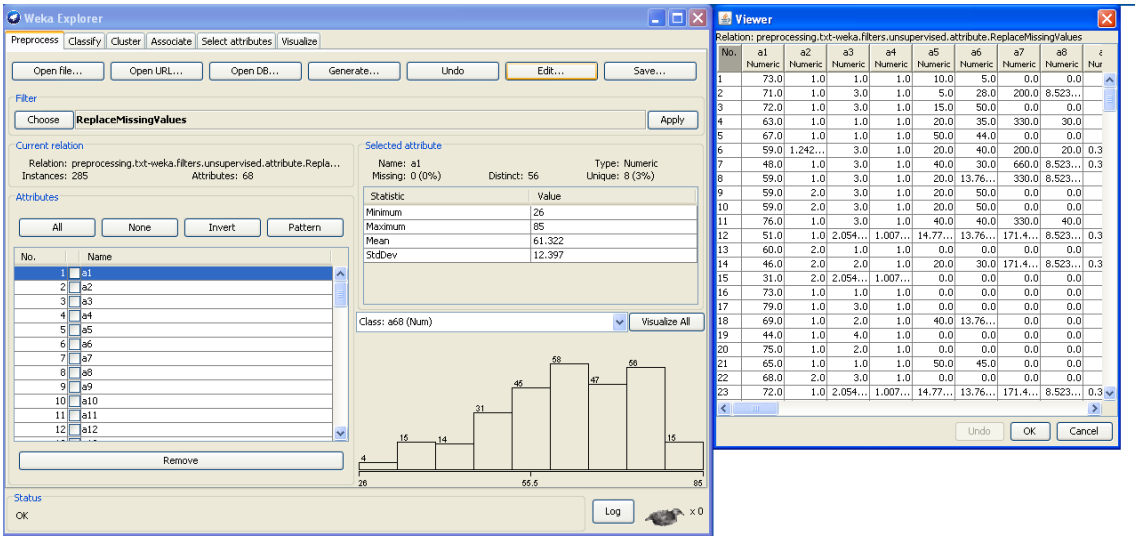
Algoritmaların karşılaştırılarak hangi algoritmanın daha iyi olduğunu bulmaya yönelik çalışmalara yapılan eleştirilerden biri uygulama sırasında yapılan veri önişleme basamağıdır. Bu adımda veri temizleme, veri birleştirme, veri dönüşümü, veri azaltma

yöntemleri kullanılarak, veri analize hazır hale getirilir. Bu işlemler oluşacak modelin başarımını etkileyebilir. Yapılan işlemler uygulamacının bakış açısına bağlıdır. Veri kümesi üzerinde yapılan bazı farklı müdahaleler farklı algoritmalarda farklı neticelere sebep olabilir. Yapılacak çalışmanın iyi sonuçlar üretmesi uygulamacının uygulama yapılan alan hakkında bilgili olmasını ya da bu alan uzmanlarıyla birlikte çalışmasını gerektirir

6.2.2. Kayıp Veriler Kayıp verilerin yaratacağı sorunları ortadan kaldırmak için kullanılan yöntemlere örnek olarak Replace MissingValues modülü kullanılmıştır (şekil 6.2). Bu yöntemle veritabanındaki kayıp değerler, ait oldukları niteliğin diğer değerlerinin ortalaması ya da moduyla değiştirilmektedir.



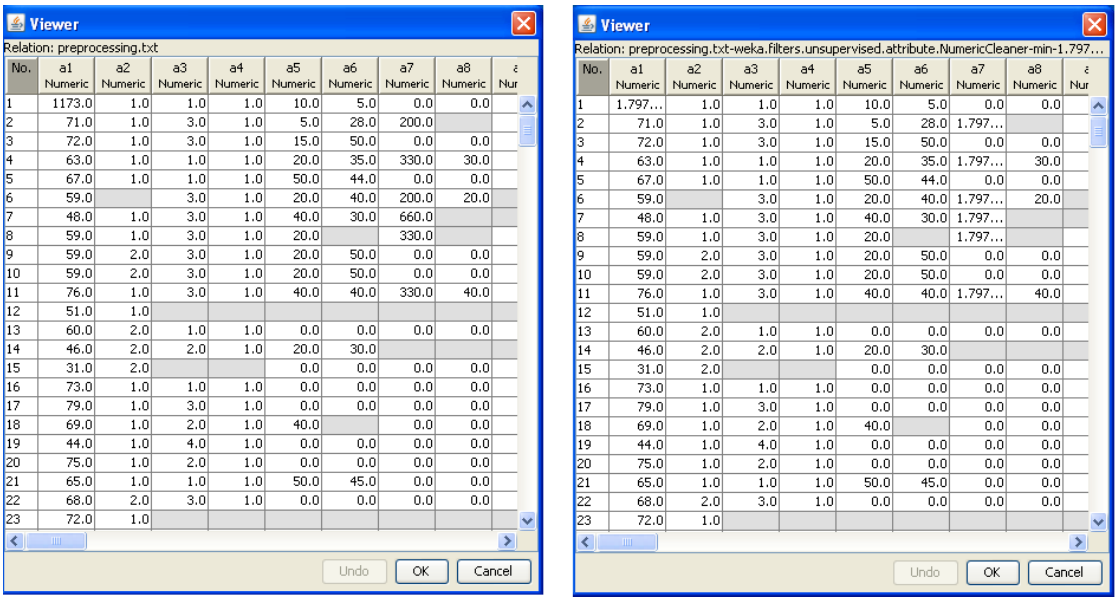

Şekil 6.2. Mevcut Veritabanındaki Kayıp Veriler



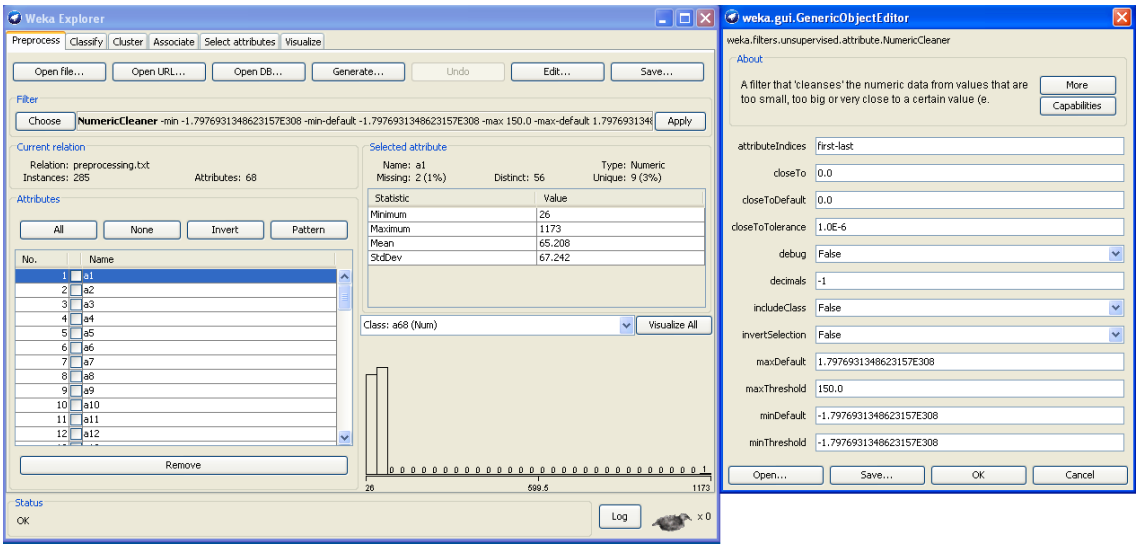

Şekil 6.3. ReplaceMissingValues Modülünün Kullanımı

### 6.2.3. Yanlış ya da Aşırı Uç Veriler

Bu tür veriler için ise Numeric Cleaner modülü ele alınmıştır (Şekil 6.4-6.5). Bu yöntem çok büyük, çok küçük ya da belli bir değere çok yakın değerlerin veritabanından silinerek bu değerlerin yerine önceden belirlenmiş başka bir değerin atanmasını içerir.



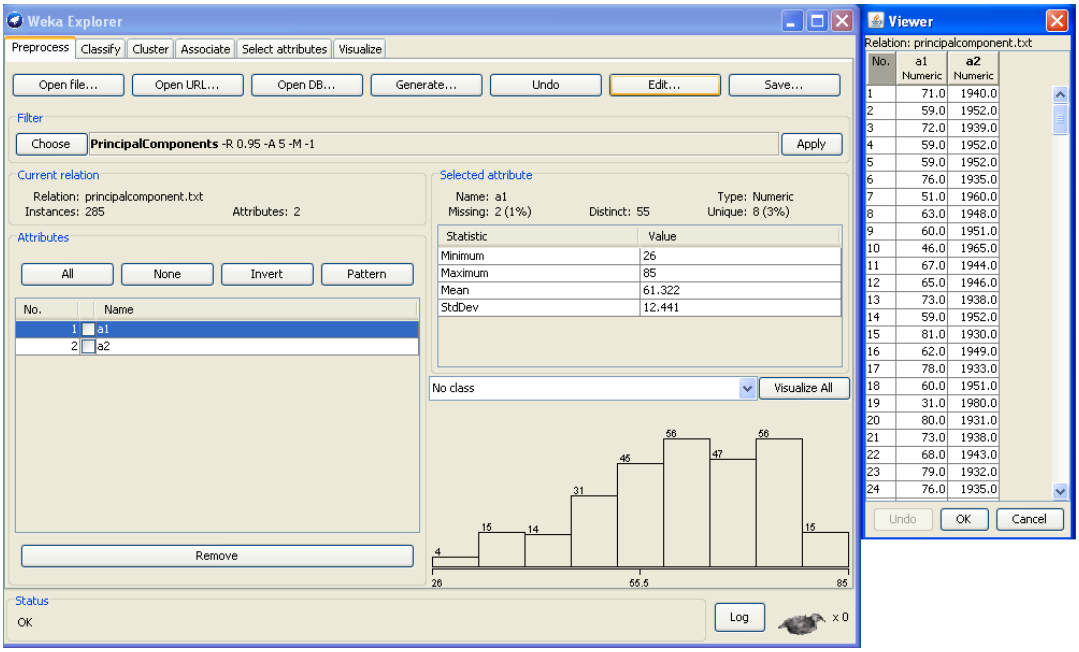
Şekil 6.4. Aşırı uç verilerin Numeric Cleaner Modülü Kullanılmadan Önce ve Sonraki Durumu göstermektedir.



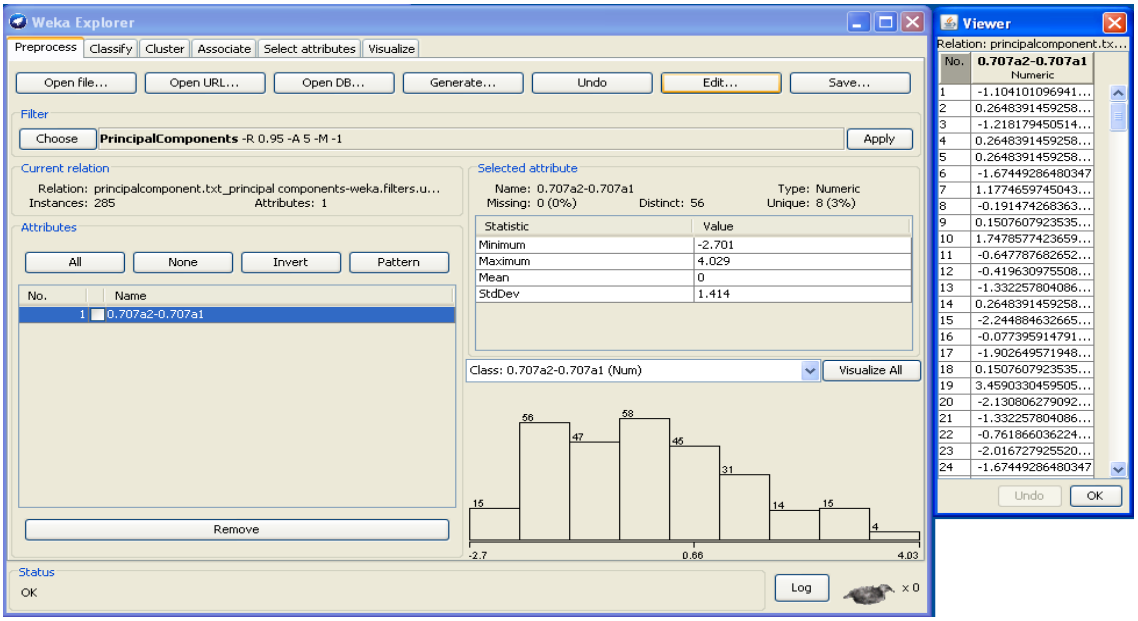

Şekil 6.5. Numeric Cleaner Modülünün Kullanımı

### 6.2.4. Gereksiz Veriler

Aynı veritabanı içinde hem yaş hem de doğum tarihi bilgisinin verilmesi durumunda oluşan gereksiz verilerin bilgisayar çalışma zamanını ve sonuçların kalitesini etkilememesi amacıyla Principal Components modülü kullanılarak veri boyutu azaltılmıştır.(Şekil 6.6-6.7).




Şekil 6.6. Principal Components Modülü Kullanılmadan Önceki Durum




Şekil 6.7. Principal Components Modülü ile Boyut indirgeme

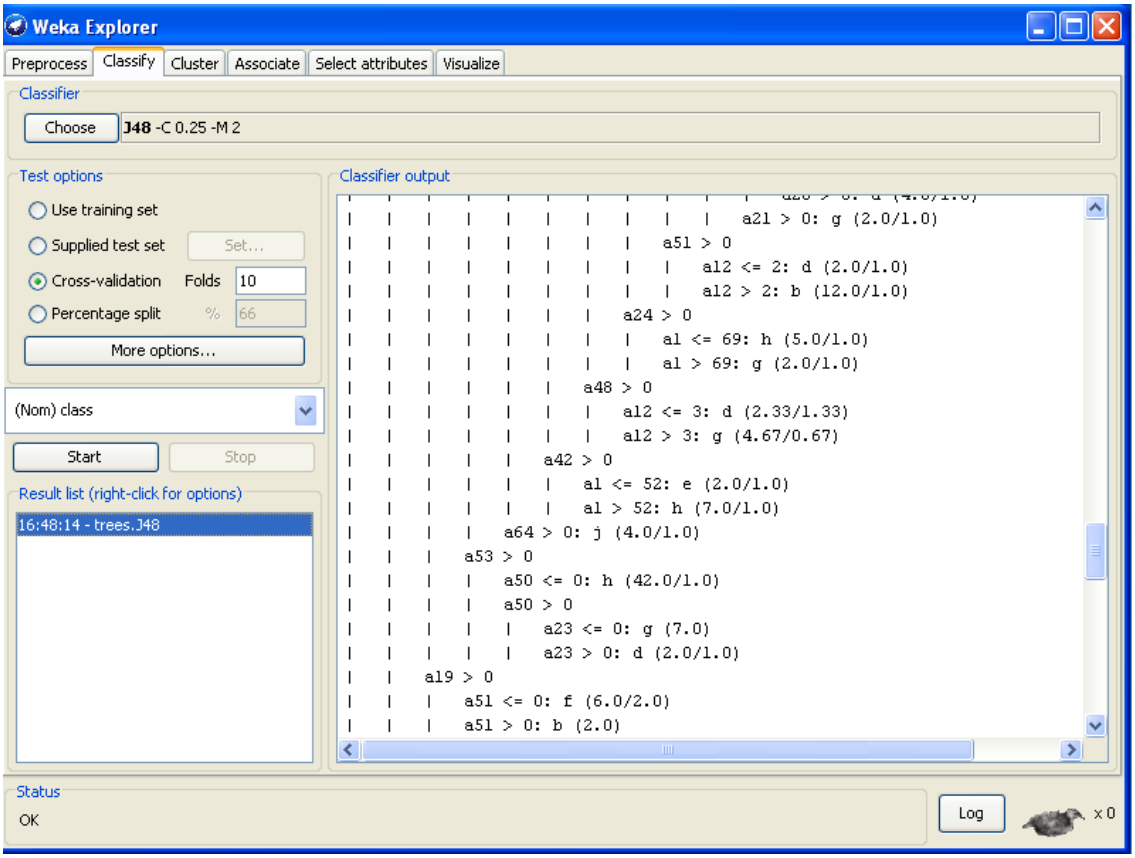
[TAPKAN, ÖZBAKIR, BAYKASOĞLU,2011]

### 6.2.5. Sınıflandırma

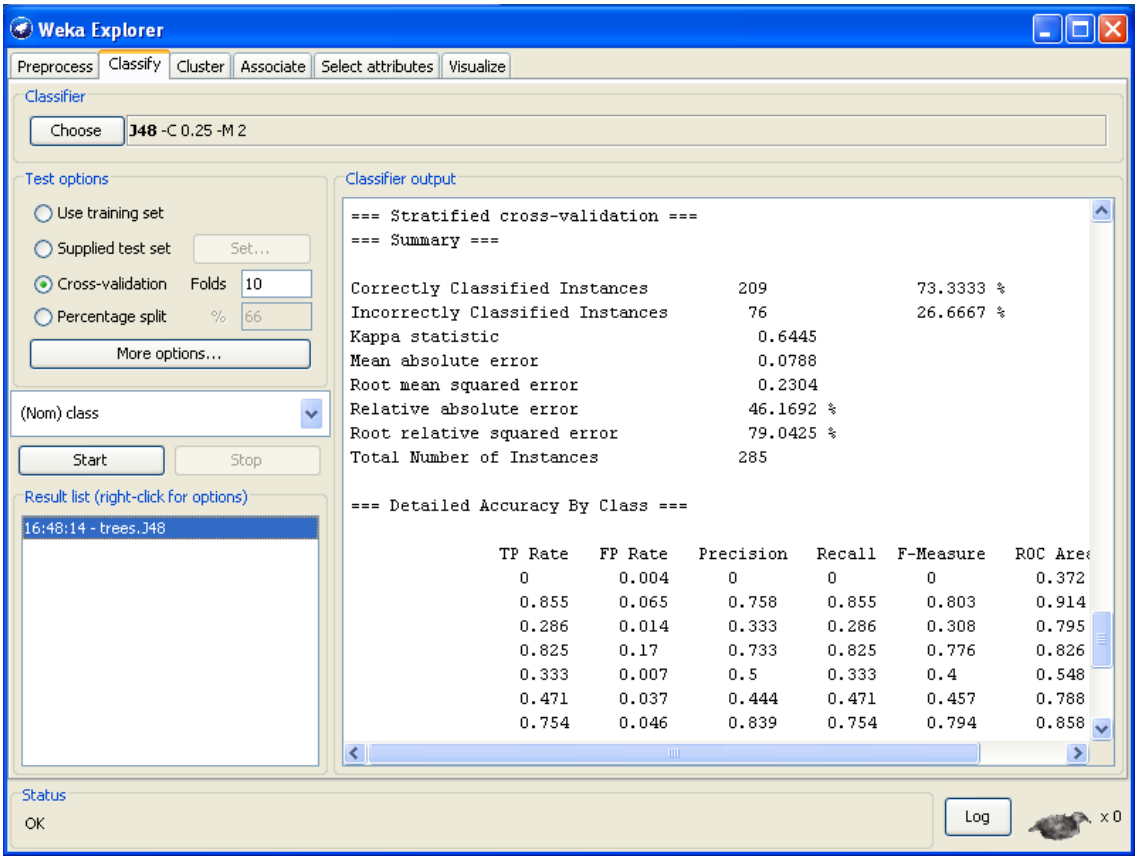
Sınıflandırma, bir örüntü tanıma sisteminin son aşamasında bulunmaktadır ve bu çalışmada Wisconsin göğüs kanseri örneklerine tanı (kötü huylu/iyi huylu tümör) koymak amacı ile kullanılmıştır. Doğrusal ayır taç analizi ile sınıflandırıcı oluşturulurken, gruplar arası varyansın en-çoklanması, grup içi var yans ortalamasının da en azlanması gerekmektedir. Bunun için bir en iyileştirme yapılmaktadır. Sınıflandırıcıların, kullanılmadan önce eğitilmeleri gerekmektedir.

Verilerin sınıflandırılması esnasında göz önünde bulundurulması gereken noktalardan biri de maliyetlerdir. Yanlış sınıflandırmanın bir maliyete tabi olduğu maliyete göre sınıflandırma yöntemlerinde belli bir hatalı sınıflandırmanın göreli önemi diğer hatalı sınıflandırmalardan daha fazla olabilmektedir. Bu bağlamda WEKA'da maliyete göre sınıflandırma yapan Cost Sensitive Classifier modülü kullanılanılır.. Bu yöntemin amacı beklenen yanlış sınıflandırma maliyetini minimize edecek en iyi sınıflandırmayı tahmin etmektir.

Sınıflandırma algoritmaları içerisinden C4.5 karar ağacına dayanan J48 modülü, uygulaması gerçekleştirilmek üzere seçilmiştir (Şekil 6.8-6.9).




Şekil 6.8. J48 Modülünün Kullanımı ile Elde Edilen Kurallar




Şekil 6.9. J48 Modülü ile Elde Edilen Sınıflandırma Doğrulukları

[TAPKAN ÖZBAKIR BAYKASOĞLU,2011]

### 6.2.5.1 Öznitelik Seçimi

Sınıflandırmada kullanılacak öznitelik vektörünü oluşturmak için, toplam 9 öznitelikten en iyi olanlar ($p < 0.001$), t-test ile kontrol edilmiştir. Burada p-seviyesi, gözlenen sonucun geçerliliğindeki hata olasılığını vermektedir. Tüm öznitelikler t-test'de başarılı olmuştur. Bu vektör içinden en iyi ayrım gücüne sahip öznitelik kombinasyonunu bulmak ve vektör boyutunu azaltmak için ardışıl ileri seçim yöntemi [ Jain, A.., Duin, R.P.W,2000] kullanılmıştır. Ardışıl ileri seçim yönteminde öncelikle en iyi öznitelik seçilmiş, daha sonra, her defasında bir öznitelik eklenerek kriter

fonksiyonunu en iyileştiren öznitelik bileşimi bulunmuştur. Elde edilen vektör 5 bileşene sahiptir ve bu öznitelikler, en iyiden başlayarak, hücre çekirdeğinin çevre sitoplazmaya oranı, topluluktaki hücre boyutlarının eşitliği, Hücre topluluğu kalınlığı, normal çekirdekçik ve kromatin dağılımı olarak sıralanmaktadır.

#### 6.2.5.2 . Sınıflandırma Algoritmalarının Karşılaştırılmasında Önemli Hususlar

Veri önişleme, parametre seçimi ve test kümesi seçimi veri madenciliği uygulama-sında ortaya çıkacak olan modelin başarımını etkiler. Dolayısı ile yapılan karşılaştırma sonuçları büyük ölçüde uygulamacıya bağlıdır.

### 6.3. Veri Önişleme

Algoritmaların karşılaştırılarak hangi algoritmanın daha iyi olduğunu bulmaya yönelik çalışmalara yapılan eleştirilerden biri uygulama sırasında yapılan veri önişleme basamağıdır. Bu adımda veri temizleme, veri birleştirme, veri dönüşümü, veri azaltma yöntemleri kullanılarak, veri analize hazır hale getirilir. Bu işlemler oluşacak modelin başarımını etkileyebilir. Yapılan işlemler uygulamacının bakış açısına bağlıdır. Veri kümesi üzerinde yapılan bazı farklı müdahaleler farklı algoritmalarda farklı neticelere sebep olabilir. Yapılacak çalışmanın iyi sonuçlar üretmesi uygulamacının uygulama yapılan alan hakkında bilgili olmasını ya da bu alan uzmanlarıyla birlikte çalışmasını gerektirir.

### 6.4. Parametre Seçimi

Veri madenciliğinde kullanılan farklı algoritmaların farklı parametreleri olabilir. Örneğin yapay sinir ağlarında gizli nöron sayısı, karar ağaçlarındaki budama işleminin parametreleri, algoritmaların kullanacağı parametrik değerleri belirler. Bu parametreler

algoritmadan algoritmaya değişebilir, ya da kullanılan veri madenciliği araç programlarında farklı olabilir. Bunların seçimi oluşacak olan modelin başarımını etkileyecektir.

### 6.5. Test Kümesinin Seçimi

Model oluşturulurken kullanılan öğrenme ve test kümelerinin belirlenmesinin de modelin başarımı üzerinde etkisi vardır. Eldeki verinin öğrenme kümesi ve test kümesi olarak ayrılmasında farklı metotlar kullanılabilir. Kullanılan veri madenciliği programında bu işlem için farklı seçenekler bulunabilir. Öğrenme kümesi ve test kümesi farklı dosyalardan programa verilebileceği gibi, programın bir veri dosyasını belirtilen bir oranda test kümesi olarak kullanması ya da n-fold metodu ile programın veri kümesini n sayıdaki parçalara ayırarak sırayla her parçayı test kümesi olarak kullanması sağlanabilir.

### 6.6. Model Başarım Ölçütleri

Model başarımını değerlendirirken kullanılan temel kavramlar hata oranı, kesinlik, duyarlılık ve F-ölçütüdür. Modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atılan örnek sayısı nicelikleriyle alakalıdır.

Test sonucunda ulaşılan sonuçların başarım bilgileri karışıklık matrisi ile ifade edilebilir. Karışıklık matrisinde satırlar test kümesindeki örneklere ait gerçek sayıları, kolonlar ise modelin tahmin edilmesini ifade eder [COŞKUN1, BAYKAL2,2008].

Tablo 6.1. İki sınıflı bir veri kümesinde oluşturulmuş modelin karışıklık matrisi

| | | Öngörülen Sınıf | |
|---|---|---|---|
| | | Sınıf =1 | Sınıf=0 |
| Doğru Sınıf | Sınıf=1 | a (TP) | b (FN) |
| | Sınıf=0 | c(FP) | d (TN) |

a: TP(True Pozitif) c: FP(False Pozitif)

b: FN(False Negatif) d: TN(True Negatif)

### 6.6.1. Doğruluk – Hata oranı

Model başarımının ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk oranıdır. Doğru sınıflandırılmış örnek sayısının (TP +TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır. Hata oranı ise bu değerin 1'e tamlayanıdır. Diğer bir ifadeyle yanlış sınıflandırılmış örnek sayısının (FP+FN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır. [COŞKUN1, BAYKAL2,2008]

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

### 6.6.2. Kesinlik

Kesinlik, sınıfı 1 olarak tahminlenmiş True Pozitif örnek sayısının, sınıfı 1 olarak tahminlenmiş tüm örnek sayısına oranıdır [COŞKUN1, BAYKAL2,2008]

$$\text{Kesinlik} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

### 6.6.3. Duyarlılık

Doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranıdır. [COŞKUN1, BAYKAL2,2008]

$$\text{Duyarlılık} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

### 6.6.4. F-Ölçütü

Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır.

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

# BÖLÜM 7.

# UYGULAMA: MEME KANSERİ VERİLERİNİN SINIFLANDIRILMASI

## 7.1.Kullanılan Meme kanseri-Wisconsin Veri Kümesi Özeti

Veri madenciliği de dijital ortamdaki büyük veri yığınlarından bilgi çıkarmayı ve bu bilgiyi değerlendirmeyi amaçlar.

Bu amaçla yapılan çalışmalarda karşılaşılan en büyük problem veri kaynağının hatalı veriler içermesi ya da çok sayıda nitelik değerinin eksik girilmiş olmasıdır. Meme kanseri-Wisconsin veri kümesi veri kaynağı son derece düzenli ve dünya çapında bilinen; istatistiksel çalışmalarda yoğun bir şekilde kullanılan veri kaynağıdır. Veri kaynağındaki nitelikler belirli bir format dahilindedir ve niteliklerin açıklamaları veri kaynağı ile birlikte kullanıma verilmektedir. Bu sebeplerle -Wisconsin veri kümesinin böyle bir uygulamada kullanılması uygulamanın sonuçlarını daha güvenilir kılacaktır.

Veri kaynağı ne kadar düzenli ve güvenilir olursa olsun, bir veri madenciliği uygulamasında kullanabilmek için veriler üzerinde önişlem yapmak gerekir. Bu çalışmada da kullanmış olduğum veri kaynağı üzerinde işlemler yapılmış, veri kaynağı analize uygun yapıya dönüştürülmüştür. Veri önişleme basamağında yapılan en önemli çalışma verinin temizlenmesi, analizde kullanılmayacak gereksiz bilgilerin veriden çıkarılarak verinin düzenlenmesi işlemidir. Yaptığım çalışmada kullandığım veri kaynağı çalışma öncesinde incelenmiş ve analizde hatalı ve eksik olan bazı nitelikler veriden

Kanser Wisconsin veri kümesi, veri tabanında verilerin kronolojik gruplandırmasını yansıtır. Literatürde veri kaynağı ile uyumlu olması için veri kümesi eksik değerlerle sahip 16 hastanın veri değerleri kaldırılmış 683 hasta verisi ile yeni bir veri kümesi oluşturulmuştur.

Meme-kanseri-Wisconsin alt dizinde bulunan dosya 699 hasta örneği içermektedir.

Bu özellikler dokuz farklı skalada dış görünüm ve kromozom değişikliklerini ölçer. Bütün veriler 1 ile 10 arasında değişen değerlere sahiptir.

9 tamsayı özellikleri öznitelik bilgileri aşağıdaki gibidir:

Alanlar

1. Clump Thickness ,Clump Kalınlığı 1 - 10 1 - 10
2. Uniformity of Cell Size, Hücre Boyutu Düzenliliği 1 - 10 1 - 10
3. Uniformity of Cell Shape, Hücre Şekli Uniformity 1 - 10 1 - 10
4. Marginal Adhesion, Marjinal Yapışma 1 - 10 1 - 10
5. Single Epithelial Cell, Size Tek Epitel Hücre Boyutu 1 - 10 1 - 10
6. Bare Nuclei ,Çıplak çekirdeklerin 1 - 10 1 - 10
7. Bland Chromatin, Bland Kromatin 1 - 10 1 - 10
8. Normal Nucleoli ,Normal nükleol 1 - 10 1 - 10
9. Mitoses ,Mitoz 1 - 10 1 - 10
10. Sınıflar

- **2 for benign**, iyi huylu
- **4 for malignant** ,huylu

Bu bir meme içinde bir kitle (kanser olmayan) malign (kanser) veya benign olup olamayacağını değerlendirmek için ince iğne aspirasyonları bir patoloji raporunda kullanılan terimlerdir.

Wisconsin veri kümesi farklı kanser gruplarını içeren ve bilimsel araştırmalarda son derece önemli bir yer tutan, güvenilir, dokümante edilmiş, eşine az rastlanır bir veri kümesidir. National Cancer Institute (NCI)'in sağladığı Amerika Birleşik Devletleri'nin belli başlı coğrafi bölgelerini kapsayan, nüfusunun %26'sını ilgilendiren ve bu kanser vakaları hakkında istatistiksel önem taşıyan bilgiler içerir. Yıllık olarak güncellenen bu veritabanı bilimsel çalışma yapanlara, sağlık sektöründe çalışanlara, halk sağlığı konusunda görevli kurumlara açık bir veri kaynağı olup, binlerce bilimsel çalışmada kaynak olarak kullanılmıştır. Veri kaynağı, kurumun web sitesinden veri kullanma talep formu doldurularak imzalandıktan sonra elektronik olarak indirilebilir.

1973 yılı itibarı ile başlanmış olan kanser verileri farklı yılları kapsayan, farklı tümör tiplerine göre gruplar altında metin formatında, 118 nitelikten oluşan, oldukça büyük veri kaynağıdır. Bazı nitelikler daha önceki yıl verilerinde yokken sonraki verilerde eklenmiş, bazı niteliklerin sonraki yıllarda değerleri alınmamış, bazı nitelikler farklı bir tümör tipinde değer taşırken bazı tiplerde bir anlam ifade etmediği için değer kullanımdan kaldırılmıştır. Her ne kadar bu veri kaynağı oldukça düzenli ve dokümante edilmiş olsa da yaptığım çalışma için bir önişlemden geçirilmesi gerekmiştir.

Bu çalışmada yıllık olarak güncellenen Wisconsin veri kümesi kaynağının 1991 yılına ait olan versiyonu kullanılmıştır.

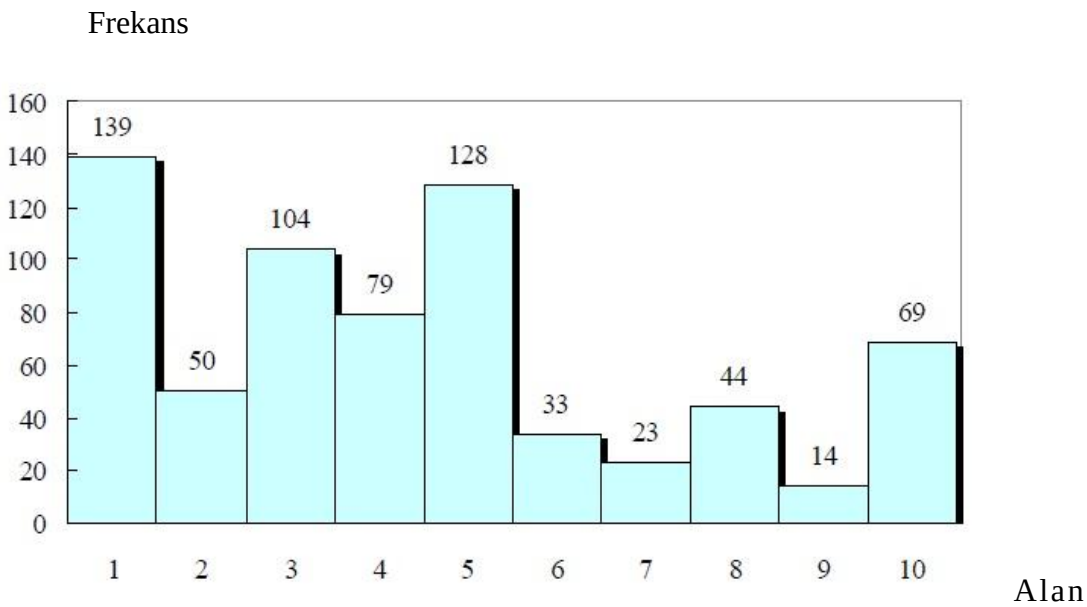
Bu bir meme içinde bir kitle (kanser olmayan) malign (kanser) veya benign olup olamayacağını değerlendirmek için ince iğne aspirasyonları bir patoloji raporunda kullanılan terimlerdir. Örneğin kanser hücreleri boyut ve şekil değişir eğilimindedir. Bu yüzden iyi huylu bir yönde hücre boyutu / şekil puan tekdüzelik. Ayrıca çıplak çekirdekler, mülayim kromatin ve normal nükleol sevecenlik işaretleridir.

Analizi muhtemelen bir üçlü test parçasıdır, diğer testler, bir veya her iki kanser işaret ederse, histolojik analiz için bir biyopsi gerekli olacaktır. Bu cerrahi işlem

sırasında frozen analizi yapılabilir olabilir. Wisconsin Meme-kanseri-alt dizinde bulunan 683 hastanın öznitelik değerleri başlıklarıyla beraber ayrıca grafik olarak da açıklanmıştır.

### 7.2. Clump Kalınlığı (Clump thickness)

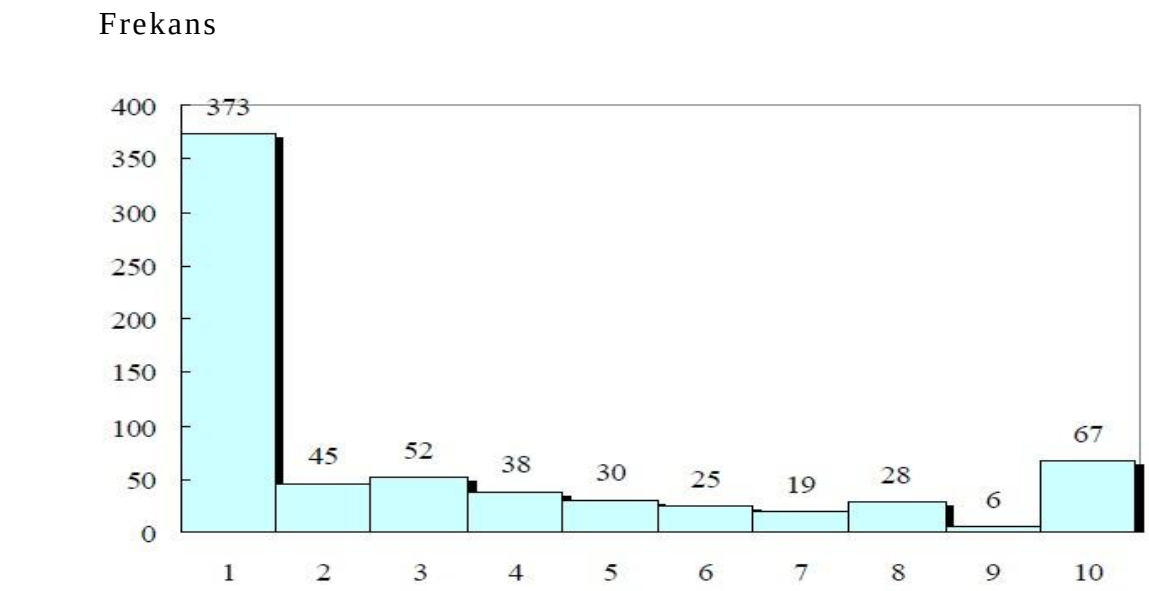
Clump kalınlığı: İyi huylu hücreler kanserli hücreleri genellikle çok katmanlı gruplandırılmıştır ederken, tek katmanlarda gruplandırılmasını eğilimindedir. Jung-Ying Wang [Jung-Ying Wang 2003]




Şekil 7.1. Clump kalınlığı

### 7.3. Hücre Boyutu Düzenliliği

Kanser hücrelerinin boyutu değişme eğilimindedir. Bu parametreler hücrelerin kanserli olup olmadığını belirlemede önemli yer teşkil eder. Jung-Ying Wang [Jung-Ying Wang 2003]

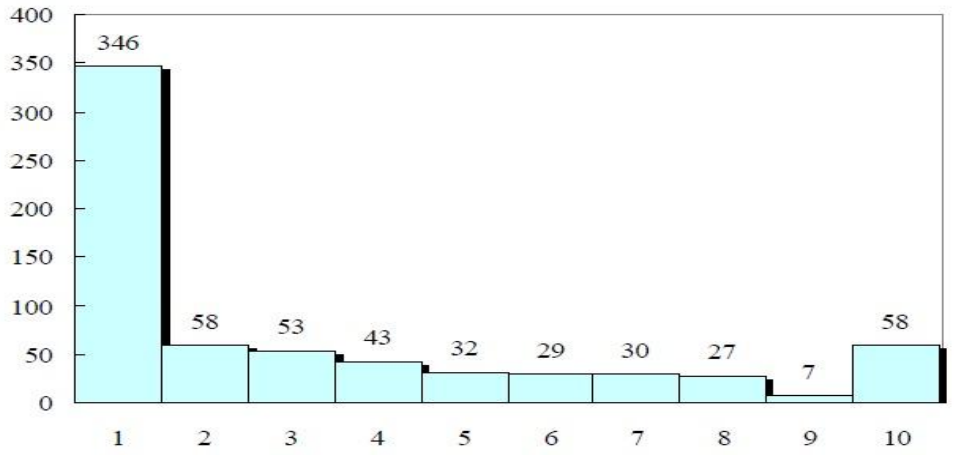



Şekil 7.2. Hücre boyutu Düzenliliği

### 7.4.Hücre Şekil Düzenliliği

Kanser hücrelerinin şekli(sınırların belirgin olup olmaması yüzeyin pürüzsüz olup olmaması) de değişme eğilimindedir. Bu parametreler hücrelerin kanserli olup olmadığını belirlemede önemli yer teşkil eder.

Frekans



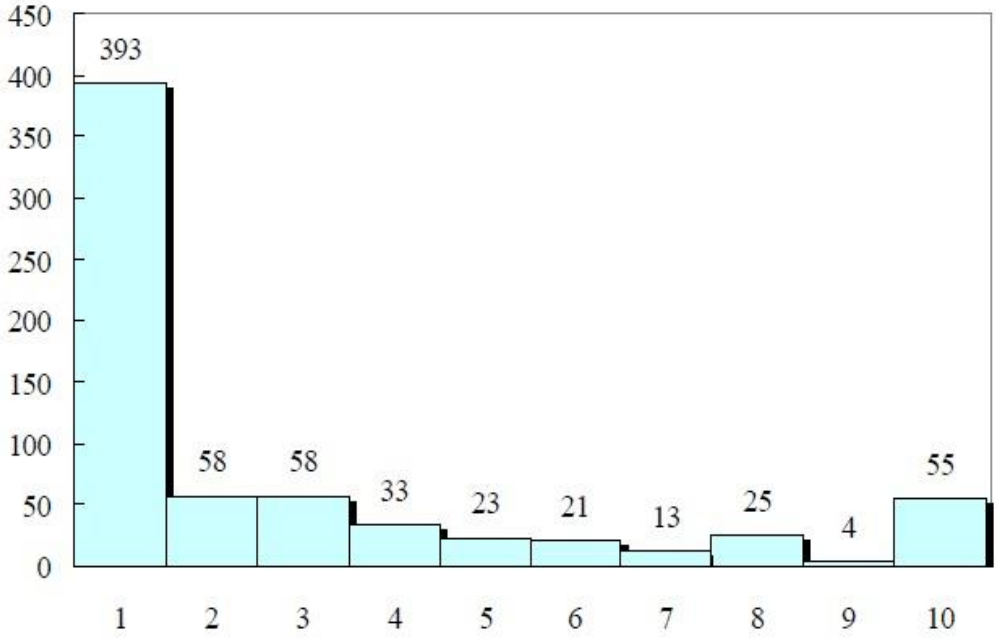

Alan

Şekil 7.3. Hücre şekil düzenliliği

### 7.5. Marjinal Yapışma

Normal hücreler birbirine yapışma eğilimi gösterirler. Kanser hücreleri bu yeteneği loos'un eğilimindedir. Bu yüzden yapışma kaybı malignite işaretidir. Jung-Ying Wang [Jung-Ying Wang 2003]

Frekans

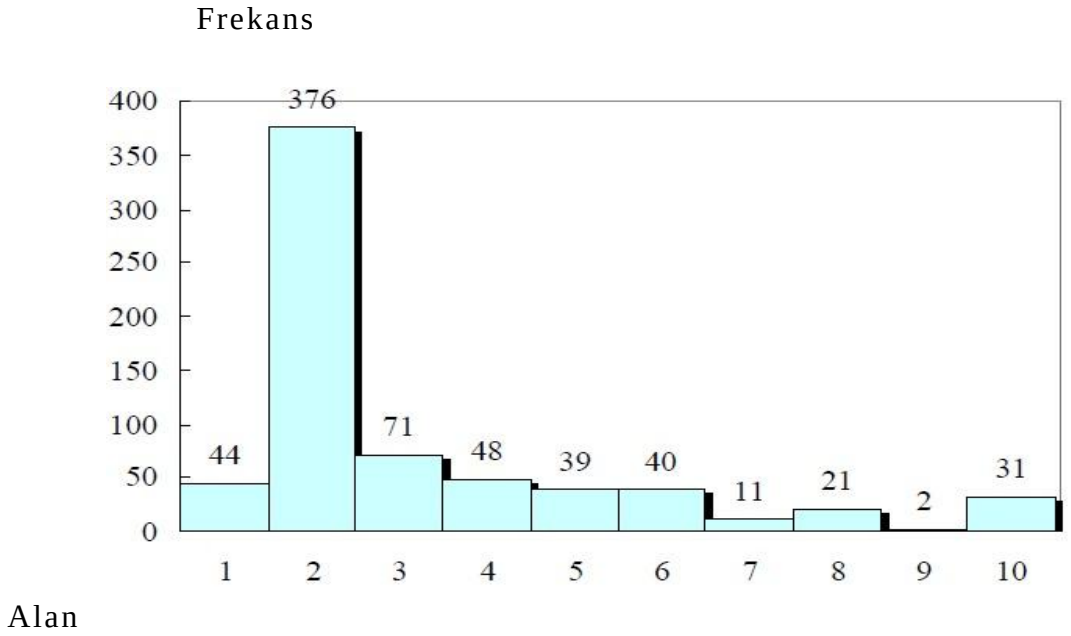



Alan

Şekil 7.4. Marjinal yapışma

### 7.6. Tek Epitel Hücre Boyutu

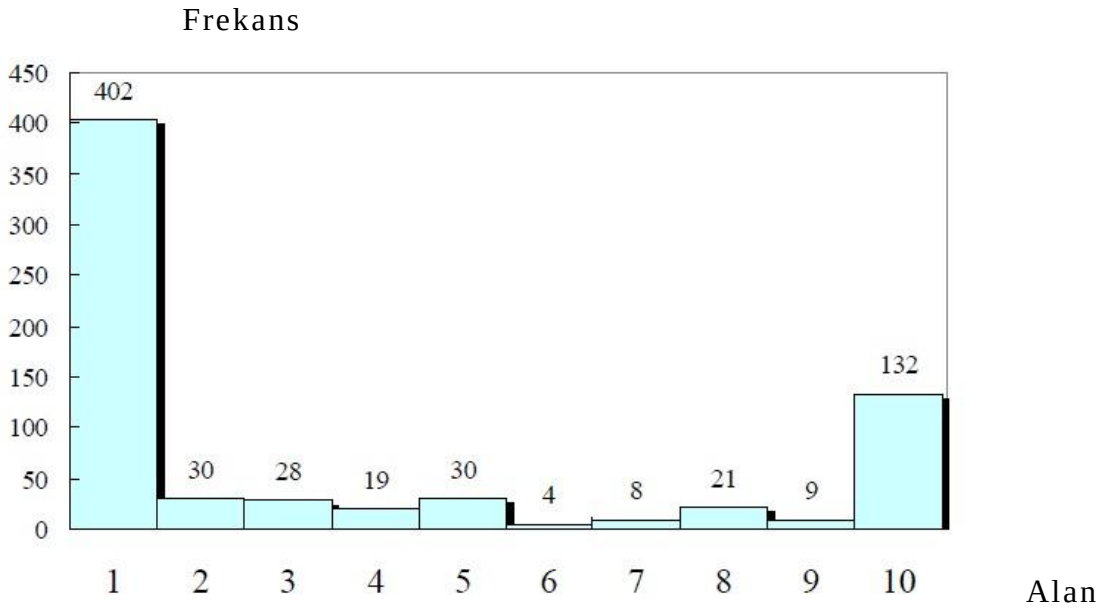
Tekdüzelik ile ilgili midir yukarıda. Önemli ölçüde büyütüldüğü epitel hücreleri habis bir hücre olabilir.




Şekil 7.5. Tek epitel hücre boyutu

### 7.7.Çıplak Çekirdekler

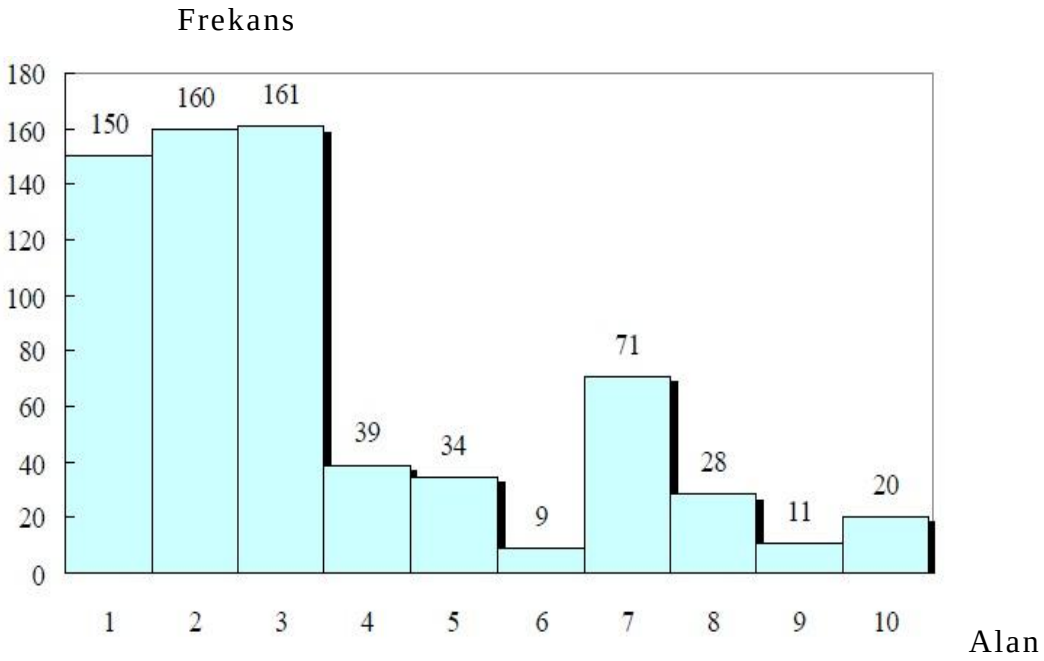
Bu sitoplazması (hücrenin kalanı) ile çevrelenmiş değildir çekirdekleri için kullanılan bir terimdir. Bunlar genellikle iyi huylu tümörler görülür.




Şekil 7.6. Çıplak çekirdekler

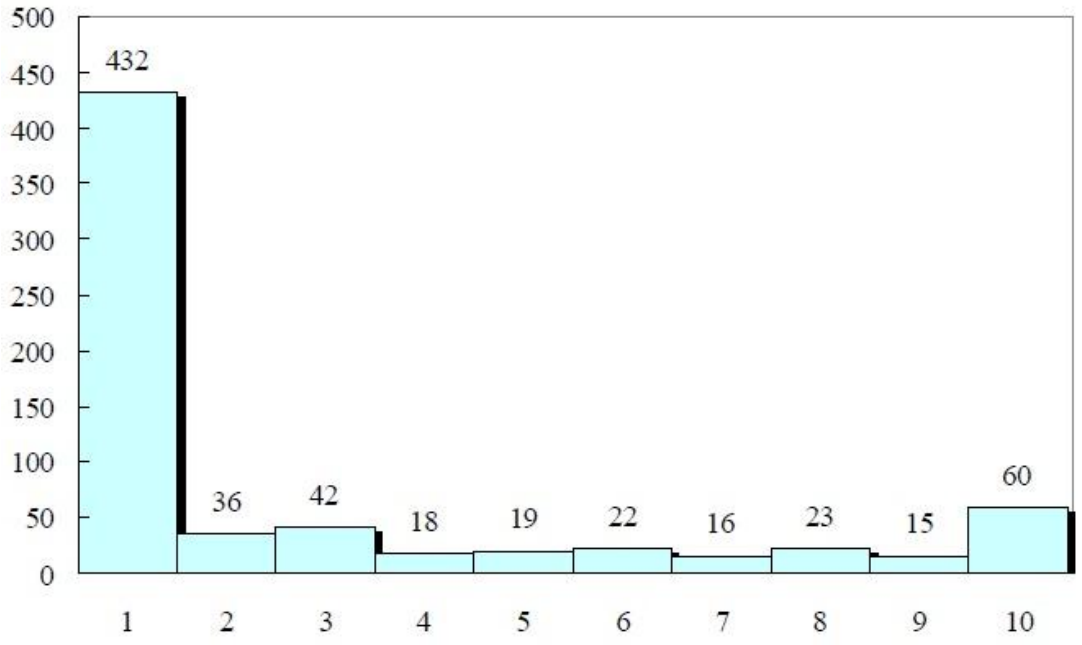
### 7.8. Bland Kromati

Huylu hücrelerde görülen çekirdek yeknesak bir "doku" açıklar. Kanser hücrelerinde kromatin daha iri olma eğilimindedirler.



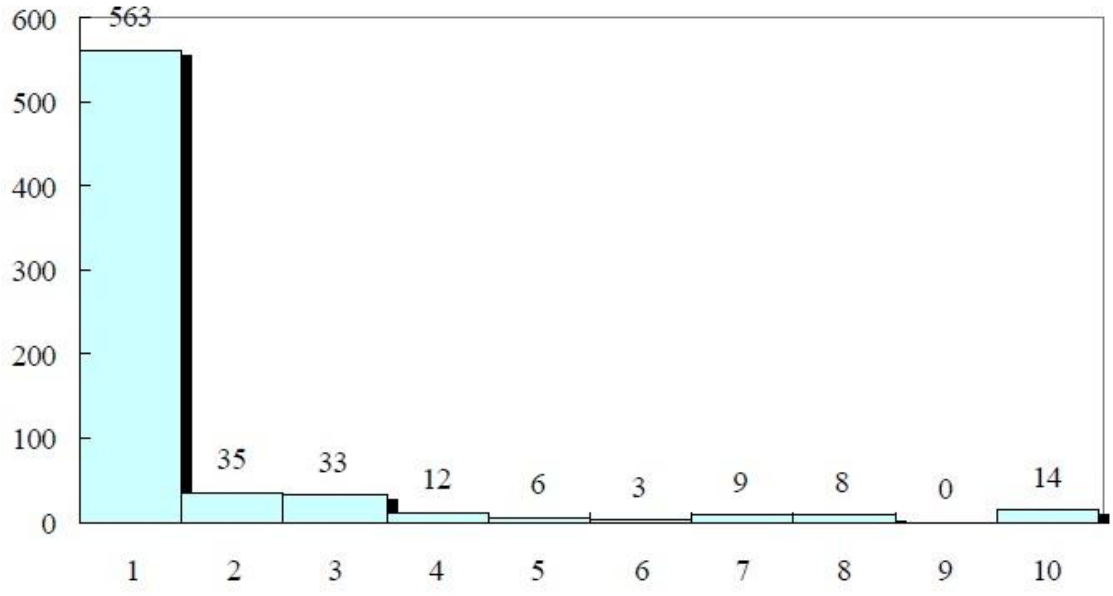

Şekil 7.7. Bland Kromati

**7.9. Normal Nükleol:** nükleol çekirdeğinde görülen küçük yapılardır. Hiç görünüyorsa, normal hücrelerde çekirdekçik genellikle çok küçüktür. Kanser hücreleri nükleol daha belirgin hale gelir ve bazen onları daha vardır. Jung-Ying Wang [Jung-Ying Wang 2003]




Şekil 7.8. Normal nükleol

**7.10.Mitoz**




Şekil 7.9. Mitoz

Tablo 7.1. Wisconsin Meme-kanseri-alt dizinde bulunan 683 hastanın öznitelik değerleri .

| ARALIK | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | TOPLAM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Clump kalınlığı | 139 | 50 | 104 | 79 | 128 | 33 | 23 | 44 | 14 | 69 | 683 |
| Hücre boyutu Düzenliliği | 373 | 45 | 52 | 38 | 30 | 25 | 19 | 28 | 6 | 67 | 683 |
| Hücre şekil düzenliliği | 346 | 58 | 53 | 43 | 32 | 29 | 30 | 27 | 7 | 58 | 683 |
| Marjinal yapışma | 393 | 58 | 58 | 33 | 23 | 21 | 13 | 25 | 4 | 55 | 683 |
| Tek epitel hücre boyutu | 44 | 376 | 71 | 48 | 39 | 40 | 11 | 21 | 2 | 31 | 683 |
| Çıplak çekirdekler | 402 | 30 | 28 | 19 | 30 | 4 | 8 | 21 | 9 |  | 683 |
| Bland Kromatin | 150 | 160 | 161 | 39 | 34 | 9 | 71 | 28 | 11 | 20 | 683 |
| Normal nükleol | 432 | 36 | 42 | 18 | 19 | 22 | 16 | 23 | 15 | 60 | 683 |
| Mitoz | 563 | 35 | 33 | 12 | 6 | 3 | 9 | 8 | 0 | 14 | 683 |
| TOPLAM | 2843 | 850 | 605 | 333 | 346 | 192 | 207 | 233 | 77 | 516 |  |

[JUNG-YİNG WANG ,2003]

# BÖLÜM 8.

# WEKA KULLANILARAK MEME KANSERİ HÜCRELERİNİN TAHMİNİ.

Veri önişleme aşamasında yaptığım nitelik seçimi, verilerin tamamlanması gibi analiz sonuçlarını etkileyici işlemlerde farkında olmadan model çıkarımını etkileyici işlemler yapılmış olabilir. Farklı ön işlemelerle oluşturulan verilerin analiz sonuçlarının farklı olması kaçınılmazdır.

Uygulamada meme kanseri hastalarının kayıtları incelenmiş, hastaların hayatta olup olmadıkları, hayatta değil iseler ne kadar süre hayatta kaldıkları ve ölüm sebepleri göz önünde tutularak herhangi bir hastanın hastalığı yenip yenemeyeceği sınıflandırılarak ileriye dönük tahminlerde bulunabilme amacı ile farklı algoritmalarla oluşturulan modellerin başarım dereceleri karşılaştırılmıştır.

Uygulamada bir karar ağacı algoritması olan ve temeli ID3 ve C4.5 algoritmalarına dayanan J48, istatistiksel bir algoritma olan Bayes sınıflandırma algoritmalarından Naive-Bayes, regresyon tabanlı algoritmalardan lojistik regresyon ve örnek tabanlı sınıflandırma algoritmalarından Kstar algoritmaları kullanılarak modeller oluşturulmuş ve oluşturulan modellerin başarım dereceleri karşılaştırılmıştır.

## 8.1.Karar Ağacı Modelinin Başarım Ölçütleri

Weka programının okuyabileceği Arff formatına çevrildikten sonra Weka ara yüzünde de önişlemden geçirilmiştir.

Önişlemler sonucunda elde edilmiş olan Arff formatındaki 683 kayıt içeren meme kanseri hastalıkları Wisconsin veri kaynağı üzerinde karar ağaçları, bayes, regresyon, örnek tabanlı sınıflandırma modellerinden birer algoritma seçilerek bunların başarıları karşılaştırılmıştır. Karşılaştırılacak algoritmalar seçilirken bu algoritmaların popülerliği ve literatürde benzer konuda yapılan çalışmalar dikkate alınmıştır.

Weka aracında model oluştururken kullanılabilecek pek çok karar ağaçları algoritmaları mevcuttur. ADTree, BFTree, Decision Stump, FT, J48, J48graft, LADTree, LMT, RBTree, RandomForest, RandomTree, RepTree algoritmalarından J48 algoritması işlenmiş olan Wisconsin veri kaynağı üzerinde çalıştırılmıştır. Tablo 8.1'de oluşturulan modelin test sonuçlarına ait istatistikler ve karışıklık matrisi görülmektedir. Tablo 8.2'de ise karışıklık matrisini kullanarak hesaplanan karşılaştırma ölçütleri verilmiştir.

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

uniformity <= 2
|   bareNuclei <= 3: 2 (394.0/2.0)
|   bareNuclei > 3
|   |   clump <= 3: 2 (11.0)
|   |   clump > 3
|   |   |   blandChromatin <= 2
|   |   |   |   marginalAdhesion <= 3: 4 (2.0)
|   |   |   |   marginalAdhesion > 3: 2 (2.0)
|   |   |   blandChromatin > 2: 4 (8.0)
uniformity > 2
|   uniformityofcellshape <= 2
|   |   clump <= 5: 2 (19.0/1.0)
|   |   clump > 5: 4 (4.0)
|   uniformityofcellshape > 2
|   |   uniformity <= 4
|   |   |   bareNuclei <= 2
|   |   |   |   marginalAdhesion <= 3: 2 (11.0/1.0)
|   |   |   |   marginalAdhesion > 3: 4 (3.0)
|   |   |   bareNuclei > 2: 4 (54.0/7.0)
|   |   uniformity > 4: 4 (174.0/3.0)
```

```
=== Summary ===

Correctly Classified Instances         653               95.7478 %
Incorrectly Classified Instances        29                4.2522 %
Kappa statistic                          0.9074
Mean absolute error                      0.0581
Root mean squared error                  0.2006
Relative absolute error                 12.7551 %
Root relative squared error             42.0437 %
Total Number of Instances              682
```

Tablo 8.1.J.48 Karışıklık matrisi

| | | Ön görülen Sınıf | |
|---|---|---|---|
| | | a=2 | b=4 |
| Doğru Sınıf | a=2 | 424 | 19 |
| | b=b | 10 | 229 |

Tablo 8.2. J48 Algoritmasına ait modelin karşılaştırma ölçütleri

| Doğruluk | Kesinlik | Duyarlılık | F-Ölçütü |
|---|---|---|---|
| % 95.74 | %97.7 | %95.7 | %96.7 |

## 8.2. Bayes (İstatistiksel) Sınıflandırma Modelinin Başarım Ölçütleri

Bayes Sınıflandırma için Weka‟ da var olan BayesNet, NaiveBayes, NaiveBayes Simple, NaiveBayesUpdateable algoritmalarından NaiveBayes algorit-ması seçilerek veri kümesi üzerinde çalıştırılmıştır. Tablo 8.3'te oluşturulan modelin test sonuçlarına ait istatistikleri karışıklık matrisi görülmektedir.

Naive Bayes Classifier

| Attribute | Class 2 (0.65) | 4 (0.35) |
|---|---|---|
| clump | | |
| mean | 2.9571 | 7.1883 |
| std. dev. | 1.6664 | 2.4328 |
| weight sum | 443 | 239 |
| precision | 1 | 1 |
| uniformity | | |
| mean | 1.3047 | 6.5774 |
| std. dev. | 0.855 | 2.7185 |
| weight sum | 443 | 239 |
| precision | 1 | 1 |
| uniformityofcellshape | | |
| mean | 1.4108 | 6.5607 |
| std. dev. | 0.9541 | 2.5637 |
| weight sum | 443 | 239 |
| precision | 1 | 1 |
| marginalAdhesion | | |
| mean | 1.3499 | 5.5858 |
| std. dev. | 0.9173 | 3.1899 |
| weight sum | 443 | 239 |
| precision | 1 | 1 |
| singlepithelial | | |
| mean | 2.1061 | 5.3264 |
| std. dev. | 0.8787 | 2.438 |
| weight sum | 443 | 239 |
| precision | 1 | 1 |
| bareNuclei | | |
| mean | 1.3521 | 7.6276 |
| std. dev. | 1.1802 | 3.1102 |
| weight sum | 443 | 239 |
| precision | 1 | 1 |
| blandChromatin | | |
| mean | 2.0813 | 5.9749 |
| std. dev. | 1.0614 | 2.2776 |
| weight sum | 443 | 239 |
| precision | 1 | 1 |
| normalNucleoli | | |
| mean | 1.2619 | 5.8577 |
| std. dev. | 0.9545 | 3.3419 |
| weight sum | 443 | 239 |
| precision | 1 | 1 |
| mitoses | | |
| mean | 1.1936 | 2.7537 |
| std. dev. | 0.4938 | 2.5199 |
| weight sum | 443 | 239 |
| precision | 1.125 | 1.125 |

Tablo 8.3. Naive Bayes karışıklık matrisi

| | | Ön görülen Sınıf | |
|---|---|---|---|
| | | a=2 | b=4 |
| Doğru Sınıf | a=2 | 424 | 19 |
| | b=4 | 6 | 233 |

Tablo 8.4. Bayes (İstatistiksel) Sınıflandırma Modelinin Algoritmasına ait modelin karşılaştırma ölçütleri

| Doğruluk | Kesinlik | Duyarlılık | F-Ölçütü |
|---|---|---|---|
| % 96.33 | %98.6 | % 95.7 | % 97.1 |

### 8.3.Regresyon Modelinin Başarım Ölçütleri

Karşılaştırma amaçlı olarak regresyon tabanlı yöntemlerden lojistik regresyon algoritması seçilerek veri kaynağına uygulanmıştır. Tablo 8.5'de oluşturulan modelin test sonuçlarına ait istatistikleri ve karışıklık matrisi görülmektedir. Tablo 8.6" da ise karışıklık matrisini kullanarak hesaplanan karşılaştırma ölçütleri verilmiştir.

```
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                           Class
Variable                       2
==================================
clump                    -0.5353
uniformity                0.0065
uniformityofcellshape    -0.3235
marginalAdhesion         -0.3302
singlepithelial          -0.0965
bareNuclei               -0.3826
blandChromatin           -0.4462
normalNucleoli           -0.2127
mitoses                  -0.5335
Intercept                10.0958


Odds Ratios...
                           Class
Variable                       2
==================================
clump                     0.5855
uniformity                1.0065
uniformityofcellshape     0.7236
marginalAdhesion          0.7188
singlepithelial            0.908
bareNuclei                0.6821
blandChromatin            0.6401
normalNucleoli            0.8084
mitoses                   0.5866


Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         661               96.9208 %
Incorrectly Classified Instances        21                3.0792 %
Kappa statistic                          0.9323
Mean absolute error                      0.0444
Root mean squared error                  0.1594
Relative absolute error                  9.7599 %
Root relative squared error             33.4087 %
Total Number of Instances              682
```

Tablo 8.5. lojistik regresyon algoritması karışıklık algoritması

| | | Ön görülen Sınıf | |
|---|---|---|---|
| | | a=2 | b=4 |
| Doğru Sınıf | a=2 | 433 | 10 |
| | b=4 | 11 | 228 |

Tablo 8.6. Lojistik regresyon algoritmasına ait modelin karşılaştırma ölçütleri

| Doğruluk | Kesinlik | Duyarlılık | F-Ölçütü |
|---|---|---|---|
| %96.92 | %97.5 | %97.7 | %97.6 |

### 8.4.Örnek Tabanlı Sınıflandırma Modelinin Başarım Ölçütleri

Örnek tabanlı yöntemlerden Weka‟ da bulunan KStar algoritması kullanılarak model oluşturulmuştur. Tablo 8.7‟ de oluşturulan modelin test sonuçlarına ait istatistikleri ve karışıklık matrisi görülmektedir. Tablo 8.8‟ de ise karışıklık matrisini kullanarak hesaplanan karşılaştırma ölçütleri verilmiştir.

| | | |
|---|---|---|
| Correctly Classified Instances | 653 | 95.7478 % |
| Incorrectly Classified Instances | 29 | 4.2522 % |
| Kappa statistic | 0.9056 | |
| Mean absolute error | 0.0514 | |
| Root mean squared error | 0.1831 | |
| Relative absolute error | 11.284 % | |
| Root relative squared error | 38.3684 % | |
| Total Number of Instances | 682 | |

Tablo 8.7. KStar algoritması karışıklık matrisi

| | | Ön görülen Sınıf | |
|---|---|---|---|
| | | a=2 | b=4 |
| Doğru Sınıf | a=2 | 434 | 9 |
| | b=4 | 20 | 219 |

Tablo 8.8. KStar Algoritmasına ait modelin karşılaştırma ölçütleri

| Doğruluk | Kesinlik | Duyarlılık | F-Ölçütü |
|---|---|---|---|
| %95.74 | % 95.6 | % 98.00 | % 96.8 |

### 8.5 Oluşturulan Modellerin Karşılaştırılması

Önişlemden geçirilen J48, NaiveBayes, Lojistik Regresyon ve KStar algoritmaları ile analiz edilerek her algoritma için oluşmuş olan modele ait test istatistiği bir önceki bölümde verilmişti. Karşılaştırma yapabilmek için her modele ait karşılaştırma ölçüt değerleri Tablo 13.‟ de genel bir tabloda yeniden verilmiştir.

Tablo 8.9. Oluşturulan modellerin karşılaştırılması

| Algoritma / Ölçüt | J48 | Naive Bayes | Lojistik Regresyon | K Star |
|---|---|---|---|---|
| Doğruluk | % 95.74 | % 96.33 | %96.92 | %95.74 |
| Kesinlik | % 97.7 | % 98.6 | %97.5 | % 95.6 |
| Duyarlılık | %95.7 | % 95.7 | %97.7 | % 98.00 |
| F-Ölçütü | % 96.7 | % 97.1 | % 97.6 | % 96.8 |

Yaptığım çalışmada, algoritmaların kullandığı parametreler varsayılan değerler olarak seçilmiştir. Bundaki amacım, algoritmalar arasında pozitif ayrımcılık denebilecek durumlara yol açmamak; amacımın, modelleri daha iyi oluşturmak olmayan bir çalışmada, çalışmanın farklı bir istikamete yönelmesini önlemektir.

Bir önceki bölümde yaptığımız karşılaştırmayı, Lojistik Regresyon algoritmasının Wisconsin veri kümesi veri kaynağındaki göğüs kanseri kayıtları üzerinde diğer algoritmalara göre daha iyi tahmin sonuçları oluşturduğu şeklinde özetleyebiliriz. Ancak, Tablo 13." deki rakamlara baktığımızda değerler arasında büyük farklar olmadığını, en azından Lojistik Regresyon ile en yakın takipçisi Naive Bayes arasında doğruluk ve F-ölçütü açısından %0.5" lik fark olduğunu görürüz.

Veri madenciliği algoritmalarının karşılaştırma yolu ile yapılan deneysel çalışmalar bilim dünyasında keskin eleştirilere maruz kalmaktadır. Doğası gereği veri madenciliği model başarımlarının veriye bağlı olduğunu, veri üzerinde yapılan önişleme işlemlerinin ve kullanılan algoritma parametrelerinin oluşan sonuç üzerinde farklı etkileri olacağını, kullanıcıya bağlı olarak aynı modelle farklı sonuçlar elde edilebileceğini belirtmiştir.

# BÖLÜM 9.

# SONUÇ VE ÖNERİLER

Bu çalışmada, Wisconsin göğüs kanseri verilerini iyi/kötü huylu olarak ayırmak için kullanılan sınıflandırıcıların başarımları ölçülmüştür

Bu çalışmada Açık Kaynak Kodlu Veri Madenciliği WEKA hakkında bilgiler verilmiş ilgili süreçlerle ilgili kullanılabilecek yöntemler tanıtılmış ve 1991yılında Meme Kanseri Wisconsin (Orijinal) Veri Seti veri kaynağındaki göğüs kanseri hasta kayıtları üzerinde seçilen yöntemlerin uygulaması WEKA yazılımı kullanılarak gerçekleştirilmiştir ve algoritmalardan çıkan farklar üzerinde durulmuştur.

Sınıflandırma algoritmalarının karşılaştırma yöntemlerini inceleyen bu tez çalışmasında veri madenciliği ve karşılaştırma ölçütleri üzerinde durulmuştur. Genel anlamda hangi algoritmanın daha iyi model ürettiği şeklinde bir çalışmada farklı veri kaynakları üzerinde, daha çok sayıda algoritma kullanarak karşılaştırma yapılması gerekecektir. Bu çalışmada, modellerin oluşturulması için ücretsiz bir yazılım olan Weka aracı kullanılmıştır. Var olan diğer veri madenciliği araçları üzerinde aynı algoritmalar çalıştırılarak farklı araçların benzer sonuçlar üretip üretmediği kontrol edilebilir.

Kullanılan göğüs kanseri veritabanı University of Wisconsin Hospitals, Madison Dr. William H. Wolberg ‘den elde edilmiştir .699 örnekten her biri dokuz özellik ve iyi huylu ya da kötü huylu olmak üzere bir sınıf bilgisi içerir. 16 örnek eksik özellik içerdiğinden simülasyonlar 683 örnek üzerinde yürütülmüştür.

Veri madenciliğinde sınıflandırma modellerinden karar ağaçları, Naivebayes, lojistik regresyon ve örnek tabanlı sınıflandırma yöntemlerinden seçilen dört algoritmanın, 1991yılında Meme Kanseri Wisconsin (Orijinal) Veri Seti veri kaynağındaki göğüs kanseri hasta kayıtları üzerinde yapılan karşılaştırması sonucunda bir karar ağacı algoritması olan lojistik regresyon algoritmasının diğer algoritmalara göre nispeten daha iyi model oluşturduğu görülmüştür.

Karşılaştırma amaçlı olarak regresyon tabanlı yöntemlerden lojistik regresyon algoritması %96.92 ile en doğru sonucu vermiştir, bu sonuçlarla 433(TP) tanesi iyi huylu 228 (TN)tanesi kötü huylu sınıfa aittir. Lojistik regresyon algoritmasının en yakın takipçisi NaiveBayes algoritması %96.33 ile ikinci en iyi sonucu çıkarmıştır, çalışmada J48 ile K Star algoritmaları doğruluk olarak %95.74 aynı sonuçları üretmektedir.

Kesinlik ölçütü bakımından NaiveBayes en iyi sonucu oluşturmuş olup, diğer algoritmalar bu ölçüte göre, J48, Lojistik Regresyon ve KStar. Ancak kesinlik ölçütü tek başına yorumlanırsa değerlendirme yanlış sonuçlara götürebilir. Bu ölçütü duyarlılık ölçütüyle beraber ele almak gerekir. Tablodan görüleceği üzere algoritmalar, duyarlılık ölçütüne göre KStar, Lojistik Regresyon,J48, NaiveBayes ve olarak sıralanabilir hatta J48, NaiveBayes duyarlılık ölçütü aynı rakamları elde etmiştir. Görüleceği üzere, kesinlik ölçütü ve duyarlılık ölçütü birbiriyle zıt bir sıralama ortaya koymuştur.

Bu kapsamda elde edilen verilerin otomatik olarak analiz edilmesi ve sınıflandırılması hem hastalar hem de sağlık sektörü açısından büyük önem taşımaktadır. İleride daha büyük veri tabanları ile gerçekleştirilecek çalışmalar bilgisayar destekli tanı sistemlerinin başarısını arttıracaktır

Algoritmaların veri kaynağı üzerinde çalıştırılması sırasında algoritma parametreleri olarak her algoritmanın o parametre için varsayılan değeri kullanılmıştır. Her algoritma ve her veri kaynağı için başarım derecesini maksimize edecek parametre değerleri tespit ederek bu parametrelerle algoritma sonuçlarını karşılaştırmak farklı sonuçlara götürebilecektir. Ancak, böyle bir karşılaştırmada yanlılık oluşabilecektir.

Bu çalışmada, algoritmaların ürettiği modellerin başarım sonuçları karşılaştırılmıştır. Benzer şekilde, algoritmaların hızı ve hafıza kullanımı ile algoritmaların performans karşılaştırması da yapılabilir.

- Bu çalışmada, algoritmaların ürettiği modellerin başarım sonuçları karşılaştırılmıştır. Benzer şekilde, algoritmaların hızı ve hafıza kullanımı ile algoritmaların performans karşılaştırması da yapılabilir.
- Bu çalışma farklı kategorilerdeki veri grupları üzerinde yapılabilir.
- Daha geniş sayıda algoritmalar kullanılarak farklı algoritmalar karşılaştırılabilir.
- Bu çalışmada Weka Aracı kullanılmıştır. Farklı Veri Madenciliği Araçları kullanılarak çalışma genişletilebilir.
- Her Algoritmanın başarımını maximize edecek parametreler bulunarak karşılaştırma bu şekilde yapılabilir.
- Algoritmaların başarımı dışında, hızı ve hafıza kullanımı gibi diğer metrikler üzerinde bir karşılaştırma da ayrı bir çalışma konusu olarak ele alınabilir.

## KAYNAKLAR


Akademik Bilişim 2008, Çanakkale Onsekiz MartÜniversitesi, Çanakkale, 30 Ocak - 01 Şubat 2008HastaneBilgi Sistemlerinde Veri Madenciliği ,Pınar YILDIRIM1, Mahmut ULUDAĞ2,Abdülkadir GÖRÜR1

Akpınar, H.: "Veri Tabanlarında Bilgi Kesfi ve Veri madenciligi", İ.Ü.İsletme Fakültesi Dergisi, Sayı :1, 1 – 22. (Nisan 2000

Bellaachia, A. ; Guven, E. ; Predictingbreast cancer survivability: a comparison of three data mining method ;Ninth Workshop on Mining Scientific and Engineering Datasets in conjunctionwith the Sixth SIAM International Conference on Data Mining (SDM 2006) ; 2006.

Berkhin, Pavel.: "Survey ofClustering Data Mining Techniques", Accrue Software Inc., San Jose, California, USA (2002).

Bilişim Teknolojileri Dergisi, Cilt: 2, Sayı:2, Mayıs 2009 21 Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları Ali Serhan Koyuncugil1, Nermin Özgülbaş2,2009

Carino., C., Jia., Y., Lambert., B., West., P., Yu.,C., "Mining Officially Unrecognized Side Effects of Drugs by Combining WebSearch and Machine Learning", CIKM'05 Oct 31- Nov-5, 2005 Bremen, Germany

Chen., Y., ve Wu., S., "Exploring Out-Patient Behaviors in Claim Database: A Case Study UsingAssociation Rules", AMIA Symposium Proceedings, 2003

Data Mining Analysis(breast-cancer data) Jung-Ying Wang Register number: D9115007, May, 2003 Jung-Ying Wang [Jung-YingWang 2003]

Data Mining Concepts andTechniques, Han, J.-Kamber, M., Morgan Kaufmann Publishers, 1st Ed., San Francisco, USA, 2000

Delen, D. ; Walker, G. ; Kadam, A. ; Predicting breastcancer survivability:a comparison of three data mining methods; ArtificialIntelligence in Medicine, Vol 34, issue 2 ; 2004; 113-127.

Determination Of Breast Cancer Using ANN Armağan Ebru Temiz1,D.Ü.ZiyaGökalp Eğitim Fakültesi Dergisi 7,95-107 (2006), Veri Madenciliği UygulamaAlanları, Application Fields of Data Mining ,Abdullah BAYKAL1

Doğan Ş Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi,Fırat Üni ,Fen Bil,Ens,2007.

Elektrik -Elektronik - Bilgisayar Mühendisliği 10. Ulusal Kongresi453,
Eğiticili Ve Eğiticisiz Nöral Algoritmalar Kullanarak Göğüs Kanseri Teşhisi ,Tüba KIYAN ,Tülay YILDIRIM,2003.

EndüstriMühendisliği Yazılımları ve Uygulamaları Kongresi | 30 Eylül-01/02 Ekim 2011 Weka İle Veri Madenciliği Süreci ve ÖrnekUygulama Pınar TAPKAN Lale ÖZBAKIR Adil BAYKASOĞLU

Farboudi,S,Tıp Bilişiminde İstatiksel Veri Madenciliği Yüksek Lisans Tezi ,Hacettepe Üni,Fen.Bil.2009) .

Fayyad, U.M.; Piatesky-Shapiro, G.; Smyth, P.;Uthurusamy, R., "Advances indata mining and Knowledge Discovery", AAAI Pres,USA (1994).

Hand, D. J. ; ClassifierTechnology and the Illusion of Progress; Statistical Science, Vol. 21;Institute of Mathematical Statistics, 2006; 1-15.

Han, J., veKamber, M., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2001

Jain, A.K., Duin, R.P.W.,Mao, J., "Statistical pattern Recognition: A Rewiew", IEEE Trans. PatternAnalysis and Machine Intelligence, Vol. 22, 4-37, 2000.

Jiawei Han ve MichelineKamber, Data Mining: Concept andTechniques, USA: Morgan Kaufmann Publishers, 2001, s.39-40.

Kaur., H.,ve Wasan., S., "Empirical Study on applications of Data Mining Techniques inHealthcare", Journal of Computer Science 2(2), 2006.

Kocabaş, Ş.,1991.A Review of learning. The Knowledge Engineering Review, Vol. 6. No.3, 195-222.

Korhan Kadir Babadag, 'VeriMadenciligi Yaklasımı ve Veri Kalitesinin Artması için Kullanılması', TÜİK 15. İstatistik ArastırmaSempozyumu Bildiriler Kitabı, Yayın No.3062, Ankara, 2006, s.242.

Kudyba, S.,"Managing Data Mining", CyberTech Publishing, 2004, 146-163.Posted by Koray Kocabaş Published in Veri Madenciliği (2012)

Sağlık Bakanlığı, www.saglik.gov.tr. ErişimTarihi: 18.05.2009.

Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi Cilt XXIX, Sayı 1, 2010, s.65-90 Veri Madenciliği ve İstatistik Selim TÜZÜNTÜRK

Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması Cengiz COŞKUN1, Yrd. Doç. Dr. Abdullah BAYKAL2

Veri Madenciliği Uygulama Alanları (Application Fields of Data Mining) Abdullah BAYKAL.2006

Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi Mustafa DANACI, Mete ÇELİK, A. Erhan AKKAYA,2010