T.C.

BEYKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MATEMATİK-BİLGİSAYAR ANABİLİM DALI

BİLGİSAYAR AĞLARI ve İNTERNET TEKNOLOJİLERİ BİLİM DALI

BİR TELEKOMİNİKASYON FİRMASINDA MÜŞTERİ SEGMENTASYONU

(Yüksek Lisans Tezi)

Tezi Hazırlayan: Emel SEYMEN TURAN

İSTANBUL, 2010

T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
MATEMATİK-BİLGİSAYAR ANABİLİM DALI
BİLGİSAYAR AĞLARI ve İNTERNET TEKNOLOJİLERİ BİLİM DALI

BİR TELEKOMİNİKASYON FİRMASINDA MÜŞTERİ SEGMENTASYONU

(Yüksek Lisans Tezi)

Tezi Hazırlayan:

Emel SEYMEN TURAN

Öğrenci No:

050861001

Danışman:

Yrd. Doç. Dr. Gökhan SİLAHTAROĞLU

İSTANBUL, 2010

YEMİN METNİ

Sunduğum Yüksek Lisans Projesi /Yüksek Lisans Tezimi, Akademik Etik İlkelerine bağlı kalarak, hiç kimseden akademik ilkelere aykırı bir yardım almaksızın bizzat kendimin hazırladığına and içerim.

19/01/2010

Emel SEYMEN TURAN

T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZLİ YÜKSEK LİSANS TEZ SINAV TUTANAĞI

02/02/2010

Enstitümüz ***Matematik-Bilgisayar*** Anabilim dalı ***Bilgisayar Ağları ve İnternet Teknolojileri*** Bilim dalı yüksek lisans öğrencilerinden 050861001 numaralı ***Emel Seymen TURAN'***ın *"Beykent Üniversitesi Lisansüstü Eğitim - Öğretim ve Sınav Yönetmeliği"*nin ilgili maddesine göre hazırlayarak, Enstitümüze teslim ettiği ***"BİR TELEKOMİNİKASYON FİRMASINDA MÜŞTERİ SEGMENTASYONU"*** tezi, Yönetim Kurulumuzun 25.01.2010 tarih ve 2010/02 sayılı toplantısında seçilen biz jüri üyeleri huzurunda, aday tarafından savunulmuş ve sonuçta adayın tezi hakkında ~~**oyçokluğu**~~***oybirliği*** ile ~~***Kabul/Red veya Düzeltme***~~ kararı verilmiştir.

DANIŞMAN
YRD.DOÇ.DR. Gökhan SİLAHTAROĞLU

ÜYE
YRD.DOÇ. Turhan KARAGÜLER

ÜYE
PROF. DR. Muhittin KARABULUT

BİR TELEKOMİNİKASYON FİRMASINDA MÜŞTERİ SEGMENTASYONU

Tezi Hazırlayan:Emel SEYMEN TURAN

ÖZET

Bu tez çalışmasında, günümüzde özellikle büyük ölçekli firmaların son dönemde sıklıkla kullanmaya başladığı veri ambarı ve veri madenciliği kavramları incelenip veri madenciliği yöntemleri ve teknikleri anlatılmıştır. Veri madenciliği basitçe verinin içinde gizlenen anlamları ortaya çıkarmak olarak tanımlanır. Veri madenciliği modelleri ile edinilen bilgiler, özellikle CRM (Customer Relationship Management) sistemlerinde sıklıkla kullanılmaktadır. Müşterilerin yapacak olduğu davranışlar hakkında tahminlerde bulunma, yapmış olduğu hareketlerle veya belirtmiş olduğu özelliklerle müşteri gruplandırma olarak nitelendirilebilecek Müşteri Segmentasyonu, veri madenciliğinin en sık rastlanan örnekleridir. Bu çalışmada, bir telekominikasyon firmasından örnek bir data seti alınarak bir veri ambarı kurulmuş, çeşitli modellemelerle segmentasyon örnekleri sunulmuş ve sonuçları anlatılmaya çalışılmıştır. Çıkan sonuçlar aslında aynı anda olması beklenilmeyen birçok örneği beraberinde getirmiştir. Bu elde edilen verilerle şirketin aleyhine sonuç veren çıkarımlar tespit edilip düzenlenecek olan kampanyalarla bu negatif etkenler düzeltilebilir.

Anahtar Kelimeler: Veri Ambarı, Veri madenciliği, CRM, Müşteri Segmentasyonu

CUSTOMER SEGMENTATION
IN A TELECOMMUNICATION FIRM

Emel SEYMEN TURAN

ABSTRACT

DataWare House and Data Mining are very important in business solutions. In big databases, analizing can be damping operational system performance. Datawarehouse and data mining are used for these solutions. We can explain data mining is finding new meanings in the data. Data mining uses two modellings; Predicting modellings and describing modellings. Data mining especially is used in CRM projects. For example predictting a cancellation prevents revenue reduction. And also classification and segmentation of customers is one of the popular applications. Customer characteristics are used in customer segmentation projects. By this way, it will be simple understanding customers without complex queries. In this project, Data WareHouse, Data Mining and segmentation algorithms were explained. It was developed a datawarehouse with the data of a telecommunication firm. The goal was segmentation of customers. According to the customer specialities three segmentation module was developed. With the results of modules, defining the customers was easy.

Key Words: Data Warehouse, Data Mining, CRM, Customer Segmentation

ÖNSÖZ

Veri ambarı ve veri madenciliği, son günlerde hızla gelişen ve her gün biraz daha fazla kullanılmaya başlayan bir konudur. Operasyonel sistemlerin zorlanmaması ve büyüklüğü açısından gerekli analizler için daha az hareket gören veri ambarları ortamları çok daha uygundur. Bu ortamlarda aynı zamanda, son kullanıcıların da rahatlıkla analizlerini yapabilmeleri açısından kullanımı da kolaylaştıracak veri madenciliği modellemeleri kullanılmaktadır. Bu çalışmada geldiğince bu konular anlatılıp örneklemeye çalışıldı.

Bu çalışmayı hazırlarken yardımlarını, fikirlerini esirgemeyen değerli hocam Yrd. Doç. Dr. Gökhan Silahtaroğlu'na teşekkür ve saygılarımı sunarım.

Ben çalışırken desteklerini, anlayışını esirgemeyen eşim Genç Osman ve kızım Aslı Duru'ya, kısıtlı zamanında yanımda olan arkadaşım Şeyda Alkan'a, kardeşim Melek Seymen'e, yazım süresinde anlayışını ve datalarını esirgemeyen müdürlerime ve tecrübelerinden faydalandığım Veri Madenciliği birim çalışanı arkadaşlarıma teşekkürü bir borç bilirim.

TABLOLAR

Sayfa No

ŞEKİLLER

Sayfa No

KISALTMALAR

CRM : Customer Relationship Management

DBMS : Database Management System

RDBMS : Relational Database Management

CVL : Comma Delimited File

ETL : Extract-Transform-Load

PSTN : Sabit Telefon

VTBK : Veri Tabanı Bilgi Keşfi

SQL **:** Structure Query Language

EBM : Enterprise Business Model

ODS : Operational Data Storage

EDW : Enterprise Data Warehouse

OLTP : Online Transaction Processing

OLAP : Online Analytical Processing

VA : Veri Ambarı

ODBC : Open Database Connection

ROLAP : Relational On-Line Analytical Process

MOLAP : Multidimensional On-Line Analytical Process

CRISP-DM : The Cross- Industry Standard Process for Data Mining

CART : Classification and Regression Trees

CHAID : Chi-Squared Automatic Interaction Detector

YSA : Yapay Sinir Ağları

AIS : Agrawal, Imielinski ve Swami

DHP : Direct Hashing and Pruning

SOM : Self Organizing Maps

TDWI : The Datawarehouse Insfracture

v.d. : ve diğerleri

b.t. : Bilinmeyen Tarih

İÇİNDEKİLER

GİRİŞ

Veri tabanlarının önemini, günümüzde artık iyice anlamakla beraber pazarının sürekli büyümesiyle birlikte ortaya öğrenilecek yeni kavramlar çıkmaktadır. Her türlü organi-zasyon büyük ya da küçük bir veritabanına sahip olmaya çalışmaktadır. Önceden sayfalarca dökümanlarla sakladığımız bilgiler artık gelişmiş teknoloji sayesinde çok daha kolay ulaşılır hale gelmiştir.

Veri tabanları, sadece yapılmış işlemleri değil o anda yapılacak olanları da tutmaktadır. Operasyonel sistemler diye isimlendirilen bu veri tabanları, özellikle büyük ölçekli firmalarda pekçok anlık işlemi başarıyla gerçekleştirmektedir. Veri tabanları sadece anlık işlemleri kayıt etmekle kalmaz aynı zamanda yapılmış işlemlerden rapor alınmasını hatta bundan sonra müşterinin yapacağı işler için bile tahminde bulunulmasına yardımcı olur. Fakat operasyonel sistemlerin çalıştığı veri tabanlarından rapor çekmek operasyonel veri tabanlarını zorlayacaktır. İşte buna engel olmak için veri ambarları denilen komplike yapılar oluşturulmuştur. Veri ambarları denilen bu yapılar, farklı kaynaklardan pekçok veriyi aynı ortamda buluşturarak operasyonel sistemleri yormadan çalışılmasını sağlar. Veri ambarları, operasyonel raporlara hitap edebildiği gibi veri madenciliği denilen çalışmalar için de kullanılmaktadır.

Veri madenciliği, elimizdeki mevcut veriyi kullanarak müşteriyi sınıflama, tanımlama veya müşteri hareketleri hakkında tahminde bulunma gibi çalışmalarda yardımcı olur. Örneğin, bir firma giderek geliri azalan müşterileri tespit ederek onların kullandığı hizmetleri belirleyerek, o müşterinin hareketlerine uygun kampanyalar geliştirip bunu müşteriye sunabilir. Veya çıkan herhangi bir kampanyadan kesinlikle alakası olmayan müşteriler ayırt edilebilir. Kaybedilen müşteri özellikleri tespit edilerek, yenilerinin kaybedilmesi engellenebilir.

Müşteri ilişkileri yönetimi ile veri madenciliği yanyana gitmektedir. Veri madenciliğinde bulunan sonuçlar kısaca CRM denilen Müşteri İlişkileri Yönetimi için kullanılabilir. CRM projeleriyle birlikte bir segmentasyona yönelik kampanyalar veya

call center arandığında müşterinin hareketi tahmin edilerek buna yönelik yeni bir öneri getirmek için kullanılabilir. Müşteri segmentasyonu denilen çalışmalar müşteriyi demografik, psikografik, coğrafik, davranışsal olarak kümelere ayırır. CRM çalışmalarında yapılan müşteri segmentasyonu (kümelemesi) ile müşteri her yönüyle tanınıp birkaç özellikten oluşmuş tanımlamalar, anlamlı cümleler oluşturulur. Bu tanımlara göre müşteri artı veya eksi yönleriyle belirlenerek negatif yönlerin düzeltilmesi için iyileştirici çalışmalar düzenlenebilir.

Bu çalışmada, bütün bu kavramlar incelendikten sonra CRM projelerinde kullanılabilecek müşteri tanımlamaları yapılması hedeflenmiştir. Amaç, klasik kümeleme yöntemlerinden farklı olarak kendi kendini düzenleyen haritalar olarak tanımlanan, denetsiz öğrenme kullanan KOHONEN algoritmasıyla müşteri segmentasyonu yapmaktır. Bu uygulama için müşteri farklı tablolardan alınacak pekçok özelliğiyle irdelenecek, veri anlaşılacak, müşteriyi abonesel, davranışsal ve genel olarak kümelemek için modellemeler yapılacaktır. Çıkan sonuçlar incelenerek, elde ettiğimiz anlamlı bilgilerin şirkete ne gibi bir faydası olacak, ne tür düzenlemeler yapılabilir diye belirlendikten sonra çalışma sonlanacaktır.

1. VERİ TABANI VE VERİ AMBARLARI

Başlıbaşına önemli bir konu olan veri tabanları ve veri ambarları, artık özellikle büyük ölçekli firmaların olmazsa olmazıdır. Müşteriler, satışlar, müşteri davranışları, ücretler, insan kaynakları gibi bilgiler hakkında sistematik bilgiler tutar. Veri ambarı ve veri madenciliği kavramları, veri tabanı kavramının gelişmesiyle doğmuştur. Birinci bölümde bu kavramlar incelenecektir.

1.1. VERİ TABANI

1.1.1. VERİ TABANI KAVRAMI

Veri tabanı, sistematik erişim imkanı olan, yönetilebilir, güncellenebilir, taşınabilir, birbirleri arasında tanımlı ilişkiler bulunabilen bilgiler kümesidir. Belirli bir amaca yönelik düzen verilmiş kayıt ve dosyaların tümüdür (İnternet Terimleri Sözlüğü, Anonim, b.t.).

Veri tabanı düzenli bilgiler topluluğudur. Kelimenin anlamı bilgisayar ortamında saklanan düzenli verilerle sınırlı olmamakla birlikte, daha çok bu anlamda kullanılmaktadır. Bilgisayar terminolojisinde, sistematik erişim imkanı olan, yönetilebilir, güncellenebilir, taşınabilir, birbirleri arasında tanımlı ilişkiler bulunabilen bilgiler kümesidir. Bir başka tanımı da, bir bilgisayarda sistematik şekilde saklanmış, programlarca istenebilecek veri yığınıdır.Bir veri tabanını oluşturmak, saklamak, çoğaltmak, güncellemek ve yönetmek için kullanılan programlara Veri Tabanı Yönetme Sistemi (DBMS) adı verilir (Usgurlu, b.t.).

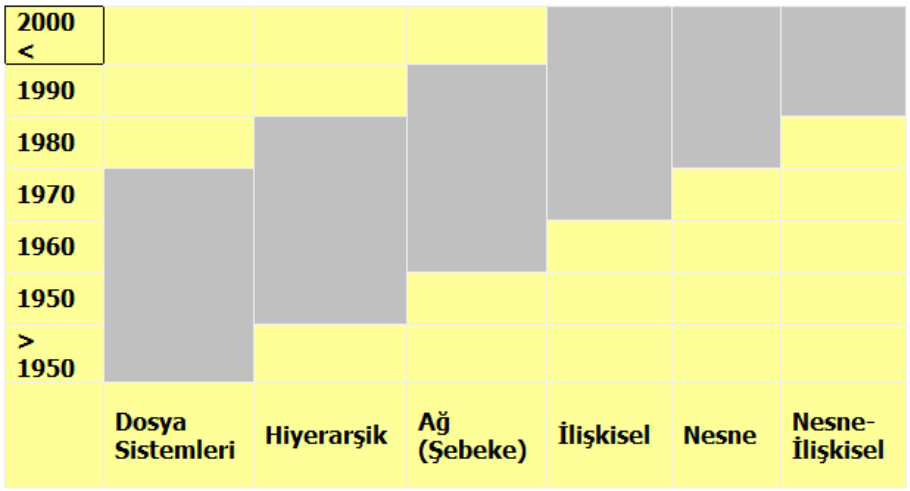
Veri Tabanında asıl önemli kavram, kayıt yığını ya da bilgi parçalarının tanımlanmasıdır. Bu tanıma Şema adı verilir. Şema veri tabanında kullanılacak bilgi tanımlarının nasıl modelleneceğini gösterir. Buna Veri Modeli (Data Model), yapılan işleme de Veri modelleme denir. En yaygın olanı, İlişkisel Model'dir (relational model). Layman'ın deyimiyle bu veriler tablolarda saklanır (Veri Tabanı(Database) Nedir? Anlamı?, Anonim,b.t.). Tablolarda bulunan satırlar (row) kayıtların kendisini, sütunlar (column) ise bu kayıtları oluşturan bilgi parçalarının ne türden olduklarını belirtir.

Veri tabanı yazılımı ise verileri sistematik bir biçimde depolayan yazılımlara verilen isimdir. Birçok yazılım bilgi depolayabilir ama aradaki fark, veritabanın bu bilgiyi verimli ve hızlı bir şekilde yönetip değiştirebilmesidir.

Veri tabanı, bilgi sisteminin kalbidir ve etkili kullanmakla değer kazanır. Bilgiye gerekli olduğu zaman ulaşabilmek esastır. İçeriği olmayan bir kütüphane ve bütün kitapların aynı kapağa sahip olduğunu düşündüğünüzde kütüphane kullanıcılarının ne kadar çok işi olacağını tahmin edersiniz. Bir veritabanı bir kütüphanenin mükemmel bir içerik sistemi olduğu gibi , aynı zamanda kütüphanenin kendisidir. Bağıntısal Veri Tabanı Yönetim Sistemleri (Relational Database Management Systems – RDBMS) büyük miktarlardaki verilerin güvenli bir şekilde tutulabildiği, bilgilere hızlı erişim imkanlarının sağlandığı, bilgilerin bütünlük içerisinde tutulabildiği ve birden fazla kullanıcıya aynı anda bilgiye erişim imkanının sağlandığı programlardır.

1.1.2. VERİ TABANI MODELLERİNİN GELİŞİMİ

Digital ortamın ilk yaygın olarak kullanılmaya başlanıldığı zamanlarda, dijital veri tabanı kavramı daha tam olarak yoktu. Veri tabanları yerine, verileri muhafaza etmek için düz-dosyalar (flat-files) kullanılıyordu. Sadece bu tip dosya türleri veri kaydı ve muhafazası yapmak için kullanıldığından, herhangi bir veri tabanı yapılanması o zamanlar mevcut değildi (Veri Tabanı Modelleri, Anonim, b.t.).



Şekil 1.1. Veri tabanı modellerinin gelişimi

Kaynak: Veri Tabanı Nedir? icinde. (18.11.2009) tarihinde http://www.netogretim.com/dokumangoster.aspx?id=177&d=Veri-Taban%C4%B1-Modelleri ' nden alındı.

1.1.2.1. Dosya Sistemleri Modeli (File Systems)

Veri tabanı modeli olarak dosya sistemlerini kullanmak, aslında bir veri tabanı modelleme tekniğinin kullanılmadığını belirtir. Böyle bir sistemde veriler flat-files olarak bilinen düz dosyalara atılır. Düz-dosya terimi ise, hiçbir format taşımayan bir text dosyasını tanımlamak için kullanılır.

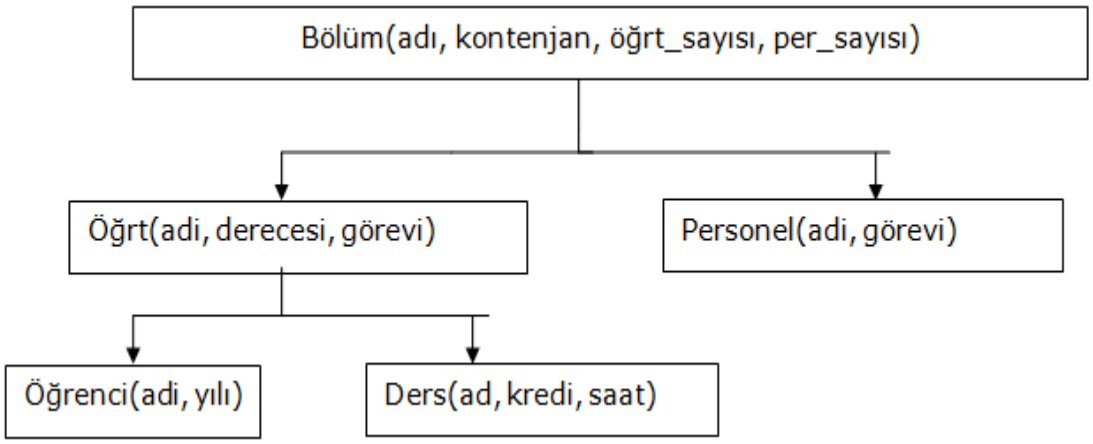
Comma delimited file (CVL) dosyaları bir yapıya sahiptirler, çünkü virgüller kullanılmaktadır. Bu dosya türleri düz dosya olarak bilinse de, geçmişte düz-dosya veri tabanları çok uzun yazılar muhafaza ediyorlardı ve bu dosyalar tek bir satırdan oluşuyor ve virgül işareti kabul etmiyorlardı. Veriler, dosya içerisindeki yerinden bulunuyorladı. Tüm bu söylenenleri göz önünde bulundurduğumuzda, Excel'de kullanılan CSV dosya türlerini düz dosya olarak tanımlamamız yanlıştır.

Text dosyalar'da veri araması yapmak için, bu işlevi belirgin bir şekilde programlamak gerekir. Bu sistemde, veriler birden fazla dosya'ya kaydedilebilinir. Fakat bu dosyalar arasındaki işlemler de belirgin bir şekilde programlanmalıdır (Veri Tabanı Modelleri, Anonim, b.t.).

1.1.2.2. Hiyerarşik Veri-tabanı Modeli (Hierarchical Database Model)

Hiyerarşik veri-tabanı modeli bir ağaç yapısına sahiptir. Bu tip veri tabanları içerisinde bulunan tablolar, child-parent ilişkisine sahiptir, ve her parent tablo birden fazla child tabloya sahip olabilir. Buna ek olarak, child tablosuna herhangi bir veri eklenirken, parent tablosunda bu veriye tekabül eden veriler bulunması gerekir. Sonuç olarak bu veri tabanı modeli one-to-many ilişkisini desteklemektedir (Veri Tabanı Modelleri, Anonim, b.t.).

Bu veri tabanı modelinin dezavantajı ise, herhangi bir arama kök tablodan başlamalıdır. Yani herhangi bir child tablosundaki verileri bulabilmek için ilk önce parent tablosundaki, o child tablosuna ait verileri bulmak gerekmektedir.



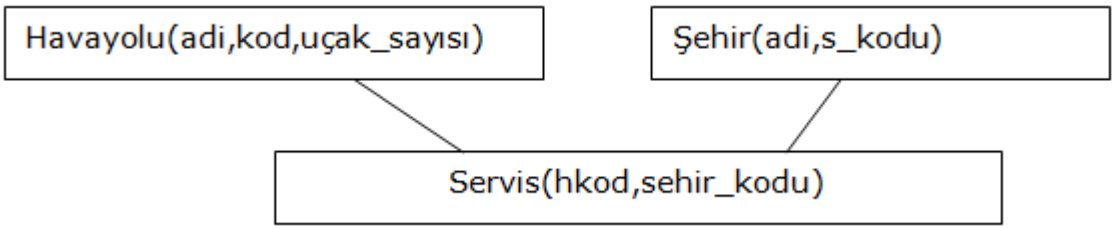
Şekil 1.2. Hiyerarşik ağ modeli

Kaynak: Veri tabanı içinde.(15.11.2009) tarihinde http://www.baskent.edu.tr/~eminec/bahar/veri_word.doc' den alınmıştır.

1.1.2.3. Şebeke Veri-Tabanı Modeli (Network Database Model)

Şebeke Veri-Tabanı Modeli esasında hiyerarşik veri tabanı modelinin geliştirilmiş bir versiyonudur. Network veri tabanı modeli child tabloların birden fazla atalarının olmasına müsade etmektedir. Niketim ortaya tablolar arasında kurulan bir şebeke çıkmaktadır. Buradaki ilişki türü many-to-many 'dir.

Bu veri tabanı modeli hiyerarşik veri tabanı modelinden çok daha esnek bir yapıya sahiptir.



Şekil 1.3. Ağ Modeli

Kaynak: Veri tabanı içinde.(15.11.2009) tarihinde http://www.baskent.edu.tr/~eminec/bahar/veri_word.doc' den alınmıştır.

1.1.2.4. İlişkisel Veri-Tabanı Modeli (Relational Database Model)

İlişkisel veri tabanı modeli, hiyerarşik modeldeki kısıtlamaya neden olan maddeleri, hiyerarşik yapıyı tamamen terk eden, elde eden bir veri tabanı modelidir. Bu modelde herhangi bir tablo üzerinde, ilk olarak parent tablo seçilmeden, arama gerçekleşebilir. Bu yapabilmeyi sağlayan püf nokta ise, arama yapacağımız veri hakkındaki birkaç detayı önceden bilmemiz gerekmesidir (Veri Tabanı Modelleri, Anonim, b.t.).

Bu modelin sunduğu bir başka avantaj ise, tabloların hiyerarşik durumu ne olursa olsun, herhangi iki tablo arasında ilişki kurulabilmesidir. Yani herhangi bir tablo, birden fazla parent tabloya, ve aynı şekilde birden fazla child tabloya ilişkilendirilebilir.

İlişkisel veri tabanı yönetim sistemi verilerin tablolarda satır ve sütunlar halinde tutulduğu ve yüksek bir veri tutarlılığına sahip veri depolama sistemidir.

İlişkisel veri tabanını çeşitli tablolar arasında organize edilmiş verilerden oluşan veri tabanı olarak açıklayabiliriz. Bu farklı tablolar arasındaki veriler, çeşitli anahtarlar vasıtası ile birbirlerine bağlanırlar. İlgili tablolarda, sütunlar arasında bir anahtar sütun yeralır. Bu anahtar sütun aracılığı ile birden çok tablo verileri birbiriyle bağlantı sağlayabilir ve herhangi bir sorgulamada birlikte görüntülenebilir. Bu tür veri tabanları arasında Mysql, Oracle, dBase, Progress, Informix, Ingres,...vb. gelmektedir.

Günümüzde en çok kullanılan veri tabanı modeli olmakla beraber, en başarılı veri tabanı modelidir.

Bolum

ogr_kod	adi	kod
1	Ahmet	INO
2	Ali	OTO
3	Ayse	SNO

Ogrenci

kod	adi
OTO	Bilgisayar
INO	İngilizce
SNO	Sınıf

Şekil 1.4. İlişkisel Veritabanı Modeli

Kaynak: Veri Tabanı içinde.(15.11.2009) tarihinde http://www.baskent.edu.tr/~eminec/bahar/veri_word.doc' den alınmıştır.

1.2.2.5. Nesne Veri-Tabanı Modeli (Object Database Model)

Nesne veri-tabanı modeli, verilerin herhangi bir noktadan çok kolayca alınabileceği, üç boyutlu bir yapıdan oluşur. İlişkisel veri tabanı verileri iki boyutlu tablolar halinde getirirken, nesne modelinde veriler tek parça olarak gelirler. Dolayısı ile birden fazla veri dönmesi arzulandığında nesne modeli performans olarak çok iyi değildir.

Fakat nesne veri-tabanı modeli, ilişkisel veri tabanı modelindeki bir iki sorunu çözmektedir. Bunlardan bir tanesi, bu veri tabanı modelinde, türlerin kullanılmasına gerek olmamasıdır.

Nesne veri-tabanı modelinin bir başka avantajı ise, çok komplex bir yapıya sahip olan büyük veri tabanı tasarımını kolaylaştırmasıdır. Bunu, nesne yöntem bilimin prensiplerine uygun olarak tasarlanmış bir model olmasından kaynaklanır.

1.1.2.6. Nesne-İlişkisel Veri-tabanı Modeli (Object-Relational Database Model)

Nesne-İlişkisel Veri-tabanı Modeli küresel bir yapıya sahiptir. Veri tabanı üzerindeki herhangi bir veriye, yüksek performansta erişim sağlar. Fakat yine de birden fazla veri istenildiği zaman bu modelde de veri tabanı performansı çok kötü bir darbe

alır (Veri Tabanı Modelleri, Anonim, b.t.). Bu veri tabanı modeli ilişkisel ve nesne veri tabanı modellerini bir şekilde aynı çatı altına almak için oluşturuldu.

1.1.3. İLİŞKİSEL VERİ TABANI YÖNETİM SİSTEMİ (RDBMS)

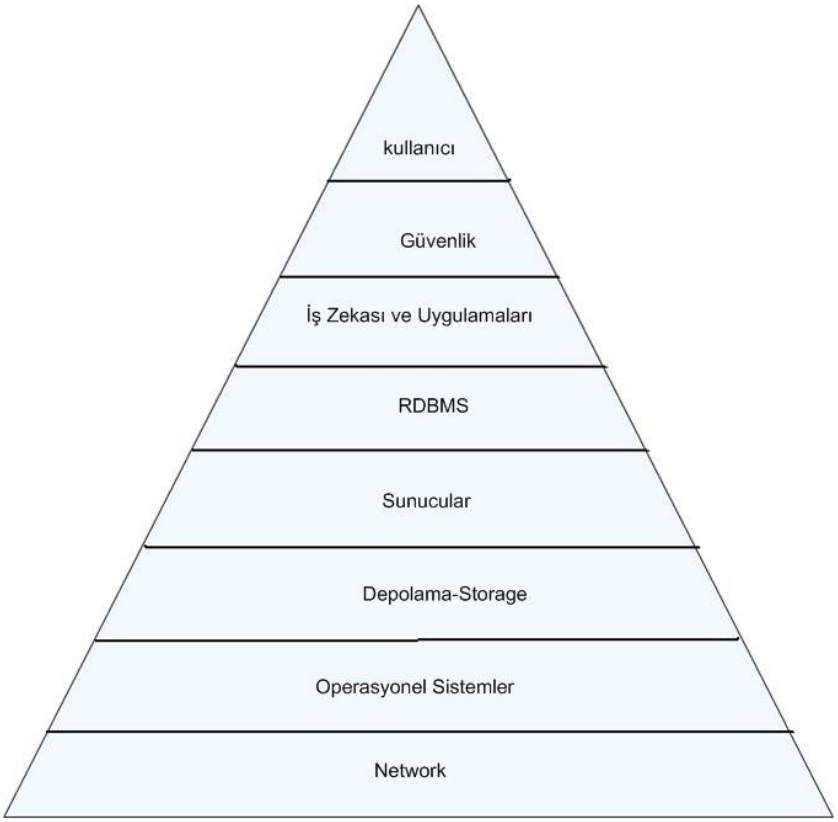
Bir veritabanı, bir kütüphanenin mükemmel bir içerik sistemi olduğu gibi , aynı zamanda kütüphanenin kendisidir. Bağıntısal Veri Tabanı Yönetim Sistemleri (Relational Database Management Systems – RDBMS) büyük miktarlardaki verilerin güvenli bir şekilde tutulabildiği, bilgilere hızlı erişim imkanlarının sağlandığı, bilgilerin bütünlük içerisinde tutulabildiği ve birden fazla kullanıcıya aynı anda bilgiye erişim imkanının sağlandığı programlardır.

Bir RDBMS aşağıdaki işlemlerden sorumludur;

• Bir veri tabanındaki veriler arasında ilişkiler kurmak

• Verileri hatasız bir şekilde saklamak ve veriler arasında tanımlanan ilişkileri bozmamak

• Bir sistem hatası durumunda tüm verileri kurtarabilmek



Şekil 1.5. Bir RDBMS şekli

Kaynak: Mark Madsen, (05.05.2009). TDWI 2008 Learning Book ' dan alınmıştır.

1.1.4. GÜNÜMÜZDE KULLANILAN VERİ SAKLAMA MODELLERİ

Günümüzde kullanılan veri tabanı modelleri ikiye ayrılmaktadır:

1.İşlemsel Veri Tabanları

2.Analitik Veri Tabanları

1.1.4.1. İŞLEMSEL VERİ TABANI MODELLERİ

İşlemsel veri tabanı modelleri, günümüzdeki birçok büyük kuruluşun ilk seçimidir. Bu tür veri tabanları, Online Transaction Processing (OLTP) senaryolarında ideal bir veri tabanı modelidir. OLTP veri tabanları, sürekli yeni veriler eklenen, verilerin modifiyesi ve müdafası yapılan durumlar için oluşturulurlar. Bu tür veri tabanlarında bulunan

veriler dinamik bir yapıya sahiptirler (sürekli değişmektedirler). Firmaların operasyonel işlemlerinde kullanılan veri tabanları bu formattadırlar.

Veriler genellikle ilişkisel tablolar içinde organize edilir. Gereksiz veri yığınları azaltır ve veri güncelleme hızını arttırır. Örnek olarak bir telekominikasyon firmasındaki ürünlerin satış ve müşteri bilgilerini içerir.

1.1.4.2. ANALİTİK VERİ TABANI MODELLERİ

Analitik veri tabanı modelleri ise, Online Analytical Processing (OLAP) için kullanılmaktadır. Bu tür veri tabanları, uzun bir zaman kemerindeki yüksek miktarda veri zamana bağlı verileri depolamak için kullanılır. Kullanılan veri türleri ya çok az değişikliğe uğrar ya da statiktir.

OLAP teknolojisi büyük verilerin organize edilmesi ve incelenmesini sağlar. Örneğin bir analist büyük verileri hızlı ve gerçek zamanlı olarak değerlendirebilir. Örneğin SQL Server Analiz Servisi toplu raporlama ve analizde, veri modelleme ve karar desteğe kadar geniş alanda çözümler sunar. Bu konular, ilerleyen bölümlerde biraz daha inceleynecektir.

1.2.VERİ AMBARI

1.2.1. VERİ AMBARI NEDİR?

Kurum yöneticileri için karar verirken doğru ve zamanında bilgi önemlidir. Bu bilgi gerçekte kurumun işleyişi sırasında toplanan verilerde mevcuttur. Karar destek sistemleri, kurum içi ve dışı verilerin, karar verme sürecinde kullanılabilecek bilgiye dönüştürülmesiyle ilgilenir.

Bir kuruma ait veri, değişik kaynaklarda bulunabilir. Bunların kolay ulaşım için tek bir havuzda toplanması istenir. Ayrıca kurumun operasyonel işlemlerini gerçekleştirdiği OLTP (OnLine Transaction Processing) sistemler (veritabanları) bilgi toplama üzerine (kayıt ekleme, çıkarma, silme gibi hareketler/Transaction) uzmanlaşmıştır. OLTP

sistemlerin karar destek faaliyetlerinde kullanılması performans açısından tavsiye edilmez. OLTP sistemlerden, karar destek faaliyetlerinde kullanılacak verilerin, (denormalizasyon gibi) performans kazandırıcı değişimlerden sonra, bu tek havuza toplanması gerekir. OLTP sistemlerde, verilerin geçmiş halleri tutulmayabilir. Aslında karar verme açısından verideki değişim, yani verinin tarihsel değişimi de önemlidir. Bunların da bu tek havuzda tutulması gerekecektir. İşte bu ihtiyaçlara cevap vermek için oluşturulan bu havuza veri ambarı (data warehouse) diyoruz. Veri ambarındaki veriler, daha çok karar destek sistemleri gibi yönetsel uygula-malar, veri madenciliği ve "Çevrimiçi Analitik İşlemler de (OnLine Analytical Proces-sing-OLAP) kullanılan, zaman boyutlu, değişken olmayan , ayrıntıdan arınmış verilerdir.

Veriler, operasyonel sistemlerden ve diğer kaynaklardan çekilerek veri ambarı dediğimiz ortamlara atılır. Verilerin veri ambarına aktarılması ETL (Extract Transform Load) olarak bilinir. Bu iş için çeşitli araçlar kullanılabilir. .

Veri Madenciliği, sık sık veri ambarlarıyla karıştırılmaktadır. En basit anlamda Veri Madenciliği ve Veri Ambarları, birbirlerinin tamamlayıcısıdır. Veri ambarları, verinin belli bir yapıda saklanması için kurulurken, veri madenciliği bu saklanan verinin bilgiye dönüştürülmesini sağlar. Kısaca veri ambarları, Veri Madenciliğinin omurgası gibidir(Harrold,D. 2000, s.9-10).

Veri ambarı (VA) için yapılmış pek çok tanım bulunmaktadır. Bunlardan bazıları şöyledir:

VA, iş zekasını ve karar verme sürecinin yönetimini desteklemek amacıyla kullanılan konuya yönelik, entegre, zamana bağlı ve kalıcı veri kümesi şeklinde tanımlanmıştır(Tezcanlar,2007).

Organizasyonların etkin bir biçimde yönetilmesi için gerekli verilerin toplandigi ve analiz sonucu bilgilerin elde edildiği ortamlar olan veri ambarlarının en önemli amaci; karar verme sürecinde kullanılmak üzere, iç ve dış kaynaklardan elde edilecek verinin toplanmasi, birleştirilmesi, dönüştürülmesi ve yorumlanmasini sağlamaktir(March ve Hevner, 2005, s.1).

Veri ambarı, organizasyonlar icin problem çözmede, sorgu oluşturma ve analiz yapmada kullanılan etkili bir araç olarak tanımlanmaktadir. Veri ambarları toplanan veriyi uyumsuzlukları giderildikten sonra, işletme uygulamasina yönelik olarak dönüştürülmesi, özetlenmesi ve bir veri tabanına yüklenmesini ifade eder. Ayrıca son kullanıcının veriye ulaşmasi ve analiz yapabilmesi için gerekli araçları ve uygulamaları da kapsamaktadir (Massa ve Testa, 2005, s. 709-718).

VA, işletme bölümünde farklı kaynaklardan toplanan verinin depolandığı, organize edildiği ve karar vericilerin sahip oldukları platform ve teknik kabiliyet seviyesine bağlı olmaksızın ulaştırıldığı karar destek ortamı olarak da tanımlanır (Singh, 1997, s.11). Yapılan bir başka tanımda VA; operasyonel sistemlerden özetlenerek veya toplu olarak alınan verilerin basitleştirilmiş biçimde saklandığı yer olarak tanımlanmıştır(Gray ve Watson,1998, s.8-9).

Günümüzde veri tabanlarının çok farklı kaynaklarda bulunması, çok büyük hacimlerde veriler içermesi ve farklı yapılara sahip olması dikkate alınırsa VA, bu zorlukların üstesinden gelmek amacıyla normal veri tabanlarından farklı olarak, analiz ve raporlama işlemlerinde kullanılmak üzere hazırlanmış verileri içermektedir. Geniş hacimli, dağınık ve farklı yapılara sahip çok sayıdaki veri tabanları ve diğer enformasyon kaynaklarındaki verilerin entegre edilmesine yönelik, gelişmiş bir yaklaşım olarak ifade edilecek veri ambarcılığı yaklaşımında, her kaynaktan alınan enformasyon, ileri aşamalarda, gerektiği şekilde süzülerek ve ilgili enformasyon ile birleştirilerek veri ambarına yüklenir. Veri ambarında sorgular orijinal veri kaynağına erişim olmaksızın yerel olarak cevaplanabilir. Ayrıca, yüksek sorgu performansı, derinlemesine analiz, veri madenciliği ve karar destek için gerekli olan karmaşık veri analizleri ile elde edilebilmektedir (Theodoratos ve Sellis, 1999, s 279-301). Karar vericilerin daha kolay ve daha doğru kararlar almalarını sağlamak için operasyonel sistemlerde depolanan verilerin farklı bir ortama alınması ve analizlerin bu ortam üzerinde yapılması bu tanımda öne çıkmaktadır. Operasyonel sistemler isimlerinden de anlaşılacakları gibi, işletmelerin günlük işlerini yürütmelerine yardımcı olurlar. Bunlar, işletmenin omurgasını oluşturan sistemlerdir.

Birçok organizasyon veri ambarlarından çeşitli araçlarla elde ettiği bilgileri, karar alma süreçlerini desteklemek amacıyla kullanır. Bunlara ek olarak veri ambarlarının diğer kullanım amaçları şunlardır(Han ve Kamber, 2000):

-Müşterilerin satın alma şekilleri üzerinde analizler yaparak satın alma alışkanlıklarını ortaya çıkarmak.

-Üretim stratejilerinde ince ayarlar yapmak için; satış performanslarını; yıllar, coğrafik bölgeler vb. gibi değişkenlere göre karşılaştırarak ürün portföylerini yönetmek.

-İşlemlerin analizi ve yeni kar alanları yaratmak.

-Müşteri ilişkilerini yönetmek.

Veri ambarları, operasyonel sistemde biriken verilerin değerlendirilmesi için kullanılmaktadır. Veri ambarlarından alınan gecmişe iliskin sonuçlar geleceğe yönelik ipuçları vereceğinden, karar mekanizmasında kullanılarak kurumun gelişiminde etkin rol oynamaktadır. Bu nedenle sayılarla, yüzdelerle, kesin değerlerle kararların alındiği günümüzde özellikle işlem hacmi büyük kurumlar için veri ambarinin önemi ortadadir (Veri Ambarı Nedir?, Anonim, b.t.). Ülkemizde de veri ambarı uygulamaları bu nedenlerden dolayı sıkça kullanılmaya başlanmıştır, gelecek günlerde de bu tip uygulamaların artması beklenmektedir.

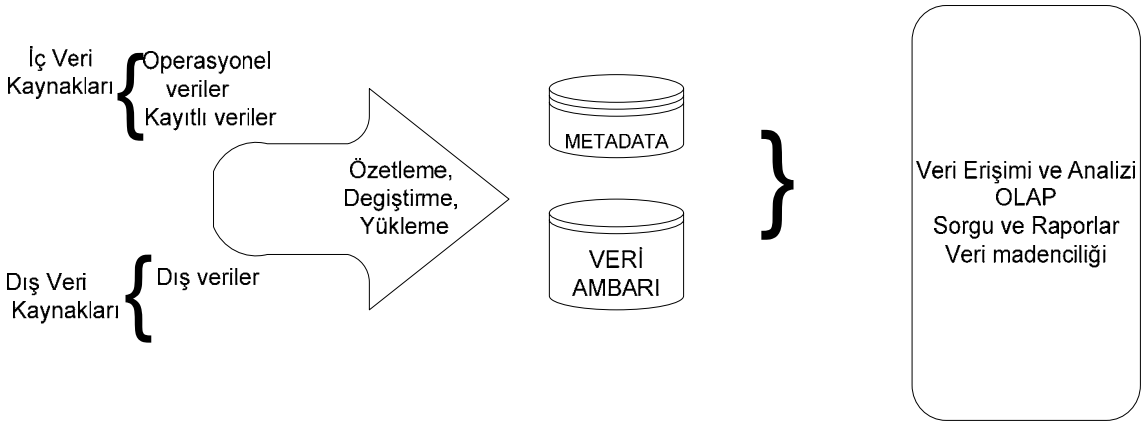
Son yıllarda perakendecilik sektöründe yer alan firmalar için operasyonel verimlilik ve müşteriyi elde tutma konuları oldukça önemli hale gelmiştir. Hızlı biçimde müşteri sadakati uygulamaları başlatılmıştır. Sektörde artan mağaza sayısı, müşteri sayısı, satış adetleri ve cirolar Türkiye'de perakendecilik sektörünü, veri ambarcılığından en çok yararlanabilecek sektörlerden biri haline getirmektedir. Türkiye'deki ilk kurumsal veri ambarı uygulamaları Migros (1997) ve Çarşı (1998) mağazalarında başlatılmıştır (Aytekin, 2002, s.181-190).

1.2.2. VERİ AMBARININ BİLEŞENLERİ

Veri ambarlarında, birbirleriyle ilişkili veriler sorgulanabilmekte ve analiz yapılabilmektedir. Genel olarak veri ambarları iki amaç etrafında çalışmaktadır.

1. Veri ambarları depo görevi görmektedir ve analitik stratejik verilerin birikimini sağlamaktadır. Bu veriler daha sonra yeniden kullanılmak üzere arşivlenmektedir. Veri ambarları verilerin sorgulanabildiği ve analizlerin yapılabildiği depolardır.

2. Veri ambarlarının pazarda yeni fırsatlar bulmaya, rekabete, iş, envanter ve ürün maliyetlerinin azalmasına katkılarının yanı sıra farklı işlere ait verilerin ilişkilendirilmesi, karar destek sistemlerine ve alınan bilgiye hızlı cevap verebilme gibi pek çok katkısı bulunmaktadır



Şekil 1.6. Veri Ambarının Bileşenleri

Kaynak: Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama içinden alınmıştır.

Pek çok organizasyonda çok büyük veri tabanları bulunmaktadır ve bu veri tabanları normal günlük bilgileri içermektedir. Bunlar "Operasyonel Veri tabanı" olarak adlandırılır. Operasyonel veri tabanları çoğu kez tarihsel veriyi saklamazlar. Günlük bilgilerle ilgili uygulamaları ve kısa vadeli taktik kararların verilmesini desteklerler. Günlük bilgiler stok yönetimi, rezervasyon yönetimi, mağazacılıkta satış yönetimi gibi alanlarda kullanılabilmektedir.

Organizasyonlarda bulunan ikinci veri tabanı tipi de veri ambarıdır. Veri ambarları stratejik kararların alınması için kurulmuşlardır. Veri ambarlarının en belirgin özelliği, çok büyük miktarda, belki milyarlarca kaydın söz konusu olduğu veriyi içermesidir. Veri ambarlarının amacı, yapısı ve çalışma biçimi alışılmış veri tabanlarından oldukça farklıdır.

Veri ambarları diğer çeşit veri tabanlarından kapasite olarak daha büyüktür. Karmaşık sorgular içeren "Çevrimiçi Hareket İşlemleri" (On-line Transaction Processing-OLTP) uygulamalarından farklı olan veri ambarı uygulamaları, farklı veri tabanı yönetim sistemleri tasarımı ve uygulama tekniklerinden iyi sonuçlar elde etmede kullanılır. Bu güçlük ve zorlukları yenmek için veri ambarları mimarisi oluşturulmaktadır. Geleneksel veri tabanlarında daha çok işletimsel veriler depolanır ve veriler üzerinde daha çok çevrimiçi hareket işlemleri gerçekleştirilir. Veri ambarlarının kullanım amaçları, içerdikleri verilerin nitelikleri ve veriler üzerinde yapılan işlemler açısından veri tabanlarından farklı olduğundan, veri ambarları için kullanılan yapılar, modeller ve teknikler de veri tabanları için kullanılanlardan oldukça farklıdır. Veri ambarındaki veriler, daha çok karar destek sistemleri gibi yönetsel uygulamalar, veri madenciliği ve "Çevrimiçi Analitik İşlemler"de (On-Line Analytical Processing-OLAP) kullanılan, zaman boyutlu, değişken olmayan, ayrıntılardan arınmış verilerdir(Yarımağan, 2000, s.291).

Veri ambarı, bir organizyonun çeşitli birimlerinin toplandığı ortak bir veritabanıdır. İlişkisel verilerin sorgulanabildiği ve analizlerin yapılabildiği bir veri deposudur. Veri ambarı, çeşitli veri kaynaklarından oluştuğu için bütün data aynı formatta olmayabilir. Bu yüzden veri ambarını oluştururken

-veri temizleme,

-veri dönüştürme,

-veri yenileme,

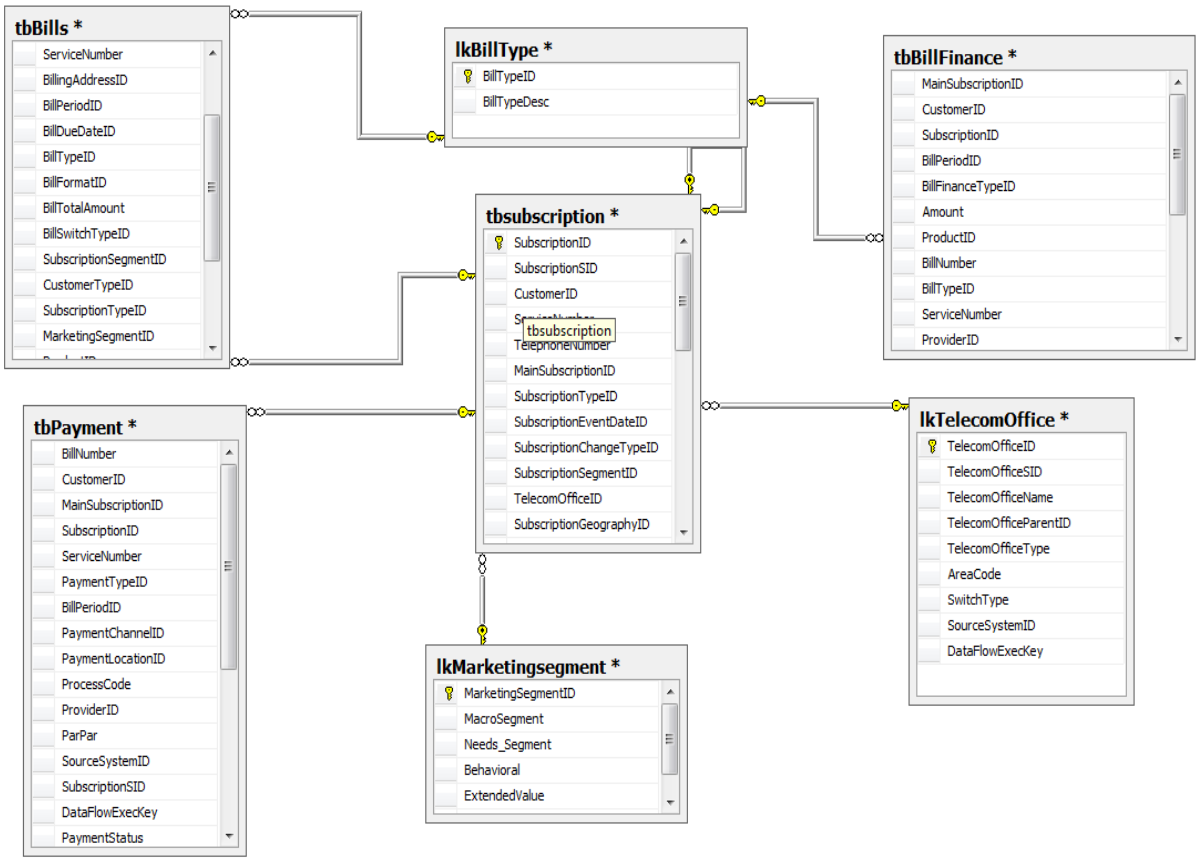
-veri yükleme işlemleri yapılır.

Veri pazarı (Data Mart); İstenilen araştırmaya uygun hale getirilmiş veri tabanına veri pazarı denir. Veri ambarı organizasyonda yer alan herşeyi içerdiğinden, seçilen nesnelere odaklanabilmek için veripazarı kullanılır.

Veri madenciliği(Data Mining); veriler arasındaki ilişkiler, etkiler, sapma ve eğilimler gibi bilgilerin ortaya çıkmasına yardım eden süreçtir. Tüme varım bir süreçtir.

OLAP; farklı seviyelere gore farklı süreçler çıkartan süreçtir. OLAP da yargı başta bellidir. Daha önceden oluşturulmuş hipotezler doğrulanmaya çalışılır. Tümden gelim bir yöntemdir.

Veri madenciliği ve veri ambarı birbirini tamamlar. Veri madenciliği ile oluşturulan hipotez OLAP ile test edilir.



Şekil 1.7. Bir veri ambarı diyagramı örneği (Bu diyagram, uygulamada kullandığımız veri ambarından alınmıştır.)

1.2.3. VERİ AMBARININ ÖZELLİKLERİ

Veri ambarı çok boyutlu modellemede depolama amaçlı kullanılan, farklı kaynaklardan gelen verilerin birleştirildiği ortamlardır. Zaman serileri ve trend analizi gibi tarihsel bilgi gerektiren yapıları desteklemektedirler. Veri ambarında bulunan veriler seyrek olarak değiştirilmektedir ve periyodik olarak güncelleme yapılmaktadır (Yıldız,2005). Inmon, "Buil-ding the Data Warehouse" (Veri Ambarı Kurmak) adlı kitabında bir veri ambarının taşıması gereken özellikleri aşağıdaki gibi sıralamıştır (2002, s. 26-27):

1. Verinin Zamana Bağlı Olması (Time-variant)

2. Verinin Kalıcı Olması (Non-volatile)

3. Veri Ambarının Konuya Yönelik Olması (Subject-oriented)

4. Verinin Entegre Edilmiş Olması (Integrated)

1.2.3.1. Veri Ambarındaki Verilerin Zamana Bağlı Olması

Veriler, geçmişten bilgi sağlamak için depolanır. Zamana bağımlı olması, verinin zamanda bir nokta ile ilişkili olması anlamına gelmektedir (Singh, 1998, s.13).

Operasyonel sistemlerde veri erişildiği anda geçerlidir. Birkaç saniye içinde yapılan işlemler nedeniyle veri geçerliliğini kaybedebilir. Enformasyonel verinin zaman boyutu vardır ve veri noktaları bu eksen boyunca karşIlaştırılabilir. Operasyonel sistemlerde veri erişildiği anda eksiksiz olmalıdır. Veri ambarında tutulan veri ise zamanın belirli bir anı için eksiksizdir. Genel olarak veri, veri ambarına yüklendiği sırada tam değildir. Zamana bağımlı olma veri am-barlarında zaman serileri analizinin yapılmasına olanak sağlar.

1.2.3.2. Verilerin Kalıcı Olması

Değişmeyen veri özelliği (nonvolatile); veri ambarına girilen verinin değiştirilemez olması demektir. Veri ambarının amacı varolan bilgiyi analiz imkanı sağlamak olduğundan, bu özellik oldukça mantıklıdır. Veri ambarına aktarılan yeni veriler, veri am-

barında mevcut bulunan verilerin güncellenmesi için kullanılamaz. Bu yüzden veri ambarındaki veriler değiştirilmemeli, güncellenememelidir. Bir veri ambarı genelde verinin yüklenmesi ve veriye ulaşma gibi iki işlem barındırır. (Han ve Kamber, 2000, s.54).

Operasyonel sistemlerdeki veriler güncellenip, temizlenip, entegre edildikten ve bütünleştirildikten sonra veri ambarına aktarılırlar. Veriler son şekillerini almadan veri ambarına aktarılmazlar. Operasyonel sistemlerde veri kayıtlarında sürekli olarak anlık güncellemeler (ekleme, silme, değiştirme gibi) yapılırken, veri ambarlarında veri yükleme belirli zaman aralıklarında yapılır (Gray ve Watson, 1998, s.14).

1.2.3.3. Veri Ambarının Konuya Yönelik Olması

Operasyonel sistemler için gereken veri, uygulamaların anlık ihtiyaçlarıyla ilgilidir ve mevcut iş kurallarına dayanır. Veri ambarı ise karar vermeye yönelik verileri içerir(Gray ve Watson, 1998, s.10-11). Veri ambarında veriler, işletmenin günlük işlemleri yerine müşteri, tedarikçi, ürün ve satış gibi ana konular üzerinde kurulmuştur. Veri ambarı, model üzerine odaklanır ve karar vericiye analiz için daha geniş olanaklar sağlar.

1.2.3.4. Veri Ambarlarının Entegre Olabilme Özelliği

Bir veri ambarı; ilişkisel veri tabanları, veri dosyaları ve çevrim içi hareket (on line transaction) kayıtları gibi ayrı ayrı birçok kaynakla bütünleşmeye olanak sağlamalıdır. Zira analiz edilecek bilginin kaynağı bunlardır. Birçok kaynaktan veri toplamak çok önemlidir, çünkü bir veri ambarının faydası kapsadığı veriden gelir.

Veri ambarlarına veri, birçok kaynaktan gelebilmektedir. Her kaynakta veriler çok çeşitli özelliklere sahip olabilmektedir ve bu özellikler kaynaklara göre farklılık gösterebilmektedir. Veri ambarlarına operasyonel kaynaklardan veya diğer dışsal veri kaynaklarından veriler aktarılırken veri entegre edilir (bütünleştirilir) ve aynı formata getirilir (Meyer ve Cannon, 1998, s.20-21). Verilerin isimlendirilmesinde, ölçü birimlerinin belirlenmesinde, kaynaklarda bulunan farklılıkların, tutarsızlıkların giderilmesi ve verilerin benzerliğinin sağlanması gerekmektedir.

Operasyonel sistemlerde geliştirilen pek çok uygulamada aynı veri birçok değişik şekilde gösterilmiş ve kodlanmış olabilmektedir. Cinsiyet, bir uygulamada bay-bayan şeklinde gösterilmişken, bir diğerinde E-K. şeklinde diğerinde ise 0-1 şeklinde gösterilebilmektedir.

1.2.4.VERİ AMBARI MİMARİSİ

1.2.4.1.FİZİKSEL MİMARİ

Pek çok organizasyon veri ambarı mimarisi ve yapısı arasındaki farkı ortaya koyamamaktadır. Veri ambarı mimarisi, uygulamadaki her bileşenin fonksiyonunu ve sorumluluklarını belirtir. Veri ambarı yapısı ise, her bileşene uygulanacak yazılım ve donanımı belirtir. Bu bileşenler aşağıda sıralanmıştır:

- Operasyonel veriyi veri ambarına transfer eden bileşenler,

- Veriyi yöneten ve veri ambarında depolayan bileşenler,

- Karar verme sürecinde kullanılacak verinin girişini, analizini ve sunumunu gerçekleştirmeyi sağlayacak olan bileşenler.

Bu bileşenlerden yola çıkarak veri ambarı mimarisinde karşımıza aşağıdaki kavramlar çıkacaktır.

Girişimci Veri Deposu (EDW): Dış veri sağlayıcıları ya da bir veya daha fazla operasyonel sistemden alınan, bütünleştirilmiş konu yönlendirmeli veriyi içerir ve ayrı bir veri deposunun veri tabanına yüklenir.

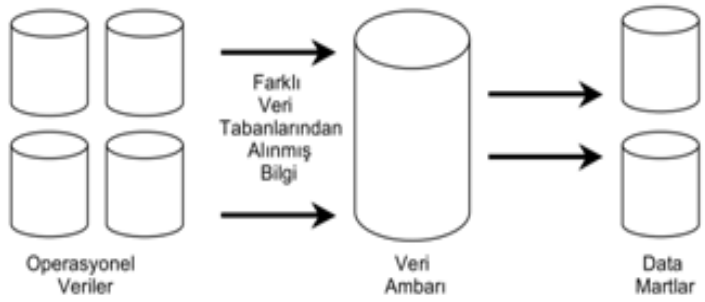
Operasyonel Veri Deposu (ODS): Düzenli, günlük işletme sorgusu ve işletme raporu için, günlük (ya da son zamanlardaki) detaylı veriyi içerir.

Data Mart : Bir grup kullanıcıya yada belirli bir departmana ait şirket verisinin alt grubunu içerir. Bu veri alt grubu bir veya daha fazla operasyonel sistemden ya da bir girişimci veri deposu (EDW)'dan alınabilir. Data Mart'lar genellikle bir kurumun belirli bir fonksiyon alanı için özetlenmiş tarihsel bilgiyi içerir ve bir EDW'den daha hızlı ve daha

ucuz kurulum avantajına sahiptir. Data Mart'ın eldeki veri tabanlarının boyutuyla değil, kullanıcıların fonksiyonel kavrama gücü ile tanımlandığını fark etmek de önemlidir. Gelecekte Data Mart'ların kullanımının artması ile boyutlarında da hızlı bir artış olacağı beklenmektedir (Pipe,1997, s.251-252).

1.2.4.1.1. BİR- İKİ VE ÜÇ KATMANLI MİMARİ

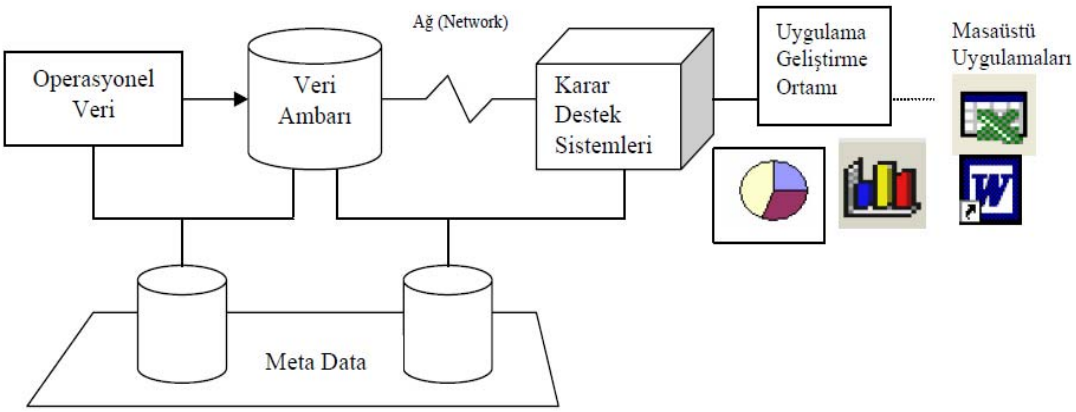
Veri ambarları bir, iki veya üç katmanlı olabilir. Veriyi içeren operasyonel sistem ve veriyi temizlemek için kullanılan yazılım bir katman, veri ambarı ikinci katman, karar destek sistemleri ise üçüncü katmanı oluşturmaktadır. Bu yapının avantajı veri ambarındaki fonksiyonların birbirlerinden ayrılmalarıdır. Bu yapı veri depoları (Data Mart) için ideal bir kullanım alanı oluşturmaktadır.



Şekil 1.8. Operasyonel veri, Veri ambarı ve data mart arasındaki ilişki
Kaynak: Müşteri İlişkileri Yönetimi içinde. http://www.boyutbilgi.com.tr 'den (10 Aralık 2005) tarihinde alınmıştır.

Veri ambarlarında operasyonel veriler genelde ayrı bir ortamda tutulur.

Veri depoları (Data Mart), küçük firmaların, firma içi bazı bölümlerin veya çalışma gruplarının ihtiyaçlarını direkt olarak karşıladıkları için basit ve fonksiyoneldirler. Veri ambarlarına göre küçük oldukları için oluşturulmaları, yönetilmeleri ve sorgulanmaları daha hızlı, ekonomiktir.



Şekil 1.9. Üç Katmanlı Veri ambarı mimarisi

Kaynak: Gray ve Watson, Decision Support in the Data Warehouse 'dan alınmıştır.

A. Alt Katman:

Veriyi içeren operasyonel sistem ve veriyi temizlemek için kullanılan yazılım bu katmanı oluşturmaktadır. İlk katman ilişkisel bir veri tabanı sistemidir. İşlevsel veri tabanlarından ve dış kaynaklardan gelen veriler uygulama programlarının arayüzleri tarafından kullanılmaktadır. Microsoft firmasına ait ODBC (Open Database Connection), ASP ve Cold Fusion gibi araçlar özellikle veri tabanlarına bağlanıp veri tabanlarını kullanabilme gibi özellikleri ile ön plana çıkmışlardır.

B. Orta Katman:

Veri ambarı bu katmanı oluşturmaktadır. İkinci katman, ya ilişkisel OLAP (ROLAP-Relational On-Line Analytical Process) ya da çok boyutlu OLAP (MOLAP- Multidimensional OLAP) modelini kullanarak başlayan bir OLAP sunucudur. İlk katmandan gelen veriler, ROLAP ya da MOLAP modellerinden birinin kullanılmasıyla raporlama, analiz ve veri madenciliği işlemleri için verileri anlamlı bir hale getirir.

C. Üst Katman:

Karar destek sistemleri bu katmanı oluşturmaktadır. Üst katman, sorgulama, raporlama araçları, analiz araçları ve veri madenciliği araçlarını içeren bir katmandır (Gray ve Watson, 1998, s.37).

İki katmanlı yapı, günümüz organizasyonlarının yaygın olarak kullandığı yapıdır. Tek katmanlı yapı, daha az kullanıcının ve sınırlı sayıda verinin bulunduğu daha küçük Data Martlar için kullanılabilir.

1.2.5. VERİ AMBARINDA VERİYE ULAŞIM ARAÇLARI

Veri ambarındaki verilerin karar destek faaliyetlerinde kullanılması aşağıdaki şekillerde olabilir.

1.Sorgulama ve raporlama

2.OLAP

3.Veri madenciliği

Veri madenciliği (Data Mining), istatistiksel bazı yöntemlerin yardımıyla veri içinde gizli olan desenlerin ortaya çıkarılması ve bu desenlerin geleceği tahmin etmekte kullanılmasıdır. Hemen bir örnek verelim. Bir telekom şirketinde, faturasını ödemeyen müşteriler tespit edilir. Hangi özellikteki müşterilerin faturalarını ödemediğine dair anlamlar çıkarılır. Bundan sonra diğer abonelere bu desen uygulanır ve muhtemel kaçak su abonelerini bulunur.

Başlıca data mining araçları olarak ORACLE, SAS, SPSS isimlerini verebiliriz. Bunlarla ilgili bir analiz sonucu aşağıdadır.

-Angoss Software Knowledge Studio 4.2 and Mining Manager 2.1

-Computer Associates CleverPath Predictive Analysis Server 3.0

-Fair Isaac Enterprise Decision Management suite

-Genalytics Predictive Suite 5.0

-IBM DB2 Intelligent Miner

-Insightful Miner 3.0

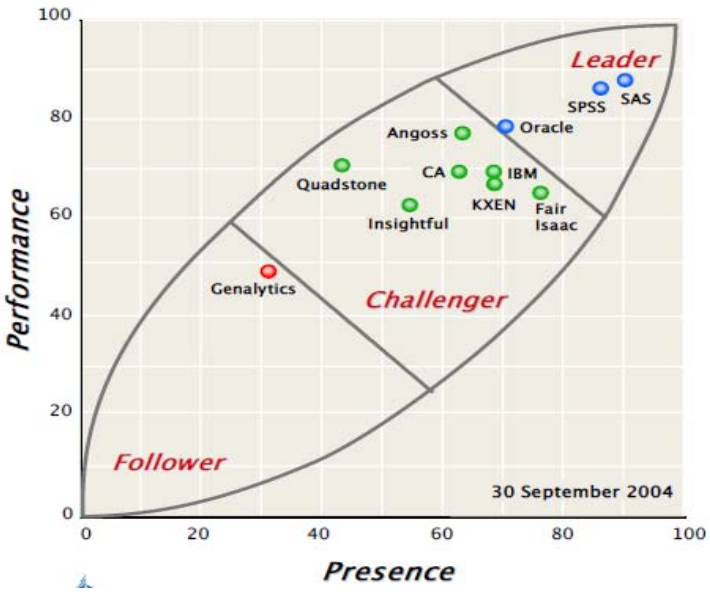
-KXEN Analytic Framework 3.0

-Oracle Data Mining

-Quadstone System V. 5

-SAS Enterprise Miner 5.1

-SPSS Clementine 8.5



Şekil 1.10.Veri madenciliği programları

Kaynak: Data Mining Tools-METAspectrumSM Evaluation, 2009, http://www.oracle.com/technology/products/bi/odm/pdf/odm_metaspectrum_1004.pdf' den alınmıştır.

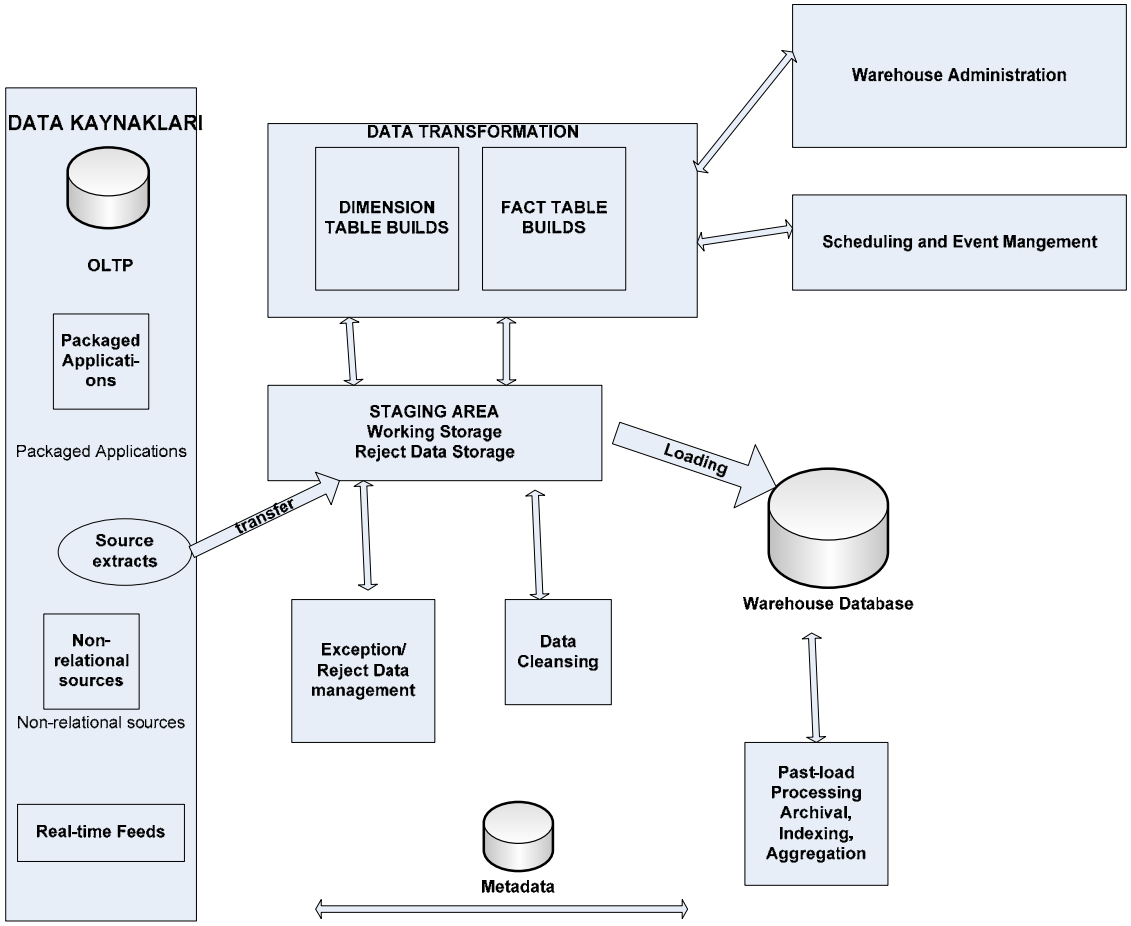
OLAP (OnLine Analytical Processing) için üzerinde görüş birliğine varılan ortak özellik çok boyutlu veri analizidir (MultiDimensional analyzing). Çok boyutlu veri analizinde, veri değişik boyutlardan incelenir. Veri ve boyutları birlikte, küp olarak adlandırılır. Mesela satış verisinin, zaman, ürün ve bölge boyutlarından bakılarak değişimleri incelenebilir. Bu boyutlarda istenilen ayrıntı ve özet seviyesine çıkılabilir. Böylece değişimin sebebi daha iyi anlaşılabilir.

OLAP araçları, çok boyutlu veri depolarını kullanıyorsa MOLAP (Multidimensional OLAP), verilere direkt ilişkili veri tabanlarından ulaşıyorsa ROLAP (Relational OLAP) ve her ikisini uygun bir şekilde kombine ederek kullanıyorsa HOLAP (Hybrid OLAP) olarak adlandırılmaktadır.

SQL(Structured Query Language): Yapısal Sorgulama Dili olan SQL (Structured Query Language), ilişkisel veritabanlarındaki bilgileri sorgulamak için standart kullanımı olan bir dildir. Standart bir dil olmasına karşılık, çeşitli veri tabanlarında SQL kullanımları arasında farklılıklar vardır. SQL komutları ile, tablolara yeni kayıt girme, varolan kayıtları sorgulama (arama ve listeleme), varolan bilgileri değiştirme ve varolan kayıtları silme işlemleri yapılabilir.

1.2.6.OPERASYONEL SİSTEMLERDEN VERİ AKTARIMI

Operasyonel sistemlerde yapacağımız analizlerin, canlı sistemi yoracağından bahsetmiştik. Bu yüzden veriambarlarında operasyonel veriler ile şekillenmiş hali farklı yapılarda tutulmaktadır. Operasyonel sitemlerden , ambara aktarım işlemini ETL (Extract, Transform,Load) olarak isimlendirebiliriz.

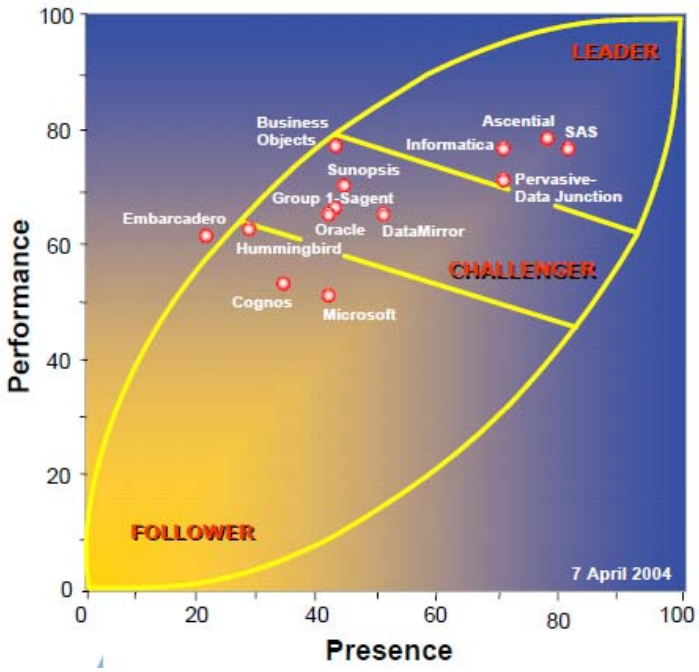


Şekil 1.11. ODS den EBM e aktarım

Kaynak: Mark Madsen, (6.05.2009). TDWI 2008 Learning Book' dan alınmıştır.

Operasyonel veri, uygulamaya yönelik, dağınık, kısa zamanda oluşan ve tekrarlayabilen veriler olarak tanımlanabilir. Güncellenen bilgiler veri ambarına aktarılır. Veri, organizasyon içinden olduğu gibi organizasyon dışı kaynaklardan da elde edilebilir.

Bu işlemi yapan çeşitli yazılımlar mevcuttur.



Şekil 1.12. ETL araçlarının karşılaştırılması

Kaynak: ETL Tools-METAspectrumSM Evaluation içinde. http://www.sas.com/offices/europe/czech/technologies/enterpriseintelligenceplatform/MetagroupETLmarket.pdf' den alınmıştır.

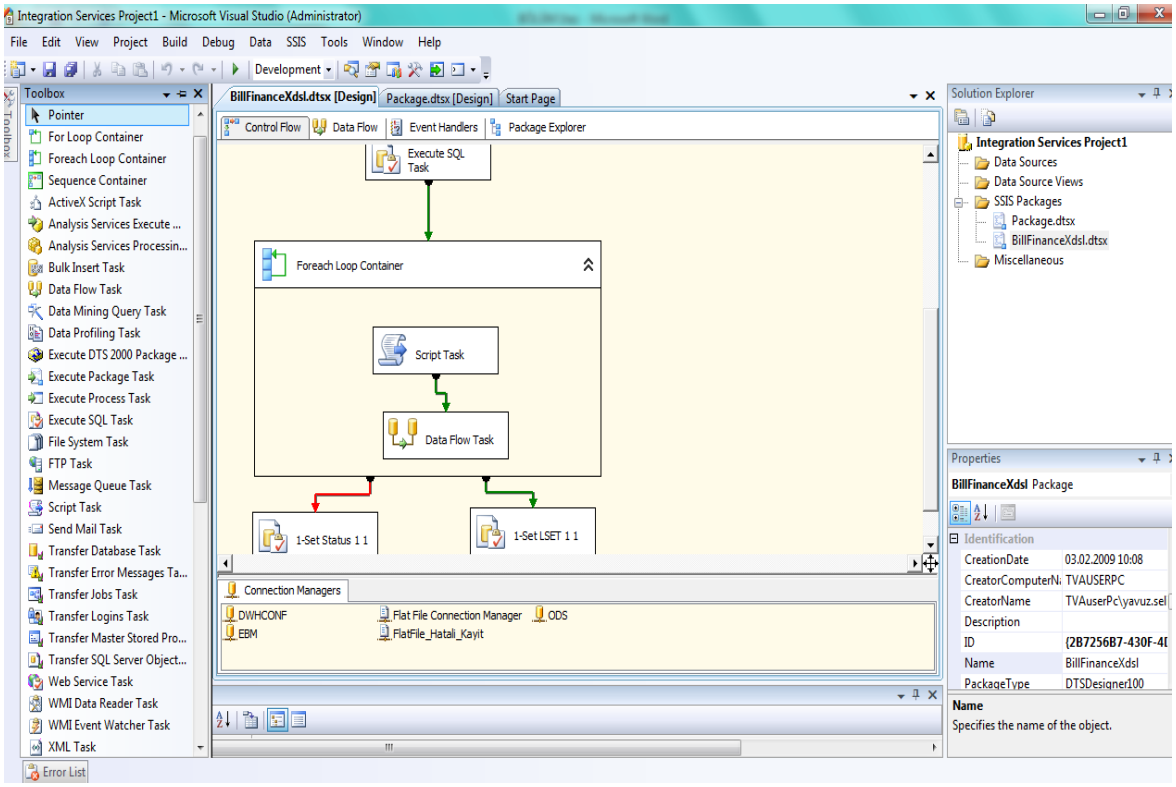
Operasyonel sistemlerden aktarım esnasında, veri ETL isminden de anlaşılacağı gibi ;

1.Extract – lazım bilgiyi seçip çeker

2.Transform- çekilen bilgiyi istenilen formata dönüştürür.

3.Load- hedef tabloya yükler.

ETL yazılımlarının bu aşamalara hitap eden araçları vardır. ETL yazılımlarına Informatica, Microsoft SSIS gibi örnekler verebiliriz. Aşağıdaki şekilde örnek bir SSIS ETL ekranı görebilirsiniz.



Şekil 1.13. Bir SSIS (Microsoft Integration Services) ekranı (Bu şekil uygulama için tasarlanan veritabanından alınmıştır.)

1.2.7. VERİ MODELLEME

Veri modellemenin amacı, verinin taşıdığı anlamı, veriler arasındaki ilişkileri, verinin niteliklerini ve verinin net tanımlarını açıkça belirtmektir. İyi bir veri modeli aşağıda belirtilenleri tanımlamalıdır(Singh, 1998, s.130):

- Varlıklar (Tablolar)

-Nitelikler (Sütunlar)

- Varlık ve niteliklerin eksiksiz olarak tanımlanması

-Veriler arasındaki ilişkiler

- Veri ilişkilerini yöneten iş kurallarını tanımlayan veri önemliliği

- Birincil, ikincil veya alternatif anahtarlar

-Geniş biçim, açık grafikler

İyi bir veri modeli;

- Kurumun ana faaliyet konularını ve

- Konular arasındaki ilişkileri verebilmelidir(Meyer ve Cannon, 1998, s.35).

Veri ambarında çok yaygın olarak kullanılan iki model vardır(Gray ve Watson, 1998, s.67):

-İlişkisel Model ve

- Çok Boyutlu Modelleme

Günümüzde operasyonel sistemlerde organizasyon dışı veri tabanlarında ve iş ortaklarında hızla artan veriler bulunmaktadır. Bu verilerin kısa sürede karar verme amacıyla kullanılması ve işletme planlarının oluşturulmasında kullanılması çok kolay olmamaktadır. Günümüzde piyasadaki pek çok uygulama, kullanıcılara ayrıntılı sorgu yapma ve rapor hazırlama olanaklarını sunmaktadır. Günümüzde veri işleme için tercih edilen yazılım olarak akla "İlişkisel Veri tabanı Yönetim Sistemi" gelmektedir. İkinci kuşak Veri tabanı Yönetim Sistemi olarak gösterilen bu yazılım E.F. Codd (1970) tarafından ortaya konulan ilişkisel veri modeline bağlıdır. İlişkisel modelde bütün veriler mantıksal olarak ilişkiler (tablolar) içinde yapılandırılmıştır. İlişkisel veri tabanları, verileri saklamak için tabloları kullanır. İlişkisel veri tabanları satır ve sütunlardan oluşan iki boyutlu veri tabanlarıdır (Daşdemir, 2004, s. 51-52).

Veri ambarlarında, organizasyonların sahip oldukları verilerin değerini arttırmak istemeleri sebebiyle, çok boyutlu modelleme araçlarından OLAP uygulamaları hızla artmaktadır. OLAP sistemleri işletmenin içerisinde birçok farklı noktada kullanım alanı bulabilir. Pazarlama ve satış verilerinin analizine dayanarak kampanyaların planlanması, finansal veriler kullanılarak finansal tahminlerde bulunulması OLAP araçları ile mümkün olmaktadır. Ayrıca üretim planlama ve analizinde de OLAP araçları kullanılabilir.

Veri ambarlarındaki ilişkisel veri tabanları iki boyutludur. Bu yüzden ikiden fazla boyutlu olan tabloları ifade etmek için farklı terimler kullanılmaktadır. Bunların en yaygın olanları OLAP, MOLAP ve ROLAP'tır.

1.2.7.1. OLAP

İlişkisel veri tabanlarının yaygınlığı ve sonrasında ortaya çıkan veri ambarlarının gelişmesi ile beraber, verilere daha hızlı şekilde erişme ve çok boyutlu analiz ihtiyaçları, bilim adamlarını ve yazılım şirketlerini, daha farklı yapılar geliştirmeye itmiştir. Bu amaçla geliştirilen bir teknoloji olan OLAP (On-line Analytical Processing), ilişkisel veri tabanları gibi, bilimsel temeller üzerine değil, OLAP ürünleri üreten firmaların desteğinde çıkan bir teknoloji olmuştur. OLAP teriminin ilk olarak ortaya çıkışı 1993 yılında Dr. E.F.Codd 'un ortaya koyduğu kurallar çerçevesinde olmuştur(Türkmen, b.t.).

OLAP, bir işletmenin elinde bulunan verileri sadece tek bir bakış açısına göre değil, çok farklı açılardan değerlendirmesine imkan veren bir veri analiz tekniğidir. Zaman kazancının dışında, OLAP üç çok önemli özelliği de beraberinde getirmektedir (Türkmen, b.t.)

Verilere Çok Boyutlu Bakabilme Özelliği: Analizler sırasında kullanılan demografik veriler (yaş, cinsiyet, eğitim durumu), sayısal veriler, adetler, işlem miktarları, gerçekleşen ve bütçelenen değerler, ürün tipleri, ürün özellikleri ve zaman gibi değişkenlere "boyut (Dimension)" adı verilmektedir. Yöneticiler ve analistler, çalışmaları sırasında, tüm bu tanımlanan verileri yatay veya düşey eksenlerde birlikte görmek isteyebilirler.

İlişkisel veri tabanları, sadece iki boyutlu oldukları için bu şekilde raporlara izin vermezler, daha karmaşık analizler söz konusu olduğunda, bir OLAP yapısı kurmadan bu raporları almak imkansız hale gelebilir.

Boyutların başka bir özelliği de hiyerarşilerin tanımlanabilmesidir. Hiyerarşiler sayesinde, hem toplamlara ulaşmak kolaylaşmakta, hem de farklı gruplar için, farklı senaryolar hazırlayabilme şansı doğmaktadır.

Karmaşık Hesaplamalar: Bir OLAP sisteminin gerçek performansı, karmaşık hesaplamaları yapma gücü ile ölçülebilir. OLAP sistemleri, toplama işleminden başka işlemler de yapabilecek güçte olmalıdır. Analiz yapanlar için, rakamlardan çok yüzdesel dağılımlar önemlidir. Birkaç yıllık satış rakamları içerisinde, binlerce ürün türü için günlük bazda satışları yüzdesel olarak analiz edip, sıraya dizebilmek bir ilişkisel veri tabanı yönetim sistemi ile saatler sürecek bir raporun çalışmasını gerektirebilir. Oysa uygun bir OLAP sistemi ile bir günlük satışlar ve birkaç yıllık satış rakamları arasında bir fark olmamalıdır.

Zaman Kavramları: Zaman boyutu neredeyse her analizin temel bileşenidir. Zaman diğer boyutlardan farklı olarak, kendine has bir sıralama içerisinde gider.

OLAP aracılığı ile ilişkisel veri tabanı ve düz dosyalardaki veriler, bazı düzenlemeler yapılarak çok boyutlu veri ambarına aktarılır. Bu çok boyutlu veri ambarlarına; eğer üç boyut söz konusuysa "veri küpü", daha fazla boyut söz konusuysa "hiper küp" adı verilmektedir(Singh, 1998, s.159).

Veri ambarında her bir boyut, bir değişkeni ifade eder. Zaman, veri ambarında hep var olan bir değişkendir. Bunun yanı sıra farklı boyutlar - değişkenler de söz konusudur. Örneğin satış verisi en az beş boyut içerebilir:

Aşağıda kullanılan telekom firması için bir örnek verilmiştir.

- Zaman

- Ürün

- Müşteri

- Müdürlük

- DetayTipi

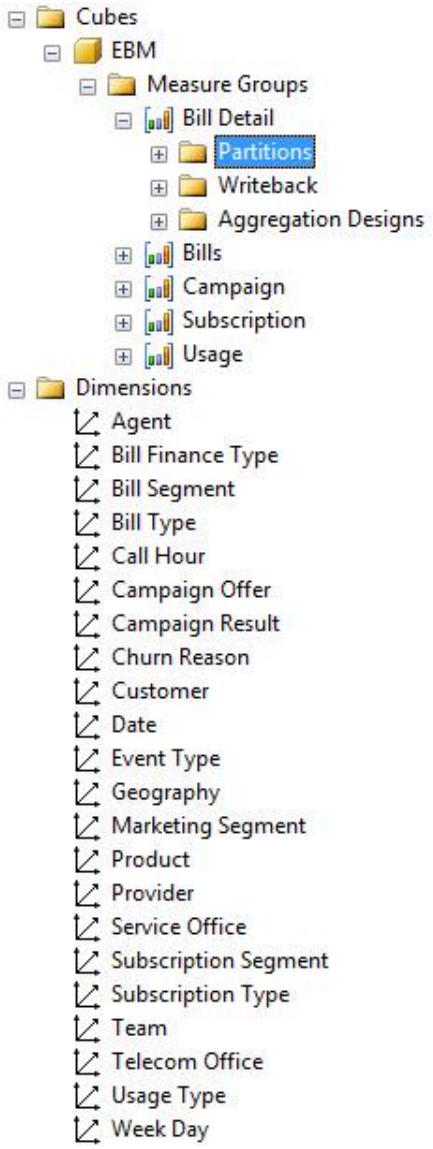
- Süre

Aşağıda beş boyutlu birgörülmektedir:

Tablo 1.1. Basit Bir İlişkisel Veri Tabanı Örneği (Bu tablo uygulamadan alınmıştır.)

Zaman	**Ürün**	**Müşteri**	**Müdürlük**	**DetayTipi**	**Süre(sn)**
Ocak	A	M1	1	D1	10
Ocak	A	M1	1	D2	5
Ocak	A	M2	3	D3	100
Şubat	A	M1	1	D4	35
Şubat	C	M3	4	D4	45
Şubat	C	M4	4	D5	56
Mart	A	M5	5	D6	33
Mart	B	M6	5	D7	200
Mart	B	M6	6	D1	55

Aşağıda kullandığımız firmanın OLAP küpünden bir örnek görüyoruz.



Şekil 1.14. Bir SQL Server OLAP küp ekranı (Bu şekil uygulamadan alınmıştır.)

1.2.7.2. MOLAP

MOLAP veri tabanı sadece uzmanlaşmış bir sorgu dili veya aracı kullanarak erişilebilen, uzmanlaşmış bir yapı içindeki tam boyutlu veri tabanını içermektedir. MOLAP çok boyutlu bir veri ambarlama biçimi kullanmaktadır. MOLAP, çok boyutlu veri sorgulamaları ile küçük ve orta boyutlu veri sorgulamaları için en iyi performansı sağlar (Berry ve Linoff, 2000, s. 7-8).

Çok boyutlu veriyi saklamanın iki yolu vardır: Çok boyutlu veri tabanları veriyi saklamak için veri küplerini kullanır. Diğer bir yol ise standart ilişkisel veri tabanı yönetim sistemlerini kullanmaktır. Bu yöntemler veriyi "Yıldız Şema" kullanarak saklarlar (Meyer ve Cannon, 1998, s. 24-25).

Boyut sayısı arttıkça hem veri hem de veri ile ilgili sorulacak sorular karmaşık hale gelebilir. Önceden hesaplanan ve çok boyutlu yapıya sahip olan veri, çoğunlukla bir küp şeklinde tanımlanmaktadır (Akpınar, 2000, s. 1-22).

1.2.7.3. ROLAP

ROLAP ambar mimarisiyle veri, kendi orijinal ilişkisel ambar yapıları içinde; kümeleme analizi sonuçları ve özetler ise ayrı ilişkisel tablolar içinde ambarlanmaktadır. ROLAP mimarisi, veri tabanı boyutu ortadan büyüğe doğru gittiği zaman uygundur.

Çoğu firma,

-Çok fazla miktarda ilişkisel veri tabanına,

-İlişkisel veri tabanları için yazılım lisanslarına,

- İlişkisel veri tabanları için tecrübeli programcılara ve veri tabanı yöneticilerine sahip olduklarından, bu firmalar MOLAP yerine ROLAP'ı tercih etmektedirler(Gray ve Watson, 1998, s. 72).

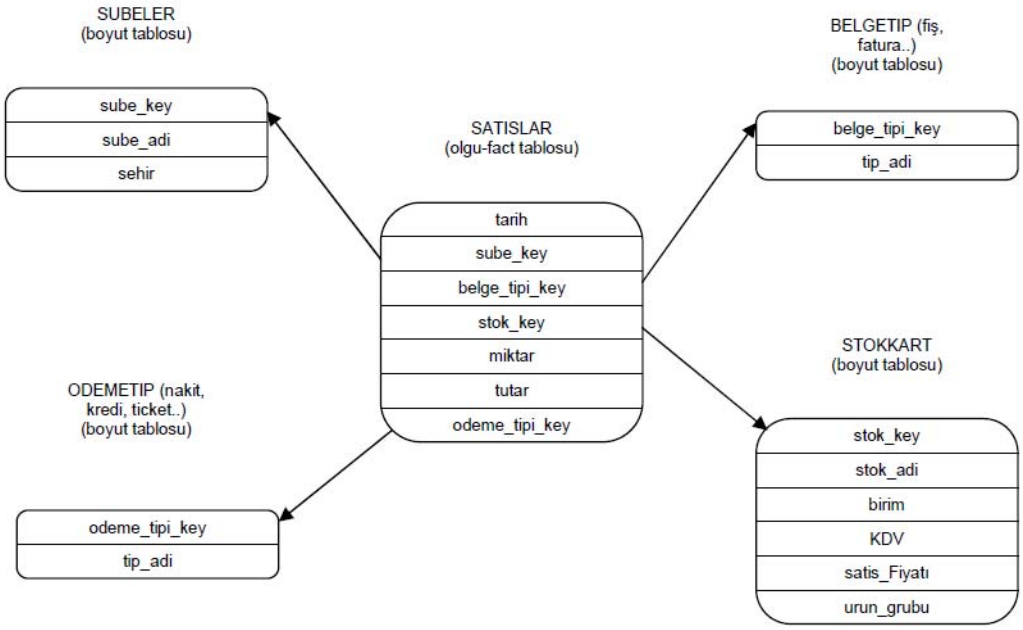
ROLAP, yıldız veya kar tanesi şemalarından birini kullanan, tamamen ilişkisel bir veri tabanı yönetim sistemine dayanan OLAP veri tabanıdır(Han ve Kamber, 2000, s. 48).

1.2.7.3.1. YILDIZ ŞEMA

Şema, belirli bir veri kümesinin tanımıdır. Başka bir ifadeyle, bir veri tabanındaki ilişkileri tanımlayan Meta Data türündeki verilerdir.

Yıldız şeması kullanan ROLAP, merkez tablo ve boyut tablolarını kullanarak, bilgiyi yeniden düzenler. Yıldız şemada bir merkez tablo ve birden fazla boyut tablosu

vardır. Merkez tablo var olan sayısal değerleri içerir. Boyut(Dimension) tablolarında ise, merkez tablodaki başlıklar ayrıntılı alt başlıklarla yer alır.



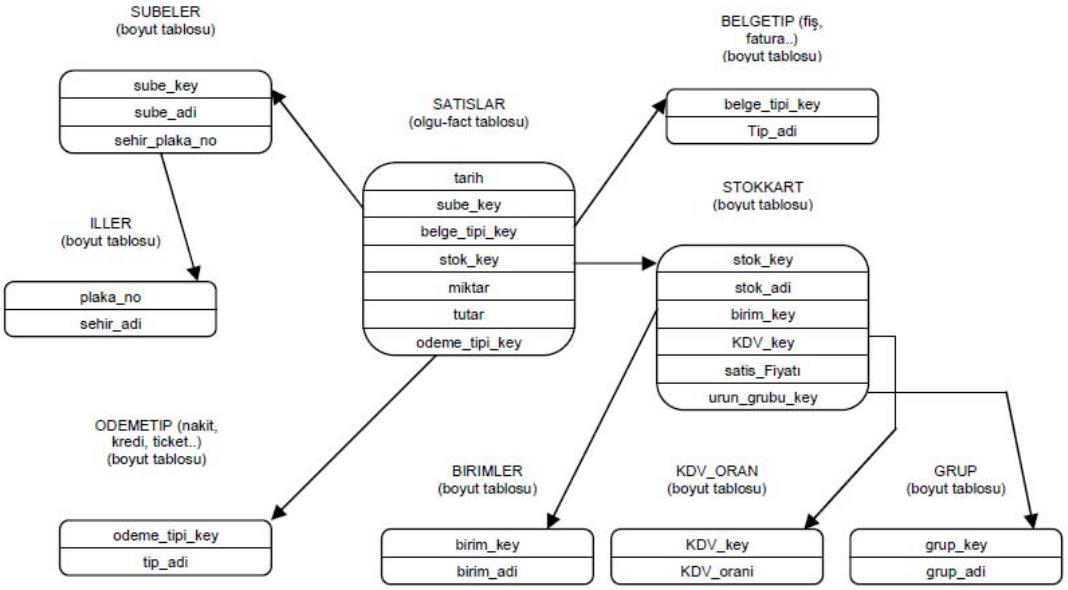
Şekil 1.15.Yıldız şema

Kaynak: Özçakır (2006) içinde. Müşteri İlişkilerindeki Birlikteliklerin Belirlenmesinde Veri Madenciliği Uygulaması' dan alınmıştır.

1.2.7.3.2. KAR TANESİ ŞEMA

Veri ambarlarında veri tabanı tasarımı, kar tanesi şema ve yıldız şema ile yapılır. Yıldız şema veri tabanı tasarımcısının bakış açısına göre sınırlayıcı olabilir.

Kar tanesi şema bir merkez tablo ve boyut tablolarından oluşan çok boyutlu bir modeldir. Kar tanesi şema boyut tablolarına bir hiyerarşi daha eklemeyi sağlar. Bu durumda boyut tablolarının sayısı artar(Levene ve Loizou, 2003, s.235-240). Kar tanesi şema yaygın olarak kullanılmaktadır.



Şekil 1.16. Kar tanesi şema

Kaynak: Özçakır (2006) içinde. Müşteri İlişkilerindeki Birlikteliklerin Belirlenmesinde Veri Madenciliği Uygulaması' dan alınmıştır.

1.3. OLTP VE VERİAMBARI ARASINDAKİ FARKLAR

Veri tabanlarında ekleme, silme, güncelleme gibi operasyonların gerçekleştirildiği bu uygulamalara "Online Transaction Processing" (OLTP) uygulamaları denir. OLTP sistemlerinin kullanıldığı veri tabanlarına erişim çoğunlukla bir tür sorgulama dili olan Yapısal Sorgulama Dili/ Structured Query Language (SQL) aracılığıyla gerçekleştirilmektedir.

Tablo 1.2.OLTP ve OLAP arasındaki farklar

Kaynak: Öztürk ve Tarımcı (b.t.) içinde, http://www.kouemk.com/makale/default.asp?set=makale&id=4' dan alınmıştır.

AMAÇ	Anlık bilgiler tutulur	Veri analizi yapılır
KULLANICI SAYISI	>1000	<100
YAPI	Veri tabanı sistemleri	Veri tabanı sistemleri
VERİ MODELİ	Normal	Çok boyutlu
GİRİŞ	SQL	SQL+Veri analiz yöntemleri
VERİ TİPİ	En eski veri 90 günlük	En eski veri yıllık
VERİNİN DURUMU	Değişken, tamamlanmamış	Tanımlayıcı, tarihsel veri
TABLOLAR	Küçük boyutlu tablolar	Geniş boyutlu tablolar
GÜNCELLEME	Sürekli	Daha uzun aralıklı

Veri ambarlarıyla birlikte ortaya çıkan bu büyük veri kümelerinin etkin bir şekilde kullanımı sorunu ise OLAP teknolojisi ile çözümlenebilmiştir. OLAP uygulamaları bir organizasyonun karar destek mekanizmasında ihtiyaç duyulan analiz edilmiş bilgiyi sağlayan uygulamalardır.

Bir firmanın farklı ürünleri için değişik yapılarda veritabanı sistemleri olabilir. OLAP, bu istemlerin tek çatı ve tek format altında birleştirilmesine yardımcı olur.

Bir OLAP veri tabanı bölgeler, ürün tipi ve satış kanalı olarak bölümlendirilmiş satış verilerini içerebilir. Tipik bir OLAP sorgusu her bir ürün tipinin her bir bölgedeki toplam satış miktarını belirlemeye çalışabilir. Sonuçları gözden geçiren bir analist, sorguyu her bir satış kanalının bölgelere göre satış hacmini bulmak için yineleyebilir veya analist her bir satış kanalının yıllar itibariyle satışlarını karşılaştırmak isteyebilir. Tüm bu süreç çevrimiçi olarak, kısa bir sürede ve analiz süreci bölünmeden incelenebilir. OLAP uygulamaları bu analizlerin kolaylıkla yapılması ve sonuçların da anlaşılır olmasını sağlar (Singh, 1998, s.179).

1.4. VERİ MADENCİLİĞİ VE VERİ AMBARI ARASINDAKİ İLİŞKİ

Önceki bölümlerde veri ambarı kavramı incelemeye çalışıldı. Veri Ambari özetle bize elimizdeki bütün veriyi gösteren, sunumunda kolaylık gösteren daha az dinamik, analize yönelik veri depolarıdır. Bu depolar, mevcut veriyi sunduğu gibi aslında içinde bizim ilk bakışta göremediğimiz bazı amlamları da içermektedir. Hangi tip müşterinin nasıl bir ürün kullandığı, veya ortalama kaç saat kullandığı; Trabzon doğumlu müşterilerin genelde hangi semtlerde oturduğu gibi bazı veri formüllerini elde edebiliriz. İşte bu formülleri çıkarmamıza yardımcı olan teknolojiye de veri madenciliği denir.

Veri madenciliği, veri ambarı sistemlerini kullanır. Bölüm 2 de veri madenciliği konusu ayrıntılı olarak incelenecektir.

Veri ambarında veri oluşturulduktan sonra bu verinin elle veya gözle analizi yapılabilir. Bunun için OLAP (*Online Analytical Processing*) programları kullanılır. Bu programlar veriye her boyutu veride bir alana karşılık gelen çok boyutlu bir küp olarak bakmayı ve incelemeyi sağlar. Böylece boyut bazında gruplama, boyutlar arasındaki korelasyonları inceleme ve sonuçları grafik veya rapor olarak sunma olanağı sağlar.

Veri madenciliğinde amaç, kullanıcının bilgi çıkarma sürecinde katkısının olabildiğince az tutulması, işin olabildiğince otomatik olarak yapılabilmesidir. Çünkü OLAP programlarını kullanırken bulunabilecek sonuçlar kullanıcının sormayı düşündüğü sorgularla sınırlıdır.

2. VERİ MADENCİLİĞİ

2.1. VERİ MADENCİLİĞİNE GENEL BAKIŞ

Veri madenciliği, yararlı bilginin büyük veri depolarından otomatik olarak keşfedilme sürecidir (Tan v.d., 2005). Diğer bir deyişle, veri madenciliği tek basına bir şey ifade etmeyen veriler içindeki gizli örüntüleri ve ilişkileri ortaya çıkarmak için istatistik, yapay zekâ ve makine öğrenmesi gibi yöntemlerin ileri veri çözümleme araçlarıyla kullanılmasını kapsayan süreçler topluluğudur (Bozkır v.d.,2008).

Veri madenciliği; büyük miktarda veri içinden, gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanılarak aranmasıdır. Veri analizi yapılarak, bir mal için bir sonraki ayın satış tahminleri yapılabilir, müşteriler satın aldıkları mallara bağlı olarak gruplandırılabilir, yeni bir ürün için potansiyel müşteriler belirlenebilir, müşterilerin zaman içindeki hareketleri incelenerek onların davranışları ile ilgili tahminler yapılabilir. Binlerce malın ve müşterinin olabileceği düşünülürse bu analizin gözle ve elle yapılamayacağı, otomatik olarak yapılmasının gerektiği ortaya çıkar ve veri madenciliği bu noktada devreye girer.

Özmen'e göre veri madenciliği; büyük miktarda ve oldukça hızlı toplanan verilerin, çeşitli analizler sonucunda anlamlı bilgilere dönüştürülmesi noktasında devreye giren bir süreçtir. Veri madenciliği tanımları incelendiğinde; bu tanımlarda ortak olan unsurlardan ilki çok fazla miktarda verinin veri ambarında tutulması, ikincisi ise bu verilerden anlamlı bilgiler elde edilmesidir (Özmen, b.t.).

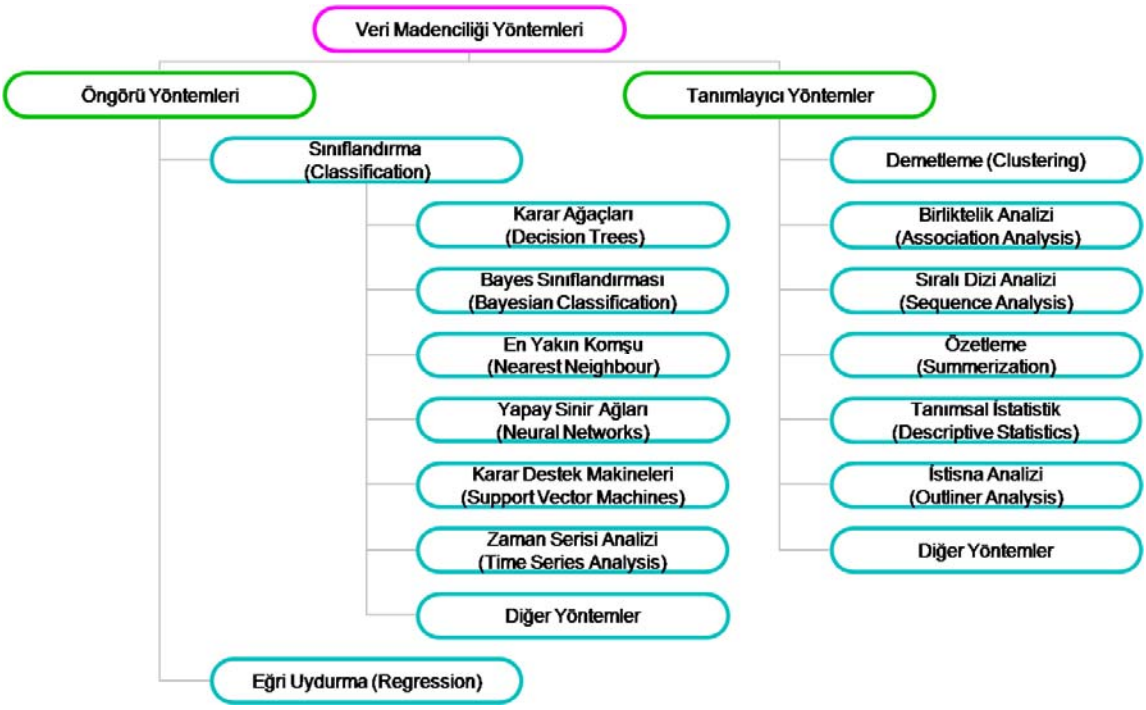
SPSS firmasının web sayfasında veri madenciliği, "büyük miktarda veri içinde gömülü olan bilgilerin çıkarılması ve bu bilgilerin kuruluşa stratejik karar desteği sağlama amaçlı kullanılması" olarak tanımlanmıştır. Müşteri ile ilgili detaylı ve yüksek miktardaki verinin, onlara sunulan hizmetlerin geri dönüşünü sağlamak ve yapılan işin değerini arttırmak amacıyla bilgiye dönüştürülmesinin, ham verideki bilinmeyen ilişkilerin ortaya çıkarılmasının, güçlü analitik gereçleri bir arada sunan veri madenciliği çözümleri ile mümkün olduğu belirtilmektedir (Data Mining, Anonim, b.t.).

SAS firmasına göre veri madenciliği, veri yığınları içinden kuruluş yöneticileri için en gerekli olan verilerin seçilmesi, düzenlenmesi ve modellenmesi süreçlerini içermektedir. Veri madenciliği, karar vericilerin kullanabileceği yeni bilgiler oluşturabilmek için yapay zeka gibi ileri teknoloji içeren yöntemler kullanmaktadır (Data Mining, Anonim, b.t.).

Intera Systems firmasının web sayfasında veri madenciliği, "arşivlenen bilgiler üzerinde yapılan analizlerle önceden bilinmeyen, değerli ve anlaşılabilir sonuçlar çıkarma süreci" olarak tanımlanmıştır. Elde edilen sonuçlar tahmin yürütme, sınıflandırma veya kayıtlar arasındaki benzerliklerin bulunması amacıyla değerlendirilmekte ve bu özellikler karar destek sistemlerinde kullanılmaktadır (Veri Madenciliği, Anonim, b.t.).

Veri ambarları veriyi kullanılabilir trend, ilişki ve profillerde sınıflandırmazlar, sadece potansiyel bilgiye sahip veri tabanlarıdırlar. Veride saklı bilgiyi keşfetmeyi sağlayan ise veri madenciliği gibi tekniklerdir. Veri ambarından veriyi çekebilmek için hangi verinin gerekli olduğunu ve bu verinin nerede olduğunu tespit etmek önemlidir. Çoğunlukla gerekli veri, farklı sistemler üzerinde olup, farklı formatlardadır. Bu nedenle, ilk aşamada veri temizleme ve düzenleme işlemi gerçekleştirilmelidir.

Veri madenciliği kümeleme, tahmin, tanımlama ve görselleştirme, sınıflandırma ve birliktelik kuralları gibi tekniklerden yararlanmaktadır.



Şekil 2.1. Veri madenciliği modelleri ve teknikleri

Veri madenciliğinde genelde 2 çesit modelleme yapılır.

1. Tahmin edici modeller (karar ağaçları, regresyon, destek vektör makineleri vs.)

2. Tanımlayıcı modeller (kümeleme, birliktelik kuralları vs.)

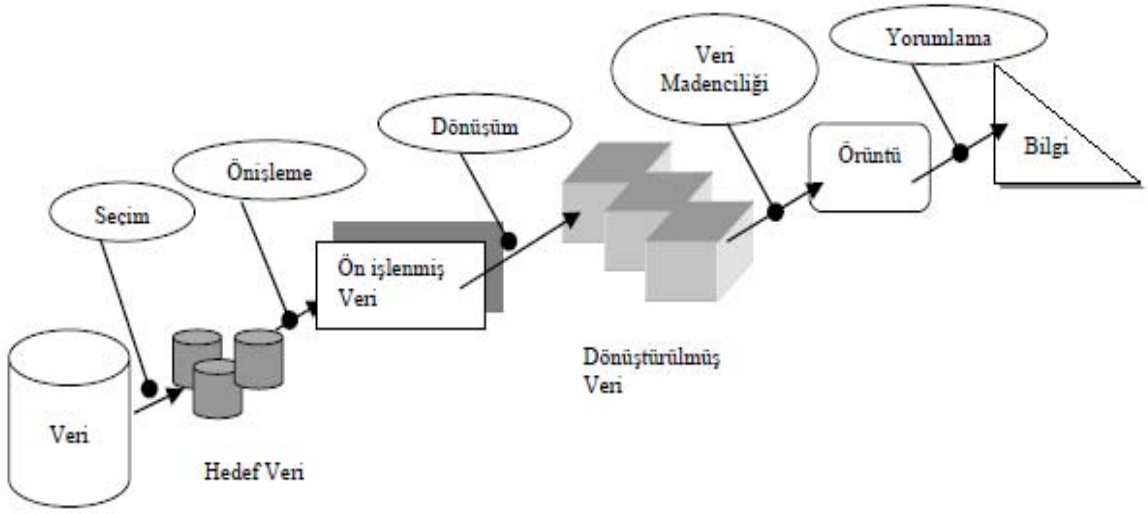
Tahmin edici modellerde amaç mevcut verileri kullanarak geleceğe yönelik kestirimler yapabilmek iken, tanımlayıcı modellerde amaç yine mevcut veri içindeki gizli ilişkileri, kümeleri ve veriyi niteleyebilecek olan özellikleri ortaya çıkarmaktır.

Bu amaçlara hitap eden veri madenciliği tekniklerini aslında 3 sınıfta da sınıflandırabiliriz.

• Sınıflama (Classification),

• Kümeleme (Clustering),

• Birliktelik kuralları ve sıralı örüntüler (Association rules and sequential patterns).



Şekil 2.2. Veri madenciliğinin veri işleme sürecindeki yeri

Kaynak: Akbulut, 2006 içinde. Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu' dan alınmıştır.

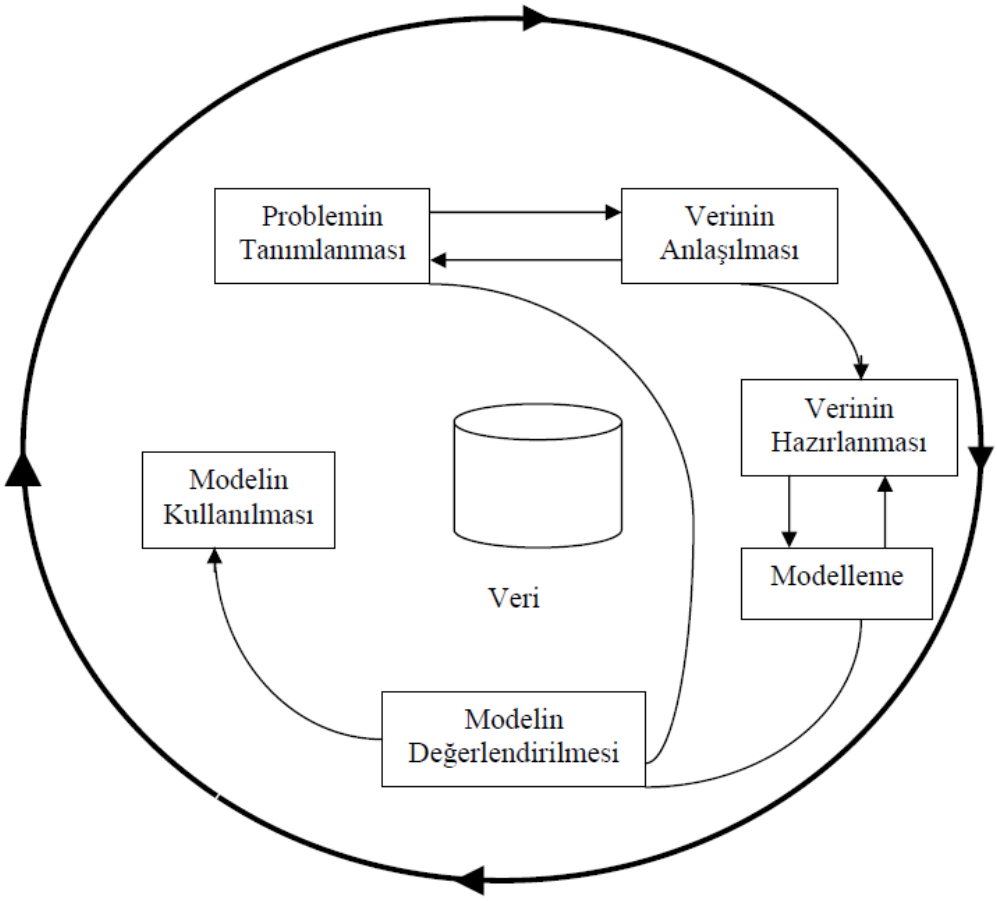
2.2. VERİ MADENCİLİĞİ SÜRECİ VE CRISP-DM

Veri madenciliği kavramını uygulayabilmek için her teknolojide olduğu gibi bazı kurallar ve teknikler doğrultusunda gidilmelidir. Veri madenciliği, bize yapacağımız bir kampanya için veri sağlayacaksa bu metodun doğru çalışması için gerekenleri yapmalıyız.

Bir firmanın A tipindeki müşterilerine bir kampanya hazırlayacağını düşünelim. Bunun için önce yapmak istediği şeyi belirler. Bir veri madenciliği çalışması için tabi ki en önemli şey veriyi elde etmektir. Fakat elde ettiğimiz veri temiz olmayabilir. Yani boş kolonlar, yanlış bilgiler olabilir. Analizlere başlamadan önce bu problemler temizlenmelidir. Veri hazırlandıktan sonra istenen yöntem seçilir ve çalışma yapılır.

Veri madenciliği süreci çok hızlı bir biçimde karmaşık hale gelebilir. Bu sebeple veri madenciliği için herkesin kullandığı bir süreç söz konusudur. Daha sonraları bu süreç bir konsorsiyum tarafından belirlenmiş ve CRISP-DM adını almıştır. The Cross-Industry Standard Process for Data Mining (CRISP-DM) konsorsiyumu, 1996 yılının sonlarına doğru genç ve olgunlaşmamış veri madenciliği pazarında üç firma tarafından kurulmuştur(Acungil, b.t.). Orjinal üyeleri Daimler-Benz, SPSS ve NCR olan bir konsorsiyum tarafından geliştirilmiştir.

İşi anlamak – Hedefleri ve gereklilikleri belirlenir.
Veriyi anlamak - Veriyi detaylı olarak inceler, saklı desenler hakkında hipotez geliştir.
Veriyi hazırlamak - Veri kaynağıyla bağlantı kur ve veri madenciliği modelinde kullanılacak veri kümesi oluşturulur.
Modelleme - Bir ya da daha fazla veri madenciliği modeli seçilir, parametreleri belirlenir, iyileştirilir.
Deneme - İş hedeflerine göre model ya da modelleri denenir, gözden geçirilir ve gerekiyorsa iyileştirmeleri yapılır.
Sahaya sürmek - Model sonuçlarını analistlere ve son kullanıcılara sunulur, model sonuçlarını iş süreçlerine yorumlanır ve uygulanacak şekilde rafine edilir.



Şekil 2.3. CRISP-DM şeması

Kaynak: SPSS Help Online içinde. (22.12.2009) tarihinde https://www.spss.com/ 'dan alınmıştır.

2.2.1. İŞ ANLAYIŞ SAFHASI

Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi kuruluş amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili kuruluş amacı, sorun üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Sorun ile tam örtüşmeyen bir veri madenciliği çalışması, sorunu çözmeye yetmeyeceği gibi sonuçta başka problemlerin de ortaya çıkmasına neden olabilecektir. Ayrıca yanlış kararlarda katlanılacak olan maliyetlere ve doğru kararlarda kazanılacak faydalara ilişkin öngörülere de bu aşamada yer verilmelidir.

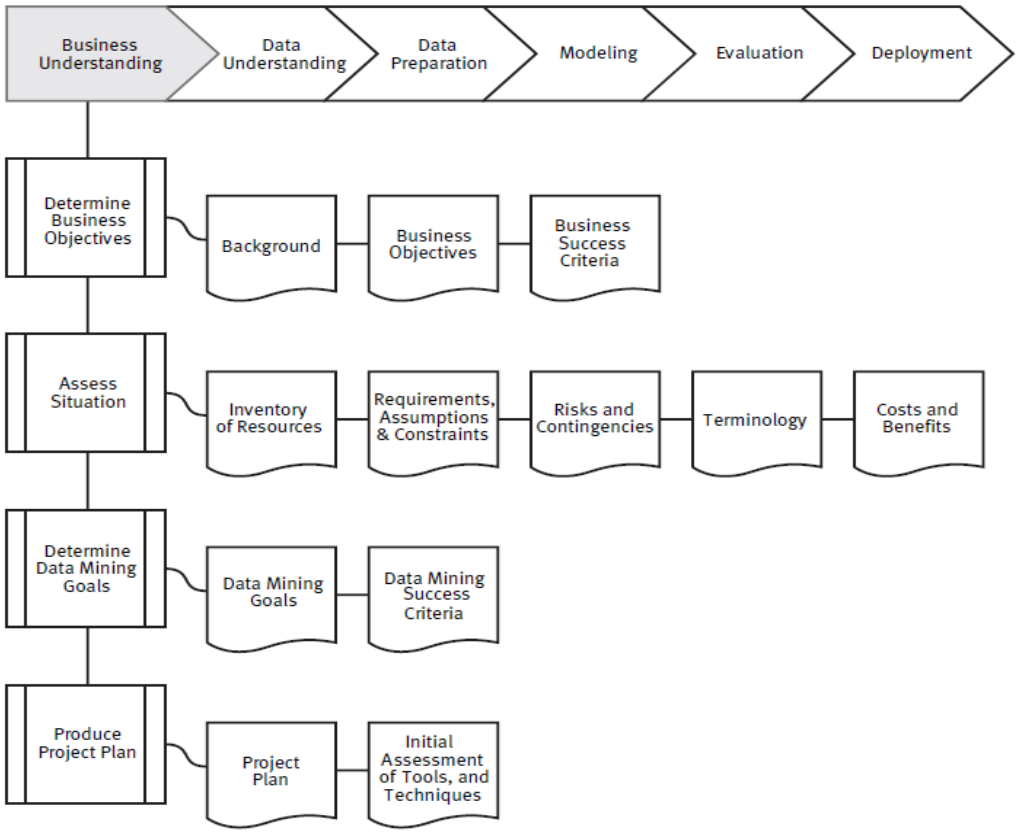
Bir veri madenciliği çalışması yapabilmek için öncelikle iş probleminin veya durumun tanımlanması gerekmektedir. Örneğin sürekli müşteri iptali yaşayan bir firma, bu kayıpların nedenini ve önceden belirleyip bunları engellemeyi istemektedir. Bu tip çalışmalara churn diyoruz. Churn, segmentasyon vs. çalışmaları için öncelikle yapılacak olan işi anlamak gerekir. İş anlayış safhası, veri madenciliği çalışmaları-nın ilk aşamasıdır. Bu aşamada; amaç,hedefler , ön strateji belirlenir.

-İş hedeflerini belirlemek

-Durum değerlendirmek

-Veri madenciliğinin hedeflerini belirlemek

-Proje planı üretmek



Şekil 2.4. İş anlayış safhası

Kaynak: SPSS Help Online, CRISP-DM, (b.t.) içinde.

-Determine Business Objects/ İş nesnelerini belirlemek

Analistin ilk hedefi, müşterinin ne istediğini bulmaktır. Analist projenin sonuçlarını etkileyecek önemli faktörleri belirler.

Ana iş öğelerini belirledikten sonra müşteriyle ilişkilendirilebilecek başka sorular olacaktır. Örneğin aboneliğini iptal eden bir tv müşterisi öncelikli izlediği kanalın düşük kaliteyle sunulmasından rahatsız mı ?

Projenin başarılı olduğunu ilan etmek için belirlenen kriterler nelerdir?

-Assess situation/Durum değerlendirmek

Bu kısımda, projenin datasını belirlerken gerekli olan bütün kaynaklar, kurallar ve diğer etkenler ayrıntılı olarak incelenir.

Projenin içerdiği personel, data ,sistemin donanım ve yazılımı incelenir.

Projenin başarısız olmasına neden olabilecek faktörler , olaylar listelenir. Terminoloji belirlenir.

Projenin maliyeti ve sonunda sağlayacağı faydalar belirlenip karşılaştırılır.

-Determine data mining goals/Veri madenciliğinin hedeflerini belirlemek

Proje için hedef net olarak belirlenir. Örneğin " mevcut müşterilerin satışını artırmak " veya " müşterilerin satın alabileceği ürünleri takip etmek " gibi.

-Produce project plan/Proje planını üretmek

Veri madenciliği çalışmasının hedeflerinin ve iş hedeflerinin belirleneceği proje planı çıkarılır.

2.2.2. VERİ ANLAYIŞ SAFHASI

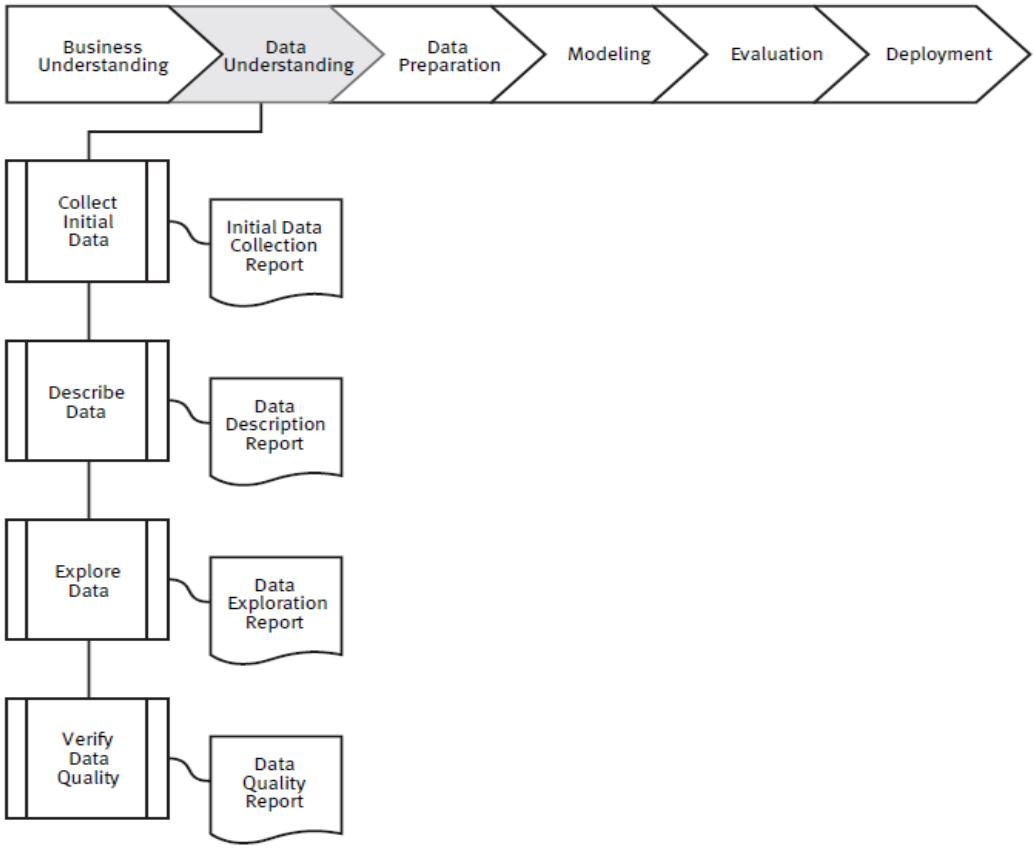
Bu aşama belirlenen proje planı doğrultusunda , proje hedefine ulaşmak için gerekli olan datanın belirlenmesi, toplanması ve kalitesinin belirlenmesi amacına yöneliktir.

-İlk veriyi toplamak

-Veriyi tanımlamak

-Veri keşfi

-Veri kalitesini doğrulamak



Şekil 2.5. Veri anlayış safhası

Kaynak: SPSS Help Online, CRISP-DM, (b.t.) içinde.

-Collect initial data-Veriyi toplamak

Bu bölümde data toplama, ilk yüklemeyi de içerir. Kullanılacak olan bir tool varsa o belirlenir. Birden fazla kaynak kullanılıyorsa bir tool kullanmak daha uygun olur. Lokasyonlarıyla beraber datasetleri listelenir. Bu esnada bazı problemler çıkabilir, çözümler bulunur.

Describe data/Datayı tanımlamak

Datanın özellikleri incelenip sonuçlarla beraber raporlanır. Datanın formatı, kayıt ve her tablodaki alan sayısı incelenir. Bağlantılı diğer ihtiyaçlar da belirlenir.

-Explore data/Datayı keşfetmek

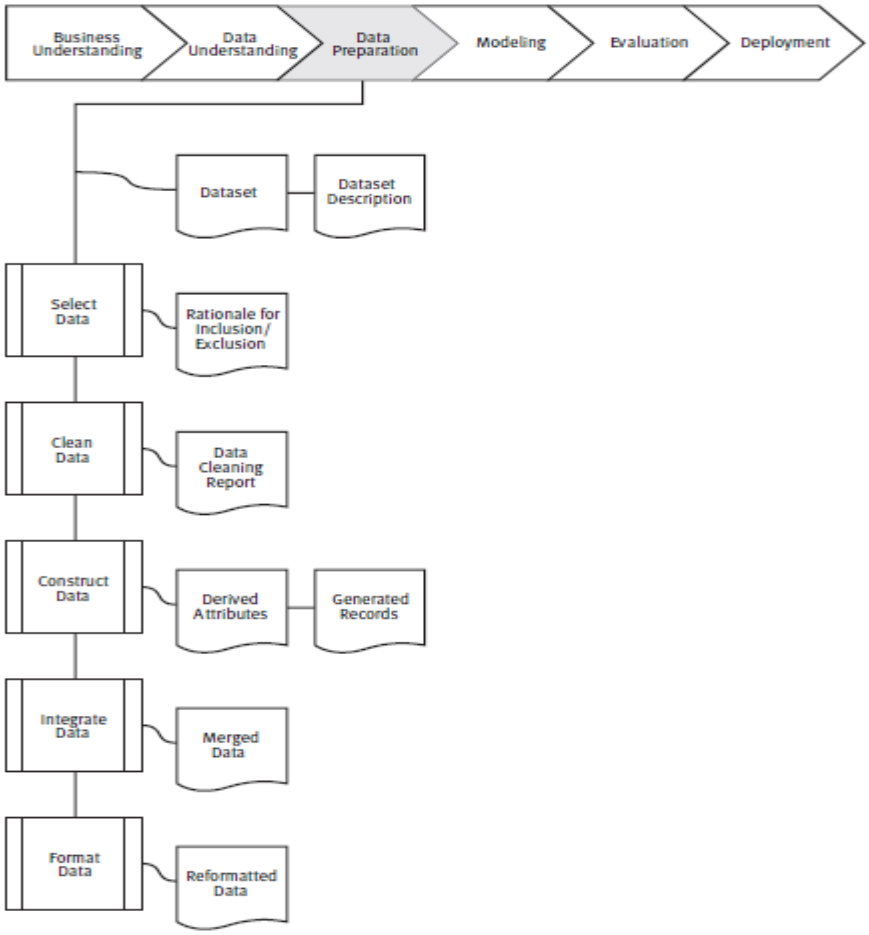
Bu kısımda, sorgulama, raporlama teknikleri kullanarak veri madenciliği sorularına ulaşılır. Bunlar arasında anahtar vasıfın dağılımı, küçük agresyonların sonucu, basit statik analizler sayılabilir. Bu analizler direk veri madenciliği hedeflerine de varabilir veya başka data hazırlama basamaklarına yön verici olabilirler. İlk bulgulara göre bir ön hipotez hazırlanır.

- Verify data quality/Data kalitesini belirlemek

Bu kısımda datanın kalitesi sorgulanır. Datanın ne kadarı dolu, hata içeriyor mu, bu hatalar genelde ortak mı gibi sorular belirlenir. Belirlenen data kalitesi hakkında listeleme yapılır. Data kalitesinde sorunlar varsa bunlar listelenir.

2.2.3. DATA HAZIRLIĞI

Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir karar vericinin veri keşfi sürecinin toplamı içerisindeki enerji ve zamanının çok büyük bir bölümünün harcanmasına neden olmaktadır. Verilerin hazırlanması aşaması kendi içerisinde toplama ve uyumlaştırma, birleştirme ve temizleme ve seçme adımlarından meydana gelmektedir.



Şekil 2.6. Veri hazırlığı

Kaynak: SPSS Help Online, CRISP-DM, (b.t.) içinde.

-Veriyi seçmek

-Veri temizliği

-Datayı entegre etmek

-Veriyi düzenlemek

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzlukların başlıcaları farklı zamanlara ait olmaları, güncelleme hataları, veri formatlarının farklı olması, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında

0/1 olarak kodlanması), farklı ölçü birimleri ve varsayım farklılıklarıdır. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır. Güvenilir olmayan veri kaynaklarının kullanımı tüm veri madenciliği sürecinin de güvenilirliğini etkileyecektir.

Şimdi bu aşamalara yakından bakalım.

-Select data/Datayı seçmek

Modellemede kullanılacak olan dataseti belirlenir. Burada projenin analiz çalışması ve modelleme için kullanılacak datasetleri hazırlanır.

Analiz için kullanılacak olan dataya karar verilir. Hedeflenen amaca yönelik olarak data tipleri ve değerleri incelenir.

-Clean data/Data temizleme

Seçtiğimiz dataya yapılacak olası temizleme çalışmaları belirlenir ve temizleme sonucunda kaybolacak data tahmini yapılır.Karşımıza çıkan sorunlara yapacağımız temizleme çalışmaları, aldığımız kararlar tanımlanır.

-Construct data/datayı toparlamak

Bu task içinde datayı toparlamak için yaptığımız eklemeler veya değişen değerler belirlenir. Elimizdeki datayı kullanarak oluşturduğumuz yeni setler veya sonradan oluşturulan datalar belirlenir. Örneğin alan=genişlik*uzunluk

-Integrate data/Datayı bütünleştirme

Bu task içinde birden fazla tablo aynı anda ilişkilendirilebilir. Birden fazlo tablo aynı nesne için farklı bilgiler içeriyorsa bunlar birleştirilir. Bir nesnenin farklı özellikleri farklı tablolarda olabilir.

-Format data/Veriyi düzenlemek

Modelleme aracı gerektiyorsa dataya yeni bir şekil verilir. Veri madenciliği çalışmasında geliştirilen modelde kullanılan veri tabanının çok büyük olması durumunda, rastgeleliği bozmayacak şekilde örnekleme yapılması uygun olabilir. Ayrıca burada seçilen örneklem kümesinin tüm popülasyonu temsil edip etmediği de kontrol edilmelidir. Halen kullanılan işletim sistemleri ve paket programlar ne kadar gelişmiş olursa olsun, çok büyük veri tabanları üzerinde çok sayıda modelin denenmesi zaman kısıtı nedeni ile mümkün olamamaktadır. Bu nedenle tüm veri tabanını kullanarak bir kaç model denemek yerine, rastgele örneklenmiş bir veri tabanı parçası üzerinde bir çok modelin denenmesi ve bunlar arasından en güvenilir ve güçlü modelin seçilmesi daha uygun olacaktır. Diğer bir deyişle modellerin performansları uygun bir karar yöntemi ile sınanmalıdır.

2.2.4. MODELLEME

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

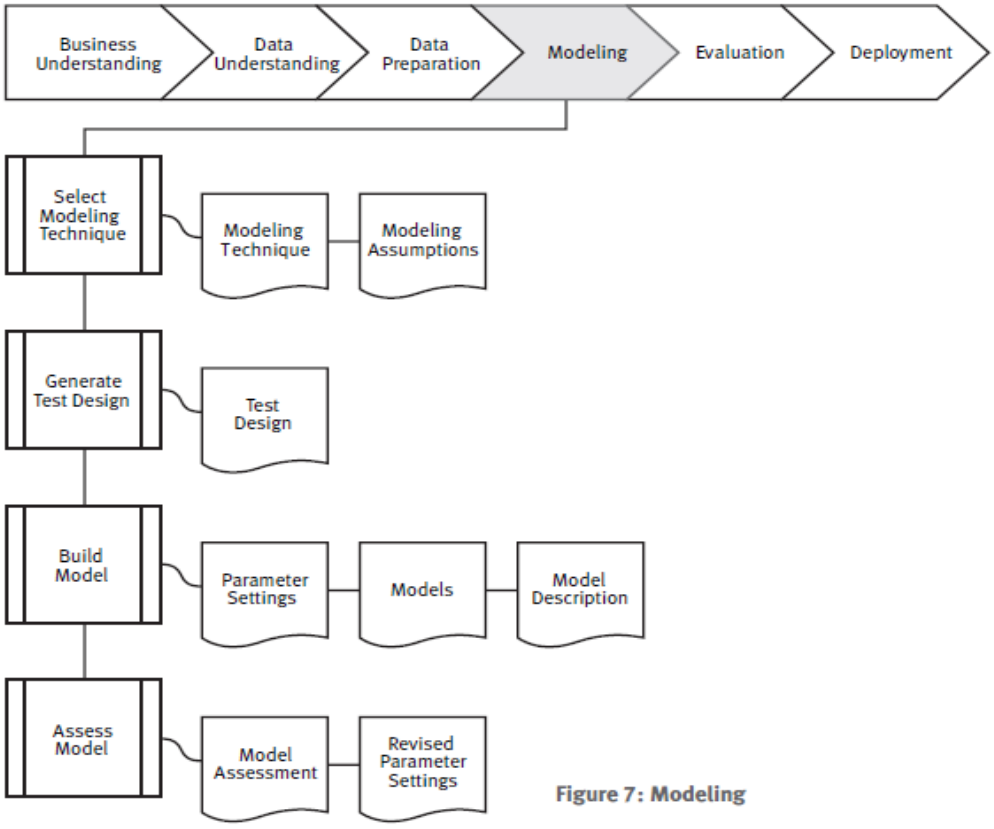
Model kuruluş süreci, denetimli ve denetimsiz öğrenmenin kullanıldığı modellere göre farklılık göstermektedir.

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

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir.

Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.

Denetimli öğrenmede seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenilmesi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenilmesi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi belirlenir.



Şekil 2.7. Modelleme

Kaynak: SPSS Help Online, CRISP-DM, (b.t.) içinde.

-Modelleme tekniğini seçmek

-Test tasarımı oluşturmak

-Modeli kurmak

-Modeli değerlendirmek

-Select modeling technique/modelleme tekniğini belirleme

Bu aşamada modelleme yapılacak teknik belirlenir. Bu iş, İş anlayış safhasında da yapılmış olabilir. Kullanılacak olan modelleme tekniği dökümanlaştırılır. Kullanılacak olan modelleme tekniği kendine has özellikler içerebilir. Örneğin boş kayıt istemeyebilir. Bunlar tespit edilir.

-Generate test design/Test dizaynı oluşturmak

Modeli oluşturmadan önce öncelikle bir test yapmamız gerekmektedir. Bu test ile hataları, eksiklikleri belirleyebiliriz. Deneyerek en iyi modelleme şekli bulunmaya çalışılır.

-Build model/Modeli oluşturmak

Belirlenen en iyi model şekli ile hazırlanan dataseti üzerinde model çalıştırılır.

-Assess model/Modeli değerlendirmek

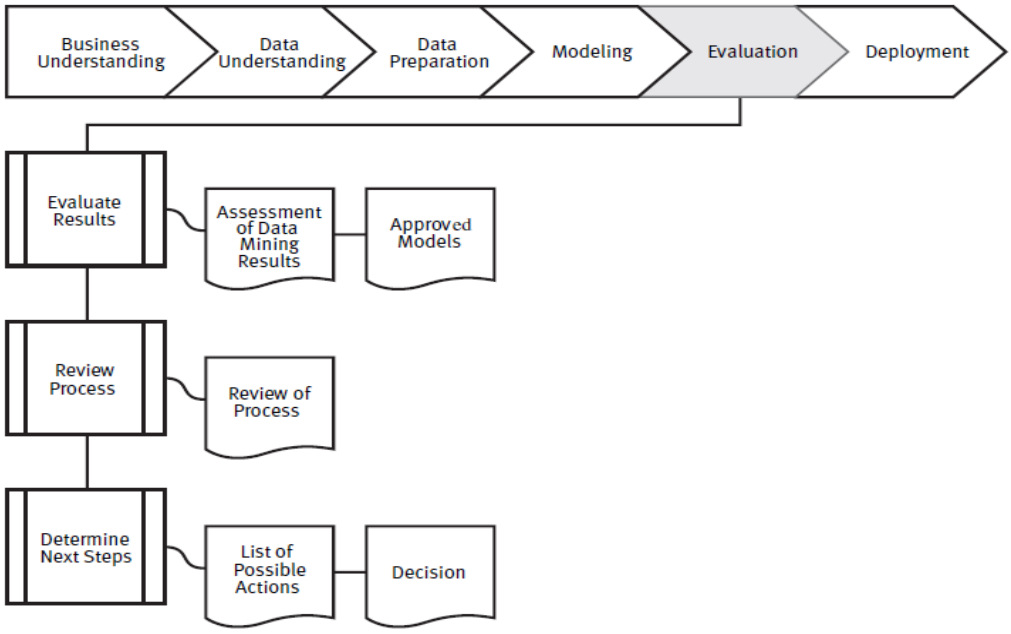
Hazırlanan model ile çıkan sonuçlar incelenerek modelin iş hedeflerine hitap edip etmediği incelenir.

2.2.5. DEĞERLENDİRME SAFHASI

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik testidir. Bu yöntemde tipik olarak 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 hesaplanır. *(Doğruluk Oranı = 1 - Hata Oranı)*

Sınırlı miktarda veriye sahip olunması durumunda, kullanılabilecek diğer bir yöntem, çapraz geçerlilik testidir. Bu yöntemde veri kümesi rastgele iki eşit parçaya

ayrılır. İlk aşamada bir parça üzerinde model eğitimi ve diğer parça üzerinde test işlemi; ikinci aşamada ise ikinci parça üzerinde model eğitimi ve birinci parça üzerinde test işlemi yapılarak elde edilen hata oranlarının ortalaması kullanılır.



Şekil 2.8. Deneme Safhası
Kaynak: SPSS Help Online, CRISP-DM, (b.t.) içinde.

-Sonuçları değerlendirmek

-Gözden geçirme süreci

-Sonraki basamakları belirlemek

Şimdi bu aşamalara yakından bakalım.

-Evaluate Results/Sonuçları değerlendirme

Bu aşamada model sonuçları, belirlediğimiz hedeflere ne kadar yaklaşmış olduğu tespit edilir. Bir eksiklik varsa onu tespit eder. Elimizdeki gerçek değerlerle karşılaştırılıp doğrulanmaya çalışılır.

-Review process/İşlemleri gözden geçirmek

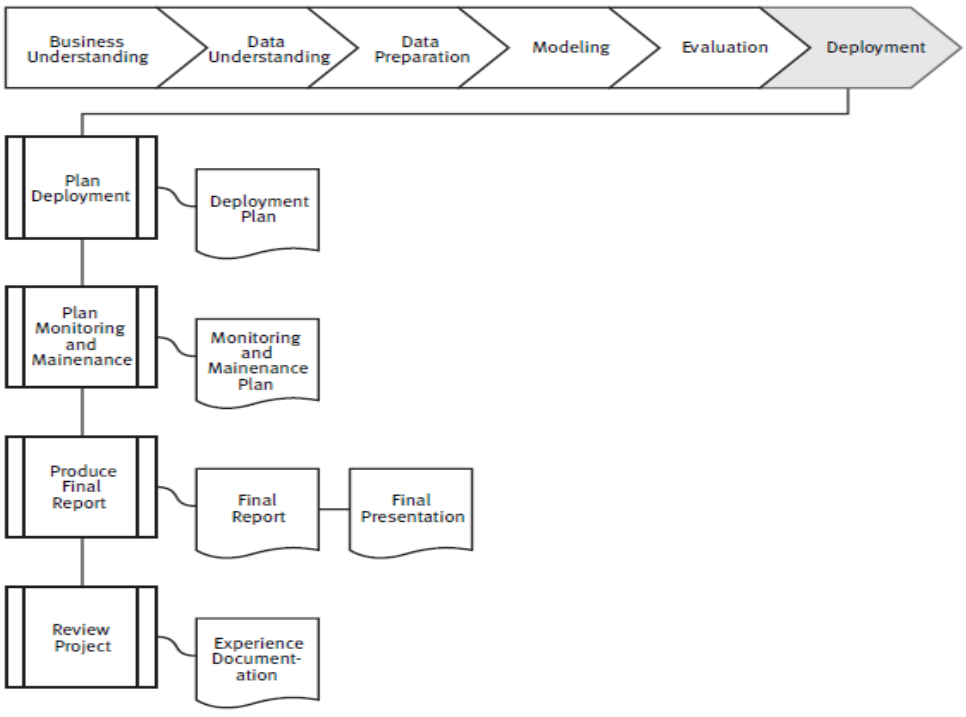
Bu aşamada çalışmanın özeleştirisi yapılır. Modeli doğru kurduk mu? Gelecekteki analizler için uygun mu vs..

-Determine next steps/Gelecek adımları belirlemek

Geçmişteki aşamaları gözeterek bu projenin geleceği konusunda karar verilir. Bu modelleme ne kadar süre kullanılacak veya yeni bir veri madenciliği çalışması gerekecek mi?

2.2.6. SAHAYA SÜRÜŞ

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir.



Şekil 2.9. Sahaya sürüş

Kaynak: SPSS Help Online, CRISP-DM, (b.t.) içinde.

-Dağıtım planı

-Plan izleme ve bakım

-Nihai rapor

-Projeyi incelemek

Bu aşamalara yakından bakalım.

-Plan deployment/Dağıtım planı

Bu task içinde, sahaya sürüş için bir strateji belirlenir. Bu çalışmanın gerekli basamaklarını içeren bir prosedür hazırlanır.

-Plan monitoring and maintenance/Plan izleme ve bakım

İzleme ve bakım uzun sureli bir proje için çok önemlidir. Ne kadar zamanda bir yenilenme yapılmalıdır, eklenen veya çıkan bir şey var mıdır gibi konulara dikkat etmek gerekir.

-Produce final report/Nihai rapor

Projenin sonunda, proje ekibi çalışmayı ve sonuçları anlatan bir rapor hazırlar. Müşteriye bir sunum veya belge il anlatılır.

- Review project/ Projeyi gözlemlemek

Bütün bu çalışmaların sonucunda bu projeyi izleyip sunmak ekibe tecrübe açısından faydalı olacaktır. Benzer projeler yaparken kolaylık sağlaması açısından ekip kendi açısından da projeyi gözlemlemelidir.

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.

2.3. VERİ MADENCİLİĞİ MODELLERİ VE MODELLEME İÇİN UYGULANAN TEKNİKLER

Daha önceki kısımlarda da bahsettiğimiz gibi veri madenciliği yapmak istediklerine göre genel olarak iki tip modelleme yapar. Veri madenciliğinin amacı, dataya bir arayüzden veya sorguyla baktığımızda bize bir sonuç ifade etmeyen yeni anlamlar çıkarmaktır. Modellemede firmanın yapmak istediği işe göre modelleme yapılır.

Firma , son dönemlerde çok sayıda müşteri kaybettiğini farkedip kaybedilen müşterilerin ortak özelliklerine bakarak ilerde kaybedebileceği müsterileri de belirlemek isteyebilir ya da müşterilerini özelliklerine göre sınıflandırabilir. Bu örnekten yola çıkarak veri madenciliğinde yapılan modellemeler iki sınıfa ayrılmıştır.

1. Tahmin edici modeller (Kaybedilecek müşterinin tahmini…)
2. Tanımlayıcı modellemeler (Müşteri Segmentasyonu…)

2.3.1.TAHMİN EDİCİ MODELLER

Tahmin edici modellerde; sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Örneğin bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır.

Tahminde bulunacağımız bir özellik için elimizde bulunan verilerden o sonuç için ortak bir formül elde edilir. Elde edilen bu formül , tahminde bulunacağımız data setine uygulanır ve sonuca gidilir. Tahminde bulunulacak özellik bağımlı değişken, onun oluşmasına etken olan diğer özellikler ise bağımsız değişkenlerdir.

Bu tahminler için formül bulunurken belli başlı yöntemler kullanılır. Bunların en önemlileri Sınıflandırma ile regresyon analizi ve zaman serileridir.

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler;

-Karar Ağaçları,

-Yapay Sinir Ağları,

-Naive Bayes,

-Bulanık Mantık,

-Bellek Temelli Nedenleme

2.3.1.1. SINIFLANDIRMA

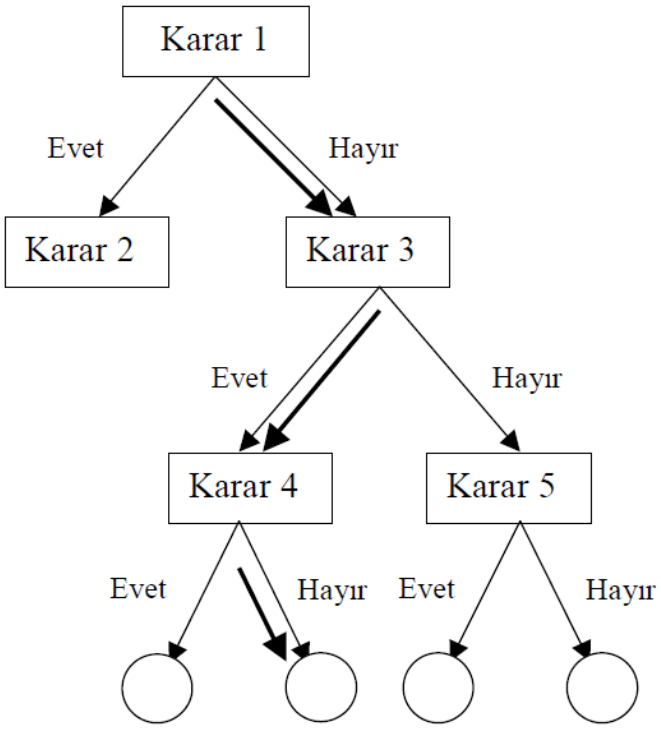
Sınıflandırmanın örüntü tanımada kaynakları vardır. Amaç yeni bir nesnenin belirli sınıflar içinde hangi sınıfa ait olup olmadığını belirleyecek bir sınıflayıcı (classifier) oluşturmaktır. Önceden belirlenmiş sınıflar, veri ambarından veya veri tabanından alınan verinin sınıflandırılması için model geliştirmede kullanılır(Bergeron, 2002, s.120).

Sınıflandırma bir ürünün özellikleri ile müşteri özelliklerinin eşlenmesi için kullanılır. Böylece bir müşteri için ideal ürün veya bir ürün için ideal müşteri profili çıkarılabilir.

Sınıflandırma modellerinin bitiminde şöyle sonuç cümleleri olur. Örneğin bir giyim firması için şehir dışında okuyan üniversite öğrencilerinin genelde indirim zamanında alışveriş yaptığı, bir digital tv yayıncısının çocuklu müşterilerinin genelde çocuk kanallarının açıldığının belirlenmesi gibi. Bu sonuçları elde eden firmalar, çok yatırım yaptığı yayınları çocukların daha az aktif olduğu bir saate koyabilir. Veya üniversite öğrencilerini kazanmak isteyen firma onlara belirli bir oranda indirim sağlayabilir.

2.3.1.1.1.KARAR AĞAÇLARI

Karar ağaçları veri madenciliğinde en sık kullanılan yöntemlerin başında gelmektedir. Bunun başlıca sebepleri ucuz olması, yorumlamalarının oldukça kolay olması ve veritabanı sistemleri ile entegre edilebilmeleridir. Karar ağaçları düğümler ve dallardan oluşan, anlaşılması oldukça kolay olan bir tekniktir. Karar ağacında bulunan her bir dalın belirli bir olasılığı mevcuttur. Bu sayede son dallardan köke veya istediğimiz yere ulaşana dek olasılıkları hesaplamamız mümkündür.

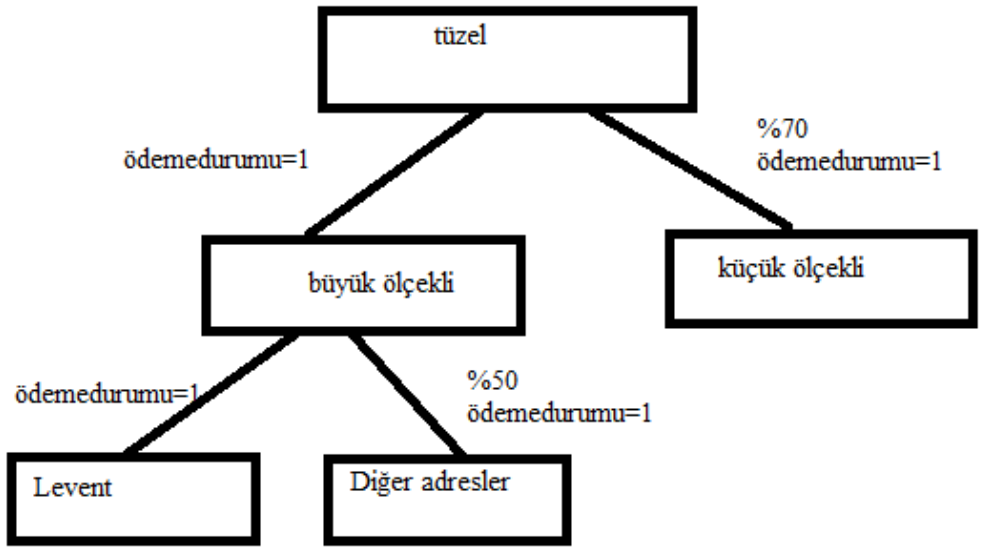


Şekil 2.10. Karar ağacı yapısı

Kaynak: Şimşek, 2006 içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

Karar ağaçlarına örnek verecek olursak, bir müşterinin faturasını zamanında ödeme ve ödememe durumunu ele alalım. ÖdemeDurumu=1 ise ödeme yapılmış, ÖdemeDurumu=0 ise ödeme yapılmamış demektir.Yaptığımız bu karar ağacı çalışmasında tüzel şirketlerin ÖdemeDurumu=1, özel kişilerin ise ödeme durmunda daha düzensiz olduğu görülmüştür. Bunun yanısıra büyük ölçekli tüzel müşterilerin daha düzenli öde-

me yaptığı , küçük ölçekli şirketlerin ise Levent tarafında olanların daha düzenli ödeme yaptığı anlaşılmıştır. Bu dallanma çeşitli özelliklere göre devam eder.



Şekil 2.11. Bir karar ağacı örneği (Bu şekil uygulama çalışmasından alınmıştır.)

Karar ağacı için çeşitli teknikler vardır. Bunlardan birkaçına aşağıda göz atalım.

2.3.1.1.1.1. CART

CART veya C&RT (Classification and Regression Trees) Breiman, Friedman, Olshen ve Stone tarafından 1984 yılında getişrilmiş ikili (binary) ağaç olarak büyüyen bir algoritmadır. C&RT veriyi iki alt kümeye ayırır. Böylece bir sonraki adımda oluşacak olan alt küme, bir öncekinden daha homojen olacaktır. Bu süreç sonuç bulunana kadar devam eden, kendini tekrarlayan bir süreçtir. C&RT Algoritması karmaşık bir algoritmadır. Büyük verilerle çalışıldığında sonuç bulması uzun sürmektedir. C&RT sınıflandırma ve regresyon analizi için kullanılan bir algoritmadır (Answer Tree 3.1 User.s Guide, b.t.).

2.3.1.1.1.2. CHAİD

CHAID (Chi-Squared Automatic Interaction Detector) algoritması 1980 yılında Kaas tarafından geliştirilmiş oldukça başarılı bir karar ağacı tekniğidir. CHAID adından

da anlaşılacağı gibi ayırma kriteri olarak ki-kare' yi kullanır. CHAID algoritması, tahmin edici değişkenin tüm değerlerini dikkate olarak analiz yapar. Hedef değişkeni dikkate alarak istatistik olarak benzer olan değişkenleri birleştirir ve farklı olan değişkenle işlemi sürdürür. Daha sonra karar ağacının ilk dalını oluşturmak için en iyi tahmin edici değişkeni seçer. Her bir düğüm seçilen değişkenin benzer değerlerinden oluşur. Bu süreç ağaç tamamıyla büyüyene kadar tekrarlanarak devam eder. Yapılacak testler hedef değişkenin türüne göre değişmektedir. Eğer değişken sürekli bir değişken ise F testi, kategorik (nominal/ordinal) bir değişken ise ki-kare testi kullanılır.

CHAID en popüler karar ağacı metotlarından biridir. CHAID algoritması ikili bir algoritma değildir, ki bu ağaçta herhangi bir seviyede ikiden çok kategori üretmesi anlamına gelir. Bu nedenle daha geniş ağaç yaratmaya eğilimlidir. Her tür değişken için kullanılan bir tekniktir(Tezcanlar,2007).

2.3.1.1.1.3. C4.5

C4.5 algoritması en iyi karar ağacı algoritmasıdır. 1993 yılında Quinlan tarafından ortaya atılmıştır. Karar ağacı oluşturulurken kayıp veriler hesaba katılmaz.196 C4.5 algoritması, kalitatif değişkenleri dikkate alır. Ayrıca kayıp verileri diğer veri ve değişkenler yardımı ile tahmin ederek, daha hassas ve daha anlamlı kurallar çıkartabilen bir ağaç üretebilir(Tezcanlar,2007).

2.3.1.1.1.4. QUEST

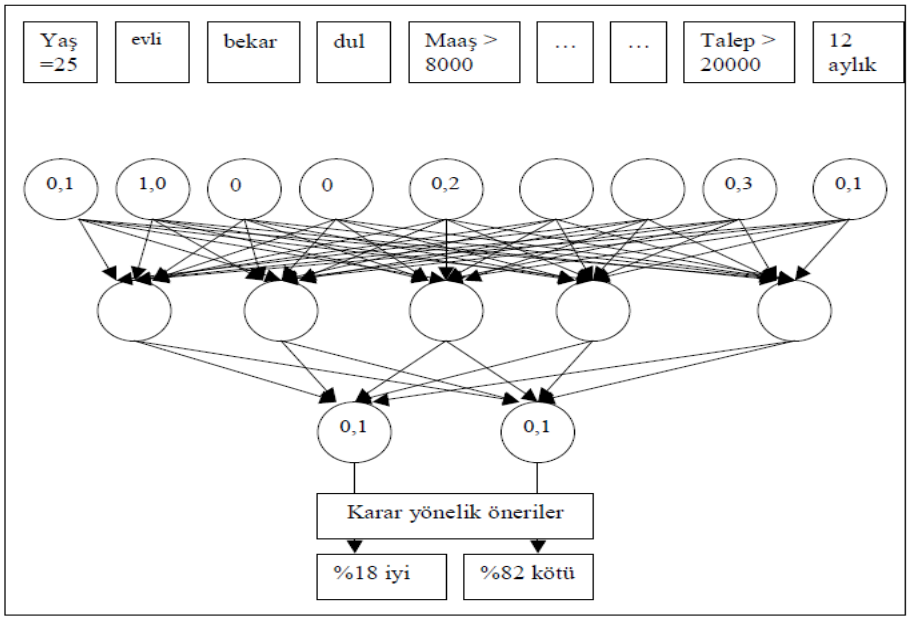
Quest (Quick, Unbiased, Efficient, Statistical Tree) 1997 yılında Loh ve Shih tarafından geliştirilmiş olan yeni bir tekniktir. Binary (ikili) büyüyen bir ağaç algoritmasıdır. Ayrı ayrı değişken seçimi ile ilgilenir. QUEST.deki birim değişken ayıracı, tahmini olarak tarafsız değişken seçimini gerçekleştirir. QUEST algoritmasının C&RT algoritmasına benzer avantajları vardır, ancak ağaçlar yavaş büyüyebilir ve ikili olduğu için karar ağacı çok geniş olabilir(Tezcanlar, 2007).

2.3.1.1.2. YAPAY SİNİR AĞLARI

Yapay sinir ağları (YSA), insan beyninin sinir sistemine ve çalışma prensibine dayanan elektriksel bir modeldir. Bir anlamda insan beyninin ufak bir kopyası gibidir. İnsan beyninin öğrenme yoluyla yeni bilgiler üretebilme, keşfedebilme, düşünme ve gözlemlemeye yönelik yeteneklerini, yardım almadan yapabilen sistemler geliştirmek için tasarlanmışlardır. Yapay Sinir ağı ile hesaplamalarda istenilen dönüşüm için, adım adım yürütülen bir yöntem gerekmez. Sinir ağı ilişkilendirmeyi yapan iç kuralları kendi üretir ve bu kuralları, bunların sonuçlarını örneklerle karşılaştırarak düzenler. Deneme ve yanılma ile, ağ kendi kendine işi nasıl yapması gerektiğini öğretir.YSA'larda bilgi saklama, verilen eğitim özelliğini kullanarak eğitim örnekleri ile yapılır. Sinirsel hesaplama, algoritmik programlamaya bir seçenek oluşturan, temel olarak yeni ve farklı bir bilgi işleme olayıdır. Uygulama imkanının olduğu her yerde, tamamen yeni bilgi işleme yete-nekleri geliştirebilir. Bu sayede de geliştirme harcamaları ile geliştirme süresi büyük ölçüde azalır (Tok, 2002).

Bir yapay sinir ağı belirli bir amaç için oluşturulur ve insanlar gibi örnekler sayesinde öğrenir. Yapay sinir ağları tekrarlanan girdiler sayesinde kendi yapısını ve ağırlığını değiştirir. Yapay sinir ağları aynen canlıların sinir sistemi gibi adapte olabilen bir yapıya sahiptir.

Aşağıda bir YSA uygulaması görülmektedir. Bir banka müşterilerinin borçlarından dolayı oluşan riskleri tahmin etmeyi hedeflemektedir. Banka, müşterilerinin adı, yaşı, medeni hali, gelir düzeyi gibi bilgiler dışında son borç bilgilerine de sahiptir. Bu veriler ışığında bir model oluşturulur.



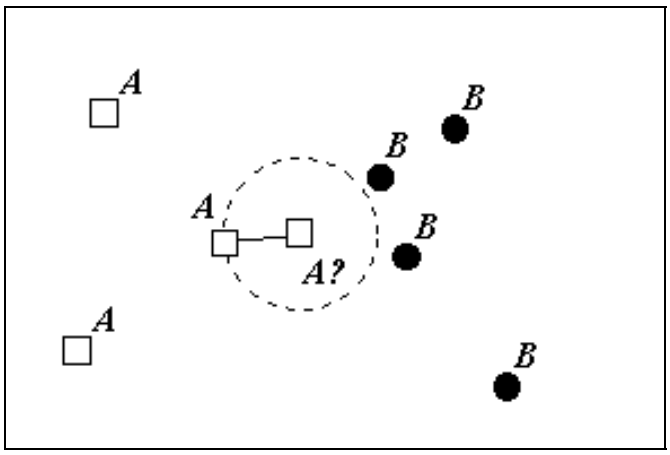
Şekil 2.12. Yapay sinir ağları uygulaması

Kaynak: Tok, 2002 içinde. Müşteri İlişkileri Yönetimi ve Veri Madenciliği' den alınmıştır.

Tez çalışmasının uygulama kısmında kullanılacak olan KOHONEN algoritması da bir sinir ağı çalışmasıdır.

2.3.1.1.3. BELLEK TABANLI YÖNTEMLER

Bellek tabanlı veya örnek tabanlı bu yöntemler istatistikte 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılamamış ama günümüzde bilgisayarların ucuzlaması ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur. Bu yönteme en iyi örnek en yakın k komşu algoritmasıdır. En yakın komşu yaklaşımı, x noktasının sınıfını, x noktasına en yakın olan noktanın sınıfı olarak belirleme yaklaşımıdır. Sınıfı belirlenen nokta ile komşu nokta aynı sınıfa ait değiller ise hata söz konusudur. Bu yaklaşım sadece en yakın komşu ile sınıflandırma yapar, önceden sınıflandırılmış diğer noktaları önemsemez.



Şekil 2.13. Bellek tabanlı yöntem

2.3.1.1.4. NAIVE BAYES

Naive Bayes algoritmasında her kriterin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır.

Özellikle müşterilerin belirli bir özelliğe göre sınıflandırılmasında kullanılan bir tekniktir. Belirsiz durumların olasılık teorisine dayalı olarak çıkarılması tekniğidir. Çeşitli sebeplerin aynı sonucu verdiği durumlarda, bazen sonuç bilindiği halde, bunun hangi sebepten dolayı meydana gelmiş olabileceği bilinmeyebilir. Söz konusu sonucun hangi olasılıkla hangi sebepten kaynaklandığı araştırılmak istendiğinde Bayes teoreminden yararlanılmaktadır (Serper, 1997, s.202).

2.3.1.1.5. BULANIK MANTIK

Belirsizliklerin anlatımı ve belirsizliklerle çalışılabilmesi için kurulmuş katı bir matematik düzen olarak tanımlanabilir. Bilindiği gibi istatistikte ve olasılık kuramında, belirsizliklerle değil kesinliklerle çalışılır ama insanın yaşadığı ortam daha çok belirsizliklerle doludur. Bu yüzden insanoğlunun sonuç çıkarabilme yeteneğini anlayabilmek için belirsizliklerle çalışmak gereklidir. Bulanık mantığın uygulama alanları çok geniştir. Sağladığı en büyük fayda ise "insana özgü tecrübe ile öğrenme" olayının kolayca modellenebilmesi ve belirsiz kavramların bile matematiksel olarak ifade edilebilmesine olanak

tanımasıdır. TAI'de araştırma gelişme kısmında bulanık mantık konusunda çalışmalar yapılmaktadır.

2.3.1.2. REGRESYON ANALİZİ VE ZAMAN SERİLERİ ANALİZİ

Regresyon Analizi: Herhangi bir değişkenin (bağımlı değişken), bir veya birden fazla değişkenle (bağımsız-açıklayıcı) arasındaki ilişkinin matematik bir fonksiyon şeklinde yazılması ve bu fonksiyon yardımıyla bağımlı değişkenin ulaşacağı değerin tahmin edilmesidir(Orhunbilge, 1996, s. 9). Örneğin spor araba aksesuarlarının satış hacmi, bir önceki ay satılan spor ara-baların sayısına bağlı olarak hesaplanabilir.

Zaman Serileri Analizi: Regresyon analizinden farklı olarak, sadece zamana bağımlı olan değişkenler tahmin edilmeye çalışılır. Örneğin tatil döneminde meydana gelen kazaların oranı, bir önceki yılın aynı tatil döneminde meydana gelmiş kazaların sayısıyla belirlenmeye çalışılır.

2.3.2.TANIMLAYICI MODELLER

Veri madenciliğinde sonuca gidilirken kullanılan bir başka model sınıfı ise tanımlayıcı modellerdir. Tanımlayıcı modellerde; işe karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. Tanımlayıcı modeller kümeleme ve birliktelik kurallarıdır.

Müşterileri gösterdiği ortak özelliklere göre kümelemek veya hangi ürünlerin beraber satıldığını belirlemek bu modelleme çeşidine örnek olabilecek çalışmalardır.

Bu model sınıfında kullandığımız teknikleri de yukarıda bahsettiğimiz gibi 2 kısıma ayırabiliriz.

-Kümeleme Analizi

-İlişki Analizi (Birliktelik Kuralları ve Ardışık Örüntüler)

2.3.2.1. KÜMELEME ANALİZİ

Tanımlayıcı modellerde kullandığımız yöntemlerden biri kümeleme analizi dediğimiz yöntemlerdir. Bu yöntemleri, basitçe elimizdeki veriyi gösterdiği benzer özelliklere göre gruplandırma olarak tanımlayabiliriz.

Kümeleme, verideki benzer kayıtların gruplandırılmasını sağlayan bir tekniktir. Hangi yöntem kullanılırsa kullanılsın süreç aynı şekilde işler. Her kayıt var olan kümelerle karşılaştırılır. Bir kayıt kendisine en yakın kümeye atanır ve bu kümeyi tanımlayan değeri değiştirir.

Kümeleme Analizi, çok boyutlu uzayda birbirine yakın olan gözlemlerden meydana gelen grupları veya kümeleri bulmayı amaçlamar. Analiz, örneklem verilerini gözlemlerin benzerliklerine göre en uygun kümelere ayırmaktadır. Diğer çok değişkenli istatistiksel analiz yöntemlerinde önemli bir yer tutan normallik varsayımı, bu analizde prensipte kalmakta ve uzaklık değerlerinin normalliği yeterli görülmektedir(Özçakır, 2006).

Kısaca, kümeleme analizi gruplar veya kümeler içine gözlemleri birleştirmede kullanılan bir tekniktir. Bu çalışmanın çıkan sonuçlarında:

1-Her bir grup veya küme belirli bir özelliğe göre homojendir. Yani, her bir gruptaki gözlemler bir diğerine benzerdir.

2-Her bir grup aynı özelliklere göre diğer gruplardan farklı olmalıdır. Yani, bir grubun gözlemleri diğer grupların gözlemlerinden farklı olmalıdır.

Kümeleme; iki gözlemin benzerlikleri (yakınlıkları) veya benzemezlikleri (uzaklıkları) temel alınarak yapılır (Johnson, Wichern, 1992, s.573). Kümeleme yöntemleri, uzaklık matrisi ya da benzerlik matrisinden yararlanarak birimler ya da değişkenleri kendi içinde homojen ve kendi aralarında heterojen gruplar oluşturmayı sağlayan yöntemlerdir.

Kümeleme algoritmaları, kümeleme yaparken izledikleri yaklaşımlara göre çeşitli gruplara ayrılırlar.

-Aşamalı Kümeleme Yöntemleri (Hierarcihal Cluster Analysis Method)

-Aşamalı olmayan Kümeleme Yöntemleri (Non-Hierarchical Cluster Analysis Method)

-Kohonen Ağları , kendi kendini düzenleyen haritalar (SOM)

olarak ayırabiliriz.

Eğer analiz edilecek veriler aralıklı veya orantılı ölçüm düzeyinde ölçülmüş ise, en çok kullanılan uzaklık ölçüleri; Öklit, Karelerialınmış Öklit, Minkowski ve Manhattan City-Blokdur. Eğer veriler sınıflayıcı veya sıralayıcı ölçüm düzeyinde ölçülmüş ise kullanılan uzaklık ölçüleri; Ki-Kare (χ^2) ve normalleştirilmiş ki-kare olarak bilinen Phi-Kare (φ) dir.

Kümeleme için kullanılan belli başlı algoritmaların arasında K-means, Kohonen, TwoStep gibi teknikleri sayabiliriz. Bir sonraki bölümde bunlara da göz atacağız.

2.3.2.2. İLİŞKİ ANALİZİ (BİRLİKTELİK KURALLARI VE ARDIŞIK ÖRÜNTÜLER)

Birliktelik kuralları, büyük veri kümeleri arasında birliktelik ilişkileri bulurlar. Toplanan ve depolanan verinin her geçen gün gittikçe büyümesi yüzünden, şirketler veritabanlarındaki birliktelik kurallarını ortaya çıkarmak istemektedirler. Büyük miktardaki mesleki işlem kayıtlarından ilginç birliktelik ilişkilerini keşfetmek, şirketlerin karar alma işlemlerini daha verimli hale getirmektedir.

Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır.Bu işlem, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını analiz eder. Bu tip birlikteliklerin keşfedilmesi, müşterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarır ve market yöneticileri de bu bilgi ışığında daha etki satış stratejileri geliştirebilirler.

Örneğin bir müşteri süt satın alıyorsa, aynı alışverişte sütün yanında bebek maması alma olasılığı nedir? Bu tip bir bilgi ışığında rafları düzenleyen market yöneticileri

ürünlerindeki satış oranını arttırabilirler. Örneğin bir marketin müşterilerinin süt ile birlikte bebek maması satın alan oranı yüksekse, market yöneticileri süt ile bebek maması raflarını yan yana koyarak mama satışlarını arttırabilirler.

Birliktelik kuralları geçmiş tarihli hareketleri analiz etmek için karar destek sistemlerinde stratejik karar verme aşamasında örüntüleri ve ilişkileri bulmada, verilen kararların kalitesini arttırmada izlenen bir yaklaşımdır. Birliktelik kuralları eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılır. Birliktelik kurallarının amacı, kullanıcı tarafından belirlenen minimum olasılık ve koşullu olasılık değerlerini sağlayan kuralların bulunmasıdır. Keşfedilen örüntüler örneklemde sıklıkla birlikte geçen nitelik değerleri arasındaki ilişkiyi gösterir. Birliktelik kurallarının uygulandığı alanlardan biri de sepet analizi çalışmalarıdır.

Birliktelik-ilişki kuralı madenciliği 2 aşamalıdır:

- Tüm sık geçen nesne kümelerinin bulunması: Tanıma gore her nesne kümesinin sık geçenler kumesinde yer alabilmesi için, her nesnenin destek değerinin onceden tanımlanmış olan minğdestek değ erinden büyük olması gerekir.

- Sık geçen nesne kümelerinden güçlü ilişki kurallarının yaratılması: Tanıma göre, bu kurallar min. destek ve min. güven durumunu sağlamalıdır(Han ve Kamber, 2000)

Birliktelik-ilişki kuralının uygulandığı algoritmaların bazıları şunlardır; AIS, Apriori, DHP, Partition.

2.3.2.2.1. AIS ALGORITMASI

AIS (Agrawal, Imielinski ve Swami) algoritmasında üretilen eşleştirme kurallarının sağ kesiminde sadece bir elemanlı ürünler kumesi yer alabilmektedir (Agrawal v.d., 1993). AIS algoritmasının tersine diğer algoritmalar birden fazla elemana sahip ku-rallar üretebilmektedir. Apriori algoritmasında k öğeli sık geçen oğe küme adayları, (*k-1*) öğeli sık geçen öğe kumelerinden faydalanılarak bulunur (Şimşek, 2006).

2.3.2.2.2. APRIORI ALGORITMASI

Apriori, boolean ilişki kuralları için geçerli bir veri madenciliği algoritmasıdır. Algoritmanın ismi, bilgileri bir önceki adımdan aldığı için "prior" anlamında Apriori'dir. Bu algoritma özünde iteratif (tekrarlayan) bir niteliğe sahiptir(Han ve Kamber, 2000) ve hareket bilgileri içeren veritabanlarında sık geçen öğe kumelerinin keşfedilmesinde kullanılır.

Sık geçen öğe kümelerini bulmak için birçok kez veritabanını taramak gerekir. İlk taramada bir elemanlı minumum destek ölçütünü sağlayan sık geçen öğe kümeleri bulunur. İzleyen taramalarda bir onceki taramada bulunan sık geçen öğe kümeleri aday kumeler adı verilen yeni potansiyel sık geçen öğe kümelerini üretmek için kullanılır. Aday kümelerin destek değerleri tarama sırasında hesaplanır ve aday kümelerinden minumum destek ölçütü sağlayan kümeler o geçişte üretilen sık geçen öğe kümeleri olur. Sık geçen öğe kumeleri bir sonraki geciş için aday küme olurlar. Bu süreç yeni bir sık geçen öğe kumesi bulunamayana kadar devam eder.

Bu algoritmada temel yaklaşım eğer k-öğe kumesi minumum destek ölçütünü sağlıyorsa, bu kümenin alt kümeleri de minimum destek ölçütünü sağlar.

2.2.2.2.3. DHP ALGORITMASI

DHP (Direct Hashing and Pruning) algoritması da k-öğe kümesi adaylarını k-1 elemanlı sık geçen öğe kümelerinden elde eder, Apriori algoritmasından farklı olarak sık geçen küme adaylarını (arama uzayını) azaltır. DHP algoritması da veritabanının bir çok kere taranmasını gerektirir(Zaki ve Ogihara, 1998).

2.2.2.2.4. PARTITION ALGORİTMASI

Partition algoritması giriş/çıkış işlemlerini, veritabanını sadece iki kez okuyarak en aza indirger. Bu algoritma veritabanını bellekte ele alınabilecek küçük parçalara böler. İlk geçişte potansiyel olarak sık geçen öğe kümelerini bulur, ikinci geçişte ise öğe kümelerinin destek değerleri hesaplanır(Chen v.d., 1996, s. 866-883).

3.KÜMELEME ANALİZİ

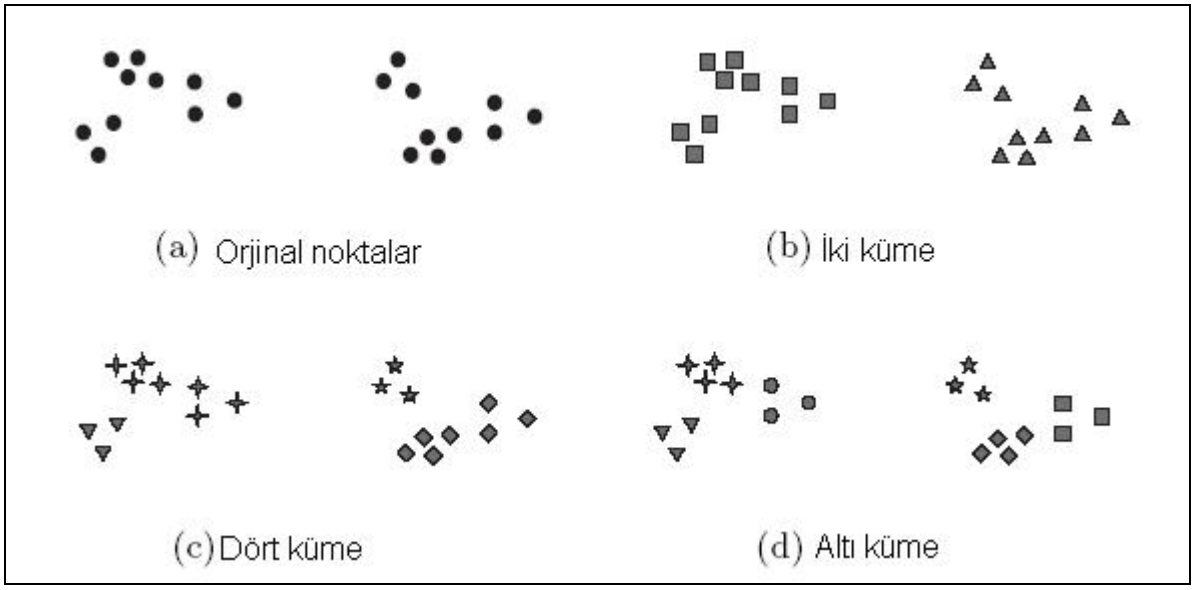
Kümeleme analizlerinin, veri madenciliğinde tanımlayıcı model olarak kullanıldığından yukarıdaki kısımda kısaca bahsedilmişti.Bu bölümde kümeleme tekniklerini, kriterlerini incelenmeye çalışılacaktır.

Kümeleme analizi veriyi anlamlı, yararlı yada hem anlamlı hem de yararlı gruplara (kümelere) ayırır. Bununla birlikte verinin özetlenmesi gibi bazı durumlarda, kümeleme analizi başka amaçlar için bir başlangıç noktasıdır.

Kümeleme analizi veri nesnelerini yalnızca nesneleri tanımlayan ve ilişkilerini ortaya koyan verilerden çıkarılacak bilgiler ışığında gruplar. Amaç aynı grup içerisindeki nesnelerin birbirine benzer veya ilişkili olması; farklı gruptakilerin ise birbirinden farklı olması yada ilişkilerinin bulunmamasıdır. Aynı gruptakilerin birbirine benzeme oranı yada farklı gruptakilerin ise birbirinden farklı olma oranları kümelemenin ne kadar iyi olduğunun yada kümelerin birbirlerinden ne kadar kesinlikle ayrıldıklarının göstergesidir (Şimşek, 2006).

Çok değişkenli istatistiksel tekniklerden birisi olan kümeleme analizi, grup sayısı kullanıcı tarafından belirlenen veya bilinmeyen ve gruplandırılmamış verilerin benzerliklerine göre sınıflandırılması amacıyla kullanılmaktadır. Kümeleme analizi verilerin birimlere veya değişkenlere göre birbirlerine benzerlikleri bakımından ayrık kümelerde toplanmasını sağlayan bir tekniktir. Kümeleme analizi birbirine benzer olan bireylerin aynı gruplarda toplanmasını amaçlaması bakımından diskriminant analizi ile, birbirine benzer değişkenlerin aynı gruplarda toplanmasını amaçlaması nedeniyle de faktör analizi ile benzerlik göstermekte olup veri indirgeme özelliği vardır.

Kümeleme işlemi yukarıda da açıklandığı gibi belirlenen amaca göre, iki gözlemin veya iki değişkenin benzerlik (yakınlık) veya uzaklık ölçüsüne bakılarak yapılır.



Şekil 3.1. Ayrı noktalardan oluşan bir setin değişik yollarla kümelenmesi
Kaynak: Kümeleme analizi:Temel Kavramar ve Algoritmalar(b.t.) içinde. www.bilmuh.gyte.edu.tr/~htakci/vm/kumeleme_analizi.doc' dan alınmıştır.

Kümeleme analizi veri nesnelerini gruplara ayıran diğer tekniklerle de ilişkilidir. Örneğin kümeleme bir çeşit sınıflandırma olarak düşünülebilir öyle ki, sınıf etiketlerine göre nesnelerin etiketlerini oluşturur. Bununla beraber, kümeleme bu bilgiyi yalnızca veriden alır. Yeni ve etiketlenmemiş nesneler, daha önceden bilinen sınıf etiketlerinden oluşturulmuş bir model aracılığıyla birer sınıf etiketine sahip olurlar. Bu sebepten, kümeleme analizi kim zaman yönetilmemiş sınıflandırma olarak adlandırılır. Eğer sınıflandırma terimi veri madenciliği bağlamında herhangi bir niteleyici ile kullanılmazsa, çoğu zaman yönetilmiş sınıflandırma kastedilir.

Kesimleme (segmentation) ve bölme (fragmantation) terimlerinin de kümeleme i-çin çoğu kez eş anlamlı kullanılmalarına karşın, bu terimler genelde kümeleme analizinin geleneksel sınırları dışındaki yaklaşımlar için kullanılırlar. Örneğin bölme terimi çoğu kez grafların(graph) alt graflara ayrılması teknikleriyle bağlantılı olarak kullanılır fakat bunlar kümelemeye doğrudan bağlı değildir. Kesimleme genelde basit teknikler kullanılarak verilerin gruplara ayrılmasına karşı düşer. Bir firmanın müşterilerini gelir durumlarına göre gruplaması buna örnek olabilir. Bu tez çalışmasında yapılacak olan uygulama bir segmentasyon çalışmasıdır.

Kümeleme analizi çok değişkenli istatistik teknikler içinde yer alıp birim veya değişkenleri benzerliklerine göre sınıflandırmaya yardımcı olan bir tekniktir (Everitt, 1974). Gruplandırma işlemi açısından bakıldığında diskriminant analizi ile kümeleme analizi ara-sında benzerlik olmakla birlikte diskriminant analizinde analizin en başından itibaren küme sayısının bilinmesinin yanında bu analizde elde edilen bilgiler gelecekte de kulla-nılabilmektedir. Oysa kümeleme analizinde başlangıçta küme sayısı bilinmemekle birlikte veriler sadece mevcut durumlar ile ilgili bilgi verdiği için gelecekte kullanılamamaktadır (Tatlıdil, 1992).

Kümeleme analizinin kullanılmasında benzerlik uzaklıklar dikkate alınarak yararlanılabilecek çok fazla alternatif ölçü ve yöntem bulunmaktadır. Örneğin sadece birimler arası uzaklıklar için Euclidyen, Kareli Euclidyen, Standardize Euclidyen, Manhattan Mahalanobis, Minkowski veya Canberra ölçüleri kullanılabilmektedir*(*Kümeleme analizi:Temel Kavramar ve Algoritmalar,Anonim,b.t*)*.

Kümeleme analizinde N adet gözlemin her birinde p adet ölçümün yapıldığı Nxp boyutlu veri matrisi aşağıdaki gibi gösterilebilir(Çakmak v.d., b.t.):

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots. & x_{2p} \\ \dots & \dots & \dots. & \dots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{bmatrix}$$

Burada x_{ij}, j. değişkenin i. birey ya da nesne için aldığı değeri gösterir.

Uzaklık fonksiyonları

d(xi,xj) fonksiyonu negatif olmayan bir fonksiyon olup; x_i ve x_j gözlem vektörleri arasındaki uzaklığı ifade eder. Uzaklık fonksiyonuna ilişkin aşağıdaki durumlar söz konusu olabilir (Duran ve Odell, 1974,3):

Tablo 3.1. Etkin ve sık kullanılan uzaklık fonksiyonları görülmektedir.
Kaynak: Çakmak v.d. (b.t.) içinde. Kümeleme Analizi Teknikleri ile İllerin Kültürel Yapılarına Göre Sınıflandırılması ve Değişimlerin İncelenmesi' den alınmıştır.

Fonksiyon	Matematiksel gösterim
Öklit	$d_2(x_i, x_j) = \left(\sum_{k=1}^{p} (x_{ki} - x_{kj})^2 \right)^{1/2}$
B_1 norm	$d_1(x_i, x_j) = \left(\sum_{k=1}^{p} \lvert x_{ki} - x_{kj} \rvert \right)$
Sup-norm	$d_\infty(x_i, x_j) = svp\{\lvert x_{ki} - x_{kj} \rvert\}$
B_p norm	$d_p(x_i, x_j) = \left(\sum_{p=1}^{k} \lvert x_{ki} - x_{kj} \rvert^p \right)^{1/p}$
Mahalanobis	$D^2(x_i, x_j) = (x_i - x_j)^T w^{-1} (xi - xj)$

Kümeleme analizi çok sayıda değişik işlevi yerine getiren yöntemler topluluğudur. Bu nedenle farklı amaçlar için farklı yöntemler uygulanmaktadır. Ayrıca değişkenlerin ölçüm birimlerinin ve ölçümleme tekniklerinin farklı olmasından dolayı, birimlerin benzerliklerinin ortaya konmasında da değişik ölçüler kullanılmaktadır.

Kümeleme analizi çalışmalarında karşımıza çıkan bazı özellikler vardır. Karşılaştırmalı olarak bu özelliklere kısaca göz atalım.

1.İç içe olan veya iç içe olmayan (bölmesel) kümeleme: Üzerinde en çok tartışmanın yapıldığı kümeleme türlerini birbirinden ayrıma kriteri onların iç içe olup olmadıkları ile ilgilidir, ya da daha geleneksel bir ifade ile hiyerarşik yada bölmesel olmaları ile ilgilidir. Bir bölmesel kümeleme basitçe veri nesnelerinin örtüşmeyen alt kümelere ayrılmasıdır öyle ki; her bir veri nesnesi yalnızca bir kümede bulunur.

2. Seçkin, örtüşen ve bulanık kümeleme. Bazı durumlarda bir noktanın birden fazla kümede yer alması mantıklı olabilir ve bu durumlar seçkin olmayan kümeleme ile daha

iyi açıklanabilir. En genel şekilde, bir örtüşen ve seçkin olmayan kümeleme bir nesnenin aynı anda birden fazla gruba(sınıfa) ait olmaları gerçeğini ortaya çıkarmada kullanılır. Örneğin; bir kişi üniversitede hem bir öğrenci hem de bir çalışan olabilir. Seçkin olmayan kümeleme aynı zamanda bir nesnenin birden fazla kümeye ait olabilmesi durumunda bunlardan herhangi birine konması için de kullanılabilir

3. Bulanık kümelemede, bir nesne belirli bir ağırlık değeriyle tüm kümelere ait olur. Bu ağırlık değeri 0(hiç ait olmama) ile 1(tamamıyla aitlik) ararsında değerler alır. Diğer bir deyişle, kümeler mantık setleri olarak ele alınırlar (Matematiksel olarak bir bulanık set içinde bir nesne herhangi bir sete 0 ile 1 arasında değerler alan bir ağırlık değeriyle aittir. Bulanık kümelemede, bir nesne için toplam ağırlık değerinin 1 olması gibi bir kısıt ortaya koyarız.). Benzer şekilde, olasılı kümeleme teknikleri de her bir noktanın her bir kümeye aitliğine dair bir olasılık hesaplar ve bu olasılıklar toplamı da 1 olmak zorundadır. Üyelik ağırlıklarının yada olasılıkları toplamının 1 olması sebebiyle, bulanık yada olasılı kümeleme gerçek birden fazla sınıflandırma(ture multiclass) durumunu açıklamazlar, örneğin bir öğrenci çalışanı durumunda bir nesne birden çok sınıfa aittir. Bunun yerine, bu yaklaşımlar bir nesnenin rasgele yalnızca bir kümeye atanmasının önüne geçildiği ve aslında birden çok kümeye yakın olduğu durumlar için elverişlidir. Pratikte, bir bulanık yada olasılı kümeleme bir seçkin kümelemeye dönüştürülür; şöyle ki bir nesne ağırlığının yada olasılık değerinin en fazla olduğu kümeye atanır.

4. **Tam kümelemeye karşın kısmi kümeleme,** Tam kümeleme her nesneyi bir kümeye atarken; kısmi kümeleme bunu yapmaz. Kısmi kümeleme ardındaki neden bir nesnenin aslında iyi tanımlanmış bir gruba ait olamayışıyla ilgilidir. Çoğu kez bir veri seti içerisindeki nesneler, bir gürültüyü(noise), küme dışında kalmayı(outlier) yada ilgi çekmeyen bir arka planı(uninteresting background) temsil edebilir. Örneğin bazı gazete makaleleri ortak bir temayı paylaşabilir, küresel ısınma gibi; fakat bazıları da çok daha genel yada tek bir çeşit tema olabilir. Bu yüzden son ay makalelerinin en önemli başlıklarını bulmak için, yalnızca ortak bir temanın geçtiği kümeler içinde aramalar yapmak isteyebiliriz. Diğer durumlarda, nesnelerin tam bir kümelemesi istenir.Örneğin, dokümanları tarama ihtiyacı için organize etmek için kümeleme kullanan bir uygulama öyle ki, bu uygulama tüm dokümanların taramamasını garanti eder.

Literatürde pek çok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi, veri tipine ve amaca bağlıdır. Genel olarak kümeleme yöntemleri şu şekilde sınıflandırılabilir(Han ve Kamber, 2000).

1- Bölme yöntemleri (Partitioning methods)
2- Hiyerarsik yöntemler (Hierarchical methods)
3- Yogunluk tabanlı yöntemler (Density-based methods)
4- Izgara tabanlı yöntemler (Grid-based methods)
5- Model tabanlı yöntemler (Model-based methods)

Bölme yöntemlerinde, *n* veri tabanındaki nesne sayısı ve *k* olusturulacak küme sayısı olarak kabul edilir. Bölme algoritması *n* adet nesneyi, *k* adet kümeye böler *(k<=n)*. Kümeler tarafsız bölme kriteri olarak nitelendirilen bir kritere uygun olusturulduğu için aynı kümedeki nesneler birbirlerine benzerken, farklı kümedeki nesnelerden farklıdırlar (Han ve Kamber, 2000).

Biz bu çalışmada kolaylık olması açısından kümeleme tekniklerini burada iki başlık altında inceleyeceğiz.

-Hiyerarşik Kümeleme yöntemleri

-Hiyerarşik olmayan Kümeleme yöntemleri

Tez çalışmasında bir de bunlardan farklı olarak tanımlanabilecek, yapay sinir ağlarından yola çıkılarak oluşturulan yeni bir kümeleme yöntemi olan KOHONEN den de bahsedilecektir.

3.1. KÜMELEME YÖNTEMLERİ

Literatürde kullanılan birçok kümeleme tekniği mevcuttur. Teknikler birbirlerinden kümelemenin oluşturuluş şekline göre ayrıldıkları gibi, kullanılan veri tipine, yapılacak çalışmanın amacına göre de farklılık gösterirler. Kümeleme teknikleri genel olarak Hiyerarşik ve Hiyerarşik Olmayan Yöntemler olarak ikiye ayrılmaktadır (Goldstein ve Dillon, b.t., s. 167).

Kümeleme analizinin amacı, bir veya birkaç özellik açısından benzer olan nesneleri veya değişkenleri belirlemektir (Aaker, 1971, s.299).

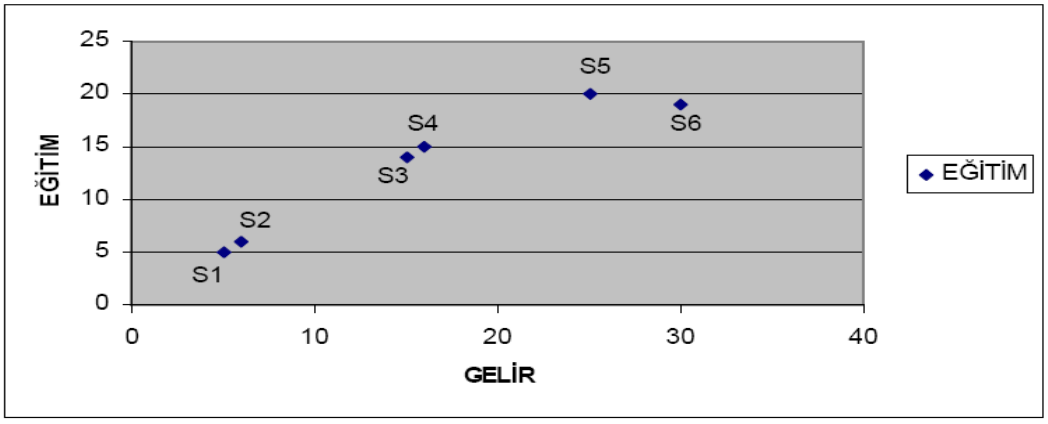
Sharma'ya göre kümeleme analizi, gözlem birimlerini grup veya kümeler olarak birleştirmek için kullanılan bir tekniktir (Sharma, 1996, s.185). Kümeleme analizinde,

1- Her bir küme homojendir veya belirli özellikler bakımından benzerdir. Dolayısıyla her bir kümedeki gözlemler birbirine benzerdir.

2- Her bir kümenin aynı özellik açısından diğer kümelerden farklı olması gerekmektedir. Böylelikle bir kümedeki gözlemler, diğer kümelerdeki gözlemlerden farklı olmaktadır.

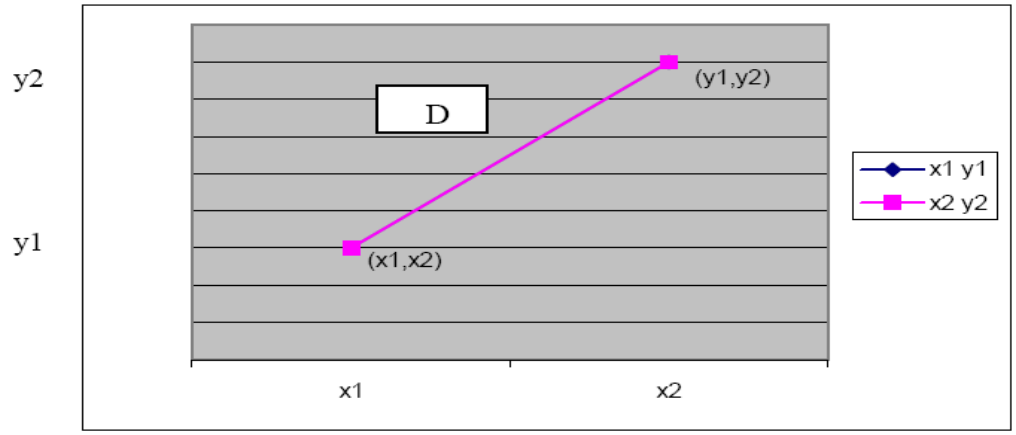
Benzerlik veya homojenlik kavramları, analizden analize farklılık göstermektedir ve yapılan çalışmanın amaçlarına bağlıdır. Amaca göre farklı sayıda ve özellikte kümeler oluşturulabilir.

Aşağıdaki şekilde bir kümeleme çalışması örneği görülmektedir.



Şekil 3.2. Serpilme diagramı

Kaynak: Şimşek, 2006 içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.



Şekil 3.3.Öklid Uzaklığı

Kaynak: Şimşek, 2006 içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

Görüldüğü gibi kümeleme analizi gözlemleri, her bir gruptaki gözlemler kümeleme değişkenini göz önüne alarak benzer olacak şekilde gruplandırmaktadır. Geometrik olarak verileri n-boyutlu gözlem uzayında göstermek ve değişkenlerin kümelerini belirlemek aynı işlemdir. Ancak çok fazla gözlem olduğunda grafik yöntemle sonuca ulaşmak pek mümkün değildir. Aynı durum gözlemlerin üçten fazla özelliği ile ilgilenildiğinde de söz konusudur (Şimşek,2006).

Kümeleme analizi yapılırken pekçok yöntem olduğundan bahsetmiştik. Kümeleme yöntemlerini 2 kısıma ayırmak istediğimizde Hiyerarşik ve Hiyerarşik olmayan yöntemler diye sınıflayabiliriz:

1. Hiyerarşik Yöntemler

- Toplaşım (Agglomerative) Kümeleme Algoritmaları

- Merkezi Kümeleme (Centroid) Yöntemi

- Tek Bağlantı (En Yakın Komşu- Single Linkage) Yöntemi

- Tam Bağlantı (En Uzak Komşu - Complete Linkage) Yöntemi

- Ortalama Bağlantı (Average Linkage) Yöntemi

- Ward Yöntemi

- Cure Yöntemi

- Agnes Yöntemi

- Bölünür (Divisive) Kümeleme Algoritmaları

- Bölünmüş Ortalamalar (Splinter- Average Distance) Yöntemi

- Otomatik Etkileşim Dedektörü (Automatic Interaction Detection- AID) Yöntemi

2. Hiyerarşik Olmayan Yöntemler

- K-ortalamalar (K-means) Yöntemi

- Metoid Parçalama Yöntemi

- Yığma Kümeleme Yöntemi

- Bulanık Kümeleme Yöntemi

3.1.1.HİYERARŞİK KÜMELEME YÖNTEMLERİ

Kümeleme yöntemleri, çıkarttığı sonuçlara Hiyerarşik(iç içe) ve Hiyerarşik olmayan yöntemler diye gruplandırılmıştır. Hiyerarşik kümeleme tekniklerinde kümeler ardarda birleştirilir ve bir grup diğeri ile bir kez birleştirildikten sonra, devam eden adımlarda bir daha ayrılmaz. Bu teknikler ele alınan değişkenler için hiyerarşik bir yapı oluştururlar. Hiyerarşik kümeleme tekniklerinde küme sayısına görsel olarak karar verilir. Bu durumda genellikle *dendogram* olarak bilinen *ağaç diyagramı* kullanılır. Çeşitli hiyerarşik kümeleme teknikleri vardır. Bunlar; Cure,Agnes, Tek Bağlantı, Tam Bağlantı, Ortanca Bağlantı, Ortalama Bağlantı ve Ward Bağlantıdır. Bu kümeleme çalışmalarında, en küçük kümedeki bir eleman aslında en üst sınıftaki kümenin de bir elemanıdır.

Hiyerarşik kümeleme algoritmaları benzerlik ve mesafe ölçülerini kullandıkları için kullanılması kolay, hemen hemen her tür veri tipine uygun ve esnek algoritmalardır. Bu yöntemde bir gözlem birimi bir kümeye dahil olduktan sonra hep bu kümede kalır.

Hiyerarşik kümelemede kümeleme işlemi, temel olarak henüz aynı kümede olmayan iki en benzer değişkeni ve onların kümelerini belirlemektir. Bu kural, kümeleme analizinin başlangıcında her değişkenin kendisi tek başına bir küme olmak üzere bütün değişkenler tek bir küme oluncaya kadar her iki kümenin birleşmesi şeklinde tekrarlanmaktadır (Şimşek, 2006).

1. Öncelikle n adet birey, n adet küme olmak üzere işleme başlanır.

2. En yakın iki küme (değeri en küçük olan) birleştirilir.

3. Küme sayısı bir indirgenerek yinelenmiş uzaklıklar matrisi bulunur.

4. 2 ve 3 numaralı adımlar n-1 kez tekrarlanır.

Kümeleme analizinde önemli sorunlardan birisi değişkenlerin birleşerek tek bir kümeye doğru gitmesi aşamasında, bu analizin hangi küme sayısında durdurulması gerektiğidir. Kümeleme analizinde ideal küme sayısı kümeler içindeki uzaklığın minimum, kümeler arasındaki farklılığın maksimum olduğu kümeleme düzeyidir. Ancak küme sayısının ne olması gerektiği konusundaki son karar, uygulamayı yapacak kişiye bırakılmıştır.

Hiyerarşik kümeleme algoritmaları benzerlik ve mesafe ölçülerini kullandıkları için kullanılması kolay, hemen hemen her türlü veri tipine uygun ve esnek algoritmalardır. Ancak özellikle bölünür algoritmalar için k küme sayısının verilmesi bir dezavantajdır. Başka bir dezavantaj ise bu kategorideki algoritmaların bir kümeyi oluşturduktan sonra, yapıları gereği oluşturdukları bu kümeyi bir daha kontrol etmemesidir.

Küme sayısına karar vermede yararlanılan en pratik yollardan birisi aşağıda verilen eşitliğin kullanılmasıdır. Hiyerarşik olmayan kümelemede, küme sayısının belirlenmesine gerek yoktur. Çünkü bu teknikte küme sayısı önceden bellidir.

$$k = \left(\frac{n}{2}\right)^2$$

n = gözlem sayısı

k = küme sayısı

Küçük örneklemler için kullanılabilir görülen bu formül örneklem hacminin büyük olması durumunda iyi sonuçlar vermemektedir. Marriot tarafından önerilen ikinci yöntemde ise küme sayısı aşağıdaki gibi hesaplanmaktadır (Çakmak v.d., b.t.).

$$M = k^2 \left| W \right|$$

W = grup içi kareler toplamı matrisi

M = küme sayısı

En küçük M sayısını veren küme sayısı gerçek küme sayısı olarak değerlendirilmektedir. Küme sayısını belirlemede kullanılan farklı yöntemler de söz konusudur. Fakat bu konuda araştırmacıların bilgi düzeyi, mesleki tecrübesi ve sonuçların anlamlı olup olmaması en önemli etkendir.

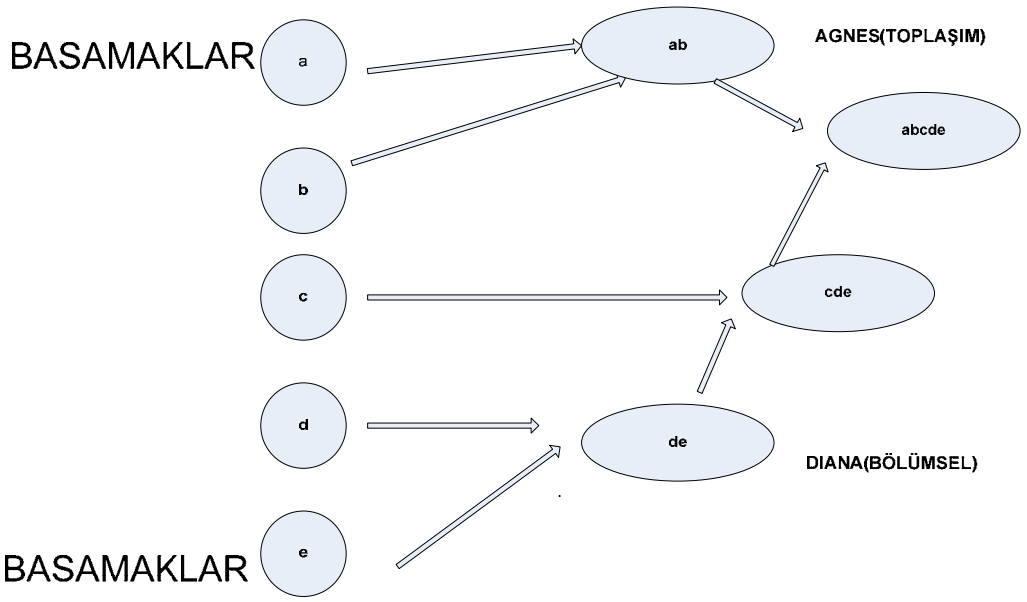
Hiyerarşik kümeleme teknikleri, toplaşım ve bölünür kümeleme algoritmaları olarak iki grupta toplanır.

Kümeleme yöntemlerinden biri olan hiyerarşik yöntemler, veri nesnelerini kümeler ağacı şeklinde gruplara ayırma esasına dayanır. Hiyerarşik kümeleme yöntemleri, hiyerarşik ayrışmanın aşağıdan yukarıya veya yukarıdan aşağıya dogru olmasına göre *agglomerative(toplasım)* ve *divisive(bölünür) hiyerarsik* kümeleme olarak sınıflandırılabilir.

Agglomerative hiyerarsik kümelemede, Sekil 3.4 de görüldüğü üzere hiyerarşik ayrışma aşagıdan yukarıya doğrudur .İlk olarak her nesne kendi kümesini oluşturur ve

ardından bu atomik kümeler birleserek, tüm nesneler bir kümede toplanıncaya dek daha büyük kümeler oluştururlar.

Divise hiyerarşik kümelemede, Sekil 3.4 de görüldüğü üzere hiyerarşik ayrışma yukarıdan aşağıya doğru olur . İlk olarak tüm nesneler bir kümededir ve her nesne tek başına bir küme olusturana dek, kümeler daha küçük parçalara bölünürler.



Şekil 3.4. Hiyerarşik Kümeleme Yöntemlerine örnek.

Kaynak: Özekeş ve Çamurcu, (2003) tarihinde Veri Madenciliğinde Karar Ağaçları Yöntemi Uygulaması' ndan alınmıştır.

Sekil 3.4., bir agglomerative hiyerarşik kümeleme yöntemi olan AGNES (Agglomerative Nesting) ve bir divise hiyerarşsik kümeleme yöntemi olan DIANA (Divise Analysis) uygulaması göstermektedir (Özekeş ve Çamurcu, 2003). Bu yöntemler beş nesneli (a,b,c,d,e) bir veri setine uygulanmaktadır. Baslangıçta AGNES her nesneyi bir kümeye yerleştirir. Kümeler, bazı kriterlere göre basamak-basamak birleşirler. Örneğin C1 ve C2 kümeleri, eğer C1 kümesindeki bir nesne ve C2 kümesindeki bir nesne ile, diğer kümelerdeki herhangi iki nesne arasında belirlenen uzaklık mesafesini karşılayacak bir mesafe varsa birleşebilirler. Bu birleşme işlemi tüm nesneler bir kümede toplanıncaya kadar devam eder (Fayyad, 1998, s.41-48). DIANA'da ise tüm nesnelerin içinde toplandığı küme, her küme bir nesne içerecek duruma gelene kadar bölünür (Fayyad, 1998, s. 41-48).

3.1.1.1. TOPLAŞIM KÜMELEME ALGORİTMALARI

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

Toplaşım kümeleme algoritmalarında başlangıçta gözlem birimleri kadar çok sayıda küme vardır. Öncelikle birbirine en çok benzeyen gözlem birimleri gruplanır. Oluşturulan gruplar benzerliklerine göre birleştirilir. Benzerlik arttıkça tüm gruplar tek bir kümede birleşme eğilimini göstermektedir (Johnson ve Wichern, 1998, s. 739).

Toplama teknikleri $\left\{\frac{1}{2}\left[n(n-1)\right]\right\}$ olası gözlem çifti arasındaki bir benzerlik veya uzaklık matrisinin hesaplanması ile başlar. Başlangıçta her gözlem bir kümedir. Benzerlik veya uzaklık matrisine göre en yakın iki küme birleştirilir. Daha sonra küme sayısı bir indirgeyerek benzerlik matrisi tekrar oluşturulur ve n birim aşamalı olarak sırasıyla n, (n-1), (n-2),...(n-r),...3,2,1 kümeye yerleştirilir.

Çeşitli toplama kümeleme yöntemlerine aşağıdaki veri seti için bir göz atalım.

Tablo 3.2. Bir veri seti

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

KONU	**GELİR ($1000)**	**EĞİTİM (YIL)**
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

3.1.1.1.1. MERKEZİ KÜMELEME YÖNTEMİ

Merkezi Kümeleme Yönteminde her bir kümenin yerine, o grubun merkezi olan bir ortalama gerekmektedir.

Örneğin 1. kümede S1 (5,5) ve S2 (6,6) noktalarıydı. Bu kümede artık gelir (5+6) / 2 = 5,5 bin dolar ve eğitim (5+6) / 2 = 5,5 yıldır.

Tablo 3.3. Merkezi Kümeleme Yöntemi:5 küme için

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

5 küme için veri seti

KÜME SAYISI	KÜMELER	GELİR	EĞİTİM
1	S1-S2	5,5	5,5
2	S3	15	14
3	S4	16	15
4	S5	25	20
5	S6	30	19

Benzerlik matrisi

	S1-S2	S3	S4	S5	S6
S1-S2	0	162,5	200,5	590,5	782,5
S3	162,5	0	2	135,96	250
S4	200,5	2	0	106	212
S5	590,5	135,96	106	0	26
S6	782,5	250	212	26	0

Minimum uzaklık değerlerini veren gözlem birimleri S3 ve S4 olduğu için, ikinci küme olarak S3-S4 kümesi seçilir. Yeniden merkezi kümeleme yöntemi uygulanır.

3.kümede S3 (15,14) ve S4 (16,15) noktalarıydı. Bu kümede artık gelir (15+16) / 2 = 15,5 bin dolar ve eğitim (14+15) / 2 = 14,5 yıldır.

TABLO 3.4.Merkezi Kümeleme Yöntemi:4 küme için

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

4 küme için veri

KÜME SAYISI	KÜMELER	GELİR	EĞİTİM
1	S1-S2	5,5	5,5
2	S3-S4	15,5	14,5
3	S5	25	20
4	S6	30	19

Benzerlik Matrisi

	S1-S2	S3-S4	S5	S6
S1-S2	0	181	590,5	782,5
S3-S4	181	0	120,5	230,5
S5	590,5	120,5	0	26
S6	782,5	230,5	26	0

Minimum uzaklık değerlerini veren gözlem birimleri S5 ve S6 olduğu için, üçüncü küme olarak S5-S6 kümesi seçilir. Yeniden merkezi kümeleme yöntemi uygulanır.

3. kümede S5 (25,20) ve S6 (30,19) noktalarıydı. Bu kümede artık gelir (25+30) / 2 = 27,5 bin dolar ve eğitim (20+19) / 2 = 19,5 yıldır.

Tablo 3.5. Merkezi Kümeleme Yöntemi:3 küme için

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

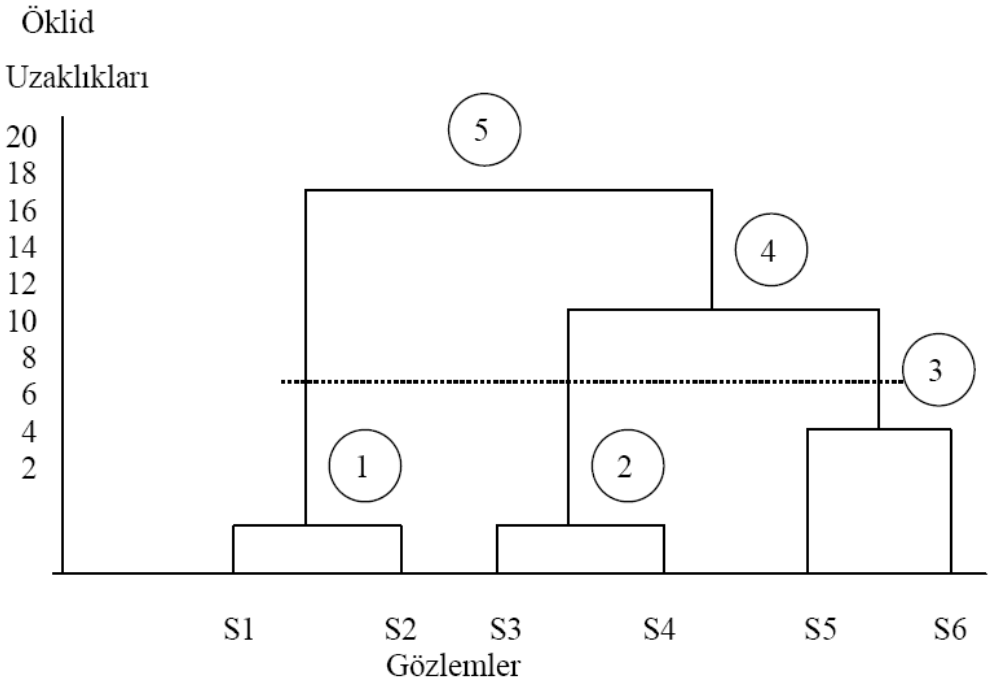
3 küme için veri seti

KÜME SAYISI	KÜMELER	GELİR	EĞİTİM
1	S1-S2	5,5	5,5
2	S3-S4	15,5	14,5
3	S5-S6	27,5	19,5

Benzerlik matrisi

	S1-S2	S3-S4	S5-S6
S1-S2	0	181	680
S3-S4	181	0	169
S5-S6	680	169	0

Hiyerarşik kümelemede tüm kümeler hiyerarşik olarak, yani sırayla oluşturulmaktadır. Dolayısıyla her aşamada oluşturulacak olan küme sayısı, bir önceki aşamadan bir eksik olacaktır. n gözlem varsa; birinci adımda n-1 küme, ikinci adımda n-2 küme, (n-1). adımda bir küme elde edilecektir. Genel olarak hiyerarşik kümeleme sürecinde bu adımlar dendogram veya ağaç (tree) adı verilen grafiklerle gösterilmektedir. 1,2,3,4 ve 5 ile gösterilen rakamlar hiyerarşik sürecin adımlarıdır. Eğer kümeleme noktalı doğrunun olduğu yerden kesilirse, S1-S2, S3-S4, S5-S6 olmak üzere üç küme elde edilir.



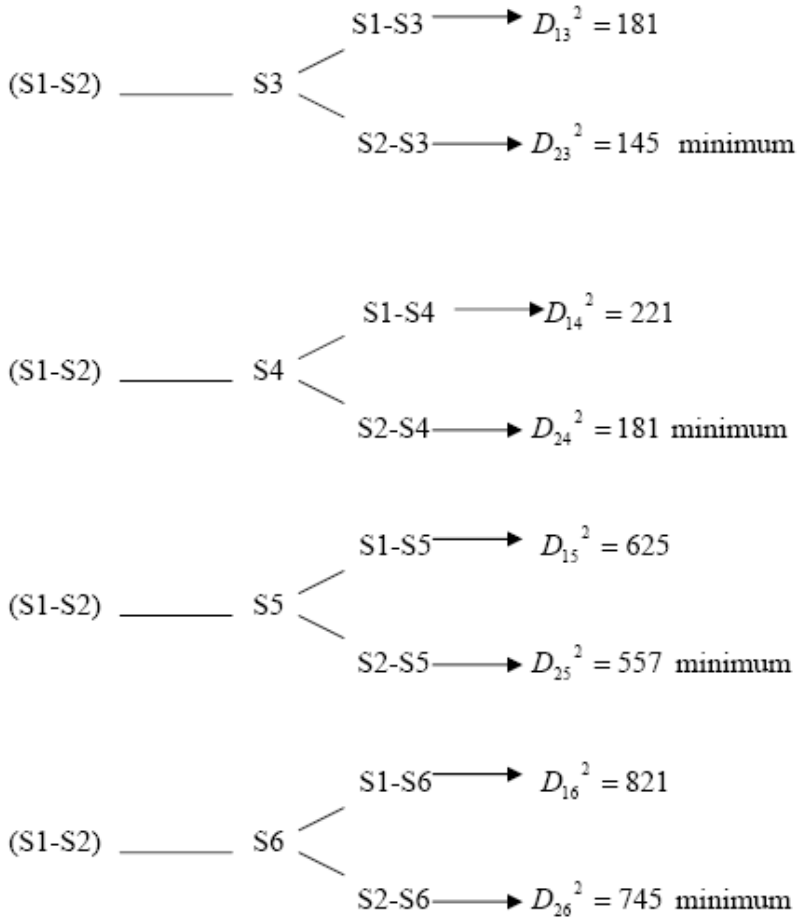
Şekil 3.5. Dendogram

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

Hiyerarşik diğer tekniklerde başka küme seçim yöntemleri de vardır. Bunların hepsinde birinci adım aynıdır. (Örneğin ilk kümenin oluşturulması süreci) Fakat kümeler arasındaki uzaklığın ölçülmesi konusunda her biri farklılaşmaktadır.

3.1.1.1.2. TEK BAĞLANTI TEKNİĞİ

Tek bağlantı yöntemi "En Yakın Komşu" yöntemi olarak da adlandırılmaktadır. Bu yöntemde ilk belirlenen kümedeki elemanların her biriyle diğer elemanlar arasındaki mesafelerden minimum olanı seçilir ve bu iki eleman ilk kümeyi oluşturur. Daha sonra ya oluşturulan ilk kümeye en yakın mesafedeki eleman seçilerek kümeye eklenir ya da daha yakın mesafede olan iki eleman belirlenerek başka bir küme oluşturmaları sağlanır. Bu işlem bütün kümeler birleşerek tek bir kümeye dönüşünceye kadar devam eder. İlk küme merkezi kümeleme yöntemindeki gibi (S1-S2) belirlenmektedir.



Bu süreç aşağıdaki matriste özetlenebilir.

Tablo 3.6. Benzerlik Matrisi

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

	S1-S2	**S3**	**S4**	**S5**	**S6**
S1-S2	0	145	181	557	745
S3	145	0	2	136	250
S4	181	2	0	106	212
S5	557	136	106	0	26
S6	745	250	212	26	0

Tablo 3.6'daki matris incelendiğinde tüm değerlerin içinde minimum olan 2 rakamı görülmektedir. Bu nedenle S3-S4, ikinci küme olarak seçilir.

S1-S2 ve S3-S4 $D_{13}^2 = 181$ $D_{23}^2 = 145$ minimum

$D_{14}^2 = 221$ $D_{24}^2 = 181$

Tablo 3.7.Benzerlik Matrisi-2

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

	S1-S2	**S3-S4**	**S5**	**S6**
S1-S2	0	145	557	745
S3-S4	145	0	106	212
S5	557	106	0	26
S6	745	212	26	0

Tablo 3.7'deki matris incelendiğinde tüm değerlerin içinde minimum olan 26 rakamı görülmektedir. Bu nedenle S5-S6, üçüncü küme olarak seçilir.

S3-S4 ve S5 $D_{35}^2 = 136$ $D_{45}^2 = 106$ minimum

S3-S4 ve S6 $D_{36}^2 = 250$ $D_{46}^2 = 212$ minimum

Karşılaştırılacak kümelerin elemanları n_1 ve n_2 ise, karşılaştırılacak uzaklık $n_1 * n_2$ sayısı kadardır.

3.1.1.1.3. TAM BAĞLANTI YÖNTEMİ

Tam bağlantı yöntemi "En Uzak Komşu Yöntemi" olarak da adlandırılmaktadır. En yakın komşu yönteminin tam tersidir. İlk aşamada birbirine en yakın iki gözlem bir küme oluşturmaktadır. İki kümedeki tüm gözlem çiftleri arasındaki uzaklıklar içinden maksimum olanı, iki küme arasındaki uzaklık olarak tanımlanmıştır.

S1-S2 ve S3 , S4 , S5 , S6 $D_{13}^2 = 181$ maksimum

$D_{23}^2 = 145$

Tablo 3.8. Benzerlik Matrisi

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

	S1-S2	S3	S4	S5	S6
S1-S2	0	181	221	625	821
S3	181	0	2	136	250
S4	221	2	0	106	212
S5	625	136	106	0	26
S6	821	250	212	26	0

Süreç en yakın komşu yöntemi ile aynıdır.

3.1.1.1.4.ORTALAMA BAĞLANTI YÖNTEMİ

Ortalama bağlantı tekniği, Sokal ve Michener tarafından önerilmiştir. Bu teknikte, iki küme arasındaki fark, bir küme arasındaki eleman çiftleri ile diğer bir kümedeki eleman çiftleri arasındaki ortalama fark olarak alınır. Bu tekniğin değiştirilmiş türleri bulunmaktadır. En yaygın kullanılan türünde gözlem çiftleri arasındaki uzaklığın aritmetik ortalaması hesaplanmaktadır. Ortalama bağlantı tekniği, yaygın olarak biyoloji biliminde kullanılmaktadır, bununla birlikte sosyal bilimlerde kullanımı da giderek artmaktadır. Genellikle tam bağlantı ve ortalama bağlantı tekniklerinde benzer dendogramlar oluşmaktadır. Ancak her bir yöntemde uzaklık farklı tanımlandığı için birleştirmeler farklı seviyelerde ortaya çıkabilmektedir[134].

S1-S2 ve S3 arasındaki uzaklık $D_{13}^{\ 2} = 181$

$D_{23}^{\ 2} = 145$ ortalaması olacaktır.

Bu değer hesaplandığında (181+145) / 2 = 163 bulunacaktır. Tüm değerlerin hesaplandığı benzerlik matrisi aşağıda Tablo 3.9'da verilmiştir.

Tablo 3.9. Benzerlik Matrisi

	S1-S2	S3	S4	S5	S6
S1-S2	0	163	201	591	783
S3	163	0	2	136	250
S4	201	2	0	106	212
S5	591	136	106	0	26
S6	783	250	212	26	0

İkinci küme S3-S4 olarak belirlenmiştir. Birinci ve ikinci küme arasındaki uzaklık;

D_{13}^2, D_{14}^2, D_{23}^2 ve D_{24}^2 'nin ortalaması alınarak hesaplanacaktır.

S1-S2 ve S3-S4 arasındaki uzaklık

$D_{13}^2 = 181 \qquad D_{14}^2 = 221$

$D_{23}^2 = 145$ ve $D_{24}^2 = 181$ in ortalaması olacaktır.

Bu değer hesaplandığında (181+221+145+181) / 4 = 182 bulunacaktır.

Tek bağlantı yöntemi sağlıklı sonuç vermesi açısından tercih edilse bile, işlemlerin uzun sürmesi açısından sakıncalıdır. Tam bağlantı yöntemi ise aynı küme içerisindeki bireylerin uzaklıklarının belli bir değerden küçük olması durumunda tüm kümelerin sağlıklı oluşturulmasını garanti etmemektedir. Son yıllarda sıkça kullanılmaya başlanan ortalama bağlantı yöntemi, bu iki uç teknik arasında sonuçlar vermesi sebebiyle bir alternatif olarak önerilmektedir(Tatlıdil, 1992, s. 336).

3.1.1.1.5. WARD BAĞLANTI YÖNTEMİ

Ward yöntemi uzaklıklar üzerinden çalışmaz. Bu yöntemde küme içindeki homojenliğin maksimum olduğu kümeler oluşturulur. Bu yöntemde grup bağlantıları yerine grup içi kareler toplamları ele alınmaktadır.

Küme-içi kareler toplamı, homojenlik ölçüsü olarak kullanılmaktadır. Dolayısıyla Ward'ın yöntemi küme-içi bütün kareler toplamını minimize etmeyi amaçlamaktadır (Hata kareleri toplamı- Error sums of square-ESS).

Bu yöntemde ilk olarak tüm mümkün küme sonuçları incelenir. 6 gözlem varken, önce ilk iki gözlem bir küme, diğer gözlemler birer küme gibi düşünülür ve tüm mümkün kombinasyonlar yazılır.

Tablo 3.10. Ward Yöntemi

AŞAMA	KÜMELER					HATA KARELERİ TOPLAMI
	1	2	3	4	5	

a) Tüm mümkün 5 küme çözümleri

1	S1-S2	S3	S4	S5	S6	1
2	S1-S3	S2	S4	S5	S6	90,5
3	S1-S4	S2	S3	S5	S6	110,5
4	S1-S5	S2	S3	S4	S6	312,5
5	S1-S6	S2	S3	S4	S5	410,5
6	S2-S3	S1	S4	S5	S6	72,5
7	S2-S4	S1	S3	S5	S6	90,5
8	S2-S5	S1	S3	S4	S6	278,5
9	S2-S6	S1	S3	S4	S5	372,5
10	S3-S4	S1	S2	S5	S6	1
11	S3-S5	S1	S2	S4	S6	68
12	S3-S6	S1	S2	S4	S5	125
13	S4-S5	S1	S2	S3	S6	53
14	S4-S6	S1	S2	S3	S5	106
15	S5-S6	S1	S2	S3	S4	13

b) Tüm mümkün 4 küme çözümleri

1	S1-S2-S3	S4	S5	S6		109,333
2	S1-S2-S4	S3	S5	S6		134,667
3	S1-S2-S5	S3	S4	S6		394,667
4	S1-S2-S6	S3	S4	S5		522,667
5	S1-S2	S3-S4	S5	S6		2
6	S1-S2	S3-S5	S4	S6		69
7	S1-S2	S3-S6	S4	S5		126
8	S1-S2	S4-S5	S3	S6		54
9	S1-S2	S4-S6	S3	S5		107
10	S1-S2	S5-S6	S3	S4		14

S1-S2 için hata kareleri toplamı şöyle hesaplanır:

$$\left[(5-5,5)^2+(6-5,5)^2\right]+\left[(5-5,5)^2+(6-5,5)^2\right]=1,0$$

S3 için hata kareleri toplamı = 0 dır. (Çünkü 1 elemanı vardır)

S4 = 0 S5 = 0 S6 = 0

Dolayısıyla birinci mümkün sonuç için toplam hata kareleri toplamı = 1 dir. Buradan en küçük hata kareleri toplamını veren mümkün sonuç seçilir. Bunlar ya 1. ya da 10. mümkün sonuçlardır.

S1-S2, S3 , S4 , S5 , S6

S3-S4, S1 , S2 , S5 , S6

Hangisinin seçileceği tamamen rassaldır. Dolayısıyla herhangi biri seçilir. Yukarıdaki örnekte S1-S2 seçilerek 4 mümkün küme oluşturma işlemi gerçekleştirilmiştir. (Tablo 3.10-b)

S1-S2-S3 , S4 , S5 , S6 için hata kareleri toplamı şöyle hesaplanır:

S1 (5,5)

S2 (6,6)

S3 (15,14)

X ort = (5+6+15) / 3 = 8,6666

Y ort = (5+6+14) / 3 = 8,3333

$$\left[(5-8,6)^2+(6-8,6)^2+(15-8,6)^2\right]+\left[(5-8,3)^2+(6-8,3)^2+(14-8,3)^2\right]=109,33$$

Tüm mümkün sonuçlar için hata kareleri toplamı hesaplanır ve minimum olan kombinasyon seçilir (Şimşek,2006).

3.1.1.1.6. CURE YÖNTEMİ

CURE algoritması, her kümenin sabit sayıda örneklem nokta ile temsil edildigi ve her adımda istenen küme sayısı elde edilene kadar örneklem noktaları en yakın olan kümelerin birlestirildigi aşağıdan yukarıya dogru çalışan hiyerarşik bir kümeleme algoritmasıdır. Her adımda yeni oluşturulan kümelerin örneklem noktalarını bulmak için birleşen kümelerin örneklem noktaları bir daraltma katsayısı ile çarpılır. Bu durumda algoritmanın dogru kümelenmeleri bulması üç parametrenin degerine baglıdır: küme sayısı (k), örneklem nokta sayısı (repisay), ve daraltma katsayısı (α).

CURE algoritmasının çalısmasındaki islem basamakları aşağıda görülmektedir (Demiralay ve Çamurcu, 2005):

1. Her küme için sabit sayıda ve küme içinde dağınık olarak yerleşmiş c adet örneklem nokta seçilir,

2. iki küme arasındaki uzaklık, bu kümelere ait örneklem noktalar arasındaki Öklit uzaklığı hesaplanarak elde edilir,

3. En yakın küme çifti birlestirilir,

4. Oluşan yeni kümenin örneklem noktaları bulunur. Bu işlem için yeni kümenin alt kümelerinden merkeze en yakın olan c adet nokta seçilir. Bu noktalar daraltma katsayısı i ile çarpılarak merkeze doğru yaklaştırılır,

5. Küme sayısı, kümeleme algoritmasında giriş parametresi olarak verilen k değerine ulaşana kadar 2, 3 ve 4. adımlar tekrarlanır.

3.1.1.2.7. AGNES YÖNTEMİ

AGNES algoritması aşağıdan yukarı doğru çalışan bir inşa yapısı izler. Başlangıçta her nesne ayrı bir küme olarak kabul edilir. Algoritmanın sonraki her adımında bu atomik kümelerden benzer özellik gösterenler birleştirilir. Her birleştirme işleminden sonra toplam küme sayısı bir azalır. İstenen sayıda küme elde edildiginde veya en yakın iki küme arasındaki uzaklık verilen eşik değere ulaştığında birleştirme işlemi sona erer. Herhangi bir sonlanma koşulu verilmezse kümeleme işlemi tamamlandığında bütün nesneler tek bir kümede toplanır (Demiralay ve Çamurcu, 2005).

Küme Sayısının (k) Algoritmadaki Etkisi : Küme sayısı parametresi, k, AGNES algoritmasının sonlanma koşulunu belirler. İstenen küme sayısı elde edildiğinde kümeleme işlemi sona erer. AGNES algoritması belirgin küresel kümelerden olusan yuvarlaklar veri setine doğru küme sayısı verilerek uygulandıgında başarılı sonuçlar vermektedir.

3.1.1.2. BÖLÜNÜR KÜMELEME ALORİTMALARI

Bölünür kümeleme algoritmaları, toplaşım algoritmalarının tersine çalışır. Başlangıçtaki gözlem birimlerinin oluşturduğu küme iki alt gruba ayrılır, bir gruptaki gözlem birimleri diğer gruptaki gözlem birimlerinden uzaktır. Daha sonra bu gruplar yine birbirine benzemeyen alt gruplara ayrılırlar. Bu süreç gözlem birimleri kadar alt grup elde edilene kadar sürdürülür.

3.1.1.2.1. BÖLÜNMÜŞ ORTALAMALAR (Splinter-Average Distance) YÖNTEMİ

Mac-Naughton-Smith ve arkadaşları 1962 yılında bu metodu geliştirmişlerdir. Bu yöntemde her bir gözlem değerinin bölünmüş gruptaki gözlem birimlerine olan ortalama uzaklıkları ve bu gözlem birimlerinin, gruptaki diğer gözlem birimlerine olan uzaklıkları hesaplanır (Johnson ve Wichern, 1998, s.739).

Süreç, diğer gözlem birimlerine en uzak olan gözlem biriminin ayrılmasıyla başlamaktadır. Böylece iki grup oluşturulur. Daha sonra iki uzaklık hesaplanır. Ana grupta yer alan her bir gözlem biriminin, ayrılmış olan gruptaki gözlem birimine olan ortalama uzaklığı ve ana grupta yer alan her bir gözlem biriminin gruptaki diğer gözlem birimlerine olan uzaklıkları hesaplanır. Eğer gözlem biriminin, ayrılmış olan gruba uzaklığı ana kümeye olan uzaklığından küçükse, o gözlem birimi ayrılmış olan gruba dahil edilir. Bu süreç bu şekilde tekrar edilir. Buradan anlaşıldığı üzere, her bir gözlem biriminin kendi kümesine olan ortalama uzaklığı, diğer kümeye olan ortalama uzaklığından küçük olmalıdır (Dillon ve Goldstein, (b.t.), s. 180-182). Aşağıda verilen matris üzerinde bu yöntem incelenmiştir.

	A	B	C	D	E
A	0	12	9	32	31
B	12	0	9	25	27
C	9	9	0	23	24
D	32	25	23	0	9
E	31	27	24	9	0

Gözlem birimleri içerisinde diğer gözlem birimlerine en uzak ortalama uzaklığı olan değişken "E" dir. Böylece başlangıçtaki ayrım "E" ve "A,B,C,D" kümeleri olarak yapılır. Bir sonraki adımda ayrılmış olan gruptaki ve ana gruptaki ortalama uzaklıklar hesaplanır. Hesaplanan değerler aşağıdaki tabloda verilmiştir:

Tablo 3.11. Gözlem Birimleri İçin Hesaplanan Ortalama Uzaklık Değerleri-1
Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

Gözlem Birimleri	Ayrılmış Gruba Olan Ortalama Uzaklık	Ana Gruba Olan Ortalama Uzaklık	Fark
A	31	17,67	-13,33
B	27	15,33	-11,67
C	24	13,67	-10,33
D	9	26,67	17,67

D gözlem birimi için ayrılmış gruba olan ortalama uzaklık, ana gruba olan ortalama uzaklıktan küçük olduğu için, E gözlem birimiyle aynı kümeye dahil edilir. Artık elde edilen iki küme "D,E" ve "A,B,C" kümeleridir. Bundan sonraki adım ise A,B ve C'nin hem ayrılmış gruba hem de ana gruba olan uzaklıklarının hesaplanmasıdır. Hesaplanan değerler aşağıdaki tabloda gösterilmişir.

Tablo 3.12. Gözlem Birimleri İçin Hesaplanan Ortalama Uzaklık Değerleri-2
Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

Gözlem Birimleri	Ayrılmış Gruba Olan Ortalama Uzaklık	Ana Gruba Olan Ortalama Uzaklık	Fark
A	31,5	10,5	-4
B	26	10,5	-15,5
C	23,5	9	-14,5

Tüm farklar negatif olduğu için, işlemler burada sona erdirilir.

3.1.1.2.2. OTOMATİK ETKİLEŞİM DEDEKTÖRÜ (Automatic Interaction Detection . AID) YÖNTEMİ

Kümeleme analizine amaç ve içerik açısından benzer olan bir başka analiz, Otomatik Etkileşim Dedektörü (A.I.D.) analizidir.

Nominal ölçekte ölçülmüş tahmin değişkenleri ile aralıklı ölçekte ölçülmüş kriter değişkenine uygulanabilecek olan bu analiz, kümeleme analizinin tersine tüm örneği kriter değerini esas alarak kendi içinde daha benzer alt gruplara ayırır. Bu ayırma her defasında ikişer ikişer yapılır ve ayırmada varyans analizinden yararlanılır. Pazar bölümlendirilmesinde çok yaygın olarak kullanılan A.I.D. analizinin en önemli kısıtlaması, oldukça büyük örneklere (genellikle 100'ün üstünde) gereksinim göstermesidir. Aksi taktirde alt gruplar kısa sürede çok küçük örneklerden oluşacaktır. Bu ise istatistiksel açıdan anlamsız olacaktır(Kurtuluş, 1998, s. 504).

3.1.2. HİYERARŞİK OLMAYAN YÖNTEMLER

Hiyerarşik olmayan kümeleme yönteminde, önceden belirlenen kriterler doğrultusunda kümeler oluşturulurken, küme sayısının önceden belirlendiği varsayılmaktadır. Genellikle hiyerarşik olmayan kümeleme analizinde, sonuçta elde edilecek küme sayısının bilindiği farzedilir. Fakat bazı yöntemler analizin aşamalarında küme sayısının değişebileceğini kabul etmektedir.

Hiyerarşik olmayan yöntemler, birimlerin kendi içinde homojen ve kendi aralarında heterojen olan kümelere ayrılmasını hedefleyen ve bu kümeler aracılığı ile alt popülasyonların parametre tahminlerini yapmayı amaçlayan yöntemlerdir. Hiyerarşik yöntemlerde hem birimler hem de değişkenler birbirleriyle değişik benzerlik düzeylerinde kümeler oluştururken, hiyerarşik olmayan yöntemlerde birimlerin uygun oldukları kümelerde toplanmaları ve n birimin k adet kümeye parçalanması hedeflenmektedir.

Hiyerarşik olmayan kümelemede veri k adet gruba ayrılır ve her bir grup bir kümeyi belirtir. Dolayısıyla hiyerarşik kümelemenin tersine, küme sayısı önceden bilinmektedir. Hiyerarşik olmayan kümeleme analizinde şu aşamalar söz konusudur;

1. k adet küme için k adet küme ortalaması keyfi olarak seçilir.

2. Gözlem birimleri hangi küme ortalamasına daha yakınsa, o kümeye dahil edilir.

3. Kümelere ait ortalama değerleri yeniden hesaplanır.

4. Küme elemanlarında herhangi bir değişiklik yoksa işlem durdurulur. Eğer değişiklik varsa ikinci adıma dönülür.

Hiyerarşik olmayan kümeleme yöntemleri arasında k-ortalamalar (k-means clustering) yöntemi, metoid kümeleme (metoid clustering), bulanık kümeleme (fuzzy clustering) ve yığma (hill climbing) kümeleme gibi pek çok teknikten söz edilmektedir. Ancak bunlardan en çok kullanılanı Mac Queen tarafından geliştirilen k-ortalamalar tekniğidir.

3.1.2.1. K –ORTALAMALAR YÖNTEMİ

K-Ortalamalar (k-means) yöntemi 1967 yılında Mac Queen tarafından sunulmuştur. Uzun yıllar boyunca pek çok uygulama alanında yoğun olarak kullanılan bir kümeleme algoritmasıdır. Bu algoritmada k sayıda küme oluşmakta ve her küme içerisinde bulunan verilerin ağırlıklı ortalamaları sonucu bir değer ortaya çıkmaktadır. Küme içerisinde bu değere en yakın olan nokta değeri küme merkezi (centroid) olarak kabul edilmektedir (Berkhin, 2002).

K-Ortalamalar yöntemi öncelikle eldeki verileri k adet kümede ve kümelerin ortalamalarına göre kümelere ayırır. k küme sayısı kullanıcı tarafından verilir. Burada bahsedilen ortalama, daha önce belirtilen küme merkezidir. Daha sonra gelen her veri değeri merkez noktaya en yakın olduğu kümeye dahil edilir. Eklendiği küme elemanlarının ağırlıklı ortalamaları tekrar hesaplanarak yeni bir küme merkezi değeri bulunur ve bu yeni değer bundan sonraki işlemlerde bu kümeyi temsil eder.

Mac Quenn en yakın değerlere sahip her elemanı, kümelere ayırabilecek algoritmayı tanımlamak için K-ortalama terimin ortaya atmıştır. Bu teknik aşağıdaki adımları izler :

1. Birimler K Adet kümeye ayrılır.

2. Birimler, değer bakımından en yakın kümeye atanarak devam edilir. Uzaklık genellikle ‘’Euclidean uzaklık’’ kullanılarak belirlenir. Daha sonra birimler hesaplanarak kümenin yeni değeri bulunur.

3. Adım 2 hiç atama yapılmayacak hale gelene kadar tekrarlanır(Demiralay ve Çamurcu, 2005).

Küme Sayısının Belirlenmesi

Kümeleme analizinden sağlıklı bir sonuç elde edilebilmesi için değişkenlerin seçimi ve küme sayısının belirlenmesi önemlidir. Küçük örneklemelerde küme sayısının belirlenmesi için aşağıdaki eşitlik sık kullanılmaktadır (Şimşek, 2006);

$$k = (n/2)^{1/2}$$

Marriot tarafından önerilen yöntemde ise ;

$$M = k^2 |W|$$

Burada en küçük M değerini veren küme sayısı gerçek küme sayısıdır. W ise grup içi kareler toplamı matrisidir. Calinsky ve Harabasz tarafından geliştirilen yöntemde ise ;

$C = [iz(B)/k - 1]/[iz(W)(n = k)]$ eşitliğini en büyükleyen k değeri küme sayısıdır. Burada B ve W sırasıyla gruplar arası ve grup içi kareler toplamı matrisleridir (Atamer, 1992).

Aşağıda bir örnek üzerinde k-ortalamalar yöntemi incelenmiştir. A,B,C ve D için x1 ve x2 değişkenlerinin ölçülmesi istenmektedir.

Tablo 3.13. Veri seti

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

	GÖZLEMLER	
	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Amacımız buradaki gözlemleri k = 2 kümeye ayırmak olsun. Yakın değerlere sahip olan gözlem birimleri aynı kümede yer almalıdır. İlk adımda k= 2-ortalama yöntemini uygulayabilmek için keyfi olarak gözlem birimleri iki kümeye ayrılır. Örneğin (AB) ve (CD) kümeleri için bu kümelerin ortalamaları hesaplanır.

Tablo 3.14.Kümelerin ortalama değerleri

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

	ORTALAMALAR	
KÜMELER	$\bar{x}_1$	$\bar{x}_2$
(AB)	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(CD)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

İkinci adımda her bir gözlem biriminin küme ortalamasına olan öklid uzaklığı hesaplanır ve gözlem birimi en yakın kümeye atanır. Eğer gözlem biriminin başlangıçta yer aldığı küme değişirse, küme ortalaması tekrar hesaplanır. Aşağıda öklid uzaklıkları hesaplanmıştır:

$$D^2(A,(AB)) = (5-2)^2 + (3-2)^2 = 10$$

$$D^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61$$

A gözlem birimi, (AB) kümesine (CD) kümesinden daha yakın olduğu için tekrar atanmaz. İşleme B gözlem birimi için devam edilir.

$$D^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10$$

$$D^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9$$

B gözlem birimi (CD) kümesine (AB) kümesinden daha yakın olduğu için (CD) kümesine atanır. Artık elimizde olan iki küme (A) kümesi ve (BCD) kümesidir. Kümelerin ortalamaları tekrar hesaplanır.

Tablo 3.15.Kümelerin ortalama değerleri

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

	ORTALAMALAR	
KÜMELER	$\bar{x}_1$	$\bar{x}_2$
(A)	5	3
(BCD)	-1	-1

Her gözlem birimi tekrar atama yapılıp yapılmaması durumu için kontrol edilir. Küme ortalamalarının uzaklık kareleri alındığında aşağıdaki tablo elde edilecektir.

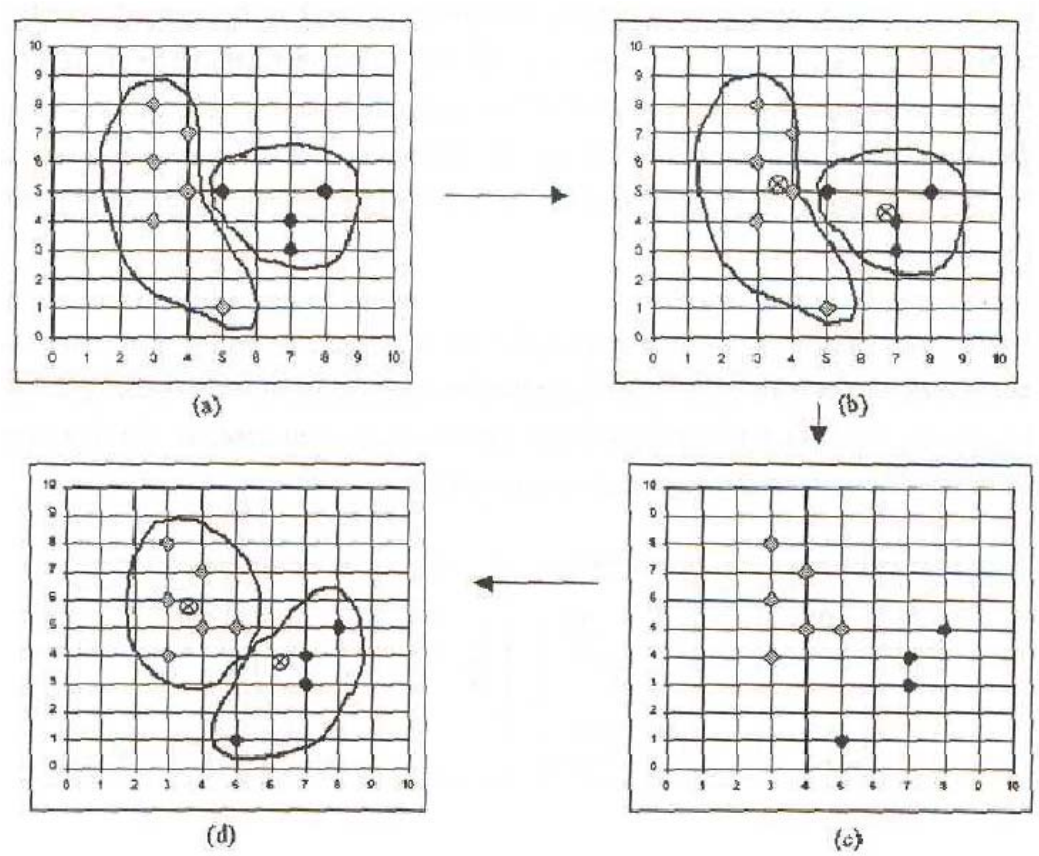
Tablo 3.16.Küme ortalamalarının uzaklık kareleri

Kaynak: Şimşek, (2006) içinde. Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama' dan alınmıştır.

	KÜME ORTALAMALARININ UZAKLIK KARELERİ			
KÜMELER	A	B	C	D
(A)	0	40	41	89
(BCD)	52	4	5	5

Tablodan görüleceği üzere her gözlem birimi, kendisine en yakın ortalama uzaklıkta olan kümeye atanmıştır. Dolayısıyla süreç sona erer. Sonuçta elde edilen k =2 küme (A) ve (BCD) kümeleridir.

Aşağıdaki şekilde bir K-Means küme dağılımı örneği görünmektedir.



Şekil 3.6. Bir K-means kümeleme örneği

Kaynak: Özekeş, (2003) içinde. Veri Madenciligi Modelleri ve Uygulama Alanları' dan alınmıştır.

3.1.2.2 METOİD YÖNTEMİ

k-metoids kümeleme yönteminin temel stratejisi, ilk olarak n adet nesnede, merkezi temsili bir medoid olan k adet küme bulmaktır(Han ve Kamber, 2000). Geriye kalan nesneler, kendilerine en yakın olan medoide göre k adet kümeye yerleştirilir. Bu bölünmelerin ardından kümenin ortasına en yakın olan nesneyi bulmak için medoid, medoid olmayan her nesne ile yer değiştirir. Bu işlem en verimli medoid bulunana kadar devam eder. Metoid yöntemi; n birimin, küme içi gözlemlerin benzer, kümelerarası gözlemlerin farklı olacağı biçimde iki veya daha fazla kümeye ayrılmasını amaçlar. Bu parçalamada metoid adı verilen kümeleri tanıtıcı merkez noktalar veya gözlemler yardımı ile n birimi k kümeye ayırma sağlanır. Metoid kümelemede en önemli sorun merkez noktaların be-

lirlenmesidir. Metoidler belirlendikten sonra her birim aralarındaki benzerliklerin maksimum ve farklılıkların minimum olduğu en yakın metoide sahip olan kümeye atanır.

3.1.2.3. YIĞMA KÜMELEME YÖNTEMİ

Yığma kümeleme yönteminde birimler, en yakın ortalamalı kümeye atanma yerine önceden belirlenen bir istatistiksel kritere göre bir kümeden diğerine hareket ederek en uygun durum sağlandığında belirli bir kümede yer alırlar. Belirlenen istatistik kritere göre veri setindeki tüm birimler hangi kümede yer alacakları kesin olmaksızın atama işlemlerine tabi tutulur. Bu istatistiksel kriter, küme içi kovaryans matrisi ve kümelerarası kovaryans matrisi ile ilgili olarak geliştirilmiştir. Toplam kovaryans matrisinin determinantının, küme içi kovaryans matrisinin determinantına oranının maksimum olduğu küme durumunun logaritması alınarak hesaplanan katsayı (c), kümeleri ayırmada bir kriter olarak alınmaktadır. n birim ardışık olarak parçalanıp k kümeye ayrılarak c katsayısının optimum olduğu değere ulaşıldığında elde edilen k adet kümenin, n birimin en uygun kümelenmesi olacağı ileri sürülmüştür (Şahin ve Hamarat, 2002).

3.1.2.4. BULANIK KÜMELEME YÖNTEMİ

Bulanık kümeleme yöntemi, kümeler birbirinden belirgin bir şekilde ayrılamıyorsa veya küme üyeliklerinde bazı birimler küme üyeliğinde kararsızsa, uygun bir yöntem olarak ortaya çıkmaktadır. Bulanık kümeler, kümedeki birimin üyeliği olarak tanımlanan 0 ile 1 arasındaki her bir birimi belirleyen fonksiyonlardır. Birbirine çok benzeyen birimler aynı kümede yüksek üyelik ilişkisine göre yer alırlar. Bundan dolayı bulanık kümeleme yöntemi, birimlerin kümeye veya kümelere ait olabilme katsayılarını hesaplar. Üyelik katsayılarının toplamı 1'e eşittir. Böylelikle birim en yüksek üyelik katsayısına sahip olduğu kümeye atanır. Üyelik fonksiyonları, kümedeki elemanlar sürekli veya süreksiz olsun bir bulanık kümedeki bulanıklığı karakterize eden fonksiyonlardır. Klasik kümeleme yöntemlerinde ise her bir birim sıfır olmayan sadece bir üyelik katsayısına sahiptir ve bu değer daima 1'dir. Dolayısıyla klasik kesin kümeleme yöntemleri, bulanık çözümlemenin sınırlı bir durumudur (Şahin ve Hamarat, 2002).

3.2. KÜMELEME ANALİZİNDE DEĞİŞİK BİR YAKLAŞIM:KOHONEN

Kümeleme analizinde, yukarda anlatılan tekniklerin yanısıra yapay sinir ağları mantığıyla çalışan KOHONEN tekniği de kullanılır. Kohonen Özörgütlemeli Harita (SOM) topoloji-korumalı bir haritadır. Bu harita yüksek boyutlu (üç veya daha fazla) bir haritadaki verileri, tipik bir iki boyutlu ızgara formundaki haritaya dönüştürür. Öz-örgütlemeli haritanın ana amacı, girdi uzayındaki komşuluk ilişkilerini mümkün olduğunca koruyan ve birimler arasındaki komşuluk ilişkilerine göre topoloji korumalı bir harita yaratmaktır (Kohonen, 1995). Böyle bir Özörgütlemeli haritanın eğitiminde başlıca zaman tüketen adımlar verilen bir örnek için kazanan düğümün (winner node) yerleştirilmesi ile ilgili alt-problem boyunca geçen adımlardır (Kohonen,1996). Bir kazanan düğüm her girdi vektörü için en iyi uyumlu birim (Best Matching Unit veya BMU) şeklinde ifade edilir.

3.2.1. SOM (Self Organizing Maps)

SOM ağları, Teuvo Kohonen tarafından geliştirilmiştir. Genel olarak sınıflandırma yapmak için kullanılmaktadır. Bu ağların en temel özelliği olayları öğrenmek için bir öğretmene ihtiyaç duymamasıdır (denetsiz). İleri besleme/geri besleme türünde olabilir ve öğrenme algoritması olarak öz-örgütlenme yöntemini kullanır. Bir girdi katmanı ve bir harita katmanı bulunur.

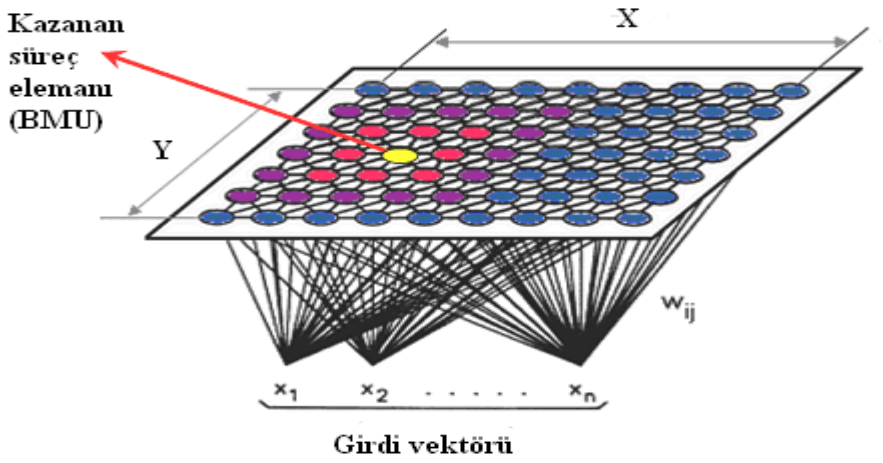
Öz-örgütlenme kullanan Kohonen Özellik Haritaları insan beynini taklit edecek şekilde tasarlanmıştır. İnsan beyninde öğrenme işlemi, sürekli tekrarlanan olaylar ve durumlar karşısında beyine iletilen sinyallerin, korteksin belli bölgelerinde yoğunlaşması sonucu bir hafıza oluşması şeklinde gerçekleşir (Yapay Sinir Ağları, Anonim, b.t.).

Benzer şekilde SOM ağlarına gönderilen sinyaller (girdi değerleri) bazı işlemlerden geçerek (iletilme-ağırlıklandırma) harita katmanına ulaşır. Bu katman 1 yada 2 boyutlu olarak dizilmiş sinir hücrelerinden oluşmaktadır.

Korteks görevini yapan bu katmana gönderilen girdiler, yapılan matematiksel hesaplamalar sonucu bir bölgede yoğunlaşırlar. Bu bölge; matematiksel işlemlerle belir-

lenen “kazanan sinir hücresi”dir. Bu sinir hücresine ait bir alan mevcuttur. Aktivasyon alanı olarak adlandırılan bu bölge öğrenme esnasında küçülür. Bu küçülme, örneğin sınıflandırma işlemlerinde kesinliğin artmasına karşılık gelir. Her sınıf için ayrı bir sinir hücresi etrafındaki toplanmalar, sonuçta sınıflara ait bölgeler oluşturur. Bu şekilde de sınıflara ait öğeler daha sonra kolaylıkla tespit edilir.

SOM ağları, hem verilerin kümelenmesinde hem de görselleştirilmesi açısından tercih edilmektedir. Bu ağlar çok boyutlu bir veriyi iki boyutlu bir haritaya indirgemektedir. Her bir küme için oluşturulan referans vektörleri bir araya geldiğinde bir haritayı meydana getirmektedir.



Şekil 3.7. Kohonen Özörgütlemeli Haritası

Kaynak: Karasulu ve Uğur (2007) içinde. Özörgütlemeli Yapay Sinir Ağı Modeli'nin Kullanıldığı Kutup Dengeleme Problemi için Paralel. Hesaplama Tekniği ile bir Başarım Eniyileştirme Yöntemi’ nden alınmıştır.

Formal olarak Özörgütlemeli Harita tanımı:

Girdi vektörü, $X=[x_1,x_2,\dots,x_n]^T \in R^n$ olsun. Bir i indisi ile düzenlenmiş birimlerin ayrık bir ızgarasını göz önüne alalım. Her düğüm ilgili ağırlık vektörü $W_i=[w_1,w_2,\dots,w_n]^T \in R^n$ içermektedir. X burada, tüm ağırlık vektörleri içerisinde ağırlık vektörü onun en yakın komşusu olan birimi göstermektedir. Buna en iyi uyumlu

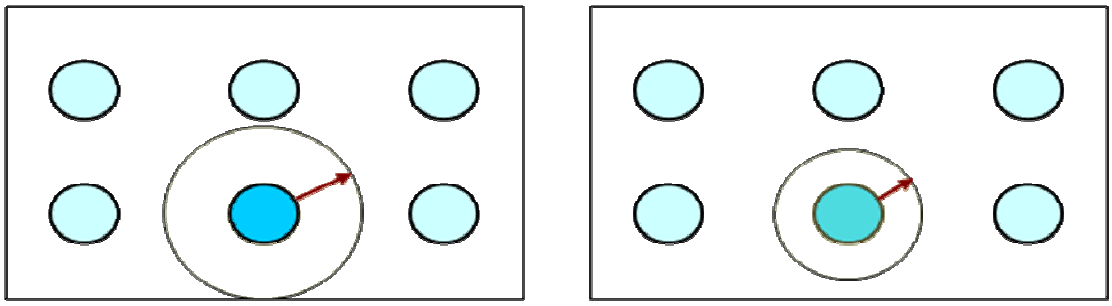
birim (BMU) denilmektedir ve şu şekilde bulunmaktadır (Vishwanathan ve Murthy, 2000).

$$\|X - W_j\| = \min_i \|X - W_i\|$$

Özörgütlemeli ağın eğitimi için her iterasyon aşağıda özetlendiği şekilde gerçekleşmektedir:

- Haritadaki düğümler arasından en yakın komşu (kazanan) her bir girdi örneği için bulunur.
- Kazananın ve tüm komşularının ağırlıkları güncellenir.

En çok zaman harcanan kısım bu komşulukları bulurken geçen süredir. Komşuluk hesapları öklid mesafesi (uzaydaki iki nokta arasındaki mesafe) uyarınca hesaplanılır. Özörgütlemeli harita temelde iki katmana sahiptir (Rauber, Tomsich ve Merkl, 2000). Giriş katmanı tamamıyla çift boyutlu Kohonen katmanına bağlanmıştır. Çıkış katmanı ise nicemleme probleminde kullanılır ve giriş vektörünün ait olabileceği üç sınıfı temsil eder. Bu çıkış katmanı tipik olarak delta kuralını uygulayarak öğrenir. Kohonen katmanı işlem elemanlarının her biri, gelen giriş değerlerinden onların ağırlıklarının öklid mesafesini ölçmektedir. Birimin ağırlık vektörü ile girdi vektörü arasındaki öklid mesafesi bu durumda aktivasyon fonk-siyonu görevi görmektedir. fiziksel uzayda iki boyutlu bir ızgara yapısı sergileyen SOM, Ağırlık/Girdi uzayında eğimli bir yapı sergilemektedir. Çalışmamızda elde ettiğimiz sonuçlarda da bu durumla karşılaşılmaktadır.



Şekil 3.8. Bir aktivasyon alanı

Kaynak: Yapay Sinir Ağları ve Uygulamaları, (b.t.) içinde. 18.11.2009 tarihinde http://mail.baskent.edu.tr/~20293638/som/ppt/sunu.ppt' dan alınmıştır.

Yukarıdaki şekilde, özellik haritasında kazanan sinir hücresine ait aktivasyon alanının algoritma ilerledikçe küçülmesi gösterilmiştir.

SOM ağları, hem verilerin kümelenmesinde hem de görselleştirilmesi açısından tercih edilmektedir. Bu ağlar çok boyutlu bir veriyi iki boyutlu bir haritaya indirgemektedir. Her bir küme için oluşturulan referans vektörleri bir araya geldiğinde bir haritayı meydana getirmektedir.

SOM ağlarında kümelemeyi etkileyen faktörler vardır.

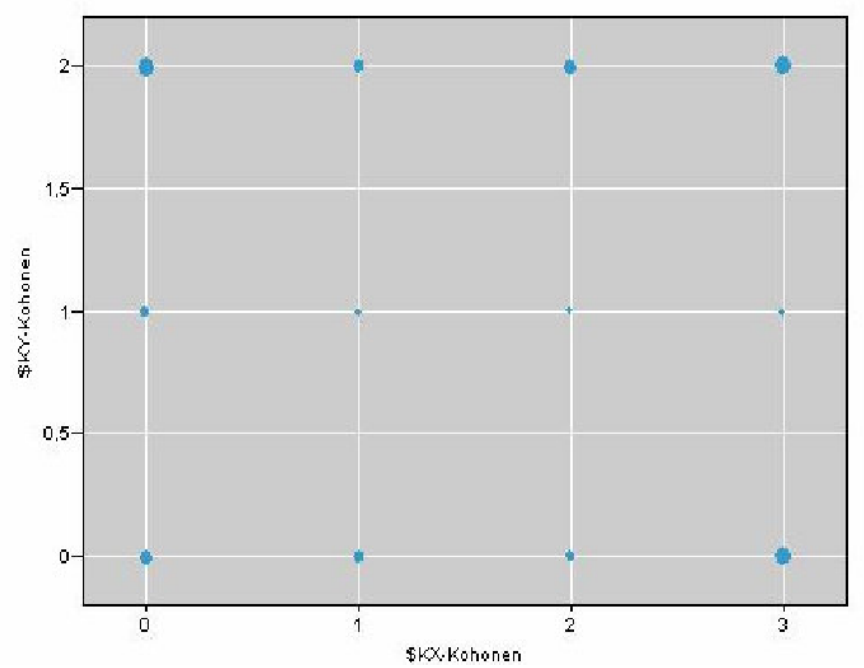
1. Çıkış katmanındaki nöron sayısı
2. Verilerin normalleştirilmesi
3. Referans vektörlerine ilk değer atanması
4. Uzaklık ölçüsü
5. Öğrenme katsayısı ve Komşuluk değişkeni

3.2.2.KOHONEN

Kohonen ağları, başlangıçta grupların bilinmediği durumlarda, verilerin kümelendirilmesini amaçlamaktadır. Kümeleme amacıyla kullanılmasının yanında, bir "veri görselleştirme aracı"dır (Flexer, 2001, s. 382). Kohonen ağlarında, kestirim yapılacak bir çıktı (bağlı) değişken bulunmadığından, denetimsiz öğrenim gerçekleştiren bir sinir ağı türüdür (Oğuzlar, 2005). Kohonen ağları, girdi (bağımsız) değişken kümesindeki örüntülerin açığa çıkarılması amacıyla kullanılmaktadır. Bu ağın çıktısında, gözlemler gruplandırılmış olarak elde edilmektedir. Bir grup veya kümenin içindeki gözlemlerin birbirine benzer olduğu, farklı gruplarda yer alan gözlemlerin ise birbirine benzer olmadığını söylemek mümkündür. Kohonen ağları, bir girdi ve iki boyutlu bir Kohonen tabakasından oluşmaktadır. Her bir sinir (nöron), girdi değişkenleri veya aynı anlamda girdi alanlarının her biri ile bağlantılıdır ve tekrar ağırlıkları (önemleri) bu bağlantıların her birinin üzerinde yer almaktadır.

Bir sinir için ağırlıklar, analizde kullanılan girdi alanlarının oluğturduğu küme için bir profili temsil etmektedir. Genellikle bir Kohonen ağı, az sayıda birim çok sayıda gözlemi özetlediğinde (güçlü birimler) veya çok sayıda birim gözlemlerin herhangi biri-

ne karşılık gelemediğinde (zayıf birimler)son bulacaktır. Kohonen veya kendini düzenleyen harita, bir çıktı olarak düğünülen sinirleri içermesine karşılık, bu ağlarda gerçek çıktı katmanı bulunmamaktadır.



Şekil 3.9. Bir Kohonen haritası

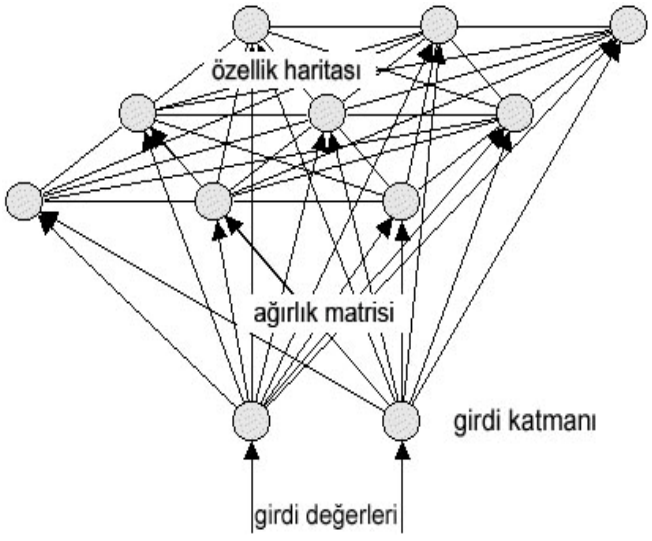
Kaynak: Oğuzlar, 2005 içinde. Kümeleme Analizinde Yeni Bir Yaklaşım: Kendini Düzenleyen haritalar:KOHONEN Ağları' dan alınmıştır.

Kohonen ağında yer alan daha düğük katmanlardaki düğümler (girdi düğümleri), örneklem veri noktaları tarafından temsil edilen girdileri alırlar. Daha yüksek katmanlardaki düğümler (çıktı düğümleri), denetimsiz öğrenim sürecinin ardından girdi örüntülerinin organizasyon haritasını temsil edecektir.

Öğretmensiz öğretme kullanılan bu teknik , kümeleme yaparak benzer grupların oluşturulması için kullanılmaktadır. KOHONEN, ileribesleme\geribesleme yapısında öğrenme metodu olarak öğretmensiz öğretme (unsupervised), öğrenme algoritması olarak ise kendi kendini düzenleyen (self-organized) algortiması kullanan bir tekniktir (Kiang ve Kumar, 2001).

Bir sinir ağının öğrenme algoritması ‘denetli’ ya da ‘denetsiz’ olabilir. İstenilen çıktısı önceden bilinen sinir ağına ‘denetli’ sinir ağı denir. İleri yayılım, denetli bir öğrenme algoritmasıdır ve bir sinir ağının girdi katma-nından çıktı katmanına doğru ‘bilgi akışı’nı açıklar. Geri yayılım, genellikle çok katmanlı perseptronların, ağın gizli katmanlarına bağlı olan ağırlıklarını değiştirmek için kullanılan denetli bir öğrenme algoritmasıdır. Geri yayılım algoritması ağırlıkları ters yönde değiştirmek için hesaplan-mış hata değerleri kullanır. Bu hatayı elde etmek için öncelikle 1 ileri yayılım safhası tamamlanmalıdır. İleri doğru yayılırken, sinir hücreleri sigmoid fonksiyonu kullanılarak etkinleştirilir (Yapay Sinir Ağları, Anonim, b.t.) .

Öz-Örgütlenme, Kohonen özellik haritaları tarafından kullanılan denetsiz bir öğ-renme algoritmasıdır. Genel olarak bilindiği gibi insan beyninin korteksi, her biri farklı işlevlere sahip bölgelere ayrılmıştır. Sinir hücreleri gelen bilgilere göre kendilerini grup-landırmıştır. Gelen bilgiler tek bir sinir hücresi tarafından alınmazlar, çevre hücreler de bu bilgiyi bir şekilde alır. Sonuç olarak bu örgütlenme bir çeşit harita yaratır. Biyolojik sinir hücrelerinin bu yapısı yapay sinir ağlarında ‘Kohonen Özellik Haritası’ kullanılarak taklit edilebilir.



Şekil 3.10.Bir Kohonen ağı

Kaynak: The Learning Process, (b.t.) içinde. 25.12.2009 tarihinde http://fbim.fhregensburg.de/~saj39122/jfroehl/diplom/e-13-text.html’ den alınmıştır.

Görüldüğü üzere girdi katmanındaki her sinir hücresi, haritadaki diğer bütün sinir hücreleri ile bağlantılıdır. Sonuçta ortaya çıkan ağırlık matrisi ağın girdi değerlerini haritadaki sinir hücrelerine aktarmak için kullanılır.

Ayrıca haritadaki bütün sinir hücreleri de kendi aralarında bağlantılıdır. Bu bağlantılar, aktivasyonun belirli bir bölgesindeki sinir hücrelerini, en büyük aktivasyona sahip sinir hücresi etrafında toplanmaya teşvik eder.

Kohonen ağları, boyut azaltma amacına bağlı olarak da kullanılabilmektedir. k adet orijinal girdi değişkeninden, grafiksel düzenleme sonucunda bulunan iki türetilmiş değişken, orijinal girdi değişkenlerinin benzerlik ilişkilerini korumaktadırlar (Kiang ve Kumar, 2001, s.178).

Kohonen ağlarının eğitimi boyunca, istenildiği takdirde bir geri bildirim grafiği (feedback graph) görüntülenebilir. Uygulama çalışmasından alınan bir geri bidirim grafiği aşağıda görülmektedir.

Şekil 3.11. Geri bildirim grafiği (Uygulama çalışmasından alınmıştır.)

Kohonen geri bildirim grafiği, eğitim süresince görüntülenmektedir. Her bir düğümün gücü, kırmızıdan beyaza doğru değişen bir renk ile temsil edilmektedir. Kırmızı renk, çok sayıda gözlem kazanan birimleri (güçlü birimler), beyaz renk ise birkaç veya hiç kayıt kazanamayan birimleri (zayıf birimler) temsil etmektedirler.

KOHONEN algoritmalarıyla çıkan sonuçları çok sayıda amaç için kullanabiliriz. Bu çalışmalardan çıkan sonuçlar, bize müşterler hakkında bilgi verir. İptal yapan müşterileri tanımamızı sağlar, az kazandıran müşterileri tespit eder. Bu bilgileri sadece öğrenmek yetmez, öğrendiğimiz bu bilgilerle iyileştirme çabasına gideriz. Pek çok şirket, bu tip uygumaları CRM (Customer Relationship Management) adı altında kullanmaktadır. CRM, arayüzler aracılığıyla müşteriye erişimi kolaylaştıran, onun hakkındaki bilgileri veren, gerektiğinde kampanyalar için kullanılan uygulamalardır. Veri madenci-liği sonucunda elde edilen cümleler CRM' de kullanılır.

4.CRM VE MÜŞTERİ SEGMENTASYONU

Günümüzde müşteriyi tanıma ve müşteriye göre hizmet verme ihtiyacı gitgide artmaktadır. Çalışmanın başından beri bahsedilen bu konular aslında bu ihtiyaçlardan dolayı oluşmuştur. Müşteriyle bu ilişkilerin tanımlandığı çalışmalara CRM (Customer Relationship Management) adı verilir.

CRM çalışmalarında, müşteriyi tanımlama, sınıflama, kümeleme veya müşterinin davranışlarına göre tahminde bulunma gibi projeler geliştirilebilir. Müşteri segmentasyonu da bu konular gibi CRM' de sıklıkla uygulanan çalışmalardan biridir.

4.1. CRM KAVRAMI

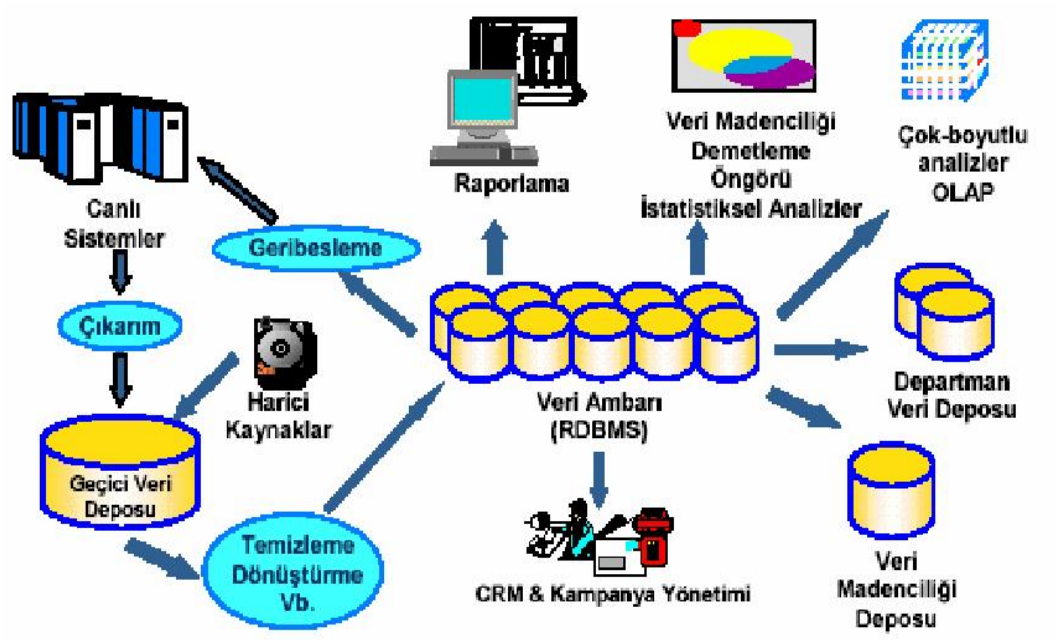
CRM (CUSTOMER RELATIONSHIP MANAGEMENT) kavramını, kurumların müşteriyi tanımlaması ve müşteriye göre hizmet vermesi olarak tanımlayabiliriz. CRM yani Müşteri İlişkileri Yönetimi, sadece satışların etkinliğini arttıran bir teknik değil, aynı zamanda ve daha önemli olarak mevcut müşterilerin elde tutulmasını da sağlayan bir tekniktir. CRM ile ürün odaklı yaklaşım son bulmuş, müşteri odaklılık önem

kazanmıştır. CRM ile firmaların sadece satış-pazarlama anlayışları değil, iç organizasyonları da değişmektedir (Müşteri İlişkileri Yönetimi, Anonim, b.t.). Müşterilerin beklenti ve ihtiyaçlarını ürün veya hizmet olarak karşılamak müşteri ilişkilerini yönetmenin en temel noktasıdır (Steihl, Anonim, b.t.).

CRM' in işleyişine yardımcı olacak dallardan biri de veri madenciliğidir. Veri madenciliği , müşteri segmentasyonu ve tahmin yöntemleriyle CRM'e yardımcı olur.

CRM, en değerli müşterileri seçmek ve yönetmek için geliştirilmiş olan bir işletme stratejisidir. CRM; etkili pazarlama, satış ve servis süreci sağlayabilmek için müşteri odaklı bir işletme felsefesi ve kültürü gerektirmektedir. CRM uygulamaları, müşteri ilişkilerinin yönetilmesini etkinleştirir.

Kitlesel pazarlama anlayışının geçerli olduğu dönemlerde müşteriyi elde etmek için, kalite ile müşteri tatmini unsurları yeterli olmaktaydı. Oysa günümüzde, müşteriyi yeniden tanımlayıp, bu yeni tanıma uygun stratejiler geliştirmek gerekli olmaktadır. Bu durum, aynı zamanda müşteri ilişkileri yönetiminin çıkış noktasını ifade etmektedir (Tekin ve Çiçek, 2003).



Şekil 4.1. Bir veri ambarı mimarisinde CRM in yeri

Üzerinde çok konuşulan CRM kavramıyla ilgili tanımları Duran aşağıdaki maddelerde özetlenmiştir.

i. CRM, müşteri ile ilişkide bulunulan her alanda müşteriyi daha iyi algılama ve onun beklentileri çerçevesinde firmanın kendisini daha iyi yönlendirmesi sürecidir.

ii. CRM, müşteri ilişkilerini yönetmek için kullanılan metodoloji ve ürünlerin genelini içermektedir.

iii. CRM, müşteri temas noktalarının entegrasyonu ve iyileştirilmesidir.

iv. CRM, müşteriyi tasarım noktasına (merkeze) yerleştiren ve müşteri ile yakın ilişki kuran bir yönetim felsefesidir.

v. CRM, satış, pazarlama ve servis süreçlerini daha etkin hale getirmek için geliştirilmiş işleme stratejisi / kültürüdür.

vi. CRM, müşteri bilgilerini kullanarak müşteri sadakatini ve sonuçta müşteri değerini artırma bilimidir.

vii. CRM, iş ve enformasyon akışlarının öncelikle müşteri ihtiyaçları, ikincil olarak ise şirket ihtiyaçlarına göre tasarlanmasıdır.

viii. CRM, kurumdaki müşteri ile ilgili her türlü bilgiyi tek bir enformasyon sistemine bağlamak ve bunu müşteri temas noktasına odaklamaktır.

ix. CRM, müşteriyi tanımak, müşteri ihtiyaçlarını anlamak, ona uygun hizmetler ve ürünler geliştirmektir (2001:2).

CRM çalışmalarının bize kazandırdığı pek çok avantaj vardır. Bunları aşağıdaki gibi sıralayabiliriz.

- Müşterilerin tam istediği ürün ve hizmetleri sağlamak ,
- Müşteriye daha iyi hizmet sunmak ,
- Daha efektif çapraz satış,

- Satışekibinin daha hızlısatışkapatması,
- Eski ve değerli müşterileri tutmak ve yenilerini kazanmaktır.
- Müşterileri sınıflandırmamızı sağlar ,
- En uygun zamanda en uygun pazarlama programıile en uygun müşteriye yaklaşma olasılığı hesaplar ,
- Müşterinin firmaya daha çabuk ulaşmasınısağlar ,
- Müşterinin daha çabuk karar vermesine olanak tanır ,
- Müşteri sadakatini artırır ,
- Başka firmalarla işbirliği yaparak yeni gelir olanaklarıyaratır ,
- Müşteri tatmin değerinin yükselmesini sağlar ,
- Birim müşteri gelirinin artmasınısağlar ,
- Müşteri sayısını arttırır ,
- Satış giderlerinin azalmasınısağlar ,
- Süreç verimliliklerini arttırır ,
- Stok yatırımlarının optimize edilmesini sağlar ,
- Rekabetten once değişimleri yakalayarak Pazar payının artırılmasını sağlar.

Bir CRM projesinin başarılı olması düzenli bir plana ve çalışmaya bağlıdır. Bu planları yaparken aşağıdakilere dikkat etmek gerekir.

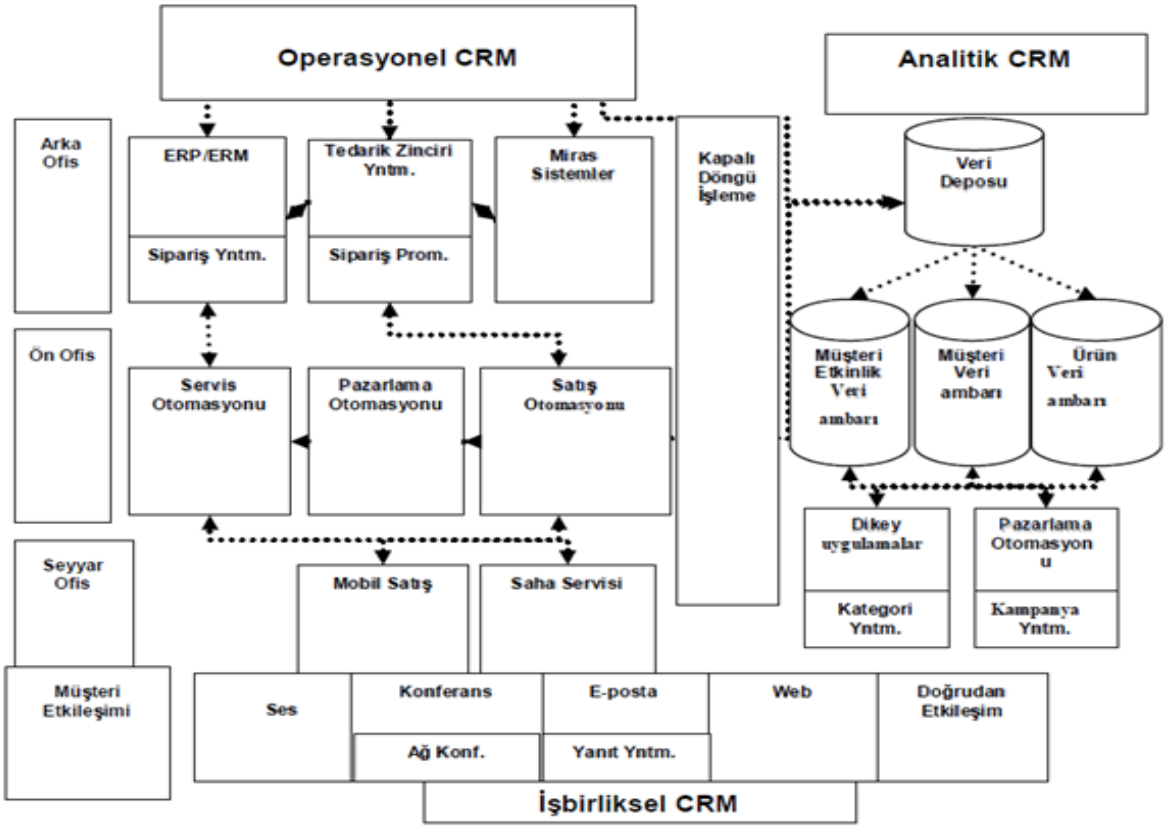
- Satış sürecinin iyi tanımlanması,
- Üst yönetimin, satışyönetiminin ve satış temsilcilerinin CRM'e bağlılığı ve kararlılığı olması,
- Etkinliklerin otomasyonuyla daha fazla satış yapılmasıve engellerin kaldırılması,

- Doğru tedarikçilerin veya diğer hizmet sağlayıcıların doğru seçilmesi,
- Yönetimin değil, satış elemanlarının ve müşterilerin önemi vurgulanmalı,
- Tüm zaman dilimlerinin, kullanıcıların ve iş tarzlarının ihtiyaçlarının karşılanması için artırılmış destek sağlanmalı,
- Saha satışları için uzaktan iletişim kurulmalı,
- Satış senaryoları üzerine kurulu bir eğitim programı planlanmalı,
- Sürdürülebilir ve geliştirilebilir teknolojiye yatırım yapılmalıdır.

4.2. CRM MİMARİSİ

Organizasyonlar CRM stratejileri geliştirirken, müşterilerinin satın alma davranışlarını da dikkate almalıdır. CRM mimarisi; Operasyonel CRM, Analitik CRM ve İşbirlikçi CRM olmak üzere, üç unsur ile tanımlanmıştır (Zengyou, 2004).

- Operasyonel CRM
- Analitik CRM
- İşbirliğine yönelik CRM



Şekil 4.2. Bir CRM mimarisi

Kaynak :(2005) http://www.erpcrm.com/crm_anasf/crm_mimarisi.htm 2003' dan alınmıştır.

- **Operasyonel CRM**

Müşteri ilişkileri yönetimi (CRM)'in bu biçimi aslında tipik iş fonksiyonlarının kapsandığı CRM çözümlerinden oluşur. Bu fonksiyonlara örnek olarak müşteri hizmetleri, sipariş yönetimi, faturalama, satış ve pazarlama otomasyonu gibi süreçler verilebilir. Bu çözümler daha çok kurumsal sistem içerisindeki finans, insan kaynakları gibi farklı iş fonksiyonlarının entegre bir yapıya kavuşturulması için kullanılmaktadır.

- **Analitik CRM**

Analitik CRM , kullanıcılara ait verilerin elde edilmesi, depolanması, işlenmesi, analiz ve tahminlere dönüştürülerek raporlanması işlemlerini gerçekleştirmektedir. Böy-

lelikle CRM'in operasyonel ve entegrasyon özellikleri üzerine analiz ve raporlama özellikleri eklenmektedir.

- **İşbirliğine yönelik CRM**

İşbirliğine yönelik CRM , aslında diğerlerinin en uygun birleşiminden oluşmaktadır. Müşteriler ile şirketler arasında tam anlamıyla bir etkileşim ve koordinasyon ağının oluşmasına imkân veren , farklı iletişim kanallarından (web, telefon, e-posta vb) gelen bilgilerin, değere dönüştürülmesini sağlayan bir süreçtir. İşbirliğine yönelik CRM çözümleri müşteri ile etkileşime imkân veren tüm fonksiyonları içermektedir.

CRM, ilk olarak müşterinin mevcut durumuyla ilgilenmiş ama pazarlamanın geldiği nokta ile artık müşterinin yapabileceği hareketleri de bilinmek istemektedir. Bu noktada mevcut datalarla bir desen oluşturmaya çalışan veri madenciliği kullanılmaktadır.

4.3. MÜŞTERİ SEGMENTASYONU

Müşterinin, firmaya kazandırdığı rakam önemlidir. Buna gore cok karlı, az karlı müşteri diye ayrılabilir. Karlı grupların karlı olmayanlardan ayrılması, pazarlama faaliyetlerinin müşteri grupları bazında farklılaştırılmasını sağlar. Müşteri segmentasyonu, müşterinin çeşitli özelliklerini değerlendirerek bir sınıflandırmaya tabi tutulması anlamına gelebilir (Akbulut, 2006)**.**

Müşteri segmentasyonu, şirket açısından müşterilerin değerini grupları itibariyle ortaya koyarak, pazarlama faaliyetlerinin uygulanmasında en değerli müşterilere öncelik verilmesini sağlar ve böylece mevcut müşterilerden elde edilen geliri maksimize eder. Doğru kitleye doğru mesajın gönderilmesini sağlayarak, kampanya yönetim maliyetlerini azaltır.

Müşteri segmentasyonu ,müşterinin pekçok özelliği kullanılarak yapılabileceğinden kullanılacak veriye göre segmentasyon türleri oluşmuştur.

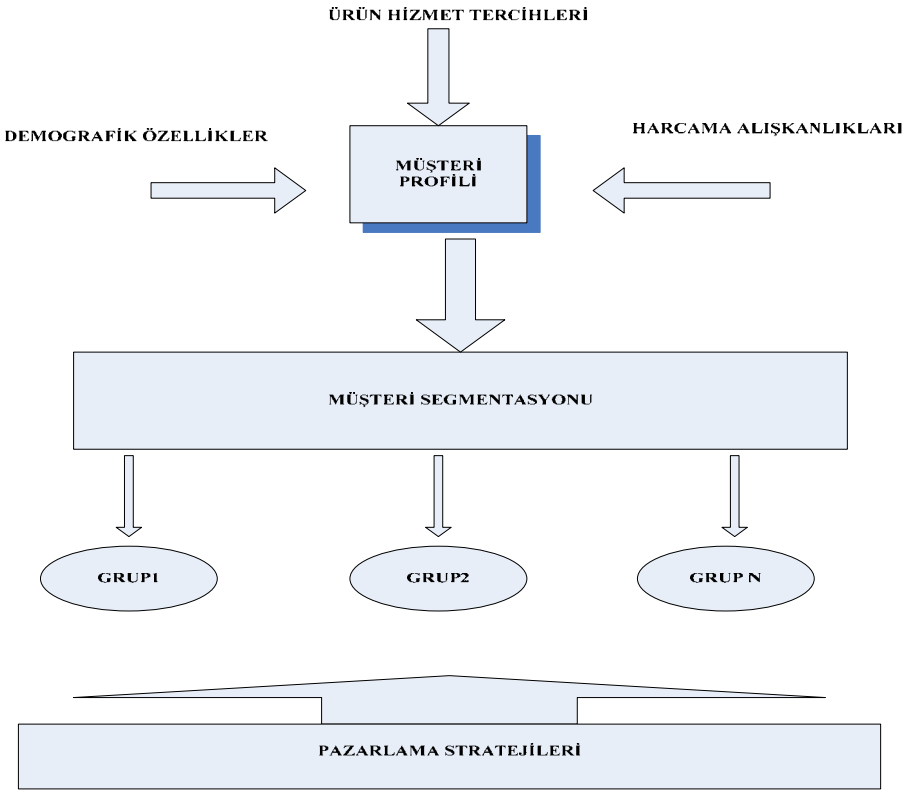
Müşteriler; coğrafi, demografik, psikografik ve davranışa dayalı özelliklerine göre gruplandırılırlar (Akbulut, 2006).

- Coğrafi segmentasyon, pazarı şehir, bölge, ülke gibi coğrafik birimlere ayırır. Günümüzde birçok şirket; reklam, promosyon ve satış faaliyetlerini farklı özellikler gösteren coğrafik birimlerin ihtiyaçları ile uyumlu hale getirmeye çalışmaktadır.

- Demografik segmentasyon; müşterileri yaş, cinsiyet, aile büyüklüğü, gelir, meslek, tahsil gibi kriterleri baz alarak gruplandırır.

- Psikografik segmentasyon; müşterileri sosyal sınıf, yaşam tarzı ve kişilik özelliklerine dayalı olarak gruplandırır.

- Davranışa dayalı segmentasyon ise müşterileri ürünü kullanma, kendisine sunulan önerilere cevap verme, satın alma sıklığı, satın alma miktarı gibi kriterlere dayalı olarak gruplandırmaktadır.



Şekil 4.3. Müşteri Segmentasyonu yapısı

Kaynak: Akbulut, (2006). Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu' ndan alınmıştır.

Müşteri segmentasyonu yapmadan önce bir müşteri profili belirlenmelidir. Bu süreci aşağıdaki gibi tanımlayabiliriz.

- Temel demografik özellikleri
- Harcama alışkanlıkları
- Hangi ürün ya da hizmetleri kullanıyorlar?
- Müşterinin alışveriş sıklığı
- Niçin sizin ürün ya da hizmetlerinizi tercih ederler?
- En değerli müşteri tipine yakın bir profile nasıl elde edebiliriz?

Segmentasyon çalışmalarının sonucunda beklediğimiz bazı faydalar vardır.

- segmentasyon çalışması sonucunda , oluşturulan kümelere özel kampanyalar geliştirilebilir.
- Yeni çıkacak bir ürünle muhtemel olarak ilgilenecek müşteri tipi belirlenebilir.
- Sonucu tahmin edilen çalışmalar için insane gücü harcanmaz.
- Kaybedilecek müşteri tahmin edilebilip önlenebilir.
- Yüksek gelir getiren müşteri profile belirlenip o tip müşteri kazanılabilir.

Bundan sonraki kısımda, bütün bu kavramları kullanarak bir telekom firmasının verileri kullanılarak, 3 çeşit segmentasyon yapılacaktır. Bu çalışmalardan ilki abonesel segmentasyon olarak isimlendirilecek şirket için en önemli abone tablosundan alınan bilgilerle, ikincisi müşterinin sabit telefon detay görüşmeleriyle, üçüncü segmentasyon ise müşterinin kişisel özelliklere göre yapılacaktır.

5.UYGULAMA

Bu çalışmada veri ambarı, veri madenciliği, CRM ve Müşteri Segmentasyonu gibi konular aktarılmaya çalışıldı. Bütün bu bahsedilen kavramlar, VTBK süresinde birbirine bağımlıdır. Birbirlerine yaslanarak çalışırlar. Burada en alttaki basamak veri ambarı, en üstte ise CRM vardır.

CRM, daha önce de bahsedildiği gibi işletmelerin işini kolaylaştıran, müşteriye ulaşmayı kolaylaştıran bir oluşumdur. CRM için veri madenciliği projeleri kullanılır. Veri madenciliği çalışmaları ile elde edilen sonuçlar da CRM' e yansır. Örneğin belli bir müşteri kesimini tek adımda bulmak veya getirisi düşen müşterileri tek bir küme ile bir hamlede bulmak gibi. Bu hedefler de bulunduktan sonra pazarlama ve kampanya sunumuna gidilebilir. Bu uygulama çalışmasında; müşteri, aynı özellikleri taşıyan kümelere ayrılacaktır. Bir başka deyişle müşteri segmentasyonu yapılacaktır.

Müşteri segmentasyonu yapılırken müşterilerin coğrafi, demografik, psikografik ve davranış özellikleri incelenerek verimli sonuçlar verebilecek kolonlar alınacaktır.

Müşteri segmentasyonu yapmak için kullanılacak olan program SPSS Clementine, kullanılacak kümeleme algoritması ise yapay sinir ağlarından türeyen sonuç değerleri olmadan denetimsiz öğrenim algoritması ile çalışan KOHONEN' dir. Bu algoritmanın seçilmesinin nedeni, büyük ölçekli datalarda da güvenilir sonuç vermesi, küme sayısının algoritma tarafından belirlenmesidir. Küme sayısının baştan belli olması, sonucu kısıtlamaya götürebilir.

5.1. UYGULAMA İÇİN VERİ AMBARI TASARIMI

Müşteri segmentasyonu uygulaması için piyasada önemli bir pazar payına sahip bir telekomunikasyon şirketinin örnek verileri kullanılacaktır. Toparlanacak olan verilerle örnek bir veri ambarı oluşturulup bu ambar üstünde çalışılacaktır.

Bir şirket içinde; ürün, satış, ücretlendirme, personel gibi bilgiler sadece bir operasyonel veri tabanında da tutulabilir veya farklı pekçok operasyonel sistemde de

tutulabilir. Çalışmada kullanılacak olan şirketin büyüklüğü, hizmet çeşitliliği, geçirdiği değişim aşamaları gözönüne alındığında pek çok farklı operasyonel sistemde veri tutulduğunun farkına varılır. Fakat bu özellikle üst düzey personel için sorun çıkarmaktadır. Her biri farklı sistemlerde tutulan bu bilgiler bütünleşik raporlamayı engellemektedir. Bu yüzden bir veri ambarı ihtiyacı doğmuş ve iki basamaklı bir uygulama kurulmuştur.

Farklı kaynaklardan gelen bütün şirket dataları, ilk olarak ODS (Operational Data Store) adı verilen yapıya aktarılır. Bu veritabanındaki datalar ya otomize edilmiş bir programla (Golden Gate vs.) gerçek sisteme paralel olarak veya çeşitli aktarım toollarıyla (ETL) belirli aralıklarla güncellenir.

Örnek ODS veritabanında aşağıdaki sistemlerden alınan kayıtlar bulunacaktır. ODS veritabanında bulunan veriler işlenmemiş ham verilerdir.

-Abonelik ve Müşteri Bilgileri

-Fatura Bilgileri

-Ödeme Bilgileri

-Görüşme ayrıntıları

Bütün verileri, ODS veritabanına aktarmak her zaman çok düzenli bir yapı sunmayacaktır. Bahsedildiği gibi ODS deki veri ham veridir. Bütün bu datalar örneğin tek bir müşteri numarasından bahsedildiğinde gelmeli, bir numarayla istenilen her bilgiyi alabilmeliyiz. Bu bize hem daha düzenli bir yapı sağlayacak hem de OLAP küpleri oluşturulmasında fayda sağlayacaktır.

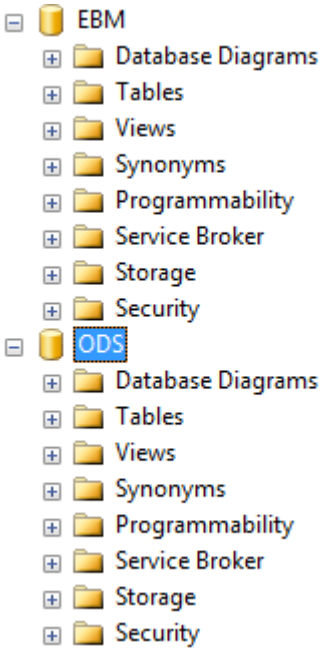
İşte bu ihtiyaçlardan dolayı operasyonel datanın düzenlenerek, belki özetlenerek belki de daha da ayrıntıya inilerek bir kaç boyutlu hale gelmesi için yeni bir veritabanı oluşturulacaktır. Bu veritabanın adı EBM' dir.

EBM' i oluşturmak için önce ODS sisteminin çok iyi analiz edilmesi gerekmektedir. ODS' deki tabloların arasındaki ilişki incelenerek tek formatlı bir yapı oluşturma-

ya çalışmalıdır. Örneğin bu şirketin farklı farklı formatlarda gelen tahakkuk bilgileri tek tabloda toplanabilmedir. Örneğin dönem bilgisi hep aynı formatta olmalıdır. Fatura tipleri aynı fakat ifade edilişleri farklıysa aynı şekilde gösterilmelidir. Lookup dediğimiz yardımcı tablolar ve ana tablolar olmalı karışıklıktan uzak durulmalıdır.

İşte böyle bir sistemde, müşteri kümeleme ihtiyacı duyulabilir. İncelenen müşterinin nasıl bir müşteri olduğunu kullandığımız tabloya bakarak anlamak zordur , her seferinde öğrenmek istediğimiz bilgiyle ilgili alana sorgu çekilmesi gerekmektedir. Veya sorgulamak istenilen kolonlar farklı sistemlerde olabilir. Bu yüzden müşteri hakkındaki istenilen bütün özellikleri kullanarak yapılacak olan segmentasyon çalışmasıyla her seferinde datayı sorgulamaya gerek kalmaz. Bu müşteri segmentasyonu çalışması sonunda o müşteriye verilmiş SegmentID ile bu özellikleri tek numara ile çözülebilir.

Bütün bu konuları gözönünde tutarak, bu şirketin veriambarından bu çalışma için gerekecek tablolardan bikaç tanesinden random olarak belli sayıda müşteri belirledikten sonra ona bağlantılı diğer tabloları da çekerek küçük çapta örnek bir veri ambarı oluşturulmuştur.



Şekil 5.1. EBM ve ODS yapısı (Bu şekil uygulama çalışmasından alınmıştır.)

Bu ilk adımdan sonra ABONE tablosunun bağıntılı olduğu başka kaynak tabloları da ODS veritabanına aktarıldı. Bu tabloların aşağıda bir özeti bulunmaktadır.

Örnek olarak kullanılan telekomünikasyon şirketi, müşteriye yaklaşık 15 hizmet türüyle ürün sunmaktadır. Milyonlarca müşterisi olan bu firmada bazı hizmet türleri için farklı sistemler kullanılmaktadır. Örneğin PSTN(Sabit Telefon) ve XDSL aboneleri farklı tablolarda tutulmaktadır. Pek çok sayıda tablo olduğu için örnek veritabanına lazım olabilecek kadarıyla kısıtlı sayıda tablo alınacaktır. Bu tablolar aşağıda kısaca anlatılmıştır.

-CRM.FIRMA: ADSL hizmeti veren anlaşmalı firmalar
-CRM.SEKTOR: ADSL sistemi için sektör kodunu tutan yardımcı tablo
-CRM.TARIFE: ADSL sistemi için tarife yardımcı tablosu
-CRM.TAHAKKUK200908: ADSL sistemi için tahakkuklar
-dbo.ABONE: PSTN ve diğer hizmetler için abone bilgileri
-dbo.DETAY200908: PSTN aboneleri için arma detayları
-dbo.DETAY200907: PSTN aboneleri için arma detayları
-dbo.DetayYonKodu: Detaylar için yönleri ifade eden yardımcı tablo
-dbo.FATURA: Tahsilat sisteminden gelen fatura bilgileri
-dbo.MUSTERI: PSTN ve diğer hizmetler için müşteri bilgileri
-dbo.TAHAKKUK200908: Tahakkuk sisteminden gelen tahakkuk bilgileri
-dbo.ODEME_KANALLARI: Fatura ödeme için kullanılan yöntemler
-dbo.XDSLABONE: ADSL isteminden gelen abonelik bilgileri
-dbo.XDSLMUSTERI: ADSL sisteminden gelen müşteri bilgileri
-TMS.BIRIM: Bütün telekom birimleri için yardımcı tablo
-TMS.FIRMA: Bütün telekom için yardımcı firmalar
-TMS.MESLEK: Müşterilerin meslek kodlarını tutan yardımcı tablo
-TMS.MUDURLUK: Telekom müdürlüklerini tutan yardımcı tablo
-TMS.MUSTERI: Müşteri tablosu
-TMS.TARIFE_PAKET: PSTN ve diğer hizmetler için tarife tablosu

Burada bahsedilen tablolar operasyonel sistemlerden gelen tablolardır. Burda da görüldüğü gibi aslında ikisi de bir ürün olan hizmetler farklı tablolarda tutulmaktadır. Bu tablolarda HIZMET_NO değerini vermesi gereken kolon farklı isimlerde farklı data tipinde olabilir. Burada bütün bu ürünleri birleştirebilecek bir database yapılmak istenmektedir. İşte bu veritabanı da EBM'dir. EBM deki tablolardan birkaç örnek aşağıdadır.

-EBM.tbAbone- Firmanın bütün müşterilerinin bir ID ile tanımlandığı tablo ve işemri hareketleri

-EBM.tbFaturaKalemleri: Faturalardaki detay kalemleri

-EBM.tbKullanımDetaylari: Hizmetlerin kullanım detayları

-EBM.lkTelekomOfisleri: Telekom ofisleri

-EBM.tbUrun: Bütün hizmetler için tarife paketleri

-EBM.lkUrunTipi: Bütün hizmetler için alt hizmet türleri

-EBM.tbFaturalar: Bütün hizmetler için faturalar

-EBM.lkServisOfisleri: Bütün hizmet veren santral ve birimler

-EBM.lkOdemeKanali: Faturaların ödenme şekli

-EBM.tbOdemeler: Faturalar ve ödeme hareketleri

- EBM.lkMüsteriKümeleri: Müşteri kümeleme yardımcı tablosu

Görüldüğü gibi EBM adı verilen yapı bizi bu dağınık veri ortamından kurtarmaya faydalı olacaktır. İşte burada uygulaması yapılacak çalışma aslında bir çeşit yardımcı (lookup) tablodur.

Bu çalışmada amaç, EBM.lkMarketingSegment adı verilen, bu firma için müşteri kümelerini ifade eden bir tablo oluşturmak ve her ID in bir grup müşteriyi ifade etmesini sağlamaktır. Bu sayede bu ID yi abonelerle eşleştirerek, o özellikleri barındıran herhangi bir kampanyada kolayca hedef müşteriler bulanabilir. Şimdi bu çalışmayı gerçekleştirmek için veri madenciliği sürecine geçilecektir.

5.2.UYGULAMA İÇİN VERİ MADENCİLİĞİ SÜRECİ

Daha önceki bölümlerde, veri madenciliği çalışmaları için standart bir sürecin işlediğinden bahsedilmişti. Bu da bir kampanya yürütmenin parçasıdır. Aslında her çalışma, bir analiz, ön araştırma gerektirir. İşte veri madenciliği sürecinde de öncelikle ihtiyaç duyulan bilgileri, sözkonusu bilginin getireceği avantajları, bu bilgilerle neler yapabileceğini ortaya koyulması gereken bu sürece veri madenciliği süreci denir. Şimdi yapılacak bu uygulama için veri madenciliği süreci anlatılacaktır.

5.2.1.İŞ ANLAYIŞ SAFHASI

Yukarıda da bahsedildiği gibi birinci safha iş anlayış safhasıdır. Bu safhada yapılmak istenen iş belirlenir ve ona göre ihtiyaçlar belirlenir. Buradan yola çıkarak segmentasyon çalışmasının ilk safhasına başlanır.

Örnek olarak kullanılan şirket, müşterilerine 10 dan fazla olmak üzere pek çok hizmet türüyle ürün sunmaktadır. Bu yüzden de farklı kaynakları vardır. Her müşteri için bir ABONE_ID kavramı vardır. Bu kavram ile müşteriyi kolayca bulunabilir. Fakat bu müşteriye ait pek çok bilgi farklı tablolardadır. Örneğin müşterinin kişisel özellikleri bir tabloda, abonelik özellikleri bir tabloda , fatura bilgileri başka tablodadır. Bu yüzden, hedeflenen müşterinin pekçok özelliğini bulmak için her defasında büyük, zahmetli sorgular yazılmaktadır. Bu nedenle, tek bir çalışmayla bu müşterilere bir sınıflandırma yapma isteği doğar. İşte bu çalışmaya da müşteri segmentasyonu denir.

Bu çalışmada kullanılacak olan hizmet türü PSTN(Sabit Telefon) hizmetleridir. PSTN müşterileri, kendi aralarında kullandıkları özelliklere göre kümelenecektir.

Bu çalışma yapılırken amaç, müşterinin birkaç özelliğini birden birarada tanımlayan kümeler bulmaktır. Örneğin herhangi bir kampanya için, GERÇEK (Bireysel Müşteri) müşteri olup, ortalamanın üstünde gelir elde edilen veya herhangi bir bölgede yaşayıp fazla sayıda şehiriçi görüşme yapan müşterileri tanımlamak gibi...

Amaç: Bu çalışmada hedeflenen, bu kümelere bir ID vermek ve bu ID yi EBM veritabanındaki en önemli tablo olan tbAbone içinde bir kolon olarak belirtmektir.

Tablo 5.1. tbAbone tablosu (Bu tablo uygulama çalışmasından alınmıştır)

tbAbone	Column1
AboneID	Algoritmanın ürettiği ID
AboneSID	Kaynak sistemden gelen ID
HizmetNumarasi	Hizmet Numarasi
CustomerID	Aboneliğin sahibi için üretilen ID
ProductID	Ürün için üretilen ID
IslemTarihi	İşemrinin girildiği tarih
AbonelikDegeri	Abonelikiptali ve diğer işemirlerine göre -1, 0 veya 1 değeri
TelekomOfisID	Telekom Müdürlüğü
ServisOfisID	İşlemi yapan Telekom birimi (Bayiler Dahil)
MüsteriKümeID	Müşteri Kümeleri
AboneSegmentID	Gerçek veya Tüzel Müşteri

Tablo 5.2. lkMüsteriKümesi tablosu (Bu tablo uygulama çalışmasından alınmıştır)

lkMüsteriKümesi
MüsteriKümesiID
ÜstSegmentID
DavranısSegmentID
AboneSegmentID

İhtiyaçlar: Bu çalışmayı yapabilmek için gerekli olan datalar vardır. İlk önce bu segmentasyon çalışmasının neler içerdiğini ve kriter olarak nelerin seçileceğini belirlemek için ODS deki tablolar incelenmiştir.

Tablo 5.3. ODS deki bazı tablolar ve alanları (Uygulamadan alınmıştır)

dbo.TAHAKKUK200908	dbo.XDSLABONE:	dbo.ABONE :	dbo.MUSTERI:	dbo.FATURA:	dbo.DETAY200908:
[FATDONEMY]	[MUSTERI_ID]	[ID]	[ID]	[HIZMET_NO]	[ID]
[IL_PLAKANO]	[CIHAZ_NO]	[HIZMET_NO]	[UST_MUSTERI_ID]	[FATURA_NO]	[HIZMET_NO]
[TMS_ABONE_ID]	[PSTN_NO]	[MUSTERI_ID]	[GERCEK_TUZEL]	[ODEME_DONEMI]	[MUSTERI_ID]
[CIHAZ_NO]	[SOZLESME_TARIHI]	[IL_MUDURLUK_ID]	[KIMLIK_CINSI]	[FATURA_TAKSIT_NO]	[IL_MUDURLUK_ID]
[HIZMET_TUR]	[HIZ_KODU]	[MUDURLUK_ID]	[AD]	[HIZMET_TURU]	[MUDURLUK_ID]
[MUD_KODU]	[STATIK_IP]	[PROJE_TAKIP_ID]	[SOYAD_UNVAN]	[ALT_HIZMET_TURU]	[PROJE_TAKIP_ID]
[ABN_TURU]	[KULLANICI_KIMLIK]	[HIZMET_TURU]	[KIMLIK_NO]	[ABONE_TURU]	[HIZMET_TURU]
[IPT_KODU]	[MODEM_FIRMA_ID]	[ALT_HIZMET_TURU]	[BABA_ADI]	[ESKI_HIZMET_NO]	[ALT_HIZMET_TURU]
[TES_TARIH]	[KULLANICI_ID]	[ABONE_TURU]	[ANA_ADI]	[TAHSILAT_TAKIP_KODU]	[ABONE_TURU]
[TES_ISMNO]	[ISLEM_TARIHI]	[TARIFE_PAKET_ID]	[CINSIYETI]	[ABONE_ID]	[TARIFE_PAKET_ID]
[ISM_TARIH]	[XDSL_TIP_KODU]	[EV_IS]	[UYRUGU]	[FATURA_TIPI]	[EV_IS]
[ISM_NO]	[EMAIL_ADET]	[TAHSILAT_TAKIP_KO	[DOGUM_YERI]	[SON_ODEME_TARIHI]	[TAHSILAT_TAKIP_KODU]
[ISM_TURU]	[EMAIL_KOTA]	[DURUM]	[DOGUM_TARIHI]	[REFERANS_NUMARASI]	[DURUM]
[BEL_KODU]	[WEB_KOTA]	[SOURCE_SYSTEM_ID]	[KIMLIK_VER_YER]	[FATURA_DURUM]	[SOURCE_SYSTEM_ID]
[MES_KODU]	[WEB_ADRES]		[KIMLIK_VER_TARIHI]	[ORJINAL_TUTAR]	[DosyaId]
[EV_IS_KODU]	[MODEM_SERI_NO]		[NF_KAYIT_IL]	[FATURA_DUZELTME_TUTAR]	[DonemKey]
[KURUM_KODU]	[SERVIS_BAS_TARIHI]		[NF_KAYIT_ILCE]	[KALAN_TUTAR]	[AlanKodu]
[ADR_DURUM]	[SERVIS_BIT_TARIHI]		[NF_KAYIT_MAHALLE]	[ODEME_TARIHI]	[HizmetNo]
[TURLA_KODU]	[FIRMA_ID]		[NF_CILT]	[DOVIZ_TUTARI]	[AboneKey]
[NESKI_KON]	[SUBE_ID]		[NF_SAYFA]	[DOVIZ_TIPI]	[AboneID]
[NYENI_KON]	[TARIFE_TURU]		[NF_KUTUK]	[DOVIZ_KURU]	[DetayTipiKey]
[FESKI_KON]	[PAKET_ID]		[OGRENIM_DURUMU]	[ADSOYAD_UNVAN]	[AramaTarihiKey]
[FYENI_KON]	[PROMOSYON_ID]		[VERGI_IL_ADI]	[TT_MUDURLUK_KODU]	[AramaSaatiKey]
[DONEM_KON]	[TAKSIT_SAYISI]		[VERGI_DAIRESI]	[EPS_ID]	[AramaSalise]
[TOPLAM_KON]	[KURULUM_EVET]		[VERGI_HESAP_NO]	[FATURA_SERI_NO]	[AramanHizmetNo]
[KREDI_BAKIYE]	[SATIS_KULLANICI_ID]		[TC_KIMLIK_NO]	[FATURA_SIRA_NO]	[AramaSuresi]
[SMS_UCR]	[SATIS_KODU]		[TICARET_ODA_SICIL]	[ICMAL_GRUP_KODU]	[DetayHKontor]
[VMS_SABIT]	[KURULUM_KULLANICI_ID]		[SANAYI_ODA_SICIL]	[REESKONT_TUTARI]	[DetayDKontor]
[FAS_TESIS]			[VAKIF_DERNEK_SICIL_NO]	[REESKONT_FLAG]	[ArayanYerKey]
[FATURA_NO]			[ANNE_KIZLIK_SOYAD]	[KAPAMA_TUTARI]	[AramanYerKey]
[SODE_TARIH]			[MESLEK_ID]	[KAPAMA_FLAG]	[TarifePaketiKey]
[OZL_ODM_GRP]			[KULLANIM_DURUM]	[GECIKME_TUTARI]	[UcretT1]
[KONUSMA_UCR]			[GELIR_DURUMU]	[GECIKME_FLAG]	[KrediInd1]
[TOP_SA_UCR]			[KURUMSAL]	[TAKSITLENDIRME_BEDELI]	[YonKodu]
[TOP_MA_UCR]			[CALISAN_SAYISI]	[TAKSITLENDIRME_BEDELI_FLA	[AYonKodu]
[TOP_GEMI_UCR]			[NUFUS_KILITLI]	[TAKSIT_GECIKME_TUTARI]	[Kademe]
[ABONMAN_UCR]			[SOURCE_SYSTEM_ID]	[TAKSIT_GECIKME_FLAG]	[TPaket]
[PAKET_UCT]			[KURUMSAL_VIP]	[FAZLA_ODEME_TUTARI]	[FatAramaSuresi]
[KAPAMA_UCR]			[MARKA_ISMI]	[FAZLA_ODEME_FLAG]	[TemizAramanNo]
[MUTFRK_UCR]			[VERGI_DAIRESI_KODU]	[DETAY_01]	
[FONO_UCR]			[SIRKET_TURU]	[DETAY_02]	
[DETAY_UCR]			[FAAL_TERK]	[DETAY_03]	
[KAMPANYA_UCR]			[UYRUK_ID]	[DETAY_04]	
[VIDEO_PAKET_UCR]				[DETAY_05]	
[VIDEO_GORUSME_UCR]				[DETAY_06]	
[TESIS_UCR]				[DETAY_07]	
				[DETAY_08]	
				[DETAY_09]	
				[DETAY_10]	
				[DETAY_11]	
				[DETAY_12]	
				[DETAY_13]	
				[DETAY_14]	
				[DETAY_15]	
				[SERBEST_BOLGE_ABONE]	
				[SYS_GONDERME_DURUM]	
				[SYS_GONDERME_ZAMANI]	
				[ISLENEN_DOSYA_ADI]	
				[TUTAR_DUZELTME_FLAG]	
				[FATURA_ID]	
				[CVG_FATURA_SIRA_NO]	
				[CVG_GONDERILEN_TUTAR]	
				[CVG_BEKLEYEN_KAYIT_SAYISI]	
				[PAKET_TIPI]	
				[TALIMAT_KURUM_KODU]	
				[TALIMAT_DOSYA_ADI]	
				[UMTH_VARMI]	
				[UMTH_GONDERILEN_TUTAR]	
				[TAKSIT_KAPAMA_TUTARI]	
				[TAKSIT_REESKONT_TUTARI]	
				[SOURCE_SYSTEM_ID]	
				[aboneID]	

Hedef: Yukarda birkaç örnek verilmiş bu tablolardan anlamlı değerler veren kümeler oluşturmak.

Proje Planı:

- ODS veritabanımızdaki tabloları ayrıntılı olarak , aralarındaki ilişkileri belirleyerek incelemek

-Örnek bir veri seti oluşturmak

-Bu veri setindeki eksiklikleri bulup düzeltmek

-Modellemeye karar vermek

-Modelleme testleri yapmak

-Bulunan en doğru sonucu sunmak

5.2.2. VERİ ANLAYIŞ SAFHASI

Bir önceki aşamada veritabanına kabataslak bakılmıştı. Bu aşamada da fazla ayrıntıya girmeden bu tabloların aralarındaki ilişkilere ve içindeki veriye göz atılacaktır. Yukarda örnek olarak verilen tablolar birbirlerine bazı kolonlarla ilişkilidir.

Bu çalışma için, firmanın bu tablolarından örnek olarak bir data seti alındı. Bu data seti, sadece PSTN hizmeti için (Bu tlf abonelerinin de sahip olduğu XDSL (internet) ürünleri de ODS ortamına alındı.) İstanbul il sınırları içindeki müşterileri kapsamaktadır.

Analiz yapılacak örnek data kümesi şöyle seçildi.

1. Yaklaşık 17 milyon küsür adet sabit telefon hizmetlerinden (HIZMET_TURU=0) İstanbul ili icinden 62842 örnek hizmet numarası alındı.

2. Bu örnek abone tablosunu kullanarak çalışmanın yapıldığı anda tesis durumunda bulunan 62007 müşteri bulunmuştur. dbo.ABONE.MUSTERI_ID=dbo.MUSTERI.ID

3. Abone ve müşteri bilgileri alındıktan sonra bu pstn abonelerinin kaç tanesinin İnternet(XDSL) hizmeti olunduğu anlaşıldığı için data setine katılmıştır.
dbo.ABONE.HIZMET_NUMARASI=CRM.ABONE.PSTN_NO

4. PSTN abonelerinin aylık dönemler halinde faturalarını , ödenip ödenmediği bilgisini tutan tablodan bu abonelerin 200906 döneminden itibaren bütün faturaları alındı.
dbo.ABONE.HIZMET_NUMARASI=TTS.HIZMET_NO

5. PSTN abonelerinin faturaya yansıyan fatura kalemlerini ayrı ayrı gösteren tahakkuk sisteminden 200908 için dbo.TAHAKKUK200908 tablosundan örnek data seçildi.
dbo.TAHAKKUK200908 .CIHAZ_NO= dbo.ABONE.HIZMET_NUMARASI

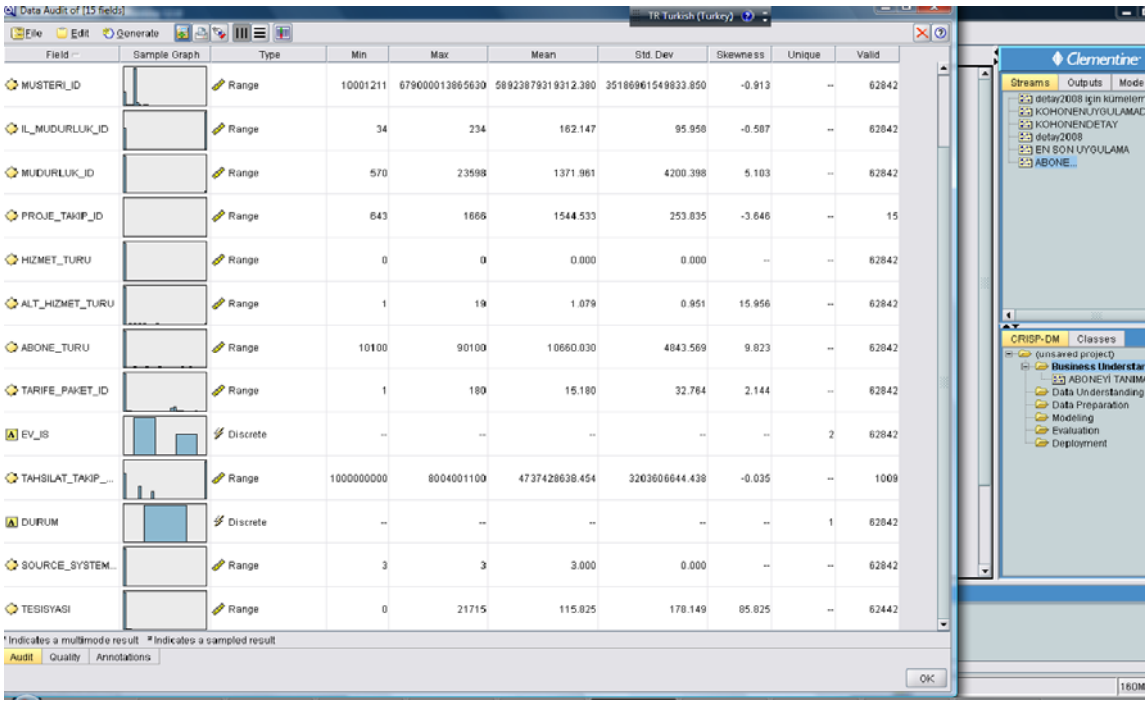
6. Bu abonelerin konuşma alışkanlıklarını da görmek için dbo.DETAY200908 tablosundan bu müşterilere ait veriler çekildi.
dbo.DETAY200908.HIZMET_NO=dbo.ABONE.HIZMET_NUMARASI

Verileri çektikten sonra bu tabloları anlamak için incelemeler yapıldı. Tabloların birbirleri ile arasındaki ilişkiler, boş olan kolonlar, bir kolonda kaç çeşit veri olabildiği konusunda analizler yapıldı. Bu analizleri yaparken bazen SQL Server da SQL dili bazen de SPSS kullanıldı.

Bu tablolarda bize bu segmentasyon çalışmasında neler ayırtedici özellik olabilir. Bunlar arasında

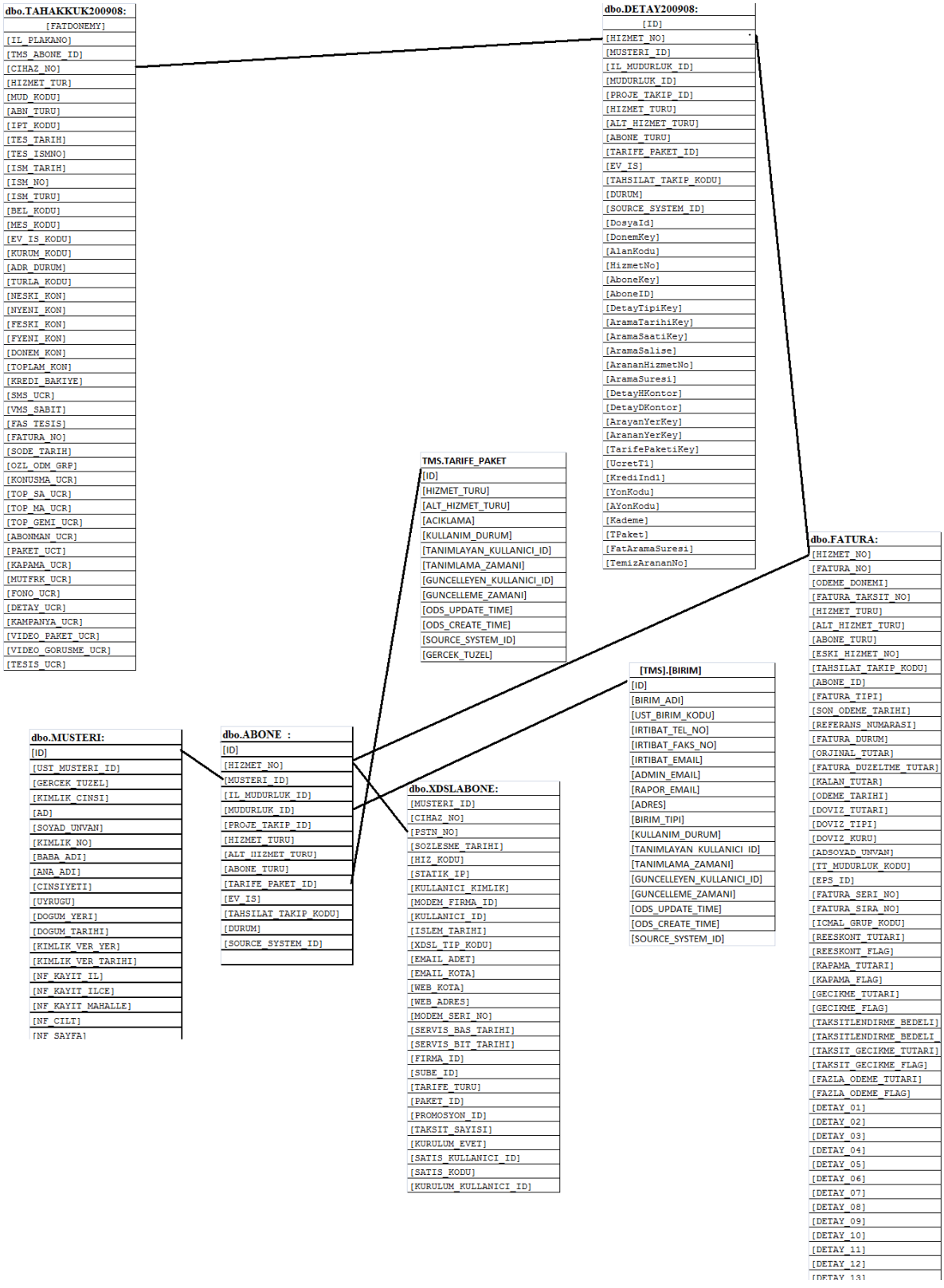
-TESIS_TARIHI (Tabloya kolaylık olması açısından ay cinsinden yaş olarak alındı)
-NF_KAYIT_IL
-KONUSMA_SÜRELERİ
-MUDURLUK_ID
-ORTALAMA_FATURA_TUTARI
-ÖDEME YAPIP YAPMADIĞI
-GERÇEK_TÜZEL
-GELİR_DURUMU
-MÜŞTERİ_YAŞI
-ARAMA_SAATI

Bu gibi kolonlara özellikle dikkat edilebilir. Örneğin ARAMA_SAATI kolonunda boş olan kayıtlar olduğu görüldüğünde sorgular buna göre yazılır.



Şekil 5.2. Datayı tanımak ve içini anlama (SPSS Data Audit)

Hangi tablodan hangi tablodaki bilgiye erişebilir, işe yarayan bişey çıkar mı düşüncesiyle aşağıdaki diagrama bakabilir.



Şekil 5.3. ODS tabloları arasındaki ilişki (Uygulama çalışmasından alınmıştır)

5.2.3. DATA HAZIRLIĞI

Verileri yükledikten sonra çeşitli sorgularla verinin durumu incelendi. Bunun için bütün tabloları inceleyerek boş , dolu kolonları veya anlamsız veri içerenleri tespit edilmeye çalışıldı. Örneğin bir servis numarası için bütün tablolara göz atıldı. Şimdi bu tablolardan hangi bilgilerin alabileceği anlatılacaktır.

dbo.DETAY200908: Bu tabloda, pstn müşterilerinin ay boyunca yaptığı görüşmeler saniye, kontür (FaturaSüresi), aranan yön, aranan numaralar, arama saati, kademe (santraller arası bir deyim) vs. bilgileri tutulmaktadır. Burada bahsedilen bu özellikler, her müşterinin ortalaması incelendiğinde milyonlarca data eder. Bu yüzden buradaki çok ayrıntılı bilgiler modellemeye koyulmadı. Örneğin bu müşteri en çok kimi aramış diye bir çalışma sadece bu amaca yönelik olmalıdır. Detay tablosundan bu müşterinin hangi yönlere, hangi saat dilimlerinde, ortalama ne kadar kontür harcayarak toplam kaç arama yaptığı gibi genel bilgiler segmentasyona katılabilir.

-hangi yön
-hangi saat dilimi
-ortalama kontur
-toplam arama sayısı

Bu bilgileri almak için tabloda bazı düzenlemeler yapılması gereklidir. Örneğin arama saati tabloda saat ve dakika olduğu için bu çalışmayı zorlayacaktır. Bu yüzden saat aralıklarını birkaç parçaya bölerek tabloya "aramasaatisınıf" adında yeni bir kolon eklendi. Aynı zamanda aranan yön de çok sayıda olduğu için 5 parçaya bölüp "YonKoduSınıf" adında bir kolon da eklenecektir.

1. `alter table dbo.DETAY200908 add aramasaatsınıf varchar (5)`

```
AramaSaatiKey between 180001 and 220000 ise aramasaatsınıf=C
AramaSaatiKey between 220001 and 240000 ise aramasaatsınıf=D
AramaSaatiKey between 20001 and 60000   ise aramasaatsınıf=E
AramaSaatiKey between 120001 and 180000 ise aramasaatsınıf=B
AramaSaatiKey between 60001 and 120000  ise aramasaatsınıf=A
AramaSaatiKey between 000001 and 20000  ise aramasaatsınıf=F
```

2. `alter table dbo.DETAY200908 add YonKoduSınıf varchar (5)`

```
YonKodu=1  ise YonKoduSınıf=A (ŞEHİRİÇİ)
YonKodu=2  ise YonKoduSınıf=B (ŞEHİRDIŞI)
YonKodu=3  ise YonKoduSınıf=C (MİLLETLERARASI)
```

```
YonKodu=4  ise YonKoduSınıf=D (GSM)
YonKodu=5  ise YonKoduSınıf=E (444)
YonKodu=6  ise YonKoduSınıf=F (822,111)
```

dbo.MUSTERI: Bu tabloda da müşterinin sosyal ve kişisel özellikleri tutulmaktadır. DOĞUM_TARIHI, DOĞUM_YERI, GELIR_SEVIYE, CINSIYETI, SIRKET_TURU vs. kolonları segmentasyon çalışmasına koyulacak alanlardır.

Burada da bu alanlarda bazı eksiklikler var. Örneğin DOĞUM_YERI kolonunda fazla çeşitli olduğu için bunu bölgelendirmeye ve NUFUSBOLGE adında kolon eklemeye karar verdik. NUFUSBOLGE alanı NF_KAYIT_IL alanı tek tek incelenerek yanlış yazılmış iller düzenlenip 7 coğrafi bölgeden biri olacak şekilde (bir de BILINMEYEN değeri var) update edildi. GELIR_SEVIYE, OGRENIM_SEVIYE gibi alanlarda da çok sayıda null ve numerik değer olduğu için bu kolonlar da sınıflara ayrıldı. MUSTERI_YASI ise değişken sayıda olduğu için YASSINIF isimli yeni bir kolon eklendi.

1.
```
update MUSTERI SET yas=CONVERT  (INT, (CONVERT (VARCHAR (4),GETDATE
(),112)))-
CONVERT  (INT, (CONVERT (VARCHAR (4), ISNULL (DOGUM_TARIHI, CONVERT
(VARCHAR (4),GETDATE (),112)))))

18-20  ise  A, 20-30 ise B, 30-40 ise C, 40-50 ise D,50 üstü ise E
```

2.
```
update MUSTERI SET DOGUM_YERI=isnull (DOGUM_YERI,'-')
```

3.
```
select sum (yas)/ ( (select COUNT (*) from MUSTERI)- (select COUNT
(*) from MUSTERI  WHERE YAS=0)) from MUSTERI
```

4.
```
UPDATE MUSTERI  SET OGRENIM_DURUMU=5 WHERE OGRENIM_DURUMU IS NULL
0=okuryazzar ve altı, 1=ilköğretim, 2=ortaogretim,3=lise,4=üniversite
```

5.
```
UPDATE MUSTERI  SET GELIR_DURUMU=100 WHERE GELIR_DURUMU IS NULL

GELIR_DURUMU-YTL cinsinden gelir Araligi: 00-50 araligi Gerçek için,
50-99 araligi Tüzel/Kurumsal için
01: 0 - 500     -A
05: 501 - 1000  -B
10: 1001 - 1500 -C
15: 1501 - 2000 -D
20: 2001 - 3000 -E
25: 3001 - 4000 -F
30: 4001 - 5000 -G
35: 5001 - 7500 -H
40: 7501 - 10000-J
45:10001-       -K
```

6. UPDATE DBO.MUSTERI SET NF_KAYIT_IL='BILINMEYEN' WHERE NF_KAYIT_IL IS NULL

```
UPDATE DBO.MUSTERI SET   NF_KAYIT_IL='AZERBEYCAN'   WHERE
NF_KAYIT_IL='BAKÜ'
```

7. ALTER TABLE DBO.MUSTERI ADD NUFUSBOLGE VARCHAR (50)

```
 NF_KAYIT_IL IN
('OSMANİYE','MERSİN','ADANA','ANTALYA','HATAY','ISPARTA','BURDUR') ise
NUFUSBOLGE='AKDENİZ'
NF_KAYIT_IL IN
('YOZGAT','ŞANLIURFA','TOKAT','SİVAS','NİĞDE','KONYA','KIRŞEHİR','KIRIK
KALE','KAYSERİ','AKSARAY','ANKARA','ÇANKIRI','ESKİŞEHİR','KARAMAN') ise
NUFUSBOLGE='İÇ ANADOLU'

NF_KAYIT_IL IN
('ADIYAMAN','DİYARBAKIR','ŞIRNAK','SİİRT','MUŞ','MARDİN','HAKKARİ','BAT
MAN','BİNGÖL','BİTLİS') ise NUFUSBOLGE='GÜNEYDOĞU ANADOLU'

NF_KAYIT_IL IN
('UŞAK','MUĞLA','MANİSA','KÜTAHYA','İZMİR','AFYONKARAHİSAR','AYDIN','ÇA
NAKKALE','DENİZLİ') ise NUFUSBOLGE='EGE'
NF_KAYIT_IL IN  ('YALOVA','SAKARYA','TEKİRDAĞ',
'KOCAELİ','KIRKLARELİ','İSTANBUL','BALIKESİR','BİLECİK','BURSA','EDİRNE
') ise NUFUSBOLGE='MARMARA'

NF_KAYIT_IL IN
('ZONGULDAK','TRABZON','SİNOP','SAMSUN','RİZE','ORDU','AMASYA','GİRESUN
','KASTAMONU','GÜMÜŞHANE','KARABÜK','ARTVİN','BARTIN','BAYBURT','BOLU',
'ÇORUM','DÜZCE') ise NUFUSBOLGE='KARADENİZ'

NF_KAYIT_IL IN
('VAN','MALATYA','NEVŞEHİR','KİLİS','KARS','KAHRAMANMARAŞ','IĞDIR','AĞR
I','ARDAHAN',
'ELAZIĞ','ERZİNCAN','ERZURUM','GAZİANTEP') ise NUFUSBOLGE='DOĞU
ANADOLU'

NF_KAYIT_IL IN
('ÖZBEKİSTAN','PEKİN','MOSKOVA','LONDRA','LARNAKA','KÖLN','KIBRIS','JAP
ONYA','İSPANYA','İRAN',
'IRAK','SURİYE'',HARLOW','GÜRCİSTAN','AZERBEYCAN','BRİNDİSİ','BULGARİST
AN','ÇİN','FRANSA','MOSTAR','NEW
YORK','NIGERIAN','TUNCELİ','YUNANİSTAN') ise NUFUSBOLGE='YABANCI'
```

8. SELECT UYRUGU,COUNT (*) FROM MUSTERI GROUP BY UYRUGU ORDER BY UYRUGU

Bu sorgunun sonucunda cok sayıda düzensiz data çıktı. Bu yuzden degerlendirmeye alınmadı. Fakat düzenlenip alınabilir.

9. SELECT * FROM DBO.MUSTERI WHERE GERCEK_TUZEL='G' AND CINSIYETI IS null

Bunun sonucunda çıkan GERÇEK abonelerin CINSIYET kolonu dolduruldu.

dbo.ABONE: Bu tabloda , abonenin firma tarafındaki özellikleri vardır. Yani TARIFE_PAKETI, MUDURLUK, TESISYASI gibi.

1. `select HIZMET_NO, BIRIM_ADI,ACIKLAMA,EV_IS,TESISYASI from odsturk2.dbo.ABONE A INNER JOIN TMS.BIRIM B ON B.ID=A.MUDURLUK_ID INNER JOIN TMS.TARIFE_PAKET P ON P.ID=A.TARIFE_PAKET_ID`

dbo.TAHAKKUK200908: Bu tabloda müşterinin faturaya yansıyan kalemlerin ne kadar olduğu görünmektedir. Bu tablodan da, NET_BORÇ tutarının meydana getiren OTO_SI, OTO-SA,OTO_MA gibi kalemler alınabilir.

TTS.FATURA: Bu tablodan da müşterinin ödeme yapıp yapmadığını, ödeme yaptıysa SON_ODEME_TARIHI' nden önce mi sonra mı, bu müşterinin aylık ortalama ORTALAMA_FATURA_TUTARI değerlerini alabiliriz.

1. `CASE WHEN TTS.FATURA.SON_ODEME_TARIHI>= TTS.FATURA.ODEME_TARIHI THEN '+' when TTS.FATURA.ODEME_TARIHI is null then '-' when TTS.FATURA.ODEME_TARIHI> TTS.FATURA.SON_ODEME_TARIHI then '0' END AS ODEMEDURUMU`

2.
```
CASE WHEN (F.MUSTERIORT/E.DONEMORT)>=1 AND
(F.MUSTERIORT/E.DONEMORT)<=2 THEN 'MEDIUM'
WHEN (F.MUSTERIORT/E.DONEMORT)>2 AND (F.MUSTERIORT/E.DONEMORT)<5 THEN
'HIGH'
 WHEN (F.MUSTERIORT/E.DONEMORT)>=5 AND (F.MUSTERIORT/E.DONEMORT)<10
THEN 'BIGGERHIGH' WHEN (F.MUSTERIORT/E.DONEMORT)<1 THEN 'SMALL' WHEN
(F.MUSTERIORT/E.DONEMORT)>=10 THEN 'ULTRA' END AS MUSTERIDURUMU
```

dbo.XDSLABONE: Bu tablodan da bu müşteriye ait XDSL olup olmadığını anlaşılabilir.

1. `CASE WHEN dbo.XDSLABONE.CIHAZ_NO IS NULL THEN 0 ELSE 1 END AS ADSL`

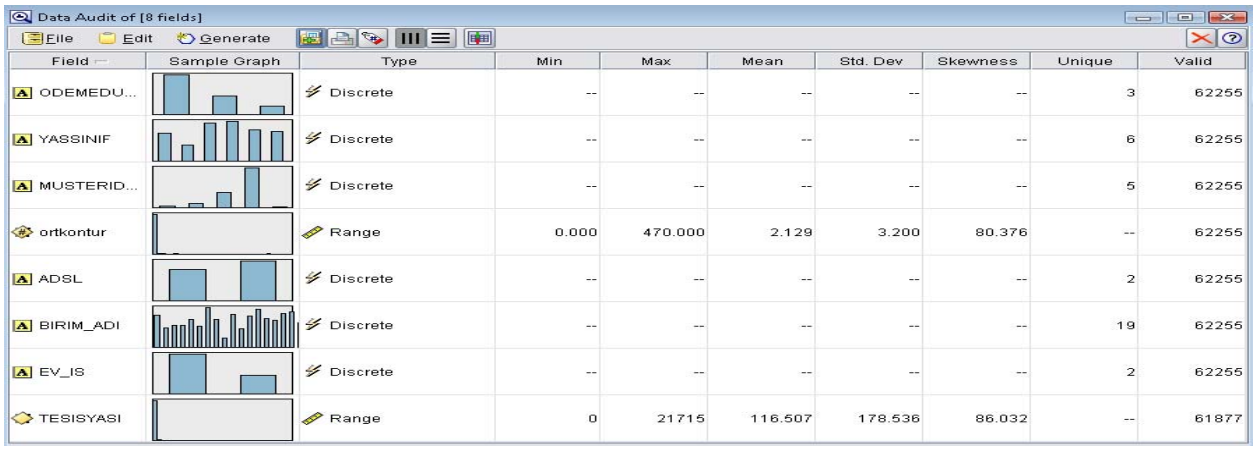
Faydası olabilecek bütün tablolara gözatıp sorun çıkarabilecek detayları düzeltmeye çalışarak modelleme kısmına başlanabilir. Bu aşamada dikkat edilmesi gereken verileri anlamlı şeyler ifade edecek hale getirmek, programın çalıştırabileceği bir vaziyete getirmektir.

5.2.4. MODELLEME

Veri madenciliğinin 4.aşaması, modelleme yapmaktır. Bu kısımda pekçok modelleme deneyerek en doğrusunu bulmak gerekir.

Modellemeleri yaparken KOHONEN tekniği kullanılacaktır. Fakat ilk aşamalarda da deneme maksatlı K-Means ve diğer yöntemlerle çeşitli örnekler yapılıp incelenip Kohonene karar verilmiştir. Pekçok kere çekilen data değiştirildi. Kullanılan özellikler değiştirildi.

1.UYGULAMA: Öncelikle sadece ABONE tablosundan genellerle küçük bir segmentasyon çalışması yapıldı.


Data Audit of [8 fields]

File Edit Generate

Field	Sample Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
ODEMEDU...		Discrete	--	--	--	--	--	3	62255
YASSINIF		Discrete	--	--	--	--	--	6	62255
MUSTERID...		Discrete	--	--	--	--	--	5	62255
ortkontur		Range	0.000	470.000	2.129	3.200	80.376	--	62255
ADSL		Discrete	--	--	--	--	--	2	62255
BIRIM_ADI		Discrete	--	--	--	--	--	19	62255
EV_IS		Discrete	--	--	--	--	--	2	62255
TESISYASI		Range	0	21715	116.507	178.536	86.032	--	61877

Şekil 5.4. Abone kümeleme

Modellemeye giren kolonlar:

ODEME_DURUMU: Müşterinin ödeme alışkanlığı

Son Ödeme Tarihi'ne kadar ödeme yapanlar '+'

Son Ödeme Tarihi'nden sonra ödeyenler '0'

Modelleme yapılan tarihe kadar ödeme yapmayanlar '-'

ADSL: ADSL hizmeti alıp almadığı

İnternet Hizmeti alanlar 'VAR'

İnternet hizmeti almayanlar 'YOK'

TESİS_YAŞI: Kaç aylık abone olduğu bilgisi

Ay cinsinden nümerik değer

BIRIM_ADI: Müdürlük adı

İstanbul içindeki müdürlükler

MUSTERI_DURUMU: Müsterinin fatura tutarına göre durumu

Toplam tahakkuk ortalamasının 2 katına kadarsa MEDIUM

Toplam tahakkuk ortalamasının 5 katına kadarsa HIGH

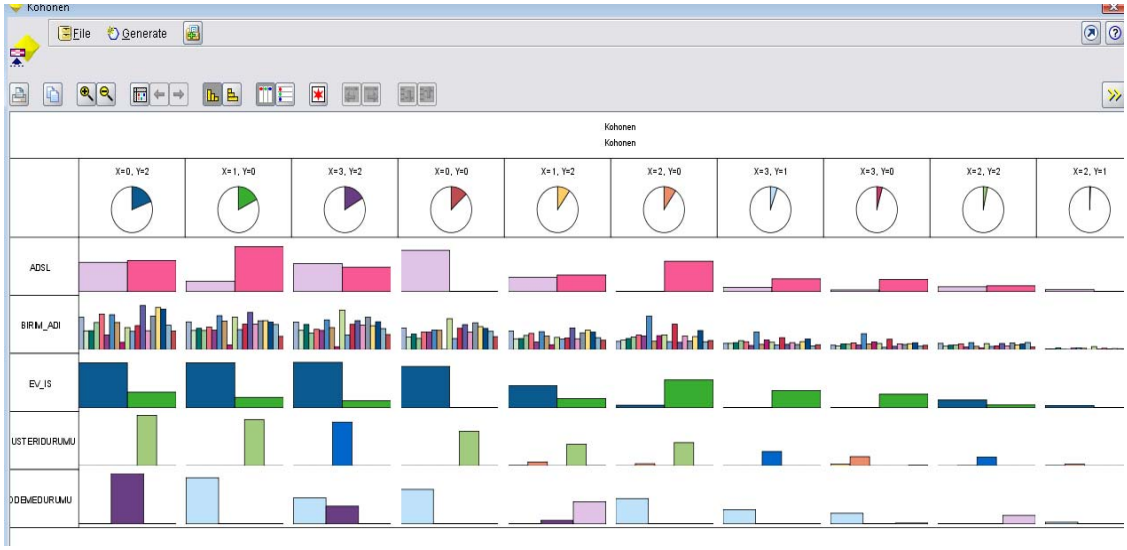
Toplam tahakkuk ortalamasının 10 katına kadarsa BIGGERHIGH

Toplam tahakkuk ortalamasından küçükse SMALL

Toplam tahakkuk ortalamasının 10 katından büyükse ULTRA

EV-IS: Ev veya İş telefonu olduğu bilgisi

Ev ise 'E', İş ise 'I'



Şekil 5.5. Abone segmentasyonu

Modelleme çalıştıktan sonra aşağıda özeti görülen kümeler ortaya çıkmıştır. Bu kümelerin bazıları, şirket için önemli, bazı kümeler ise önemsizdir. Ayrıca her kümede , bütün birimlerden örnek olduğu için kümelemeler için BIRIM_ADI kolononun bir önemi olmadığı farkedilmiştir. Bu kolondan o kümenin en çok hangi bölgede olduğu şeklinde faydalanılabilir. Aşağıdaki çizelgede birim bilgileri yoktur.

küme X-Y	KAYIT SAYISI	ODEME_DURUMU	BIRIM_ADI	ADSL	EV_IS	MUSTERI_DURUMU	ORT. GÜV.
0-0	8173	ÖDEME YAPMIŞ	DEĞİŞKEN	VAR	EV	%99,93 SMALL,%0,07 5 BIGGERHIGH	0,998
0-2	11956	ÖDEME YAPMAMIŞ	DEĞİŞKEN	%51 VAR, %49 YOK	%74 EV %26 İŞ	SMALL	1
1-0	10955	DÜZGÜN ÖDENMİŞ	DEĞİŞKEN	% 81 YOK %19 VAR	%81 EV %19 İŞ	SMALL	1
1-2	6130	%86 GEÇ ÖDENMİŞ,%14 ÖDENMEMİŞ	DEĞİŞKEN	%53 YOK %47 VAR	%70 EV, %30 İŞ	%83 SMALL,%17 BIGGER HIGH	0,82
2-0	6006	DÜZGÜN ÖDENMİŞ	DEĞİŞKEN	YOK	%91 İŞ,%9 EV	%91 SMALL, %9 HIGH	1
2-1	461	%92 düzenli ödenmiş,%8 geç ödenmiş	DEĞİŞKEN	%85 VAR, %15 YOK	%92 EV %8 İŞ	%83 HIGH, %13 BIGGERHIGH,%3 ULTRA	0,95
2-2	2108	%96 GEÇ ÖDENMİŞ,%4 ÖDENMEMİŞ	DEĞİŞKEN	%46 VAR, %54 YOK	%87 EV, %13 İŞ	% 97 MEDIUM %2,5 HIGH	1
3-0	2723	% 92,88 ÖDENMİŞ,%7,12 GEÇ ÖDENMİŞ	DEĞİŞKEN	var 13% , yok 87%	%99,38 EV, %0,62 İŞ	%80 HIGH, % 15 BIGGER HIGH, %5 ULTRA	0,84
3-1	3384	DÜZGÜN ÖDENMİŞ	DEĞİŞKEN	%74 VAR,%26YOK	İŞ	MEDIUM	0,97
3-2	10359	%60 ÖDENMİŞ	DEĞİŞKEN	%53 VAR, %47 YOK	%87 EV, %13 İŞ	MEDIUM	0,99

Şekil 5.6. Abone Segmentasyonu Kümeleme Sonuç çizelgesi

Oluşan kümeler, yukarıdaki çizelgeyle genel olarak ifade edilebilir. Buradan da anlaşıldığı gibi bazı kümeler kesin sonuçlar vermemiştir. Aynı kolona bikaç türden müşteri girildiği görülmüştür. Oluşan bu kümelerden en önemlileri, aşağıda açıklanmıştır. Bu kümelerin ayrıntılarını EK-1' de bulabilirsiniz.

Küme 0-0: % 99,98 güven seviyesinde, ortak özelliği EV telefonu ve ÖDEME YAPMIŞ ve %99' u MUSTERIDURUMU'nu SMALL diye tabir edilmiş ortalama fatura değeri, şirketin aylık telefon geliri ortalamasının altında olan kalanı BIGGERHIGH denilen ortalaması 10 katına kadar olan ve tamamı ADSL hizmeti alan 8173 müşteriden oluşmaktadır. Bu toplam kayıttaki SMALL müşterilerin %20 sini oluşturmaktadır. BIRIM_ADI kolonu kümelerde belirleyici olmamakla beraber bu kümede en fazla sayıda müşterinin EMİNÖNÜ'nde olduğu görülmüştür.

Küme 0-2: 1 güven seviyesinde, ortak özellikleri ödeme yapmamış olan tamamı SMALL diye tabir edilmiş -ortalama fatura değeri, şirketin aylık telefon geliri ortalamasının altında olan-, %74 ü Ev telefonu olan 11956 müşteriden oluşmuştur. Bu kümedeki belirleyizi özellikler SMALL olup ödeme yapmamasıdır.

Küme 1-0: 1 güven seviyesinde, tamamı düzgün ödenmiş, tamamı SMALL diye tabir ettiğimiz ortalama fatura değeri, şirketin aylık telefon geliri ortalamasının altında olan, % 81' i ADSL hizmeti almayan veya ADSL hizmeti alıp İŞ telefonu olan DÜZGÜN ÖDEME yapan 10955 kayıttan oluşmuştur. Bu sayı SMALL müşterilerin %25 idir.

Küme 2-0:1 güven seviyesinde tamamı ADSL hizmeti almayan, % 9' u MUSTERI_DURUMU; HIGH (Donem ortalamasının 2 ile 5 katı arası) olan kalanı SMALL olan tamamı DÜZGÜN ÖDEMİŞ 6006 kayıttan oluşmuştur.
Küme 3-1: % 97 güven derecesinde tamamı DÜZGÜN ÖDENMİŞ, MUSTERI_DURUMU MEDIUM diye tabir edilmiş ortalama faturası , donem ortalamasının 2 katına kadar olan tamamı İŞ telefonu olan 3384 kayıttan oluşmuştur.
Küme 3-0: % 84 güven seviyesinde, HIGH –ortalaması 5 katı ve fazlasından büyük olan- toplam 2723 kayıttan oluşmuştur.

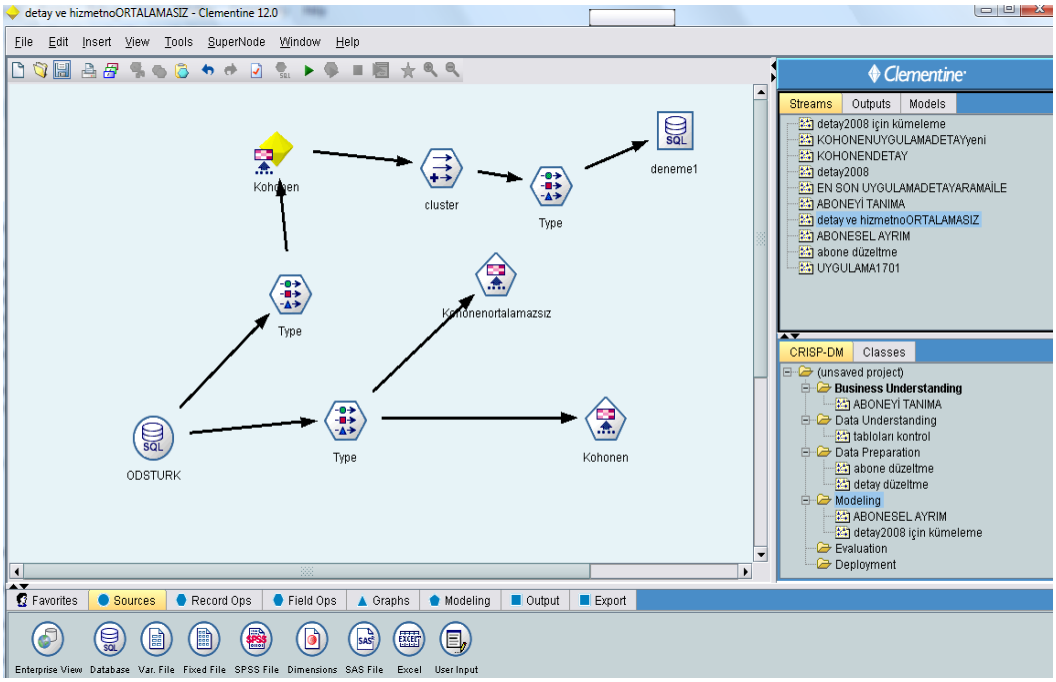
Bu küme sonuçlarına daha sonra kural tanımlama algoritmalarıyla kural tanımlanmış, küme ve güven değerleriyle bir tabloya atılmıştır. Bu çalışma ile AboneSegmentID=Küme olmuştur. Bu kurallar EK-2' de bulanabilir.

2.UYGULAMA: 2. Kısımda sadece detay tablosundan müşterilerin kullanım alışkanlıklarını ayıran bir çalışma yapılmıştır.

Field	Type	Values	Missing	Check	Direction
Asaatliarama	Range	[0,10]		None	In
Ayonluarama	Range	[0,1]		None	In
Bsaatliarama	Range	[0,15]		None	In
Byonluarama	Range	[0,0]		None	In
Csaatliarama	Range	[0,7]		None	In
Cyonluarama	Range	[0,0]		None	In
Dsaatliarama	Range	[0,4]		None	In
Dyonluarama	Range	[0,0]		None	In
Esaatliarama	Range	[0,1]		None	In
Eyonluarama	Range	[0,0]		None	In
Fsaatliarama	Range	[0,2]		None	In
Fyonluarama	Range	[0,0]		None	In
HIZMET_NO	Typeless			None	None
ortkontur	Range	[0.089552...		None	In
toplamarama	Range	<Curre...		None	In

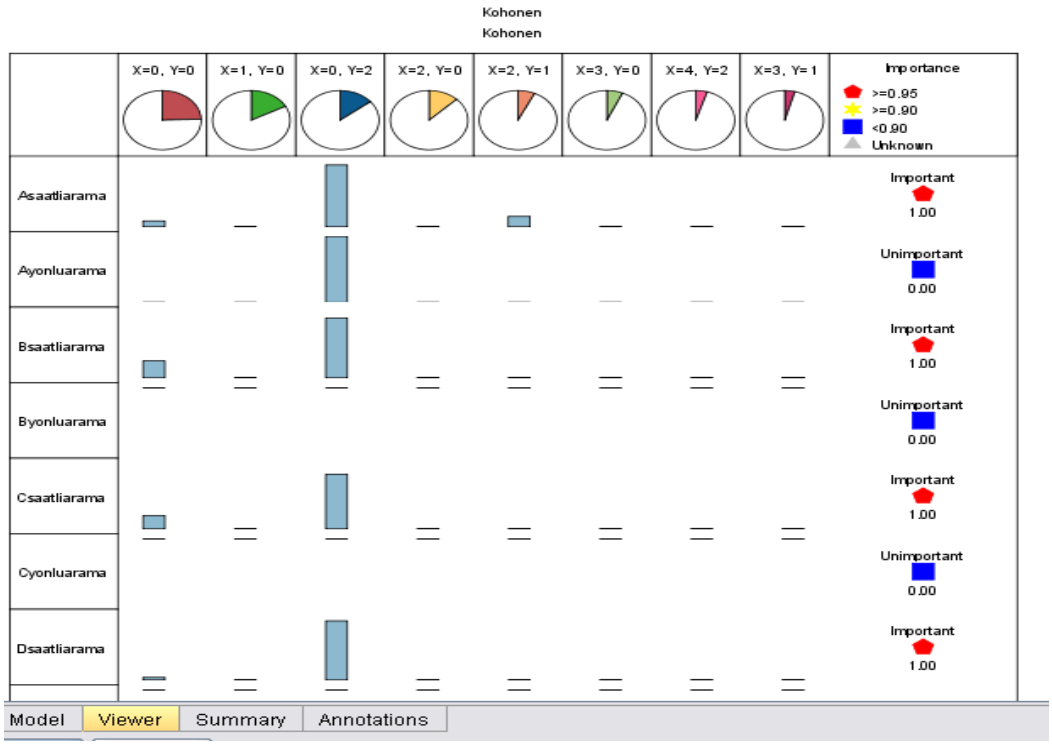
Şekil 5.7. Detay kümeleme için alanlar

Buradaki kolonlar, her müşterinin yaptığı toplam arama, ortalama kontür ve bunların arama saati ve yön kodlarına göre aylık toplam görüşme adedine göre dağılımıdır.



Şekil 5.8. Detay için segmentasyon ekranı

Bu çalışmanın sonucunda çıkan kümeleme ekranı Şekil 5.7' de görülmektedir.



Şekil 5.9. Detay kümeleme sonuç ekranı

Şekilden de anlaşıldığı burada Cyonluarama gibi bazı kolonların segmentasyon çalışması için önemsiz olduğu görülmüştür.

Çıkan kümeleme bilgileriyle her HIZMET_NO bir kümeye dahil oldu. Bu kümeleme çalışması için çıkarılan kurallara göre, müşterinin hangi yöne aradığı bu segmentasyon çalışması için pek önemli değildir. Burdaki dallanma, arama saati ve toplam arama adetine göre sınıflandırılmıştır. KOHONEN algoritması sonucunu EK-3'de bulabilirsiniz. Bu çalışmaya bir kural seti çıkarıldığında daha düzenli bir sonuca ulaşılmıştır. Bu çalışmanın kuralları EK-4 de bulanabilir.

Örnek verirsek ToplamArama<7,5 gibi. Diğer kırılımlar kendi aralarında oluştu. Burada AramaYonu A >0,5 olanların bir kümeye, küçük veya eşit olanların ise birkaç kümeye dağıldığı gözlendi. Burada 0,5 ifadesi, abonenin yaptığı toplam görüşme adedinin yarısı anlamına gelmektedir. Diğer kümeler de toplam görüşme adedi 39 a kadar olanlar veya 53 e kadar olanlar diye değişmektedir.

küme X-Y	KAYIT SAYISI	Ayonlı	Asaati	Byonu	Bsaati	Csaati	Cyönü	Dyönü	Dsaat	Eyönü	Esaati	Fsaati	Fyonu	ortkontu	toplamarama	Ort.Guv.
0-0	12842	0	0,005	0	0,036	0,008	0	0	0	0	0	0	0	3,095	17,945	0,75
0-2	7253	1	0,056	0	0,127	0,033	0	0	0,01	0	0,002	0,001	0	2,429	43,87	0,99
1-0	8773	0	0	0	0	0	0	0	0	0	0	0	0	2,423	41,549	0,9
2-0	6374	0	0	0	0	0	0	0	0	0	0	0	0	2,448	65,571	0,8
2-1	3589	0	0,009	0	0	0	0	0	0	0	0	0	0	2,522	87,094	0,74
2-2	1416	0	0	0	0	0	0	0	0	0	0	0	0	2,686	136,068	0,25
3-0	3370	0	0	0	0	0	0	0	0	0	0	0	0	2,227	111,03	0,6
3-1	2146	0	0	0	0	0	0	0	0	0	0	0	0	2,084	204,199	0,25
3-2	946	0	0,001	0	0	0	0	0	0	0	0	0	0	2,082	267,602	0,25
4-0	1468	0	0	0	0,001	0	0	0	0	0	0	0	0	1,945	158,511	0,26
4-1	753	0	0	0	0,001	0	0	0	0	0	0	0	0	1,899	323,87	0,26
4-2	2409	0	0	0	0	0	0	0	0	0	0	0	0	1,918	719,617	0,26

Şekil 5.10. Detay kümeleri sonuç çizelgesi

Bu çalışma kümeleme açısından sonuçlara baktığımızda da çok ayırtedici özellik kullanmamıştır. Aşağıda ayrıntılı açıklamaları EK-3 'de verilmiş olan bu kümelerin en büyük ve güvenilir olanlarından bikaç tanesi anlatılmıştır.

Küme 0-0: % 75 güven değeriyle, ORTKONTUR değerleri 3,095, TOPLAMARAMA sayısı 17,45 (standart sapma 9,7) olan, saat 06-12 arası yaptığı aramalar, toplamaramanın % 0,5' i olan, saat 12-18 arası % 0,36 olan,18-22 arası %0,8 olan toplam 12842 kayıttan oluşmuştur. Bu toplam detay kayıtların yaklaşık %20 sidir.

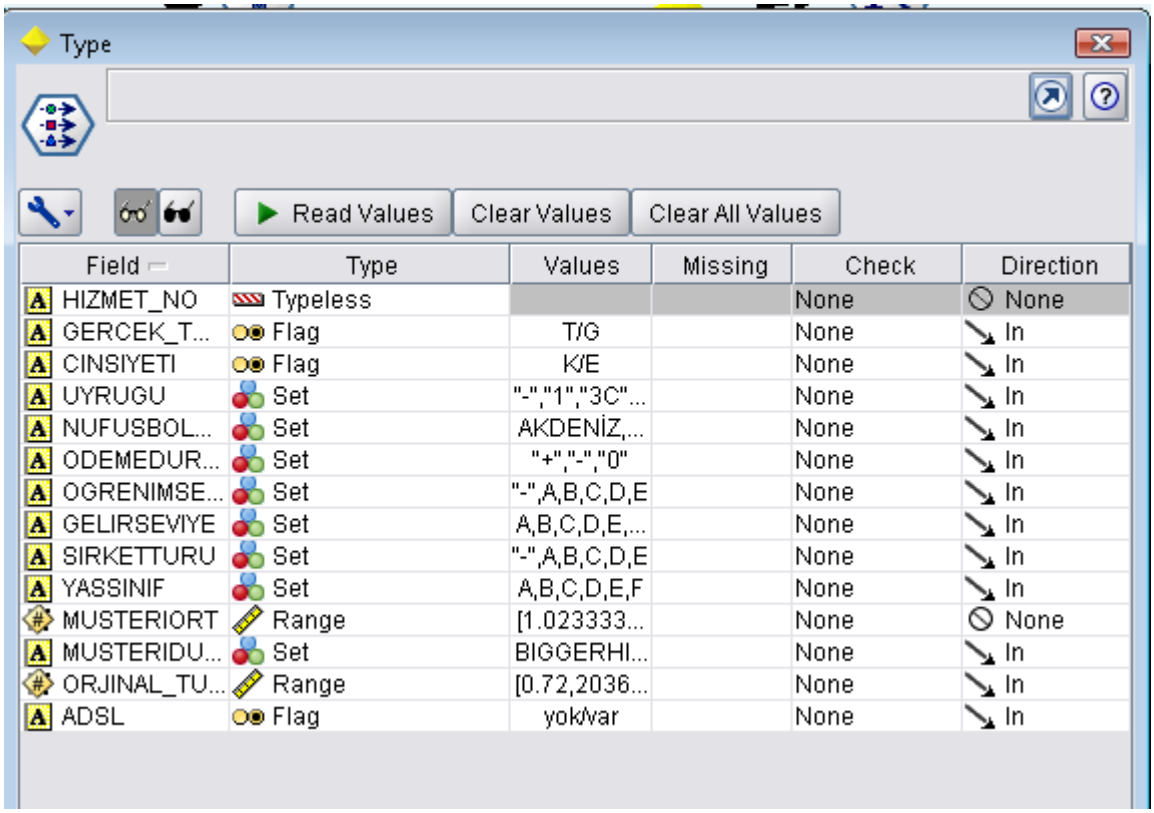
Küme 0-2: % 99 güven değeriyle 06-12 saatleri arasında % 0,56 arama yapan,12-18 arasında % 12, 18-22 arasında % 0,33, 22-24 arasında %0,1, 00-02 arasında %0,1, 02-06 arasında % 0,1 arama yapan, 2,4 ORTKONTUR kullanıp, 43 TOPLAMARAMA yapan 7253 kayıt yerleşmiştir.

Küme 1-0: % 90 güven değeriyle 2,423 ORTKONTUR kullanıp 41 TOPLAMARAMA yapan 8773 kayıt yeralmıştır.

Küme 2-1: % 74 güven değeriyle ORTKONTUR 2,542 olan 87,094 TOPLAMARAMA yapan 3589 kayıttan oluşmuştur.

Bu çalışmanın sonuçlarını EK-3'de , bunun için CART algoritmasıyla tanımlanmış kuralları ise EK-4' de bulanabilir.

3.UYGULAMA: En son olarak UstSegmentID değerini verecek çalışma yapılmıştır. Müşterinin detay özelliklerini kullanmadan müşteri ve tahakkuk sistemlerinden gelen özellikleriyle uygulanan segmentasyondur. Bu çalışma için de aşağıdaki kolonları kullanılmıştır.



Şekil 5.11. Üst segmentasyon için kullanılan alanlar

Çeşitli sorgularla denenen segmentasyon çalışmalarında çok anlamlı sonuçlar alınmadı. Bunların nedeni ;

1. Bazı telefonların şirket bazılarınsa ev telefonu olması. Bu nedenle şahsa özel bilgilerde doğru sonuçlar gelmedi.
2. Bazı 1ve 0 değerlikli kolonlar için flag kolonlar seçilmesi faydalı olur.

Nihai modellemede aşağıdaki kolonlar vardı.

GERCEK_TÜZEL
CİNSİYETİ
GELİRSEVİYE
ADSL
MUSTERIDURUMU
NUFUSBOLGE
ODEMEDURUMU
OGRENIMSEVIYE
ORJINAL_TUTAR
ŞİRKET_TÜRÜ

Modelleme çalıştırıldıktan sonra çıkan küme sonuçları aşağıdadır. Burada da TÜZEL denilen şirket telefonlarının yoğunlukla 1 kümede olduğu, diğer kümeleri GERÇEK denilen bireysel müşterilerin oluşturduğu görülmüştür. GERÇEK müşterilerin kümelerinde ŞİRKET_TÜRÜ kolonu boştur. Bu, kümelemenin başarılı olduğuna dair bir örnektir. Aşağıda sonuçların ayrıntısını EK-6' da bulunmaktadır.

küme X-Y	KAYIT S	orjınaltut	adsl	cinsiyeti	gerçektüz	gelirseviy	müsteridurun	nüfusbolge	ödemedurumu	öğrenin	şirket	yassınıf	Ort.Guv.
0-0	12029	21,199	%50 var	%70 E,%3	G	%90 bilinr	ortalamadan k	%39 karaden	%72 ödenmemiş,%28 geç ödeme	%46 E	boş	%29 C	0,93
0-1	651	100	%52 yok	%77 E	G	%93 M	%91 Hıgh ve ü	karadeniz%3	%76 ödenmemiş gerisi geç ödenmiş	%43 E	boş	%28 C	0,9
0-2	7885	41	%66 var	%66 E	G	% 88 M	dönem ortala	%38 marmar	%37 ödenmiş	%36 A	boş	31%	0,8
1-0	3775	21	%50 var	%61 E	%99,97 G	%92 M	dönem ortala	%99 marmar	%65 ödenmemiş,%35 geçödeme	%48 E	boş	%23 C	0,92
1-1	336	87	%54 yok	%50 E	G	% 91 M	% 94 5 katı ve	%97 marmar	%40 ödenmiş	%33 E	boş	% 26 E,F	0,87
1-2	5681	43	%67 yok	%71 E	G	%97 M	%94 M	%31 marmar	% 86 ödenmiş	%71 E	boş	%24 C	0,7
2-0	7333	22	% 58 yol	%64 E	G	%93 M	%98 small	marmara	ödenmiş	%41 E	boş	%33 F	0,93
2-1	595	83	%70 yok	%75 E	G	%99 M	% 72 Hıgh ve ü	%51 marmar	%88 ödenmiş	%85 E	boş	%30 E	0,82
2-2	590	61	%88 yok	%83 E	%98 G	%99 M	%56 Medıum v	%65 bilinme	%87 +,	%99 E	boş	%30 F	0,85
3-0	11584	22	%54 var	%70 E,%3	G	%91 M	%98 small	%43 karaden	ödenmiş	%38 E	boş	%28 D	0,91
3-1	1858	21	%91 yok	%88 E	%99 G	%99 M	%98 small	534 bilinmey	ödenmiş	%98 E	boş	%31 F	0,91
3-2	9938	54	%81 yok	%99 E	%98 T	%99 M	%53 small	%99 bilinme	%78 ödenmiş	%99 E	değiş	a	0,93

Şekil 5.12. Üst segmentasyon kümeleme sonuç çizelgesi

Yukarıdaki kümelerden bazıları şöyle tanımlanabilir. Çizelgeye baktığımızda GELIRSEVIYE kolonunun her kümede aynı oranda yeraldığı görülmüştür. Burdan bu kolona anlam cıkarmanın doğru olmayacağı çıkarılabilir.

Küme 0-0: %93 güven değeriyle tamamıyle SMALL denilen ortalamadan küçük ve GERÇEK(bireysel) müşterilerden oluşan faturalarını ödemeyen veya geç ödeyen 12029 kayıttan oluşmuştur.

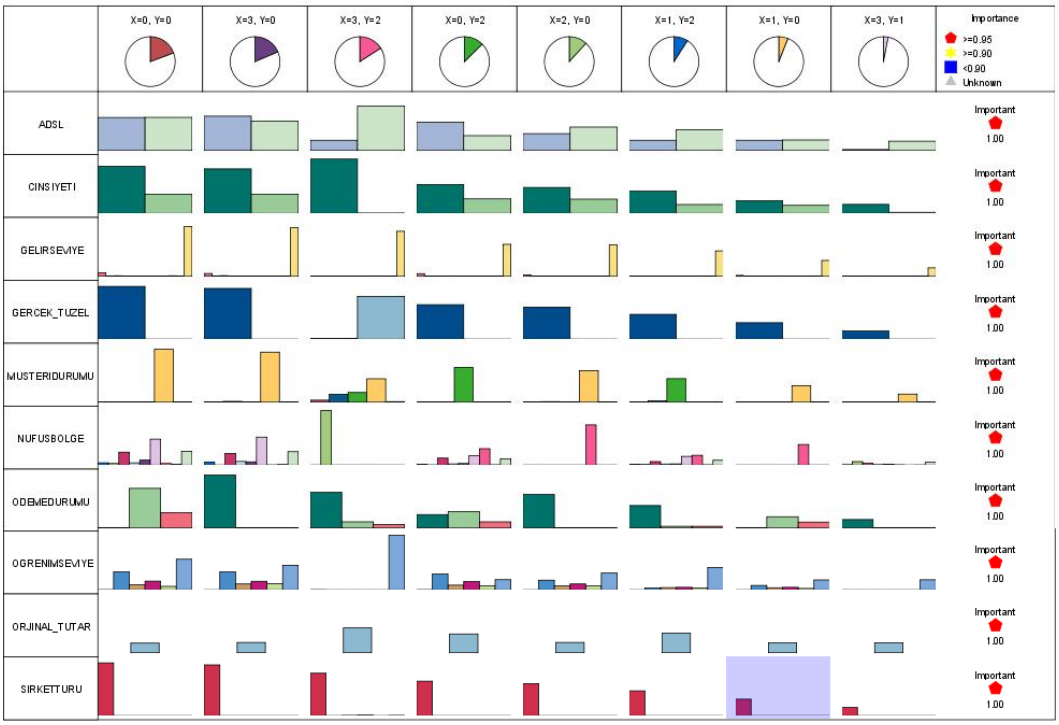
Küme 0-1: % 90 güven değeriyle MUSTERIDURUMU ortalamanın 5 katı ve daha üstü olup faturasını geç ödeyen veya ödemeyen GERÇEK 651 kayıttan oluşmaktadır. Azmiktardaki yüksek getirili müşterilerin fatura ödemede sorun yaşayan müşteriler bu kümede toplanmıştır.

Küme 2-0: %92 güven değeriyle tamamı tamamı GERÇEK müşteri olan, MARMARA bölgesi kayıtlı, DÜZGÜN ÖDEME yapmış ,% 98' i ortalamadan küçük müşterilerden oluşan 7333 kayıttan oluşmuştur.

Küme 3-0: %91 güven değeriyle tamamiyle DÜZGÜN ÖDEME yapmış % 98'i ortalamadan küçük olan, tamamiyle GERÇEK olan 11584 kayıttan oluşmuştur. Bu grupta KARADENIZ bölgesinin yoğunluğu görülmüştür.

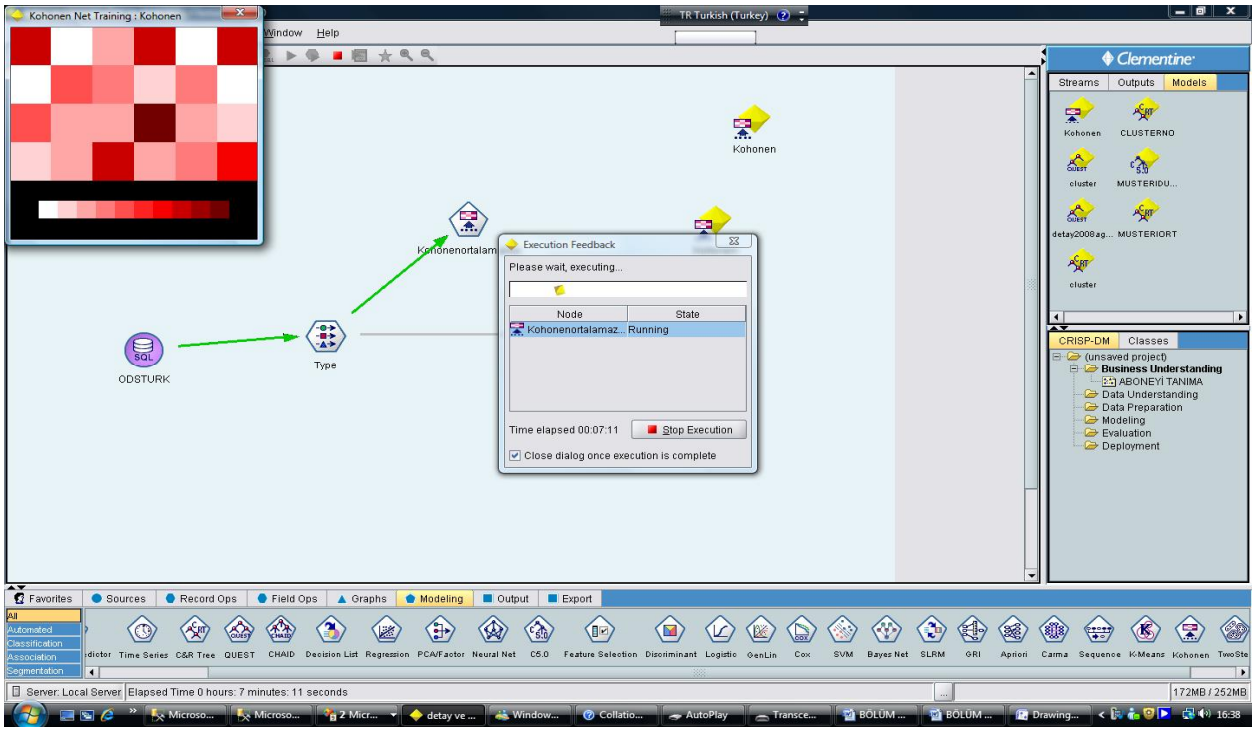
Küme 3-2: %93 güven değeriyle % 98 TÜZEL (şirket) telefonu olup her müşteri durumunda 9938 kayıttan oluşmuştur. Bu tüzel müşterilerin %92' si bu kümede toplanmıştır. Geri kalanı ödeme yapılmamış bir kümededir.

Bu kümeleme sonucunda da GELIRDURUMU, OĞRENIMSEVIYE GİBİ bazı kolonların bekleneni vermediği gözlenmiştir. Bu kolonlar kümelerdeki ağırlıkları tespit etmek için kullanılabilir. Bu sonuçlar, firmaların operasyonel sistemleri için girişler açısından büyük fayda sağlayacaktır.



Şekil 5.13. Üst segmentasyon sonucu kümeler

Bu çalışmanın ardından CART algoritmasıyla kural tanımlanmıştır. Bu kurallar , EK-6' da bulunabilir.



Şekil 5.14. Kohonen ile segmentasyon ekranı

Bu son çalışmayla da modelleme aşaması bitirilmiştir. Burada biri abone özellikleri, biri genel olarak müşteriyi, biri de detayları barındıran 3 segmentasyon çalışması yapılmıştır. Bu özelliklerin aynı anda yapılmasının nedeni, herhangi birine yönelik bir çalışma yapılmak istendiğinde karmaşıklık olmasını engellemektir. Bütün kümeler, belli olduktan sonra EBM.lkMüsteriKümesi tablosuna bu kümeleri girerek özel bir ID elde edilmiştir. EBM.lkAbone tablosuna bu segmentasyon değerlerinden herhangi birini veya 3 durumdaki değeri veren MüşteriKümesiId değeri girilebilir.

5.2.5.DEĞERLENDİRME

Değerlendirme aşamasında uygulanan modeller test edilecektir. Bir önceki bölümde bahsedildiği gibi modeli test etmek için çeşitli yöntemler mevcuttur. Bunların en sık kullanılanı eldeki datanın bir kısmıyla modeli aynı şekilde kurup, bir diğer kısmıyla da bu modeli uygulamaktır.

1.Abone cluster: Bunun için datanın %50 si tekrar model yapmak için kullanılmıştır. Diğer kalan yarısına da çıkan model uygulanmıştır. Sonuçları bir tabloya atıp modelleme ile karşılaştırılıp örnek numaralar alarak cluster özellikleri kontrol edilmiştir. Çoğunluk içeren gruplarda başarının daha fazla olduğu görülmüştür.

Örneğin ilk segmentasyon çalışmasında 0-2 clusterına giren bir abone değerlendirme safhasında da nerdeyse benzer özelliklerde bir kümeye girmiştir.

ID=82000007461108

Şekil 5.15. Abone segmentasyonu değerlendirme

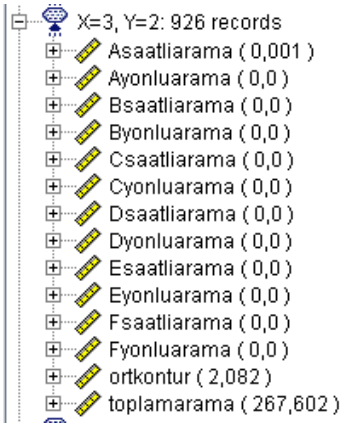
Ayrıca daha önce oluşturulmuş model, datanın %10 luk bir kısmına değiştirmeden uyguladığında hata payınının %23, dogruluk payının ise 1-0,23=0,77 olduğu görülmüştür.

2.Detay cluster: PSTN müşterilerinin detay görüşmelerinden yola çıkarak yapılan bu modelleme için de aynı yöntem uygulanmıştır. Bu yöntemle yaklaşık %60 lık doğru sonuç elde edilmiştir.

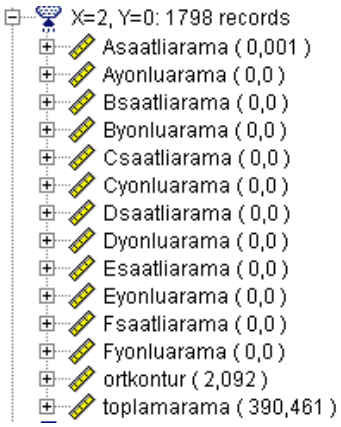
SELECT A.*,xdetaydegerlendirme,ydetaydegerlendirme FROM newods.dbo.clusterdetaylar a inner join dbo.detaydegerlendırme2 b on a.HIZMET_NO=b.HIZMET_NO

Örnek bir numara alıp deniyoruz.

İlk segmentasyon

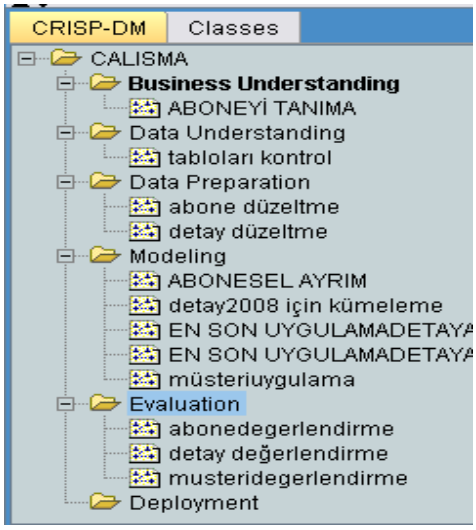


Değerlendirme segmentasyonu



Şekil 5.16. Detay segmentasyonu değerlendirme

3. Müşteri cluster: Müşterinin daha çok kişisel ve standart özelliklerini kullanarak yapılan bu modelleme için de % 50 datayı modelleme, %50 sini de test için kullanılmıştır. Bu modellemenin diğerlerine oranla daha başarılı olduğu görülmüştür. Modeli dataya uyguladığımızda, müşterilerin en düşük % 80 civarında bir doğrulukla kümelendiği görülmüştür.



Şekil 5.13. Segmentasyon projesine genel bakış

5.2.6. SAHAYA SÜRÜŞ

Bu proje ile segmentasyon yapılmaya çalışıldı. 3 çeşitli şekilde yapılan bu çalışmada amaç, müşteriyi birkaç özelliğiyle beraber tanımlayabilmekti. Bu müşteri kümeleri düzenlendikten sonra bahsedilen EBM ortamına koyulabilir. Son kullanıcıya yönelik OLAP küpleri, portal gibi ortamlarla hızla sadece istenen kümeye yönelik analizler yapılabilir.

Elde edilen bu küme numaralarını EBM veritabanındaki tbAbone tablosuna ekleyerek bu KümeID'lerle abone veya müşteri bir ID ile tanımlanabilir. Hatta çok başarılı veri madenciliği çalışmalarında bu ID ler birkaç kolon yerini tutabilir. İşi anlama kısmında bahsettiğimiz AboneSegmentID, DavranışSegmentID, UstSegmentID değerlerimiz bu çalışmayla oluşturulabilir.

Uyglama sonuçlarından anlaşıldığı gibi aslında yüksek getirisi olup faturasını zamanında ödemeyen müşteriler kimlerdir, yüksek öğrenim görmüş ama evinde internet kullanmayan müşteriler kimlerdir bunlar tespit edilebilir. Bütün bunları tek bir KümeID değeriyle bulanabilir. Elde edilen bu değerlerle, CRM uygulamaları için kampanyalar düzenlenebilir. Örneğin düzenli ödeme yapılmayan bölgelerdeki müşteriler aranarak onların düzenli ödeme yapabilmeleri için öneriler sunulabilir.

SONUÇ

Bu tez çalışmasında, veri tabanı ve veri ambarı kavramlarından yola çıkarak veri madenciliği kavramı, yöntemleri, süreçleri, amaçları, araçları, bu kavramlarla beraber gelen CRM ve Müşteri Segmentasyonu konuları da incelenip örnek bir müşteri segmentasyonu yapmaya çalışılmıştır.

Veri madenciliği projesi için önce bir veri ambarına ihtiyaç vardı. Veri ambarı ortamı, 2 aşamadan oluşmaktadır. Bu ortam için öncelikle operasyonel sistemdeki verilerin ham olarak atıldığı ODS veritabanını oluşturulmuştur. Diğer veritabanı da operasyonel sistemlerden alınan dataları yorumlayarak daha az karmaşık, daha düzenli, temiz dataların oluşturduğu EBM dir. EBM ortamında tbAbone tablosu en önemli tablodur. Amaç, bu bulunan sonuçları AboneSegmentID, DavranışSegmentID, UstSegmentID adıyla EBM ortamına oturtmaktır.

İlk olarak AboneSegmentID değeri için aşağıdaki özellikleri kullanarak abone segmentasyonu yapılmıştır.

ODEME_DURUMU
ADSL
BIRIM_ADI
MUSTERI_DURUMU
EV-IS

Yapılan kümeleme çalışması sonucunda aşağıdaki küme kuralları oluşmuştur.
Küme 0-0: % 99,98 güven seviyesinde, ortak özelliği EV telefonu ve ÖDEME YAPMIŞ ve %99' u MUSTERIDURUMU'nu SMALL diye tabir edilmiş ortalama fatura değeri, şirketin aylık telefon geliri ortalamasının altında olan kalanı BIGGERHIGH denilen ortalaması 10 katına kadar olan ve tamamı ADSL hizmeti alan 8173 müşteriden oluşmaktadır. Bu toplam kayıttaki SMALL müşterilerin %20 sini oluşturmaktadır. BIRIM_ADI kolonu kümelerde belirleyici olmamakla beraber bu kümede en fazla sayıda müşterinin EMİNÖNÜ'nde olduğu görülmüştür.
Küme 0-2: 1 güven seviyesinde, ortak özellikleri ödeme yapmamış olan tamamı SMALL diye tabir edilmiş -ortalama fatura değeri, şirketin aylık telefon geliri

ortalamasının altında olan-, %74 ü Ev telefonu olan 11956 müşteriden oluşmuştur. Bu kümedeki belirleyizi özellikler SMALL olup ödeme yapmamasıdır.

Küme 1-0: 1 güven seviyesinde, tamamı düzgün ödenmiş, tamamı SMALL diye tabir ettiğimiz ortalama fatura değeri, şirketin aylık telefon geliri ortalamasının altında olan, % 81' i ADSL hizmeti almayan veya ADSL hizmeti alıp İŞ telefonu olan DÜZGÜN ÖDEME yapan 10955 kayıttan oluşmuştur. Bu sayı SMALL müşterilerin %25 idir.

Küme 2-0:1 güven seviyesinde tamamı ADSL hizmeti almayan, % 9' u MUSTERI_DURUMU; HIGH (Donem ortalamasının 2 ile 5 katı arası) olan kalanı SMALL olan tamamı DÜZGÜN ÖDEMİŞ 6006 kayıttan oluşmuştur.

Küme 3-1: % 97 güven derecesinde tamamı DÜZGÜN ÖDENMİŞ, MUSTERI_DURUMU MEDIUM diye tabir edilmiş ortalama faturası , donem ortalamasının 2 katına kadar olan tamamı İŞ telefonu olan 3384 kayıttan oluşmuştur.

Küme 3-0: % 84 güven seviyesinde, HIGH –ortalaması 5 katı ve fazlasından büyük olan toplam 2723 kayıttan oluşmuştur.

Bu kümelerden içinden %100 doğruluklu bazı kurallar da çıkardık (EK-2).

1.İŞ telefonu olup MUSTERIDURUMU dönem ortalamasından düşük olan, düzgün ödeme yapan ve ADSL sahibi müşteriler bir kümede toplanır.

2. İŞ telefonu olup MUSTERIDURUMU, dönem ortalamasının 2 katından daha büyük ve ödeme yapan bütün müşteriler bir kümede toplanır.

3. EV telefonu olup MUSTERIDURUMU,donem ortalamasından küçük olup düzgün ödeme yapan ve ADSL sahibi bütün müşteriler bir kümede toplanır.

4.MÜŞTERİDURUMU, Dönem Ortalamasının 2 ile 5 katı arasında olan müşteriler içinde İŞ telefonu olanların büyük çoğunlukla ADSL hizmeti almayan abonelerden genelde faturalarını düzenli olarak ödediği ve bu çoğunluğun genelde BEYOĞLU civarında olduğu görülmüştür.

5. MUSTERIDURUMU, Dönem Ortalamasının 2 ile 5 katı arasında olan müşterilerin içinde EV telefonu kullanıp düzenli ödeme yapan müşteri yoğunluğunun en çok ERENKÖY en az da İKİTELLİ tarafında olduğu görülmüştür.

İkinci segmentasyonda ise telefonların detay aramaları incelendi. Burda da konuşma alışkanlıklarını kümeleyeceğimiz DavranısSegmentID değeri bulundu. Burda da 300 den fazla konusma yaptığı halde harcanan ortalama kontur sayısı sadece 2 olan

müşterileri veya sadece 20 tane arama yapan müşteriler ayrıştırıldı. Kullanılan kolonlar aşağıdadır:

Toplam Kontör
Ortalama Kontor
Şehir içi arama
Şehir dışı Arama
Milletler arası arama
GSM aramaları
Diğer aramalar

Bu kolonların girdiği segmentasyon çalışmasından çıkan örnekler aşağıdadır. Bu çalışma, datanın %10'u kullanarak tekrar kümeleme yoluyla test edilmiştir. Doğruluk payı %77 çıkmıştır.

Küme 0-0: % 75 güven değeriyle, ORTKONTUR değerleri 3,095, TOPLAMARAMA sayısı 17,45 (standart sapma 9,7) olan, saat 06-12 arası yaptığı aramalar, toplamaramanın % 0,5' i olan, saat 12-18 arası % 0,36 olan,18-22 arası %0,8 olan toplam 12842 kayıttan oluşmuştur. Bu toplam detay kayıtların yaklaşık %20 sidir.

Küme 0-2: % 99 güven değeriyle 06-12 saatleri arasında % 0,56 arama yapan,12-18 arasında % 12, 18-22 arasında % 0,33, 22-24 arasında %0,1, 00-02 arasında %0,1, 02-06 arasında % 0,1 arama yapan, 2,4 ORTKONTUR kullanıp, 43 TOPLAMARAMA yapan 7253 kayıt yerleşmiştir.

Küme 1-0: % 90 güven değeriyle 2,423 ORTKONTUR kullanıp 41 TOPLAMARAMA yapan 8773 kayıt yeralmıştır.

Küme 2-1: % 74 güven değeriyle ORTKONTUR 2,542 olan 87,094 TOPLAMARAMA yapan 3589 kayıttan oluşmuştur.

Yukarıdan da anlaşıldığı gibi detay kümeleri sonucunda pek verimli sonuçlar alınamamıştır. Genelde toplam arama ve ortalama kontur üstünden ayrıştırma yapılmaktadır.

1.En büyük kümenin Şehiriçi, şehirdışı ve Milletlerarası görüşme yapanların oluşturduğu görülmüştür.

2.En küçük küme toplam araması 700 kadar olup, ortalama kontür 2 civarında olan müşterilerden oluşmuştur.

Bu çalışmada, arama saatlerinin kümelemede pek öneminin olmadığı genelde toplam aramalara ve kontür sayısına göre ayrım yapıldığı görülmüştür.

Üçüncü segmentasyonda genel olarak bütün tablolardan aldığımız daha çok sayıda özellikle oluşturulan UstsegmentID değerini oluşturmak içindi. Bu çalışmada ortaya bazı kurallar çıktı. Bu uygulama için, pekçok kere denemeler yapıldı. En doğru sonuçların genel özellikler için alındığı görüldü. ŞİRKET TÜRÜ gibi şirketlere yönelik kolonların kümelemelerde pek etkili olmadığı görülmüştür. Hatta GERÇEK müşteriler için boş gelen değerler aynı kümede bulunuduğundan burada KOHONEN ' in başarısı anlaşılabilir. Bu çalışmada aşağıdaki kolonlar kullanıldı.

GERCEK_TÜZEL
CİNSİYETİ
GELİRSEVİYE
ADSL
MUSTERIDURUMU
NUFUSBOLGE
ODEMEDURUMU
OGRENIMSEVIYE
ORJINAL_TUTAR
ŞİRKET_TÜRÜ

Bu çalışma sonunda çıkan model kümeleri aşağıdadır.

Küme 0-0: %93 güven değeriyle tamamıyle SMALL denilen ortalamadan küçük ve GERÇEK(bireysel) müşterilerden oluşan faturalarını ödemeyen veya geç ödeyen 12029 kayıttan oluşmuştur.

Küme 0-1: % 90 güven değeriyle MUSTERIDURUMU ortalamanın 5 katı ve daha üstü olup faturasını geç ödeyen veya ödemeyen GERÇEK 651 kayıttan oluşmaktadır. Azmiktardaki yüksek getirili müşterilerin fatura ödemede sorun yaşayan müşteriler bu kümede toplanmıştır.

Küme 2-0: %92 güven değeriyle tamamı tamamı GERÇEK müşteri olan, MARMARA bölgesi kayıtlı, DÜZGÜN ÖDEME yapmış ,% 98' i ortalamadan küçük müşterilerden oluşan 7333 kayıttan oluşmuştur.
Küme 3-0: %91 güven değeriyle tamamiyle DÜZGÜN ÖDEME yapmış % 98'i ortalamadan küçük olan, tamamiyle GERÇEK olan 11584 kayıttan oluşmuştur. Bu grupta KARADENIZ bölgesinin yoğunluğu görülmüştür.
Küme 3-2: %93 güven değeriyle % 98 TÜZEL (şirket) telefonu olup her müşteri durumunda 9938 kayıttan oluşmuştur. Bu tüzel müşterilerin %92' si bu kümede toplanmıştır. Geri kalanı ödeme yapılmamış bir kümededir.

Bu kümeleme sonuçları, EK-5' de bulunmaktadır. Bu kümeler için oluşturalan kuralları ise EK-6'da yeralmaktadır.

Detay kümelerinde daha çok toplam arama sayılarının önemli olduğu, hangi yöne veya hangi saatte arama yaptığının fazla önemli olmadığı, müşteri kümelerinde Öğrenim seviyesi üniversite olup ADSL kullanmayan müşterilerin faturalarını da zamanında ödemediği, ortalamanın çok üstünde fatura geldiği halde faturasını ödemeyen müşteriler olduğu anlaşılmıştır.

Üst segmentasyon için kullanılan pekçok alanın küme belirlemede çok etkili olmadığı, kümelerin sayısal olarak başarılı olduğu fakat çok kolon kullanılmadığı görülmüştür. Örneğin ŞİRKET_TURU, GELIRSEVIYESI gibi kolonlar segmentasyonda etkisiz gibi davranmıştır. Yapılmış bu 3 segmentasyon çalışmasından da anlaşıldığı gibi sayısal değer içeren kolonlarda tekrar yapılan modellemelerde sapmalar olabilmektedir. Bu çalışmalarla datanın durumu da anlaşılmıştır. Daha profesyonel bir çalışma için yapılması gerekenler daha kolayca tespit edilebilir.

Bütün yapılan bu çalışmalar, aslında CRM programlarında kullanılmak içindir. Bu çalışmalar sonunda firmaya bazı önerilerde bulunabiliriz. Veriden mantıklı kuramlar çıkarabilmek için datanın temiz, dolu olması çok önemlidir. Bu tip çalışmalar, genelde doluluk oranı yüksek kolonlar için yapılmalıdır. Bu çalışma sonucunda, şu öneriler yapılabilir.

1. MUSTERIDURUMU, donem ortalamasından büyük olan işyeri telefon faturalarının zamanında ödendiği tespit edildiği için bu müşterilere ADSL hizmeti indirimli fiyatlarla verilebilir. Buradan gelecek ADSL faturalarının da düzenli ödenem ihtimali yüksektir. CRM projesiyle bu SegmentID' ye sahip müşteriler aranarak ADSL önerimi yapılabilir.

2. MUSTERIDURUMU, dönem ortalamasından küçük olan EV telefonu sahibi olup ADSL hizmeti alıp, düzgün ödeme yapan müşteriler bir kümededir. Bu müşterilere, ayda belli konuşma karşılığı bir fiyat önerilebilir. Böylece bu müşterilerden gelecek olan gelirin artması beklenir.

3. Toplam arama sayısı çok fazla olduğu halde ortalama kontürü sadece 2 olan müşterilere, kontür daha uygun olarak verilebilir. Bu, müşterinin de avantajına olacak bir durumdur.

4. Kümeleme çalışmasını hazırlarken bazı kolonların işe yaramadığı farkedilmiştir. Örneğin ŞİRKET_TÜRÜ, GELIRSEVIYE, ÖĞRENIMDURUMU VS.. Bunun en önemli sebebi, bu bilgilerin düzgün doldurulmamasıdır. Bu yüzden operasyonel sistemlere kayıt girerken, o bilgilere dikkat edilmesi sağlanabilir.

5. Şirkete kazandırdığı gelir, ortalamanın 2 katından fazla olan müşteri çok az sayıdadır. Bu müşteriler, genelde TÜZEL müşterilerdir. Bu müşterilere özel sabit anlaşmalar düzenlenebilir. Böylece tahakkuk sistemlerinin işi kolaylaşır.

6. Ortalama kontür sayısı çok fazla olmadığı halde arama sayısı 700 olan kümeler görülmüştür. Bu müşterilere, x adet 1 konturlu konuşma bedava şeklinde bir kampanya düzenlenerek müşterinin arama alışkanlığı artırılabilir.

7. Gerçek VE tüzel müşteriler için ayrı ayrı segmentasyon çalışmaları yürütülebilir.

8. Bu sonuçlardan yola çıkarak yeni veri madenciliği kampanyaları gerçekleştirilebilir. Örneğin ADSL kullanan müşteri tipi belirlenip potansiyel ADSL müşterileri belirlenebilir.

Bu tez çalışmasında yapay sinir ağı mantığıyla çalışan KOHONEN kullanılmıştır. KOHONEN, özellikle alfanumerik kolonlarda oldukça başarılıdır. Bu çalışmada, müşteri kümeleri oluşturulmuş ve incelenmiştir. Veri madenciliğinde yaratıcılık bize kalmıştır. Müşterinin geri dönüşlerine göre bu çalışmanın ne kadar fayda getirip getirmeyeceği anlaşılacaktır.

EKLER

EK-1- Abone segmentasyonu kümeleri

X=0, Y=0
8173 Records
  * ADSL
    * var (100%)
    * var 100%
      yok 0%
  * EV_IS
    * E (100%)
    * E 100%
      I 0%
  * BIRIM_ADI
    * ERENKÖY TELEKOM (8,98%)
    * AVCILAR TELEKOM 6,31%
      BAHÇELİEVLER TELEKOM 3,67%
      BAKIRKÖY TELEKOM 4,5%
      BAYRAMPAŞA (ESENLER) TELEKOM 3,29%
      BAĞCILAR TELEKOM 5,15%
      BEBEK TELEKOM 5,13%
      BEYOĞLU TELEKOM 5,64%
      BÜYÜKÇEKMECE TELEKOM 5,62%
      EMİNÖNÜ (SURIÇI) TELEKOM 0,09%
      ERENKÖY TELEKOM 8,98%
      FATİH TELEKOM 3,25%
      GAYRETTEPE TELEKOM 6,19%
      GAZİOSMANPAŞA TELEKOM 7,27%
      KADIKÖY TELEKOM 5,14%
      KÜÇÜKYALI TELEKOM 7,79%
      PENDİK TELEKOM 6,71%
      ÜMRANİYE TELEKOM 6,14%
      ÜSKÜDAR TELEKOM 5,33%
      İKİTELLİ TELEKOM 3,79%
  * MUSTERIDURUMU
    * SMALL (99,93%)
    * BIGGERHIGH 0,07%
      HIGH 0%
      MEDIUM 0%
      SMALL 99,93%
      ULTRA 0%
  * ODEMEDURUMU
    * + (100%)
    * + 100%
      - 0%

0 0%

X=0, Y=2
11956 Records
* ADSL
* yok (51,56%)
* var 48,44%
yok 51,56%
* EV_IS
* E (74,13%)
* E 74,13%
I 25,87%
* BIRIM_ADI
* GAZİOSMANPAŞA TELEKOM (8,77%)
* AVCILAR TELEKOM 6,42%
BAHÇELİEVLER TELEKOM 3,73%
BAKIRKÖY TELEKOM 3,76%
BAYRAMPAŞA (ESENLER) TELEKOM 5,39%
BAĞCILAR TELEKOM 7,06%
BEBEK TELEKOM 2,94%
BEYOĞLU TELEKOM 6,93%
BÜYÜKÇEKMECE TELEKOM 5,39%
EMİNÖNÜ (SURIÇI) TELEKOM 1,46%
ERENKÖY TELEKOM 4,38%
FATİH TELEKOM 3,65%
GAYRETTEPE TELEKOM 4,77%
GAZİOSMANPAŞA TELEKOM 8,77%
KADIKÖY TELEKOM 3,76%
KÜÇÜKYALI TELEKOM 6,66%
PENDİK TELEKOM 8,37%
ÜMRANİYE TELEKOM 8,04%
ÜSKÜDAR TELEKOM 4,84%
İKİTELLİ TELEKOM 3,69%
* MUSTERIDURUMU
* SMALL (100%)
* BIGGERHIGH 0%
HIGH 0%
MEDIUM 0%
SMALL 100%
ULTRA 0%
* ODEMEDURUMU
* - (100%)
* + 0%
- 100%
0 0%

X=1, Y=0

10955 Records
  * ADSL
    * yok (81,31%)
    * var 18,69%
      yok 81,31%
  * EV_IS
    * E (81,31%)
    * E 81,31%
      I 18,69%
  * BIRIM_ADI
    * GAZİOSMANPAŞA TELEKOM (8,05%)
    * AVCILAR TELEKOM 6,01%
      BAHÇELİEVLER TELEKOM 4,16%
      BAKIRKÖY TELEKOM 4,62%
      BAYRAMPAŞA (ESENLER) TELEKOM 4,16%
      BAĞCILAR TELEKOM 4,87%
      BEBEK TELEKOM 4,22%
      BEYOĞLU TELEKOM 7,36%
      BÜYÜKÇEKMECE TELEKOM 6,19%
      EMİNÖNÜ (SURIÇI) TELEKOM 1,03%
      ERENKÖY TELEKOM 7,07%
      FATİH TELEKOM 4,35%
      GAYRETTEPE TELEKOM 5,5%
      GAZİOSMANPAŞA TELEKOM 8,05%
      KADIKÖY TELEKOM 5,56%
      KÜÇÜKYALI TELEKOM 6,24%
      PENDİK TELEKOM 6,27%
      ÜMRANİYE TELEKOM 6%
      ÜSKÜDAR TELEKOM 5,46%
      İKİTELLİ TELEKOM 2,87%
  * MUSTERIDURUMU
    * SMALL (100%)
    * BIGGERHIGH 0%
      HIGH 0%
      MEDIUM 0%
      SMALL 100%
      ULTRA 0%
  * ODEMEDURUMU
    * + (100%)
    * + 100%
      - 0%
      0 0%

X=1, Y=2
6130 Records
  * ADSL
    * yok (53,56%)

* var 46,44%
yok 53,56%
* EV_IS
* E (70,57%)
* E 70,57%
I 29,43%
* BIRIM_ADI
* GAZİOSMANPAŞA TELEKOM (8,42%)
* AVCILAR TELEKOM 7,23%
BAHÇELİEVLER TELEKOM 4,24%
BAKIRKÖY TELEKOM 4,31%
BAYRAMPAŞA (ESENLER) TELEKOM 5,51%
BAĞCILAR TELEKOM 6,28%
BEBEK TELEKOM 2,99%
BEYOĞLU TELEKOM 6,97%
BÜYÜKÇEKMECE TELEKOM 5,27%
EMİNÖNÜ (SURIÇI) TELEKOM 1,92%
ERENKÖY TELEKOM 4,73%
FATİH TELEKOM 4,13%
GAYRETTEPE TELEKOM 4,5%
GAZİOSMANPAŞA TELEKOM 8,42%
KADIKÖY TELEKOM 4,24%
KÜÇÜKYALI TELEKOM 6,54%
PENDİK TELEKOM 7,55%
ÜMRANİYE TELEKOM 6,84%
ÜSKÜDAR TELEKOM 4,67%
İKİTELLİ TELEKOM 3,67%
* MUSTERIDURUMU
* SMALL (83,13%)
* BIGGERHIGH 1,57%
HIGH 14,91%
MEDIUM 0%
SMALL 83,13%
ULTRA 0,39%
* ODEMEDURUMU
* 0 (86%)
* + 0%
- 14%
0 86%

X=2, Y=0
6006 Records
* ADSL
* yok (100%)
* var 0%
yok 100%
* EV_IS

* I (91,58%)
* E 8,42%
I 91,58%
* BIRIM_ADI
* BEYOĞLU TELEKOM (13,24%)
* AVCILAR TELEKOM 3,3%
BAHÇELİEVLER TELEKOM 3,78%
BAKIRKÖY TELEKOM 4,41%
BAYRAMPAŞA (ESENLER) TELEKOM 5,04%
BAĞCILAR TELEKOM 5,76%
BEBEK TELEKOM 5,44%
BEYOĞLU TELEKOM 13,24%
BÜYÜKÇEKMECE TELEKOM 3,31%
EMİNÖNÜ (SURIÇI) TELEKOM 5,53%
ERENKÖY TELEKOM 5,81%
FATİH TELEKOM 3,16%
GAYRETTEPE TELEKOM 10,02%
GAZİOSMANPAŞA TELEKOM 2,9%
KADIKÖY TELEKOM 5,34%
KÜÇÜKYALI TELEKOM 3,45%
PENDİK TELEKOM 5,19%
ÜMRANİYE TELEKOM 7,53%
ÜSKÜDAR TELEKOM 3,13%
İKİTELLİ TELEKOM 3,65%
* MUSTERIDURUMU
* SMALL (91,58%)
* BIGGERHIGH 0,2%
HIGH 8,23%
MEDIUM 0%
SMALL 91,58%
ULTRA 0%
* ODEMEDURUMU
* + (100%)
* + 100%
- 0%
0 0%

X=2, Y=1
461 Records
* ADSL
* var (85,68%)
* var 85,68%
yok 14,32%
* EV_IS
* E (92,19%)
* E 92,19%
I 7,81%

* BIRIM_ADI
  * ERENKÖY TELEKOM (15,84%)
  * AVCILAR TELEKOM 4,77%
  BAHÇELİEVLER TELEKOM 3,04%
  BAKIRKÖY TELEKOM 9,11%
  BAYRAMPAŞA (ESENLER) TELEKOM 1,95%
  BAĞCILAR TELEKOM 1,3%
  BEBEK TELEKOM 7,81%
  BEYOĞLU TELEKOM 9,76%
  BÜYÜKÇEKMECE TELEKOM 6,94%
  EMİNÖNÜ (SURIÇI) TELEKOM 0,22%
  ERENKÖY TELEKOM 15,84%
  FATİH TELEKOM 2,6%
  GAYRETTEPE TELEKOM 7,16%
  GAZİOSMANPAŞA TELEKOM 3,25%
  KADIKÖY TELEKOM 5,86%
  KÜÇÜKYALI TELEKOM 4,99%
  PENDİK TELEKOM 3,25%
  ÜMRANİYE TELEKOM 2,82%
  ÜSKÜDAR TELEKOM 7,16%
  İKİTELLİ TELEKOM 2,17%
* MUSTERIDURUMU
  * HIGH (83,51%)
  * BIGGERHIGH 13,45%
  HIGH 83,51%
  MEDIUM 0%
  SMALL 0%
  ULTRA 3,04%
* ODEMEDURUMU
  * + (92,19%)
  * + 92,19%
  - 0%
  0 7,81%

X=2, Y=2
2108 Records
* ADSL
  * yok (54,17%)
  * var 45,83%
  yok 54,17%
* EV_IS
  * E (73,1%)
  * E 73,1%
  I 26,9%
* BIRIM_ADI
  * ÜSKÜDAR TELEKOM (8,16%)
  * AVCILAR TELEKOM 7,07%

BAHÇELİEVLER TELEKOM 3,89%
BAKIRKÖY TELEKOM 5,03%
BAYRAMPAŞA (ESENLER) TELEKOM 5,22%
BAĞCILAR TELEKOM 6,31%
BEBEK TELEKOM 2,99%
BEYOĞLU TELEKOM 7,92%
BÜYÜKÇEKMECE TELEKOM 3,8%
EMİNÖNÜ (SURIÇI) TELEKOM 1,94%
ERENKÖY TELEKOM 6,78%
FATİH TELEKOM 3,65%
GAYRETTEPE TELEKOM 4,32%
GAZİOSMANPAŞA TELEKOM 7,02%
KADIKÖY TELEKOM 4,03%
KÜÇÜKYALI TELEKOM 7,4%
PENDİK TELEKOM 5,31%
ÜMRANİYE TELEKOM 6,21%
ÜSKÜDAR TELEKOM 8,16%
İKİTELLİ TELEKOM 2,94%
* MUSTERIDURUMU
* MEDIUM (97,3%)
* BIGGERHIGH 0,43%
HIGH 2,23%
MEDIUM 97,3%
SMALL 0%
ULTRA 0,05%
* ODEMEDURUMU
* 0 (96,25%)
* + 0%
- 3,75%
0 96,25%

X=3, Y=0
2723 Records
* ADSL
* yok (87%)
* var 13%
yok 87%
* EV_IS
* I (99,38%)
* E 0,62%
I 99,38%
* BIRIM_ADI
* BEYOĞLU TELEKOM (13,88%)
* AVCILAR TELEKOM 3,93%
BAHÇELİEVLER TELEKOM 3,38%
BAKIRKÖY TELEKOM 5,1%
BAYRAMPAŞA (ESENLER) TELEKOM 4,77%

BAĞCILAR TELEKOM 5,99%
BEBEK TELEKOM 4%
BEYOĞLU TELEKOM 13,88%
BÜYÜKÇEKMECE TELEKOM 5,07%
EMİNÖNÜ (SURIÇI) TELEKOM 6,83%
ERENKÖY TELEKOM 5,36%
FATİH TELEKOM 2,06%
GAYRETTEPE TELEKOM 9,11%
GAZİOSMANPAŞA TELEKOM 2,9%
KADIKÖY TELEKOM 4,99%
KÜÇÜKYALI TELEKOM 4,59%
PENDİK TELEKOM 5,18%
ÜMRANİYE TELEKOM 6,06%
ÜSKÜDAR TELEKOM 2,97%
İKİTELLİ TELEKOM 3,82%
* MUSTERIDURUMU
* HIGH (80,46%)
* BIGGERHIGH 14,98%
HIGH 80,46%
MEDIUM 0%
SMALL 0%
ULTRA 4,55%
* ODEMEDURUMU
* + (92,88%)
* + 92,88%
- 0%
0 7,12%

X=3, Y=1
3384 Records
* ADSL
* yok (74,32%)
* var 25,68%
yok 74,32%
* EV_IS
* I (100%)
* E 0%
I 100%
* BIRIM_ADI
* BEYOĞLU TELEKOM (12,35%)
* AVCILAR TELEKOM 4,4%
BAHÇELİEVLER TELEKOM 4,4%
BAKIRKÖY TELEKOM 4,96%
BAYRAMPAŞA (ESENLER) TELEKOM 6,35%
BAĞCILAR TELEKOM 5,56%
BEBEK TELEKOM 2,98%
BEYOĞLU TELEKOM 12,35%

BÜYÜKÇEKMECE TELEKOM 3,75%
EMİNÖNÜ (SURIÇI) TELEKOM 6,77%
ERENKÖY TELEKOM 5,23%
FATİH TELEKOM 3,31%
GAYRETTEPE TELEKOM 8,45%
GAZİOSMANPAŞA TELEKOM 3,75%
KADIKÖY TELEKOM 4,96%
KÜÇÜKYALI TELEKOM 3,63%
PENDİK TELEKOM 5,61%
ÜMRANİYE TELEKOM 7,48%
ÜSKÜDAR TELEKOM 2,57%
İKİTELLİ TELEKOM 3,46%
* MUSTERIDURUMU
* MEDIUM (100%)
* BIGGERHIGH 0%
HIGH 0%
MEDIUM 100%
SMALL 0%
ULTRA 0%
* ODEMEDURUMU
* + (100%)
* + 100%
- 0%
0 0%

X=3, Y=2
10359 Records
* ADSL
* var (53,3%)
* var 53,3%
yok 46,7%
* EV_IS
* E (86,68%)
* E 86,68%
I 13,32%
* BIRIM_ADI
* ERENKÖY TELEKOM (9,05%)
* AVCILAR TELEKOM 6,32%
BAHÇELİEVLER TELEKOM 4,42%
BAKIRKÖY TELEKOM 5,77%
BAYRAMPAŞA (ESENLER) TELEKOM 3,79%
BAĞCILAR TELEKOM 4,64%
BEBEK TELEKOM 4,41%
BEYOĞLU TELEKOM 6,81%
BÜYÜKÇEKMECE TELEKOM 5,06%
EMİNÖNÜ (SURIÇI) TELEKOM 0,68%
ERENKÖY TELEKOM 9,05%

FATİH TELEKOM 3,38%
GAYRETTEPE TELEKOM 5,79%
GAZİOSMANPAŞA TELEKOM 6,63%
KADIKÖY TELEKOM 5,65%
KÜÇÜKYALI TELEKOM 7,43%
PENDİK TELEKOM 5,48%
ÜMRANİYE TELEKOM 6,39%
ÜSKÜDAR TELEKOM 4,83%
İKİTELLİ TELEKOM 3,46%

* MUSTERIDURUMU
  * MEDIUM (100%)
  * BIGGERHIGH 0%
  HIGH 0%
  MEDIUM 100%
  SMALL 0%
  ULTRA 0%
* ODEMEDURUMU
  * + (59,27%)
  * + 59,27%
  - 40,73%
  0 0%

EK-2- Abone Segmentasyonu için kurallar

```
Rules for 0 - 0 - contains 1 rule (s)
        Rule 1 for  0 - 0
                if MUSTERIDURUMU in [ "BIGGERHIGH" "HIGH" "SMALL"
"ULTRA" ]
                and ODEMEDURUMU in [ "+" "0" ]
                and ODEMEDURUMU in [ "+" ]
                and ADSL in [ "var" ]
                and EV_IS in [ "E" ]
                then 0 - 0
Rules for 0 - 2 - contains 1 rule (s)
        Rule 1 for  0 - 2
                if MUSTERIDURUMU in [ "BIGGERHIGH" "HIGH" "SMALL"
"ULTRA" ]
                and ODEMEDURUMU in [ "-" ]
                and MUSTERIDURUMU in [ "SMALL" ]
                then 0 - 2
Rules for 1 - 0 - contains 2 rule (s)
        Rule 1 for  1 - 0
                if MUSTERIDURUMU in [ "BIGGERHIGH" "HIGH" "SMALL"
"ULTRA" ]
                and ODEMEDURUMU in [ "+" "0" ]
                and ODEMEDURUMU in [ "+" ]
                and ADSL in [ "var" ]
```

```
                and EV_IS in [ "I" ]
                then 1 - 0
        Rule 2 for  1 - 0
                if MUSTERIDURUMU in [ "BIGGERHIGH" "HIGH" "SMALL"
"ULTRA" ]
                and ODEMEDURUMU in [ "+" "0" ]
                and ODEMEDURUMU in [ "+" ]
                and ADSL in [ "yok" ]
                and EV_IS in [ "E" ]
                then 1 - 0
Rules for 1 - 2 - contains 2 rule (s)
        Rule 1 for  1 - 2
                if MUSTERIDURUMU in [ "BIGGERHIGH" "HIGH" "SMALL"
"ULTRA" ]
                and ODEMEDURUMU in [ "-" ]
                and MUSTERIDURUMU in [ "BIGGERHIGH" "HIGH" "ULTRA" ]
                then 1 - 2
        Rule 2 for  1 - 2
                if MUSTERIDURUMU in [ "BIGGERHIGH" "HIGH" "SMALL"
"ULTRA" ]
                and ODEMEDURUMU in [ "+" "0" ]
                and ODEMEDURUMU in [ "0" ]
                then 1 - 2
Rules for 2 - 0 - contains 1 rule (s)
        Rule 1 for  2 - 0
                if MUSTERIDURUMU in [ "BIGGERHIGH" "HIGH" "SMALL"
"ULTRA" ]
                and ODEMEDURUMU in [ "+" "0" ]
                and ODEMEDURUMU in [ "+" ]
                and ADSL in [ "yok" ]
                and EV_IS in [ "I" ]
                then 2 - 0
Rules for 2 - 2 - contains 1 rule (s)
        Rule 1 for  2 - 2
                if MUSTERIDURUMU in [ "MEDIUM" ]
                and ODEMEDURUMU in [ "0" ]
                then 2 - 2
Rules for 3 - 1 - contains 1 rule (s)
        Rule 1 for  3 - 1
                if MUSTERIDURUMU in [ "MEDIUM" ]
                and ODEMEDURUMU in [ "+" "-" ]
                and EV_IS in [ "I" ]
                and ODEMEDURUMU in [ "+" ]
                then 3 - 1
Rules for 3 - 2 - contains 2 rule (s)
        Rule 1 for  3 - 2
                if MUSTERIDURUMU in [ "MEDIUM" ]
```

and ODEMEDURUMU in ["+" "-"]
and EV_IS in ["E"]
then 3 - 2

Rule 2 for 3 - 2
if MUSTERIDURUMU in ["MEDIUM"]
and ODEMEDURUMU in ["+" "-"]
and EV_IS in ["I"]
and ODEMEDURUMU in ["-"]
then 3 - 2

Default: 0 - 2

EK-3

Kohonen

X=0, Y=0
12482 Records
* Asaatliarama
  * Mean = 0,005
  * Standard Deviation = 0,11
* Ayonluarama
  * Mean = 0,0
  * Standard Deviation = 0,0
* Bsaatliarama
  * Mean = 0,036
  * Standard Deviation = 0,268
* Byonluarama
  * Mean = 0,0
  * Standard Deviation = 0,0
* Csaatliarama
  * Mean = 0,008
  * Standard Deviation = 0,117
* Cyonluarama
  * Mean = 0,0
  * Standard Deviation = 0,0
* Dsaatliarama
  * Mean = 0,0
  * Standard Deviation = 0,018
* Dyonluarama
  * Mean = 0,0
  * Standard Deviation = 0,0
* Esaatliarama
  * Mean = 0,0
  * Standard Deviation = 0,009
* Eyonluarama
  * Mean = 0,0

        * Standard Deviation = 0,0
    * Fsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,009
    * Fyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * ortkontur
        * Mean = 3,095
        * Standard Deviation = 4,93
    * toplamarama
        * Mean = 17,945
        * Standard Deviation = 9,737

X=0, Y=2
7253 Records
    * Asaatliarama
        * Mean = 0,056
        * Standard Deviation = 0,259
    * Ayonluarama
        * Mean = 1,0
        * Standard Deviation = 0,0
    * Bsaatliarama
        * Mean = 0,127
        * Standard Deviation = 0,355
    * Byonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Csaatliarama
        * Mean = 0,033
        * Standard Deviation = 0,189
    * Cyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dsaatliarama
        * Mean = 0,007
        * Standard Deviation = 0,098
    * Dyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Esaatliarama
        * Mean = 0,002
        * Standard Deviation = 0,041
    * Eyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fsaatliarama

        * Mean = 0,001
        * Standard Deviation = 0,041
    * Fyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * ortkontur
        * Mean = 2,429
        * Standard Deviation = 3,017
    * toplamarama
        * Mean = 43,87
        * Standard Deviation = 199,643

X=1, Y=0
8773 Records
    * Asaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Ayonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Bsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Byonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Csaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Cyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Esaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Eyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0

* Fyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* ortkontur
    * Mean = 2,423
    * Standard Deviation = 1,045
* toplamarama
    * Mean = 41,549
    * Standard Deviation = 7,791

X=2, Y=0
6374 Records
* Asaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Ayonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Bsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Byonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Csaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Cyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Dsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Dyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Esaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Eyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Fsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Fyonluarama
    * Mean = 0,0

        * Standard Deviation = 0,0
    * ortkontur
        * Mean = 2,448
        * Standard Deviation = 1,083
    * toplamarama
        * Mean = 65,571
        * Standard Deviation = 7,38

X=2, Y=1
3589 Records
    * Asaatliarama
        * Mean = 0,009
        * Standard Deviation = 0,097
    * Ayonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Bsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Byonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Csaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Cyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Esaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Eyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * ortkontur

        * Mean = 2,522
        * Standard Deviation = 1,267
    * toplamarama
        * Mean = 87,094
        * Standard Deviation = 9,929

X=2, Y=2
1416 Records
    * Asaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Ayonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Bsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Byonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Csaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Cyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Esaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Eyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * ortkontur
        * Mean = 2,686
        * Standard Deviation = 1,376

* toplamarama
    * Mean = 136,068
    * Standard Deviation = 7,163

X=3, Y=0
3370 Records
* Asaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Ayonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Bsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Byonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Csaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Cyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Dsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Dyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Esaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Eyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Fsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Fyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* ortkontur
    * Mean = 2,227
    * Standard Deviation = 0,891
* toplamarama
    * Mean = 111,03

        * Standard Deviation = 8,626

X=3, Y=1
2146 Records
    * Asaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Ayonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Bsaatliarama
        * Mean = 0,001
        * Standard Deviation = 0,031
    * Byonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Csaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Cyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Esaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Eyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * ortkontur
        * Mean = 2,084
        * Standard Deviation = 1,104
    * toplamarama
        * Mean = 204,199
        * Standard Deviation = 19,977

X=3, Y=2
926 Records
  * Asaatliarama
    * Mean = 0,001
    * Standard Deviation = 0,033
  * Ayonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Bsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Byonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Csaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Cyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Dsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Dyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Esaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Eyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Fsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * Fyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
  * ortkontur
    * Mean = 2,082
    * Standard Deviation = 1,036
  * toplamarama
    * Mean = 267,602
    * Standard Deviation = 15,436

X=4, Y=0
1468 Records

* Asaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Ayonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Bsaatliarama
    * Mean = 0,001
    * Standard Deviation = 0,026
* Byonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Csaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Cyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Dsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Dyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Esaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Eyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Fsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Fyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* ortkontur
    * Mean = 1,945
    * Standard Deviation = 0,644
* toplamarama
    * Mean = 158,511
    * Standard Deviation = 9,183

X=4, Y=1
753 Records
* Asaatliarama
    * Mean = 0,0

        * Standard Deviation = 0,0
    * Ayonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Bsaatliarama
        * Mean = 0,001
        * Standard Deviation = 0,036
    * Byonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Csaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Cyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Dyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Esaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Eyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fsaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Fyonluarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * ortkontur
        * Mean = 1,899
        * Standard Deviation = 0,625
    * toplamarama
        * Mean = 323,87
        * Standard Deviation = 18,404

X=4, Y=2
2409 Records
    * Asaatliarama
        * Mean = 0,0
        * Standard Deviation = 0,0
    * Ayonluarama

    * Mean = 0,0
    * Standard Deviation = 0,0
* Bsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Byonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Csaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Cyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Dsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Dyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Esaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,02
* Eyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Fsaatliarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* Fyonluarama
    * Mean = 0,0
    * Standard Deviation = 0,0
* ortkontur
    * Mean = 1,918
    * Standard Deviation = 0,546
* toplamarama
    * Mean = 719,617
    * Standard Deviation = 512,911

EK-4

Rules for 0 - 0 - contains 2 rule (s)

Rule 1 for 0 - 0

if Ayonluarama <= 0,500

and toplamarama <= 29,500

then 0 - 0

Rule 2 for 0 - 0

if Ayonluarama <= 0,500

and toplamarama > 29,500

and toplamarama <= 53,500

and ortkontur > 3,962

and toplamarama <= 39,500

then 0 - 0

Rules for 0 - 2 - contains 1 rule (s)

Rule 1 for 0 - 2

if Ayonluarama > 0,500

then 0 - 2

Rules for 1 - 0 - contains 2 rule (s)

Rule 1 for 1 - 0

if Ayonluarama <= 0,500

and toplamarama > 29,500

and toplamarama <= 53,500

and ortkontur <= 3,962

then 1 - 0

Rule 2 for 1 - 0

if Ayonluarama <= 0,500

and toplamarama > 29,500

and toplamarama <= 53,500

and ortkontur > 3,962

and toplamarama > 39,500

then 1 - 0

Rules for 2 - 0 - contains 1 rule (s)

Rule 1 for 2 - 0

if Ayonluarama <= 0,500

and toplamarama > 29,500

and toplamarama > 53,500

and toplamarama <= 78,500

then 2 - 0

Rules for 2 - 1 - contains 1 rule (s)

Rule 1 for 2 - 1

if Ayonluarama <= 0,500

and toplamarama > 29,500

and toplamarama > 53,500

and toplamarama > 78,500

and toplamarama <= 97,500

then 2 - 1

Rules for 3 - 0 - contains 1 rule (s)

Rule 1 for 3 - 0

if Ayonluarama <= 0,500

and toplamarama > 29,500

and toplamarama > 53,500

and toplamarama > 78,500

and toplamarama > 97,500

then 3 - 0

Default: 0 – 0

EK-5

Kohonen

X=0, Y=0
12029 Records
* ORJINAL_TUTAR
  * Mean = 21,199
  * Standard Deviation = 9,225
* ADSL
  * yok (50,11%)
  * var 49,89%
  yok 50,11%
* CINSIYETI
  * E (71,13%)
  * E 71,13%
  K 28,87%
* GERCEK_TUZEL

    * G (100%)
    * G 100%
      T 0%
* GELIRSEVIYE
    * M (90,55%)
    * A 6,33%
      B 0,64%
      C 1,09%
      D 0,25%
      E 0,06%
      F 0,02%
      G 0,02%
      H 0,01%
      J 0,05%
      K 0,98%
      L 0,02%
      M 90,55%
* MUSTERIDURUMU
    * SMALL (100%)
    * BIGGERHIGH 0%
      HIGH 0%
      MEDIUM 0%
      SMALL 100%
      ULTRA 0%
* NUFUSBOLGE
    * KARADENİZ (39,33%)
    * AKDENİZ 3,49%
      BİLİNMEYEN 2,49%
      DOĞU ANADOLU 19,53%
      EGE 3,4%
      GÜNEYDOĞU ANADOLU 7,45%
      KARADENİZ 39,33%
      MARMARA 2,47%
      YABANCI 1,02%
      İÇ ANADOLU 20,82%
* ODEMEDURUMU
    * - (72,12%)
    * + 0%
      - 72,12%
      0 27,88%
* OGRENIMSEVIYE
    * E (46,5%)
    * - 0,16%
      A 27,26%
      B 7,56%
      C 13,3%
      D 5,23%

E 46,5%
* SIRKETTURU
* - (100%)
* - 100%
A 0%
B 0%
C 0%
D 0%
E 0%
* YASSINIF
* C (29,11%)
* A 0,67%
B 16,58%
C 29,11%
D 28,24%
E 15,84%
F 9,56%

X=0, Y=1
651 Records
* ORJINAL_TUTAR
* Mean = 100,393
* Standard Deviation = 105,563
* ADSL
* yok (52,07%)
* var 47,93%
yok 52,07%
* CINSIYETI
* E (77,42%)
* E 77,42%
K 22,58%
* GERCEK_TUZEL
* G (100%)
* G 100%
T 0%
* GELIRSEVIYE
* M (93,23%)
* A 4,46%
B 0,31%
C 0,92%
D 0%
E 0%
F 0%
G 0%
H 0%
J 0,15%
K 0,77%

L 0,15%
M 93,23%
* MUSTERIDURUMU
  * HIGH (91,69%)
  * BIGGERHIGH 6,77%
  HIGH 91,69%
  MEDIUM 0%
  SMALL 0%
  ULTRA 1,54%
* NUFUSBOLGE
  * KARADENİZ (31,69%)
  * AKDENİZ 3,54%
  BİLİNMEYEN 0,62%
  DOĞU ANADOLU 19,69%
  EGE 3,69%
  GÜNEYDOĞU ANADOLU 5,69%
  KARADENİZ 31,69%
  MARMARA 14,46%
  YABANCI 1,85%
  İÇ ANADOLU 18,77%
* ODEMEDURUMU
  * - (76,15%)
  * + 0%
  - 76,15%
  0 23,85%
* OGRENIMSEVIYE
  * E (43,69%)
  * - 0%
  A 23,38%
  B 7,85%
  C 17,08%
  D 8%
  E 43,69%
* SIRKETTURU
  * - (100%)
  * - 100%
  A 0%
  B 0%
  C 0%
  D 0%
  E 0%
* YASSINIF
  * C (28,31%)
  * A 0,15%
  B 18,92%
  C 28,31%
  D 26,77%

E 15,08%
F 10,77%

X=0, Y=2
7885 Records
* ORJINAL_TUTAR
* Mean = 41,114
* Standard Deviation = 19,264
* ADSL
* var (65,66%)
* var 65,66%
yok 34,34%
* CINSIYETI
* E (66,48%)
* E 66,48%
K 33,52%
* GERCEK_TUZEL
* G (100%)
* G 100%
T 0%
* GELIRSEVIYE
* M (88,86%)
* A 7,29%
B 0,84%
C 1,27%
D 0,34%
E 0,09%
F 0,04%
G 0%
H 0%
J 0,05%
K 1,15%
L 0,06%
M 88,86%
* MUSTERIDURUMU
* MEDIUM (100%)
* BIGGERHIGH 0%
HIGH 0%
MEDIUM 100%
SMALL 0%
ULTRA 0%
* NUFUSBOLGE
* MARMARA (38,06%)
* AKDENİZ 2,09%
BİLİNMEYEN 0,33%
DOĞU ANADOLU 16,2%
EGE 2,85%

GÜNEYDOĞU ANADOLU 3,83%
KARADENİZ 21,33%
MARMARA 38,06%
YABANCI 0,77%
İÇ ANADOLU 14,53%

* ODEMEDURUMU
  * - (45,39%)
  * + 37,48% , - 45,39% , 0 17,13%
* OGRENIMSEVIYE
  * A (36,53%)
  * - 0,18%
    A 36,53%
    B 10,96%
    C 19,32%
    D 9,5%
    E 23,53%
* SIRKETTURU
  * - (100%)
  * - 100%
    A 0%
    B 0%
    C 0%
    D 0%
    E 0%
* YASSINIF
  * D (31,44%)
  * A 0,24%
    B 10,08%
    C 25,02%
    D 31,44%
    E 19,04%
    F 14,18%

X=1, Y=0
3775 Records
* ORJINAL_TUTAR
  * Mean = 21,746
  * Standard Deviation = 9,913
* ADSL
  * yok (50,57%)
  * var 49,43%
    yok 50,57%
* CINSIYETI
  * E (61,17%)
  * E 61,17%
    K 38,83%
* GERCEK_TUZEL

    * G (99,97%)
    * G 99,97%
      T 0,03%
* GELIRSEVIYE
    * M (92,32%)
    * A 5,25%
      B 0,53%
      C 0,98%
      D 0,24%
      E 0,11%
      F 0,03%
      G 0%
      H 0%
      J 0,08%
      K 0,48%
      L 0%
      M 92,32%
* MUSTERIDURUMU
    * SMALL (100%)
    * BIGGERHIGH 0%
      HIGH 0%
      MEDIUM 0%
      SMALL 100%
      ULTRA 0%
* NUFUSBOLGE
    * MARMARA (99,63%)
    * AKDENİZ 0,03%
      BİLİNMEYEN 0,26%
      DOĞU ANADOLU 0%
      EGE 0,08%
      GÜNEYDOĞU ANADOLU 0%
      KARADENİZ 0%
      MARMARA 99,63%
      YABANCI 0%
      İÇ ANADOLU 0%
* ODEMEDURUMU
    * - (65,83%)
    * + 0%
      - 65,83%
      0 34,17%
* OGRENIMSEVIYE
    * E (48,21%)
    * - 0,08%
      A 21,75%
      B 10,33%
      C 12,19%
      D 7,44%

E 48,21%
* SIRKETTURU
* - (100%)
* - 100%
A 0%
B 0%
C 0%
D 0%
E 0%
* YASSINIF
* C (23,58%)
* A 0,42%
B 9,09%
C 23,58%
D 22,41%
E 22,91%
F 21,59%

X=1, Y=1
336 Records
* ORJINAL_TUTAR
* Mean = 87,038
* Standard Deviation = 58,629
* ADSL
* yok (54,17%)
* var 45,83%
yok 54,17%
* CINSIYETI
* E (50,3%)
* E 50,3%
K 49,7%
* GERCEK_TUZEL
* G (100%)
* G 100%
T 0%
* GELIRSEVIYE
* M (91,96%)
* A 4,76%
B 0,3%
C 0,89%
D 0%
E 0,3%
F 0,3%
G 0%
H 0%
J 0,3%
K 0,89%

L 0,3%
M 91,96%
* MUSTERIDURUMU
  * HIGH (94,64%)
  * BIGGERHIGH 4,76%
  HIGH 94,64%
  MEDIUM 0%
  SMALL 0%
  ULTRA 0,6%
* NUFUSBOLGE
  * MARMARA (97,62%)
  * AKDENİZ 0%
  BİLİNMEYEN 0%
  DOĞU ANADOLU 0%
  EGE 0,3%
  GÜNEYDOĞU ANADOLU 0,3%
  KARADENİZ 0,89%
  MARMARA 97,62%
  YABANCI 0%
  İÇ ANADOLU 0,89%
* ODEMEDURUMU
  * + (40,77%)
  * + 40,77%
  - 28,57%
  0 30,65%
* OGRENIMSEVIYE
  * E (33,63%)
  * - 0%
  A 28,57%
  B 8,33%
  C 10,42%
  D 19,05%
  E 33,63%
* SIRKETTURU
  * - (100%)
  * - 100%
  A 0%
  B 0%
  C 0%
  D 0%
  E 0%
* YASSINIF
  * E,F (26,19%)
  * A 0%
  B 5,95%
  C 22,92%
  D 18,75%

E 26,19%
F 26,19%

X=1, Y=2
5681 Records
* ORJINAL_TUTAR
* Mean = 43,082
* Standard Deviation = 24,084
* ADSL
* yok (67,07%)
* var 32,93%
yok 67,07%
* CINSIYETI
* E (71,34%)
* E 71,34%
K 28,66%
* GERCEK_TUZEL
* G (100%)
* G 100%
T 0%
* GELIRSEVIYE
* M (97,85%)
* A 1,32%
B 0,19%
C 0,21%
D 0,09%
E 0,07%
F 0,02%
G 0%
H 0,02%
J 0%
K 0,23%
L 0%
M 97,85%
* MUSTERIDURUMU
* MEDIUM (94,37%)
* BIGGERHIGH 0,14%
HIGH 5,47%
MEDIUM 94,37%
SMALL 0%
ULTRA 0,02%
* NUFUSBOLGE
* MARMARA (31,65%)
* AKDENİZ 3,06%
BİLİNMEYEN 1,87%
DOĞU ANADOLU 11,28%
EGE 3,59%

GÜNEYDOĞU ANADOLU 4,05%
KARADENİZ 27,97%
MARMARA 31,65%
YABANCI 0,86%
İÇ ANADOLU 15,67%
* ODEMEDURUMU
* + (86,22%)
* + 86,22%
- 6,92%
0 6,86%
* OGRENIMSEVIYE
* E (71,99%)
* - 0,12%
A 6,51%
B 7,25%
C 8,33%
D 5,79%
E 71,99%
* SIRKETTURU
* - (100%)
* - 100%
A 0%
B 0%
C 0%
D 0%
E 0%
* YASSINIF
* C (24,31%)
* A 0,07%
B 6,35%
C 24,31%
D 24,08%
E 23,69%
F 21,49%

X=2, Y=0
7333 Records
* ORJINAL_TUTAR
* Mean = 22,64
* Standard Deviation = 18,489
* ADSL
* yok (58,3%)
* var 41,7%
yok 58,3%
* CINSIYETI
* E (64,5%)
* E 64,5%

K 35,5%
* GERCEK_TUZEL
* G (100%)
* G 100%
T 0%
* GELIRSEVIYE
* M (93,28%)
* A 4,36%
B 0,64%
C 0,72%
D 0,19%
E 0,14%
F 0,03%
G 0%
H 0%
J 0,04%
K 0,56%
L 0,04%
M 93,28%
* MUSTERIDURUMU
* SMALL (98,04%)
* BIGGERHIGH 0,25%
HIGH 1,66%
MEDIUM 0%
SMALL 98,04%
ULTRA 0,05%
* NUFUSBOLGE
* MARMARA (100%)
* AKDENİZ 0%
BİLİNMEYEN 0%
DOĞU ANADOLU 0%
EGE 0%
GÜNEYDOĞU ANADOLU 0%
KARADENİZ 0%
MARMARA 100%
YABANCI 0%
İÇ ANADOLU 0%
* ODEMEDURUMU
* + (100%)
* + 100%
- 0%
0 0%
* OGRENIMSEVIYE
* E (41,87%)
* - 0,03%
A 24,06%
B 9,59%

C 14,58%
D 9,89%
E 41,87%
* SIRKETTURU
* - (100%)
* - 100%
A 0%
B 0%
C 0%
D 0%
E 0%
* YASSINIF
* F (33,86%)
* A 0%
B 4,88%
C 15,89%
D 22,04%
E 23,33%
F 33,86%

X=2, Y=1
595 Records
* ORJINAL_TUTAR
* Mean = 83,284
* Standard Deviation = 101,304
* ADSL
* yok (70,08%)
* var 29,92%
yok 70,08%
* CINSIYETI
* E (75,63%)
* E 75,63%
K 24,37%
* GERCEK_TUZEL
* G (100%)
* G 100%
T 0%
* GELIRSEVIYE
* M (99,66%)
* A 0%
B 0%
C 0,17%
D 0,17%
E 0%
F 0%
G 0%
H 0%

J 0%
K 0%
L 0%
M 99,66%
* MUSTERIDURUMU
* HIGH (72,44%)
* BIGGERHIGH 5,55%
HIGH 72,44%
MEDIUM 0%
SMALL 21,01%
ULTRA 1,01%
* NUFUSBOLGE
* MARMARA (51,6%)
* AKDENİZ 7,56%
BİLİNMEYEN 11,93%
DOĞU ANADOLU 4,37%
EGE 5,71%
GÜNEYDOĞU ANADOLU 2,35%
KARADENİZ 8,57%
MARMARA 51,6%
YABANCI 0,84%
İÇ ANADOLU 7,06%
* ODEMEDURUMU
* + (88,07%)
* + 88,07%
- 0%
0 11,93%
* OGRENIMSEVIYE
* E (85,04%)
* - 0%
A 0%
B 4,54%
C 4,03%
D 6,39%
E 85,04%
* SIRKETTURU
* - (100%)
* - 100%
A 0%
B 0%
C 0%
D 0%
E 0%
* YASSINIF
* E (30,59%)
* A 0,34%
B 5,04%

C 13,78%
D 29,92%
E 30,59%
F 20,34%

X=2, Y=2
590 Records
* ORJINAL_TUTAR
* Mean = 61,875
* Standard Deviation = 44,571
* ADSL
* yok (88,14%)
* var 11,86%
yok 88,14%
* CINSIYETI
* E (83,39%)
* E 83,39%
K 16,61%
* GERCEK_TUZEL
* G (98,31%)
* G 98,31%
T 1,69%
* GELIRSEVIYE
* M (99,66%)
* A 0,34%
B 0%
C 0%
D 0%
E 0%
F 0%
G 0%
H 0%
J 0%
K 0%
L 0%
M 99,66%
* MUSTERIDURUMU
* MEDIUM (56,78%)
* BIGGERHIGH 1,69%
HIGH 41,36%
MEDIUM 56,78%
SMALL 0%
ULTRA 0,17%
* NUFUSBOLGE
* BİLİNMEYEN (65,93%)
* AKDENİZ 0,68%
BİLİNMEYEN 65,93%

DOĞU ANADOLU 9,32%
EGE 2,54%
GÜNEYDOĞU ANADOLU 1,53%
KARADENİZ 11,19%
MARMARA 1,02%
YABANCI 0,17%
İÇ ANADOLU 7,63%

* ODEMEDURUMU
  * + (87,97%)
  * + 87,97%
  - 4,24%
  0 7,8%
* OGRENIMSEVIYE
  * E (99,66%)
  * - 0,34%
  A 0%
  B 0%
  C 0%
  D 0%
  E 99,66%
* SIRKETTURU
  * - (100%)
  * - 100%
  A 0%
  B 0%
  C 0%
  D 0%
  E 0%
* YASSINIF
  * F (30%)
  * A 14,41%
  B 3,22%
  C 13,22%
  D 20,34%
  E 18,81%
  F 30%

X=3, Y=0
11584 Records
* ORJINAL_TUTAR
  * Mean = 22,625
  * Standard Deviation = 19,717
* ADSL
  * var (54,06%)
  * var 54,06%
  yok 45,94%
* CINSIYETI

    * E (70%)
    * E 70%
      K 30%
* GERCEK_TUZEL
    * G (100%)
    * G 100%
      T 0%
* GELIRSEVIYE
    * M (91,57%)
    * A 5,25%
      B 0,64%
      C 1,16%
      D 0,33%
      E 0,11%
      F 0,03%
      G 0%
      H 0%
      J 0,03%
      K 0,85%
      L 0,03%
      M 91,57%
* MUSTERIDURUMU
    * SMALL (98,3%)
    * BIGGERHIGH 0,13%
      HIGH 1,52%
      MEDIUM 0%
      SMALL 98,3%
      ULTRA 0,05%
* NUFUSBOLGE
    * KARADENİZ (43,97%)
    * AKDENİZ 4,47%
      BİLİNMEYEN 0,02%
      DOĞU ANADOLU 18,44%
      EGE 5,79%
      GÜNEYDOĞU ANADOLU 5,09%
      KARADENİZ 43,97%
      MARMARA 0,16%
      YABANCI 0,89%
      İÇ ANADOLU 21,17%
* ODEMEDURUMU
    * + (100%)
    * + 100%
      - 0%
      0 0%
* OGRENIMSEVIYE
    * E (38,81%)
    * - 0,08%

A 28,43%
B 9,31%
C 13,44%
D 9,94%
E 38,81%
* SIRKETTURU
* - (100%)
* - 100%
A 0%
B 0%
C 0%
D 0%
E 0%
* YASSINIF
* D (28,76%)
* A 0,02%
B 9,11%
C 27,62%
D 28,76%
E 18,75%
F 15,74%

X=3, Y=1
1858 Records
* ORJINAL_TUTAR
* Mean = 21,913
* Standard Deviation = 13,109
* ADSL
* yok (91,39%)
* var 8,61%
yok 91,39%
* CINSIYETI
* E (88,91%)
* E 88,91%
K 11,09%
* GERCEK_TUZEL
* G (99,95%)
* G 99,95%
T 0,05%
* GELIRSEVIYE
* M (99,52%)
* A 0,43%
B 0%
C 0%
D 0%
E 0%
F 0%

G 0%
H 0%
J 0%
K 0,05%
L 0%
M 99,52%
* MUSTERIDURUMU
* SMALL (98,71%)
* BIGGERHIGH 0%
HIGH 1,29%
MEDIUM 0%
SMALL 98,71%
ULTRA 0%
* NUFUSBOLGE
* BİLİNMEYEN (34,07%)
* AKDENİZ 3,39%
BİLİNMEYEN 34,07%
DOĞU ANADOLU 16,79%
EGE 4,41%
GÜNEYDOĞU ANADOLU 8,61%
KARADENİZ 2,85%
MARMARA 0,38%
YABANCI 1,4%
İÇ ANADOLU 28,09%
* ODEMEDURUMU
* + (100%)
* + 100%
- 0%
0 0%
* OGRENIMSEVIYE
* E (98,39%)
* - 0%
A 0,05%
B 0,27%
C 0,65%
D 0,65%
E 98,39%
* SIRKETTURU
* - (100%)
* - 100%
A 0%
B 0%
C 0%
D 0%
E 0%
* YASSINIF
* F (31,05%)

* A 1,51%
B 7,32%
C 19,54%
D 19,16%
E 21,42%
F 31,05%

X=3, Y=2
9938 Records
* ORJINAL_TUTAR
* Mean = 54,392
* Standard Deviation = 88,543
* ADSL
* yok (81,31%)
* var 18,69%
yok 81,31%
* CINSIYETI
* E (99,58%)
* E 99,58%
K 0,42%
* GERCEK_TUZEL
* T (98,43%)
* G 1,57%
T 98,43%
* GELIRSEVIYE
* M (99,09%)
* A 0,74%
B 0,01%
C 0,04%
D 0,01%
E 0,02%
F 0,02%
G 0,01%
H 0,02%
J 0,02%
K 0,01%
L 0%
M 99,09%
* MUSTERIDURUMU
* SMALL (53,67%)
* BIGGERHIGH 4,52%
HIGH 18,2%
MEDIUM 22,27%
SMALL 53,67%
ULTRA 1,34%
* NUFUSBOLGE
* BİLİNMEYEN (99,98%)

* AKDENİZ 0%
BİLİNMEYEN 99,98%
DOĞU ANADOLU 0,02%
EGE 0%
GÜNEYDOĞU ANADOLU 0%
KARADENİZ 0%
MARMARA 0%
YABANCI 0%
İÇ ANADOLU 0%

* ODEMEDURUMU
* + (78,52%)
* + 78,52%
- 13,73%
0 7,76%

* OGRENIMSEVIYE
* E (99,84%)
* - 0,16%
A 0%
B 0%
C 0%
D 0%
E 99,84%

* SIRKETTURU
* - (97,42%)
* - 97,42%
A 0,01%
B 0,73%
C 1,3%
D 0,03%
E 0,5%

* YASSINIF
* A (100%)
* A 100%
B 0%
C 0%
D 0%
E 0%
F 0%

EK-6

Rules for 0 - 0 - contains 1 rule (s)

Rule 1 for 0 - 0

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["BIGGERHIGH" "HIGH" "SMALL" "ULTRA"]

and ODEMEDURUMU in ["-" "0"]

and NUFUSBOLGE in ["AKDENİZ" "BİLİNMEYEN" "DOĞU ANADOLU" "EGE" "GÜNEYDOĞU ANADOLU" "KARADENİZ" "YABANCI" "İÇ ANADOLU"]

then 0 - 0

Rules for 0 - 2 - contains 3 rule (s)

Rule 1 for 0 - 2

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["MEDIUM"]

and OGRENIMSEVIYE in ["-" "A" "B" "C" "D"]

and ADSL in ["var"]

then 0 - 2

Rule 2 for 0 - 2

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["MEDIUM"]

and OGRENIMSEVIYE in ["-" "A" "B" "C" "D"]

and ADSL in ["yok"]

and ODEMEDURUMU in ["-" "0"]

then 0 - 2

Rule 3 for 0 - 2

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["MEDIUM"]

and OGRENIMSEVIYE in ["E"]

and ODEMEDURUMU in ["-" "0"]

and ADSL in ["var"]

then 0 - 2

Rules for 1 - 0 - contains 1 rule (s)

Rule 1 for 1 - 0

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["BIGGERHIGH" "HIGH" "SMALL" "ULTRA"]

and ODEMEDURUMU in ["-" "0"]

and NUFUSBOLGE in ["MARMARA"]

then 1 - 0

Rules for 1 - 2 - contains 3 rule (s)

Rule 1 for 1 - 2

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["MEDIUM"]

and OGRENIMSEVIYE in ["-" "A" "B" "C" "D"]

and ADSL in ["yok"]

and ODEMEDURUMU in ["+"]

then 1 - 2

Rule 2 for 1 - 2

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["MEDIUM"]

and OGRENIMSEVIYE in ["E"]

and ODEMEDURUMU in ["+"]

then 1 - 2

Rule 3 for 1 - 2

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["MEDIUM"]

and OGRENIMSEVIYE in ["E"]

and ODEMEDURUMU in ["-" "0"]

and ADSL in ["yok"]

then 1 - 2

Rules for 2 - 0 - contains 1 rule (s)

Rule 1 for 2 - 0

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["BIGGERHIGH" "HIGH" "SMALL" "ULTRA"]

and ODEMEDURUMU in ["+"]

and NUFUSBOLGE in ["MARMARA"]

then 2 - 0

Rules for 3 - 0 - contains 1 rule (s)

Rule 1 for 3 - 0

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["BIGGERHIGH" "HIGH" "SMALL" "ULTRA"]

and ODEMEDURUMU in ["+"]

and NUFUSBOLGE in ["AKDENİZ" "BİLİNMEYEN" "DOĞU ANADOLU" "EGE" "GÜNEYDOĞU ANADOLU" "KARADENİZ" "YABANCI" "İÇ ANADOLU"]

and NUFUSBOLGE in ["AKDENİZ" "DOĞU ANADOLU" "EGE" "GÜNEYDOĞU ANADOLU" "KARADENİZ" "YABANCI" "İÇ ANADOLU"]

then 3 - 0

Rules for 3 - 1 - contains 1 rule (s)

Rule 1 for 3 - 1

if GERCEK_TUZEL in ["G"]

and MUSTERIDURUMU in ["BIGGERHIGH" "HIGH" "SMALL" "ULTRA"]

and ODEMEDURUMU in ["+"]

and NUFUSBOLGE in ["AKDENİZ" "BİLİNMEYEN" "DOĞU ANADOLU" "EGE" "GÜNEYDOĞU ANADOLU" "KARADENİZ" "YABANCI" "İÇ ANADOLU"]

and NUFUSBOLGE in ["BİLİNMEYEN"]

then 3 - 1

Rules for 3 - 2 - contains 1 rule (s)

Rule 1 for 3 - 2

if GERCEK_TUZEL in ["T"]

then 3 - 2

Default: 0 – 0

KAYNAKLAR

Aaker, D.(1971). *Multivariate Analysis in Marketing: Theory and Application*, Wadswort Publishing, California.

Acungil, M. (b.t.) *Veri madenciliği.* 20.12.2009, http://mustafaacungil.blogspot.com/

Agrawal, R., Imielinski, T., Swami, A. (**1993**). *Mining association rules between sets of items in large databases.* ACM SIGMOD Conference on Management of Data, Washington.

Akbulut, S. (2006). *Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu*, Yayınlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi

Akpınar, H. (Nisan 2000). *Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği*, İ.Ü.İşletme Fakültesi Dergisi, Sayı:1, İstanbul

Alpaydın, E. (b.t.). *Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri*. Bilişim 2000 Eğitim Semineri, 10.04.2004, http://www.cmpe.boun.edu.tr/~ethem

Atamer, B. (1992). *Kümeleme Analizi (Cluster Analysis) ve Kümeleme Analizinin İlaç Sektörüne Uygulanması*, Yayınlanmamış Yüksek lisans Tezi, İstanbul

Aytekin, G. (2002). *Perakendecilik Sektöründe Veri Ambarı Uygulamaları Üzerine Bir Araştırma*. Cilt 5, Sayı 17 .

Bergeron, B. (2002). *Bioinformatics Computing*, Prentice Hall PTR, U.S.A.

Berkhin, P. (2002). *Survey of Clustering Data Mining Techniques*, California, U.S.A.

Berry, M., Linoff, G. (2000). *The Art and Science of Customer Relationship Management*, Wiley Computer Publishing, U.S.A.

Bozkır, A.S., Gök, B., Sezer, E. (2008). *Üniversite Öğrencilerinin İnterneti Eğitimsel Amaçlar İçin Kullanmalarını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti*, BUMAT 2008: Bilimde Modern Yöntemler Sempozyumu.

Chen, M.S., Han, J., Yu, P. S. (1996). *Data Mining: An Overview from a Database Perspective.* IEEE Transactions on Knowledge and Data Engineering,8(6)

Çakmak, Z., Uzgören, N., Keçek, G. (b.t.). *Kümeleme Analizi Teknikleri ile llerin Kültürel Yapılarına Göre Sınıflandırılması ve Değişimlerin İncelenmesi*

Daşdemir, Y. *Veri Tabanı ve Yönetim Sistemleri*. (2004). Türkmen yayınevi, İstanbul

Data Mining. (b.t.). metinler. 25.08.2004, http://www.sas.com/technologies/analytics/datamining/index.html

Data Mining (Veri Madenciliği). (b.t.). metinler. 25.08.2004, http://www.spss.com.tr/Clementine.asp

Data Mining Tools-METAspectrumSM Evaluation. (b.t.). metinler. 19Aralık 2009, http://www.oracle.com/technology/products/bi/odm/pdf/odm_metaspectrum_1004.pdf

Demir, F., Kırdar, Y. (b.t.). Müşteri İlişkileri Yönetimi:CRM. 28.11.2009.

Demiralay, M. Çamurcu, A.Y. (2005). *Cure, Agnes ve K-Means Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması.* İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Yıl:4 Sayı:8 ,s.1-18.

Dillon R. W., Goldstein, M. (b.t.). *Multivariate Analysis:Methods and Applications*.18.12.2009, http://www.amazon.ca/gp/reader/0471083178/ref=sib_fs_top?ie=UTF8&p=S00K&checkSum=lAdt6MOqLkOurBcLk40%2B3stmISqleGR%2F20cvtkkrmm8%3D#reader-link

Duran, B.S. and P.L. Odell (1974). *Cluster Analysis (Lecture Notes in Economics and Mathematical Systems, Econometrics*; *Managing Editors: M. Beckmann and H.P. Kunzi*). Springer-Verlag: New York

Duran, M. (2002). Veri Tabanı Pazarlama. 27.02.2005. www.danismend.com

Everitt, B. (1974). *Cluster Analysis*. Heinemann Educational Books Ltd, London.

ETL Tools-METAspectrumSM Evaluation. (b.t.). metinler.19 aralık 2009, http://www.sas.com/offices/europe/czech/technologies/enterpriseintelligenceplatform/MetagroupETLmarket.pdf

Fayyad, U. (March 1998) *Mining Databases: Towards Algorithms for Knowledge Discovery*, IEEE Bulletin of the Technical Committee on Data Engineering, Vol.21 No1

Flexer, A. (2001). *On the Use of Self-Organizing Maps for Clustering and Visualization*. Intelligent Data Analysis 5.

Gray, P., Watson, H. (1998). *Decision Support in the Data Warehouse*, Prentice Hall PTR, New Jersey.

Gray, P. Watson, H.J. (1998). *Decision Support in The Data Warehouse.* Prentice Hall, USA

Han, J., Kamber, M. (2000). *Data Mining Concepts and Techniques*. USA: Morgan Kaufmann Publishers.

Harrold,D. (2000), *What.s Your Data Telling You?*,Control Engineering

http://www.erpcrm.com/crm_anasf/crm_mimarisi.htm. 2005

Inmon, W.H. (2002). *Building the Data Warehouse*. Wiley, U.S.A.

İnternet Terimleri Sözlüğü.(b.t.). metinler. 19.12.2009, http://www.ttnet.com.tr/web/208- 997-1-1/tr/ttnet/internet_terimleri_sozlugu/_internet_terimleri_sozlugu

Johnson A. R., Wichern W. D. (1998). *Applied Multivariate Statistical Analysis*. Prentice Hall, U.S.A.

Kiang M.Y., Kumar, A. (2001). *An Evaluation of Self Organizing Map Networks as a Robust Alternative to Factor Analysis in Data Mining Applications.* Information Systems Research, Vol.12. No.2.

Karakaş, M. (b.t.) Veri Ambarları Genel Yapısı, 07.10.2004, http://www.bilgiyonetimi.com

Karasulu, B., Uğur, A. (2007). *Özörgütlemeli Yapay Sinir Ağı Modeli'nin Kullanıldığı Kutup Dengeleme Problemi İçin Paralel Hesaplama Tekniği İle Bir Başarım En iyileştirme Yöntemi,* Dumlupınar Üniversitesi, Akademik Bilişim 07, Kütahya.

Kohonen T. (1995). *Self Organization Maps*. Springer, Berlin.

Kohonen, T. (1996). *The speedy SOM. Technical Report A33*. Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland

Kurtuluş, K. (1998). *Pazarlama Araştırmaları.* Avcıol Basım, İstanbul

Kümeleme analizi:Temel Kavramar ve Algoritmalar. (b.t.).metinler. 15.10.2009. www.bilmuh.gyte.edu.tr/~htakci/vm/kumeleme_analizi.doc

Levene, M., Loizou, G. (2003). *Why is the Snowflake Schema a Good Data Warehouse Design**?**", Information Systems***,** Volume 28, Issue 3.

Mark Madsen. (5.05.2009). *The Datawarehouse Insfracture*. TDWI 2008 Learning book.

Mark Madsen. (6.05.2009). *Evaluating ETL Tools and Tecnologies*.

TDWI 2008 Learning book.

Massa, S., Testa, S. (2005). *Data-Warehouse in Practice: Exploring the Function of Expectations in Organizational Outcomes". Information & Management* (Volume 42 Issue 5) .

Meyer, D., Cannon, C. (1998). *Building a Better Data Warehouse*, Prentice Hall PTR, New Jersey.

Müşteri İlişkileri Yönetimi. (23 Mart 2003). metinler. 10 Aralık 2005. http://www.boyutbilgi.com.tr

Müşteri İlişkileri Yönetimi: CRM Alt Yapısının Oluşturulması. (b.t.). 05.01.2010, http://www.teknoturk.org

Oğuzlar, A. (2005). *Kümeleme Analizinde Yeni Bir Yaklaşım: Kendini Düzenleyen haritalar:KOHONEN Ağları.* İktisadi ve İdari Bilimler Dergisi, Cilt: 19 Eylül 2005 Sayı:2

Orhunbilge, N. (1996). *Uygulamalı Regresyon ve Korelasyon Analizi*, İ.Ü.İşletme Fakültesi Yayınları, , İstanbul.

Özçakır, F. (2006). Müşteri İlişkilerindeki Birlikteliklerin Belirlenmesinde Veri Madenciliği Uygulaması, Yayınlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi

Özekeş S., Çamurcu Y., (*2003). Veri Madenciliğinde Karar Ağaçları Yöntemi Uygulaması.* Pamukkale Üniversitesi - Bilgi Teknolojileri Kongresi II.

Özekeş, S. *(2003). Veri Madenciligi Modelleri ve Uygulama Alanları,* İstanbul Ticaret Üniversitesi Dergisi, 2, No. 3.

Özmen, Ş. *İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor*, Marmara Üniversitesi İ.İ.B.F, 04.05.2004. http://idari.cu.edu.tr/sempozyum/bil38.htm

Öztürk, B. Tarımcı, A. *Veri Ambarlama*, 17.01.2005, http://www.kouemk.com/makale/default.asp?set=makale&id=4

Pipe,P. (1997). *The Data Mart: A New Approach to Data Warehousing*. International Review of Law, Computers & Technology, Volume 11.

Rauber, A., Tomsich, P., Merkl, D. (2000). parSOM: A Parallel Implementation of the Self-Organizing Map Exploiting Cache Effects: Making the SOM Fit for Interactive High-Performance Data Analysis. HPCN Europe.

Salvatore, M., Hevner, A. (2005).

Integrated Decision Support Systems: A Data Warehousing Perspective. http://datawarehouse.ittoolbox.com/documents.asp?i=2231

Singh, H. (1997). *Data Warehousing: Consepts, Technologies, Implementations, and Management*. Prentice Hall, USA.

Singh, H. (1998). *Data Warehousing- Concepts, Technologies, Implementations, and Management*. Prentice Hall PTR, ABD.

Serper, Ö. (1997). *Uygulamalı İstatistik*, 2. Baskı, Marmara Kitabevi, Bursa

Sharma, S. (1996). *Applied Multivariate Techniques*, John Wiley & Sons, U.S.A.

SPSS, CRISP-DM 1.0, 22.12.2009, https://www.spss.com/

Statistics and Algorithms. (b.t.), Answer Tree 3.1 User.s Guide

Stiehl, C. *Find Out What the Customer Wants, First*. 14.12.2004. http://www.crmguru.com/articles/2004

Şahin, M., Hamarat, B. (2002). *G-10 Avrupa Birliği ve OECD Ülkelerinin Sosyo-Ekonomik Benzerliklerinin Fuzzy Kümeleme Analizi ile Belirlenmesi*. International Conference in Economics VI, ODTÜ, Ankara.

Şimşek, T. (2006). *Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama*. Yayınlanmamış Doktora Tezi, İstanbul Üniversitesi.

Tan P.T., Steinbach, M. Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley, Boston

Tatlıdil H, (1992) *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Hacettepe Üniversitesi, Ankara

Tekin, M., Çiçek, E. (2003). Değişim Yönetimi Sürecinde Müşteri İlişkileri Yönetimi ve Önemi. III. Ulusal Üretim Araştırmaları Sempozyumu.

Tezcanlar. P, (2007), *Müşteri İlişkileri Yönetimi, Veri Madenciliği ve Bir Uygulama.* Yayınlanmamış Yüksek Lisans Tezi, İstanbul Üniversitesi

Theodoratos, D., Sellis, T. (1999). *Designing Data Warehouses,* Data&Knowledge Engineering, Volume 31, Issue 3.

The Learning Process. (b.t.). 25.12.2009, http://fbim.fhregensburg.de/~saj39122/jfroehl/diplom/e-13-text.html

Tok, A. (2002). *Müşteri İlişkileri Yönetimi ve Veri Madenciliği*. Yayınlanmamış Yüksek Lisans Tezi, Uludağ Üniversitesi

Türkiye'nin CRM Tarifleri. (2001) . CRM Institute Turkey Konferans Notları

Türkmen, E. (b.t.). *OLAP Nedir?,* 07.08.2005, http://www.danismend.com/konular/bilgiveteknoyon/bilgi_olap1.htm

Usgurlu.Ü, b.t. *Veri Tabanı, Veri Ambarı, Veri Madenciliği, Veri Pazarı*, 22.12.2009, http://mail.baskent.edu.tr/~20394676/0302/bil483/HW2.pdf

Veri Amban Nedir? .(b.t.). 21.12.2009. http//www.proyltd.com/html/veriambari.htm

Veri Madenciliği.(b.t.). 05.09.2004, , http:// www.interasystems.com/iq3.asp

Veri Tabanı. (b.t.). metinler. 15.11.2009, http://www.baskent.edu.tr/~eminec/bahar/veri_word.doc

Veri Tabanı (Database) Nedir? Anlamı?. (b.t.). metinler. 18.11.2009, http://www.bilgipasaji.com/forum/hosting-391/367735-veri-tabani-database-nedir-anlami.html

Veri Tabanı Modelleri. (b.t.). metinler. 18.11.2009, http://www.netogretim.com/dokumangoster.aspx?id=177&d=Veri-Taban%C4%B1-Modelleri

Vishwanathan S. V. N., Murty M. N. (2000). *Kohonen's SOM with cache*, Pattern Recognition, 33(11):1927–1929

Yarımağan, Ü. (2000). *Veri Tabanı Sistemleri*. Akademi & Türkiye Bilişim Vakfı , Ankara.

Yıldız, T. (26.12.2005). *Veri Ambarı ve Veri Madenciliği*. 18.11.2009, http://e-learning.bahcesehir.edu.tr/coursecontent

Yapay Sinir Ağları ve Uygulamaları. (b.t.). 25.12.2009, http://mail.baskent.edu.tr/~20293638/som/ppt/sunu.ppt

Yapay Sinir Ağları (Artificial Neural Networks). (b.t.). 03.01.2010. http://www.yapay-zeka.org/modules/wiwimod/index.php?page=ANN&back=SOM

Zaki, M.J., Ogihara, M. (June **1998**). *Theorical foundations of association rules*. 3rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discover (DMKD), Seattle, WA.

Zengyou, H. *Mining Class Outliers: Concepts, Algorithms and Applications in CRM*. Expert Systems with Applications, Volume 27, Issue 4, 2004, ss.681-697

ÖZGEÇMİŞ

24 Haziran 1977 tarihinde Trabzon'da doğdum. İlk, orta ve lise öğrenimi Trabzon ilinde bitirdikten sonra lisans eğitimimi de Karadeniz Teknik Üniversitesi Elektrik/Elektronik Fakültesinde tamamladım. 2003 yılından beri bir telekominikasyon firmasında ça-lışmaktayım. Çeşitli birimlerde görev aldıktan sonra şu anda Veri Ambarı ve İş Zekası Müdürlüğü'nde görevime devam ediyorum.

Yabancı dilim İngilizce olup, evli ve bir çocuk annesiyim.

Aday : Emel SEYMEN TURAN