Süleyman Demirel Üniversitesi
İktisadi ve İdari Bilimler
Fakültesi Dergisi
Y.2009, C.14, S.1 s.31-52.

Suleyman Demirel University
The Journal of Faculty of Economics
and Administrative Sciences
Y.2009, Vol.14, No.1 pp.31-52.

VERİ MADENCİLİĞİ: KARAR AĞACI ALGORİTMALARI VE İMKB VERİLERİ ÜZERİNE BİR UYGULAMA*

DATA MINING: DECISION TREE ALGORITHMS AND AN APPLICATION ON ISE DATA

Yrd.Doç.Dr.Ali Sait ALBAYRAK**
Öğr.Gör.Şebnem KOLTAN YILMAZ***

ÖZET

Veri madenciliği, verilerden önceden bilinmeyen anlamlı bilgileri tanımlama ya da tahmin etme tekniklerini içermektedir. Çalışmada, İMKB 100 endeksinde sanayi ve hizmet sektörlerinde faaliyet gösteren 173 işletmenin 2004–2006 yıllarına ait yıllık finansal göstergelerinden yararlanarak veri madenciliği tekniklerinden birisi olan karar ağaçları tekniği uygulanmıştır. Seçilen finansal göstergelere göre sanayi ve hizmet sektörlerinde faaliyet gösteren firmaları ayıran en önemli değişkenler saptanmıştır.

ABSTRACT

Data mining include descriptive and predictive techniques for meaningful knowledge which is unknown early from data. In this study, decision trees technique which is one of the data mining techniques is applied on 173 firms that operate in industry and service sectors in ISE 100 Index using their annual financial indicators from 2004 to 2006. According to selected financial indicators, the most important factors of firms that operate in industry and service sectors are determined.

Veri Madenciliği, CHAID Analizi, İMKB–100
Data Mining, CHID Analysis, ISE–100

GİRİŞ

Ham veri kendi başına değersizdir. Veri, bilgisayar sistemleriyle belirli bir amaca yönelik işlenerek bilgiye dönüşmektedir. Organizasyonlarda bu amaca yönelik kurulan bilgi sistemleri, satışlar, faturalar, ödemeler gibi işlemlerin takip edilmesini sağlarken, karar almaya yönelik ihtiyaçlara da girdi oluşturmaktadır. Ancak bilginin olağanüstü artışıyla birlikte kurum içi ve dışı bilgilerin yanı sıra önceden tahmin edilemeyen sorulara yönelik cevap

---

* Bu çalışma, ZKÜ, SBE, İşletme Anabilim Dalında Tamamlanan "Veri Madenciliği: İstanbul Menkul Kıymetler Borsası Örneği," adlı Yüksek Lisans Tezinden Üretilmiştir.

** Zonguldak Karaelmas Üniversitesi, İİBF, İşletme Bölümü Öğretim Üyesi.
*** İnönü Üniversitesi, Akçadağ MYO, Bilgisayar Tek. ve Prog. Bölümü Öğretim Görevlisi.

bulan, ileriye dönük tahmin sistemlerine ihtiyaç duyulmuştur. Bu bağlamda veri madenciliği (VM) teknikleri diğer birçok alanda olduğu gibi işletmecilik alanında da yaygın bir şekilde kullanılmaktadır. VM, bilgi teknolojilerinin doğal gelişim sürecinin sonucu olarak da değerlendirilebilir. Çok büyük ölçekli veriler, farklı alanlardaki büyük ölçekli veri tabanları içinde değerli verileri bulunduran bir veri madeni gibi düşünülebilir. VM ise bu verilerden önceden bilinmeyen anlamlı bilgileri üretme süreci olarak tanımlanmaktadır. VM, bu süreci bilgisayarı, makine öğrenmesini (machine learning), veri tabanı veya veri ambarı yönetimini, matematiksel algoritmaları ve istatistik teknikleri kullanarak gerçekleştirmektedir.

VM ile keşfedilen bilgi, çevrimiçi (online) veri takibi, iş yönetimi, ürün kontrol sistemleri, pazarlama, tıp, finans, kimya, coğrafi bilgi sistemleri (CBS), mühendislik, görüntü tanıma ve robot görüş sistemleri, uzay bilimleri, sosyal bilimler, davranış bilimleri, meteoroloji gibi alanlarda yaygın bir şekilde kullanılmaya başlanmıştır.

VM yerine veri tabanlarından bilgi keşfi kavramı (VTBK) da kullanılmaktadır. VTBK ifadesi ilk defa Piatetsky-Shapiro [1] tarafından 1989 yılında gerçekleştirilen ilk VTBK çalışma grubu toplantısında kullanılmış, konuyla ilgili kavram ve tanımlar ortaya konulmuştur. VM terimi de VTBK'nin bir bileşeni olarak tanımlanmıştır. Daha sonra VM yöntemleri geliştirmeye yönelik çalışmalar devam etmiş, Agrawal vd. [2], nicel birliktelik kurallarının madenciliği için hızlı bir algoritma olan Apriori'yi önermişlerdir. VTBK'nin temel süreçleri üzerinde hiyerarşi arayışının sonucunda, Fayyad vd. [3], veri tabanlarında bilgi keşfinin süreçlerine ve VM'nin bu süreçteki yerine yönelik bir akış sunmuşlardır. Ayrıca, VM'nin temel özelliklerini irdelemişlerdir. VM'nin uygulama yaygınlığı kazanmasıyla birlikte uygulamaya yönelik çalışmalar ağırlık kazanmaya başlamış ve Berson vd. [4], VM'nin en yaygın kullanım alanı olan müşteri ilişkileri yönetimi kapsamında VM yöntemleri ve uygulamalarına yer vermişlerdir. Ayrıca, yeni nesil yöntemlerden olan karar ağacı yöntemlerinin başlıcalarından CRT (Classification and Regression Tree) ve CHAID (Chi-squared Automatic Interaction Detection) yöntemlerini incelemişlerdir. VM bilişim ve istatistik olmak üzere iki bakış altında incelenmektedir. Ziarko [5], Elder ve Pregipon [6], VM'yi, istatistik alanındaki birçok yöntemi kullanmasına karşın, nesnelerin nitelik değerlerine bağlı çıkarım yapmada bilinen istatistik metotlardan ayırmaktadır.

---

[1] Gregory PİATETSKY-SHAPİRO, Knowledge Discovery in Real Databases: A Workshop Report, **AI Magazine,** C. 11, S. 5, 1991, s. 68-70.

[2] R. AGRAWAL, H. MANNİLA, R. SRİKANT, H. TOİVONEN, ve A. I. VERKAMO, Fast Discovery of Association Rules, Advances in Knowledge Discovery and Data Mining, **AAAI/MIT Press,** Chapter 12, 1995, s. 307-328.

[3] Usama FAYYAD, Gregory, PIATETSKY-SHAPIRO ve Padhraic SMYTH, From Data Mining to Knowledge Discovery in Databases, **AI Magazine,** C. 17, S. 3, 1996, s. 37-54.

[4] A. BERSON, S. SMİTH, ve K. THEARLİNG, Building Data Mining Applications for CRM, **McGraw Hill,** USA, 1999.

[5] W. ZİARKO, The Discovery, Analysis, and Representation of Data Dependencies in Databases, Knowledge Discovery in Databases, **AAAI/MIT Press,** 1991, s.195-212.

[6] J. F. ELDER, ve D. PREGİPON, "A Statistical Perspective on KDD", **The 1st International Conference on Knowledge Discovery and Data Mining,** 20-21 Ağustos, Montreal, 1995, s. 87-93.

Hastie vd. [7], VM'ye istatistik perspektiften yaklaşmış ve VM ile istatistik arasındaki ilişkiyi 'İstatistik Öğrenme' kavramı ile kurmuşlardır. Moss ve Atre [8], VM ile istatistik analiz yöntemlerini karşılaştırmış ve aralarındaki farklılıkları ortaya koymuşlardır.

VM'de kullanılan modeller, tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında incelenmektedir. VM modellerini gördükleri işlevlere göre sınıflama (classification) ve regresyon (regression) modelleri, kümeleme (clustering) modelleri ve birliktelik kuralları (association rules) ve ardışık zamanlı örüntüler (sequential patterns) olmak üzere üç ana başlık altında incelemek de mümkündür. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir.

Bu çalışmada, VM modellerinden biri olan sınıflama modelinin karar ağacı tekniği kullanılmıştır. Sınıflama modeli, sınıfı tanımlanmış mevcut verilerden yararlanarak, sınıfı belli olmayan verilerin sınıfını tahmin etmek için kullanılan VM modelidir. Karar ağacı, çok sayıda kayıt içeren bir veri kümesini, bir dizi karar kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yapıdır.

VM'de karar ağacı oluşturmak için çeşitli algoritmalar geliştirilmiştir. Bu algoritmalardan biri olan CHAID algoritması 1980 yılında Kaas tarafından en iyi bölmeyi hesaplamak için istatistik olarak anlamlı bir farklılığın olmadığı, hedef değişkene uyan çiftlerde tahmin değişkeninin olası kategori çiftini birleştirmesiyle oluşturulmuştur.[9] Haughton ve Oulabi [10], çalışmasında bir doğrudan pazarlama modelini CHAID algoritması ile gerçekleştirmiş, CRT ve CHAID algoritmalarının sonuçlarını karşılaştırmıştır. Akpınar [11], uygulamasında kişilerin geri ödemelerini düzenli veya düzensiz yapmalarına bağlı olarak, kredi değerlendirmesini iyi ve kötü şeklinde sınıflandırmak üzere CHAID algoritmasını kullanmıştır. Grobler vd. [12], okul etkinliği ve çeşitli bağımsız değişkenler arasındaki ilişkiyi tahmin etmede CHAID yönteminden yararlanmıştır. Doğan ve Özdamar[13], CHAID analizi yardımıyla ailelerin çocuk isteğine etki eden faktörlere ulaşmada bağımsız değişkenlerin birleşmiş kategorilerini ve alt gruplarını tahmin eden bir çalışma yapmıştır. Van Diepen ve Franses [14], CHAID algoritmasından

---

[7] T. HASTİE, R. TİBSHİRANİ, ve J. FRİEDMAN, The Elements of Statistical Learning; Data Mining, Inference and Prediction, **Springer Series in Statistics,** USA, 2001.

[8] L.T. MOSS, ve S. ATRE, Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications, **Addison-Wesley Publishing,** USA, 2003.

[9] SPSS, AnwerTree Algorithm Summary, **SPSS White Paper,** USA, 1999, s.2.

[10] Dominique HAUGHTON, ve Samer OULABI, "Direct marketing modeling with CART and CHAID", **Journal of Direct Marketing,** C. 11, S. 4, 1999 s. 42-52.

[11] Haldun AKPINAR, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İstanbul Üniversitesi İşletme Fakültesi Dergisi,** C. 29, S. 1, 2000, s.1-22.

[12] B. R. GROBLER, T.C. BISSCHOFF ve K.C. MOLOI, "The Chaid-Technique and the Relationship between School Effectiveness and Various Independent Variables", **International Studies in Educational Administration**, C. 30, S. 3, 2002, s. 44-56.

[13] Nurhan DOĞAN ve Kazım ÖZDAMAR, "Chaid Analizi ve Aile Planlaması ile İlgili Bir Uygulama", **Türkiye Klinikleri Tıp Bilimleri Dergisi,** C. 23, S. 5, 2003, s. 392-397.

[14] Merel VAN DIEPEN ve Philip Hans FRANSES, "Evaluating Chi-Squared Automatic Inter-

tahmin edilen sonuçların güven sınırlarını hesaplamak için önerdiği yöntemle, benzetim tekniğiyle elde edilmiş veriler kullanmıştır. Laura vd. [15], market üyelik kartlarının sınıflamasını CHAID yöntemi kullanarak yapmıştır. Koyuncugil ve Özgülbaş [16], CHAID algoritmasını kullanarak İstanbul Menkul Kıymetler Borsası (İMKB)'nda işlem gören Küçük ve Orta Ölçekli İşletmelerin (KOBİ'lerin) finansal profilleri ile güçlü ve zayıf yönlerinin belirlenmesini amaçlamıştır. McCarty vd. [17], iki farklı veri seti kullanarak doğrudan pazarlama bölümlemesi için RFM (Recency, Frequency, and Monetary Value), CHAID ve lojistik regresyon modellerinin karşılaştırmasını yapmıştır. Koyuncugil [18], CHAID algoritması ile İMKB şirketleri özelinde risk ölçümlemesindeki değişimleri otomatik olarak değerlendirmeyi ve öncü risk göstergelerini sektörel bazda belirlemeyi sağlayacak VM'ye dayalı bir sektörel risk modeli oluşturmuştur. Türe vd. [19], meme kanserli hastalarda yinelemesiz sağkalım süresini etkileyen risk faktörlerinin belirlenmesinde karar ağacı yöntemlerinden CRT, CHAID, QUEST (Quick Unbiased Efficient Statical Tree), C4.5 ve ID3 ile Kaplan- Meirer analizini birlikte kullanmıştır. Dolgun [20], Türkiye'deki hastaneleri karar ağaçları tekniğini kullanarak durumlarına ait sonuçlar çıkarmış ve bu sonuçlardan elde ettiği verilerle kümeleme analizi yaparak gruplandırmıştır.

Bu çalışmada, karar ağaçları kullanılarak İstanbul Menkul Kıymetler Borsası (İMKB) örneği incelenmiştir.

1. VERİ TABANLARINDA BİLGİ KEŞFİ

Verinin bilgiye dönüşümünün geleneksel yöntemi, klasik çözümleme ve yorumlamaya dayanmaktadır. Uygulamalar için veri setlerinin klasik araştırma şekli yavaş, pahalı ve subjektif olmaktadır. Veri tabanı sistemlerinin artan kullanımı ve hacimlerindeki olağanüstü artış, organizasyonların elde toplanan bu verilerinin pek çok alanda uygulanamaz olmasına sebep olmaktadır. Geleneksel sorgu veya raporlama araçlarının veri yığınları karşı-

---

action Detection", **Informatıon Systems,** C. 31, S. 8, 2006, s. 814-831.

15 Galguera LAURA, David LUNA ve M. Paz MENDEZ, "Predictive Segmentation in Action: Using CHAID to Segment Loyalty Card Holders", **International Journal of Market Research,** C. 48, S. 4, 2006, s. 459- 479.

16 Ali Serhan KOYUNCUGİL ve Nermin ÖZGÜLBAŞ, "İMKB'de İşlem Gören KOBİ'lerin Finansal Başarısızlığına Etki Eden Faktörlerin Veri Madenciliği İle Belirlenmesi", **3. KOBİ ve Verimlilik Kongresi,** 2006, 17–18 Kasım, İstanbul.

17 John A. MCCARTY ve Manoj HASTAK, "Segmentation Approaches in Data-Mining: A Comparison of RFM, CHAID and Logistic Regression", **Journal of Business Research,** C. 60, S. 6, 2007, s. 656–662.

18 Ali Serhan KOYUNCUGİL, Borsa Şirketlerinin Sektörel Risk Profillerinin Veri Madenciliğiyle Belirlenmesi, **Sermaye Piyasası Kurulu Araştırma Raporu,** Araştırma Dairesi, Ankara, 2007.

19 Mevlut TÜRE, Füsun TOKATLI ve İmran KURT, "Using Kaplan-Meirer Analysis Together With Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 and ID3) In Determining Recurrence-Free Survival of Breast Cancer Patients", **Expert Systems with Applications**, 2008, Article in Press.

20 M. Özgür DOLGUN, "Türkiye'deki Hastanelerin Veri Madenciliğiyle Gruplandırılması", **XI. Biyoistatistik Kongresi,** 27–30 Mayıs, Malatya, 2008.

sında yetersiz kalması, Veri Tabanlarında Bilgi Keşfi – VTBK adı altında sürekli ve yeni arayışlara neden olmaktadır.[21]

1.1. Veri Tabanlarında Bilgi Keşfi Aşamaları

Fayyad vd. [22], VTBK aşamalarını, veriyi anlama, hedef veri seti oluş-turma, veri temizleme ve önişleme, veri indirgeme, amaçları seçme, model analizi ve hipotez seçme, VM, örüntüleri yorumlama, bilgi üzerinde harekete geçme olarak 9 aşamada incelemektedir. Fu [23]'ya göre VTBK aşamaları, veri temizleme, veri dönüştürme, VM, modelleme, yorumlama ve değerlendirme aşamalarından oluşmaktadır. VM ise bu incelemelerde VTBK'nın odak nok-tasıdır. Fayyad [24], VTBK'nın veri hazırlama, örüntü arama, veri analizi, bilgi değerleme, yenileme gibi yinelemeli adımlardan oluştuğunu belirtmiş, VM'yi ise bilgi ya da örüntülerin algoritmik anlamıyla ilgilenen VTBK aşaması olarak değerlendirmiştir. Hegland [25], VTBK'yı veri sorgulama, veri temizle-me, veri analizi, bilginin sunumunu içeren prosedürler olarak ele almıştır.

Veri seçme işlemi, üzerinde çalışılacak veritabanından veya diğer veri kaynaklarından verilerin seçilerek veri dosyası oluşturulması işlemidir. Önişleme, uç (extreme), sapan (outlier), eksik (missing) veya hatalı birim değerlerinin düzeltilmesi gibi işlemleri içerir.

Veri dönüştürme; düzeltme, birleştirme, genelleştirme ve normalleştirme gibi değişik işlemlerden birini veya bir kaçını içerebilir. Veri normalleştirme en sık kullanılan veri dönüştürme işlemlerinden birisidir. [26] Bilgi keşfi sürecinin son aşaması olan VM, veriyi özetlemek ve gözlemlenemeyen ilişkileri bulmak için incelenen veri gruplarının analizidir. [27]

1.2. Veri Madenciliği

1.2.1. Veri Madenciliği Tanımı ve Amaçları

VM genel anlamda; büyük miktarda veri içerisinden, gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılması biçiminde ifade edilmektedir. VM'ndeki amaç, toplanmış olan bilgilerin bir takım istatistik yöntemlerle incelenip ilgili kurum ve yönetim destek sistemlerinde kullanılmak üzere değerlendirilmesidir. Veri madencisinin geleneksel yöntemlerde olduğunun aksine başlangıçta herhangi bir amacı ya da varmak istediği bir kavram yoktur. Yapılacak analizlerden sonra elde edilen verilerin bir istatistikçi gözü ile

---

[21] FAYYAD, 1996, s.37

[22] FAYYAD, 1996, s.37

[23] Yongjian FU, "Data Mining Tasks, Techniques and Applications", **IEEE,** No: 6648/97, 1997, s. 18-20.

[24] Usama FAYYAD, "Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases", **Ninth International Conference on Scientific and Statistical Database Management,** 11-13 Ağustos, Olympia, 1997, s. 2-11.

[25] Markus HEGLAND, Data Mining Techniques, **Acta Numerica,** Cambridge University Press, 2001, s. 313–355.

[26] R. J. ROIGER ve M. W. GEATZ, Data Mining: A Tutorial-Based Primer, **Addison Wesley**, USA, 2003.

[27] David HAND, Heikki MANİLA ve Padhraic SYMTH, Principles of Data Mining, **The MIT Press,** London, 2001.

incelenip daha önceden düşünülmemiş kavramların ortaya çıkarılması, başarılı bir VM süreci olarak kabul edilmektedir. [28]

VM, birleşik verilerdeki gizli bilgileri bulmak ve iş uzmanlığını arttırmak amacıyla yapılan yeni bir karar destek analiz işlemidir. Bazı anahtar kelimeler kullanılarak 4 aşamalı ayrıntılı VM tanımı şöyledir: [29]

1. VM, bir süreçtir.

2. VM, karar destek araçlarının niteliğini yüceltir.

3. VM, gizlenmiş bilgileri bulur.

4. VM, iş uzmanları için kavrayış dağıtıcı bir sistemdir.

VM uygulamalarının etkili olabilmesi için gerçekleştirmek zorunda olduğu 3 koşul vardır. Birincisi, VM, bölüme özgü veri yerine organizasyon çapında veriye ulaşmalıdır. Organizasyonun verisi, sık sık açık kaynaklı ya da maliyetli verilere eklenmektedir. Bu şekilde oluşturulan veri tabanı veri ambarı olarak adlandırılmaktadır. Veri entegrasyonu süresince uygulama, verileri, türetilmiş özellikleri (çift tanımlamaları kaldırarak, eksik değerleri doldurarak) temizlemektedir. İkinci olarak, bir VM uygulaması veri ambarlarındaki bilgiyi işlemek zorundadır. Son aşamada ise VM işlenmiş veriyi karar vermeye imkân verecek şekilde düzenlemeli ve sunmalıdır. [30]

1.2.2. Veri Madenciliği Süreci

VM altı adımlı bir süreç olarak incelenebilir: [31]

(1) Araştırma Probleminin Tanımlanması (Business Understanding)

Bu aşama veri madenciliği sürecinin en önemli aşamasıdır. Araştırma probleminin (konusunun) tanımlanması aşaması araştırmanın amacını, mevcut durumun değerlendirilmesini, veri madenciliğinin amaçlarını ve proje planlama sürecinin belirlenmesini kapsamaktadır.

(2) Verileri Tanıma Aşaması (Data Understanding)

Veri anlama aşaması veri toplamakla başlamaktadır. Daha sonra benzer verileri bir araya getirme, veri niteliklerini tanımlama, verileri keşfetme, gizli bilgileri sınıflandırma ile sürece devam etmektedir.

(3) Veri Hazırlama Aşaması (Data Preperation)

Veri hazırlama aşaması, ham veriden başlayarak son veriye kadar yapılması gereken bütün düzenlemeleri içermektedir. Veri hazırlama, tablo,

---

[28] Alper VAHAPLAR, "Bir Coğrafi Veri Madenciliği Uygulaması", **Yayınlanmamış Yüksek Lisans Tezi**, Ege Üniversitesi, Fen Bilimler Enstitüsü, İzmir, 2003.

[29] Charly KLEİSSNER, "Data Mining for Enterprise", **31st Annual Hawai International Conference on System Screnes**, 1060–3425/98, 1998, s. 295–304.

[30] Evangelos SİMOUDİS, "Reality Check for Data Mining", **IEEE Expert: Intellegent Systems and Their Applications**, C. 11, S. 5, 1996, s. 26-33.

[31] Daniel T. LAROSE, Discovering Knowledge in Data: An Introduction to Data Mining, **John and Wiley Sons Incorporated,** USA, 2005.

kayıt, veri dönüşümü ve modelleme araçları için veri temizleme gibi özellikleri içermektedir.

(4) Modelleme Aşaması (Modelling)

Bu aşamada, verilerden bilgi çekmek için ileri çözümleme yöntemleri kullanıldığından VM sürecinin en gösterişli aşamasıdır. Bu aşama uygun modelleme tekniğinin seçimi, test tasarımının üretimi, model geliştirme ve tahmin işlemlerini içermektedir. Uygun modellerin seçilip uygulanmasıyla birlikte parametreler en uygun değişkenlere dönüştürülmektedir. VM, her problem tipi için farklı yöntemler içermektedir. Bazı yöntemler, veri tipi için uygun değildir ya da özel tanımlamalar gerektirmektedir. Bu nedenle gerekli olduğunda 3. aşama olan veri hazırlama aşamasına geri dönülür.

(5) Değerlendirme Aşaması (Evaluation)

Değerlendirme aşamasında, uygun model ya da modeller kurulduktan sonra, VM sonuçlarının araştırma probleminin amaçlarını gerçekleştirip gerçekleştirmediği değerlendirilir. Bu aşama sonuçların değerlendirilmesi, veri madenciliği sürecinin gözden geçirilmesi ve sonraki adımların ne olacağı hususlarını içermektedir. Bu aşamanın sonunda VM sonuçlarının kullanımı üzerindeki karara varılmaktadır.

(6) Uygulama Aşaması (Deployment)

Son aşama olan uygulama aşaması, araştırmacının tüm emeklerinin karşılığını aldığı bir aşamadır. Bu aşamada VM süreciyle üretilen bilgiler, pratik işletme problemlerinin çözümünde kullanılmaktadır. Bu aşamada elde edilen bilgilerin uygulanabilmesine yönelik bir plan hazırlama, gözden geçirme ve bakım faaliyetlerini içerir. Ayrıca bu aşamada nihai araştırma raporunun yazılması ve projenin gözden geçirilmesi işlemleri yer almaktadır.

VM'de veri kümesinin büyüklüğünden kaynaklanan en fazla zaman alıcı aşama, verilerin ön işlemden geçirilmesi aşamasıdır. VM uygulamalarında kaynakların %80'i verilerin ön işlemden geçirilmesi ve temizlenmesi süreçleri için harcanmaktadır. [32]

1.2.3. Veri Madenciliği İşlevleri

Genel olarak VM işlevleri tahmin edici ve tanımlayıcı veri madenciliği olarak ikiye ayrılabilir.

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. [33] Tanımlayıcı VM modellerinde ise tahmin edici modelin tersine karar

[32] Selwyn PİRAMUTHU, "Evaluating Feature Selection Methods for Learning in Data Mining Applications", **European Journal of Operational Research,** C. 156, S. 2, 2004, s. 483-494.

[33] Serhat ÖZEKES, "Veri Madenciliği Modelleri ve Uygulama Alanları", **İstanbul Ticaret Üniversitesi Dergisi,** Yıl 2, S. 3, 2003, s. 65–82.

vericilere yol göstermek için kullanılan mevcut verideki örüntüler tanımlanmaktadır. [34]

1.3. Veri Madenciliği Yöntem ve Teknikleri

Bir VM modeliyle aşağıdaki işlemlerden bir veya birkaçı gerçekleştirilebilir:

Sınıflama (Classification) ve Regresyon (Regression) Modelleri,

Kümeleme (Clustering) Modelleri ve

Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns).

Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir.

1.3.1. Sınıflama ve Regresyon Modelleri

En yaygın uygulanan VM tekniklerinden biri olan sınıflama, sınıfı tanımlanmış mevcut verilerden yararlanarak sınıfı belli olmayan verilerin sınıfını tahmin etmek için kullanılan VM modelidir. Sınıflama iki adım içeren bir işlemdir Birinci adımda tahmin için kullanılacak bir model oluşturulmaktadır. İkinci adımda, oluşturulan bu model sınıfı belli olmayan veriler üzerinde uygulanarak sınıflar tahmin edilmektedir. [35] Başlıca sınıflandırma teknikleri, Yapay Sinir Ağları (Artificial Neural Networks), Genetik Algoritmalar (Genetic Algorithms), K- En Yakın Komşu (K–Nearest Neighbour), Bellek Temelli Nedenleme (Memory Based Reasoning), Naive – Bayes, Lojistik Regresyon (Logistic Regression) ve Karar Ağaçlarıdır (Decision Trees).

1.3.2. Kümeleme

Kümeleme analizi, nesnelerin alt dizinlere gruplanmasını yapan bir işlemdir. Böylece nesneler, örneklenen kitle özelliklerini iyi yansıtan etkili bir temsil gücüne sahip olmaktadır. Sınıflamanın aksine, yeniden tanımlanmış sınıflara dayalı değildir. Kümeleme, bir denetimsiz öğrenme (unsupervised learning) yöntemidir. [36]

1.3.3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Birliktelik kuralları ile bir ilişkide yer alan niteliklerin değerleri arasındaki bağımlılıklar, anahtarda yer almayan diğer niteliklerin gruplandırılması ile bulunur. Bu kurallar ilk olarak Agrawal tarafından 1994'te geliştirilmiştir. Birliktelik kurallarının analizi süreci market sepeti analizi olarak da adlandırılır. Market sepeti analizinde müşteri ile ilgili veri hareketlerinden

---

[34] Herb EDELSTEIN, Mining Large Database - A Case Study, **Two Crows Corporation**, 2000.

[35] J. HAN ve M. KAMBER, Data Mining: Concepts and Techniques, **Morgan Kaufmann Publishers,** USA, 2001.

[36] Fatih AYDOĞAN, "E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi", **Yayınlanmamış Yüksek Lisans Tezi,** Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 2003.

hareketlerinden gelecekte müşterinin nasıl bir tercih yapacağına dair sonuçlar tahmin edilmektedir. [37] Çok sayıda verinin depolandığı bir veri tabanı içinde çeşitli nitelikler arasında hemen fark edilmeyen birtakım ilişkilerin ortaya çıkartılması stratejik kararların alınmasına yardımcı olabilir. Ancak, bu ilişkilerin çok sayıda verinin içinden elde edilmesi basit bir süreç değildir. Bu süreç birliktelik kuralı madenciliği (association rule mining) olarak adlandırılmaktadır. Veriler arasındaki ilişkiler, eğer-sonra ifadeleri ile aşağıdaki gibi gösterilmektedir. [38]

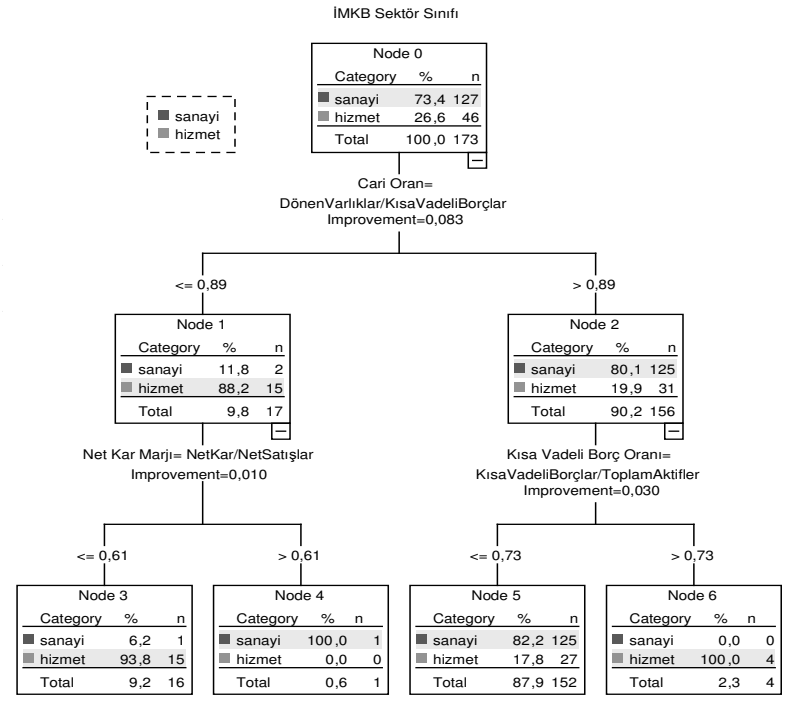
Eğer <bazı şartlar sağlanırsa> sonra <bazı niteliklerin değerlerini tahmin et>

En yaygın birliktelik kuralı algoritmaları arasında GRI (The Generalized Rule Induction), Apriori sayılabilir.

2. KARAR AĞAÇLARI (DECISION TREES)

Karar ağaçları, sınıfları bilinen örnek veriden tümevarım yöntemiyle öğrenilen ağaç şekilli bir karar yapısı çeşididir. [39] Bir karar ağacı, basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, çok küçük kayıt gruplarına bölerek kullanılan bir yapıdır. Her başarılı bölme işlemiyle, sonuç gruplarının üyeleri bir diğeriyle çok daha benzer hale gelmektedir. [40] Büyük veri tabanlarının kullanıldığı pek çok sınıflama probleminde ve karmaşık ya da hata içeren bilgilerde karar ağaçları yararlı bir çözüm olmaktadır. [41] Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları, VM'de kuruluşlarının kolay olması, yorumlanmalarının kolay olması, veri tabanı sistemlerine kolayca entegre edilebilmeleri, güvenilirliklerinin daha iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip olan bir tekniktir. [42]

Şekil 2.1: Karar Ağacı Örneği



Karar ağacı temelli analizlerin yaygın olarak kullanıldığı alanlar şunlardır: [43]

- Belirli bir sınıfın olası üyesi olacak elemanların belirlenmesi,
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok sayıdaki değişkenden en önemlilerinin seçilmesi,
- Gelecekteki olayların tahmin edilebilmesi için kurallar oluşturulması,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması,
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikli değişkenlere dönüştürülmesidir

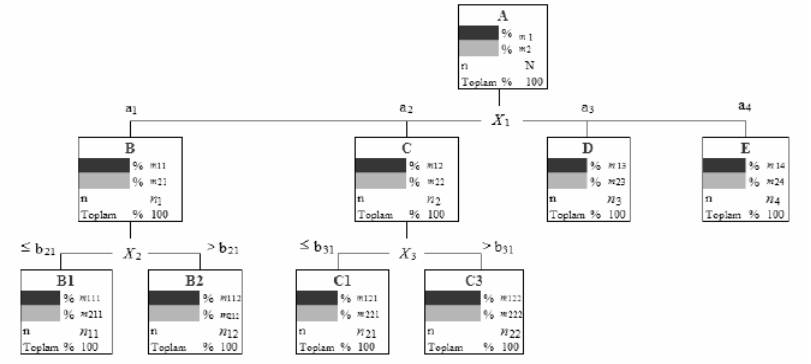
Karar ağacı oluşturmak için geliştirilen bu algoritmalar arasında CHAID (Chi- Squared Automatic Interaction Detector), Exhaustive CHAID, CRT (Classification and Regression Trees), ID3, C4.5, MARS (Multivariate Adaptive Regression Splines), QUEST (Quick, Unbiased, Efficient Statistical Tree), C5.0, SLIQ (Supervised Learning in Quest), SPRINT (Scalable Paralleizable Induction of Decision Trees) yer almaktadır.

[43] AKPINAR, 2000, s. 17.

2.1.1. CHAID Algoritması

CHAID metodu 1980'de Kaas tarafından en iyi bölmeyi hesaplamak için istatistik olarak anlamlı bir farklılığın olmadığı, hedef değişkene uyan çiftlerde tahmin değişkeninin olası kategori çiftini birleştirmesiyle oluşturulmuştur. [44] En uygun bölümleri seçmek için kullanılan entropy veya gini metrikleri yerine chi-square testi kullanılmaktadır. [45] En iyi bölmeyi hesaplamak için tahmin değişkenleri hedef değişkene uyan bir çiftin içinde istatistik olarak anlamlı bir fark kalmayıncaya kadar birleştirilmektedir. CHAID ile diğer yöntemler arasındaki en önemli farklılıklardan birisi, ID3, C4.5, ve CRT ikili ağaçlar türetirken, CHAID çoklu ağaçlar türetmektedir. [46]

Şekil 2.2: Karar Ağacı Değişken İlişkisi



Kaynak: Ali S. Koyuncugil, *Borsa Şirketlerinin Sektörel Risk Profillerinin Veri Madenciliğiyle Belirlenmesi*, Sermaye Piyasası Kurulu Araştırma Raporu, Araştırma Dairesi, Ankara, 2007.

Y'nin hedef değişken olduğu varsayılarak, Şekil 2.2: Karar Ağacı Değişken İlişkisi dikkate alındığında: [47]

- X1, X2 ve X3 olmak üzere sadece 3 değişken hedef Y değişkeniyle istatistik olarak önemli ilişkiye sahiptir.
- X1 değişkeni Y hedef değişkeniyle istatistik olarak en önemli ilişkiye sahiptir.
- X2 değişkeni X1 değişkeniyle X1=a1 olması koşuluyla istatistik olarak önemli ilişkiye sahiptir.
- X3 değişkeni X1 değişkeniyle X1=a2 olması koşuluyla istatistiksel açıdan önemli ilişkiye sahiptir.

---

[44] SPSS, 1999, s. 2

[45] Tuba İMAMOĞLU, "Veri Madenciliğinde Karar Ağaçları ile Bir Öğrenci Ders Başarısı Tahmin Aracı", **Yayınlanmamış Yüksek Lisans Tezi,** Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Kocaeli, 2005.

[46] TÜRE vd., 2008, s. 3

[47] KOYUNCUGİL, 2007, s. 17.

3. VERİ MADENCİLİĞİ UYGULAMASI: İMKB VERİLERİ

Bu çalışmada, İMKB100 sanayi ve hizmet sektöründe faaliyette bulunan 173 işletmenin 2004–2006 yıllarına ait yıllık finansal göstergelerinden yararlanarak VM tekniklerinden birisi olan karar ağaçları tekniği, CHAID algoritması uygulanarak seçilen finansal göstergelere göre firmaları ayıran en önemli değişkenler araştırılmaktadır. Bu şirketlerin, 2004–2006 yıllarına ait bilanço ve gelir tablolarına ilişkin veriler İMKB'nin ağ sayfasından elde edilmiştir.

3.1. Çalışmanın Amacı ve Önemi

Bu çalışmanın temel amacı, karar verme konumunda olanların, işletmeleri belirli yönlerden performanslarını değerlendirmede dikkate alması gereken en önemli ölçütlerin (değişkenlerin) saptanmasında veri madenciliği tekniklerinden karar ağacı algoritmalarının uygun bir yöntem olarak kullanılabileceğini göstermektir. Bu amaçla bugüne kadar daha çok regresyon analizi, diskriminant analizi ve lojistik regresyon analizi gibi çok değişkenli istatistik tekniklerden daha çok yararlanıldığı görülmektedir. Ülkemizde karar ağacı algoritmalarının kullanımı ise oldukça yenidir. Karar ağacı algoritmalarının en önemli avantajları, parametrik olmayan yöntemler arasında olması nedeniyle diğer çok değişkenli tekniklerde sağlanması gereken istatistik varsayımların olmamasıdır. Ayrıca karar ağacı algoritmalarının bağımlı ve bağımsız değişkenler arasındaki ilişkilerin yönünü, önem sırasını görselleştirmesi diğer avantajları arasındadır. Bu özelliği özellikle elde edilen sonuçların yorumunu oldukça basitleştirerek daha somut ve kullanışlı hale getirmektedir.

Bu bağlamda bu çalışmada İMKB100 sanayi ve hizmet sektöründe faaliyet gösteren firmaları ayırmada etkili olan en önemli karlılık, likitide, varlık kullanım etkinliği, sermaye yapısı ve işletme büyüklüğü değişkenleri araştırılmaktadır.

3.2. Çalışmanın Yöntemi

Profillerin oluşturulması için görsel, kolay anlaşılır, kolay yorumlanabilir ve kural çıkarımına izin vermesi nedeniyle VM yöntemlerinden karar ağaçları kullanılmıştır. Çalışmada İMKB100 endeksindeki şirketlerin profilleri, karar ağacı yöntemlerinden CHAID ile belirlenmiştir. CHAID'in analiz aracı olarak seçilme nedenlerinin başında sürekli ve kategorik tüm değişken tipleriyle çalışabilmesi gelmektedir.

Bununla beraber, sürekli bağımlı değişkenler otomatik olarak analizin amacına uygun olarak kategorize edilmektedir. CHAID, Ki-Kare metriği vasıtasıyla, ilişki düzeyine göre farlılık rastlanan grupları ayrı ayrı sınıflamakta ve ağacın yaprakları, ikili değil, verideki farklı yapı sayısı kadar dallanmaktadır.

3.2.1. Veri Anlama ve Veri Seçimi Aşaması

En uygun veri tabanı oluşturulurken hisse senedi ulusal pazarda işlem görmekte olan firmaların adları ve faaliyet konuları ile ilgili bilgiler

İMKB tarafından yayınlanmış olan İMKB100 endeksindeki firmalardan sağlanmıştır. Örneğin oluşturulabilmesi için kapsamdaki firmaların hesap dönemleri de incelenmiş, oranların hesaplanıp değerlendirilmesinde daha sağlıklı bir sonuca ulaşılabilmesi için İMKB'de kaydı bulunan, Ulusal Pazarda işlem gören ve finansal hesap dönemi 1 Ocak- 31 Aralık olan firmalar değerlendirmeye alınmıştır.

3.2.2. Veri Önişleme Aşaması

Bir ekonomideki finansal firmalar, birikim fazlası olan kişi ve kuruluşlarla belli amaçlar için bu birikimleri kullanmak isteyenlerin karşılaşmalarını sağlarlar. Finansal firma faaliyetleri, finansal olmayan firmalardan farklı olarak belli bir üretim süreci içermemektedir. Çeşitli araçlar kullanarak, fon aktarımları konusunda hizmet veren bu firmaların finansal yapıları diğer üretim ve hizmet firmalarından son derece farklıdır. [48]

Finansal firmaların müşterilerinden topladıkları mevduatların finansal tablolarındaki gösterim şekli, bu firmalarda yüksek kaldıraç oranlarına sebep olmaktadır. Bu firmaların finansal analizlerinde kullanılan birçok oranın, kendilerine özgü bir şekilde yorumlanması gerekmektedir ve tüm sektörlerde geçerli olabilecek değerlendirme ölçütlerinin oluşturulması mümkün değildir. Çalışmada daha tutarlı ve anlamlı bulgular elde edebilmek ve sağlıklı sonuçlara ulaşmak için finansal firmalar değerlendirme dışı bırakılmıştır.

Çalışma örneğinin belirlenmesi aşamasında ilk olarak ulusal pazarda işlem görmekte olan İMKB100 endeksindeki firmaların ad ve faaliyet konuları veri tabanına alınmıştır. Daha sonra çalışmada temel alınan finansal firma tanımının kapsamında bulunduğu saptanan firmalar listeden elenerek temizlenmiştir. Söz konusu 173 işletmeye ait finansal göstergeler veri tabanına aktarılmıştır.

3.2.3. Veri Dönüştürme Aşaması

Verilerin analizinde yıllar arası enflasyon etkisini yok etmek için finansal tablolardan alınan finansal göstergelere enflasyon düzeltmesi yapılmıştır. Daha sonra çalışmada kullanılan finansal oranlar hesaplanarak veriler düzenlenmiştir. Tablo 3.1'de çalışmada kullanılan finansal oranlar verilmiştir.

Çalışmada bağımlı değişken olarak sektör kukla değişkeni (sanayi=1 ve hizmet=2) alınmıştır.

3.2.4. Modelleme Aşaması: Karar Ağacı Tekniği CHAID Algoritmasının Uygulanması

Karar ağaçları algoritmaları uygulama bakımından bir hedef değişken (bağımlı değişken) ve hedef değişkeni açıklamaya yönelik kullanılacak açıklayıcı değişkenler (bağımsız değişkenler) olmak üzere iki grup değişken

[48] Arzum ERKEN, "Başlıca Fiyat Bazlı Oranların Hisse Senedi Getirisi Üzerindeki Etkileri ve İstanbul Menkul Kıymetler Borsası'nda Bir Uygulama", **Yayınlanmamış Yüksek Lisans Tezi,** Ankara Üniversitesi Sosyal Bilimler Enstitüsü, Ankara, 1998.

ile gerçekleştirilmektedir. SPSS 15 yazılımı ile CHAID algoritması kullanılarak $\alpha_{birleştir} = \alpha_{böl}$=0,10 olarak belirlenmiş ve karar ağaçları elde edilmiştir.

Tablo 3.1: Çalışmada Kullanılan Finansal Oranlar Tablosu

Değişken Sınıfı	Simge	Değişkenin Tanımı
Karlılık Oranları	*NK/TA*	Aktiflerin Karlılığı=Net Kar/Toplam Aktifler
	NK/ÖS	Özsermayenin Karlılığı=Net Kar/Özsermaye
	NK/NS	Net Kar Marjı=Net Kar/Net Satışlar
	ERO	Ekonomik Rantabilite Oranı=(VÖK+Finansman Giderleri)/Toplam Aktifler
Likitide Oranları	*DÖV/KVB*	Cari Oran=Döner Varlıklar/Kısa Vadeli Borçlar
	ATO	Likitide Oranı=(Döner Varlıklar-Stoklar)/Kısa Vadeli Borçlar
	NİS/NS	İşleme Sermayesinin Net Satışlara Oranı=(DÖV-KVB)/Net Satışlar
	KVAL/DÖV	Kısa Vadeli Alacaklar/Döner Varlıklar=KVAL/DÖV
Varlık Kullanım Etkinliği	*NS/TA*	Aktif Devir Hızı=Net Satışlar/Toplam Aktifler
	SDH	Stok Devir Hızı=Satışların Maliyeti/Stoklar
	NS/DÖV	İşletme Sermayesi Devir Hızı (İSDH)=Net Satışlar/Döner Varlıklar
	NS/ÖS	Özsermaye Devir Hızı=Net Satışlar/Özsermaye
Sermaye Yapısı Oranları	*KVB/TA*	Kısa Vadeli Borç Oranı=Kısa Vadeli Borçlar/Toplam Aktifler
	TB/TA	Toplam Borç Oranı=Toplam Borçlar/Toplam Aktifler
	DÖV/TA	Döner Varlıklar/Toplam Aktifler
	DV/ÖS	Duran Varlıklar/Özsermaye
İşletme Büyüklüğü	*İB1*	Toplam Aktifler
	İB2	Net Satışlar

3.2.5.Değerlendirme Aşaması: CHAID Modeli Sonuçları

Çalışmada karar ağaçları algoritması bakımından hedef değişken olan sektör bulgularına göre 173 işletmenin CHAID model sonuçları ve profilleri verilmiştir.

Risk değeri, sonuç modeli ve modelleri oluşturmak için bilgi sağlayan CHAID model tablosunun en uygun ağaç yapısını belirlemede bir göstergedir. Terminal düğüm sayısı da en uygun ağaç yapısı için bir gösterge olarak kullanılmaktadır. Aşağıda verilen tablo, hedef değişkene ait 12 farklı ağaç yapısından oluşmaktadır. Ağaç yapılarının derinliği ve düğüm sayıları, ağaç 1'den ağaç 12'ye kadar artmaktadır. Risk değerleri ise standart hata, yerine koyma (resubstitution) ve çapraz geçerlilik (cross validation) yüzdelerinden oluşmaktadır. En uygun modellerin seçiminde toplam ve nihai düğüm sayıları, ağaç derinliği, yerine koyma, çapraz geçerlilik ölçütlerinden yararlanılmıştır. Hedef değişken olarak sektör sınıfını tanımlayan modelde riskin ölçüsü yanlış sınıflandırma oranını göstermektedir. Bu nedenle risk yüzdesi, en uygun ağaç yapısının seçiminde dengeli ve minimum olmalıdır.

Modelin uygulanması sırasında çapraz geçerlilik kuralı seçilmiş ve farklı ağaç yapılarını izlemek için algoritmanın durdurma kurallarına (stopping rules) uygun farklı düğüm sayıları denenmiştir. Tablolarda sektör bağımlı değişkeni için yerine koyma ve çapraz geçerlilik riskleri birbirine en yakın olan ağaç yapısı seçilmiştir.

Tablo 3.2: Sektör Sınıfı CHAID Algoritma Sonuçları

Tüm Olası Ağaçlar	Karar Düğüm Sayısı			Yerine Koyma (%)		Çapraz Geçerlilik (%)	
	Toplam	Terminal	Ağaç Derinliği	Risk	Standart Hata	Risk	Standart Hata
Ağaç 1	1	1	0	26,6	3,4	26,6	3,4
Ağaç 2	3	2	1	26,6	3,4	26,6	3,4
Ağaç 3	4	3	1	25,4	3,3	30,1	3,5
Ağaç 4	7	5	2	17,3	2,9	18,5	3,0
Ağaç 5	8	5	2	19,7	3,0	30,6	3,5
Ağaç 6	13	9	2	12,1	2,5	24,3	3,3
Ağaç 7	10	7	3	17,3	2,9	17,9	2,9
Ağaç 8	12	8	3	16,2	2,8	24,3	3,3
Ağaç 9	16	10	3	13,3	2,6	21,4	3,1
Ağaç 10	17	11	3	11,0	2,4	21,4	3,1
Ağaç 11	23	14	4	8,7	2,1	22,5	3,2
Ağaç 12	25	15	5	8,7	2,1	20,8	3,1

Tablo 3.3'de CHAID karar ağacı ile elde edilen İMKB100 endeksindeki 7 işletme profiline, profili oluşturan düğümlere ve sektör değişkenine etki eden değişkenlere yer verilmiştir. Profilleri ve sektör durumunu değerlendirmek için Tablo 3.3 ile Şekil 3.1 birlikte incelenmektedir. Şekil 3.1, 4 ana daldan oluşan CHAID karar ağacını ve profillerin tamamını tek bir şekilde göstermektedir.

Çalışmada, karar ağaçları algoritması bakımından hedef değişken seçilen sektör bulgularına göre araştırma kapsamında İMKB100 endeksinde yer alan 173 işletmeden %73,4'ünün (127 işletme) sanayi sektöründe, %26,6'sının (46 işletme) hizmet sektöründe olduğu anlaşılmaktadır. Şekil 3.1'de verilen CHAID karar ağacı ile işletmelerin sektör profilleri oluşturulmuş ve işletmeler 7 farklı profilde gruplanmıştır. İşletmeleri klasik yöntemle "sanayi" ya da "hizmet" sektörü olarak iki grupta toplamak mümkünken, bu çalışma ile sektör durumu ve özelliklerine göre 7 farklı profilde gruplandırmak söz konusu olmuştur.

İMKB100 endeksindeki şirketlerin sektörler itibariyle profillerinin oluşturulmasında, en önemli değişkenin işletme sermayesinin net satışlara oranı (p=0,000) olduğu anlaşılmaktadır. İşletme sermayesinin net satışlara oranı değerlerine göre ağacın dalları dört ana dalda (Node 1-Node 4) toplanmıştır. Bu dört düğümü oluşturan alt düğümler de profilleri oluşturmaktadır. İMKB100 endeksindeki firmaların sektör profillerinin oluşturulmasında etkili

olan en önemli değişkenler sırasıyla işletme sermayesinin net satışlara oranı, stok devir hızı ve ekonomik rantabilite oranı değişkenleridir.

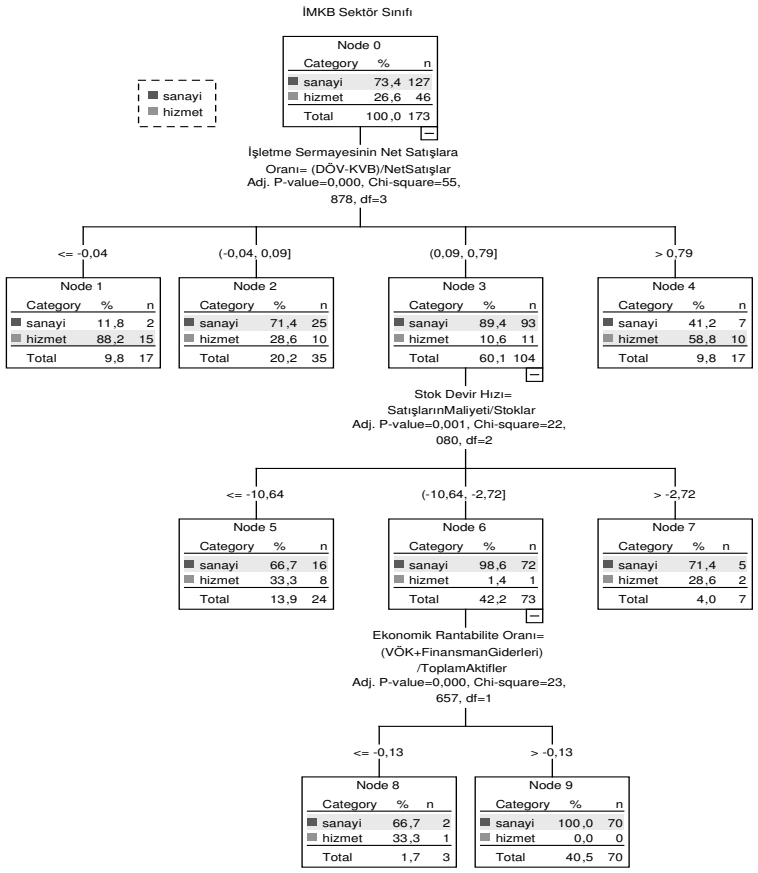
Profil 1 (Node 1) ve 2 (Node 2): Tablo 3.3 ve Şekil 3.1'de görüldüğü gibi 1. ve 2. profil, işletme sermayesinin net satışlara oranı değişkeninden oluşmaktadır. İşletme sermayesinin net satışlara oranı -0,04'e eşit ya da küçük olan işletmeler birinci profili, -0,04 ile 0,09 arasında olan işletmeler ikinci profili oluşturmaktadır. Birinci profildeki işletmelerin %11,8'i (2 işletme) sanayi sektöründe, %88,2'si (15 işletme) hizmet sektöründe yer alırken; ikinci profildeki işletmelerin %71,4'ü (25 işletme) sanayi sektöründe; %28,6'sı (10 işletme) hizmet sektöründe yer almaktadır. İşletme sermayesinin net satışlara oranı negatif olan birinci profil, işletmelerin kısa vadeli borçlarının yüksek olduğunu göstermekte ve faaliyetleri süresince tam kapasitede karlı ve verimli çalışabilmesini güçleştirmektedir. Net işletme sermayesinin büyüklüğü, işletmelerin üretimini devam ettirebilmesi ve yükümlülüklerini karşılayamama riskini azaltması açısından önem taşımaktadır. İkinci profilde yer alan işletmelerin işletme satış oranlarının sermayeye bağımlı olmadan, yeterli bir satış düzeyine ulaşabildiği ve diğer finansal oranlarının da yeterli olması durumunda satışların, işletme faaliyetleriyle gerçekleşebileceği görülmektedir. Buna karşın işletmeler, işletme sermayesine bağlı olmadan satışlarını yeterli düzeye getirmek için önlemler almalı ve bu düzeyde kendi faaliyetlerinin dönüş hızını koruyabilmelidir.

Tablo 3.3: Sektör Sınıfı Profilleri ve Profilleri Oluşturan Değişkenler

Profiller	Düğümler	NİS/NS	SDH	ERO
Profil 1	1	≤ -0,04		
Profil 2	2	-0,04 – 0,09		
Profil 3	3, 5	0,09 – 0,79	≤ -10,64	
Profil 4	3, 6, 8	0,09 – 0,79	(-10,64) – (-2,72)	≤ -0,13
Profil 5	3, 6, 9	0,09 – 0,79	(-10,64) – (-2,72)	> -0,13
Profil 6	3, 7	0,09 – 0,79	> (-2,72)	
Profil 7	4	> 0,79		

Profil 3 (Node 3): Bu profil, işletme sermayesinin net satışlara oranı ve stok devir hızı değişkenlerinden oluşmaktadır ve işletmelerin %66,7'si (16 işletme) sanayi sektöründe, %33,3'ü (8 işletme) hizmet sektöründe yer almaktadır. Bu profili oluşturan işletmelerin işletme sermayesinin net satışlara oranı 0,09 ile 0,79 arasında iken, stok devir hızı -10,64'e eşit ya da küçüktür. Stok devir hızının negatif olması işletmede fazla stok bulundurulduğunu, satışların net işletme sermayesine bağlı olduğunu ve bu nedenle işletme satışlarının azaldığını göstermektedir. Bu da stok tutma maliyetinin artmasına ve stoklarda bozulma riskine sebep olmaktadır. İşletme sermayesinin net satışlara oranının yüksek olması, satışların net işletme sermayesine bağlı olması stok devir hızının yavaş olmasında da önemli bir etkendir.

Şekil 3.1: İMKB Sektör Sınıfı Karar Ağacı *



Profil 4 (Node 8) ve 5 (Node 9): Toplam 73 işletmenin olduğu dördüncü ve beşinci profiller, işletme sermayesinin net satışlara oranı, stok devir hızı ve ekonomik rantabilite oranı değişkenlerinden oluşmaktadır. Dördüncü profildeki işletmelerin %66,7'si (2 işletme) sanayi sektöründe, %33,3'ü (1 işletme) hizmet sektöründe yer alırken beşinci profildeki işletmelerin tamamı (70 işletme) sanayi sektöründe yer almaktadır. Bu profili oluşturan işletmelerin, işletmelerin işletme sermayesinin net satışlara oranı 0,09 ile 0,79 arasında, stok devir hızı -10,64 ile -2,72 arasında değişmektedir. Dördüncü profildeki işletmelerin ekonomik rantabilite oranları -0,13'ten eşit ya da küçük iken beşinci profilde ise -0,13'ten büyüktür. Ekonomik rantabilite oranının negatif olması firmaların faaliyetlerinin finansmanında ağırlıklı olarak yabancı kaynak kullandığını göstermektedir. Ayrıca söz konusu firmaların stok devir hızının negatif olması stok tutma maliyetini

* Şekilde geçen Node, Category, Total İngilizce sözcükleri sırasıyla düğüm, kategori ve toplam anlamına gelmektedir.

yükselterek, firmanın finansman gereksinimini arttırabilir ve stokta bulunan malların çeşitli nedenlerle satış kabiliyetini yitirme olasılığını yükselebilir. İşletme sermayesinin net satışlara oranı pozitif olan söz konusu firmaların satışlarının işletme sermayesini karşılayacak düzeyde olmadığını ya da ihtiyaçtan fazla net işletme sermayesi bulunduğunu, alacakların ve stokların dönüş hızının yeterli olmadığını göstermektedir.

Profil 6 (Node 7): Bu profil, işletme sermayesinin net satışlara oranı ve stok devir hızı değişkenlerinden oluşmaktadır. Altıncı profili oluşturan işletmelerin %71,4'ü (5 işletme) sanayi sektöründe, %28,6'sı (2 işletme) hizmet sektöründe yer almaktadır. Bu profilde yer alan işletmelerin işletme sermayesinin net satışlara oranı 0,09 ile 0,79 arasında iken, stok devir hızları -2,72'den büyüktür. İşletmelerin, satışlarından çok özsermayesine bağlı kalması ve stok devir hızının düşük olması kredi veren kurumlar açısından riskli bir durum oluşturmaktadır. İşletmelerin stok devir hızını arttırıcı önlemler alması gerekmektedir.

Profil 7 (Node 4): Toplam 17 işletmenin yer aldığı yedinci profili, işletme sermayesinin net satışlara oranı 0,79'dan büyük olan işletmeler oluşturmaktadır. Bu profildeki işletmelerin %41,2'si (7 işletme) sanayi sektöründe, %58,8'i (10 işletme) hizmet sektöründe yer almaktadır. İşletme sermayesinin net satışlara oranının çok yüksek olması işletmelerin satışlarının çok düşük olduğunu ya da sermayesinin ihtiyacın çok üzerinde olduğunu göstermektedir. Karlılığını satışlarından çok sermayeye bağımlı olması işletmelerin faaliyetlerini devam ettirmesi ve kredibilitesini koruması bakımından yüksek risk oluşturmaktadır.

SONUÇ

Verinin bilgiye dönüşümünün geleneksel yöntemi, klasik çözümleme ve yorumlamaya dayanmaktadır. Ancak günümüzde veri miktarındaki olağanüstü artış, sağlık, eğitim, ticaret, askeri alanlar, alışveriş, devlet sektörü, özel sektör ve pek çok alanda verilerin işlenmesi ve bu verilerin değerlendirilerek bilgiye dönüştürülmesi bir zorunluluk haline gelmiştir. Veri madenciliği bu noktada uygun teknikler kullanarak gizli kalmış bilgileri ortaya çıkarma özelliğiyle gereklidir.

Bu çalışmada, veri madenciliği tekniklerinden karar ağaçları tekniği kullanılarak İstanbul Menkul Kıymetler Borsası'ndan elde edilen veriler üzerinde yöntemin uygulanabilirliği gösterilmiştir. Karar ağacı algoritmalarının en önemli avantajları, parametrik olmayan yöntemler arasında olması nedeniyle diğer çok değişkenli tekniklerde sağlanması gereken istatistik varsayımların söz konusu olmamasıdır. Çünkü veri madenciliği sonucunda oluşturmuş olduğumuz tüm kurallar kesinlikle kullanılabilir veya kullanılamaz diye bir yargıya önceden varmamız mümkün değildir. Ayrıca karar ağacı algoritmalarının bağımlı ve bağımsız değişkenler arasındaki ilişkilerin yönünü, önem sırasını görselleştirmesi diğer avantajları arasındadır. Bu özelliği özellikle elde edilen sonuçların yorumunu oldukça basitleştirerek daha somut ve kullanışlı hale getirmektedir.

Çalışma kapsamına İstanbul Menkul Kıymetler Borsası Ulusal 100 endeksi sanayi ve hizmet sektörlerinde faaliyet gösteren 173 şirket alınmış ve şirketlere ait finansal bilgiler kullanılarak elde edilen verilere CHAID algoritması uygulanmış işletmelerin birbirlerine göre konumları ortaya konmuştur. CHAID ile diğer algoritmalar arasındaki en önemli farklılıklardan birisi, ağacın yapraklarının ikili değil verideki farklı yapı sayısı kadar dallanmasıdır. Bu özelliği nedeniyle, daha fazla alt gruplarla değerlendirme yapmak ve daha homojen gruplardan sonuç çıkarmak mümkün olmaktadır. Bu bağlamda bu çalışmada İMKB100 sanayi ve hizmet sektöründe faaliyet gösteren firmaları ayırmada etkili olan en önemli karlılık, likitide, varlık kullanım etkinliği, sermaye yapısı ve işletme büyüklüğü değişkenleri araştırılmıştır.

Çalışmada incelenen ve yorumlanan en uygun ağaç yapısına göre elde edilen bulguların değerlendirilmesi sonucunda şu hususlar ortaya konmuştur:

Çalışmada, sanayi ve hizmet sektörlerinde faaliyet gösteren firmaları ayırmada etkili olan değişkenler işletme sermayesinin net satışlara oranı, stok devir hızı ve ekonomik rantabilite oranı değişkenlerinden oluşmaktadır. 1. ana dalda işletme sermayesinin net satışlara oranı negatif olması, 1. profildeki 17 işletmenin (sanayi: 2, hizmet:15) kısa vadeli borçların yüksek olduğunun göstergesidir ve yükümlülüklerini karşılayabilme riskini arttırmaktadır. 2. ana daldaki 35 işletmenin (sanayi: 25, hizmet: 10) oluşturduğu 2. profilde, işletme sermayesinin net satışlara oranı değerinin yükselmesiyle birlikte işletme sermayesi satışlara bağımlı olmaktan çıkmakta ve satışlar artmaktadır. 3. ana daldaki toplam 104 işletmenin (sanayi: 93, hizmet: 11) yer aldığı profillerde elde edilen bulgular değerlendirildiğinde, 4. ve 5. profillerde ekonomik rantabilite oranının negatif olması işletme faaliyetlerinin finansmanında ağırlıklı olarak yabancı kaynak kullanıldığını göstermektedir. 3., 4. ve 5. profillerde stok devir hızının negatif olması ise işletmelerde fazla stok bulundurulduğunu, satışların net işletme sermayesine bağlı olduğunu ve bu nedenle işletme satışlarının azaldığını göstermektedir. 4. ana dalda yer alan 7. profildeki 17 işletmenin (sanayi:7, hizmet:10), ihtiyaçlarının çok üzerinde işletme sermayesi bulunmakta ve kredi verecek kurumlar tarafından riskli bir yapıyı oluşturmaktadır.

Çalışmanın sonuçlarıyla görüldüğü gibi karar ağaçları tekniğiyle işletmelerin birbirlerine göre konumları ortaya konmuş ve sektör değişkenini etkileyen en önemli değişkenler saptanmıştır.

Sınıflandırmanın geniş olması ve sonuçların görselliği ve kolay yorumlanabilir olması nedeniyle bulgular karar ağacı tekniğinin ihtiyaca yönelik cevapları ortaya çıkarmakta etkili olduğunu göstermektedir. Bununla birlikte karar ağaçlarıyla ortaya çıkan sonuçlar farklı yöntemlerde veri ve değişken olarak da kullanılabilmekte ve bu yolla yeni bilgiler elde edilebilmektedir.

KAYNAKÇA

1. AGRAWAL, R., H. MANNILA, R. SRIKANT, H. TOIVONEN, ve A. I. VERKAMO, "Fast Discovery of Association Rules, Advances in Knowledge Discovery and Data Mining", **AAAI/MIT Press**, Chapter 12, 1995.
2. AITKENHEAD, M. J., "A Co-Evolving Decision Tree Classification", **Expert Systems with Applications,** 34, No. 1, 2008.
3. AKPINAR, Haldun, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İstanbul Üniversitesi İşletme Fakültesi Dergisi,** 29, No. 1, 2000.
4. ALTINIŞIK, Umut, "Öğrenci Bilgi Sisteminde Veri Madenciliğinin Uygulanması", Yayınlanmamış Yüksek Lisans Tezi, **Kocaeli Üniversitesi Fen Bilimleri Enstitüsü,** Kocaeli, 2006.
5. AYDOĞAN, Fatih, "E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi", Yayınlanmamış Yüksek Lisans Tezi, **Hacettepe Üniversitesi Fen Bilimleri Enstitüsü,** Ankara, 2003
6. BERRY, Michael J., "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management", **John Wiley and Sons Incorporated**, USA, 2004
7. BERSON, A., S. SMITH, ve K. THEARLING, "Building Data Mining Applications for CRM", **McGraw Hill,** USA, 1999.
8. DOĞAN, Nurhan Ve Kazım ÖZDAMAR, "Chaid Analizi ve Aile Planlaması ile İlgili Bir Uygulama", **Türkiye Klinikleri Tıp Bilimleri Dergisi**, 23, No. 5, 2003.
9. DOLGUN, M. Özgür, "Türkiye'deki Hastanelerin Veri Madenciliğiyle Gruplandırılması", **XI. Biyoistatistik Kongresi,** 27 – 30 Mayıs, Malatya, 2008.
10. EDELSTEIN, Herb, "Mining Large Database - A Case Study", **Two Crows Corporation,** 2000.
11. ELDER, J. F. ve D. PREGIPON, "A Statistical Perspective on KDD", **The 1st International Conference on Knowledge Discovery and Data Mining,** 20–21 Ağustos, Montreal, s. 87–93, 1995.
12. ERKEN, Arzum, "Başlıca Fiyat Bazlı Oranların Hisse Senedi Getirisi Üzerindeki Etkileri ve İstanbul Menkul Kıymetler Borsası'nda Bir Uygulama", Yayınlanmamış Yüksek Lisans Tezi, **Ankara Üniversitesi Sosyal Bilimler Enstitüsü**, Ankara, 1998.
13. FAYYAD, Usama, Gregory PIATETSKY-SHAPIRO ve Padhraic SMYTH, "From Data Mining to Knowledge Discovery in Databases", **AI Magazine,** 17, No.3, 1996.
14. FAYYAD, Usama, "Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases", **Ninth International**

Conference on Scientific and Statistical Database Management, 11–13 Ağustos, Olympia, 1997.

15. FU, Yongjian, "Data Mining Tasks, Techniques and Applications", No: 6648/97, **IEEE,** 1997.
16. GHOSH, A. ve B. NATH, "Multi-Objective Rule Mining Using Genetic Algorithms", **Information Sciences,** 163, No. 1–3, 2004.
17. GROBLER, B. R., T.C. BISSCHOFF ve K.C. MOLOI, "The Chaid-Technique and the Relationship between School Effectiveness and Various Independent Variables", **International Studies in Educational Administration,** 30, No. 3, 2002.
18. HAN, J. ve M. KAMBER, "Data Mining: Concepts and Techniques", **Morgan Kaufmann Publishers,** USA, 2001.
19. HAND, David, Heikki MANILA ve Padhraic SYMTH, "Principles of Data Mining", **The MIT Press,** London, 2001.
20. HASTIE, T., R TIBSHIRANI, ve J. FRIEDMAN, "The Elements of Statistical Learning; Data Mining, Inference and Prediction", **Springer Series in Statistics,** USA, 2001.
21. HAUGHTON, Dominique ve Samer OULABI, "Direct marketing modeling with CART and CHAID", **Journal of Direct Marketing**, 11, No. 4, 1999.
22. HEGLAND, Markus, "Data Mining Techniques", **Acta Numerica,** Cambridge University Press, 2001.
23. İMAMOĞLU, Tuba, "Veri Madenciliğinde Karar Ağaçları ile Bir Öğrenci Ders Başarısı Tahmin Aracı", Yayınlanmamış Yüksek Lisans Tezi, **Kocaeli Üniversitesi Fen Bilimleri Enstitüsü**, Kocaeli, 2005.
24. İMKB (2007); www.imkb.gov.tr, (Erişim Tarihi:06.04.2007).
25. KLEISSNER, Charly, "Data Mining for Enterprise", **31st Annual Hawai International Conference on System Screnes,** 1060–3425/98, 1998.
26. KOLTAN YILMAZ, Şebnem, "Veri Madenciliği: İstanbul Menkul Kıymetler Borsası Örneği," Yayınlanmamış Yüksek Lisans Tezi (Danışman: Yrd.Doç.Dr. Ali Sait ALBAYRAK), ZKÜ SBE, Zonguldak, 2008.
27. KOYUNCUGİL, Ali Serhan, "Borsa Şirketlerinin Sektörel Risk Profillerinin Veri Madenciliğiyle Belirlenmesi", **Sermaye Piyasası Kurulu Araştırma Raporu,** Araştırma Dairesi, Ankara, 2007.
28. KOYUNCUGİL, Ali Serhan ve Nermin Özgülbaş, "İMKB'de İşlem Gören KOBİ'lerin Finansal Başarısızlığına Etki Eden Faktörlerin Veri Madenciliği İle Belirlenmesi," **3. KOBİ ve Verimlilik Kongresi**, 17–18 Kasım, İstanbul, 2006.
29. LAROSE, Daniel T., "Discovering Knowledge in Data: An Introduction to Data Mining", **John and Wiley Sons Incorporated,** USA, 2005.

30. LAURA, Galguera, David LUNA ve M. Paz MENDEZ, "Predictive Segmentation in Action: Using CHAID to Segment Loyalty Card Holders", **International Journal of Market Research,** 48, No. 4, 2006.

31. MCCARTY, John A. ve Manoj HASTAK, "Segmentation Approaches in Data-Mining: A Comparison of RFM, CHAID and Logistic Regression", **Journal of Business Research,** 60, No. 6, 2007.

32. MOSS, L.T. ve S. ATRE, "Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications", **Addison-Wesley Publishing,** USA, 2003.

33. OLAFSSON, Sigurdur, Xiaonan LI ve Shuning WU, "Operations Research and Data Mining", **Europen Journal of Operational Research,** 187, No. 3, 2008.

34. ÖZEKES, Serhat, "Veri Madenciliği Modelleri ve Uygulama Alanları", **İstanbul Ticaret Üniversitesi Dergisi,** 2, No. 3, 2003

35. IİATETSKY-SHAPIRO, Gregory, "Knowledge Discovery in Real Databases: A Workshop Report", **AI Magazine,** 11, No. 5, 1991.

36. PIRAMUTHU, Selwyn, "Evaluating Feature Selection Methods for Learning in Data Mining Applications", **European Journal of Operational Research,** 156, No. 2, 2004.

37. ROIGER, R. J. ve M. W. Geatz, "Data Mining A Tutorial-Based Primer", **Addison Wesley,** USA, 2003.

38. SIMOUDIS, Evangelos, "Reality Check for Data Mining", **IEEE Expert: Intellegent Systems and Their Applications**, 11, No. 5, 1996.

39. SPSS, "AnwerTree Algorithm Summary", **SPSS White Paper,** USA, 1999.

40. SUN, Jie ve Hui LI, "Data Mining Method for Listed Companies, Financial Distress Prediction", **Knowledge-Based Systems,** 21, No. 1, 2008.

41. TÜRE, Mevlut, Füsun TOKATLI ve İmran KURT, "Using Kaplan-Meirer Analysis Together With Decision Tree Methods (C&RT, CHAID, QUEST, C4.5 and ID3) In Determining Recurrence-Free Survival of Breast Cancer Patients", **Expert Systems With Applications,** Article in Pres, 2008.

42. VAHAPLAR, Alper, "Bir Coğrafi Veri Madenciliği Uygulaması", Yayınlanmamış Yüksek Lisans Tezi, **Ege Üniversitesi Fen Bilimler Enstitüsü**, İzmir, 2003.

43. VAN DIEPEN, Merel ve Philip Hans FRANSES, "Evaluating Chi-Squared Automatic Interaction Detection", **Informatıon Systems,** 31, No. 8, 2006.

44. ZIARKO, W., "The Discovery, Analysis, and Representation of Data Dependencies in Databases, Knowledge Discovery in Databases", **AAAI/MIT Press,** 1991.