**İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ**

# LOJİSTİK ALANINDA BİR VERİ MADENCİLİĞİ UYGULAMASI

**YÜKSEK LİSANS TEZİ**
**Mat. Müh. Sevcan TİRYAKİ**
**(509011206)**

**Tezin Enstitüye Verildiği Tarih : 8 Mayıs 2006**
**Tezin Savunulduğu Tarih : 13 Haziran 2006**

**Tez Danışmanı : Doç.Dr. Ali ERCENGiZ**
**Diğer Jüri Üyeleri Doç.Dr. Fikret BALTA**
**Doç.Dr. Metin Orhan KAYA**

**HAZİRAN 2006**

## ÖNSÖZ

Bu çalışmayı hazırlarken yaptığı değerli katkılardan ve manevi desteğinden ötürü Sayın Hocam Yrd. Doç. Dr. Ali ERCENGİZ'e, veri madenciliği konusu ile ilgilenmeme aracı olan Sayın Prof. Dr. Gazanfer ÜNAL'a, bana olan güven ve sevgilerini her zaman yanımda hissettiğim sevgili aileme çok teşekkür ederim.

Mayıs, 2006 Sevcan TİRYAKİ

**İÇİNDEKİLER**

## KISALTMALAR

**COM** : Component Object Model
**ÇAİ** : Çevrimiçi Analitik İşleme
**DVM** : Destek Vektör Makineleri
**OLAP** : Online Analytical Processing
**VA** : Veri Ambarı
**VM** : Veri Madenciliği
**VT** : Veri Tabanı
**VTBK** : Veri Tabanlarında Bilgi Keşfi

## TABLO LİSTESİ

## ŞEKİL LİSTESİ

## LOJİSTİK ALANINDA BİR VERİ MADENCİLİĞİ UYGULAMASI

## ÖZET

Günümüzde bilgisayar sistemleri her geçen gün ucuzlamakta ve aynı zamanda güçleri de artmaktadır. Bilgisayar sistemlerindeki bu gelişmeyle birlikte kullanımı da büyük ölçüde yaygınlaşmaktadır. Bu gelişmeyle birlikte işletmelerde üretilen sayısal bilgi miktarının arttığını buna paralel veri tabanlarının daha fazla veriyi saklayabilecek boyutlara ulaştığını ve bilgisayar sistemlerindeki gelişme ile veriye ulaşmanın kolaylaştığını görmekteyiz. Bu sayede doğru ve daha detaylı bilgiye ulaşmamız mümkün hale gelmekte fakat bu durum başka bir sorunu ortaya çıkarmaktadır. Bu sorun, oluşan bu büyük sayısal veri yığınlarının yönetilmesi ve anlamlı hale getirilmesi sorunudur.

Şirketlerin bilgi sitemleri üzerinden ürettiği bilgi miktarının büyük artış gösterdiğini ve firmaların veri tabanlarının boyutlarının 1 milyon gigabyte (GB) ulaştığını görmekteyiz. İşte veri tabanlarında ki bu teknolojik gelişme ve hacimlerindeki bu olağanüstü artış, firmaları elde toplanan bu verilerden nasıl faydanılacağı ve bu verilerin nasıl anlamlı hale getirileceği sorunuyla karşı karşıya bırakmıştır.

Bilgisayar sistemleri ile üretilen bu veriler tek başlarına değersizdirler çünkü çıplak gözle bakıldığında verilerin bir anlam ifade etmediğini söyleyebiliriz. Bu veriler belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmektedir. İşte ham veriyi bilgiye veya anlamlı hale dönüştürme işini veri madenciliği ile yapabiliriz.

Bu çalışmanın ilk bölümünde veri madenciliği ile ilgili genel bilgilendirme yapılmakta, ikinci bölümünde de veri madenciliği algoritmalarından sınıflandırma metodu kullanılarak bir lojistik firmasının verileri üzerinde ele alınan bir problem veri madenciliği adımları tek tek ele alınarak incelenmektedir.

## A DATA MINING APPLICATION ON LOGISTICS AREA


## SUMMARY

In our time, computer systems are getting cheaper and getting stronger at the same time. With this improvement in computer systems, the use of them is becoming widespread. Besides this, it is seen that the amount of numerical data produced by the companies is increasing, the databases are now capable of hiding much more data and it is getting easier to reach the data. Now it is possible to reach at the right and more detailed information but this causes another problem. This problem is how to manage this huge data mountains and make them meaningful.

We see that the amount of the data that the companies produced by the computer systems is increasing rapidly and the companies' databases' dimensions is now almost 1 gigabyte (GB). The technological improvement in databases and the extraordinary increase in their dimensions make the companies to be face to face with the problem of how to get benefit from these datas and how to make these datas valuable.

The datas that are produced by the computer systems are worthless alone because we can say that they mean nothing when we look at them with the naked eye. This datas get valuable when they are processed with a clear aim. So we can manage transforming raw data to valuable data with data mining.

The first section of this study contains general information about what data mining is. In the application section, a logistic problem on a logistics company's datas is examined detailed explaining the data mining steps using classification algorithm.

# 1. GİRİŞ

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

Başka bir deyişle, veri madenciliği, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir.

Temel olarak veri madenciliği, veri setleri arasındaki desenlerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir [10].

Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Veri madenciliğinde vurgulanan unsurlar istatistiğin tanımı içerisinde zaten yer almaktadır. İstatistik, verilerin toplanması, sınıflandırılması, özetlenmesi, grafik ve tablolarla sunulması, analiz edilerek ana kütle hakkında anlamlı bilgilerin elde edilmesi ve yorumlar yapılmasıdır. Veri madenciliğinde ulaşılmak istenen amaç aslında istatistik biliminin amacı ile aynı doğrultudadır: Verilerden bilgiyi keşfetmek. Zaten veri madenciliğinde kullanılan temel aracın istatistiksel yöntemler olduğu birçok tanımda ve uygulamada vurgulanmaktadır. Her ikisinde de temel olan öğeler, veri ve bilgidir. Bu nedenle birbiriyle oldukça örtüşen konulardır [6]. Ancak veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır. Bu bağlamda, veri madenciliği insan merkezlidir ve bazen insan – bilgisayar arayüzü birleştirilir.

Veri madenciliği; önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak da tanımlanabilir [4]

Gartner Grup tarafından yapılan başka bir tanımda ise veri madenciliği, istatistik ve matematik tekniklerle birlikte ilişki tanıma teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni ilişki ve eğilimlerin keşfedilmesi süreci olarak tanımlanmıştır.

Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği; analistin'e, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir.

## 2. VERİ TABANINDA BİLGİ KEŞFİ SÜRECİ

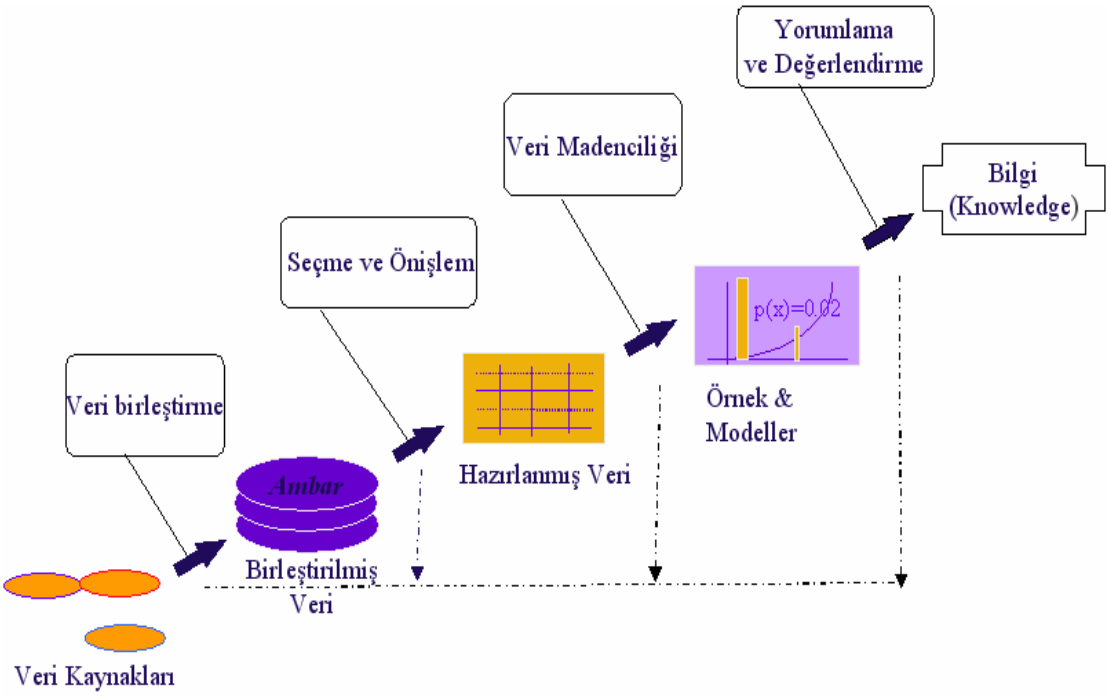
VT'lerde tutulan büyük miktarlardaki verinin VM teknikleriyle işlenmesine veri tabanında bilgi keşfi denir (VTBK). Büyük hacimli olan ve genelde veri ambarlarında tutulan verilerin işlenmesi yeni kuşak araç ve tekniklerle mümkün olabilmektedir. Bundan dolayı bu konularda yapılan çalışmalar güncelliğini korumaktadır. Bazı kaynaklara göre; VTBK daha geniş bir disiplin olarak görülmektedir ve VM terimi sadece bilgi keşfi (BK) metotlarıyla uğraşan VTBK sürecinde yer alan bir adımdır. Şekil 2.1'de de görüldüğü gibi VTBK sürecinde yer alan adımlar şu şekilde sıralanmaktadır:

1. Veri Seçimi: Bu adım veri kümelerinden sorguya uygun verilerin seçilmesidir. Elde edilen verilere örneklem kümesi denmektedir.

2. Veri Temizleme ve Ön İşleme: Örneklem kümesi elde edildikten sonra, örneklem kümesinde yer alan hatalı tutanakların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır. Bu aşama, seçilen veri madenciliği sorgusunun çalışma zamanını iyileştirir.

3. Veri Madenciliği: Veri temizleme ve ön işlemden geçen örneklem kümesine VM sorgusunun uygulanmasıdır. Örnek VM sorguları: kümeleme, sınıflandırma, ilişkilendirme vb. sorgulardır.

4. Yorumlama: VM sorgularından ortaya çıkan sonuçların yorumlanma kesimidir. Burada geçerlilik, yenilik, yararlılık ve basitlik açılarından üretilen sonuçlar yorumlanır.




**Şekil 2. 1.** Veri Tabanlarında Bilgi Keşfi Süreci

## 2.1. VTBK İle Diğer Disiplinler Arasındaki İlişki

### 2.1.1. VTBK ile Makine Öğrenimi Arasındaki İlişki

Makine öğrenimi gözlem ve deneye dayalı ampirik kuralların otomatik biçimde bulunması olan VTBK sistemleri ile yakından ilgilidir. Genel olarak makine öğrenimi ve örüntü tanıma alanlarında yapılan çalışmaların sonuçları VTBK'de veri modelleme ve örüntü çıkarmak için kullanılmaktadır. Bu çalışmalardan bazıları örneklerden öğrenme, düzenli örüntülerin keşfi, gürültülü ve eksik veri ve eksik belirsizlik yönetimi olarak sayılabilir.

VTBK'nın makine öğreniminden en büyük farkı aşağıda sıralanmıştır:

- VTBK büyük veri kümeleriyle çalışabilir,
- VTBK gerçek dünya verileriyle uğraşır.

Veri görselleştirmede kullanılan yöntemler, VTBK sistemi ile elde edilen örüntülerin, kullanıcıya grafikler aracılıyla sunumunu sağlar.

### 2.1.2. VTBK ile İstatistik Arasındaki İlişki

İstatistik ile VTBK arasındaki ilişkinin ana sebebi veri modelleme ve verideki gürültüyü azaltmadan kaynaklanmaktadır. İstatistiğin VTBK'de kullanılan tekniklerinden bazıları aşağıda sıralanmıştır:

- Özellik seçimi,
- Veri bağımlılığı,
- Tanıma dayalı nesnelerin sınıflandırılması,
- Veri özeti,
- Eksik değerlerin tahmini,
- Sürekli değerlerin ayrımı

### 2.1.3. VM ile Veri Tabanı Arasındaki İlişki

VM sorgularına girdi sağlamak amacıyla VT kullanılmaktadır. VT'deki sorgu cümlecikleri VM'nin istediği örneklem kümesini elde etmek amacıyla kullanılmaktadır. Özellikle ilişkilendirme sorgusunda fazla miktarda VT sorgusu yapmak gerekmektedir.

VM, VT'den farklıdır, çünkü VT'de var olan örüntüler için sorgular çalıştırılırken, VM'deki sorgular genelde keşfe dayalı ve ortada olmayan örüntüleri keşfetmeye dayalıdır.

## 3. VERİ MADENCİLİĞİNE GENEL BAKIS

VM yaklaşımı ortaya çıkmadan önce, büyük veri tabanlarından faydalı örüntüler elde etmek için, çevrim-dışı veri üzerinde çalışan istatistiksel paketler kullanılırdı. İstatistiksel yaklaşımların kullanımında bu paketlerin dezavantajları ortaya çıkmaktaydı. Bu dezavantajlardan en önemlisi; istenen verilerin toplanmasından ve amacın belirlenerek istatistiksel yaklaşımların uygulanmasından sonra bir uzman tarafından değerlendirilmesi gerekliliğidir. Başka bir dezavantajı ise her farklı ihtiyaç için bu işlemlerin tekrarlanmasıdır. Bu sorun VTBK'de kısmen aşılmıştır. VTBK, çok büyük hacimli verilerden anlamlı ilişkileri otomatik keşfeder [2].

Araştırmacıların, geniş, çok hacimli ve dağınık veri setleri üzerinde yapmış oldukları çalışmalar sonucu aşağıdaki sonuçlara varılmıştır.

- Veri madenciliği ve bilgi keşfi (data mining & knowledge discovery), özellikle elektronik ticaret, bilim, tıp, iş ve eğitim alanlarındaki uygulamalarda yeni ve temel bir araştırma sahası olarak ortaya çıkmaya başlamıştır. Veri madenciliği, eldeki yapısız veriden, anlamlı ve kullanışlı bilgiyi çıkarmaya yarayacak tümevarım işlemlerini formüle analiz etmeye ve uygulamaya yönelik çalışmaların bütününü içerir. Geniş veri kümelerinden desenleri, değişiklikleri, düzensizlikleri ve ilişkileri çıkarmakta kullanılır. Bu sayede, web üzerinde filtrelemeler, DNA sıraları içerisinde genlerin tespiti, ekonomideki eğilim ve düzensizliklerin tespiti, elektronik alışveriş yapan müşterilerin alışkanlıkları gibi karar verme mekanizmaları için önemli bulgular elde edilebilir.

- Sayısal verinin miktarı, son 10 yılda bir patlama yaşayarak tahminlerin dışında bir artış göstermiştir. Buna karşılık, bilim adamlarının, mühendislerin ve analistlerin sayısı değişmemektedir. Bu orantısızlığı gidermek için yeni araştırma problemlerinin çözümleri birkaç gruba ayrılabilir :

1. Geniş hacimli ve çok boyutlu veri madenciliği için yeni algoritma ve sistemlerin geliştirilmesi,

2. Yeni veri tiplerinin madenciliği için yeni algoritma, teknik ve sistemlerin geliştirilmesi,

3. Dağıtık veri madenciliği için algoritma, protokol ve altyapıların geliştirilmesi,

4. Mevcut veri madenciliği sistemlerinin kullanımının ilerletilip geliştirilmesi,

5. Veri madenciliği için özel gizlilik ve güvenlik modellerinin geliştirilmesi.

- Tüm bu uğraşların başarıya ulaşması ve sonuç verebilmesi için hükümetin ve çok disiplinli ve disiplinler arası çalışan iş sahalarının desteği gereklidir.
- İlgili sistemlerin, ölçülmüş altyapıların ve test ortamlarının oluşturulmasını gerektiren önemli deneysel bileşenlerin gerçekleştirilmesi gerekir.

# 4. VERİ MADENCİLİĞİNİN KULLANIM AMACI VE KULLANIM ALANLARI

## 4.1. Veri Madenciliğinin Kullanım Amaçları

İstatistiğin amacı nasıl ana kütle hakkında anlamlı bilgiler elde etmek ve yorum yapmaksa veri madenciliğinin amacı da anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır [6]. Buradaki temel amaç, değişkenler arasındaki ilişkilerden çok, geleceğe yönelik sağlıklı öngörülerin üretilmesidir. Bu anlamda VM, özbilginin keşfedilmesi anlamında bir "kara kutu" bulma yaklaşımı olarak kabul edilmektedir ve bu doğrultuda yalnızca keşifsel veri analizi tekniklerini değil, sinir ağı tekniklerinden hareketle geçerli öngörüler yapmak ve öngörülen değişkenler arasındaki ilişkilerin belirlenmesi mümkün olduğu için aynı zamanda sinir ağı tekniklerini de kullanmaktadır [9].

Yöntemin işletmelerde kullanımı sonucunda sağlanabilecek faydalar aşağıdaki gibi özetlenebilir:

- Bir işletme kendi müşterisiyken rakibine giden müşterilerle ilgili analizler yaparak rakiplerini tercih eden müşterilerinin özelliklerini elde edebilir ve buradan hareketle gelecek dönemlerde kaybetme olasılığı olan müşterilerin kimler olabileceği yolunda tahminlerde bulunarak onları kaybetmemek, kaybettiklerini geri kazanmak için farklı stratejiler geliştirebilir.
- Mevcut müşterilerin işletme tarafından daha iyi tanınmasını sağlayabilir. Özellikle finans sektöründe mevcut müşterilerinin segmentlere ayrılarak çıkarılacak kredi risk davranış modellerinin yeni başvuruda bulunan müşterilere uygulanmasını sağlayarak riski minimize edebilir. Bir anlamda kredi risk skorlamasının altyapısının oluşturulmasında kullanılabilir.
- Mevcut müşterilerin ödeme performansları incelenerek kötü ödeme performansı gösteren müşterilerin ortak özellikleri belirlenerek, benzer

- özelliklere sahip tüm müşteriler için yeni risk yönetim politikaları oluşturulabilir.
- En karlı mevcut müşteriler belirlenerek, potansiyel müşteriler arasından en karlı olabilecekler belirlenebilir. Karlı müşteriler tespit edilerek onlara özel kampanyalar uygulanabilir. En masraflı müşteriler daha masrafsız müşteri haline dönüştürülebilir. Örneğin en çok bankacılık işlemi yapanlar ortaya çıkarılıp bunlar şube bankacılığı yerine daha masrafsız internet bankacılığına yönlendirilebilir.
- Mevcut müşteriyi tanıyarak işletmelerin müşteri ilişkileri yönetimlerinde düzenleme ve geliştirmeler yapılabilir. Bu sayede firmanın müşterilerini daha iyi tanıyarak müşteri gibi düşünme kapasitelerinin arttırılması sağlanabilir. Bunun da işletmelere pazarda avantaj sağlayacağı unutulmamalıdır.
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminlerde bulunulabilir. Özellikle ciro, karlılık, pazar payı gibi analizlerde veri madenciliği çok rahat kullanılabilir.
- Mevcut müşteriler üzerinde firma ürünlerinin çapraz satış kapasitesinin arttırılması sağlanabilir. Mesela firmanın X ürününü alan müşterilerin çok büyük bir bölümünün Y ürününü de aldıkları biliniyorsa, buna yönelik pazarlama stratejileri geliştirilebilir.
- Piyasada oluşabilecek değişikliklere mevcut müşteri portföyünün vereceği tepkinin firma üzerinde oluşturabileceği etkinin tespitinde kullanılabilir.
- Operasyonel süreçte oluşabilecek olası kayıpların veya suistimallerin tespitinde kullanılabilir.
- Kurum teknik kaynaklarının en optimal şekilde kullanılmasını sağlamakta kullanılabilir.
- Firmanın finansal yapısının, makro ekonomik değişmeler karşısındaki duyarlılığı ve oluşabilecek risklerin tespitinde kullanabilir.
- Günümüzde var olan yoğun rekabet ortamında firmaların hızlı ve kendisi için en doğru kararı almalarını sağlayabilir.

## 4.2. Veri Madenciliğinin Kullanım Alanları

Ülkemizde son yıllarda yeni yeni tanınmaya başlayan VM kavramının, Avrupa ve Kuzey Amerika ülkelerinde birbirinden çok farklı alanlarda kullanıldığı görülmektedir [5]. Pazarlama ve satış alanında, hedef pazarların tespitinde, müşteri ilişkilerinin yönetiminde, sepet analizinde, çapraz satışlarda, pazar segmentasyonlarında ve müşteri hatırlamada sık sık veri madenciliğinden yararlanılmaktadır. Veri kaynaklarını işlemek için müşteri kartı bilgilerinin kaydedilmesinde, müşteri şikayetlerinin incelenmesinde, e-ticarette oldukça büyük işlevlere sahiptir. Diğer taraftan satış kampanyalarının, verimlilik analizlerinin yapılması, reklamcılık, indirim kartları ve bonuslandırmaları, karlılığın artırılması gibi daha bir çok kullanım alanı bulunmaktadır.

Sayılan bu kullanım alanlarının yanında, astronomi, biyoloji, finans, sigorta, tıp gibi bir çok başka alanda da uygulanmaktadır. Son 20 yıldır Amerika Birleşik Devletleri'nde çeşitli veri madenciliği algoritmalarının gizli dinlemeden, vergi kaçakçılıklarının ortaya çıkarılmasına kadar çeşitli uygulamalarda da kullanıldığı görülmektedir [3].

Özellikle, son yıllarda, risk analizi ve yönetiminde de, doğru ve etkin kredi kararı verebilme, kredi geri ödemesi yapmamaya meyilli müşterileri belirleme, risk derecelendirme, finansal işlemlerde sahtekarlığa yönelik eğilimleri izleme, ekonomik ve finansal yatırımları kararlaştırma, iflas / başarısızlık tahmini gibi alanlarda da yaygın olarak kullanılmaya başlanmıştır [8].

Görüldüğü gibi veri madenciliği teknikleri çok çeşitli alanlarda kullanılmaktadır. Bu uygulama alanları ana başlıklar altında aşağıdaki gibi özetlenebilir:

Pazarlama

- Müşteri segmentasyonunda,
- Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında,
- Çeşitli pazarlama kampanyalarında,
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında,

- Pazar sepeti analizinde,
- Çapraz satış analizleri,
- Müşteri değerleme,
- Müşteri ilişkileri yönetiminde,
- Çeşitli müşteri analizlerinde,
- Satış tahminlerinde,

Bankacılık

- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,
- Kredi kartı dolandırıcılıklarının tespitinde,
- Müşteri segmentasyonunda,
- Kredi taleplerinin değerlendirilmesinde,
- Usulsüzlük tespiti,
- Risk analizleri,
- Risk yönetimi,

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesinde,
- Sigorta dolandırıcılıklarının tespitinde,
- Riskli müşteri tipinin belirlenmesinde.

Perakendecilik

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

Borsa

- Hisse senedi fiyat tahmini,
- Genel piyasa analizleri,

- Alım-satım stratejilerinin optimizasyonu.

Telekomünikasyon

- Kalite ve iyileştirme analizlerinde,
- Hisse tespitlerinde,
- Hatların yoğunluk tahminlerinde,

Sağlık ve İlaç

- Test sonuçlarının tahmini,
- Ürün geliştirme,
- Tıbbi teşhis
- Tedavi sürecinin belirlenmesinde

Endüstri

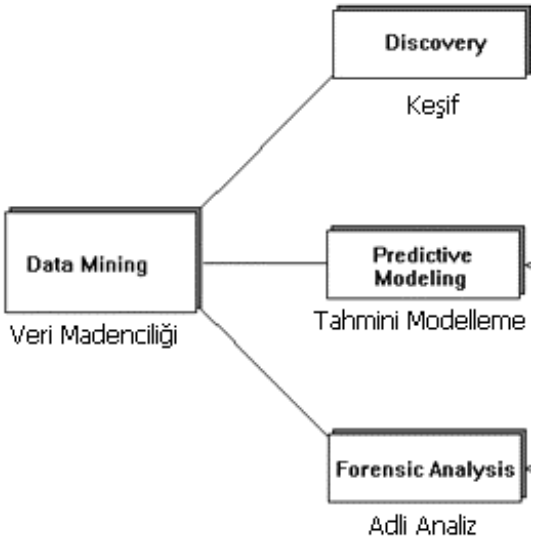
- Kalite kontrol analizlerinde
- Lojistik,
- Üretim süreçlerinin optimizasyonunda,

Bilim ve Mühendislik

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

## 5. VERİ MADENCİLİĞİNİN İŞLEVLERİ

Veri madenciliğine işlevleri açısından bakılacak olursa, veri madenciliği aktiviteleri Şekil 5.1'de gösterildiği gibi 3 sınıf altında toplanmıştır. : Keşif (discovery), tahmini modelleme (predictive modeling) ve adli analiz (forensic analysis).




**Şekil 5.1** : Veri Madenciliği Aktiviteleri

Keşif, ne olabileceği konusunda önceden belirlenmiş bir fikir ya da hipotez olmadan, veri tabanı içerisinde gizli desenleri arama işlemidir. Geniş veri tabanlarında kullanıcının pratik olarak aklına gelmeyecek ve bulmak için gerekli doğru soruları bile düşünemeyeceği birçok gizli desen olabilir. Buradaki asıl amaç, bulunacak desenlerin zenginliği ve bunlardan çıkarılacak bilginin kalitesidir.

Basit bir örnek vermek gerekirse, bir ülkenin nüfus kayıtlarını düşünelim. Kullanıcı, eldeki bu veri tabanına "Bankacıların yaş ortalaması nedir?" şeklinde bir ilk soru sorabilir. Sistemin bu soruya 47 olarak cevap verdiğini varsayalım.

Kullanıcı, artık "yaş"la ilgili daha ilginç veriler bulma yoluna gidebilir. Sistem, bu andan itibaren, bir analist gibi hareket edecek ve kurallar çıkarmaya çalışacaktır. Örneğin "Eğer Meslek=Sporcu ise, Yaşı %71 kesinlikle 30'dan küçüktür." kuralının anlamı, eğer veri tabanından 100 adet sporcu seçilirse, bunların 71 adedinin yaşı, 30'dan küçüktür demektir. Benzer olarak sistem, "Eğer Meslek=Sporcu ise, Yaşı %97 kesinlikle 60'dan küçüktür" sonucunu da çıkarabilir. Bu da 100 sporcudan en az 97 sinin 60 yaşından küçük olduğunu belirtir.

Tahmini modellemede, veri tabanından çıkarılan desenler, geleceği tahmin için kullanılır. Bu model, kullanıcının bazı alan bilgilerini bilmese bile kayıt etmesine izin verir. Sistem, bu boşlukları, önceki kayıtlara bakarak tahmin yoluyla doldurur. Keşif, verideki desenleri bulmaya yönelikken, tahmini modelleme, bu desenleri yeni veri nesnelerini bulmak için uygular.

Az önceki örneği baz alırsak, artık mesleği sporcu olan birinin yaşını yaklaşık olarak tahmin edebilmekteyiz. Kayıtlar arasında yaşı bilinmeyen fakat mesleği sporcu olan birini bize söylediklerinde, yaşının %71 kesinlik oranıyla 30'dan küçük, hatta %97 kesinlikle de 60'dan küçük olduğunu tahmin edebiliriz. Burada keşif, genel bilgiyi bulmamıza yardımcı olur ama tahmini modelleme, daha spesifik bilgileri tahmin etmekte kullanılır.

Adli analiz, normal olmayan ya da sıra dışı veri elemanlarını bulmak için, çıkarılmış desenleri uygulama işlemidir. Sıra dışı olanı bulmak için ilk önce sıradan kısmı tespit etmek gerekir. Örneğimize göre 60 yaşından sonra hala spor yapan %3'lük bir kesimin olduğunu biliyoruz ancak sebebini bilmiyoruz. Bunlar sıra dışı eleman olarak kabul edilmektedirler. Kimisi normalin dışında sağlıklı olabilir ya da yaş ile ilgisi olmayan sporlarla (örneğin golf) uğraşıyor olabilirler. Ya da bu veri tabanındaki bilginin yanlış olabileceğini de gösteriyor olabilir. Göründüğü gibi adli analiz, keşifte aranılan genel bilginin tersine, sıra dışı ve özel durumları araştırır [10].

### 5.1. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modelleri, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında toplayabiliriz.

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.

Tanımlayıcı modellerde; ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. 25 yaş altı bekar kişiler ile, 25 yaş üstü evli kişiler üzerinde yapılan ve ödeme performanslarını gösteren bir analiz tanımlayıcı modellere örnek olarak verilebilir.

# 6. VERİ MADENCİLİĞİ ALGORİTMALARI

VM süreci sonunda elde edilen örüntüler kurallar biçiminde ifade edilir. Elde edilen kurallar, (1) koşul yan tümcesi ile sonuç arasındaki eşleştirme derecesini gösterir (if <koşul tümcesi>, then <sonuç>, derece (0..1)), (2) veriyi önceden tanımlanmış sınıflara bölümler (partition); veya (3) veriyi bir takım kriterlere göre sonlu sayıda kümeye ayırır. Bu kurallar veri üzerinde belirli bir tekniğin (algoritmanın) sonlu sayıda yinelenmesiyle elde edilir. Elde edilen bilginin kalitesi veri analizi için kullanılan algoritmaya büyük ölçüde bağlıdır [7].

## 6.1. Hipotez Testi Sorgusu

Hipotez testi sorgusu algoritması, doğrulamaya dayalı bir algoritmadır. Bir hipotez öne sürülür ve seçilen veri kümesinde hipotez doğruluğu test edilir. Öne sürülen hipotez genellikle belirli bir örüntünün veri tabanındaki varlığıyla ilgili bir tahmindir. Bu tip bir analiz özellikle keşfedilmiş bilginin genişletilmesi veya damıtılması (refine) işlemleri sırasında yararlıdır.

Hipotez ya mantıksal bir kural ya da mantıksal bir ifade ile gösterilir. Her iki biçimde de seçilen veri tabanındaki nitelik alanları kullanılır. X ve Y birer mantıksal ifade olmak üzere “IF X THEN Y” biçiminde bir hipotez öne sürülebilir.

Verilen hipotez, seçilen veri tabanında doğruluk ve destek kıstasları baz alınarak sistem tarafından sınanır.

## 6.2. Sınıflama Sorgusu

Sınıflama sorgusu yeni bir veri elemanını daha önceden belirlenmiş sınıflara atamayı amaçlar. Veri tabanında yer alan çoklular, bir sınıflama fonksiyonu yardımıyla kullanıcı tarafından belirlenir veya karar niteliğinin bazı değerlerine göre anlamlı ayrık alt sınıflara ayrılır. Bu yüzden sınıflama, denetimli öğrenmeye

(supervised learning) girer. Sınıflama algoritması bir sınıfı diğerinden ayıran örüntüleri keşfeder. Sınıflama algoritmaları iki şekilde kullanılır:

• Karar Değişkeni ile Sınıflama: Seçilen bir niteliğin aldığı değerlere göre sınıflama işlemi yapılır. Seçilen nitelik karar değişkeni adını alır ve veri tabanındaki çoklular karar değişkeninin değerlerine göre sınıflara ayrılır. Bir sınıfta yer alan çoklular, karar değişkeninin değeri açısından özdeştir.

• Örnek ile Sınıflama: Bu biçimdeki sınıflamada veri tabanındaki çoklular iki kümeye ayrılır. Kümelerden biri pozitif, diğeri negatif çokluları içerir.

Yaygın kullanım alanları, banka kredisi onaylama işlemi, kredi kartı sahteciliği tespiti ve sigorta risk analizidir.

### 6.3. Kümeleme Modelleri

Kümeleme modellerinde amaç, üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Sınıflandırmaya benzer. Farkı, grupların önceden belirlenmemiş olmasıdır. Temel özellikleri:

- Oluşacak küme sayısı belirsizdir
- Kümeler hakkında bir ön bilgi olmayabilir
- Küme sonuçları dinamiktir

Kümeleme algoritması veritabanını alt kümelere ayırır. Her bir kümede yer alan elemanlar dahil oldukları grubu diğer gruplardan ayıran ortak özelliklere sahiptir. Kümeleme modellerinde amaç, küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir.

Yaygın kullanım alanları nüfusbilimi, astronomi vb.dir.

Kümeleme analizinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı konunun uzmanı olan bir kişi tarafından belirtilebileceği gibi veri tabanındaki kayıtların hangi kümelere ayrılacağını, geliştirilen bilgisayar programları da yapabilmektedir.

## 6.4. Ardışık Örüntüler

Ardışık örüntü keşfi, bir zaman aralığında sıklıkla gerçekleşen olaylar kümelerini bulmayı amaçlar.

- Bir yıl içinde Orhan Pamuk'un "Benim Adım Kırmızı" romanını satın alan insanların %70'i Buket Uzuner' in "Güneş Yiyen Çingene" adlı kitabını satın almıştır.
- X ameliyatı olanlarda, 15 gün içinde % 45 ihtimalle Y enfeksiyonu oluşacaktır,
- İMKB endeksi düşerken A hisse senedinin değeri % 15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri % 60 ihtimalle artacaktır,
- Çekiç satın alan bir müşteri, ilk üç ay içerisinde % 15, bu dönemi izleyen üç ay içerisinde % 10 ihtimalle çivi satın alacaktır.

Bu tip örüntüler perakende satış, telekomünikasyon ve tıp alanlarında yararlıdır.

## 6.5. İstisna (Outlier) Analiz

Bir veri tabanı, tüm veri modelinin davranışını sergilemeyen veriler içeriyor olabilir. Bu tür veriler "outliers" olarak adlandırılırlar. Birçok veri madenciliği tekniği outlier'ları gürültü yada istisna olarak adlandırır. Buna rağmen bazı uygulamalarda örneğin hile tesbitinde (fraud detection), daha seyrek oluşmuş olan olaylar sık oluşmuş olanlara göre daha ilginç ve önemli olabilirler. Outlier verinin analizi, outlier analiz olarak adlandırılır.

Outliers, istatistiksel testler kullanılarak saptanabilir. Örnek olarak kredi kartı sahteciliklerinin tesbitinde kullanılan model verilebilir.

Outlier (istisna) analizinde iki yöntem söz konusudur:

*1. İstatistik Tabanlı Yöntem:*

Dağılım analizi ya da standart sapma hesabı gibi istatistik yöntemlerle istisna olabilecek noktalar tespit edilir. Fakat çok büyük veri yığınlarında yoğun hesaplama gücü gerektirdikleri için performansları sınırlıdır.

*2. Yoğunluk Tabanlı Yöntem:*

Bu yöntemde her noktanın çevresindeki komşuları ile olan yakınlığı hesaplanır. Yakınlık hesaplamada genelde öklit uzaklığı kullanılsa da veri türüne göre yakınlık hesaplama yöntemi farklılık gösterebilir. Bu yöntemin temel prensibi "yeterince komşusu olmayan noktaları" tespit etmektir.

## 6.6. Evrimsel Analiz

Evrimsel analiz, zamanla davranışları değişen nesnelerin düzenlilik ya da eğilimlerini ortaya çıkarmayı amaçlar. Evrimsel analiz, tanımlama, ayrımlama, birliktelik analizi, sınıflama ve kümeleme metodlarını içerse de asıl amacı verinin zaman ile olan ilişkisini ortaya çıkarmaktır. Bunun için zaman serileri ardışıklık ve periyodiklik örüntüsü bulma, benzerlik analizi gibi yöntemleri kullanır.

Evrimsel analiz, birçok kaynakta bağımsız bir kategori olarak yer almamaktadır. Evrimsel analizin kullandığı her bir yöntem evrimsel analiz adı altında değil kendi başına bağımsız bir yöntem olarak kabul görmektedir.

## 6.7. İlişki Analizi

İlişki analizi, belirli bir datasette yüksek sıklıkta birlikte görülen attribute değerlerine ait ilişkisel kuralların keşfidir. Market-Basket analizi ve transaction veri analizinde sıkça kullanılır.

İlişki ya da birliktelik analizi, bir veri kümesinde kendiliğinden, sıklıkla gerçekleşen, birlikte ya da aynı süre içinde alınma, yapılma, oluşma gibi etkileri keşfetme temeline dayanır. Bu yöntem bankacılık işlemlerinin analizinde ya da sepet analizi tekniğinde yaygın olarak kullanılır. Sepet analizi, bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın alma eğiliminde olduğunun belirlenmesiyle müşteriye daha fazla ürün satılması yollarından biridir. Sepet Analizi ile, örneğin müşteriler bira satın aldığında %75 ihtimalle cips de alırlar şeklinde bir ilişki ortaya çıkarılabilir. Bunun sonucunda bira ile cips yan yana raflara yerleştirilebilir veya bira alanlar cips aldığında cips fiyatında indirim yapılacak şekilde kampanyalar oluşturularak satışlar arttırılabilir.

## 6.8. Bellek Tabanlı Yöntemler

Bellek tabanlı veya örnek tabanlı bu yöntemler (memory-based, instance-based methods; case-based reasoning) 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 (k-nearest neighbor).

## 6.9. Yapay Sinir Ağları(YSA)

1980'lerden sonra yaygınlaşan yapay sinir ağlarında (artificial neural networks) amaç fonksiyon, birbirine bağlı basit işlemci ünitelerinden oluşan bir ağ üzerine dağıtılmıştır. Yapay sinir ağlarında kullanılan öğrenme algoritmaları veriden, üniteler arasındaki bağlantı ağırlıklarını hesaplar. YSA istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz yani uygulama alanı daha geniştir ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez.

## 6.10. Karar Ağaçları

İstatistiksel yöntemlerde veya yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçları ise veriden oluşturulduktan sonra ağaç kökten yaprağa doğru inilerek kurallar (IF-THEN rules) yazılabilir. Bu şekilde kural çıkarma (rule extraction), veri madenciliği çalışmasının sonucunun doğrulanmasını sağlar. Bu kurallar uygulama konusunda uzman bir kişiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. Sonradan başka bir teknik kullanılacak bile olsa karar ağacı ile önce bir kısa çalışma yapmak, önemli değişkenler ve yaklaşık kurallar konusunda analiste bilgi verir ve daha sonraki analizler için yol gösterici olabilir.

Etkin bir VM algoritması geliştirebilmek için aşağıdaki hususlara dikkat edilmesi gerekmektedir:

1. Veri gizliliği ve güvenliğinin sağlanması: Bir VTBK sisteminde keşfedilen bilgi pek çok farklı açıdan ve soyutlama düzeyinden izlenebildiği için, gizlilik ve

veri güvenliği, VM sistemini kullanan kullanıcının haklarına ve erişim yetkilerine göre sağlanmalıdır.

2. Sonuçların yararlılık, kesinlik ve anlamlılık kıstaslarını sağlaması: Elde edilen sonuçlar analiz için kullanılan VT'yi doğru biçimde yansıtmalıdır. Bunun yanı sıra gürültülü ve aykırı veriler işlenmelidir. Bu işlem elde edilen kuralların kalitesini belirlemede önemli bir rol oynar.

3. Farklı tipdeki verileri ele alma: Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri üzerinde değil, aynı zamanda tamsayı, kesirli sayı, çoklu ortam verisi ve coğrafi veri gibi farklı tipteki veriler üzerinde de işlem yapılmasını gerektirir. Kullanılan verinin saklandığı ortam, düz bir kütük veya ilişkisel VT'de yer alan tablolar olabileceği gibi, nesneye yönelik VT'ler, çoklu ortam VT'leri ve coğrafi VT'ler vb. de olabilir. Saklandığı ortama göre veri, basit tipte olabileceği gibi karmaşık veri tipleri (çoklu ortam verisi, zaman boyutlu veri, yardımlı metin, coğrafi veri vb.) de olabilir. Bununla birlikte veri tipi çeşitliliğinin fazla olması bir VM algoritmasının tüm veri tiplerini ele alabilmesini olanaksızlaştırmaktadır. Bu yüzden veri tipine özgü adanmış VM algoritmaları geliştirilmektedir.

4. Farklı ortamlarda yer alan veri üzerinde işlem yapabilme: Kurumlar yerel ağlar üzerinden pek çok dağıtık ve heterojen VT üzerinde işlem yapmaktadır. Bu VM'nin farklı kaynaklarda birikmiş biçimli ya da biçimsiz veriler üzerinde analiz yapabilmesini gerektirir. Veri büyüklüğünün yanı sıra verinin dağıtık olması, yeni araştırma alanlarının ortaya çıkmasına sebep olmuştur. Bunlar, koşut ve dağıtık VM algoritmalarıdır.

5. Veri madenciliği algoritmasının etkinliği ve ölçeklenebilirliği: Çok büyük hacimli veri içinden bilgi elde etmek için kullanılan VM algoritmasının etkin ve ölçeklenebilir olması gerekir. Bu, VM algoritmasının çalışma zamanının tahmin edilebilir ve kabul edilebilir bir süre olmasını gerektirir. Üssel veya çok terimli bir karmaşıklığa sahip bir VM algoritmasının uygulanması kullanışlı değildir.

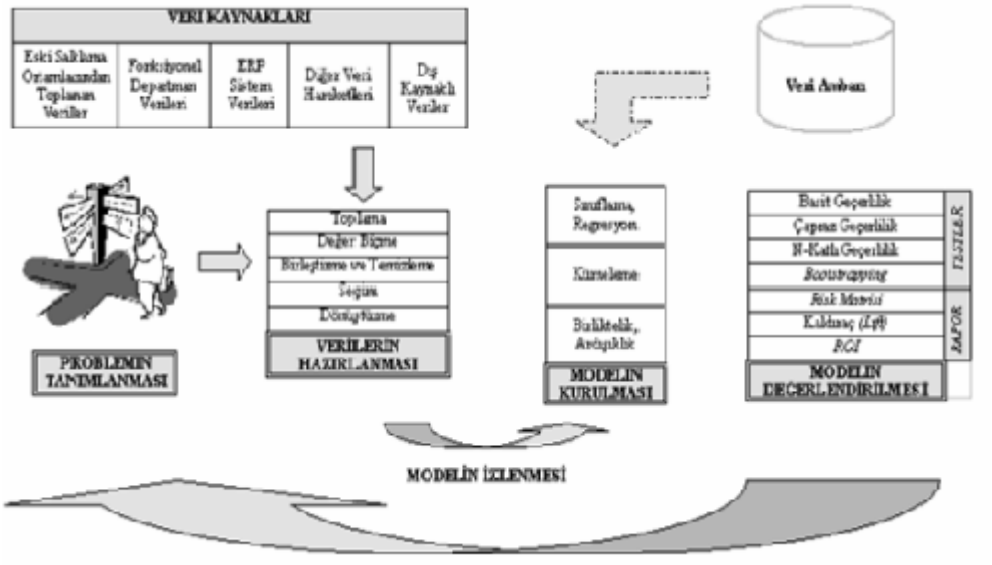
6. Keşfedilen kuralların çeşitli biçimlerde gösterimi: Bu özellik keşfedilen bilginin gösterim biçiminin seçilebilmesini sağlayan yüksek düzeyli bir dil tanımının yapılmasını ve grafik arayüzünü gerektirir.

7. Farklı bir kaç soyutlama düzeyi ve etkileşimli veri madenciliği: Büyük VT'lerden VM sorgularıyla elde edilecek bilginin edinilmesi güçtür. Bu yüzden VM sorgusu, elde edilen bilgilere göre kullanıcıya etkileşimli olarak sorgusunu değiştirebilmeyi, farklı açılardan ve farklı soyutlama düzeylerinden keşfedilen bilgiyi inceleyebilme esnekliğini sağlamalıdır.

## 7. VERİ MADENCİLİĞİ SÜRECİ

Ne kadar etkin olursa olsun hiç bir veri madenciliği algoritmasının üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda fayda sağlaması mümkün değildir. Bu nedenle yukarıda tanımlanan tüm aşamalardan önce, iş ve veri özelliklerinin öğrenilmesi / anlaşılması başarının ilk şartı olacaktır. Başarılı veri madenciliği projelerinde izlenmesi gereken yol Şekil 7.1'de de gösterildiği gibi aşağıdaki gibidir:

1. Problemin Tanımlanması,
2. Verilerin Hazırlanması,
3. Modelin Kurulması ve Değerlendirilmesi,
4. Modelin Kullanılması,
5. Modelin İzlenmesi




**Şekil 7.1.** Veri Madenciliği Süreci

1. Problemin Tanımlanması : Veri madenciliği çalışmalarında başarılı olmanın en önemli şartı, projenin hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.

Bu aşamada mevcut iş probleminin nasıl bir sonuç üretilmesi durumunda çözüleceğinin, üretilecek olan sonucun fayda - maliyet analizinin başka bir değişle üretilen bilginin işletme için değerinin doğru analiz edilmesi gerekmektedir. Analistin işletmede üretilen sayısal verilerin boyutlarını, proje için yeterlilik düzeyinin iyi analiz edilmesi gerekmektedir. Ayrıca analistin işletme konusu hakkındaki iş süreçlerinin de iyi analiz edilmesi gerekmektedir.

2. Verilerin Hazırlanması : Veri madenciliğinin en önemli aşamalarından bir tanesi olan verinin hazırlanması aşaması analistin toplam zaman ve enerjisinin %50 - %75'ini harcamasına neden olmaktadır. Bu aşamada firmanın mevcut bilgi sistemleri üzerinde ürettiği sayısal bilginin iyi analiz edilmesi, veriler ile mevcut iş problemi arasında ilişki olması gerektiği unutulmamalıdır. Proje kapsamında kullanılacak sayısal verilerin, hangi iş süreçleri ile yaratıldığı da bu veriler kullanılmadan analiz edilmelidir, bu sayede analist veri kalitesi hakkında fikir sahibi olabilir.

Verilerin hazırlanması aşaması kendi içerisinde toplama, birleştirme ve temizleme, dönüştürme adımlarından meydana gelmektedir.

i.Toplama: Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, nüfus sayımı, hava durumu, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir.

ii.Birleştirme ve Temizleme : Bu adımda toplanan verilerde bulunan farklılıklar giderilmeye çalışılır. Hatalı veya analizin yanlış yönlenmesine sebep olabilecek verilerin temizlenmesine çalışılır. Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın

gerçekleşmesinden kaynaklanan verilerin, önemli bir uyarıcı enformasyon içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir. Ancak basit yöntemlerle ve baştan savma olarak yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır.

iii.Dönüştürme : Kullanılacak model ve algoritma çerçevesinde verilerin tanımlama veya gösterim şeklinin de değiştirilmesi gerekebilir. Örneğin kredi riski uygulamasında iş tiplerinin, gelir seviyesi ve yaş gibi değişkenlerin kodlanarak gruplanması faydalı olacaktır.

3. Modelin Kurulması ve Değerlendirilmesi :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.

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik (Simple Validation) 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)

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir çok işletme uygulamasında ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Çok ender olarak yorumlanamayacak kadar karmaşıklaşsalar da, genel olarak karar ağacı ve kural temelli sistemler model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir.

4. Modelin Kullanılması : 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 veya tahmin edilen envanter düzeyleri yeniden

sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.

6. Modelin İzlenmesi : 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.

# 8. KRİTİK BAŞARI FAKTÖRLERİ

Verinin Önemi : Veri madenciliğinde amaç çok büyük miktardaki ham veriden değerli bilginin çıkarılmasıdır Çok miktarda güvenilir (hata ve eksiklerin olmadığı) veri önşarttır çünkü çözümün, yani çıkarılan kuralların kalitesi öncelikle verinin kalitesine bağlıdır. Veri madenciliği simya değildir; taşı altına çeviremeyiz.

Uzmanı Önemi : Veri madenciliği çalışması bilgisayarcıların ve uygulama konusundaki uzmanların ortak çalışmasıdır. Her ne kadar olabildiğince otomatik olmasını istesek de uzmanların yardımı ve desteği olmadan başarılı olmak söz konusu değildir. Uzmanlar amacı tanımlar. Uygulama ile ilgili sonuca yararlı olabilecek her tür bilginin sisteme verilmesi gerekir ve bunları da ancak uzmanlar bilir. Ayrıca çalışma ile alınan sonuçların yorumlanması ve geçerlenmesi uzmanlar tarafından yapılır.

Sabrın Önemi : Veri madenciliği tek aşamalı bir çalışma değildir, tekrarlıdır. Sistem ayarlanana dek birçok deneme gerekebilir. Çalışma uzun olabilir. Buna çalışan ekibin ve yönetimin hazırlıklı olması, kısa vadede çok büyük beklentilere sahip olunmaması gerekir [1].

## 8.1. Veri Madenciliğindeki Problemler

Küçük veri kümelerinde hızlı ve doğru bir biçimde çalışan bir sistem, çok büyük veri tabanlarına uygulandığında tamamen farklı davranabilir. Bir VM sistemi tutarlı veri üzerinde mükemmel çalışırken, aynı veri grubuna hatalı veri eklendiğinde kayda değer bir biçimde kötüleşebilir.

Veri madenciliği girdi olarak ham veriyi sağlamak üzere veri tabanlarına dayanır. Bu da veri tabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda sorunlar doğurur. Diğer sorunlar da verinin konu ile uyumsuzluğundan doğabilir.

Sınıflandırmak gerekirse başlıca sorunlar şunlardır :

- Sınırlı Bilgi : Veri tabanları genel olarak veri madenciliği dışındaki amaçlar için tasarlanmışlardır. Bu yüzden, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir.
- Veri Tabanı Boyutu : Veri tabanı boyutları inanılmaz bir hızla artmaktadır. Pek çok makine öğrenimi algoritması birkaç yüz tutanaklık oldukça küçük örneklemleri ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüzbinlerce kat büyük örneklemlerde kullanılabilmesi için azami dikkat gerekmektedir. Örneklemin büyük olması, örüntülerin gerçekten var olduğunu göstermesi açısından bir avantajdır ancak böyle bir örneklemden elde edilebilecek olası örüntü sayısı da çok büyüktür. Bu yüzden VM sistemlerinin karşı karşıya olduğu en önemli sorunlardan biri veri tabanı boyutunun çok büyük olmasıdır. Dolayısıyla VM yöntemleri ya sezgisel bir yaklaşımla arama uzayını taramalıdır, ya da örneklemi yatay/dikey olarak indirgemelidir.
- Gürültülü Veri : Büyük veri tabanlarında pek çok niteliğin değeri yanlış olabilir. Bu hata, veri girişi sırasında yapılan insan hataları veya girilen değerin yanlış ölçülmesinden kaynaklanır. Veri girişi veya veri toplanması sırasında oluşan sistem dışı hatalara gürültü adı verilir. Günümüzde kullanılan ticari ilişkisel veri tabanları, veri girişi sırasında oluşan hataları otomatik biçimde gidermek konusunda az bir destek sağlamaktadır. Hatalı veri, gerçek dünya veri tabanlarında ciddi problem oluşturabilir. Bu durum, bir VM yönteminin kullanılan veri kümesinde bulunan gürültülü verilere karşı daha az duyarlı olmasını gerektirir. Gürültülü verinin yol açtığı problemler tümevarımsal karar ağaçlarında uygulanan metodlar bağlamında kapsamlı bir biçimde araştırılmıştır. Eğer veri kümesi gürültülü ise sistem bozuk veriyi tanımalı ve ihmal etmelidir.
- Boş Değerler : Bir veri tabanında boş değer, birincil anahtarda yer almayan herhangi bir niteliğin değeri olabilir. Boş değer, tanımı gereği kendisi de dahil olmak üzere hiç bir değere eşit olmayan değerdir. Bir çokluda eğer bir nitelik değeri boş ise o nitelik bilinmeyen ve uygulanamaz bir değere sahiptir. Bu durum ilişkisel veri tabanlarında

sıkça karşımıza çıkmaktadır. Bir ilişkide yer alan tüm çoklular aynı sayıda niteliğe, niteliğin değeri boş olsa bile sahip olmalıdır. Örneğin, kişisel bilgisayarların özelliklerini tutan bir ilişkide bazı model bilgisayarlar için ses kartı modeli niteliğinin değeri boş olabilir. Boş değerli nitelikler veri kümesinde bulunuyorsa, ya bu çoklular tamamıyla ihmal edilmeli ya da bu çoklularda niteliğe olası en yakın değer atanmalıdır [7].

- Eksik Veri : Evrendeki her nesnenin ayrıntılı bir biçimde tanımlandığı ve bu nesnelerin alabileceği değerler kümesinin belirli olduğu varsayılsın. Verilen bir bağlamda her bir nesnenin tanımı kesin ve yeterli olsa idi sınıflama işlemi basitçe nesnelerin alt kümelerinden faydalanılarak yapılırdı. Bununla birlikte, veriler kurum ihtiyaçları göz önünde bulundurularak düzenlenip toplandığından, mevcut veri bilgi keşfi açısından uygun olmayabilir. Örneğin hastalığın tanısını koymak için kurallar sadece çok yaşlı insanların belirtilerinin bulunduğu bir veri kümesi kullanılarak üretilseydi, bu kurallara dayanarak bir çocuğa tanı koymak pek doğru olmazdı. Bu gibi koşullarda bilgi keşfi modeli belirli bir güvenlik (veya doğruluk) derecesinde tahmini kararlar alabilmelidir.

- Artik Veri : Verilen veri kümesi, eldeki probleme uygun olmayan veya artık nitelikler içerebilir. Bu durum pek çok işlem sırasında karşımıza çıkabilir. Örneğin, eldeki problem ile ilgili veriyi elde etmek için iki ilişkiyi ortak nitelikler üzerinden birleştirirsek, sonuç ilişkide kullanıcının farkında olmadığı artık nitelikler bulunur. Artık nitelikleri elemek için geliştirilmiş algoritmalar özellik seçimi olarak adlandırılır. Özellik seçimi, tümevarıma dayalı öğrenmede bir ön işlem olarak algılanır. Başka bir deyişle, özellik seçimi, verilen bir ilişkinin içsel tanımını, dışsal tanımın taşıdığı (veya içerdiği) bilgiyi bozmadan onu eldeki niteliklerden daha az sayıdaki niteliklerle (yeterli ve gerekli) ifade edebilmektir. Özellik seçimi yalnızca arama uzayını küçültmekle kalmayıp, sınıflama işleminin kalitesini de artırır

- Dinamik Veri : Kurumsal çevrim içi veri tabanları dinamiktir, yani içeriği sürekli olarak değişir. Bu durum, bilgi keşfi metodları için önemli sakıncalar doğurmaktadır. İlk olarak sadece okuma yapan ve uzun süre çalışan bilgi keşfi metodu, bir veri tabanı uygulaması olarak mevcut veri

tabanı ile birlikte çalıştırıldığında mevcut uygulamanın da performansı ciddi ölçüde düşer. Diğer bir sakınca ise, veri tabanında bulunan verilerin kalıcı olduğu varsayılıp, çevrim dışı veri üzerinde bilgi keşif metodu çalıştırıldığında, değişen verinin elde edilen örüntülere yansıması gerekmektedir. Bu işlem, bilgi keşfi metodunun ürettiği örüntüleri zaman içinde değişen veriye göre sadece ilgili örüntüleri yığmalı olarak günleme yeteneğine sahip olmasını gerektirir. Aktif veri tabanları tetikleme mekanizmalarına sahiptir ve bu özellik bilgi keşif metodları ile birlikte kullanılabilir.

- Farklı Tipteki Verileri Ele Alma : Gerçek hayattaki uygulamalar makine öğreniminde olduğu gibi yalnızca sembolik veya kategorik veri türleri değil, fakat aynı zamanda tamsayı, kesirli sayılar, çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılmasını gerektirir. Kullanılan verinin saklandığı ortam, düz bir kütük veya ilişkisel veri tabanında yer alan tablolar olacağı gibi, nesneye yönelik veri tabanları, çoklu ortam veri tabanları, coğrafik veri tabanları vb. olabilir. Saklandığı ortama göre veri, basit tipte olabileceği gibi karmaşık veri tipleri (çoklu ortam verisi, zaman içeren veri, yardımlı metin, coğrafi, vb.) de olabilir. Bununla birlikte veri tipi çeşitliliğinin fazla olması bir VM algoritmasının tüm veri tiplerini ele alabilmesini olanaksızlaştırmaktadır. Bu yüzden veri tipine özgü adanmış VM algoritmaları geliştirilmektedir
- Belirsizlik : Yanlışlıkların şiddeti ve verideki gürültünün derecesi ile ilgilidir. Veri tahmini bir keşif sisteminde önemli bir husustur.
- Ebat, güncellemeler ve konu dışı sahalar : Veri tabanlarındaki bilgiler, veri eklendikçe ya da silindikçe değişebilir. Veri madenciliği perspektifinden bakıldığında, kuralların hala aynı kalıp kalmadığı ve istikrarlılığı problemi ortaya çıkar. Öğrenme sistemi, kimi verilerin zamanla değişmesine ve keşif sisteminin verinin zamansızlığına karşın zaman duyarlı olmalıdır.

## 8.2. Veri Madenciliğini Etkileyen Eğilimler

Temel olarak veri madenciliğini 5 ana harici eğilim etkiler :

a) Veri : VM'nin bu kadar gelişmesindeki en önemli etkendir. Son yirmi yılda sayısal verinin hızla artması, VM'deki gelişmeleri hızlandırmıştır. Bu kadar fazla veriye bilgisayar ağları üzerinden erişilmektedir. Diğer yanda bu verilerle uğraşan bilim adamları, mühendisler ve istatistikçilerin sayısı hala aynıdır. O yüzden, verileri analiz etme yöntemleri ve teknikleri geliştirilmektedir.

b) Donanım : VM, sayısal ve istatistiksel olarak büyük veri kümeleri üzerinde yoğun işlemler yapmayı gerektirir. Gelişen bellek ve işlem hızı kapasitesi sayesinde, birkaç yıl önce madencilik yapılamayan veriler üzerinde çalışmak mümkün hale gelmiştir.

c) Bilgisayar Ağları : Yeni nesil internet, yaklaşık 155 Mbits/sn lik hatta belki de daha da üzerinde hızları kullanmamızı sağlayacak. Bu da günümüzde kullanılan bilgisayar ağlarındaki hızın 100 katından daha fazla bir sürat ve taşıma kapasitesi demektir. Böyle bir bilgisayar ağı ortamı oluştuktan sonra, dağıtık verileri analiz etmek ve farklı algoritmaları kullanmak mümkün olacaktır.

d) Bilimsel Hesaplamalar : Günümüz bilim adamları ve mühendisleri, simülasyonu bilimin üçüncü yolu olarak görmekteler. VM ve bilgi keşfi, bu 3 metodu birbirine bağlamada önemli rol almaktadır : teori, deney ve simülasyon.

e) Ticari Eğilimler : Günümüzde ticaret ve işler çok karlı olmalı, daha hızlı ilerlemeli ve daha yüksek kalitede servis ve hizmet verme yönünde olmalı, bütün bunları yaparken de minimum maliyeti ve en az insan gücünü göz önünde bulundurmalıdır. Bu tip hedef ve kısıtların yer aldığı iş dünyasında veri madenciliği, temel teknolojilerden biri haline gelmiştir. Çünkü veri madenciliği sayesinde müşterilerin ve müşteri faaliyetlerinin yarattığı fırsatlar daha kolay tespit edilebilmekte ve riskler daha açık görülebilmektedir.

# 9. VERİ MADENCİLİĞİ SİSTEMLERİ ÜZERİNE YAPILAN ÇALIŞMALAR

VM tekniklerinin bir çok alanda gerekli olan bilgiye erişmek için uygulanabilir olması VM teknikleriyle hem genel hem de özel amaçlı bir çok uygulamanın geliştirilmesini sağlamıştır.

1. Özel Amaçlı Sistemler: VM algoritmalarının spesifik problem çözümleri için kullanılmasıdır. Bu uygulamaların çıkış amacı VM'nin kullanıcıdan bağımsız bir şekilde çalıştırılarak kullanıcının istediği bilgilerin keşfedilmesi ve/veya keşfedilen bilgilerin gömülü (embedded systems) bir uygulama içinde direkt karar alınmasında faydalanılmasını sağlamaktır.

VM algoritmalarının özel amaçlı uygulandığı yerlerden ilk göze çarpanlar: astronomi, işletmelerdeki satış analizleri, pazarlama, borsa, sigorta vb. alanlardır.

2. Genel Amaçlı Sistemler: Bu tür sistemlerde amaçlanan VM sorgularının problemden bağımsız olarak tanımlanması ve bu özelliğinden dolayı istenen problemde bu sorguların kullanılabilmesidir.

Genel amaçlı sistemlerden ön plana çıkmış ürünlerden bazıları şunlardır [2]:

**Analysis Manager**

Analysis Manager, Microsoft firmasının VM için üretmiş olduğu ürünüdür. Kümeleme analizi ve karar ağaçları için hazırlanmıştır. Analysis Manager, OLAP (çevrim içi analitik işlem) küp desteği sunmaktadır. Analysis Manager'ın güçlü olduğu taraf kullanıcı-dostu (user friendly) bir ara yüze sahip olması ve uygulama kolaylığıdır. Aracın SQL SERVER 2000'le bütünleşik çalışabilmesi bu aracı etkin hale getirmektedir. Analysis Manager'ın bir VM sorgusu için farklı algoritmaları desteklememesi en büyük eksikliğidir. Kaynak kodun açık olmaması uygulama geliştiriciler için büyük zorluklar oluşturmaktadır. Kaynak kod yerine Microsoft kümeleme ve karar ağacı için COM (Bileşen nesne modeli - Component Object Model) desteği sunsa da bu destek bir çok gömülü sistem uygulamalarında geliştiriciler için eksik bir hizmet olarak görülmektedir.

Analysis Manager üretilen sonuçları farklı bir çok gösterim şekliyle kullanıcıya sunabilmektedir. Mesela karar ağaçları için karar ağacını gösterebildiği gibi sonuçları kural tabloları şeklinde yorumlama imkanı vermektedir.

**Darwin**

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

**Clementine**

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

**Enterprise Miner**

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

# 10. VERİ MADENCİLİĞİN UYGULANDIĞI VERİTABANLARI

Veri madenciliği birçok depolama birimi üzerinde uygulanabilir. Bunlar, ilişkisel veritabanları, veri ambarları, geleneksel veri tabanları, gelişmiş veri tabanları, dosyalar ve worl wide web olabilir. Gelişmiş veri tabanı sistemleri arasında, nesneye yönelik, nesne ilişkisel, text veri tabanları, multimedya veri tabanları sayılabilir. Veri madenciliği tekniklerinin avantajları, üzerinde uygulandığı depolama sistemlerine göre değişiklik gösterebilir [2].

## 10.1. İlişkisel Veri Tabanları

İlişkisel veri tabanları, tablolardan oluşmaktadır. Her tablonun tekil bir adı vardır ve attribute(columns, fields) değerlerinden oluşmaktadır. Ve genelde geniş bir satır kümesi içerir (records, rows). İlişkisel veri tabanlarındaki her satır, attribute değerleri ile tanımlanan bir nesneyi temsil eder. Veri tabanındaki entity ve ilişkileri modelleyen ER diagramları mevcuttur.

İlişkisel veri, SQL gibi yapısal sorgu dilleri ile yazılan sorgular ile ya da grafik kullanıcı arayüzleri ile erişilebilen verilerdir. Kullandığınız sorgu dili ya da kullanıcı arayüzünün size sağladığı olanaklar çemberinde, veriler ile istediğiniz soruların karşılıkları alınmaktadır.

Veri madenciliği, ilişkisel veri tabanlarındaki kayıtlara ait trendleri analiz etmek için ya da veri örüntülerini bulabilmek için kullanılabilir. Örneğin müşterilere ait kredi durumlarını analiz ederek yeni müşterilerin kredi risk durumlarını tesbit edebilir. Hangi yılda hangi ürünlerin satıldığı ya da satılması gerektiği gibi tahminler yapabilir.

Veri tabanları en sık kullanılan veri madenciliği uygulama platformlarından birisidir.

## 10.2. Veri Ambarları

Bir işletmenin değişik bölümleri tarafından toplanan bilgilerin, ileride değerlendirilmek üzere arka plandaki sistemde birleştirilmesinden oluşan geniş ölçekli veri deposudur.

Günümüz ticari işletmelerini iki başlıkta toplayabiliriz.

1-Canlı Sistemler :

Bu sistemlerde güncel veriler bulunur. Gündelik işleri gerçekleştirebilmek ve alınan sonuçları saklamak için geliştirilmişlerdir. Stok takibi, satış işlemleri, üye hareketlerinin takibi gibi. Bu sistemlerde veriye en kısa sürede ulaşmak ve işlemleri en kısa sürede sona erdirmek hedeflenir.

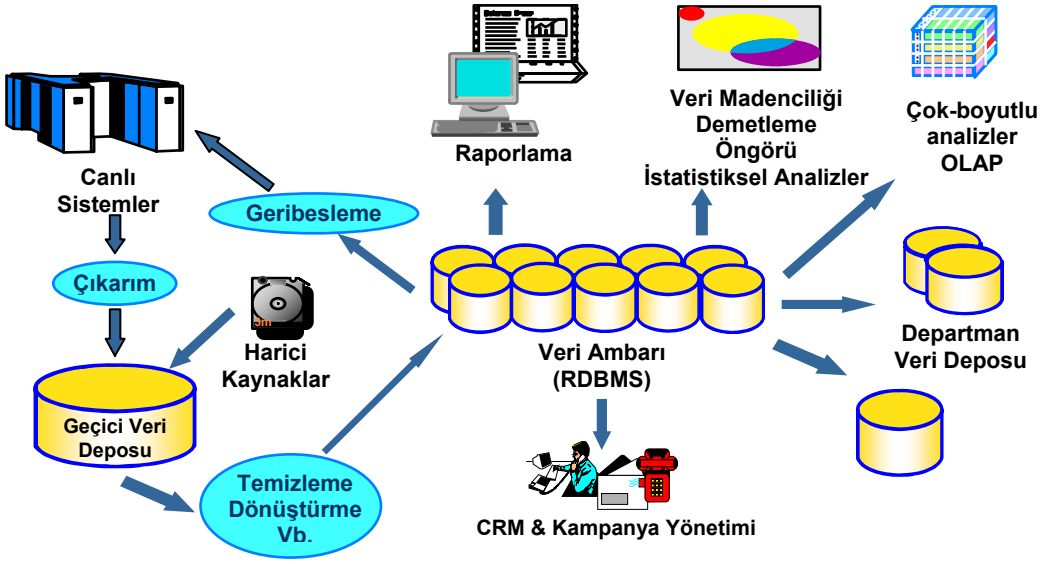
2-Karar Destek Sistemleri :

Bu sistemlerdeki bilgiler inceleme ve araştırmadan geçirilerek ileride yönetimin işletmenin verimliliğini artırmasını, izlenecek politikaların belirlenmesini ve benzeri yönetimsel kararların alınmasını kolaylaştırır. Veriler canlı sistemdekilerden çok daha büyük boyutlardadırlar. Asıl olarak hedeflenen performanstır.

Bu durumda bu iki modeli değerlendirecek olursak veri ambarı karar destek sistemi olarak değerlendirilebilir. Veri ambarları günlük işlemlerin gerçekleştirildiği sistemlerin arkasındadır. Bu sistemlerde oluşan veriler işletmenin şeçimine göre belirlenen periyotlarla veri ambarına aktarılırlar. Veri ambarları, veri temizleme, veri dönüştürme, veri yükleme ve periyodik veri transferi işlemlerinden inşa edilmişlerdir (Şekil 10.1).

Bir veri ambarı, genelde çokboyutlu veri tabanı yapısı ile modellenmiştir. Her boyut bir attribute ya da attribute kümesidir. Ve her hücre attribute lere bir ölçüm değeri taşır. Bir veri ambarının fiziksel yapısı, ilişkisel bir veri deposu olabilir ya da çok boyutlu veri küpü olabilir.

Veri ambarlarının perspektifinden veri madenciliği, Online Anaytical Prosesin (OLAP) advanced bir adımı olarak görülebilir. Veri madenciliği, veriyi anlayabilmek için veri ambarı sistemlerinin online analizini gerçekleştirir.




**Şekil 10.1** Veri Ambarının Yapısı

Şekil 10.1'de yer aldığı gibi veri ambarı üzerinde bir çok işlem gerçeklenebilmektedir. Bunlardan başlıcaları veri madenciliği, çok boyutlu analizler (OLAP), müşteri ilişkileri yönetimi (MİY – CRM), kampanya yönetimi, istatistiksel analizler ve raporlamadır. OLAP ile veri ambarı içerisinde yer alan kayıtlar yöneticiler tarafından istenilen boyutlarda ve biçimlerde raporlar haline getirilebilmekte ve çok boyutlu analizler yapılabilmektedir. Veri madenciliği ile verinin doğasında yatan kümelenmeleri, kayıtlar arasındaki ilişkileri bulmak, karar verme sürecinde yer alan soruların cevaplarını çıkarmak olası hale getirilmiştir. Aynı zamanda veri ambarı içerisinde yer alan kayıtlar üzerinde istatistiksel yaklaşımlarla raporlar oluşturmak ve istatistiksel sonuçlara varmak mümkündür.

## 10.3. Transactional(İşlemsel) Veri Tabanları

Genelde, transactional veri tabanı, her kaydın bir transactionu temsil ettiği dosyadan oluşur. Bir işlem (transaction), tekil bir işlem tanıtım numarası (trans_ID gibi) ve ilgili işlem içerisinde gerçekleşen olayların listesini içerir (bir alışverişte alınan malzeme listesi gibi).

Yapılan işlemler, kendilerine ait malzeme kümeleri ile bir kayıt olarak tutulabilirler .

## 10.4. Gelişmiş Veri Tabanı Sistemleri ve Uygulamaları

İlişkisel veri tabanı sistemleri ticari uygulamalarda çokca kullanılmaktadırlar. Veri tabanı teknolojisinin gelişimi ile birçok yeni kuşak veri tabanı teknikleri uygulamalarda kullanılmaya başlanmıştır.

Yeni veri tabanı uygulamaları, uzaysal veri (haritalar gibi), mühendislik dizayn verileri (bina dizaynları, sistem bileşenleri, devreler), multimedya datalar, zaman eksenli datalar, www datalar gibi veriler üzerinde işlem yapmaktadırlar. Bu tür uygulamalar, karmaşık nesne yapıları, değişken boyutlu kayıt yapıları, test-multimedya datalar için, verilerdeki dinamik değişimler için daha etkin veri yapıları gerektirmektedirler.

Bu ihtiyaçlara cevap verebilmek için, gelişmiş veri tabanı sistemleri geliştirilmiştir.

## 10.5. Nesneye Yönelik Veri Tabanları

Nesneye yönelik veri tabanları, nesneye yönelik programlama mantığına dayanmaktadır. Genel anlamda bir entity, nesne olarak kabul edilmektedir. Her nesne şu şekilde ortaya çıkmıştır.

- Nesneyi tanımlayan değişkenler kümesi. Değişkenler, entity-relationship modelindeki attribute'lere uyarlar.
- Nesnenin diğer nesneler ya da veri tabanının kalan kısmı ile haberleşebilmesi için gereken mesajlar kümesi
- Bir mesajı tamamlamak için gerekli kodları içeren methodlar kümesi. Bir mesajı yanıtlamanın üzerine method değeri response olarak döndürür.

Aynı özellikleri paylaşan nesneler aynı nesne sınıflarına ayrılabilirler. Nesne sınıfları, class/sub class hiyerarşisi ile yapılanabilirler.

## 10.6. Nesne İlişkisel Veri Tabanları

Nesne ilişkisel veri tabanları, nesne ilişkisel model baz alınarak yapılandırılmıştır. Nesne ilişkisel model aslında ilişkisel modelin üzerine

yapılanmıştır. Fakat ek olarak nesneye yönelik programlamadaki var olan kavramlar da işin içine girmektedir.

### 10.7. Uzaysal Veri Tabanları

Uzaysal veri tabanları, uzay ilişkili veriler içerirler. Örnek olarak coğrafik harita veri tabanları, VLSI chip dizayn veri tabanları, medical ve uydu imaj veri tabanları gibi.

Uzaysal veri tabanları üzerinde ne tür veri madenciliği işlemleri gerçekleşebilir sorusu sorulabilir. Mesela verilen bir yerin mesela bir parkın yakınlarında bulunan evin karakteristiklerini tanımlayarak örüntüleri keşfedebilir. Başka örüntüler, çeşitli yüksekliklerdeki dağlık alanların iklimlerini tanımlıyor olabilir.

### 10.8. Time Series-Temporal Veri Tabanları

Time-series database yada temporal database, ikisi de zaman ile ilişkili verileri depolarlar. Bir temporal database, zaman-ilişkili ilişkisel olan verileri depolar. Bu attributeler birçok timestamp içerirler. Time series database ise, zaman ile değişen bir dizi veri depolar. Veri stokları gibi örneğin.

Veri madenciliği teknikleri bunlar gibi veri tabanlarında verilere ait trendleri yakalamak amaçlı kullanılırlar.

### 10.9. Text ve Multimedya Veri Tabanları

Text veri tabanları, nesneler için metin tanımları içerirler. Bu metin tanımları genelde basit anahtar kelimeler değil uzun cümle ya da paragraflardır. Ürün ayrıntısı, hata kodları, uyarı mesajları, rapor özetleri gibi.

# 11. VERİ MADENCİLİĞİNDE YENİ YAKLAŞIMLAR

## 11.1. Yapay Bağışıklık Sistemi

Bağışıklık sistemindeki etkileşimleri daha iyi anlayabilmek için bağışıklık sisteminin bir modelini oluşturmak ve sistemdeki olayları hesapsal araç olarak kullanabilmek amacıyla ortaya atılmıştır

Bilgisayar bilimcileri, mühendisler, matematikçiler, filozoflar ve diğer araştırmacılar, karmaşıklığı beyne benzer olan bu sistemin özellikle yetenekleri üzerine ilgi duymaktadırlar. Yapay bağışıklık sistemi ile VM tekniklerinden sınıflama ve kümeleme için yeni yeni birkaç çalışma yapılmıştır. Birliktelik kurallarının keşfi için de tek bir çalışma vardır. Ancak diğer teknikler için de çalışmaların olacağı açıktır. Özellikle sistemin sahip olduğu özelliğiyle dağıtık ve paralel VM'de etkili olarak kullanılabilecektir.

## 11.2. Karınca Koloni Optimizasyonu

Karıncalar, yuvalarından bir gıda kaynağına giden en kısa yolu, herhangi görsel ipucu kullanmadan bulma yetisine sahiptirler. Koloni halinde yaşayan karıncalar yiyecek bulmak için ilk olarak öncü karıncaları tek başına gönderirler. Bu öncüler çevreyi araştırarak uygun yiyecek kaynağını bulmaya çalışır. Öncüler yiyecek bulursa, koloninin olduğu yere geri dönerken arkalarında özel bir koku izi bırakarak ilerler. Bu iz sayesinde diğer karıncalar da bu yiyecek kaynağını bulabilirler. Araştırmacılar bu arkalarında iz bırakarak ilerleyen öncü karıncaların uyguladığı yöntemi "sanal karıncalar" oluşturarak bilgisayarlarla simüle ettiklerinde çoğu problemin daha kolay çözülebileceğini gösterdiler.

Karınca koloni optimizasyon algoritması VM'de sınıflama tekniğinde yeni yeni kullanılmaya başlanmıştır. Elde edilen sonuçlar tatmin edicidir. Daha optimize parametrelerle dağıtık ya da paralel gerçeklemelerle çok daha iyi sonuçların alınacağı kesindir. Diğer tekniklerde de bu yaklaşım kullanılabilir.

## 11.3. Destek Vektör Makineleri

Destek vektör makineleri yeni bir öğrenme metodudur. Çekirdek tabanlı doğrusal olmayan sınıflandırıcıların sinyal işleme, yapay öğrenme ve VM alanındaki pratik problemlerde iyi sonuçlar verdiği bulunmuştur.

V. Vapnik tarafından önerilen destek vektör makineleri (DVM) ileri yönde beslemeli yeni bir ağ kategorisidir. İstatistiksel öğrenme teorisinde iyi şekilde kurulmuş bir teoriye sahiptir ve sınıflandırma problemlerine yaklaşım için uygundur. Özellikle iki sınıf sınıflandırma probleminde, DVM iki sınıf arasındaki sınırı büyükleyen optimal ayırt etme yüzeyini belirlemekte, yani eğitim kümesi ile ayırt etme yüzeyine en yakın noktaların arasındaki mesafeyi en büyüklemektedir.

Kısaca DVM, doğrusal olmayan bir şekilde ayrılabilen öbekler için optimal hiperdüzlemi bulmaya çalışır. Bu yüzden DVM'nin VM'deki uygulamaları özellikle sınıflama tekniğinde ortaya çıkmıştır. Elde edilen sonuçlar bu yöntemin sınıflama tekniğinde oldukça başarılı olduğunu göstermiştir .

## 11.4. Kaos

Kaos kelimesi insanda pek de hoş olmayan çağrışımlar yapar. Karmaşıklık, belirsizlik ve hatta anarşi. Bilimde ise kaos kelimesi belirlenemezlik olarak kabul edilir. Yani günlük yaşamda kullanımı ile bilimde kullanımı oldukça farklıdır.

Kaos teorisi engin uygulama alanına sahip olan bir yaklaşımdır. Her türlü alanda uygulanabilme yeteneğinden dolayı, kaos teorisinin bilim dallarını birbirinden soyutlayan engelleri aştığı söylenebilir. Çok küçük görünen bir nedenin kendisinden çok daha büyük sonuçlara yol açabileceği mantığından hareket eden kaos kuramı, düzensizlik ve karmaşadan çok, bu düzensizlik içerisinde belli bir düzeni, düzenli düzensizliği anlamaya yöneliktir.

Endüstriyel alanlarda çoğu işlemlerin doğrusal olmamasından dolayı kaotik işlemlerin tahmini yapılmaya çalışılmaktadır. VM'de kaos kuramı, kümelemede ve zaman verisinin de kullanıldığı durumlarda birliktelik kural keşfinde kullanılmıştır.

# 12. LOJİSTİK SEKTÖRÜNDE VERİ MADENCİLİĞİ UYGULAMASI

Bu uygulamada uluslararası lojistik firması olan Omsan Lojistik A.Ş. firmasının Ocak 2004 - Ocak 2006 operasyonel verileri kullanılmıştır.

## 12.1. Şirket Hakkında Genel Bilgi

Omsan Lojistik A.Ş., 1978 yılında bir OYAK iştiraki olarak kurulmuştur. Lojistik sektörünün lider kuruluşu olan şirket, dünyadaki ve Türkiye'deki gelişmelere paralel olarak faaliyet alanlarını "entegre lojistik hizmet - tedarik zinciri yönetimi" kavramı doğrultusunda genişletmiştir. Şirketin faaliyetleri arasında Kara, Hava, Deniz ve Demiryolu taşımacılığı faaliyetleri, ilgili danışmanlık ve takip hizmetleri; Türkiye'nin değişik noktalarında ve kritik endüstriyel bölgelerinde depo, antrepo ve gümrükleme hizmetleri; nakliye, depolama, gümrükleme, danışmanlık gibi çeşitli faaliyetlerin bir arada sunulduğu, müşteri ihtiyaçlarına tek elden cevap verebilecek entegre lojistik çözümler; tedarik zinciri boyunca sipariş ve stok yönetimi vb. gibi faaliyetler bulunmaktadır.

## 12.2. Şirket IT Yapısı ve Uygulamada Kullanılan Araçlar Hakkında Genel Bilgi

Omsan Lojistik A.Ş., 2001 yılında kapsamlı bir değişim sürecinden geçerek, yönetim birimlerini tek çatı altında toplamış, vizyonu olan entegre lojistik hizmetlerini en kapsamlı ve ekonomik şekilde sağlayabilecek yeni bir organizasyona kavuşmuştur. Bu yapısal değişim sürecine paralel olarak, şirket içerisinde kullanılan IT yazılımları da gözden geçirilmiş ve operasyonel süreçlerdeki istekleri karşılayan hazır bir ERP (Kurumsal Kaynak Planlama) paketi bulunamadığı için dışarıdan bir yazılımcı firma ile anlaşılmış ve şirket ihtiyaçlarını karşılayacak OPUM adı verilen in-house bir çözüm geliştirilmiştir. Şirket halen tüm operasyonel süreçlerinde Opum sistemini kullanmaktadır.

Şirket, veritabanı olarak Oracle 9i veritabanını kullanmaktadır. Uygulamada, veritabanından yapılan her tür işlem için Oracle'ın TOAD toolu kullanılmıştır.

Kullanılan Toad toolunun versiyonu 8.5'tir. Uygulama sonunda veritabından yapılan sorgulamaların raporlama şeklinde kullanıcıya sunulabilmesi için Oracle'ın raporlama toolu olan Oracle Discoverer kullanılmıştır. Kullanılan Discoverer toolunun versiyonu 9.0.4'tür.

### 12.3. Sistem Üzerinde Süreç İşleyişi Hakkında Genel Bilgi

Veri madenciliği çalışmasına başlamadan önce çalışmanın anlaşılabilmesi için, sürecin OPUM üzerinde nasıl işlediği ile ilgili kısa bir bilgilendirme yapmak faydalı olacaktır.

Süreç, müşteriden siparişin alınması ile başlar. Lojistik hizmet departmanına bağlı müşteri temsilcileri, müşteriden almış oldukları bu siparişleri sisteme girerler. Sipariş girişi, üç adımlı bir işlemdir. Birinci adımda sipariş tarihi, sipariş müşterisi, siparişin ait olduğu grup (oto taşıma, evlojistik v.b.) gibi sipariş ile ilgili ana bilgiler girilir. İkinci adımda bu siparişte müşterinin bizden talep ettiği hizmetler ile ilgili bilgiler sisteme girilir. Müşteri bizden hem karayolu taşımacılığı hem de gümrükleme hizmeti talep etmiş olabilir Bu iki hizmet talebi ilgili siparişin altına iki ayrı sipariş satırı olarak girilir. Her bir sipariş satırı ile ilgili gerekli detay bilgileri girilir. Üçüncü adımda her bir sipariş satırının altına Aktivite adı verilen ve işin fiili olarak yapılabilmesi için iş emri niteliği taşıyan kayıtlar açılır. Açılan bu aktiviteler, özmal araçlarımızı yöneten FILO departmanının önüne düşer. Her bir aktivite üzerinde işin niteliği ile ilgili tüm bilgileri mevcuttur. Örnek olarak karayolu taşımacılığı için aracın nerelerde yükleme yapıp nerelerde boşaltma yapacağı, müşterinin işin başlamasını ve bitmesini istediği tarih gibi bilgiler mevcuttur. FILO departmanı bu bilgilere bakarak elindeki boş kaynakları (şoför, araç) bu aktiviteler üzerine atar ve işi başlatır. İş başladıktan sonra aktivitelerin işleyişi ile ilgili önemli detay bilgiler de sisteme girilir. Aktivitenin ne zaman başladığı, her bir yükleme-boşaltma yerine ne zaman varıldığı, ne kadar sürede yükleme-boşaltma işleminin gerçekleştiği, aktivitenin fiili olarak ne zaman bittiği gibi önemli bilgiler sisteme girilir. Bu bilgilerin sisteme girilmesi, bizim raporlama ve analiz yapmamıza olanak vermektedir.

## 12.4. Veri Madenciliği Adımlarının Uygulanması

Bu bölümde veri madenciliği adımlarının her biri üzerinde tek tek durulacak ve yapılan uygulamanın her bir aşaması tek tek incelenecektir.

### 12.4.1. Problemin Tanımlanması

Lojistikte en önemli iki kavram tahsis ve termin tarihlerine uyumdur. Tahsis tarihi, iş emrine (aktiviteye) kaynağın (aracın,şoförün) atandığı tarih ya da kısaca işin başladığı tarih olarak düşünülebilir. Termin tarihi ise malın istenilen son noktaya iletildiği yani iş emrinin fiili olarak bittiği tarihtir. Tahsis tarihinin müşterinin o işin başlamasını istediğe tarihe uyumu tahsis uyumu, termin tarihinin müşterinin o işin bitmesini istediği tarihe uyumu termin uyumu olarak nitelendirilir.

Bu uygulamada rakip firmalara kaptırdığımız müşterilerin siparişlerine ait iş emirleri tahsis ve termin tarihlerine uyum açısından incelenecektir. Amacımız, bu iki önemli kriter bazında kaybedilen müşterilerle bizim sürekli müşterilerimiz arasında bir ilişki yakalayabilmek, bir kural bulabilmektir. Eğer böyle bir kural bulabilirsek çalıştığımız yeni bir müşteriye ait gerçekleşen iş emirlerini analiz ederek müşterinin potansiyel kaybedilecek bir müşteri olup olmadığını anlayabilir ve önlemlerimizi ona göre alabiliriz.

### 12.4.2. Verilerin Hazırlanması

Bu aşamada öncelikle tanımlanan problem için gerekli olacak verilerin toplanacağı veri kaynakları belirlenir. Bizim örneğimizde gerekli tüm bilgiler sistem üzerinde bulunmaktadır. Bu nedenle bizim veri kaynağımız veri tabınımızdır. Veri tabanında mevcut problemimizle ilgili sayısal verilerin tutulduğu tablolar belirlenir. Tablo12.1'de problemle ilgili tablolar ve açıklamaları verilmiştir :

**Tablo 12.1** Problem ile İlgili Veri Tabanında Bulunan Tablolar ve Açıklamaları

| | |
|---|---|
| TBLSIPARISLER | Müşterilerden alınan siparişlere ait ana bilgilerin tutulduğu tablodur. |
| TBLSIPARISSATIR | Müşterilerin verdikleri siparişte bizden talep ettikleri hizmet bilgilerinin tutulduğu tablodur. |
| TBLSIPARISAKTIVITELERI | Her bir sipariş satırına ait iş emirleri (aktivite) bilgilerinin tutulduğu tablodur. |
| TBLAKTIVITELER | İş emirleri ve bu iş emirlerine ait detay bilgilerin tutulduğu tablodur. |
| TBLFIRMALAR | Bir şekilde ilişki içerisinde bulunduğumuz tüzel ya da gerçek kişilerin tutulduğu tablodur. |
| TBLFATURABASLIK | Sisteme girilen gelir fatura bilgilerinin tutulduğu tablodur. |

İkinci aşamada, verilerin temizlenmesi ve birleştirilmesi işlemi yapılır. Temizlik aşamasında öncelikle sistemin işlemeye başladığı 2001 yılına ait veriler göz ardı edilmiştir. Sebebi, sistemin o yıl içinde ilk kez kullanılmış olması ve kullanıcı hatalarına imkan bırakan çok sayıda bugın(eksikliğin) sistem üzerinde ilk zamanlarda mevcut olmasıdır. İlk yıl girilen verilerin temizliği şüphelidir. Yapılacak analizde bizi yanlış yönledirmesini engellemek adına 2001 yılı verileri göz ardı edilecektir.

Problemi inceleyebilmek için öncelikle "kaybedilen müşteri" tanımının yapılması gerekir. Müşterilerin bir kısmı spot (anlık) işler için bizimle çalışan müşterilerdir. Bunlar bir defaya mahsus olmak üzere bize iş veren ve bunun devamını getirmeyen müşterilerdir. Örnek olarak ev lojistik taşımlarını verebiliriz. Bu nedenle kaybedilen müşteriyi tanımlarken bizimle en azından bir ay çalışmış olması zorunluluğunu arayabiliriz. Müşterinin bizi terk ettiğine karar verebilmemiz için de bizimle 6 aydan az çalışmış olması koşulunu aramak yerinde olacaktır. Bu durumda bizimle bir aydan fazla, 6 aydan az çalışan müşteriler "kaybedilen

müşteri"lerimizdir, diyebiliriz. Bir müşterinin bizimle ne kadar süre çalıştığı bilgisini o müşteriye kesilen fatura tarihi bilgisinden elde etmemiz mümkündür. Müşteriye kestiğimiz ilk faturanın tarihi ile son faturanın tarihi arasında bir aydan çok 6 aydan az olan müşteriler benim kaybedilen müşterilerimdir. Bu tespitlerden sonra veritabanında bana bu bilgiyi verebilecek yapıyı kurma aşamasına geçebiliriz. Bu yapı, bir ara tablodur. Problemin çözümünde ihtiyacımız olan müşteri grubunu verebilecek ara tabloyu oluşturan sorgu şu şekildedir:

```
CREATE TABLE TBLMUSTERI_DM AS
SELECT f.firmakodu,
        fc_musteriilkfaturatrh(f.firmakodu)   ilkfaturatrh,
       --bu firmaya kesilen ilk faturanın tarihi gösterir
        fc_musterisonfaturatrh(f.firmakodu) sonfaturatrh
        --bu firmaya kesilen son  faturanın tarihi gösterir
FROM    tblfirmalar f
WHERE  f.firmatipi='01'  -- tipi müşteri olan firma kayıtları
      and fc_musteriilkfaturatrh(f.firmakodu) is not null
       and fc_musterisonfaturatrh(f.firmakodu) is not null
```

Bu sorguda iki tane fonksiyon kullanılmıştır: fc_musteriilkfaturatrh ve fc_musterisonfaturatrh. Bu fonksiyonlar, sırasıyla, ilgili firmaya ait sistemde kesilen ilk gelir faturasının ve son gelir faturasının tarihlerini getirmektedirler.

Artık elimizde her bir müşteri için ne kadar süre çalışıldığı bilgisi mevcuttur.

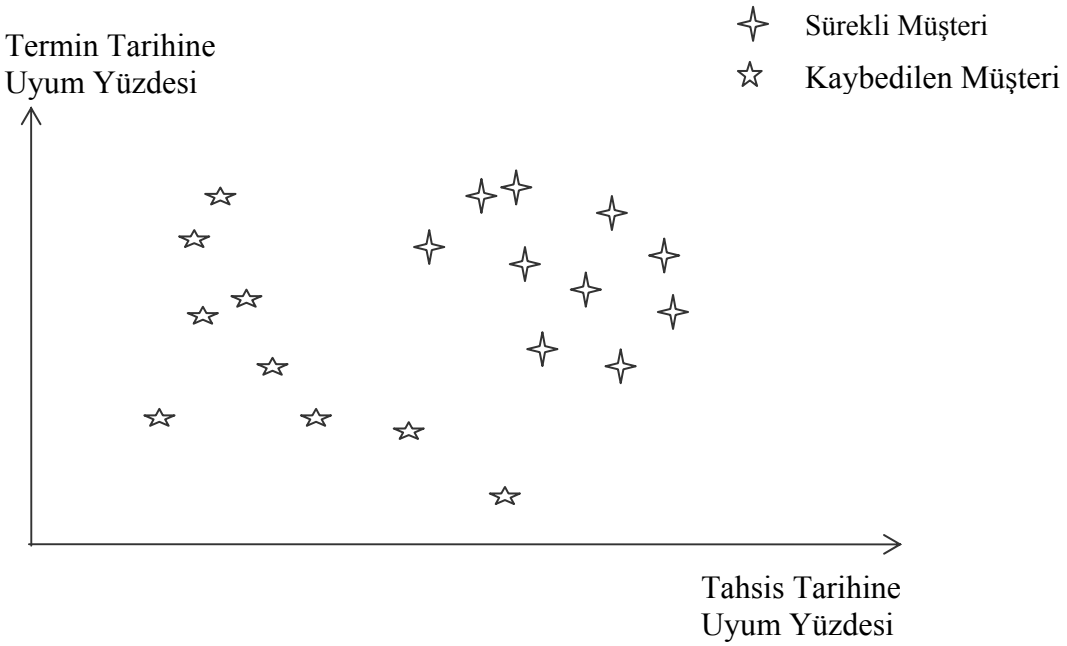
Ikinci olarak, her bir müşteri ile ilgili olarak alınan sipariş ve bu siparişlere istinaden yapılan iş emirlerinin detay bilgisine ihtiyacımız vardır. Bizim ilgilendiğimiz hizmet tipi karayolu taşımacılığı ve özmal  taşımacılık olduğu için ilgilendiğimiz tablolarda tutulan gereksiz verilerin bizim ilgi alanımızdan çıkarılması gerekmektedir. Bunun için de ilgilendiğimiz verileri derleyip toplayan ve bir ara tabloya atan aşağıdaki sorgu yazılmıştır :

```
CREATE TABLE TBLAKTIVITEBILGILERI_DM AS
SELECT   s.firmakodu,
         a.aktivitekodu,
         a.tahminibaslangictarihi,
         a.tahminibitistarihi,
         a.baslangictarihi,
         a.bitistarihi,
         a.baslangictarihi-a.tahminibaslangictarihi tahsis_farki,
         a.bitistarihi-a.tahminibitistarihi termin_farki,
         m.ilkfaturatrh,
         m.sonfaturatrh
FROM     tblsiparisler s,
         tblsiparissatir ss,
         tblsiparisaktiviteleri sa,
         tblaktiviteler a,
         tblmusteri_dm m
WHERE    s.siparisno=ss.siparisno
         and ss.siparisno=sa.siparisno
         and ss.satirno=sa.satirno
         and ss.musterihizmettipikodu='KTasima'
         and ss.iptalflg=0
         and sa.aktivitekodu=a.aktivitekodu
         and s.firmakodu=m.firmakodu
         and a.kesinflg=-1
         and a.iptalflg=0
         and a.ozmalflg=-1
         and a.aktivitetipikodu=’KARAYOLU'
         and s.siparistarihi>=’01.01.2001’
```

Bu ara tablo bize her bir müşteriye ait yapılan karayolu-özmal iş emirlerine ait fiili başlangıç-bitiş ve müşterinin talep ettiği başlangıç-bitiş tarihlerini vermektedir.

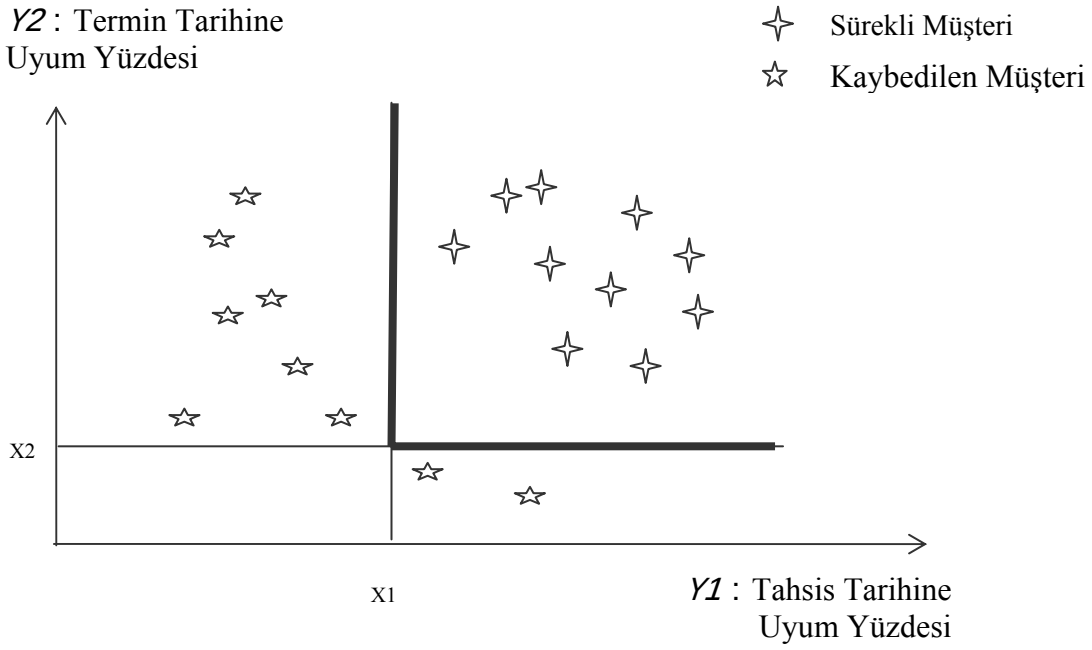
### 12.4.3. Modelin Kurulması ve Değerlendirilmesi

Tanımlanan problemin çözümünde kullanılacak yöntem sınıflandırma - karar ağaçları yöntemidir. Amacımız iki boyutlu uzayda tahsis ve termin tarihlerine uyum yüzdesine bakarak kaybettiğimiz ve bizimle çalışmaya devam eden müşterilere karşılık gelen noktaları birbirinden ayıran bir sınır bulabilmektir. Şekil 12.1’de problemimiz için veri örneği kümesi iki boyutlu grafik üzerinde gösterilmiştir.

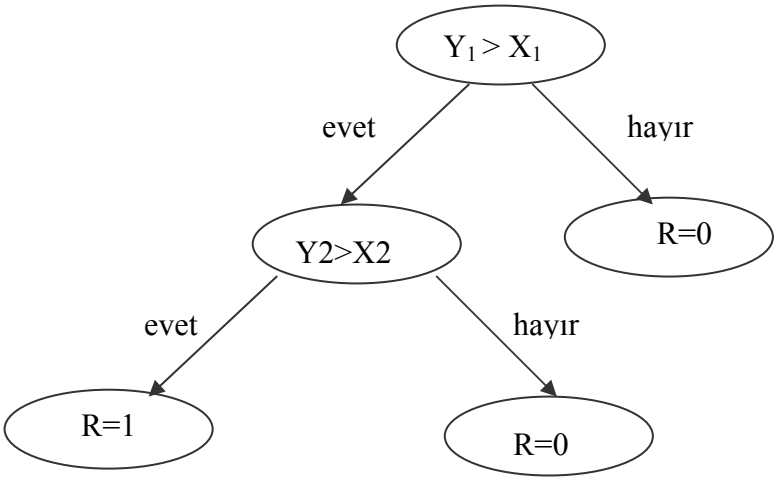



**Şekil 12.1** Veri örneği

Bu örnek üzerinde karar ağacı kullanılınca bulunabilecek sınır Şekil 12.2'de, karşılık gelen karar ağacının yapısı da Şekil 12.3'de verilmiştir.




**Şekil 12.2** Karar Ağacı Tarafından Tanımlanan Sınır




**Şekil 12.3** Şekil 12.2'de Verilen Sınırları Tanımlayan Karar Ağacının Yapısı.
X1: Tahsis Tarihine Uyum Yüzdesi X2: Termin Tarihine Uyum Yüzdesi
R=0 Devamlı Müşteri R=1 Potansiyel Kaybedilecek Müşteri.

Bu karar ağacı şu kurala karşılık gelir:

EĞER Tahsis Tarihine Uyum Yüzdesi>X1 veya Termin Tarihine Uyum Yüzdesi > X2 ise Sürekli Müşteri değilse Potansiyel Kaybedilecek Müşteri

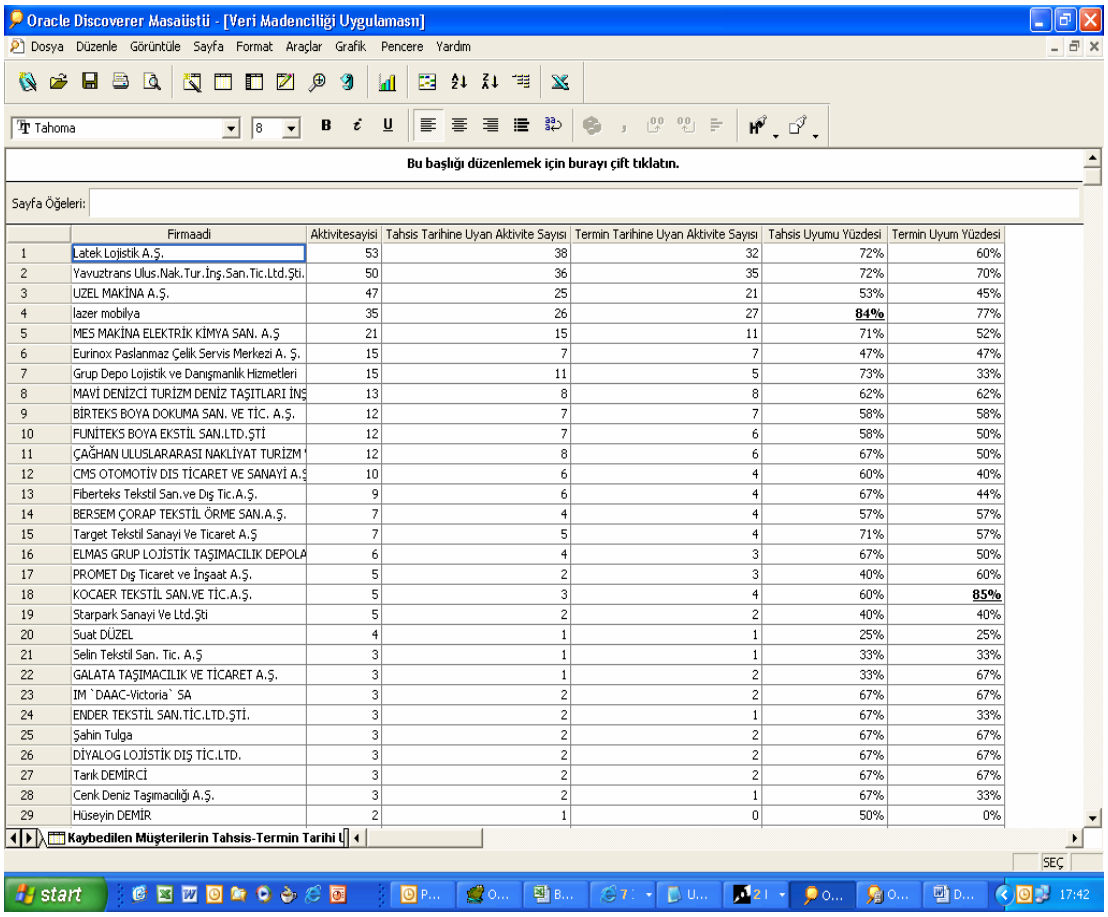
Karar ağaçlarının en büyük yararı veriden öğrenilen kuralın anlaşılır bir şekilde yazılabilmesidir. X1 ve X2 iki boyutlu eşik değerleridir. Karar ağacı ve bu eşik değerleri karar ağacı öğrenme algoritması tarafından veriden otomatik hesaplanır.

Modelimizi ve hedefimizi belirledikten sonra bize bu iki eşik değerini verebilecek raporlama aşamasına geçebiliriz.

Öncelikle yarattığımız ara tablolardan faydalanarak kaybettiğimiz müşterilerin iş emirlerini, tahsis ve termin uyum başarısı bakımından inceleyen bir rapor hazırlayalım. Bu rapor Oracle'ın raporlama toolu Oracle Discoverer yardımıyla hazırlanmıştır. Yarattığımız tabloları kullanarak aşağıdaki gibi bir sorgu çekelim :

```
select   firmakodu,
count(aktivitekodu) aktivitesayisi,
(
        select count(*) from tblaktivitebilgileri_dm f where f
        .firmakodu=d.firmakodu and trunc(f.TAHSIS_FARKI)=0
) tahsis,
(
        select count(*) from tblaktivitebilgileri_dm f where
        f.firmakodu=d.firmakodu and trunc(f.termin_FARKI)=0
) termin,
(       (select count(*) from tblaktivitebilgileri_dm f where
        f.firmakodu=d.firmakodu and
        trunc(f.TAHSIS_FARKI)=0)/count(aktivitekodu)
)*100 tahsis_uyumu,
(       (select count(*) from tblaktivitebilgileri_dm f where
        f.firmakodu=d.firmakodu and
        trunc(f.termin_FARKI)=0)/count(aktivitekodu)
)*100 termin_uyumu
from tblaktivitebilgileri_dm d
where  (sonfaturatrh - ilkfaturatrh)>30 and
        (sonfaturatrh - ilkfaturatrh)<180
group by firmakodu
```

Discoverer raporlama arayüzünde bu raporun çıktısı aşağıdaki şekilde olacaktır :

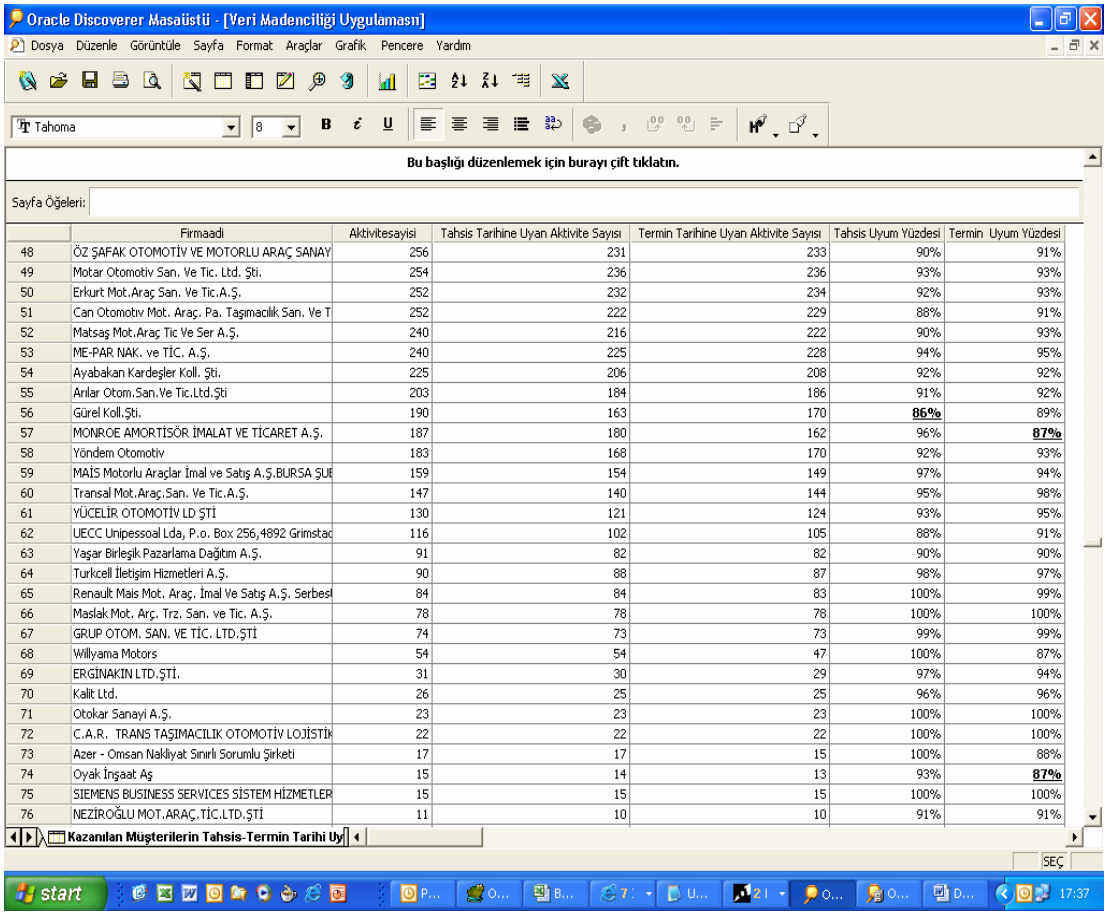



| | Firmaadi | Aktivitesayisi | Tahsis Tarihine Uyan Aktivite Sayısı | Termin Tarihine Uyan Aktivite Sayısı | Tahsis Uyumu Yüzdesi | Termin Uyum Yüzdesi |
|---|---|---|---|---|---|---|
| 1 | Latek Lojistik A.Ş. | 53 | 38 | 32 | 72% | 60% |
| 2 | Yavuztrans Ulus.Nak.Tur.İnş.San.Tic.Ltd.Şti. | 50 | 36 | 35 | 72% | 70% |
| 3 | UZEL MAKİNA A.Ş. | 47 | 25 | 21 | 53% | 45% |
| 4 | lazer mobilya | 35 | 26 | 27 | **84%** | 77% |
| 5 | MES MAKİNA ELEKTRİK KİMYA SAN. A.Ş | 21 | 15 | 11 | 71% | 52% |
| 6 | Eurinox Paslanmaz Çelik Servis Merkezi A. Ş. | 15 | 7 | 7 | 47% | 47% |
| 7 | Grup Depo Lojistik ve Danışmanlık Hizmetleri | 15 | 11 | 5 | 73% | 33% |
| 8 | MAVİ DENİZCİ TURİZM DENİZ TAŞITLARI İNŞ | 13 | 8 | 8 | 62% | 62% |
| 9 | BİRTEKS BOYA DOKUMA SAN. VE TİC. A.Ş. | 12 | 7 | 7 | 58% | 58% |
| 10 | FUNİTEKS BOYA EKSTİL SAN.LTD.ŞTİ | 12 | 7 | 6 | 58% | 50% |
| 11 | ÇAĞHAN ULUSLARARASI NAKLİYAT TURİZM | 12 | 8 | 6 | 67% | 50% |
| 12 | CMS OTOMOTİV DIS TİCARET VE SANAYİ A.Ş | 10 | 6 | 4 | 60% | 40% |
| 13 | Fiberteks Tekstil San.ve Dış Tic.A.Ş. | 9 | 6 | 4 | 67% | 44% |
| 14 | BERSEM ÇORAP TEKSTİL ÖRME SAN.A.Ş. | 7 | 4 | 4 | 57% | 57% |
| 15 | Target Tekstil Sanayi Ve Ticaret A.Ş | 7 | 5 | 4 | 71% | 57% |
| 16 | ELMAS GRUP LOJİSTİK TAŞIMACILIK DEPOLA | 6 | 4 | 3 | 67% | 50% |
| 17 | PROMET Dış Ticaret ve İnşaat A.Ş. | 5 | 2 | 3 | 40% | 60% |
| 18 | KOCAER TEKSTİL SAN.VE TİC.A.Ş. | 5 | 3 | 4 | 60% | **85%** |
| 19 | Starpark Sanayi Ve Ltd.Şti | 5 | 2 | 2 | 40% | 40% |
| 20 | Suat DÜZEL | 4 | 1 | 1 | 25% | 25% |
| 21 | Selin Tekstil San. Tic. A.Ş | 3 | 1 | 1 | 33% | 33% |
| 22 | GALATA TAŞIMACILIK VE TİCARET A.Ş. | 3 | 1 | 2 | 33% | 67% |
| 23 | IM `DAAC-Victoria` SA | 3 | 2 | 2 | 67% | 67% |
| 24 | ENDER TEKSTİL SAN.TİC.LTD.ŞTİ. | 3 | 2 | 1 | 67% | 33% |
| 25 | Şahin Tulga | 3 | 2 | 2 | 67% | 67% |
| 26 | DİYALOG LOJİSTİK DIŞ TİC.LTD. | 3 | 2 | 2 | 67% | 67% |
| 27 | Tarık DEMİRCİ | 3 | 2 | 2 | 67% | 67% |
| 28 | Cenk Deniz Taşımacılığı A.Ş. | 3 | 2 | 1 | 67% | 33% |
| 29 | Hüseyin DEMİR | 2 | 1 | 0 | 50% | 0% |



**Şekil 12.4** Kaybedilen Müşterilerin Termin-Tahsis Uyumu

Bu raporda, kaybettiğimiz müşterilerde elde ettiğimiz en yüksek tahsis ve termin uyum yüzdesine bakalım. Tahsis uyumunda en fazla %84 ve termin uyumunda en fazla %85 başarı göstermişiz ve bu değerler müşteriyi elimizde tutmaya yetmemiş. Bulduğumuz bu iki değerin tutarlılığını sınamak için bu sefer de benzer raporu sürekli müşterilerimiz için çekelim ve onlardaki bu iki boyuttaki başarı oranımıza bakalım.




| | Firmaadi | Aktivitesayisi | Tahsis Tarihine Uyan Aktivite Sayısı | Termin Tarihine Uyan Aktivite Sayısı | Tahsis Uyum Yüzdesi | Termin Uyum Yüzdesi |
|---|---|---|---|---|---|---|
| 48 | ÖZ ŞAFAK OTOMOTİV VE MOTORLU ARAÇ SANAY | 256 | 231 | 233 | 90% | 91% |
| 49 | Motar Otomotiv San. Ve Tic. Ltd. Şti. | 254 | 236 | 236 | 93% | 93% |
| 50 | Erkurt Mot.Araç San. Ve Tic.A.Ş. | 252 | 232 | 234 | 92% | 93% |
| 51 | Can Otomotiv Mot. Araç. Pa. Taşımacılık San. Ve T | 252 | 222 | 229 | 88% | 91% |
| 52 | Matsaş Mot.Araç Tic Ve Ser A.Ş. | 240 | 216 | 222 | 90% | 93% |
| 53 | ME-PAR NAK. ve TİC. A.Ş. | 240 | 225 | 228 | 94% | 95% |
| 54 | Ayabakan Kardeşler Koll. Şti. | 225 | 206 | 208 | 92% | 92% |
| 55 | Anlar Otom.San.Ve Tic.Ltd.Şti | 203 | 184 | 186 | 91% | 92% |
| 56 | Gürel Koll.Şti. | 190 | 163 | 170 | **86%** | 89% |
| 57 | MONROE AMORTİSÖR İMALAT VE TİCARET A.Ş. | 187 | 180 | 162 | 96% | **87%** |
| 58 | Yöndem Otomotiv | 183 | 168 | 170 | 92% | 93% |
| 59 | MAİS Motorlu Araçlar İmal ve Satış A.Ş.BURSA ŞUB | 159 | 154 | 149 | 97% | 94% |
| 60 | Transal Mot.Araç.San. Ve Tic.A.Ş. | 147 | 140 | 144 | 95% | 98% |
| 61 | YÜCELİR OTOMOTİV LD ŞTİ | 130 | 121 | 124 | 93% | 95% |
| 62 | UECC Unipessoal Lda, P.o. Box 256,4892 Grimstac | 116 | 102 | 105 | 88% | 91% |
| 63 | Yaşar Birleşik Pazarlama Dağıtım A.Ş. | 91 | 82 | 82 | 90% | 90% |
| 64 | Turkcell İletişim Hizmetleri A.Ş. | 90 | 88 | 87 | 98% | 97% |
| 65 | Renault Mais Mot. Araç. İmal Ve Satış A.Ş. Serbest | 84 | 84 | 83 | 100% | 99% |
| 66 | Maslak Mot. Arç. Trz. San. ve Tic. A.Ş. | 78 | 78 | 78 | 100% | 100% |
| 67 | GRUP OTOM. SAN. VE TİC. LTD.ŞTİ | 74 | 73 | 73 | 99% | 99% |
| 68 | Willyama Motors | 54 | 54 | 47 | 100% | 87% |
| 69 | ERGİNAKIN LTD.ŞTİ. | 31 | 30 | 29 | 97% | 94% |
| 70 | Kalit Ltd. | 26 | 25 | 25 | 96% | 96% |
| 71 | Otokar Sanayi A.Ş. | 23 | 23 | 23 | 100% | 100% |
| 72 | C.A.R. TRANS TAŞIMACILIK OTOMOTİV LOJİSTİK | 22 | 22 | 22 | 100% | 100% |
| 73 | Azer - Omsan Nakliyat Sınırlı Sorumlu Şirketi | 17 | 17 | 15 | 100% | 88% |
| 74 | Oyak İnşaat Aş | 15 | 14 | 13 | 93% | **87%** |
| 75 | SIEMENS BUSINESS SERVICES SİSTEM HİZMETLER | 15 | 15 | 15 | 100% | 100% |
| 76 | NEZİROĞLU MOT.ARAÇ.TİC.LTD.ŞTİ | 11 | 10 | 10 | 91% | 91% |

**Şekil 12.5** Sürekli Müşterilerin Termin-Tahsis Uyumu

Bu raporda, sürekli müşterilerimizde elde ettiğimiz en düşük tahsis ve termin uyum yüzdesine bakalım. Tahsis uyumunda en düşük %86 ve termin uyumunda en düşük %87 başarı göstermişiz ve bu değerleri elde ettiğimiz müşteri bizimle çalışmaya devam etmiş.

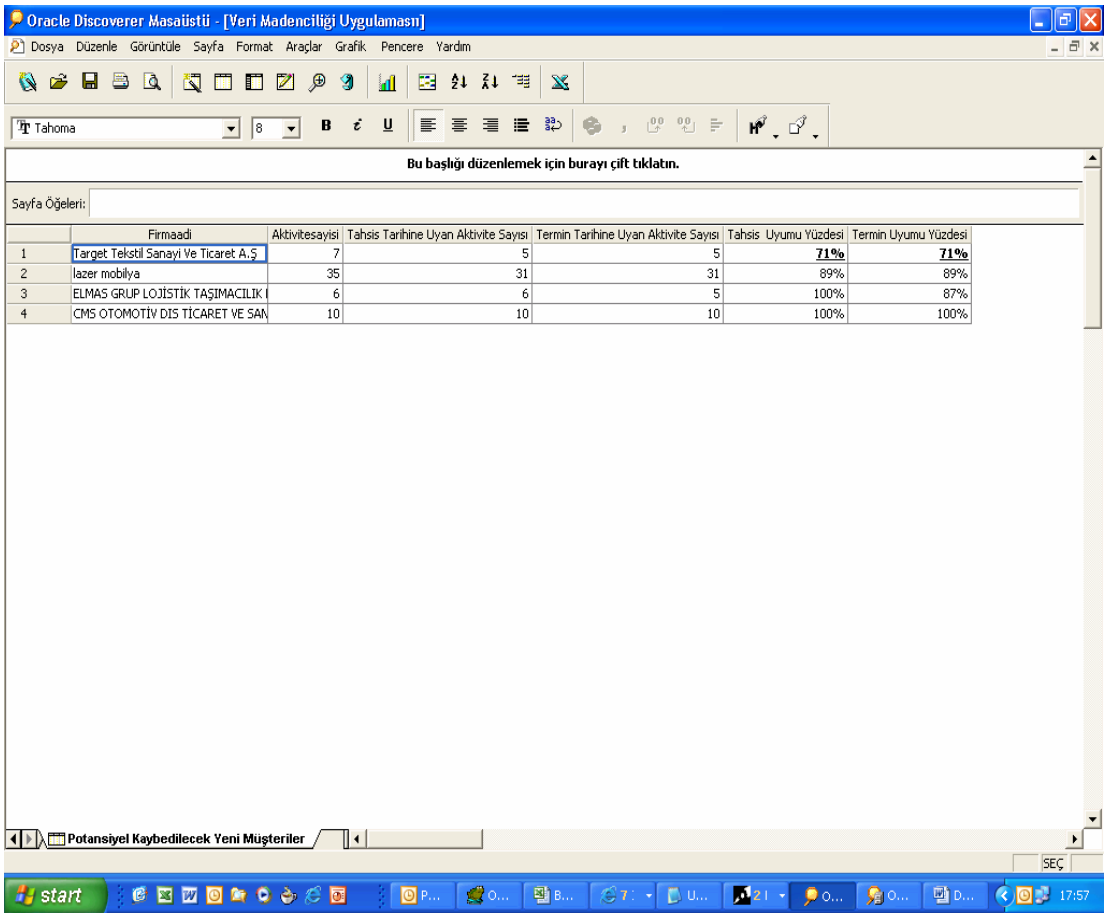
Bu durumda bu iki rapor sonucunda sürekli müşteri ve kaybedilen müşteri arasındaki ince sınır çizgisini belirlemiş olduk. Buna gore termin uyum eşiğimiz %86, tahsis uyum eşiğimiz de %85 diyebiliriz.

### 12.4.4. Modelin Kullanılması

Veri madenciliğinin son aşaması, kurulan model sonucunda elde edilen kuralın geleceğe yönelik tahminlerde ve alınacak kararlarda kullanılmasıdır.

Yukarıda kurduğumuz modeli 2006 yılında edindiğimiz yeni müşterilere uygulayalım ve kaybedilme potansiyeli olan müşterilerimizi belirlemeye çalışalım.

Aşağıdaki rapor, 2006 yılında çalışmaya başladığımzı yeni müşterilerin, siparişlerine ait iş emirlerinin tahsis ve termin uyum yüzde değerlerini vermektedir.




| | Firmaadi | Aktivitesayisi | Tahsis Tarihine Uyan Aktivite Sayısı | Termin Tarihine Uyan Aktivite Sayısı | Tahsis Uyumu Yüzdesi | Termin Uyumu Yüzdesi |
|---|---|---|---|---|---|---|
| 1 | Target Tekstil Sanayi Ve Ticaret A.Ş | 7 | 5 | 5 | 71% | 71% |
| 2 | lazer mobilya | 35 | 31 | 31 | 89% | 89% |
| 3 | ELMAS GRUP LOJİSTİK TAŞIMACILIK | 6 | 6 | 5 | 100% | 87% |
| 4 | CMS OTOMOTİV DIS TİCARET VE SAN | 10 | 10 | 10 | 100% | 100% |

**Şekil 12.6** Yeni Kazanılan Müşterilerin Termin-Tahsis Uyumu

Bu rapor sonucuna gore 2006 yılı başından beri edindiğimiz yeni müşteriler arasında Target Tekstil hem termin hem de tahsis tarihlerine uyumda sadece %71'lik başarı gösterildiği ve bu değer bizim bulduğumuz eşik değerlerinin altında olduğu için potansiyel kaybedilecek bir müşteridir, diyebiliriz. Elde ettiğimiz bu bilgiden yola çıkarak bu müşterimizi kaybetmemek adına bu müşterinin önceliğini artırabiliriz. Örneğin, elimizde sadece tek bir araç var ancak biri Target Tekstil diğeri Elmas Gruba ait iki tane siparişimiz varsa bu tek aracı Target Tekstile vererek tahsis ve termin uyumlarını artırmaya ve bu müşterimizi sürekli müşterilerimiz arasına katmaya çalışabiliriz. Elmas Grup'taki başarı oranımız çok yüksek olduğu için bir iki siparişteki bu tür gecikmelerden zaten etkilenmeyecektir.

## 13. SONUÇLAR

Veri madenciliği ve bilgi keşfi, bilime, mühendisliğe, tıp sahasına, eğitime ve bilhassa ticari hayata yeni uygulamalar kazandıran bir disiplin olarak ortaya çıkmaya başlamıştır. Bilhassa dijital veri miktarında artış patlaması ve buna karşılık, bu veriler üzerinde araştırma ve uygulama yapan kişilerin sayısının değişmemesi, çalışmaları veri madenciliğine doğru zorlamıştır.

Ancak, veri madenciliği basite alınmayacak kadar ciddi bir çalışmadır. Bu konudaki çalışmalara hükümet ya da ticaret kollarının önde gelen kuruluşları destek vermeli ve gerekli imkanları sağlamalıdırlar. Özellikle veri yataklarının temini konusunda hassas davranan firmalar, bu konuda daha açık görüşlü olmaya davet edilmektedirler.

Veri madenciliğinin ileride varacağı hedef ve gereksinimler olarak şu an için yeni ve hızlı algoritmalar, akıllı sistemler, yapay sinir ağlarının bu konu ile uygulamaları, özel güvenlik mekanizmaları geliştirilmelidir. Ayrıca yeni veri modelleri üzerinde de çalışmalar yapılmalıdır.

Bu tez çalışmasında veri madenciliği yöntemlerinden biri kullanılarak lojistik sektöründe kritik problemlerden biri ele alınmış ve kurulan model yardımıyla bulunan kural geleceğe yönelik tahminlerde bulunmamıza imkan sağlamıştır. Yapılan bu çalışma şirket kurumsal portalinde yer alan ve şirket genelinin takip ettiği raporlar arasına konmuş ve ilgili kişilere yeni bir bakış açısı kazandırmıştır. Bu şekilde lojistik sektöründe dar boğaz yaratan bir çok konu veri madenciliği teknikleri kullanılarak araştırılabilir ve geleceğe yönelik olarak doğru karar vermemizi kolaylaştıracak kurallar çıkarılmaya çalışılabilir.

Elimizin altında bulunan büyük veri yığınları içerisinde çıplak gözle göremediğimiz ancak keşfedilmeyi bekleyen ve belki de şirketlere milyon dolarlar kazandıracak olan değerli bilgiler keşfedilmeyi beklemektedirler. Geleceğin, en azından yakın geleceğin, geçmişten çok fazla farklı olmayacağını varsayarsak geçmiş veriden çıkarılmış olan kurallar gelecekte de geçerli olacak ve ilerisi için doğru tahmin yapmamızı sağlayacaktır.

## KAYNAKLAR


**[1] Alpaydın, E.,** 2000, Ham Veriden Altın Bilgiye Ulaşma Yöntemleri, www.cmpe.boun.edu.tr/~ethem/files/papers/veri-maden_2k notlar.doc Boğaziçi Üniversitesi, İstanbul

**[2] Aydoğan, F.,** 2003, E-Ticarette Veri Madenciliği Yaklaşimlariyla Müşteriye Hizmet Sunan Akilli Modüllerin Tasarimi Ve Gerçekleştirimi, http://www.iticu.edu.tr/kutuphane/dergi/d3/M00041.pdf, Hacettepe Üniversitesi, Ankara

**[3] Dilly, R.,** 2004, Data Mining: An Introduction, www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html University of Ulster Jordanstown

**[4] Eker, H.**, 2001, Veri Madenciliği veya Bilgi Keşfi, http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538, Family Finans

**[5] Karakaş, M.,** 2005, Veri Madenciliği Üzerine, http://www.bilgiyonetimi.org/cm/pages/mkl-gos.php?nt=132

**[6] Özmen, Ş.,** 2001. İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını YenidenKeşfediyor, http://www.ceterisparibus.net/kongre/cukurova_5.htm, Marmara Üniversitesi, İstanbul.

**[7] Quinlan, J. R.,** 1986, Induction of decision trees. *Machine Learning*, **l**, pp 81-106.

**[8] Toktaş, P. ve Demirhan, M.B.,** 2004, Risk Analizinde Veri madenciliği Uygulamaları, http://yaem2004.cukurova.edu.tr/bildiriler/ , YA/EM-2004- Yöneylem Araştırması/Endüstri Mühendisliği XXIV Ulusal Kongresi, Gaziantep-Adana

**[9] Tunçsiper, B.,** 2005, Küreselleşme Sürecinde Veri Madenciliği Ve Ekonomik Kararlardaki Etkinliği Açisindan Bir Değerlendirme, http://bsy.marmara.edu.tr/TR/konferanslar/2005/2005tebligleri/12.doc Balıkesir Üniversitesi, Balıkesir

**[10] Vahaplar, A. ve İnceoğlu, M.,** 2000. Veri Madenciliği ve Elektronik Ticaret, http://www.bayar.edu.tr/bid/dokumanlar/inceoglu.doc, Ege Üniversitesi, İzmir.

## ÖZGEÇMİŞ

Bu tezi hazırlayan Sevcan TİRYAKİ, 01.09.1980 İstanbul doğumludur. Lise öğrenimini Pertevniyal Anadolu Lisesi'nde tamamladıktan sonra, 1997 yılında İstanbul Teknik Üniversitesi Matematik Mühendisliği Bölümü'nde lisans eğitimine başlamıştır. 2001 yılında lisans eğitimini tamamlayıp, aynı yıl içinde İstanbul Teknik Üniversitesi Sosyal Bilimler Enstitüsü İşletme Bölümü'nde ve Fen Bilimleri Enstitüsü Sistem Analizi Bölümü'nde yüksek lisans eğitimlerine başlamıştır. 2003 yılında İşletme yüksek lisans eğitim programından mezun olmuştur. 2002 yılında Teknosa İç ve Dış Ltd. Şti.'nde IT bölümünde çalışmaya başlamış, 2004 yılında Omsan Lojistik A.Ş.'de IT departmanına geçmiştir. Halen bu şirkette görevini sürdürmektedir.