T.C.
İSTANBUL TİCARET ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME YÜKSEK LİSANS PROGRAMI

VERİ MADENCİLİĞİNDE APPRIORI, TAHMİNCİ APPRIORI VE TERTIUS ALGORİTMALARININ WEKA VE YALE PROGRAMLARI İLE KARŞILAŞTIRILMASI VE BİR UYGULAMA

Yüksek Lisans Tezi

Gültekin ARABACI

0650Y38223

İSTANBUL, 2007

T.C.
İSTANBUL TİCARET ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME YÜKSEK LİSANS PROGRAMI

VERİ MADENCİLİĞİNDE APPRIORI, TAHMİNCİ APPRIORI VE TERTIUS ALGORİTMALARININ WEKA VE YALE PROGRAMLARI İLE KARŞILAŞTIRILMASI VE BİR UYGULAMA

Yüksek Lisans Tezi

Gültekin ARABACI

0650Y38223

Danışman: Yrd. Doç. Dr. Dicle CENGİZ

İSTANBUL, 2007

T.C.

İSTANBUL TİCARET ÜNİVERSİTESİ

SOSYAL BİLİMLERİ ENSTİTÜSÜ

ONAY SAYFASI

Yüksek Lisans Öğrencisi Gültekin ARABACI'nın "Veri Madenciliğinde Appriori, Tahminci Appriori ve Tertius Algoritmalarının Weka ve Yale Programları İle Karşılaştırılması ve Bir Uygulama" konunulu tez çalışması jurimiz tarafından İşletme Yüksek Lisans tezi olarak oybirliği/oyçokluğu ile başarılı bulunmuştur.

		İmza
Tez Danışman	: Yrd.Doç.Dr. Dicle Cengiz	.....
Jüri Üyesi	: Yrd.Doç.Dr. Ünal H. Özden	.....
Jüri Üyesi	: Prof.Dr. Münevver Turanlı	.....

ONAYLI

Yukarıdaki juri kararı Enstitü Yönetim Kurulunun ...../...../2007 tarih ve ..... kararı ile onaylanmıştır.

Prof. Dr. Kerem ALKİN
Müdür

İstanbul Ticaret Üniversitesi
Sosyal Bilimleri Enstitüsü Müdürlüğüne
Eminönü Yerleşkesi- İstanbul

Halen Enstitünüzün İşletme Yönetimi yüksek lisans programı öğrencisiyim. Hazırlamakta olduğum tez tümüyle özgün bir çalışma olup YÖK ve İTİCU Lisansüstü Yönetmeliklerine uygun olarak hazırlanmıştır. Ayrıca bu çalışmayı yaparken bilimsel etik ve kurallarına tamamıyla uyduğumu; yararlandığım tüm kaynakları gösterdiğimi ve hiçbir kaynaktan yaptığım ayrıntılı alıntı olmadığını beyan ederim.

ÖZET

Günümüzde veri, birçok depolama aracında artan bir hızla birikmektedir. Biriken veri, içinde değerli bilgiler barındırmaktadır. Bilgiye ulaşmada kullanılacak verinin büyüklüğü veri analiz tekniklerinin kullanılmasını zorunlu kılmaktadır. Büyük veri tabanlarından yararlı bilgiler elde etmek olarak tanımlabilecek veri madenciliği, klasik istatistiksel metodlarla açık ve net olarak gösterilemeyen bağlantıları ortaya çıkarabilecek teknikleri kullanarak elde edilemeyen örüntü ve eğimleri keşfetme işlemidir. Veri madenciliği teknikleri genel olarak pazarlama, bankacılık, sigortacılık, telekomünikasyon, endüstri, tıp ve mühendislik alanlarında kullanılmaktadır.

İnfeksiyon hastalıklarının tanısında mikrobiyoloji önemli bir yer tutmaktadır. Çalışmamızda hastanedeki servislerden gönderilen idrar, kan vb. örneklerin incelenmesi sonucu elde edilen gram negatif basillere ait veri kullanılacaktır.

Çalışmamızın konusunu, örnek veri tabanında bilgi keşfi sürecinin yapılması ve modelin değerlendirilmesinde kullanılan iki ayrı programın karşılaştırılması oluşturmaktadır.

ABSTRACT

Nowadays data is being collected in many kind of storage devices at an incremental amount. Collected data has valuable information inside. Data analysis techniques have to be used in order to obtain information from great amount of data. Defined as obtaining useful information from large data sets, data mining is the process of obtaining correlation and patterns that can not be done by classic statistical methods. Data mining techniques are generally used in marketing, banking, insurance, telecommunication, industry, healthcare and engineering. Our study is related with application of data mining in healtcare.

Microbiology has an important role in the diagnosis of infectious diseases. In our study, data obtained from a sample gram negative basil database, consisting of results derived by examining urine, blood etc.samples of gram negative basilli will be used.

The subject of our study is to generate necessary steps in data mining process and to compare two softwares used in evaluation of the model.

İÇİNDEKİLER

Sayfa No

TABLOLAR LİSTESİ

Sayfa No

ŞEKİLLER LİSTESİ

Sayfa No

KISALTMALAR LİSTESİ

a.g.e. **:** Adı Geçen Eser

GNB : Gram Negatif Basil

s. **:** Sayfa

S. **:** Sayı

VTBK **:** Veri Tabanında Bilgi Keşfi

WEKA : Waikato Environment for Knowledge Analysis

YALE : Yet Another Learning Environment (Yale) Programı

YSA **:** Yapay sinir ağları

GİRİŞ

Günümüzde veri akıl almaz bir hızda şirketlerin, okulların, kurumların ve bireysel kullanıcıların veri depolama birimlerinde birikmektedir. Veri bazen işlenmeye hazır bir şekilde bulunurken, bazen de işlenmeye hazır hale getirilebilmesi için bir hayli uğraş verilmesi gereken bir şekilde depolanmaktadır.

Verinin bilgi haline getirilmesi pamuk, ipek, naylon vb. maddelerden elbise üretilmesi sürecine benzemektedir. Veriden bilginin üretilebilmesi için verinin bazı safhalardan geçmesi gerekir. Pamuklu elbise üretiminde pamuğun toplanması, biriktirilmesi, üzerindeki yabancı maddelerden ayıklanması, iplik haline getirilmesi, boyanması, kumaş haline getirilmesi ve kumaştan elbise dikilmesi şeklinde birbirini izleyen bir süreç bulunmaktadır. Veride benzer bir süreçten geçirilerek faydalı, kullanılabilir bilgiye dönüştürülür. Nüfusun her geçen gün arttığı günümüz dünyasında artık el tezgahlarında pamuğun işlenerek kumaş haline getirilmesi ile giyim ihtiyaçlarının karşılanması mümkün değildir. Aynı şekilde büyük ölçüdeki verinin klasik istatistik yöntemleri ile yorumlanarak, bağıntı ve kuralların ortaya çıkarılabilmesi mümkün değildir. Büyük miktardaki verinin işlenmesi için veri madenciliği programlarına, yüksek bellek ve depolama ünitelerine, toplanmış veriye ihtiyaç bulunmaktadır. Veri madenciliği için gerekli olan tüm koşullar son 10 yılda oluşmuştur.

Bilgisayar işlemcilerinin hızlanması, hafızaların büyümesi, verinin sayısal ortamda saklanması ile birlikte verinin bir bölümünden çıkarım yapmak yerine verinin tümünün işlenmesi ve örüntülerin, ilişkilerin bu şekilde bulunması fikri ortaya çıkmıştır. Verinin bütününün işlenmesi için veri madenciliği araçları geliştirilmiş ve birçok alanda başarılı ile kullanılmaya başlanmıştır.

1 GENEL BİLGİLER

1.1 Veri

Ham, yapısal olmayan, kendi başına değersiz, belli bir forma sokulmamış girdilerdir. Veri rakam, harf olabileceği gibi, daha değişik formatlarda da toplanmış olabilir.

Veri nesnelerin ve bu nesnelere ait niteliklerin toplanması olarak tanımlanabilir. Nitelik nesneye ait özellik veya karakteristiktir. Nitelik değerleri tanımlı (Göz rengi, plaka numaraları), sıralı (notlar, boy ölçüleri-kısa, orta, uzun), aralık belirten (tarih bilgisi, Fahrenhayt cinsinden ısı değeri), oransal (uzunluk, zaman) olarak gruplanabilir.[1]

1.2 Bilgi

Verinin belirli bir amaç için çeşitli aşamalardan geçirilerek işlenmesi ve kullanıma hazır hale getirilmesi süreci sonucu ortaya çıkan çıktıya bilgi adı verilmektedir. Bilgi anlamlı ve kullanılabilir olmalıdır.

Bilgi verinin, veriyi kullanacak olan kişiye bilgi ekleyecek şekilde düzenlenmesi ve işlenmesi olarak da tanımlanabilir.

1.3 Veri Ambarı

Veri ambarı yapısı ve araçları yöneticilerin verilerini sistematik olarak düzenlemelerini, anlamalarını ve stratejik kararlar vermede verilerini kullanmalarını sağlar. İşletme ve kurumların bir çoğu bugünün rekabetçi, çabuk değişim gösteren dünyasında veri ambarı araçlarının değerini anlamışlardır. Son 10 yılda birçok firma dünya çapında veri ambarları oluşturmak için milyonlarca dolar harcamışlardır. Birçok insan rekabetçi ortamda veri ambarının müşteri ihtiyaçlarını saptamada kullanılabilecek geçerli bir silah olarak görmüşlerdir. Veri ambarı, işletmenin aktif olarak kullanılan veri tabanından ayrı olarak tutulan bir veri tabanıdır. Veri ambarı sistemi birçok

[1] Pang-Ning Tan, Michael Steinbach ve Vipin Kumar, Introduction to Data Mining, Boston, Addison Wesley, 2005, s.2-30

uygulamanın bütünleştirilmesi ile geçmişe dair birleşik ve tutarlı bir bilgi işlem ortamı sağlamaktadır.[2]

Veri ambarı karar destek sistemleri, kaynak olarak birden fazla işlevsel veri tabanı kullanan bir veri tabanı sistemidir. Perakendeci için, veri tabanı pazar sepetinden tedarikçi, müşteri vb.ait verileri kapsamalıdır. Karar verme sürecinde çok önemli değilse, maaş bordrosu veri ambarı içinde yer almamalıdır. Veri ambarı birçok farklı veri tabanının bir diske kopyalanması ile elde edilmez. Bu süreçte veri tabanları arasındaki alanlardaki tutarsızlıkları gidermek için birkaç bütünleştirme işleminin yapılması gerekebilir. Veri ambarı oluşturulması, uzman personel müdahalesi gerektirdiği ve sürecin ayrıntılı olarak çözümlenmesine ihtiyaç duyulduğu için masraflı bir iştir. [3]

Veri ambarı, "bir işletmenin ya da kurumun çeşitli birimleri tarafından canlı sistemler aracılığı ile toplanan bilgilerin, ileride değerlendirmeye alınabilecek olanlarının arka planda yer alan bir sistemde birleştirilmesinden oluşan büyük ölçekli bir veri deposu" olarak ta tanımlanabilir.[4]

Veri ambarı, karar destek sistemi olarak görülebilir. Günlük işlemlerin gerçekleştiği canlı sistemlerin arka planında bulunan veri ambarına, canlı sistemlerde oluşan veriler belli zaman aralıkları ile aktarılırlar. Zaman aralıklarının ne kadar olacağının seçimi tamamen veri ambarını kullanan işletmenin ihtiyaçları doğrultusunda belirlenir. Veri ambarı çevrimdışı olarak çalışmaktadır. Bu yüzden veri ambarı içerisindeki kayıtlar güncel olmayabilir. Bir işletmenin içerisindeki farklı bölümler ve farklı günlük işlemleri gerçeklemek üzere tasarlanmış birbirlerinden habersiz ve bağımsız çalışan değişik canlı sistemlerdeki veriler, veri ambarındaki bir tek yapı içerisinden erişilebilecek bir şekilde toplanıp veri ambarına aktarılabilir. Bu aktarım süreci içerisinde veriler üzerinde veri ambarında önceden bulunan kayıtları aktarmama, kayıtlar içerisindeki bazı bilgileri değiştirmek, silmek ve benzeri bir takım işlemler de gerçekleştirilebilir. Bu ön işlem

2 Jiawei Han ve Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000, s.44

3 David Hand, Heikki Mannila ve Padharic Smyth, Principles of Data Mining, MIT Press, Londra, 2001, s.417

4 Ahmet Cüneyd Tantuğ, Veri Madenciliği ve Demetleme, İstanbul Teknik Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul, 2002, s.1

sonrasında değişik kaynaklardan yani değişik canlı sistemlerden toplanan veriler, anlamsal bütünlüğü sağlayacak şekilde veri ambarına yerleştirilirler. Veri ambarına aktarılan veriler, işletmeye ait canlı sistemlerin yanısıra işletmeye ait olmayan kaynaklardan da elde edilebilir. Veri ambarında yer alan bilgiler, bu bilgilerin kullanılacağı alanlara göre ayrı alt depolara dağıtılabilirler. Çoğunlukla işletme içindeki departmanların kullanımına göre bölmelenen veri ambarlarında alt depolar (data-mart) oluşturulur. [5]

İşlevsel sistemlerin yarattığı ham verileri daha sonra üzerinde işlem yapacak şekilde değişik ve güvenli bir ortamda saklamak, geçmişe dönük olarak verileri biriktirmek ve bu sayede karar destek sistemlerine girdi temin etmek, işlevsel ve analitik ihtiyaçlar doğrultusunda farklı veri yapıları oluşturmak veri ambarı kullanım amaçları arasında sayılabilir.

1.4 Veri Madenciliği Tanımı

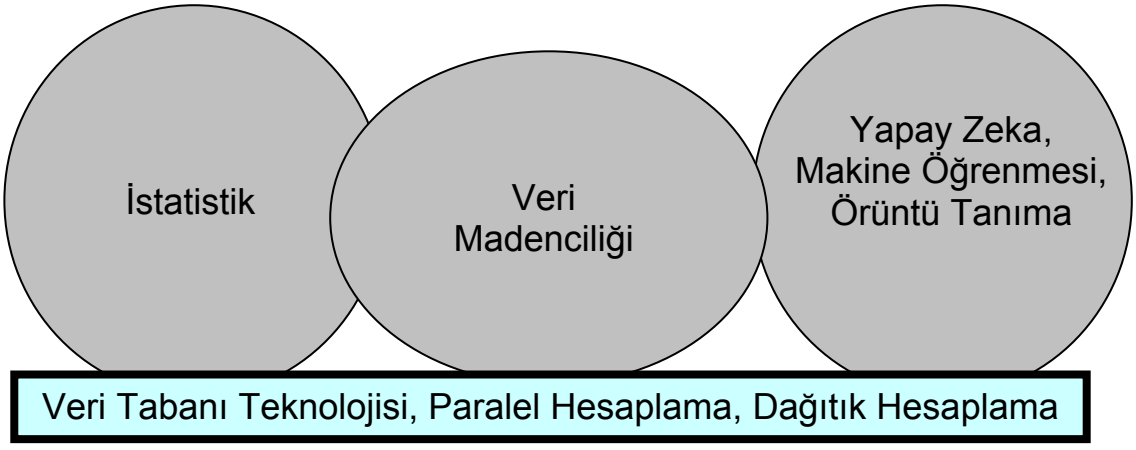
Büyük veri tabanlarından yararlı bilgiler elde etmek olarak tanımlabilecek veri madenciliği, açık ve net olarak gösterilemeyen bağlantıları ortaya çıkarabilecek teknikleri kullanarak elde edilemeyen örüntü ve eğilimleri keşfetme işlemidir.

Veri madenciliği büyük veri tabanlarından, veri ambarlarından ve diğer büyük veri havuzlarından otomatik veya uygun yöntemlerle bilgiyi temsil eden örüntülerin elde edilmesidir. Veri madenciliği, veri tabanı teknololojisi, yapay zeka, makina öğrenmesi, yapay sinir ağları, istatistik, örüntü tanıma, bilgi edinme, yüksek başarım hesaplamaları, veri görüntüleme gibi çok farklı disiplinlerin kullanıldığı bir alandır. [6]

Veri madenciliğinin sadece işletmelerin pazar ve kâr payını nasıl arttıracakları ile ilgili kullanılmasına ilişkin bir hayli kaynak bunmaktadır. Bunlara örnek olarak "Data Mining Cookbook (Veri Madenciliği Yemek Kitabı)",[7] "Data Mining Techniques For

[5] a.g.e., s.2-3
[6] Jiawei Han ve Micheline Kamber, a.g.e., s.1
[7] Olivia Parrud, Data Mining Cookbook, John Wiley & Sons, New York, 2001.

Marketing, Sales and Customer Relationship Management (Pazarlama, Satış ve Müşteri ilişkileri için Veri Madenciliği Teknikleri)"[8] verilebilir.



Şekil 1. Farklı Disiplinlerden Oluşan Veri Madenciliği

Kaynak: Jiawei Han ve Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000, s.15

Veri madenciliği veri tabanlarındaki mevcut verinin analiz edilerek problemlerin çözülmesi olarak tanımlanabilir. Kararsız müşterilerin ikna edilmesi sürecinde, geçmiş dönemde ürüne sadık kalan ve ürün değiştiren müşterilerin bilgileri ve örüntüleri araştırılarak başka bir ürün kullanmaya meyilli müşteriler tespit edilir. Bu gruptaki müşterilere özel ilgi gösterilerek müşterilerin aynı üründen kullanmaya devam etmesi sağlanabilir. Bugünün rekabetçi, müşteri merkezli, hizmete yönelik dünyasında veri, işletmelerin büyümelerini sağlayan hammaddedir. Veri madenciliği verideki faydalı ve üstünlük (genelikle ekonomik) sağlayacak örüntülerin, otomatik veya yarı otomatik yöntemlerle ortaya çıkarılmasıdır. Veri madenciliği ile makine öğrenmesi arasındaki

[8] Michael J.A. Berry ve Gordon S. Linoff, Data Mining Techniques For Marketing, Sales and Customer Relationship Management, Wiley Publishing, Indiana, 2004

fark, öğrenme sürecinde gelecekte gerçekleştirilecek uygulamaların daha iyi icra edilmesini sağlayacak şekilde davranışsal değişiklik yaratılmasıdır.[9]

Son 20 yılda veri tabanı sistemleri alanında büyük başarılar elde edilmiştir. Her geçen gün daha fazla veri toplanmakta ve biriktirilmektedir. Bu veri tabanlarında faydalı bilgi bulabilmek, birçok işletmenin üzerinde önemle durduğu bir konu olarak ortaya çıkmış ve veri madenciliği faydalı bilgiye ulaşmada anahtar rol oynayan bileşen olarak gün geçtikçe artan bir oranda önem kazanmıştır. Veri madenciliği algoritmaları ve veri madenciliği araçları, verideki önemli örüntülerin bulunmasında ve faydalı tahminlerin yapılmasında kullanılmaktadır. Bu teknoloji sanal olarak bankacılık, telekomünikasyon, üretim, pazarlama ve elektronik bankacılık alanlarına uygulanmaktadır.

Veri madenciliğinin diğer bir tanımı da, çoğunlukla büyük, gözlemsel veri tabanlarında, var olduğu bilinmeyen ilişkilerin ortaya çıkarılması ve verinin alışılmışın dışındaki şekilde özetlenerek, sahibine anlaşılabilir ve yararlı olacak şekilde sunulmasıdır.[10]

1.5 Veri Tabanında Bilgi Keşfi.(VTBK)

Veri tabanı sistemlerinin artan kullanımı ve verilerin katlanarak artması, organizasyonları elde toplanan bu verilerden nasıl faydalanılabileceği problemi ile karşı karşıya bırakmıştır. Geleneksel sorgu veya raporlama araçlarının veri yığınları karşısında yetersiz kalması, Veri Tabanlarında Bilgi Keşfi-VTBK (Knowledge Discovery in Databases) adı altında, sürekli ve yeni arayışlara neden olmaktadır. VTBK süreci içerisinde, modelin kurulması ve değerlendirilmesi aşamalarından meydana gelen Veri Madenciliği en önemli kesimi oluşturmaktadır. Bu önem, bir çok araştırmacı tarafından VTBK ile veri madenciliği terimlerinin eş anlamlı olarak da kullanılmasına neden olmaktadır.[11]

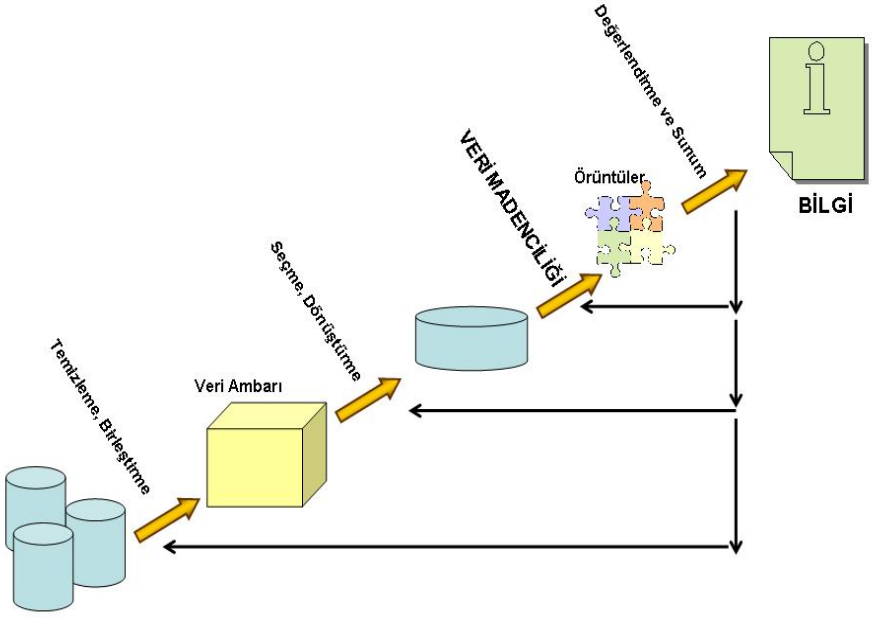
---

[9] Ian H. Witten ve Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques, Elsevier, Boston, 2005, s.7-9

[10] David Hand, Heikki Mannila ve Padharic Smyth, a.g.e.,s.1

11 Haldun Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İstanbul Üniversitesi İşletme Fakültesi Dergisi, C.29, S.1 (Nisan 2000), s: 1-22

VTBK ile veri madenciliği 1990 yıllarında aynı anlamda kullanılmasına rağmen 2000'li yıllara gelindiğinde VTBK sürecinin en önemli halkası olan veri madenciliği VTBK sürecinin tümü için kullanılmaktadır.[12]



Şekil-2 Veri Tabanında Bilgi Keşfi Süreci

Kaynak: Jiawei Han ve Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000, s.7

İzleyen bölümlerde VTBK süreci aşamaları olan problemin tanımlanması, verinin toplanması, verinin temizlenmesi ve dönüştürülmesi, modelin kurulması, modelin değerlendirilmesi, raporlama, tahmin, kullanılan uygulamalarla entegrasyon, modelin yönetilmesi ayrıntılı olarak sunulacaktır.

[12] Erkan Kıyak, CRISP-DM Yöntembilim Kullanılarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması, Kocaeli Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Kocaeli, 2006, s.10

1.6 VTBK Proje Aşamaları

ZhaoHui Tang ve Jamie MacLennan tarafından, VTBK proje aşamaları sırasıyla verinin toplanması, verinin temizlenmesi ve dönüştürülmesi, modelin kurulması, modelin değerlendirilmesi, raporlama, tahmin, kullanılan uygulamalarla entegrasyon, modelin yönetilmesi olarak belirlenmiştir. [13] Fakat problemin tanımlanması aşamasının sürecin daha iyi hale gelmesi için gereklidir.

1.6.1 Problemin Tanımlanması

Her uygulamada olduğu gibi veri madenciliğinde istediğimiz sonuca ulaşabilmek için, projenin ne amaçla yapılacağının açık bir şekilde tanımlanmalıdır. Amaç, açık bir dille ifade edilmiş olmalıdır. Elde edilecek sonuçların nasıl yorumlanacağına ilişkin çalışmalar yapılmalıdır.

Bu aşamada sonucu etkileyebilecek faktörlerin ortaya konması gerekir. Bu aşamaya gerekli önem verilmezse yanlış sorulara doğru cevaplar bulunabilir. Problemin açık bir şekilde tanımlanması, ilerideki aşamalarda kullanılacak yöntemin saptanmasında kolaylık sağlayacaktır.[14]

Kullanılacak alan ile ilgili bilgiler uzman kişilerle istatistikçi veya veri analistlerinin birlikte çalışmalaları, bilginin keşfedilmesinde önemlidir. Elde edilen bilginin yeni ve faydalı olmasına ait bilgi ancak bu konuda uzman kişilerden elde edilebilir.

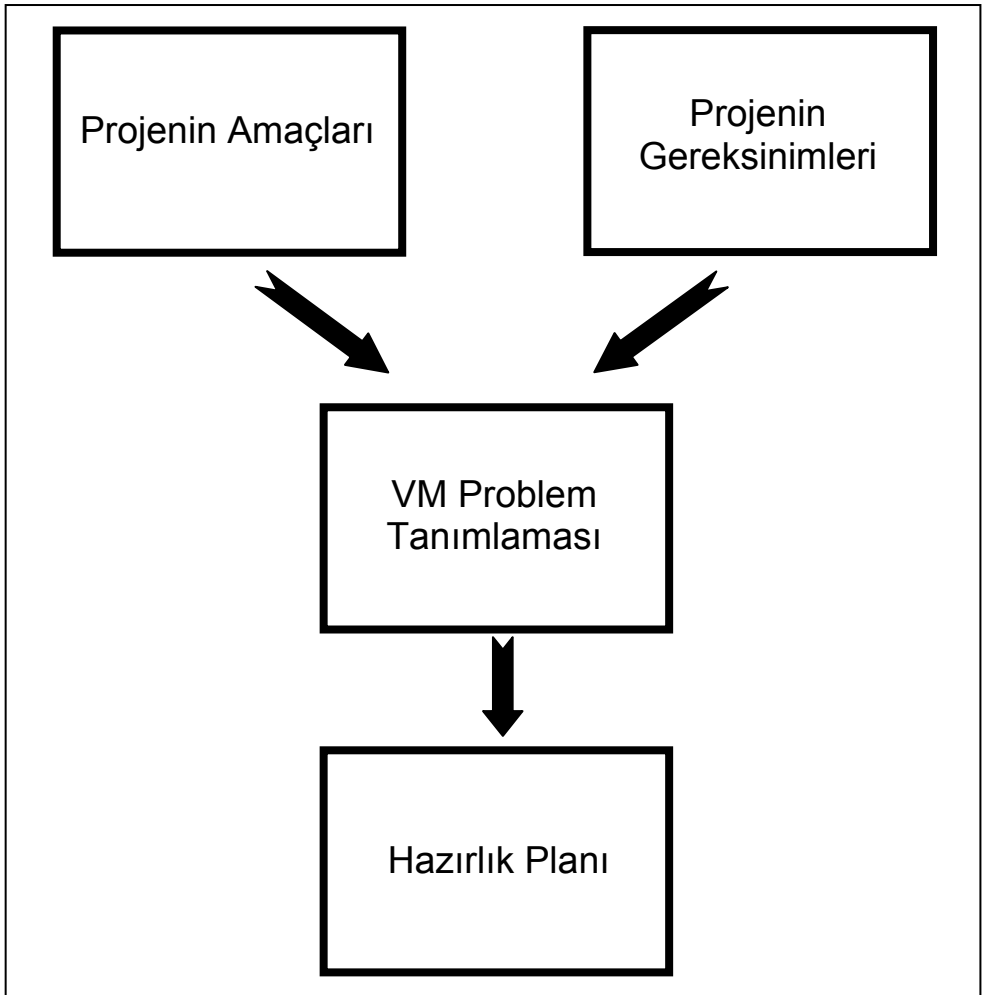
Veri analisti, işletme/kurumda üretilen sayılan verilerin büyüklüğü, proje için yeterlilik derecesi ve verinin kimler tarafından ne şekilde üretildiği konularını çok iyi incelemelidir.

Problemin tanımlanması aşamasına ait süreç Şekil 3'te sunulmuştur.

---

[13] ZhaoHui Tang ve Jamie MacLennan, Data Mining with SQL Server 2005, Wiley Publishing, Indianapolis, 2005, s.13-17

[14] Erkan Kıyak, **a.g.e.**,s.42



Şekil 3 Problemin Tanımlanması

Kaynak: Erkan Kıyak, CRISP-DM Yöntembilim Kullanılarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması, Kocaeli Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Kocaeli, 2006, s.43

1.6.2 Verinin Toplanması

Veriler bir kuruluşta birçok sistemde depolanabilirler. Örneğin Microsoft'ta 70'in üzerinde veri ambarı bulunmaktadır. İlk aşama ilgili verinin veri analizinin uygulanacağı bir veri tabanı veya **data marta** aktarılmasıdır. Eğer web sayfası kullanma eğilimini araştıyorsak ve firmamızın bir düzine web sunumcusu varsa, ilk aşama her web sunumcudan web günlük verilerini indirmektir. Bazen analiz yapmak istediğimiz veri, veri ambarında bulunabilir. Fakat veri istediğimiz tüm alanları kapsamayabilir. Bu durumda diğer kaynaklardan veri toplamak zorunda kalabiliriz. Farzedelim web sayfasındaki her tıklamanın kaydedildiği bir sunumcumuz olsun. Müşterilerin temel

dolaşma hareketleri hakkında bilgimiz olmasına rağmen, daha doğru bir model yaratmak için demografik verilere de ihtiyaç duyuyorsak, ilgili veriyi satın alabilir veya toplayabiliriz. Veri toplandıktan sonra eğitme verilerinin hacmini azaltmak için örnek veri tabanı oluşturuşabilir. [15]

Veriyi anlamak için ilk yapılması gereken verinin toplanmasıdır. Veri madenciliği için verinin tek bir tabloda toplanması en uygunu olacaktır.

1.6.3 Verinin Temizlenmesi ve Dönüştürülmesi

Verinin temizlenmesi ve dönüştürülmesi veri madenciliği projesinde en fazla kaynak harcanan aşamadır.Veri temizlemenin nedeni, veri setinden gürültü ve konu ile ilgisi olmayan bilginin ayıklanmasıdır. Veri dönüştürülmesinin nedeni ise kaynak verinin veri tipleri ve değerleri olarak değişik biçimlerde değiştirilmesidir.[16]

Veri madenciliği projesi için çalışmaya başladığımızda bazı gerçeklerle yüzyüze gelebiliriz. Veri tabanı sisteminde bazı kayıtlarda hatalar, olağandışı değerler, tutarsızlıklar ile karşılaşabiliriz. Diğer bir deyişle veri madenciliği ile analiz etmek istediğimiz veri eksik (Nitelik verilerinden, sadece bazılarının bulunması veya hiç bulunmaması, sadece kümelenmiş verilerin bulunması), gürültülü (Beklenenin dışında sapma göstermiş veya hatalı veri içeren), tutarsız (Örnek: Parçaları sınıflandırmada kullanılan bölüm kodlarından farklı kodlar içeren öğeler) olabilir. Eksik, gürültülü, tutarsız veri, büyük, gerçek veri tabanlarının ve veri ambarlarının genel özellikleridir. Veri birkaç sebepten dolayı eksik olabilir. Satış işlem verisindeki müşteri bilgileri gibi ilgili niteliklerden bazıları elde mevcut olmayabilir. Kayıt zaman, diğer veri önemsiz olabilir. Yanlış anlaşılmalar yüzünden veya donanımın düzgün çalışmaması yüzünden ilgili veri kaydedilmemiş olabilir. Tutarsızlığa, diğer verilerle çelişen verilerin silinmemesi, verinin kayıt zamanı ve veri üzerinde yapılan değişikliklere dikkat edilmemesi, neden olabilir. Verinin gürültülü olmasına, kullanılan veri toplama gereçlerinin hatalı olması, veri girmede personel veya bilgisayar hatası bulunması, verinin iletilmesindeki hatalar, kullanılan veri kodları veya isimlendirmedeki

[15] ZhaoHui Tang ve Jamie MacLennan, **a.g.e.**, s.13
[16] A.g.e., s.13

tutarsızlıklar, geçici bellek büyüklüğünün sınırlı olması gibi teknolojik yetersizlikler neden olabilir. [17]

1.6.3.1 Eksik Veriler

Birkaç nitelik için eksik veri girişi varsa bazı yöntemlerle eksik veriler tamamlanır. Bu amaçla, eksik verilerin elle doldurulması (Zaman alıcıdır ve eksik veri sayısı fazla olan büyük veri tabanlarında uygulanabilir değildir), eksik verinin tümü için "bilinmeyen" veya "-∞" gibi aynı değerin girilmesi (Eğer örneğin tüm eksik veriler "bilinmeyen" olarak tamamlanırsa veri madenciliği programı yanlışlıkla ilginç bir örüntü bulabilir, bu yüzden önerilen bir yöntem değildir), eksik değerler için ortalama değerin girilmesi, eksik değerlerin tüm veri kümelere ayrıldıktan bu kümelerin ortalama değerleri ile doldurulması, karar ağacı vb. yöntemlerl kullanarak en olası değerin atanması gibi yöntemler kullanılır.[18]

1.6.3.2 Gürültülü Veri

Gürültülü veri ölçülebilir bir değişkende rastgele hata veya sapmadır. Gürültülü verinin düzeltilmesi için kutulama yöntemi (Sıralanan değerler kutulara dağıtılır, herbir kutu kendi içinde olmak üzere, tüm değerler ortalama değerlerle, medyanla, minimum-maksimum değerlerden birine yakın olmasına göre minimum veya maksimum değerle değiştirilir), kümeleme yöntemi (Kümeleme yöntemi ile sapan değerler tespit edilir), insan ve bilgisayarın beraber çalışmasından oluşan yöntem (Bilgisayarın saptamadığı değerler için insan gücü kullanılır, anlamsız veriler çıkarılır, bu yöntem tüm veri tabanının insan gücü ile incelenmesinden çok daha etkindir.), regresyon (regresyon yöntemi verinin tahmin edilmesinde etkin rol oynayabilir) kullanılabilir.[19]

1.6.3.3 Tutarsız Veri

Bazı veri tutarsızlıkları belgeler, faturalar vb. dış kaynakların kullanılması ile düzeltilebilir. Veri bütünleşmesinden dolayı aynı öznitelik değişik veri tabanlarında farklı isimlerle tanımlanabilir. Tekrarlar görülebilir. Tekrarların önlenmesi için veri

[17] Jiawei Han ve Micheline Kamber**, a.g.e**., s.3
[18] **A.g.e**., s.5-6
[19] **A.g.e**., s.6-7

hakkında veri anlamına gelen "metada" kullanılabilir. Bazı tekrarlar korelasyon analizi ile saptanabilirler.

1.6.3.4 Verinin Dönüştürülmesi

Verinin dönüştürülmesi veri tipinin dönüştürülmesi ile yapılabilir. Veri madenciliği yazılımının gereksinimlerine uygun olarak bazen evet/hayır tipindeki bir verinin 1 veya 0 şekline dönüştürülmesi (bazı durumlarda da tersi uygulanabilir.)gerekir.

Yaş vb. verinin süreklilik gösterdiği durumlarda ilgili kolon belirlenmiş gruplara ayrılır (Çocuk, genç, orta yaşlı, yaşlı, çok yaşlı). Diğer bir yöntemde normalizasyon kullanılarak tüm sayısal değerlerin [0, +1] yada [-1, +1] aralığında gösterilmesidir.[20]

Bazı verilerin detaylarının atılarak sadece gerekli bölümünün kullanılması uygun olacaktır. Buna birleştirme adı verilir.

Bazen veri tabanına yeni bir öznitelik eklenebilir.[21]

1.6.4 Modelin Kurulması

Verinin temizlenmesi ve değişkenlerin dönüştürülmesinin ardından, modelin kurulması safhası gelmektedir. Modelin kurulmasında problemin tanımlanması aşamasında tespit edilen amaç büyük rol oynar. Bize verilen veri madenciliği tekniğinin çeşidi de (Küme Analizi, Birliktelik Kuralları ve Ardısık Zamanlı Örüntüler, Karar Agaçları, Yapay Sinir Ağları (YSA) vb.)önemlidir. Bu aşamada, bu alanda tecrübe sahibi kişilerle takım çalışması yapmak önemlidir. Borsa verileri ile ilgili bir çalışma yapılacaksa ekonomi alanında bilgi ve tecrübe sahibi bir yetkili ile birlikte verilerin ve sonuçların incelenmesi gereksiz çalışmaların yapılmasını önleyecektir.

Modelin kurulması verinin dönüştürülmesi işlemi kadar emek yoğun değildir, fakat VTBK sürecinin özüdür. Veri madenciliği tekniğini belirledikten sonra doğru algoritmanın seçilmesi nispeten daha kolaydır. Herbir veri madenciliği tekniği için geçerli birkaç algoritma mevcuttur. Çoğu zaman örneklem ile denemeler yapmadan en uygun algoritmanın ortaya çıkarılması olası değildir. Algoritmanın doğruluğu verinin

[20] ZhaoHui Tang ve Jamie MacLennan, **a.g.e.**, s.10
[21] Jiawei Han ve Micheline Kamber**, a.g.e**., s.114

yapısı ile ilişkilidir. Herbir öznitelik değerinin dağılımı, nitelikler arasındaki ilişkiler vb. belirleyici rol oynar. Örneğin tüm girdi değerleri ile tahmin edilebilir değerler arasındaki ilişki doğrusal ise, karar ağaçları iyi bir seçim olacaktır, eğer nitelikler arasındaki ilişkiler karmaşıksa, yapay sinir ağlarının kullanılması düşünülebilir. Doğru yaklaşım değişik algoritmalar kullanarak birkaç model yapmak ve aynı yazılımı kullanarak bu modellerin doğruluğunu karşılaştırmak olabilir. Kullanılan aynı algoritma ile değişik parametre ayarları ile farklı modeller oluşturarak denemeler yapmak, modelin doğruluğunun düzenlenmesinde kullanılabilir. [22]

1.6.5 Modelin Değerlendirilmesi

Modelin değerlendirilmesi aşamasında keşfedilen bilgi geçerlilik, yenilik, yararlılık kriterlerine göre değerlendirilir. Bu kriterler aynı zamanda veri madenciliğinin tanımında bulunan kriterlerdir. Geçerlik, uygulamada elde edilen bulguların, yeni veri kümelerinde de uygulanabilirliğidir. Yenilik ise elde edilen bulgunun tahmin edilmeyen ve bilinmeyen bir bulgu olma durumudur. Örneğin "İlişki = Sinek => Türü = Böcek % 100 güvenirlilikle" ifadesi geçerli olmakla beraber hiçbir yenilik getirmemektedir. Bulguların yararlı olması ise, bulguların potansiyel olarak fayda yaratma gücüne sahip olmasını göstermektedir.

Modelin değerlendirilmesi aşamasında modelin doğruluğunun yanısıra bulunan örüntülerin anlamları konunun uzman personeli ile birlikte değerlendirilir. Örüntü geçerli olmasına rağmen ekonomik değeri olmayabilir. Birlikte çalışılan uzman kişinin konusunda yeterli tecrübe ve bilgi birikimine sahip olması önemlidir. Modelden yararlı bir örüntü elde edilemesinin bazı nedenleri olabilir. Bunlardan bir tanesi verinin tamamen rasgele girdilerden oluşmasıdır. Çoğu zaman gerçek verilerden oluşan veri tabanları zengin bilgi kaynağıdır. İkinci neden olan modeldeki değişken kümesinin uygun küme olmaması daha çok karşılaşılan bir nedendir. Bu durumda geçerli değişken kümesinin bulunması için verinin temizlenmesi ve dönüştürülmesi aşamasının tekrarlanması gerekebilir. VTBK çevrimsel bir süreçtir. Doğru modelin bulunması birden fazla tekrarın yapılması gerekebilir.[23]

[22] ZhaoHui Tang ve Jamie MacLennan, **a.g.e.**, s.15
[23] **A.g.e**., s.16

1.6.6 Raporlama

Raporlama veri madenciliği bulguları için önemli bir dağıtım aracıdır. Kuruluşların birçoğunda veri madencilerinin görevi pazarlama bölümü yöneticilerine raporların verilmesidir. Birçok veri madenciliği yazılımı kullanıcılara önceden belirlenmiş raporlama şekline göre metin veya grafik olarak çıktı üretirler. Örüntüler ve tahmin hakkında olmak üzere iki tip rapor bulunur.[24]

Veri madenciliğinin etkin olabilmesi için, veri madenciliği sistemleri kullanıcıya ortaya çıkarılmış örüntüleri kurallar, tablolar, grafikler, karar ağaçları gibi birçok şekilde sunabilmelidir. [25]

Raporlamada kullanılan araçlar Ian H. Witten ve Eibe Frank tarafından ayrıntılı olarak sunulmuştur. Bunlar arasında karar tabloları, karar ağaçları, sınıflama kuralları, birliktelik kuralları, istisnai kurallar, karşılıklı ilişkilere ait kurallar, sayısal tahmine ait kurallar, olay bazlı sunum ve kümelemedir.[26]

1.6.7 Tahmin

Veri madenciliği projelerinin birçoğunda örüntülerin bulunması VTBK sürecinin % 50'lik bölümünü oluşturmaktadır. Amaç bu modellerin tahmin için kullanılmasıdır. Tahmin yapabilmek için eğitim veri tabanına ve yeni durumlardan oluşan verilere ihtiyaç vardır. Kredi verilmesi kararının alınmasında tahmin sıklıkla kullanılır.[27]

1.6.8 Kullanılan Uygulamalarla Entegrasyon

Veri madenciliği ile ilgili yazılımların, işletmede veya kurumda gerçek zamanlı kullanılan yazılımlarla tümleşik kullanılması veri madenciliğinin geniş çaplı kullanımında önemli rol oynayacaktır. Amazon.com web sitesinden elektronik alışveriş yaparken seçtiğiniz ürünün yanında bu ürüne benzer diğer ürünlerin seçilen ürüne göre otomatik olarak sunulması uygulaması, bu konuda uygun bir örnek olarak karşımıza çıkmaktadır.

---

[24] **A.g.e.**
[25] Jiawei Han ve Micheline Kamber**, a.g.e**., s.130
[26] Ian H. Witten ve Eibe Frank, **a.g.e**., s.61-81
[27] ZhaoHui Tang ve Jamie MacLennan, **a.g.e.**, s.16

1.6.9 Modelin Yönetilmesi

VTBK süreci durağan değildir. VTBK yinelemeli ve devimsel bir süreçtir. VTBK sürecinin işin özelliğine göre tekrarlanması gerekir. Bu tekrar sürecin tümünü kapsayabileceği gibi, sadece bazı bölümleriyle de sınırlı kalabilir.

Her veri madenciliği modelinin bir yaşam süresi vardır. Bazı işletmelerde örüntüler kalıcıdır fakat birçok işletmede örüntüler sık sık değişir. Örneğin çevrimiçi kitap satan sitelerde, hergün yeni kitaplar satış için listeye eklenir. Bu hergün yeni birliktelik kurallarının ortaya çıkması anlamına gelir. Veri madenciliği modelinin kullanım süresi sınırlıdır. Modelin yeni sürümü geliştirilmelidir. Modelin doğruluğunun belirlenmesi ve yeni model geliştirme süreci otomatik hale getirilmelidir. Veri modellerinin kullanılmasında veri güvenliğine dikkat edilmesi gerekmektedir. Veri modelleri örüntüleri, örüntülerde hassas verinin özetini içermektedir. Bu modellere erişim haklarının düzenlenmesi sırasında veri güvenliğine önem verilmesi uygun olacaktır.

2 VERİ MADENCİLİĞİ

2.1 Veri Madenciliğinin Tarihsel Gelişimi

Bugünün veri madenciliği teknikleri matematik, mantık ve bilgisayar mühendislerinin yapay zeka ve makina öğrenmesi konularını ortaya çıkarmak için bir araya geldikleri 1950 yıllarına dayanır.[28]

1960'larda yapay zeka ve istatistik uzmanları regresyon analizi, sinir ağları, doğrusal sınıflama modelleri gibi yeni algoritmalar geliştirdiler.[29] Veri madenciliği bu yıllarda bulunmasına rağmen, veri madenciliği terimi veriden güçlükle araştırmalar yapmak ve hiçbir faydası olmayan örüntüler bulmak anlamında kullanılan küçültücü bir kavramdı.[30]

1960'larda bilgi erişimi alanında, kümeleme teknikleri ve benzerlik ölçüleri bulundu. Bu teknikler metin tipindeki belgelere uygulandı. Daha sonra bu çalışmalar veri madenciliği alanında kullanılacaktı.[31]

Önceleri veri tabanı sistemleri yapısal verinin sorgulanması ve işlenmesine, bilgi erişimi ise bilginin birçok metin tabanlı büyük veri tabanından elde edilmesi ve düzenlenmesine odaklanmıştı.1960 yılının sonunda veri tabanı sistemleri ve bilgi erişimi aynı doğrultuda faaliyet göstermeye başlamışlardır.[32]

1970'ler, 80'ler, 90'larda çeşitli disiplinlerin (yapay zeka, bilgi erişimi, istatistik, ve veri tabanları) etkisi ve kişisel bilgisayarların kullanımının yaygınlaşması ile veriye erişim ve verinin analizi daha kolay hale gelmiştir. Bu aralıkta yeni programlama dilleri

---

[28] B.G. Buchanan, Brief History of Artificial Intelligence, "Brief History of Artificial Intelligence", (Çevrimiçi) http://www.aaai.org/AITopics/bbhist.html, (Erişim Tarihi: 07.09.2007), par.15.
[29] Margaret H. Dunham, Data mining Introductory and Advanced topics, Prentice Hall, 2003,s 13
[30] Usama Fayyad, Gregory Piatetsky-Shapiro, ve Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, c.17,S.3(1996). s.40
[31] Margaret H. Dunham, **a.g.e.**
[32] Jiawei Han ve Micheline Kamber**, a.g.e**., s.428

geliştirilmiş ve genetik algoritmalar, karar ağacı algoritmaları ve kümeleme tekniklerini kapsayan yeni hesaplama tekniklerinin kullanılmasına başlanmıştır.[33]

1990'ların başlarında, VTBK bulunmuş ve ilk VTBK semineri düzenlenmiştir.[34] Büyük miktardaki verinin birikmesi, bu veriden elde edilmesi hedeflenen bilginin işlenmesi için yeni tekniklerin bulunmasını gerektirmiştir.

Veri ambarları 1990 yıllarında kullanılmaya başlanmıştır. Veri ambarlarının kullanımı ile beraber karar destek sistemleri ve birliktelik kuralları algoritmalarının kullanılması da gündeme gelmiştir.[35]

1990'lar veri madenciliğinin, ilginç yeni bir teknoloji olmaktan çıkıp günlük iş hayatının bir parçası olduğu yıllardır. Bu gelişme bilgisayar depolama alanlarının ucuzlaması, işlemcilerin gücünün artması ve veri madenciliğinin faydalarının daha açık olarak görülmeye başlanması sayesinde gerçekleşmiştir.

2000'li yıllara gelindiğinde veri madenciliği bankacılık, sigortacılık, perakendecilik, borsa, telekomünikasyon, ilaç, sağlık ve internet alanlarında yoğun olarak kullanılmaktadır.[36]

2.2 Veri Madenciliği Kullanım Alanları

Veri madenciliği Kuzey Amerika ülkeleri başta olmak üzere, Avrupa ülkeleri, Japonya gibi ekonomik olarak gelişmiş ülkelerde yoğun olarak kullanılmaktadır. Ülkemizde veri madenciliği yeni tanınmaya başlamıştır ve büyük ölçekli firmalarda veri madenciliği yazılımları kullanılmaktadır. Aşağıdaki maddelerde veri madenciliğinin kullanıldığı alanlar ve kullanılma amaçları ayrıntılı olarak sunulmuştur.

2.2.1 Bankacılık

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,[37]
- Kredi kartı dolandırıcılıklarının tespiti,

[33] Margaret H. Dunham, **a.g.e.**
[34] Usama Fayyad, Gregory Piatetsky-Shapiro, ve Padhraic Smyth, **a.g.e.**
[35] Margaret H. Dunham, **a.g.e.**
[36] Usama Fayyad, Gregory Piatetsky-Shapiro, ve Padhraic Smyth, **a.g.e.**, s.37-38
[37] Haldun Akpınar, **a.g.e.**

- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi,
- Risk Yönetimi,[38]
- Firma derecelendirme,
- Faiz oranlarının tahmini,
- Borçlanma ve iflas tahminleri,
- Müşterilerin bölümlenmesi,[39]
- Müşteri kârlılığının analizi,
- Döviz kurlarının tahmini.

2.2.2 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,[40]
- Kampanya ürünlerini belirleme,
- Müşteri değerleme,
- Müşteri ilişkileri yönetimi,
- Satış tahminleri,[41]

---

[38] Ümmühan Altıntop, Internet Tabanlı Öğretimde Veri Madenciliği Tekniklerinin Uygulanması, Kocaeli Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Kocaeli, 2006, s.15

[39] Gülçin Buruncuk, Data Mining For Customer Segmentation And Profiling: A Case Study For A Fast Moving Consumer Goods (Fmcg) Company, Boğaziçi Üniversitesi, Bilgi Yönetim Sistemleri Enstitüsü Yüksek Lisans Tezi, İstanbul, 2006, s.7

[40] Haldun Akpınar, **a.g.e.**

- Ürünlerin rafa yerleştirilmesi,
- Satış ve mevsimsel farklar arasındaki örüntülerin tespit edilmesi.[42]

2.2.3 Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri örüntülerinin belirlenmesi.[43]
- Müşteri kaybı sebeplerinin belirlenmesi,
- Usulsüzlüklerin önlenmesi,
- Ana giderlerin azaltılması,
- Poliçe fiyatlarının belirlenmesi.[44]

2.2.4 Borsa

- Hisse senedi fiyat tahmini,
- Genel piyasa analizleri,
- Alım-satım stratejilerinin uygunluğu,[45]
- Portföy yönetimi,[46]

2.2.5 Telekomünikasyon

- Kalite ve iyileştirme analizleri,
- Hile tespitleri,
- Hatların yoğunluk tahminleri,[47]
- Müşteri kazanma ve elde tutma analizleri,[48]

---

[41] Ümmühan Altıntop, **a.g.e.**
[42] Gülçin Buruncuk,, , **a.g.e.**
[43] Haldun Akpınar, **a.g.e.**
[44] Erkan Kıyak, **a.g.e.**,s.20
[45] Ümmühan Altıntop, **a.g.e.**
[46] Gülçin Buruncuk,, , **a.g.e.**
[47] Ümmühan Altıntop, **a.g.e.**
[48] Erkan Kıyak, **a.g.e.**,s.20

2.2.6 İnternet

- Metin madenciliği,
- Web pazarlama,[49]
- Arama motorları.[50]

2.2.7 Üretim

- Envanter kontrolü,
- Donanım arızası analizi,
- Kaynak yönetimi,
- Süreç / kalite kontrol,
- Kapasite yönetimi,[51]
- Lojistik uygulamalar,
- Üretim süreçlerinin uygunluğu.[52]

2.2.8 Sağlık ve İlaç

- Test sonuçlarının tahmini,
- Ürün geliştirme,
- Tıbbi teşhis,
- Tedavi sürecinin belirlenmesi,
- Yeni ilaç türlerinin keşfi ve sınıflandırılması.[53]

2.2.9 Bilim ve Mühendislik

- Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesi,
- Yeni virüs türlerinin keşfi ve sınıflandırılması,
- Gen haritasının analizi ve genetik hastalıkların tespiti,

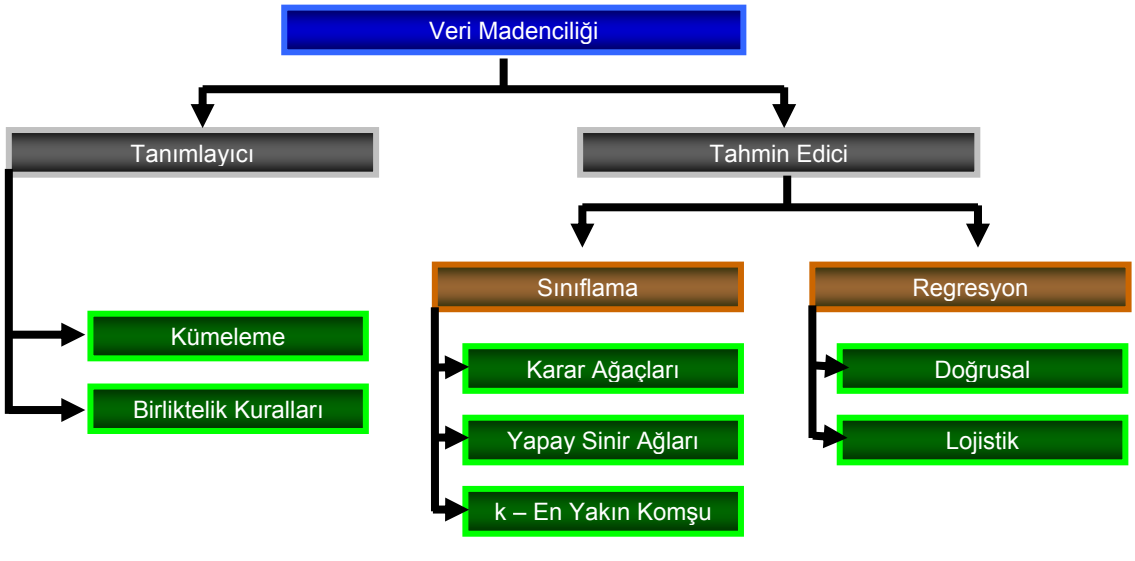
[49] Gülçin Buruncuk,, , **a.g.e.**
[50] Erkan Kıyak, **a.g.e**
[51] Gülçin Buruncuk,, , **a.g.e.**
[52] Ümmühan Altıntop, **a.g.e.**
[53] **A.g.e.**

- Kanserli hücrelerin tespiti,
- Gezegen yüzey şekillerinin, gezegen yerleşimlerinin ve yeni galaksilerin keşfi.[54]

2.3 Veri Madenciliğinde Kullanılan Modeller

Veri madenciliğinde kullanılan modeller, tahmin edici (Tahminci) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında incelenmektedir. [55]

Tahmin edici modeller veri tabanındaki bazı değişkenleri veya alanları kullanarak, ilgilenilen diğer değişkenlerin bilinmeyen veya gelecekteki değerlerini öngörmeyi hedeflemektedir. Tanımlayıcı modeller ise, verinin tanımlanmasında, insanlar tarafından yorumlanabilen örüntüler bulmaya odaklanmıştır. Modeller arasındaki sınırlar çok kesin olmamasına rağmen, ayrım bütün keşif hedefinin anlaşılması için yararlıdır. [56] Veri madenciliği modellerine ait sınıflandırma Şekil-4'te sunulmştur.



Şekil-4 Veri Madenciliği Modelleri

[54] **A.g.e.**
[55] Haldun Akpınar, **a.g.e.**
[56] Usama Fayyad, Gregory Piatetsky-Shapiro, ve Padhraic Smyth, **a.g.e.**, s.44

Şekil-4'te sunulan sınıflandırmanın çizgileri kesin sınırlar değildir. .Örneğin bazen yapay sinir ağları regresyon sınıfının altında incelenebilir.

2.3.1 Tahmin Edici Modeller

"Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Örneğin bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır".[57]

Bazı test sonuçlarına dayanarak hastaya tanı konması, bilinen bir grup ürünle müşterinin A ürününü de alma olasılığının tahmin edilmesi, geçmiş altı aylık indekse bakılarak Dow Jones indeksinin tahmin edilmesi tahmin edici modellere örnek olarak verilebilir.[58]

2.3.1.1 Sınıflama

Sınıflama bir veri maddesini önceden belirlenmiş sınıflardan birine eşleyen, öğrenen bir fonksiyondur.

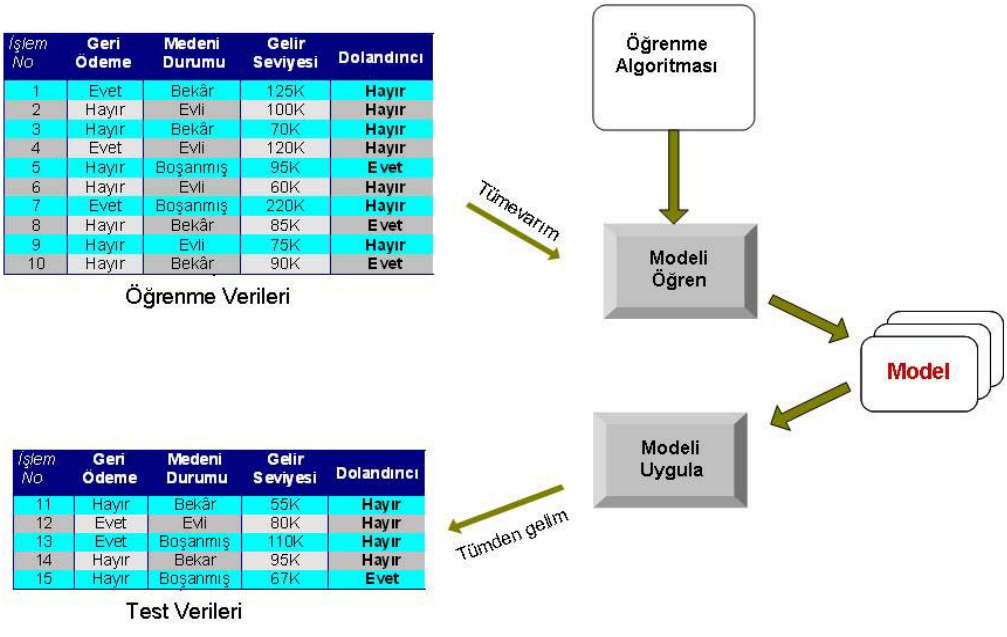
Sınıflama bölümsel değerlerin tahmin edilmesinde kullanılmaktadır. Tahmin edilmesinde öğrenme verisine dayanmaktadır. Sınıflamaya ait model şekil-5'te sunulmuştur.

Sınıflamanın kullanılma amaçları arasında

- Tümör hücrelerinin iyi veya kötü huylu olmasına göre bölümlenmesinde,
- Kredi kartı işlemlerinin yasal veya sahte olarak ayrılmasında,

[57] Haldun Akpınar, **a.g.e.**
[58] David Hand, Heikki Mannila ve Padharic Smyth, a.g.e.,s.195-196

- Haber içeriklerinin finans, hava, magazin ve spor olarak sınıflandırılmasında kullanılır.[59]



Sekil 5 **Sınıflama Modeli**

Kaynak: Pang-Ning Tan, Michael Steinbach ve Vipin Kumar, Introduction to Data Mining, Boston, Addison Wesley, 2005, s.148

Bir kaydın önceden belirlenmiş bir gruba girebilmesi için sınıflama algoritması ile öğrenme verileri kullanılarak hangi sınıfların var olduğu ve bu sınıflara girmek için bir kaydın hangi özelliklere sahip olması gerektiği otomatik olarak keşfedilir. Test verileriyle de bu öğrenmenin testi yapılarak ortaya çıkan kurallar optimum sayısına getirilir. Sınıflama algoritması, denetimli öğrenme kategorisine giren bir öğrenme biçimidir. Denetimli öğrenme, öğrenme ve test verilerinin hem girdi hem de çıktıyı içerecek şekilde olan verileri kullanmasıdır.[60]

[59] Pang-Ning Tan, Michael Steinbach ve Vipin Kumar, a.g.e.,s.148
[60] Fatih Aydoğan, E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi, Hacettepe Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Ankara, 2003, s.17

2.3.1.1.1 Karar Ağaçları

Karar ağacı, görünümünün ağaç şeklinde olmasından dolayı bu şekilde isimlendirilmiş, tahmin edici bir tekniktir. Yapısından dolayı programlamaya uygulanması kolaydır. Görsel açıdan da uygulama sonrası problemin çözümüne ait kuralların kullanıcılara rahatlıkla sunulabilir. Uygulanması fazla maliyet gerektirmez, güvenirlikleri iyi seviyededir. Bu yüzden veri madenciliğinde geniş kullanım alanı bulmaktadır.

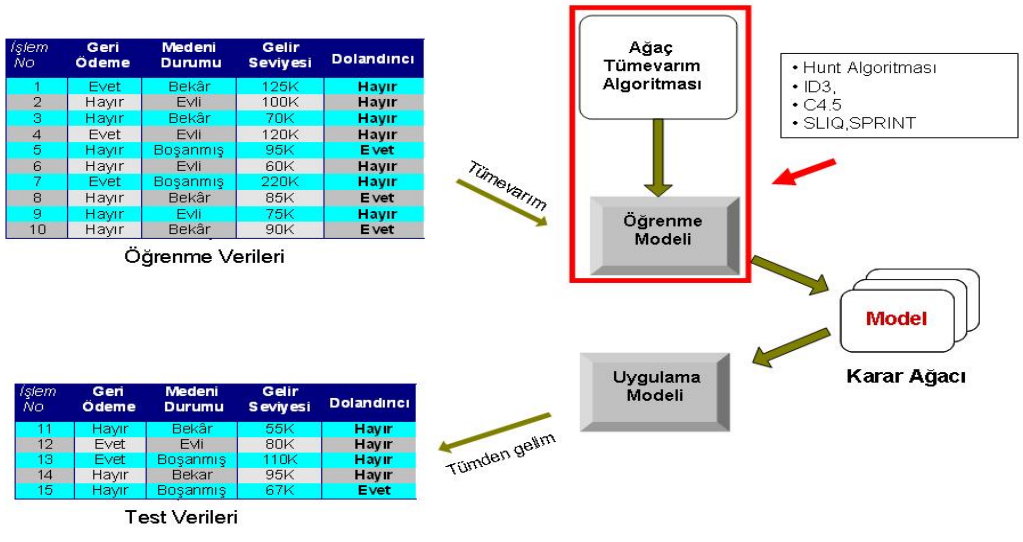
Karar ağacını kök, karar düğümleri, dallar ve yapraklardan oluşturmaktadır.[61] Kök, veri alanlarının içinde en önemli olandır. Karar düğümü, yapılacak testi temsil eder. Bu testin sonucu, veri, ağacın dalları arasında ayrılır. Ayrılma işlemleri bütün düğümlerde ardışık olarak gerçeklesir. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa, orada bir karar düğümü oluşur. Ancak gruba ayrılma sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden başlar ve yukarıdan asağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir. [62]

Karar ağacı tekniği sınıflamanın genel özelliklerini taşır. İki aşamalı bir tekniktir. Birinci basamakta öğrenme gerçekleştirilir. Öğrenme işlemi, sınıflama algoritması tarafından öğrenme verisi üzerinde gerçekleştirilir. Bu aşamadan sonra uygulanacak model oluşturulur. Öğrenilen model karar ağacını oluşturur. Model oluşturuşurken çıkan anlamsız dallar incelenerek modelden çıkarılır. Bu işleme budama adı verilir. Modelin oluşturulmasından sonra sıra test edilecek verinin sınıflanmasına gelir. Test verisi sınıflama kurallarının doğruluğunu test etmek amacıyla kullanılır. Eğer elde edilen sonuç belirlenen sınırlar içinde ise model daha büyük çapta yeni verilerin sınıflandırılması amacıyla kulanılabilir.

---

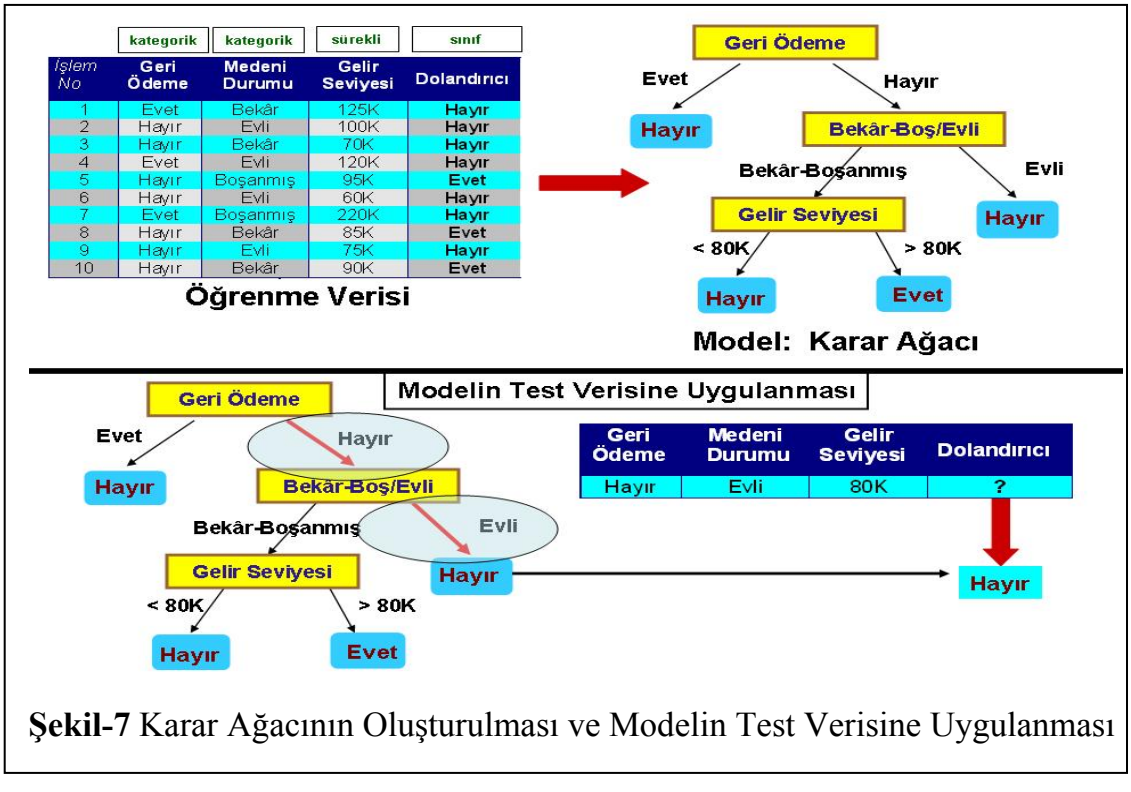
[61] Jiawei Han ve Micheline Kamber**, a.g.e**., s.221

[62] Serhat Özekeş, "Veri Madenciliği Modelleri ve Uygulama Alanları", **İstanbul Ticaret Üniversitesi Dergisi**, S.3 (Haziran 2003), s.65-82



Şekil 6 Karar Ağacının Uygulanması

Kaynak: Pang-Ning Tan, Michael Steinbach ve Vipin Kumar, Introduction to Data Mining, Boston, Addison Wesley, 2005, s.148



Şekil-7 Karar Ağacının Oluşturulması ve Modelin Test Verisine Uygulanması

Kaynak: Pang-Ning Tan, Michael Steinbach ve Vipin Kumar, Introduction to Data Mining, Boston, Addison Wesley, 2005, s.153

Karar ağaçlarının bakımı ve anlaşılması verinin karmaşıklığının artmasıyla zorlaşır. Eksik verilerin olması durumunda bölünme, değişkenlerinin birisinin değeri bilinmiyorsa karara varılması mümkün değildir. Karar ağacı algoritması elimizdeki veriyi bölümlere ayırırken dikkat edilecek nokta, bağımlı değişkenin değerini en çok belirleyen bağımsız değişkenleri ayırmaktır. Algoritmaya ait adımlar:

- Veri içinden ilgilendiğimiz bağımlı ve bağımsız değişkenlerin belirlenmesi,
- Hedef bağımlı değişkeni en çok etkileyen bağımsız değişkenin bulunması, bu maksatla her değişkenin hedefi ne kadar etkilediğinin bulunması ve en çok etkileyen değişkenin seçilmesi (Amaç bölünmeden sonra kalan parçaların bölünme öncesine oranla daha sade olmasını sağlamaktır.),
- Bölünme sonrasında kalan verilere aynı bölünme testlerinin yapılması ve daha sade gruplara ulaşılıncaya kadar bu işleme devam edilmesidir. [63]

2.3.1.1.2 Yapay Sinir Ağları (YSA)

YSA'da amaç fonksiyon birbirine bağlı basit işlemci birimlerden oluşan bir ağ üzerine dağıtılmıştır. YSA sinir ağlarında kullanılan öğrenme algoritmaları, veriden birimler arasındaki bağlantıya ait ağırlık değerlerini hesaplar, uygulama alanı geniştir ve bellek tabanlı yöntemler kadar yüksek miktarda işlem ve bellek gerektirmeyen bir modeldir.[64]

YSA üzerindeki çalışma, insan beyninin bilinen dijital bilgisayardan tamamen farklı bir şekilde çalıştığının farkına varılması ile başlamıştır. Farklı alanlardan birçok bilim adamı insan beyninin işlemsel sürecini modelleme konusunda araştırmalar yapmıştır. Beyin oldukça karmaşık, doğrusal olmayan, bilgiyi paralel işleyebilen bir sistemdir. İçeriğinde bulunan bileşenleri düzenleyerek bazı alanlarda bugün bilinen en hızlı bilgisayardan birkaç kat hızlı ve yüksek kalitede işlem yapma kapasitesine sahiptir. Bu alanlar arasında örüntü tanıma, algılama ve motor kontrol işlevleri sayılabilir. YSA insan beyninin soyut bir hesaplama modelidir. İnsan beyni yaklaşık 10^{11} adet, nöron adı

[63] Emrah Yılmaz, Kütahya İlinde Sosyal Sınıfların Belirlenmesi ve Veri Madenciliği ile Tüketici Profilinin Çıkarılmasına Yönelik Bir Uygulama, Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Yüksek Lisans Tezi, Kütahya, 2006, s.104

[64] H.Uğuz vd., "Apriori Algoritması Kullanılarak Web Kullanım Madenciliği Yönteminin Web Log Kayıtlarına Uygulanması", IJCI Proceeding of International Conference on Signal Processing, C.1, S.2 (2000), s: 499-501

verilen küçük birime sahiptir. Bu birimler arasında yaklaşık olarak 10^{15} adet bağlantı bulunmaktadır. Gerçek yapıya benzer olarak YSA'da yapay nöronlardan (işlem birimleri) ve bunlar arasındaki bağlantılardan oluşmaktadır. Bu ağa grafiksel olarak yaklaşırsak, nöronların düğüm noktaları ve bağlantıların da kenarlar olarak görüntülendiğini söyleyebiliriz. YSA adından da anlaşılacağı gibi birçok düğüm noktasının birbirine yönsel bağlantılarla bağlandığı bir ağ yapısıdır. Bu yapıda her düğüm noktası bir işlemci birimi, düğüm noktaları arasındaki bağlantılar da nedensel ilişkileri temsil ederler. Her düğüm noktası uyarlanabilir özelliktedir. Uyarlanabilir olması, bu düğüm noktalarının çıktılarının, düğüm noktaları ile ilgili değiştirilebilir parametrelere dayalı olduğunu göstermektedir. YSA küçük işlemcilerden oluşmuş büyük dağıtık paralel bir işlemcidir. YSA birimleri arası bağlantı güçleri ile ifade edilen deneysel bilgiyi öğrenme ve bu bilgiyi kullanıma sunma kabiliyetine sahiptir.[65]

Yapay bir nöron YAS'nın temelini oluşturan bilgi işleme ünitesidir. Şekil 8'de gösterildiği gibi yapay nöron üç ana birimden meydana gelmektedir.

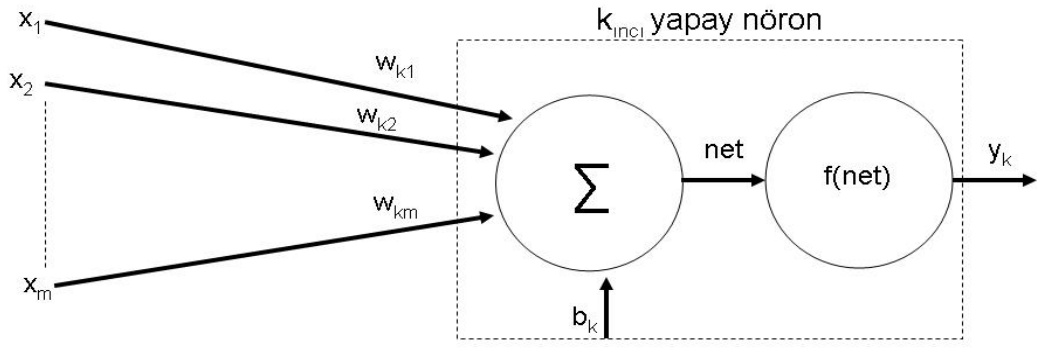
- Değişik girdilerden gelen bir grup bağlantı $\mathbf{x_i}$ (veya sinapsler), bağlantılar bir ağırlık veya güç değeri ile karakterize edilmiştir. Bu değer $\mathbf{w_{ki}}$ ile gösterilmektedir. Birinci indeks ilgili nörona, ikinci indeks ise ilgili bağlantının ağırlık değerine karşılık gelmektedir. Genelde yapay nöronun ağırlık değeri eksi veya artı değerler arasında olabilir.

- Girdi sinyallerin ağırlık değerlerine göre toplanmasını sağlayan bir toplama fonksiyonu bulunmaktadır. Bu işlem doğrusal bir birleştiricidir.

- Nöronun çıktısının büyüklüğünün belirli limitler içinde olmasını sağlayan f fonksiyonu vardır.

- Ayrıca b_k olarak adlandırılan ve önyargıyı temsil eden bir etki değeri bulunmaktadır. Artı veya eksi değerde olmasına göre f fonksiyonuna girdi değerine arttırıcı veya azaltıcı etki yapar.[66]

[65] Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, Danvers, 2003, s.206
[66] Mehmed Kantardzic, **a.g.e.**,s.208



Şekil 8 Yapay Nöron Modeli

Kaynak: Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, Danvers, 2003, s.208

2.3.1.1.3 k-En Yakın Komşu

En yakın komşu sınıflandırıcıları benzerlik yöntemi ile öğrenmeyi esas alır. Eğitim örnekleri n-boyutlu sayısal nitelik ile tanımlanırlar. Her bir örnek n-boyutlu uzayda bir noktayı temsil eder. Bu şekilde tüm eğitim örnekleri n-boyutlu uzayda depolanır. Bilinmeyen bir örnek geldiğinde, bir k-en yakın komşu sınıflandırıcısı bilinmeyen örneğe en yakın k eğitim örneğini bulmak için örüntü uzayını tarar. K eğitim örnekleri bilinmeyen örneğin k-en yakın komşularıdır. Yakınlık Öklit mesafesi kullanılarak ölçülür. Öklit mesafesi X = (x_1;x_2; ;x_n) ve Y= (y_1;y_2; ;y_n) olarak adlanrılan iki nokta arasında;

d (X, Y)= $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ formülü ile bulunur.

Bilinmeyen örnek, örüntü uzayında kendisine en yakın eğitim örnekleri kümesine atanır. En yakın komşu sınıflandırıcıları tüm eğitim örneklerini depoladıkları için örnek tabanlıdır. Sınıflandırılmamış bir örnek karşılaştırılmak istendiğinde eğer olası komşularının sayısı fazlaysa hesaplama zamanı oldukça yüksektir. Bu durumda indeksleme tekniklerinin kullanılması gerekebilir. Karar ağacındaki tümevarım ve tümdengelimin sürecinde uygulananın aksine, en yakın komşu sınıflandırıcıları her bir niteliğe eşit ağırlık verirler. Bu durum, veride çok fazla ilgisiz nitelik bulunduğunda karışıklığa sebep olabilir. [67]

2.3.1.2 Regresyon Analizi

Regresyon Analizi süreklilik gösteren değerlerin tahmin edilmesinde kullanılır.[68] Regresyon sınıflamaya benzemektedir. Regresyon Analizi'nin sınıflamadan en önemli farkı tahmin edilebilir değişkenin sürekli bir sayı olmasıdır. Regresyon Analizi teknikleri yüzyıllardan beri istatistiğin geniş çapta .çalışmalar yapılan bir alanıdır. Doğrusal regresyon, lojistik regresyon en çok bilinen regresyon çeşitleridir. Dağıtım metodları, hava ısısına bağlı rüzgar hızının ve nem oranının tahmininde Regresyon Analizi kullanılmaktadır.[69]

Regresyon Analizi varolan veriye formüllerin uygulanması ile tahminler yapmada kullanılmaktadır. Doğrusal veya lojistik regresyon tekniklerini kullanarak, varolan veriden fonksiyon elde edilmektedir. Yeni veri varolan fonksiyona uygulanarak tahmin yapmada kullanılmaktadır. Değerleri bilinen değişkenlerin kullanılarak diğer değişkenlerin tahmininde kullanılır. Regresyon terminolojisinde, tahmin edilecek olan değişken "bağımlı değişken", bağımlı değişkeni tahmin etmek için kullanılan değişken ya da değişkenler de "bağımsız değişken" olarak adlandırılır.[70]

---

[67] Jiawei Han ve Micheline Kamber**, a.g.e**., s.247-248
[68] Ümmühan Altıntop, **a.g.e.,** s.21
[69] ZhaoHui Tang ve Jamie MacLennan, **a.g.e.**, s.8
[70] Joseph F. Hair vd., Multivariate Data Analysis, Prentice Hall, New Jersey, 1998, s.680-695

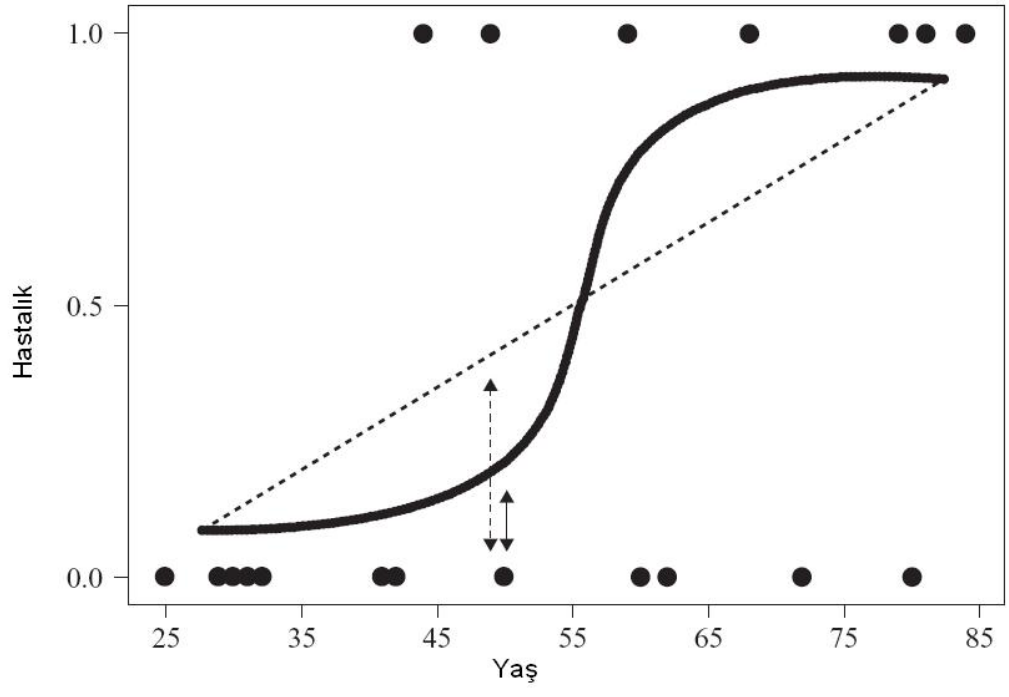
- **Doğrusal Regresyon**

Bir bağımlı bir bağımsız değişkenden oluşan en basit regresyon analizi "basit doğrusal regresyon", iki ya da daha fazla bağımsız değişken içeren regresyon analizi de "çoklu regresyon analizi" olarak adlandırılır.

- **Lojistik Regresyon**

Doğrusal regresyon sürekli değer fonksiyonlarını modellemede kullanılır. Genelleştirilmiş regresyon modelleri ise kategoriksel cevap değişkenlerinin modellemesinde doğrusal regresyon yaklaşımının kullanılmasının kuramsal altyapısını temsil eder. Genelleştirilmiş doğrusal regresyon modelinin en yaygın tipi lojistik regresyon modelidir. Lojistik regresyon bir grup tahmin edici değişkenin doğrusal fonksiyonu olaral gerçekleşmiş bir olayın olasılığını modeller. Bağımlı değişkenin değerini hesaplamak yerine, bağımlı değişkenin verilen bir değeri alma olasılığını hesaplar. Örneğin bir müşterinin kredibilitesinin iyi veya kötü olduğunu tahmin etmek için, lojistik yöntem iyi kredibilite olasılığını tahmin etmeye çalışır. Bağımlı değişkenin güncel durumu tahmin edilen olasılığa bakarak tespit edilir. Eğer tahmin edilen olasılık değeri 0.50 değerinden büyükse tahmin EVET (iyi kredibilite), diğer durumda ise HAYIR (kötü kredibilite) değerine yakındır. Bu yüzden lojistik regresyonda kredibilite "p" başarı olasılığı olarak adlandırılır. Diğer yönden, girdilerin bazılarının sayısal olması veya olmaması lojistik regresyon modeli için önemli değildir. Bu yüzden lojistik regresyon daha genel veri çeşitlerinin kullanımını destekler.

Doğrusal regresyonun aksine, lojistik regresyon çizgisi doğrusal değildir. Bununla ilgili örnek Şekil-9'da sunulmuştur.



Şekil 9 Örnek Lojistik Regresyon Eğrisi

2.3.2 Tanımlayıcı Modeller

Tanımlayıcı modeller ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanmasında kullanılmaktadır.[71] Tanımlayıcı modeller arasında Kümeleme.(Clustering), Birliktelik Kuralları (Association Rules), Ardışık Zamanlı Örüntüler sayılabilir.

2.3.2.1 Kümeleme.Analizi

Fiziksel veya soyut nesnelerden oluşan bir grubun, birbirine benzer nesnelerin aynı sınıflarda kalacak şekilde daha küçük gruplara ayrılması işlemine kümeleme adı verilir. Küme içinde bulunduğu gruptaki nesnelere benzeyen fakat diğer gruptaki nesnelere

[71] Haldun Akpınar, **a.g.e.**

benzemeyen veri nesnelerinin biraraya toplanmış halidir. Aynı kümede bulunan veri nesneleri üzerinde birçok uygulama daha kolay ve hızlı olarak yapılabilir.[72]

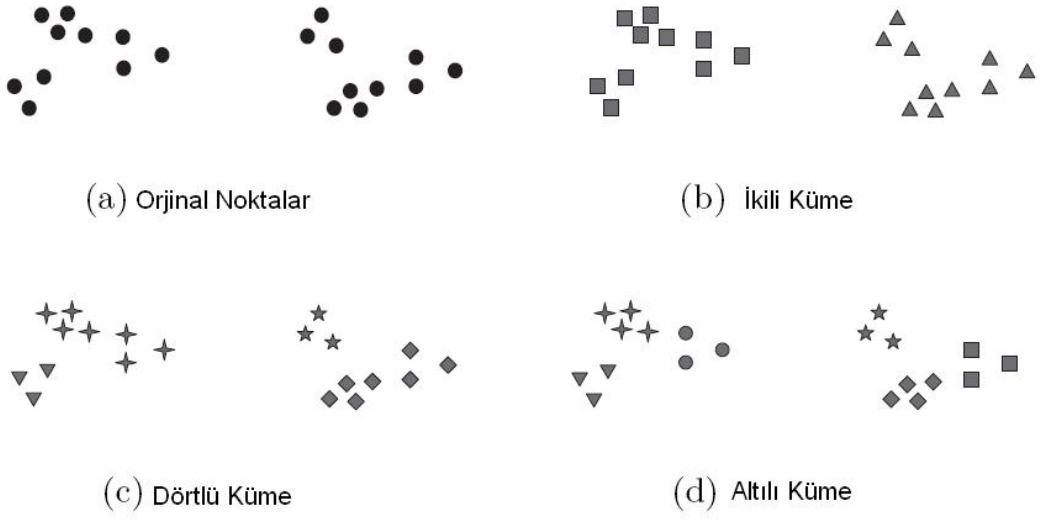
Kümeleme günlük hayatımızda sıklıkla başvurduğumuz bir yöntemdir. Bitki hayvan sınıflandırılmasının yapılması, hayvanların etçil, otçul olarak sınıflandırılması vb. şekilde yapılan sınıflandırmalar sürekli, gelişen ve daha ayrıntılı olacak şekilde yapılmaya devam eder. Kümeleme yöntemi örüntü tanıma, veri analizi, görüntü işleme, pazar analizi gibi sayısal uygulamalarda yaygın olarak kullanılmıştır. Kümeleme yapılarak kalabalık ve seyrek bölgeler ortaya çıkarılabilir. Böylece veri alanları arasındaki ilginç korelasyon ve örüntü dağılımları ortaya çıkarılabilir.World Wide Web (WWW)'de bulunan belgelerin bilgiye daha kolay ulaşım için gruplanması, benzer işlevlere sahip genlerin sınıflandırılması, bitki ve hayvan cinslerinin kökünün araştırılması, bir yerleşim birimindeki benzer özellikteki evlerin gruplanması gibi birçok alanda kullanılabilir. Sınıflamanın aksine, kümeleme sınıfların önceden belirlenmesine veya öğrenme verilerine ihtiyaç duymaz. [73] Şekil-10'da aynı verinin farklı şekilde kümelenme şekilleri görülmektedir.

2.3.2.2 Birliktelik Analizi (Association Rules)

Birliktelik kuralları, büyük veri kümelerinde, veriler arasındaki ilginç yerel örüntüleri, bağlantıları ve kuralları bulmaya yönelik, denetimsiz (unsupervised) veri madenciliği şeklidir. [74] Örüntünü verinin belli bir yönüyle ilgili bilgiler veren bir yerel kavram, model ise verinin tüm betimlemesini yapan genel bir kavramdır.[75]

---

[72] Jiawei Han ve Micheline Kamber**, a.g.e**., s.270
[73] **A.g.e.**
[74] Jiawei Han ve Micheline Kamber**, a.g.e**., s.186
[75] David Hand, Heikki Mannila ve Padharic Smyth, a.g.e.,s.254



Şekil 10 Aynı Grup Verinin Farklı Şekillerde Gruplanması

Kaynak: Pang-Ning Tan, Michael Steinbach ve Vipin Kumar, Introduction to Data Mining, Boston, Addison Wesley, 2005, s.491

Birliktelik analizi insanların veri madenciliğini anlamak için kafalarında tasarladıkları veri madenliciğine en yakın olanıdır. Birliktelik analizi büyük bir veri tabanında altın aramaktır. Veri tabanındaki altın, bizim veri tabanı ile ilgili daha önce bilmediğimiz ve muhtemelen açıkça ifade edemediğimiz bir kuraldır. Bu yöntembilim veri tabanındaki tüm ilginç örüntüleri bulur. Her taşın altına bakılır. Bu aynı zamanda modelin zayıflığıdır. Kullanıcı ortaya çıkan bilginin büyüklüğü karşısında bunalabilir ve bu miktardaki bilginin kullanılabilirliğinin analizi zor ve zaman alıcıdır. [76]

Veri madenciliği sonucu ortaya çıkarılan ilişkiler birliktelik kuralları olarak gösterilebilir. Birliktelik analizi, satın alma eğilimlerinin tanımlanması, müşterinin hangi mal veya hizmetleri almaya eğilimli olduğunun saptanması ve bu yolla müşteriye daha fazla mal satılmasını sağlamak için sıklıkla başvurulan bir modeldir.

[76] Mehmed Kantardzic, **a.g.e.**, s.175

Tablo-1'de birlikteklik kurallarının en tipik örneklerinden biri olan pazar sepeti uygulamasına ait veriler bulunmaktadır.

Tablo-1 Alışveriş Verileri

İşlem No	Satın Alınan Ürün Alt Kümeleri
1	**Ekmek, Süt**
2	**Ekmek, Çocuk Bezi, Bira, Yumurta**
3	**Süt, Çocuk Bezi, Bira, Kola**
4	**Ekmek, Süt, Çocuk Bezi, Bira**
5	**Ekmek, Süt, Çocuk Bezi, Kola**

Kaynak: Pang-Ning Tan, Michael Steinbach ve Vipin Kumar, Introduction to Data Mining, Boston, Addison Wesley, 2005, s.327

Açığa çıkarılan örüntüler birliktelik kuralları şeklinde gösterilebilirler. Aşağıdaki kural Tablo-1'den elde edilmiştir.

- {Süt, Çocuk Bezi}→{Bira}

Bu kural çocuk bezi satışı ile bira satışı arasında güçlü bir ilişki olduğunu göstermektedir. [77] Aşağıdaki örnekler birliktelik kurallarına örnek olarak gösterilebilir.

- Yaş(X, "30 - 34") Λ Gelir (X, "42K - 48K") => satılalınan (X; "Plasma TV")
- Satılalınan (X; "bilgisayar") => Satılalınan (X; "finansal_yönetim_yazılımı") [78]

Birliktelik analizi Pazar sepeti analizinin yanısıra aşağıda sunulan alanlarda da kullanılmaktadır:

- Biyoinformatik, tıbbi teşhis,
- Web madenciliği,

[77] Pang-Ning Tan, Michael Steinbach ve Vipin Kumar, a.g.e.,s.328
[78] Jiawei Han ve Micheline Kamber**, a.g.e**., s.188

- Bilimsel veri analizi (Dünya'ya ait okyanus, kara ve atmosferik verilerin bilimsel analizinde),

- Terör olaylarının tespit edilmesi.

2.3.2.2.1 Birliktelik Analizi ile İlgili Tanımlar

I= $\{i_1,i_2,i_3,...,i_d\}$ kümesinin pazar sepeti analizinde kullanılan tüm öğelerin kümesi ve T=$\{t_1, t_2, t_3, ,t_N\}$ bu analizdeki tüm işlemlerin kümesi olsun. Herbir t_i işlemi I öğe kümesinin alt kümelerini barındırmaktadır. Birliktelik analizinde öğe kümesi bir veya daha fazla öğe barındıran kümeyi temsil etmektedir.

- **Birliktelik Kuralı:** $X \rightarrow Y$ iken $X \subset I$, $Y \subset I$, ve $X \cap Y=\emptyset$

- **Öğe Kümesi** bir veya daha fazla öğeden meydana gelen kümedir. { Süt, Çocuk Bezi, Bira, Kola }

- **k-öğe kümesi** içinde k adet öğe bulunduran kümedir.

- **Destek sayısı (σ)** öğe kümesinin görülme sıklığıdır. Örnek: σ ({Süt, Ekmek, Çocuk Bezi})=2

- **Destek (Support)** $X \rightarrow Y$ kuralındaki X ve Y öğelerinin/öğe setlerinin her ikisini de kapsayan işlemlerin toplam işlemlere oranıdır. Örnek:

  {Süt, Çocuk Bezi}→{Bira}

  s = ({Süt, Çocuk Bezi, Bira})/ |T| = 2/5=0.4

  $s(X \rightarrow Y)=\frac{\sigma(X \cup Y)}{N}$

- **Güven (Confidence)** $X \rightarrow Y$ kuralında Y öğe kümesindeki elemanların X öğe kümesi elemanlarının bulunduğu işlemlerde hangi sıklıkta bulunduğunu göstermektedir.

  {Süt, Çocuk Bezi}→{Bira}

$$c = \frac{\sigma(\text{Süt, Çocuk Bezi, Bira})}{\sigma(\text{Süt, Çocuk Bezi})} = \frac{2}{3}$$

$$c\ (X \rightarrow Y) = \frac{\sigma(X\ U\ Y)}{\sigma(X)}$$

Destek önemli bir ölçü birimidir. Eğer bir kural düşük desteğe sahipse bu durumda kural şans eseri ortaya çıkmış olabilir. Destek kuralın kullanıbilirliğini, güven ise doğruluğunu gösterir.

Birliktelik kurallarını bulmak için sık tekrarlanan öğelerin bulunması, bu öğelerin önceden belirlenen minimum destek sayısı kadar tekrarlanması gerekir. Daha sonra tekrarlanan öğelerden güçlü birliktelik kuralları oluşturulur. Bu kurallar minimum destek ve minimum güven değerlerini karşılamalıdır.Sık tekrarlanan öğeleri bulmak için kullanılan en temel yöntem Apriori Algoritmasıdır. [79]

2.3.2.2.2 Appriori Algoritması

Apriori Algoritması 1994 yılında bulunmuştur.[80] Appriori Algoritması birkaç döngü ile veri tabanında sık bulunan öğe kümelerini hesaplar. Döngü *i*, i eleman sayısına sahip kümeleri hesaplar. Her bir döngü iki aşamadan oluşur: birinci aşamada aday kümeler üretilir, ikinci aşamada ise aday kümeler sayılır ve belirlenen değerlere göre seçim yapılır.[81]

İlk döngünün ilk aşamasında, yaratılan aday alt kümelerin hepsi veri tabanındaki tüm öğeleri kapsamaktadır. Sayma aşamasında, algoritma bu adayların destek sayılarını belirlemek için tüm veri tabanını tarar. Sonunda sadece belirtilen destek değerinden yukarıda destek değerine sahip olanlar sık tekrarlananlar grubuna seçilirler. Böylece ilk döngü sonunda tüm sık kullanılan öğe setleri saptanmış olur.[82]

İkinci basamakta, tüm ikili alt kümeler seçilmek için adaydır. Fakat önceki döngülerden elde edilen bilgilerin ışığında Apriori algoritması destek sayısı az olanları

[79] Serhat Özekeş, **a.g.e.**, s.13
[80] Rakesh Agrawal ve Ramakrishnan Srikant, "Fast Algorithms For Mining Association Rules", Proceedings of the 20th VLDB Conference, Santiago, Şili, 1994, s.487-499
[81] Mehmed Kantardzic, **a.g.e.**, s.178
[82] **A.g.e.**

budar. Budama eğer bir ürün kümesi sık tekrarlanıyorsa, bu kümenin tüm alt kümeleri sık tekrarlanıyor prensibi doğrultusunda yapılır. Aynı şekilde eğer bir ürün kümesi sık tekrarlanmıyorsa bu ürünün alt kümeleri de sık tekrarlanmıyor anlamına gelir ve bu alt kümeler atılır.[83]

Konunun daha anlaşılabilir olması için Apriori algoritması Tablo-2'de görülen pazar sepeti işlemleri üzerinde uygulamalı olarak anlatılacaktır.

Tablo-2 Basit İşlem Veri Tabanı Modeli

İşlem No	Satın Alınan Ürün Alt Kümeleri
001	A, C, D
002	B, C, E
003	A, B, C, E
004	B,E

Kaynak: Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, Danvers, 2003, s.176

Veri tabanı birliktelik kuralları iki aşamada incelenebilir. Bu aşamalar, önceden belirlenmiş belli eşik değeri üzerinde olan kümelerin bulunması ve bulunan bu kümelerden önceden belirlenmiş belli bir güven değeri üzerindeki kuralların üretilmesidir.

Apriori algoritmasının uygulanmasına ait aşamalar Tablo-3'te sunulmuştur. Farzedelim istenen minimum destek değeri s = % 50, minimum güvenilirlik değeri c = % 50 olsun. İlk aşamaya ait değerler Tablo-3'teki apriori algoritması birinci tekrar bölümünde görülmektedir. İlk aşamada 1 elemanlı kümeler yaratılmaktadır. İkinci aşamanın birinci bölümünde bunların kaç adet olduğu ve minimum destek oranları hesap edilmekte, ikinci bölümde ise % 50 destek değerinin altındaki kümeler atılmaktadır. Bu hesaplamalar algoritmanın gerektirdiği döngü sayısı kadar yapılır.

[83] **A.g.e.**

Tablo-3 Apriori Algoritmasının Uygulanması

Apriori Algoritması Birinci Tekrar

1-elemanlı alt küme C_1
{A}
{C}
{D}
{B}
{E}

a. Yaratma Aşaması

1-elemanlı alt küme	Adet	s (%)
{A}	2	50
{C}	3	75
{D}	1	25
{B}	3	75
{E}	3	75

b.1 Sayma Aşaması

Büyük1-elemanlı alt küme L_1	Adet	s (%)
{A}	2	50
{C}	3	75
{B}	3	75
{E}	3	75

b.2 Seçme Aşaması

Apriori Algoritması İkinci Tekrar

2-elemanlı alt küme C_2
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

a. Yaratma Aşaması

2-elemanlı alt küme	Adet	s (%)
{A, B}	1	25
{A, C}	2	50
{A, E}	1	25
{B, C}	2	50
{B, E}	3	75
{C, E}	2	50

b.1 Sayma Aşaması

Büyük 2-elemanlı alt küme	Adet	s (%)
{A, C}	2	50
{B, C}	2	50
{B, E}	3	75
{C, E}	2	50

b.2 Seçme Aşaması

Apriori Algoritması Üçüncü Tekrar

3-elemanlı alt küme C_3
{B, C, E}

a. Yaratma Aşaması

3-elemanlı alt küme	Adet	s (%)
{B, C, E}	2	50

b.1 Sayma Aşaması

Büyük 3-elemanlı alt küme	Adet	s (%)
{B, C, E}	2	50

b.2 Seçme Aşaması

Kaynak: Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, Danvers, 2003, s.178

Bu aşamadan sonra yapılması gereken güvenilirlik ölçütlerine göre kuralların bulunmasıdır. Üçüncü tekrar sonucu elde edilen 3-elemanlı alt kümeden ve ikinci tekrar sonucu elde edilen iki elemanlı alt kümeden faydalanılarak güvenilirlik katsayısı hesaplanır.

- c ({B,C}→)= $\frac{s(B,C,E)}{s(B,C)}=\frac{2}{2}=1$ (% 100) ;Güvenilirlik arzu edilen %50'lik orandan daha yüksek olduğu için kural kabul edilir.

2.3.2.2.3 Tahminci Appriori Algoritması

Tahminci Apriori algoritması Schaffer tarafından bulunmuştur. [84]

Tahminci Apriori algoritması standart apriori algoritmasından birliktelik kuralına uyguladığı değişik ölçü değeri nedeni ile ayrılır. Tahminci Apriori ve Apriori, destek verilerine göre bir arama yapar ve destek değerinin aşağı doğru kapanmasını kullanarak üssel arama alanınındaki olası birliktelik kurallarının budanmasını sağlar. Apriori güvenilirliği esas alır ve kuralların derecelendirmesini buna göre yapar. Tahminci Apriori ise kuralların güvenilirliğini destek değerlerine göre değerlendirir. İlginçlik ölçütü, keşfedilmemiş veride beklenen doğruluk değerini en yüksek değerine ulaştırmaktır. Tahminci Apriori kullanıldığında yüksek destek değerine sahip çok genel kurallar veya yüksek güvenilirlik ve düşük desteğe sahip çok özgün kurallar bulmak mümkündür. Apriori algortimasında, budama süreci destek değerine göre yapılmakta ve sonrasında bulunan kurallar güvenilirlik derecesi kontrol edilmektedir. Tahminci Apriori algoritmasının ilginçlik ölçüsü önceden belirlenmiş doğruluk değeri olduğu için, algoritma n adet en iyi birliktelik kuralını çıktı olarak verir. n rakamı kullanıcı tarafından belirlenmektedir. Ayrıca tahminci apprioride eklenen bir aşamada, genel kuralların daha özel tipleri olan kuralların budanmasıdır.[85]

2.3.2.2.4 Tertius Algoritması

Peter A. Flach Nicolas Lachiche tarafından geliştirilen Tertius algoritması 7500 satırlık C kodundan oluşmaktadır. Tertius, k adet en çok onaylanmış hipotezin bulunmasında, uygun en iyi birinci değer aramasını kullanan bir makine öğrenme aracıdır. Bu durum kuralın yeniliği ve kullanılabilirliği ölçülerinin dengeye getirilmesini sağlamaktadır. Arama esnasındaki tekrarların önlenmesi için artıksız düzeltme operatörü kullanılmaktadır. Bölümlere ait verinin analizinde Wickens. tarafından düzenlenen olasılık tablosu kullanılmaktadır. Gözlemlenen ve tahmin edilen değerler arasındaki ilişkinin ortaya çıkarılmasında Pearson χ^2 istatistiksel yöntemi kullanılmaktadır. Tertius diğer tümcesel kural bulma yöntemlerine yenilik olarak

[84] Tobias Scheffer, "Finding Association Rules that Trade Support Optimally Against Confidence", Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases Conference, Freiburg, Almanya, Eylül 2001, s.424-435

[85] Stefan Mutter, Classification using Association Rules, Waikato Üniversitesi Yüksek Lisans Tezi, Yeni Zelanda, 2004, s.15-16

doğrulama değeri kavramını getirmiştir. Tertius algoritmasında doğrulama değeri buluşsal olarak ölçülmektedir. Bulunan örneklerin budama işlemi karşıt örneklerin incelenmesi ile yapılır.[86]

Tertius algoritmasının çalıştırılması sonucu elde edilen çıktıya ait alanların açıklaması Tablo-4'te gösterilmiştir.

Tablo-4 WEKA programı Tertius Algoritması Çıktısı

```
Tertius
=======
1. /* 0.633754 0.071429 */ play = yes ==> outlook = overcast or humidity = normal
2. /* 0.607625 0.000000 */ humidity = normal ==> temperature = cool or play = yes
3. /* 0.607625 0.000000 */ temperature = cool ==> humidity = normal
4. /* 0.594071 0.214286 */ humidity = normal ==> temperature = cool
5. /* 0.590214 0.000000 */ outlook = sunny and humidity = high ==> play = no
6. /* 0.555556 0.000000 */ play = no ==> outlook = sunny or windy = TRUE
7. /* 0.486606 0.000000 */ humidity = normal ==> outlook = rainy or play = yes
8. /* 0.486606 0.000000 */ outlook = sunny and play = no ==> humidity = high
Number of hypotheses considered: 1353
Number of hypotheses explored: 481
Time: 00 min 01 s 680 ms
```

• Kurallarla beraber verilen **ilk rakam** doğrulama değeridir.
• Kurallarla beraber verilen **ikinci rakam**, karşıt örneklerin frekans değeridir.
• **Number of hypotheses considered** (Hesaba katılan varsayımların sayısı) ise filtre işlemi sonrası üretilen kuralların sayısıdır.
• **Number of hypotheses explored** (Keşfedilen varsayımların sayısı)" ise potansiyel olarak ilginç kuralların sayısıdır.

Kaynak: Amelie Deltour, Tertius extension to WEKA(Technical Report), University of Bristol, Bristol, 2001, s.9

[86] Peter A.Flach ve Nicolas Lachiche, "Confirmation-Guided Discovery of First-Order Rules with Tertius", Machine Learning, S.42, 2001, s.61-95

3 UYGULAMA

3.1 Uygulamanın Amacı

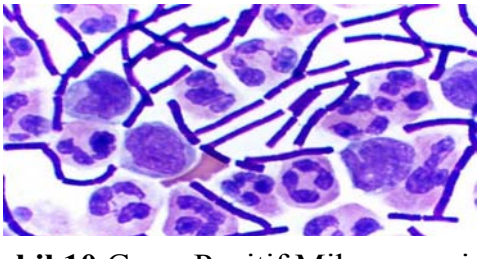
Veri madenciliği çalışmalarında bilgisayar teknolojilerinin gelişmişliğinin önemi göz ardı edilemez. Günümüzde VM uygulamaları için geliştirilmiş programlar bulunmaktadır.

Çalışmanın bu bölümünde, VM için geliştirilmiş bulunan WEKA ve YALE programları örnek bir veri tabanı üzerinde uygulanarak programlar arasında karşılaştırma yapılacaktır.

3.2 Tanımlar

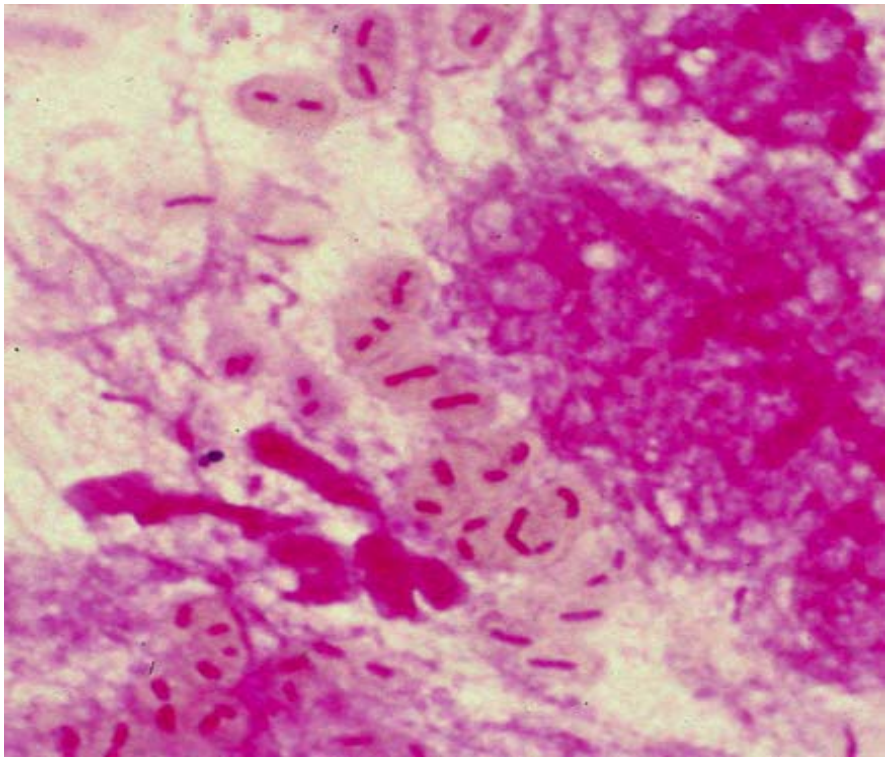
3.2.1 Gram Boyama

Gram boyama mikroorganizmaları, şekillerine, boyutlarına, hücre morfolojilerine ve gram reaksiyonlarına göre sınıflamak için kullanılır. Test 1884'te Christian Gram tarafından geliştirilmiştir. Mikroorganizmalar, hücre duvarı, bileşim ve şekil farklılıklarına göre gram pozitif veya gram negatif olarak boyanırlar. Yöntem bakterilerin tanımlanmasında kullanılan ilk testtir. Belirli aşamalardan geçtikten sonra (kristal viyole, iodin, alkol ve sulu fuksin) mikroskop altında, gram pozitif mikroorganizmalar koyu mor, gram negatif mikroorganizmalar ise pembe-kırmızı renkte görünür.



Şekil 10 Gram Pozitif Mikroorganizma

Kaynak:http://en.wikipedia.org/wiki/ Image:Gram_Stain_Anthrax. jpg



Şekil 11 Gram Negatif Mikroorganizma

Kaynak:http://www.uphs.upenn.edu/bugdrug/antibiotic_manual/Gramstains/small/tocframeset1.htm

3.2.2 Gram Negatif Basil (GNB)

Bakteriler ışık veya elektro mikroskobunda incelendiklerinde yuvarlak, çomakcık veya sarmal şeklinde görülürler. [87] GNB'ler çomakçık şeklinde ve gram yöntemi ile boyandığında pembe-kırmızı renkte olan mikroorganizmalardır.

[87] Ayşe Wilke Topçu, Güner Söyletir ve Mehmet Doğanay, İnfeksiyon Hastalıkları ve Mikrobiyolojisi, İstanbul, Nobel , 2002, s.13

3.2.3 Epidemiyoloji

Epidemiyoloji bir toplumda sağlık ve hastalık göstergelerine göre o toplumun sağlıkla ilgili durumunu ve hastalıkların dağımını inceleyen bir bilim olarak tanımlanabilir. Sağlık ve hastalık ilişkileri belirlenirken kişi, yer, zaman özelliklerinden faydanılır.

3.2.4 GNB'lerle Oluşan Hastalıklar

- *Escherichia coli* (ishal, idrar yolu enfeksiyonu, yeni doğanlarda menenjit, hastane enfeksiyonları)
- *Shigella* türleri (dizanteri, tifo)
- *Campylobacter jejuni* (bağırsak iltihabı)
- *Pseudomonas aeruginosa* (fırsatçı enfeksiyonlar: idrar yolu, solunum sistemi, yanık, yara, deri, göz ve kulak enfeksiyonları)
- *Citrobacter* türleri, *Enterobacter* türleri, *Klebsiella* türleri, *Proteus*, *Serratia* (Hastane kaynaklı sounum sistemi, üriner sistem enfeksiyonları, mikroorganizmanın kana geçmesi ile oluşan bakteriyemi, özellikle hastane enfeksiyonları ve düşkün hastalarda)
- Acinetobacter (Hastane kaynaklı genitoüriner sistem, solunum sistemi enfeksiyonları, yara, yumuşak doku enfeksiyonları ve bakteriyemi)
- *Stenotrophomonas maltophilia* (Bakteriyemi, yara enfeksiyonları, zatürre, idrar yolu enfeksiyonları) [88]

3.3 Problemin Tanımlanması

Enfeksiyon hastalıkları günümüzde insan ölümlerinin en önemli nedenlerinden biridir. Yakın gelecektede enfeksiyon hastalıklarının önemini sürdüreceği öngörülmektedir. Bu çalışmada, gram negatif basillerle oluşan hastalıklara ait elde edilen veriler ışığında, epidemiyolojik bilgi girdisi yapılması amaçlanmaktadır. Bu nedenle ilgili veri tabanındaki örüntüler saptanmaya çalışılacaktır.

[88] Betty A. Forbes, Daniel F. Sahm ve Alice S. Weissfeld, Diagnostic Microbiology, Missouri, Mosby, 2002, s.365-389

3.4 Verinin Toplanması

Veri GNB'lere ait bir veri tabanından elde edilmiştir. Veri tabanında 1968 kayıt bulunmaktadır. Veriler, verinin kolay bir şekilde incelenebilmesi için tek bir tabloda toplanmıştır. Veri tabanın başlangıç durumuna ait özellikleri Tablo-5'te sunulmuştur.

Tablo-5 Veri Tabanı Başlangıç Durumuna Ait Özellikler

VERİ ALANININ ADI	ÖZELLİKLERİ
Sıra No	Sayısal
Adı Soyadı	Metin
Protokol No	Sayısal
Örnek Türü	Metin (İdrar, Kan, Sürüntü vb.)
Örneğin Geldiği Servis	Metin (Ortopedi, Genel Cerrahi, Dahiliye vb.)
Örneğin Geliş Tarihi	Tarih
Hastanın Yaşı	Sayısal (3, 45, 65 vb.)
Barkod No	Sayısal
Protokol No	Sayısal
Saptanan Mikroorganizma Türü	Metin (Escherichia coli, Campylobacter jejuni vb)

3.5 Verinin Temizlenmesi ve Dönüştürülmesi

Veri temizleme aşamasında, veri tabanından konu ile ilgisi olmayan bilgi ayıklanmış ve eksik veriler tamamlanmıştır. Bu kapsamda veri madenciliği çalışmasında yarar sağlamayacağı değerlendirilen sıra no, barkod no, protokol no alanları veri tabanından çıkarılmıştır.

Veri dönüştürülmesi aşamasında ise amaç kaynak verinin veri tipleri ve değerleri olarak değişik biçimlerde değiştirilmesidir. Bu nedenle adı soyadı alanı cinsiyet olarak değiştirilmiş ve hastaların cinsiyetleri yazılmış, hastanın yaşı alanı belli yaş aralıklarını belirten metin olarak değiştirilmiş, tarih alanındaki gün ve ay bilgileri metin olarak ayrıca girilmiş ve yıl bilgisi tüm değerler için aynı olduğundan çıkarılmıştır. Ay bilgileri daha sonra mevsim değerlerine dönüştürülmüştür.

Gün ve mevsim değerlerinin ayrı olarak veri tabanına eklenmesi ile gün ve mevsimlerle diğer alanlar arasındaki örüntülerin tespit edilmesi hedeflenmektedir. Aynı şekilde adı soyadı bilgisinin cinsiyet bilgisi ile değiştirilmesi ile cinsiyet ile diğer alanlar

arasındaki ilişkilerin incelenmesi planlamıştır. Aslında cinsiyet bilgisi bir dönüştürme işleminden çok yeni bir nitelik ekleme olarak değerlendirilebilir.

Veri temizlenmesi ve dönüştürülmesinden sonraki veri tabanının son hali Tablo-6'da sunulmuştur. Apriori algoritması kullanan programlar verinin metin şeklinde olmasının gerektirdiği için sayısal veriler metne çevrilmiştir.

Tablo-6 Verinin Temizlenmesi ve Dönüştürülmesinden Sonra Veri Tabanının Son Hali

VERİ ALANININ ADI	ÖZELLİKLERİ
Cinsiyet	Metin (Erkek,Bayan)
Örnek Türü	Metin (İdrar, Kan, Sürüntü vb.)
Örneğin Geldiği Servis	Metin (Ortopedi, Genel Cerrahi, Dahiliye vb.)
Örneğin Geliş Günü	Metin (Pazartesi, Salı vb.)
Örneğin Geldiği Mevsim	Metin (İlkbahar, yaz, sonbahar, kış)
Hastanın Yaş Aralığı	Metin (bir_onbeş, onaltı_otuz vb.)
Saptanan Mikroorganizma Türü	Metin (Escherichia coli, Campylobacter jejuni vb)

3.6 Modelin Kurulması

Veri madenciliği yapılırken veri niteliklerine göre gruplara ayrılarak, bu nitelikler arasındaki ilişkiler incelenmiştir. Niteliklerine göre verinin bölünmesi işlemi sonucu ortaya çıkan gruplar Tablo-7'de sunulmuştur.

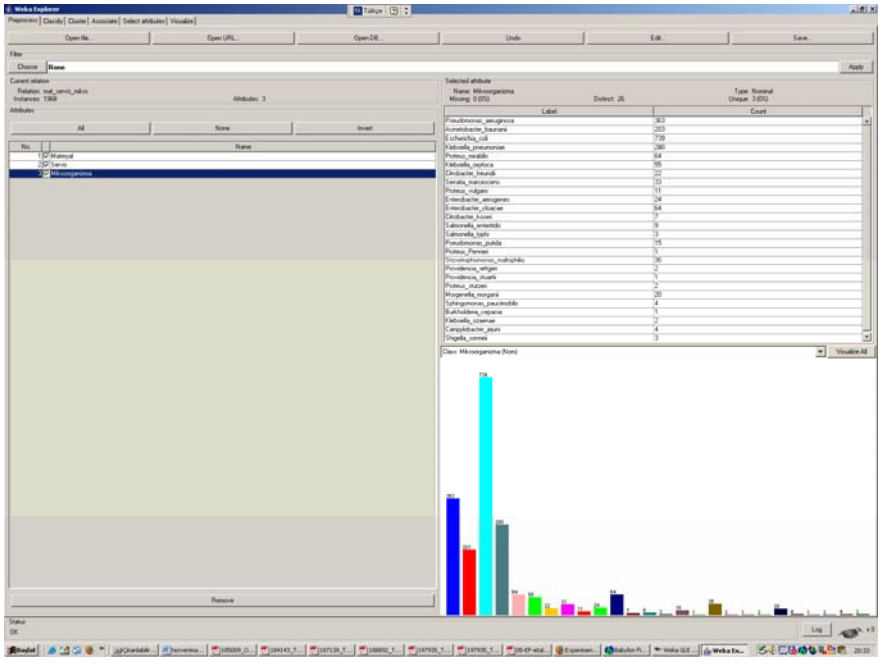
Tablo-7 Veri İnceleme Grupları

VERİ ALANI ADI	OLUŞTURULAN GRUPLAR							
	Grup 1	Grup 2	Grup 3	Grup 4	Grup 5	Grup 6	Grup 7	Grup 8
Cinsiyet		X						
Örnek Türü	X	X			X		X	X
Örneğin Geldiği Servis	X	X		X	X		X	
Örneğin Geliş Günü						X		X
Örneğin Geldiği Mevsim			X			X		
Hastanın Yaş Aralığı			X	X	X			
Saptanan Mikroorganizma Türü	X		X	X		X		

Nitelikler arasındaki ilişkileri incelemek için Birliktelik Kuralları modeli kullanı caktır.

3.7 Modelin Değerlendirilmesi

Modelin değerlendirilmesi için WEKA ve YALE programları kullanılacaktır. Birliktelik kurallarının incelenmesinde her iki programda bulunan Apriori, Tahminci Apriori ve Tertius algoritmaları ile yapılan veri madenciliği sonuçları incelenmiştir.



Şekil-12 Weka Explorer Arayüzü

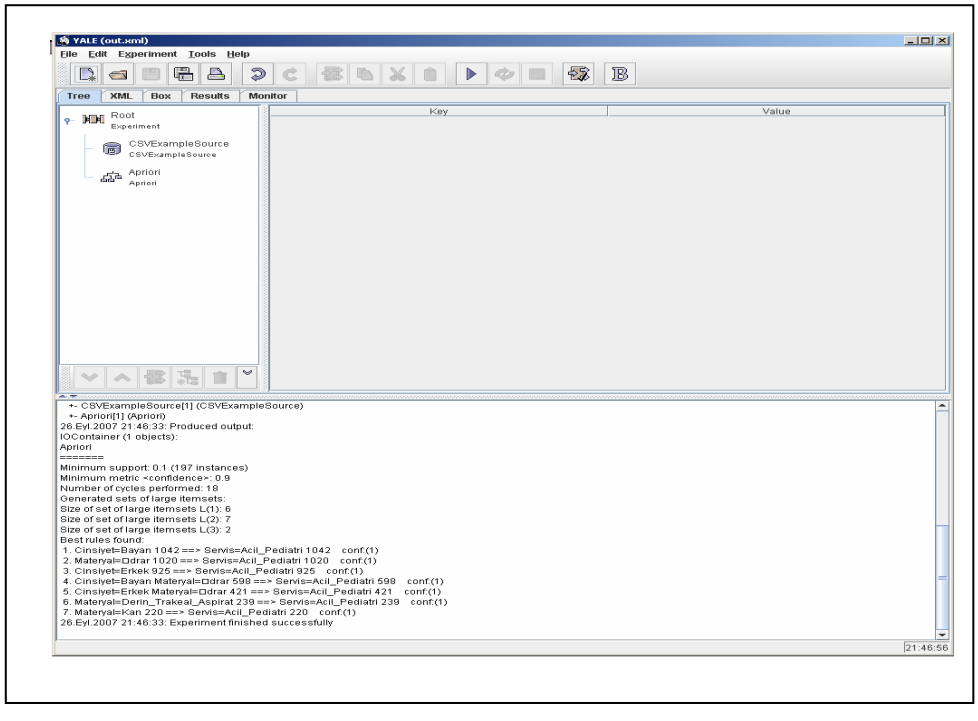
3.7.1 Waikato Environment for Knowledge Analysis (WEKA) Programı

Weka makina öğrenme algoritmalarının ve veri ön işleme araçlarının bir araya getirildiği, akademik çevrelerde sıklıkla kullanılan, açık kaynak kodlu bir veri madenciliği programıdır. Yeni Zelanda'nın Waikato Üniversitesinde geliştirilmiş ücretsiz bir yazılımdır. Yazılım, Java yazılım dili ile geliştirilmiştir. Büyük veya dağıtık veri tabanlarında kullanılabilir. Weka ile verinin hazırlanması, sınıflama, kümeleme, birliktelik analizi, nitelik değerlerinin seçilmesi yapılabilmektedir. Arff, csv, c45

biçimindeki dosyalar kullanılabilmektedir.[89] Weka programı arayüzü Şekil-12'de görülmektedir.

3.7.2 Yet Another Learning Environment (Yale) Programı

Yale makina öğrenmesi uygulamaları için kullanılan bir yazılımdır. Birimsel yapısı sayesinde karmaşık problemlerin çözülmesinde kullanılmaktadır. Arff, csv, c45, bibtex, excel, spss, dat formatındaki verilerle çalışılabilir. Akademik çevrelerde kullanılmaktadır. Ücretsiz bir yazılımdır. XML dilini kullanmaktadır. Almanya'nın Dortmund Üniversitesi'nde geliştirilmiştir. [90] Yale programı arayüzü Şekil-13'de görülmektedir.



Şekil-13 Yale Programı Arayüzü

[89] David Scuse ve Peter Reutemann, WEKA Experimenter Tutorial for Version 3-4, Yeni Zelanda Waikato Üniversitesi, 2007, s.1-6

[90] Ingo Mierswa. vd. "Yale: Rapid Prototyping for Complex Data Mining Tasks", In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006)

3.7.3 Weka ve Yale Programları Apriori Algoritması Uygulama Sonuçları

3.7.3.1 Cinsiyet, Örnek Türü ve Örneğin Geldiği Servis Verisi

Weka ve Yale programları programları minimum destek değeri 0.3, minimum güvenilirlik değeri 0.4 ve diğer tüm değerler aynı olmak koşulu ile çalıştırıldığında elde edilen sonuçlar sırasıyla YALE için Tablo-10'da, WEKA için Tablo-11'de sunulmuştur.

Tablo-8 YALE Programı ile Cinsiyet, Örnek Türü ve Örneğin Geldiği Servis Verisinde Apriori Algoritması Uygulama Sonuçları

```
YALE
Apriori
=======
Minimum support: 0.3 (590 instances)
Minimum metric <confidence>: 0.4
Number of cycles performed: 14
Generated sets of large itemsets:
Size of set of large itemsets L(1): 4
Size of set of large itemsets L(2): 4
Size of set of large itemsets L(3): 1
Best rules found:
1. Cinsiyet=Bayan 1042 ==> Servis=Acil_Pediatri 1042 conf:(1)
2. Materyal=idrar 1020 ==> Servis=Acil_Pediatri 1020 conf:(1)
3. Cinsiyet=Erkek 925 ==> Servis=Acil_Pediatri 925 conf:(1)
4. Cinsiyet=Bayan Materyal=idrar 598 ==> Servis=Acil_Pediatri 598 conf:(1)
5. Materyal=idrar 1020 ==> Cinsiyet=Bayan 598 conf:(0.59)
6. Materyal=idrar Servis=Acil_Pediatri 1020 ==> Cinsiyet=Bayan 598 conf:(0.59)
7. Materyal=idrar 1020 ==> Cinsiyet=Bayan Servis=Acil_Pediatri 598 conf:(0.59)
8. Cinsiyet=Bayan 1042 ==> Materyal=idrar 598 conf:(0.57)
9. Cinsiyet=Bayan Servis=Acil_Pediatri 1042 ==> Materyal=idrar 598 conf:(0.57)
10. Cinsiyet=Bayan 1042 ==> Materyal=idrar Servis=Acil_Pediatri 598 conf:(0.57)
```

Tablo-9 WEKA Programı ile Cinsiyet, Örnek Türü ve Örneğin Geldiği Servis Verisinde Apriori Algoritması Uygulama Sonuçları

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.4 -D
0.05 -U 1.0 -M 0.3 -S -1.0
Relation:     cinsiyet_mat_servis
Instances:    1968
Attributes:   3
              Cinsiyet
              Materyal
              Servis
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.3 (590 instances)
Minimum metric <confidence>: 0.4
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Size of set of large itemsets L(2): 1

Best rules found:

 1. Materyal=İdrar 1020 ==> Cinsiyet=Bayan 598    conf:(0.59)
 2. Cinsiyet=Bayan 1042 ==> Materyal=İdrar 598    conf:(0.57)
```

Tablolardan görüleceği gibi iki yazılımın verdiği sonuçlar arasında farklar bulunmaktadır. YALE yazılımı tarafından 10 adet kural üretilirken, WEKA yazılımı tarafından 2 adet kural üretilmiştir.

YALE yazılımı tarafından üretilen kurallardan 8 adedi geçerli görünmemektedir. Güven formülünün (c (X → Y) = σ(X U Y)/σ(X) olduğunda, YALE'de bulunan ilk kural için bu değer σ(Bayan U Acil Pediatri)/σ(Bayan) olmaktadır. Bayanların sayısı 1052, Bayan ve Acil Pediatri olanların sayısı ise 28'dir. Buradan 28/1052=0.026 değerine ulaşılır. Bu değer program tarafından bulunan 1 değerinden oldukça farklıdır. Tüm Acil Pediatriden gelen örnek sayısı 40 adettir. Bu değerler Acil Pediatri geçen tüm kuralların güvenilirlik değerlerinin YALE yazılımı tarafından bulunan değerlerden oldukça farklı olduğunu göstermektedir. Bu yüzden 5 ve 8'inci kurallar hariç diğer kurallar geçerli görünmemektedir.

WEKA yazılımı tarafından üretilen kurallar güvenilirliği az olmakla beraber geçerlidir. Bu kurallar YALE yazılımı tarafından üretilen 5 ve 8'inci kurallar ile aynı kurallardır. Kural aynı zamanda anlamlıdır. Bayanların idrar yolu kanalı daha kısa olduğu için idrar yolu enfeksiyonlarına yakalanma ihtimalleri daha fazladır.

3.7.3.2 Hastanın Yaş Aralığı ve Örneğin Geldiği Servis Verisi

Bu örnekte amaç bilinen bir gerçekten yola çıkarak algoritmaları test etmektir. Bu örnekte beklenen sonuç pediatri (çocuk) servisinden gelen örneklere ait hasta yaş aralığının ilk aralık olan bir_onbeş yaş aralığında bulunmasıdır.

Weka ve Yale programları minimum destek değeri 0.1, minimum güvenilirlik değeri 0.9 ve diğer tüm değerler aynı olmak koşulu ile çalıştırıldığında elde edilen sonuçlar sırasıyla YALE için Tablo-10'da, WEKA için Tablo-11'de sunulmuştur.

Tablo-12'de görülen YALE yazılımı ile üretilen kurallar anlamlı olmaktan uzaktır. Birinci kural incelenecek olursa pediatri servisinden araştırılması istenen örneğin alındığı altmışbir_yetmişbeş yaş arası sadece bir hasta bulunmaktadır. Bu örneğe ait güvenilirliğin 1.0 çıkması bir hata olduğunu göstermektedir.

Tablo-10 YALE Programı ile Hastanın Yaş Aralığı ve Örneğin Geldiği Servis Verisinde Apriori Algoritması Uygulama Sonuçları

```
7.Eyl.2007 10:30:20: Experiment:
  Root[0] (Experiment)
  +- CSVExampleSource[0] (CSVExampleSource)
  +- Apriori[0] (Apriori)
27.Eyl.2007 10:30:22: Experiment finished after 1 seconds
27.Eyl.2007 10:30:22: Experiment:
  Root[1] (Experiment)
  +- CSVExampleSource[1] (CSVExampleSource)
  +- Apriori[1] (Apriori)
27.Eyl.2007 10:30:22: Produced output:
IOContainer (1 objects):
Apriori
=======
Minimum support: 0.1 (197 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 5
Size of set of large itemsets L(2): 4
Best rules found:
1. Servis=Pediatri 346 ==> yas=altmisbir_yetmisbes 346 conf:(1)
2. Servis=Yoğun_Bakım_Ünitesi 332 ==> yas=altmisbir_yetmisbes 332 conf:(1)
3. Servis=Acil_Yetişkin 236 ==> yas=altmisbir_yetmisbes 236 conf:(1)
4. Servis=Dahiliye 207 ==> yas=altmisbir_yetmisbes 207 conf:(1)
```

Tablo-11 WEKA Programı ile Hastanın Yaş Aralığı ve Örneğin Geldiği Servis Verisinde Apriori Algoritması Uygulama Sonuçları

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Relation:     yas_servis
Instances:    1968
Attributes:   2
              Servis
              yas
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (197 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 1

Best rules found:

 1. Servis=Pediatri 346 ==> yas=bir_onbes 344    conf:(0.99)

12:53:27: Started weka.associations.Apriori
12:53:28: Finished weka.associations.Apriori
```

YALE'de ikinci örnekte bulunan kurala göre, güvenilirlik katsayısı 1.0 olduğu için tüm yoğun bakım ünitesi hastalarının altmışbir_yetmişbeş yaş arası olması gerekir. Yoğun bakım hastalarının sayısı 332'dir. Bunlardan 138'i altmışbir_yetmişbeş yaş aralığındadır. Bu durumda güvenilirlik katsayısının 0.42 olması gerekir.

Tablo-12 WEKA Programı ile Hastanın Yaş Aralığı ve Örneğin Geldiği Servis Verisinde Apriori Algoritması Uygulama Sonuçları (minimum destek değeri 0.05, minimum güvenilirlik değeri 0.4)

```
=== Run information ===

Scheme:     weka.associations.Apriori -N 10 -T 0 -C 0.4 -D 0.05 -U 1.0 -M 0.05 -S -1.0
Relation:   yas_servis
Instances:  1968
Attributes: 2
            Servis
            yas
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.05 (98 instances)
Minimum metric <confidence>: 0.4
Number of cycles performed: 19

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 2

Best rules found:

 1. Servis=Pediatri 346 ==> yas=bir_onbes 344    conf:(0.99)
 2. yas=bir_onbes 645 ==> Servis=Pediatri 344    conf:(0.53)
 3. Servis=Yoğun_Bakım_Ünitesi 332 ==> yas=altmisbir_yetmisbes 138    conf:(0.42)

13:13:01: Started weka.associations.Apriori
13:13:02: Finished weka.associations.Apriori
```

WEKA'da aynı değerlerle program çalıştırıldığında pediatri servisinden örnek gönderilen hastaların yaş aralığının 0.99 güvenilirlik katsayısı ile bir_onbeş arası olduğu görülmektedir. Bu zaten bizim beklediğimiz bir sonuçtur. WEKA programı aynı veri ile, minimum destek değeri 0.05, minimum güvenilirlik değeri 0.4 olarak çalıştırıldığında Tablo-13'deki kurallar elde edilmektedir. Bu kurallardan Servis= yoğun bakım ünitesi ➔ yaş=altmışbir_yetmişbeş kuralı 0.42 güvenilirlikle belirlenmiştir.

3.7.4 Weka ve Yale Programları Tahminci Apriori Algoritması Uygulama Sonuçları

Yale programı Tablo-13'de görüleceği gibi veri setinde herhangi bir hata olmamasına rağmen 1968 kayıttan oluşan veri kümesinde yetersiz hafıza hatası

vermektedir. Çalışmalar Mobile AMD Sempron 3500+791MHz işlemci ve 448 MB hafızaya sahip bilgisayarda yapılmıştır. Yetersiz hafıza hatası alınması üzerine Pentium 4 işlemci ve 1GB hafızaya sahip bilgisayarda denemeler tekrarlanmış fakat tekrar yetersiz hafıza hatası alınmıştır. Bunun üzerine 1968 kayıtlık veri 100 kayıta indirilmiş ve denemeler Pentium 4 işlemci ve 1GB hafızaya sahip bilgisayarda tekrar edilmiştir. Yetersiz hafıza hatası tekrarlanmıştır.

Tablo 13 YALE Programı ile Örnek Türü,Örneğin Geldiği Servis ve Saptanan Mikroorganizma Verisinde Tahminci Apriori Algoritmasının 25 Kural için Uygulama Sonuçları

```
26.Eyl.2007 17:02:19: Initialising experiment
26.Eyl.2007 17:02:19: [Warning] No filename given for result file, using stdout for logging results!
26.Eyl.2007 17:02:19: Creating temp directory C:\Program Files\YALE\yale-3.4\out.xml.tmp
26.Eyl.2007 17:02:19: Checking properties...
26.Eyl.2007 17:02:19: Properties are ok.
26.Eyl.2007 17:02:19: Checking experimental setup...
26.Eyl.2007 17:02:19: Inner operators are ok.
26.Eyl.2007 17:02:19: Checking i/o classes...
26.Eyl.2007 17:02:19: i/o classes are ok. Experiment output: WekaAssociator.
26.Eyl.2007 17:02:19: Experiment ok.
26.Eyl.2007 17:02:19: Experiment initialised
26.Eyl.2007 17:02:19: Experiment starts
26.Eyl.2007 17:02:19: Experiment:
  Root[0] (Experiment)
  +- CSVExampleSource[0] (CSVExampleSource)
  +- PredictiveApriori[0] (PredictiveApriori)
26.Eyl.2007 17:03:04: [Exception] OutOfMemoryError occured in 1st application of Root
(Experiment)
26.Eyl.2007 17:03:04: [Exception] Java heap space
java.lang.OutOfMemoryError: Java heap space
        Root[1] (Experiment)
        +- CSVExampleSource[1] (CSVExampleSource)
 here ==> +- PredictiveApriori[1] (PredictiveApriori)
26.Eyl.2007 17:03:04: [Fatal] Experiment failed!
```

Weka programında tahminci appriori algoritması örnek türü,örneğin geldiği servis ve saptanan mikroorganizma verisinde 25 kural için çalıştırılmıştır. Elde edilen sonuçlar Tablo-'te sunulmuştur. Elde edilen sonuçların destek değerlerinin yüksek olması için 0.05 değere yakın veya üzerindeki kurallar kırmızı ile işaretlenmiştir. Konunun uzmanı ile yapılan değerlendirmeler sonucunda, bulunan kuralların hastanedeki epidemiyolojik verilere uygun olduğu tespit edilmiştir. Tablo-14'te 18 numaralı aşağıda sunulmuştur.

- Materyal=Derin_Trakeal_Aspirat Mikroorganizma=Acinetobacter_baunanii 105 ==> Servis=Yoğun_Bakım_Ünitesi 89 acc:(0.84111)

Derin Trakeal Aspirat soluk borusundan alınan örnektir. Acinetobacter_baunanii özellikle hastane kaynaklı enfeksiyonlara neden olan bir mikroorganizmadır ve yoğun bakım ünitelerindeki bağışıklık sistemi baskılanmış hastalarda enfeksiyonlara yol açmaktadır. WEKA Tahminci Apriori tarafından ortaya çıkarılan kural, geçerli bir kuraldır.

Tablo-14 WEKA Programı ile Örnek Türü,Örneğin Geldiği Servis ve Saptanan Mikroorganizma Verisinde Tahminci Apriori Algoritması Uygulama Sonuçları

PredictiveApriori (WEKA)
==========

Best rules found:

1. Servis=Laboratuvar Mikroorganizma=Escherichia_coli 99 ==> Materyal=İdrar 99 acc:(0.99494)
2. Servis=Acil_Pediatri 40 ==> Mikroorganizma=Escherichia_coli 40 acc:(0.99464)
3. Servis=Pediatri Mikroorganizma=Proteus_mirabilis 16 ==> Materyal=İdrar 16 acc:(0.99298)
4. Servis=Kadın_Doğum Mikroorganizma=Escherichia_coli 14 ==> Materyal=İdrar 14 acc:(0.99239)
5. Servis=Nefroloji 12 ==> Materyal=İdrar Mikroorganizma=Escherichia_coli 12 acc:(0.99142)
6. Servis=Fizik_Tedavi Mikroorganizma=Escherichia_coli 21 ==> Materyal=İdrar 20 acc:(0.97553)
7. Servis=Laboratuvar Mikroorganizma=Klebsiella_pneumoniae 18 ==> Materyal=İdrar 17 acc:(0.96646)
8. Servis=Acil_Pediatri 40 ==> Materyal=İdrar Mikroorganizma=Escherichia_coli 38 acc:(0.9623)
9. Materyal=Derin_Trakeal_Aspirat Mikroorganizma=Proteus_mirabilis 5 ==> Servis=Yoğun_Bakım_Ünitesi 5 acc:(0.96149)
10. Materyal=Nefrostomi_Sıvısı 4 ==> Servis=Üroloji 4 acc:(0.93949)
11. Mikroorganizma=Campylobacter_jejuni 4 ==> Materyal=Dışkı 4 acc:(0.93949)
12. Servis=Fizik_Tedavi 28 ==> Materyal=İdrar 26 acc:(0.92558)
13. Servis=Üroloji Mikroorganizma=Klebsiella_pneumoniae 25 ==> Materyal=İdrar 23 acc:(0.9105)
14. Servis=Pediatri Mikroorganizma=Escherichia_coli 97 ==> Materyal=İdrar 87 acc:(0.88889)
15. Servis=Kadın_Doğum 19 ==> Materyal=İdrar 17 acc:(0.86291)
16. Servis=Laboratuvar 176 ==> Materyal=İdrar 151 acc:(0.85393)
17. Servis=Üroloji 157 ==> Materyal=İdrar 134 acc:(0.84906)
18. Materyal=Derin_Trakeal_Aspirat Mikroorganizma=Acinetobacter_baunanii 105 ==> Servis=Yoğun_Bakım_Ünitesi 89 acc:(0.84111)
19. Materyal=Kemik 2 ==> Mikroorganizma=Escherichia_coli 2 acc:(0.83649)
20. Materyal=Göbek_Sıvısı 2 ==> Mikroorganizma=Escherichia_coli 2 acc:(0.83649)
21. Materyal=Nazofarengeal_Aspirat 34 ==> Servis=Pediatri 29 acc:(0.82986)
22. Materyal=İdrar Servis=Acil_Yetişkin 179 ==> Mikroorganizma=Escherichia_coli 147 acc:(0.81768)
23. Servis=Acil_Yetişkin Mikroorganizma=Escherichia_coli 183 ==> Materyal=İdrar 147 acc:(0.79999)
24. Servis=Acil_Yetişkin 236 ==> Mikroorganizma=Escherichia_coli 183 acc:(0.77307)
25. Servis=Pediatri_Acil Mikroorganizma=Klebsiella_pneumoniae 31 ==> Materyal=İdrar 25 acc:(0.77145)

20:19:05: Started weka.associations.PredictiveApriori
20:19:17: Finished weka.associations.PredictiveApriori

Tablo-14'de 17'nci kuralı inceleyecek olursak;

- Servis=Üroloji 157 ==> Materyal=İdrar 134 acc:(0.84906)

Üroloji idrar yolları hastalıkları ile ilgilenen bilim dalıdır. Bu nedenle, bu servisten gelen örnek türü idrardır.

3.7.5 Weka ve Yale Programları Tertius Algoritması Uygulama Sonuçları

Weka ve Yale programları programları tüm değerler aynı olmak koşulu ile 10 kural bulunması için çalıştırıldığında, tamamen aynı kurallar elde edilmiştir. Fakat YALE programı Tertius algoritmasının uygun çalışmasını sağlamak amacıyla, veri tabanına tüm değerleri aynı olan, metin tipinde yeni bir alan eklenmiştir. Elde edilen sonuçlar aynı olmasına rağmen YALE ve WEKA programları arasında performans bakımından farklar bulunmaktadır. Uygulamaya ait sonuçlar sırasıyla Tablo-15'de ve Tablo-16'da sunulmuştur.

Tablo-15 ve Tablo-16 incelendiğinde elde edilen tüm kuralların aynı olduğu görülmektedir. Tertius algoritması ile elde edilen kurallar incelendiğinde bu kuralların gerçeğe uygun kurallar olduğu görülmektedir. Örneğin 1 numaralı kural ele alınacak olursa servislerden gönderilen 1968 örnek türünden 1020 adedi idrardır. Saptanan 1968 mikroorganizmadan 739 adedi Escherichia_coli tipindedir. Örneklerin geldiği servislerden sadece 157 adedi üroloji servisinden gelmesine rağmen, idrar ve üroloji ile üroloji ve Escherichia_coli arasında ilişki yakın bir ilişki bulunmaktadır.

İşlemin tamamlanma zamanı bakımından incelendiğinde WEKA programının YALE programından yaklaşık 4 kat daha yavaş çalıştığı tespit edilmiştir. Fakat bunda YALE programının son alanıda hesaplamaya dahil etmesi için veri tabanına metin tipinde eklenen yeni alanın etkili olabileceği değerlendirilmektedir.

Tertius algoritması ile bulunan kurallar anlamlıdır. WEKA ve YALE tamamen aynı sonuçları bulduğu için veri tabanının tüm alanlarını kapsayacak şekilde yapılan çalışma sadece WEKA programı ile yapılmıştır. 1968 kayıt ve 7 ayrı veri tabanı için AMD Sempron 3500+791MHz işlemci ve 448 MB hafızaya sahip bilgisayarda yapılan çalışmada hesaplama süresi 20 dakika gösterilmesine rağmen Tablo-17'ye eklenen değerlerden kolaylıkla görülebileceği gibi gerçek hesaplama süresi 1 saat 20 dakika olarak gerçekleşmiştir. Alan sayısının artması hesaplama süresinin katlanarak artmasına yol açmaktadır.

Tablo-15 YALE Programı ile Örnek Türü,Örneğin Geldiği Servis ve Saptanan Mikroorganizma Verisinde Tertius Algoritması için Uygulama Sonuçları

```
Tertius
Tertius(YALE)
=======
1. /* 0,445027 0,194106 */ Materyal = idrar ==> Servis = Üroloji or Mikroorganizma = Escherichia_coli
2. /* 0,433194 0,065041 */ Mikroorganizma = Escherichia_coli ==> Servis = Acil_Yetişkin or Materyal = idrar
3. /* 0,425272 0,198171 */ Materyal = idrar ==> Servis = Pediatri_Acil or Mikroorganizma = Escherichia_coli
4. /* 0,419111 0,199695 */ Materyal = idrar ==> Servis = Laboratuvar or Mikroorganizma = Escherichia_coli
5. /* 0,407443 0,077236 */ Mikroorganizma = Escherichia_coli ==> Servis = Üroloji or Materyal = idrar
6. /* 0,404584 0,209858 */ Materyal = idrar ==> Servis = Acil_Yetiikin or Mikroorganizma = Escherichia_coli
7. /* 0,404386 0,222561 */ Materyal = idrar ==> Servis = Radyoloji or Mikroorganizma = Escherichia_coli
8. /* 0,402852 0,211382 */ Materyal = idrar ==> Servis = Pediatrik_Cerrahi or Mikroorganizma = Escherichia_coli
9. /* 0,402316 0,223069 */ Materyal = idrar ==> Servis = Fizik_Tedavi or Mikroorganizma = Escherichia_coli
10. /* 0,402263 0,080793 */ Mikroorganizma = Escherichia_coli ==> Servis = Kemik_iliği_Transplant or Materyal =
idrar
Number of hypotheses considered: 36061
Number of hypotheses explored: 16026
Time: 04 min 50 s 538 ms
```

Tablo-16 WEKA Programı ile Örnek Türü,Örneğin Geldiği Servis ve Saptanan Mikroorganizma Verisinde Tertius Algoritması için Uygulama Sonuçları

```
=== Run information ===

Scheme:      weka.associations.Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -p -P 0
Relation:    mat_servis_mikro
Instances:   1968
Attributes:  3
             Materyal
             Servis
             Mikroorganizma
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0,445027 0,194106 */ Materyal = İdrar ==> Servis = Üroloji or Mikroorganizma = Escherichia_coli
 2. /* 0,433194 0,065041 */ Mikroorganizma = Escherichia_coli ==> Materyal = İdrar or Servis = Acil_Yetişkin
 3. /* 0,425272 0,198171 */ Materyal = İdrar ==> Servis = Pediatri_Acil or Mikroorganizma = Escherichia_coli
 4. /* 0,419111 0,199695 */ Materyal = İdrar ==> Servis = Laboratuvar or Mikroorganizma = Escherichia_coli
 5. /* 0,407443 0,077236 */ Mikroorganizma = Escherichia_coli ==> Materyal = İdrar or Servis = Üroloji
 6. /* 0,404584 0,209858 */ Materyal = İdrar ==> Servis = Acil_Yetişkin or Mikroorganizma = Escherichia_coli
 7. /* 0,404386 0,222561 */ Materyal = İdrar ==> Servis = Radyoloji or Mikroorganizma = Escherichia_coli
 8. /* 0,402852 0,211382 */ Materyal = İdrar ==> Servis = Pediatrik_Cerrahi or Mikroorganizma = Escherichia_coli
 9. /* 0,402316 0,223069 */ Materyal = İdrar ==> Servis = Fizik_Tedavi or Mikroorganizma = Escherichia_coli
10. /* 0,402263 0,080793 */ Mikroorganizma = Escherichia_coli ==> Materyal = İdrar or Servis =
Kemik_İliği_Transplant

Number of hypotheses considered: 18506
Number of hypotheses explored: 10031
Time: 00 min 48 s 881 ms
```

Tablo-17 Weka Programı ile Tüm Veri Alanlarını Kapsayan Uygulama Sonuçları

```
=== Run information ===

Scheme:       weka.associations.Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -p -P 0
Relation:     veri27eylülson
Instances:    1968
Attributes:   7
              Cinsiyet
              Materyal
              Servis
              Mikroorganizma
              Gun
              Mevsim
              yas
=== Associator model (full training set) ===


Tertius
=======

 1. /* 0,509658 0,003557 */ Servis = Pediatri ==> yas = bir_onbes
 2. /* 0,486054 0,003049 */ Servis = Pediatri ==> Materyal = Balgam or yas = bir_onbes
 3. /* 0,472661 0,144817 */ yas = bir_onbes ==> Materyal = Dışkı or Servis = Pediatri or Mikroorganizma = Morgenella_morganii
 4. /* 0,471916 0,146341 */ yas = bir_onbes ==> Materyal = Dışkı or Servis = Pediatri or Mikroorganizma = Proteus_vulgaris
 5. /* 0,471222 0,146850 */ yas = bir_onbes ==> Materyal = Endotrakeal_Aspirat or Servis = Pediatri or Mikroorganizma = Morgenella_morganii
 6. /* 0,470900 0,140752 */ yas = bir_onbes ==> Materyal = Dışkı or Servis = Pediatri or Mikroorganizma = Enterobacter_cloacae
 7. /* 0,470552 0,146850 */ yas = bir_onbes ==> Materyal = Nazofarengeal_Aspirat or Servis = Pediatri or Mikroorganizma = Morgenella_morganii
 8. /* 0,470485 0,148374 */ yas = bir_onbes ==> Materyal = Endotrakeal_Aspirat or Servis = Pediatri or Mikroorganizma = Proteus_vulgaris
 9. /* 0,469814 0,148374 */ yas = bir_onbes ==> Materyal = Nazofarengeal_Aspirat or Servis = Pediatri or Mikroorganizma = Proteus_vulgaris
10. /* 0,469435 0,142785 */ yas = bir_onbes ==> Materyal = Endotrakeal_Aspirat or Servis = Pediatri or Mikroorganizma = Enterobacter_cloacae

Number of hypotheses considered: 1559034
Number of hypotheses explored: 940132
Time: 20 min 40 s 781 ms

08:39:50: Started weka.associations.Tertius
10:00:31: Finished weka.associations.Tertius
```

4 SONUÇ

Çalışmada örnek bir veri tabanında bilgi keşfi sürecinin yapılması anlatılmış ve modelin değerlendirilmesinde kullanılan WEKA ve YALE programları örnekler üzerinde test edilerek, kurallar elde edilmiştir.

Veri olarak birliktelik kurallarının incelenmesinde kullanılabilecek gram negatif basillerle oluşan hastalıklara ait veriler kullanılmıştır.

Çalışmada kullanılan programların birliktelik kuralları modelinde kuralların bulunmasında kullanılan Apriori, Tahminci Apriori ve Tertius algoritmaları incelenmiş ve bunlara ait sonuçlar yorumlanmıştır. Her iki programda sayısal veri kabul etmediği için, sayısal veri tipleri, metinsel veri tiplerine dönüştürülmüştür.

WEKA ve YALE algoritmalarının karşılaştırılmalarına ait bilgiler Tablo-18'de sunulmuştur.

Tablo-18 WEKA ve YALE Programları Birliktelik Kuralları Algoritmalarının Karşılaştırılması

Algoritma	Sonuçların Güvenilirliği		Programın Çalışması		Programın Çalışma Hızı	
	WEKA	YALE	WEKA	YALE	WEKA	YALE
Apriori	Çok İyi	Orta	İyi	İyi	Hızlı	Orta Hızlı
Predictive Apriori	Çok İyi	Tespit Edilemedi	İyi	Çalıştırılamadı	Hızlı	Tespit Edilemedi
Tertius	Çok İyi	Çok İyi	İyi	Orta Hızlı fakat veriye fazla alan eklenmesi gerekti	Hızlı	Yavaş

WEKA ve YALE programları Apriori ve Tertius algoritmalarına ait performansların karşılaştırılmaları Tablo-19'da sunulmuştur. Yapılan denemelerde WEKA programının YALE'a göre Apriori algoritması için 1'e 2 ila 1'e 5, Tertius algoritması için ise 1'e 3 ila 1'e 8 arasında değişen oranlarda daha hızlı çalıştığı ortaya çıkarılmıştır.

Tablo 19 WEKA ve YALE Programları Apriori ve Tertius Algoritmaları Performans Zamanları

SERVİS	PROGRAM	Zaman (Saniye)	
		Apriori	Tertius
cinsiyet_materyal_servis	WEKA	1	24
	YALE	2	96
gun_materyal	WEKA	1	2
	YALE	1	7
gun_mevsim_materyal	WEKA	2	11
	YALE	2	82
gun_mevsim_mikroorganizma	WEKA	1	11
	YALE	1	72
materyal_mikroorganizma	WEKA	1	3
	YALE	2	15
materyal_servis	WEKA	1	3
	YALE	2	14
materyal_servis_mevsim	WEKA	1	24
	YALE	3	144
materyal_servis_mikroorganizma	WEKA	1	49
	YALE	4	290
mevsim_yaş_mikroorganizma	WEKA	1	12
	YALE	2	80
servis_mikroorganizma	WEKA	1	3
	YALE	2	15
servis_mikroorganizma_yaş	WEKA	1	19
	YALE	3	125
yas_servis	WEKA	1	2
	YALE	1	5
7 alanlı_tam_veri	WEKA	1	4800
	YALE	5	24700

Program Apriori için 0.4 güvenilirlik, 0.1 destek değeri, Tertius için ise 10 kural değeri ile çalıştırılmıştır.

Veri madenciliği henüz Türkiye'de yaygın olarak kullanılmamaktadır. Genelde bankacılık, sigortacılık, borsa gibi alanlarda ve büyük ölçekli işletmeler tarafından kullanılmaktadır. Fakat henüz tıp alanında çok sınırlı olarak kullanılmamaktadır.

Gelecekte tıp alanında kullanılmasının ve günlük olarak kullanılan programlara bütünleşik hale getirilmesinin faydalı olacağı düşünülmektedir.

Yapılan uygulamalarda farklı sonuçlara ulaşılması, kuralların yorumlanmasında konunun uzmanları ile birlikte çalışılması gerektiğini göstermiştir.

KAYNAKÇA

AGRAWAL, Rakesh ve Ramakrishnan Srikant, "Fast Algorithms For Mining Association Rules", Proceedings of the 20th VLDB Conference, Santiago, Şili, 1994, s.487-499

AKPINAR, Haldun, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İstanbul Üniversitesi İşletme Fakültesi Dergisi, C.29, S.1 (Nisan 2000), s: 1-22

ALTINTOP, Ümmühan, Internet Tabanlı Öğretimde Veri Madenciliği Tekniklerinin Uygulanması, Kocaeli Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Kocaeli, 2006

AYDOĞAN, Fatih, E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi, Hacettepe Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Ankara, 2003, s.17

BERRY, Michael J.A., Gordon S. Linoff, **Data Mining Techniques For Marketing, Sales and Customer Relationship Management**, Wiley Publishing, Indiana, 2004

BUCHANAN, B.G., Brief History of Artificial Intelligence, "Brief History of Artificial Intelligence", (Çevrimiçi) http://www.aaai.org/AITopics/bbhist.html, (Erişim Tarihi: 07.09.2007)

BURUNCUK, Gülçin, Data Mining For Customer Segmentation And Profiling: A Case Study For A Fast Moving Consumer Goods (Fmcg) Company, Boğaziçi Üniversitesi, Bilgi Yönetim Sistemleri Enstitüsü Yüksek Lisans Tezi, İstanbul, 2006

DUNHAM, Margaret H, **Data mining Introductory and Advanced topics**, Prentice Hall, 2003

FAYYAD, Usama, Gregory Piatetsky-Shapiro, ve Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, c.17,S.3(1996). s.40

FLACH, Peter A. ve Nicolas Lachiche, "Confirmation-Guided Discovery of First-Order Rules with Tertius", Machine Learning, S.42, 2001, s.61-95

FORBES, Betty A., Daniel F. Sahm ve Alice S. Weissfeld, **Diagnostic Microbiology**, Missouri, Mosby, 2002

HAIR, Joseph F. vd., **Multivariate Data Analysis**, Prentice Hall, New Jersey, 1998, s.680-695

HAN, Jiawei ve Micheline Kamber, **Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, 2000

HAND, David, Heikki Mannila ve Padharic Smyth, **Principles of Data Mining**, MIT Press, Londra, 2001

KANTARDZIC, Mehmed, **Data Mining: Concepts, Models, Methods, and Algorithms**, John Wiley & Sons, Danvers, 2003

KIYAK, Erkan, CRISP-DM Yöntembilim Kullanılarak Deniz Kuvvetleri Verisi Üzerinde Veri Madenciliği Sınıflandırma Tekniklerinin Karşılaştırılması, Kocaeli Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Kocaeli, 2006,

MIERSWA, Ingo vd. "Yale: Rapid Prototyping for Complex Data Mining Tasks", In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006)

MUTTER, Stefan, Classification using Association Rules, Waikato Üniversitesi Yüksek Lisans Tezi, Yeni Zelanda, 2004, s.15-16

ÖZEKEŞ, Serhat, "Veri Madenciliği Modelleri ve Uygulama Alanları", İstanbul Ticaret Üniversitesi Dergisi, S.3 (Haziran 2003), s.65-82

PARRUD, Olivia, **Data Mining Cookbook**, John Wiley & Sons, New York, 2001.

SCHEFFER, Tobias, "Finding Association Rules that Trade Support Optimally Against Confidence", Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases Conference, Freiburg, Almanya, Eylül 2001, s.424-435

SCUSE, David ve Peter Reutemann, **WEKA Experimenter Tutorial for Version 3-4**, Yeni Zelanda Waikato Üniversitesi, 2007

TAN, Pang-Ning, Michael Steinbach ve Vipin Kumar, **Introduction to Data Mining**, Boston, Addison Wesley, 2005

TANG, ZhaoHui ve Jamie MacLennan, **Data Mining with SQL Server 2005**, Wiley Publishing, Indianapolis, 2005

TANTUĞ, Ahmet Cüneyd, Veri Madenciliği ve Demetleme, İstanbul Teknik Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul, 2002

TOPÇU, Ayşe Wilke, Güner Söyletir ve Mehmet Doğanay, **İnfeksiyon Hastalıkları ve Mikrobiyolojisi**, İstanbul, Nobel , 2002

UĞUZ, H. vd., "Apriori Algoritması Kullanılarak Web Kullanım Madenciliği Yönteminin Web Log Kayıtlarına Uygulanması", IJCI Proceeding of International Conference on Signal Processing, C.1, S.2 (2000), s: 499-501

WITTEN, Ian H. ve Eibe Frank, **Data Mining Practical Machine Learning Tools and Techniques**, Elsevier, Boston, 2005

YILMAZ, Emrah, Kütahya İlinde Sosyal Sınıfların Belirlenmesi ve Veri Madenciliği ile Tüketici Profilinin Çıkarılmasına Yönelik Bir Uygulama, Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Yüksek Lisans Tezi, Kütahya, 2006