

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

HİZMET SEKTÖRÜNDE BİR VERİ MADENCİLİĞİ UYGULAMASI

YÜKSEK LİSANS TEZİ
Songül ŞEKEROĞLU

Anabilim Dalı : Endüstri Mühendisliği

Programı : Mühendislik Yönetimi

EYLÜL 2010

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

HİZMET SEKTÖRÜNDE BİR VERİ MADENCİLİĞİ UYGULAMASI

YÜKSEK LİSANS TEZİ
Songül ŞEKEROĞLU
(507071222)

Tezin Enstitüye Verildiği Tarih : 13 Eylül 2010

Tezin Savunulduğu Tarih : 24 Eylül 2010

Tez Danışmanı : Prof. Dr. Cengiz GÜNGÖR (YALOVA)
Diğer Jüri Üyeleri : Prof. Dr. Fethi ÇALIŞIR (İTÜ)
Öğr. Gör. Dr. Halil Halefşan SÜMEN
(İTÜ)

EYLÜL 2010

Anneme ve babama,

ÖNSÖZ

Veri Madenciliği konusunda araştırma olanağı sağlayan, bu çalışma sırasında ilgisini ve desteğini esirgemeyen tez danışmanım Sayın Prof. Dr. Cengiz GÜNGÖR'e, bu çalışmadaki katkılarından dolayı *TÜBİTAK*'a ve son olarak da bana ömrümün ilk gününden beri her konuda destek olan aileme sonsuz teşekkürlerimi sunarım.

Eylül 2010

Songül Şekeroğlu
(Endüstri Mühendisi)

İÇİNDEKİLER

Sayfa

ÖNSÖZ.....	v
İÇİNDEKİLER	vii
KISALTMALAR	ix
ÇİZELGE LİSTESİ.....	xi
ŞEKİL LİSTESİ.....	xiii
ÖZET.....	xv
SUMMARY	xvii
1. GİRİŞ	1
2. VERİ MADENCİLİĞİ	3
2.1 Verinin Yararlı Bilgiye Dönüşüm Süreci	3
2.2 Veri Tabanlarında Bilgi Keşfi	7
2.3 Veri Madenciliği Tanımı	9
2.4 Veri Madenciliğinin Gelişimi.....	11
2.5 Veri Madenciliğinin Önemi	15
2.6 Veri Madenciliğinin Uygulama Alanları.....	17
3. VERİ MADENCİLİĞİ SÜRECİ	23
3.1 İş Sorusunu Anlama	24
3.2 Veriyi Anlama	24
3.3 Veri Hazırlığı.....	24
3.3.1 Veri temizleme	26
3.3.2 Veri dönüştürme.....	29
3.4 Modelin Kurulması	31
3.5 Değerlendirme	32
3.6 Uygulama	33
3.7 İzleme	34
4. VERİ MADENCİLİĞİ MODELLERİ	35
4.1 Sınıflandırma	37
4.1.1 Sınıflandırma süreci	37
4.1.2 Karar ağaçları ile sınıflandırma.....	37
4.1.3 İstatistiksel sınıflandırma modelleri.....	44
4.1.4 Mesafeye dayalı sınıflandırma modelleri.....	48
4.1.5 Yapay sinir ağları	50
4.1.5.1 Yapay sinir ağlarının temel özellikleri	54
4.1.5.2 Öğrenme şekillerine göre yapay sinir ağları	55
4.1.6 Genetik algoritmalar	56
4.1.7 Destek vektör makineleri	59
4.1.8 Yapısal risk minimizasyonu	61
4.2 Kümeleme	61
4.2.1 Kümeleme analizinin sınıflandırılması	66
4.2.2 Hiyerarşik yöntemler	66

4.2.3 Bölümlemeli yöntemler	70
4.2.4 Grid Temelli Algoritmalar	75
4.2.5 Genetik algoritmalar.....	77
4.3 Birliktelik Kuralları ve İlişki Analizi	78
4.3.1 Destek ve güven ölçütleri.....	78
4.3.2 Birliktelik analizinde kullanılan algoritmalar	79
4.4 Ardışlık Keşfi	81
5. UYGULAMA.....	83
5.1 Giriş	83
5.2 Uygulamada Kullanılan Paket Program	83
5.3 Mevcut Durum Analizi.....	84
5.3.1 Kuyumculuk sektörü giriş	84
5.3.2 Kuyumculuk sektörü tarihçe	85
5.3.3 Dünyada kuyumculuk sektörü.....	87
5.3.4 Türk kuyumculuğunun yıllar itibariyle gelişimi	90
5.3.5 ABC şirketi hakkında genel bilgi	94
5.4 Uygulama	95
5.4.1 İş sorusunu anlama	95
5.4.2 Veriyi anlama ve hazırlama.....	96
5.4.3 Modelleme.....	96
5.4.4 Uygulama	97
5.4.5 İzleme	105
6. SONUÇ VE ÖNERİLER.....	107
KAYNAKLAR.....	109
ÖZGEÇMİŞ.....	113

KISALTMALAR

VTBK	: Veri Tabanlarında Bilgi Keşfi
ENIAC	: Electrical Numerical Integrator and Calculator
PCA	: Principal Component Analysis
VC	: Vapnik-Chervonenkis
OLAP	: Online Analytical Processing
CRISP-DM	: Cross Industry Standard Process for Data Mining
SLIQ	: Supervised Learning in Quest
CART	: Classification and Regression Trees
SPRINT	: Scalable Parallelizable Induction of Decision Trees
YSA	: Yapay Sinir Ağları
SVM	: Support Vector Machines
SRM	: Yapısal Risk Minimizasyonu
CURE	: Clustering Using Representatives
BIRCH	: Balanced Iterative Reducing and Clustering Using Hierarchies
PAM	: Partitioning Around Medoids
CLARA	: Clustering Large Applications
STING	: Statistical Information Grid
VLDB	: Very Large Database Endowment
OCDB	: Offline Candidate Determination
CARMA	: Continuous Association Rule Mining Algorithms

ÇİZELGE LİSTESİ

	<u>Sayfa</u>
Çizelge 2.1 : Veri madenciliğinin gelişimi.	13
Çizelge 4.1 : Bilinen YSA mimarilerinin tarihsel gelişimi.....	51

ŞEKİL LİSTESİ

Sayfa

Şekil 2.1 : İş zekasının kullanıldığı sektörler.....	4
Şekil 2.2 : İş zekasının yararları.	5
Şekil 2.3 : Veri, enformasyon ve bilgi ilişkisi.	7
Şekil 2.4 : Veri tabanlarında bilgi keşfi süreci.	9
Şekil 2.5 : Veri madenciliğine katkıda bulunan disiplinler.	15
Şekil 2.6 : Veri madenciliğinin önemi.	16
Şekil 2.7 : Veri madenciliğinin kullanım alanları ve yüzdeleri.	17
Şekil 3.1 : CRISP_DM veri madenciliği süreci.	23
Şekil 4.1 : Veri madenciliği model ve teknikleri.	36
Şekil 4.2 : Örnek karar ağacı.	38
Şekil 4.3 : Biyolojik sinir ağının yapısı.	52
Şekil 4.4 : Yapay sinir ağı.	52
Şekil 4.5 : 3 katlı sinir ağı örneği.	53
Şekil 4.6 : Doğrusal destek vektör makineleri.	60
Şekil 4.7 : Kümeleme örneği.	62
Şekil 4.8 : Tek bağlantı kümeleme yöntemi örneği.	65
Şekil 4.9 : Tam bağlantı kümeleme yöntemi örneği.	65
Şekil 4.10 : Ortalama bağlantı yöntemi.	66
Şekil 4.11 : CHAMELEON algoritması çalışma mantığı.	69
Şekil 4.12 : CF ağaç yapısı.	69
Şekil 4.13 : Bir veritabanının k-ortalama yöntemiyle kümelenmesi.	71
Şekil 4.14 : Bir veri tabanının K-medoids yöntemi ile kümelenmesi.	72
Şekil 4.15 : STING kümelemenin hiyerarşik yapısı.	76
Şekil 5.1 : Yıllara göre dünya toplam mücevherat ihracat değerleri.	87
Şekil 5.2 : Dünya mücevherat ihracatının ülkelere göre dağılımı.	87
Şekil 5.3 : 2008 yılı ihracat oranının ülkelere göre dağılımı.	88
Şekil 5.4 : Ülkelere göre dünya mücevherat ithalat değerleri.	89
Şekil 5.5 : Yıllara göre dünya mücevherat ithalat hacmi.	89
Şekil 5.6 : Ülkelere göre dünya mücevherat ithalat oranı.	90
Şekil 5.7 : Yıllara göre Türkiye mücevherat ithalat değerleri.	93
Şekil 5.8 : Yıllara göre Türkiye mücevherat ihracat değerleri.	94
Şekil 5.9 : İki adım algoritması sonucu.	97
Şekil 5.10 : 1 numaralı kümeye ait bölge bilgileri.	98
Şekil 5.11 : 1 numaralı kümeye ait müşteri genel bilgileri.	99
Şekil 5.12 : 2 numaralı kümeye ait müşteri bölgeleri.	100
Şekil 5.13 : 2 numaralı kümeye ait müşteri genel bilgileri.	100
Şekil 5.14 : 3 numaralı kümeye ait müşteri bölgeleri.	101
Şekil 5.15 : 3 numaralı kümeye ait müşteri genel bilgileri.	102
Şekil 5.16 : Net gelire göre müşteri kümelerinin karşılaştırılması.	103
Şekil 5.17 : Sipariş başına düşen gelire göre kümelerin karşılaştırılması.	104

HİZMET SEKTÖRÜNDE BİR VERİ MADENCİLİĞİ UYGULAMASI

ÖZET

Son günlerde bilgisayarların insan yaşamında kullanılması ile beraber bilgi ve bilginin işlenmesi önem kazanmıştır. Teknolojinin gelişmesiyle beraber ucuzlayan donanım ve yazılımlar sonucunda kayıt altına alınan veri miktarı artmaya başlamıştır. Bu çalışmada teknolojinin gelişimiyle beraber iş zekasının artan önemi anlatılmıştır. Veri madenciliğinin ayrıntılı tanımı yapılmıştır. Veri madenciliğinin günümüzde neden bu kadar önemli ve gerekli olduğu açıklanmıştır. Verinin yararlı bilgiye dönüşüm süreci anlatılmış, bilgiye ulaşmanın gerekliliğinden bahsedilmiştir. Veri tabanlarında bilgi keşfi olarak adlandırılan bilgiye ulaşma yolları anlatılmıştır. Farklı sektörlerdeki veri madenciliği uygulamalarından örnekler verilmiştir. Veri madenciliği sürecinde yaygın olarak kullanılan CRISP DM adımları ayrıntılı olarak açıklanmıştır.

Veri madenciliği modelleri olan sınıflandırma, kümeleme, birliktelik kuralları ve ilişki analizi ve ardışık keşfi modelleri kullanılan algoritmalarla beraber ayrıntılı olarak anlatılmıştır. Bahsedilen modellerin işleyiş süreçleri de aktarılmıştır. Algoritmalar arasında günümüzde yaygın olarak kullanılan yapay sinir ağları, karar ağaçları ve genetik algoritmalar da açıklanmıştır.

Yapılan çalışmada veri madenciliği kullanılarak yapılan bir uygulamada yer almaktadır. Yapılan uygulamada Kuyumculuk Sektörünün dünyada ve Türkiye'deki tarihçesi ve mevcut durumu incelenmiştir. Kuyumculuk Sektöründe bir firmanın müşterileri kümeleme analizi yapılarak segmentlere ayrılmıştır. Uygulamada paket program kullanılmıştır. Müşteriler segmentlere ayrılırken firmanın stratejileri ve kuyumculuk sektörünün şartları göz önünde bulundurulmuş, analiz kriterleri bu iki hususa göre belirlenmiştir. Analiz sonucunda elde edilen müşteri segmentlerine yönelik izlenecek politikalar ve kampanya önerileri oluşturulmuştur.

A DATA MINING APPLICATION IN SERVICE SECTOR

SUMMARY

Recently, information and operating information have become important due to the use of computers in everyday life. Cheaper hardware and software appeared in the markets with technology development, therefore data amount that is recorded, has been increased. In this study, importance of business intelligence and development of technology described. A detailed definition of data mining is made. It is clarified that why data mining is very important and necessary in today's world. It is explained that the process of data's transformation to beneficial information, the necessity of reaching knowledge. Called information exploration in databases is the ways of reaching knowledge, is described. In variety of sectors, implementation samples are given in this study. CRISP DM steps, which are widely used in data mining process, are described in detail.

Classification, Clustering, Association Rules and Correlation Analysis, which are basic models of data mining, are described with the algorithms that are in use, such as Artificial Neural Networks, Decision Trees, Genetic Algorithms are some of the algorithms which are widely used. Process of the above mentioned models are described.

In this study, an application is made by using data mining techniques, is also described. The history of Jewelry Industry's is investigated in Turkey, in the world as well as Jewelry industry's current status. One of the companies' customers in the Jewelry Industry is divided into segments by clustering analysis. During application process, a package program is used. Strategies of the Company and the Jewelry Industry's conditions are considered to determine criteria for determining segment of customers. After clustering, the segments are investigated. Policies and recommendation for marketing campaigns for each segments is developed.

1. GİRİŞ

Bilgisayarların insan yaşamında kullanılması ile beraber bilgi ve bilginin işlenmesi önem kazanmıştır. Bilgi ve teknolojinin birlikte kullanılmasını ifade eden bilişim, bilgi ve teknoloji arasında köprü kurarak hayatı kolaylaştırmaktadır. Yoğun rekabetin yaşandığı ve bilginin öneminin her geçen gün arttığı günümüzde, doğru ve geçerli bilgiyi elde eden organizasyonlar büyük rekabet avantajı elde ederler. Bilişim teknolojisinin 4 ayağı olan yazılım, donanım, kullanıcı ve toplumun gelişmesiyle kayıt altına alınan verilerin miktarı artmıştır. Depolanan verilerin içerisinden yararlı bilgiyi elde etmek organizasyonlara büyük rekabet avantajları sağlamaktadır. İş zekası ve veri madenciliği teknolojileri verilerden yararlı bilgi elde etmeyi sağladığından her geçen gün önemi artmakta ve doğal sonucu olarak kullanımı yaygınlaşmaktadır.

Günümüzde organizasyonların hayatta kalabilmesi müşteri memnuniyetini sağlamasına bağlıdır. Müşterilerin memnuniyetlerini üst seviyede tutabilmek ancak müşteri talep ve isteklerine en doğru zamanda cevap vermekle mümkün olabilir. Bu durumda organizasyonların hayatta kalabilmesi esnek bir yapıya sahip olmalarıyla mümkündür. Organizasyonların esnek bir yapıya sahip olmaları ancak müşterilerin talep ve isteklerini önceden tahmin etmekle gerçekleşebilir. Veri madenciliği teknikleriyle esnekliği sağlayabilmek için gerekli bilgiler elde edilebilmektedir.

Kuyumculuk sektörü ülkemizde gün geçtikçe artan öneme sahiptir. Çok sayıda müşteriye sahip olan kuyumculuk firmaları, kendilerine yüksek miktarda getiri sağlayan müşterilerini ayırt etmeksizin, tüm müşterilerine eşit şekilde davranmakta bu durum da kimi durumlarda karlı müşterilerin kaybedilmesine neden olmaktadır.

Yapılan çalışma giriş, veri madenciliği, veri madenciliği süreci, veri madenciliği modelleri, uygulama, sonuç ve öneriler olmak üzere 6 ana bölümden oluşmaktadır.

Veri madenciliği bölümünde veri, enformasyon, bilgi ve yararlı bilginin tanımları yapılmış, verinin yararlı bilgi haline dönüşmesi sürecinden bahsedilmiştir. Veri madenciliğinin önemi ve tanımı anlatılmış, kullanımının günümüzde zorunluluk

haline gelmesinin nedenlerinden, hangi amaçlarla hangi alanlarda kullanıldığından bahsedilmiştir.

Veri madenciliği süreci bölümünde CRISP DM tarafından hazırlanan veri madenciliği adımları ayrıntılı olarak açıklanmıştır.

Veri madenciliği modelleri bölümünde tahminleyici ve tanımlayıcı modeller olmak üzere veri madenciliği modelleri anlatılmıştır. Modellerin gelişim süreçleri ve modellerde kullanılan algoritmalarından bahsedilmiş; veri madenciliğinde kullanılan tekniklerden yaygın kullanılanları ayrıntılı olarak açıklanmıştır.

Uygulamada kuyumculuk sektöründe yer alan bir firmanın müşterilerinin şirkete sağladığı yarar göz önünde bulundurularak kümelemesi yapılmıştır.

Sonuç ve öneriler kısmında ise uygulama sonucu elde edilen kazanımların neler olduğundan, uygulamanın devamında hangi çalışmaların yapılabileceği anlatılmış, önerilerde bulunulmuştur.

2. VERİ MADENCİLİĞİ

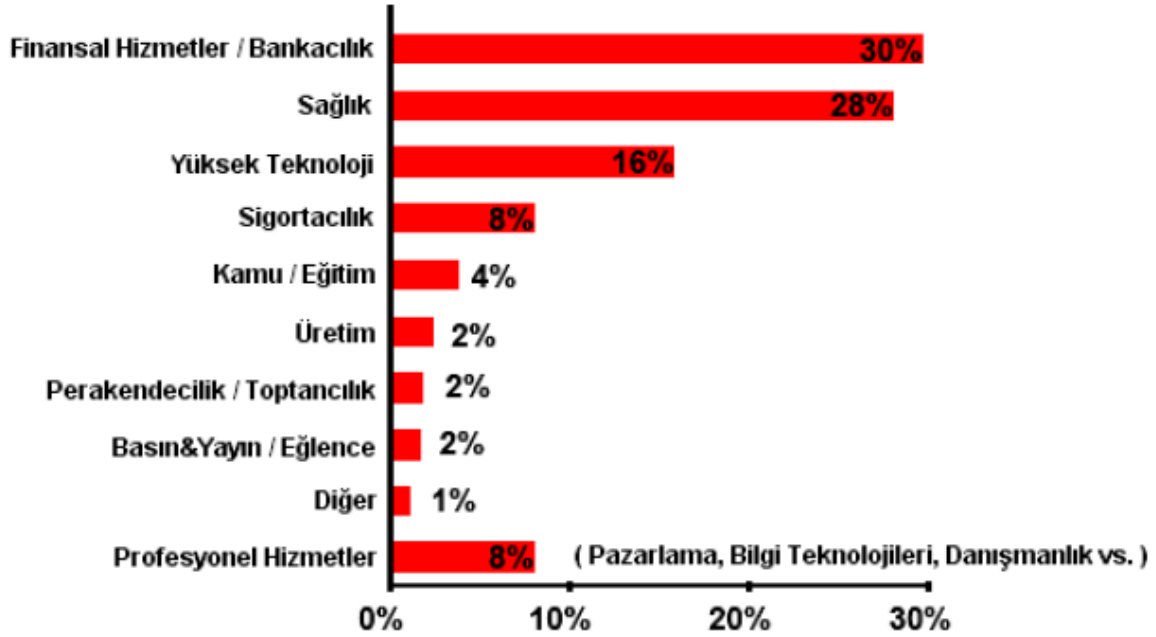
2.1 Verinin Yararlı Bilgiye Dönüşüm Süreci

Yaşanan teknolojik gelişmeler toplumlarda sosyal ve ekonomik açıdan büyük değişikliklere neden olmuştur. Yaşanan bu değişimler toplumların bilgi toplumuna dönüşmesini sağlamıştır. Günümüzde bilgi çok önemli ve etkili bir kavram olarak her alanda yer almaktadır. İş dünyası da bu akımdan büyük ölçüde etkilenmiştir ve bilgi iş dünyasının da en önemli faktörlerinden birisi olmuştur. Bilhassa yöneticiler için doğru karar verme gereksinimini karşılamak, doğru ve geçerli bilgiye sahip olmayı zorunluluk haline gelmiştir.

Bilişim alanında yaşanan baş döndürücü değişimler ve gelişimler, bilgisayar teknolojilerinin ucuzlaşmasını ve yaygınlaşmasını sağlamaktadır. Bunun sonucu olarak yapılan her işlem kaydedilmekte; örneğin markette yapılan satışlar, müşteri bilgileri ve kamera kayıtları vs. sürekli kaydedilmektedir. Bu sayede farklı sistemler aracılığıyla her gün yüzlerce veri girişi yapılmakta; müşteri ve firmalar arası sistemlerin çoğalmasıyla, bütün bu verilerin aynı çatı altında toplanabilmesi, ayrıştırılıp anlamlandırılabilmesi daha da güçleşmektedir. Bu veriler o halleriyle organizasyonlara sadece yük oluşturmaktadır. Bu tarz büyük çapta verilerden yararlanarak kurumlar için karar destek sistemleri oluşturulabilir. Verilerin karar verme aşamasında kullanılması için çeşitli işlemlerden geçmesi gerekmektedir. Bu verilerin saklanması, analiz edilmesi ve işletmenin bu verilerden kendine özgü sonuçlar çıkarıp kendini yönlendirmesi gerekmektedir. İşletmelere bu yolda rehberlik edecek en güncel teknoloji İş zekası (Business Intelligence) olarak adlandırılmaktadır.

İş zekası, bütün kaynaklardan toplanan verileri, bilgiyi elde etmek için yeni formlara dönüştürmeyi amaçlayan, bilinçli, sistemli, işle ilgili ve sonuç odaklı işlemlerin bütünüdür (Biere, 2003).

İşletmelere karlılık, müşteri memnuniyeti, performans ve kalite gibi konularda gelişmesine katkıda bulunan iş zekası birçok sektörde kullanılabilir. İş zekası son 15-20 yıldır ilgi duyulan bir kavram haline gelmiştir. Ülkemizde ise son 3-4 yıldır yaygınlaşmaya başlamıştır. Gantry Group'un Amerika' da farklı sektörlerde yaklaşık 75000 kişi üzerinde yaptığı çalışma sonucunda iş zekası kullanımının sektörlere göre dağılımı Şekil 2.1'de gösterilmektedir (Gantry Group, 2005).



Şekil 2.1 : İş zekasının kullanıldığı sektörler.

İş zekasının 4 temel özelliği bulunmaktadır. Bu özellikler özet olarak şöyle sıralanmaktadır. (Businessobjects, 2007).

1. Bilgiye tek bir noktadan erişim: İş zekası kullanan organizasyonlarda, veriler ortak bir veri tabanında saklanmaktadır. Bu da bilginin entegrasyonunu kolaylaştırmakta, tüm kullanıcıların bilgiye istediği zaman istediği durumda ulaşma imkanı sağlamaktadır.
2. İşletmenin bütün bölümlerinde kullanılabilir olması: İş zekası kavramından önce her bir departman kendilerine ait bilgileri saklamaktaydılar. Bu durumda bilgilerin entegrasyonunu zorlaştırmakta, aynı verilerin birkaç kez tutulup gereksiz bilgi yüküne neden olmaktaydı. İş zekası organizasyondaki bütün bölümlerin verilerini bir araya toplayarak, istenen bilgiye daha kolay ve çabuk ulaşılabilmesini sağlamıştır.

3. Ortaya çıkan sorulara anında cevap verebilmesi: Kullanıcılar iş zekası sistemlerine basit sorgular girerek aradıkları cevaplara ulaşabilmektedirler.

4. İnternetin olanaklarından yararlanılabilmesi: Şirket alanı dışındaki bir kullanıcı, şirketin bilgi ağına bağlanıp bazı verilere internet sayesinde ulaşabilmektedir. İş zekası kavramı bu duruma olanak sağlamaktadır.

İş zekasının getirdiği bilgiye ulaşabilme özelliği, kullanıcılara karar vermede kolaylık sağlamaktadır. İş zekası sisteminde her bir kullanıcı ihtiyaçlarına göre özet ya da detaylı raporlar hazırlayıp bunları işletme içine ya da dışına dağıtabilmektedir. Rapor hazırlamanın fazla teknik bilgi gerektirmeyen, kolay ve hızlı olması, insandan ve zamandan kazanç sağlamaktadır. İş zekasının en büyük yararlarından biri maliyetleri azaltmasıdır. Bu çözümler ile , işletme içerisinde paranın nerelerde daha fazla harcandığı izlenebilmekte; harcamaların gerekli olup olmadığı analiz edilebilmekte; maliyet avantajı sağlayacak alanlar tespit edilebilmektedir. İş zekası maliyetleri azaltmakla birlikte gelirleri de artırmaktadır. İş zekası çözümlerini kullanan bir işletmede pazarlama yöneticisi , ürününe daha fazla ödemeye gönüllü müşterileri tespit edebilir. Ürün dışında bilginin satışından da kazanç sağlanabilmektedir. Bazı firmalar, iş zekası kullanarak elde ettikleri bilgiyi müşterileri ya da tedarikçileri ile belirli ücretler karşılığında paylaşmaktadır (Ateş, 2008). Şekil 2.2’ de İş zekasının yararları özet halinde gösterilmektedir (Ericsson, 2004).



Şekil 2.2 : İş zekasının yararları.

Veri (Data), ham haldeki yani işlenmemiş kayıtlardır. Oldukça esnek yapıdadır. Veri haliyle kayıtlar düzenlenmemiş yani gerekli ilişkilendirme ve anlamlandırma işlemi yapılmamış haldedir. Veri tabanı yönetim sistemlerinin keşfi ve veri saklama teknolojilerindeki ilerleme ile organizasyonlara büyük miktarlarda veri toplanmakta ve depolanmaktadır. Verilerin büyük bir kısmı organizasyonun fonksiyonel prosesleri ile ilişkilidir (Fayyad, 1996). Örneğin markette çalışan kasiyerin müşterinin almış olduğu ürünleri kasadan geçirerek, müşterinin hangi ürünleri hangi miktarda aldığı, ödemeyi hangi yöntemle yaptığı şeklindeki verileri kaydetmektedir.

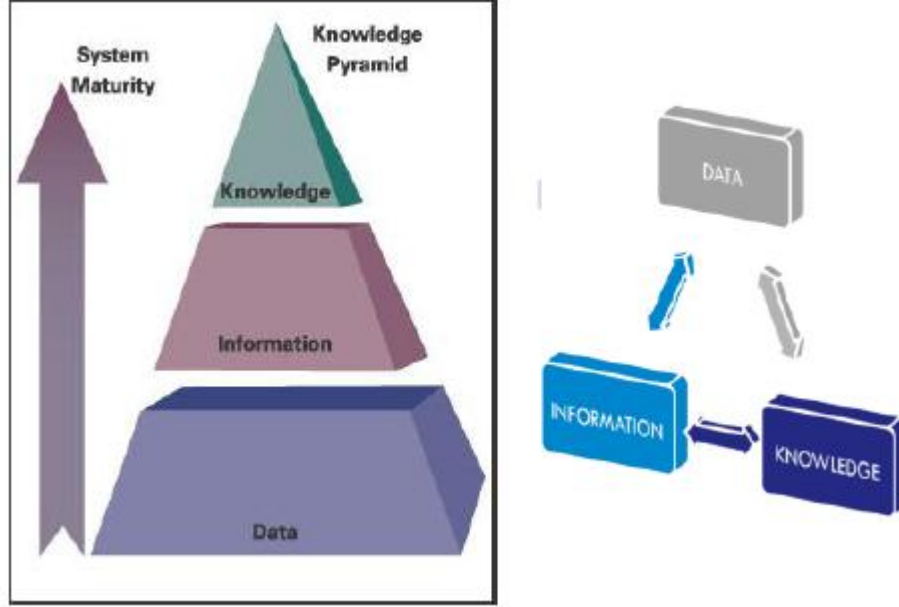
Enformasyon (Information), verilerin düzenlenmiş, ilişkilendirilmiş ve anlamlandırılmış haline olarak tanımlanır. Enformasyonlar belirli bir amaç doğrultusunda yapılmaktadır. O nedenle enformasyon başka bir amaç için veri halini korumaktadır. İş zekası uygulamalarının sorgu ve raporlama yetenekleri sayesinde veritabanındaki verinin enformasyona dönüşümü sağlanmaktadır (Fayyad, 1996). Son bir haftada hangi ürünlerin satıldığı, satılmış olan ürünlerin miktarları ve tutarları vs. şekline dönüştürülmüş veriler, enformasyona örnek olarak verilebilir. Örnekte satışlar ve satışların nitelikleri belirli bir zaman diliminde gerçekleşmelerine göre düzenlenmiştir.

Bilgi (Knowledge), enformasyonun birey tarafından algılanması ve sonuç çıkarmasıyla oluşur. Birey bilgiye ulaşmak için Veri Madenciliği (data mining) teknolojisi içeren uygulamalar kullanarak, veri içerisindeki gizli eğilim ve örüntüleri belirleyebilir (Fayyad, 1996). Örneğin bir perakende şirketinin yaptığı veri madenciliği araştırmasının sonucuna göre özellikle Cuma günleri bira ve çocuk bezi satışları arasında güçlü bir ilişki olduğu sonucu ortaya çıkmıştır (Cabena, 1998). Bu sonuç tahmin edilebilmesi zor bir çıkarımdır, o nedenle bilinmeyeni ortaya koymaktadır. Enformasyonun bilgiye dönüşmesinde çalışmayı yapan bireyin algılama yeteneği, yaratıcılığı, deneyimi vb. kişisel özellikleri de çıkan sonuçlarda etkili olmaktadır.

Yararlı bilgi (wisdom), ulaşılmaya çalışılan noktadır. Bilgilerin kişiler tarafından toplanıp özümsemesiyle ortaya çıkar. Sentez sonucunda elde edilen bilgi yarar sağlayacak şekilde kullanılır. Örneğin bira çocuk bezi örneğinde alışılmışın dışındaki örüntüler keşfedilip, bunun nedeni araştırılarak gerekli önlemler alınabilir ya da önerilerde bulunabilir.

Operasyonel sistemdeki verileri biriktiren organizasyonlar, kendilerine değer katacak verideki potansiyeli anlama fırsatına sahip olurlar (Göral, 2007).

Veri, enformasyon ve bilgi ilişkisi Şekil 2.3' de gösterilmektedir.



Şekil 2.3 : Veri, enformasyon ve bilgi ilişkisi.

2.2 Veri Tabanlarında Bilgi Keşfi

Veri Tabanlarında Bilgi Keşfi, veriden faydalı bilginin keşfedilmesi sürecinin tamamını ifade etmekte kullanılmaktadır. Veri Madenciliği ise bu sürecin bir adımına karşılık gelmektedir. Veri Madenciliği, veriden örüntülerin aktarımı için özel algoritmaların uygulanması adınının ifadesidir.

VTBK süreci, veritabanlarını kullanarak veritabanlarında istenilen seçim, ön işleme, alt örnekleme, dönüşüm, örüntülerin açığa çıkarılması için Veri Madenciliği yöntemlerinin uygulanması ve açığa çıkarılan örüntülerin tanımlanması için Veri Madenciliği ürünlerinin yorumlanması süreçlerini içermektedir. VTBK sürecinin, Veri Madenciliği bileşeni, veriden hangi örüntülerin aktarılıp, dikkate alınacağını ifadesi olarak değerlendirilmelidir (Koyuncu, 2006).

VTBK süreci interaktif ve yinelemeli, kullanıcı tarafından kararların verilmesini gerektiren adımların birleşmesinden oluşmaktadır. Brachman ve Anand, sürecin interaktif yapısına vurgu yapan pratik bir görünüm vermişlerdir (Brachman ve Anand, 2006).

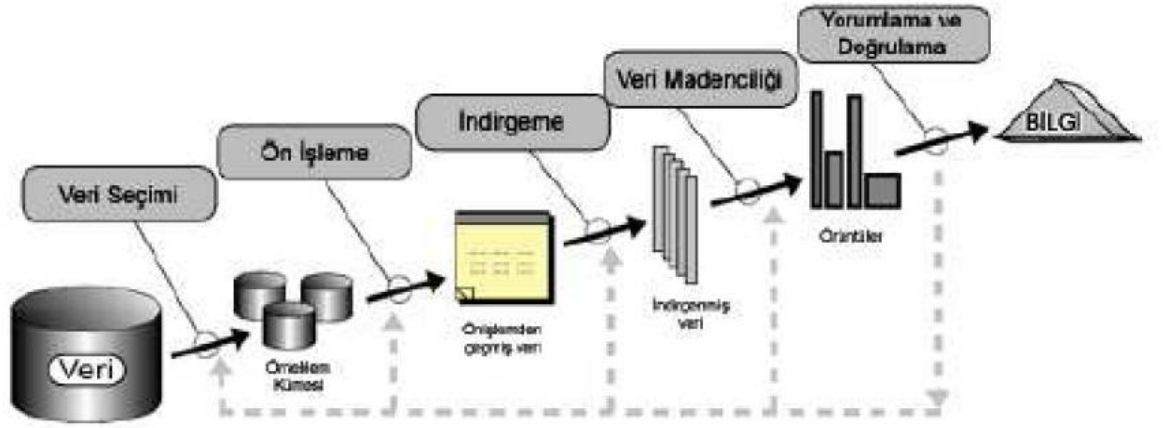
Sürecin bazı temel adımlarının çerçevesi aşağıda verilmiştir (Koyuncugil, 2006)

1. Adımda uygulama alanı ile ilgili önsel bilgi ile bir anlayış geliştirmek ve müşterinin bakış açısından VTBK sürecinin hedefini tanımlanır.
2. Adımda hedef veri kümesi yaratılır. Keşfin uygulanacağı veri kümesi seçilir veya değişkenlerin bir alt kümesi veya veri örnekleri üzerine odaklanılır.
3. Adımda Veri temizleme ve ön işleme prosesleri yapılır. Eğer uygunsa gürültünün kaldırılması, model için gerekli enformasyonun toplanması, kayıp veri alanları için stratejilere karar vermeyi içeren temel operasyonlardır.
4. Adımda veri indirgeme ve projeksiyon prosesleri yapılır. Bu adımdaki amaç hedefine bağlı veriyi temsil edecek faydalı özellikleri bulmaktır. Boyut indirgeme veya dönüşüm yöntemleriyle göz önüne alınan değişken sayısı indirgenebilir veya verinin değişmez temsili bulunabilir.
5. Adımda VTBK sürecinin hedefleri ile (1. Adımda belirlenmiş), Veri Madenciliği yönteminin eşleştirilir. Özetleme, sınıflandırma, regresyon, kümeleme vb. yöntemler uygulanmaktadır.
6. Adımda Veri Madenciliği algoritma(larının)sının seçimi yapılır. Açıklayıcı analizler, model ve hipotez seçimi de bu adımda yapılmaktadır. Tercih edilen Veri Madenciliği algoritmaları ve seçilen yöntemler veri örüntülerini araştırmak için kullanılır. Bu süreç, hangi modelin ve parametrelerin uygun olabileceğine ve Veri Madenciliği yönteminin VTBK sürecinin bütün kriterleriyle eşleşip eşleşmediğine karar verilmesini içermektedir.
7. Adımda Veri Madenciliği özel bir temsili form veya temsili küme içerisinde ilgilenilen örüntüler; sınıflandırma kuralları ve ağaçları, regresyon ve kümelemeyi içererek araştırılır.
8. Adımda Veri Madenciliği ile çıkarılan örüntülerin yorumlanmaktadır. Sonraki iterasyonlarda, Adım 1- 7'den herhangi birine dönülmesi ihtimaliyle veri madenciliği ile çıkarılan örüntüler yorumlanır.
9. Adımda Keşfedilen bilgiler birleştirilir. Keşfedilen bilgi sonraki çalışmalar için bir başka sistem altında toplanabilir veya basitçe dökümantasyonu yapıp, raporlanarak ilgili birimlere iletilir. Bu aynı zamanda, önceden inanılan veya aktarılan bilgilerin

doğruluğunu kontrol etme ve olası farklılıkların ayrıştırılmasını da içerir (Fayyad, 1996) (Zaine, 1999). VTBK sürecinin adımları Şekil 2.4’ de gösterilmektedir.

Veri Ambarları, Veri Madenciliği ile eşanlı olarak anılan ve Veri Madenciliği sürecinin gerçekleştirildiği veriyi sağlayan özel bir veri tabanıdır. Tanım olarak Veri Ambarı, pek çok farklı kaynaktan ve genellikle de farklı yapıda verinin depolandığı ve hepsinin de aynı birleşik çatı altında kullanılmasının ümit edildiği yapılardır. Ayrıca, Veri Ambarı pek çok farklı kaynaktan elde edilen veriyi aynı çatı altında analiz etme imkânı tanımaktadır (Fayyad, 1996).

Veri ambarcılığı, veri kümelerine VTBK aşaması için veri temizleme ve veri erişimi konularında yardımcı olmaktadır. VTBK süreci Şekil 2.4’te gösterilmektedir Han, 2000).



Şekil 2.4 : Veri tabanlarında bilgi keşfi süreci.

2.3 Veri Madenciliği Tanımı

Veri Madenciliği veri depolama araçları, barkot ve birçok teknolojiye paralel olarak gelişmektedir. Genel bir tanım olarak ifade edildiğinde veri madenciliği; biriken verilerden organizasyon için yararlı olanlarının çekilip ortaya çıkarılması işlemidir. Çok sayıda organizasyon tarafından kabul gören bir süreç haline gelen veri madenciliğinin birçok değişik tanımı yapılmaktadır.

“Veri Madenciliği önceleri bilinmeyen, geçerli ve etkin bilginin büyük veri tabanlarından çekilmesi ve daha sonra bu bilginin son iş kararlarını almak için kullanılmasını kapsayan bir süreçtir (Cabena, 1998).”

“Veri madenciliği, VTBK sürecinde bir adımdır ve verideki örüntüleri ortaya çıkarmak için kullanılan algoritmaları kapsar. Ortaya çıkarılan bilgi daha sonra bir öngörü (prediction) veya sınıflandırma (classification) modeli kurmak, eğilimleri ve birliktelikleri belirlemek, mevcut bir modeli yenilemek veya üzerinde madencilik çalışması yapılmış bir veri tabanının özetini çıkarmak için kullanılabilir (Fayyad, 1996). ”

“Veri madenciliği, anlamlı örüntüler ve kurallar keşfetmek için büyük miktardaki veriyi, otomatik veya yarı otomatik yöntemlerle araştırma ve analiz etme sürecidir (Berry ve Linolf, 2000).”

“Veri madenciliği, organizasyonların veri tabanlarında bulunan en önemli bilgilere odaklanabilmesine olanak sağlar. Bu sayede yöneticiler gelecekteki eğilimleri ve davranışları öngörerek daha bilgili kararlar alabilirler (Chopoorian, 2001).”

“Veri madenciliğini amacı, mevcut veri içindeki geçerli, alışılmamış, kullanışlı ve anlaşılır korelasyonları ve örüntüleri saptamaktır (Chung ve Gray, 1999).”

“Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir (Grossman, 2001).”

“Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Ancak veri madenciliği, geleneksel istatistikten birkaç yönden farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır (Dönmez, 2008).”

“Veri madenciliği, kullanıcının açık ve faydalı sonuçlar elde edebilmesi için, çok büyük miktardaki verinin içinden başlangıçta görülmeyen, bilinmeyen bazı ilişkiler ve düzenler keşfedilmesi amacıyla, verinin seçilmesi, araştırılması ve modellenmesinden oluşan bir süreçtir (Giudici,2003).”

“Veri madenciliği, büyük miktarlardaki verinin içinden geleceğin tahmin edilmesinde yardımcı olacak anlamlı ve yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir. Ayrıca veri madenciliği, çok büyük miktardaki verilerin içindeki ilişkileri inceleyerek aralarındaki bağlantıyı bulmaya yardımcı olan veri analizi tekniğidir (Akpınar,2000).”

Veri madenciliğiyle ilgili en önemli unsur elde edilen bilginin önceden bilinmiyor ve tahmin edilemiyor oluşudur. Daha önce akla gelmemiş düşünülmemiş sonuçları ortaya çıkarması veri madenciliğini diğer yöntemlerden ayıran en önemli unsurdur. Elde edilen bilgiler organizasyonların karar destek sistemleri için önemli bir yere sahiptir. Nihayetinde amaç bilgiyi keşfederek ona ulaşmak ve bu yolla organizasyonlara fayda sağlamaktır. Veri madenciliği başka yöntemlerle çıkarılan sonuçların ispatı için kullanılmamaktadır.

Veri madenciliği, aynı zamanda bir süreçtir. Veri yığınları arasında bilgiyi ortaya çıkarmanın yanı sıra bilgi keşfi sürecinde elde örüntü ve bağlantıları süzerek bir sonraki aşamaya hazır hale getirmek bu sürecin bir parçasıdır.

Veri madenciliği klasik istatistiksel uygulamalardan çalıştırdığı kayıtlar konusunda da ayrılır. İstatistiksel uygulamalarda özet ve aşırı düzenlenmiş veriler çalıştırılır fakat veri madenciliği milyonlarca veri ve çok daha fazla değişken ile çalışır.

2.4 Veri Madenciliğinin Gelişimi

Veri madenciliğinin kökeni ilk sayısal bilgisayar olan ENIAC (Electrical Numerical Integrator And Calculator)’a kadar dayanmaktadır. 1946 yılında geliştirilen ve bugün kullanılan kişisel bilgisayarların atası olan ENIAC, II. Dünya Savaşı sırasında ABD ordusu için ABD’li bilim adamları John Mauchly ve J. Presper Eckert tarafından geliştirilmiştir. Bugün kullanılan bilgisayarlarla ENIAC kıyaslandığında ilk bilgisayarın geçirmiş olduğu evrimin boyutunu görmek mümkün olacaktır. Bu 60 yıllık süre içerisinde geliştirilen donanımların yazılımlarla hayat bulması evrimin en büyük itici gücünü oluşturmaktadır.

Önceleri sadece hesaplamalar yapmak için geliştirilen bilgisayarlar, ilerleyen zamanlarda kullanıcı ihtiyaçları doğrultusunda, veri depolama amacıyla da kullanılmaya başladı. Bu sayede veri tabanları ortaya çıktı. Veri tabanları genişledikçe donanımların da genişlemesi gerekti ve bu durum veri ambarı kavramının ortaya çıkmasını sağladı. Veri ambarlarının kullanımının yaygınlaşması sonucunda depolanan ve saklanan veri miktarları büyük miktarlarda artmaya başladı. Büyüyen veri tabanları, verilerin organizasyonu, düzenlenmesi ve yönetimi gibi işlemleri gibi eskiye oranla çeşitli zorlukları beraberinde getirdi.

Bu zorlukları yenebilme çabaları veri modelleme kavramının ortaya çıkmasını sağladı. İlk olarak hiyerarşik ve şebeke veri modelleri geliştirildi. Temelinde kök olan ve kök aracılığıyla üst kısmında bir, alt kısmında ise n adet düğüm bulunan, ağaç yapısına sahip veri modelleri, hiyerarşik veri modelleri olarak adlandırılmaktadır. Şebeke modelleri, kayıt tipi ve bağlantıların olduğu; kayıt tiplerinin varlık tipini, bağlantıların ise ilişki tiplerini belirlediği bir veri modelidir. Bu iki modelin kullanıcıların ihtiyaçlarını karşılayamaması üzerine Geliştirilmiş Veri Modelleri geliştirildi. Bu modeller Varlık – İlişki, İlişkisel ve Nesne – Yönelimli veri modelleri olarak adlandırılır. İlişkisel veri modelleri günümüzde kullanımı en yaygın olan modeldir. Nesne – Yönelimli veri modelleri ise gelişim sürecine devam etmektedir. Veri madenciliğinin gelişimi Çizelge 2.1’ de özet halinde gösterilmektedir (Dunham, 2003).

Çizelge 2.1 : Veri madenciliğinin gelişimi.

Zaman	Alan	Katkı
1700'lerin sonu	İstatistik	Olasılığa ait Bayes teoremi
1900'lerin başı	İstatistik	Regresyon analizi
1920'lerin başı	İstatistik	Maksimum olasılık tahmini
1940'ların başı	Yapay zeka	Sinir ağları
1950'lerin başı	Yapay zeka	En yakın komşu, Tek bağlantı
1960'ların başı	Veri tabanı	Toplu raporlar
1960'ların ortaları	Veri tabanı	Karar ağaçları
1960'ların ortaları	İstatistik	Sınıflama için lineer modeller, Kümeleme
1960'ların sonları	Veri tabanı	İlişkisel veri modeli
1970'lerin ortaları	Yapay zeka	Genetik algoritmalar
1970'lerin sonları	İstatistik	Eksik veri ile tahmin
1970'lerin sonları	İstatistik	K-ortalama kümelemesi (K-means)
1980'lerin başları	Yapay zeka	Kohonen kendini düzenleyen haritalar
1980'lerin ortaları	Yapay zeka	Karar ağacı algoritmaları
1990'ların başı	Veri tabanı	Birliktelik kuralları algoritmaları, Web ve arama motorları
1990'lar	Veri tabanı	Veri Depolama (data warehousing)
1990'lar	Veri tabanı	Çevrimiçi analitik işleme (OLAP)

Veri madenciliği, kavramsal olarak 1960lı yıllarda, bilgisayarların veri analiz problemlerini çözmek için kullanılmasıyla ortaya çıktı. Ana bilgisayarlar ve COBOL (Common Business Oriented Language) ile üretilen listeler bilgisayar destekli analizin ilk örnekleri olarak gösterilir. Bu dönemde bilgisayarlarla yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşılabileceği kabul edildi. Bu işlemlere veri

madenciliği yerine veri taraması (data dredging) ve veri yakalaması (data fishing) gibi isimler kullanıldı.

1980lere gelindiğinde geleneksel algoritmalara dayalı istatistik araçları verilerin analiz edilmesinde kullanılmaya başlandı. İstatistik araçları mükemmel sonuçlar veriyordu fakat kullanımının ve anlaşılmasının zorluğundan dolayı sadece deneyimli istatistik uzmanları tarafından kullanılmaktaydılar. Bu yöntemlerin kullanılmasında yaşanan bir zorluk da veri boyutu büyüdükçe modellerin güvenilirliğinin azalmasıydı.

Daha sonra PCA (Principal Component Analysis) veya yapay sinir ağları (artificial neural networks) gibi yöntemler ortaya çıktı. Bu yöntemler yüksek miktarda verilerde de çok iyi sonuçlar vermesine rağmen; kapalı kutu olmaları ve sonuçların nedenlerini açıklayamamaları, yaygınlaşamamalarına neden olmuştur.

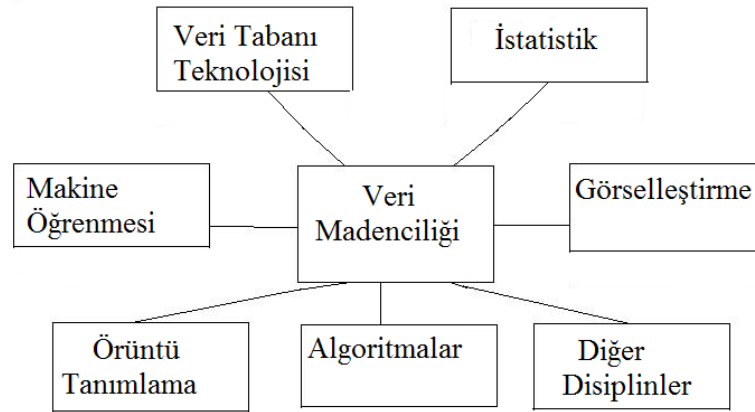
1990lı yıllarda veri madenciliği ismi bilgisayar mühendisleri tarafından kullanılmaya başlandı. Bu isimlendirmenin amacı; veri analizinin geleneksel istatistiksel yöntemler yerine, algoritmik bilgisayar modülleri tarafından yapılması gerekliliğini vurgulamaktır. Daha sonra veri madenciliğinde değişik yaklaşımlar kullanılmaya başlandı. Bu yaklaşımların temelinde istatistik, makine öğrenimi (machine learning), veri tabanları, otomasyon, pazarlama, araştırma gibi disiplinler ve kavramlar yatmaktaydı.

Veri madenciliği, müşteri odaklı büyük veritabanlarından bilgi örüntülerini çıkaran uygun teknoloji kümelerini kullanır. Bununla birlikte, veri madenciliği tek bir teknoloji değildir. Tam tersine veriden bilgi çekmekte kullanılan araçların bir kümesidir (Applied Technology Group, 1997). Veri madenciliğin birçok bilim dalı ve disiplinin katkılarıyla gelişmektedir. İstatistik alanında regresyon, faktör, kümeleme, ayırma (Discriminant) ve zaman serileri analizleri; yapay zeka alanında makina öğrenimi, yapay sinir ağları, genetik algoritmalar, zeki ajan sistemleri (Intelligent Agent Systems), bayes ağları, örüntü tanıma (Pattern Recognition) modelleri veri madenciliğine önemli katkılarda bulunmaktadır. Bilgisayar dilbilimi (Computer Linguistics) alanında ise web madenciliği (Web Usage Mining), metin madenciliği (Text Mining) ve vaka temelli çıkarım (Case Based Reasoning) veri madenciliğinde önemli rol oynayan alanlardır (Akpınar, 2004). İlgili analiz çalışmalarının yapılması çok güçlü veri tabanı yönetim sistemleri ile

gerçekleşebilmektedir. Çok boyutlu verilerin gösterilebilmesi için etkin görselleştirme teknikleri kullanımını gerekli kılmaktadır. Veri tabanı yönetim sistemlerinin ve etkin görselleştirme tekniklerinin gelişimi veri madenciliğinin de gelişmesine büyük katkılar sağlamaktadır.

Bilgisayar destekli veri analizinde bu gelişmeler yaşanırken, 1970’lerde Vladimir Vapnik ve bir grup Rus bilim adamı, İstatistik Öğrenme Teorisi’nin temelinde olan Vapnik-Chervonenkis (VC) boyutunu geliştirdiler. Modellenen verinin dağılımına bağlı olmaksızın, VC boyutu ile modellerin güvenilirliği arasında bir bağlantı kuruldu. Vapnik, önce sınıflandırma, daha sonra da regresyon problemlerini ele almış ve VC boyutunu kontrol ederek modelin kullanılan veriye uygunluğu ve yeni veride doğru sonuç vermesi arasında en iyi dengeyi sağlayan bir teknik geliştirmiştir (Bera, 2001).

Veri madenciliğinin gelişmesinde ve ilerlemesinde birçok disiplinin katkısı olmuştur. Katkıda bulunan disiplinler Şekil 2.5’ de özet halinde yer almaktadır.

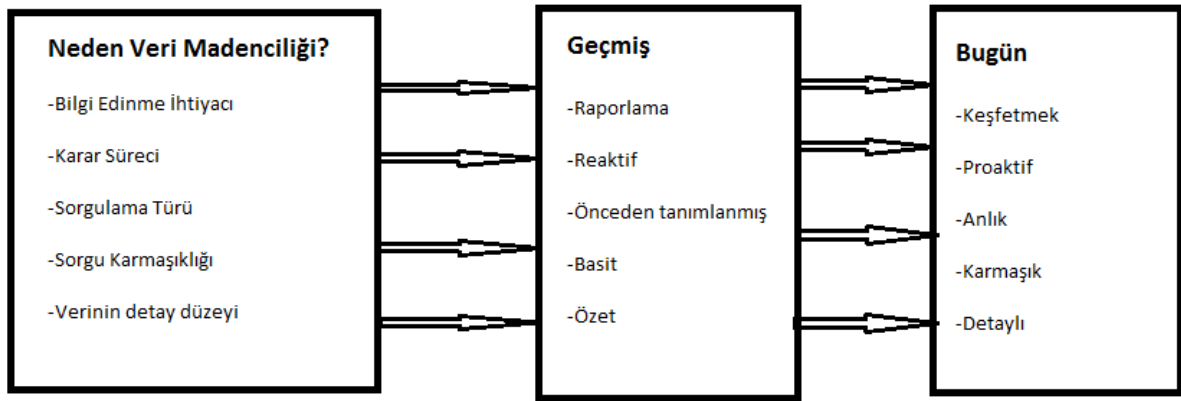


Şekil 2.5 : Veri madenciliğine katkıda bulunan disiplinler.

2.5 Veri Madenciliğinin Önemi

Teknolojinin gelişmesi ve daha ucuza imal edilmesinin neticesinde veri ambarlarının hacimleri artmakta ve kullanımı yaygınlaşmaktadır. Geçmişte veriler veri tabanlarının dar olmasından dolayı özet halinde tutulurken, günümüzde veri tabanlarının genişlemesinden dolayı daha detaylı bir şekilde tutulmaktadır. Saklanan bu veriler geçmişte çeşitli raporların çıkarılması için kullanılmaktaydı. Az hacimde özet veriler olduğundan dolayı çok bir çaba harcamaya gerek kalmadan bağlantıları tespit edebilmek mümkündü. Fakat günümüzde, detaylı olarak büyük hacimlerde

tutulan verilerin içerisindeki yararlı bilgi, bağlantı ve örüntüyü tespit edebilmek ancak bilgisayar algoritmaları kullanarak gerçekleştirilmektedir. Veri madenciliği algoritmaların kullanılarak yararlı bilgi elde etme sürecidir. Verilerin daha detaylı tutulması, birçok değişkeni ve durumu ifade etmesinden dolayı basitliğini kaybetmiş, daha komplike hale gelmiştir. Geçmişte özet olarak depolanan verilerin sorgulaması değişken sayısının azlığından ötürü daha basit olurken, daha çok değişken içeren detaylı verilerin sorgulaması daha karmaşık olmaktadır. Düünden bugüne bilginin önemindeki değişiklik ve veri madenciliğine etkileri Şekil 2.6’ da özet halinde gösterilmektedir.



Şekil 2.6 : Veri madenciliğinin önemi.

Geçmişte kararlar reaktif (bir olay ya da durum sonucu) verilmekteyken, günümüzde bir olay gerçekleşmeden onu fark etmek ve koruyucu önlemler almak önem kazanmaktadır. Rekabetin arttığı günümüzde müşterilerin ihtiyaçlarını fark edebilmek; ihtiyaç ve beklentilerine göre ürün veya hizmet sunabilmek büyük avantaj sağlamaktadır. İhtiyaç ve beklentilerin doğru ve güvenilir tahmin edilmesi ancak yüksek bilgi akışıyla sağlanabilir. Şartların çok hızlı değiştiği ve geliştiği günümüzde piyasalara, pazarlara veya teknolojik gelişmelere hızlı uyum sağlayabilmek, organizasyonların hayatta kalmasında en önemli faktörü oluşturmaktadır. Organizasyonlarda esnekliğin sağlanabilmesi ancak gelecekte oluşacak durumların öngörülmesiyle ve durum oluşmadan gerekli hazırlıkların yapılmasıyla sağlanabilir.

Veri madenciliği, tanımlarda da belirtildiği üzere veri tabanlarındaki bilgilere odaklanarak yöneticilerin gelecekteki eğilimleri ve davranışları öngörmesini sağlamaktadır. Bu yönüyle veri madenciliği yönetimde proaktif yaklaşım için gerekli bilginin elde edilmesine olanak verir.

2.6 Veri Madenciliğinin Uygulama Alanları

Kitlesel veri toplama, güçlü çok işlemcili bilgisayarlar ve veri madenciliği algoritmaları alanlarında kullanılan teknolojiler geliştikçe veri madenciliğinin kullanım alanları genişlemekte ve yaygınlaşmaktadır. Veri madenciliği perakende, telekomünikasyon, üretim, bankacılık, finans ve sağlık sektörlerinde pazarlama yönetimi, sadakat yönetimi, müşteri ilişkileri yönetimi, risk yönetimi ve dolandırıcılık saptama vb. amaçlarla kullanılmaktadır. Veri Madenciliğinin kullanım alanları ve yüzdelik dağılımları Şekil 2.7’ de yer almaktadır (Kayaalp, 2007).

Son 3 yıl içinde veri madenciliğinin uygulandığı alanlar	
Bankacılık (51)	12%
Biyoteknoloji / Genetik (11)	3%
Kredi skorlama (35)	8%
CRM (52)	12%
Doğrudan pazarlama (34)	8%
e-Ticaret (11)	3%
Eğlence/ Müzik (4)	1%
Sahtekarlık tespiti (31)	7%
Şans oyunu (2)	0,01 %
Kamu uygulamaları (12)	3%
Sigortacılık (24)	6%
Yatırım / Hisse senedi (5)	1%
Junk email / Anti-spam (5)	1%
Sağlık/ İK (15)	4%
İmalat (19)	5%
Tıp/ Farmakoloji (12)	3%
Perakende (25)	6%
Bilim (17)	4%
Güvenlik / Anti-terörizm(5)	1%
Telekomünikasyon (23)	5%
Seyahat (8)	2%
Web (9)	2%
Diğer (11)	3%

Şekil 2.7 : Veri madenciliğinin kullanım alanları ve yüzdeleri.

Günümüzde hemen hemen bütün sektörlerde organizasyonların en büyük yaşadığı problem müşteri kaybıdır. Organizasyonlar bu problemin önüne geçmek için müşteri ilişkileri yönetimi ve sadakat yönetimi gibi kavramlar kullanmaktadırlar. Organizasyonların; müşterilerinin rakiplerine geçmesini engellemek için çeşitli pazarlama taktikleri uygulamaları, ürünleri sürekli yeni sunular ile çeşitlendirmeleri ve yenilemeleri gerekmektedir. Kendi müşterisiyken rakiplerin müşterileri olanlarla ilgili çeşitli analizler yapılarak rakiplerini tercih eden müşterilerinin özelliklerini ve tercih nedenlerini elde etmek mümkün olabilmektedir. Bu durumun sonucunda gelecek dönemlerde kaybetme olasılığı olan müşterilerin kimler olabileceği konusunda tahminlerde bulunarak, müşterilere özgü ürünler ve servisler geliştirebilirler. Organizasyonlar hangi müşterilerini kaybedebileceklerini önceden belirleyebildikleri zaman, bu müşterilerini elde tutma için çeşitli stratejiler geliştirebilirler.

Yapılan arařtırmalar sonucunda müşteriyi elde tutmak için yapılacak maliyet her zaman için müşteriyi geri kazanmak için yapılacak maliyetten düşük olduđu sonucuna varılmıştır. Mevcut ve potansiyel müşteriler hakkında detaylı bilgiye sahip olmak rekabetçi kalmak için önemli bir gerekliliktir. Farklı müşteri grupları için en uygun ürünleri bulmak, hangi tip müşterilerin ne tür ürün aldıklarını belirlemek, müşteri tabanını gruplara ayırmak, bu grupların karlılıklarını belirlemek ve buna göre farklı seviyelerde hizmet sunmak mümkün olabilir. Ürün veya hizmette hangi özelliklerin ne derecede müşteri memnuniyetini etkilediğı, hangi özelliklerinden dolayı müşterin bunları tercih ettiğı ortaya çıkarılabilir. Ürün veya hizmet sunumuna kimin yanıt vereceğini tahmin etmek maliyet düşürmek açısından önemli yöntemlerden biridir. Bir ürün veya hizmet ile ilgili bir kampanya programı oluşturmak için hedef kitlenin seçiminden başlayarak bunun hedef kitleye hangi kanallardan sunulacağı kararına kadar olan süreçte veri madenciliğı kullanılabilir. Aynı grubun geçmiş davranışlarına dayandırılabilceğı gibi, mantıksal bir alternatif popülasyonun davranışlarına da dayandırılabilir. Tüm bu anlatılan müşteriyle ilgili bilgilerin analiz edilmesi işlemlerinde veri madenciliğı tekniklerinden yararlanılabilmektedir. Finans sektöründe müşterilerle ilgili yüksek miktarda kayıt tutulmaktadır. Tutulan bu kayıtlar işletmenin amaçları ve politikaları doğrultusunda servisleri geliştirme ve rekabet avantajı sağlamak amacıyla kullanılabilir fakat veri miktarının çok olması verilerin analizini ve yorumlanmasını zorlaştırmaktadır. Verilerin analizi için veri madenciliğı teknikleri kullanılmaktadır. Finans sektöründe veri madenciliğı pazarlama alanında olduğı gibi müşteri profili ve en iyi müşteri segmentlerinin belirlenmesinde kullanılabilir. Elde edilen müşteri segmentleri doğrultusunda; müşteri gruplarına göre pazarlama kampanyaları oluşturulabilir. Örneğin New York'taki Chase Manhattan Bankası müşterilerini rakiplerine kaybetmeye başlayınca, müşteri hesaplarını analiz etmek ve kendi hesap gereksinimlerinde değişiklikler yapabilmek için veri madenciliğı kullanmaya başlamış, bu sayede karlı müşterilerinin hangi grup olduğunu anlamış, o gruba özel kampanyalar yaparak; karlı müşteri grubunu elinde tutabilmiştir. Pazarlama haricinde risk yönetimi konusunda da veri madenciliğı tekniklerinden yararlanılabilir. Mevcut müşterileri verilerinden kredi risk davranış modelleri oluşturarak, yeni başvurularda riskin en aza indirilmesini sağlamakta veri madenciliğı tekniklerinden yararlanılabilir. Riskli müşterilere kredi verilmesi engellenip, olası ödememe durumlarına karşı önleyici bir yaklaşım sergilenmiş olur. Kredi kartı ödemelerini

aksatan, gecikmeli olarak yapan veya hiç yapmayanların özelliklerinden yola çıkarak bundan sonra aynı duruma düşebilecek muhtemel kredi sahiplerini saptamada; bunun sonucunda kötü ödeme performansı gösteren müşterilerin ortak özelliklerini belirleyerek, benzer özelliklere sahip tüm müşteriler için politikalar geliştirmekte de yararlanılabilir. Firma için finansal kayıp oluşturabilecek müşterileri veya müşteri adaylarını belirleyerek bunlar ile çalışılmamasını sağlanabilir. Dolandırıcılık tespitinde de veri madenciliği teknikleri kullanılabilir. Aykırı değer analizi ile verilerin analizi yapılarak istisnai durumlar tespit edilir ve dolandırıcılık olup olmadığı anlaşılır. Kredi kartı kaybolduğunda bankalar kaybedilme sırasında oluşan zararın bir kısmını üstlenmektedirler. Bu dönemde oluşan zararları azaltmak için sahtekarlığı tespit edici sistemler vurgulanmaktadır. Müşterilerin tipik harcama biçimlerini önceden tanımlayıp, harcama eğilimlerinde oluşan ani değişiklikleri tespit etmek ve bu doğrultuda satın alma işlemlerini onaylamayı durdurmak kullanılan yöntemlerden biridir. Bu konulara ek olarak finansal değişkenler arası bağımlılık ölçümleri ve müşteri teminat yapılarının belirlenmesi gibi konularda da veri madenciliği teknikleri kullanılabilir. Üretici kuruluşlar veri madenciliği ile tahmin analiz çözümlerinin analitik gücünü arkasına alarak, başarılarını etkileyen hem kurum içi konular hem de dış faktörleri kolayca yönetebilirler. Üretim sektöründe çok boyutlu analizler ile hammadde seçimi ve tedariği, üretim sorunlarının nedenleri, ürünler müşterilere gönderilmeden önce istatistiki kalite testleri ve kalite sorunlarını çözümlenebilir. Üretim hatalarının nedenlerinin bulunmasında, makine bozulma sebeplerinin tespit edilmesinde, müşterinin tercih ettiği modellerin ve ürünlerin analiz edilmesinde, ürün fiyatlandırmanın teknik açıdan analizi, stok maliyetleri ve hacminin analizi stok maliyetleri ve hacminin analizi, ürün fiyatlandırmalarının teknik açıdan analizi ve müşteri beklentilerinin teknik analizi gibi konularda veri madenciliği teknikleri kullanılabilir. Tedarikçi kararı verilmesinde, ürün portföyünün genişletilmesi veya daraltılması gibi soruların cevaplandırılmasında da veri madenciliğinden yararlanılabilir. Telekomünikasyon sektörü hızla artan teknoloji, abone sayısındaki artış, katma değerli servislerle kazandığı zenginlik gibi sebeplerden dolayı çok hızlı büyüyen bir veri hacmine sahiptir. Telekomünikasyon sektöründe veri madenciliği teknikleri müşterilerin operatör değişikliğinin erkenden tespiti amacıyla kullanılabilir. Bu sayede elde tutulan pazar payının kaybedilmemesi sağlanarak rekabette güçlü bir pozisyon sağlanabilir. Dolandırıcılık tespitinde olduğu gibi hiçbir ödeme yapmama niyetiyle abone olanların erkenden tespiti sağlanarak

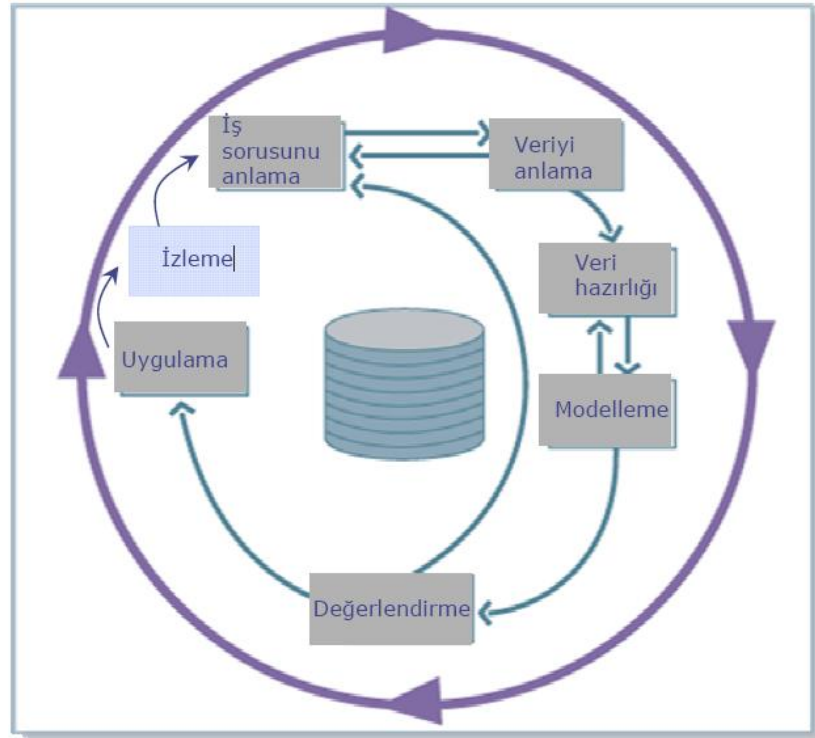
maddi kayıp önlenmiş olur. Bu tespit de aykırı değer analizi yardımıyla yapılabilir. Abonelik klonlamalarını saptayan veri madenciliği yöntemleri sayesinde abonelerin mağduriyeti sonucu memnuniyetsizliğinin giderilmesinde de veri madenciliği tekniklerinde yararlanılabilmektedir. Uluslararası dolaşım/roaming anlaşmalarının optimize edilmesinde de Veri madenciliği tekniklerinden yararlanılabilmektedir. Operatörlerin sahip oldukları uluslararası görüşen kullanıcılarının kullanım alışkanlıklarına ve anlaşılabilir operatörlerin fiyatlama ve servis sağlama özelliklerine göre ülkelere göre stratejiler belirlenebilir. Yine veri madenciliği ticari her sektörde olduğu gibi müşterilerin ayrıştırılması, segmentasyonu gibi konularda da etkili bir şekilde kullanılmaktadır. Telekomünikasyon sektöründe veri madenciliğinin en sık kullanıldığı alan ağ hatalarının tespitidir. Bu tespit 4 aşamayı barındıran bir veri madenciliği tabanlı akıllı bir bakım sistemi kurulabilmekte; bu aşamalar ise (i) Hata veritabanının oluşturulması, (ii) Hata veritabanının işlenmesi, (iii) örüntü belirleyen kuralların çıkarılması ve (iv) kuralların yorumlanması olarak sıralanabilir (Klemettinen ve diğ., 1999).

Sigortacılık sektöründe de veri madenciliği tekniklerinden yararlanılmaktadır. Finans sektöründe olduğu gibi sigortacılıkta da dolandırıcılık tespiti ve riskli müşterilerin tespit edilmesi alanlarında yararlanılabilir. Ona ek olarak veri madenciliği teknikleri kullanılarak talep edilecek poliçeler tespit edilebilir.

3. VERİ MADENCİLİĞİ SÜRECİ

Veri madenciliğinin birçok disiplini barındıran yapısı ve farklı uygulama alanlarındaki görevlerle prosedürlerin çeşitliliği, standart bir endüstri metodolojisi oluşturma yolunda karşımıza çıkan başlıca problemlerdir. Standart bir uygulama metodolojisi, teknoloji uygulamasını daha ucuz, daha güvenilir, daha kullanışlı ve daha hızlı bir hale getirebilir. Bunların dışında bir metodoloji, veri madenciliği teknolojisini daha kolay uyum sağlayabilir ve anlaşılabilir kılacaktır (Wirth ve Hipp, 2001).

Veri madenciliği süreciyle ilgili en yaygın kullanılan model CRISP-DM (Cross Industry Standard Process for Data Mining) modelidir. CRISP-DM süreç modeli, Daimler Chrysler AG, SPSS ve NCR gibi lider veri madenciliği kullanıcıları ve tedarikçilerinden oluşan bir konsorsiyum tarafından geliştirilmiştir (Göral, 2007). Şekil 3.1’de CRISP-DM süreci gösterilmektedir (Chapman ve diğ., 1996).



Şekil 3.1 : CRISP_DM veri madenciliği süreci.

3.1 İş Sorusunu Anlama

İş sorusunu anlama veri madenciliği sürecinin ilk adımını oluşturmaktadır. Bu adımda projenin amaçları ve iş gereksinimleri belirlenir. Belirlenen proje amacı, işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalıdır. Elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği öncesinde tanımlanmalıdır. Bu adım süresince sonuçların nasıl kullanılacağını bilmek büyük önem taşımaktadır. Bu safha veri madenciliği sürecinin beklentilerinin ve standartlarının saptandığı adımdır. Ayrıca çalışılan verilerin kalitesini öğrenmeye yardımcı olmaktadır. Eğer çok fazla iterasyona gereksinim duyuluyorsa ve sonuçlar kabul edilemeyecek kadar belirsizlikle sonuçlanıyorsa, problemin tanımı yerine verinin kalitesi adımı üzerinde yoğunlaşılmalıdır.

3.2 Veriyi Anlama

Veriyi anlama adımı verileri toplamakla başlar. Daha sonra verinin içindeki değişkenler tespit edilir ve bu değişkenlerin neyi ifade ettikleri anlaşılmaya çalışılır. Eğer analist veriler hakkında bilgi sahibi değilse, veri hakkında bilgi sahibi olan bir kişiden yardım almalıdır. Veri hakkında bilgiye sahip olmadan veriyi anlaması ve neticesinde doğru bir model kurabilmesi mümkün değildir. Bu da projenin başarısızlığına neden olur. Gerekli veriler iç ve dış kaynaklar olmak üzere iki türlü temin edilebilir. İç kaynaklar, işletmenin veri tabanlarıdır. Müşteri kayıtları, geçmiş teklifler ve satın almalar, işlem kayıtları vs. iç kaynaklara örnektir. Dış kaynaklar ise işletme dışından elde edilen verilerdir. Bu verilere örnek olarak nüfus sayımı, demografik analizler, merkez bankası kayıtları, hava durumu raporları, Türkiye İstatistik Kurumu ve pazar araştırma şirketleri veri tabanları vs. verilebilir.

3.3 Veri Hazırlığı

Veri tabanlarındaki bilgilerin tamamının gerçek ve doğru bilgiler olduğu kesin değildir; ayrıca bu bilgilerin, mevcut haliyle yapılan çalışmaya hizmet edeceği garanti edilemez. O nedenle elimizdeki verilerin çeşitli işlemlerden geçmesi gerekmektedir. Verilerin hazırlanma aşaması analistin toplam süresinin % 70-80 ini almaktadır.

Verilerin kaynağı daha öncede belirtildiği gibi farklı kaynaklardan olabilir. Bu şekilde farklı kaynaklardan elde edilen bilgilerin belirli bir düzen dahilinde birleştirilmesi gerekmektedir. Belirli bir standarda dönüştürülmeyen veriler ileride büyük uyumsuzlukların yaşanmasına neden olabilmektedir. Verilerin içerisinde çok sayıda değişken yer almaktadır ve bu değişkenlerin saklanma şekilleri birbirlerinden farklılık gösterebilir. Verilerin incelenmesinde değişkenlerin saklanma şekilleri;

- Dizi (String): Değişken karakter dizileri içeriyorsa seçilir. Örneğin isim, adres, e-mail.
- Sayı (Number): Hesaplanabilir sayılar içeren değişkenlerdir. Örneğin satış bilgileri, mil puanları vs. Telefon numarası gibi değişkenler bu kategoride yer almamalıdır.
- Tarih (Date): Değişken sadece tarih bilgisi içeriyorsa seçilir. (Dönmez, 2008)

Veri türleri nitel ve nicel veriler olmak üzere ikiye ayrılır. Nitel veriler kategorize edilmiş değişkenlere olarak tanımlanabilir. Nitel veriler nominal ve ordinal veriler olmak üzere ikiye ayrılırlar. Değişkenlerin birbirinden farklı olduğu ve bu farklılık da herhangi bir üstünlük olmadığı durumdaki değişkenler nominal verileri oluşturmaktadır. Nominal veriler hem sayısal hem de karakter dizileri şeklinde olabilir. Nicel veriler sayısal büyüklükler biçiminde ifade edilen verilerdir. Nicel veriler aralık ve oran veriler olmak üzere ikiye ayrılırlar. Verilerin belirli bir sıra izlediği ve aralarındaki farkın anlamlı olduğu verilerdir. Örneğin hava sıcaklığı Ankara’da 10°C, İstanbul’da 12°C ve Eskişehir’de 6°C olsun. Sıcaklıklarına göre sıralama yaparsak en sıcak il İstanbul; en soğuk il Eskişehir olmaktadır. Ankara ve İstanbul arasındaki sıcaklık farkı 2°C’dir ve sıcaklık farkı bir anlam ifade etmektedir. Oran verilerde ise aralık verilere ek olarak iki verinin oranı da anlamlıdır. Örneğin fiyatları 8 Türk Lirası (TL) ve 4 TL olan 2 bisikletten 8 TL’lik olan diğerinden 4 TL daha pahalıdır, aralarındaki fiyat farkı 4 TL’dir ve 8 TL’lik bisikletin fiyatı diğerinin 2 katıdır. Yukarıda yapılan sınıflandırmalara göre verinin standart yapısına karar verilir.

Verinin standart yapısına karar verildikten sonra veri tabanlarında ileride problem çıkarabilecek veri türlerine dikkat etmek, ayıklamak ve gerekli düzeltmeler yapmak gerekmektedir. Örneğin veri tabanında bazı kayıtlarda eksiklikler olabilir. Bu eksiklik kayıp veriler (missing data) olarak isimlendirilmektedir. Diğer bir örnek olarak ek olarak veri tabanlarındaki bazı kayıtlar aşırı uç değerler ya da yanlış

girilmiş değerler olabilir. Bu gibi verilere gürültü yada gürültülü veri denir. Bazı durumlarda ise aynı anlama gelebilecek birden fazla veri olabilir. Müşterilerin hem doğum tarihlerinin hem de yaşının kayıtlarının tutulması bu duruma örnek olarak verilebilir. Bu durumda verilerden bir tanesi kesinlikle fazladır. Bazı durumlarda ise eldeki değişkenlerin bir ya da birkaçı birleştirilip tek bir değişken halinde ifade edilebilir. Tek değişken olarak ifade edilmesi elde edilecek sonuçların hem güvenilirliğini ve kalitesini artıracaktır hem de bilgisayar çalışma zamanı karmaşıklığını azaltacaktır. Bu tür veriler artık veriler olarak adlandırılmaktadır. Bazı durumlarda ise zaman ve koşullar değiştikçe bazı verilerin güncellenmesi gerekmektedir. Bu şekildeki veriler eskimiş veri olarak adlandırılmaktadır. Bu durumların dışında bazı veriler gerçekten yanlış ve anlamsız bilgiler içerebilir. Veri tabanları genel olarak veri madenciliği dışındaki amaçlar için tasarlandığından amaca hizmet edecek veriler eksik olabilir. Bu da veritabanlarında sınırlı bilgi bulunmasına sebep olur. Bu durumda dış kaynaklardan veri seti tamamlanacağından veri standart yapısı bozulacaktır ve hazırlık süresi artacaktır (Dönmez, 2008).

Büyük miktarda veriler projenin bilgisayar zaman karmaşıklığını artırmaktadır. Zaman zaman bu karmaşıklığı önlemek için tıpkı istatistik çalışmalarında olduğu gibi ana küleden bir örneklem alınarak verinin boyutu düşürülebilir. Ancak bu örnekleme yapılıbilmesi için istatistik çalışmalarından farklı olarak elimizde ana külenin verilerinin tamamının bulunması gerekmektedir. Örnekleme yapmak analiste projenin bitiminde karşılaşılabileceği sorunlar, hangi değişkenlerin daha önemli olduğu vb. konularda ön bilgi verir.

Veri hazırlığı adımı verilerin temizlenmesi ve verilerin yeniden yapılandırılması olmak üzere iki başlık altında incelenmektedir.

3.3.1 Veri temizleme

Verilerin temizlenmesi adımı, gürültülü, yanlış ve aşırı uçta bulunan verilerin etkileri ortadan kaldırılmaya çalışılır.

Kayıp verilerin oluşturabileceği sorunları ortadan kaldırmak için kullanılan teknikler:

- Eğer kayıp verili kayıt sayısı, toplam kayıt sayısına oranlandığında sonuçları etkilemeyecek kadar önemsizse; kayıp verinin bulunduğu kayıtlar veri kümesinden çıkarılabilir ya da bu tarz kayıtlar iptal edilebilir. Eğer kayıp

veriler sonuçları etkileyecek kadar önemli bir orandaysa, bu yöntem kullanışsız olup, gerçekten uzak sonuçlar verecektir.

- Kullanılan veri tabanları küçükse, kayıp verilere ulaşabilmek mümkünse ve verilere gerçekten ihtiyaç duyuluyorsa, kayıp veriler elle tek tek doldurulabilir. Fakat bu yöntemin kullanılabilmesi için zamana ihtiyaç vardır.
- Tüm kayıp verilere aynı bilgi girilebilir. Örneğin öğrenim durumu boş olanlara B harfi girilebilir. Bu yöntemin bazı olumsuz yönleri olabilir, örneğin veri madenciliği algoritmaları, öğrenim durumu B olanları ayrı bir küme olarak düşünüp ayırt edici bir özellik gibi bir sonuç çıkabilir.
- Kayıp olan verilere tüm verilerin ortalama değeri verilebilir. Bu yöntemin kullanılması için verilerin nicel veriler olması gerekmektedir. Kategorik verilerde bu yöntemi kullanmak mümkün değildir.
- Eksik verilere, değerleri kendisine çok benzeyen diğer kayıtlardaki değerler verilebilir.
- Eksik veriler tahmin yöntemleri aracılığıyla, mevcut veriler kullanılarak tahmin edilebilir. Regresyon, zaman serileri analizi, Bayesyen sınıflandırma, karar ağaçları ve maksimum beklenti (expectation maximization) tahmin yöntemlerine örnek olarak verilebilir.

Kayıp verilerin dışında temizlenmesi veya düzgünleştirilmesi gereken yanlış veya tutarsız girilmiş veriler, gürültülü veriler ve aşırı uçlarda olan veriler vardır. Bu tür verilerin düzgünleştirilmesinde kullanılan yöntemler:

- Veriler sıraya dizilir ve belirlenen küme sayısına göre eşit derinlikte alt kümeler ayrılır. Bu aşamalardan sonra 3 değişik yöntem uygulanmaktadır.

Örneğin D: {2, 1, 8, 12, 11, 5, 16, 9}

Veriler sıraya dizildikten sonra D: {1, 2, 5, 8, 9, 11, 12, 16}

Küme sayısı 2 olsun. Her bir kümenin eleman sayısı 4 olur.

$D_1 = \{1, 2, 5, 8\}$

$D_2 = \{9, 11, 12, 16\}$

İlk yöntemde, sonraki adım olarak her bir kümenin aritmetik ortalaması alınır ve küme içindeki veriler bu aritmetik ortalamayla değiştirilir.

1. Kümenin aritmetik ortalaması = $(1+2+5+8) / 4 = 4$

2. Kümenin aritmetik ortalaması = $(9+11+12+16) / 4 = 12$

Bu durumda düzgünleştirilmiş veri tabanı aşağıdaki gibi olur:

$D = \{4, 4, 4, 4, 12, 12, 12, 12\}$

İkinci yöntemde sınırlar yardımıyla düzgünleştirme yapılmaktadır. Her bir küme içindeki alt ve üst sınırlar belirlenir. Küme elemanları alt veya üst sınırdan hangisine yakınsa o değeri alır.

1. Kümenin alt ve üst sınırı = 1 ve 8

2. Kümenin alt ve üst sınırı = 9 ve 16

Kümelerin düzgünleştirilmiş veri tabanı aşağıdaki gibi olur:

$D = \{1, 1, 8, 8, 9, 9, 9, 16\}$

Üçüncü yöntemde ise kümelerde yer alan uçtaki verilerin aralarındaki farkın küme sayısına bölümüyle elde edilen değerler, o kümenin elemanları yerine atanır.

1. Küme için değer = $(8-1) / 4 = 1,75$

2. Küme için değer = $(16-9) / 4 = 1,75$

Kümelerin düzgünleştirilmiş veri tabanı aşağıdaki gibi olur:

$D = \{1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 1,75\}$

Kümeleme yöntemi kullanılarak uçtaki veriler belirlenip bunlara yeni değerler atanabilir. Bu yöntemde mevcut veriler birbirlerine olan yakınlıklarına göre kümeler ayrılır. Bu kümeleme işlemi sırasında uç değerler herhangi bir kümeye dahil olamayacaklardır. Bu uç değerler her birine en yakın olduğu kümenin ortalama, en küçük veya en büyük değerlerinden biriyle değiştirilebilir. Böylece verilerin hepsi düzgünleştirilmiş olur.

- Uçta bulunan veriler tahmin yöntemleri aracılığıyla, mevcut veriler kullanılarak düzgünleştirilebilir. Regresyon, zaman serileri analizi, Bayesyen sınıflandırma, karar ağaçları ve maksimum beklenti tahmin gibi yöntemler kullanılarak tahmini değerler bulunur. Ve bu tahmini değer uçtaki değerler yerine atanır.
- Bu yöntemlerin haricinde Hızlı Fourier transformasyon yöntemiyle de verilerin düzgünleştirilmesi sağlanabilir (Kosarev, 1983)

3.3.2 Veri dönüştürme

Veri madenciliği çalışmasında kullanılan algoritma, model veya tekniklerin belirli türdeki verilerle çalışabilip bazı türdeki verilerle çalışamamaktadırlar. Bu durumda eldeki verilere, uygulanacak algoritmaya uygun hale getirmek için bazı dönüşümler yapmak gerekmektedir. Veri dönüştürmede kullanılan teknikler aşağıda özetlenmektedir.

- Bazı algoritmalar sürekli değerler yerine aralıklı değerler kullanırlar. Örneğin hız değişkeni, 20 ile 140 arasında değişen çeşitli değerler alabiliyorsa, bu değerler 20-50; 51-80; 81-110; 111-140 gibi aralıklara bölünerek aralıklı değerler elde edilmiş olunur. Karar ağaçları aralıklı değerler kullandığından, sürekli değerlere bu dönüşümün yapılması gerekmektedir.
- Yapay sinir ağları gibi algoritmalar 0,0-1,0 değerleri arasında çalışmaktadır. Eldeki verilerin 0,0-1,0 gibi aralıklara indirilmesi işlemine normalizasyon denir. Normalizasyon işlemi için çeşitli yöntemler vardır, en yaygın olarak kullanılan yöntem min-maks normalizasyonudur.
Min-maks normalizasyonu yönteminde; veriler doğrusal olarak normalize edilir. *min* bir verinin alabileceği en küçük değeri, *max* ise verinin alabileceği en büyük değeri ifade eder. Veriyi 0,0-1,0 aralığına indirmek için min-maks yöntemi 3.1'deki formül ile kullanır.

$$s' = (s - \min) / (\max - \min) \quad (3.1)$$

- *s* verinin normalize halini, *s* ise verinin orijinal halini ifade etmektedir. *max* en büyük değeri, *min* ise en küçük değeri ifade etmektedir.
Eğer normalizasyon aralığı 0,0-1,0 dan farklı bir aralıksa bu durumda yeni formül 3.2'deki gibi olur.

$$s' = (s - \min) / (\max - \min) * (\text{yeni}_{\max} - \text{yeni}_{\min}) \quad (3.2)$$

yeni_{max} normalizasyon aralığının en büyük değerini, *yeni_{min}* ise en küçük değerini ifade etmektedir.

- İstatistik çözümlemelerinde sıkça kullanılan dönüşüm biçimi Z-score olarak adlandırılmaktadır. Bu yöntemde, verilerin ortalaması ve standart hatası kullanılarak yeni veriler elde edilir. Dönüşümde kullanılan formül 3.3'de yer almaktadır.

$$X^* = \frac{X - \bar{X}}{\sigma_x} \quad (3.3)$$

- X^* dönüştürülmüş değerleri, X gözlem değerlerini, $\bar{X}$ verilerin ortalama değerini ve σ_x de verilerin standart sapmasını ifade eder.
- Klasik veri madenciliği algoritmalarının birlikte çalışabildikleri değişken sayısı sınırlıdır. Bununla birlikte veri kümesine eklenen her bir değişken, analiz sürecini karmaşıktırıp, modelleme süresinin uzamasına sebep olacaktır. Dolayısıyla bu tür problemlerle karşılaşmamak için analiz öncesinde değişkenler üzerinde kapsamlı çalışmalar yapılması gerekmektedir. İndirgeme işlemi değişken sayısının azaltılmasıyla gerçekleşir. Gereksiz tutulan bir değişkenin kaldırılması, birden fazla değişkenin tek bir değişken şeklinde ifade edilmesi indirgeme işlemlerine örnektir. Algoritmanın yapısı ve çıkacak sonuçların hassasiyeti açısından, belirli değişkenlerin birleştirip tek değişken olarak ifade edilmesi mümkün olabilmektedir. Bu birleştirme işleminde yaygın olarak kullanılan yöntemler dalga dönüşümü (wavelet transform) ve temel bileşenler analizidir (principal component analysis). Karhunen Loeve yöntemi olarak da bilinen temel bileşenler analizi bir değişkenler kümesinin varyans-kovaryans yapısını, bu değişkenlerin doğrusal birleşimleri vasıtasıyla açıklayarak boyut indirgenmesi ve yorumlanmasını sağlayan, çok değişkenli bir istatistik yöntemidir (Silahtaroğlu, 2008). Bu yöntemde p adet değişken; doğrusal, ortagonal ve birbirinden bağımsız olma özelliklerini taşıyan k ($k \leq p$) tane yeni değişkene dönüştürülmektedir (Yamaç, 2002). Dönüştürülen k adet değişken, p adet değişkeni temsil edebilmektedir. Yeni oluşturulan k adet değişken gerçek değişkenlerin doğrusal bileşimidir. Bu indirgemedede önemli bir varyans kaybı olmamaktadır.
- Bazı durumlarda değişkenlerden yeni değişkenler türetip eklemek modele güç katabilir. Müşterilerin geçen yıla göre harcamalarında görünen artış yüzdesi örnek olarak verilebilir. Değişkenin değeri bu yıl yapılan harcamaların geçen yıl yapılan harcamalarına oranlanarak elde edilir. Bu şekilde türetilen değişkenler faydalı ilişkiler, örüntüler keşfetmek için yeni parametreler sağlar.

3.4 Modelin Kurulması

Veri setinin hazırlanması ile çeşitli veri madenciliği modelleri uygulanır.

En uygun modeli bulabilmek için olabildiğince çok sayıda model kurulup denenmelidir. Bazı modeller için verileri yeniden çeşitli işlemlere sokmak gerekebilir. O nedenle modelin kurulması ve veri hazırlama aşamaları en uygun model elde edilinceye kadar tekrarlanan süreçtir.

Model kuruluş süreci denetimli (supervised) ve denetimsiz (unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir (Giudici, 2003).

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilmektedir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir (Akpınar, 2000).

Öğrenme süreci bittiğinde tanımlanan kural cümleleri yeni örneklerle uygulanır, yeni örneklerin sınıfları bu şekilde belirlenmiş olur. Denetimli öğrenimde verilerin bir kısmı öğrenmeyi test etmek için kullanılabilir. Verilerin bir kısmı test verileri olarak ayrılır. Öğrenme sonucu elde edilen kural cümleleri test verilerine uygulanır ve modelin doğruluk derecesi bu şekilde belirlenir.

Denetimsiz öğrenimde, veriler benzer özelliklerine göre gruplanarak sınıflama tanımları yapılır. Kümeleme analizi denetimsiz öğrenmeye örnektir.

Bazı durumlarda öğrenme çok ileri seviyede gerçekleşmektedir. Böyle durumlarda kurulan modelden elde edilen sonuçlar yalnızca öğrenme kümesindeki veriler için geçerli olmaktadır. Bu durum aşırı öğrenme olarak adlandırılmaktadır. Aşırı öğrenmeye öğrenme kümesindeki verilerin azlığı, öğrenme kümesindeki uç veriler gibi unsurlar neden olmaktadır.

3.5 Değerlendirme

Projenin bu aşamasında analistin elinde kalitesi yüksek bir model mevcuttur. Modelin uygulama aşamasına geçmeden önce modelin eksiksiz olarak değerlendirilmesi ve iş amaçlarına ulaşıp ulaşılmadığından emin olmak için model oluşturulana kadar yürütülen adımların tekrar gözden geçirilmesi büyük önem taşımaktadır. Buradaki temel amaç, analiz süresince gözden kaçan önemli bir noktanın var olup olmadığını belirlemektir (Göral, 2007).

Bir modelin doğruluğunun test edilmesi için kullanılan en basit yöntem, basit geçerlilik (Simple Validation) testidir. Bu yöntemde verilerin %5 ile %33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanmaktadır (Akpınar, 2000).

Sınırlı miktarda veriye sahip olunmadığı durumlarda, kullanılabilecek diğer bir yöntem çapraz geçerlilik (Cross Validation) testidir. Bu yöntemde veri kümesi tesadüfi olarak iki eşit parçaya ayrılır. İlk aşamada A parçası üzerinde model eğitimi ve B parçası üzerinde test işlemi; ikinci aşamada ise B parçası üzerinde model eğitimi ve A parçası üzerinde test işlemi yapılarak, elde edilen hata oranlarının ortalaması kullanılmaktadır (Akpınar, 2000).

Bir kaç bin veya daha az satırdan meydana gelen küçük veri tabanlarında, verilerin n gruba ayrıldığı n katlı çapraz geçerlilik (N-Fold Cross Validation) testi tercih edilebilir. Verilerin örneğin 10 gruba ayrıldığı bu yöntemde, ilk aşamada birinci grup test, diğer gruplar ise öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır (Akpınar, 2000).

Bootstrapping, küçük veri kümeleri için modelin hata düzeyinin tahmininde kullanılan bir başka tekniktir. Çapraz geçerlilikte olduğu gibi model bütün veri kümesi üzerine kurulmaktadır. Daha sonra en az 200, bazen 1000'in üzerinde olmak üzere çok fazla sayıda öğrenim kümesi tekrarlı örneklemlerle veri kümesinden oluşturularak hata oranı hesaplanmaktadır (Akpınar, 2000).Önemli diğer bir değerlendirme kriteri, modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir çok kuruluş uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşırsalar da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir (Yaralıoğlu, 2008).

Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir (Göral, 2007).

3.6 Uygulama

Veri madenciliği modeli oluşturulup, modelin çıktılarının elde edilmesiyle süreç tam olarak tamamlanmamış olur. Sonuçların değerlendirilmesi ve yorumlanması; elde edilen sonuçlar neticesinde belirlenen eksiklikler, stratejiler, sürecin eksik tarafları gerçek hayatta uygulanır. Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak da kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabilmesi gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine de gömülebilir (Eker, 2005).

3.7 İzleme

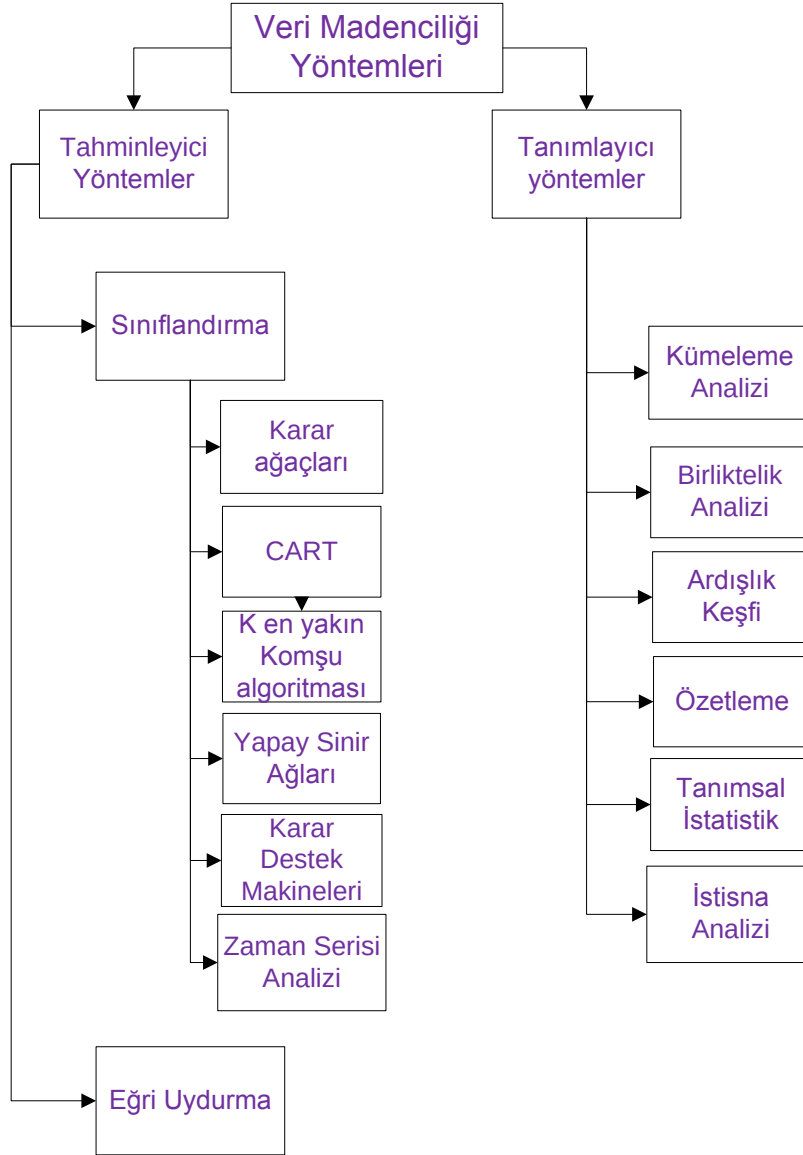
İlerleyen zamanla beraber bütün sistemlerde değişiklikler meydana gelir. Bu değişimler sistemin ürettikleri verilerin de değişmesine yol açmakta ve bu durumda modelleri geçersiz kılmaktadır. Geçerliliğini yitirmiş modellerin güncellenmesi gerekmektedir. Hangi modellerin güncellenmesi gerektiğini anlamak için ise modeller izlenmeli meydana gelen değişim gözlenmelidir.

4. VERİ MADENCİLİĞİ MODELLERİ

Veri madenciliği yöntemleri tahminleyici modeller ve tanımlayıcı modeller olmak üzere ikiye ayrılırlar.

Tahminleyici modeller; sonuçları bilinen verilerden bir model geliştirmekte, kurulan model kullanılarak sonuçları bilinmeyen veriler için sonuç tahmini yapmakta kullanılır. Örneğin ders geçme notu üzerinde not alan öğrenciler dersi geçmektedir. Ders geçme notu; dersi veren öğretim görevlisi, sınavların zorluk derecesi ve dersi alan öğrenci sayısına göre değişmektedir. Öğretim görevlisi, sınavların zorluk derecesi ve dersi alan öğrenci sayısı bağımsız değişkenler; ders geçme notu ise bağımlı değişken olmak üzere model kurulur. Öğrencinin notu ve diğer değişkenlere göre öğrencinin dersi geçip geçmediği tahmin edilir.

Tanımlayıcı modellerde karar vermeye ise karar vermeye yardımcı olmak için verilerdeki örüntülerin tanımlanması sağlanmaktadır. Örneğin tanımlayıcı model sonucunda “Çocuk maması alan bir müşterinin, çocuk bezi alma olasılığı diğerlerine oranla 3 kat fazladır.” gibi ilişkileri ortaya çıkarır. Veri madenciliği model ve teknikleri Şekil 4.1’de gösterilmiştir.



Şekil 4.1 : Veri madenciliği model ve teknikleri.

Veri madenciliği modelleri temel olarak; sınıflandırma (classification), kümeleme (clustering) ve birliktelik kuralları (association rules) olmak üzere 3 şekilde gruplandırılmaktadır. 3 modele ek olarak tahmin (estimation), öngörme (prediction), zaman serisi analizleri (time series analysis) ve ardışlık keşfi (sequence discovery) gibi modeller de bulunmaktadır.

4.1 Sınıflandırma

Sınıflandırma en çok bilinen veri madenciliği modellerinden birisidir, resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları sınıflandırmanın bolca kullanıldığı alanlardır. Sınıflandırma tahminleyici bir modeldir; havanın bir sonraki gün nasıl olacağı ya da bir kutuda ne kadar mavi top olduğunun tahmin edilmesi aslında bir sınıflandırma işlemidir (Dunham, 2003).

Sınıflandırma bir öğrenme algoritmasına dayanır. Öğrenmenin amacı bir sınıflandırma modelinin oluşturulmasıdır. Sınıflandırma bir başka deyişle sınıflandırması belli olmayan bir kayıt için sınıf belirleme sürecidir.

Sınıflandırma işleminde mevcut sınıf veya istatistiksel tabiriyle bağımlı değişken hem sınıfsal hem de sürekli değer taşıyabilir; bu anlamda regresyon ve çok terimli regresyona yaklaşmaktadır (Akpınar, 2000). Veri madenciliği çerçevesinde bu istatistiksel yöntemlerin dışında sınıflandırma işleminde “Bayesyen sınıflandırma algoritması”, “karar ağaçlarına dayalı algoritmalar”, “yapay sinir ağları” temelli algoritmalar ve “k-en yakın komşu algoritması” gibi birçok teknik ve algoritma geliştirilmiştir (Lipmann, 1987).

4.1.1 Sınıflandırma süreci

Verilerin sınıflandırma süreci iki adımdan oluşur (Han, 2006).

İlk adım, veri kümelerine uygun bir modelin ortaya konulmasıdır. Söz konusu model, veri tabanındaki kayıtların nitelikleri yani değişkenler kullanılarak gerçekleştirilir. Sınıflandırma modelinin kurulması için veri tabanının bir kısmı rastsal olarak seçilerek eğitim verileri olarak kullanılır. Kalan veriler ise test verileri olarak kullanılır. Sonrasında eğitim verileri üzerinde bir algoritma uygulanarak sınıflama modeli elde edilir.

İkinci adımda ise test verileri üzerinde sınıflandırma kuralları belirlenir. Söz konusu kurallar bu sefer test verilerine uygulanarak sınanır. Test sonucu elde edilen modelin doğruluğu kabul edilecek olursa, bu model diğer veriler üzerinde uygulanır.

4.1.2 Karar ağaçları ile sınıflandırma

Verileri sınıflandırma yöntemlerinden biri olan karar ağaçları ile sınıflandırma, uygulamalı istatistikte makine öğrenmesi adı altında birçok şekilde uygulanmıştır.

Örneklerden oluşan veri tabanı kullanılarak karar ağacının oluşturulmasını sağlayan çok sayıda öğrenme yöntemi vardır. Karar ağaçları sınıflandırma problemlerinde en çok kullanılan algoritmalarından biridir. Diğer yöntemlerle kıyaslandığında karar ağaçlarının yapılandırılması ve anlaşılması daha kolaydır denilebilir.

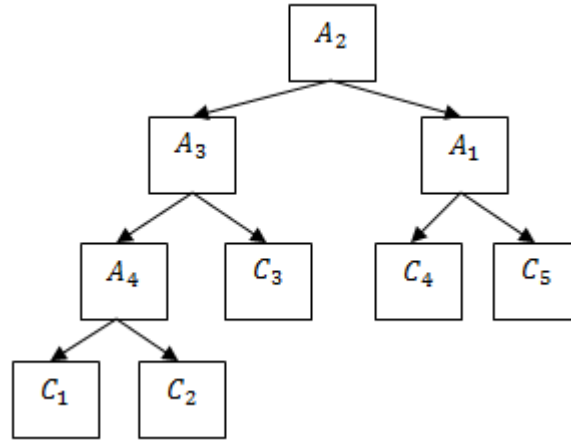
Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı yaprak, en üst yapı kök ve bunların arasında kalan yapılar ise dal olarak adlandırılır (Quinlan, 1993).

Örnek karar ağacı Şekil 4.2’de yer almaktadır.

Karar ağaçları tekniğinde; sınıflandırma için bir ağaç oluşturulur, daha sonra veri tabanındaki her bir kayıt bu ağaca uygulanır ve çıkan sonuçlara göre kayıtlar sınıflandırılır.

Bir karar ağacı aşağıdaki gibi tanımlanabilir:

- Her bir düğümü A_i alanıyla adlandırılmış
- Her düğümün ayrılan kollar bu alanla ilgili soruya yanıt veren
- Her yaprağın sınıf olduğu bir ağaçtır (Dunham, 2003).



Şekil 4.2 : Örnek karar ağacı.

Ağaçtaki $A_1, A_2, \dots, A_n$ ‘dan her biri bir düğümü oluşturmaktadır. Her düğüm kendinden sonra iki dala ayrılmaktadır. Her düğüm kendinden sonra dallara ayrılmaktadır. Bu ayrılma süreci, A_i düğümü hakkında cevabı veri tabanında bulunacak bir soru sorulmaktadır ve verilen cevaba göre bir dal izlenmektedir.

Ağaçtaki $C_1, C_2 \dots \dots C_n$ 'lerin her biri birer yapraktır ve bir sınıfı temsil etmektedir (Silahtaroglu, 2008).

Karar ağaçları oluşturulurken hangi algoritmanın kullanıldığı önemlidir çünkü kullanılan algoritmaya göre ağacın şekli değişmekte ve değişik ağaç yapıları da farklı sınıflandırma sonuçları vermektedir. Kök düğümün değişmesi, en uçtaki yapraklara ulaşmada izlenecek yolu değiştirmekte ve dolayısıyla sınıflandırma da değişecektir. Kök düğümün belirlenmesindeki en büyük kriter, o noktadan dallara ayrıldığında veri tabanının eşit parçalara ayrılmış olmasıdır. Yani kök düğüm belirlenirken öyle bir soru sorulmalıdır ki alınan cevaba göre veri tabanı cevap sayısına göre eşit parçaya ayrılabilmelidir. Örneğin cevap evet/hayır ise 2 eşit parçaya, evet/hayır/belki ise 3 eşit parçaya ayrılmalıdır. Buradaki amaç en kısa yoldan istenen sınıflara ulaşmaktır. Kök düğümünden sonra ağacın alt dallarını oluşturacak düğümler belirlenir. Bu tarz düğümlerin belirlenmesinde de kök düğüm belirleme işlemleri uygulanır. Daha önce de belirtildiği gibi her ağaç kendi sınıflandırmasını oluşturacağından bu düğümler belirlenirken kullanılan algoritmalar çok önemlidir.

Karar ağacı kurulurken eldeki veri tabanının bir kısmı öğrenme işlemi için kullanılmaktadır. Ağaç çalışırken kurulan sistemin çalışıp çalışmadığı belirlenir. Eğer ağaç belirlenen düzeyde çalışırsa dallanma durdurulur ve sınıflandırma tamamlanır. Durdurma kriteri ağacın hassasiyetini ortaya koyar. Geç durdurulan bir ağaç daha fazla dallanacak ve ağaç daha geniş olacak çalışma süresi uzayacaktır. Bunun karşılığında ise duyarlı sonuç verecektir. Erken durdurulan ağaç ise her ne kadar hızlı çalışsa da tam öğrenmenin gerçekleşmeme olasılığını her zaman taşıyacaktır (Dunham, 2003).

Ağaç oluşturmada yapılan işlemlerden bir tanesi de budama işlemidir. Budama ağaçta oluşmuş sonucu etkilemeyen ve sınıflamaya herhangi bir katkısı olmayan dalların ağaçtan ayrılmasıdır kısaca ağaçtaki gereksiz ayrıntıların sonuçtan çıkartılmasıdır. Ağaçta birçok dal ve düğüm oluşursa, ağacın alt dallar ve yapraklarına ulaşan veri sayısı da azalacaktır; bu da ağacın hassasiyetini azaltacaktır (Cabena, 1998).

Karar ağaçlarına dayalı olarak geliştirilen birçok algoritma vardır; bu algoritmalar kök, düğüm ve dallanma kriterleri belirlemede izledikleri yol bakımından birbirlerinden farklılık göstermektedirler. Karar ağaçlarında en önemli sorunlardan

birisi dallanmanın hangi kritere göre yapılacağını belirlenmesidir. Dallanmada kullanılan algoritmalar 3 grupta toplanmaktadır. Bu gruplar entropiye dayalı algoritmalar, Sınıflandırma ve regresyon ağaçları (CART) ve bellek tabanlı sınıflandırma algoritmalarıdır. Entropiye dayalı bölümlmeyi kullanan algoritmalara örnek olarak ID3 ve onun daha gelişmiş biçimi olan C4.5 algoritmaları verilebilir. Sınıflandırma ve regresyon ağaçları konusunda ise Twoing ve Gini algoritmaları kullanılmaktadır. Bellek tabanlı sınıflandırma yöntemleri arasında k-en yakın komşu algoritması sayılmaktadır (Özkan,2008).

Entropi; bir sistemdeki belirsizliğin ölçüsüdür şeklinde tanımlanmaktadır. Örneğin eldeki bütün veriler aynı bölüm mezunu olsun; bu durumda aynı sınıfa ait olmuş oluyorlar. Rastgele seçilen bir kimseye hangi bölüm mezunu olduğunu sorduğumuzda aldığımız yanıt bizi şaşırtmayacaktır. Bu durumda entropi 0'dır. Eğer eldeki tüm veriler farklı bölüm mezunu olsalardı yani sınıfları birbirlerinden farklı olsaydı; entropi en büyük değeri olan 1 olacaktı.

Entropinin hesaplanması aşağıda gösterilmektedir.

S bir kaynak olsun. Bu kaynağın $\{m_1, m_2, \dots, m_n\}$ olmak üzere n adet mesaj ürettiğini varsayalım. Tüm mesajlar birbirlerinden bağımsız olarak üretilmektedir ve m_i mesajlarının üretilme olasılıkları p_i 'dir. $P = \{p_1, p_2, \dots, p_n\}$ Olasılık dağılımına sahip mesajları üreten S kaynağının entropisi $H(S)$ şeklinde gösterilmektedir. Entropi miktarı formül 4.1 kullanılarak hesaplanır (Shannon, 1948).

$$H(S) = \sum_{i=1}^n p_i * \log_2 (p_i) \quad (4.1)$$

Karar ağaçlarının oluşturulması esnasında dallanmaya hangi nitelikten başlanacağı büyük önem arz etmektedir. Çünkü sınırlı sayıda kayıttan oluşan bir eğitim kümesinden yararlanarak olası tüm ağaç yapılarını ortaya çıkarmak ve içlerinden en uygununu seçerek ondan başlamak kolay değildir. Örneğin 5 değişken ve 20 elemanlı bir eğitim kümesindeki verilerden yararlanarak çizilebilecek karar ağaçlarının sayısı 10^6 dan daha büyüktür (Kantardzic, 2003). O nedenle karar ağaçları algoritmalarının büyük bir kısmında bir takım değerler hesaplanır ve eldeki bu değerlere göre ağaç oluşturma yoluna gidilmektedir. Gerekli değerleri hesaplama entropi kavramı kullanılabilir ve bu şekilde karar ağaçları oluşturulabilir.

ID3 algoritması ilk olarak J. Ross tarafından Sydney Üniversitesi'nde geliştirilmiştir (Quinlan, 1986) (Collin ve Journal, 1996). ID3 makine öğrenmesi ve bilişim teorisine dayanarak, verilen örnekler içinde en ayırıcı değişkeni bulan bir algoritmadır (Mitchell, 1997). ID3 algoritması veri tabanı bölünmeden önce doğru sınıflandırmayı yapmak için gelen bilgiyle, veri tabanı bölündükten sonra doğru sınıflandırma için gelen bilgi arasındaki farkı kullanarak, öncelikli düğüme ve dallanmalara karar verir. Bu aradaki fark kazanım olarak adlandırılmaktadır. Verilerin ham halinin entropisi yani başlangıçtaki entropiyle her bir alt bölümün entropilerinin ağırlıklı toplamı arasındaki fark alınmaktadır. Bu fark hangi alt bölüm için büyükse o bölüme doğru dallandırma yapılmaktadır. Kazanım (K) 4.2 deki formül kullanılarak hesaplanır (Silahtaroglu, 2008).

$$K(D;S)= H(D)- \sum_{i=1}^n P(D_i)* H(D_i) \quad (4.2)$$

C4.5 algoritması ilk olarak 1993 yılında Quinlan tarafından ortaya atılmıştır (Joshi). C4.5 algoritması ID3 algoritmasına bazı konularda üstünlük sağlamaktadır. ID3 algoritmasında kazanım oranı hesaplanırken yalnızca eksik olmayan veriler kullanılmaktadır yani eksik verileri hesaba katmaz. C4.5 algoritması ise eksik verileri diğer veri ve değişkenler yardımıyla öngörerek kazanım oranının hesaplanmasında kullanır (Dunham, 2003). Bu üstünlüğüyle daha duyarlı ve daha anlamlı kurallar çıkartabilen bir ağaç üretilebilir. Diğer bir üstünlük olarak ise ID3 algoritması değişkenleri birçok alt bölüme ayırır ve bu ayırma işlemi aşırı öğrenmeye neden olabilir. Aşırı öğrenmeye engel olmak için Quinlan kazanım yerine kazanım oranını kullanmaktadır. Kazanım oranı 4.3'deki , Ayırma Bilgisi (A) 4.4'deki gibi hesaplanmaktadır.

$$K(D;S)= K(D;S) \div A(D;S) \quad (4.3)$$

$$A(D;S)= \left(\frac{|D_1|}{|D|} , \frac{|D_2|}{|D|} , \frac{|D_3|}{|D|} , \dots , \frac{|D_p|}{|D|} \right) \quad (4.4)$$

ID3 algoritması sadece kategorik nitelikler için kullanılmaktadır. ID3 algoritmasından farklı olarak C4.5 algoritması sayısal nitelikler için de kullanılmaktadır.

CART (Classification and Regression trees) yani sınıflandırma ve regresyon ağaçları yöntemi 1984'te Breiman tarafından ortaya atılmıştır. CART karar ağacı, her bir karar düğümünden itibaren ağacın iki dala ayrılması ilkesine dayanmaktadır. (Breiman, 1998). CART algoritmasında, her bir düğümde belirli bir kriter uygulanarak bölünme işlemi gerçekleştirilmektedir. CART tekniği de ID3 algoritmasında olduğu gibi en iyi dallara ayırma kriterini seçmek için entropiden yararlanır (Dunham, 2003). Fakat ayırma kriterini belirlemek için C4.5 ve ID3 ten farklı bir formül kullanır. CART dallara ayırma kriterini hesaplarken ID3 algoritmasında olduğu gibi kaybolan verileri önemsemez.

Gini algoritması ikili bölünmelere dayalı sınıflandırma yöntemlerinden birisidir. Gini algoritması, nitelik değerlerinin solda ve sağda olmak üzere iki bölüme ayrılması esasına dayanmaktadır. Gini algoritmasının uygulanma adımları aşağıda anlatılmaktadır.

1. Adımda her nitelik değerleri ikili olacak biçimde gruplanır. Bu şekilde elde edilen sol ve sağ değerlere karşılık gelen sınıf değerleri gruplandırılır.
2. Adımda her bir nitelik ile ilgili sol ve sağ taraftaki bölünmeler için $Gini_{sol}$ ve $Gini_{sağ}$ değerleri hesaplanır. Bu hesaplamalar 4.5 ve 4.6'daki gibi yapılmaktadır.

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2 \quad (4.5)$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sağ}|} \right)^2 \quad (4.6)$$

k: Sınıfların sayısı

T: Bir düğümdeki örnekler

$|T_{sol}|$: Sol taraftaki örneklerin sayısı

$|T_{sağ}|$: Sağ taraftaki örneklerin sayısı

L_i : Sol taraftaki i kategorisindeki örneklerin sayısı

R_i : Sağ taraftaki i kategorisindeki örneklerin sayısı

3. Adımda; her j niteliği için, n eğitim kümesindeki satır sayısı olmak üzere 4.7’de yer alan bağıntının değeri hesaplanmaktadır.

$$Gini_j = \frac{1}{n} (|T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ}) \quad (4.7)$$

4. Adımda her j niteliği için hesaplanan $Gini_j$ değerleri arasından en küçük olanı seçilir ve bölünme bu nitelik üzerinden gerçekleştirilir. Bölünme gerçekleştikten sonra ilk adımdan itibaren diğer adımlar bir sonraki bölünecek niteliği belirlemek için gerçekleştirilir (Özkan, 2008).

SLIQ (Supervised Learning in Quest) algoritması 1996 yılında IBM Almaden araştırma merkezinde Mehta M., Agrawal R. Ve Rissanen J. Tarafından önerilmiştir (Manish, 1996). SLIQ algoritması hem sayısal veriler hem de kategorik verilerin sınıflandırılmasında kullanılmaktadır. Sayısal verilerin değerlendirilmesindeki maliyeti azaltmak amacıyla ağacın oluşturulması esnasında önceden-sıralama tekniği kullanılır. Sayısal verilerle işlem yapıldığında verileri sıraya dizmek yapılacak işlemleri kolaylaştırmaktadır. SLIQ algoritmasında kullanılan teknik ise verileri sıralama işlemi her düğümde yapmak yerine, öğrenme verileri sadece bir kere, o da ağacın büyüme aşamasının başlangıcında yapılarak gerçekleştirilir (Silahtaroglu, 2008). ID3 ve C4.5 gibi algoritmalar önce derinlik ilkesiyle çalışırken, SLIQ algoritması önce genişlik düşüncesiyle hareket eder. SLIQ algoritmasında aynı anda birçok yaprak oluşturulur ve bu durumda mevcut ağacın yapraklara ayrılma işlemi veri üzerinden bir kere geçmesiyle tamamlanmış olur. SLIQ bu özelliği haricinde kategorik verileri alt kümelerle ayırmada da hızlı bir algoritma kullanır. SLIQ algoritmasına göre verileri en iyi temsil edecek model, tanımlanma ve oluşturulma maliyeti en küçük olanıdır. SLIQ algoritması veri üzerinden tek seferde geçmesiyle hızlıdır ve buna ek olarak çok iyi sonuçlar veren karar ağaçları da üretebilmektedir.

SPRINT (Scalable Parallelizable Induction of Decision Trees) de SLIQ algoritması gibi önce genişlik ilkesiyle çalışmaktadır. Ancak farklı veri yapılarına kullanarak SLIQ algoritmasından ayrılır (Shafer, 1996). SPRINT ilk olarak her bir değişken için ayrı bir değişken listesi hazırlar. Her tabloda kullanılacak olan değişken için sınıf ve sıra numaraları yer almaktadır. Bu durumda veri tabanındaki değişken sayısı kadar tablo oluşur. Sürekli değer taşıyan tablolar sürekli değer değişkenine göre sıraya dizilirken kategorik veriler taşıyan tablolar sıra numaralarına göre sıralı olarak kalacaktır. Eğitim kümelerinden elde edilen ilk listeler sınıflandırma ağacının kökleriyle

ilişkilendirilir. Ağaçlar büyüyüp düğümler yeni dallara bölündükçe her düğüme ait değişken listeleri de bölünerek yeni dallarla ilişkilendirilir. Bölünme aşamasına gelmiş düğümler için $C_{üst}$ ve C_{alt} adı verilen ve düğümlerdeki sınıf dağılımlarını elde etmek için kullanılan histogramlar belirlenir. Düğümlerden alt dallara ayırma kriteri için de SLIQ algoritmasında olduğu gibi Gini indeksi kullanılır (Silahtaroglu, 2008).

4.1.3 İstatistiksel sınıflandırma modelleri

Veri Madenciliğinde verilerin önceden verilen sınıflara göre ayrılması, gelecekte elde edilecek sonuçların tahmini olarak düşünülebilir. Yani gelecekteki sınıfların tahmin edilmesidir. Regresyon, lojistik regresyon, zaman serileri analizi ve Bayesyen yaklaşım istatistiksel sınıflandırma tekniklerine örnek olarak verilebilir.

Regresyon analizi

Bağımlı değişken ile bir veya daha çok bağımsız değişken arasındaki ilişkiyi incelemek amacıyla kullanılan bir analiz yöntemi regresyon analizi olarak adlandırılmaktadır. Regresyon analizi sonucu elde edilen denkleme ise regresyon denklemi denilmektedir. Bir bağımlı değişken bir tek bağımsız değişkenle açıklanabiliyorsa basit regresyon analizi, birden fazla değişkenle bağımsız değişkenle açıklanabiliyorsa çoklu regresyon analizi olarak adlandırılmaktadır. Oluşturulan denklemin türüne göre de regresyon analizi doğrusal regresyon ve doğrusal olmayan regresyon olmak üzere ikiye ayrılmaktadır. En küçük kareler yöntemiyle elde edilen doğrusal regresyon denklemi 4.8’de gösterilmektedir.

$$y = a + bx + e \quad (4.8)$$

a doğrusal fonksiyonun sabiti, b ise doğrusal fonksiyonun eğimidir. Fonksiyonun eğimi regresyon katsayısı olarak da adlandırılmaktadır. x teki bir birimlik değişimin, y üzerinde yine y cinsinden yaratacağı değişikliği gösteren katsayıdır. y bağımlı değişkeni yani tahmin edilecek değişkeni temsil etmektedir. y değişkeni sürekli değişkendir. Veri madenciliği açısından bakıldığında ise y sınıfları temsil etmektedir. x ise bağımsız değişkeni yani nitelikleri temsil etmektedir. Verilen x nitelik değerine göre y yani hangi sınıfa ait olacağı tahmin edilmektedir. Birden fazla nitelik var ise yani birden fazla x değeri mevcut ise; çoklu regresyon olarak adlandırılır. Çoklu regresyon denklemi ise 4.9’da gösterilmektedir.

$$y = a + b_1x_1 + b_2x_2 + ... + e \quad (4.9)$$

Bu regresyon analizine bir örnek verelim. Değişkenler cinsiyet, kilo, boy ve beden olsun. Cinsiyet boy ve kilo bağımsız değişkenler yani x leri ifade etmektedir. Beden ise x e bağlı değişken yani y yi ifade eder. Regresyon analizinde ilk adım diğer sınıflandırma modellerinde olduğu gibi öğrenme adımıdır. Bu adımda regresyon analiz denklemi elde edilir. Daha sonraki adımda ise hangi bedene ait olduğu bilinmeyen; boy, kilo ve cinsiyeti verilmiş bir kimsenin hangi sınıfa yani hangi bedene sahip olduğu tahmin edilir.

Elde edilen regresyon denkleminin derecesi değişebilir. Örneğin ikinci dereceden bir regresyon denklemi 4.10'daki gibidir.

$$y = a + bx + cx^2 + e \quad (4.10)$$

3. dereceden bir regresyon denklemi 4.11'de gösterilmektedir.

$$y = a + bx + cx^2 + dx^3 + e \quad (4.11)$$

Lojistik regresyon analizi

Bağımlı değişkenin yani y'nin nitel veri olduğu durumlarda lojistik regresyon yöntemi kullanılmaktadır. Lojistik regresyon, bağımlı değişken üzerinde hangi bağımsız değişken/ değişkenlerin önemli risk faktörü olduğunu ve bu bağımsız değişken/değişkenlerin bağımlı değişkenin değerlerinin tahmininde ne düzeyde etkide bulunduklarını belirlemekte kullanılmaktadır. Özet olarak bağımsız değişkenlerin etkilerine dayanarak verilerin sınıflandırılmasında kullanılan bir yöntemdir. Bu yöntem olasılık kurallarına uygun olarak verilerin belirli sınıflara atanmasını sağlar (Kurt ve diğ., 2005). 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 tanımlayabilen ve biyolojik olarak kabul edilebilir bir model kurmaktır (Bircan, 2004). Örneğin bir kişinin hasta ya da sağlam olması gibi durumlarda kullanılmaktadır. Elde edilen y değeri 0 ile 1 aralığında yer almakta olup, her bir gözlem için hasta olma olasılığını ifade etmekte kullanılır. Bir başka deyişle lojistik regresyonda bağımlı değişkenin alabileceği iki değerden birinin gerçekleşme olasılığı elde edilir.

İleri parametrik olmayan bir istatistiksel yöntem olan lojistik regresyon analizi, bağımlı değişken mutlaka ikili sonucu olan değişken olduğu durumlarda kullanılır. Ayrıca zorunlu olmamakla beraber bağımsız değişkenler genellikle sürekli olurlar.

Bu avantajından dolayı gözlemlerin gruplara ayrılmasında ve yeni gözlemlerin bu uygun gruplara atanmasında sıkça tercih edilen bir yöntem olmaktadır (Atan ve diğ., 2004).

Tek bağımsız değişken olduğunda kullanılan ikili lojistik regresyon modeli 4.12’de gösterilmektedir.

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (4.12)$$

Birden fazla bağımsız değişken olduğunda kullanılan ikili lojistik regresyon modeli ise 4.13’deki gibidir.

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

(4.13)

Lojistik regresyon analizinin uygulamadaki adımları aşağıdaki gibidir (Ünsal, 2005).

1. adımda önsel grup üyelikleri belirlenir.
2. adımda modele girecek değişkenler belirlenir. Bu amaçla önsel bilgidен ya da istatistiksel tekniklerden yararlanılabilir.
3. adımda modelin parametreleri Newton-Raphson yöntemi ile tahmin edilir. Ardından modelin tümünün anlamlılığı olabilirlik oranı ile test edilir. Model anlamlı değilse analize son verilir. Eğer model anlamlı bulunursa diğer aşamaya geçilir.
4. adımda tahmin edilen model parametrelerinin tek tek anlamlılığı incelenir. Bu amaçla olabilirlik oranı ya da Wald istatistiği kullanılabilir. Her katsayının anlamlılığı incelendikten sonra, teklik oranları incelenerek, açıklayıcı değişkenlerin bağımlı değişken üzerindeki etkileri yorumlanabilir.
5. adımda tahmin edilen model parametreleri kullanılarak, her bir gözlemin hangi gruptan geldiği tahmin edilir.

6. adımda modelin uyum iyiliğini incelemek amacıyla doğru sınıflandırma yüzdesi ve yapay R2 ölçütleri kullanılır. Modelin uyum iyiliği kabul edilebilir düzeyde ise beşinci adımdaki grup tahminleri kullanılabilir. Aksi halde ikinci aşamaya geçilerek modele girecek değişkenler yeniden gözden geçirilir ve işlemler tekrar edilir.

Zaman Serileri Analizi

Belirli zamanlarda bir cevap değişkeninin gözlemlenmesi yoluyla elde edilen veriler zaman serisi olarak adlandırılmaktadır.

Veriler eşit aralık zaman noktalarından elde edilmektedir. Zaman serileri analizi ise özet olarak bir zaman serisinin olasılıksal yapısının keşfedilmesi ve gelecekteki durumunun öngörülmesi olarak tanımlanmaktadır. Zaman serileri, gelecekteki operasyonların uzun dönem planlamasında ve tahmin edilmesinde kullanılmak amacıyla analiz edilmektedir. Genel olarak zaman serisi analizi geçmişe ait kayıtlardan, geleceğe dair öngörüler yapmaktır. Zaman serilerinin bileşiminde 4 unsur yer almaktadır bunlar; trend bileşeni, mevsim bileşeni, konjonktürel bileşen ve rastsal bileşendir.

Trend Bileşeni: Zaman serisi gözlem değerinin uzun zaman döneminde (en az 7 yıl) artma ya da azalma yönünde gösterdiği genel eğilime “trend” adı verilir. Bu eğilimi açıklayan bileşene de “Trend bileşeni” denir. Trend bileşeni, zamana bağlı değişken üzerindeki genel eğilime neden olan uzun dönemli etkileri açıklar.

Mevsimsel Bileşen: Mevsimsel bileşen birbirini izleyen dönemlerin, yılların, mevsimlerin, ayların ya da günlerin aynı zaman noktalarında zaman serisi gözlem değerlerindeki bir artma ve bir azalma şeklindeki düzenli değişimleri, var olan mevsimsel değişimleri açıklar. Mevsimsel değişimler, genellikle iklimle, vakitle ya da alışkanlıklarla ilişkilidir.

Konjonktürel Bileşen: Ekonomi ve iş idaresi alanlarıyla ilgili değişkenlerde sabit bir oranda artış ya da azalış gözlemlenmez. Trend düzeyi etrafında, iki ile on yıl ya da daha fazla yıl zaman aralıklarıyla, herhangi bir dönemde, artma ya da azalma şeklinde tekrarlanabilen değişimler gözlenir. Konjonktürel değişme adı verilen bu değişimlerin etkisini açıklayan bileşen “konjonktürel bileşen” olarak adlandırılmaktadır.

Rastal Bileşen: Zaman serilerindeki düzensiz ve ani değişmelere “rastal değişme” adı verilir. Rastal değişmeler, beklenmedik olayların zaman serileri üzerindeki etkisiyle meydana gelen değişmeler olarak da tanımlanmaktadır.

Bayesyen Sınıflandırma

Bayesyen sınıflandırma tekniği, mevcut bulunan, sınıfları belirlenmiş verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine girme olasılığını hesaplayan bir yöntemdir. Olasılıkları hesaplamada Bayesyen kuralı kullanıldığı için bu isimle adlandırılmaktadır. Eldeki verilerin belirlenmiş olan sınıflara ait olma olasılıklarını öngörür. İstatistikteki Bayes teoremine dayanır.

Bu teorem; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar. Belirsizlik taşıyan durumlarda karar verme konusunda çok kullanışlıdır. En önemli zafiyeti değişkenler arası ilişkinin modellenmiyor olması ve değişkenlerin birbirinden tamamen bağımsız olduğu varsayımıdır. Bayes yöntemi koşullu olasılık durumları ile ilgilidir. Her hangi bir koşullu olasılık durumu $P(X=x | Y=y) = R$ şeklinde tanımlanır. Bu ifade; “Eğer $Y = y$ doğru ise, $X = x$ olma olasılığı R ’dir.” anlamına gelmektedir. X ve Y ’nin alabileceği değerlerin her kombinasyonu için koşullu olasılıkları belirleyen tabloya koşullu olasılık dağılımı adı verilir ve $P(X|Y)$ ile ifade edilir (Argüden ve diğ., 2008).

4.1.4 Mesafeye dayalı sınıflandırma modelleri

Sınıflandırma yapılırken eldeki verilerin birbirlerine olan uzaklığı veya benzerlikleri kullanılarak da sınıflandırma yapılabilir. Veriler arasındaki mesafe ölçümünde en yaygın olarak Öklid uzaklık kullanılmaktadır. En yaygın kullanılan yöntem K-en yakın komşu (K-nearest neighbour) algoritmasıdır.

K-en yakın komşu yönteminde, örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine uzaklıklarının hesaplanması ve en küçük uzaklığa sahip k sayıda gözlemin seçilmesi esasına dayanmaktadır. Uzaklıkların hesaplanmasında i ve j noktaları için 4.14’deki Öklid uzaklık formülü kullanılmaktadır (Özkan, 2008).

$$d(i, j) = \sqrt{\sum_{k=1}^p (x_{ij} - x_{jk})^2} \quad (4.14)$$

Yöntemde k değeri önceden seçilmektedir. K-en yakın komşu yönteminde aşağıdaki adınlar sırasıyla uygulanmaktadır.

1. adımda k parametresi belirlenir. Tipik k değeri 3, 5 ve 7 dir (Khan, 2002). Değerinin yüksek olması birbirlerine benzemeyen noktaların bir araya toplanmasına, çok küçük seçilmesi ise birbirine benzeyen yani aynı sınıfa dahil olan noktaların başka sınıflarda yer almasına neden olur.
2. adımda verilen bir noktaya en yakın komşular belirleneceği için, söz konusu nokta ile diğer tüm noktalar arasındaki uzaklıklar tek tek hesaplanmaktadır. Hesaplama yaygın olarak Öklid mesafesi kullanılmaktadır. Öklid mesafesini bulmak için kullanılan formül yukarıda anlatılmıştır.
3. adımda; bir önceki adımda hesaplanan uzaklıklara göre satırlar sıralanır ve bunlar arasında en küçük olan k tanesi seçilir.
4. adımda 2 farklı yol izlenebilmektedir. 1. Yol olarak seçilen satırların hangi kategoriye ait oldukları belirlenir ve en çok tekrarlanan kategori değeri seçilir. 2 yolda ise ağırlıklı oylama yöntemine başvurulmaktadır. Bu yolda en çok tekrarlanan sınıf seçme yöntemi yerine ağırlıklı uzaklıkların hesaplanması esasına dayanmaktadır. Ağırlıklı uzaklıkların hesaplanması 4.15’deki formüle göre yapılmaktadır.

$$d(i, j)' = \frac{1}{d(i, j)^2}$$

(4.15)

Formülde yer alan $d(i,j)$ ifadesi i ve j gözlemleri arasındaki Öklid uzaklığıdır. Her bir sınıf değeri için bu uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri elde edilmektedir. En büyük ağırlıklı ortalama değerine sahip olan sınıf değeri yeni gözlemin ait olduğu sınıf olarak kabul edilmektedir.

5. adımda seçilen kategori, tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir (Özkan, 2008)

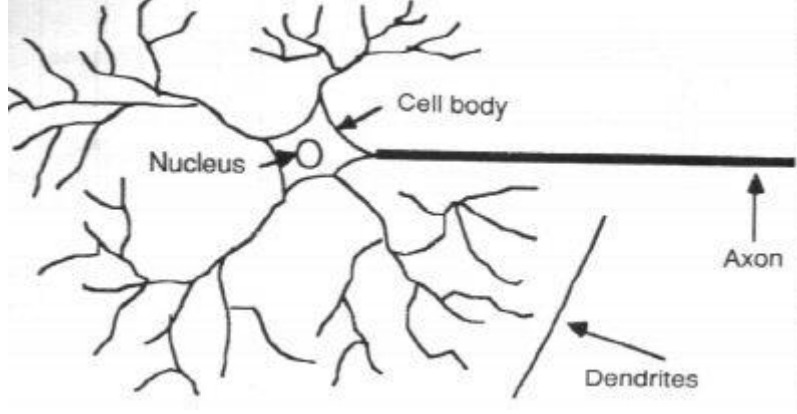
4.1.5 Yapay sinir ağıları

Yapay sinir ağıları (YSA) biyolojik sinir ağlarından esinlenerek geliştirilmiş bir bilgi işleme sistemidir. Yapay sinir ağlarının geçmişi 1942 yılına kadar gitmektedir. 1942 yılında McCulloch ve Pitts ilk hücre modelini geliştirmiştir o nedenle yapay sinir ağlarının başlangıcı kabul edilmektedir. 1949 yılında Hebb tarafından hücre bağlantılarını ayarlamak için ilk öğrenme kuralı önerilmiştir. 1958 yılında ise Rosenblatt, algılayıcı modeli ve öğrenme kuralını geliştirerek, bugün kullanılan kuralların temelini ortaya koymuştur. 1969 yılında Minsky ve Papert algılayıcının kesin analizini yaptı ve algılayıcının karmaşık lojik fonksiyonlar için kullanılamayacağını ispatladılar. 1982-1984 yılında Kohonen kendi kendini düzenleyen haritayı (self-organizing map) tanımladı. Kendi adıyla anılan denetimsiz öğrenen bir ağ geliştirdi. 1986 yılında Rumelhart geriye yayılımı tekrar ortaya çıkarttı. 1988 yılında Chua ve Yang hücresel sinir ağlarını geliştirdiler. Yapay sinir ağlarının gelişimini gösteren tablo Çizelge 4.1’de yer almaktadır (Url-1).

Çizelge 4.1 : Bilinen YSA mimarilerinin tarihsel gelişimi.

<i>Yıl</i>	<i>Ağ Mimarisi</i>	<i>İcat Eden Bilim Adamı</i>
1942	Mc Culloh-Pitts Hücresi	Mc Culloh-Pitts
1957	Algılayıcı (Perceptron)	Roserblatt
1960	Madaline	Widrow
1969	Cerebellatron	Albus
1974	Geriye Yayınım (Backpropagation)	Werbos, Parker, Rumelhart
1977	Bir Kutu İçinde Zeka	Anderson
1978	Neocognitron	Fukushima
1978	Adaptif Rezonans Teorisi	Carpenter, Grossberg
1980	Ön Düzenlemeli Harita	Kohonen
1982	Hopfield	Hopfield
1985	İki Yönlü Bileşik Hafıza	Kosko
1985	Boltzman Makinesi	Hinton, Sejnowsky, Szu
1986	Sayıcı Yayılım (Counterpropagation)	Hecht-Nielsen
1988	Hücrel Sinir Ağı	Chua, Yang

Yapay sinir ağları, yapay sinir hücrelerinin birbirleriyle çeşitli şekillerde birleşmesinden oluşmakta ve katmanlar şeklinde düzenlenmektedir. YSA' ların en belirgin özellikleri birbirlerine bağlı nöronlar, bağlantılar arasındaki aralıkların belirlenmesi ve ateşleme fonksiyonudur. Şekil 4.3'de biyolojik sinir ağının yapısı görülmektedir.



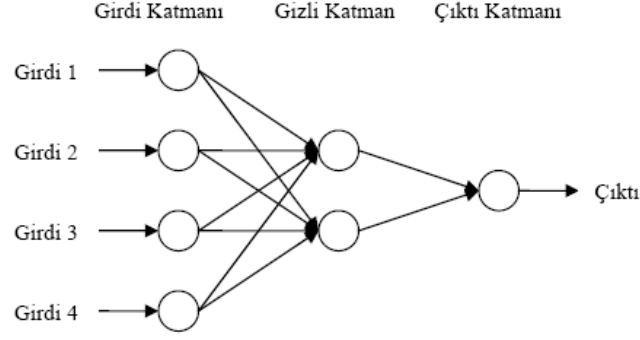
Şekil 4.3 : Biyolojik sinir ağının yapısı.

Yapay sinir ağını oluşturan nöronun bir iç hali mevcuttur. Bu iç hale aktivasyon veya aktivasyon seviyesi denilmektedir. Bu seviye gelen girdileri tanımlayan bir fonksiyondur. Ağ içerisindeki bir nöron aynı doğal nöronların yaptığı gibi diğer nöronlara bir sinyal göndermekte; gönderdikleri bu sinyaller ise gönderilen nöronlar için giriş fonksiyonları oluşturmaktadır. Bir nöron birden fazla nörona sinyal gönderebilmektedir (Fausett, 1994). Şekil 4.4’de yapay sinir ağının yapısı görülmektedir.



Şekil 4.4 : Yapay sinir ağı.

Sinir ağları iki ya da üç katmandan oluşmaktadırlar. Bu katmanlar girdi, gizli ve çıktı katmanları olarak adlandırılmaktadır. İki katmanlı sinir ağlarından gizli katman yer almamaktadır. Şekil 4.5’de üç katmanlı sinir ağı örneği gösterilmektedir.



Şekil 4.5 : 3 katlı sinir ağı örneği.

Şekilde de görüldüğü üzere ilk katman girdi katmanıdır. Giren her bir verinin 0 ile 1 arasında olması zorunludur. Gizli katman daha önce de belirtildiği üzere tüm sinir ağlarında bulunmak zorunda değildir. Gizli katman, daha fazla örüntünün tanınmasını mümkün kılmasından dolayı ağı daha güçlü kılmaktadır. Genellikle tek bir gizli katman yeterli olmaktadır. Son katman çıktı katmanıdır. Birden fazla çıktı katmanı olabilir fakat çoğunlukla sinir ağları tek bir değer hesaplamaktadır. Bu değer 0 ile 1 arasında yer almaktadır. Eğer gizli katmanı var ise tüm girdiler gizli katmanlara bağlanmaktadır, yok ise çıktı katmanına bağlanmaktadır. Gizli katmanından ise girdilerin tamamı çıktı katmanına bağlanmaktadır. Gizli katmanın aktif hale gelebilmesi için fonksiyonun değerinin belirli bir eşiğin üzerinde olması gerekmektedir.

YSA, sınıflandırma, öngörü ve kümeleme modellerinde doğrudan uygulanabilen çok güçlü bir yöntemdir. Mali serilerin tahmininden sağlık durumlarının teşhisine, değerli müşterilerin belirlenmesinden kredi kartı sahtekarlıklarının tespitine, el yazısı formlarının değerlendirilmesinden makine arızalanma oranının öngörülmesine kadar birçok alanda uygulanmaktadır (Berson ve Smith, 1997).

Yapay sinir ağları tıpkı insanların yaptığı deneyimlerden bilgi çıkarma işlemini yapmaktadır. Yapay sinir ağları da kendilerine verilen örneklerden bir takım bilgiler çıkarmaktadır. Öncelikle bir veri kümesi üzerinde öğrenme algoritmaları çalıştırılarak eğitilir.

Bu eğitim sonucunda yapay sinir ağının içerisinde bir takım ağırlıklar belirlenir. Bu ağırlıklar kullanılarak yeni gelen veriler üzerine işlenir ve aşamalar sonucunda bir sonuç elde edilir. Yapay sinir ağlarının en olumsuz tarafı ise bu ağırlıkların neden ilgili değerleri aldıklarının bilinmemesidir. Çıkan sonucun nedenleri açıklanamamaktadır. Açıklayamaması kullanım alanını daraltmaktadır.

Yapay sinir ağlarını kullanmak için en iyi yaklaşım, onları içi bilinmeyen bir şekilde çalışan kara kutular olarak düşünmek olacaktır (Göral, 2007). Yapay sinir ağlarının veri madenciliğinde kullanılması çok fazla miktarda avantaj sağlamaktadır. Yapay sinir ağları kullanımı sayesinde çok geniş yelpazedeki problemlerin çözümü sağlanabilmektedir. Öğrenme işlevi sayesinde çok karmaşık durumlarda dahi çok iyi sonuçlar üretmektedir. Hem sayısal hem de kategorik veriler üzerinde de işlem yapabilmektedir. Bütün bu avantajlarına rağmen dezavantajları da mevcuttur. Girdi verilerinin 1 ile 0 arasında olması gereklidir ve bunu sağlamak için verilerde dönüşüm işlemi yapılması gerekmektedir. Bu durum zaman kaybına yol açmaktadır. Yapay sinir ağları ürettikleri sonuçların nedenleri hakkında açıklama yapmazlar. Bu açıdan bakıldığında kapalı kutulara benzetilebilirler. Örneğin müşterinin riskli müşteri grubunda olduğu sonucu elde edilmiş olsun. Müşterinin neden riskli olduğunu açıklamaz ve bu durumda bankacıları zor durumda bırakabilir. Elde edilen sonuçların en iyisi olduğuna dair bir garanti yoktur. Yapay sinir ağlarını kullanmak belirli bir uzmanlık gerektirmektedir o nedenle kullanımı zordur.

4.1.5.1 Yapay sinir ağlarının temel özellikleri

Yapay Sinir Ağlarının hesaplama ve bilgi işleme gücünü, paralel dağılmış yapısından, öğrenebilme ve genelleme yeteneğinden aldığı söylenebilir (Ergezer ve diğ., 2003). Genelleme, eğitim ya da öğrenme sürecinde karşılaşılmayan girişler için de YSA' nın uygun tepkileri üretmesi olarak tanımlanır. Bu üstün özellikleri, YSA' nın karmaşık problemleri çözebilme yeteneğini gösterir. Günümüzde birçok bilim alanında YSA, aşağıdaki özellikleri nedeniyle etkin olmuş ve uygulama yeri bulmuştur.

Doğrusal Olmama; YSA' nın temel işlem elemanı olan hücre, doğrusal değildir. Dolayısıyla hücrelerin birleşmesinden meydana gelen YSA da doğrusal değildir ve bu özellik bütün ağa yayılmış durumdadır. Bu özelliği ile YSA, doğrusal olmayan karmaşık problemlerin çözümünde en önemli araç olmuştur.

Öğrenme; YSA'nın arzu edilen davranışı gösterebilmesi için amaca uygun olarak ayarlanması gerekir. Bu, hücreler arasında doğru bağlantıların yapılması ve bağlantıların uygun ağırlıklara sahip olması gerektiğini ifade eder.

YSA'nın karmaşık yapısı nedeniyle bağlantılar ve ağırlıklar önceden ayarlı olarak verilemez ya da tasarlanamaz. Bu nedenle YSA, istenen davranışı gösterecek şekilde ilgilendiği problemten aldığı eğitim örneklerini kullanarak problemi öğrenmelidir.

Genelleme; YSA, ilgilendiği problemi öğrendikten sonra eğitim sırasında karşılaşmadığı test örnekleri için de arzu edilen tepkiyi üretebilir. Örneğin, karakter tanıma amacıyla eğitilmiş bir YSA, bozuk karakter girişlerinde de doğru karakterleri verebilir ya da bir sistemin eğitilmiş YSA modeli, eğitim sürecinde verilmeyen giriş sinyalleri için de sistemle aynı davranışı gösterebilir.

Uyarlanabilirlik; YSA, ilgilendiği problemdeki değişikliklere göre ağırlıklarını ayarlar. Yani, belirli bir problemi çözmek amacıyla eğitilen YSA, problemdeki değişimlere göre tekrar eğitilebilir ve değişimler devamlı ise gerçek zamanda da eğitime devam edilebilir. Bu özelliği ile YSA, uyarlamalı örnek tanıma, sinyal işleme, sistem tanılama ve denetim gibi alanlarda etkin olarak kullanılır.

Hata Toleransı; YSA, çok sayıda hücrenin çeşitli şekillerde bağlanmasından oluştuğu için paralel dağılmış bir yapıya sahiptir ve ağına sahip olduğu bilgi, ağıdaki bütün bağlantılar üzerine dağılmış durumdadır. Bu nedenle, eğitilmiş bir YSA'nın bazı bağlantılarının hatta bazı hücrelerinin etkisiz hale gelmesi, ağına doğru bilgi üretmesini önemli ölçüde etkilemez. Bu nedenle, geleneksel yöntemlere göre hatayı tolere etme yetenekleri son derece yüksektir.

4.1.5.2 Öğrenme şekillerine göre yapay sinir ağları

Sinir ağlarının en önemli özelliği, öğrenme yeteneğidir. Bir sinir ağında öğrenmenin anlamı, ağına belirli bir probleme ait doğru çıktıları üretmesini sağlayacak optimum ağırlık değerlerinin bulunmasıdır.

Öğrenme, ağırlık değerlerinin nasıl değiştirilmesi gerektiğini ifade eden bir öğrenme kuralına dayanır. Bir öğrenme kuralının temel ilkesi ise, benimsenen öğrenme stratejisi ile tanımlanır. Literatürde, üç tip öğrenme stratejisinden söz edilmektedir.

Denetimli öğrenme metoduyla, yapay sinir ağının eğitimi için eğitici veriler (eğitim kümesi) kullanılmaktadır. Eğitim kümesi, giriş bilgileri ve istenen (hedef) bilgiler olmak üzere iki ayrı vektör gibi düşünülebilir. Vektörlerin her bir karşılıklı elemanları bir eğitim çiftini oluşturmaktadır. Eğitim kümesi, ağına eğitime başlamadan önce belirlenmektedir.

Ağın eğitimi için, öncelikle bağlantı ağırlıklarına rastgele değerler atanmaktadır. Daha sonra eğitim çiftlerine bağlı olarak bir algoritma dahilinde ağırlıklar yenilenmektedir. İstenilen bilgiler ve ağın çıkışı arasındaki fark (hata) azalınca kadar eğitim sürdürülmektedir.

Ağ çıkışındaki hatanın azalması, ağırlıkların kararlılık kazanması demektir. Ağırlıklar istenilen kararlılığa ulaştığında eğitim bitirilmektedir.

Denetimsiz öğrenme metoduna, “kendi kendine öğrenilebilen metot” da denilmektedir. Bu öğrenme metodunda eğitim kümesi kullanılmamaktadır. Ağ, birbirine benzer giriş bilgilerini gruplamakta veya giriş bilgisinin hangi gruba ait olduğunu göstermektedir. Ağ eğitimi için sadece giriş bilgileri yeterli olmakta, referans alınacak (eğitici) bilgiye ihtiyaç duyulmamaktadır. Ağın performansını kendiliğinden izlenmesi söz konusudur. Ağ, giriş sinyallerinin yönüne veya düzenine bakmakta ve ağın fonksiyonuna göre ayarlama yapmaktadır. Bu nedenle, denetimsiz öğrenme stratejisini kullanan ağlar, kendi kendine organize olan ağlar olarak adlandırılır.

Takviyeli Öğrenme kuralı denetimli öğrenmeye yakın bir metottur. Denetimsiz öğrenme algoritması istenilen çıkışın bilinmesine gerek duymaz. Hedef çıktıyı vermek için "öğretmen" yerine, burada YSA' ya bir çıkış verilmemekte fakat elde edilen çıkışın verilen girişe karşılık iyiliğini değerlendiren bir kriter kullanılmaktadır.

4.1.6 Genetik algoritmalar

Genetik algoritmalar evrimsel hesaplama (evolutionary computing) metotlarının örnekleridir ve optimizasyon tipli algoritmalar (Dunham,2003). Genetik algoritmalarının kaynağı yapay sinir ağlarında olduğu gibi biyolojidir. Genetik algoritmalar doğadaki bilinen evrim yasalarından ya da öne sürümlerinden etkilenecek geliştirilmiş algoritmalar.

Algoritma nüfus olarak adlandırılan ve kromozomlar tarafından temsil edilen bir dizi sonuçla işlemlere başlamaktadır. Eldeki sonuçlar kullanılarak yeni bir sonuç elde edilmektedir. Elde edilen her yeni sonucun bir öncekinden daha iyi olması beklenmektedir. Aynı şekilde istenilen durma kriterine ulaşıncaya kadar yeni nesil üretimini devam edilir.

Temel bir genetik öğrenme algoritması şu şekilde ifade edilebilir (Roeger ve diğ., 2003):

1. n tane elemandan oluşan bir popülasyon, P, seçilir. Buradaki elemanlar genelde kromozom olarak anılır.
2. Belirlenmiş bir bitim koşulu sağlanana kadar
 - a. Geçerli çözümün her elemanını değerlendirmek için bir uygunluk fonksiyonu kullanılır. Eğer bir eleman uygunluk kriterini geçerse, popülasyonun içinde kalır.
 - b. Popülasyon şu an m adet elemandan oluşmaktadır ($m \leq n$). (n-m) adet yeni eleman üretmek için genetik operatörleri kullanır. Yeni elemanları popülasyona ekler.

Veri madenciliği açısından bakıldığında kromozom, veri tabanındaki her bir kaydı ifade etmekte kullanılmaktadır. Bu kromozomlar üretilecek yeni sonuçlar hakkında bilgiler içermektedir. Bu bilgilerin kullanılabilmesi için kromozomların kullanılabilir formlar haline dönüştürülmesi gerekmektedir, bu işlem kromozomların çözümlenmesi olarak adlandırılmaktadır. Kromozomların çözümlenmesi ikili, permütasyon, değer çözümü ve ağaç çözümü olmak üzere dört şekilde yapılmaktadır. İkili çözümleme yönteminde tüm kromozomlar ikili sayı sisteminde gösterilirler. Örneğin Kromozom 1: 00010011100. Permütasyon yönteminde ise her bir kromozom bir sayıyla temsil edilmektedir. Örneğin Kromozom 1: 1539847. Değer çözümleme yönteminde ise her bir kromozom değerler dizisidir. Bu değerler her türlü sayı ve nesne olabilir. Örneğin Kromozom 1: AKOLJGPHBUTRNUTR. Ağaç çözümleme yöntemi ise program ya da ifade geliştirmekte kullanılmaktadır.

Kromozomlar istenilen yöntemle çözümlendikten sonra sıra üretme işlemine gelmiştir. Üretme iki veya daha fazla kromozomdan yararlanılarak yeni nesil elde etme işlemidir. Tek nokta üretiminde anne ve baba kromozomların belirli bir kısmı alınarak üretilir. İki noktalı üretimde ise hem anne hem de baba kromozomdan iki ayrı parça alınarak üretim gerçekleştirilir. Bir başka yöntemde ise özellikler anne ve babadan rastgele seçilerek kopyalanır. Bu üç yöntemin dışında önceden belirlenmiş bir takım aritmetik işlemler yapılarak da yeni kromozom üretme işlemi yapılabilir.

Üretme işleminden sonra mutasyon işlemine gelmektedir. Mutasyon işleminin amacı nüfus içindeki tüm sonuçların bir kısır döngüye girip aynı türler üzerinde gezinmesini engellemektir.

Mutasyon deęiřim olarak da açıklanabilir. Mutasyon iřleminde üretim iřleminden elde edilen yeni kromozom üzerinde rastgele iřlemler yapmaktır.

Üretme ve mutasyona ek olarak nüfus büyüklüęü de genetik algoritmalarında etkili bir parametredir. Tüm nüfusta kaç adet kromozom bulunmakta, her bir üretilen yeni nesilde kaç adet kromozom bulunacak vs. sayılar genetik algoritmanın çalışma hızını da arttırıp azaltacaktır.

Bu anlamda çok büyük miktarda nesiller üretmek hız açısından çok avantajlı olmayacak zaman kaybına neden olacak, az miktardaki kromozom ise yeteri miktarda hassas sonuçlar vermeyecektir.

Veri madenciliğinde genetik algoritmalar kümeleme ve öngörme problemleri hatta birliktelik kuralları için kullanılabilirler. Bu teknikler, veriyi tanımlamak için kurulan modellere ait kümeden en uygun modelleri bulmak olarak düşünülebilir. Bu yaklaşımda öncelikle bir başlangıç modeli kabul edilir ve bir çok iterasyondan sonra modeller yeni modeller oluşturmak üzere birleştirilir. Bu modeller arasından bir uygunluk fonksiyonu (fitness function) tarafından tanımlanan en iyi model, bir sonraki iterasyon için girdi olarak kullanılır. Algoritmalar modelin nasıl tanımlandığına, modeldeki farklı bireylerin / kromozomların nasıl birleştirildiğine ve uygunluk fonksiyonunun nasıl kullanıldığına göre farklılık göstermektedir (Dunham, 2003).

Genetik algoritmalar açıklanabilir sonuçlar üretirler. Deęişik tiplerdeki verileri iřleme özelliğine sahiptirler. Ayrıca genetik algoritmalar yapay sinir aęları ile çalışarak başarılı sonuçlar üretmektedirler. Genetik algoritmalar bu kadar avantaja sahipken; dezavantajlara da sahiptir. Genetik algoritmaları anlayıp son kullanıcıya anlatmanın çeşitli zorlukları mevcuttur. Bazı sorunlar çok karmařık olduğundan genetik kodlamasını yapmak genetik algoritma kullanımını zorlařtırmaktadır. En iyi uygunluk fonksiyonunu belirlemek zordur çünkü sürekli yeni fonksiyonlar elde edilmektedir. Çaprazlama ve mutasyon iřlemleri için çok fazla olasılık olduğundan nasıl yapılacağını belirlemek zordur. Genetik algoritmalarda elde edilen sonucun optimal olduğuna dair bir kanıt bulunmamaktadır.

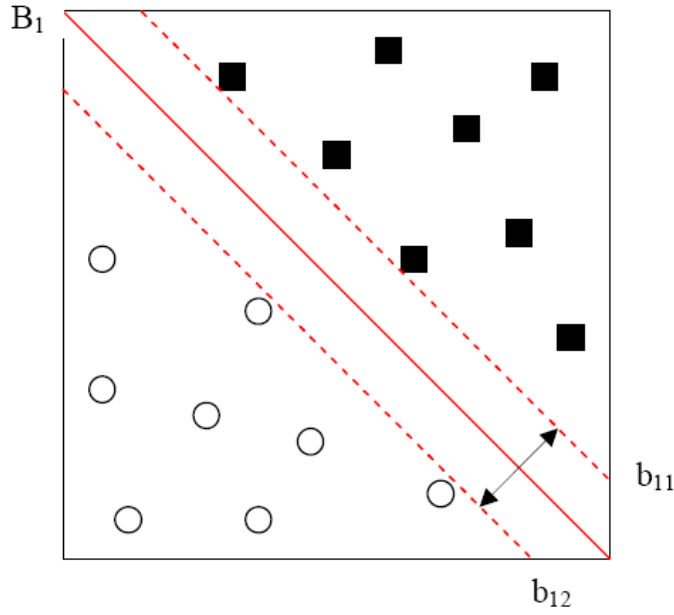
4.1.7 Destek vektör makineleri

İstatistikteki, genelleme kuramındaki hesapsal öğrenme ve makina öğrenme kuramlarındaki yeni ilerlemeler, model yapım/öğrenme/sağlamlaştırma sürecinin genel karakterine ve doğasına yeni izleme metotları ve derin özellikler sağlamıştır. Bazı araştırmacılar, istatistiksel ve makine öğrenme modellerin kavramsal olarak tamamıyla farklı olmadıklarına dikkat çekmektedir. Yeni hesapsal ve makine öğrenme yöntemlerinin birçoğu, istatistikte parametre tahmini fikrini genellemektedir. Geçtiğimiz bir kaç yılda bu yeni yöntemlerden en çok ilgi çeken Destek Vektör Makineleri (Support Vector Machines, SVM) olmuştur (Hsieh, 2004).

1960'ların sonunda Vapnik tarafından önerilen destek vektör makineleri; sınıflandırma ve doğrusal olmayan fonksiyon yaklaşımı problemlerinin çözümü için önerilen bir öğrenme algoritmasıdır. Son yıllarda daha yaygın olarak kullanılmaya başlanan destek vektör makineleri; yazı tanıma, nesne tanıma, ses tanıma, yüz tanıma gibi örüntü tanıma uygulamalarında kullanılmıştır (Burges, 1998). Doğrusal olmayan problemleri çözmedeki başarısı diğer yöntemlere göre çok daha fazladır.

Günümüzde başarısı sayesinde kullanım alanları yaygınlaşmıştır. Temelde lineer olarak ayrıştırılabilir iki sınıfın karar yüzeyinin destek vektörler olarak tanımlanan ve sınıf sınırlarını belirleyen örnekler arasında maksimum marjının oluşturulması ilkesine dayanmaktadır. Marjın maksimizasyonu işlemi bir kuadratik sınırlamalı optimizasyon problemi şeklinde yazılarak, Lagrangian fonksiyonu şeklinde ifade edilerek dual forma dönüştürülür. Doğrusal problemler için gerçekleştirilen bu yaklaşım doğrusal olmayan ayrıştırma problemleri için kernel dönüşümleri kullanılarak genelleştirilebilir (Polat ve diğ., 2007). Destek vektör makineleri tekniği, sınıfları birbirinden ayıran marjını en büyük, doğrusal bir ayırt edici fonksiyon bulunmasını amaçlamaktadır. Doğrusal olarak ayrılamayan örnekler için, örnekler doğrusal olarak ayrılabilirdikleri daha yüksek boyutlu başka bir uzaya taşınır ve sınıflandırma o uzayda yapılmaktadır (Amasyalı ve diğ., 2006).

Destek vektör makinelerinde karşılaşılabilecek iki durum bulunmaktadır. Birinci durumda veriler doğrusal olarak ayrılabilir. İkinci durumda ise veriler doğrusal olarak ayrılamayan bir yapıya sahiptir. Birinci durumun çözümü destek vektör makineleri tekniğinin en basit modelini oluşturmaktadır. Bu veriler arasında maksimum sınırın bulunması çok kolaylıkla sağlanmaktadır. Bu şartlarda klasik yöntemler kullanılarak analizler yapıp, sonuçlar elde edilir. Ancak gerçek dünya problemlerinin büyük çoğunluğu birçok farklı bileşenden oluşan karmaşık problemler olmakta ve bu durumun sonucu olarak doğrusal olarak ayrılmış bir form halinde olmamaktadır. Böyle problemlerde de doğrusal olmayan sınıflandırma yöntemi kullanılmaktadır. Bu durumda doğrusal olarak ayrılamayan veriler öncelikle doğrusal olarak ayrılacakları farklı bir uzaya aktarılmalıdır. Doğrusal destek vektör makinelerinin işleyişi Şekil 4.6’ da gösterilmiştir.



Şekil 4.6 : Doğrusal destek vektör makineleri.

4.1.8 Yapısal risk minimizasyonu

Yapısal Risk Minimizasyonu'nun (SRM) temeli Vladimir Vapnik tarafından ispatlanan İstatistik Öğrenme Teorisi'ne dayanmaktadır (Bera, 2001). 1995'de Vapnik tarafından ispatlanan teori ise öngörüye yönelik modellemenin tanımını sorgulayarak yeni bir çözüm sistemi yaratmıştır. O zamana kadar kullanılan çözümlerin aksine, bu çözüm tam olarak ispatlanmış bir istatistik teorisine dayanmaktadır. Böylece Vapnik'in çalışması Fisher parametrelerini bırakıp, Glivenko-Cantelli-Kolmogorov'un genel yaklaşımları çerçevesinde teorik istatistiğin temellerine geri dönüşü gerçekleştirmiştir.

Geleneksel modelleme yaklaşımı:

- Verinin bilinmeyen istatistik dağılımı üzerine bir hipotez ileri sürmek
- Yüksek boyutlu verinin çok fazla değiştire ve uzun hesaplama zamanı gerektireceğini kabul etmek veya tutarlılık sorunu olan bazı değişkenleri önceden seçerek değişken sayısını azaltmak
- Daha doğru bir model bulup geçerli olduğunu ispatlamak adımlarından oluşmaktadır.

Fakat SRM yaklaşımı ile

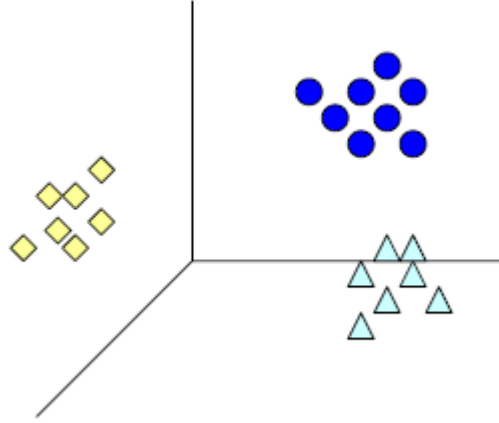
1. YRM bakış açısı ile, VC boyutunu kontrol ederek en iyi F model ailesini bulmak
2. Tanım olarak modelin tutarlılığı kontrol altında olduğundan tüm parametrelerle çalışmak
3. Doğruluk ve tutarlılık arasındaki en iyi dengeyi bulmak

Adımlarından oluşan süreç kullanılmaya başlanmıştır.

4.2 Kümeleme

Kümeleme analizi (Clustering) sınıflandırmada olduğu gibi sahip olunan verileri gruplara ayırma işlemidir. Sınıflandırma işleminde sınıflar önceden belirli iken, kümelemede sınıflar önceden belirli değildir. Sınıflandırmadan farklı olarak kümeleme analizinde ne kadar grup oluşacağı da belirli değildir.

Kümeleme modellerinde amaç, tanımlanan verileri kullanarak nesnelerin birbirlerine olan benzerlik ve farklılıklarına göre gruplara ayrılmasıdır. Kümeleme analizinde aynı grup elemanlarının olabildiğince birbirine benzer yani homojen, farklı grup elemanlarının birbirinden farklı yani heterojen olması istenmektedir. Belirlenen her bir grup küme olarak adlandırılmaktadır. Kümeleme örneği Şekil 4.7’ de gösterilmektedir.



Şekil 4.7 : Kümeleme örneği.

Veri tabanındaki veriler kümeler ayrılırken uzaklık ve benzerlik kavramlarından yararlanılmaktadır. Veri tabanındaki her bir kaydın diğer bir kayıtla olan benzerliği ya da her bir kaydın veri tabanındaki diğer kayıtlardan olan uzaklığı gibi oluşturulan gerçek ve aday kümeler arasındaki mesafe ve benzerliği içermektedir (Silahtaroglu, 2008). Uzaklık kavramı iki veri arasındaki mesafeyi göstermektedir. Benzerlik kavramı ise uzaklık kavramının tersine yakınlık kavramını ifade etmektedir.

Benzerlik kavramı genel olarak 4.16’deki gibi ifade edilmektedir.

$$ben(X_m, X_j) = \frac{1}{1 + mes(X_m, X_j)} \quad (4.16)$$

Kümelemede birçok uzaklık bağıntısı kullanılmaktadır. Bunlardan en yaygın kullanılanları Öklid uzaklığı, Manhattan uzaklığı ve Minkowski uzaklığıdır.

Öklid uzaklık bağıntısı uygulamada en çok kullanılan uzaklık ölçüsüdür. İki boyutlu uzayda Pisagor teoreminin bir uygulamasıdır. Öklid uzaklık bağıntısı 4.17’de yer almaktadır.

$$d(i, j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \quad (4.17)$$

Manhattan uzaklığı, gözlemler arasındaki mutlak uzaklıkların toplamı alınarak hesaplanmaktadır. Manhattan uzaklık bağıntısı 4.18’de yer almaktadır.

$$d(i, j) = \sum_{k=1}^p (|x_{ik} - x_{jk}|) \quad i, j = 1, 2, \dots, n; k = 1, 2, \dots, p \quad (4.18)$$

Minkowski uzaklığı p sayıda değişken göz önüne alınarak gözlem değerleri arasındaki uzaklık hesaplamasında kullanılmaktadır. Minkowski uzaklık bağıntısı 4.19’da yer almaktadır.

$$d(i, j) = \left[\sum_{k=1}^p (|x_{ik} - x_{jk}|^m) \right]^{1/m} \quad i, j = 1, 2, \dots, n; k = 1, 2, \dots, p \quad (4.19)$$

Veriler arasındaki mesafe veya benzerlik ölçülerek, birbirine benzeyen veriler aynı küme içinde toplanırlar. Kümeler bu şekilde oluşturulurken elde edilen kümelerin birbirlerinden farklı olup olmadığının da hesaplanması gerekmektedir. Örneğin bir veri hesaplamalar sonucunda K kümesine dahil olmuş olsun. Bu durumda K kümesinin diğer kümelere olan uzaklığı veya benzerliği ne kadar değişeceğinin, K kümesinin yeterli büyüklüğe ulaşp ulaşmadığının, K kümesinin daha küçük kümelere bölünüp bölünmemesi gerektiğinin anlaşılması gerekmektedir. Bu sorulara yanıt verilebilmesi için kümeler arasında da bir benzerlik ölçümünün yapılması gerekmektedir. Kümeler arasındaki mesafe ve benzerlik kavramlarının tanımları aşağıda yer almaktadır (Zhang, 1996).

Merkez (centroid) kümenin ortasını temsil etmektedir. Kümenin ortasında gerçekten böyle bir nokta bulunmasına gerek yoktur. Kümenin ortalamasını ifade etmektedir. Denklemi 4.20’de yer almaktadır.

$$Merkez = X_0 = \frac{\sum_{i=1}^N x_{mi}}{N}$$

(4.20)

Bazı algoritmalar ise kümeyi temsil etmesi için merkez yerine ortacı (medadoid) denilen bir temsilci kullanırlar. Algoritma içinde bu temsilci, kümenin elemanının durumuna göre sürekli değişmektedir.

Yarıçap (radius), R, küme elemanlarının merkeze olan uzaklığını ifade etmektedir. Denklemi 4.21’ de yer almaktadır.

$$R = \sqrt{\frac{\sum_{i=1}^N (x_{mi} - x_0)^2}{N}} \quad (4.21)$$

Çap (diameter) küme içerisindeki iki nokta arasındaki ortalama mesafeyi ifade etmektedir. Denklemi 4.22’de yer almaktadır.

$$\text{Çap} = D = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^N (x_{mi} - x_{mj})^2}{N(N-1)}} \quad (4.22)$$

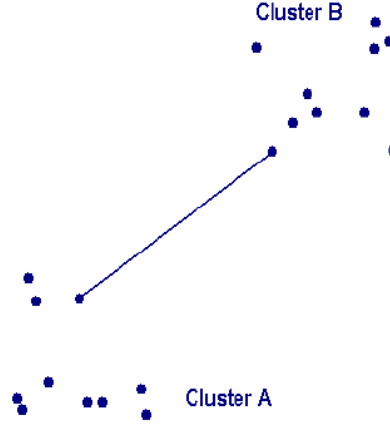
İki küme arasındaki mesafenin hesaplanması için 2 grup yöntem kullanılmaktadır. Bu yöntemler; Geometrik model ve şebeke modelleridir (Olson,1993).

Geometrik modelin de 3 çeşidi bulunmaktadır. İlk modelde iki kümenin merkezi hesaplanarak, iki küme temsilcisi arasındaki Öklid mesafesi bulunmaktadır. İkinci modelde ise kümeleri temsil etmesi için bir temsilci seçilir. Seçilen temsilci kümenin merkezine en yakın olan gerçek bir noktadır. Kümeye her katılım olduğunda kümenin merkezi yeniden hesaplanmakta ve mevcut noktalar arasında merkeze en yakın olan nokta o kümenin temsilcisi seçilmektedir. İkinci model ortacı yöntemi olarak da adlandırılmaktadır. Üçüncü modelde ise kümelerin merkezi temsilcisi olarak kullanılmakta, iki küme arasındaki mesafe ise her noktadan merkeze olan uzaklıkların kareleri toplamındaki değişim olarak hesaplanmaktadır. Üçüncü model ise minimum varyans yöntemi olarak da adlandırılmaktadır.

Şebeke modelinde ise kümelerin her bir elemanını bir düğüm olarak kabul etmekte ve kümeler arasındaki mesafeyi bu düğümlerin birbirlerine olan mesafesine göre tayin etmektedir. Bu yöntemler ise tek bağlantı, tam bağlantı ve ortalama bağlantı yöntemleridir (Olson, 1993).

Tek bağlantı yöntemi, iki küme arasındaki en kısa mesafeyi kümeler arası mesafe olarak kabul etmektedir. Yani kümeler arası mesafe hesaplanırken birbirine en yakın iki nokta arasındaki mesafe hesaplanır. Formülü 4.23’de yer almaktadır. Yapısı da Şekil 4.8’ de gösterilmektedir.

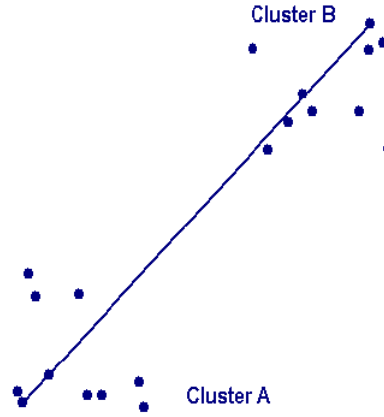
$$\text{mes}(K_1, K_2) = \min(\text{mes}(x_{1i}, x_{2i})) \quad (4.23)$$



Şekil 4.8 : Tek bağlantı kümeleme yöntemi örneği.

Tam bağlantı yönteminde ise birbirine en uzak iki düğüm arasındaki mesafe kümeler arası mesafe olarak kabul edilir. Formülü 4.24’de gösterilmektedir. İşleyişi ise Şekil 4.9’ da gösterilmektedir.

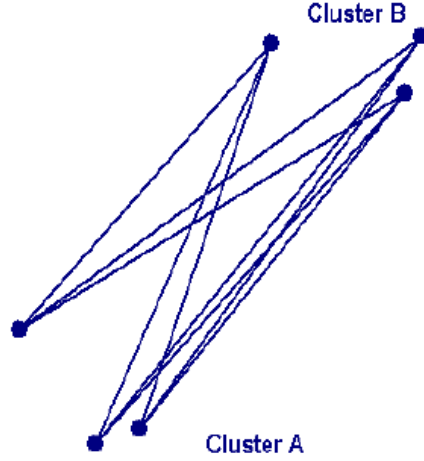
$$mes(K_1, K_2) = maks(mes(x_{1i}, x_{2i})) \quad (4.24)$$



Şekil 4.9 : Tam bağlantı kümeleme yöntemi örneği.

Ortalama bağlantı yönteminde ise her iki kümedeki noktalarının her birinin diğerine olan mesafelerinin ortalaması iki küme arasındaki mesafe olarak kabul edilmektedir. Formülü 4.25’de gösterilmektedir. İşleyişi ise Şekil 4.10’ da gösterilmektedir.

$$mes(K_1, K_2) = ortalama(mes(x_{1i}, x_{2i})) \quad (4.25)$$



Şekil 4.10 : Ortalama bağlantı yöntemi.

4.2.1 Kümeleme analizinin sınıflandırılması

Veri Madenciliğinde kullanılmakta olan çok sayıda kümeleme algoritması bulunmaktadır. Hangi algoritmanın kullanılacağı analiz edilecek verinin yapısına göre belirlenmektedir. Kümeleme yöntemleri genel olarak aşağıdaki gibi sınıflandırılmaktadır.

Hiyerarşik Yöntemler; toplasım kümeleme algoritmaları ve bölünür kümeleme algoritmaları olmak üzere ikiye ayrılırlar.

Bölümlemeli Yöntemler; K-medadoid yöntemler, k-means yöntemler ve yoğunluğa dayalı algoritmalar olmak üzere sınıflandırılmaktadırlar (Berkhin).

Grid Temelli Yöntemler

Kategorik Verinin Yinelenmesine Dayanan Yöntemler

Genetik Algoritmalar

Kısıtlara Dayanan Yöntemler

Makine Öğrenmesi Alanında Kullanılan Yöntemler

4.2.2 Hiyerarşik yöntemler

Hiyerarşik kümeleme yöntemleri, kümelerin bir ana küme olarak ele alınması ve sonra aşamalı olarak içerdiği alt kümelere ayrılması veya ayrı ayrı ele alınan kümelerin aşamalı olarak bir küme biçiminde birleştirilmesi esasına dayanmaktadır (Özkan, 2008). Toplasım kümeleme algoritmaları ve bölünür kümeleme algoritmaları olmak üzere ikiye ayrılırlar.

Toplařım kümeleme algoritmaları, bařlangıçta veri tabanındaki her bir noktayı ayrı bir küme olarak düşünür. Bu kümeleri birleřtirerek birbirinden ayrı kümeler oluřturur. Bölünür kümeleme algoritmaları ise bařlangıçtaki veri tabanındaki tüm noktaları tek bir kümedeymiş gibi kabul eder. Veri tabanını taradıkça birbirinden farklı olan noktaları kümeden dıřarı atarak, önceden belirlenmiş k adet kümeye daęıtır. Hiyerarřik kümeleme yönteminde benzerlik ve mesafe ölçütleri kullanılmakta; o neden tüm veri türlerine uygulanmakta kolaylık ve esneklik saęlamaktadır. Mesafe ve benzerlik ölçütlerini ifade etmesi için mesafe/benzerlik matrisi oluřturmaktadır. Kümeleme analizinde algoritmaların karmařıklıęını en çok artıran bu benzerlik/mesafe matrisleridir. Karmařıklıęı azaltmak için algoritmalar genellikle belirli bir eřik deęerinin altındaki deęerleri yok sayarlar.

Hiyerarřik kümeleme ařaęıdaki özelliklere sahiptir:

- Bir veri tabanını bir kaç kümeye ayrıřtırır.
- Bu ayrıřtırma dendogram adı verilen bir aęaç sayesinde yapılır.
- Bu aęaç, yapraklardan gövdeye doęru veya gövdeden yapraklara doęru kurulabilir.
- Ařaęıdan-yukarıya yaklařım (toplayıcı (agglomerative)) hiyerarřik kümeleme řu şekildedir:
 - Her bir nesne için farklı bir grup oluřturarak bařla,
 - Bazı kurallara göre grupları birleřtir: örn.; merkezler arasındaki uzaklık,
 - Bir sonlandırma durumuna ulařılıncaya kadar devam et.
- Yukarıdan ařaęıya yaklařımı (bölücü (divisive)):
 - Aynı kümedeki bütün nesnelerle bařla,
 - Bir kümeyi daha küçük kümelere böl,
 - Bir sonlandırma durumuna ulařılıncaya kadar devam et.

Slink algoritması ve tek baęlantı teknięi

Slink algoritması tek baęlantı ya da en yakın komřu teknięini kullanmaktadır (Sibson, 1973). Tek baęlantı yöntemi daha öncede anlatıldıęı üzere, kümeler arası mesafe ölçülürken iki küme içinde birbirine en yakın iki elemanın uzaklıęını kabul

eder. İlk aşamada mevcut verilerin mesafe/benzerlik matrisi elde edilir; bu matrisi bir ağaç haline dönüştürür. Şebeke modellerinden en küçük maliyetli ağaç elde edilerek verilen eşik değerine göre kümeler oluşturulur. Tek bağlantı tekniği ve SLINK algoritması toplasımli algoritmaların tipik özelliklerini taşımaktadır. Teknik literatürde en yakın komşu kümesi olarak adlandırılmaktadır (Dunham, 2003).

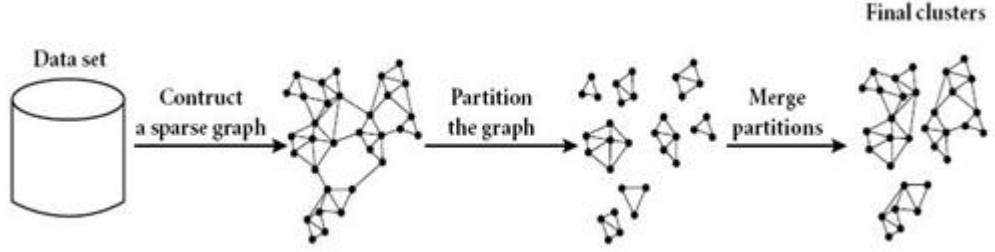
CURE algoritması

Kümeleme işlemi esnasında kümelerin kalitesini en çok etkileyen unsur, veri topluluğu içinde diğer verilerden uzakta bulunan, sayıları az olup aslında hiçbir kümeye ait olmaması gereken uç verilerdir. CURE (Clustering Using Representatives- Temsilciler kullanarak kümeleme) algoritması, uç verilerin oluşturulan kümelerin kalitesini etkilememesi amacıyla 1998 yılında geliştirilmiş bir algoritmadır. CURE algoritması başlangıçta her girdiyi sanki ayrı bir kümeymiş gibi ele alır. Her adımda bu küme temsilcilerin birbirlerine olan yakınlıklarına göre birleştirilir ya da ayrı küme olarak tutulur. Her bir küme için c adet iyi dağıtılmış temsilci nokta seçilir ve seçilen bu noktalara göre kümelerin fiziksel şekli ortaya çıkarılmış olur. Daha sonra dağıtılmış noktalar bir α katsayısıyla kümenin merkezine doğru kaydırılır. Dağıtılmış noktalar, bu kaydırma işleminden sonra artık o kümenin temsilcileri olarak kabul edilirler. Bu aşamadan sonra iki küme arasındaki uzaklık, her biri bir kümeye ait olan en yakın temsilci çifti arasındaki uzaklıktır. Temsilcilerin bir α katsayısıyla çarpılması sonucu kümenin merkezine kaydırılması, kümedeki yüzey anomalilerini tolere ettiği gibi uç verilerin etkisini de azaltmaktadır (Silahtaroglu, 2008). α nın alacağı değer 0-1 arasında yer alacaktır. Küçük değerli α ; dağıtılmış noktaların çok az yer değiştirmesine neden olacakken, büyük değerli α ; ise küme merkezine daha çok yaklaştıracığı için toplu halde kümeler oluşacaktır.

CHAMELEON algoritması

Chameleon algoritması 1999 yılında Karypis ve arkadaşları tarafından geliştirilmiş bir algoritmadır. Chameleon algoritması iki küme arasındaki benzerliği dinamik bir model kullanarak belirlemektedir. Diğer algoritmalarından farklı olarak iki alt kümenin birbirlerine olan benzerliği ve yakınlığı bu iki kümeden her birinin kendi iç benzerlikleri ve yakınlıkları ile kıyaslanarak belirlenmektedir. Yapılan karşılaştırmalar sonucunda bu iki alt küme birbirlerine yakınsa birleştirilmektedir. Bu yöntem sayesinde daha kaliteli ve homojen kümeler oluşturulmaktadır.

Benzerlik/mesafe matrisinin oluşturulabildiği tüm veri türleri ve veri kümeleri için uygulanabilecek bir algoritmadır (Silahtaroglu, 2008). Algoritmanın çalışma mantığı Şekil 4.11’ de gösterilmektedir.

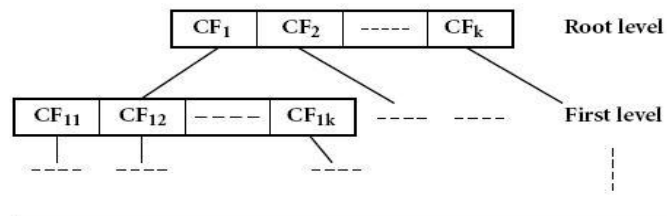


Şekil 4.11 : CHAMELEON algoritması çalışma mantığı.

BIRCH

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) çok büyük veri tabanlarının kümelmesi için geliştirilmiş gürültülü verilerin kontrol edilmesi için de bu alanda öne sürülen ilk algoritmadır (Zhang, 1996). BIRCH algoritması sadece sayısal veriler üzerinde kullanılabilir. BIRCH algoritması kümeleme işlemini bir ağaç yapısı oluşturarak gerçekleştirir. Burada belirtilen ağaç yapısına CF ağacı olarak tanımlanmaktadır. CF = (n,LS,SS) olarak 3 tane bilgiyi barındırır. Burada ki “n” kümedeki nokta sayısı, “LS” kümedeki noktaların toplam değeridir. “SS” kümedeki noktaların değerlerinin karelerinin toplamına eşittir. CF ağacı yani toplama 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 yerde bir alttaki düğüme geçilir. Aşağıdaki şekilde CF ağacının yapısı görülebilir.

CF ağacının yapısı Şekil 4.12’de gösterilmektedir.



Şekil 4.12 : CF ağaç yapısı.

4.2.3 Bölümlemeli yöntemler

Bölümlemeli yöntemlerde n adet nokta önceden verilen k adet küme sayısına ($k < n$) göre kümeler ayrılmaktadır. Hiyerarşik yöntemlerden farklı olarak yaratılacak küme sayısı önceden belirlenmektedir. Kullanıcı aynı zamanda algoritmaya kümeler arasındaki maksimum/minimum mesafeyi ve kümelerin iç benzerlik kriterlerini de vermek zorundadır (Giudici, 2004).

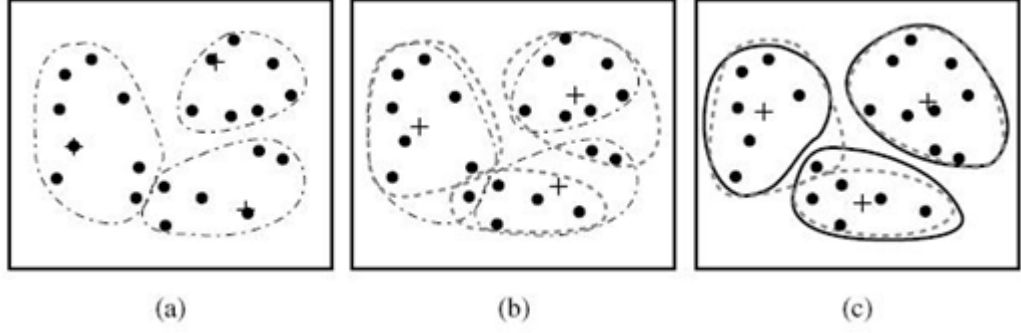
Bölümlemeli algoritmalar hiyerarşik algoritmalarındaki gibi benzerlik/uzaklık matrisi kullanmak zorunda olmadıklarından daha hızlı çalışmaktadırlar. Bu neden büyük veri tabanlarının kümelenmesinde hiyerarşik yöntemlere göre daha kolaylık sağlamaktadırlar. Bölümlemeli yöntemlerde bazı kriterler kullanıcı tarafından belirlendiği için birden fazla sonuç elde etmek mümkündür bu nedenle algoritmanın gerçekten en uygun çözümü bulup bulmadığından emin olunamayacaktır. Hangi sonucun en iyi olduğunun anlaşılabilmesi için algoritmanın değişik kriterlerle, dağıtılarak ve sıra ve yerleri değiştirilerek tekrar ve tekrar çalıştırılması gerekecektir. Çıkan sonuçların da birbiriyle kıyaslanıp en iyi sonucun hangisi olduğuna karar vermek maliyetin çok fazla artmasına neden olacaktır.

K-Ortalama (K-means) Algoritması

K-Ortalama algoritmasında; kümeler sürekli olarak yenilenmekte ve en uygun sonuç elde edilene kadar çalışmaktadır. Bölümlemeli algoritmaların tipik özelliklerini taşımaktadırlar. İlk olarak 1967 yılında MacQueen tarafından ortaya atılmıştır. K-ortalama algoritması mevcut verileri kümelerin ortalamalarına göre k adet kümeye ayırmaktadır. Ortalama denildiğinde anlaşılması gereken küme merkezleridir. k küme sayısı kullanıcı tarafından belirlenmektedir. Bir veri tabanının k -ortalama yöntemiyle kümelenmesi Şekil 4.13' de yer almaktadır.

Algoritmanın çalışma şekli aşağıda yer almaktadır.

1. Adımda keyfi olarak $m_1, m_2, \dots, m_k$ ortalamaları belirlenir.
2. Adımda her bir veri en yakın olduğu m_i kümesine atanır.
3. Adımda kümelere ait $m_1, m_2, \dots, m_k$ değerleri yeniden hesaplanır.
4. Adımda küme elemanlarında herhangi bir değişiklik yoksa algoritma durur.
Değişiklik var ise 5. Adıma devam eder.
5. Adımda ise ilk adımdan itibaren yeniden döngü devam eder.



Şekil 4.13 : Bir veritabanının k-ortalama yöntemiyle kümelenmesi.

Algoritmanın girdilerini eldeki veri tabanı ve ayrılması gereken küme sayısı, k , oluşturur. Daha önce de belirtildiği gibi kaç adet küme oluşturulacağı kullanıcı tarafından belirlenmektedir. Algoritmanın sonucunda da k adet küme elde edilmektedir.

Ortalamalar 4.26'daki formül kullanılarak hesaplanmaktadır.

$$m_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik} \quad (4.26)$$

n_k k kümesindeki eleman sayısını, X_{ik} ise k kümesindeki verilerin değerlerini ifade etmektedirler.

Her bir küme için kare-hata hesaplaması yapılmaktadır. Kare-hata ise küme içerisinde yer alan her bir elemanın, kümenin merkezine olan Öklid uzaklıklarının toplamına eşittir. Bu hata küme içi değişim olarak da adlandırılmaktadır. Küme içi değişim hesaplanması 4.27'de gösterildiği gibi olmaktadır.

$$e_i^2 = \sum_{k=1}^k (x_{ik} - M_k)^2 \quad (4.27)$$

Bütün kümeler için kare hatanın hesaplanması ise tüm küme içi değişimlerin toplamına eşittir. Formülü 4.28'de gösterilmektedir.

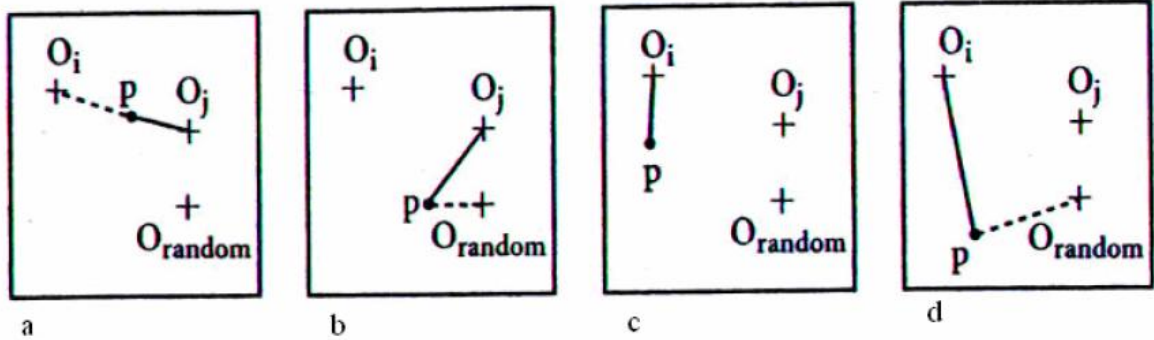
$$E_k^2 = \sum_{k=1}^K e_k^2 \quad (4.28)$$

Kare- hata kümeleme yönteminin amacı verilen k değeri için E_k^2 değerini minimize eden k kümelerini bulmaktır. Bu durumda her bir iterasyonda E_k^2 değerinin azalması gerekmektedir. Her bir elde edilen sonucun diğerinden daha iyi olması beklenmektedir.

K-ortalama algoritması yalnızca sayısal verilerde kullanılabilir. K- ortalama algoritmasında bir kümeyi diğerinden ayıran en önemli özellik ortalamalar olduğundan dolayı kategorik verilerde kullanılması anlamsız olmaktadır. Bu algoritma gürültülü ve uçtaki verilerden çok fazla etkilenmektedir.

K-medoids Algoritması

PAM (Partitioning Around Medoids) (Temsilciler etrafında Bölümleme) algoritması olarak da adlandırılan K-medoids algoritması Kaufman ve Rousseeuw tarafından 1990 yılında geliştirilmiştir. PAM algoritması k adet kümeyi oluşturmak için seçilen temsilcilerin etrafına ana kümedeki mevcut elemanları toplayarak, her defasında temsilcileri değiştirerek kümeleme işlemini tamamlamaktadır. PAM algoritmasının temsilci olarak seçtiği nokta medoid olarak adlandırılmaktadır. Temsilci yani medoid seçiminden kastedilen, kümenin merkezine en yakın noktanın belirlenmesidir. K sayıda küme oluşturulacağı için k sayıda temsilci seçilmektedir. K sayıda temsilcinin seçilmesinden sonra mevcut verilerden temsilciye en çok benzeyen veriler temsilcilerin etrafına toplanmaktadırlar. PAM algoritmasında daha önce anlatılan benzerlik ve uzaklık ölçümlerinin hepsi kullanılabilmektedir. Tercih yapma işlemi kullanıcıya bırakılmaktadır. Bir veri tabanının K-medoid yöntemi ile kümeleneceği Şekil 4.14’ de gösterilmektedir.



Şekil 4.14 : Bir veri tabanının K-medoid yöntemi ile kümeleneceği.

Algoritmanın çalışma şekli aşağıda yer almaktadır.

- 1.Adımda k adet temsilci ana veri tabanından rastgele seçilerek elde edilir.
- 2.Adımda veri tabanındaki temsilci olmayan diğer noktalar kendilerine en çok benzeyen temsilcilerin etrafında toplanır.
- 3.Adımda seçilen her temsilci seçilmeyen bir veriyle yer değiştirir.

4. Adımda temsilci değıştirmenin küme kalitesi üzerinde küme kalitesinde oluşturacağı iyileştirme hesaplanmaktadır.
5. Eğer daha performanslı sonuç elde ediliyorsa diğeri yerine yeni medoid bu nesne olur (yer değıştirilir) (örneğin a kümesinden bir nesne seçerek b ve a kümeleriyle karşılaştır ve eğer daha kaliteli bir duruma gelecekse yer değıştirilir.)
6. Bir değışiklik olmayana kadar iterasyona devam edilir.

Yapılan deneysel çalışmalar sonucunda PAM algoritmasının 100 nesneli ve 5 ayrı kümeli küçük veri tabanlarında daha iyi çalıştığı sonucu elde edilmiştir. PAM algoritması gürültülü ve uç verilerden k-ortalama yöntemine göre daha az etkilenmektedir.

CLARA algoritması

CLARA (Clustering Large Applications-Geniş Uygulamaların Kümelenmesi) büyük veri tabanlarının daha kısa sürede kümelenmesi amacıyla Kaufman ve Rousseeuw tarafından 1990 yılında geliştirilmiş bir algoritmadır. PAM algoritmasıyla kıyaslandığında CLARA algoritması daha geniş veri tabanlarında daha güvenli çalıştığı sonucuna varılmıştır. Algoritmanın çalışması aşağıdaki gibi olmaktadır.

1. Adımda bütün veri tabanı taranarak rastgele bir örnek küme elde edilir.
2. Adımda oluşturulan örnek küme üzerinde PAM algoritması uygulanır.
3. Adımda ana kümeyi oluşturan veri tabanından bir örnek küme daha seçilir.
4. Adımda bir önceki aşamada belirlenmiş temsilciler kullanılır.

Bu algoritma sayesinde temsilci değışimi azalmış, algoritma daha hızlı bir şekilde işlenmiş ve bu durumun sonucu olarak daha kaliteli sonuçlar elde edilmiş olacaktır.

CLARANS Algoritması

CLARANS (Clustering LArge Applicationsa based on randomized search- Rastgele Aramaya Dayalı Geniş Uygulamaları Kümeleme) Algoritması PAM ve CLARA algoritmalarının gelişmiş halidir. CLARANS algoritması veri tabanının temsilciler aracılığıyla ve bir şebeke diyagramından yararlanılarak k adet kümeye ayrılması olarak tanımlanmaktadır.

Yoğunluğa Dayalı Algoritmalar

Farklı şekillerdeki kümeleri belirlemek için yoğunluğa dayalı algoritmalar geliştirilmiştir. Bu tarz şekillerdeki kümelemenin mesafe ölçümüyle yapılması mümkün olmamaktadır ve bu tür kümelerdeki uç verilerin de ayıklanması gerekmektedir. Bu tarz durumlarda kümeleme işlemi yoğunluğa dayalı olarak yapılabilmektedir. Yoğunluktan kast edilen, analiz edilen nesnelerin sayısıdır. Genel olarak yapılan şey bir arada yoğunluk oluşturan verileri küme olarak kabul etmektir. Yoğunluğa dayalı algoritmalara örnek olarak DBSCAN, OPTICS ve DENCLUE algoritmaları verilebilir.

DBSCAN algoritmasında veriler 3 grupta toplamaktadır. Çekirdek noktalar yoğunluk-tabanlı kümenin içersinde yer alan noktalar olarak tanımlanmaktadır. Kenar noktalar ise çekirdek nokta değildir ancak çekirdek noktaların komşuluk alanları içersinde yer alırlar. Gürültü noktalar ise Çekirdek nokta kümesine de kenar nokta kümesine de girmeyen noktalar gürültü noktalardır. DBSCAN algoritması şöyle açıklanabilir: birbirine yeteri kadar yakınlıkta olan iki çekirdek nokta aynı kümeye konur. Aynı şekilde çekirdek noktaya yeteri kadar yakınlıkta olan bir kenar nokta çekirdek noktayla aynı kümeye yerleştirilir. Yerleştirme işleminde bir kenar noktanın başka bir kümedeki çekirdek noktaya olan uzaklığı da dikkate alınmalıdır. Son aşamada gürültülü noktalar çıkartılır.

OPTICS (Ordering Points to Identfyng Clustering Structure) algoritması küme sayısının önceden belirlenmesi gibi kullanıcının belirlediği parametrelerle ilgili olumsuzluğu ortadan kaldırmak için geliştirilmiş bir algoritmadır.

DENCLUE algoritması (Density Based Clustering), veri tabanında bulunan noktaların etki fonksiyonlarının toplamından yararlanılarak elde edilen genel bir yoğunluk fonksiyonunun, yerel maksimumlarının kullanılmasıyla yoğunluğa dayalı bir kümeleme gerçekleştirilir. DENCLUE algoritması temel olarak 3 adımdan oluşmaktadır. İlk adımda her bir noktanın kendi çevresi kapmasındaki etki fonksiyonu hesaplanır. 2. adımda bu fonksiyonların toplamı hesaplanır, elde edilen toplam fonksiyon veri tabanının genel yoğunluğunu verir. 3. adımda ise kümeler matematiksel olarak yoğunluk çekicileri (yerel maksimumlar) yardımıyla tespit edilir (Silahtaroglu, 2008).

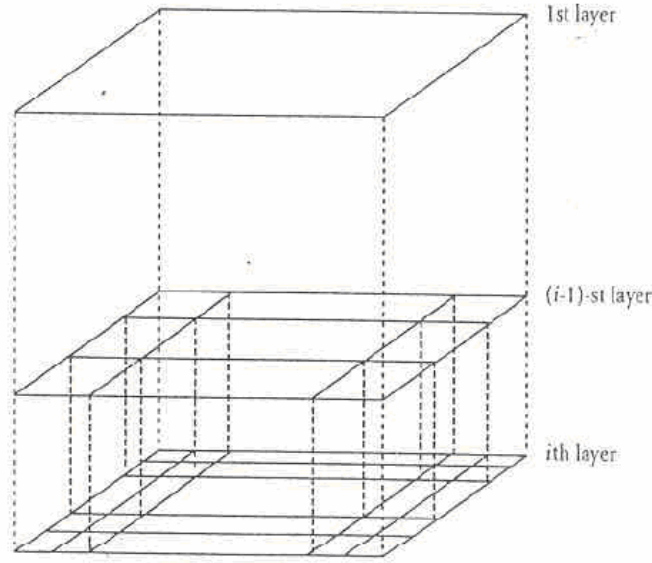
4.2.4 Grid Temelli Algoritmalar

Kümelemenin yapılabilmesi için çok yüksek miktarda bellek gerektiren büyük veri tabanlarının, kümelere ayrılması için numaralandırılmış çizgilerden oluşan hücresel yapılar yardımıyla yapılan kümeleme grid temelli algoritmalar olarak adlandırılmaktadır (Silahtaroglu, 2008).

STING

STING (Statistical Information Grid- İstatiksel bilgi Grid) algoritmasında ele alınan alan dikdörtgen hücrelere bölünmekte ve hiyerarşik bir yapı oluşturulmaktadır. Üst seviyede yer alan bütün hücreler bir sonraki seviyede parçalanmış şekilde yer almaktadır. Her bir grid hücresindeki niteliklerle ilgili istatistiksel bilgi kaydı tutulmaktadır. İstatistiksel bilgiler minimum, maksimum, var olan nokta sayısı, ve hücrede bulunan tüm sayısal değerlerin ortalaması ve standart sapması gibi bilgileri içermektedir.

Üst seviyede hücre istatistiksel parametreleri, alt seviyedeki hücre istatistiksel parametrelerden kolayca hesaplanabilmektedir. Bu parametreler: bağımsız parametre, sayı (count); nitelik-bağımlı parametreler, m (mean), s (standart sapma), min (minimum), max (maksimum) ve hücrelerin nitel dağılımlarındaki dağılım tipi: normal, tek biçimli, üstel, veya hiçbirisi (dağılım bilinmiyor) şeklinde olmaktadır. Veri, veritabanına kaydedilirken, en alt seviyede hücrelerdeki count, m, s, min ve max parametreleri direkt olarak hesaplanmaktadır. Dağılım değeri eğer dağılım tipi biliniyorsa kullanıcı tarafından önceden hesaplanabilir veya X^2 testi gibi hipotez testleri tanımlanabilir. Üst seviyedeki dağılım tipi, aynı alt seviyedeki hücrelerin birleştirilerek eşik filtreleme işleminden geçirilerek bulunabilir. Eğer alt seviye hücrelerdeki dağılım birbiriyle uyuşmuyorsa, eşik testi boşa gider ve üst seviyedeki dağılım tipi hiçbirisi olur. “İstatistiksel bilgi, sorgu cevabı için nasıl yarar sağlar?” İlk olarak, sorgu-cevap işleminin başlayacağı hiyerarşik yapıdaki seviye belirlenir. Bu katman genelde küçük sayıda hücre içerir. Katmandaki bütün hücreler için, verilen sorguya ilgisine göre güven aralığı hesaplanır. İlgisiz hücreler ileriki adımlar için silinir. Bu işlem en alt seviyeye ulaşılan kadar devam eder. Sorgu şartı sağlanırsa hücrelerdeki ilgili bölgeler döndürülür. İlgili veri, sorgunun gereklerini yerine getirene kadar yeniden düzeltilmekte ve işlenmektedir (Altıntaş, 2006). Sting kümelemenin hiyerarşik yapısı Şekil 4.15’ de yer almaktadır.



Şekil 4.15 : STING kümelemenin hiyerarşik yapısı.

STING'in kalitesi grid yapısındaki en alt seviye taneciğine bağlı olarak değişmektedir. Tanecikler hassas ise işlem maliyeti artmakta; bununla beraber, en alt seviyedeki grid yapısının kalın olması kümeleme analiz kalitesini azaltabilmektedir. STING, ana hücrenin çocuk ve komşuları ile ilişkilerini göz önünde bulundurmaz. Kümeleme sınırları düşey veya yataydır, diagonal sınır yoktur. Bu da kalite ve doğruluğu düşürmektedir. (SAS,1998).

Dalga Kümeleme

Dalga kümeleme (Wave cluster) algoritması büyük, hassas, değişik şekil ve biçimlerde veri tabanlarının kümelenmesi için kullanılmaya uygundur. Oluşturulacak küme sayısını algoritmanın kendisi belirlediğinden, dalga kümeleme denetimsiz öğrenme sağlamaktadır. Dalga kümelemede genel olarak aşağıdaki algoritma kullanılmaktadır (Silahtaroğlu, 2008).

1. Adımda özellik uzayı sayılaştırılır ve nesneler birimlere atanır.
2. Adımda dalga dönüşüm işlemi özellik uzayına uygulanır.
3. Adımda dönüştürülmüş özellik uzayının alt bantlarındaki birbirine bağlı kümeler elde edilir.
4. Adımda birimler isimlendirilir.
5. Adımda kontrol tablosu yaratılır.
6. Adımda nesnelerin kümelere atanma işlemi gerçekleştirilir.

Kümeleri belirlemek için şapka şekilli filtreler kullanılmakta ve eş zamanlı olarak zayıf bilgileri kendi sınırları içinde bastırmaktadır. Ana veri grubu içerisindeki kümeler ortaya belirgin bir şekilde çıkmakta ve kümenin etrafı berrak ve temiz bir hal almaktadır.

CLIQUE Algoritması

CLIQUE (Clustering in Quest) algoritması yoğunluğa dayalı ve grid temelli algoritmanın birleştirilmesinden oluşmaktadır. Çok yüksek veri gruplarının kümelenmesi amacıyla geliştirilmişlerdir.

CLIQUE algoritması çok boyutlu veri uzayının alt uzaylarında çalışmakta ve bu sayede performansı daha yüksek kümeleme gerçekleştirmektedir. Algoritma genel olarak yoğun olan bölgeleri diğer bölgelerden ayırmaktadır.

CLIQUE algoritması genel olarak 3 adımda tamamlanmaktadır:

1. Adımda kümeleri kapsayan alt uzaylar tanımlanmakta
2. Adımda kümeler tanımlanmakta
3. Adımda kümeler için minimum tanım genelleştirilmektedir.

4.2.5 Genetik algoritmalar

Sınıflandırma anlatılırken genetik algoritmalar konusu anlatılmıştı. Genetik algoritmalar kümeleme için de kullanılabilir. Genetik algoritmalar konu olarak şu bölümde anlatıldığı için bu kısımda sadece kümeleme analizinde kullanılan algorithmadan bahsedilecektir.

Genetik algoritmanın kümeleme analizinde çeşitli tekniklerle kullanımı mevcuttur. Bu teknikler arasında kullanımı en yaygın olan Mauli ve Bandyopadhyay tarafından geliştirilmiş olan tekniktir. Bu teknikte veritabanını k adet kümeye ayırmak için gerekli olan kümelerin merkezini hesaplama işlemi genetik algoritma yardımıyla hesaplanmakta ve kümelerin merkezini minimum yapmak hedeflenmektedir. Kümelerin merkezi ise kümeyi oluşturan elemanların birbirleriyle olan Öklid mesafesi toplamına eşittir. Bu algoritmanın adımları aşağıdaki gibidir (Maulik ve Sanghamitra, 2000):

1. adım $t=0$
2. adım ilk nüfusu başlat $P(t)$
3. adım $P(t)$ için uygunluk hesapla

4. adım $t=t+1$
5. adım Durma kriterine erişildiyse 10. adıma git
6. adım $P(t-1)$ 'den $P(t)$ 'yi al
7. adım $P(t)$ 'yi üretme (crossover) işlemi uygula
8. adım $P(t)$ 'yi mutasyon işlemine uygula
9. adım 3. Adıma dön
10. adım Dur

İki numaralı adımda nüfusun başlatılmasından kastedilen, rastgele olarak k adet küme merkezinin tayin edilmesi ve tüm nüfusu temsil eden P kadar kromozom için tekrar edilmesidir. 3. Adımda kast edilen uygunluk ise kümeleme değerini ifade etmektedir. Sonraki adımlarda bahsedilen üretme ve mutasyon işlemleri önceden bahsedilen yöntemlerden herhangi birisiyle yapılabilmektedir.

4.3 Birliktelik Kuralları ve İlişki Analizi

Bir arada gerçekleşen olayları çözümlemek de veri madenciliğinin konuları kapsamı içine girmektedir. Olayların birlikte gerçekleşme durumlarını çözümleyen veri madenciliği yöntemleri birliktelik kuralları (association rules) olarak adlandırılmaktadır. Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları, pazarlama amaçlı olarak pazar sepeti analizi (Market Basket Analysis) adı altında veri madenciliğinde yaygın olarak kullanılmaktadır (Göral, 2007). Alış-veriş merkezlerinde ürün yerleştirmelerinde, market alanı tasarımında, markette sergilenecek ve satılacak ürünlerin tespit edilmesinde ve katalog tasarımı gibi birçok konuda pazar sepeti analizi sonuçlarından yararlanılabilmektedir. Bu yöntemler birlikte olma kurallarını belirli olasılıklarla ortaya koymaktadırlar. Bu olasılıklarla ilgili ölçütler diğer bölümlerde anlatılacaktır.

4.3.1 Destek ve güven ölçütleri

Pazar sepeti analizinde satılan ürünler arasındaki ilişkileri ortaya koymak için destek ve güven gibi iki ölçütten yararlanılmaktadır.

Destek sayısı A ve B ürün gruplarını birlikte içeren alışveriş sayısını ifade etmektedir. Gösterimi $sayı(A \rightarrow B)$ şeklindedir.

Birliktelik kuralı $A \rightarrow B$ biçiminde gösterilmektedir.

N tüm alışverişlerin sayısını göstermektedir.

Kural destek ölçütü bir ilişkinin tüm alışverişler içerisinde hangi oranda tekrarlandığını ifade eder. Bu destek ölçütü 4.29'daki gibi hesaplanmaktadır.

$$destek(A \rightarrow B) = \frac{sayı(A, B)}{N}$$

(4.29)

Kural güven ölçütü, A ürün grubunu alan müşterilerin B ürün grubunu da alma olasılığını ortaya koymaktadır. Bu güven ölçütü 4.30'daki gibi hesaplanmaktadır.

$$güven(A \rightarrow B) = \frac{sayı(A, B)}{sayı(A)} \quad (4.30)$$

Eşik değer, destek ve güven ölçütlerini karşılaştırmak amacıyla kullanılmaktadır. Hesaplanan destek ve güven ölçütlerinin destek (eşik) ve güven(eşik) değerlerinden büyük olması beklenmektedir.

Hesaplanan destek ve güven ölçütleri ne kadar büyükse birliktelik kuralları o kadar güçlüdür kanısına varılır.

4.3.2 Birliktelik analizinde kullanılan algoritmalar

AIS algoritması

AIS algoritması 1993 yılında Agrawal tarafından geliştirilmiştir. Veri tabanındaki isimlerin en başta A'dan Z'ye sıralanması kısıtlamasını taşımaktadır. AIS algoritması veri tabanını ilk kez taradığında hangi verilerin geniş nesneler olduğunu belirler. Geniş olan nesneleri aday nesne kümeleri olarak işaretler. İkinci taramada ise ilk taramada geniş oldukları belirlenen nesne kümeleriyle, o işlemin nesneleri arasındaki ortak nesne kümeleri belirlenir. Belirlenmiş olan ortak nesne kümeleri ile işlemde mevcut bulunan diğer nesnelerle birleştirilerek yeni aday kümeler oluşturulur. Herhangi bir nesne kümesi, bir işlemdeki nesnelerle birleşip aday kümelerden birini oluşturabilmesi için, birleşeceği nesnenin hem geniş hem de harf sırası açısından nesne kümesindeki tüm nesnelerden sonra geliyor olması gerekmektedir. AIS algoritması bu adımı gerçekleştirmek için bir budama tekniği kullanılmaktadır.

Budama tekniğinin amacı gereksiz kümelerin silinmesidir. Bu adımdan sonra her aday kümesinin desteği hesaplanır, destek seviyeleri eşikten büyük yada eşit olanlar geniş nesne kümesi olarak işaretlenir. Bir sonraki taramada geniş işareti taşıyan kümeler, yukarıda anlatılan aşamalardan tekrar geçerek bir sonraki aday kümesinin belirlenmesi için kullanılır.

SETM algoritması

SETM algoritmasında geniş nesne kümesinin her bir elemanı iki parametreden oluşmaktadır. Bunlardan ilki nesnenin ismi ikincisi ise bu nesneyi ayırt etmeye yarayan bir özellik numarasıdır. SETM algoritmasının AIS algoritmasından farkı özellik numarasını da kayıta tutmasıdır. Bu durumda algoritmanın yer ve zaman karmaşıklığını artırmaktadır.

Apriori Algoritması

Birliktelik kurallarının üretilmesinde en yaygın kullanılan algoritmadır. Agrawal ve Srikant tarafından 1994 yılında 20. VLDB (Very Large Database Endowment) konferansında sunulmuştur. Bu algoritma aşağıda verilen aşamalara sahiptir (Özkan, 2008):

1. Aşamada birliktelik çözümlemesinin yapılabilmesi için destek ve güven ölçütlerini karşılaştırmak üzere eşik değerleri belirlenir. Uygulamalardan elde edilen sonuçların eşik değerinden büyük veya eşit olması beklenir.
2. Aşamada veri tabanı taranarak çözümlemeye dahil edilecek her ürün için tekrar sayıları, yani destek sayıları hesaplanır. Elde edilen destek sayıları eşik destek sayısı ile karşılaştırılır. Eşik destek sayısından küçük değerlere sahip olan satırlar çözümlemeden çıkarılır. Uygun olan kayıtlar göz önüne alınır.
3. Aşamada yukarıdaki adımda seçilen ürünler bu kez ikişerli gruplandırılarak, bu grupların tekrar sayıları, yani destek sayıları elde edilir. Bu sayılar eşik destek sayıları ile karşılaştırılır. Eşik değerden küçük değerlere sahip satırlar çözümlemeden çıkarılır.
4. Aşamada üçerli, dörderli vb. gruplandırmalar yapılarak bu grupların destek sayıları elde edilir, eşik değerlerle karşılaştırılır, eşik değerin üstünde ve eşit olduğu müddetçe işleme devam edilir.

5. Aşamada ürün grubu belirlendikten sonra kural destek ölçütüne bakılarak birliktelik kuralları türetilir ve bu kuralların her birisiyle ilgili olarak güven ölçütleri hesaplanır.

Apriori algoritmasının SETM ve AIS algoritmasından farkı aday nesneleri üretirken veri tabanındaki işlemleri işin içine dahil etmeden sadece bir önceki taramada geniş olduğu tespit edilmiş nesne kümelerini kullanarak oluşturur yani Apriori algoritması geniş bir nesne kümesinin herhangi bir alt kümesinin de geniş olacağı varsayımına dayanmaktadır.

Literatürdeki diğer bağlantı analizi algoritmaları ise şunlardır (Silahtaroglu, 2008):

- AprioriTid Algoritması
- Apriori ve AprioriTid algoritmasının bir karışımı olan Apriori-Hybrid algoritması
- Geniş nesne kümelerini belirlemek için veri tabanından alınmış küçük örneklerin çok iyi sonuçlar vereceği fikrine dayanan OCD (Offline Candidate Determination-Sıradışı Aday Belirleme) Algoritması (Mannila, 1994)
- Veri tabanını küçük parçalara bölerek, bellekte işgal edilen yeri azaltıp daha hızlı sonuca ulaşma sağlayan bölümlleme (partitioning) tekniği (Toivonen, 1996)
- 1996'da Toivonen tarafından ortaya atılan ve veri tabanındaki tarama sayısını azaltan örnekleme (Savasere, 1995) tekniği
- Kullanıcıya her taramadan sonra oluşan kuralları gösterip, minimum destek seviyelerini değiştirme olanağı veren CARMA (Continuous Association Rule Mining Algorithm-Sürekli bağlantı kuralı madenciliği (Hidber, 1999) gibi algoritmalarıdır.

4.4 Ardışık Keşfi

Ardışık keşfi, verideki ardışık zamanlı örüntüleri belirlemek için kullanılmaktadır. Ardışık keşfi birliktelik analizine benzerlik göstermektedir ancak aralarındaki ilişki zamana dayanmaktadır. Pazar sepeti analizinde ürünlerin aynı zamanda alınması gereksinimi vardır lakin ardışık keşfinde ürünler zaman içerisinde herhangi bir sırada satın alınabilirler.

Ardışık keşfinin pratikte kullanımı ile ilgili bazı örnekler (Göral, 2007):

- Web sitesi davranışlarını öngörmek:

Müşteri şimdi ne yapacak?

Müşterinin bir sonraki en muhtemel alışverişi ne olacak?

- Segmentler arası geçişleri inceleyerek müşterinin hayat boyu değerini hesaplamak

Firmadan ayrılmak üzere olan bir müşterinin bulunduğu duruma gelmesine nasıl bir olaylar sırası neden oldu?

- Sahtekarlık tespiti

Sahtekarlık tespitinde en anlamlı işlem sırası nedir?

5. UYGULAMA

5.1 Giriş

Uygulama kuyumculuk sektöründe yer alan ABC adlı bir şirkette yapılmıştır. Öncelikli olarak kuyumculuk sektörünün tarihi ve gelişimi dünya ve Türkiye’de incelenmiştir. Dünya ve Türkiye’de kuyumculuk sektörünün mevcut durumu araştırılmış, kuyumculuk için pazarlar ve yerleri hakkında bilgi edinilmiştir. Şirket hakkında genel bilgiler verilmiş ve mevcut durumda müşterilere karşı sergiledikleri davranış ve sundukları kampanyaları incelenmiştir. Müşterilere ait sistemlerde saklı tutulan kayıtlar kullanılmıştır. Kayıtlar kullanılarak, firma stratejileri doğrultusunda müşterilerin şirkete sağladığı yararlar ilgili bilgi elde etme amacıyla veri madenciliği tekniklerinden faydalanılmıştır. Kümeleme analizi teknikleri SPSS Clementine paket programı kullanılarak düzenlenmiş müşteri verilerinde kullanılmıştır.

5.2 Uygulamada Kullanılan Paket Program

Uygulamada Clementine programı kullanıldı. Clementine, SPSS Inc. Şirketi tarafından veri madenciliği uygulamaları için tasarlanmış ve veri madenciliği dünyasının yazılımları arasında tercih sıralamasında üç kez birincilik ödülünü almış bir yazılımdır. Görselliğe önem verilerek tasarlanmıştır. Çalışma ekranında sürükleyip bırak ile nesne yerleştirme ve nesneleri birbirine bağlama işlemleri kolaylıkla yapılabilmektedir. Clementine ile veri madenciliği adımları olan verinin hazırlanması, veri temizleme, veri birleştirme, seçme, dönüştürme, veri kalitesini belirleme, hata ayıklama, model kurma, modelin değerlendirilmesi ve modelin izlenmesi konularını geliştirmiş bir teknoloji ile gerçekleştirme imkanı sunmaktadır.

Clementine de veri modelleme aşamasında zengin bir içerik sunmaktadır. Clementine içerisinde yer alan modelleme yöntemleri 3 ana grup altında toplanmaktadır.

- Tahminleyici Modeller: Neural Networks, iki farklı rule induction tekniği C5.0 ve C&R tree, regresyon, lojistik regresyon ve sequence detection olmak üzere 6 ayrı

teknik içermektedir. Tahminleyici modellerde bir dizi input değeri baz alınarak bir “sonuç” değerinin tahmin edilmesi amaçlı modeller söz konusudur.

- Sınıflama Amaçlı Modeller: Benzer nitelik gösteren segmentlerin belirlenmesi amaçlıdır. Kohonen ağları, K-ortalama, iki adımlı sınıflama olmak üzere üç ayrı sınıflama yöntemi bulunmaktadır.

- Birliktelik Teknikleri: Genelleştirilmiş tahminleyici yöntemler olarak da tanımlanmakta olup, belirli bir sonucu bir dizi kural ile ilişkilendirmeye çalışırlar. Clementine içerisinde APRIORI ve GRI olmak üzere iki ayrı ilişkisel kural belirleme yöntemi vardır.

İş problemlerinin irdelenmesi aşamasında is deneyimi önemlidir. Bu ilk adımda projenin amaç ve gerekliliklerinin is perspektifi ile anlaşılması, bu bilginin veri madenciliği problem tanımı olarak netleştirilmesi ve hedeflere ulaşma amaçlı ilk planların oluşturulması söz konusudur. Clementine ile birlikte opsiyonel olarak lisanslanan uygulama şablonları SPSS in farklı veri madenciliği projelerine dair ciddi bir is deneyimini kullanıcılarına aktarmayı amaçlayarak hazırlanmış bir programdır. Verinin anlaşılması aşamasında veri kaynaklarına bağlanma, veriyi tanıma, verinin kalitesini anlama ve verinin grafiksel olarak incelenmesi, hipotezleri oluşturma amaçlı veri gruplarını değerlendirme çalışmalarında Clementine grafikler ve tablolar belli bölgelerin seçimini yapma seçeneği sunmaktadır. Clementine içerisinde yer alan histogram, line plot, point plot, web associaion graphs, statistics, distribution graphs, data audit işlemcileri verinin ön incelemesinde sıkça kullanılan işlemcilerden bazılarıdır.

5.3 Mevcut Durum Analizi

5.3.1 Kuyumculuk sektörü giriş

Altın, gümüş gibi kıymetli metal veya alaşımların; eritildikten sonra işlenerek ziynet eşyasına dönüşmesi işlemlerinin yapıldığı meslek dalına Kuyumculuk denmektedir.

Günümüzde kuyumculuk, gelişmiş teknoloji kullanılarak; insanlığın eski çağlardan bugüne taşıdığı bilgi birikimiyle görsellik göz önünde bulundurularak icra edilen bir meslektir. İnsanların çeşitli zevk ve tercihlerini düşünerek, insanların daha iyi ve estetik hissetmelerini sağlamalarına yardımcı olacak ürünler tasarlanması, üretilmesi

ve insanların beğenilerine sunulması kuyumculuk mesleğinin çalışma alanlarını oluşturmaktadır.

Kuyumculuk sektöründe yıllardan beri değerli madenlerle çalışılmaktadır. Bu durum ekonomik olarak önem arz etmektedir. Paranın olmadığı ilk çağlarda bu değerli madenler bir ödeme aracı olarak kullanılmaya başlanmıştır.

Değerli madenlerin kullanılması, alım satım işlemlerinin daha rahat yapılabilmesine olanak sağlamaktadır. Alım satım işlemleri uluslar arası platformda da yapıldığından toplumlararası ilişkilerin gelişmesi üzerinde de oldukça etkili olmuştur. İnsanları asırlar boyu peşinden sürükleyen, savaş ve barışlara neden olan, estetik ve güzel görünümü ve maddi değerinden dolayı bazen takı eşyası, bazen külçe ve bazen de para olarak şekilden şekle giren bu madenler bütün zamanların en değerli ekonomik kaynakları oluşturmaktadır.

5.3.2 Kuyumculuk sektörü tarihçe

Kuyumculuğun tarihsel gelişimi çok eski çağlara dayanmaktadır. Kuyumculuk mesleği madenlerin keşfedilmesiyle başlamıştır. Günümüze kadar kendini koruyarak ve geliştirerek gelmiştir.

Kazılar sonucunda M.Ö 3000’lerde Mezopotamya’nın Ur kentinde kuyumculuk sanatının ilk örneklerine rastlanmıştır. Özellikle Ur kenti mezarlarındaki hükümdar mezarlarından çıkan ürünler gerdanlık, küpe, saç tokası, bilezik ve müzik aletleri gibi takı ve eşyalardır.

İlkel toplumlarda insanların vücutlarını takı ile süslemelerinde törelerin etkisi olduğu kadar dinsel veya bütünsel inançların da katkısı vardır. Anadolu’ da Arkaik Dönem takıları eski yerleşim bölgesi olan Diyarbakır yöresinde Çayönü Tepesi, Orta Anadolu ‘da Çatalhöyük, Aşıklar Höyük ve Köşk Höyük kazılarında çıkarılmıştır. Takıların üretilmeye başlanması yani kuyumculuk mesleğinin başlangıcı bakır, gümüş ve altın madenlerinin bulunmasıyla başlamıştır denilebilir. M.Ö. 4000 yıllarında değerli madenlerin bulunmasıyla kuyumculuğun tarihi başlamıştır. Bu tarihte kuyumculuğun ve kuyumculuktaki tasarımın çok gelişmiş olduğunun en önemli kanıtı kazılarla ortaya çıkan değerli örnekler olan eserlerdir. M.Ö 7. yüzyılda Anadolu’nun batı yarısında birbirinden değerli pek çok takı bulunmaktadır. Lydia devletinin egemen olduğu İç Batı Anadolu ‘da ise başkent Sardes kuyumculuğunun merkezi olarak bilinmektedir. Daha sonraları ise Antiokhia (Antakya) ilse

Aleksandria (İskenderiye) faaliyete geçmiş olmasına rağmen M.Ö 2. yüzyılın yarısında başlayan, birinci yüzyılda yoğunlaşan ekonomik sıkıntı, takı üretimini de sıkıntıları da beraberinde getirmiştir. Anadolu' un, Roma'ya bağlı bir eyalet olduğu dönemde ise konu edilen ekonomik sıkıntılar Roma dönemi kuyumculuğunda da yaşanmıştır.

Anadolu dünya kuyumculuğunun başlangıç yeri sayılabilir. İlk örnekleri tunç çağı eserleri arasında yer alıp, bu topraklarda yaşayan sanatkarlar tarafından yapılmıştır. Alacahöyük, Boğazköy, Truva, Eskiypar hazineleri bunu doğrulamaktadır. Anadolu da yaşamış olan uygarlıklarda fildişi ve değerli taşların işlenmesi biliniyordu ve bu taşlar altından yapılmış takılar üzerine yerleştirilmiştir.

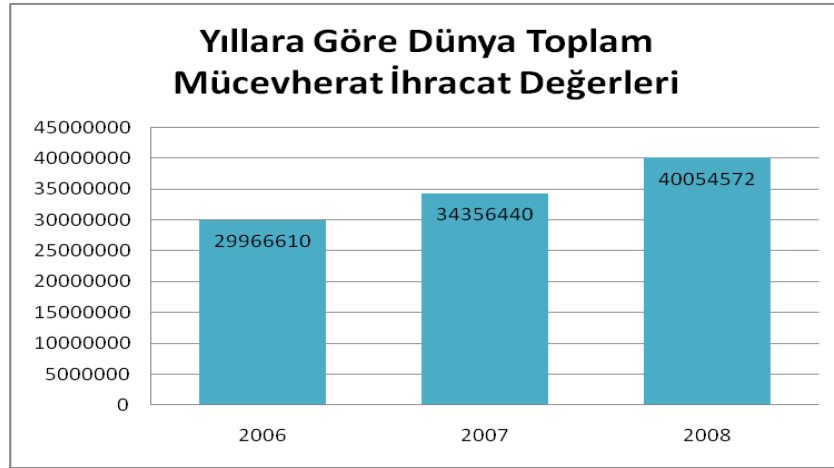
Selçuklular, Bizans kuyumcuları ile doğu ustalarının tekniklerini birleştirerek yeni bir sentez yaratmışlardır. Selçukluların Horasan ve Herat' da ki kuyumculuk merkezleri, başkent olan Konya 'ya da katılmıştır. Artuklu beyliğinin kuyumculukta önemli şehirleri olan Mardin, Şanlıurfa ve Diyarbakır ustalarıyla ünlüydü.

Osmanlı dönemine gelindiğinde, eski zamanların bütün ustalarından ve kuyumculuk tekniklerinden yararlanılmıştır. Osmanlılar'da kuyumculuğun en görkemli günleri Yavuz Sultan Selim ve Kanuni Sultan Süleyman dönemlerinde yaşanmıştır. Padişahlar kuyumculuğa büyük ilgi göstermişler, kuyumculukla ilgilenenlere sermaye yardımı yapılmıştır. Bu dönemdeki eserlerin en önemli özellikleri arasında, altın yüzeylerde savat işçiliğiyle beraber, demir, tutya, yeşim, necef üzerine altın kakma rumi, hatayi, çiçek desenleri, yine altınla yapılan çok kademeli kabartma ve oyma süslemeler sayılmaktadır. XVII. yüzyılda Osmanlı kuyumculuğunda tekrar bir sadeleşme görülür. Avrupa etkisinin görülmeye başlandığı XVII. yüzyıldan itibaren, eserlerde kalem işi, kabartma tekniği ve mine işçiliğiyle kaplı yüzeyler, elmas, yakut, zümrüt, turmalin, Seylan taşı ve incilerle süslenmiştir.

Günümüzde ise kuyumculuk sektörünün büyüme hızı özellikle 1990larda artmıştır. Sektörde işletmelerin gittikçe büyümesi ve ihracata yönelik üretimin artması ile el emeği ağırlıklı üretimden, teknoloji ağırlıklı üretime doğru bir yönelim olmuş ve el emeği ile çalışanların sayısında azalma görülmüştür. Yurt içi ve yurt dışı pazarlarında söz sahibi olabilmek için takı dizaynına önem verilerek bu alanda eğitim kurumları açılmıştır.

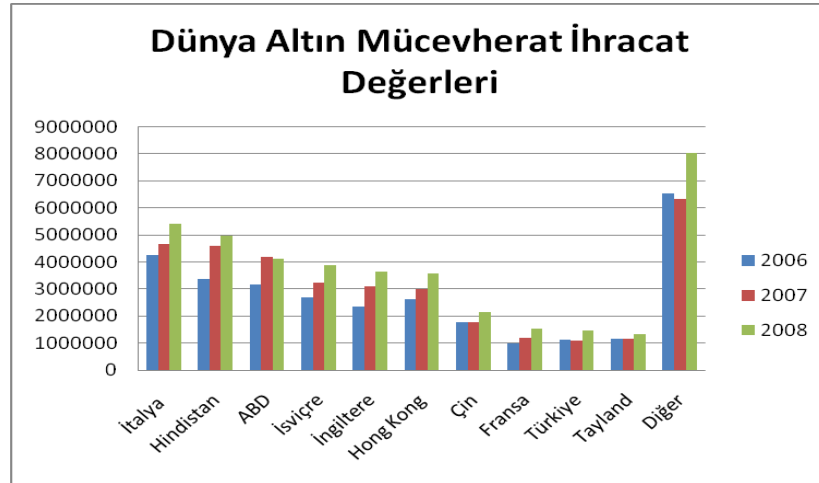
5.3.3 Dünyada kuyumculuk sektörü

Dünya mücevherat ticaret hacmi her geçen gün artmaktadır. 2007 yılında 2006 yılına göre %25,4 oranında artarak 76 milyar dolar seviyesine ulaşmıştır. 2007 yılında ihracat değeri 40 milyar dolar, ithalat değeri ise 36,3 milyar dolar değerinde gerçekleşmiştir. Son 3 yıllık mücevherat ihracat değeri Trademap ITC'den alınan bilgilere göre Şekil 5.1'de gösterilmektedir.



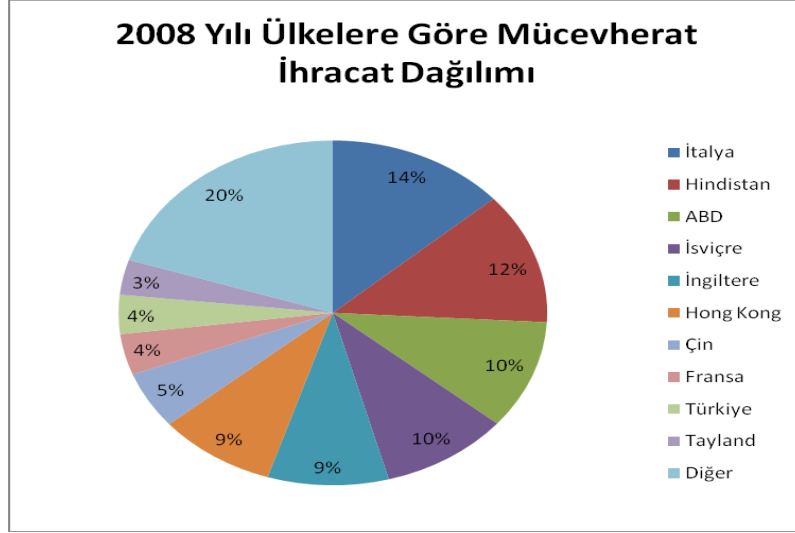
Şekil 5.1 : Yıllara göre dünya toplam mücevherat ihracat değerleri.

Trademap ITC'den alınan bilgilere göre dünya mücevherat ihracatının ülkelere göre dağılımı Şekil 5.2'de gösterilmektedir.



Şekil 5.2 : Dünya mücevherat ihracatının ülkelere göre dağılımı.

2008 yılı mücevherat ihracatının ülkelere göre oranı aşağıdaki grafikte yer almaktadır. Türkiye dünya mücevherat ihracat sıralamasında 9. Sırada yer almaktadır. Trademap ITC'den alınan bilgilere göre 2008 yılı ülkelere göre mücevherat ihracat dağılım grafiği Şekil 5.3'te gösterilmektedir.



Şekil 5.3 : 2008 yılı ihracat oranının ülkelere göre dağılımı.

Mücevher sektöründe dünyanın en önemli pazarları Hindistan, ABD, Orta doğu ve AB pazarlarıdır.

Hindistan 2006 yılı itibariye dünyanın en büyük altın mücevherat pazarıdır. Altın alımı özellikle festival ve düğün dönemlerinde artış göstermektedir. Hindistan’da kullanılan altınların birçoğu ithal edilirken üretim Hindistan’da kendilerine has tasarımlarla gerçekleştirilmektedir. Son yıllarda dünya altın mücevherat ihracat sıralamasında 2. sıraya yükselmişlerdir.

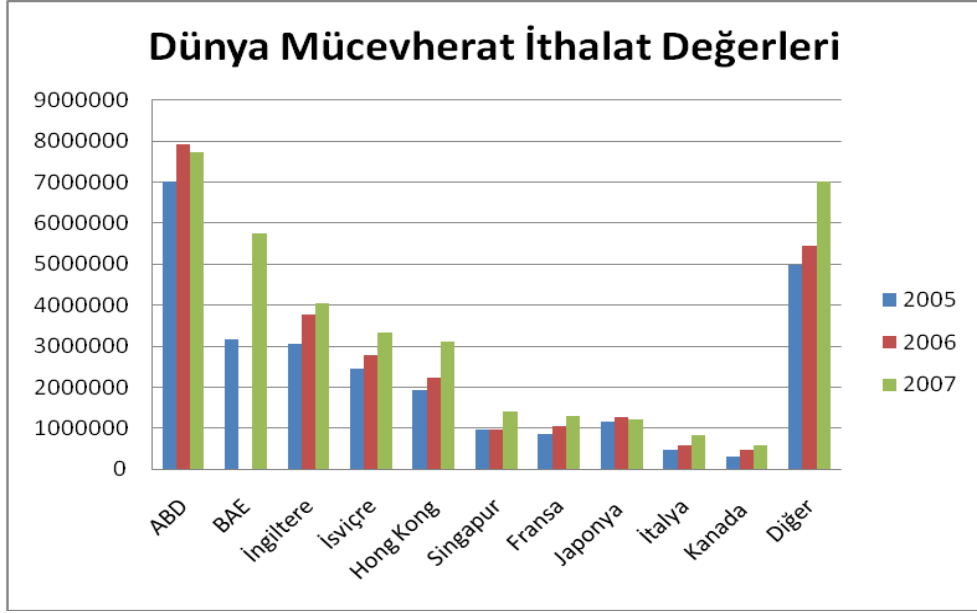
ABD pazarı, tüketim talebi açısından 2. sırada yer almaktadır. Rekabetin yoğun olarak yaşandığı bir pazardır. Noel gibi tatillerde ve önemli günlerde tüketim daha fazlalaşmaktadır. Talepler Hindistan, İtalya ve Türkiye gibi ülkeler tarafından karşılanmaktadır.

Orta Doğu pazarında altın mücevherat her zaman önem taşımaktadır. Özellikle Dubai dünya altın ticaretinde önemli bir yer haline gelmiştir. Sektördeki firmalar açısından Dubai pazarında yer almak büyük önem arz etmektedir çünkü dünyanın her bölgesine ulaşabilmeye olanak sağlamaktadır. Türkiye’nin altın mücevherat ihracatında Birleşik Arap Emirlikleri 2008 yılı itibariyle birinci sırada yer almaktadır.

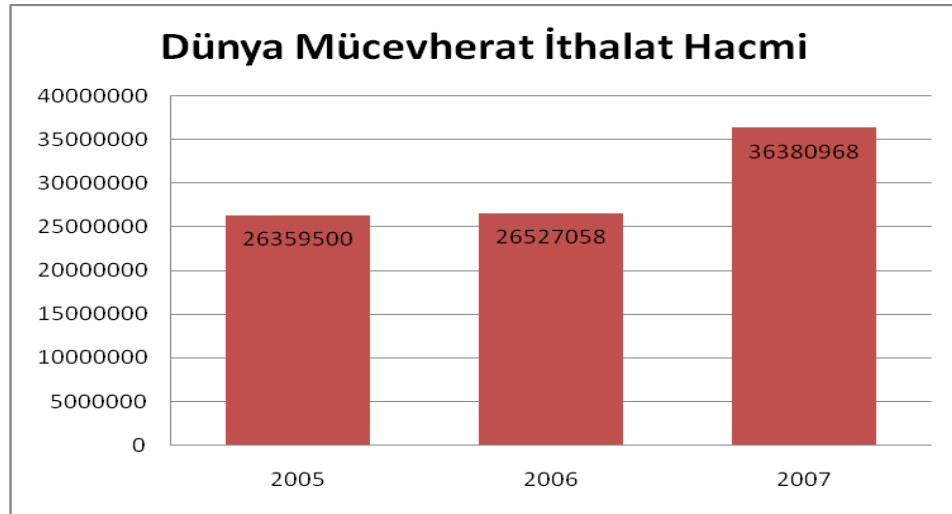
AB pazarı, ABD pazarından sonra tüketim açısından 2. sırada yer almaktadır. Dünya mücevher pazarının yaklaşık % 20’sini oluşturmaktadır.

Altın ithalat değeri 2008 yılında 426,3 milyon dolar olarak gerçekleşmiştir. İthalat yapılan başlıca ülkeler ise İtalya, Birleşik Arap Emirlikleri, ABD, Hong Kong, Endonezya ve Çin’dir. Trademap ITC’den alınan bilgilere göre dünya mücevherat

ithalat deęerleri Şekil 5.4'te, dünya mücevherat ithalat hacmi Şekil 5.5'te gösterilmektedir.

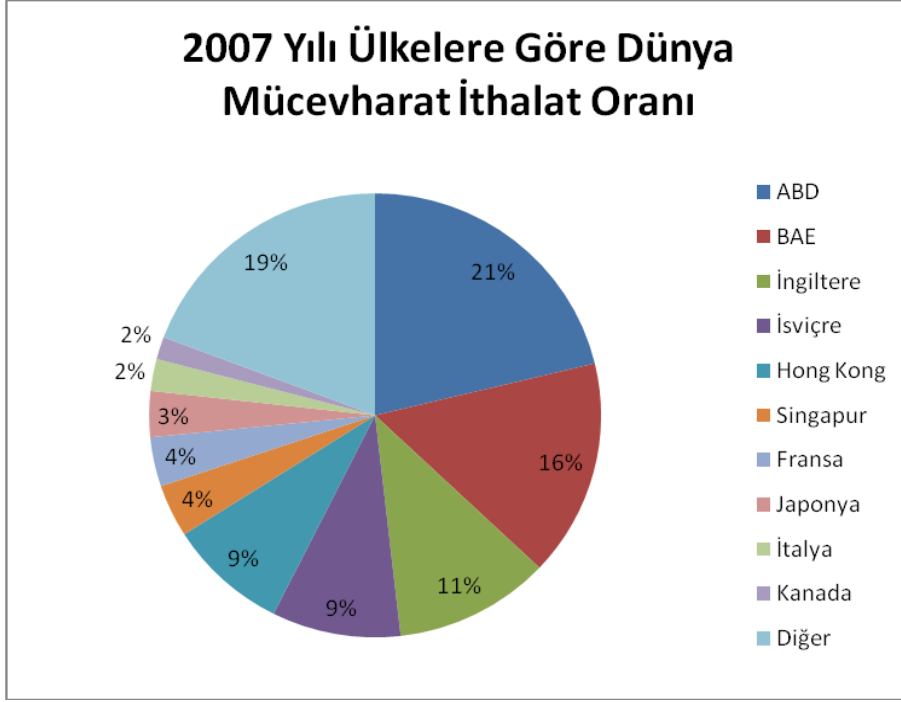


Şekil 5.4 : Ükelere göre dünya mücevherat ithalat deęerleri.



Şekil 5.5 : Yıllara göre dünya mücevherat ithalat hacmi.

Trademap ITC'den alınan bilgilere göre 2007 yılı dünya mücevherat ithalat oranının ülkelere göre dağılımı Şekil 5.6'da gösterilmektedir.



Şekil 5.6 : Ülkelere göre dünya mücevharat ithalat oranı.

5.3.4 Türk kuyumculuğunun yıllar itibariyle gelişimi

Cumhuriyet döneminde Türk kuyumculuğunun gelişiminde en etkili adım altın ticaretinin serbestleşmesiyle atılmıştır. 1980 öncesi altın ithalatının ve ihracatının yasak olması, yurt dışından yasal olmayan yollarla yurt içine getirilen altın miktarının sınırlı olmasından dolayı üretim kısıtlı olmaktaydı, doğal olarak bu durumda altın ticaretinin kısıtlı olmasına yol açmaktaydı. 24 Ocak 1980 kararlarıyla ekonominin tümünde görülmeye başlayan serbestleşme kararlarından biri de altın ticaretinin serbestleşmesiydi. Serbestleşmeden sonra kuyumculuk sektörü için büyük önem arz eden altının serbestleşmesi; kuyumculuk sektörü için bir dönüm noktası olmuştur. 1985 yılında külçe altının ithaline Merkez Bankası nezdinde izin verilmesiyle gelişimi daha da artmıştır. 1995 yılında kurulan Altın Borsasıyla altının hareket alanı genişlemiş, bu durumda kuyumculuk sektöründe serbestliğin artmasını sağlamıştır. Bu girişimler kuyumculuk sektörünün altının dünya fiyatlarına paralel bir şekilde alabilmesine imkan sağlamıştır (Özbek, 2009). Bu sayede Türkiye yurt dışıyla rekabet edebilecek şartlara sahip olmaya başlamıştır.

Kuyumculuk sektöründe ihracat yasağının 1983 yılında kaldırılmasından sonra üretimin ihracata yönelik kısmı önem kazanacak büyüklükte artmıştır. 1986 yılına kadar Ortadoğu pazarına yönelik çalışılmıştır. Gereken deneyim ve güveni bu şekilde sağladıktan sonra 1987 yılında Amerika pazarında yer almaya başlamıştır. Sektörün

büyürken iç pazarda talebin karşılanması için gerçekleştirdiği üretim, sektörün yaşadığı krizle darboğaza girmiş, zarardan kurtulabilmek için firmalar ihracata ağırlık vermeye başlamıştır (Özbek, 2009).

Türkiye 1994 yılına oranla %36,8’lik üretim artışıyla 1995 yılında önemli bir büyüme yakalamıştır. Buna rağmen Körfez Savaşı öncesi ulaştığı 130,90 tonluk üretimin ve 1994 ekonomik krizi öncesi ulaştığı 126,60 tonluk üretimin gerisinde kalmıştır. 1995 yılında 110,4 ton seviyesinde altın mücevherat üretimi gerçekleştirmiştir. Bu miktarla kuyumculuk alanında dünyanın en büyük yedinci ülkesi konumuna gelmiştir Aynı yılda dünyanın en büyük altın mücevherat üreticisi konumundaki İtalya’nın yıllık üretimi 446 ton düzeyinde gerçekleşmiştir. Bu rakam Türkiye’nin aynı yılda gerçekleştirdiği üretimin yaklaşık olarak 4 katına denk gelmektedir. Bu dönemde Dünya Altın Konseyi’nin yaptığı araştırmalara göre kuyumculuk sektörü Türkiye genelinde 5.000 atölye ve 40.000 dolayında kuyumcu dükkanı ve yaklaşık 130.000 dolaylarında kişiye istihdam sağladığı bilgileri elde edilmiştir. Aynı döneme ilişkin İstanbul Kuyumcular Odası’nın verileriyle Dünya Altın Konseyi araştırma sonuçlarına göre farklılık göstermektedir. İstanbul Kuyumcular Odası’nın verilerine göre ülke genelinde bulunan atölye sayısı 6.500 düzeyinde, sektörün yarattığı ortalama istihdam ise 250.000 düzeyindedir. Rakamların farklılıklarına rağmen kuyumculuk sektörünün Türkiye’de 1990’ların başında önemli bir büyüklüğe ulaştığını söylemek sonucuna varılabilir. Diğer taraftan 1985 öncesinde kuyumculuk sektörü usta çırak ilişkisine bağlı yürürken, 1985 sonrasında yöntem değişmiştir. Kurumlarında sektörle ilgili bölümlerin açılması ile kaliteli ve eğitimli işgücü sektöre girmeye başlamış, böylece sektördeki nitelikli iş gücünde artış gözlenmiştir (Özbek, 2009).

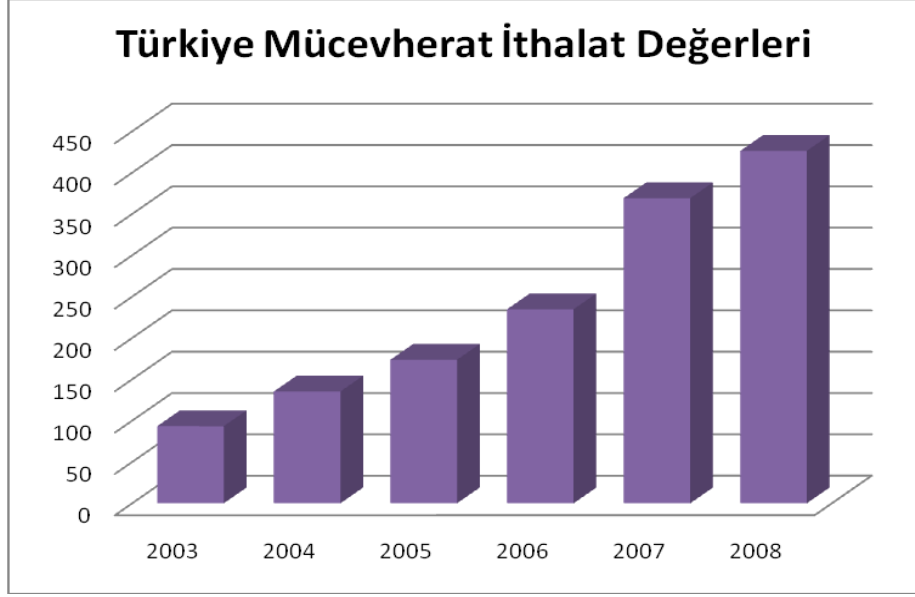
1995 yılında Türk altın sektörünü serbestleştirmek ve uluslar arası piyasalara uyum sağlamasını sağlamak amacıyla İstanbul Altın Borsası kurulmuştur.

Dünya Altın Konseyi’nin verilerine göre 1995 yılında gerçekleştirilen 110,40 tonluk üretimin %10’luk kısmı sektördeki büyük firmalara satıldıktan sonra o firmaların direkt ihraç etmesiyle yurt dışına çıkmıştır. %45’lik kısmı yabancı turistlere satılarak dolaylı olarak yurt dışına çıkmış, geri kalan %45’lik kısım ise iç piyasaya satılmıştır. Yapılan araştırmalar sonucunda İstanbul ve İzmir’in hem iç piyasa hem de yabancı turist talebinin fazla olduğu bölgeler olduğu, Ankara’nın iç talebin, Antalya ise yabancı turist talebinin fazla olduğu ve İstanbul ve İzmir’in hem iç piyasa hem de

yabancı turist talebinin fazla olduđu bölgeler olduđu ortaya çıkmıştır. Yapılan diğerk araştırma sonucunda ise Türk altın mücevherat piyasasının yılın farklı dönemlerine göre taleplerde farklılık oluştuđu gözlemlenmiştir. Buna göre yılın ilk döneminde iç piyasa talebinde artış olduđu gözlenirken, turizm sezonun yoğunlaştığı Mayıs-Eylül döneminde ise yabancı turist taleplerinde artış olduđu gözlemlenmiştir. 1980 sonrasında turizm sektöründe yaşanan gelişmeler de kuyumculuk sektörünü olumlu yönde etkilemiştir. Türkiye'ye turist olarak gelenler kuyumculuk sektörünün canlanmasını, taleplerin artmasını sağlamıştır (Özbek, 2009).

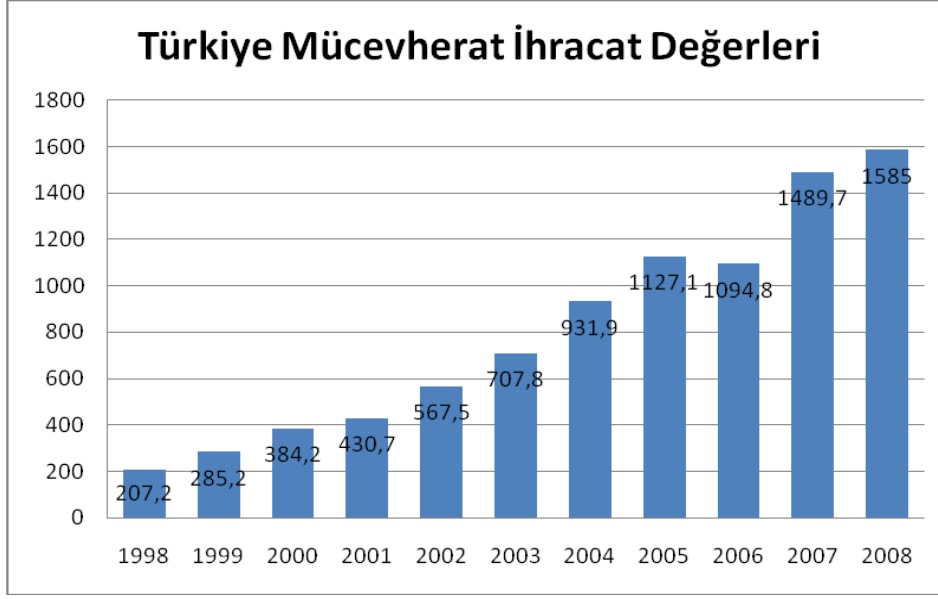
İstanbul Altın Rafinerisi 1995 yılında kurulmuş, 2002 yılı Şubat ayında ise faaliyete geçmiştir. Altın rafinerilerinin işlevi hurda altını külçe altına dönüştürerek ihracata yönlendirmek ya da yeniden kullanımını sağlamaktır. İstanbul Altın rafinerisinin kurulmasıyla yastık altı altınlar yeniden üretime veya ihracata katılmıştır, bu durum da sektörün hareketlenmesini sağlamıştır.

Günümüze gelindiğinde 2009 yılı itibariyle İstanbul Altın Borsası üye sayısı 62 olmuştur. Üyeler arasında bankalar, yetkili müesseseler, kuyumculuk sektörü üretici ve pazarlamacıları yer almaktadır. Türkiye son 12 yılda ortalama 100-200 ton altın ithal etmektedir. 2008 yılı itibariyle bu rakam 180 tondur. Altın ithalatı ise yalnızca gerekli izin belgesine sahip İstanbul Altın Borsası üyeleri tarafından yapılabilmektedir. Altın mücevherat ithalatının değeri 2008 yılında 426,3 milyon dolardır. İthalatın yapıldığı başlıca ülkeler İtalya, Birleşik Arap Emirlikleri, ABD, Hong Kong, Endonezya ve Çin'dir. Dış Ticaret Müsteşarlığı'ndan alınan bilgilere göre, yıllara göre Türkiye Mücevherat İthalat değerleri milyon dolar olarak Şekil 5.7'de gösterilmektedir.



Şekil 5.7 : Yıllara göre Türkiye mücevherat ithalat değerleri.

İhraç edilen ve turistlere satılan mücevherat miktarı, Türkiye’de sektörün % 70’ini oluşturmaktadır. Yurdumuza gelen turist artışıyla sektörün canlanacağı düşünülmektedir. Türkiye’de her sene 4 adet fuar organize edilmektedir. Bu fuarlar dünyadan birçok firma ve ziyaretçinin katılımını sağlayarak ihracat miktarının artırılması hedeflenmektedir. Türkiye’nin ihracat geçmişi son 10 yıla dayanmaktadır. Son 10 yılın ihracat değerleri değerlendirildiği bir artış eğilimi olduğu gözlemlenmektedir. Türkiye’nin ihracat miktarı 2008 yılında da 2007 yılına oranla % 6,4 artış göstermiştir. İhracat sıralamasında dünyada 9. Sırada yer almaktadır. Türkiye’nin başlıca ihracat yaptığı ülkeler Birleşik Arap Emirlikleri, ABD, Rusya, İtalya ve Almanya’dır. Dış Ticaret Müsteşarlığı’ndan alınan bilgilere göre Türkiye mücevherat ihracat değerleri Şekil 5.8’de gösterilmektedir.



Şekil 5.8 : Yıllara göre Türkiye mücevherat ihracat değerleri.

5.3.5 ABC şirketi hakkında genel bilgi

ABC, geleneksel el işçiliği ile en son teknolojinin getirdiği üretim tekniklerinin birleşmesiyle yaratılan nadide ürünlerin satışını yapmaktadır. Kalite ve kapasite, ABC' in pazardaki seçkin yerini açıklayan sözcüklerdir.

Yurdumuzdaki kuyumculuk sektörü gelişiminde liderliği hedefleyen ABC şirketi ABD, Kanada, İsrail, Avusturya, Hong Kong, İtalya, Portekiz, İspanya, Meksika, Arjantin, Rusya ve Almanya pazarlarına ürünlerini ihraç etmektedir. ABC ürünlerinin %40'ını dış pazarlara ihraç etmektedir.

Kuyumculuk dünyasındaki en yeni modaları yansıtan taşlı ve taşsız ABC bileklik, kolye, kolye ucu, mini set, yüzük, küpe ve el veya makine yapımı, içi boş zincirleri yaratıcılık ve kalitenin bulunduğu çarpıcı ürünlerdir.

ABC, şirket misyonunu, “sürekli değişen rekabet koşullarında, tüm müşterilerinin beklentilerini aşan ve rakiplerden daha üstün yenilikçi ürün ve hizmetler sunmak, çalışanlarının kendilerini sürekli geliştirmelerini sağlamak, sektörel sorunların çözümüne önderlik yaparak sektörün dünya çapında rekabet edebilir hâle gelmesine katkıda bulunmak ve toplumsal sorumluluk bilincinin gereği olarak toplumsal sorunlara duyarlılık göstermek” olarak belirtmektedir.

ABC' in satış faaliyetlerinin temelinde kuyumculuk dünyasının nefes kesen görüntülerini müşterilerine ulaştırmak yatmaktadır. Ürün ve hizmetleri ile birlikte, çalışanlarının yetenek ve bilgilerini geliştirmek üretim felsefesinin en önemli ana

başlıklarından birini oluşturmaktadır. Yurdumuzda sektörünün gelişimine inanan bir firma olarak temel hedefi, yurtiçi ve yurtdışı pazarlarda öncü olan bir ad olarak zihinlerde kalmaktır.

Takıda son moda ürünler pazarlamak, ömür boyu garantili hizmet sunmak, satış yapılan ürünlerin tam ayar ve üstün kalitede olmasını sağlamak, satış yapılan takıların hayatın her anına uygunluğunu sağlamak şeklinde faaliyetleri olan ABC şirketi, bu misyonları gerçekleştirirken yüzde yüz müşteri hoşnutluğu, ilkeli ve dürüst yönetim anlayışı, toplumsal sorumluluk bilinci gibi politikalarından taviz vermemektedir. Satış yapılan takıların belirlenmesinde; takıların dayanıklılığı, aynı model eski satışı yapılan ürünlerle eş değerli olması, takıların estetikliği ve takıların sağlamlığı gibi özellikler denetlenerek müşterilere adil bir hizmet vermeyi hedeflemektedir.

5.4 Uygulama

Veri Madenciliği'nde amaç manuel olarak işlem yapılması mümkün olmayan çok büyük miktarlardaki verilerden; otomatik olarak yararlı bilgi edilmesidir. Elde edilen yararlı bilgi doğru karar vermede etkili bir role sahiptir. Uygulamada; kuyumculuk sektörüne ait bir pazarlama şirketinin, mevcut müşterilerini gruplayarak, her bir gruba özel kampanyalar oluşturulması için altyapının hazırlanması uygulaması yapılmıştır.

Literatür bölümünde ayrıntılı olarak anlatıldığı üzere CRISP DM adımları veri madenciliğinde uygulanan adımları genel olarak ifade etmektedir. Uygulama yapılırken de bu adımlardan yararlanıldı.

5.4.1 İş sorusunu anlama

İş sorusunu anlama veri madenciliğinin özet olarak hangi problemin cevabı olarak yapıldığını ifade eder.

ABC firmasının çok fazla sayıda müşterisi vardır. Fakat bu müşterilerin bir kısmı şirkete fazla gelir getirmemekle beraber sipariş sayısı ve sıklığının fazlalığından dolayı şirkete zararı daha fazla olmaktadır. Kimi durumlarda o tarz müşterilere öncelik tanınması asıl geliri sağlayan müşterilerin kaybedilmesine neden olmaktadır.

Müşteriler şirkete sağladıkları katkı göz önünde bulundurularak gruplandırılırsa bu türlü kayıpların önüne geçilecektir.

5.4.2 Veriyi anlama ve hazırlama

Müşteri kayıtlarına bakıldığında müşterilere ait çok fazla sayıda data tutulmaktadır. Müşterilerin satış kayıtları müşteri baz alınarak tutulmamaktadır. Her bir siparişe göre kayıtlar tutulduğu için, her bir müşteriye belirli bir sürede ne kadar satış yapıldığı bilgilerine ulaşılabilmesi için müşteri birleştirmesi yapılması gerekmektedir. Çalışmanın en zor ve zaman alıcı kısmı verilerin anlaşılması ve hazırlanması adıımıdır. Yaklaşık 1500 müşteri ve 1500 müşterinin toplam 35000 siparişi, müşteri bazında tek tek birleştirilmiştir.

Müşterilere ait her bir sipariş için satış organizasyon kanalı, dağıtım kanalı, bölüm, bölge, sipariş veren, malzeme, dönem, faturalanan miktar, geri iadeler ve iskontolar, net satış miktarı, brüt satış miktarı gibi değişkenlerden oluşmaktadır.

Siparişler her bir müşteri için birleştirildikten sonra sipariş veren, net satış miktarı, faturalanan miktar, fatura kalem sayısı, bölge gibi değişkenlerin önemli olduğuna ve analizlerde bu değişkenlerin kullanılmasına karar verilmiştir.

Elde edilen bu değişkenlerden yeni bazı değişkenler türetilerek analize katkı sağlanmıştır. Bu değişkenler ise sipariş başına gelir miktarı, sipariş başına net satış miktarı ve sipariş başına ortalama faturalanan miktardır.

Bu adımda son 2 yıllık satış verileri göz önünde bulundurulmuş, 2 yıl içerisinde hiç satış yapılmayan müşteriler analize dahil edilmemiştir. Tüm kirli verilerin temizlenmesi ve verilerin hazırlanması işlemleri sonucunda her bir satır bir müşteriyi temsil etmek üzere 685 müşteri için veriler elde edilmiştir.

5.4.3 Modelleme

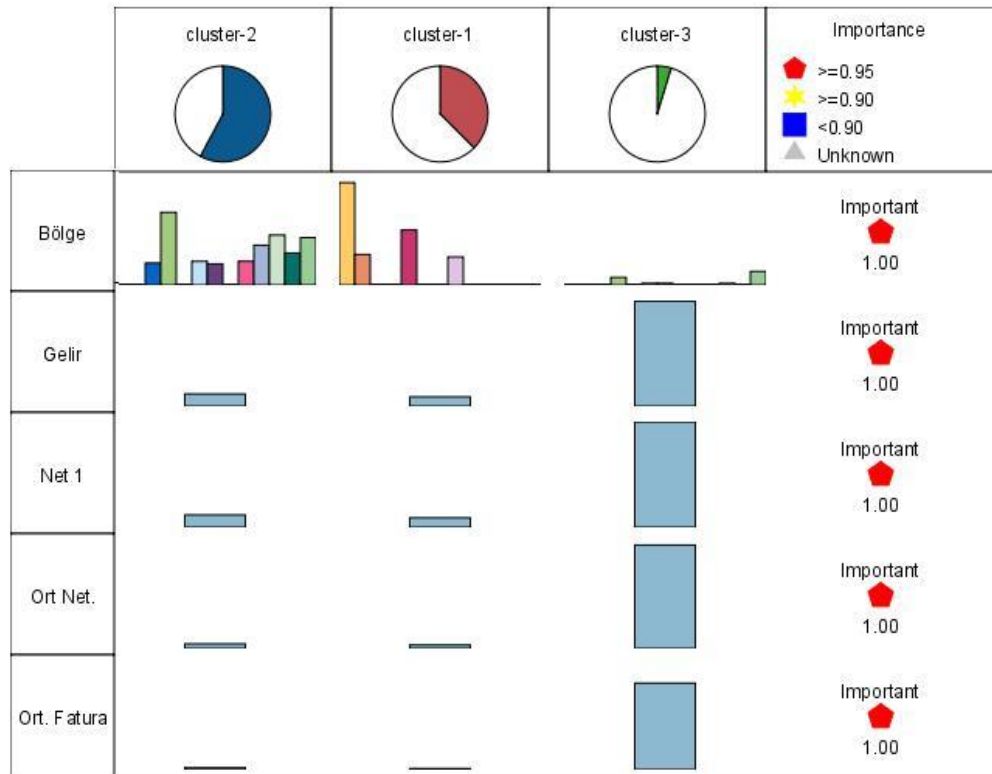
Analizin amacı müşteri segmentasyonunun gerçekleştirilmesi olduğu için kümeleme analizi yapılmıştır.

Kümeleme analizi SPSS Clementine paket program kullanılarak yapılmıştır. Analiz modeli olarak Two Step Algorithm seçilmiştir. Seçilme nedeni, segmentasyon

yapılırken kaç adet küme oluşturulacağı bilinmemekte, o nedenle algoritma seçilirken küme sayısını da belirlemesi gerekiyordu. Two Step algoritması bu duruma olanak sağlamaktadır. Analiz yapılırken 2 ile 15 küme arasında şeklinde geniş bir aralık sunarak kaç küme olması gerektiği algoritma sayesinde bulunmaktadır.

5.4.4 Uygulama

Modelleme sonucunda müşteriler 3 kümeye ayrılmıştır. Modelleme sonucu Şekil 5.9'da gösterilmektedir.

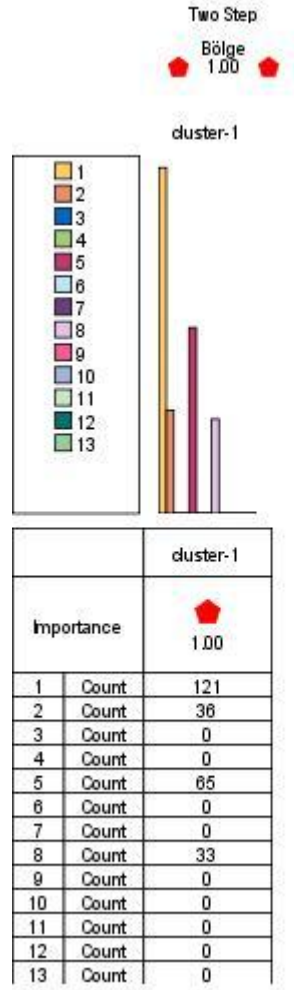


Şekil 5.9 : İki adım algoritması sonucu.

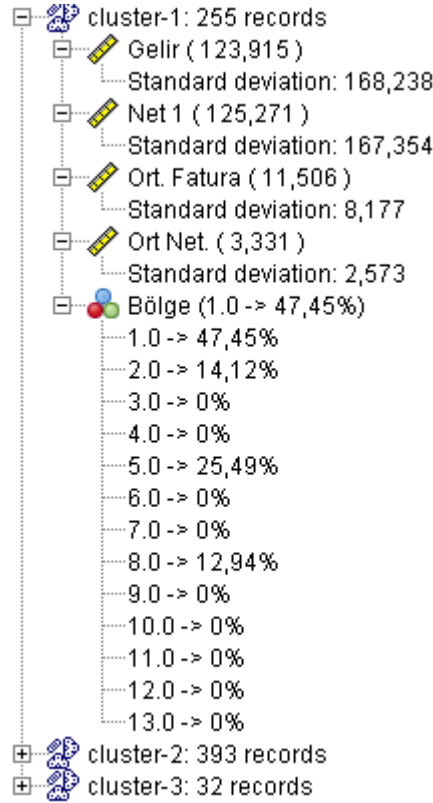
Değişkenler arasında bölge, müşterilerin getirdiği gelir, her bir müşteriye yapılan net satış miktarı, ortalama net gelir ve ortalama faturalanan miktar gibi değişkenlerin müşteri kümelemede çok etkili olduğu sonucuna varılmıştır.

1 Numaralı Küme

255 müşteriden oluşmaktadır. Ortalama sağladığı gelir miktarı en az olan kümedir. Her bir fatura başına faturalanan miktar ve net satış miktarı da diğer gruplara oranla daha azdır. 1 numaralı kümeye ait genel bilgiler şekil 5.11'de gösterilmektedir. Müşterilerin hangi bölgelere ait olduğu şekil 5.10'da gösterilmektedir.



Şekil 5.10 : 1 numaralı kümeye ait bölge bilgileri.

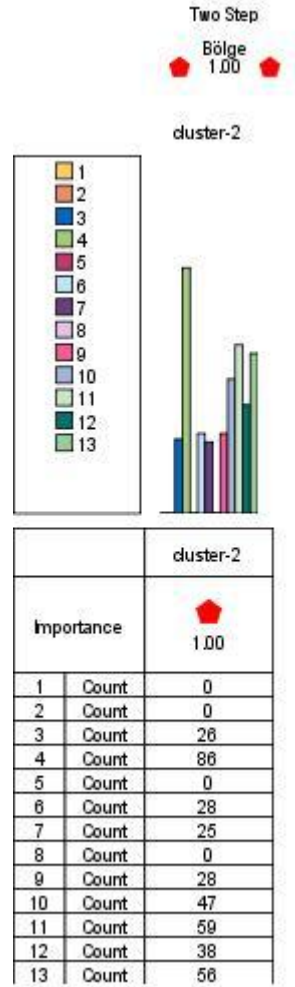


Şekil 5.11 : 1 numaralı kümeye ait müşteri genel bilgileri.

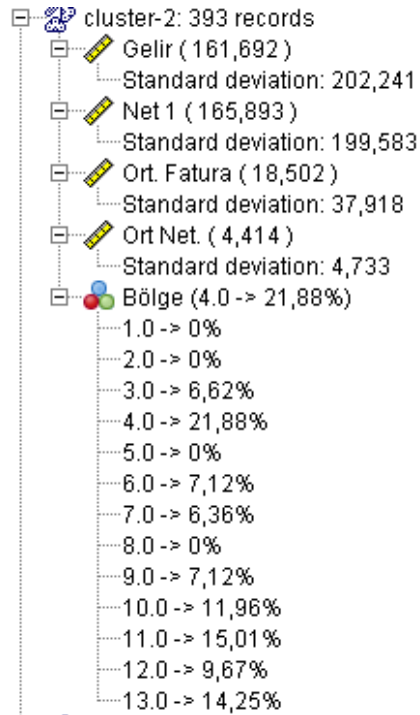
Müşteriler 1,2,5 ve 8 numaralı bölgelere aittir. Yani 1 numaralı küme müşterileri Anadolu, Ankara, Karadeniz ve personele yapılan satışlardan oluşmaktadır.

2 Numaralı Küme

393 kayıttan oluşmaktadır. Şirkete sağladığı gelir açısından 2. Sırada yer alan kümedir. Sağladığı gelir değerlendirildiğinde, 1 numaralı kümeyle arasında az fark var iken 3 numaralı küme ile yüksek oranda fark vardır. Her bir fatura başına faturalanan miktar ve net satış miktar da değerlendirildiğinde benzer bir oran mevcuttur. 2 numaralı küme müşterileri hakkında genel bilgiler şekil 5.13’de yer almaktadır. Müşterilerin hangi bölgelere ait olduğu şekil 5.12’de gösterilmektedir.



Şekil 5.12 : 2 numaralı kümeye ait müşteri bölgeleri.

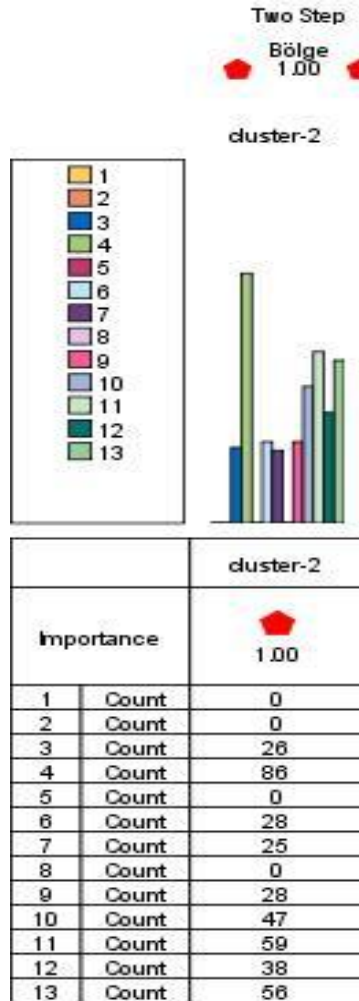


Şekil 5.13 : 2 numaralı kümeye ait müşteri genel bilgileri.

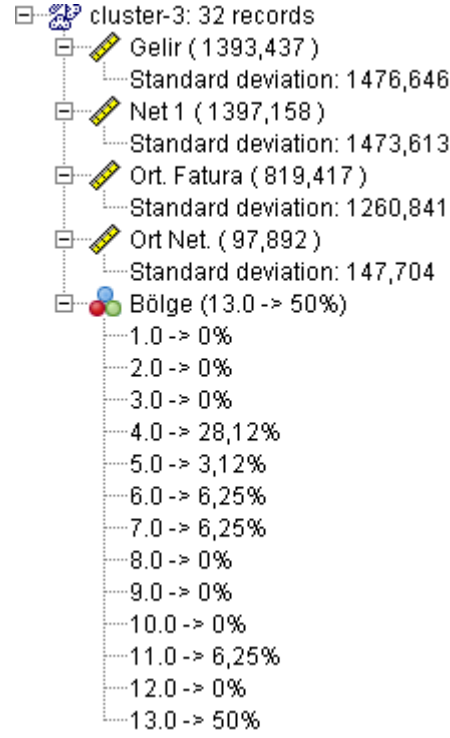
Müşteriler 3,4,6,7,9,10,11,12,13 numaralı bölgelere aittir. Yani müşteriler daha çok Bursa, Güney, Kapalıçarşı, Marmara, Trakya, İstanbul Anadolu, İstanbul Avrupa, İzmir ve yurtdışı müşterileridir.

3 Numaralı Küme

32 kayıtle en az müşteriye sahip olan kümedir. Diğer kümelerle sağladığı gelir, net satış değerleri vs. ile kıyaslandığında aralarında büyük bir fark olduğu gözlemlenmiştir. Sağladığı gelir açısından 2 numaralı kümenin yaklaşık 10 katıyken, sipariş başına düşen ortalama satış açısından yaklaşık 20 katı civarındadır. Bu durum göstermektedir ki 3 numaralı küme az sayıda sipariş vereni verdiği sipariş tutarları ise yüksek olan en değerli müşteri grubunu oluşturmaktadır. Bu durum şekil 5.15’de açıkça gözlenmiştir. Müşterilerin hangi bölgelere ait olduğu şekil 5.14’de gösterilmektedir.



Şekil 5.14 : 3 numaralı kümeye ait müşteri bölgeleri.



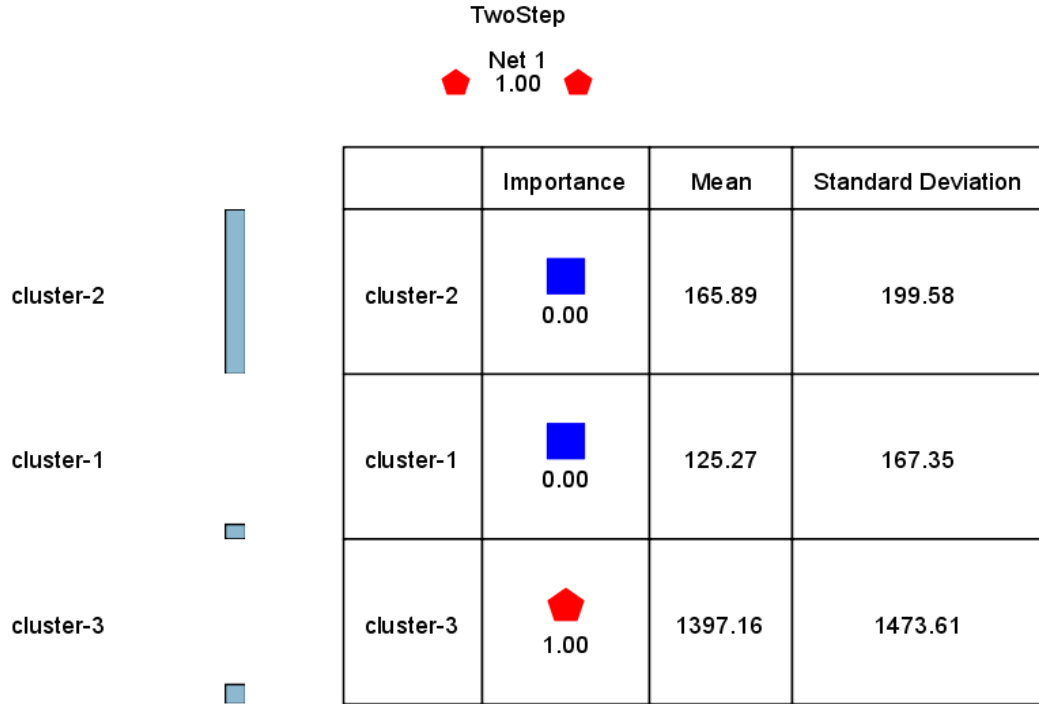
Şekil 5.15 : 3 numaralı kümeye ait müşteri genel bilgileri.

Sağladığı gelir açısından 2 numaralı kümenin yaklaşık 10 katıyken, sipariş başına düşen ortalama satış açısından yaklaşık 20 katı civarındadır. Bu durum göstermektedir ki 3 numaralı küme az sayıda sipariş vereni verdiği sipariş tutarları ise yüksek olan en değerli müşteri grubunu oluşturmaktadır.

Müşteriler 4,5,6,7,11 ve 13 numaralı bölgelere aittir. Müşteriler Güney, Karadeniz, Kapalıçarşı, Marmara, İstanbul Avrupa ve Yurt dışı müşterileridir.

Müşteri kümelerinin karşılaştırılması

Müşteriler şirkete sağladığı net gelirlere göre kümelerin karşılaştırılmasına şekil 5.16'da görülmektedir. Şekil 5.16 incelendiğinde ortalama net gelir miktarının en fazla 3 numaralı kümeye ait olduğuna sonucuna ulaşılır. 3 Numaralı kümeyle onu takip eden 2 numaralı kümenin ortalama net gelirleri arasında çok büyük farklılıklar vardır. 2 numaralı takip eden 1 numaralı kümeyle aralarında çok az fark bulunmaktadır. 3 numaralı kümede net gelir ortalamasının yüksek olma nedenleri arasında kümede yer alan müşteri sayısının az olmasıdır. 3 numaralı küme şirkete sağladığı gelir açısından en değerli ve az müşteri grubunu temsil etmektedir.



Şekil 5.16 : Net gelire göre müşteri kümelerinin karşılaştırılması.

Sipariş başına düşen ortalama net gelire göre müşteri kümelerinin karşılaştırılması

ABC şirketinde yaşanan en büyük sıkıntılardan biri küçük ve geliri az, maliyeti çok olan siparişler yüzünden geliri fazla olan siparişlerin kaybedilmesi idi. Bu açıdan değerlendirildiğinde sipariş başına düşen gelir miktarına göre müşteriler önem kazanmaktadır. Müşteri kümelerinin sipariş başına düşen ortalama gelirlere göre kıyaslanması sonucu şekil 5.17’de yer almaktadır. Şekil incelendiğinde 3 numaralı kümenin net gelir kıyaslamasında olduğu gibi daha yüksek bir değere sahip olduğu görülecektir. Bu yönüyle de 3 numaralı küme en değerli müşteri kümesini oluşturmaktadır.



Şekil 5.17 : Sipariş başına düşen gelire göre kümelerin karşılaştırılması.

Uygulama önerileri aşağıda yer almaktadır.

- Yeni gelen bir müşteri net satış miktarı, faturalanan miktar, fatura kalem sayısı, fatura başına düşen ortalama gelir miktarı, fatura başına düşen net gelir miktarı (son 6 aylık) göz önünde bulundurularak 3 kümeden birine yerleşmelidir.
- Sistemlerde müşteriler bulunduğu kümelere göre tanımlanmalı, sipariş vb diğer işlemler için sistemlere girildiği zaman otomatik olarak önceliklendirme yapılmalıdır.
- 3 numaralı kümeye ait olan müşteriler en öncelikli müşteriler olmalı, sonrasında 2 numaralı kümeye ait müşteriler gelmeli en sonda ise 1 numaralı kümeye ait müşteriler yer almalıdır.
- 3 numaralı kümeye ait olan müşteriler daha yoğun olarak Güney, Karadeniz, Kapalıçarşı, Marmara, İstanbul Avrupa ve Yurt dışı müşterileri olduğundan dolayı o bölgelerde o kümedeki müşterilere hitap edecek yetkinlikte satış temsilcileri atanmalıdır.
- 3 numaralı küme müşterilerin işlemleri sürekli analiz edilerek varsa satışlardaki değişiklikler gözlemlenmeli, herhangi bir düşüş sırasında müşteri elde tutma kampanyaları düzenleyerek müşteri kaybı engellenmelidir.

- Aynı işlemler 1 ve 2 numaralı kümeler için de gerçekleştirilmelidir, 3 numaralı kümenin sağladığı gelir daha yüksek olması dolayısıyla müşteri kaybetme riski yüksektir.
- 3 numaralı kümeye ait müşterilerin ayrıntılı olarak ciro vb. bilgilerine ulaşılabilirdiği durumlarda, müşteriye ait potansiyelin ne kadarının kullanıldığı anlaşılmalıdır. Cüzdan payı olarak da nitelendirilen bu durumda müşteri satışlarının yüzde kaçının rakipler tarafından gerçekleştirildiği bilgisine ulaşılmalıdır. Bu oran çok yüksek ise müşteriye özel teklif ve kampanyalar sunularak, tüm satışlarının şirket üzerinden yapılması sağlanabilir ve bu sayede müşterinin şirkete bağımlılığı ve bağlılığı arttırılabilir.
- 2 numaralı kümeye ait müşterilerden sağladığı net gelir vb. diğer kriterler açısından 3 numaralı kümeye en yakın değerlere sahip olanlar için satış temsilcileri atanmalı, ve satış değerleri yükselterek en değerli müşteri kümesi olan 3 numaralı kümeye ait müşteri haline getirilebilir.
- 1 numaralı kümeye ait, kümeleme analizinde kullanılan kriterler açısından en düşük değerlere sahip olan 1 numaralı küme müşterileri yok sayılabilir.
- Müşteri kümelerin bölgelere dağılımları göz önünde bulundurularak satış personeli için iş planlaması yapılabilir.

5.4.5 İzleme

Kurulan modelin izlenmesi aşamasını içerir. Müşteri davranışları sürekli değişeceği için kümeleme analizi en az yılda bir kez tekrarlanmalı, eğer yapıldıysa sistemlerdeki tanımlamalar yenilenmelidir. Kümeleme analizinde kullanılan kriterler ise sektör şartları ve günümüz şartları sürekli değiştiği için 2 senede 1 yenilenmelidir.

6. SONUÇ VE ÖNERİLER

Bu çalışmada Türkiye’de ve dünyada kuyumculuk sektörü incelenmiştir. Şirket stratejisindeki değişikliğe paralel olarak müşteriler şirkete sağladıkları gelir ve işlem sayısı göz önünde bulundurularak veri madenciliği teknikleriyle kümelere ayrılmıştır. Müşteri kümeleri arasındaki geçişler belirlenmiştir. Elde edilmiş müşteri kümelerinin yarar sağlayacak şekilde kullanılabilmesi doğrultusunda hayata geçirilmesi için alınması gereken aksiyonlar belirlenmiştir. Öneri olarak ise çalışmada en çok zaman kaybını sistemlerde kayıtların işlem bazlı tutulması oluşturmuştur. Sistemlerde müşteri birleştirilmesinin yapılması, kümeleme analizin daha kolay yapılmasını sağlayacaktır. Elde edilen kümeler ve kümeler arasındaki geçişler en az senede bir tekrarlanmalı, kümelemede kullanılan kriterler ise 2 senede bir yeniden belirlenmeli bu sayede güncelliğini devam ettirmelidir.

KAYNAKLAR

- Akpınar, H.**, 2000. Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, *İÜ. İşletme Fakültesi Dergisi*, İstanbul.
- Akpınar, H.**, 2004. *Business Intellegience & Data Mining*, Dönence ve Basın Yayın Hizmetleri, İstanbul.
- Amasyalı M.F., Diri B., Türkoğlu F.**, 2006. Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi, *15. Türkiye Yapay Sinir Ağları Sempozyumu*, İstanbul.
- Argüden, Y., Erşahin, B.**, 2008. *Veriden Bilgiye Masraftan Değere*, ARGE Danışmanlık, İstanbul.
- Atan M., Çatalbaş E.**, 2004. Çok Değişkenli İstatistiksel Analiz Yöntemleri ile Türk Bankacılık Sektöründe Çok Boyutlu Mali Başarısızlık Tahmin Modelleri Oluşturulması, *4. İstatistik Günleri Sempozyumu*, İzmir.
- Ateş H.**, 2008. Karar Vermede İş Zekasının Önemi: Tekstil Sektöründe Bir Araştırma, *Yüksek Lisans Tezi*, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü, İzmir.
- Bera, M.**, 2001. *The New Approaches to Predictive Modeling with a Very Great Number of Variables*, KXEN Inc., USA.
- Berkhin P.**, 2009. *Survey of Clustering Data Mining Techniques*, <<http://citeseer.nj.nec.com/berkhin02survey.html>>, alındığı tarih 23.02.2010.
- Berry, M.J.A. and Linoff, G.S.**, 2000. *Mastering Data Mining*, John Wileys&Sons, New York.
- Berson, M.A. and Smith, S.**, 1997. *Data Warehousing, Data Mining and OLAP*, McGraw Hill, New York.
- Bircan H.**, 2004. Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama, *Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, Kocaeli.
- Biere, M.**, 2003. *Business Intelligence for Enterprise*, Prentice Hall, New Jersey.
- Brachman R., Anand T.**, 1996. The Process of Knowledge Discovery in Databases: A Human-Centered Approach Advances in Knowledge Discovery and Data Mining, *MIT Press*, Cambridge.
- Burges C.J.C.**, 1998. Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, *İÜ. İşletme Fakültesi Dergisi*, İstanbul.
- Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Kamber, M.**, 1998. *Discovering Data Mining from Concept to Implemantation*, Prentice Hall, New Jersey.

- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R.,** 1996. CRISP-DM 1.0: Step-by-Step Data Mining Guide, <<http://www.crispdm.org/CRISPWP-0800.pdf>> , alındığı tarih 18.12.2009.
- Chopoorian, J.A., Witherell, R., Khalil, O.E.M and Ahmed, M.,** 2001. Mind Your Business by Mining Your Data, *SAM Advanced Management Journal*, Texas.
- Chung, H., Gray, M.,** 1999. Special Section: Data Mining, *Journal of Management Information Systems*, New York.
- Colin A., ve Journal D.,** 1996. *Building Decision Trees with the ID3 Algorithm*.
- Dönmez, Z., S.,** 2008. Bayi Performans Değerlendirmesinde Bir Veri Madenciliği Uygulaması, *Yüksek Lisans Tezi*, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.
- Dunham, M.H.,** 2003. *Data Mining Introductory and Advanced Topics*, Pearson Education Inc., New Jersey.
- Eker, H.,** 2009. Veri Madenciliği veya Bilgi Keşfi, <<http://www.bilgiyonetimi.org/cm/pages/mklgos.php?nt=538>>, alındığı tarih 05.03.2010.
- Ergezer, H., Dikmen, M. ve Özdemir, E.,** 2003. Yapay Sinir Ağları ve Tanıma Sistemleri, *PİVOLKA*, Ankara.
- Ericsson, R.,** 2004. *Building Business Intelligence Applications for NET*, Charles River Media, USA.
- Fausett L.,** 1994. *Fundamentals of Neural Networks*, Prentice-Hall, USA.
- Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.,** 1996. Advances in Knowledge Discovery and Data Mining, *MIT Press*, Cambridge.
- Fayyad U., Piatetsky-Shapiro G., Symth, P.,** From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, California.
- Giudici, P.,** 2003. *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons Ltd., England.
- Göral, M. A.,** 2007. Kredi Kartı Başvuru Aşamasında Sahtecilik Tespiti İçin Bir Veri Madenciliği Modeli, *Yüksek Lisans Tezi*, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.
- Grossman, R.L., Kamath, C., Kegelmeyer P., Kumar, V., Namburu R. R.,** 2001. *Data Mining For Scientific and Engineering Applications*, Kluwer Academic Publishers, Netherlands.
- Hsieh, N.,** 2004. An Integrated Data Mining and Behavioral Scoring Model for Analyzing Bank Customers, *Expert Systems with Applications*.
- Jack, L.B. and Nandi, A.K.,** 2004. Fault Detection Using Support Vector Machines and Artificial Neural Networks, Augmented By Genetic Algorithms, *Mechanical Systems and Signal Processing*.
- Kantardzic, M.,** 2003. *Data Mining: Concepts, Methods, and Algorithms*, John Wileys&Sons, New York.

- Kayaalp, K.**, 2007. Asenkron Motorlarda Veri Madenciliği İle Hata Tespiti, *Yüksek Lisans Tezi*, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, Isparta.
- Khan M.**, 2002. K-Nearest Neighbor Classification Spatial Data Streams Using P-Trees, 6. *Pasifik Asya Knowledge Discovery and Data Mining Konferansı*, Taiwan.
- Koyuncuğil A.S.**, 2006. Bulanık veri madenciliği ve sermaye piyasalarına uygulanması, *Doktora Tezi*, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.
- Kurt İ., Türe M.**, 2005. Yapay Sinir Ağları ile Lojistik Regresyon Analizi'nin Karşılaştırılması, *Doktora Tezi*, Trakya Üniversitesi Tıp Fakültesi, Edirne.
- Manish M.**, 1996. SLIQ: A Fast Scalable Classifier for Data Mining, 5. *Uluslararası Extending Database Technology Konferansı*, Avignon, Fransa
- Mannila H.**, 1994. *Efficient Algorithms for Discovering Association Rules*, AAAI Workshop on Knowledge Discovery in Databases.
- Maulik U., ve Sanghamitra B.**, 2000. Genetic Algorithm-based Clustering Technique, *Journal of the Pattern Recognition*, Pergamon
- Özbek O.**, 2009. Kuyumculuk Sektöründe Dünya Lideri Olabiliriz, *İzmir Ticaret Odası Ar&Ge Bülten*, İzmir.
- Özkan Y.**, 2008. *Veri Madenciliği Yöntemleri*, Papatya Yayıncılık, İstanbul
- Polat G., Altun H.**, 2007. Ses Öznetelik Gruplarının Duygu Tespitinde Etkinliklerinin Belirlenmesi, *IEEE 15. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, Eskişehir.
- Quinlan J. Ross**, 1986. Induction of Decision Trees, *Journal of Machine Learning*.
- Roiger, R.J. and Geatz, M.W.**, 2003. *Data Mining: A Tutorial-Based Primer*, Pearson Education Inc., USA.
- SAS**, 1998. Institute Inc. Data Mining And The Case For Sampling.
- Savasere A.**, 1995. An Efficient Algorithm for Mining Association Rules in Large Databases, 21. *International Conference on Very Large Databases*, İsviçre.
- Shannon, C.E.**, 1948. A Mathematical Theory of Communication, *The Bell System Technical Journal*, USA.
- Sibson R.**, 1973. An Optimally Efficient Algorithm for the Single Link Cluster Method, *The Computer Journal*.
- Silahtaroglu G.**, 2008. Kavram ve Algoritmalarıyla Temel Veri Madenciliği, *Papatya Yayıncılık*, İstanbul.
- Url-1** <<http://www.backpropagation.netfirms.com>>, alındığı tarih 08.04.2010.
- Ünsal A, Güler H**, 2005. Türk Bankacılık Sektörünün Lojistik Regresyon ve Diskriminant Analizi ile İncelenmesi, *VII. Ulusal Ekonometri ve İst. Sempozyumu*, İstanbul.

- Yaralıoğlu, K.,** 2008. Veri Madenciliği, http://www.deu.edu.tr/userweb/k.yaralioglu/dosyalar/ver_mad.doc, alındığı tarih 04/09/2009.
- Zhang T.,** 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases, *ACM International Conference on Management of Data*, USA.
- Zaine O. R.,** 1999. Principles of KDD, *Doktora Tezi*, University of Alberta Department of Computing Science, USA.

ÖZGEÇMİŞ

Ad Soyad: Songül ŞEKEROĞLU

Doğum Yeri ve Tarihi: Tarsus/ 25.09.1985

Lisans Üniversite: Fatih Üniversitesi Mühendislik Fakültesi Endüstri Mühendisliği Bölümü

Yayın Listesi: -

