

VERİ MADENCİLİĞİ VE BİR E-TİCARET UYGULAMASI

Anarbek KALIKOV

**YÜKSEK LİSANS TEZİ
ELEKTRONİK VE BİLGİSAYAR EĞİTİMİ**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**MAYIS 2006
ANKARA**

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

Anarbek Kalikov

VERİ MADENCİLİĞİ VE BİR E-TİCARET UYGULAMASI**(Yüksek Lisans Tezi)****Anarbek KALIKOV****GAZİ ÜNİVERSİTESİ****FEN BİLİMLERİ ENSTİTÜSÜ****Mayıs 2006****ÖZET**

Sayısal veri miktarının her geçen gün durmadan artması bu verilerin verimli analizi sorununu ortaya çıkarmaktadır. Birçok sayısal veri insanların fark edemeyeceği kadar az değişiklik gösterir ve bu değişikliklerin sebeplerinin ne olduğunu standart istatistiksel yöntemlerle anlatmak bazen mümkün olmamaktadır. Bazı durumlarda herhangi bir verinin değişikliği başka verilerin de ortaya çıkmasına yol açabilmektedir. Bu bağlantıları ortaya çıkarabilmek için klasik istatistiksel metotların yanı sıra yeni bağlantıları açık ve net gösterebilecek tekniklerin kullanılması gerekmektedir. Literatürde bu teknikler bütünü Veri Madenciliği(Data Mining) adıyla anılmaktadır.

Veri Madenciliğinin en önemli hedeflerinden biri anlamsız görünen veri yığınlarının arasındaki gizli bağımlılıkları, desenleri tespit etmek ve elde edilen bilgiyi yararlı bir sonuç üretmek için kullanmaktır.

Günümüzde en başarılı kurumlar Veri Madenciliğini vazgeçilmez araç olarak görmeye başlamışlardır. Veri Madenciliği günümüzde işletme, sağlık, emniyet, spor gibi çok çeşitli alanlarda başarılı bir şekilde kullanılmaktadır.

Bu tezde, Veri Madenciliği tekniklerini kullanarak e-ticaret amaçlı kurulan bir yayınevi web sitesinin veri tabanında tutulan verilerin analizleri yapılmıştır. Bu

teknikler, klasik istatistiksel tekniklerin yanı sıra Birliktelik Kuralları ve Karar Ağaçlarıdır. Tekniklerin uygulanması sonucunda, veri tabanında bulunan sanal ürünlerin(kitapların) kategorilerine göre doğru yerleştirilmesinde yardımcı olacak bilgiler keşfedilmiştir. Böylece hem sistem yöneticisi açısından hem de site üzerinden alışveriş yapacak kullanıcılara ilgi alanlarına göre kitap tavsiyesinde bulunacak ortam yaratılması amaçlanmıştır.

Bilim Kodu : 704 - 3.006
Anahtar Kelimeler :Veri Madenciliği, Web Madenciliği, Karar Ağaçları, Birliktelik Kuralları, E-Ticaret.
Sayfa Adedi : 94
Tez Yöneticisi : Yrd. Doç. Dr. Nursal Arıcı

DATA MINING AND AN E-COMMERCE APPLICATION**(M.Sc. Thesis)****Anarbek Kalykov****GAZI UNIVERSITY****INSTITUTE OF SCIENCE AND TECHNOLOGY****May 2006****ABSTRACT**

Amount of digital data grows rapidly daily and thus a problem of analysis of such a data occurs. Many of these data change in unnoticeable amounts and sometimes it's not possible to explain these changes by standart statistical methods. Sometimes change in data leads to change in some another data. To reveal these dependencies new techniques should be used among with classic statistical methods. All of these techniques are called Data Mining in literature.

One of the main targets of Data Mining is to discover unnoticed dependencies, patterns within collection of seemingly meaningless data and to use collected information in producing useful result.

Nowadays the most successful corporations consider Data Mining as an indispensable instrument. At present time Data Mining is successfully used in such areas as management, health, security, sports.

In this thesis analysis of data, stored in publisher's website was made, using Data Mining's common techniques. These techniques, among with classic statistical thechniques, are Association Rules and Decision Trees. As a result of using these techniques information, that should help in locating virtual products(books) to right kategories, was discovered. Thus, creation of

environment to advice boks to website users or system administrator, depending on their areas of interests, was intended.

Science Code : 704 - 3.006

Key Words : Data Mining, Web Mining, Decision Trees, Association Rules, e-Commerce.

Page Number : 94

Adviser : Assistant Prof. Dr. Nursal Arıcı

TEŞEKKÜR

Bu çalışmanın gerçekleşmesine katkılarından dolayı ve danışmanım olarak tezin yazılmasında yol gösterdiği için sayın hocam Yrd. Doç. Dr. Nursal Arıcı'ya içtenlikle teşekkür ederim. Ayrıca çalışmalarda kullanılan alışveriş verilerini sağlayan ve çalışmalarımız boyunca desteğini veren PegemA yayınevinin yöneticilerinden Servet Sarıkaya'ya teşekkürleri borç bilirim.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT	vi
TEŞEKKÜR	viii
İÇİNDEKİLER	ix
ÇİZELGELERİN LİSTESİ	xii
ŞEKİLLERİN LİSTESİ	xiii
SİMGELER VE KISALTMALAR	xv
1. GİRİŞ	1
2. VERİ MADENCİLİĞİ.....	7
2.1. Veri Madenciliğinin Farklı Tanımları	8
2.2. Veri Madenciliğinin Kullanım Alanları	9
2.3. Veri Madenciliği Uygulamalarından Örnekler	11
2.3.1. Araba arızalanmalarının incelenmesi	14
2.3.2. YSA kullanarak sıra dışı pazar davranışlarının önceden kestirilmesi	15
2.3.3. Yasal Veritabanından Birliktelik Kurallarının çıkartılması	15
2.3.4. Karar ağaçları kullanılarak toplu iflas etmeleri kestirme.....	16
2.3.5. K-Means tekniğini kullanarak turizm pazarının profillenmesi	16
2.3.6. Kütüphane kullanıcılarının erişim desenlerinin keşfi	16
2.3.7. Ağ madenciliğine ilişki	17

Sayfa

2.3.8. Birliktelik kuralları ve ardışık zamanlı örüntüler	17
2.3.9. Birliktelik kuralı madenciliği kavramlarına dayandırılmış sınıflandırma	19
2.3.10. Sepet analizi	20
2.4. Veri Madenciliğinde Kullanılan Yöntemler	21
2.4.1. En yakın k-komşu algoritması.....	23
2.4.2. Genetik algoritmalar	23
2.4.3. Lineer ve katlı regresyon.....	24
2.4.4. Kümeleme modelleri.....	27
2.4.5. Yapay sinir ağları(YSA)	37
2.4.6. Karar Ağaçları	38
2.5. Veri Madenciliği Süreci	39
3. YAZILIMIN TASARIMI	42
3.1. Yazılımın Tasarımı.....	42
3.1.1. Veri tanımı	42
3.1.2. Veri yapısı	45
3.2. Veriye Uygulanan Ön İşlemler	50
3.3. Birliktelik Kurallarını Tekniğiyle Alınan Bilgileri Değerlendirme	56
3.3.1. Sipariş tablosu için sıra dışı olan ürünler.....	60
3.3.2. Sepet tablosu için sıra dışı olan ürünler	61
3.4. Karar Ağacı Tekniği ve Bu Teknikle Hazırlanan Model	62

Sayfa

3.5. Karar Ağacı Tekniğıyle Alınan Bilgileri Değerlendirme	73
3.6. İstatistiksel Veriler	74
4. SONUÇLAR VE DEĞERLENDİRME	87
KAYNAKLAR	88
EKLER	91
EK-1 Apriori algoritması	92
EK-2 “Apriori Tid” algoritması	93
ÖZGEÇMİŞ	94

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1. “Apriori Tid” algoritması	32
Çizelge 2.2. Apriori algoritmasının aşamaları.....	36
Çizelge 3.1. “Birliktelik Kuralları” için giriş tablosu	51
Çizelge 3.2. “Pegem’A” ya uygulanan “Birliktelik Kuralları” yönteminin sonucu.....	56
Çizelge 3.3. Sepet için oluşturulan birliktelik kuralları	60
Çizelge 3.4. Karar ağacının tablo şeklinde gösterimi	72

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 2.1. Yapay sinir ağları	38
Şekil 3.1. “Pegem’ A VM Uygulama Programı” uygulamasının genel görünümü	43
Şekil 3.2. “Analysis Sevices” uygulamasının genel görünümü	43
Şekil 3.3. “ShoppingCart” tablosunun yapısı.....	45
Şekil 3.4. “ShoppingCart” tablosunun örnek verileri	46
Şekil 3.5. “Orders” tablosunun yapısı.....	46
Şekil 3.6 “Orders” tablosunun örnek verileri	47
Şekil 3.7 “Orderdetails” tablosunun yapısı	47
Şekil 3.8 “Orderdetails” tablosunun örnek verileri.....	48
Şekil 3.9 “Ürün” tablosunun yapısı	48
Şekil 3.10. “Ürün” tablosunun örnek verileri.....	49
Şekil 3.11. Oluşan tablo yapısı(OrderDetails Transactions).....	52
Şekil 3.12. Oluşan tablo yapısı(ShoppincCart Transactions	53
Şekil 3.13. Yeni veritabanı bağlantısını oluşturma.....	64
Şekil 3.14. Veri kaynağı belirtme	65
Şekil 3.15. Yeni VM modelini oluşturma	65
Şekil 3.16. Veri tabanı yapısını belirtme.....	66
Şekil 3.17. Tabloları belirtme	67
Şekil 3.18. VM yöntemini belirtme	67
Şekil 3.19. Tablolar arasındaki ilişkileri gösterme	68

Şekil	Sayfa
Şekil 3.20. Anahtar sütununu belirtme	69
Şekil 3.21. Giriş ve tahmin edilecek sütunları belirtme.....	70
Şekil 3.22. Modele isim verme.....	70
Şekil 3.23. Karar ağacı.....	71
Şekil 3.24. “Pegem’A VM Uygulama Programı” uygulamasının genel görünüm	75
Şekil 3.25. Toplam sipariş sayısı grafiği.....	75
Şekil 3.26. Toplam yapılan işlem sayısı grafiği	76
Şekil 3.27. Birliktelik kurallarını hesaplama paneli	77
Şekil 3.28. Sipariş tablosunun birliktelik kuralları	77
Şekil 3.29. Sepet tablosunun birliktelik kuralları	78
Şekil 3.30. En çok satılan 20 ürün ve onlara ilişkin satış sayıları grafiği	79
Şekil 3.31. 157 No’lu ürünün satış grafiği	80
Şekil 3.32. 157 No’lu ürünün satışlarına bağlı meslek dağılımı grafiği	81
Şekil 3.33. İlgi alanları grafiği	82
Şekil 3.34. İşlem sayısı grafiği	82
Şekil 3.35. Kitap tavsiye etme paneli	83
Şekil 3.36. Kitap tavsiye etme işleminin sonucu.....	84
Şekil 3.37. Meslek-ürün sayısı dağılımı grafiği	85
Şekil 3.38. Meslek dağılım grafiği	86
Şekil 3.39. İlgi alanları grafiği.....	86

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
VM	Veri Madenciliği
WVM	Web Veri Madenciliği
YSA	Yapay Sinir Ağları
BK	Birliktelik Kuralları
BA	Bayez Ağları

1. GİRİŞ

Günümüz teknolojisi hızla ilerlemekte ve her geçen gün gücü de artmaktadır. Bilgisayarların bilgi saklama kapasitelerinin artmasıyla birlikte bilgi kaydı yapılan alanların sayısı da artmaktadır(alışveriş kayıtları, bankacılık işlemleri, telefon kayıtları, uzaktan algılayıcı ve uydu verileri, devlet ve işletme yönetiminde yapılan işlem kayıtları v.s). Bilgisayar sistemlerinin her geçen gün hem daha ucuzluyor, hem de güçlerinin artıyor olması, bilgisayarda daha büyük miktarda verinin saklanabilmesine imkan vermektedir. Bunların hepsi veri fazlalığına yol açmaktadır [1,2]. “Sadece uydu ve diğer uzay araçlarından elde edilen görüntülerin saatte 50 Gygabyte düzeyinde olması bu artışın boyutlarını daha açık bir şekilde göstermektedir” [2].

“Dünyadaki en büyük işletme veri tabanlarının belirlenmesi amacı ile Winter Corporation tarafından yapılan bir araştırmada, Sears, Roebuck and Co.’nun sadece karar destek amaçlı kullanılan veri tabanının 1998 yılında 4630 gigabyte’a eriştiği görülmüştür” [2]. İnsanlar bu bilgilerin arasında çok azını kendileri görebileceklerdir. Bu veri seli ile ne yapılabilir? İşte bu veri fazlalığı kurumları bu verilerden nasıl faydalanılacağı ve bu verilerin anlamlı hale getirilmesi sorunuyla karşı karşıya bırakmaktadır [1]. Bilgisayar sistemleri ile üretilen bu veriler tek başlarına değersizdir, çünkü çıplak gözle bakıldığında bir anlam ifade etmezler. Bu veriler belli bir amaç doğrultusunda işlendiği zaman bir anlam ifade etmeye başlar. Bu yüzden büyük miktardaki verileri işleyebilen teknikleri kullanabilmek büyük önem kazanmaktadır. Bu ham veriyi bilgiye veya anlamlı hale dönüştürme işlemleri Veri Madenciliği(VM) ile yapılabilmektedir [1]. Veri madenciliği, bu gibi durumlarda kullanılan büyük miktardaki veri setlerinde saklı durumda bulunan örüntü ve eğilimleri keşfetme işlemidir [3].

SQL dilinin yetersiz kalması, Veri Tabanında Bilgi Keşfi(VTBK) adında yeni bir veri analizi yönteminin gelişmesinin nedenlerinden biri olmuştur. Veri Madenciliği(VM), bu analizin içinde bir aşamayı oluşturmaktadır. Ama son zamanlarda VTBK ile VM eş anlamlı olarak da kullanılmaktadır [2]. VTBK süreci,

çeşitli veri kaynaklarından verinin toplanması ile başlar. Bir sonraki adımda bu veriler analiz için uygun hale getirilir. Bu adımı Veri Madenciliği takip eder.

Genel anlamda, VM, ham verinin bilgiye dönüştürülmesidir [2]. VM, verinin incelenerek, gizli ve değerli bilginin keşfedilmesini sağlayan bir işlemdir [4]. VM, büyük miktardaki verinin içinden geleceği tahmin edilmesinde yardımcı olacak anlamlı, yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir [2]. Konunun uzmanlarından olan Piatetsky-Shapiro, VM’ni şöyle tanımlamaktadır: “Verilerden daha önceden bilinmeyen , zımnî, muhtemelen faydalı bilginin monoton olmayan bir süreçte çıkartılması”. Bu süreç farklı bir çok yaklaşımı kapsamaktadır, bunlardan bazıları:

- Kümeleme (Clustering),
- Veri özetleme (Data Summarization),
- Sınıflama kurallarının (Classification Rules) öğrenilmesi,
- Birliktelik Kuralları (Dependency Networks) bulunması,
- Değişikliklerin analizi (Analysing Changes) ve
- Anomali tespiti (Detecting Anomaly) ‘dir.

VM’nin temellerini yapay zeka ve istatistik disiplinlerindeki gelişmeler oluşturmaktadır. Veri Madenciliği, çeşitli alanlarda kendine yer bulmuş durumdadır ve her geçen gün bu alanlara yenileri eklemektedir. Bu alanların çeşitliliğini göstermek için bazı örnekler verilebilir [2].

- Makine Öğrenimi
- Uzman Sistemleri
- Veri Tabanları
- Optimizasyon
- Görselleştirme
- Yüksek performanslı paralel işlemciler v.s [2].

Günlük hayatımızda da VM birçok alandaki araştırmalarda çok önemli yer kazanmıştır. VM, astronomi, biyoloji, finans, sigorta, ulusal güvenlik v.s gibi alanlarda başarıyla kullanılmaktadır [2,5].

Son zamanlarda yaygın olarak kullanılan bir diğer araştırma alanı VM ile WWW(World Wide Web)'i birleştiren alandır. Bu alana Web Veri Madenciliği(WVM) denilmektedir. WVM içindeki araştırmalar Veri Tabanı, İstatistik, Yapay Zeka ve Görsellik gibi diğer alanlardaki uzmanlarının ilgilerini çekmektedir. WVM yaygın olmadığından, her ülkede farklı ve yeni yöntemler kullanılmaktadır. Bu yüzden bir çok WVM kavramı netlik kazanmamıştır. Ama yine de bu alandaki çalışmaları 3 ana başlık altında toplamak mümkündür [6]:

- Web içerik madenciliği
- Web yapı madenciliği
- Web kullanım madenciliği

Web içerik madenciliği, web'deki veri ve belgelerden faydalı bilgi bulmaya odaklanır. Web yapı madenciliği ise bağlantı(link) yapılarını bulmaya ve bunların arasındaki ilişkileri modellemeye çalışır. Ama bu iki alan arasındaki fark her zaman net değildir [6].

Web kullanım madenciliği diğer iki alana göre daha bağımsızdır. Web kullanım madenciliği konusunda şu çalışmalar sürdürülmektedir:Genel Web Kullanım Madenciliği, Site Değişikliği (Modifikasyonu), Sistem İyileştirme ve Kişiselleştirme. Büyük sitelerde her bir çalışma çok önemli sonuçlar etmeye yardımcı olabilir [6].

Genel Web Kullanım Madenciliği log dosyalarını inceleyerek içindeki desenleri bulmaya çalışır. Site Değişikliği (Modifikasyonu) sistemleri sitelerin içerik ve konu yapısında iyileştirmeler önererek site tasarımını iyileştirmeye çalışır. Sistem İyileştirme, web trafiğini iyileştirmek için web madenciliği yapar. Son olarak, kişiselleştirme sistemleri kullanıcılarının tercihlerini anlamayı ve web sitelerini kişisel hale getirmeyi amaçlar [6].

Günümüzün rekabetçi ortamında, kurumların etkili bir şekilde mücadele edebilme yetenekleri en uygun olanı belirleme ve seçmeye bağlıdır. Bu sebeple mevcut verileri incelemek ve öngörülen amaçlar doğrultusunda ilerlemek, kurumlara yüksek kalite, üretkenlik ve iş dağılımı konularında büyük faydalar sağlar. İnternet kullanımının artması ve her geçen gün sanal ortamdaki alışveriş sayılarının artması bu verilerin incelenmesi gerekliliğini arttırmaktadır, çünkü bu alışverişler bazen gerçek mağaza satışlarından da fazla olabilmektedir. Bu yüzden WVM uygulamaları da Türkiye’de de her geçen gün önemini arttıracaktır.

Tezin amacı günümüzdeki alışveriş sitelerinde VM Tekniklerinin uygulanmasının yararını göstermek ve gerçek bir sistem üzerinde bilgi analizi yapmaktır [Pegem’A web sitesi : <http://www.pegema.com.tr>]. PegemA yayınevi için hazırlamış olduğumuz web sitesi üzerinde VM teknikleri uygulanarak, yayınevi için faydalı olabilecek bilgileri elde etmektir. Ayrıca bu çalışma, Web kullanım Madenciliği Sistemi, Genel Web kullanım Madenciliği alanında esnek SQL(Yapısal Sorgulama Dili)sorgulama yeteneklerini kullanan veritabanı yaklaşımına bir örnek teşkil etmektedir.

PegemA Yayınevi web sitesinin alışveriş kayıtları, kullanıcı tabloları üzerinde VM ve veri tabanı analizi teknikleri uygulanacaktır. Çalışmada Birliktelik Kuralları(Association Rules) ve Karar Ağaçları(Decision Trees) teknikleri uygulanarak bunlardan yararlı sonuçlar elde edilmeye çalışılacaktır. Bunlara ek olarak sitedeki alışveriş kayıtlarının genel istatistiksel analizi yapılarak sitenin kullanımı hakkında ek sonuçlar elde etme hedeflenmektedir.

Tüm veri temizleme ve kayıt tanıma işlemleri C# ve VB.NET kod yazılımları ile yapılacaktır. Kayıtlar ön filtreleme(temizleme) işleminden geçtikten sonra Web kullanım Madenciliği ilişkisel veri tabanına aktarılmaktadır.

Veriden bilgi elde etmeye yönelik sorgulamalara ek olarak SQL’in esnek özelliklerini ve en sık Apriori algoritması uygulanmaktadır.

Tezin ilk bölümünde VM hakkında genel bilgiler verilerel uygulama alanları açıklanmaktadır.

İkinci bölümünde Veri Madenciliğinin tanımı yapılacak ve genel özellikleri hakkında bilgi verilecektir. Hangi VM yöntemleri hangi bilim dallarda daha yaygın olduğu ve uygulama alanları anlatılacaktır. Ayrıca Web VM hakkında bilgi verilecek ve VM ile yapılabilecekler açıklanmaktadır.

Üçüncü bölümde VM Teknikleri ele alınarak her birinin temel özellikleri açıklanacak ve çeşitli ülkelerde gerçekleşen başarılı VM kullanımları hakkında bilgi verilecektir. Aynı bölümde Birliktelik Kuralları'nın uygulandığı yazılım modülü hazırlanması açıklanmaktadır.

Bu yazılım modülü ile:

1. İnternet üzerinden satış yapan PegemA yayınevinin sepet ve sipariş tabloları incelenecektir. Böylece sistem yöneticisine hangi ürünlerin kategorisinin değiştirilmesi uygun olacağı konusunda faydalı bilgiler sağlanmaktadır.

2. Geleneksel istatistik metotlar kullanılarak sistemin genel verileri incelenecektir.

Bunlar:

- Sistemdeki kullanıcıların meslek, ilgi alanı dağılımları.
- Meslek ve ilgi alanlarına göre satış grafikleri.
- Belli bir ürün(kitap) için o ürünü satın alan ziyaretçiler hakkında bilgiler grafik şeklinde gösterilmektedir.

Üçüncü bölümde ayrıca Birliktelik Kuralları(BK) tekniği kullanılarak elde edilen bilgilerin bir değerlendirmesi yapılmaktadır. Aynı bölümde Microsoft Karar Ağaçları tekniği veriye uygulandığında elde edilen sonuçların değerlendirilmesi yapılmaktadır.

“Microsoft Analysis Services” uygulaması aracılığıyla :

“Karar Ağacı”(Microsoft Decision Trees) tekniği kullanılarak sisteme giriş yapan kullanıcının alışveriş yaparken hangi ödeme seçeneği(havale yada kredi kartı) kullanacağı konusunda tahminde bulunarak kurallar belirlenmektedir.

Dördüncü bölümde yapılan çalışmadan elde edilen sonuçlar değerlendirilerek öneriler yapılmaktadır.

2. VERİ MADENCİLİĞİ

Ağdaki sürekli çoğalan verilerden yararlı bilgi çıkarabilmek için o verilerin kazılması (analiz edilmesi) gerekmektedir [3]. Uydulardan her saniye gelen resimlerin analiz edilmesi gerek. Her gün siyasi ve ekonomik hayatta birbirine bağlı bir sürü olay gerçekleşmektedir. Bunların hepsinin incelenmesi ve bunlardan yararlı sonuçların elde edilmesi gerekiyor. VM ayrı sorgular vererek büyük miktarda olan veriden yararlı bilgi, desenler, ve eğilimleri (genelde önceden belli olmayan) çıkarabilmektir. Temel olarak, birçok kurum veri madenciliğinin amacı olarak geçmiş deneyimlere ve şimdiki eğilimlere dayanarak kurum kapasitesini geliştirmek, anormal desenleri bulmak ve geleceği öngörebilmek sayılır. Anlaşıyor ki bu teknoloji için bir ihtiyaç vardır. Günümüzde çok büyük miktarda günlük ve tarihi veriler depoda saklanmaktadır. Onun için gittikçe artan veri tabanlarını kullanarak karar vermekte zorluklar yaşanmaktadır. Ek olarak, veri birçok kaynaktan da gelmekte olabilir. Sonuçta planlamayı ve kurumun başka fonksiyonlarını desteklemek için veriyi analiz etmek gerekmektedir [3].

Veri Madenciliği, büyük miktarlardaki verinin içinden geleceği tahmin edilmesinde yardımcı olacak anlamlı ve yararlı bağlantı ve kuralların bilgisayar programlarının aracılığıyla aranması ve analizidir. Ayrıca Veri Madenciliği, çok büyük miktardaki verilerin içindeki ilişkileri inceleyerek aralarındaki bağlantıyı bulmaya yardımcı olan veri analizi tekniğidir [2].

VM aracılığıyla, büyük veri kümelerinden oluşan veritabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veritabanı teknolojisi ve çeşitli bilgisayar programları kullanılarak yapılır [2].

Veri madenciliği yerine her türlü terimler kullanılmıştır. Bazıları VM'yi belli olmayan bilgiyi çıkarma anlamına ve bazıları toplanan informasyondan bir sonuç çıkarmak anlamına gelmekte olduğunu belirtmektedirler. Belirli bir tekniğin veri madenciliğinin tekniği olmasını fark etmek ve belirlemek zordur. Örneğin,

bazıları istatistiğe ait analiz tekniklerini veri madenciliğinin tekniklerine ait olduğunu söyleyerek tartışır. Diğerleri buna karşı çıkarlar. Örnek olarak, veri madenciliğiyle, tıp mallarını satan şirket reklamcılığında belli skorlara baskı yaparak kendi ürünlerinin satışını artırabilir ve bir kredi bürosu uygun ve ödemelerinde kusur olmayan adayları seçerek kendi kayıplarını sınırlayabilir. Bunun gibi olaylar türlü gazetelerde yayınlandı. Ek olarak, VM ayrıca anormal davranışı bulmak için kullanılabilir. Örneğin, bir casus ajansı bu teknolojiyi kullanarak onun çalışanları arasında anormal davranışı belirleyebilir [3].

2.1. Veri Madenciliğinin Farklı Tanımları

Bazı veri madenciliğinin teknikleri tümevarım mantığı(inductive logic) proglamlamasını, makinevi öğrenmeyi v.b. kapsar. VM sınıflandırmayı, birliktelik ve verilerin arasındaki sıralamayı yapmak için kurallar bulmayı içerir. Temel olarak, birisi gözlenen örneklerden ve desenlerden çıkarılan bilgi olan hipotezlere varır. Bu desenler verilen sorgulardan elde edilir ve her sorgu önce verilen sorguya bağlıdır [3].

Gartner Grup tarafından yapılan 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 sürecidir [1,2,3,7].

VM, temel olarak bilgisayar destekli bir *bilgi çözümleme* işlemidir [5].

VM, ayrı sorgular vererek büyük miktarda olan veriden yararlı bilgi, desenler, ve eğilimler (genelde önceden belli olmayan) çıkarabilmektir [3].

VM, verinin sahibine anlamlı ve yararlı olacak şekilde veri kümesinin içinde şüphe uyandırmayan ilişkileri bulmak ve veriyi yeni bir şekilde özetlemek için veri kümelerinin incelenmesidir [7,8].

2.2. Veri Madenciliğinin Kullanım Alanları

Günümüzde VM teknikleri başta işletmeler olmak üzere çeşitli alanlarda başarı ile kullanılmaktadır. Bu uygulamaların en sık görülenleri ilgili alanlara göre aşağıda özetlenmektedir [1].

Pazarlama

- Müşterilerin satın alma örüntülerinin belirlenmesi
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması
- Posta kampanyalarında cevap verme oranının artırılması
- Pazar sepeti analizi (Market Basket Analysis)
- Müşteri ilişkileri yönetimi (Customer Relationship Management)
- Müşteri değerlendirme (Customer Value Analysis)
- Satış tahmini (Sales Forecasting) [2].
- Müşteri dağılımında
- Çeşitli pazarlama kampanyalarında
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında
- Çapraz satış analizleri
- Müşteri ilişkileri yönetiminde
- Çeşitli müşteri analizlerinde [1].

Bankacılık

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

- Risk yönetimi [2].

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespiti
- Riskli müşteri örüntülerinin belirlenmesi [2].

Perakendecilik

- Satış noktası veri analizleri
- Alış-veriş sepeti analizleri
- Tedarik ve mağaza yerleşim optimizasyonu
- Hisse senedi fiyat tahmini [1]
- Genel piyasa analizleri
- Alım-satım stratejilerinin optimizasyonu

Telekomünikasyon

- Kalite ve iyileştirme analizleri
- Hisse tespitleri
- Hatların yoğunluk tahminleri

Sağlık Ve İlaç

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

Endüstri

- Kalite kontrol analizleri
- Lojistik
- Üretim süreçlerinin optimizasyonu [2].

Bilim ve Mühendislik

- Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesi [1].

VM kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği, gözetleyicisine, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir [1].

2.3. Veri Madenciliği Uygulamaları

Veri madenciliğinde araştırma çok farklı disiplinlerin içinde uygulanmaktadır. Veri tabanını yöneten araştırmacılar veri madenciliğinin avantajını sorgu işleminden almaktadır. İlgi çeken alanlardan biri de sorgu işlemini genişletmek ve veri madenciliğini kolaylaştırmaktır. Veri depolama başka bir veri yönetiminin teknolojisidir ve bu ayrı veri kaynakları tamamlamakta, veriyi kolay bir şekilde organize edebilmektedir [3].

İstatistik analiz araştırmacıları VM için daha geliştirilmiş ve sofistike teknikleri üretmek için kendi tekniklerini makineyi öğrenme teknikleri ile bütünleştirmektedir [3].

İstatistiğe ait ayrı analiz paketleri şimdi veri madenciliğinin araçları olarak görülmektedir. Bunun üzerinde de bazı tartışmalar var ama yine de istatistik veri madenciliğine çok büyük bir katkıda bulunmaktadır [3].

Makineyi öğrenme demek ayrı gözlenen desenlerden kuralların öğrenilmesi ve sorunları çözmek için öğrenilen kuralları uygulamaktır [3].

Genellikle makineyi öğrenmede ve veri madenciliğindeki ilkeler hemen hemen aynı ve benzeridir ama VM deyince büyük miktardaki veri söz konusudur [3].

Bilgisayar canlandırmasında çalışan araştırmacılar veri madenciliğine başka bir perspektiften yaklaşmaktadır. Onların ilgi duyduğu alanların biri de bilgisayar canlandırmasındaki yaparak veri madenciliğine yardım etmek, onu desteklemektir. Yani etkileşim VM canlandırma toplumunun amacıdır [3].

Veri madenciliğinin asıl amacı veri yığınlarından anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmak olduğuna göre örnek birkaç kullanım alanı aşağıda özetlenmektedir [1]:

- İşletme alanındaki uygulamalar: Müşterilerin rakiplerine gitmelerini önlemek için bir şirket Veri Madenciliğine başvurabilir. Burada Veri Madenciliğinin amacı müşterilerin özelliklerini elde etmek ve rakip şirketlere gidebilecek müşterileri tespit etmek. Sonra da VM sonuçlarından faydalanarak onları kaybetmemek için strateji geliştirilebilir.

Şirketin müşterileriyle olan ilişkilerini daha anlamlı bir şekilde takip edilmesi ve sonuçlar çıkarılması. En çok şikayet alınan konular, ortalama şikayet çözüm süreleri, şirketin müşterilere en çok etkileşim içinde bulunduğu konular saptanabilir [9].

Ürün ve hizmetlerin tercih edilme sebeplerini ve hangi özelliklerin müşterileri ne kadar etkilediğini belirleme.

Banka müşterilerinin kredi durumları ve ödemeleri incelenerek aralarında riskli olan müşterilerin tespit edilmesi ve aynı risk grubuna düşebilecek diğer müşterilerin önceden tahmin edilebilmesi [4].

Ürün talebi bazında müşteri profillerini çıkararak, her profil için uygun ürün belirleyip ürün satışını arttırma.

Piyasada oluşabilecek değişikliklere mevcut müşterilerin vereceği tepkinin firma üzerinde yaratabileceği etkinin tespiti [1].

Karlı müşteriler tespit edilerek onlara özel kampanyalar uygulanabilir. Firma için zararlı müşteriler daha az masraflı hale dönüştürülebilir [1]. Örneğin havale ile ödemeyi tercih eden ama kredi durumu iyi olan müşterileri kredi kartlarını kullandırmaya yönelik kampanya başlatılabilir.

Bir ürün veya hizmetle ilgili bir kampanya programı oluştururken hedef kitlenin kim olacağının saptanması ve bu kampanyanın hedef kitleye hangi kanallardan sunulması daha uygun olacağını belirlemek için VM kullanılabilir [1].

“Firmanın makro ekonomik değişimler karşısındaki duyarlılığı ve oluşabilecek risklerin tespitinde kullanılabilir” [1].

- Eğitim Alanındaki uygulamalar:

- Öğrencileri performanslarına göre branşlara ayırma ve kişi-branş uyumunun arttırılması [10].

- Öğrencilerin performans ve memnuniyetlerinin arttırılması [11].

- Tıp alanındaki uygulamalar:

- Bir ilacın hangi yaş grup hastalarında nasıl etki yaratacağı konusunda tahminde bulunabilir.

- Yeni bulunan kanser tedavi yöntemi için en uygun adayı belirleme [7].

- Spor alanında:

- Bir basketbol takımına alınacak oyuncunun gelecekteki performansı ve takıma sağlayacağı yararlar tahmin edilebilir [7].

- Kütüphanecilik alanında:

- Bir müşterinin aldığı kitabı ne zaman getireceği ve bir sonraki gelişinde hangi kitabı seçeceği konusunda tahminler yapılabilir.

- Turizm alanında:

- Bir bölgeye turistlerin niçin geldiğini tespit ederek reklam kampanyası düzenlemek ve bir sonraki sezonda turist sayısını arttırmak için [7].

- Web alanında

- Kullanıcıların profilini belirleyip onlara uygun ürünlerin reklam kampanyalarını sayfalara koyma.

- En sık beraber ziyaret edilen çift sayfalar [11].

- Geçmişteki ve şu anki veriler analiz edilerek geleceğe yönelik tahminlerde bulunulabilir. Bu durum özellikle karlılık, ciro, pazar payı, hava durumu gibi analizlerde çok rahat kullanılabilir [1].

Yukarıda açıklanan uygulamalara ek olarak, VM tekniklerinin nasıl uygulanabileceğini daha iyi açıklayabilmek için literatür incelemesi yoluyla ulaşılan bazı çalışmalar da aşağıda özetlenmektedir:

2.3.1. Araba arızalanmalarının incelenmesi

DC (Daimler Chrycler) firması VM araştırmasında araba arızalarındaki bağımlılıklar incelemiştir. Bu bağımlılıkları incelemek için Bayez Ağları ve Birliktelik Kuralları algoritmalar en uygunudur. Bu sebeple DC araştırmacıları da bu yöntemleri tercih etmişlerdir.

İnceleme yaptıktan sonra araştırmacılar bazı sonuçlara varmışlardır. Bunlar:

Bazı yapı özelliklerinin beraber kullanılmaları, arabalarda kabloların elektrik sorunlarını çoğalttığı sonucuna varmışlardır.

Araştırmacılar bu kablo sorunlarının garajlarla bağlı olup olmadığını incelemişlerdir. Onların araştırmaları, Birliktelik Kurallarının “Eğer X garajıysa o zaman kablo sorunu” kuralı için güvenilirlik seviyelerinin garajdan garaja önemli ölçüde değiştiğini göstermişlerdir [7,12].

2.3.2. YSA kullanarak sıra dışı pazar davranışlarının önceden kestirilmesi

Ülke içindeki stokların çoğu anormal sonuçlar verdiği için, ülkedeki stok pazarlarının sonuçlarının kestirilebilmesi gerekmektedir.

Alan M. Safer adındaki araştırmacı, bu konuyu ele alarak incelemiştir. Önceden 343 kurumun 1993 - 1997 yılları arasındaki incelemeye alınmıştır . Başka araştırmacılar kurumlarının sırasını(rank) araştırma sonucunda büyük rol oynadığını düşünse de, Safer bu değişkeni araştırmasına katmamıştır.

Veri öğrenme (%80) ve onaylama (%20) kümeleri olarak ikiye ayrılmış ve Y.S.A. metodu kullanılarak VM yapıldığında aşağıdaki sonuçlar çıkmıştır. Belli endüstriler kestirilebilecek stok sonuçlarına sahipti. Bunlar:

- 36 endüstri grubu: elektronik donanım endüstrisi (bilgisayar sektörü hariç), 28 endüstri grubu: kimyasal ürünler, 37 endüstri grubu: ulaştırma donatımı, 73 endüstri grubu: business servisler.
- 9-12 ay arasında ve daha uzun süreler için stok kestirmeleri yapılabilirdi.
- Büyük çaplı firmalara göre küçük çaplı firmalar için kestirimler daha kolay yapılabilmekteydi [7,13].

2.3.3. Yasal veritabanından birliktelik kurallarının çıkartılması

Avustralya’daki Stranieri La Trobe Üniversitesinden bir ekip, Avustralya’nın maddi yardımı için başvurular hakkında bilgiler içeren veritabanında çalışmalar yapmışlardır. Bu araştırmanın ana hedefi legal hizmetin gereken yerlere daha kaliteli bir şekilde ulaştırılmasını sağlamaktır. Araştırmacılar 380.000 kaydı

incelemiştir. Bu veriler 1997-1999 seneleri arasındaki verileri içeriyordu. Veriler “18 altı”, “50 yaş üstü” gibi kümelerle ayrılmıştı. Bu araştırmadan bazı hipotezler ortaya atılmıştır.

- Hipotez A: Vietnamlı başvuranlar sadece suç yasası için başvuruda bulunmuşlardır.
- Hipotez B: Vietnamlı başvuranlar başka gruplara göre daha çok suç işlemiştir.
- Hipotez C: Vietnamlı erkekler kadınlara göre daha çok başvuruda bulunmuştur. Erkekler suç yasasıyla daha çok bağlantılı.
- Hipotez D: Vietnamlıların yardım kurumunun reklam materyallerine hazır ulaşımı yoktu [7,14].

2.3.4. Karar ağaçları kullanılarak toplu iflas etmeleri kestirme

Doğu Asya’daki krizden sonra, firmaların üretme kapasitesini ele alarak onların iflas etme oranını hesaplayan Karar Ağacı kullanılarak bir VM uygulaması yapılmıştır [7,15].

2.3.5. K-Means tekniğini kullanarak turizm pazarının profillenmesi

K-means sınıflama tekniği kullanılarak 1999 senesinde Kanada’daki Alberta eyaletine iç turistin arttırılmasını amaçlayan VM uygulaması yapılmıştır. VM sonucu 5 turist sınıfı oluşturulmuş ve her bir sınıf için ayrı bir reklam kampanyası düzenlenerek turist sayısında artışa ulaşılmıştır [7,16].

2.3.6. Kütüphane kullanıcılarının erişim desenlerinin keşfi

Web alışveriş sitelerinde kullanılan Web İçerik Madenciliği teknikleri kullanılarak (log dosyaları incelemeleri v.s), sanal kütüphane kullanıcılarının gezdikleri sayfa desenleri de çıkarılabilmekte ve okuyucular için daha iyi hizmet verilebilmesi yararlanılmaktadır. Erişim desenlerinin keşfi için en uygun olan yöntemlerden biri Birliktelik Kurallarıdır [17].

2.3.7. Ağ madenciliğine ilişki

Ağ veri yönetiminde ve makineyi öğrenmede araştırmacıların karşısına çıkan en önemli araştırma konularından biri Web Madenciliğidir. Ağda o kadar çok veri ve bilgi vardır ki kullanıcı için yararlı olan ve konuyla ilgili bilgiyi bulmak oldukça zor bir işlemdir. Ağları arama sıkıcı olmasıyla beraber kullanıcı fazla veriyle karşılaşılıyor. Burdan çıkan soru şu: bu kadar veriyi kullanıcı nasıl bilgiye çevirir ve sonra nasıl bunu yararlı bilgiye çevirir? Ayrıca, önceden belirsiz bilgiyi nasıl ağdaki veriden bulunabilir ve kurumun yararına nasıl faydalanılabilir? Başka bir deyişle, elektronik ticaret ağ madenciliğinden en çok yararlanan dallardan biridir [3].

2.3.8. Birliktelik kuralları ve ardışık zamanlı örüntüler

Birliktelik Kuralları büyük veri kümeleri arasındaki ilginç ilişkileri veya korelasyonları bulmak için kullanılır. Birliktelik Kuralları, verilen veri kümesi içindeki sıkça görülen özellik değer durumlarını tespit eder. Birliktelik kurallarının tipik ve genel kullanılan alanı market sepet analizidir(Market Basket Analysis) [18].

Örneğin marketlerde veri, barkod okuyucuları kullanılarak toplanır. Böyle ‘pazar sepetleri’ çok sayıda işlem(transaction) kayıtlarından oluşmaktadır. Her işlem kaydı bir müşterinin hangi ürünleri aldığı hakkında bilgiyi içerir. Pazar yöneticileri belli ürünlerin beraber alındıkları hakkında bilgi edinmek isteyebilirler. Onlar bu bilgileri stokların görünümünü düzeltmek(ürünleri birbirine göre optimal olarak yerleştirmek), çapraz satışlar, promosyonlar, katalog tasarımı ve alışveriş desenlerine göre müşteri kesimini belirlemek için kullanabilirler [18].

Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler veri madenciliğinde yaygın olarak kullanılmaktadır. Aynı zamanda bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır [19].

Aşağıda getirilen örneklerde görüldüğü gibi Birliktelik kuralları aynı zamanda gerçekleşen ilişkilerin tanımlanmasında da kullanılır.

- Müşteriler birayla % 75 ihtimalle patates cipsi de alır.
- Düşük yağlı peynir ve yağsız yoğurt alan müşteriler, %85 ihtimalle diyet sütünü de satın alır.

Ardışık zamanlı örüntüler ise birbirleri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır.

- X ameliyatı yapıldığında, 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 [2].

Birliktelik kuralları verinin potansiyel ilişkilerini tanımlar. Kurallar destek ve güvenilirlik seviyelerine göre karakterize edilir. Güçlü bir kural çok büyük bir destek ve yüksek seviyeli bir güvenilirliğe sahiptir. Farklı işlemsel veri kaynaklarından çıkarım, dönüştürme, temizleme ve güçlendirme ile kurallar türetilir. Aşağıda bir veri tablosundaki verinin destek ve güvenilirlik seviyelerinin hesaplanması gösterilmektedir [2].

Görüldüğü gibi bir birliktelik kuralı : $A_1, A_2, \dots, A_n$ olayları gerçekleştiğinde B olayı da gerçekleşmektedir. Bu durumları formülize ettiğimizde veri madenciliğine özel iki yeni terimle karşılaşırız. Bir kuralın desteği (Support) ve güvenilirliği (Confidence). Destek tüm kayıtlar içerisinde $A_1, A_2, \dots, A_n, B$ nesnelerinin beraber olma ihtimalidir. Güvenilirlik ise $A_1, A_2, \dots, A_n$ nesnelerini içeren bir kayıtlarla karşılaşıldığında B ile karşılaşma ihtimalidir. Bu birliktelik kuralının destek ve güvenilirlik seviyeleri: $\text{Destek} = \frac{\text{Kayıt Sayısı}(A_1 \cap A_2 \cap \dots \cap A_n \cap B)}{\text{Kaynak Verinin Kayıt Sayısı}} \times 100\%$ $\text{Güvenilirlik} = \frac{\text{Kayıt Sayısı}(A_1 \cap A_2 \cap \dots \cap A_n \cap B)}{\text{Kayıt Sayısı}(A_1 \cap A_2 \cap \dots \cap A_n)} \times 100\%$ [3].

2.3.9. Birliktelik kuralı madenciliği kavramlarına dayandırılmış sınıflandırma

Birliktelik kuralı madenciliğinden bazı düşünceler sınıflandırmaya uygulanabilir mi? Birliktelik kuralı madenciliği, VM araştırmasında önemli ve aktiflik derecesi yüksek bir alandır. Bu kesimde, kronolojik sıraya göre 3 metottan bahsedilecektir. İlk ikisi ARCS ve birliktelik sınıflandırması, birliktelik kurallarını sınıflandırma için kullanır; üçüncü metot CAEP, madencilik birlikteliklerinde kullanılan destek kavramlarını içeren “örnek ortaya çıkarma” madenciliğini yapar [19].

İlk metot kümeleme tabanlı birliktelik kurallarının madenciliğini yapar ve bu kuralları sınıflandırma için kullanır. ARCS (Association Rule Clustering System), birliktelik kurallarında madenciliği $A_{quan1} \wedge A_{quan2} \rightarrow A_{cat}$ formunda yapar. Burada A_{quan1} ve A_{quan2} , nicel nitelik aralıklarında test edilir ve A_{cat} elimizdeki verilerden bir kategorik nitelik için sınıf etiketi saptar, Birliktelik Kuralları iki boyutlu grid olarak işaretlenir. Algoritma gibi tarar, dikdörtgen kümeleme kurallarını arar. Bu yolla, nicel niteliklerin bitişik aralıkları bir kuralla meydana gelen bir kümeyle birleştirilir. Kümelenmiş birliktelik kuralları, sınıflandırmaya uygulanan ARCS’le genelleştirilir. ARCS’nin doğruluğu, veritabanının boyutu sabit bir hafıza alanı daha yüksek çalışma zamanını gerektirir [19].

İkinci metottan, birleşik sınıflandırma şeklinde söz edilebilir. İkinci metot, condset Y formundaki kurallarda madencilik yapar. Burada y bir sınıf etiketidir ve condset item’ler kümesidir. Eğer kurallar kümesi aynı condset’e sahipse kümeyi temsil etmek için en yüksek güvence değeri olası kural(possible rule - PR) olarak seçilir. İlk adım en sık ve doğru PR’lerin kümesini bulur. İkinci adım, heuristic bir metotla sınıflandırıcı kurar. Algoritma en uzun kuralın bulunması için veri kümesinin üzerinden birkaç sefer geçebilir. Ayrıca sınıflandırıcı kendi içinde başka bir kurala uydurulmamış herhangi yeni örnek için default bir sınıf tanımlanmamış, düşük önceliğe sahip default bir kural içerir [19].

Üçüncü metot, CAEP(Classification by aggregating emerging patterns) madenciliğini destekleyen itemset görüşünü kullanır. EP, bir veri sınıfından diğerine

doğru artan support değerli itemset'tir. İki supportun oranına EP'nin artış oranı denir [19].

CAEP EP'yi bir sınıflandırıcı yapılandırırken nasıl kullanır? Her C sınıfı için, EP's'e uygun verilen support ve artış oranı bulunur. Burada artış oranı, bütün C olmayan örneklerin kümesine uyularak hesaplanır. Sınır tabanlı(border-based) algoritmalar bu amaç için kullanılabilir. CAEP, C4.5 ve birliktelik-tabanlı sınıflandırmaya göre daha doğru bulunmuştur. Bu verinin hacmini ve boyutluluğunu arttırır. Alternatif bir sınıflandırıcı da JEP-sınıflandırıcısıdır. JEP, jumping-emerging-patterns(JEPs) üzerine kurulmuştur. İki sınıflandırıcı da tamamlayıcı olarak düşünülmektedir [19].

2.3.10. Sepet analizi

Sepet analizinde amaç kümenin(sayı kümesi,kelime kümesi) içindeki alanlar arasındaki ilişkileri bulmaktır. Bu ilişkilerin bilinmesi şirketin kârını arttırmak yada bazı bilimsel tahminlerde bulunmak için kullanılabilir. Eğer X ve Y malını alanların Z malını da yüksek oranla aldıkları biliniyorsa ve eğer bir müşteri X ve Y malını alıyor ama Z malını almıyorsa o potansiyel bir Y müşterisidir ve bu müşteriye Z malının reklamı yapılabilir. Eğer X'in Y ile beraber olma olasılığı yüksek ise ve bir bölgede Y bulunmuşsa o zaman bölgede X'in de aranması gerekir [20].

Örnek olarak internet üzerinden kitap satan Amazon şirketi [<http://www.amazon.com>] *BookMatcher* adlı programıyla müşterilerine okudukları ve sevdikleri kitaplara göre satın almaları için kitap tavsiyesinde bulunmaktadır [20]. Buna benzer modül geliştirdiğimiz projede de yapılmaktadır. Ancak, www.amazon.com'da kullanılan mantığın aksine siteye giren kullanıcıya kitap tavsiyesinde bulunmak için doğrudan onun ilgilendiği alanlar alınarak ve o alandaki satın alınan kitaplardan en fazla satış yapılan kitaplar tavsiye edilmektedir. Başka bir örnek : Süt satın alanlar genelde beraberinde şekeri de alırlar. Onun için bu iki ürün aynı rafta olursa daha iyi.

2.4. Veri Madenciliğinde Kullanılan Yöntemler

VM modelleri, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere ikiye ayrılmaktadır.

Tahmin edici modellerin amacı sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesidir. Burada tahmin etme, yargıya varma, sınıflandırma v.s benzer işlevleri görecektir şekilde ve çalıştırılabilir kod olacak bir model üretme amaçlanmaktadır [21]. “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” [1].

Tanımlayıcı modellerde amaç, büyük veri kümelerindeki desen ve ilişkileri tespit ederek, incelenen sistemin anlamını kavramaktır [21]. “25 yaş altı bekar kişiler ile, 25 yaş üstü evli kişiler üzerinde yapılan ve ödeme performanslarını gösteren bir analiz tanımlayıcı modellere örnek olarak verilebilir” [1].

VM modelleri gördükleri işlevlere göre, altı ana başlık altında incelenebilir:

- a. Sınıflama (Classification) – Veri birimini birkaç önceden tanımlı sınıflara bölen tahmin edici ve kendini öğreten fonksiyon tanımlama.
- b. Regresyon (Regression) – Veri birimini gerçek değerli tahmin değişkenine götüren tahmin edici ve kendini öğreten fonksiyon tanımlama.
- c. Kümeleme (Clustering) – Veriyi tanımlamak için sonlu (belli sayıdaki) kategori kümesini tanımlama görevi.
- d. Özetleme (Summarization) – Veri kümesi için kısa tanımlama bulma metotlarını içeren ek tanımlayıcı görevi.

- e. Bağlantı modelleme(Dependency Modeling) – Değişkenler veya veri kümesi yada veri kümesinin bir parçası içindeki özelliklerin değerleri arasındaki dikkate değer bağlantıları açıklayan yerel modeli tanımlama.
- f. Değişim ve sapma saptanması(Change and Deviation Detection) – Veri kümesi içindeki en büyük değişimleri belirleme [21].

Sınıflama ve Regresyon Modelleri: Sınıflama ve Regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Ancak bazı tekniklerde her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır. Burada , kayıtlar bazı anlamlı alt gruplar içine gruplanır [3]. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler:

- Genetik Algoritmalar (Genetic Algorithms),
- K-En Yakın Komşu (K-Nearest Neighbor)
- Bayesian Ağları(Naïve-Bayesian Network)
- Çoklu Regresyon, Lojistik Regresyon (Logistic Regression)
- Faktör ve Ayırma analizleri (factor and discriminant) [3].
- Özetleme (Summarization).
- Bağlantı modelleme (Dependency Modeling).
- Değişim ve sapma saptanması (Change And Deviation Detection) [21].

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) [1,2].

2.4.1. En yakın k-komşu algoritması

Bu yöntemin tercih edilme sebebi, sayısı bilinen veri kümeleri için hızlı ve verimli olmasıdır [7]. Aynı kaynak bu yöntemin turizm marketini profillemeye için kullanılan örneğini getirmektedir. Bu metotta eğitim örnekleri n -boyutlu sayısal niteliklerle tanımlanır. Her örnek n -boyutlu uzayda bir noktayı gösterir. Bu yolla, bütün eğitim örnekleri n -boyutlu örnek uzayda depolanır. Bu k -eğitim örnekleri bilinmeyen örneğe k en yakın komşudur. Yakınlık Euclidean uzaklığının terimlerinde tanımlanır. Euclidean uzaklığı $X=(x_1, x_2, \dots, x_n)$ ve $Y=(y_1, y_2, \dots, y_n)$ arasındadır.

Bilinmeyen örnek k -komşularının arasında en sık olan sınıfa verilir. $k=1$ olduğunda, bilinmeyen örnek uzayında en yakın eğitim örneğinin sınıfıdır.

En yakın komşu sınıflandırıcısı örnek-tabanlı ya da tembel-öğrenenlerdir(lazy-learners); bunlar bütün eğitim örneklerini depolarlar ve yeni bir örnek sınıflandırmaya gerek olmadıkça bir sınıflandırıcı yapılandırmazlar. Tembel-öğrenenler, potansiyel komşularının sayısı etiketlenmemiş örneklerin sayısına göre büyükse hesaplama maliyetini arttırabilir. Bu nedenle etkili indeksleme teknikleri gerektirir. Beklendiği gibi lazy-learning metotları eager metotlardan daha hızlıdır; fakat bütün hesaplamalar geciktirildiğinde sınıflanma yavaşlar. En yakın komşu sınıflandırıcısı kestirim içinde kullanılabilir [19].

2.4.2. Genetik algoritmalar

Genetik algoritmalar doğal evrim görüşlerini birleştirmeye çalışmaktadır. Genellikle genetik öğrenme şöyle olur: Rasgele oluşturulmuş kuralları içeren ilk populasyon(sayı kümesi) oluşturulur. Her kural, bir bit dizisi şeklinde gösterilir. Basit bir örnekle açıklamak gerekirse; eğitim kümesinde A_1 ve A_2 boolean(evete veya hayır) niteliklerinin ve $C_1 - C_2$ sınıflarının verildiğini varsayalım. Kural “IF A_1 AND NOT A_2 THEN C_2 ” bit dizisiyle 100 olarak ifade edilir. A_1 ve A_2 soldaki 2 bitle sınıfta sağdaki bitle gösterilir. Benzer şekilde “IF NOT A_1 AND NOT A_2 THEN C_1 ” kuralıda 001 şekline kodlanır. Eğer bir nitelik k ($k > 2$) değerlerine

sahipse, niteliklerin deęerleri k bitleri kullanılarak kodlanabilir. Sınıflar benzer şekilde kodlanabilir [19].

Yeni nesil (offspring), mutasyon ve crossover gibi genetik operatörleri uygulanarak yaratılabilir. Crossover’da kural çiftlerinden yarı diziler, yeni kural çiftleriyle yer deęiştirilir. Mutasyonda da ,kural dizisinden rasgele seçilmiş bitlerin tersi alınır(invert) [19].

Genetik algoritmalar, optimizasyon problemlerinde olduęu gibi sınıflandırma için de kullanılabilir. Veri madencilięinde, dięer algoritmaların uygunluęunu deęerlendirmek için kullanılabilirler [19].

2.4.3. Lineer ve katlı regresyon

Lineer regresyonda, veriler düz bir çizgi kullanılarak modellenir. Lineer regresyon, kestirim(prediction) çeşitlerinden en basit olanıdır. İki deęişkenli(bivariate) lineer regresyon rasgele deęerler üretir; Y ve lineer fonksiyonun dięer deęişkeni olan X . Lineer regresyon aşıęıdaki formülle tanımlanabilir:

$$Y = \alpha + \beta X \quad (2.1)$$

Bu fonksiyonda Y’nin bir sabit olması varsayılmaktadır; α ve β , sırayla Y’nin eğilimli ve durdurulabilir olmasını belirlemektedir. Bu katsayılar, asıl veride hataları en aza indirgeyen ve doęruyu deęerlendiren en küçük kareler yöntemiyle çözülebilir. $(x_1, y_1), (x_2, y_2), \dots, (x_s, y_s)$ formunda s tane örnek yada veri verilmiş olsun:

X ; $x_1, x_2, \dots, x_s$ ‘lerin ortalamasıdır. Y ; $y_1, y_2, \dots, y_s$ ‘lerin ortalamasıdır. A ve β dięer regresyon eşitliklerine göre daha iyi bir yaklaşım sunar [19].

Bulanık küme yaklaşımı(Fuzzy set approach)

Sınıflandırmada kullanılan kural tabanlı sistemler, sürekli niteliklerde ani kesimler içerdiğinden dezavantajlı olurlar. Örneğin, aşağıdaki örnek için müşteri kredi uygulama onayını değerlendirelim. Kural esasen, iki veya daha fazla yıldan beri çalışan ve yüksek bir maaşa sahip olan müşteriler için uygulamayı onaylamaktadır [19].

IF (years_employed $\geq$ 2) AND(income $\geq$ 50k)THEN credit ="approved" (2.2)

2.2 kuralına göre, bir müşteri 2 yıldan daha uzun süredir çalışıyorsa ve maaşı 50k 'nın üzerindeyse kredisi onaylanacak, ama maaşı 49k ise onaylanmayacaktır. Bulanık mantık ise sisteme "bulanık" başlangıç yada sınırların tanımlanması için izin verebilir. Kategori ve kümeler arasından kesin sınırlar yerine bulanık mantık kategoride verilen kesin değerin ilişki derecesini belirtmek için 0.0 ve 1.0 değerleri arasındaki gerçek değerleri kullanır. Bu nedenle 49k kadar maaş alan birini, 50k maaş alan biri kadar yüksek bir derece olmasa da derecelendirme kapsamına alınabilir [19].

Bulanık mantık, VM sistemlerinde sınıflandırma için yararlıdır. Kural-tabanlı sistemlerde bulanık mantığın kullanımı şunları içerir:

- Nitelik değerleri bulanık değerlere dönüştürülür.
- Bir örnekte sadece bir bulanık kural geçerlidir. Her uygun kural, kategorilerdeki ilişkilerden birini kullanır. Tipik olarak, her kategori için önceden belirlenmiş doğru değerler toplanır.
- Elde edilen toplamlar sistem tarafından bir değere dönüştürülür. Bu işlem, her kategorinin gerçek toplamı ölçülerek yada gerçek değerlerin çarpılmasıyla yapılabilir. Hesaplamalar, bulanık ilişki şekillerine dayanan karmaşıklıkları içerebilir [19].

Bulanık mantık sistemleri, sınıflandırmada sağlık, finans gibi birçok alanda kullanılmaktadır [19].

Kestirim(Prediction)

Süreklilik özelliğine sahip değerler kategorik tablodan alınmadan direk kestirilmek istenirse ne yapılmalı? İstatistik tekniklerinin regresyonu ile sürekli değerlerin kestirimi modellenir. Bu yöntemle 10 yıl tecrübesi olan adam için maaş tahmini yada yeni gelecek ürünün piyasadaki fiyatı kestirilebilir. Lineer olmayan bir fonksiyon, değişkenlere transformasyonlar uygulanarak rahat bir şekilde lineer fonksiyona dönüştürülebildiği için çoğu problem lineer regresyon metoduyla çözülebilmektedir. Kestirim problemleri için birçok yazılım paketleri bulunmaktadır. SAS, SPSS, S-Plus bunlardan birkaçıdır [19].

Duruma bağlı sonuçlandırma

Duruma bağlı sonuçlandırma(DBS) sınıflandırıcıların örnekleme tabanlı örneğidir(CBR). En yakın K komşu sınıflandırıcı yönteminde olduğu gibi eğitim örneklerini Euclidean uzayında noktalar olarak depolamak yerine CBR, örnek yada durumları karmaşık sembol tanımlamasıyla depolar. CBR ayrıca, mühendislik ve hukuk gibi alanlarında da uygulanabilmektedir. Sınıflandırılmak üzere bir durum verildiğinde DBS metodu öncelikle bu durum tanımlanmış bir eğitim durumu olup olmadığını kontrol eder. Eğer tanımlanmış bir durum bulunursa, duruma çözüm de eklenerek gönderilir. Tanımlanmış bir durum bulunamamışsa, duruma bağlı sonuçlandırıcı, yeni duruma benzer bileşenleri olan eğitilmiş durumları arar. Bu eğitim durumları yeni durumun komşuları olarak düşünülebilir. Eğer durumlar graph yapısında ise, yeni duruma benzer alt grafikler(subgraph'lar) aranır. DBS, arka planda mantıklı çözümler önermek için bilgi ve problem çözme stratejilerini çalıştırabilir [19].

DBS uğraştırıcı olan iyi bir benzerlik metriği bulma, eğitim durumları için etkili teknikler geliştirme, birleştirilmiş sonuçlar için metotları bulmayı içerir [19].

Diğer regresyon modelleri

Lineer regresyonun yaygın kullanılmasının nedeni basit olmasıdır. Lineer regresyon sürekli-değerli fonksiyonları modellemek için kullanılmaktadır. Genelleştirilmiş lineer modeller, hangi lineer modelin kategorik değişkenler için kullanılabileceğini teorik bir temel sunarak göstermektedir. Genelleştirilmiş lineer modelde Y lineer modeldeki gibi sabit değil, fonksiyonun değişkeni anlamındadır. En yaygın genelleştirilmiş lineer modeller: *logistic* ve *poisson* regresyonlarıdır. Logistic regresyon, predictor(kestirimci) değişkenler kümesinin lineer fonksiyon şeklinde oluşabilecek durumların olasılığını modeller. Toplam veri genellikle poisson dağılımını gösterir ve *poisson* regresyonu kullanılarak modellenir [19].

Log-lineer modeller, ayrık çok-boyutlu olası dağılıma yakınlaşmaktadırlar. Veri küp hücreleriyle birleştirilmiş olasılık değerlerini hesaplamak için kullanılır [19].

2.4.4. 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. 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 ayrıldığını geliştirilen bilgisayar programları da yapabilmektedir [1]. Bu modelin sonucu da desenleri gözlemleyerek sıklığı hesaplamadır [3].

Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Müşteriye daha fazla ürün satabilmek için, bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi gibi birçok yöntem vardır. Veri madenciliğinde pazarlama amaçlı olarak pazar sepeti analizi (Market Basket Analysis) adı altında müşterilerin

satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler kullanılmaktadır. Aynı zamanda bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının incelendiği alanlarda da önemli yere sahiptir [1,2]. Birliktelik Kurallarına sapma analizi de denebilir. Örneğin: John bankaya Pazar günü gitti ama sonra bakkala gitmedi. Onun yerine, o bir futbol oyununa gitti. Bu madencilik türüyle, düzensiz örnekler ve çelişkiler bulunur. Bu da bu tekniğin çok enteresan bir madencilik türü olduğu anlaşılır [3].

Birliktelik Kuralları, büyük veri kümeleri arasında birliktelik ilişkilerini bulmaya yarar [19,22-24]. Toplanan ve depolanan verinin her geçen gün gittikçe artmasının sonucu olarak, şirketler veritabanlarındaki birliktelik kurallarını ortaya çıkarmak istemektedir [19,22,25]. Şirketlerin karar alma işlemlerini daha verimli hale getirilmesindeki bir diğer unsur, büyük miktardaki mesleki işlem kayıtlarından ilginç birliktelik ilişkilerini keşfetmektir [19,26].

Pazar sepeti uygulaması, birliktelik kurallarının kullanıldığı en tipik örnektir. Müşterilerin yaptıkları alışverişteki ürünler arasındaki birliktelikleri bularak, pazar sepeti uygulaması müşterilerin satın alma alışkanlıklarını analiz eder ve müşterilerin hangi ürünleri bir arada aldıkları hakkındaki bilgiyi ortaya çıkarır. Daha sonra pazar yöneticileri bu bilgiyi daha etkili satış stratejileri geliştirmek için kullanabilirler [19,27].

“Örneğin bir müşteri süt satın alıyorsa, aynı alışverişte sütün yanında ekmek alma olasılığı nedir?” Rafları düzenleyen market yöneticileri bu tip bir bilgi ışığında ürünlerindeki satış oranı artırabilirler. Örneğin bir markette müşterilerin süt ile ekmeğin birlikte satın alınma oranı yüksekse, süt ile ekmek raflarını yan yana koyarak satışlar arttırılabilir [19,27].

Diğer bir örnek: “Bir A ürününü satın alan müşteriler aynı zamanda B ürününü de satın alıyorsa, bu durum Birliktelik Kuralı ile gösterilir:

$A=B$ [destek=2%,güven=60%]”

Buradaki destek ve güven kuralın ilginçlik ölçülerini gösteren ifadelerdir. Yukarıdaki kuralda : destek - keşfedilen kuralın kullanışlılığını, güven - doğruluğunu gösterir. Örnekteki %2 oranındaki bir Birliktelik Kuralının destek değeri, analiz edilen tüm alışverişlerden %2 sinde B ürününün A ile birlikte satıldığını belirtir [28-30]. %60 oranındaki güven değeri ise aynı alışverişte A ürününü satın alan müşterilerinin %60'ının B ürününü de satın aldığını belirtir [28-30]. Büyük veri tabanlarında birliktelik kuralları bulunurken, genelde aşağıda belirtilen adımlar takip edilir [18,19]:

1-Sık tekrarlanan küme üyeleri bulunur: Bu öğelerin her biri en az, önceden belirlenen minimum destek sayısı kadar sık tekrarlanırlar.

2-Sık tekrarlanan küme üyeleri güçlü birliktelik kuralları oluşturur: Bu kurallar minimum destek ve minimum güven değerlerini karşılamalıdır.

Sık tekrarlanan öğeleri bulmak için kullanılan en temel yöntem Apriori Algoritmasıdır [19,27-31].

Birliktelik kurallarını çıkarmak için hızlı algoritmalar

Bar kod teknolojisindeki gelişmeler pazarlama kurumlarına “sepet” verisi olarak adlandırılan büyük miktardaki satış verilerini toplama ve kaydetme imkanını verdi. Bu veriler genel olarak işlem tarihi ve işlem sırasında satın alınan mallar hakkında bilgiler içermektedir. Başarılı kurumlar bu tür veri tabanlarını pazarlama altyapısının önemli bir parçası olarak görmektedirler. Onlar veri tabanı teknolojisi tarafından yönetilen bilgi yönelimli pazarlama sürecini kurumlaştırmayla ilgileniyorlar. Bu yenilik pazarlama elemanlarına geleneksel pazarlama program ve stratejilerini geliştirmeye ve onları yerine getirmeye imkan sağlamaktadır [32,23].

BK'nın bir örneğini gösterelim: araba yedeklerini satın alan müşterilerin %98'i aynı zamanda araba bakım servisinden de faydalanırlar. Bunun gibi kuralları ortaya çıkarma pazarlama ve mesaj atma uygulamaları için çok önemlidir. Aynı zamanda

BK stokları düzenleme, satın alma desenlerine göre müşteri dağılımı, katalog tasarımı v.s gibi uygulamalarda kullanılmaktadır. Bu uygulamalarda kullanılan veri tabanlarının boyutları çok büyüktür [32].

Birliktelik Kuralları genel olarak aşağıda açıklanmaktadır:

$I = (i_1, i_2, \dots, i_m)$ parça(item) olarak adlandırılan harfler grubu olsun. D - her biri birer *parça dizisinden(itemset)* oluşan işlem dizisi olsun. Her bir işleme $T(Transaction)$ adı verilmektedir. Her bir işlemin $TID(Transaction ID)$ adı verilen işlem numarası vardır [32].

BK madenciliğinin amacı, D işlem dizisi verildiğinde, kullanıcı tarafından verilen ve *en az destek(misup)* ve *en az güven(minconf)* sayılarından daha büyük destek ve güven sayısına sahip tüm birliktelik kurallarını üretmektir. Burada D herhangi bir veri tabanı, veri dosyası veya bir veri tabanını sorgulama sonucu da olabilir [32].

Birliktelik Kurallarını bulma problemi iki alt gruba bölünebilir [32,34]:

1) İşlem desteği(*support*) en az'dan büyük olan tüm veri kümelerinin(*itemset*) bulunması gerekiyor. Veri kümelesi için destek sayısı(*support*) bu kümenin içerdiği parça sayısını göstermektedir. En az desteğe sahip veri kümeleri Büyük Veri Kümeleri(*large Itemsets*), diğerleri ise Küçük Veri Kümeleri(*small Itemsets*) olarak adlandırılmaktadırlar.

2) İstenilen kuralları üretmek için Büyük Veri Kümelerinin kullanılması gerek. Bunu yapmak için gereken işlemi açıklayalım:

Her bir büyük L veri kümesi için, içi boş olmayan tüm altkümeler bulunur. Bunun gibi her bir a altkümesi için eğer *destek(a)* en az *minconf*'a eşitse, o zaman $a \Rightarrow (L - a)$ şeklinde bir kural üretilir. Çok sonuçlu kurallar üretmek için tüm altkümeleri hesaba katmamız gerekmektedir.

Büyük veri kümelerini ortaya çıkarma

Büyük veri kümelerini ortaya çıkarma algoritmalarında veri üzerinden çok sayıda geçiş yapılır. İlk geçişte her bir verinin destekleri sayılır ve desteği en az olan büyük veri kümeleri ortaya çıkarılır. Sonraki geçişlerde bulunmuş büyük veri kümeleri ile başlama. Her geçişte bulunmuş kümeleri yeni aday veri kümesi dediğimiz daha da büyük kümeleri yaratmakta kullanırız ve bunların desteklerini sayarız. Geçiş sonunda en büyük aday veri kümeyi ortaya çıkarır ve sonraki geçişte kullanmak için hazır tutarız. Bu işlem yeni büyük veri kümesi bulununcaya kadar devam edilir [32].

Apriori ve AprioriTid algoritmalarının AIS [32,35] ve SETM [32,36] algoritmalarından farkı geçişlerde farklı aday veri kümelerinin sayılması ve onların yaratılmalarıdır. AIS ve SETM algoritmalarında aday veri kümelerinin yaratılması veri okuma sırasında yapılır. Ayrıca işlem okunduktan sonra önceki geçişte bulunmuş veri kümeleri işlemde belli olmuş olur. Yeni aday veri kümeleri önceki büyük veri kümeleriyle işlemdeki başka verilerle ile yaratılır. Bunun dezavantajı çok sayıda ve gereksiz aday veri kümelerini yaratmasıdır.

En önemlisi büyük veri kümenin alt kümesinin de büyük olması. Bu yüzden içerisinde K veri bulunan aday veri kümeleri içerisinde K- 1 veri bulunan büyük veri kümelerinin birleşmesi ile yapılır. Bu işlemde daha da az aday veri kümeleri yapılmış olur.

AprioriTid algoritmasının önemli özelliği ilk geçişten sonra veri tabanının hiç kullanılmamasıdır. Sonraki geçişlerde veri tabanına göre daha da az okunur daha sonra algoritmalarla detaylı anlatılır. Not: Her işlemde verilerin leksikografik şeklinde sıralandığı farz edilir. Veri taban normal haldeyken bu algoritmalar direk uygulanabilir ve veri tabandaki veriler $\langle TID, veri \rangle$ çift halde olduğu varsayılır. Veri kümesindeki veri sayısı kümenin büyüklüğü olur ve K büyüklüğündeki veri kümesine K-veri kümesi denir. K-veri kümesini $C(1)C(2)...C(K)$ olarak gösterilir. C her bir veriye işaret eder ve $C(1) < C(2) \dots C(k)$. Her bir veri kümesinin kendine ait

destek yüklenecek yeri olur ve veri kümesi yeni yaratılınca orası sıfırdır. Tablo 3.1’de algoritmanın özetini görülebilir. C k kümesi AprioriTid kümesidir.

Apriori algoritması

Apriori Algoritması Çizelge 2.1’de gösterilmiştir. İlk geçişte algoritma sadece bileşenleri sayarak büyük L - veri kümesini ortaya çıkarır, sonraki K’inci geçişinde iki işlem yapılır. Lk – veri kümeleri (K-1) – veri kümelerini kullanarak yapılır. 2 K – aday veri kümelerinin desteği veri tabanını tarayarak sayılır. Hızlı sayabilmek için t işlemindeki K – veri kümelerinin adaylarının hızlı bir şekilde ortaya çıkarılması gerekmektedir.

Çizelge 2.1: “Apriori Tid” algoritması

K-Veri kümesi	K verisini kapsayan veri kümesi
Lk	k-veri kümesinin büyük seti (azınlık destekli olanı) Bu setin her üyesinin iki alanı vardır: i)veri kümesi ve ii)destek sayısı
Ck	k-adayları’nın veri kümesinin seti (özellikle büyük veri kurumları). Bu setin her üyesinin iki alanı vardır: i)veri kümesi ve ii)destek sayısı
Ck	k-adayları’nın veri kümesinin seti ancak TIDin aday’larla birlikte ortak işlem yapması sonucudur.

Apriori aday yaratma

Apriori – gen fonksiyonu ($K-1$) - veri kümelerinin kümesini, $L_k - 1$, argüman olarak alır ve tüm büyük K – veri kümelerinin üst kümesini döner. Fonksiyon böyle çalışır. İlk önce birleştirme kısmında $L_k - 1$ ile $L_k - 1$ ile birleştirilir [32].

Insert into C_k

select p.item1, p.item2,...,p.item k-1, q.item k-1

from L_{k-1} with L_{k-1} q

where p.item1 = q.item1,...,p.item k-2 = q.item k-2, p.item k-1 < q.item k-1

Budama kısmında $L_k - 1$ de olmayan ($K - 1$) alt kümelerin tümünü siliyoruz.

forall itemsets $c \in C_k$ do

forall (k-1)-subsets s of c do

if ($s \notin L_{k-1}$) then

delete c from C_k

Örnek $L_3 = \{\{1\ 2\ 3\}, \{1\ 2\ 4\}, \{1\ 3\ 4\}, \{1\ 3\ 5\}, \{2\ 3\ 4\}\}$. Birleştirilme kısmından sonra $C_4 = \{\{1\ 2\ 3\ 4\}, \{1\ 3\ 4\ 5\}\}$. Budama kısmında $\{1\ 3\ 4\ 5\}$ silinir çünkü $\{1\ 4\ 5\}$ L_3 içerisinde yoktur. Bu algoritmaların K geçişlerinde veri tabanın T işlemi okunur ve T de olan $L_k - 1$ büyük veri kümeleri ortaya çıkarılır, sonra her büyük veri kümeleri T de bulunan büyük verilerle büyütülür ve leksikografik olarak sıralanır. $\{1\ 2\ 3\ 4\ 5\}$ işlemine bakılırsa dördüncü geçişte AIS ve SETM $\{1\ 2\ 3\ 4\}$ ve $\{1\ 2\ 3\ 5\}$ 'leri $\{1\ 2\ 3\}$ veri kümesini ekleyerek yaratır. Böylece üç tane daha aday veri kümeleri yaratılır ve hepsi beş tane olur. Ama Apriori sadece $\{1\ 3\ 4\ 5\}$ yaratır.

Doğruluk: Büyük veri kümesinin bütün alt kümelerinin destekleri de en az olması lazım. Biz $L(k-1)$ deki her veri kümelerini olabilecek bütün veriler ile büyüttüğümüz için ve sonra ($K-1$) – alt kümeleri $L(k-1)$ de olmayanların tümünü sildiğimiz için L_k da sadece üst küme kalır. Birleştirme işlemi aynı anda $L_k - 1$ bire veri tabandaki her veri ile büyötmek ve sonra L_k -1de olmayan ($K-1$) verileri silerek oluşturulan ($K-1$) – veri kümelerini silme demektir. p. item ($K-1$) < q.item($K-1$) denksizliği

benzerlerin yaratılmalığını ispat eder. O yüzden butama aşamasında Lk bulunan veri kümeleri silinmemektedir [32].

Değişkenlik: Tek geçişte değişik büyüklükteki adayları sayma.

K'ncı geçişte K büyüklüğündeki aday'ları sayarken onlardan oluşan K+1 adayları da sayılabilir. Bu işlem daha sonraki geçişlerde veri tabanı tarama zorluğunu gidermek için kullanılır [32].

Alt küme fonksiyonu

Aday veri kümeleri hash ağacında(hash tree) saklanır. Hash ağacı düğümü veri kümelerinin listesini tutar veya alt düğümler tablosunu tutar. Hash ağacı düğümünün derinliği 1'dir. d derinliğindeki düğüm d+1 derinliğindeki düğümleri gösterir. Veri kümeleri yapraklarda saklanır. C veri kümesi eklediğimizde ilk düğümden başlayarak yapraklara ulaşmaya kadar iner. d derinliğine gelince ne tarafa ineceğini bulmak için her fonksiyon denenir. Her düğüm ilk önce yaprak olarak oluşur. Yaprakların sayısı bir sınırı aşınca yaprak düğümü alt düğüme dönüşür. İlk düğümden başlayarak alt küme fonksiyonu bütün *aday'ları* böyle bulur. Yaprakların hangisi T işleminde varsa onların cevap kümesine referans eklenir. Eğer alt kümedeysek t'nin içerisinde olduğu sürece hash yapmaya devam edilir. İlk düğümden her t için hash yapılır. Her C veri kümesinin T işleminde olması için C'nin ilk verisinin t'de olması gerekir. İlk düğümden her t verisine hash yaparak t de olmayan verilere dokunulmadığına emin olunur. Bütün derinliklere aynı şey geçerlidir. Ek olarak veri kümedeki her verinin sıralanmış olduğu için her i verisine hash yaparak geldiği düşünülürse t'de olan ve sadece i'den sonra gelen verilere bakılması gerekir [32].

AprioriTid algoritması: EK-1'deki AprioriTid Algoritması da Apriori – gen fonksiyonunu kullanır ve kullanarak geçişten önce aday veri kümelerine bakar. Bu algoritmanın en önemli tarafı ilk geçişten sonra veri tabanı destek saymak için kullanılmayıp ve onun yerine Ck kullanılır. Ck'nın her bir elemanı $\langle TID, \{X_k\} \rangle$

formunda olur, $k=1$ için $C1$ veri tabanı işaret eder, $k > 1$ için (10ncu adım) algoritması ile yapılır. Eğer t işleminde hiç k –*aday* k veri kümesi yoksa C_k 'nın bu işleme hiç girişi olamaz o yüzden C_k 'nın giriş sayısı veri tabanındaki işlem sayısından az olabilir ve çoğunlukla k sayısı büyük olunca olur. Ek olarak k büyüyünce her giriş işleminden küçük olabilir çünkü işlemde çok az aday kalmış olabilir ama k küçükken C_k bütün *candidate* k – veri kümelerini işlem içinde barındırdığı için işleminden büyük olabilir [32].

Örnek: Örnek olarak Çizelge 2.2'de gösterilen tabloyu ele alalım. İstenilen destek sayısı 2. 4'ncü adımda $L1$ 'i değişken olarak “Apriori-Gen” algoritmasını çalıştırdığımızda sonuç olarak $C2$ 'yi elde ederiz. 6-10 adımlar arasında $C2$ 'nin içindeki adayların destek sayısını hesaplamak için $\hat{C}1$ 'in içindeki kayıtları tek tek kontrol ederek $\hat{C}2$ 'yi elde ederiz. $\hat{C}1$ 'deki ilk kayıt $\{\{1\}\{3\}\{4\}\}$ 'dür, bu da “100” no'lu işleme denk gelmektedir. 7'nci adımdaki bu kayda denk gelen $C1$ $\{\{1\}3\}\}$ 'dür, çünkü aynı zamanda $\{\{1\}3\}\{1\}\}$ ve $\{\{1\}3\}\{3\}\}$ $C2$ 'nin ve aynı zamanda $t.set-of-itemsets$ 'in üyeleridir.

$L2$ 'yi değişken olarak “Apriori-Gen”de kullandığımızda $C3$ 'ü elde ederiz. $\hat{C}2$ 'yi ve $C3$ 'ü kullanarak veriyi işleminden geçirdiğimizde $\hat{C}3$ 'ü elde ederiz. $\hat{C}3$ 'te TID'leri “100” ve “400” olan kayıtlarının olmamasının sebebi TID'si “100” ve “400” olan işlemlerin $C3$ 'ün üyelerini(sayılarını) içermemesidir. $C3$ 'teki $\{\{2\}3\}5\}$ adayı $L3$ 'ün en büyük ve tek üyesi olarak elimizde kalmaktadır. $L3$ 'ü kullanarak $C4$ 'ü elde etmeye çalıştığımızda $C4$ boş çıkar ve hesaplama sonuna gelir.

Çizelge 2.2 Apriori algoritmasının aşamaları

Veri Tabanı		$\hat{C}_1$	
TID	Items	TID	Set-of-itemsets
100	1 3 4	100	{{1}, {3}, {4}}
200	2 3 5	200	{{2}, {3}, {5}}
300	1 2 3	300	{{1}, {2}, {3}, {5}}
400	2 5	400	{{2}, {5}}

L1		C2	
Itemset	Support	Itemset	Support
{1}	2	{1 2}	1
{2}	3	{1 3}	2
{3}	3	{1 5}	1
{5}	3	{2 3}	2
		{2 5}	3
		{3 5}	2

$\hat{C}_2$		L2	
TID	Set-of-itemsets	Itemset	Support
100	{{1 3}}	{1 3}	2
200	{{2 3}, {2 5}, {3 5}}	{2 3}	2
300	{{1 2}, {1 3}, {1 5}, {2 3}, {2 5}, {3 5}, {2 5}}	{2 5}	3
400	{{2 5}}	{3 5}	2

C3		$\hat{C}_3$	
Itemset	Support	TID	Set-of-itemsets
{2 3 5}	2	200	{{2 3 5}}
		300	{{2 3 5}}

L3	
Itemset	Support
{2 3 5}	2

Veri biçimi

Her bir adayın veri kümelerine özel numara vermekteyiz ve ID olarak adlandırılmaktadır. C_k olan her bir adayın veri küme seti, ID tarafından düzgün sıra şeklinde tutulmaktadır. C_k üyesi formu: $\langle TID, \{ID\} \rangle$ şeklindedir. Her bir C_k sıra biçiminde biriktirilmiştir [32].

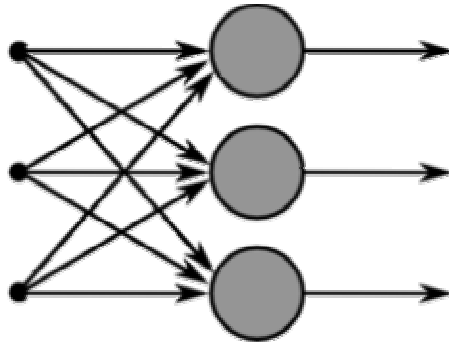
Apriori-gen fonksiyonu iki geniş $(k - 1)$ -veri kümelerini birleştirerek k -veri kümesi olan C_k 'yı oluşturmaktadır. Bu sırada iki candidate veri kümeler için iki farklı alan sunmaktayız: i) üretici ve ii) uzatma. Aday'ın üretici alanı iki büyük $(k-1)$ -veri kümelerinin ID'lerini biriktirmektedir ve birleştirmesi C_k 'yı meydana getirmektedir. C_k veri kümesi uzatma alanında bütün $(k+1)$ - aday'ların ID'lerin biriktirmektedir [32].

Şimdi yukarıda belirtilmiş olan veri biçimlerini kullanarak Şekil 2.2'nin 7'nci basamağının açıklayalım. C_{k-1} 'de t geçişinin alanı $t.set-of-itemsets$ tüm $(k-1)$ aday'ların ID'lerini $t.TID$ işlem sırasında vermekte olduğunu hatırlayalım. Bütün buna benzer C_{k-1} adaylarına büyüme alanı T_k 'leri verir. Bunlar bütün C_{k-1} büyüme alanından k -veri kümelerdir. Her C_k 'deki T_k 'ye üretim alanı iki veri kümesinin ID'lerini verir ve C_k 'yı ilerletmiş olur. Eğer bu veri kümeleri $t.set-of-itemsets$ geçitinde bulunursa biz sadece C_k $t.TID$ işleminde bulunduğunu sonuçlandırıp ve C_k 'yi C_t 'ye ekleyebiliriz [32].

2.4.5 Yapay sinir ağları(YSA)

Yapay Sinir Ağları, insanlığın doğayı araştırma ve taklit etme çabalarının en son ürünlerinden bir tanesi olan teknolojidir. 1980'lerden itibaren yaygınlaşan ve Yapay Sinir Ağları(artificial neural networks) adı verilen programlar, basit biyolojik sinir sisteminin çalışma şeklini simge etmek(canlandırmak) için tasarlanmışlardır [10]. Yapay sinir ağlarında amaç fonksiyon birbirine bağlı basit işlemci ünitelerinden oluşan bir ağ üzerine dağıtılmıştır. Bu ünitelere “nöron” ismi verilmektedir. Yapay sinir ağlarında kullanılan öğrenme algoritmaları veriden üniteler arasındaki bağlantı

ağırlıklarını hesaplar. Bu ağlar öğrenme, hafızaya alma ve veriler arasındaki ilişkiyi ortaya çıkarma kapasitesine sahiptirler. YSA istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz yani uygulama alanı daha genişir ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmez [2]. Biyolojik özelliği taşıdığı için, yani insanın beynindeki sinir ağlarının modelini oluşturabildiği için YSA bir insanın düşünme ve gözlemlemeye yönelik doğal yeteneklerini gerektiren problemlere çözüm üretebilmektedir [9].



Şekil 2.1 Yapay sinir ağı

2.4.6 Karar ağaçları

Yapay sinir ağlarında olsun istatistiksel yöntemlerde olsun veriden fonksiyon çıkarıldıktan sonra bu fonksiyonun insanlar tarafından anlaşılabilir bir kural olarak yorumlanması zordur. Karar ağaçları ise veri oluşturulduktan sonra ağ kökten yaprağa doğru inilerek kurallar (Eğer-O zaman kuralları(“if-then rules”)) yazılabilir. Karar ağaçlarında kök ve her düğüm bir soruyla etiketlenir [2]. Düğümlerden ayrılan dallar ise ilgili sorunun olası yanıtlarını belirtir. Her dal düğümü de söz konusu sorunun çözümüne yönelik bir tahmini temsil eder [11]. Bunun gibi kural çıkarma, VM çalışmasının sonucunu doğrulamak için kullanılır. 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 analiz bilgi verir ve daha sonraki analizler için yol gösterici olabilir [2].

Karar ağaçları (decision trees), bir tahmin tekniğidir . Genelde sınıflandırma, kümeleme ve tahmin modellerinde kullanılır ve sorunla ilgili araştırma alanını alt gruplara ayırmak için kullanılır.

Karar ağacı modeli, üç bölümden oluşur:

1. Karar ağacı.
2. Ağacı oluşturacak bir algoritma.
3. Ağacı veriye uygulayacak ve söz konusu sorunu çözecek bir algoritma .

Karar ağaçları eğitici örnekteki veriyi sınayan bir algoritma aracılığıyla gerçekleştirilir yada alanın bir uzmanı tarafından oluşturulur. Oluşturma tekniğine bağlı olarak karar ağaçları birbirlerinden ayrılır [5].

2.5. Veri Madenciliği Süreci

Üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda hiç bir VM algoritmasının fayda sağlaması mümkün değildir. Bu sebeple, VM sürecine girilmeden önce, başarının ilk şartı, iş ve veri özelliklerinin detaylı analiz edilmesidir. VM sürecinde genelde aşağıdaki adımlar izlenir: [19].

- Problemin Tanımlanması ve Hipotezin Açıklanması,
- Verilerin Toplanması
- Verilerin Hazırlanması
- Modelin Belirlenmesi [21]
- Modelin Kurulması ve Değerlendirilmesi
- Modelin Kullanılması
- Modelin İzlenmesi [19]

Problemin tanımlanması ve hipotezin açıklanması : VM’de başarılı olmanın en önemli şartı, projenin hangi kurum için yapılacağıının tanımlanmasıdır. Elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir [1].

İş probleminin nasıl bir sonuç üretilmesi durumunda çözüleceğinin, üretilen sonucun fayda - maliyet analizinin kurum için değerinin doğru analiz edilmesi gerekmektedir. Analizin proje için yeterlilik düzeyinin ve kurum konusu hakkındaki iş süreçlerinin de iyi analiz edilmesi gerekmektedir [1].

Verilerin hazırlanması : Verinin hazırlanması veri madenciliğinin toplam zaman ve enerjisinin %50 - %75 ini harcamasına neden olmaktadır. Bu aşamada kurumun mevcut bilgi sistemleri sayısal bilginin, veriler ile iş problemi arasındaki ilişki unutulmamalıdır. Proje kapsamında kullanılacak sayısal verilerin, hangi iş süreçleri ile yaratıldığı da analiz edilmelidir.

Verileri hazırlama aşaması toplama, birleştirme, temizleme ve dönüştürme adımlarından meydana gelmektedir [1].

Toplama: Tanımlanan problem için gerekli olan verilerin ve veri kaynaklarının belirlenmesi adıımıdır. Verilerin toplanmasında, nüfus sayımı, hava durumu, merkez bankası kara listesi gibi veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir [1].

Birleştirme ve temizleme: Toplanan verilerdeki farklılıklar giderilir. Analizin yanlış yönlendirilmesine sebep olabilecek veriler temizlenir. Genellikle yanlış veri girişinden kaynaklanan veriler kontrol edildikten sonra veri kümesinden atılır. Basit yöntemlerle sorun giderme işlemlerinin, büyük sorunların kaynağı olacağı unutulmamalıdır [1].

Dönüştürme : Gösterim şeklinde kullanılacak model ve algoritma verileri tanımlama. Örneğin kredi riski uygulamasında iş tiplerinin, gelir seviyesi ve yaş gibi değişkenlerin kodlanarak gruplanması faydalı olacaktır [1].

Modelin kurulması ve değerlendirilmesi: Tanımlanan problem çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, yinelenen bir süreçtir [1].

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 bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile doğruluk oranı hesaplanır. (Doğruluk Oranı = 1 - Hata Oranı) [1].

Önemli diğer bir değerlendirme kriteri modelin anlaşılabilirliğidir. Bazı uygulamalarda doğruluk oranlarındaki küçük artışlar çok önemli olsa da, bir çok kurumsal uygulamada ilgili kararın niçin verildiğinin yorumlanabilmesi çok daha büyük önem taşıyabilir. Genel olarak karar ağacı ve kural temelli sistemler, model tahmininin altında yatan nedenleri çok iyi ortaya koyabilmektedir [1].

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 kurumsal uygulamalarda doğrudan kullanılabileceği gibi, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir [1].

Modelin izlenmesi: Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir[1].

3. YAZILIMIN TASARIMI

3.1. Yazılımın Tasarımı

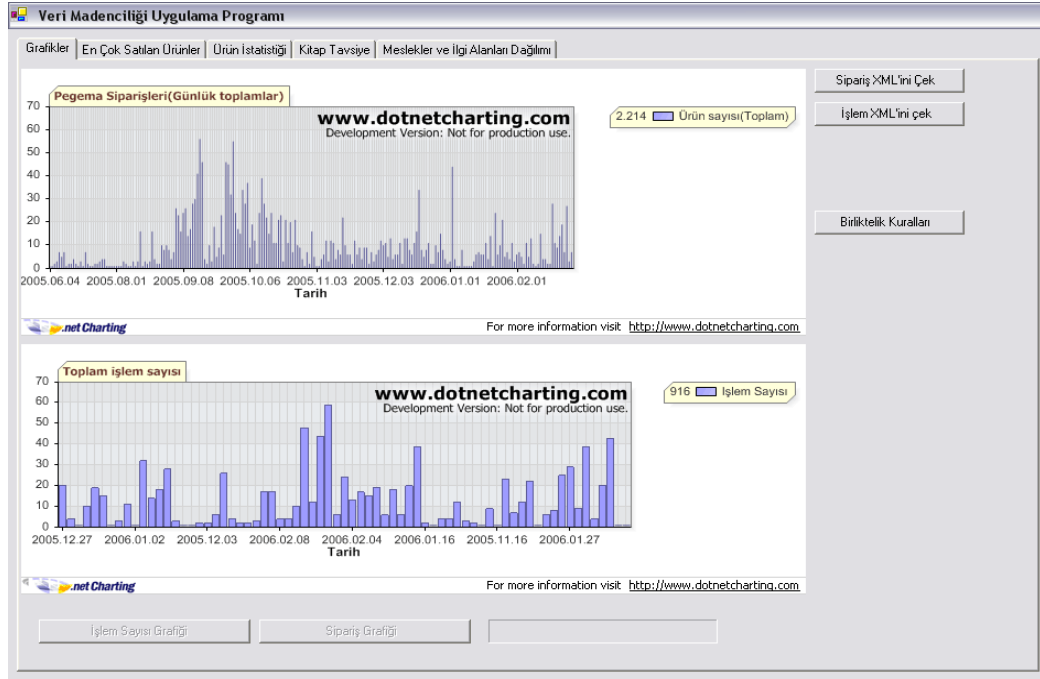
Bu bölümde tez çalışmamızın amacına uygun olarak, geliştirilen yazılım hakkında bilgiler verilmekte ve elde edilen Veri Madenciliği tekniklerinin sonuçlarının analizleri yapılmaktadır.

3.1.1. Veri tanımı

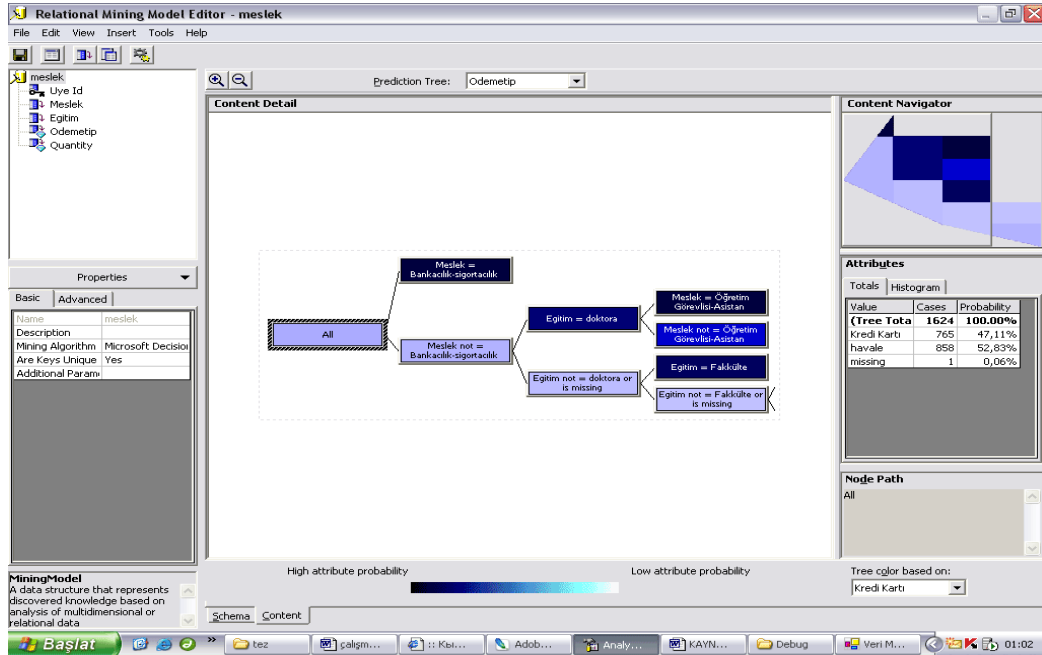
Bu çalışmada kullanılan veriler, Pegem'A yayınevinin bir alışveriş sitesi olan web sitesindeki veri tabanı kayıtlarından elde edilmektedir. Bu kayıtlar toplam 12000 satırdan(işlem) oluşmaktadır. Veriler hangi kullanıcılar hangi tarihte hangi ürünleri sepete attıkları hakkında bilgileri içermektedir. Bu tablolar internette bir SQL Server 2000 veri tabanında tutulmaktadır.

Bu çalışmada VM araçları olarak kendimizin geliştirdiği “Pegem'A VM Uygulaması”(Şekil 3.1) ve Microsoft SQL Server 2000 ve onun “Analysis Services“(Şekil 3.2) bileşeni kullanılmıştır.

“Analysis Services” verideki ilişkisel kuralları çıkarmanın yanı sıra, veriden kümeleme, karar ağaçlarını çıkarmaya yarayan uygulama aracıdır.



Şekil 3.1 “Pegem’A VM Uygulama Programı” uygulamasının genel görünümü.



Şekil 3.2 “Analysis Sevcices” uygulamasının genel görünümü.

“Pegem’A VM Uygulaması” müşteri dağılımı, sistemde yapılan sipariş özelliklerini çıkaran ve VM tekniklerini kullanarak ilgi alanlarına göre kitap tavsiyesinde bulunabilen bir programdır. Ayrıca sistemde yapılan tüm siparişlerin istatistiksel özelliklerini de gösterebilmektedir ve müşteri sepetlerinin analizini yaparak “Birliktelik Kuralları”nı kullanarak ürünleri doğru kategorilendirmeye yardımcı olan programdır.

Programları çalıştırmak için en az donanım gereksinimleri aşağıdaki özellikleri taşımaktadır:

- Pentium işlemci : 1 GHz ve daha yüksek.
- 256 RAM.
- 80-160 MB sabit disk alanı.

Tercihen daha yüksek hızda çalışan işlemciler kullanılmalıdır.

Yazılım gereksinimleri:

- Windows Server 2003 Veya Windows XP Professional Service Pack 2
- Microsoft İnternet Explorer 6.0 veya daha üst versiyonları.

Microsoft SQL Server 2000 Analysis Services gereksinimlerinin hepsini Windows XP Professional Service Pack 2 karşılayabilir.

“Analysis Services” aşağıdaki veri kaynaklarını desteklemektedir.

- Text dosyalar
- OLE DB protokolü aracılığıyla erişilebilen veri tabanları(Microsoft Access,SQL Server 7.0, SQL Server 2000)
- SQL Client protokolü aracılığıyla erişilebilen veri tabanları(SQL Server 7.0, SQL Server 2000)
- Microsoft Excel 97/2000

3.1.2. Veri yapısı

Çalışmada kullanılacak olan Pegem'A web sitesinden alınan 6 aylık alış veriş sepet tablosunun ve ona bağlı ürün tablolarının yapıları Şekil 3.3 ve Şekil 3.9'da verilmiştir.

Bu verilerden sadece Sepet tablosundaki veriler temizleme ve dönüştürme işleminden geçmiştir. Çünkü siteye üyelik kaydını yapmayan ama ürün seçimleri yapılmış olan kayıtların sayısı çok fazla ve programın çalışma süresini etkilemektedir. Programlarda gerçek ve güncel tablolar kullanılmıştır.

shoppingcart	
	recordID
	cartID
	quantity
	productID
	datecreated

Şekil 3.3 “ShoppingCart” tablosunun yapısı

“*ShoppingCart*” tablosu sipariş edilmek üzere kullanıcı tarafından seçilen ürünler hakkında bilgileri içermektedir. Tablonun kayıt yapısı aşağıda verilmektedir:

recordID : Kayıt Numarası.

cartID: Sepet Kodu.

Quantity: Sepete eklenen bir ürün koduna sahip ürün Sayısı.

productID: Ürün Kodu.

Datecreated: Kaydın girilme tarihi.

recordID	cartID	quantity	productID	datecreated
12777	2b604a54-b291-4a25-a65d-2a25991c0ff0	1	157	29.08.2005 17:25:
12778	fdb54763-3d52-4758-976f-1cfc28383edf	1	55	29.08.2005 17:37:
12780	407d301e-ba7f-464e-979b-3a0683482edf	1	194	29.08.2005 17:43:
12783	5455	1	157	29.08.2005 17:50:
12784	1cb563c5-4d48-4d6f-923a-56aca005ae37	1	155	29.08.2005 18:20:
12786	4687	1	106	29.08.2005 19:20:
12787	4687	1	159	29.08.2005 19:21:
12788	9836da4f-153e-4ce3-8297-01e8dcfc3c7a	1	14	29.08.2005 19:21:
12789	4687	1	138	29.08.2005 19:22:
12791	c52854f1-3eec-4202-8f4f-08aa2219dd08	1	9	29.08.2005 20:24:
12794	5433	1	420	29.08.2005 21:15:
12795	3999ca3a-b1eb-43a1-b14f-c249fd9033e9	1	127	29.08.2005 21:37:
12796	b5b63f73-c7f5-4bba-a762-a3eb1e048476	1	180	29.08.2005 21:37:
12797	5459	1	463	29.08.2005 21:43:
12798	5459	1	502	29.08.2005 21:43:
12799	f1282058-f55f-440d-b11b-4b60ad2253e4	1	157	29.08.2005 22:08:
12800	4357966a-721a-4f5c-b442-9c7f0e7b9f7b	1	397	29.08.2005 22:47:
12801	41409369-76be-4c25-8d94-447457c3024a	1	318	29.08.2005 23:13:
12802	beffad01-9d91-4ee9-b3c1-1131c03b8f1d	1	358	29.08.2005 23:18:
12803	6fd274f8-af32-40ca-acb9-f70f69c4794c	1	364	29.08.2005 23:19:
12804	6061f2a1-fe18-41b1-bf67-8c21784c9ab5	1	137	29.08.2005 23:30:
12805	f3ef3612-bd70-4c1c-87f8-60caa12f9cfe	1	270	29.08.2005 23:31:
12807	e4f84d76-259e-4181-9caa-d98a4311820e	1	372	29.08.2005 23:36:
12808	77b9862d-75ce-41a9-bbd6-f784634cc3e6	1	4	29.08.2005 23:40:
12810	77b9862d-75ce-41a9-bbd6-f784634cc3e6	1	520	29.08.2005 23:43:

Şekil 3.4 “ShoppingCart” tablosunun örnek verileri

orders
orderID
customerID
orderDate
shipDate
Onay
kargoref
adsoyad
adres
telefon
sehir
semt
pkod

Şekil 3.5. “Orders” tablosunun yapısı

“Orders” tablosu sipariş yapan kullanıcının adresi, telefonu gibi detaylı bilgilerini içermektedir. Tablodan yazılım için kullanılan kayıtların yapısı aşağıdaki gibidir:

orderID: Sipariş Kodu, kullanıcı siparişlerini bu numarayla takip eder.

customerID: Müşteri Kodu (Üyeler Tablosunda UyeID olarak geçer).

orderDate: Sipariş Tarihi.

shipDate: Teslimat Tarihi.

Onay: Sipariş Onayı.

orderID	customerID	orderDate	shipDate	Onay	kargoref	provkod	odemetip
220	5357	07.08.2005 13:37:	08.08.2005 13:37:	3		<NULL>	havale
221	5357	07.08.2005 14:56:	09.08.2005 14:56:	0		<NULL>	havale
224	4687	07.08.2005 22:23:	09.08.2005 22:23:	0		1123442473425-212.174.160.129-50948	Kredi Kartı
226	5363	08.08.2005 12:38:	09.08.2005 12:38:	3		1123493789777-212.174.160.129-51922	Kredi Kartı
227	2835	09.08.2005 10:47:	10.08.2005 10:47:	3		1123573509526-212.174.160.129-53883	Kredi Kartı
228	5367	09.08.2005 22:13:	10.08.2005 22:13:	3		<NULL>	havale
231	5371	10.08.2005 16:11:	11.08.2005 16:11:	3		1123679386030-212.174.160.129-57219	Kredi Kartı
233	5365	11.08.2005 16:09:	13.08.2005 16:09:	3		1123765645749-212.174.160.129-1205	Kredi Kartı
234	5280	12.08.2005 00:39:	13.08.2005 00:39:	3		1123796249318-212.174.160.129-2053	Kredi Kartı
235	5377	12.08.2005 11:17:	13.08.2005 11:17:	3		1123834544903-212.174.160.129-2753	Kredi Kartı
237	5379	12.08.2005 15:14:	12.08.2005 15:14:	3		1123848769222-212.174.160.129-3356	Kredi Kartı
238	5222	12.08.2005 23:43:	12.08.2005 23:43:	3		1123879288448-212.174.160.129-4358	Kredi Kartı
239	5383	14.08.2005 14:45:	15.08.2005 14:45:	3		1124019789625-212.174.160.129-6713	Kredi Kartı
241	5388	15.08.2005 13:48:	16.08.2005 13:48:	3		1124102793310-212.174.160.129-8592	Kredi Kartı
242	5389	15.08.2005 15:34:	15.08.2005 15:34:	3		<NULL>	havale
244	5392	15.08.2005 18:34:	16.08.2005 18:34:	3		1124119954386-212.174.160.129-9814	Kredi Kartı
245	5395	16.08.2005 11:10:	16.08.2005 11:10:	3		1124179726193-212.174.160.129-11206	Kredi Kartı
249	5401	16.08.2005 17:35:	17.08.2005 17:35:	3		1124202772494-212.174.160.129-12685	Kredi Kartı
251	5408	17.08.2005 14:58:	19.08.2005 14:58:	3		1124279769405-212.174.160.129-14771	Kredi Kartı
252	5410	17.08.2005 17:11:	19.08.2005 17:11:	3		1124287780775-212.174.160.129-15200	Kredi Kartı
253	5411	17.08.2005 18:23:	19.08.2005 18:23:	3		1124292074566-212.174.160.129-15364	Kredi Kartı
254	5377	18.08.2005 09:54:	19.08.2005 09:54:	3		1124347952683-212.174.160.129-16429	Kredi Kartı
255	5383	18.08.2005 10:57:	19.08.2005 10:57:	3		1124351737179-212.174.160.129-16617	Kredi Kartı

Şekil 3.6 “Orders” tablosunun örnek verileri

orderDetails	
🔑	orderID
🔑	productID
	quantity
	unitcost

Şekil 3.7 “Orderdetails” tablosunun yapısı

“Orderdetails” tablosu hangi üründen kaç tane sipariş edildiği hakkında bilgileri içermektedir. Tablonun kayıt yapısı aşağıdaki gibidir:

OrderID: Sipariş Kodu.

productID: Ürün Kodu.

Quantity: Ürün Sayısı.

Unitcost: Ürünün Fiyatı

orderID	productID	quantity	unitcost
57	6	1	4,8
58	8	1	7,2
59	8	1	7,2
60	8	1	7,2
60	19	1	10,8
61	6	1	4,8
61	145	1	15
65	8	1	7,2
66	6	1	4,8
67	8	1	7,2
68	8	1	7,2
70	8	1	7,2
71	137	2	6,4
72	8	1	7,2
72	26	1	6,8
72	27	1	6,8
73	8	1	7,2
73	9	2	6,8
73	13	2	6,8
73	131	1	16,45
73	132	1	7,2
73	137	1	6,4
73	138	1	7,2
75	7	1	7,2
75	8	1	7,2

Şekil 3.8 “Orderdetails” tablosunun örnek verileri

urun
id
baski_yeri
barkod
baski_sayi
ekleme_tarih
isbn
kod
isim
yazar
ceviren
ebat
sayfa_bilgi

Şekil 3.9 “Ürün” tablosunun yapısı.

“Ürün” tablosunda sistemdeki kitapların bilgileri yer almaktadır. Siparişler bu tablodaki “id” sütunundan yararlanarak yapılmaktadır. Herhangi bir siparişteki “ProductID” sütunu direk “Ürün” tablosundaki “id” no’lu kitabı göstermektedir. Bu işlem için ele alınan sütunlar:

Id: Ürün Kodu, ShoppingCart ve OrderOetails tablolarında “productID”.

İsim: Ürün İsmi.

Kategori: Ürünün ait olduğu kategori. Bu sütuna göre kullanıcıların hangi kategorideki kitapları daha çok ilgi gösterdikleri hakkında bilgiler edinilebilir.

id	baski_yeri	barkod	baski_sayı	ekleme_tarih	isbn	kod	isim	yazar	cevirer	ebat
224	ANKARA		4	03.01.2000	975-8322-06-0	0	Çağdaş Eğitim Deni	Mustafa Aydın		16X24
225	ANKARA		3	04.01.2001	975-7527-55-6	0	Dış Ticaret ve İhrac	Mehmet Tomanbay		16X24
226	ANKARA		10	05.01.2004	975-7527-73-4	0	Beslenme	Prof.Dr.Ayşe Baysı		16X24
227	İSTANBUL		2	02.01.2005	9752105610	0	4' ün Kuralı	Dustin Thomason, J Pinar Ç		14x20
228	ANKARA		12	03.08.2005	975-8640-06-2	0	Gelişim Öğrenme ve	Nuray Senemoğlu		16X24
229	ANKARA		1	02.01.2005	975-8895-58-3	0	Ticari İşletme Hukuk	Ahmet Battal		14x20
230	ANKARA		1	02.01.2005	975-8895-57-5	0	Kıymetli Evrak Huku	Ahmet Battal		14x20
231	ANKARA		1	02.01.2005	975-8895-42-7	0	Anayasa Hukuku	Hasan Tunç, Faruk		14x20
232	ANKARA		1	02.01.2005	975-8895-41-9	0	İcra İflas Hukuku	Erdal Tercan, Sühe		14x20
233	ANKARA		1	02.01.2005	975-8895-59-1	0	Şirketler Hukuku	Ahmet Battal		14x20
234	Ankara		3	05.01.2005	975-6859-54-7	0	Türkiye Coğrafyası	Cemalettin Şahin,H		16X24
235	Ankara		2	02.01.2003	975-6859-06-7	0	Ülkeler Coğrafyası	Cemalettin Şahin		16X24
237	Ankara		2	02.01.2003	975-6859-12-1	0	Coğrafyaya Giriş	Cemalettin Şahin		16X24
238	Ankara		1	02.01.2004	975-6858-52-4	0	Komşu Ülkeler Coğr	Prof.Dr.Enrullah Gi		16X24
239	istanbul		5	02.01.2005	975-6448-43-1	0	Sivil Örmceğin Ağı	Atilla İlhan		14X20
241	İSTANBUL		2	09.01.2004	975-8621-82-3	0	Vatikan sırları	Paul L.Williams	Handa	14x20
242	Ankara		2	04.01.2005	9789752201255	0	Emperyalizmin Uşak	Vural Savaş		14x20
243	İSTANBUL		4	02.01.2004	975-8509-86-1	0	BEBEKLERDE VE ÇÇ	DR.RÜÇHAN ALTUĞ		14x20
244	Ankara		3	04.01.2005	975-6114-04-5	0	BİZİM HÜCRE"BEN	ETEM KARAÜZÜM		14x20
245	İSTANBUL		25	12.01.2004	975-362-332-1	0	Armagedon- Türkiy	Aydoğan Vatandaş		14x20
246	İSTANBUL		1	05.01.2005	975-8658-93-X	0	Hız.Muhammed Giz	Tolstoy	Prof.Dr	14x20
247	İSTANBUL		1	08.01.2004	975-263-057-X	0	Büyük Ortadoğu Ku	Atilla Akar		14x20
248	Ankara		1	03.01.2005	975-8163-57-4	0	Ortadoğu Siyasetin	Türel Yılmaz-Mehme		14x20
249	Ankara		1	10.01.2004	975-8163-48-5	0	Ortadoğu Siyasetin	Türel Yılmaz-Mehme		14x20
250	Ankara		1	01.01.2005	975-8163-58-2	0	KARADENİZ-Uluslar	Prof.Dr.Osman Met		14x20
251	Ankara		2	01.01.2005	975-8163-53-1	0	Türk ve Dünya Ede	Cevdet Yalçın		14x20
252	Ankara		10	05.01.2005	975-8163-50-7	0	Yeni Kamu Yönetimi	Dr.m.Akif Özer		14x20
253	Ankara		1	01.01.2005	975-8163-54-X	0	Bağımsız Devletler	Dr.Fırat Purtaş		14x20
254	Ankara		8	03.01.2005	975-8163-61-2	0	Türk ve Dünya Ede	Cevdet Yalçın		14x20
255	Ankara		3	08.01.2004	975-8163-49-3	0	İş Yerinde Psikolojik	Dr.Hasan Tutar		14x20

Şekil 3.10 “Ürün” tablosunun örnek verileri

Birliktelik Kuralları Tekniği kullanılan bu çalışmadaki amaç, yukarıda açıklanan verileri kullanarak hangi kitapların daha çok beraber tercih edildiklerini ve bu tercihlerin destek sayılarına göre tablolarını oluşturmaktır (destek sayısı 2den 9a kadar). Yapılan alışverişlere dayanarak siteyi ziyaret eden müşteriye onun ilgi alanına göre kitap tavsiyesinde bulunacak bir sql sorgulaması geliştirmek ve bunu sitede verimli olacak şekilde kullanmaktır.

Çalışmanın sonucunda kullanıcıların kitap tercih etme desenleri belirlenecektir. Bu desenler “Birliktelik Kuralları” tekniği aracılığıyla belirlenmektedir.

Birçok kişi siteyi ziyaret eder ama arama zamanı olmadığı için onun ilgisini çekebilecek ürünleri(kitapları) gözden kaçırabilir. Bunun için kitap tavsiyesinde

bulunacak bir program geliştirilmiştir. Böylece siteyi ziyaret eden kullanıcılar kendi ilgi alanlarına göre kitap tavsiyeleri alabileceklerdir.

Buna ek olarak istatistiksel metotları kullanarak 6 ayda yapılan alışveriş bilgilerine göre kitapların satış dağılımını, tıklanma istatistiklerini, sitede yapılan işlemlerin satışı etkileyip etkilemeyeceği, belirlenmektedir. Sitede kayıtlı üyelerin mesleği ve aldığı eğitim arasındaki ilişki belirlenmektedir. Site ziyaretçi bilgileri de varolan bilgiler üzerinde incelemeye alınmaktadır.

Karar Ağacı tekniği aracılığıyla sitenin kullanıcılarının ödeme eğilimleri belirlenecektir. Bu inceleme sonucunda kredi kartı kullanmaya yatkın kullanıcılar tespit edilerek onlara reklam kampanyalar düzenlenerek satış arttırılması amaçlanmaktadır.

3.2. Veriye Uygulanan Ön İşlemler

“Birliktelik Kuralları”nda kullanılacak olan “orderdetails” ve “shoppingcart” tablolarının dışındaki veriler bir ön işlemten geçmeden kullanılacaktır. “OrderDetails” ve “ShoppingCart” tablolarındaki veriler sıralı olmadığı için önce onları Sipariş takip(OrderDetails) ve Sepet (“ShoppingCart”) numaralarına göre sıralandırılması gerekmektedir. “Birliktelik kuralları”nı çıkartmak için veri tabanından alınan veriler Çizelge 3.1’de gösterilen formata dönüştürülmelidir. Bunu yapmak için “Orders” ve “OrderDetails” tabloları kullanılarak SQL ile aşağıdaki veri yapısı elde edilebilir. Sepet tablosunda ise sisteme üye olmadan sepete eklenen ürünler hakkındaki bilgiler silinmiş ve böylece “Sepet” tablosundaki kayıt sayısı 3600’e kadar indirilmiştir. Bu sayı şu an kullanıcıların sepetlerinde olan ama siparişi yapılmayan ürün sayısını göstermektedir. Bu bilgilerin incelenmesi de sipariş kayıtlarının incelenmesi kadar önemlidir.

Çizelge 3.1. “Birliktelik Kuralları” için giriş tablosu

TransactionID	Items
100	1,3,4
200	2,3,5

“OrderDetails”ve “ShoppingCart” tablolarındaki verileri alınarak bu formata dönüştürülmesi gerekmektedir. Bunu yapmak için bir sql sorgusu kullanıldı ve aşağıdaki formatta tablo elde edildi, tabloya “Transactions” ismi verildi. “Transactions” tablosu sistemdeki tüm müşteri hareketlerinin tutulduğu bir veri tabanı olarak algılanabilir.

“OrderDetails” tablosundan “Transactions” tablosunu elde etmek için kullanılan SQL sorgusu:

Sql Sorgusu (Orderdetails): “SELECT orderDetails.orderID, orderDetails.productID, orderDetails.quantity, (SELECT isim FROM urun WHERE urun.id = orderdetails.productID) AS isim FROM orderDetails INNER JOIN orders ON orderDetails.orderID = orders.orderID ORDER BY orders.orderID”

orderID	productID	quantity	isim
198	110	1	Deneysel Desenler
198	442	1	Örneklerle Çocuk Edebiyatı
199	85	1	Okulda Çoklu Zeka Kuramı
199	130	1	Eğitimde Yeni Değerler
200	85	1	Okulda Çoklu Zeka Kuramı
200	130	1	Eğitimde Yeni Değerler
200	218	1	Kendini Keşfet
201	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
203	106	1	Örnek Grup Rehberliği Etkinlikleri
206	326	1	Bilimsel Araştırmaya Giriş
207	49	1	Eğitimde Değişim Yönetimi
208	109	1	Sosyal Bilimler için Veri Analizi El Kitabı (CD’li)
210	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
211	404	1	Bağımsız Denetim
212	206	1	İleri Düzey Spk Lisanslama Sınavlarına Hazırlık
214	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
215	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
216	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
217	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
218	155	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık
219	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
220	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
221	156	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Tamamı

Şekil 3.11 Oluşan tablo yapısı(OrderDetails Transactions)

Burada:

orderID: Sipariş Kodu.

productID: Ürün Kodu.

Quantity: Ürün Sayısı.

İsim: Ürün ismi.

Yukarıdaki oluşan veri yapısı “Birliktelik Kuralları için Hızlı Algoritmalar” metodu için geçerli yapı oluşturmaktadır.

“ShoppingCart” tablosundan “Transactions” tablosunu elde etmek için kullanılan SQL sorgusu:

Sql Sorgusu(ShoppingCart): “SELECT cartID, productID, quantity, (SELECT isim FROM urun WHERE urun.id = shoppingcart.productID) AS isim FROM shoppingcart ORDER BY cartID”

cartID	productID	quantity	isim
5941	155	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık
5942	597	1	Kuramdan Uygulamaya Fen ve Teknoloji Öğretimi
5946	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
5947	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
5950	601	1	Uzman Öğretmenlik Başöğretmenlik Sınavlarına Hazırlık (KBY
5952	591	1	Öğretim Teknolojileri ve Materyal Geliştirme
5952	591	1	Öğretim Teknolojileri ve Materyal Geliştirme
5952	211	1	Taocu Sevişme ve Seks / Eski Çinlilerin Cinsel Esrime Yöntem
59567a7a-cc43-42f8-94c7-478a0043b3d4	227	1	4' ün Kuralı
5956ff78-fdf6-44bb-8fa2-d2b4b500bfd6	601	1	Uzman Öğretmenlik Başöğretmenlik Sınavlarına Hazırlık (KBY
59589ffc-458a-4c27-a650-98c5ea3cba4e	237	1	Coğrafyaya Giriş
5959	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
595a8575-9bf8-4fb3-a808-1eb8408c8ad7	592	1	İmparatorluktan Ulus Devlete Türk İnkılap Tarihi
595e3f82-e6e9-487c-81ed-3d86231ea4c1	164	1	HUKUK SÖZLÜĞÜ (İngilizce-Türkçe)
5960bf29-a899-43b0-b8ae-1773493de076	908	2	Building Skills for Proficiency
59624afc-b167-4f79-948a-49913c00d071	346	1	Vergilemede Kanunilik ve Türk Vergi Sistemi
5966	661	1	Kuram ve Uygulamada Eğitim Yönetimi
5967	360	3	Mühendislik Kaya Mekaniği ilkelere giriş
59674e84-e205-4fa4-b56d-d8881045777c	394	1	Kaya Mekaniğine Giriş
5971e205-d3cc-4cfe-b6e6-0fa546cd1cb1	221	1	Sosyal Psikoloji
5974f6f9-eb47-4ab3-972c-85b13e5ffb8e	416	1	İşletmelerde Finansal Yönetim
597564e9-b142-4098-9e95-5643acc0ebb8	446	1	Örgütsel Psikoloji
5979	591	1	Öğretim Teknolojileri ve Materyal Geliştirme
5979	156	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Tamamı
597a89f2-ae41-41a9-bc94-059908f2d964	264	4	Ergenler İçin Grupla Psikolojik Danışma Uygulamaları
597a89f2-ae41-41a9-bc94-059908f2d964	106	1	Örnek Grup Rehberliği Etkinlikleri
5980	601	1	Uzman Öğretmenlik Başöğretmenlik Sınavlarına Hazırlık (KBY
5980bbcd-545e-40e3-b395-633706d1cdf5	214	1	Hititler
5983	157	2	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
5989fbed-68d3-46d9-8065-50c20ac73b34	360	1	Mühendislik Kaya Mekaniği ilkelere giriş
598afd3c-60c4-450b-8000-62748cce0bde	125	1	Ev ve Ailede Yaşam Yönetimi
5995	157	1	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti
5995a973-3ac4-4ba8-8783-074bc8d44e21	177	1	Redhouse Sözlüğü (İngilizce - Türkçe)

Şekil 3.12 Oluşan tablo yapısı(ShoppincCart Transactions)

Oluşan tablo da “Birliktelik Kuralları için Hızlı Algoritmalar” metodu için geçerli veri yapısını oluşturmaktadır.

Burada:

CartID: Sepet Numarası:

productID: Ürün kodu.

Quantity: Ürün Sayısı.

İsim: Ürün İsmi.

Bu tablolar bir DataSet(Veri Kümesi) şeklinde işlemlerde kullanılmaktadır. Bu DataSet “Fast Algorithms for Discovering Association Rules”(Agrawal) Algoritmasını kullanarak programda kullanılmaktadır.

Birliktelik Kurallarını çıkarma

Programda kullanılacak değişkenler:

k-ItemSet : k üyesi olan sayı kümesi.

Lk: Büyük k-Itemset kümesi.(En az destek sayısı olanlar). k-ItemSetin her bir üyesi iki değişkenden oluşur

- *Lk*'ye ait sayı kümesi (Bizim uygulamamızda ürün kod numaraları).
- *Lk*'ye ait destek sayısı (desen veri tabanında kaç kez ortaya çıkmaktadır).

Ck: k-ItemSet aday kümesi.(Potansiyel olarak büyük kümeler). *Ck*-ItemSetin her bir üyesi iki değişkenden oluşur

- Sayı kümesi(*Ck* için).
- Destek sayısı(*Ck* için).

$\overline{Ck}$: Oluşturulan işlemlerin(Transactions) TIDleri(TransactionID's) adaylarla ilişkilendirildiğinde oluşan k-ItemSets kümesi.

Programın genel algoritmasını şu şekildedir:

$L1 = \{\text{tek üyeli büyük kümeler}\};$

For ($k=2; Lk-1 < > 0; k++$) do begin

$Ck = \text{apriori-gen}(Lk-1);$ // Yeni adaylar

Forall transactions $t \in D$ do begin

$Ct = \text{Subset}(Ck, t)$ // t 'nin içindeki tüm adaylar

Forall candidates $c \in Ct$ do

$C.\text{count}++;$

End

Cevap = $UkLk$;

Yazdığımız programda:

- L1'e Transaction
- Ck'ya Transactions
- $\overline{Ck}$ 'ya ise SItemSets isimleri verilmiştir.

Program 3 parçadan oluşmaktadır. Birincisi L ile ilgili hesaplamaları yapmaktadır. İkincisi alınan L'lerden Ck oluşturmakta ve Ck ile ilgili işlemleri gerçekleştirmek için kullanılır ve üçüncüsü ise Ck'lardan $\overline{Ck}$ oluşturarak $\overline{Ck}$ onunla ilgili hesaplamaları yapar.

Programın çalışmasını bir örnekle açıklamak gerekirse;

Program önce veritabanından oluşturulan *sayı kümesini* okuyarak ve hafızada bir veritabanı örneği oluşturmaktadır.

DataSet ds = a1.CreateDataSetSql(Sql);

Burada Sql: Yukarıda belirttiğimiz sqllerden biri olabilir.

Sayı Kümesi içindeki veri herhangi bir yerden çekilebilir(Sql Server 2000,XML dosyası) veya elle de oluşturulabilir.

Sonraki adımlarda programın kullanabilmesi için oluşturulan *sayı kümesinden* bir SItemSets oluşturulup ona C ismi verilmekte. Sonuç olarak bize birimi L olan bir veri gelmesi gerekiyor. Onun için Ck'yı giriş değişkeni olarak alan ve sonuç olarak tekrar L gönderen bir fonksiyon yazılmıştır. Onun ismi de Calculate. "Calculate" fonksiyonu "Birliktelik kuralları"nı çıkartan ana fonksiyondur.

3.3. Birliktelik Kurallarını Tekniğiyle Alınan Bilgileri Değerlendirme

Yukarıdaki anlatılan programı Pegem’A web sitesinin verilerine uyguladığımızda sepet ve siparişleri için farklı kurallar elde edilmiştir. Toplam Sepet için 3 ve Siparişler için 8 olmak üzere 11 kural çıkartılmıştır. Bu kurallar Çizelge 3.2’de gösterilmektedir. Kural sayısının az olmasının sebebi kayıt sayısının az olmasına ve güçlü Birliktelik Kurallarının olmamasına bağlıdır. Yine de elde edilen sonuçlardan faydalanılarak satışları etkileyebilecek stratejiler üretilebilir. Çünkü kullanılan algoritma kesin ve net ürün kümelerini vermektedir. Destek sayısı istenilenden az olanları da hesaplama dışında bırakmaktadır.

Çizelge 3.2 “Pegem’A”ya uygulanan “Birliktelik Kuralları” yönteminin sonucu

Destek Sayısı	TransactionID	Ürün No:	Ait olduğu kategori.
2	14	85	Eğitim Bilimleri
		106	Eğitim Bilimleri
		114	Eğitim Bilimleri
		138	Eğitim Bilimleri
		159	Eğitim Bilimleri
	15	85	Eğitim Bilimleri
		106	Eğitim Bilimleri
		114	Eğitim Bilimleri
		138	Eğitim Bilimleri
		159	Eğitim Bilimleri
3	18	23 “Öğretim Teknolojileri ve Materyal Geliştirme”	Eğitim Bilimleri

Çizelge 3.2 (Devam) “Pegem’A”ya uygulanan “Birliktelik Kuralları” yönteminin sonucu

3	18	26 “Çatışma Yönetimi”	Yönetim
		30 “Nümerik Analiz ”	Matematik
	19	138	Eğitim Bilimleri
		159	Eğitim Bilimleri
		264	Eğitim Bilimleri
	20	155	Sınavlara Hazırlık
		156	Sınavlara Hazırlık
		157	Sınavlara Hazırlık
	21	106	Eğitim Bilimleri
		138	Eğitim Bilimleri
		159	Eğitim Bilimleri
4	23	23 “Öğretim Teknolojileri ve Materyal Geliştirme”	Eğitim Bilimleri
		26 “Çatışma Yönetimi”	Yönetim
	24	8 “Etkili Okul ve Okul Geliştirme”	Eğitim Bilimleri
		9 “Temel Bilgi Teknolojisi Kullanımı”	Bilgisayar - İnternet

Çizelge 3.2 (Devam) “Pegem’A”ya uygulanan “Birliktelik Kuralları” yönteminin sonucu

4	25	7 “Örgütsel Gelişme”	Yönetim
		8 “Etkili Okul ve Okul Geliştirme”	Eğitim Bilimleri
5	8	157 “Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti”	Sınavlara Hazırlık
		500 “Şu Çılgın Türkler”	Tarih-İnkılap Tarihi
	10	157	Sınavlara Hazırlık
		907	Sınavlara Hazırlık
	11	156	Sınavlara Hazırlık
		907	Sınavlara Hazırlık
	12	42 “Okul Yönetimi”	Yönetim
		134 “Eğitim Yönetiminde Teori ve Uygulama”	Eğitim Bilimleri
	13	58	Sınavlara Hazırlık
		113	Sınavlara Hazırlık

Çizelge 3.2 (Devam) “Pegem’A”ya uygulanan “Birliktelik Kuralları” yönteminin sonucu

6	3	138	Eğitim Bilimleri
		159	Eğitim Bilimleri
	9	106	Eğitim Bilimleri
		138	Eğitim Bilimleri
7	2	138	Eğitim Bilimleri
		264	Eğitim Bilimleri
	4	155	Sınavlara Hazırlık
		157	Sınavlara Hazırlık
	6	156	Sınavlara Hazırlık
		157	Sınavlara Hazırlık
	7	159	Eğitim Bilimleri
		264	Eğitim Bilimleri
8	16	26 “Çatışma Yönetimi”	Yönetim
		30 “Nümerik Analiz ”	Matematik
	17	155	Sınavlara Hazırlık
		156	Sınavlara Hazırlık

Çizelge 3.3 Sepet için oluşturulan birliktelik kuralları

Destek Sayısı	TransactionID	Ürün No:	Ait olduğu kategori.
2	1	156	Sınavlara Hazırlık
		157	Sınavlara Hazırlık
		601	Sınavlara Hazırlık
3	2	138	Eğitim Bilimleri
		159	Eğitim Bilimleri
	3	138 ” Sınıf İçi Rehberlik Uygulamaları”	Eğitim Bilimleri
		562 ” Psikolojik Sağlığı Koruyucu Rehberlik ve Psikolojik Danışma”	İnsan Psikolojisi
4	4	157	Sınavlara Hazırlık
		601	Sınavlara Hazırlık
	5	156	Sınavlara Hazırlık
		157	Sınavlara Hazırlık
	6	157	Sınavlara Hazırlık
		946	Sınavlara Hazırlık

3.3.1. Sipariş tablosu için sıra dışı olan ürünler

Elde edilen tablolar incelenerek satış sayısı yüksek olup ama Birliktelik Kuralları bağımlılıkları gösteren ürünlerin listesi elde edilmiştir. Bu ürünlerin kod ve isimlerinden oluşan liste aşağıda listelenmiştir:

- 7 – “Örgütsel Gelişme” (Yönetim)
- 8 – “Etkili Okul ve Okul Geliştirme” (Eğitim Bilimleri)
- 9 – “Temel Bilgi Teknolojisi Kullanımı” (Bilgisayar - İnternet)
- 23- “Öğretim Teknolojileri ve Materyal Geliştirme”(Eğitim Bilimleri)
- 26 – “Çatışma Yönetimi” (Yönetim)
- 30 – “Nümerik Analiz ” (Matematik)
- 42 - “Okul Yönetimi” (Yönetim)
- 134 – “Eğitim Yönetiminde Teori ve Uygulama” (Eğitim Bilimleri)
- 157 – “Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık Seti” (Sınavlara Hazırlık)
- 500 – “Şu Çılgın Türkler” (Tarih-İnkılap Tarihi)

3.3.2. Sepet tablosu için sıra dışı olan ürünler

- 138 – “Sınıf İçi Rehberlik Uygulamaları”
- 562 – “Psikolojik Sağlığı Koruyucu Rehberlik ve Psikolojik Danışma”

Burada aynı işleme ait olup ama farklı kategorilerden gelen ürünler dikkate alınması gerekmektedir. Asıl aranan desen bunlardır. Eğer bu ürünler bir kategoriye, yani bir rafa konulursa onların satışlarında artış beklenecektir. Bu da satışları etkileyecektir. Çünkü bu ürünler bir arada satın alınma eğilimine sahip oldukları için onlar yan yana konulursa satılma şansları daha da artacaktır. Eğer bir kategoriye konulmayacaksa o zaman bu ürünlerin satışları farklı yaklaşımla arttırılabilir. Mesela ürünler sayfada gösterilirken o ürünle aynı Transaction IDsine sahip ürünleri reklam aracılığıyla o ürünün yanına yada tavsiye olarak o sayfaya konulabilir.

Yukarıda alınan bilgilerden yararlanarak aşağıdaki sonuçlara varılmıştır:

- 9 numaralı ürünün (“Temel Bilgi Teknolojisi Kullanımı” (Bilgisayar - İnternet)) “Eğitim Bilimleri” kategorisi altına taşınması daha uygun, çünkü sitede daha çok

tercih edilen kategori “Eğitim Bilimleri”dir, bu bilgiyi nasıl elde ettiğimiz hakkında da bir sonraki bölümde detaylı şekilde açıklanmaktadır. Bununla birlikte 9 No’lu ürün kendisi “bilgisayar İnternet” kategorisine ait olduğu halde daha çok (destek sayısı kadar) “Eğitim Bilimleri” kategorisine ait olan 8 No’lu ürün ile (“Etkili Okul ve Okul Geliştirme” (Eğitim Bilimleri)) birlikte alınmıştır. Ama aynı anda 8 No’lu ürünün kategorisini “Yönetim” olarak uygun gördüğümüze göre niye bu iki ürün genelde satın alındığını siteyi ziyaret eden kişilerin kim olduğundan yola çıkarak şu şekilde izah edilebilir. İstatistiksel veriler incelendiğinde siteyi daha çok eğitimci, öğretmen ve yöneticilerin ziyaret ettiği görülmüştür. Bu kişilerin bir eğitimci ve aynı anda da yönetici olduğunu varsayarsak (mesela okul müdürü, öğretmen) niye bu iki ürünü bir arada aldığını anlaşılabılır.

23, 8, 134 No’lu ürünlerin ise isimlerine ve yukarıdaki “Birliktelik Kuralları” uygulamasının sonucuna bakılırsa “Yönetim” kategorisi altında yer alması daha uygundur.

Sepet için ise ürün No’su 138 olan ” Sınıf İçi Rehberlik Uygulamaları” adlı kitap “Eğitim Bilimleri” kategorisine ait olduğu halde kategorisi “İnsan Psikolojisi” olarak değiştirilmesi önerilmektedir.

Aynı anda Siparişler için destek sayısının en fazla 8, Sepet için ise destek sayısının en fazla 4 olabileceği de ortaya çıkmıştır. Bu durum yapılan siparişlerden alınan sonuçların(Birliktelik Kuralların) daha da güvenilirli olacağını gösteriyor.

3.4. Karar Ağacı Tekniği ve Bu Teknikle Hazırlanan Model

“Microsoft Analysis Services “ modülünün “Karar Ağaçları” yöntemi sitedeki kullanıcıların eğitim ve meslek bilgilerine dayanarak ödeme yaparken hangi ödeme şeklini(kredi kartı yada havale) kullanacaklarını tahmin edecek modelin elde edilmesi için kullanılmıştır.

Karar Ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde, tahmin edici bir tekniktir [4]. Karar Ağacı, kolay anlaşabilecek kurallar yaratabilen, bilgi teknoloji işlemleri ile kolay entegre olabilen en popüler sınıflama tekniğidir [4].

Karar Ağacı karar düğümleri, dallar ve yapraklardan oluşur. Karar düğümü, gerçekleştirilecek düğümü belirtir. Her düğümde test ve dallar ayırma işlemleri ardışık olarak gerçekleşir ve bu ayrılma üst seviyedeki ayrılmalara bağlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşmiyorsa, o dalın sonucunda bir karar düğümü oluşur. Ancak dalın sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır [4]. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir.

Karar ağacı tekniğini kullanarak verinin sınıflanması iki basamaklı bir işlemdir. İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından analiz edilir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır [4].

Test verisinde uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karşılaştırılır. Eğer modelin doğruluğu kabul edilebilir bir değer ise model, sınıfı bilinmeyen yeni verileri sınıflama amacıyla kullanılabilir [4].

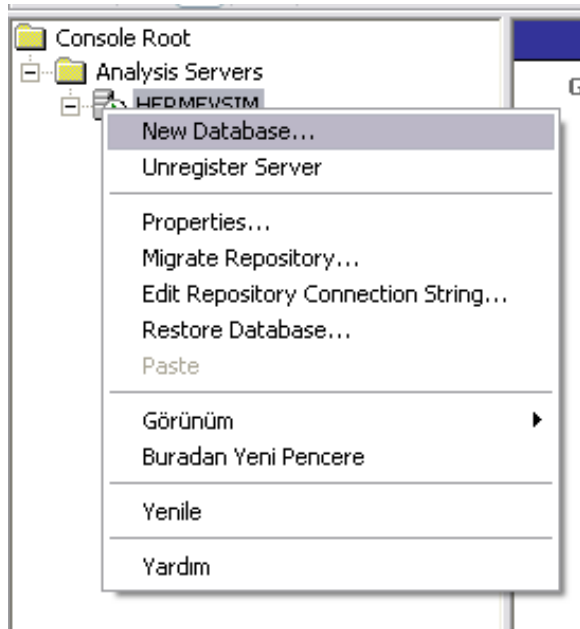
Şekil 3.23’de (bkz Sayfa 71) verilen modelde eğitim verisi incelenerek, müşteri hangi ödeme şeklini kullanacağı tahmin edilmiştir.

“IF Meslek not = Bankacılık-sigortacılık and Egitim not = doktora and not = Fakkülte or is missing and Quantity = 1 THEN odemetip = havale” şeklindedir.

Bu kurala göre mesleği bankacılık-sigortacılık olan bir kişi genelde “kredi kartı ” ödeme şeklini seçeceği görülmektedir.

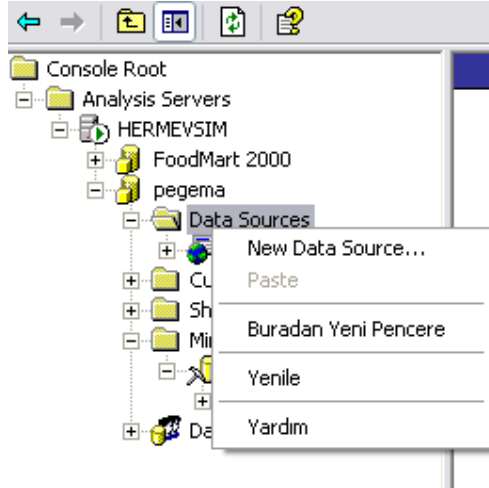
“Microsoft Analysis Services” programında karar ağacını oluşturmak için aşağıdaki adımlar izlenir:

1) Önce yeni veritabanı bağlantısı oluşturulur:



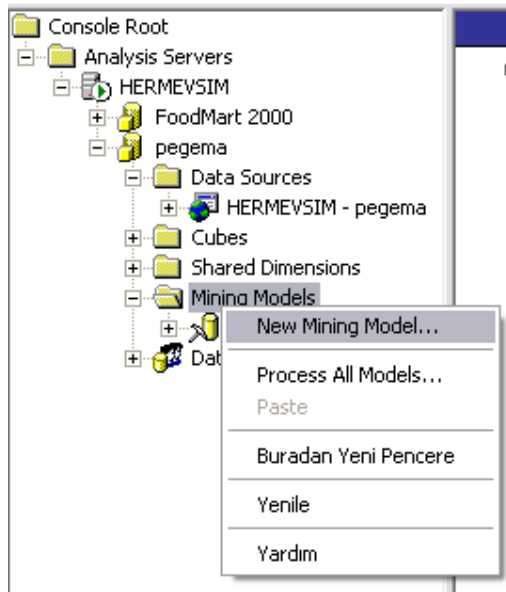
Şekil 3.13 Yeni veritabanı bağlantısını oluşturma

2) Oluşturulan bağlantının veri kaynağı belirlenir



Şekil 3.14 Veri kaynağı belirtme

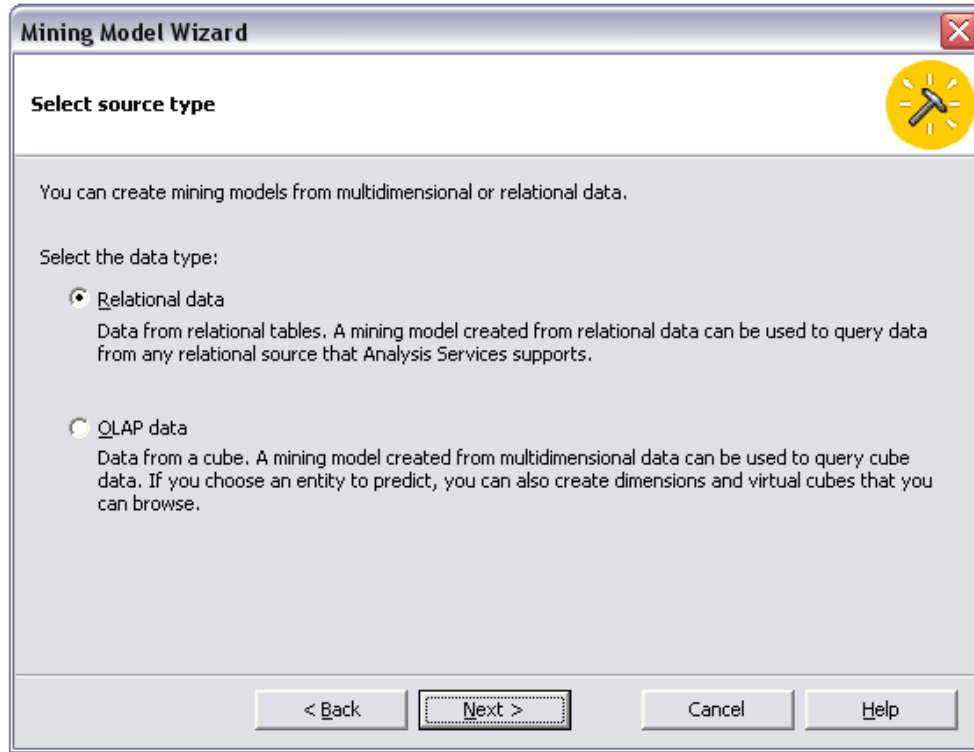
3) Yeni VM modeli oluşturulur.



Şekil 3.15 Yeni VM modelini oluşturma

4) Ekrana çıkan menüde “Next” butonu basılır.

5) Karşımıza aşağıdaki menü çıkacaktır.



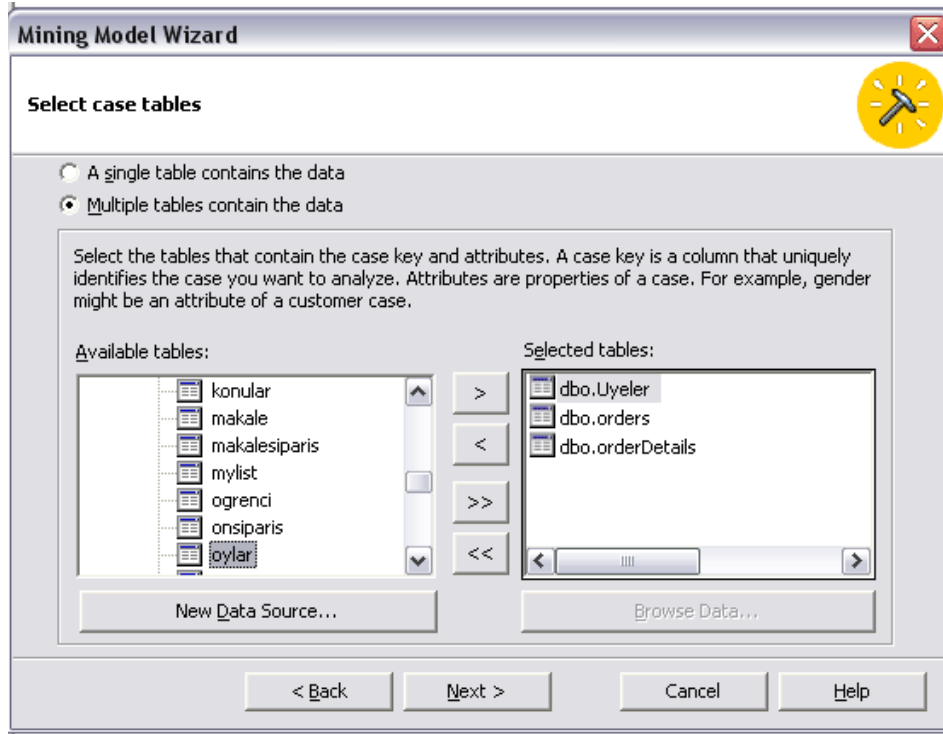
Şekil 3.16 Veri tabanı yapısını belirtme

Burada “Relational Data” seçilmelidir.

6) Bir sonraki menüde veriler kaç tablodan seçileceği belirtilmesi gerekir. Bu durumda:

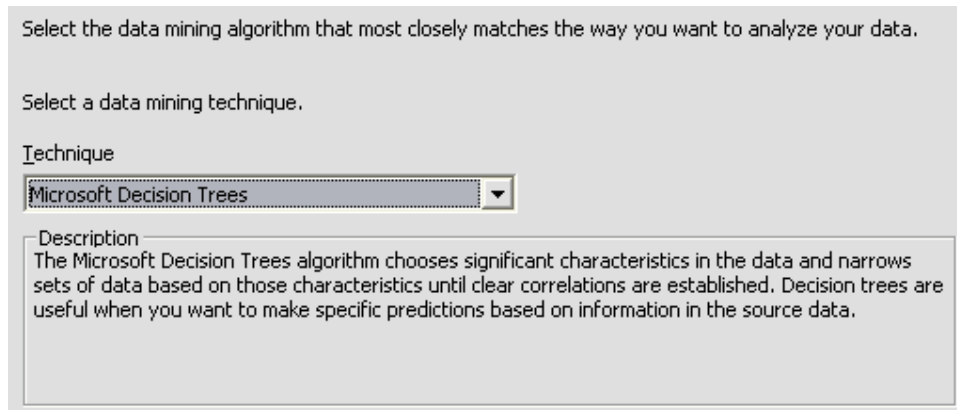
Kullanılacak tablolar :

- “Üyeler”
- “Orders”
- “OrderDetails”



Şekil 3.17 Tabloları belirtme

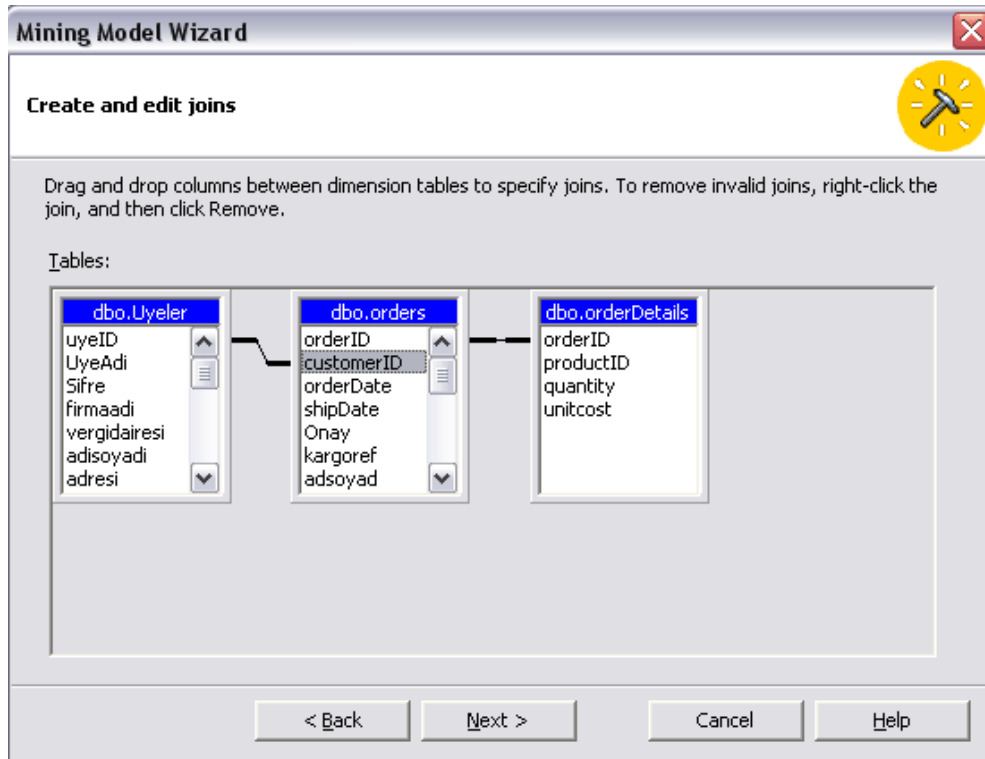
- 7) Sonraki adıma geçmek için “Next” butonu basılacak ve karşımıza çıkan panelde hangi VM Yöntemi kullanılacağına karar verilmesi gerekir. Bu durumda “Microsoft Decision Trees” olacak:



Şekil 3.18 VM yöntemi belirtme

8) “Next” butonuna tıkladığımızda karşımıza çıkan panelde tabloların arasındaki ilişkiyi göstermemiz istenecektir. Burada “Üyeler” tablosundaki “üye ID”

sütununu “Orders” tablosunun “customer ID” sütununun üzerine sürükleyerek iki tablo arasındaki ilişki elle bir daha gösterilmeli. Bunu yapmak için anahtar sütunların fareyle birbirinin üzerine taşınması yeterli olmaktadır.



Şekil 3.19 Tablolar arasındaki ilişkileri gösterme

9) Sonraki adımda anahtar sütunu göstermemiz istenmektedir. Burada “üye ID” sütununu göstererek “Next” butonuna tıklanmalıdır.

Select the key column

Select the table that contains the case key column, and then select the case key column.
a column that uniquely identifies the case you want to analyze.

Case key table:

Case key column:

< Back Next > Cancel

Şekil 3.20 Anahtar sütununu belirtme

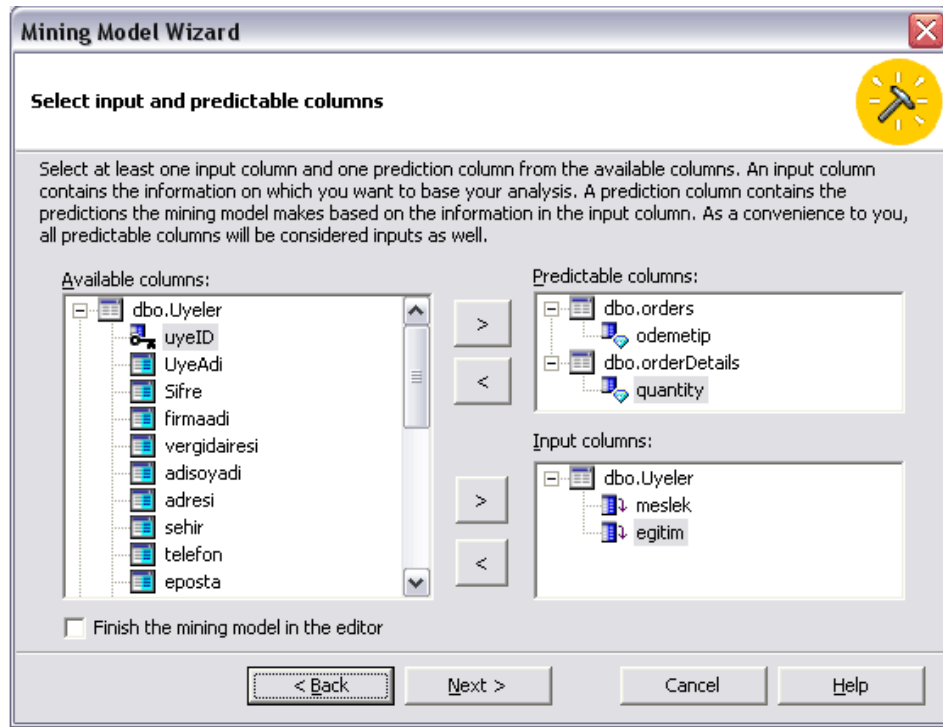
10) Bu adımda modelin giriş ve çıkış(sonuç) sütunlarını belirtir.

Giriş sütunları:

- Meslek
- Eğitim

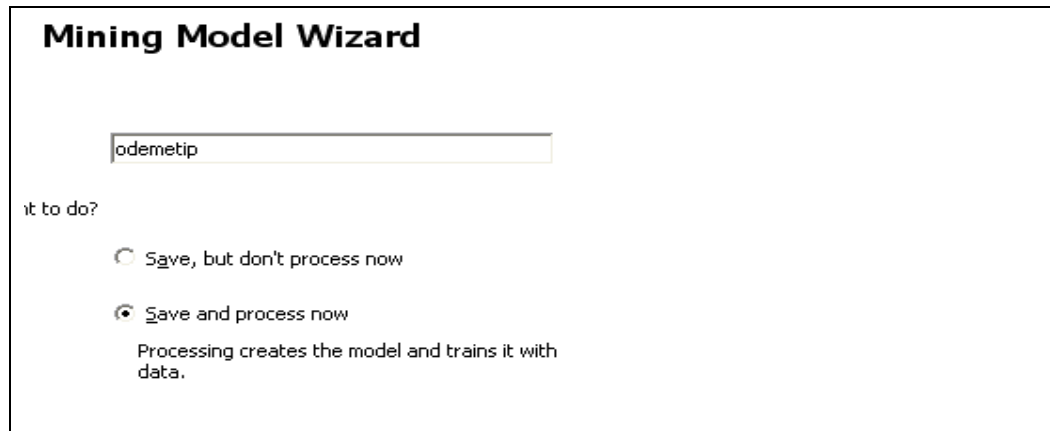
Çıkış sütunları :

- Ödeme şekli
- Ürün sayısı



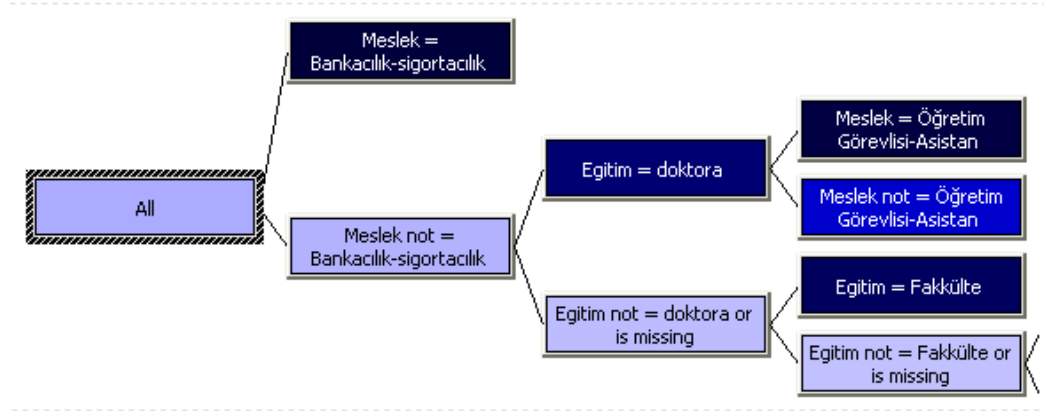
Şekil 3.21 Giriş ve tahmin edilecek sütunları belirtme

11) Son adım olarak da modelin ismi verilir ve modeli oluşturmak için “Save and Process Now” denir (Şekil 3.22):



Şekil 3.22 Modele isim verme

12) Sonuç olarak bir “Karar ağacı” çıkıyor:



Şekil 3.23 Karar ağacı

Bu karar ağacına bakarak sisteme giriş yapan müşterinin eğitim ve meslek bilgilerine göre alışveriş yaparken hangi ödeme şeklini kullanacağı hakkındaki bilgiler edinilebilir.

Bu karar ağacını tablo şeklinde gösterdiğimizde aşağıdaki kurallar ve kredi kartı veya havale seçeneklerinin yüzdeleri çıkmaktadır.

Çizelge 3.4 Karar ağacının tablo şeklinde gösterimi

Kural	Güvenirlilik		
	Kredi Kartı	Havale	Belirsiz
Meslek = Bankacılık-sigortacılık	90,00%	5,00%	5,00%
Meslek not = Bankacılık-sigortacılık	46,52%	53,35%	0,12%
Meslek not = Bankacılık-sigortacılık and Egitim = doktora	81,58%	15,79%	2,63%
Meslek = Öğretim Görevlisi-Asistan and Egitim = doktora	88,89%	5,56%	5,56%
Meslek not = Bankacılık-sigortacılık and not = Öğretim Görevlisi-Asistan and Egitim = doktora	69,57%	26,09%	4,35%
Meslek not = Bankacılık-sigortacılık and Egitim not = doktora or is missing	45,65%	54,22%	0,13%
Meslek not = Bankacılık-sigortacılık and Egitim = Fakkülte	84,62%	7,69%	7,69%

Çizelge 3.4 (Devam) Karar ağacının tablo şeklinde gösterimi

Meslek not = Bankacılık-sigortacılık and Egitim not = doktora and not = Fakülte or is missing	45,30%	54,57%	0,13%
Meslek not = Bankacılık-sigortacılık and Egitim not = doktora and not = Fakülte or is missing and Quantity = 1	45,15%	54,71%	0,14%
Meslek not = Bankacılık- sigortacılık and Egitim not = doktora and not = Fakülte or is missing and Quantity not = 1	46,67%	52,59%	0,74%

3.5. Karar Ağacı Tekniğiyle Alınan Bilgileri Değerlendirme

Karar Ağacı incelendikten sonra müşterinin eğitim durumuna ve mesleğine bakılarak kredi kartı kullanıp kullanmayacağı tahmin edilebilir. Bu tabloya bakıldığında eğer müşteri:

- bankacılık sektöründe çalışıyorsa yada
- doktorasını bitirmiş bir öğretim görevlisi ise veya

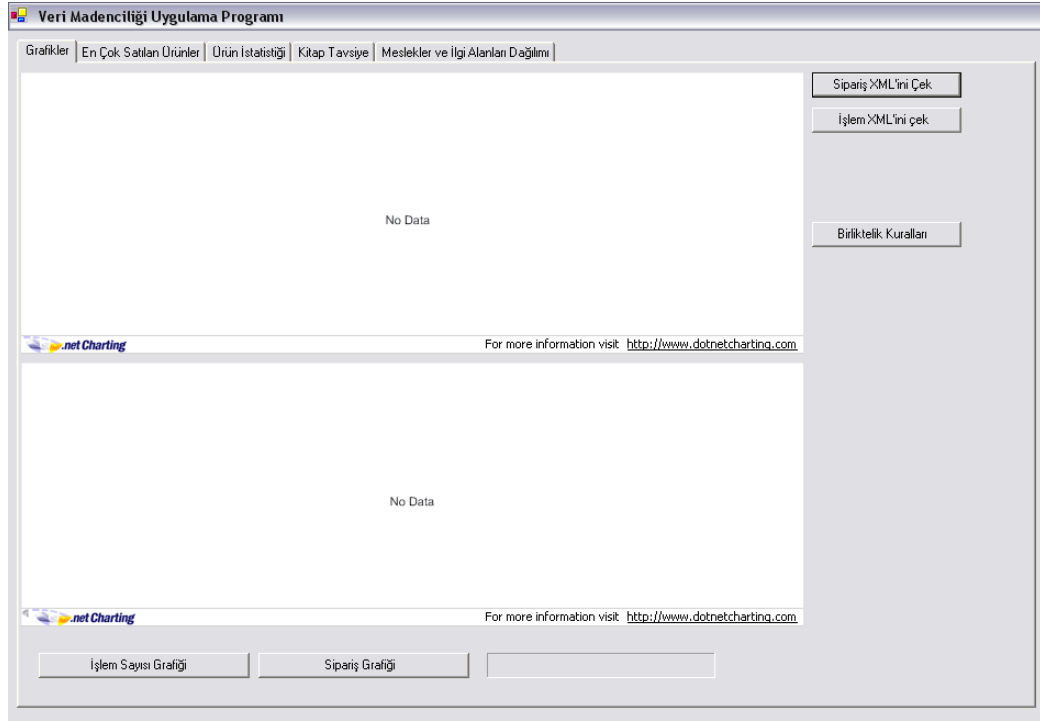
- mesleği bankacılık sektörü değil ama bir fakülte mezunu ise o zaman bu müşterinin ortalama %87 kredi kartı kullanacağı söylenebilir.

3.6. İstatistiksel Veriler

Bu kesimde ileri seviye SQL kullanılarak yapılan Veri Madenciliği açıklanmaktadır.

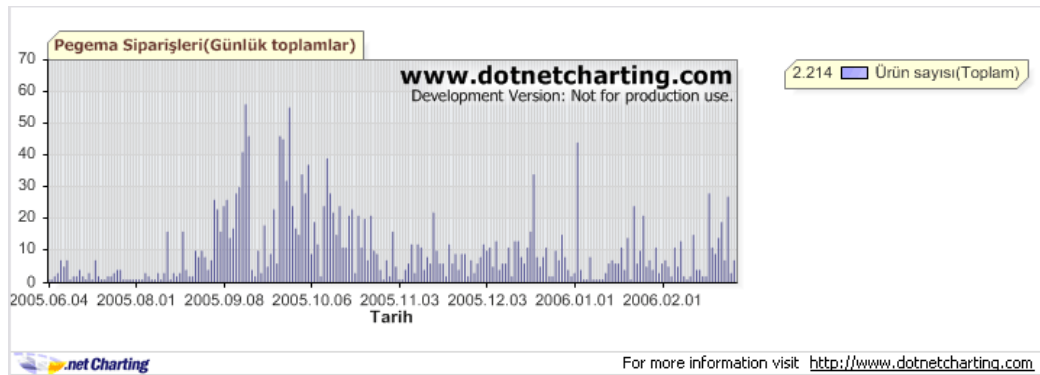
İstatistiksel metotların(İM) bu çalışmada kullanılmasının amacı İM Veri madenciliği metotları kadar karmaşık olmaması ve genel bilgileri kolayca elde edilmesine yardımcı olmasıdır. Örneğin bir kitabın aylar boyunca satış grafiklerini gösterebilmek için İM kullanılması daha uygun olur. İM'den elde edebileceğimiz sonuçlar da VM sonuçları kadar önemlidir. Yazılım sayesinde kitap satış istatistiklerini ayrıntılı inceleyebiliriz ve aynı anda VM metotlarının sonuçları ile karşılaştırabilme imkanımız olmuştur.

Programı çalıştırıldığında bir aşağıdaki(Şekil 3.24) pencere karşımıza çıkmaktadır.

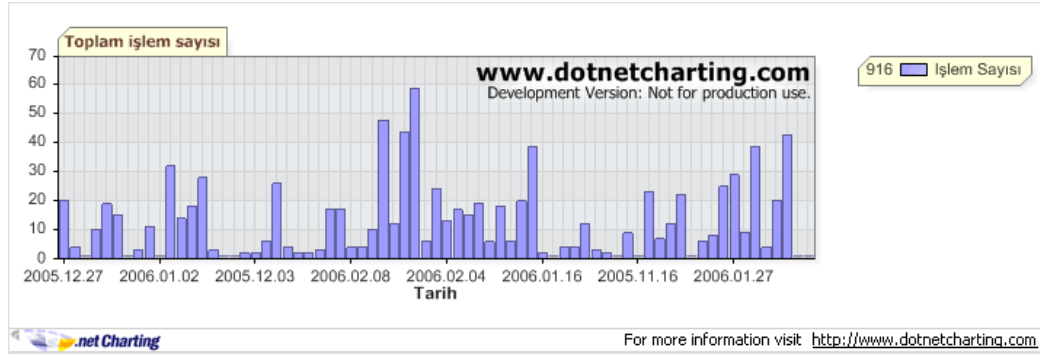


Şekil 3.24 “Pegem’A VM Uygulama Programı” Uygulamasının genel görünümü

Veri Tabanını fazla meşgul etmemek için bilgileri önce birer XML dosyasına çekilmesi daha uygundur. O nedenle sağdaki “Sipariş XML’ini çek ” ve “İşlem XML’ini çek” butonları kullanılıyor. XML’leri aldıktan sonra Sipariş ve ürünlerin üzerinde yapılan işlemlerin sayılarının grafiklerini görmek için “İşlem Sayısı Grafiği” ve “Sipariş Grafiği” butonları kullanılır. “İşlem Sayısı” ve “Sipariş” grafikleri Şekil 3.25 ve Şekil 3.26’da gösterilmektedir.



Şekil 3.25 Toplam sipariş sayısı grafiği



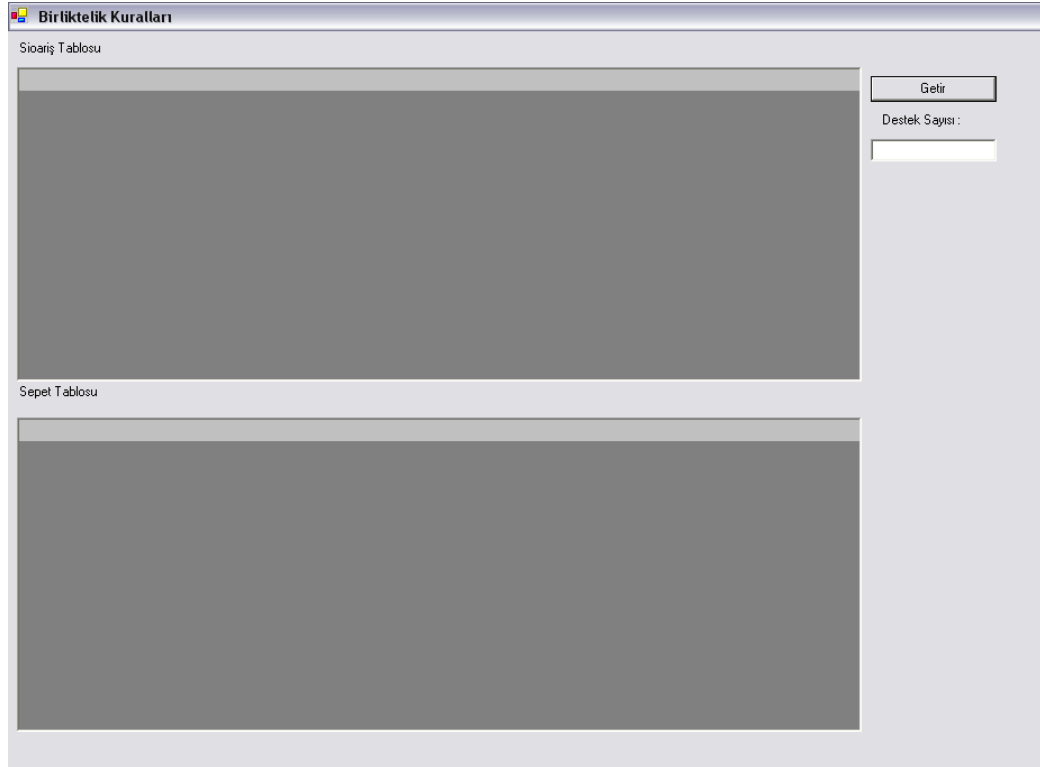
Şekil 3.26 Toplam yapılan işlem sayısı grafiği:

İlk grafik(Şekil 3.25) 6 ay süresi içindeki sipariş sayısının dağılımını göstermektedir. Bu grafiği inceleyerek sipariş sayılarının aylık ortalamasının son 6 ayda değiştiğini (yukarıya doğru) net bir şekilde görülebilir. Ayrıca bu grafikteki 09.08.2005 ve 10.06.2005 tarihleri arasındaki satışların bir ara ani bir şekilde düşüş gösterdiğini görülebilir. Bu aralığı incelediğimizde o aralar kredi kartı satışlarında bir sorun oluştuğunu öğrendik ve ilgili birimlere bu düşüşün sebebini açıkladı.

İkinci grafik(Şekil 3.26) ise site yöneticisinin aktifliğini göstermektedir. Ne zaman site yöneticisi bir ürün üzerinde bir değişiklik yaparsa o an veri tabanına o işlem hakkında bilgi girilmektedir ve sonra da site yöneticisi kendi aktifliğini bu grafikte görsel olarak inceleyebilmektedir.

Birliktelik Kuralları butonunu tıkladığımızda karşımıza Şekil 3.27'deki pencere çıkmaktadır.

Burada önceden açıkladığımız "Birliktelik Kuralları" yöntemini kullanarak bize bu kuralları veren panel yer almaktadır.



Şekil 3.27 Birliktelik kurallarını hesaplama paneli

Burada Destek Sayısı kutusuna istediğimiz destek sayısını girdikten sonra “Getir” butonuna tıkladığımızda sol taraftaki “Sipariş Tablosu” ve “Sepet Tablosu” yerlerine Sipariş ve Sepet için o destek sayısı için Birliktelik Kuralları oluşturulacak ve karşımıza Şekil 3.28 ve Şekil 3.29’daki bilgiler gelecektir.

Sipariş Tablosu

	TransID	Support	ProductID	urunisim	kategori
▶	11	8	26	Çatışma Yönetimi	Yönetim
	1	8	30	Nümerik Analiz	Matematik
	2	4	23	Öğretim Teknolojileri ve Materyal Geliştirme	Eğitim Bilimleri
	2	4	26	Çatışma Yönetimi	Yönetim
	3	4	8	Etkili Okul ve Okul Geliştirme	Eğitim Bilimleri
	3	4	9	Temel Bilgi Teknolojisi Kullanımı	Bilgisayar - İnternet
	4	4	7	Örgütsel Geliştirme	Yönetim
	4	4	8	Etkili Okul ve Okul Geliştirme	Eğitim Bilimleri
	5	7	138	Sınıf İçi Rehberlik Uygulamaları	Eğitim Bilimleri
	5	7	264	Ergenler İçin Grupla Psikolojik Danışma Uygul	Eğitim Bilimleri
	6	6	138	Sınıf İçi Rehberlik Uygulamaları	Eğitim Bilimleri
	6	6	159	Deneyssel Olarak Sınanmış Grupla Psikolojik	Eğitim Bilimleri
	7	7	155	Uzman Öğretmen Başöğretmen Sınavlarına H	Sınavlara Hazırlık

Şekil 3.28 Sipariş tablosunun birliktelik kuralları

Sepet Tablosu

	TransID	Support	ProductID	urunisim	kategori
►	1	4	157	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık	Sınavlara Hazırlık
	1	4	601	Uzman Öğretmenlik Başöğretmenlik Sınavlarına Haz	Sınavlara Hazırlık
	2	4	156	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık	Sınavlara Hazırlık
	2	4	157	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık	Sınavlara Hazırlık
	3	4	157	Uzman Öğretmen Başöğretmen Sınavlarına Hazırlık	Sınavlara Hazırlık
	3	4	946	Milli Eğitim Bakanlığı Sınavlarına Hazırlık El Kitabı "E	Sınavlara Hazırlık
*					

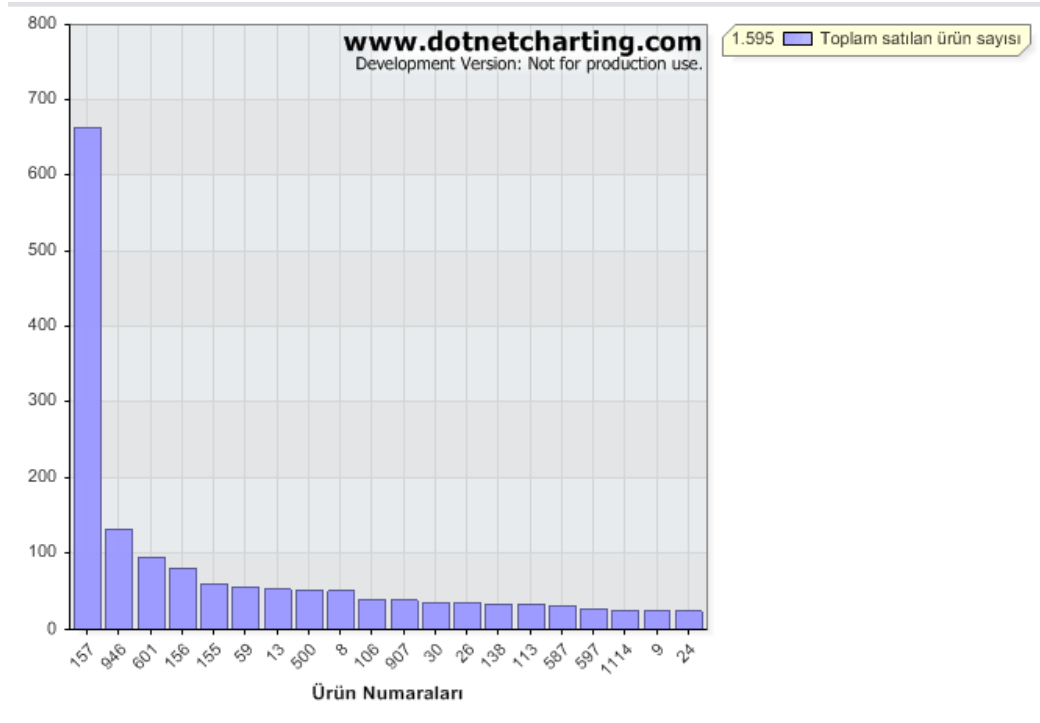
Şekil 3.29 Sepet tablosunun birliktelik kuralları

Bu bilgileri inceleyip gerekli kararları aldıktan sonra programın ikinci Tab'ına geçiyoruz. Burada sistemde en çok satılan 20 ürün ve onlara ilişkin satış grafikleri getirilmektedir.

Bu grafikleri hazırlayan SQL sorgusu şöyledir:

```
"SELECT TOP 20 SUM(quantity) AS mysum, productID FROM      orderDetails
GROUP BY productID ORDER BY SUM(quantity) DESC"
```

Bu sorguyu kullandığımızda Şekil 3.30'da yer alan grafik hazırlanmaktadır. Sorgu satış sayılarına göre ürünlerin grafiklerini getirmektedir.



Şekil 3.30 En çok satılan 20 ürün ve onlara ilişkin satış sayıları grafiği

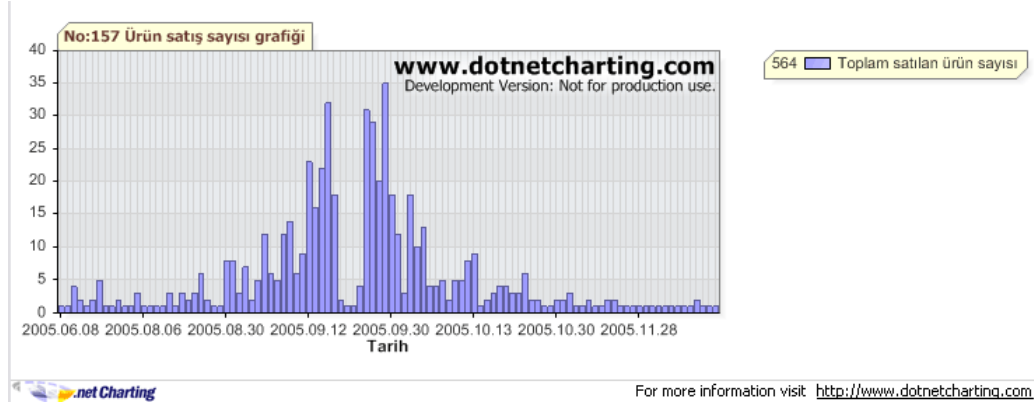
Şekil 3.30'da Y ekseninde ürünlerin satış sayıları, X ekseninde ise onları ürün kodları getirilmiştir. Bu ürün numaralarını kullanarak bir sonraki tab'ta her bir ürün için ayrıntılı satış istatistiklerini inceleyebiliriz. Ürün numaralarına bakıldığında en çok eğitime yönelik kitapların daha başarılı satış sayılarına sahip olduğunu görmekteyiz.

Üçüncü tab'a geçerek bir önceki adımda aldığımız sonuçlardan birini alarak (ürün kod numarasını) o ürün için ayrıntılı satış grafiğini elde ediyoruz. Örnekte 157 No'lu ürünün grafikleri verilmiştir :

1) Satış grafiği:

Bu grafiği hazırlamak için kullanılan SQL sorgusu: "SELECT convert(varchar, orders.orderDate,102) as mydate, SUM(orderDetails.quantity) AS toplam FROM orders INNER JOIN orderDetails ON orders.orderID = orderDetails.orderID where

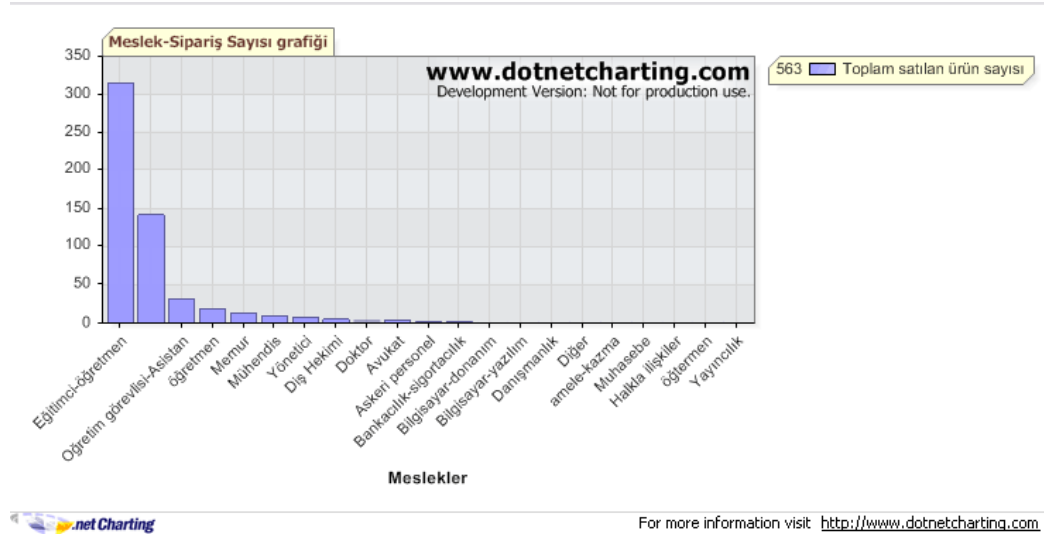
orderDetails.ProductID=" + urun_No.Text + " GROUP BY convert (varchar, orders.orderDate, 102) order By convert (varchar, orders.orderDate, 102) "



Şekil 3.31 157 No'lu ürünün satış grafiği

2) Meslekler ve o mesleklerle göre satış sayılarının grafiği(Hangi meslekteki kullanıcılar kaç kitap satın almış).

Bu grafiği getirmek için kullanılan SQL sorgusu : "SELECT Uyeler.meslek, sum(orderDetails.quantity) AS kacsiparis FROM Uyeler INNER JOIN orders ON orders.customerID = Uyeler.uyeID inner join OrderDetails on OrderDetails.OrderID=orders.OrderID where orderDetails.ProductID=" + urun_No.Text + " GROUP BY Üyeler.meslek ORDER BY COUNT(Uyeler.meslek) DESC"

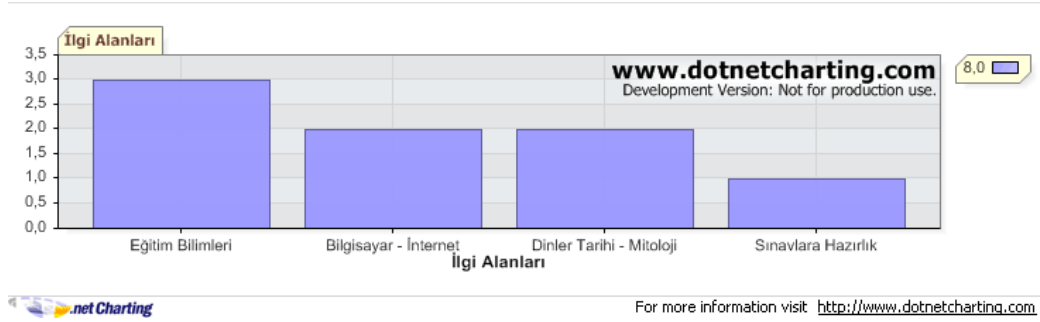


Şekil 3.32 157 No’lu ürünün satışlarına bağlı meslek dağılımı grafiği

Grafiğe bakıldığında(Şekil 3.32) sistemdeki 157 No’lu kitabı en çok “eğitimci öğretmenler ” tercih ettikleri gözükmemektedir. Aynı şekilde istediğimiz ürün için bu meslek dağılımı grafiklerini elde edebiliriz.

3) İlgi alanları (Hangi ilgi alanındaki kullanıcılar kaç kitap almış).

Bu grafiği hazırlamak için kullanılan SQL sorgusu: "Select count(Uyeilgialankategori.kategoriID) as toplam,(select isim from kategori where kategori.id = Uyeilgialankategori.kategoriID) as isim from uyeler inner join Uyeilgialankategori on Uyeilgialankategori.UyeID = Uyeler.UyeID inner join orders on orders.CustomerID=Uyeler.UyeID inner join orderDetails on OrderDetails.OrderID = orders.OrderID where OrderDetails.ProductID = " + urun_No.Text + " group by Uyeilgialankategori.kategoriID order by toplam desc"

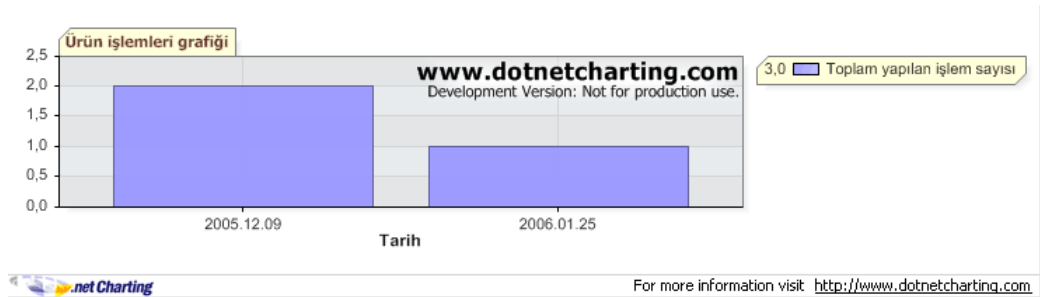


Şekil 3.33 İlgi alanları grafiği

Yukarıdaki grafikte(Şekil 3.33) aynı meslek dağılımları için olduğu gibi ilgi alanları da 157 No'lu ürün için aynı özellikleri taşıdığını görmekteyiz.

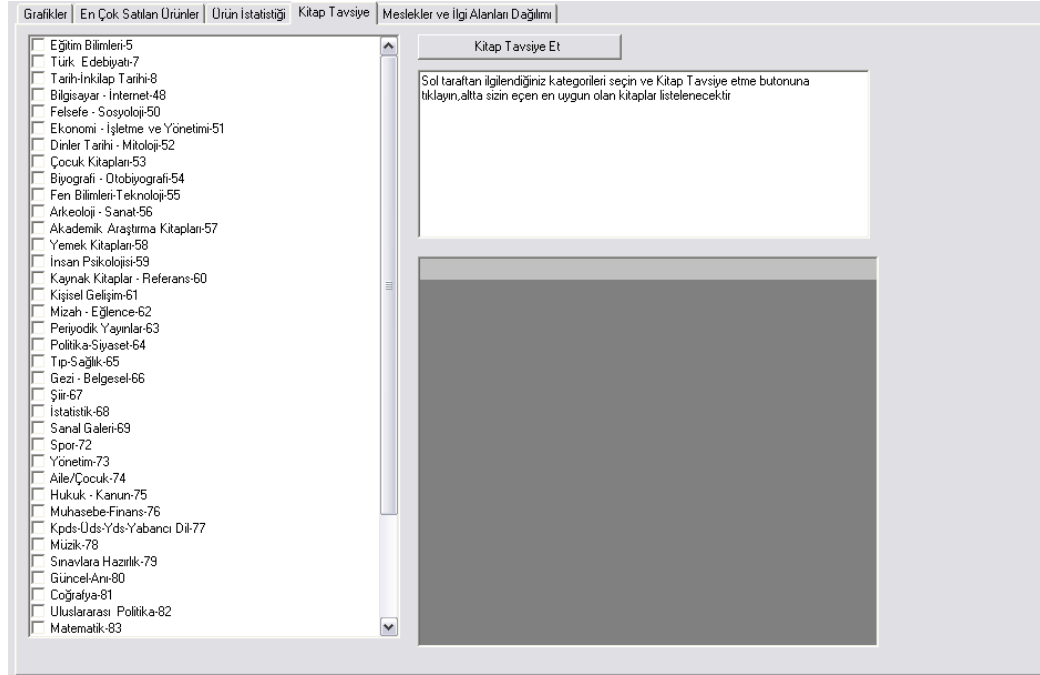
4) İşlem grafiği(Bu ürün üzerinde sistem yöneticisi tarafından yapılan işlemler ve günlere göre sayıları.) Şekil 3.34'te 157 No'lu ürün için grafikler getirilmektedir. Bu grafiği getirmek için kullanılan SQL sorgusu :

"Select convert(varchar,tarih,102) as tarih ,(select count(islemID)) as mycount from islemler where islemler.urunID = " + urun_No.Text + " group by convert(varchar,tarih,102) order by tarih"



Şekil 3.34 İşlem sayısı grafiği

Bir sonraki tab'ta kitap tavsiye etme paneli yer almaktadır. Bu panelin görüntüsü Şekil 3.35'te verilmiştir.



Şekil 3.35 Kitap tavsiye etme paneli

Sol tarafta yer alana kategorilerden herhangi birisini seçerek “Kitap Tavsiye Et” butonuna tıkladığımızda Sonuçlar sağdaki panelde yer alacaktır. Kitap Tavsiye etme işlemi için kullanılan SQL sorgusu:

```
“Select top 20 OrderDetails.ProductID,sum(OrderDetails.quantity) as mysum,
(select isim from urun where urun.id = OrderDetails.ProductID) as myisim from
Uyeilgialankategori inner join Orders on orders.customerID =
Uyeilgialankategori.UyeID inner join OrderDetails on OrderDetails.OrderID =
Orders.OrderID inner join urun on urun.id = OrderDetails.ProductID mystr Group
by OrderDetails.ProductID order by mysum desc,myisim desc”
```

Yukarıdaki SQL sorgusunda aslında birkaç işlem bir araya getirilmiştir. Sonucu elde edebilmek için önce kullanıcının ilgi alanı kodları alınır. Sonra ise tüm alışveriş tablosu ele alınarak alışveriş sahipleri arasında ilgi alanlarının seçilenlerle aynı olan kullanıcıların aldıkları ürünler seçilir. Bu ürünler arasında ise en çok satış sayısına sahip ürünler seçilerek son kullanıcıya gösterilmektedir. Aşağıda örnek

olarak ilgi alanları “Eğitim Bilimleri ve Bilgisayar-İnternet” seçildiğinde elde edilen sonuç ekranı gösterilmiştir.

Sonuç(ilgi alanları “Eğitim Bilimleri ve Bilgisayar-İnternet olduğunda”):

	ProductID	mysum	myisim
►	228	3	Gelişim Öğrenme ve Öğretim
	587	2	Türkçe İlk Okuma Yazma Öğretimi
	591	2	Öğretim Teknolojileri ve Materyal Gelişti
	88	2	Eğitimde Program Geliştirme
	29	1	Öğretimi Planlama Uygulama ve Değer
	127	1	Öğrenme ve Öğretme
	1479	1	Öğrenme Öğretme Süreci
	1217	1	Okulu Yeniden Kurmak
	597	1	Kuramdan Uygulamaya Fen ve Teknoloj
	429	1	Hayat Bilgisi ve Sosyal Bilgiler Öğretimi
	19	1	Hayat Bilgisi ve Sosyal Bilgiler Öğretim
	448	1	Exel XP Makro Ütesi
	1413	1	Etkili ve Verimli -Nitelikli-Eğitim
	590	1	Eğitimde Yeni Yönelimler
	945	1	Çoklu Zeka Teorisi Uygulama Rehberi
	937	1	Aktif Öğrenme Yöntemleri
*			

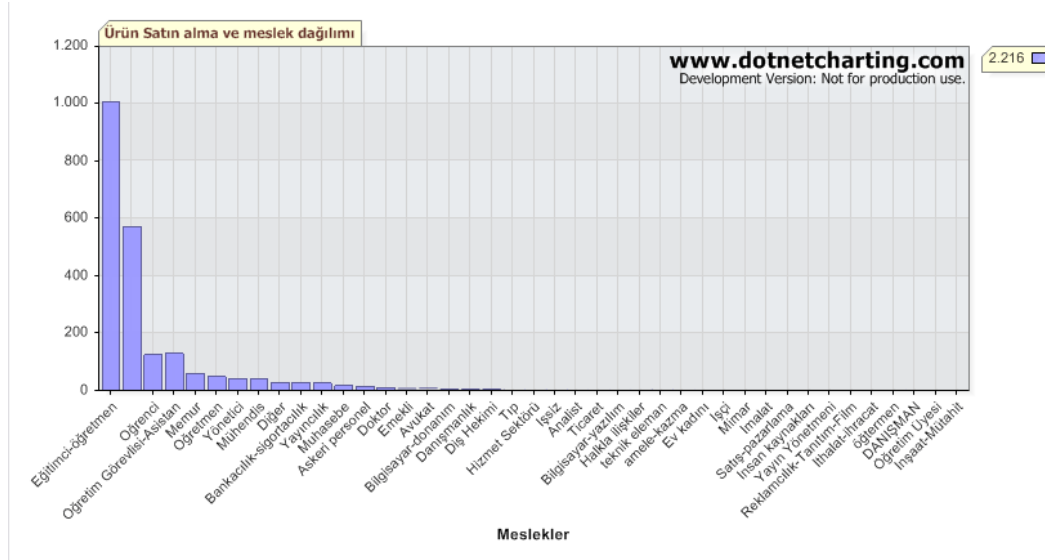
Şekil 3.36 Kitap tavsiye etme işleminin sonucu.

En son tab'ta ise sitede kullanıcılarının meslek,ilgi-alanları ve ürün satışları arasındaki ilişkileri gösteren grafikleri verilmiştir.

1) Bu grafik hangi meslekteki müşteriler kaç ürün satın aldığını gösteren bir grafikdir.

Burada kullanılan SQL sorgusu:

```
"SELECT    Uyeler.meslek, sum(orderDetails.quantity) AS kacsiparis
FROM        Uyeler INNER JOIN orders ON orders.customerID = Uyeler.uyeID
inner join  OrderDetails on OrderDetails.OrderID=orders.OrderID GROUP BY
Uyeler.meslek ORDER BY COUNT(Uyeler.meslek) DESC"
```



Şekil 3.37 Meslek-ürün sayısı dağılımı grafiği

Teker alınan ürün örneklerinde görünen meslek dağılımı bu grafiğe de yansımaktadır. Şekilde görüldüğü gibi sitede alışveriş yapanların çoğunlukla eğitim sektörüne ait oldukları görünmektedir.

2) Bu grafikte site kullanıcılarının meslek dağılımları hakkında bilgiler verilmektedir. Kullanılan SQL sorgusu :

```
"SELECT COUNT(kategoriID) AS SUM,(SELECT kategori.isim FROM kategori
WHERE kategori.id = Uyeilgialankategori.kategoriID) AS isim FROM
Uyeilgialankategori GROUP BY kategoriID ORDER BY SUM DESC"
```


4. SONUÇLAR VE DEĞERLENDİRME

Bu tez çalışması kapsamında hazırlanan program aracılığı ile aşağıdakiler gerçekleştirilmiştir.

“Birliktelik Kuralları” tekniği kullanılarak sistemdeki sepet ve sipariş tabloları incelenmiş ve hangi ürünlerin kategorisinin değiştirilmesi uygun olacağı belirlenerek bunun sebepleri açıklanmıştır.

İstatistiksel metotlar kullanılarak sistemin genel verileri toplanmıştır. Bunlar:

- Sistemdeki kullanıcıların meslek, ilgi alanı dağılımları.
- Meslek ve ilgi alanlarına göre satış grafikleri.
- Belli bir ürün(kitap) için o ürünü satın alan adamlar hakkında bilgiler (meslek, ilgi alanları).

“Microsoft Analysis Services” uygulaması aracılığıyla “Karar Ağacı”(Microsoft Decision Trees) tekniği kullanılarak sisteme giriş yapan kullanıcının alışveriş yaparken hangi ödeme seçeneği(havale yada kredi kartı) kullanacağı konusunda kurallar belirlenmiştir.

Bu çalışma sonucu Birliktelik Kuralları tekniğinin her aşaması gerçek bir sistemde uygulanarak tüm ayrıntılarıyla incelenmiştir. Bu işlemlerin yapılabilmesi için SPSS, SQL BI gibi hazır yazılımlar kullanılarak yapılabilse de, onlar BK tekniğinin nasıl çalıştığını son kullanıcıya göstermez. Tezin amacı ise bu tekniğin çalışmasının tüm aşamalarını inceleyebilmek ve verileri nasıl etkilediğini görebilmek idi. Bu sebeple, bu çalışmada BK’nın önde gelen uzmanlarından Rakesh Agrawal’ın *Hızlı Apriori algoritması* baz alınarak bir yazılım geliştirilmiştir.

Bu çalışmadaki birliktelik kurallarının tespiti için geliştirdiğimiz sadece bir veri tabanı ve bir çeşit tablolarıyla çalışabilmektedir. İleride yazılım çeşitli veri kaynaklarını değerlendirmeye alabilecek şekilde geliştirilebilir.

KAYNAKLAR

1. İnternet: Eker H., Veri Madenciliği Veya Bilgi Keşfi. http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538 (2002).
2. İnternet: Akpınar H., Veri Tabanlarında bilgi keşfi ve Veri Madenciliği. <http://www.isletme.istanbul.edu.tr/dergi/nisan2000/1.htm> (2000).
3. Thuarisingham B. M.(Author),Web Data Mining and Applications in Business Intelligence and counter Terrorism. **Auerbach Publishers, incorporated**, Boca Raton,FL,USA, 35 (2003).
4. Özeken S.,”Veri Madenciliği Uygulaması”, Yüksek LisansTezi, **İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü**, İstanbul, 10-28 (2002).
5. B. Nazife., “Veri Tabanı ve Veri Madenciliği” Sunumu, **Orta Doğu Teknik Üniversitesi**, Ankara, Ekim, 25 (2003).
6. Belgin Ö., “Finding and Evaluating Patterns in Web Repository Using Database Technology and Data Mining Algorithms.”, Yüksek Lisans Tezi, **İzmir Teknoloji Enstitüsü**, İzmir, 29-53 (2002).
7. Larose D.T., Discovering Knowledge in Data: An Introduction to Data Mining, **John Wiley & Sons Inc.**, 42-70 (2005).
8. Akkuş E., “Designing a Customer Relationship Management Model for an Insurance Company.”, Yüksek Lisans Tezi, İstanbul, 10-25 (2002).
9. Yurtoğlu H., “Yapay Sinir Ağları Metodolojisi İle Öngörü Modellemesi: Bazı Makroekonomik Değişkenler İçin Türkiye Örneği”, **Ekonomik Modeller Ve Stratejik Araştırmalar Genel Müdürlüğü**, 12-32 (2005).
10. Yılmaz L., “A Decision Support System Using Data Mining”, Yüksek Lisans Tezi, **Yeditepe Üniversitesi**, İstanbul, 16-22 (2002).
11. Güvenç E., “Student Performance Assesment in Higher Education Using Data Mining”, Yüksek Lisans Tezi, **Boğaziçi Üniversitesi**, İstanbul, 5-14 (2001).
12. Jochen H. and Guido L., Analyzing Warranty Claimsof automobiles: an application description following thw CRISP-DM data mining process, in **Proceedings of the5th International Computer Science Conference(ICSC’99)**, Hong Kong, 31-40 (1999).
13. Alan M. S., A comparision of two data mining techniques to predict abnormal stock market returns, **Intelligent Data Analysis**, Volume:7: 3-13 (2003).

14. Ivkovivh S., John Yearwood, and Andrew S., Discovering interesting association rules from legal databases, *Information and Communication Technology Law*, 11, 50-65 (2002).
15. Tae K. S., Namsik C., and Gunhee L., Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction, *Journal of Management Information Systems*, 16, 63-85 (1999).
16. Simon H. and Brent R., Understanding the domestic market using cluster analysis: a case study of the marketing efforts of Travel Alberta, *Journal of Vacation Marketing*, 8(3): 263-276 (2002).
17. Takcı H., İbrahim S. “Kütüphane Kullanıcılarının erişim desenlerinin keşfi.”, *Gebze İleri Teknoloji Enstitüsü, Bilgisayar Mühendisliği Bölümü*, Gebze, 1-25 (2003).
18. İnternet: Association Rules, Introduction. http://www.resample.com/xlminer/help/Assocrules/associationrules_intro.htm
19. Han J. and Kamber M. Data Mining: Concepts and Techniques. *Morgan Kaufmann*, 45-53 (2001).
20. Alpaydın E., “Ham Veriden Altın Bilgiye Ulaşma Yöntemleri”, *Boğaziçi Üniversitesi, Bilgisayar Mühendisliği Bölümü*, 25-42 , İstanbul (2000).
21. Kantardzic M., “General Data Analysis and Data Mining”, *Wiley-IEEE Press*, 1-18 (2002).
22. Holsheimer M. and Siebes A., “Data mining: The search for knowledge in databases.”, *Technical Report* , CWI, Netherlands, 12 (1994).
23. Piatetsky-Shapiro G., Knowledge Discovery in Databases. *AAAI/MIT Press*, 23 (1991).
24. Breiman L., H. Friedman J., Olshen R. A., and Stone C. J.. “Classification and Regression Trees.” *Wadsworth*, Belmont, 87-92 (1984).
25. Han J., Cai Y., and Cercone N. “Knowledge discovery in databases: An attribute oriented approach.” *In Proc. of the VLDB Conference*, Vancouver, British Columbia, Canada, 547-559 (1992).
26. Quinlan J. R.” Programs for Machine Learning” *Morgan Kaufman*, 225-251 (1993).
27. Pearl J., “Probabilistic reasoning in intelligent systems: Networks of plausible inference”, *Morgan Kaufman*, 130-147 (1992).

28. Piatetsky-Shapiro G., "Discovery, analysis, and presentation of strong rules, knowledge discovery in databases. *AAAI/MIT Press*, (1991).
29. Holsheimer M. and Siebes A., "Data mining: The search for knowledge in databases", *Technical Report CS-R9406*, CWI, Netherlands, (1994).
30. Cheeseman P. et al. Autoclass: "A bayesian classification system.", In 5th Int'l Conf. On Machine Learning., *Morgan Kaufman*, (1988).
31. Fisher D. H. "Knowledge acquisition via incremental conceptual clustering. Machine Learning", *Kluwer Academic Publishers* , (1987).
32. Agrawal R., Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules", *IBM Almaden Research Center*, (1994).
33. Associates D. S., "The new direct marketing." *Business One Irwin*, Illinois, (1990).
34. Agrawal R., T. Imielinski, and A. Swami."Database mining: A performance perspective.IEEE Transactions on Knowledge and Data Engineering", *Special Issue on Learning and Discovery in Knowledge Based Databases*, 5(6):914-925 (1993).
35. Agrawal R., Imielinski T., and Swami A. "Mining association rules between sets of items in large databases." *In Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., (1993).
36. Houtsma M. and Swam A. "Set-oriented mining of association rules". Research Report RJ 9567,*IBM Almaden Research Center*, San Jose, California, (1993).

EKLER

EK-1. Apriori algoritması

```
1)L1={large 1-itemsets};
2)for (k = 2; Lk-1≠∅; k++) do begin
3)    Ck = apriori-gen (Lk-1);// yeni adaylar
4)    forall transactions t ∈ D do begin
5)        Ct = subset (Ck,t); // t'nin içindeki tüm adaylar
6)        forall candidates c ∈ Ct do
7)            c.count ++;
8)    end
9)    Lk = {c ∈ Ck | c.count ≥ minsup}
10)end
11)Cevap =Uk Lk;
```

EK-2. “Apriori Tid” algoritması

- 1) $L_1 = \{\text{large 1-itemsets}\};$
- 2) $C_1 = \text{database } D;$
- 3) for ($k=2; L_{k-1} \neq \emptyset; k++$) do begin
- 4) $C_k = \text{apriori-gen}(L_{k-1});$ // New candidates
- 5) $C_k = \emptyset;$
- 6) forall entries $t \in C_{k-1}$ do begin
- 7) // determine candidate itemsets in C_k contained
 // in the transaction with identifier $t.TID$
 $C_t = \{c \in C_k \mid (c-c[k]) \in t.\text{set-of-itemsets}$
 $(c-c[k-1]) \in t.\text{set-of-itemsets}\};$
- 8) forall candidates $c \in C_t$ do
- 9) $c.\text{count}++;$
- 10) if ($C_t \neq \emptyset$) then $C_k += \langle t.TID, C_t \rangle;$
- 11) end
- 12) $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
- 13) end
- 14) Answer = $\bigcup_k L_k$

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : Kalıkov, Anarbek
 Uyuğu : Kırgızistan
 Doğum tarihi ve yeri : 27.07.1980 Isık-Göl
 Medeni hali : Evli
 Telefon : 0 (312) 435 26 74
 e-mail : anarych1@yahoo.com.

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek lisans	Gazi Üniversitesi /Elektronik Bilgisayar Eğitimi Bölümü	2006
Lisans	Kırgız Teknik Üniversitesi/ Teknik Sistemlerin Otomasyonu Bölümü	2002
Lise	Isık-Göl Hüseyin Karasaev Kırgız-Türk Erkek Lisesi	1998

Yabancı Dil

İngilizce, Rusça

Hobiler

Basketbol, Fotoğrafçılık, Resim, Bilgisayar teknolojileri, Futbol.