T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ ALGORİTMALARINI KULLANARAK ÖĞRENCİ VERİLERİNDEN BİRLİKTELİK KURALLARININ ÇIKARILMASI

Ufuk EKİM

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Ekim-2011
KONYA

TEZ KABUL VE ONAYI

Ufuk EKİM tarafından hazırlanan "Veri Madenciliği Algoritmalarını Kullanarak Öğrenci Verilerinden Birliktelik Kurallarının Çıkarılması" adlı tez çalışması 31/10/2011 tarihinde aşağıdaki jüri tarafından oy birliği ile Selçuk Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Yrd.Doç.Dr. Ömer Kaan BAYKAN	.....
Danışman Yrd.Doç.Dr. Gülay TEZEL	.....
Üye Yrd.Doç.Dr. Hasan Erdinç KOÇER	.....
Üye Unvanı Adı SOYADI	.....
Üye Unvanı Adı SOYADI	.....

Yukarıdaki sonucu onaylarım.

Prof. Dr. ....
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

İmza

Ufuk EKİM

Tarih:

ÖZET

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ ALGORİTMALARINI KULLANARAK ÖĞRENCİ VERİLERİNDEN BİRLİKTELİK KURALLARININ ÇIKARILMASI

Ufuk EKİM

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yrd.Doç.Dr. Gülay TEZEL

2011, 47 Sayfa

Jüri
Yrd.Doç.Dr. Gülay TEZEL
Yrd.Doç.Dr. Ömer Kaan BAYKAN
Yrd.Doç.Dr. Hasan Erdinç KOÇER
Diğer Üyenin Unvanı Adı SOYADI
Diğer Üyenin Unvanı Adı SOYADI

Günümüzde veri tabanları büyük miktarlarda veri depolamaktadır. Bu işlenmemiş, ham verilerden bilgi elde etmek çok zordur. Bilgiye ulaşmak için ise verilerin analiz edilerek, anlaşılabilir veriler haline dönüştürülmesi gerekir.

Büyük miktarlardaki verilerin, kullanılan büyük veri tabanlarında bilgisayar programları vasıtasıyla aranarak, bulunan sonuçlar kullanılarak gelecekle ilgili tahmin yapılması işlemlerine veri madenciliği denilmektedir. Geleceğe dair tahmin yapılabilmesi için geçmişe dönüp, geçmişte bu konularla ilgili ne gibi bilgiler olduğunu ve ne gibi uygulamalar yapıldığını görmek gerekir. Günümüzde bu amaçla birçok algoritma ve yazılım geliştirilmiştir. Bu algoritma ve yazılımlar sayesinde, analistlerin işleri oldukça kolaylaşmıştır.

Bu tez çalışmasında, halen Selçuk Üniversitesinde kullanılan öğrenci işleri otomasyonundan elde edilen veriler üzerinden, öğrenciler hakkında gelecekle ilgili tahmin yapılabilmesi için gerekli birliktelik kuralları çıkarılmıştır. Bu amaçla, bu tezde apriori algoritması ve karar ağacı algoritması kullanılmıştır. Bu kurallar sayesinde, Selçuk Üniversitesini yeni kazanan bir öğrencinin, üniversitedeki başarısına etki eden faktörler araştırılmıştır.

Bu çalışma sonucunda, ailenin eğitim seviyesinin ve gelir düzeyinin öğrencinin başarısında en etkili faktörler olduğu görülmüştür.

Anahtar Kelimeler: Veri Madenciliği, Apriori Algoritması , Karar Ağacı Algoritması,.

ABSTRACT

DEDUCING OF ASSOCIATION RULES FROM STUDENTS DATABASE USING DATA MINING ALGORITHMS

Ufuk EKİM

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF MASTER
IN COMPUTER ENGINEERING

Advisor: Yrd.Doç.Dr. Gülay TEZEL

2011, 47 Pages

Jury
Asst.Prof.Dr. Gülay TEZEL
Asst.Prof.Dr. Ömer Kaan BAYKAN
Asst.Prof.Dr. Hasan Erdinç KOÇER
Diğer Üyenin Unvanı Adı SOYADI
Diğer Üyenin Unvanı Adı SOYADI

In this day and time, data bases are storing the data in great amounts. It is so difficult to get information from these raw datum. For reaching information, datum are to be analyzed and transformed into comprehensible datum. The processes, which datum in great amounts are searched in the great data bases by computer programs and the results found provide that we make a prediction about the future, are called as "*data mining*". To be able to see the future and to make a prediction about the future, we should look at the past and should see what sort information about these subjects were present and what sort applications were done in the past. In our day, in these matters, many algorithms and software have been improved. Thanks to these algorithms and software, analysts' works have been gotten easy considerably.

In this thesis work, it has been reached necessary information to make future predictions about students by the datum acquiring from the student affairs automation that is still used at Selcuk University. By this information and with this aim, in this thesis, it has been used a priori algorithm and decision tree algorithm, Thanks to these rules, it has been researched the factors that affect the success at the university of a newly-registered student to Selcuk University.

At the end of this work, it has been seen that the level of education and income of a family are the most effective factors in the success of a student.

Key Words: Data Mining, Apriory Algorithm, Decision Tree Algorithm.

ÖNSÖZ

Bu tez çalışmasında, bana yol gösteren, rehberlik eden, kısıtlı zamanını benden esirgemeyen, olumlu yaklaşımları ile beni sürekli teşvik eden, her konuda destek veren tez danışmanım Yrd.Doç.Dr.Gülay TEZEL'e teşekkürlerimi sunarım.

Ufuk EKİM
KONYA-2011

İÇİNDEKİLER

ŞEKİLLER

ÇİZELGELER

1. GİRİŞ

Her geçen gün, bilgi miktarı sürekli artmaktadır. Dolayısıyla bu bilgileri saklama ve depolama kapasitelerinde de artışlar yaşanmaya başlamıştır. Bilgi miktarları, yaklaşık olarak her iki yılda bir iki katına çıktığı tahmin edilmektedir. (Frawley, ark., 1991). Bütün sektörlerde veri artışlarının olduğu görülmektedir. Otomasyon sistemlerindeki ve bu sistemlerde kullanılan barkod teknolojisindeki gelişme, büyüyen verinin oluştuğu anda veritabanlarında saklanmasına imkan vermiştir. Çünkü telefon konuşmaları, tıbbi test sonuçları, marketlerde satın alınan ürünler gibi en basit işlemler bile bilgisayar ortamına kaydedilmektedir. Günümüz insanının her bankacılık işleminde, her telefon edişinde kaydedilen veriler her an inanılmaz boyutlarda artmaktadır. Bu artışı daha iyi anlayabilmek için uydu ve uzay araçlarından gelen görüntülerin boyutuna bakma yeterlidir. Bu görüntüler, saatte 50 gigabyte'lık kapasiteyle gelir. Veri tabanı sistemlerinin hacimlerindeki bu olağanüstü artış, organizasyonları elde toplanan bu verilerden nasıl faydalanılabileceği problemi ile karşı karşıya bırakmıştır. Veri Madenciliği, veri tabanlarında bilgi keşfi süreci içerisinde, modelin kurulması ve değerlendirilmesi aşamalarından meydana gelir. Bunun içindir ki, bir çok araştırmacı tarafından Veri Tabanlarında Bilgi Keşfi ile Veri Madenciliği terimlerinin eş anlamlı olarak da kullanılmasına neden olmaktadır.

Özel şirketlerde ve hükümete ait kuruluşlarda bulunan veritabanları da hızla büyümektedir. Örneğin NASA gibi büyük kuruluşlar şimdiden analiz edebileceği veriden daha fazlasını veritabanlarında saklamaktadırlar (Fayyad ve ark., 1996a). Bir terabyte veri üreten dünya gözlem uyduları vardır. Bir günde üretilen resimlere bakılması gerekseydi, her resme bir saniye ayırarak bakan bir kişi, yalnızca resimlere bakma işlemi o kişinin hafta sonları ve geceleri de çalışmak suretiyle bir kaç yılını alacaktı. Farklı sektörlerde bulunan büyük veritabanları, içinde değerli bilgileri barındıran bir veri madeni olarak görülebilir. Fakat bu büyüklükteki veriyi analiz ederek bilgi elde etmek insan yeteneğinin sınırlarını aşmaktadır.

Yeni kuşak donanım ve yazılımlar, büyük miktardaki ham verileri yorumlama alanındaki yetersizlikten dolayı ortaya çıkmıştır. Veritabanlarında bilgi keşfi, çok büyük boyuttaki verileri tam ya da yarı otomatik bir biçimde yorumlayan yeni kuşak araç ve tekniklerin geliştirilmesidir.

Literatürde, çok büyük verilerin, çözümlenip kullanılabilir bilgi haline dönüştürülmesi için kullanılmaya hazır veri içinden faydalı örüntülerin bulunması işlemine birçok terim karşılık gelmektedir. Bunlardan bazıları veri tabanında bilgi keşfi

(VTBK), veri madenciliği (VM), bilgi harmanlamadır. VTBK'nın tanımı ve faaliyet alanı için, gelişmekte olan her araştırma dalında farklı görüşlerin olduğu gibi, bu alanda da birçok görüş vardır. Bazı kaynaklara göre, VM terimi sadece bilgi keşfi metotlarıyla uğraşan VTBK sürecinde yer alan bir adımdır (Fayyad ve ark., 1996). Çeşitli ortamlarda veya çeşitli biçimlerde, kullanılmaya hazır çok büyük verilerin, yaygın bilgisayar kullanımı sonucu biriktiği görülmüştür. İki şekilde büyüme gösteren veri tabanı boyutundan ilki, veritabanında yer alan tutanak sayısının artması ya da nesne sayısının artmasıdır. Diğeri ise her nesne için tanımlı nitelik sayısının artmasıdır. Örneğin astronomi veritabanlarında tutanak sayısı 109'lara ulaşırken sağlık sektöründeki uygulamalarda öz nitelik sayısı 102 ila 103 arasında değişmektedir. Winter Corporation tarafından yapılan bir araştırmada, dünyadaki en büyük veri tabanının belirlenmesi amaçlanmıştır. Sears, Roebuck and Co.'nun sadece karar destek amaçlı kullanılan veri tabanının 1998 yılında 4630 gigabyte'a eriştiği görülmektedir (Akpınar, 1997).

Elektronik ticaret ve online alışveriş mekanizmalarının da artmasıyla birlikte, günümüzde VM'nin önemi ön plana çıkmaktadır. Araştırmacılar, çok hacimli, büyük ve dağınık veri setleri üzerinde yapmış oldukları çalışmalar sonucu aşağıdaki sonuçlara ulaşmışlardır (Akpınar, 2000).

- Yeni bir araştırma sahası olarak ortaya çıkmaya başlayan VM ve bilgi keşfi, özellikle eğitim, tıp ve elektronik ticaret gibi alanlarda kullanılmaya başlamıştır. VM, kullanışlı ve anlamlı bilgiyi, anlamsız veriden çıkarmaya yarayan analiz ve uygulamaya yönelik çalışmaların tamamını kapsar. Geniş veri kümelerinden değişiklikleri, düzensizlikleri ve ilişkileri çıkarmakta kullanılır. Bunun sonucunda, elektronik alışveriş yapan müşterilerin alışkanlıkları, web üzerinde filtrelemeler, genlerin tespiti, gibi karar verme mekanizmaları için önemli bulgular elde edilebilir (Akpınar, 2000).
- Son yıllarda, sayısal veri miktarında bir patlama yaşanmış ve tahminlerin dışında bir artış göstermiştir. Fakat mühendislerin, bilim adamlarının ve analistlerin sayısı değişmemektedir. Bu orantısızlığı gidermek için yeni araştırma problemlerinin çözümleri birkaç gruba ayrılabilir :
  - Yeni algoritma ve sistemlerin geliştirilmesi,
  - Dağıtık VM için altyapıların ve algoritmaların geliştirilmesi,
  - Günümüzde çalışan VM sistemlerinin ilerletilip geliştirilmesi,
  - VM için özel gizlilik ve güvenlik modellerinin geliştirilmesi.

Belirtilen bu tür girişimlerin başarılı sonuç verebilmesi, ancak iş sahalarının ve hükümetin desteği ile olabilir. Yani alt yapılarının, test ortamlarının ve deneysel bileşenlerinin geliştirilmesi gereklidir.

.

2. KAYNAK ARAŞTIRMASI

Sıramkaya (2005), çalışmasında, veri madenciliği çalışmasında Bulanık Mantık çalışması, kişi-kişi ilişkilerini bulmakta uygulanmıştır. Bu uygulamadaki amaç kullanıcıların arama yapmak istedikleri kişilerin isimlerini yazarken yapabilecekleri yazım hatalarını elemektir. İsimlerdeki harflerin konumlarının birbirlerine göre uzaklıklarını temel alarak bulanık mantık kurallarının uygulandığı bir veri madenciliği algoritması kullanılmıştır.

Özçakır (2006), çalışmasında veritabanındaki veriler üzerinde Apriori algoritması uygulamıştır. Uygulama yazılımının çalışması esnasında algoritmanın her aşaması izlenmiştir. Uygulanan Apriori algoritması ile farklı zaman dilimi, farklı satış lokasyonu girdi değerleri doğrultusunda birlikte satın alınan ürünler ile ilgili bağıntılar olduğu gözlemlenmiştir. Genelde aynı ürün grubuna ait ürünlerin en sık birlikte satın alınan ürünler olduğu görülmüştür.

Kalıkov (2006), çalışmasında, 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 Microsoft'un SQL Server 2000'in Analysis Services 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.

Dolgun (2006), çalışmasında Veritabanlarında Bilgi Keşfi, Veri Madenciliği ve Birliktelik Kuralları'nı ayrıntılı olarak incelemiş, veri madenciliğinde istatistiksel çözümlemeye ağırlık vererek bir pazar sepeti çözümlemesi uygulaması gerçekleştirip sonuçları değerlendirmiştir.

Tosun (2006), çalışmasında kredi kartı müşterilerinin kaybedilme nedenleri incelenmiştir. Bu incelemeyi veri madenciliği tekniklerinden karar ağacı algoritması ve C programlama dili kullanılarak yapılmıştır. İnceleme sırasında 30,000 adet müşterinin bilgileri üzerinde bu yöntemle çalışmalar yapılmıştır. Bu yöntemle ortaya çıkarılan kurallar test edilerek doğruluk oranları ortaya çıkarılmış, bunlar istatistiksel tablolarla göz önüne konmuştur. Sonuçta bulunan değerlere bakılarak, müşteri memnuniyetsizliği ve kredi kartı kullanmama nedenleri bulunmaya çalışılmıştır.

Arabacı (2007), çalışmasında örnek veri tabanında bilgi keşfi sürecinin yapılması ve modelin değerlendirilmesinde kullanılan iki ayrı programın karşılaştırılması olmuştur. Bu çalışmasında, hastanedeki servislerden gönderilen idrar, kan vb. örneklerin incelenmesi sonucu elde edilen gram negatif basillere ait verileri, WEKA ve YALE programlarını kullanarak karşılaştırmış ve sonuçları tartışılmıştır.

Şen (2008), çalışmasında Veritabanlarında Bilgi Keşfi, Veri Madenciliği ve Birliktelik Kuralları'nı ayrıntılı olarak incelemiş, veri madenciliğinde istatistiksel çözümlemeye ağırlık vererek bir pazar sepeti çözümlemesi uygulaması gerçekleştirip sonuçlarını değerlendirmiştir. Sonuçta birliktelik kuralları yöntemi ile Pazar Sepeti Çözümlemesi uygulaması yapılmış ve elde edilen sonuçlar tartışılmıştır

Döşlü (2008), çalışmasında, veri madenciliği ile ilgili kavramlar ve özellikle market sepet analizinde kullanmak üzere birliktelik kuralları üreten temel algoritmalar detaylı bir şekilde ele alınmış ve birbiriyle karşılaştırmıştır. Ayrıca, örnek veri setlerinden iki farklı algoritma ile birliktelik kurallarını bulan bir uygulama geliştirmiştir.

Gürgen (2008), çalışmasında, Türkiye'deki market zincirlerinden birinin yedi günlük fişleri kullanılmıştır. Bu fişlerdeki ürünlerin birbirleri ile olan ilişkileri, Birliktelik Kuralları ile Sepet Analizi uygulaması ve Apriori algoritması ile belirlenmiştir. Ürünler arasında bulunan birliktelikler ürün satışlarının, dolayısı ile gelirin arttırılması için kullanılması amaçlanmıştır.

Onat (2008), çalışmasında web madenciliği yöntemi kullanılarak internet üzerinden insanların yaş, cinsiyet, yaşadığı yer, lisans düzeyi gibi özelliklerine bakılarak insanların birbirleriyle olan uyumluluklarının bulunması amaçlanmıştır. Bunun için veri madenciliğinde kullanılan Apriori Algoritması uygulanmış ve sonuçları tartışılmıştır.

Ezerçe (2008), çalışmasında hazır giyim perakende sektöründe faaliyet gösteren bir firmanın alışveriş kayıtları ile alışverişi gerçekleştiren müşteri verileri ele alınmıştır. Öncelikle Birliktelik Kuralları Analizi ile müşterilerin alışveriş alışkanlıkları belirlenmeye çalışılmıştır. Daha sonra Kümeleme Analizi ile müşteriler demografik özellikleri dikkate alınarak bölümlendirilmiş ve Karar Ağaçları ile alışveriş alışkanlıklarını en çok etkileyen değişkenler analiz edilerek, bu konuda bir uygulama sunulmuştur.

Ergun (2008), çalışmasında ürün kategorileri ve ürün sınıfları arasındaki satış ilişkisinin tespit edilmesi amaçlanmıştır. Bu amaçla perakendeci bir işletmenin bir yıllık süre boyunca topladığı alışveriş fişi verileri üzerinde birliktelik kuralları analizi ve

hiyerarşik kümeleme gibi veri madenciliği yöntemleri uygulanmıştır. Sonuçta ise en büyük kural desteğin içecekler=>tatlı ürünler bağlantısında olduğunu görmüştür.

Üçgün (2009), çalışmasında okul otomasyon yazılım hazırlanması ve bu yazılımda öğrenci verileri üzerinde veri madenciliği uygulaması anlatılmıştır. İ.M.K.B ticaret meslek lisesi öğrenci verileri üzerinde yapılan bir veri madenciliği çalışması ile ilgili bilgiler verilmiştir. Yazılım VB.Net programlama dilinde hazırlanmıştır. Veri tabanındaki veriler üzerinde apriori algoritması uygulanarak her aşaması kullanıcı tarafından gözlenmiştir. Elde edilen sonuçlardan öğrencilerin başarısız olduğu dersler arasındaki ilişkiler ortaya çıkarılmıştır.

Bu tez çalışması ve literatür taramasındaki uygulamalar, genel olarak veri madenciliği ve veri madenciliği yöntemleri ile yapıldığı görülmektedir. Veri madenciliği çok geniş bir konu olduğu için, literatür taramasında market-sepet ilişkisi, öğrencinin başarısız olduğu derslerin bulunması, web madenciliği, metin madenciliği, doku-sınıflama gibi konular üzerinde çalışıldığı anlaşılmaktadır. Bu çalışmalarda, genellikle apriori algoritması kullanılarak, sonuçlar karşılaştırılmıştır. Bu tezde, Selçuk Üniversitesi, öğrenci işleri veri tabanında tutulan verilerden yola çıkılarak, hazırlık sınıfında okuyan öğrencilere web üzerinden bir anket yapılmıştır. Yapılan bu anketi, veri madenciliği yöntemlerinden apriori ve karar ağacı algoritmaları kullanılarak, öğrencilerin başarılarını etkileyen faktörler bulunmaya çalışılmıştır. Her iki algoritma da, Matlab programında hazırlanmıştır. Veri tabanından çekilen anket sonuçları, Matlab'da yapılan her iki algoritma ile birliktelik kuralları çıkarılmıştır. Sonuçta, öğrencinin başarısını etkileyen faktörler bulunmuş ve sonuçlar karşılaştırılmıştır.

3. MATERYAL VE YÖNTEM

3.1.Materyal

Bu uygulamada üniversitemizin Oracle veri tabanında tutulan öğrenci işleri verileri kullanılarak, Matlab (2009R) programında kodlanmış apriori ve karar ağacı algoritmaları kullanılarak, öğrenci verilerini çözümleyip, gelecekte bu öğrenciler hakkında tahmin yapmamızı sağlayacak iki yazılım gerçekleştirilmiştir. Bu yazılımlarla hazırlık sınıfında okuyan öğrencilerin başarılarını etkileyen faktörler araştırılmıştır.

Uygulamaların yapıldığı bilgisayar özelikleri :

Pentium(4) işlemci 3.2 GHz
1 GB Ram
80 GB Sabit Disk

3.2.Yöntem

Günümüzde, büyük miktardaki verilerin işlenerek otomatik veya yarı otomatik bir biçimde kullanılabilir bir bilgi haline dönüştürülmesi veri işleme algoritma ve yazılımların önemini oldukça artmıştır.

Yazılan bu tezde, öğrenci işleri veri tabanındaki anlamsız veriler, birliktelik kuralları algoritmalarıyla işlenerek, öğrencilerin gelecekteki eğitim-öğretimiyle ilgili tahmin yapılmasını sağlayacak anlamlı bilgiler çıkarılmış ve tartışılmıştır.

4. VERİ MADENCİLİĞİ

4.1. Veri Madenciliği Nedir?

Son yılların gözde araştırma konularından biri olan VTBK, çok büyük ve karmaşık verileri analiz eder ve bununla ilgili araç ve tekniklerin üretilmesi ile ilgilenir. VTBK, veri seçimi, veri temizleme ve ön işleme, veri indirgeme, VM ve değerlendirme aşamalarından oluşan bir süreçtir. VTBK süreci içindeki bir adım olan VM, önceden bilinmeyen anlamlı ve yararlı verilerin, veritabanından otomatik biçimde elde edilmesini sağlar (Fayyad ve ark.,1996).

Veri ambarındaki karışık ve anlamsız verilerden, kullanışlı bilgilerin çıkarılmasına veri madenciliği denir. Bu da; kümeleme, veri özetleme, değişikliklerin analizi, sapmaların tespiti gibi belirli sayıda teknik yaklaşımları içerir (Frawley ve ark., 1991).

Bir başka değişle veri madenciliği, desenlerin ve düzensiz yapıların yarı otomatik olarak keşfedilmesine de veri madenciliği denilmektedir. Yazılım tekniklerini kullanan veri madenciliği, verilerin analizi ile ilgilenir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir.

İstatistiksel bir yöntem olarak görmek mümkün olan veri madenciliği, birkaç yönde geleneksel istatistikten farklılık gösterir. Görsel sunumlara çevrilebilen veya mantıksal kurallara dönüştürülebilen nitel modellerin çıkarılması amacında olan veri madenciliği, insan ve bilgisayar ara yüzünün birleşmesi gibi düşünülebilir (Vahaplar ve İnceoğlu, 2001).

VM, paralel programlama, veritabanı yönetim sistemleri, makine öğrenimi ve veri ambarlama ve istatistik gibi farklı alanlarda kullanılan işlemleri birleştirmektedir. VM, istatistik ve makine öğrenimi arasında yakın bir bağ vardır (Şen, 2008). Bunlar, veri içindeki düzenlilikleri ve örüntüleri bulmayı amaçlar. VM algoritmalarında kullanılan yöntemlerin çekirdeğini, makine öğrenimi oluşturur. Pek çok VM algoritmalarında kullanılan karar ağacı ve kural tümevarımı, makine öğreniminde de kullanılmaktadır. Makine öğreniminde kullanılan veri boyutuna göre çok büyük olan ve VM algoritmalarında kullanılan örneklem boyutuyla, ikisinin arasında benzerliklerin olduğu kadar, farklılıkların da olduğunu göstermektedir. VM algoritmaları milyonlarca gerçek hayat nesneleriyle uğraşırken, makine öğreniminde kullanılan örneklem boyu,

genelde 100 ile 1000 arasında değişmektedir. Bunların karakteristiği ise artık, eksik, gürültülü ve boş değerler olarak belirlenebilir. Bununla birlikte bilgi keşfetmeye uygun nesne niteliklerinin elde edilme sürecindeki karmaşıklıkla da, yine VM algoritmaları baş etmek zorundadır (Raghavan ve Sever, 1994b).

İstatistikte birçok metodu kullanan VM, nesnelerin değerlerine bağlı çıkarım yapmada bilinen istatistiksel metotlardan ayrılmaktadır. Örneğin, x-kare veya t testi gibi istatistiksel test yöntemleri birden fazla nitelik arasında korelasyon derecesini belirli bir güvenlik arasında verebilmesine karşılık, belirli nitelik değerleri arasındaki ilişkinin derecesini açığa çıkaramazlar. VM disiplini ortaya çıkmadan önce, karar verme mekanizmasında istatistiksel yöntemler çok sık kullanılırdı. Fakat, uygun verinin bulunması sürecindeki en güç adımı oluşturan bu yöntem, VM algoritmalarının uygulama kolaylığı ile karşılaştırıldığında, kullanımı oldukça zordur (Sever ve Oğuz, 2002).

Büyük miktardaki bilgiyi saklama ve etkin bir biçimde erişim sağlayan veritabanı yönetim sistemlerinde, veri düzenlemesi ilgili organizasyonun işletimsel veri ihtiyacı doğrultusunda gerçekleştirilir. Bu ise her zaman bilgi keşfi perspektifi ile bire-bir çakışmaz. VM kullanımına sunulacak olan veritabanındaki veriler, öncelikle temizleme, transfer, vb. işlemlerinden geçirilir. Daha sonra çalıştırılacak olan VM teknikleri, tek başına kullanılacağı gibi bir VTYS ile de entegre olabilirler (örneğin, data mart, çevrim içi analitik işleme: OLAP).

VM algoritmalarının paralel programlama ile beraber gelişinin sebebi, algoritmalarda girdi olarak kullanılan verinin çok büyük olması ve işletim süresinin ise kısıtlı olmasıdır.

Etkin bir VM uygulayabilmek için dikkat edilmesi gereken noktalar aşağıdaki gibi özetlenebilir.

· Farklı tipteki verileri ele alma: Çoklu ortam verisi, coğrafi bilgi içeren veri gibi farklı tipteki veriler üzerinde işlem yapılan gerçek hayattaki uygulamalar, makine öğreniminde olduğu gibi sembolik değildir. Çoklu ortam veritabanları, coğrafi veritabanları ve nesneye yönelik veritabanlarında saklanabilen kullanışlı veriler, düz bir kütük veya ilişkisel veritabanındaki tablolarda da saklanabilir. Bu veriler, saklandığı ortama göre, basit tipte olabileceği gibi karmaşık veri tipleri de olabilir. VM algoritmasının bütün veri tiplerini ele alabilmesi için, veri tiplerinin çeşitliliğinin az olması gerekmektedir. Bu yüzden veri tipine özgü adanmış VM algoritmaları geliştirilmektedir (Sever ve Oğuz, 2002).

· VM algoritmasının etkinliği ve ölçeklenebilirliği: VM algoritması, etkin ve ölçeklenebilir olduğunu zaman, çok büyük boyutlu veri içinden bilgi elde edebilir. Bunun için ise algoritmanın zamanının tahmin edilip, kabul edilebilir bir süre olması gerekir. Üssel veya çok terimli bir karmaşıklığa sahip bir VM algoritmasının uygulanması kullanışlı değildir.

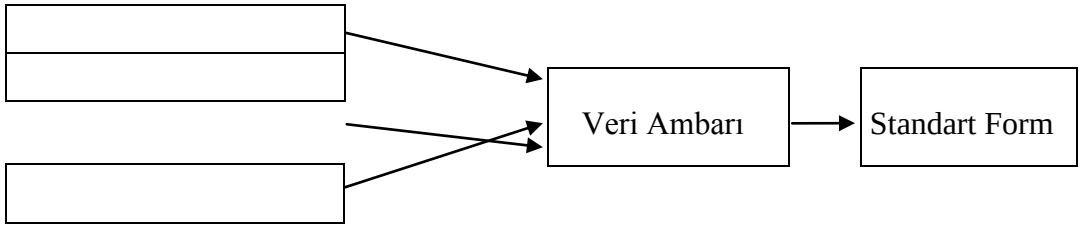
· Sonuçların yararlılık, kesinlik ve anlamlılık kıstaslarını sağlaması: Elde edilen sonuçlar analiz için kullanılan veritabanını doğru biçimde yansıtmalıdır. Bunun yanı sıra gürültülü ve aykırı veriler ele alınmalıdır. Bu işlem elde edilen kuralların kalitesini belirlemede önemli bir rol oynar.

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

· Farklı birkaç soyutlama düzeyi ve etkileşimli VM: Büyük veritabanlarından elde edilecek bilginin tahmin edilmesi güçtür. Bu yüzden VM sorgusu, elde edilen bilgilere göre kullanıcıya etkileşimli olarak sorgusunu değiştirebilmeyi, farklı açılardan ve farklı soyutlama düzeylerinden keşfedilen bilgiyi inceleyebilme esnekliğini sağlamalıdır.

· Farklı ortamlarda yer alan veri üzerinde işlem yapabilme: Kurumlar yerel ağlar üzerinden pek çok dağıtık ve heterojen veritabanı üzerinde işlem yapmaktadır. Bu VM' nin farklı kaynaklarda birikmiş formatlı ya da formatsız veriler üzerinde analiz yapabilmesini gerektirir. Verinin büyüklüğünün yanı sıra dağıtık olması, yeni araştırma alanlarının ortaya çıkmasına sebep olmuştur. Bunlar, paralel ve dağıtık VM algoritmalarıdır.

· Gizlilik ve veri güvenliğinin sağlanması: Gizlilik ve veri güvenliği, VM sistemini kullanan kullanıcının haklarına ve erişim yetkilerine göre sağlanmalıdır. VM büyük miktarda veri inceleme amacı üzerine kurulmuş olduğu için veri tabanları ile yakından ilişkilidir. Gerekli verinin, hızla ulaşılabilecek şekilde amaca uygun bir şekilde saklanması ve gerektiğinde hızla ulaşılabilmesi gerekir. Günümüzde yaygın olarak kullanılmaya başlanan veri ambarları günlük kullanılan veri tabanlarının birleştirilmiş ve işlemeye daha uygun bir özetini saklamayı amaçlar. Şekil 4.1 de görüldüğü gibi günlük veri tabanlarından istenen özet bilgi seçilerek ve gerekli ön işlemeden sonra veri ambarında saklanır. Ardından amaç doğrultusunda gerekli veri ambardan alınarak VM çalışması için standart bir forma çevrilir. (Tiryaki, 2006)



Şekil 4.1 Veritabanı - veri ambarı - standart form

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

Kullanıcının bilgi çıkarma sürecinde katkısının az tutulması, işin olabildiğince otomatik olarak yapılabilmesi, VM'nin en önemli amacıdır. Çünkü OLAP programlarını kullanırken bulunabilecek sonuçlar kullanıcının sormayı düşündüğü sorgularla sınırlıdır. Ama veri içinde çocuk bezi ile bira örneğindeki bağıntı gibi kullanıcının hiç aklına gelmeyecek bilgiler de olabilir. Zaten VM' de esas amaç bu tip bilgileri bulabilmektir (Weiss ve Indurkhya, 1998).

Şekil 4.2' de görüldüğü gibi çeşitli veri kaynaklarından verilerin toplanması ile başlayan VTBK süreci, toplanan verilerin analiz için uygun hale getirilmesi aşaması ile devam etmektedir.

Ancak veri ambarına sahip olan kuruluşlarda, gerekli verilerin Data Mart olarak isimlendirilen işleve özel veri tabanlarına aktarılması ile doğrudan VM işlemlerine başlanabilmesi de mümkündür.

VTBK sürecinde yer alan adımlar şöyledir

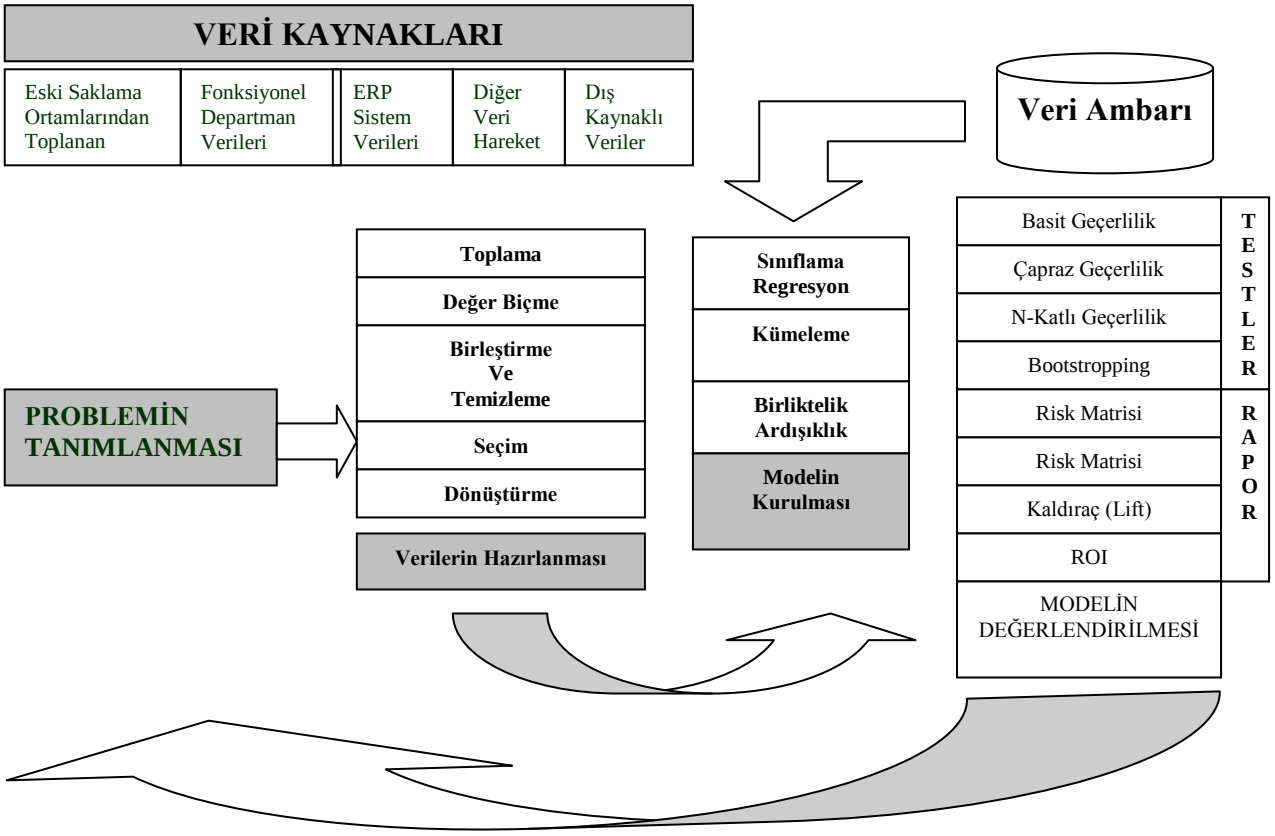
• Veri Seçimi: Bu adım birkaç veri kümesini birleştirerek, sorguya uygun örneklem kümesini elde etmeyi gerektirir.

• Veri Temizleme ve Ön işleme: Seçilen örneklemde yer alan hatalı tutanakların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır. Bu aşama keşfedilen bilginin kalitesini arttırır.

• Veri İndirgeme: Seçilen örneklemden ilgisiz niteliklerin atıldığı ve tekrarlı tutanakların ayıklandığı adımdır. Bu aşama seçilen VM sorgusunun çalışma zamanını iyileştirir (Sever ve Oğuz, 2002).

• Veri Madenciliği: Verilen bir VM sorgusunun (sınıflama, kümeleme, birliktelik, vb.) işletilmesidir.

• Değerlendirme: Keşfedilen bilginin geçerlilik, yenilik, yararlılık ve basitlik kıstaslarına göre değerlendirilmesi aşamasıdır.



Şekil 4.2 Veri tabanlarında bilgi keşfi süreci ve veri madenciliği (Akpınar,2000)

VM birçok dalda uygulanmaktadır. Günümüzde VM teknikleri özellikle işletmelerde çeşitli alanlarda başarı ile kullanılmaktadır. Son 20 yıldır ABD'de çeşitli VM algoritmalarının gizli dinlemeden, vergi kaçakçılıklarının ortaya çıkartılmasına kadar çeşitli uygulamalarda kullanıldığı bilinmektedir. VM tekniklerinin birkaçı ilgili alanlara göre aşağıda özetlenmiştir (Akpınar, 2000).

Pazarlama

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,

- Posta kampanyalarında cevap verme oranının artırılması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi,
- Müşteri ilişkileri yönetimi,
- Müşteri değerlendirme,
- Satış tahmini.

Bankacılık

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri örüntülerinin belirlenmesi

4.2. Yeni Uygulamalar

VM disiplini, bugünkü teknoloji ile tam olarak desteklenemeyen yeni yeteneklere sahip uygulamaları ihtiyaç doğurmuştur. Bu uygulamalar, genel olarak 3 ana kategoride toplanmaktadır (Vahaplar ve İnceoğlu, 2001).

a) İş ve Elektronik Ticaret Verileri: Geri ofis, ön ofis ve ağ uygulamaları iş süreçleri sırasında geniş çaplarda veri üretirler. Bu veriyi karar verme mekanizmalarında efektif olarak kullanmak, ilgili ticari kuruluşun temel yapı taşlarından olmalıdır.

b) Bilimsel, Mühendislik ve Sağlık Bakım Verileri: Günümüzde bilimsel veriler, iş sahası verilerinden daha da karmaşık hale gelmişlerdir.

Buna ek olarak, bilim adamları ve mühendisler uygulama sahası bilgilerini kullanarak simülasyon ve sistem kullanımının arttırılması hedefindedirler.

c) Web Verileri: Internet ve web üzerindeki veriler hem hacim hem de karmaşıklık olarak hızla artmaktadır. Düz metinsel belgeler ve bunlar üzerindeki

uygulamalar (Saraçoğlu ve ark., 2007) önemli bir yer tutmakla beraber resim ve nümerik veriler de web verileri arasında yer almaktadır.

4.3. Veri Madenciliğinde Karşılaşılan Problemler

Bir veri kümesinde çalışan bir sistem, veritabanı büyütüldüğünde tamamen farklı davranabilir. Gürültüsüz ve güzel çalışan bir VM sistemine, gürültü eklendiği zaman, gözle görülür bir biçimde kötüleşme olabilir. İzleyen kesimde günümüz VM sistemlerinin karşı karşıya olduğu problemler incelenecektir (Sever ve Oğuz, 2002).

4.3.1. Veritabanı Boyutu

Veritabanı boyutunun çok büyük olması, VM'nin en önemli sorunlarından biridir. Dolayısıyla VM yöntemleri ya sezgisel/buluşsal bir yaklaşımla arama uzayını taramalıdır ya da örneklemi yatay/dikey olarak indirgemelidir (Sever ve Oğuz, 2002).

Önceden belirlenen genelleme sıradüzenine göre bir üst nitelik değeriyle değiştirilme işlemi yapıldıktan sonra, aynı olan değerlerin çıkarılması işlemine Yatay indirgeme denir. Özellik seçimi yöntemleri veya nitelik bağımlılık çizelgesi uygulanarak yapılan işleme de Dikey indirgeme denir (Han ve ark., 1992).

4.3.2. Gürültülü Veri

Birçok niteliğin değeri, büyük veritabanlarında yanlış olabilir. Bu yanlışlık, değerin yanlış ölçülmesinden veya veriyi giren kişiden kaynaklanan bir hata olabilir. Sistemden kaynaklanmayan ve insan hatasından kaynaklanan hatalara Gürültü denir. Fakat veri girişi sırasında oluşabilecek hataları, ilişkisel veritabanlarının hataları otomatik biçimde gidermesi çok zordur. Gerçek hayatta hatalı veri çok ciddi problem oluşturabilir. Bu ise, gürültülü verilere karşı, VM yöntemlerinin daha az duyarlı olmasını gerektirir. Kapsamlı bir biçimde araştırılan gürültülü verinin yol açtığı problemler, tümevarımsal olarak, karar ağaçlarında uygulanan metotlar bağlamında değerlendirilmiştir. Sistem veri kümesini taramalıdır. Eğer veri kümesi gürültülü ise, sistem bozuk veriyi tanımalı ve ihmal etmelidir. Quinlan (1986a) gürültünün sınıflama üzerindeki etkisini araştırmak için bir dizi deney yapmıştır. Deneysel sonuçlar, etiketli

öğrenmede etiket üzerindeki gürültü öğrenme algoritmasının performansını doğrudan etkileyerek düşmesine sebep olmuştur (Sever ve Oğuz, 2002).

En çok %10'luk gürültülü verinin, eğitim kümesinin içindeki veriden ayıklanabilmesi mümkün olabilmektedir. Chan ve Wong (1991) gürültünün etkisini analiz etmek için istatistiksel yöntemler kullanmışlardır.

4.3.3. Artık Veri

Birçok işlemde, karşımıza probleme uygun olmayan veya artık nitelikler içeren veriler çıkabilir. Bu durum pek çok işlem sırasında karşımıza çıkabilir. Bir problemle ilgili veri elde edebilmek için, iki kümeyi birleştirdiğimizde, sonuçta kullanıcının farkında olmadığı artık nitelikler bulunur. Özellik seçimi olarak adlandırılan, artık nitelikleri elemek için algoritmalar geliştirilmiştir (Sever ve Oğuz, 2002).

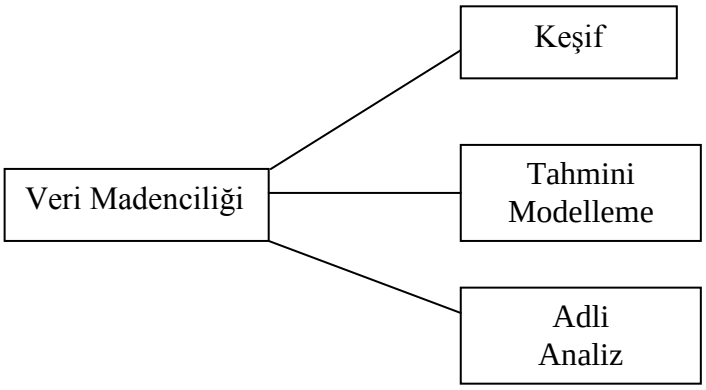
Tanımlamak için, yeterli ve gerekli niteliklerin küçük bir alt kümesinin seçimine özellik seçimi denir. Sınıflama işleminin kalitesini artıran özellik seçimi, aynı zamanda arama uzayını da küçültür (Kira ve Rendell, 1992).

4.3.4. Dinamik Veri

İçeriği sürekli değişen veritabanları dinamik olarak tanımlanır. Bilgi keşfi metodu içim, bu işlemin birçok sakıncası vardır. Öncelikle mevcut veritabanıyla birlikte çalışan ve okuma yapan bilgi keşfi metodunun performansı oldukça düşer. Bilgi keşfi metotları aktif veritabanlarıyla birlikte çalışabilir (Sever ve Oğuz, 2002).

4.4. Veri Madenciliği İşlevleri

İşlevleri açısından VM aktiviteleri 3 sınıfta toplanmıştır: Adli analiz, keşif ve tahmini modelleme (Şekil 4.3) (Tiryaki, 2006).



Şekil 4.3 Veri madenciliği aktiviteleri

Veritabanında, fikir veya hipotez olmadan gizli desenleri arama işlemine keşif denir. Kapasiteli büyük veritabanlarında, kullanıcının bilmeyeceği veya aklına gelmeyecek değişik birçok gizli desenler olabilir. Asıl amaç, buradan çıkarılacak bilginin kalitesi ve desenlerin zenginliğidir.

Kullanıcı herhangi bir bilgiye ulaşabilmek isterse, kurulu olan sisteme sorular yöneltir. Sistem, bu sorulara cevap verebilmek için veritabanındaki verilerle çeşitli kurallar çıkartır. Çıkarılan bu kurallarla sistem, veritabanındaki bilgilere paralel olarak, ileride olabilecek olay, nesne veya kişilerle ilgili tahmin yapmamıza yardımcı olur. Dolayısıyla bu veriler tahmini modellemede kullanılır.

4.5. Birliktelik Kuralları

Genellikle ürün satışlarında, müşteriye daha fazla mal satabilmek veya yapılan alışverişlerde müşterilerin hangi ürünleri alma eğiliminde olduğunun belirlenmesidir. VM'de yaygın olarak kullanılan Pazar sepeti adı altındaki birliktelik kuralları, satın alma eğilimlerinin tanımlanmasını sağlar. Dolayısıyla birliktelik kuralları, değerli bilgi kazanmanın söz konusu olduğu ortamlarda, birbiriyle ilişkili farklı olayların birlikte belirlenmesinde önem taşımaktadır.

Niteliklere göre gruplama yapılmış verileri kullanarak, ilişkide yer alan bir niteliğin alabileceği değerler arasındaki bağımlılıklar bulunur (Agrawal ve ark.,1993).

Eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında birliktelik kuralları kullanılır.

- Müşteriler çocuk bezi satın aldığında, % 75 ihtimalle süt de alırlar,
- Düşük yağlı peynir ve yağsız yoğurt alan müşteriler, % 85 ihtimalle diyet süt de satın alırlar.

Yaygın kullanım alanları katalog tasarımı, mağaza ürün yerleşim planı, müşteri kesimleme, telekomünikasyon vb.dir.

5. VERİ TABANLARINDA BİLGİ KEŞFİ SÜRECİ

Verilerin ve yapılan işin özelliklerinin bilinmemesi, VM algoritmasının ne kadar iyi olursa olsun, fayda vermesi mümkün değildir. Dolayısıyla ilk önce verilerin ve işin özelliklerinin öğrenilmesi gerekir. Daha sonra aşağıdaki aşamalar gerçekleştirilir.

- Problemin Tanımlanması,
- Verilerin Hazırlanması,
- Modelin Kurulması ve Değerlendirilmesi,
- Modelin Kullanılması,
- Modelin İzlenmesi veri tabanlarında bilgi keşfi sürecinde izlenmesi gereken temel aşamalardır (Tiryaki, 2006).

5.1. Problemin Tanımlanması

Uygulamanın hangi işletme amacı için yapılacağı, VM çalışmalarında başarılı olmanın ilk şartıdır.

İşletme problemi üzerine odaklanmış olan işletme amacı, açık bir dille ifade edilmelidir. Sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Bu aşamada doğru veya yanlış tahminlerin faydalarında ilişkin tahminlere de yer verilmelidir (Akpınar, 1997).

5.2. Verilerin Hazırlanması

Verilerin yeniden düzenlenmesine ve bu aşamaya sık sık dönülmesine, modelin kurulma aşamasında çıkabilecek sorunlar neden olabilir. Dolayısıyla bu durum, analizcinin verilerin hazırlanması ve modelin kurulması aşamaları için toplam zamanın yarıdan fazlasında harcanmasına neden olur. Birleştirme, temizleme, toplama, seçme ve dönüştürme gibi adımlar, veri hazırlanma aşamalarıdır (Dolgun, 2006).

5.2.1. Toplama

Bu aşamada verilerin ve veri kaynaklarının toplanacağı adımdır. Nüfus sayımı, merkez bankası kredi veri kaynağı ve hava durumu gibi kuruluşların veri kaynaklarından faydalanılabilen kuruluş, kendi veri kaynağından faydalanılamaz.

5.2.2. Değer Biçme

Farklı kaynaklardan toplanan veriler, VM'de kullanılırken, veriler arasında bir uyumsuzluk olacaktır. Örnek verecek olursak kodlama farklılıkları, farklı zamanlara ait olmaları ve farklı ölçü birimidir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır.

Verilerin ne kadar uyumlu olup olmadıkları incelenerek değerlendirilmelidir. Çünkü uyumlu ve iyi verilerin üzerine kurulmuş modeller ancak iyi sonuç verebilir (Dolgun, 2006).

5.2.3. Birleştirme ve Temizleme

Birleştirme ve temizleme kısmında, veriler tek bir veritabanında toplanır. Bunun için önceki adımdaki sorunlar, büyük ölçüde giderilmelidir. İleriki aşamalarda sorun yaşamamak için, bu adımda yapılacak sorun giderme işlemleri titizlikle çok dikkatli yapılmalıdır (Dolgun, 2006).

5.2.4. Seçim

Veri kümesinin seçimi, modelin eğitimden, bağımlı ve bağımsız değişkenlerde kullanılarak, tahmin edici bir model oluşturulması sağlanır.

Anlamsız değişkenler, diğer değişkenlerin ağırlığını da azaltır. Dolayısıyla modele kimlik numarası ve sıra numarası gibi anlamsız değişkenler sokulmamalıdır. Bu tip anlamsız değişkenleri, kullanılan bazı algoritmalar elese dahi, yine de bu işlem yazılıma bırakılmamalıdır.

Bağımsız değişkenlerin seçilmesine, pratik araçlar ve bunların sunduğu ilişkiler önemli yararlar sağlayabilir.

Veri kümesinden atılacak verilerin, özenle kontrol edilmesi gerekir. Çünkü yanlış veri girişi olsa bile, girilen bu verilerin önemli bir uyarıcı enformasyon içerebilir.

Örnekleme yapılması, büyük veritabanlarında ancak tesadüfiliği bozmamasıyla mümkün olabilir. Çok büyük veritabanlarında birçok modelin denenmesi zaman açısından, hesaplama olanaklarının artmasına rağmen mümkün olamamaktadır. Bunun için rastgele seçilen bir veritabanı parçasıyla, birçok modelin denenmesi mümkündür (Akpınar, 2000).

5.3. Modelin Kurulması ve Değerlendirilmesi

Birçok modelin kurularak denenmesi, tanımlanan problem için en uygun modelin hangisi olduğunu, daha kolay bulmamızı sağlar. Dolayısıyla en iyi modeli buluncaya kadar, sürekli veri hazırlama ve model kurma aşamaları denenir. Denetimli ve denetimsiz öğreniminin kullanıldığı modellere göre, modelin kuruluş süreci farklılık gösterir.

Örnek olarak, önceden belirlenmiş bir kritere göre ilgili sınıflar, bir denetçi tarafından ayrılır. Ayrılan bu sınıfların her biri için çeşitli örnekler verilir. Burada amaç, her bir sınıf için özelliklerin bulunmasıdır. Özellikler için verilen örnekler baz alınır. Bulunan bu özellikler kural cümleleriyle ifade edilir.

Sonuçta verilen yeni örneklere tanımlanan kural cümleleri uygulanır. Kurulan model tarafından yeni örneklerin hangi sınıfa ait olduğu belirlenir. Sınıfların tanımlanması, denetimli öğrenmenin amacıdır. Örneklerin gözlenmesi ve bunlar arasındaki benzerlikler kümeleme analizinde olduğu gibi denetimli öğrenmede de bulunmaya çalışılır (Akpınar, 2000).

Öncelikle ilgili veriler, seçilen algoritmaya uygun bir şekilde, denetimli öğrenim için hazırlanır. İlk olarak verinin bir kısmı modelin öğrenimi için ayrılır. Verinin diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Öğrenim kümesi kullanılarak, modelin öğrenimi gerçekleştirilir. Daha sonra modelin test kümesiyle modelin doğruluk derecesi belirlenir.

Basit geçerlilik testi, bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntemdir. Bu yöntemde verinin yaklaşık olarak %5 ile %33'i, test verisi olarak kullanılır. Modelin öğrenimi kalan veri üzerinde gerçekleştirilir. Daha sonra bu veriler üzerinde test işlemi yapılır. Hata oranı sınıflama modelinde yanlış olarak sınıflanan olay sayısının tüm olay sayısına bölünmesiyle bulunur. Doğruluk oranı ise doğru

olarak sınıflanan olay sayısının tüm olay sayısına bölünmesiyle hesaplanır. Doğruluk oranı ile hata oranının toplamı 1'e eşittir. Bir başka yöntem ise, geçerlilik testidir. Bu teste, sınırlı miktarda veriye sahip olunması durumunda başvurulur. Veri rastgele iki eşit parçaya ayrılır. İlk aşamada birinci parça üzerinde model eğitimi ve ikinci parça üzerinde ise test işlemi uygulanır. Sonraki aşamada ise ilk parça üzerinde test işlemi, ikinci parça üzerinde model eğitimi yapılır. Daha sonra elde edilen hata oranlarının ortalaması alınır ve sonuçta bu ortalama kullanılır (Akpınar, 2000).

n katlı çapraz geçerlilik testi, birkaç bin veya daha az satırdan meydana gelen küçük veritabanlarında tercih edilir. Bu yöntemde veri kümesi n parçaya bölünür. İlk aşamada birinci grup test, diğerleri öğrenim için kullanılır. Sonraki aşamalarda, her defasında bir grubun test, diğerlerinin öğrenim amaçlı olarak kullanılması sürdürülür. Kurulan modelin tahmini hata oranı ise, elde edilen bu n tane hatanın ortalamasıyla hesaplanır (Akpınar, 1997).

Bir başka teknik de Bootstrapping tekniğidir. Bu teknik de, modelin hata düzeyinin tahmin edilmesinde kullanılır. Bütün veri kümesi üzerine kurulu olan model, çapraz geçerlilikteki gibidir. Veri kümesinden oluşturulan hata oranı bazen 200, bazen de binin üzerinde öğrenim kümesi tekrarlı örneklemlerle hesaplanır. Aynı teknikle farklı parametrelerin kullanıldığı veya başka araç ve algoritmaların denendiği değişik modeller kurulabilir. Hangi tekniğin en uygun olduğuna, imkânsız olmasa da, model kuruluş çalışmasında başlamadan önce karar vermek güçtür. Dolayısıyla, farklı modeller kurarak, doğruluk derecesine göre en uygun modeli bulmak için birçok deneme yapılması gerekir (Alpaydın, 2000).

Basit fakat yararlı bir araç olan risk matrisi, kurulan modellerin doğruluk derecelerinin değerlendirilmesinde kullanılan yöntemlerden biridir.

Modelin anlaşılabilirliği, değerlendirmede farklı bir kriterdir. Birçok işletme uygulamasında, doğruluk oranlarındaki küçük artışların yorumlanması büyük önem taşır. Birçok uygulamada ise doğruluk oranındaki küçük artışlar çok önemli olabilir. Model tahmininin altında yatan nedenleri, kural temelli sistemler ve karar ağacı çok iyi ortaya koyabilmektedir (Akbulut, 2006).

Önemli bir yardımcı olarak, kaldıraç oranı ve grafiğini verebiliriz. Çünkü bu oran ve grafik, modelin sağladığı faydanın değerlendirilmesinde kullanılır. Örnek olarak, tesâdüfi seçilen 100 müşterinin 5'i kredi kartını iade eder ve modelin belirlediği 100 kişinin 35'i de iade ederse, kaldıraç oranı 7 olarak bulunacaktır.

Model tarafından önerilen uygulamada elde edilecek kazancın katlanılacak maliyete bölünmesiyle elde edilecek yatırımın geri dönüş oranı kullanılan farklı bir ölçü birimidir. Gerçek dünyayı modelleyebilmek mümkün değildir. Bunun için kurulan modelin doğruluk derecesi çok yüksek olsa bile, gerçek dünyayı yine de tam anlamıyla modelleyemez. Model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmaması, testler sonucunda geçerli bir modelin doğru olamayacağını doğurur. Örnek olarak, bireyin satın alma davranışını varsayılan enflasyon oranının zaman içerisinde değişmesi büyük ölçüde etkileyecektir (Akpınar, 1997).

5.4. Modelin Kullanılması

Bir uygulamanın alt parçası veya doğrudan bir uygulama olarak kullanılabilen modelin, kurulmuş ve geçerliliği kabul edilmiş olması gereklidir. Örneğin, dolandırıcılık tespiti için kurulan bir model promosyon planlaması simülasyonuna entegre olabileceği gibi, işletme uygulamalarında da doğrudan kullanılabilir (Sever ve Oğuz 2002).

5.5. Modelin İzlenmesi

Sürekli olarak kurulan modellerin izlenmesi ve gerekiyorsa yeniden düzenlenmesi sistemlerin özelliklerinde ve bunların ürettikleri verilerde ortaya çıkabilecek değişikliklerden kaynaklanmaktadır. Model sonuçlarının izlenmesinde kullanılan grafikler yararlı bir yöntemdir. Bu grafikler gözlenen ve tahmin edilen değişkenler arasındaki farklılığı gösterir (Akbulut, 2006).

6. BİRLİKTELİK KURALLARI

6.1. Birliktelik Kuralı Çıkarım Algoritmaları

Minimum güvenirlik ve destek metriklerini sağlayan birliktelik kuralı çıkarım problemi iki adıma bölünmüştür (Agrawal ve ark., 1993; Agrawal ve ark., 1994).

1. Kullanıcı tarafından belirlenmiş minimum destek kıstasını sağlayan ürün kümelerinin bulunması: Sık geçen öğe kümesi adı verilen bu kümelerde, verilen örneklemde N adet ürün varsa, potansiyel olarak 2^N adet sık geçen öğe kümesi olabilir. Sık geçen öğe kümelerini bulmak için üstel arama uzayını etkili bir biçimde tarayan yöntemler kullanılmalıdır.

2. Sık geçen öğe kümeleri kullanılarak minimum güvenlik kıstasını sağlayan birliktelik kurallarının bulunması: Burada işlem şöyle yapılmaktadır: l' nin boş olmayan alt kümeleri a ile gösterilsin. Her sık geçen l öğe kümesi için, boş olmayan l' nin tüm alt kümeleri üretilir.

Her a kümesi için a $\Rightarrow$ (l - a) gerektirmesi, l kümesinin destek ölçütünün a kümesinin destek ölçütüne oranı minimum güvenirlik eşiği ölçütünü sağlıyorsa a $\Rightarrow$ (l - a) birliktelik kuralı olarak üretilir. Minimum destek eşiğine göre üretilen çözüm uzayında minimum güvenirlik eşiğine göre taranarak bulunan birliktelikler kullanıcının ilgilendiği ve potansiyel olarak önemli bilgi içeren birlikteliklerdir.

Birinci adım birliktelik sorgusu algoritmalarının performansını belirler. Sık geçen öğe kümeleri belirlendikten sonra, birliktelik kurallarının bulunması düz bir adımdır (Sever ve Oğuz., 2002).

literatürdeki birliktelik sorgusunu yukarıda bahsedildiği biçimde ele alan ve en bilinen algoritma Apriori algoritmasıdır (Agrawal ve ark., 1994).

6.1.1 Karar Ağacı

Karar ağacı öğrenmesi, etkin sonuç çıkarımı için yaygın bir şekilde kullanılan pratik bir yöntemdir. Karar ağaçları, örnekleri kökten yaprağa doğru bir ağaç gibi sıralamayı sağlar (Mitchell, 1997).

Veri madenciliğinde, tanımlayıcı ve tahmin edici özelliklere sahip olan karar ağaçlarının özellikleri aşağıdaki şekilde ifade edilebilir.

- Ucuz maliyetle kurulabilmeleri
- Kolay bir şekilde yorumlanabilmeleri
- Entegre olarak kullanılmak istendiğinde, veritabanı sistemleriyle kolayca uyum sağlayabilmeleri
- Güvenilirliklerinin daha iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahiptir.

Yaygın olarak, karar ağacı temelli analizlerin kullanıldığı sahalar

- Muhtemel olarak belirli bir sınıfın üyesi olacak elemanların belirlenmesi
- Düşük, orta, yüksek risk grupları gibi birçok vakanın çeşitli kategorilere ayrılması
- Oluşturulan kurallarla, gelecekteki olayların tahmin edilebilmesi
- Veri kümesinden faydalı olanların ve parametrik modellerin kurulmasında kullanılmak üzere birçok değişkenin seçilmesi
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması,
- Değişkenlerin sürekli kesikliye dönüştürülerek, kategorilerin birleştirilmesi.

Çeşitli algoritmalarla sürdürülen karar ağacı modellerinin temelleri AID'le (Automatic Interaction Detector) atılmıştır. Geliştirilen bu algoritmalar içerisinde CHAID, ID3, Exhaustive bulunmaktadır (Akpınar, 2000).

6.1.2. ID3 Algoritması

Entropi kavramından yararlanan ID3 algoritması, diğer değişkenler içerisinde sınıflamada en ayırıcı özelliğe sahip değişkeni bulabilmektedir. Eldeki bilginin sayısallaştırılması olarak bilinen Entropi kavramı, veri kümesi içindeki belirsizlik ve rasgeleliği ölçmek için kullanılır. 0-1 arasında bir değer alan Entropi, olasılıkların tamamı eşit olduğunda en büyük değerine ulaşır (Silahtaroğlu, 2008).

Matematiksel olarak entropi şöyle ifade edilir. p değişkeni, 1'den n'e kadar olan olasılıkları ifade ettiğinde

$H(p_1, p_2, \ldots, pn) = \sum_{i=1}^{n} pi \log(1/pi)$ şeklinde olacaktır. (6.1)

Öncelikle veritabanına kaydedilmiş olumlu ve olumsuz verilerin tamamı dikkate alınarak, veritabanının tamamının entropisi hesaplanır (Denk 6.1). Eğer veritabanı alt bölümlere de ayrılıyorsa, daha sonra da bu alt bölümlerin ayrı ayrı entropisi hesaplanır. Veritabanının entropisi bulunduktan sonra, ağaç yapısının kök kısmı ve dalları bulunur. Veritabanının tamamı için bulunan entropi değeriyle, veriler içindeki her bir farklı değişkenler için bulunan değerler, ayrı ayrı çıkarılır. Bulunan her bir sonuca ise kazanım denir (Denk 6.2).

$$\text{Kazanım (D,S)} = \text{H(D)} - \sum_{i=1}^{n} P(Di)H(Di) \qquad (6.2)$$

Bulunan kazanımlar arasında en büyük olan ağacın kök kısmı olarak seçilir. Yine aynı denkleme göre (Denk 6.2), ağacın dalları bulunur. Böylece ağaç yapısı oluşturulur.

6.1.3. Apriori Algoritması

Veritabanının pek çok kez taranarak sık geçen öğe kümelerinin bulunduğu apriori algoritmasında, birinci taramada bir elemanlı minimum destek metriğini sağlayan sık geçen öğe kümeleri bulunur. Sonraki taramalarda bir önceki taramada bulunan sık geçen öğe kümeleri aday kümeler adı verilen yeni potansiyel sık geçen öğe kümelerini üretmek için kullanılır.

Tarama sırasında aday kümelerin destek metriği hesaplanır. Sık geçen ve minimum destek metriğini sağlayan kümeler, aday kümelerden çıkarılır. Sık geçen öğe kümeleri bir sonraki geçiş için aday küme olurlar. Sık geçen öğe kümesi bulunmayana kadar bu süreç devam eder

k-öğe kümesi minimum destek metriğini sağlıyorsa, apriori algoritmasındaki temel yaklaşım bu kümenin alt kümeleri de minimum destek metriğini sağlar (Sever ve Oğuz, 2002).

6.1.3.1. Algoritmaya İlişkin Varsayımlar

Sayısal olan ürün kodları, market sepeti verisinde küçükten büyüğe doğru sıralı olarak kullanılır. Eleman sayılarıyla birlikte anılan öğe kümeleri, k adet ürüne sahip ise, k-öğe kümesi ile gösterilir. Ürün kodları küçükten büyüğe sıralı olan öğe

kümeleri için her öğe kümesine destek metriğini tutmak üzere bir sayaç değişkeni iliştirilmiştir. Öğe kümesi ilk kez oluşturulduğunda sayaç değişkeni sıfırlanır.

Algoritmada kullanılan değişkenler Çizelge 6.1' de özetlenmiştir. Apriori Algoritması Şekil 6.1' de verilmiştir.

Çizelge 6.1 Apriori algoritmasında kullanılan değişkenler

k-öğeküme	K adet öğe içeren küme
Lk	Sık geçen k-öğeküme kümesi (Bu kümeler minimum destek kıstasını sağlar.) Bu kümenin her üyesi iki alandan oluşur. i) öğe kümesi ii) destek sayacı
Ck	Aday k-öğeküme kümesi (Bu kümeler potansiyel olarak sık geçen öğe kümeleridir.) Bu kümenin her üyesi iki alandan oluşur. i) öğe kümesi ii) destek sayacı

C_k : k boyutlu kümenin aday öğesi
L_k : sık geçen öğre kümeleri
L_1 : sık geçen parça

For ($k=1$; $L_k <> \{\}$; k++) do
Begin

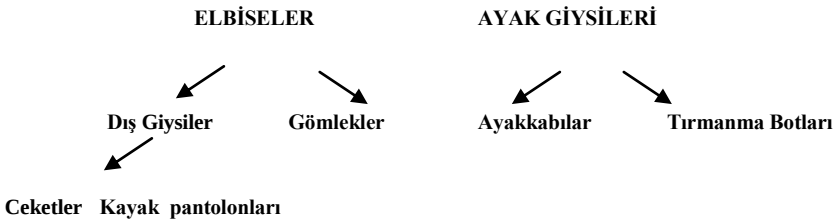
C_{k+1} = L_k oluşan aday
For (Her bir t işlemi için) do
t işlemlerindeki C_{k+1} adaylarının sayısını bir artır
L_{k+1} *minimum destekten geçen* C_{k+1} adayları
End;
L_k Döndür

Şekil 6.1 Apriori Algoritması örneği

Apriori algoritmasının problemi: Bir işlem birçok aday içerebilir. Hash tabanlı öğe küme sayımı, işlem azaltma, bölümleme, örnekleme ve dinamik öğe küme sayımı, apriori algoritmasının etkinliğini artırmak için sunulmuş farklı yaklaşımlardır. (Ding ve Perrizo, 2001).

6.2. Çok Düzeyli Birliktelik Kuralı Çıkarımı

Verinin seyrek olması nedeniyle, birçok uygulamada veri öğeleri arasındaki ilginç ve güçlü birliktelik kurallarını çıkarmak güçtür. Barkod seviyesinde olmayan göreli olarak daha soyut kavram düzeyinde güçlü birliktelik kuralları karşımıza çıkar. Örnek olarak, barkod düzeyinde bulunan bir birlikteliklerden, üst soyutlama düzeyindeki birliktelikler, satış hareketlerine ilişkin bilgilerin tutulduğu veritabanında daha genel bir yapıya sahiptir. Dolayısıyla birliktelik sorguları, farklı soyutlama düzeyleri arasında kolaylıkla geçiş yapabilmeli ve çok düzeyli soyutlamayı da ele alabilmelidir. Pek çok durumda öğeler arasında taksonomi (is-a ilişkisi) mevcuttur. Örnek bir taksonomi Şekil 6.2' de verilmiştir (Sever ve oğuz., 2002).



Şekil 6.2 Örnek bir taksonomi

Çok sayıda taksonomi, veri üzerinde bulunabilir. Örneğin ürün fiyatları kullanılacağı gibi, ürünlerin is-a ilişkisi kullanılarak da taksonomi kurulabilir. Kavram sıralı düzeni de işin içine katacak bir biçimde, birliktelik sorguları için genişletilmiş algoritmalar önerilmiştir (Agrawal ve ark.,1995). Algoritmalarda kullanılan yaklaşımları iki ana grupta toplanabilir. İlk yöntem Basic algoritmasıdır. Bu algoritmada girdi olarak kullanılan örneklemde yer alan her T hareketi, is-a ilişkisi yer alacak biçimde genişletilmiş T′ hareketiyle değiştirilir. T hareketinde yer alan ürünlerin tüm ataları birleştirilerek T′ hareketi elde edilir. Daha sonra eldeki tek boyutlu birliktelik sorgu algoritmalarından biri uygulanır. Buradan da anlaşılacağı gibi, Basic algoritması çok yavaştır. Diğer bir yöntem ise, Basic algoritması üzerinde iyileştirmeler yapılarak elde edilmiş algoritmalardır (Agrawal ve ark., 1995).

T hareketine, T' de yer alan ürünlerin atalarının hepsinin eklenmemesi, Basic algoritmasından yapılan iyileştirmelerden biridir. Ayrıca ata ürünlerin bulunması sırasında ön-hesaplama yapılması ve elde edilen birliktelik kurallarından artık olanları ayıklamak için ilginç eşik değeri kullanılmıştır (Sever ve Oğuz., 2002).

6.3. Birliktelik Sorgusunun Uygulama Alanları

Tıp, finans, mühendislik ve ticaret gibi pek çok farklı alanda uygulanabilen birliktelik sorgusunun kökeni, market sepeti analizi problemine dayanır (Fayyad ve ark.,1996). Veri modelleme, karar destek sistemleri ve gelecekte oluşacak sonuçların tahmini için önemli rol oynayan birliktelik sorgusu, VM'de yer alan temel algoritmalardan biridir.

6.3.1. Market-Sepeti Analizi

Market sepeti analizi, satıcı için hangi ürünlerde indirim uygulanacağı, rafların nasıl düzenleneceği, müşterilerin satın alma alışkanlılarını ve tercihlerini anlamak ve promosyon politikalarının nasıl belirleneceği gibi konularda karar verebilmesi için gereklidir. Etkin bir VM uygulaması, mağazacılık ortamında kullanılırsa market sepeti analizi olarak belirlenir. Birliktelik sorgusu kullanılarak market sepeti analizindeki satış noktalarında oluşan veriler analiz edilir. Etkili reklam ve promosyon faaliyetleri keşfedilen kurallar ışığında başlatılabilir. Örnek olarak, diyet kola içeren kurallar kullanılarak diyet kola satışları artırılabilir (Döşlü, 2008).

6.3.2. Çapraz Satış

Birçok firma, pek çok ürün ve hizmet satmaktadır. Bunun için yeni müşterilerin ilgisini çekmek ve eski müşterileri de kaybetmemek için günümüzdeki hizmet sektöründe güçlü bir rekabet ortamı oluşmuştur. Ürün hizmetlerini geçmişte satın almış müşterileri hakkında firmalar yeterli bilgiye sahiptir. Firma, ürettiği ve satılmamış ürünlere mevcut olan müşteriyi yönelterek hızlı bir şekilde kar edebilir. Firmanın müşteriye ürettiği diğer bir ürünü satma çabası çapraz satış terimi olarak tanımlanır. Bir uzmanla birlikte büyük veritabanlarında, birliktelik sorgusu algoritması uygulanarak çapraz satış problemi çözülebilir (İnan, 2003).

6.3.3. Kısmi Sınıflama

Verilerin sınıflandırılmasını gerektiren, gerçek hayatta birçok problem karşımıza çıkar. Birçok durumda sınıflama işlemi kısmi bir sınıflama yapmayı gerektirir. Bir başka değişle kısmi sınıflama, veri sınıflarının tipik özelliklerini modelleyen bir yöntemdir. Fakat verilerin bir sınıftaki bütün çokluları veya sınıfları içermeyebilir. Çok

sayıda niteliğe sahip olan sıradan sınıflayıcılarda her niteliğin birçok değeri kayıp iken, ekili bir sınıflama yapamazlar.

Birçok sayıda niteliğe sahip bir örneklemde tam sınıflama, çoğu kez mümkün olmayabilir. Bununla birlikte tam sınıflama işlemi istenmeyebilir. Hastalara uygulanan tıbbi testlere ilişkin bilgiler bu tip verilerin örneklerinden biridir.

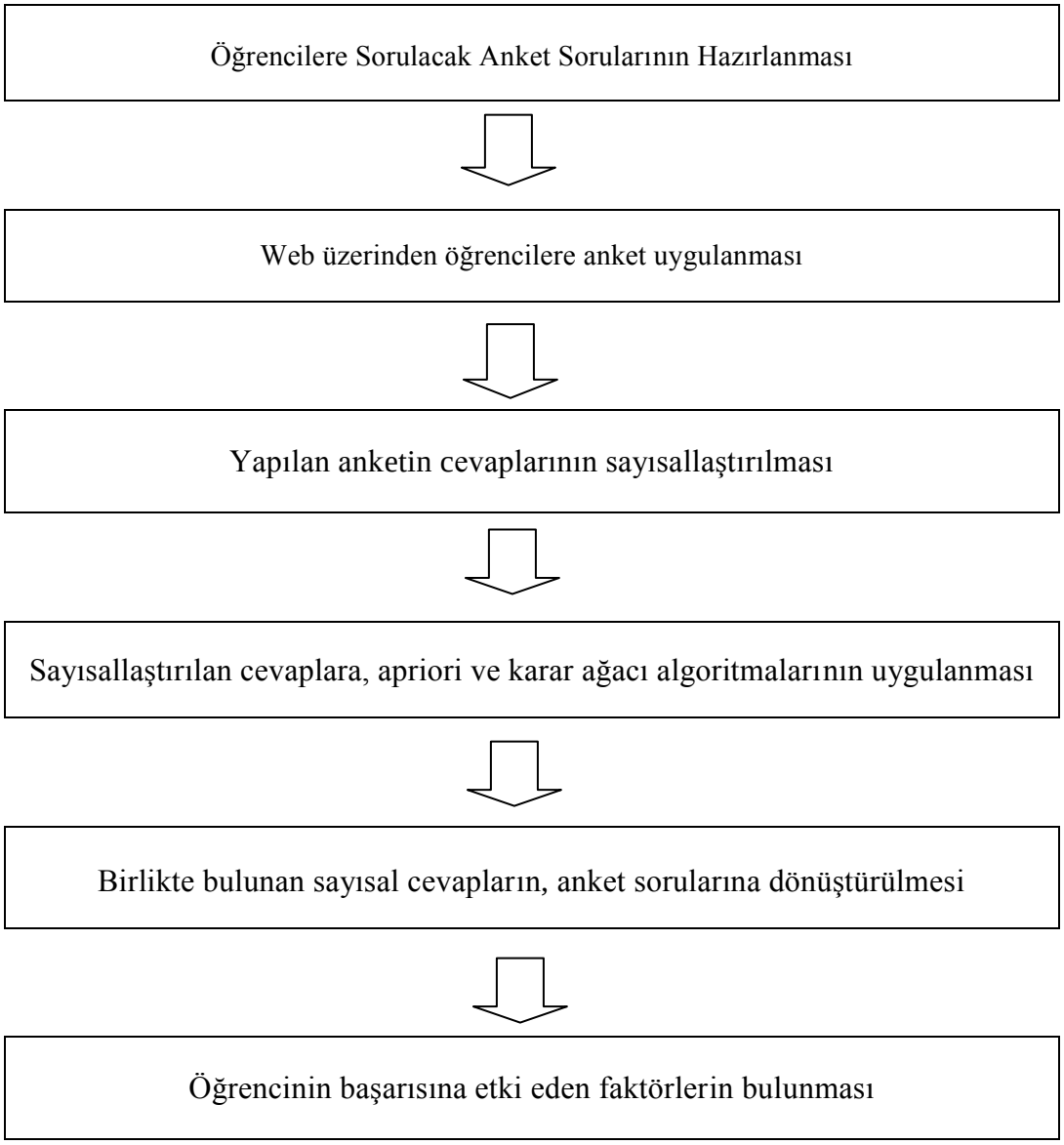
Birçok tıbbi test verisi içinden, ancak çok azı belirgin bir hastaya uygulanır. Hangi test sonucunun diğer test sonuçlarının kombinasyonlarından çıkarılabileceğini birliktelik analizini kullanan doktorlar üretilen kuralları kullanarak bulabilirler. Bunun sonucunda hastalara karmaşık testlerin yerine, daha basit bir test uygulanır veya fazladan test uygulanmaz. Tıp ve iletişim alanlarına iki farklı örnek olay incelemesi yapılarak kısmi sınıflama işleminin birliktelik sorgusu kullanılarak gösterilmiştir. Bütün bunların sonucu olarak, tıbbi testlerin belirlenmesinde ve iletişim araçları sipariş iptalini azaltmada önemli rol oynar (Ali ve ark., 1997).

7. ARAŞTIRMA SONUÇLARI VE KARŞILAŞTIRMA

Eskiden bilgili insan denince, başkalarının ürettiği veya çevreden bir şekilde edindiği bilgileri kafasında tutan ve diğer insanlara bunu sözlü aktaran insan akla gelirdi. Bunun içindir ki eğitim, geçmiş yüzyıllarda var olan bilgi birikiminin, yaşamsal değerlerin ve kültürel değerlerin yeni yetişen kuşaklara anlatım şeklinde aktarılması olarak görülmüştür. Bugün ise bilgili insan denince, bilginin farkında olan, bilgiye ulaşmanın yollarını bilen ve ulaştığı bu bilgilerle yeni bilgiler üretebilerek sorunları çözebilen insan akla gelmektedir. Büyük miktarlarda bilgilerin tutulduğu veri ambarlarından gerekli bilgileri doğrudan alabilmek, bilgilerin çokluğu dikkate alındığında nerdeyse imkânsızdır. Ülkemizde her resmi kurum ve kuruluşa ait bir veri ambarı mevcuttur.

Selçuk Üniversitesi'nde de, yeni gelen öğrencilerin kaydedildiği ve kaydı silinen veya mezun olan öğrencilerin arşivlendiği, şahsi bilgilerinin saklandığı, aldığı derslerin ve notlarının takip edildiği (aldığı veya sildirdiği derslerin kaydının tutulması), yatırmak zorunda oldukları harç miktarlarının tutulduğu büyük bir veri ambarı oluşturulmuştur. Ayrıca bu veri ambarı kullanılarak, öğrencilerin cep telefonlarına SMS yoluyla notları ve aldığı dersler iletilebilmektedir. Bu veri ambarının amacı büyük miktarlardaki verileri depo etmek ve uygun analiz yöntemlerini kullanarak verileri çözümlenebilecek, yorumlanabilecek ve üniversitede karar alma konumunda bulunan kişilere kaynak oluşturabilecek şekilde güçlendirmektir.

Bu tez çalışmasında, üniversitenin veri ambarında bulunan ve hazırlık sınıfında okuyan öğrencilere uygulanan anket kullanılarak, başarılarında etkili ortak bir takım etmenler olup olmadığı araştırılmıştır. Bunun için veri madenciliği yöntemlerinden Apriori ve Karar ağacı algoritmalarıyla, sayısallaştırılmış anket verilerinden birliktelik kuralları çıkarılmıştır. Bu tez çalışmasında yapılan işlemler akış diyagramı olarak Şekil 7.1 de gösterilmiştir.



Şekil 7.1.Yapılan çalışmanın akış diyagramı

Appriori ve Karar ağacı algoritmalarının kodları Matlab programında yazılmıştır. Aynı zamanda veri olarak hazırlık sınıfında okuyan öğrencilere uygulanan anketin sonuçları değerlendirilmiştir. Bu ankette her öğrenciye yedi soru sorulmuştur. Bu öğrencilere uygulanmış anketler arasından rasgele seçilen 200 ve 500 kişinin anket verileri bu tez çalışmasında kullanılmıştır. Böylece öğrenci sayısındaki farklılıkların, çalışmadan elde edilen sonuçlara olan etkisinin ölçülmesi hedeflenmiştir.

Bu anket çalışmasında, her soruda beş cevap şıkkı vardır. Öğrencilerin sorulara verdiklerin cevapların toplam sayılarına göre birliktelik uygulanmış ve bu yöntemle hangi soruların birlikte ön plana çıktığı gözlemlenmiştir.

7.1 Hazırlık sınıfında okuyan öğrencilere uygulanan anket soruları

Bu çalışmada kullanılan ve hazırlık sınıfında okuyan öğrencilere uygulanan ankette, aşağıdaki sorular sorulmuştur.

1) Babanızın Mesleği Nedir?
   a) İşçi b) Memur c) Serbest Meslek d)Çiftçi e)Eğitimci

2) Annenizin Mesleği Nedir?
   a) Ev Hanımı b) Memur c) Serbest Meslek d)Çiftçi e)Eğitimci

3) Babanızın Eğitim Düzeyi Nedir?
   a) İlköğretim b) Lise c) Üniversite d) Lisans Üstü e)Doktora

4) Annenizin Eğitim Düzeyi Nedir?
   a) İlköğretim b) Lise c) Üniversite d) Lisans Üstü e)Doktora

5) Ailenin Aylık Geliri Ne Kadardır?
   a) 250-500 b)500-750 c)750-1000 d)1000-1500 e)1500'in üzeri

6) Mezun Olduğunuz Lise Türü Nedir?
   a) Düz Lise b)Meslek Lisesi c)Fen-Anadolu d)Süper Lise e)Özel Lise

7) Konya'da Nerede Kalıyorsunuz?

a)Ailemin yanında b)Arkadaşlarımla c)Devlet Yurdu d)Özel Yurt e)Akrabalarımla

7.2 Apriori ve Karar Ağacı Uygulamaları

Bu çalışma için kullanılan anketin soruları ve bu sorulara verilen cevaplar harflerden oluşmaktadır. Öncelikle bu sorular ve sorulara verilen cevaplar sayısallaştırarak, program için hazır hale getirilmiştir. Bu amaçla, ankette bulunan toplam yedi soru ve her soruya ait beş cevap seçeneği, toplam otuz beş cevap seçeneği olarak değerlendirilmiştir. Çizelge 7.1'den görüldüğü gibi birinci sorunun beş seçeneği olduğundan, ilk seçenek için 1, ikinci seçenek için 2, üçüncü seçenek için 3, dördüncü seçenek için 4 ve beşinci seçenek için ise 5 rakamı kullanılmıştır. (Yani ilk beş seçenek birinci soruya verilen cevapları temsil etmektedir. 6'dan 10'a kadar ise ikinci soruya verilen cevapları göstermektedir.) Benzer şekilde diğer cevaplar da numaralandırılmıştır. Her öğrencinin verdiği cevaplar bu veri yapısı kullanılarak, aralarında virgülle ayrılmış rakamlardan oluşan bir veri dosyasına atanmıştır.

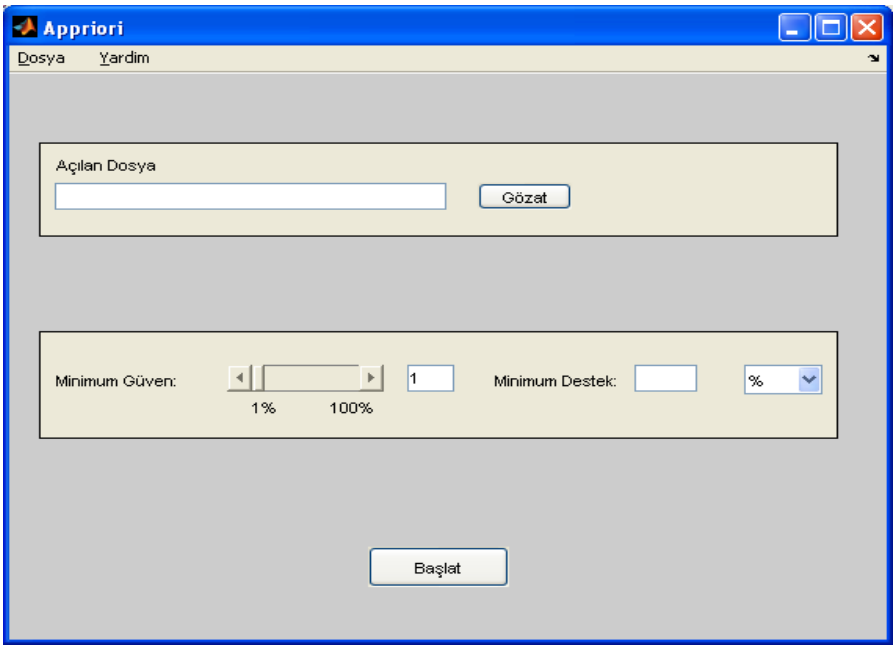
Çizelge 7.1 Anket soru numaraları ve temsil ettiği cevaplar

Soru No	**İçerdiği Cevap No**
1	1-2-3-4-5
2	6-7-8-9-10
3	11-12-13-14-15
4	16-17-18-19-20
5	21-22-23-24-25
6	26-27-28-29-30
7	31-32-33-34-35

7.3 Apriori Algoritması ile Elde Edilen Sonuçlar

Sayısallaştırılmış anket verileri kullanılarak birliktelik kurallarını çıkarmak için uygulanan apriori ve karar ağacı algoritmaları Matlab programı kullanılarak yazılmıştır.

Apriori algoritmasının Matlab programında yazılmış arayüzü Şekil 7.2'de görülmektedir. Kural çıkarımı yapabilmek için, Şekil 7.2'de Gözat butonu tıklanarak veri dosyası seçilir. Daha önce de belirtildiği gibi aralarında birer virgülle ayrılmış bu veri dosyası rakamlardan oluşmaktadır. En küçük destek değeri, istenen düzeyde seçildikten sonra başlat butonu ile uygulama başlatılır. Böylece apriori algoritması ile öğrenci başarısını etkileyen faktörleri bulmak için belirlenen güven aralığı ve minimum destek düzeyinde öğrencilerin verdiği cevaplara göre birliktelik kuralları çıkartılacaktır. Bu uygulama ile önce apriori algoritmasının genel mantığıyla verilen cevapların sayısı bulunur. Daha sonra minimum destekten az olanlar elenir.



Şekil 7.2 Matlab'da yazılan apriori programının genel görünümü

Daha sonra ikili olarak birlikte verilmiş cevaplar taranır ve bu şekilde devam ederek verilen cevaplardan, en çok hangi sorularda daha çok birliktelik olduğu anlaşılmış olur. Çıkan sonuçlarda ifade edilecek olan başarı oranı ise, öğrenci başarısı değil, soruların birlikte bulunma oranlarının yüksek çıkması olarak tanımlanacaktır.

Ankete katılan öğrenciler hazırlık sınıfında okuyor olup, aynı zamanda kendi bölümlerinden de ders almaktadırlar. Buna göre bölüm derslerinin hiçbirinden başarısız olmayan, yani bölüm derslerini geçen öğrenciler başarılı, diğer öğrenciler başarısız kabul edilmiştir.

Apriori algoritmasında minimum destek 2 alınarak program çalıştırılır. Veri tabanından rastgele seçilen 200 kişiye ait anket verileri, minimum destek değeri 2 alınarak çalıştırılan programda, "Annenizin mesleği nedir?" sorusu ile "Anne ve babanın eğitim düzeyleri nedir?" ve "Mezun olduğunuz lise türü nedir?" sorularının öğrencinin başarısında büyük bir etmen olduğu gözlenmiştir. Buna göre elde edilen sonuçlar Çizelge 7.2'de görüldüğü gibidir. Cevapları birlikte bulunma oranını x değişkeni ile, bu oranın bulunmasında sayılan öğrenci sayısını y değişkeni ile ve toplam öğrenci sayısını da z değişkeni ile gösterecek olursak, bu oranın hesaplanması denklem 7.1'deki gibi olacaktır.

$$X = \frac{y}{z} * 100 \qquad (7.1)$$

y : Cevapları Birlikte Bulunan Öğrenci sayısı

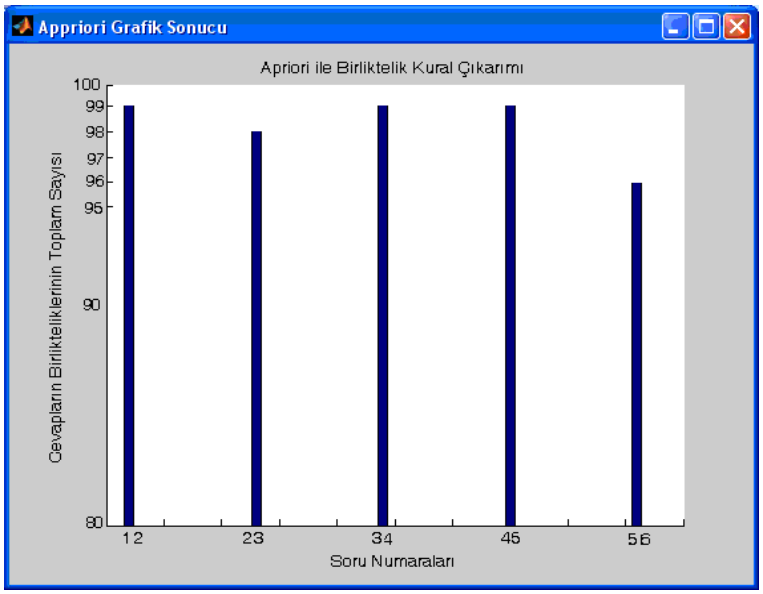
z : Cevapları Birlikte Bulunan Toplam Öğrenci Sayısı

Uygulanan anket sonucuna göre, başarılı öğrencilerin verdiği cevapların genel olarak birlikte bulunma oranı % 54,1 olup, başarısız öğrencilerin verdiği cevapların genel olarak birlikte bulunma oranı ise % 45,8 olduğu görülmektedir. Çizelge 7.2'ye göre babası eğitimci ve annesi ev hanımı olan öğrencilerin 99'u % 53,8'lik oranla başarılı, 85 öğrenci ise % 46,1'lik oranla başarısız olduğu sonucu çıkmaktadır. Annesi ev hanımı olup, babasının ise üniversite mezunu olduğu öğrencilerin 98'i, % 57,3'lük bir oranla başarılı, 73 öğrencinin ise % 42,6'lık bir oranla başarısızdır. Babası lise mezunu olup, annesi ilkokul mezunu olan öğrencilerin 99'u, % 49,5'lik oranla başarılı, 101 kişi ise başarısızdır. Annesi lisans mezunu ve ailenin geliri 2000 TL üzeri olan öğrencilerin 99'u, % 54,3 ile başarılı, 83'ü ise % 45,6'lık bir oranla başarısızdır. Ailenin aylık geliri 2000 TL üzeri ve fen lisesi mezunu öğrencilerin 96'sı, % 56,8'lik bir oranla başarılı, 73'ü ise % 43,1'lik bir oranla başarısızdır.

Çizelge 7.2 200 Kişilik gruba minimum destek 2 alındığında çıkan sonuçlar

Birlikte Bulunan Sorular	Durum	Öğrenci Sayısı	Cevapları Birlikte Bulunma Oranı %
Babası eğitimci ve annesi ev hanımı	Başarılı	99	53,8
	Başarısız	85	46,1
Annesi ev hanımı olup, babasının ise üniversite mezunu	Başarılı	98	57,3
	Başarısız	73	42,6
Babası lise mezunu olup, annesi ilkokul mezunu	Başarılı	99	49,5
	Başarısız	101	50,5
Annesi lisans mezunu ve ailenin geliri 2000 TL	Başarılı	99	54,3
	Başarısız	83	45,6
Ailenin aylık geliri 2000 TL üzeri ve fen lisesi mezunu öğrenciler	Başarılı	96	56,8
	Başarısız	73	43,1
Genel	Başarılı	491	54,1
	Başarısız	415	45,8

İki yüz kişilik bir gruba, minimum destek değeri 2 alınarak uygulanan anket sonuçlarının grafiği Şekil 7.3 gösterilmiştir.



Şekil 7.3 200 kişilik anket grubu için minimum destek değeri 2 olduğunda çıkan grafik

Şekil 7.3'deki grafikteki çubukların soru numarası satırında bulunan değerin anlamı, iki sorunun birlikte bulunmasıdır. Soru numarası yan yana yazdırılmıştır. Örneğin ilk çubuğa göre 1. ve 2. sorunun birlikte bulunduğunu gösterir. Düşey satır ise, bu iki soruya birlikte verilen cevapların toplam sayısını göstermektedir. Yani bu duruma göre, Çizelge 7.2'de görüldüğü gibi, 1.ve 2. soruya toplam 99 kişi aynı cevap vermiş olup, cevap verilme oranı ise % 53,8 olarak çıkmıştır.

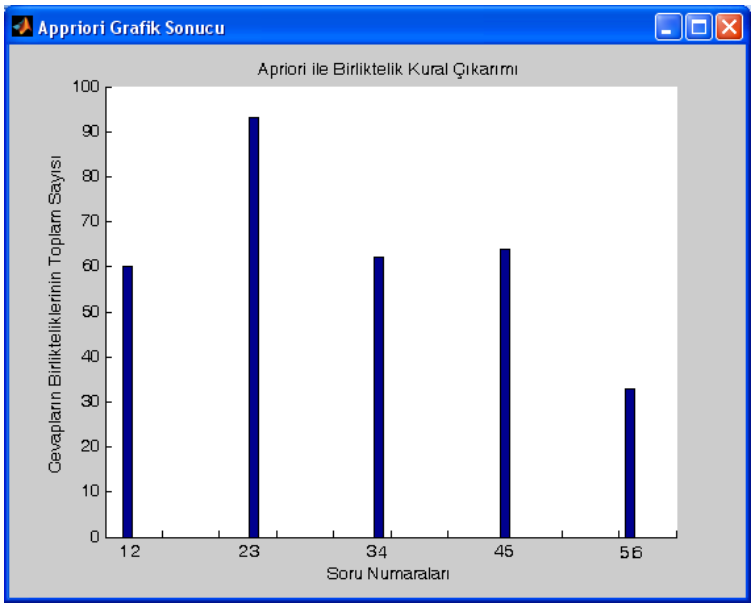
Çizelge 7.2'ye göre minimum % 49,5'lik oranla babasının ve annesinin eğitim düzeyi sorusuna birlikte verilen 99 cevap olmuştur. Yine bu çizelgeye göre maximum % 57,3'lük bir oranla annesinin mesleği ve babasının eğitim düzeyi sorusuna, birlikte 98 kişinin cevap verdiği görülmüştür.

Benzer şekilde veri tabanından rasgele seçilen 200 kişiye ait anket verileri, minimum destek değeri bu defa 3 alınarak çalıştırılan programda elde edilen sonuçlar Çizelge 7.3'deki gibidir.

Çizelge 7.3 200 Kişilik gruba minimum destek 3 alındığında çıkan sonuçlar

Birlikte Bulunan Sorular	Durum	Öğrenci Sayısı	Cevapları Birlikte Bulunma Oranı %
Babası eğitimci ve annesi ev hanımı	Başarılı	60	53,5
	Başarısız	52	46,4
Annesi ev hanımı olup, babasının ise üniversite mezunu	Başarılı	93	55,3
	Başarısız	75	44,6
Babası lise mezunu olup, annesi ilkokul mezunu	Başarılı	62	39,7
	Başarısız	94	60,2
Annesi lisans mezunu ve ailenin geliri 2000 TL	Başarılı	64	55,1
	Başarısız	52	44,8
Ailenin aylı geliri 2000 TL üzeri ve fen lisesi mezunu öğrenciler	Başarılı	33	51,5
	Başarısız	31	48,4
Genel	Başarılı	312	50,6
	Başarısız	304	49,3

200 kişilik bir gruba, minimum destek değeri 3 alınarak uygulanan anket sonuçlarının grafiği Şekil 7.4 gösterilmiştir.



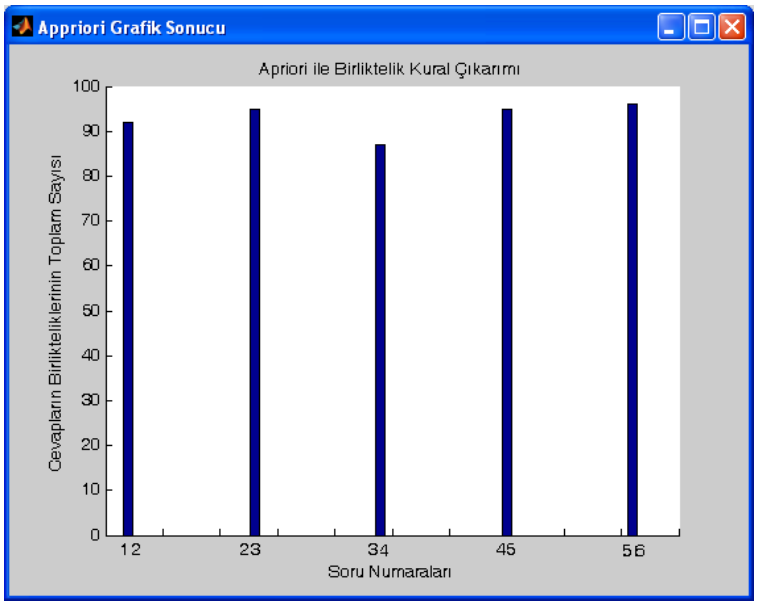
Şekil 7.4 200 kişilik anket grubu için minimum destek değeri 3 olduğunda çıkan grafik

Benzer şekilde yine veri tabanından rasgele seçilen 500 kişiye ait anket verileri, minimum destek değeri 2 alınarak çalıştırılan programdan elde edilen anket sonuçları Çizelge 7.4'deki gibi çıkmaktadır. Burada da verilen cevaplar ve çıkan oranlar, öğrencilerin birlikte verdiği cevapların sayıları ve oranlarıdır.

Çizelge 7.4 500 Kişilik gruba minimum destek 2 alındığında çıkan sonuçlar

Birlikte Bulunan Sorular	Durum	Öğrenci Sayısı	Cevapları Birlikte Bulunma Oranı %
Babası eğitimci ve annesi ev hanımı	Başarılı	92	53,1
	Başarısız	81	46,8
Annesi ev hanımı olup, babası ise üniversite mezunu	Başarılı	95	55,8
	Başarısız	75	44,1
Babası lise mezunu olup, annesi ilkokul mezunu	Başarılı	87	47,2
	Başarısız	97	52,7
Annesi üniversite mezunu ve ailenin geliri 2000 TL	Başarılı	95	53,9
	Başarısız	81	45,7
Ailenin aylı geliri 2000 TL üzeri ve fen lisesi mezunu öğrenciler	Başarılı	96	54,9
	Başarısız	73	45,0
Genel	Başarılı	465	53,3
	Başarısız	407	46,6

500 kişilik bir gruba, minimum destek değeri 2 alınarak uygulanan anket sonuçlarının grafiği Şekil 7.5 gösterilmiştir.



Şekil 7.5 500 kişilik anket grubu için minimum destek değeri 2 olduğunda çıkan grafik

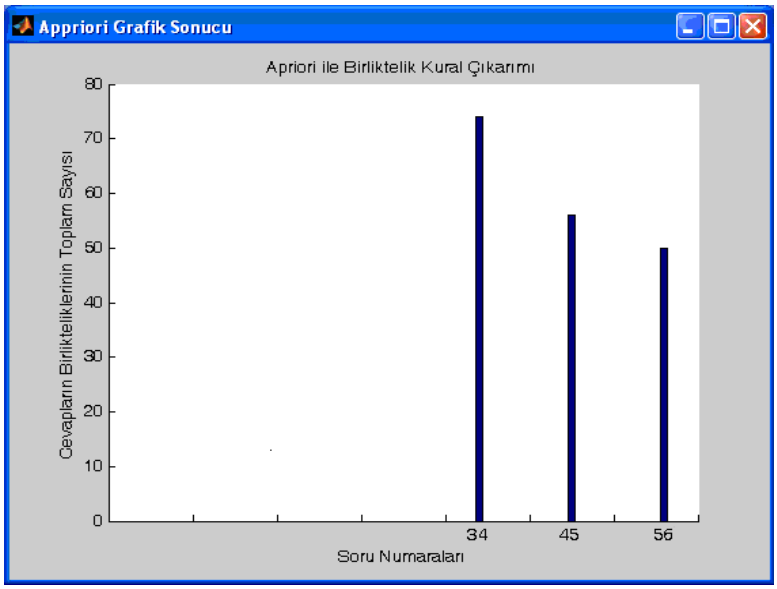
Çizelge 7.4'te, başarılı öğrenciler için, annesinin mesleği ile babasının eğitim düzeyi sorularının 95 öğrencinin aynı cevapları vermesiyle, % 55,8'lik bir başarı ortalaması (Denk. 7.1) bulunmuştur. Yine aynı çizelgede, başarısız öğrenciler için, babasının eğitim düzeyi ile annesinin eğitim düzeyi sorularına 97 öğrencinin aynı cevapları vermesiyle, %52,7 lik bir başarısızlık ortalaması (Denk. 7.1) bulunmuştur. Bulunan bu oranla, birlikte bulunan sorular içinde, oranı en yüksek çıkan soru grubu olduğu görülmektedir. Bunun anlamı ise, annesinin mesleği ve babasının eğitim düzeyinin, öğrencinin başarısında büyük etmen olduğu görülmektedir. Bu sorudan daha az cevapların birlikte bulunduğu diğer soru gruplarından 4. ve 5. sorular ile 5. ve 6. soruların da 95 ve 96 kişilik cevaplama sayısıyla öğrenci başarısında etkili olduğu görülmektedir. Bunun anlamı ise annenin eğitim düzeyi ile ailenin aylık geliri ve ailenin aylık geliri ile mezun olduğu lise, öğrencinin başarısına etki ettikleri görülmektedir.

Veri tabanından rasgele seçilen 500 kişiye ait anket verileriyle, minimum destek değeri 3 alınarak çalıştırılan programdan elde edilen sonuçlar Çizelge 7.5'de görüldüğü gibi çıkmaktadır.

Çizelge 7.5 500 Kişilik gruba minimum destek 3 alındığında çıkan sonuçlar

Birlikte Bulunan Sorular	Durum	Öğrenci sayısı	Cevapları Birlikte Bulunma Oranı %
Babası lise mezunu olup, annesi ilkokul mezunu	Başarılı Başarısız	74 98	43,0 56,9
Annesi lisans mezunu ve ailenin geliri 2000 TL üzeri	Başarılı Başarısız	56 42	57,1 42,8
Ailenin aylı geliri 2000 TL üzeri ve fen lisesi mezunu	Başarılı Başarısız	50 38	56,8 43,1
Genel	Başarılı Başarısız	180 178	50,2 49,7

500 kişilik bir gruba, minimum destek değeri 3 alınarak uygulanan anket sonuçları Şekil 7.6'da gösterilmiştir.



Şekil 7.6 500 kişilik anket grubu için minimum destek değeri 3 olduğunda çıkan grafik

Çizelge 7.5'de, 500 kişi için minimum destek 3 alınarak elde edilmiş sonuçta, 3. ve 4. soruların, birlikte verilmiş 74 cevaplama sayısı ve % 43'lük bir oranla babanın ve annenin eğitim düzeyinin ne kadar etkili olduğu görülmektedir.

Hazırlık sınıfında okuyup, aynı zamanda kendi bölümünden de ders alan ve ilk dönemin sonunda bu derslerin hiçbirinden kalmayan öğrenciler başarılı, diğerleri başarısız kabul edilmiştir. Genel olarak 500 kişi için minimum destek 2 alındığında, minimum destek 3'e göre daha iyi (Denk. 7.1) sonuç ortaya çıktığı görülmektedir. 200 kişilik grup için ise yine minimum destek 2'nin, minimum destek 3'e göre birlikte bulunma oranının daha yüksek olduğu görülmüştür. Ayrıca 200 kişi ile 500 kişinin genel olarak bir karşılaştırması yapıldığı takdirde, 200 kişilik grubun her iki desteğe göre 500 kişilik gruba göre daha yüksek birlikte bulunma oranı verdiği görülmektedir. Örneğin, babası eğitimci ve annesi ev hanımı olan öğrencilerin, 200 kişilik grup için minimum destek 2'de 99 kişinin % 53,8'le başarılı olduğu, minimum destek 3'te ise 60 kişi ve % 53,5 gibi bir başarı oranlarının ortaya çıktığı görülmektedir. 500 kişilik gruba baktığımızda, babası lise mezunu olup, annesi ilkokul mezunu olan öğrencilerin, minimum destek 2'de, 87 kişinin % 47,2 ile başarılı olduğu, minimum destek 3'te ise 74 kişinin % 43'lük bir başarı oranı elde ettikleri görülmektedir.

7.4. Karar Ağacı ile elde edilen sonuçlar

200 kişilik gruba karar ağacı algoritmalarından ID3 algoritması uygulandığında elde edilen sonuçlar Çizelge 7.6'daki gibidir.

Çizelge 7.6 200 Kişilik gruba ID3 algoritması uygulandığında çıkan sonuçlar

Birlikte Bulunan Sorular	Durum	Öğrenci sayısı	Cevapları Birlikte Bulunma Oranı %
Babası eğitimci ve annesi ev hanımı	Başarılı	72	52,5
	Başarısız	65	47,4
Annesi ev hanımı olup babası lisans mezunu	Başarılı	68	56,1
	Başarısız	53	43,8
Babası lise mezunu olup, annesi ilkokul mezunu	Başarılı	47	43,1
	Başarısız	62	56,8
Annesi lisans mezunu ve ailenin geliri 2000 TL üzeri	Başarılı	52	53,6
	Başarısız	45	46,3
Genel	Başarılı	239	51,5
	Başarısız	225	48,4

500 kişilik gruba karar ağacı algoritmalarından ID3 algoritması uygulandığında elde edilen sonuçlar Çizelge 7.7'deki gibidir.

Çizelge 7.7 500 Kişilik gruba ID3 algoritması uygulandığında çıkan sonuçlar

Birlikte Bulunan Sorular	Durum	Öğrenci sayısı	Cevapları Birlikte Bulunma Oranı %
Babası eğitimci ve annesi ev hanımı	Başarılı	89	51,4
	Başarısız	84	48,5
Annesi ev hanımı olup babası lisans mezunu	Başarılı	91	53,5
	Başarısız	79	46,4
Babası lise mezunu olup, annesi ilkokul mezunu	Başarılı	83	45,1
	Başarısız	101	54,8
Annesi lisans mezunu ve ailenin geliri 2000 TL üzeri	Başarılı	91	51,7
	Başarısız	85	48,2
Genel	Başarılı	354	50,3
	Başarısız	349	49,6

Karar ağacı algoritmalarından ID3 algoritması, 200 ve 500 kişi için uygulandığında birlikte bulunan soruların başarılı öğrencilerde, cevapları birlikte bulunma oranlarına göre, genel olarak 200 kişilik grupta daha fazla birlikte bulunma oranının (Denk. 7.1) ortaya çıktığı görülmektedir. Burada iyi bir sonuç çıkmasının anlamı öğrenci başarısı olarak düşünülmemelidir. Başarılı bir sonuçtan kastedilen, genel olarak birlikte bulunma oranlarının yüksek çıkmasıdır. Yani başarılı öğrencilerin birlikte verdiği cevapların sorularının, cevapları birlikte bulunma oranlarına göre yüksek çıkmasıdır. Birlikte bulunan sorulardan, her iki grupta da annesi ev hanımı olup, babası lisans mezunu olan öğrencilerin daha başarılı oldukları, yapılan anket sonuçlarının değerlendirilmesi sonucunda görülmektedir. Bunun yanında babası eğitimci olup annesi ev hanımı olan öğrencilerle, annesi lisans mezunu olup ailenin aylık geliri 2000 TL üzeri olan öğrencilerin de, genel olarak yine başarılı oldukları, yapılan uygulamanın sonucunda ortaya çıkmaktadır.

Genel olarak, her iki algoritmadan çıkan sonuçları karşılaştırıldığında, yine birlikte bulunan soruların, başarılı öğrencilerde, cevapları birlikte bulunma oranlarına göre apriori algoritmasından çıkan sonuçların, karar ağacından çıkan sonuçlara göre daha yüksek olduğu görülmektedir. Bunun yanında algoritmaların çalışma hızına bakılacak olursa, karar ağacı algoritmasının apriori algoritmasına göre daha hızlı çalıştığı görülmüştür. Bunun en önemli sebebi ise, apriori algoritmasının sürekli veritabanı ile bağlantılı çalışmasıdır. Yani algoritmanın genel yapısı itibariyle, bulunan değerlerin tekrar veritabanına bağlanılarak sonraki taramalarda eldeki kaynak veri olarak kullanılmasıdır. Her iki algoritmadan çıkan genel sonuçlar, Çizelge 7.8'de gösterilmiştir.

Çizelge 7.8 Apriori ve Karar Ağacı Algoritmasından çıkan genel sonuçlar

Gruplar	Başarılı	Başarısız	Genel Olarak Cevapları Birlikte Bulunma Oranı %
200 kişilik grup için Minimum Destek = 2	491		54,1
		415	45,8
200 kişilik grup için Minimum Destek = 3	312		50,6
		304	49,3
500 kişilik grup için Minimum Destek = 2	465		53,3
		407	46,6
500 kişilik grup için Minimum Destek = 3	180		50,2
		178	49,7
200 kişilik gruba karar ağacı algoritması uygulandığında	239		51,5
		225	48,4
500 kişilik gruba karar ağacı algoritması uygulandığında	354		50,3
		349	49,6

Bu çalışmada, İnan (2003)'ün yaptığı çalışmadan farklı olarak, hem karar ağacı hem de apriori algoritması kullanılmış ve genel olarak İnan'ın çalışmasında % 48,97 ile başarılı öğrencilerin verdiği cevapların, % 50,97 başarısız öğrencilerin verdiği cevapların birlikte bulunduğu gözlenmiştir. Bu çalışmada ise 200 kişilik grup için % 51,5 ve 500 kişilik grup için % 50,3 ile başarılı öğrencilerin verdiği cevaplar

birlikte bulunur iken, yine 200 kişilik grup için % 48,4 ve 500 kişilik grup için % 49,6 ile başarısız öğrencilerin verdiği cevapların birlikte bulunduğu gözlenmiştir.

İnan (2003)'ün çalışmasında, "Annenizin mesleği nedir?" sorusu ile "Mezun olduğu lise türü nedir?" sorularının, öğrenci başarısına etki eden önemli faktörler olduğu görülürken, bu çalışmada ise "Annenizin mesleği nedir?" sorusu ile "Babanızın eğitim düzeyi nedir?" sorularının öğrenci başarısında daha çok ön plana çıktığı görülmüştür. Ayrıca "Ailenin aylık geliri nedir?" sorusu ile "Mezun olduğu lise türü nedir?" sorularının da anket sonuçlarının değerlendirilmesi sonucu, öğrenci başarısına etki eden faktörler arasında çıktığı görülmüştür.

8. SONUÇLAR VE ÖNERİLER

Gelişmiş ülkelerde birçok kurum ve kuruluş artık yol haritalarını veri madenciliği teknikleri ile belirlemektedir ve sadece bu işle ilgili personel istihdam etmeye başlamışlardır. Hekim ve bilim adamlarının karar vermelerinde destek sistemi olarak veri madenciliği uzman sistemleri yoğun şekilde kullanılmaya başlanmıştır. Tıbbi cihazlar, veri madenciliği yazılım ve donanımları ile üretilir duruma gelmiştir. Veri madenciliğinde kullanılan teknikler, günlük yaşantımızın da kaçınılmaz bir parçası haline gelmiştir. Yapılan uygulamada, Selçuk üniversitesi otomasyonunda tutulan anket verilerine, apriori ve karar ağacı algoritmalarını uygulayıp, sonuçları karşılaştırılmış ve bu sonuçlara göre öğrenci başarısına etki eden faktörlerin birlikte bulunması amaçlanmıştır.

8.1. Sonuçlar

Uygulanan apriori algoritması için, 200 ve 500 kişilik grupta genel olarak birlikte bulunma oranı, ankete katılanların yarısından daha az olduğu görülmektedir. Ankete katılan öğrencilerin başarılı olup olmadıkları önceden bilinmektedir. Anket sonuçları, algoritmalar ile değerlendirildikten sonra çıkan cevapların birlikte bulunma oranları, dolayısıyla soruların birlikte bulunma oranlarının ne kadar yüksek olduğu ifade edilmek istenmektedir. Bir başka deyişle, sorulardan aynı cevapları veren öğrenci sayılarının fazla olması ve dolayısıyla genel oranının yüksek çıkması, birlikte bulunma olarak ifade edilmektedir.

200 kişilik grubun, diğer gruba göre genel birlikte bulunma başarı oranı daha yüksek çıkmıştır. 200 kişilik grubun, minimum destek 2 ve 3‘e göre genel birlikte bulunma oranı birbirine yakın çıkmış, diğer grupta ise minimum destek 2‘nin minimum destek 3’e göre daha iyi olduğu görülmüştür. Karar ağacı algoritmasında ise, 200 kişilik grubun, diğer gruba göre birlikte bulunma oranının daha yüksek çıktığı, fakat genel olarak her iki grubun da genel birlikte bulunma oranının apriori algoritmasında olduğu gibi, yarısından az olduğu görülmektedir.

Birlikte bulunan soruların cevapları birlikte bulunma oranlarına göre, her iki algoritma ile yapılan programlarda çıkan sonuçlarda, genel birlikte bulunma oranının yarısından az çıktığı görülmüştür. Fakat apriori algoritmasından çıkan sonuçların ise karar ağacı algoritmasından çıkan sonuçlara göre birlikte bulunma oranının daha

yüksek çıktığı görülmektedir. Sonuçların her ikisinde de genellikle "annenin mesleği", "babanın ve annenin eğitim düzeyi" ve "mezun olduğu lise nedir" soruları ön plana çıkmaktadır. Burada öğrencilerin üniversite hayatından önceki aile yaşantıları, onların gelecekteki başarılarına ne kadar etki ettiği görülmektedir. Özellikle annelerinin mesleği, çocuklarının gelecekteki başarılarında büyük etmen olduğu anlaşılmaktadır. 200 kişilik grupta minimum destek 2 alınarak çıkan sonuçlara bakıldığında, annesi ev hanımı olup, babası üniversite mezunu olan 98 öğrencinin, % 57,3'lük gibi bir birlikte bulunma oranı ortaya çıkmıştır. Yine bu grupta minimum destek 3 alındığında çıkan sonuçlara bakıldığında, annesi ev hanımı olup, babasının ise üniversite mezunu olduğu 93 öğrencinin % 55,3'lük gibi bir birlikte bulunma oranı elde ettiği görülmektedir. 500 kişilik gruba bakıldığında, minimum destek 2'de, annesi ev hanımı olup babası üniversite mezunu olan 95 öğrencinin % 55,8'lik bir birlikte bulunma oranı elde etmiştir. Yine aynı grupta minimum destek 3 alındığında annesi lisans mezunu ve ailenin geliri 2000 TL üzeri olan 56 öğrencinin, % 57,1'lik bir birliktelik elde ettiği yapılan anket sonuçlarının değerlendirilmesinden görülmüştür. 200 kişilik gruba karar ağacı algoritmalarından ID3 uygulandığında, yine annesi ev hanımı olup babası lisans mezunu 68 öğrencinin % 56,1'lik bir birliktelik ile grubunda en yüksek birlikteliği elde ettiği, bu algoritma 500 kişilik gruba uygulandığında 200 kişilik grupta olduğu gibi annesi ev hanımı olup babası lisans mezunu 91 öğrencinin % 53,5'lik bir birliktelik ile bu grupta en yüksek birliktelik oranını elde ettiği görülmüştür.

Sonuç olarak, bu çalışmada hazırlık sınıfında okuyan öğrencilerin başarılarını etkileyen faktörler, veri madenciliği algoritmalarından apriori ve karar ağacı algoritmaları ile bulunmaya çalışılmıştır. Hazırlık sınıfında okuyan öğrencilere uygulanan anket çalışmasının soruları ve sonuçları bu amaçla kullanılmıştır. Bu anketin soruları ve verilen cevaplar sayısal değerlere çevrilerek, Matlab programında, apriori ve karar ağacı yöntemleri uygulanmış ve verilen cevapların yoğunluğuna göre, hangi faktörlerin öğrencilerin öğrenim hayatı boyunca, onları olumlu veya olumsuz yönde etkilediği araştırılmıştır.

Burada karşılaşılan başlıca sorunlardan bir tanesi, apriori algoritmasının bu tür uygulamalarda yavaş çalışmasıdır. Çünkü bu algoritma sürekli veri tabanı ile çalıştığı için yani algoritmanın çalışma esnasında yapılan sorgulamalarda sürekli veri tabanına bağlanması gerektiği ve bu veri tabanını her defasında sorgulaması gerektiği için karar ağacına göre daha yavaş çalıştığı görülmüştür. Karşılaşılan bir başka sorun ise, öğrencilerin internet üzerinden yapılan anketlere özen göstermeden verdiği cevaplar

olarak karşımıza çıkmaktadır. Çünkü uygulama sonuçlarında, ankette bulunan sorulardan 7. sorunun, öğrenci başarısına etki eden faktörler arasında bulunmadığı görülmüştür. Oysaki öğrencinin yaşadığı yer ve ortamın, öğrenci için çok önemli olduğu bilinen bir gerçektir. Sonucun bu şekilde çıkmasının nedeni ise web üzerinden yapılan anketlerin, güvenilirliğinin yüksek olmamasıdır.

8.2. Öneriler

İleride veri madenciliği için, hızlı algoritmalar, akıllı sistemler, özel güvenlik mekanizmaları ve yapay sinir ağlarının bu konu ile ilgili yeni uygulamalar geliştirmesi gerekecektir. Çünkü dünyada her geçen gün bilgi katarları daha da karmaşık ve büyük boyutlara ulaşmaya başlamıştır. Gerek yazılım gerekse donanım olsun, eldeki sistemlerin ileride bu tür uygulamalar için yetersiz kalacağı açık olarak görülmüştür. Hali hazırda kullanılan yazılımların bile bugünkü veri ambarlarında tutulan verileri işleyerek, anlaşılabilir ve karar verilebilir hale getirmekte oldukça zorlandıkları görülmüştür.

Bu çalışmada, öğrencilerin başarılarını etkileyen faktörler arasında ailenin eğitim düzeyi, anne ve babanın mesleği, mezun olduğu lise ve ailenin gelir düzeyinin başta geldiği görülmüştür. Web üzerinden yapılan bu anketin değerlendirilmesi sonucunda, yedinci sorunun öğrenci başarısına etki eden faktörler arasında çıkmaması, internetten yapılan anketlerin güvenilirliğinin sorgulanmasını akla getirmiştir. Dolayısıyla öğrenci başarısına etki eden faktörlerin bulunabilmesi için daha iyi şartlarda uygulanmış anketler kullanılabilir. Geliri az olan başarısız öğrenciler için, burs verilerek başarılı olmalarına katkı sağlanabilir.

KAYNAKLAR

Agrawal, R., Imielinski, T. ve Swami, A., 1993, Mining association rules between sets of items in large databases. ACM SIGMOD Conference on Management of Data.

Agrawal, R. ve Srikant, R., 1994, Fast Algorithms for Mining Association Rules. 20th International Conference on Very Large Databases, VLDB'94 Bildiri Kitabı, Santiago, Chile.

Agrawal R. ve Srikant R., 1995, "Mining Sequential Patterns", 11th International Conference on Data Engineering, Taipei, Taiwan.

Akbulut, S., 2006, Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu, Yüksek Lisans Tezi, *Gazi Üniversitesi Fen Bilimleri Enstitüsü,* Ankara

Akpınar H., 1997, Enformasyon Teknolojisi ve İşletmecilik Öğretimine Etkileri, http://www.isletme.istanbul.edu.tr/akpinar/content/Enformasyon%20Teknolojileri.pdf, İstanbul Üniversitesi İşletme Fakültesi, s.:1-45.

Akpınar H., 2000, Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İ.Ü. İşletme Fakültesi Dergisi, Cilt: 29, S:1, s. 1-22.

Ali, K., Manganaris, S., ve Srikant, R., 1997, Partial Classification using Association Rules. Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California.

Alpaydın, E., 2000, Zeki veri madenciliği: ham veriden altın bilgiye ulaşma yöntemleri, Bilişim2000eğitimsemineri,www.deu.edu.tr/userweb/k.yaralioglu/dosyalar/ver_mad.doC, [Erişim tarihi: 29.11.2010].

Arabacı, G., 2007, Veri Madenciliğinde Apriori, Tahminci Apriori ve Tertius Algoritmalarının Weka ve Yale Programları ile Karşılaştırılması ve Bir Uygulama, Yüksek Lisans Tezi, *İstanbul Ticaret Üniversitesi Sosyal Bilimler Enstitüsü*, İstanbul

Chan, K.C.C. ve Wong, A.K.C., 1991, A statistical technique for extracting classificatory knowledge from databases, Knowledge Discovery in Databases, Cambridge, 107-123.

Ding, Q., and Perrizo, W., 2001, Association Rule Mining on Remotely NDSU-CSOR-TR-01-1, North Dakota State University, Fargo.

Dolgun, M.,Ö., 2006, Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları, Yüksek Lisans Tezi, *Hacettepe Üniversitesi Fen Bilimleri Enstitüsü,* Ankara

Döşlü, A., 2008, Veri Madenciliğinde Market Sepet Analizi ve Birliktelik Kurallarının Belirlenmesi, Yüksek Lisans Tezi, *Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü,* İstanbul

Ergun, E., 2008, Ürün Kategorileri Arasındaki Satış İşilkisinin Birliktelik Kuralları Ve Kümeleme Analizi İle Belirlenmesi Ve Perakende Sektöründe Bir Uygulama, Doktora Tezi, *Afyon Kocatepe Üniversitesi,* Afyon

Ezerçe, A., 2008, Müşteri İlişkileri Yöntemi (Crm) Ve Veri Madenciliği (Data Mining) ,Tekstil Sektöründe Bir Uygulama, Yüksek Lisans Tezi, *İstanbul Üniversitesi Sosyal Bilimler Enstitüsü,* İstanbul

Fayyad, U., M., iatetsky-Shapiro, G. ve Uthurusamy, R. (1996). Advances in Knowdedge Discovery and Data Mining. Cambridge, MA: MIT Press.

Frawley, W.; Piatetsky-Shapiro, G.; and Matheus, C., 1991, Knowledge discovery databases: An overview. G. Piatetsky-Shapiro ve W. J. Frawley (eds.), Knowledge Discovery in Databases, Cambridge, MA: AAAI/MIT Press, 1-27.

Gürgen, G., 2008, Birliktelik Kralları ile Sepet Analizi ve Uygulaması,Yüksek Lisans Tezi, *Marmara Üniversitesi Sosyal Bilimler Enstitüsü,* İstanbul

Han, J., Cai, Y. ve Cercone, N., 1992, Knowledge discovery in databases, An attribute-oriented approach. Bildiri Kitabı, 18. VLDB Konferansı, Vancouver, British Columbia, Canada, s. 547-559.

İnan, O., 2003, Öğrenci İşleri Veritabanı Üzerinde Veri Madenciliği Uygulamaları, Yüksek Lisans Tezi, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*, Konya

Kalıkov, A., 2006, Veri Madenciliği ve Bir E-Ticaret Uygulaması, Yüksek Lisans Tezi, *Gazi Üniversitesi Fen Bilimleri Enstitüsü,* Ankara

Kira, K., ve Rendell, L., 1992, A Practical Approach to Feature Selection. In Proceedings of The Tenth International Conference on Machine Learning, Amherst, Massachusetts, *Morgan Kaufmann*.

Mitchell, T.M., 1997, Machine Learning, C.L., Liu, Allen, B. T, McGraw-Hill Companies, Inc., Carnegie Mellon University, Pensilvanya, s. 52-53

Onat, A., 2008, Veri Madenciliğinin Web Tabanlı Uygulamalarda İnsan Uyumluluklarının Tespiti Üzerine Bir Çalışma, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü,* Konya

Özçakır, F. C., 2006, Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı Ve Uygulaması, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 6(12),21-37

Quinlan, J., R., 1986a, Induction of decision trees. Machine Learning. *Morgan Kaufmann,* Netherlands, cilt.1, 81-106..

Raghavan, V.,V., ve Sever, H., 1994b, The State of Rough Sets For Database Mining Applications, T.Y. Lin editör, 23rd Computer Science Conference Workshop on Rough Sets and Database Mining, San Jose, California, 1-11.

Saraçoğlu, R., Tutuncu, K., & Allahverdi, N., 2007, A fuzzy clustering approach for finding similar documents using a novel similarity measure. Expert Systems with Applications, 33(3), 600–605.

Sever, H. Ve Oğuz B., 2002, Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım, *Bilgi Dünyası*, 3(2), 173-204.

Sıramkaya, E., 2005, Veri Madenciliğinde Bulanık Mantık Uygulaması, Yüksek Lisans Tezi, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü,* Konya

Silahtaroğlu G., 2008, Kavram ve Algoritmalarıyla Temel Veri Madenciliği, *Papatya Yayıncılık*. İstanbul

Şen, F., 2008, Veri madenciliği ile birliktelik kurallarının bulunması, Yüksek Lisans Tezi, *Sakarya Üniversitesi Fen Bilimleri Enstitüsü,* Sakarya

Tiryaki, S., 2006, Lojistik Alanın Bir Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, *İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü*, İstanbul

Tosun., T., 2006, Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi, Yüksek Lisans Tezi, *İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü,* İstanbul

Vahaplar, A., İnceoğlu, M.M., 2001, Veri madenciliği ve elektronik ticaret, http://www.bayar.edu.tr/bid/dokumanlar/inceoglu.doc., [Erişim Tarihi : 01.06.2011].

Üçgün, K., 2009, Orta öğretim okuları için öğrenci otomasyonu tasarımı ve öğrenci verileri üzerine veri madenciliği uygulamaları, Yüksek Lisans Tezi, *Marmara Üniversitesi Fen BilimleriEnstitüsü,* İstanbul

Weiss, S., ve Indurkhya, N., 1998, Predictive data mining: A practical guide. *Morgan Kaufmann*, DMSK Software: www.data-miner.com., [Erişim Tarihi : 13.02.2011].

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Ufuk EKİM
Uyruğu : T.C.
Doğum Yeri ve Tarihi : Akşehir – 1977
Telefon : 223 25 43
Faks : 241 00 64
e-mail : ufuk@selcuk.edu.tr

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	Akşehir Lisesi, Akşehir, Konya	1994
Üniversite	:	Selçuk Üniversitesi,Selçuklu,Konya	2000
Yüksek Lisans	:		
Doktora	:		

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2003	Selçuk Üniversitesi	Uzman

UZMANLIK ALANI : Veri Tabanı, yazılım

YABANCI DİLLER : İngilizce