İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE BAĞLILIK

YÜKSEK LİSANS TEZİ

Yıldız KOÇTÜRK

Anabilim Dalı : Bilgisayar Mühendisliği

Programı : Bilgisayar Mühendisliği

Tez Danışmanı: Prof. Dr. Eşref ADALI

HAZİRAN 2010

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE BAĞLILIK

YÜKSEK LİSANS TEZİ

Yıldız KOÇTÜRK
(504031532)

Tezin Enstitüye Verildiği Tarih : 07 Mayıs 2010
Tezin Savunulduğu Tarih : 14 Haziran 2010

Tez Danışmanı : Prof. Dr. Eşref Adalı (İTÜ)
Diğer Jüri Üyeleri : Prof. Dr. Coşkun SÖNMEZ (YTÜ)
Doç. Dr. Zehra ÇATALTEPE (İTÜ)

HAZİRAN 2010

ÖNSÖZ

Tez çalışmam süresince bana yol göstererek, yardımlarını esirgemeyen hocam Sayın Prof. Dr. Eşref Adalı' ya, tüm hayatım boyunca her konuda benden manevi desteklerini esirgemeyen anneme, babama, ablalarıma, tüm aileme ve desteklerini benden esirgemeyen tüm arkadaşlarıma sonsuz teşekkürlerimi sunarım.

Mayıs 2010

Yıldız Koçtürk

Bilgisayar Mühendisi

İÇİNDEKİLER

Sayfa

ÖNSÖZ ..... **iii**
İÇİNDEKİLER ..... **v**
KISALTMALAR ..... **vii**
ÇİZELGE LİSTESİ ..... **ix**
ŞEKİL LİSTESİ ..... **xi**
ÖZET ..... **xiii**
SUMMARY ..... **xv**
1. GİRİŞ **1**
1.1 Giriş ve Çalışmanın Amacı 1
1.2 Literatür Özeti 2
2. VERİ TABANLARINDA BİLGİ KEŞFİ **5**
2.1 Veri Tabanlarında Bilgi Keşfi Aşamaları 6
3. VERİ MADENCİLİĞİ **7**
3.1 Veri Madenciliği Tanımı 7
3.2 Veri Madenciliği Süreci 8
3.3 Veri Madenciliği Kullanım Alanları 9
3.4 Veri Madenciliği Modelleri 11
3.4.1 Tahmin edici modeller 12
3.4.2 Tanımlayıcı modeller 12
4. MÜŞTERİ İLİŞKİLERİ YÖNETİMİ **13**
4.1 Müşteri İlişkileri Yönetiminde Müşteri Kavramı 13
4.1.1 Müşteri için değer yaratma 15
4.1.2 Müşteri ömür boyu değeri 17
4.1.3 Müşteri tatmini 18
4.1.4 Müşteri sadakati 20
4.1.5 Yeni müşteri kazanmak ve eldeki müşteriyi tutmak 22
5. KARAR AĞAÇLARI **25**
5.1 Karar Ağacı Tanımı 25
5.2 Karar Ağacı Oluşturma 25
5.2.1 Böl ve elde et (divide and conquer) 25
5.2.2 ID3 algoritması 26
5.2.3 C4.5 algoritması 29
5.2.3.1 Bölünme – dallanma bilgisi 29
5.2.3.2 Sayısal özellikler 29
5.2.3.3 Kayıp veriler 30
5.3 Ağacın Testi 30
5.4 Ağaç Budama Metotları 31
6. VERİ MADENCİLİĞİNDE KULLANILAN YAZILIMLAR **35**
6.1 SPSS Clementine 35

6.2 STATISTICA Data Miner 35
6.3 Darwin 36
6.4 DBMiner 36
6.5 Enterprise Miner 37
6.6 Intelligent Miner 37
6.7 WEKA 37
7. UYGULAMA 39
7.1 Veri Kümesinin Tanımlanması 39
7.2 Veri Temizleme ve Dönüştürme 42
7.3 Veri Modelleme ve Karar Ağacı Algoritmasının Uygulanması 42
8. SONUÇ VE TARTIŞMA 69
KAYNAKLAR 71

KISALTMALAR

CRM	**:** Customer Relationship Management
EM	**:** Enterpise Miner
OLAP	**:** Online Analytical Processing
OLAM	**:** Online Analytical Mining
TARP	**:** Technical Assistance Research Programs
VTBK	**:** Veri Tabanından Bilgi Keşfi
WEKA	**:** Waikato Environment for Knowledge Analysis

ÇİZELGE LİSTESİ

Sayfa

Çizelge 7.1 : Veri kümesinde kullanılan nitelikler ve açıklamaları 40
Çizelge 7.2 : Veri kümesinde kullanılan niteliklerin özellikleri 41
Çizelge 7.3 : Eğitim ve test kümesi büyüklük değişimine göre doğruluk oranları .. 43
Çizelge 7.4 : Chi-Square dallanma yöntemi kullanan karar ağacında düğümler 44
Çizelge 7.5 : Entropy dallanma yöntemi kullanan karar ağacında düğümler 45
Çizelge 7.6 : Gini index dallanma yöntemi kullanan karar ağacında düğümler 46
Çizelge 7.7 : Karar ağacı yakınsaklık matrisi 48
Çizelge 7.8 : Entropy-3 dallanma yöntemi kullanan karar ağacında düğümler 49
Çizelge 7.9 : Entropy-4 dallanma yöntemi kullanan karar ağacında düğümler 50
Çizelge 7.10 : Entropy tabanlı karar ağaçlarının yakınsaklık matrisi 51

ŞEKİL LİSTESİ

Sayfa

Şekil 3.1 : Veri madenciliği uygulama alanları 11
Şekil 4.1 : Müşteriye sunulan değerlerin belirleyicileri (Kotler, 2005)..... 16
Şekil 4.2 : Müşteri karlılığında 80/20/30 kuralı (Günaydın, 2005) 23
Şekil 5.1 : Hunt' ın ağaç oluşturma metodu 26
Şekil 5.2 : ID3 algoritmasının akış diyagramı 28
Şekil 5.3 : Eğitim kümesi ile test kümesinin ağaç boyuna göre doğruluk performansı 32
Şekil 7.1 : Chi-Square hatalı sınıflandırma oranı – yaprak sayısı dağılımı 44
Şekil 7.2 : Entropy hatalı sınıflandırma oranı – yaprak sayısı dağılımı 46
Şekil 7.3 : Gini index hatalı sınıflandırma oranı – yaprak sayısı dağılımı 47
Şekil 7.4 : Entropy-3 hatalı sınıflandırma oranı – yaprak sayısı dağılımı 49
Şekil 7.5 : Entropy-4 hatalı sınıflandırma oranı – yaprak sayısı dağılımı 50
Şekil 7.6 : Entropy-4 dallanma yöntemini ile oluşturulan karar ağacı 53

VERİ MADENCİLİĞİNDE BAĞLILIK

ÖZET

Veri madenciliği, büyük veri kümeleri içindeki anlamlı bilgiyi ortaya çıkarma sürecidir. Veri madenciliğinin en yaygın kullanıldığı alanlardan biri, ayrılma eğilimi gösteren müşteri kesitini belirlemedir. Ayrılma eğilimini gösteren müşteri kesitini belirleme, şirketlerin bu müşterilere özel kampanyalarını geliştirmelerini sağlamaya yöneliktir.

Yapılan bu çalışmada, bireysel emeklilik müşterilerinin çeşitli niteliklerdeki bilgilerini inceleyerek, kaybedilmiş müşteri profili veri madenciliği yöntemleriyle ortaya çıkarılmaya çalışılacaktır.

Çalışmada, karar ağaçları yöntemi kullanılmıştır. Karar ağaçları ile veri kümesindeki olaylar özelliklerine göre sorgulanır ve kurallar oluşturulur. Amaç, birbirinden bağımsız gözüken veriler arasındaki ilişkiyi çözmektir. Veri kümesinde gizlenmiş kurallar mevcuttur ve karar ağacı oluşturulurken bu kurallar ortaya çıkarılır. Karar ağacının doğruluğunu hesaplamak için ilgili veri kümesi ikiye bölünür. Bu kümelerden biri ile ağaç oluşturulur ve diğeri ile ağaç test edilir. Böylelikle karar mekanizmasına ne kadar güvenileceği bilgisi elde edilir.

İnceleme sırasında 26.858 adet müşterinin bilgileri üzerinde bu yöntemle çalışmalar yapılmıştır. Bu yöntemle ortaya çıkan kurallar test edilerek doğruluk oranları ortaya çıkarılmış, bunlar istatiksel tablolarla göz önüne konmuştur.

Karar ağaçları genelde bu tip konularda sebep ortaya çıkarma konusunda oldukça güçlüdür. Bu sebeple, çalışmanın en son bölümünde kural tabloları incelenerek, müşteri kayıplarının sebepleri ve ne zaman gerçekleştiği bilgisine ulaşılmaya çalışılacaktır.

CHURN ANALYSIS USING DATA MINING TECHNIQUES

SUMMARY

Data mining is the process of finding hidden and unknown patterns in huge amount of data. Data mining has a wide application area such as marketing, banking and finance, medicine and manifacturing. Churn analysis is a common application area of data mining. Churn modeling is predicting which customer will leave the company. This allows companies to increase customer loyalty by producing special campaings.

In this study, some data attributes of Individual Pension Savings and Investment System customers are investigated and churned customer profile is stated by using data mining methods.

The decision tree method is used in this study. The instances in data sets are inquired according to their attributes and the rules are created. The aim of this inquiry is to figure out the relationship between data that seem to be totally distinct from each other. In the data sets some rules are hidden inherently; and in the course of creating the decision tree, these rules are revealed. In order to calculate the accurancy of the desicion tree, the related data set is divided into two. One of these sets serves the creation of the tree and the other is used to test it.

Total number of data, which is examined through this analysis, is 26858. The rules obtained from the analysis are tested on the test data and the error and correctness rates are found and statistically measured.

Desicion tree algorithms are powerful to find the causes of certain problems associated with their human – readable rule tables. These rule tables are examined to learn the cause, and time of customer churn.

1. GİRİŞ

1.1 Giriş ve Çalışmanın Amacı

Globalleşme ve artan rekabet, müşteri satın alma davranışlarını ve beklentilerini, büyük ölçüde değiştirmiştir. Değişen müşteri davranışlarını, işletme amaç ve hedeflerine uygun olarak yönetebilmek; değişim yönetimi ve müşteri ilişkileri yönetimi kavramlarını anlayabilmek ve doğru bir şekilde yorumlayabilmek ile mümkün olabilecektir. İşletmelerde başarılı bir müşteri ilişkileri yönetimiyle, küresel bir dünyada ve giderek artan rekabetçi piyasalarda, işletmeler için yaşamsal önem taşıyan, müşteri için değer yaratmak, müşteri bağlılığını sağlamak ve bu konularda kurumsallaşmayı gerçekleştirmek mümkün olabilecektir.

Bilgi sistemleri ve iletişim teknolojilerindeki gelişmeler sayesinde tıp ve sağlık alanındaki birçok veri sayısal ortamda saklanabilir ve kolaylıkla erişilebilir hale gelmiştir. Kayıt altına alınan veriler sayesinde organizasyonlar veri yığınlarını anlamlandırarak ve içerisindeki gizli bilgileri tespit ederek organizasyon yararına bilgiler elde edebilmektedirler. Bu bilgileri elde etmek klasik sorgulama yöntemleriyle veya basit gözlemlerle elde edilemeyeceği için geliştirilmiş olan veri madenciliğinden faydalanılmaktadır.

Veri madenciliği büyük verilerden yararlı örüntüler çıkarma sürecidir. Birçok veri madenciliği yöntemi bulunmaktadır. Ulaşılmak istenen bilgi ve veri şekline en uygun veri madenciliği yöntemi seçilerek uygulanmaktadır.

Veri madenciliğinin en yaygın kullanıldığı uygulama alanlarından biri, ayrılarak rakip şirketi tercih etme eğilimi gösteren müşteri profilini belirlemedir. Ayrılacak müşteri tahmininin temel amacı karlı müşterileri elde tutmaktır. Hangi müşterilerin ayrılma eğilimi gösterdiği tespit edildikten sonra, bu müşteriler üzerinde belirli bir pazarlama faaliyeti hedeflenerek müşteri tutmak için etkin stratejiler geliştirmek mümkündür.

Burada yapılan çalışmanın amacı, bireysel emeklilik sistemi verilerine veri madenciliği tekniklerinden biri olan karar ağacını uygulayarak, bağlılığını yitirme

eğiliminde olan müşterilerin önceden tespit edilmeye çalışılmasıdır. Çalışma kapsamında; sektöründe Türkiye' de ilk 5 arasında yer alan şirkete ait bireysel emeklilik sistemi verileri kullanılmıştır. Müşterilerin sistemdeki statü bilgisinden yola çıkarak sonuçlara ulaşılmıştır. Böylelikle, müşterinin bağlılığını neden yitirdiği bilgisinin yanı sıra, hangi özelliklere sahip müşterilerin bağlılıklarını daha sık kaybettikleri tahmin edilmeye çalışılmıştır.

1.2 Literatür Özeti

Veri madenciliği 1990' lı yıllarda ortaya çıkmıştır. Veri madenciliğinin özellikle 2000 yılından bu yana büyük bir gelişme gösterdiği göze çarpmaktadır. Aşağıda, veri madenciliği müşteri bağlılığı konusunda farklı alanlarda gerçekleştirilen uygulama örnekleri yer almaktadır.

Jonker ve diğ. (2003), CHAID ve Markov karar sürecini kullanarak uzun dönem karlılığı için birleşik müşteri segmentasyonu ve pazarlama politikası çalışması gerçekleştirmişlerdir.

Hsieh (2004), bir bankanın kredi kartı müşterileri için sinir ağları ve birliktelik kurallarını kullanarak bir davranışsal skorlama modeli yaratmış; müşterileri davranış ve karakteristiklerine göre farklı gruplara ayırarak her grubun karakteristiklerine uygun yönetim stratejileri önermek suretiyle müşteri sadakatini arttırmayı hedeflemiştir.

Hwanng ve diğ. (2004), müşterinin geçmişte yarattığı potansiyel değeri ve rakip şirketi tercih etme olasılığını dikkate alarak, lojistik regresyon, sinir ağları ve karar ağacı tekniklerini kullanmak suretiyle bir müşteri yaşam ömrü modeli geliştirmişlerdir.

Chen ve diğ. (2005), perakende endüstrisinde etkin pazarlama kampanyaları düzenlemek amacıyla davranışsal müşteri verileri madenciliği çalışmasını gerçekleştirmişlerdir.

Liu ve Shih (2006), AHP ve birliktelik kuralı yöntemlerini birleştiren bir öneri methodolojisi geliştirerek müşterileri yaşam ömrü değerlerine göre segmentlere ayırmışlardır.

Bayram (2001), veri madenciliği tekniklerinden karar ağaçları, kümeleme analizleri ve lojistik regresyon analizlerini kullanarak kablosuz iletişim endüstrisinde müşteri segmentasyonu ve ayrılacak müşterilerin tahmin modelini ortaya koymuştur.

Tosun (2006), Yapı Kredi Bankası'nın kredi kartı müşterilerinin çeşitli niteliklerdeki bilgilerini incelemiş, kaybedilmiş bir müşterinin profilini veri madenciliği yöntemleriyle ortaya çıkarmaya çalışmıştır.

Gazi (2007), GSM operatörleri tarafından yapılan kampanyaların, cep telefonu kullanıcıları üzerindeki etkisini analiz etmiştir. Kampanya dönemine ait müşteri bilgileri ile kampanyasız döneme ait müşteri bilgilerinin karşılaştırılmasını yapıp, müşterilerin yapılan kampanyaya olan duyarlılığını analiz etmiştir. Müşterilerin yapılan kampanya dönemine ait verileri ile kampanyasız döneme ait verilerini, karşılaştırmıştır. Bu verilere dayanarak parametrelerin birbirlerine karşı olan duyarlılıkları ile tüm veri içindeki duyarlılığını hesaplamıştır.

Akbulut (2007), bir kozmetik markasının müşteri gruplarını ve ayrılma eğilimi gösteren müşteri profilini belirlemiş; bu müşterilere özel pazarlama stratejileri geliştirilmesini hedeflemiştir. Segmantasyon için kümeleme tekniklerini, ayrılacak müşteri profilini belirlemek için sınıflama tekniklerini kullanmıştır.

Bin ve diğ. (2007), karar ağaçları algoritmasını kullanarak telekomünikasyon sektöründe müşteri kaybetme analizi yapmışlardır. Eğitim datasını farklı alt periyotlara bölerek ve büyüklüğünü farklılaştırarak doğru sınıflandırmaya etkisini incelemişlerdir.

Fontana (2008), karar ağaçları algoritmaları ile öğrenme temelli algoritmaların birlikte kullanıldığında daha yüksek performans göstereceğini söyleyerek, WEKA ile örnek veriler üzerinde çalışarak bu modeli oluşturmuştur.

Qi ve diğ. (2008), müşteri kaybetme analizinde ADTrees ve lojistik regresyon modelinin bir arada kullanılmasının daha başarılı sonuçlar vereceğini ortaya çıkarmaya çalışmıştır.

Wang ve diğ. (2009), kablosuz ağ şirketinde müşteri sadakatini sağlamak amacıyla karar ağaçları algoritmasını kullanmışlardır. İlk 9 haftalık veri eğitim, 2 aylık veri de test amaçlı kullanılmıştır.

Xin ve diğ. (2009), Least Square Vector Machine ve Principal Component Analizini bir arada kullanarak müşteri sadakatini ölçümlemişlerdir.

Popovic ve Basic (2009), bir bankaya ait veriler üzerinden yaptıkları müşteri kaybetme analizinde fuzzy c-mean (FCM) bulanık methodunu kullanarak k en yakın ortalama algoritmasına göre doğruluk seviyesi daha yüksek sonuçlar verdiğini ortaya çıkarmıştır.

Huang ve diğ. (2010), telekomünikasyon sektöründe çoklu özellik seçimi için önerdikleri yöntem ile NSGA-II methodunu optimize etmişlerdir. Kümelerden birden fazla lokal özellik seçip, baskın olmayan özelliklerin arama yöntemi ile çıkarılması yaklaşımını benimsemişlerdir.

2. VERİ TABANLARINDA BİLGİ KEŞFİ

Bilgi toplumlarının temel hammaddesi bilgidir. Bilgisayar ve iletişim teknolojileri geliştikçe bilginin üretilmesi, taşınması ve depolanması kolaylaşmıştır. Miniwatts' a (2006) göre elektronik ortamların gün geçtikçe yaygınlaşması ve kullanımının artmasıyla birlikte bilgi miktarında da hızlı bir artış gözlenmektedir.

Bilgisayarların ucuzlaması ile bilgisayar ve otomatik veri toplama araçları yaygın bir şekilde kullanılmaya başlamıştır. Bu yaygın kullanım sonucunda insanların yaptığı işlemler kayıt altına alınmakta ve veriler doğrudan sayısal olarak toplanmakta ve saklanmaktadır (Alpaydın, 2000). Bunun sonucu olarak da büyük ölçekli işletimsel veriler oluşmaktadır.

Çok çeşitli alanlarda ve organizasyonlarda biriken veri yığınları içinden anlamlı, değerli, önceden bilinmeyen bilgiyi keşfetmek için geleneksel sorgulama veya raporlama araçlarının yetersiz hale gelmesi, verilerin toplanması ve etkileyici şekilde birikmesi karşında, hızla büyüyen dijital verileri kullanışlı bilgilere dönüştürmede insanlara yardımcı olacak, yeni nesil hesaplama teorilerine ve araçlara duyulan acil ihtiyaç, Veri Tabanından Bilgi Keşfi (VTBK) adı altında yeni arayışlara neden olmaktadır. Aktif araştırma alanlarından biri olan VTBK disiplini, çok büyük boyutlu verileri tam ya da yarı otomatik bir biçimde analiz eden yeni kuşak araç ve tekniklerin üretilmesi ile ilgilenen son yılların gözde araştırma konularından biridir (Sever ve Oğuz, 2000). VTBK uygulamaları, faaliyet alanına yönelik karar destek mekanizmaları için gerekli ön bilgileri temin etmek için veri madenciliğini kullanır (Uskurlu, 2006).

VTBK süreci verilerden kullanışlı tüm bilgi keşif süreçlerini kapsamaktadır. VTBK süreçlerinde ek olarak veri hazırlama, veri seçimi, veri temizleme, daha önce elde olan uygun bilgiler ile birleştirme ve madencilik sonuçlarının uygun yorumu, verilerden faydalı bilgi çıkarımını sağlamak için gereklidir (Çetinyokuş ve Gökçen, 2008). Veri madenciliği, bu aşamaların en önemli kısmını oluşturur.

Uygulanacak veri madenciliği algoritması ne kadar etkin olursa olsun, inceleme yapılacak işin ve verilerin özelliklerinin bilinmemesi durumunda sağlıklı ve faydalı bilgi elde edilmesi mümkün değildir.

2.1 Veri Tabanlarında Bilgi Keşfi Aşamaları

Fayyad (1997), VTBK aşamalarını, veriyi anlama, hedef veri seti oluşturma, veri temizleme ve önişleme, veri indirgeme, amaçları seçme, model analizi ve hipotez seçme, veri madenciliği, örüntüleri yorumlama, bilgi üzerinde herekete geçme olarak 9 aşamada incelemektedir. Fu (1997)' ya göre VTBK aşamaları, veri temizleme, veri dönüştürme, veri madenciliği, modelleme, yorumlama ve değerlendirme aşamalarından oluşmaktadır. Veri madenciliği bu incelemelerde VTBK' nin odak noktasıdır. Fayyad (1997), VTBK' nin veri hazırlama, örüntü arama, veri analizi, bilgi değerlendirme, yenileme gibi yinelemeli adımlardan oluştuğunu belirtmiş, veri madenciliğini ise bilgi ya da örüntülerin algoritmik anlamıyla ilgilenen VTBK aşaması olarak değerlendirmiştir. Hegland (2001), VTBK' ni veri sorgulama, veri temizleme, veri analizi, bilginin sunumunu içeren prosedürler olarak ele almıştır.

3. VERİ MADENCİLİĞİ

3.1 Veri Madenciliği Tanımı

Veri madenciliği genel anlamda; büyük miktarda veri içerisinden, gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılması biçiminde ifade edilmektedir. Veri madenciliğinde amaç, toplanmış olan bilgilerin bir takım istatistiksel yöntemler ile incelenip ilgili kurum ve yönetim destek sistemlerinde kullanılmak üzere değerlendirilmesidir.

Literatürde, veri madenciliği için yapılan tanımlamalardan birkaçı şöyledir: Fayyad'a (1996) göre veri madenciliği önceden bilinmeyen, veri içinde gizli, anlamlı ve yararlı örüntülerin büyük ölçekli veri tabanlarından otomatik bir biçimde elde edilmesini sağlayan VTBK süreci içinde bir adımdır. Gartner Group' a göre ambarlarda depolanmış verileri, istatiksel ve matematiksel teknikler yardımıyla inceleyerek anlamlı yeni korelasyonların, örüntülerin ve eğilimlerin ortaya çıkarılması sürecidir (Larose, 2005). Raghavan ve Sever' e (1994) göre ise, veri madenciliği büyük veri kümesi içinde saklı olan genel örüntülerin ve ilişkilerin bulunmasıdır. Deogun ve diğ.' ne (1997) göre veri madenciliği, makine öğrenimi, istatistik, veri tabanı yönetim sistemleri, veri ambarlama, koşut programlama gibi farklı disiplinlerde kullanılan yaklaşımları birleştirmektedir. Swift (2001), veri madenciliğini veri kaynağında tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş verileri ortaya çıkarmak, bunları karar vermek ve gerçekleştirmek için kullanma süreci olarak tanımlamaktadır. Hand (2001) ise büyük veri kümelerinden yararlı bilgilerin çıkarılması olarak ifade etmiştir (Shmueli, 2007).

Veri madenciliği, eldeki yapısız veriden, anlamlı ve kullanışlı bilgiyi çıkarmaya yönelik çalışmaların bütünü olmuştur. Yıllar ilerledikçe ortaya çıkan veri yığınlarına bir düzen verme, başka bir deyişle potansiyel olarak kullanışlı bilgi haline getirme amacıyla veri madenciliği ortaya çıkmıştır.

Veri madenciliği, büyük miktardaki veriler arasından gelecek ile ilgili tahminde bulunabilmemizi sağlayabilecek bağlantıların bilgisayar programı kullanarak aranmasını sağlayan bilgiye ulaşma işidir. Yani veri işlenerek amaçlanan bilgi elde edilir (Doğan, 2007).

3.2 Veri Madenciliği Süreci

Veri madenciliği, altı adımlı bir süreç olarak incelenebilir (Larose, 2005).

1. Araştırma Probleminin Tanımlanması (Business Understanding)

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

2. Veri 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, 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 veri madenciliği sürecinin en gösterişli aşamasıdır. Bu aşama uygun modelleme tekniğinin seçimi, test tasarımının üretimi, model geliştirme ve tahmin işlemlerini içermektedir. Uygun modellerin seçilip uygulanması ile birlikte parametreler en uygun değişkenlere dönüştürülmektedir. Veri madenciliği, her problem tipi için farklı yöntemler sunmaktadır. 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, veri madenciliği 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, veri madenciliği sonuçlarının kullanımı üzerindeki karara varılmaktadır.

6. Uygulama Aşaması (Deployment)

Son aşama olan uygulama aşaması, araştırmacının tüm emeklerinin karşılığını aldığı aşamadır. Bu aşamada veri madenciliği 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.

Veri madenciliğinde veri kümesinin büyüklüğünden kaynaklanan en fazla zaman alıcı aşama, verilerin ön işlemden geçirilmesidir. Veri madenciliği uygulamalarında kaynakların, %80' i verilerin ön işlemden geçirilmesi ve temizlenmesi için harcanmaktadır (Piramuthu, 2004).

3.3 Veri Madenciliği Kullanım Alanları

Veri madenciliği pazarlama, bankacılık, sigortacılık gibi sektörlerde yaygın olarak kullanılmaktadır. Karar verme süreçlerinde başarılı sonuçlar elde edildiği için taşımacılık-ulaşım-konaklama, eğitim öğretim, perakendecilik gibi konularda da kullanılabilir. Kullanım alanlarını genel olarak aşağıdaki gibi gruplayabiliriz:

İş ve Elektronik Ticaret Verileri: Her türlü ofis uygulamaları ve iş süreçleri sonucunda büyük çaplarda veriler üretilmektedir. Bunların karar verme süreçlerinde etkili olarak kullanılabilmesi ticari kuruluşların temel yapı taşlarından biri olmalıdır.

Bilimsel, Mühendislik Verileri: Günümüzde bilimsel veriler daha karmaşık hale gelmişlerdir. Bilim adamları ve mühendisler bu verileri kullanarak simülasyon ve sistem kulllanımının arttırılması hedefindedirler.

Web Verileri: İnternet ve web üzerindeki veriler hem hacim hem de karmaşıklık olarak hızla artmaktadır. Web verilerinde düz metin ve resimden başka akan (streaming) ve nümerik veriler gibi farklı yapılarda veriler de yer alabilmektedir (Vahaplar ve İnceoğlu, 2001).

Pazarlama:

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasında bağlantıların bulunması,
- Düzenlenen kampanyalarda cevap verme oranlarının arttırılması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepet analizi
- Müşteri ilişkileri yönetimi (CRM),
- Müşteri değerlendirmesi,
- Satış tahmini,

Bankacılık:

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

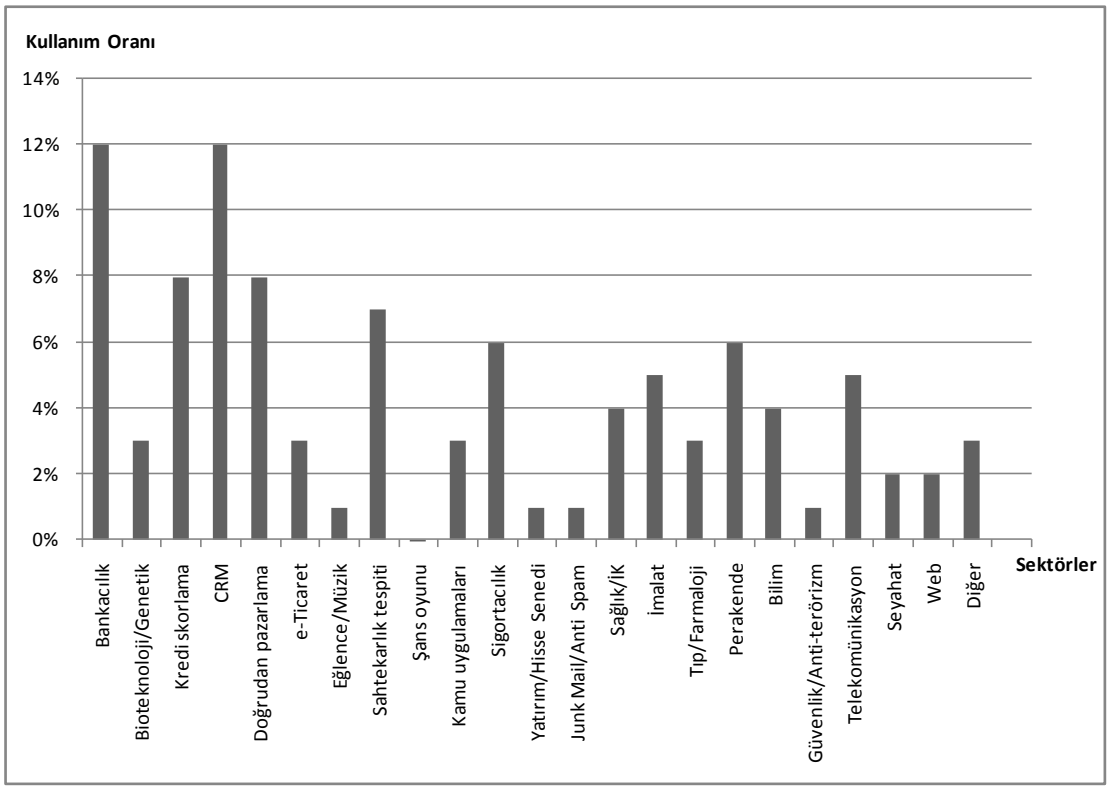
Sigortacılık:

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri örüntülerinin belirlenmesi (Akpınar, 2000),

Sağlık:

- Belirli bir hastalığa sahip kişilerin ortak özelliklerinin tahmin edilmesi,
- Tıbbi tedaviden sonra hastaların durumlarının tahmin edilmesi,
- Hastane maliyetlerinin tahmin edilmesi,
- Ölüm oranları ve salgın hastalıkların tahmin edilmesi (Yıldırım ve diğ., 2007)

Şekil 3.2' de veri madenciliğinin sektörde kullanım oranları ile ilgili özet bir tablo gösterilmektedir. Uygulama alanları incelendiğinde CRM, bankacılık ve pazarlama ilk sıralarda yer almaktadır.



Şekil 3.1 Veri madenciliği uygulama alanları.

3.4 Veri Madenciliği Modelleri

Veri madenciliği modelleri, tahmin edici (predictive) ve tanımlayıcı (descriptive) modeller olmak üzere ikiye ayrılır.

3.4.1 Tahmin edici modeller

Sonuçlanmış veri kümelerinden yola çıkarak model geliştirilmesinin ve bu modeli sonucu bilinmeyen veriler üzerinde uygulayarak bir tahmin sonucunun elde edilmesinin amaçlandığı modeldir. Örneğin, bir banka önceki dönemde kredi alan müşterilerinin verilerine sahiptir. Bu verilerde müşteri bilgileri bağımsız değişkenler olurken kredinin geri ödenip ödenmediği bağımlı değişkendir. Bu veriler incelenerek oluşturulan model yardımı ile gelecek dönemlerde kredi talebinde bulunan müşterilerin krediyi geri ödeyip ödemeyeceğinin tahmini yapılabilir. Tahmin edici model tabanlı geliştirilen modeller, Sınıflandırma (Classification) ve Regresyon (Regression) Modelleridir.

3.4.2 Tanımlayıcı modeller

Karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanmasını sağlayan modellerdir. Belirli bir gelir düzeyindeki çocuklu ailelerin ve bu gelir düzeyinden düşük çocuksuz ailelerin aynı satın alma eğilimlerinde olmalarının belirlenmesi tanımlayıcı modele bir örnektir. Tanımlayıcı modeli temel alan modeller, Kümeleme (Clustering), Ardışık Zamanlı Örüntüler (Sequential Pattern) ve Birliktelik Kuralları (Association Rules) Modelleridir.

4. MÜŞTERİ İLİŞKİLERİ YÖNETİMİ

4.1 Müşteri İlişkileri Yönetiminde Müşteri Kavramı

Müşteri kavramı, ticaretin ve değiş tokuşun başladığı ilk zamanlardan günümüze dek değişmeyen bir unsur olarak kalmış ve zamanla değişen pazarlama anlayışı ile birlikte işletmelerin en değerli öz kaynağı haline gelmeye başlamıştır. İşletmeleri ayakta tutacak en önemli varlıklardan birisi olan müşteriler; pazarlama, satış, üretim, hizmet, karlılık, zaman ve kaynak dağılımı gibi kavramların odak noktası olmuştur. Müşteriler, işletmelerin var oluş nedenlerinden ve devamlılığının sağlanmasındaki en önemli temel taşlarından birisidir. Bu durumda işletmelerin, varlıklarını sürdürmek için sürekli müşteri kazanma, mevcut müşterilerini tatmin etme ve elde tutma gibi her türlü çabayı harcamaları gerektiği görülmektedir. Son yılların en popüler işletme trendlerinden biri olan Müşteri İlişkileri Yönetimi kavramının ne olduğunu anlayabilmek için öncelikle müşteri kavramı ile ne ifade edilmeye çalışıldığının çok iyi anlaşılması gerekmektedir (Soyaslan, 2006). Çünkü Müşteri İlişkileri Yönetimi'nin en önemli ilgi alanı müşteridir.

Pazarlama literatüründe müşteri kavramı birçok farklı şekilde tanımlanmaktadır. Odabaşı ve Gülfidan (2003), müşteriyi, belirli bir mağazadan ya da işletmeden düzenli olarak alışveriş yapan kişi; bunun yanında ticari amaçla ürün ya da hizmetleri satın alan müşterileri ise ticari müşteri olarak tanımlamıştır. Başka bir tanıma göre ise müşteri, belirli bir işletmenin belirli bir marka malını, ticari veya kişisel amaçları için satın alan kişiler olarak tanımlamıştır (Taşkın, 2000). Bazı kaynaklarda ise müşteri, işletmelerin kendisi için ürettiği ürün ya da hizmetleri satın alan, ürün ya da hizmetlerin pazardaki yerini ve konumunu belirleyen, hedeflere ulaşmada işletmelere en önemli desteği veren kişi olarak tanımlanmıştır (Soyaslan, 2006). Bu tanımlamalar paralelinde genel bir müşteri tanımı yapılacak olursa, müşteri, ürün ya da hizmeti kabul eden kuruluş, kişi ya da kişiler olarak tanımlanabilir. Müşteri kavramı, genel olarak iki grupta incelenmektedir (Griffin, 1997):

Bireysel müşteri: Nihai ürünü alan ya da hizmetten faydalanan birey ya da ailelerdir.

Endüstriyel müşteri: Hizmet ya da organizasyonel karlılığını arttırmak amacıyla, bir işletmenin ürün ya da hizmetini satın alarak kendi ürününe gerekli eklemeleri yaptıktan sonra diğer tüketicilere satan müşterilerdir.

Müşteri kavramıyla ilgili iki kavram daha vardır: İç müşteri ve dış müşteri. Dış müşteriler, bir işletmeden ürün ve hizmet alan kişi ya da kuruluşlardır (Tek, 1999). İç müşteriler ise söz konusu ürün ve hizmetlerin üretilmesinde ve dış müşterilere ulaştırılmasında doğrudan veya dolaylı olarak katkı sağlayan tüm işletme çalışanlarıdır (Çoban, 2004). İç ve dış müşteri kavramlarının birbirinden ayrı düşünülmesi olanaksızdır. İşletmenin, yalnızca dış müşterilerin tatminine odaklanarak, yani iç müşterilerin ihtiyaç ve beklentilerini göz ardı ederek başarıya ulaşması mümkün değildir. Eğer iç müşteriler tatmin edilemiyorsa, dış müşterilerin tatmin edilmesini beklemek doğru olmayacaktır (Gündaş, 2003). Dış müşteri tatminini maksimize ederek işletmenin karını ve pazar payını artırabilmek için iç müşterilerin tatmin oldukları bir süreç yaratılmalıdır. Böyle bir ortam yaratmanın yolu, iç müşterilerin yani çalışanların müşteri olarak algılanması ve iç müşterilere yönelik pazarlama uygulamalarının benimsenmesidir. Bu bağlamda işletme, hizmet kalitesini arttırmak ve etkili dışsal pazarlama yapmak amacıyla çalışanların tatminini ve motivasyonunu arttırmaya çalışmalı yani içsel pazarlama uygulamalarına yönelmelidir. Çünkü içsel pazarlama faaliyetleri, müşteri odaklı çalışanları işletmeye çekmek, onları işletmede tutmak ve onların motivasyonlarını arttırmaya çalışmak üzerine yoğunlaşmaktadır. Başka bir anlatımla, içsel pazarlama çabalarıyla etkili dışsal pazarlama davranışı için uygun iklim sağlanmaktadır (Çoban, 2004).

Taşkın'a göre (2000) müşteriler; mevcut müşteri, muhtemel müşteri, eski müşteri, yeni müşteri, hedef müşteri olmak üzere beşe ayrılır. Mevcut müşteri, işletmenin ürününü veya hizmetini çoğunlukla satın alan müşteri iken; muhtemel müşteri, işletmenin satış için görüştüğü, fakat halen işletmenin müşterisi olmamış, satın alma gücü ve isteği olan müşteri adayıdır. Eski müşteri işletmenin daha önce müşterisi olmasına rağmen çeşitli nedenler ile artık müşterisi olmayan kişi veya kuruluş olarak tanımlanırken; yeni müşteri kavramı ise bir işletmenin ürününü veya hizmetini ilk defa satın alan müşteridir. Hedef müşteri ise, belirli bir işletmenin belirli mallarını satın alabileceği amaçlanan kişi veya kurumlar olarak tanımlanır.

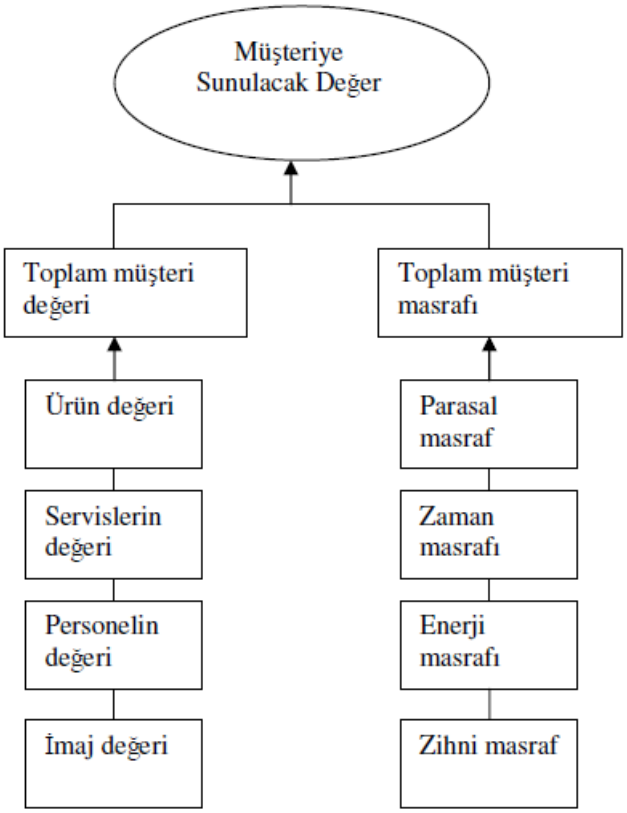
İşletme için en önemli müşteri grubu mevcut müşteri grubudur. Mevcut müşteriler, memnun, memnun olmayan ve tatmin olmuş müşteriler olmak üzere bir sınıflandırmaya tabi tutulabilir. Memnun olmayan müşteriler, beklentileri karşılanmadığı için büyük olasılıkla bir sonraki alışverişlerinde aynı işletmeyi tercih etmeyecektir. Memnun müşteriler, sadece beklentileri karşılanan müşterilerdir. Tatmin olmuş müşteriler ise beklentilerinin üzerinde bir satın alma deneyimi yaşayan müşterilerdir (Gündaş, 2003). Tatmin olmuş mevcut müşterilerin gelecekte yapacakları satın almaları ve işletme ile olumlu propagandaları düşünülürse, işletmenin hedefinin müşteri tatmini sağlamak zorunluluğu olduğu daha iyi kavranacaktır.

Müşteri varlığının tümünü aynı derecede karlı kabul etmek doğru bir yaklaşım olarak değerlendirilmemektedir. Çünkü müşteri kitlesinin sadece belirli bir bölümü karlıdır. İçinde bulunulan bu yeni dönemde gerçek fiyatlar ve gerçek maliyetler rasyonel bir şekilde analize tabi tutulduğunda pek çok müşterinin kar getirmediği görülmektedir. Bu bağlamda üretici işletmelerin yapması gereken; mevcut ve potansiyel karlarını müşteri bazında incelemek, potansiyel taşıyan müşterileri seçmek ve belirlenen bu müşteriler üzerine, ne kadar yatırım yapılacağı konusunda karar vermektir (Kırım, 2001). Yani hedef, doğru felsefe ile sadık ve karlı bir müşteri profili oluşturmaktır (Law ve diğ., 2003).

4.1.1 Müşteri için değer yaratma

Günümüzde müşterilerin eğitim düzeyinin artması sonucunda, müşterilerin kalite tanımları, ürün ve hizmetten beklentileri yükselmiştir. Kalite olmazsa olmaz bir koşul haline gelmiştir. Sadece kalite yeterli olmamakta, müşteri bunun yanında başka beklentilerini de dile getirmektedir. Bu noktada "değer" kavramı ortaya çıkmaktadır. Değer dış müşterilerin bekledikleri, istedikleri, algıladıkları bir şeydir ve ürün kalitesi, hizmet kalitesi, fiyat birbiriyle uyum içinde ortaya çıkıyorsa ve müşterinin beklentilerini aşıyorsa sağlanabilir (Güzel, 2001). Müşterilerin herhangi bir işletme ile iş yapmaya devam etmesinin temel nedeni kendilerine sunulan değerdir. Müşteri için değer yaratma, müşterilerin ne istedikleri ve ürünü satın alıp kullandıktan sonra ne elde ettikleri ile ilgili bir yaklaşımdır. Müşteri açısından değer yaratma kavramı, müşterinin ürün ve hizmete ödediği karşılığında beklediğinden fazlasını elde ettiği zamanki durumu ve anlamı içermektedir. Müşterinin bir ürün ya da hizmetten ne kazandığı ile ürün ve hizmeti elde ederken ne gibi ödünlerde bulunduğu arasındaki

farkı kapsamaktadır (Odabaşı, 2005). Yani müşteriye sunulan değer, toplam müşteri değeri ve toplam müşteri masrafı arasındaki farktır. Toplam müşteri değeri, müşterilerin belirli bir ürün veya servis karşılığındaki beklentilerinin toplamıdır. Toplam müşteri masrafı, müşterilerin ürün veya servisi değerlendirmek, elde etmek, kullanmak ve elden çıkarmak için yapacakları masrafların toplamıdır (Kotler, 2000). Müşteriler burada yararları ve ödünleri kendilerine göre karşılaştırırlar. Yararlar ödünleri aşınca yüksek müşteri değerine ulaşılmaktadır. Müşteri değeri ne kadar yüksekse, müşterilerin satın alma istekleri de o kadar fazla olacaktır. Müşterilerin beklentileri yeterli ölçüde veya daha yüksek bir seviyede karşılanıyorsa yüksek müşteri tatmini elde edilecektir (Güzel, 2001).



Şekil 4.1: Müşteriye sunulan değerlerin belirleyicileri (Kotler, 2005).

Şekil 4.1'de görüldüğü gibi müşteri ürünü veya hizmeti değerlendirirken ürün için katlandığı tüm maliyetleri ve sağladığı tüm yararları göz önüne almaktadır. Aradaki fark müşteri için değeri ifade etmektedir. Müşteri için değer yaratma ek yararları herhangi bir bedel ödemeden müşterilere sunmaktır. Buradaki önemli nokta; beklenen, umut edilen, algılanan durumdur. Örneğin, bir paket tuz aldığınızda kimsenin bunu arabanıza kadar taşıyabileceğini düşünemez ve bekleyemezsiniz. Ancak hepimiz arabamıza benzin aldığımızda bir kahve ikramını ya da camların

temizlenmesini bekleriz. Yaratılan değer, müşteri ile işletme arasında duygusal bir bağın kurulmasına olanak sağlamalıdır. Bu durum ise müşteri açısından yeniden satın alma ile müşteri sadakatini yaratmayı ortaya çıkartabilmektedir (Odabaşı, 2005). Müşteri için değer yaratmak sadece müşteriyi memnun etmekle yeterli olmamaktadır. Müşterinin hayallerini gerçekleştirmek de gerekmektedir. Değer yaratmak için bilinen birçok yöntem vardır. Çekirdek ürüne daha fazla değer katarak (ürünün kalitesinin arttırılması, destek hizmetlerinin verilmesi gibi), işletmeler müşteri tatminini arttırmaya çalışırlar. Arabalarda kullanılan çeliğin değiştirilmesi, çamaşır makinelerinin düğmelerinin geliştirilmesi bu duruma örnek olarak verilebilir. Eğer bunlar, müşteri açısından değer yaratmıyorsa hiçbir anlamı olmadığı gibi başarı şansları da çok düşük olabilmektedir. İşletmeler, müşterinin algıladığı ödünü azaltacak ve böylece ilişki maliyetini en aza indirecek bir değer sunabilirse, başarılı olabilme şansı yüksek olabilecektir (Odabaşı, 2005).

4.1.2 Müşteri ömür boyu değeri

Değer kavramının diğer ve önemli bir boyutu, işletme için müşterinin ömür boyu değerinin ölçülmesidir. Bir müşterinin ömür boyu değeri, müşterinin gelirleri, harcamaları ve müşteriyle işletme arasındaki ilişkinin beklenen ömrü arasındaki bağlantıyı gösteren net bugünkü değer hesaplamalarıdır (Akalın, 2004).

Müşteri ömür boyu değerini ölçebilmek için işletmelerin aşağıdaki bilgilere ihtiyacı vardır (Odabası, 2005):

- Müşterinin satın almalarından elde edilen tüm gelirler ve müşterinin talebini karşılarken katlanılan tüm değişken maliyetler,

- Müşterinin satın alma sıklığı,

- Müşterinin işletmeden aktif olarak satın almayı sürdürme zamanının genişliği,

- Müşterinin diğer ürünleri satın alma düzeyi ve bu tür satışlardan elde edilebilecek gelirler,

- Müşterinin, çevresine tavsiyelerde bulunma özelliği,

- Müşterilere yapılan uygun indirimler.

Müşteri değerini belirlemek için gereken bu bilgilerin birçoğu mevcut bilgilerdir. Burada yapılması gereken işletmenin içinde bulunduğu sektörü ve hedef müşteri

grubunu göz önüne alarak, sahip olduğu müşterilerini, işletme için yarattıkları değere göre sınıflandırmasıdır.

Müşteriler değerlerine göre üç sınıfa ayrılırlar (Kırım, 2001):

- En Değerli Müşteriler (EDM): En değerli müşteriler, en yüksek müşteri değerine sahip olanlardır. Bu müşterileri elde tutmak ve aynı zamanda onların müşteri payını arttırmak işletmenin stratejik hedefi olmalıdır.
- En Çok Büyüme Potansiyelli Müşteriler (EBM): Bunlar, ömür boyu değerleri EDM'den daha düşük ama büyüme potansiyelleri olan müşterilerden oluşur. Amaç bu müşterilerin işletmelerden aldıkları ürün veya hizmetlerin payını arttırarak daha değerli bir hale getirmek ve bir üst gruba geçirmek olmalıdır.
- Sıfırın Altı Müşteriler (SA): Sıfırın altı olanlar muhtemelen hiç kar getirmeyen ve hatta zarar ettiren müşterilerdir. Bu müşteriler karlı hale getirilmeye çalışılmalıdır. Eğer müşteriler karlı hale getirilemiyorsa bu müşterileri terk etmek alınacak en doğru karardır.

Bu sınıflandırmalar sonucunda işletme için yüksek değer ve düşük maliyet yaratan müşteri grubu belirlenir. Diğer bir deyişle, müşteri ömür boyu değeri belirlenerek, pazarlama faaliyetleri yüksek değer yaratan müşteriler üzerinde yoğunlaştırılır. Müşterilerin değerlerine göre bölümlendirilmesi işletmelerin karlılığı için çok önemlidir. Çünkü bunun sonucu olarak işletmeler uygun müşterileri elde tutma ve kazanma şansına sahip olabilmektedirler.

4.1.3 Müşteri tatmini

Müşteri istek ve ihtiyaçlarının tatmin edilmesi olgusu, modern pazarlama anlayışının temelini oluşturmaktadır. Bu nedenle birçok alanda, müşteri tatmininin yapısı ve süreci araştırmacıların temel ilgi alanlarından biri olmuştur. Öncelikli hedefleri varlığını devam ettirebilmek ve karlılık sağlamak olan işletmelerin memnun müşteriler yaratması rakipler karşısında önemli bir avantaj yakalaması anlamına gelmektedir. Bunu başaran ve pazarın değişen koşullarına uyum sağlayabilen işletmeler orta ve uzun vadeli rekabet üstünlüğü ve uzun dönemli karlılık elde edebilmektedirler.

Satın alınan ürün ya da hizmetin sağladığı faydalar ile müşterinin beklentilerinin uyuştuğu noktada müşteri tatmini ortaya çıkmaktadır. Müşterinin bir ürün veya hizmeti satın almadan önceki beklentileri ile o ürün veya hizmeti satın aldıktan ve kullandıktan sonra algıladığı performansı arasındaki farklılık tatmin olarak ele alınabilir. Diğer bir tanımda müşteri tatmini, tüketim sırasında elde edilen duygularla, tüketim öncesi duyguların değerlendirmesi sonuçlarının bir bileşimi olarak değerlendirilmektedir (Tütüncü ve Doğan, 2003).

Müşterinin tatmin olup olmaması, kendine sunulanı kendi beklentileriyle kıyaslamasına bağlıdır. Eğer sunulan ürün ya da hizmet, beklentilere uyar ise müşteri tatmin olacak, uymaz ise müşteri tatmin olmayacaktır. Hatta sunulan ürün ya da hizmet müşterilerin beklentilerinin üstünde ise müşteri yüksek ölçüde tatmin ve memnun olacaktır. Sadece tatmin olmuş müşteriler, daha iyi bir ürün ve hizmet sunumu karşısında kolaylıkla o sunuma yönelebileceklerinden pek çok işletme, müşterilerinin yüksek ölçüde tatmin olmalarını hedeflemektedir. Yüksek ölçüde tatmin olanlar, bir diğer sunum karşısında kolay kolay yer değiştirmeyeceklerdir. Müşteriler kendilerini yüksek ölçüde tatmin eden sunumu sadece rasyonel bir tercih olarak değerlendirmemekte aynı zamanda o sunuma karşı hissi bir yakınlık da duymaktadırlar. Bunun sonucunda da müşteri tatmini ortaya çıkmaktadır (Kotler, 2000).

Her yıl işletmeler ortalama %10 ile %30 arasında değişen oranlarda müşteri kaybına uğramaktadırlar. Ancak işletmelerin bir kısmı, müşterilerini niçin kaybettiklerini, ne zaman kaybettiklerini veya hangi müşteriyi kaybettiklerini bilmemektedir. Müşteri kayıpları genel olarak müşteri tatminin sağlanamamasından kaynaklanmaktadır. Bunun nedenini anlamanın en iyi yolu sormak, araştırmak ve müşteriler ile ilişki kurmaktan geçer. Müşteri kaybı, müşteri isteklerini ve ihtiyaçlarını tam anlayamamaktan da kaynaklanabilir. Bunu anlamanın yolu ise yine anketler ve ziyaretlerden geçmektedir. Ayrıca müşteri tatminini güçlü bir şekilde arttırmanın diğer bir yolu da çalışanlardan geçmektedir. Çalışanların motivasyonu arttırılmalı ve en üstten en alta kadar herkesin faaliyetlere bilinçli bir şekilde tam katılımı sağlanmalıdır (Kalder Grubu, 2000).

İşletmelerin başarıya ulaşmaları için esas alacakları en önemli husus "müşteri tatmini esastır" ilkesinin benimsenmesidir. Bu ilke işletmelerin tüm yönetim politikalarında ve organizasyon yapısının tüm kademelerinde ana prensip olarak dikkate alınmalı ve

uygulanmalıdır. Müşterilerin şikayetleri ve dilekleri dinlenmeli, bunlar dikkate alınmalı ve ürün ve hizmetle ilgili müşteri tatminine etki eden özellikler belirlenerek müşteriden işletmeye bir geri besleme olan müşteri tatmini araştırması yapılmalıdır. Bu araştırma sayesinde müşteriler, sunulmuş müşteri değerinin en doğru değerlendiricisi haline gelecektir.

4.1.4 Müşteri sadakati

Son yıllarda hızla artan rekabet, müşterilerin tercihlerindeki ve demografik özelliklerindeki değişim sonucunda müşteri elde etme ve tutma konusunda zor durumda kalan işletmeler fiyat düşürme, dağıtım kanallarını genişletme ve reklam yapma gibi değişik stratejiler kullanarak hem var olan müşterilerini elde tutmak hem de yeni müşteriler edinme yönünde çeşitli çabalar göstermektedir. Rekabetin ve diğer zorlayıcı faktörlerin yoğun olarak yaşandığı bir ortamda, bir işletmenin sadık müşterilere sahip olması, işletmenin devamlılığı bakımından çok önemli bir etmendir (Hançer, 2003).

Müşteri sadakati değişik yazarlar tarafından farklı şekilde tanımlanmıştır. Müşteri sadakati müşterilerin tercihlerinde bir değişikliğe neden olabilecek durumlarda ve pazarlama çabalarına rağmen, müşterinin sürekli olarak tercih ettiği ürün ve hizmetleri tekrar satın alması, tekrar tekrar o işletmenin müşterisi olarak kendini adaması olarak tanımlanmaktadır (Oliver, 1999). Griffin (1997) ise müşteri sadakatini, müşterinin alışveriş yaptığı sektördeki toplam harcama kapasitesinin büyük çoğunluğunu aynı işletmeden yapması olarak tanımlamakta ve müşteri sadakatinin müşterinin alışveriş alışkanlıklarının incelenmesiyle birlikte tanımlanabilmekte olduğunu savunmaktadır.

İşletmelerin sahip oldukları en değerli varlıklardan birisi kuvvetli bir müşteri sadakatidir. Çünkü müşterilerin işletmeye olan sadakatinin güçlülüğü, rakiplerin söz konusu müşteriyi kendilerine çekmeye yönelik uygulamalarını güçleştirmekte ve oldukça masraflı kılmaktadır. Sadık müşteriler işletmeler açısından satış işlemini kolaylaştırmakta, pazar istikrarı ve fiyatlar üzerinde kontrol olanağı sağlamakta ve işletmeye dağıtım kanalları konusunda yardım etmektedir (Madran ve Canbolat, 2006). Müşterilerin sadık olmamasının işletme performansı üzerindeki etkileri ise hemen görülmekte ve işletme için çok önemli potansiyel tehlike oluşturmaktadır. Hemen görülebilen etkilerden satışların azalması, müşteri kaybı, gelirlerin azalması, pazar payının düşmesi ve karlılığın azalmasını sayılabilmektedir. Müşterilerin sadık

olmamasının memnuniyetsizlikten kaynaklandığı hallerde, durum daha da ciddi boyutlar almaktadır. Müşterilerin olumsuz yaklaşım ve sözleri işletmenin imajını düşürmekte ve muhtemel satışları engellemektedir (Kalder Grubu, 2000). Ryals ve Knox (2001) sadık müşterilerin işletme için her zaman karlı olma nedenlerini şu şekilde sıralamaktadır:

- Müşteri elde etme maliyetleri yüksek olması nedeniyle, müşteriler birkaç sene işletmeye bağlı kalmadan bu müşterilerden kar elde etmek zordur.

- Müşterilerin karlılığı her yıl bir önceki yıla göre artmaktadır. Örneğin, bir sigorta işletmesinde ilk iki yıl işletmenin müşteriden çok fazla karı olmamakta ancak daha sonra müşteriden elde ettikleri kar oranı artmaktadır.

- Tatmin edilmiş müşteri, memnuniyetini çevresine anlatarak ve tavsiye ederek işletmeye yeni müşteri kazandırmaktadır.

- İşletmeye sadık olan ve işletmeyle iyi ilişki kuran müşteri bir değer yaratır. Böylece müşteri fiyata karşı daha az duyarlı olmaya başlar ve fiyatlardaki küçük değişikliklerle o kadar fazla ilgilenmez.

Müşteri sadakati, müşterilerin firma ile çeşitli deneyimler sonucunda edindiği pozitif izlenimlerin firmaya geri dönüşüdür. Mahalle bakkalları bu konuda örnek olarak düşünülebilir. Bakkallar müşterilerle tek tek ilgilenme şansına sahip oldukları için onların tüm taleplerini bilmektedir. Bu da müşteriyi sadık yapan bir unsurdur. Bakkalın veresiye satması, kredi açması, yiyeceklerden tatmaya izin vermesi, kötü çıkan yumurtayı geri alması, müşteri üzerinde olumlu bir izlenim bırakmaktadır (Güzel, 2001). Yani müşteriler genellikle alışveriş yaptıkları işletmelerin kendilerini tanımalarını, hatırlamalarını ve özel hissettirmelerini beklemektedir. Müşterileri rakip işletmelere kaptırmamak için müşteri beklentileri her şekilde karşılanmaya çalışılmalıdır. Bunun için de müşteri ilişkilerini sağlam tutarak müşteri sadakati oluşturmak gerekmektedir.

Müşteri sadakatini arttırabilmek için önerilen bir yaklaşım; müşterileri, karlılıklarına göre farklılaştırmaya tabi tutmaktır. Müşterilerini sadık ve karlı bir yapıya kavuşturmak isteyen işletmeler, müşterileriyle öğrenen bir ilişkiye girmek durumundadırlar. Bunun için ise, müşteri davranışlarını etkili bir biçimde analiz etmeleri gerekmektedir. Müşterilerinden elde edecekleri ham verileri öncelikle anlamlı ve kullanılabilir bilgiye sonra ise bu bilgiden etkili müşteri ilişkilerine

ulaşmaya çalışmalıdırlar. İşletmelerde, müşteri bilgi sistemi, sürekli olarak güncel hale getirilip, geliştirilmek durumundadır. Ayrıca işletmelerde müşteri sadakati iklimleri yaratılmalıdır. Bunun için izlenecek yollar; işletmenin tatminkar bir ürün sunmasından, müşterilerin ürünü, markayı ya da mağazayı tanımasının sağlanmasından, ürünün ulaşılabilir olmasından, işletmenin sözüne sadık olmasından ve satın alma işlemlerinin kolaylaştırılmasından geçmektedir (Madran ve Canbolat, 2006).

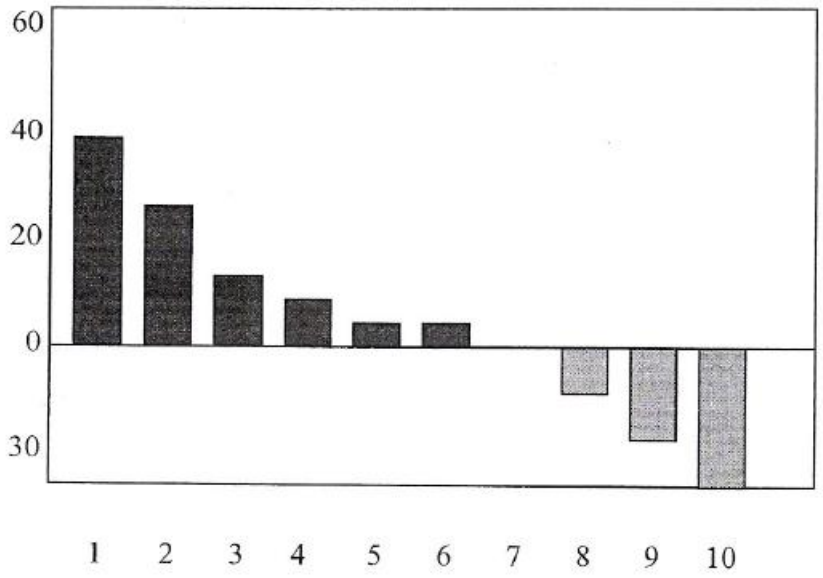
4.1.5 Yeni müşteri kazanmak ve eldeki müşteriyi tutmak

Yeni müşterileri bulma ve onlarla ilişkileri kurup geliştirme, birçok kuruluşta mevcut müşterilerle ilişkilerden daha çok ilgi çekebilmektedir. Satışlarını mümkün olduğunca arttırmak isteyen işletmeler daha fazla müşteri kazanmaya yönelmektedir. Aslında pek anlamlı olmayan bu durumdan kuruluşların canları da yanabilmektedir. Yeni müşteriler elde ederken işletme onları işletmeye çekebilmek için ilave maliyetlere de katlanmak zorundadır. Bu nedenle işletmelerin yeni müşterileri bulmak için girişimde bulunduğu uygulamaların maliyeti ile var olan müşterileri elde tutmaya yönelik uygulamaların maliyeti aynı değildir. Yeni bir müşteri kazanmanın maliyetinin mevcut müşterileri elde tutma maliyetinden ortalama 5 misli fazla olmaktadır (Reicheld ve Sasser, 1990).

Müşterilerin işletmeye olan ilgilerinin ve devamlılıklarının süreceğini düşünmek çok riskli bir varsayımdır. Gerçek ise, müşterilerin ürün ve hizmetlerde daha geniş bir tercih sunan işletmeleri seçtiği, diğerlerinden farksız ya da zayıf hizmet sunan ve müşteri ilişkileri kopuk olan işletmeleri ise terk ettiğidir (Odabaşı, 2005). Müşteriler işletmelerden memnun kalmadıklarında sorunlarını işletmeye bildirme ya da sessiz kalma olmak üzere iki çeşit davranışta bulunmaktadırlar. Technical Assistance Research Programs (TARP) tarafından yapılan araştırmalara göre, müşterilerin %95'i, bir sorunu olduğunda sessiz kalmayı tercih etmektedir. Aynı zamanda bu müşteriler, işletmeden bir daha ürün satın almamakta yani işletmeyi terk etmektedirler. Ayrıca TARP'ın araştırma sonuçları, sessiz kalmayı tercih eden müşterilerin %91'inin etrafindakilere olumsuz deneyimlerini anlattıklarını göstermektedir. Bu durumla karsılaşmak istemeyen işletmeler, müşterilerinin sessiz kalmasından ziyade konuşmasını tercih etmektedirler. Çünkü tatmin olmayan müşterinin durumu işletmeye bildirmesiyle ilk aksaklığı işletme öğrenecek ve bunu düzeltme imkanı kazanacaktır. Ayrıca şikayetlerin işletmede toplanması, müşteriyi

psikolojik olarak rahatlatmaktadır. Ürünü satın almadan önce, bir sorunu olduğunda işletmeyi arayabileceğini bilen ve sorunun çözüleceği garantisi verilen müşteri, endişe duymadan daha kolay alışveriş yapabilmektedir (Timur ve Sarıyer, 2004). Bu nedenle müşterilerini elde tutmak isteyen işletmeler, müşterilerden gelen şikayetlerin doğrudan kendilerine gelmesini sağlamak için bir takım yöntemler geliştirmeli ve bu şikayetleri değerlendirmelidirler.

İşletmeler için bütün müşteriler önemlidir. Fakat işletmenin varlığını sürdürebilmesi için karlılık oranı yüksek müşteriler daha fazla dikkate alınmalıdır. Karlılık açısından bakıldığında müşteriler farklı gruplarda ele alınmalıdır. Karlı müşterileri karlı olmayan müşteri grubundan ayırmak için çeşitli uygulamalar yapılabilir. Bunlardan en basit olanı Pareto analizidir. Araştırmalar, Pareto ilkesinin işletmelerin sahip oldukları müşteriler için de geçerli olduğunu ortaya koymaktadırlar. Pareto analizi müşterilerin karlılığını incelemek için kullanılır (Rosenfield, 1999). Analize göre karın %80'i müşterilerin %20'sinden gelmektedir. Buna karşılık işletmelerin en dipteki %30'u diğer müşterilerce sağlanan karın yarısını tüketebilir.



Şekil 4.2 : Müşteri karlılığında 80/20/30 kuralı (Günaydın, 2005).

Şekil 4.2'de 1. ve 2. çubuklarda yer alan en üst %20'lik müşteri dilimi, şirketin toplam karının %80'ini sağlarken, 8., 9. ve 10. çubuklarda yer alan alt %30'luk dilim ise diğer müşterilerin yarattığı karın yarısını israf eder. Bu nedenle karlı müşteriler işletmeler için çok önemlidir. Bu nedenle işletmeler karlı müşteriye daha iyi hizmet verme yöntemini benimsemektedir. Buna en güzel örnek olarak bankalarda kuyruğa giren müşteriler verilebilmektedir. Bankalar kendi müşterilerine ayrı ve daha

avantajlı bir sıra numarası vermektedirler. Hatta karlı ve portföyü büyük müşteriye daha öncelikli sıra verilebilir ve bu kişiler sıra beklemeden direk müşteri temsilcileriyle görüştürülebilir (Karaman, 2006).

Müşterileri elde tutmayı güçlendirmenin iki yolu vardır. Biri, onların başka tarafa geçmelerini engelleyecek yüksek duvarlar örmektir. Eğer müşterilerin işletmeyi terk etmeleri, kendilerinin sermaye ve araştırma masraflarını arttıracaksa veya sadık müşterilerin indirimli hesaplarını kaybetmelerine yol açacaksa, onlar bir başka tedarikçiye gitmeyi pek istemeyeceklerdir. Bu nedenle, müşteriyi daha yüksek derecede tatmin etmek gereklidir. O zaman rakiplerin, sadece düşük fiyatlar veya müşterileri kendilerine çekmek için bazı tekliflerle bu duvarları aşmaları zorlaşacaktır.

5. KARAR AĞAÇLARI

5.1 Karar Ağacı Tanımı

Karar ağaçları, ağaç şeklinde sınıflandırıcılardır. Bu ağaçtaki her düğüm bir yaprağı veya karar düğümünü belirtir. Yaprak düğümü hedef niteliğin değeridir. Karar düğümü ise, bir nitelikte uygulanacak olan test değeridir, bu düğümü, o niteliğe ait olan tüm olası nitelik değerleri izler, bu değerler ise ağacın dallarını oluşturur (Tosun, 2006). Bir başka deyişle, düğüm soruları ifade eder, dal bu soruların cevaplarını ve yaprak ise kararın verildiği sınıfı ifade eder. Ağacın ilk düğümü ile sorular sorulmaya başlar ve dalları olmayan düğümler ya da yapraklara gelene kadar bu sorular devam eder. Karar ağacının if – then yapıları gibi kurallarla ifade edilmesi son derece kolaydır. Bu noktada kaç tane yaprak varsa o kadar kural oluşur.

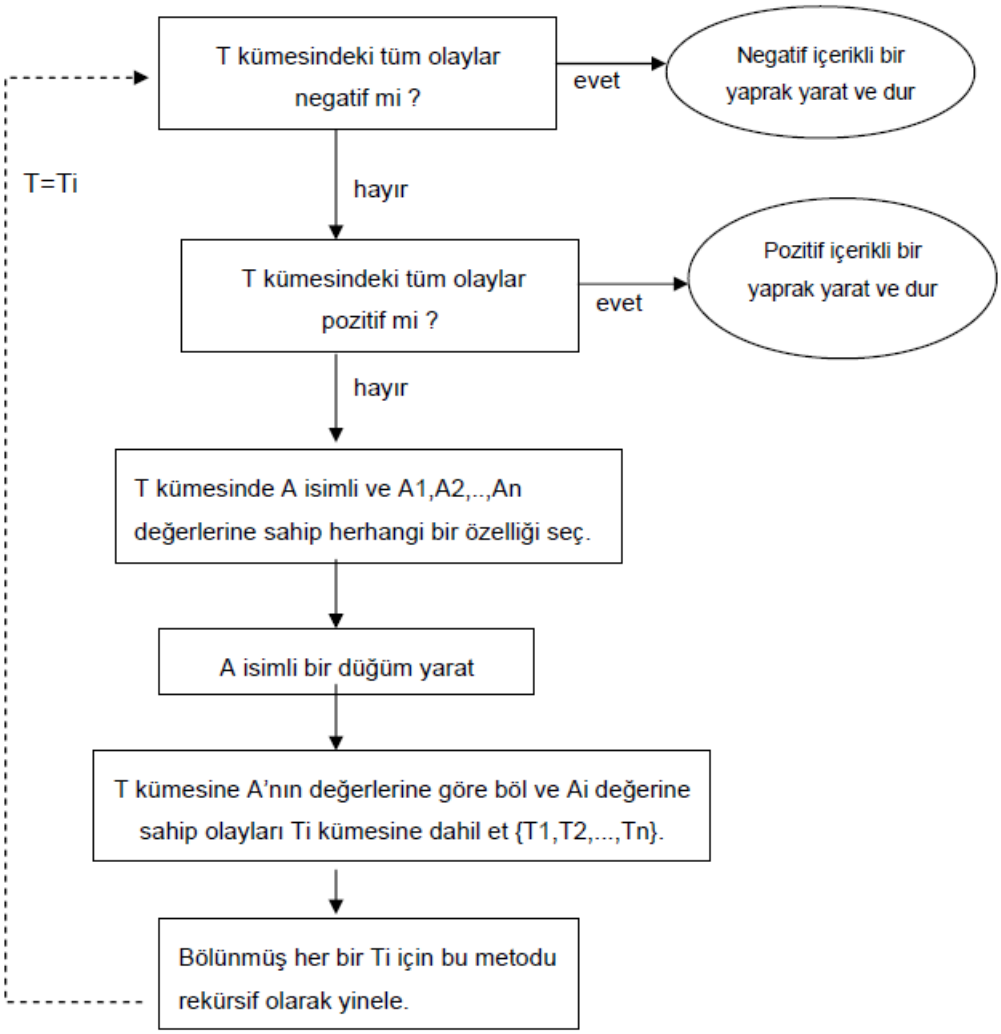
5.2 Karar Ağacı Oluşturma

Ağacın oluşturulmasına yönelik olarak çeşitli ağaç oluşturma metotları bulunmaktadır. Ağacı oluşturmadaki en önemli kriter belli özelliklere göre toplanmış, güvenilir ve yeterli sayıda olay örneklerinin varlığıdır. Bu iki faktör ağaç oluşturmanın temelini oluşturur. Ağaç oluşturmadaki en önemli adım ise, böl ve elde et aşamasıdır.

5.2.1 Böl ve elde et (divide and conquer)

Hunt'un uyguladığı bir metottur. Metotta örnek uzay T, sınıflar pozitif ve negatif olarak düşünülürse karar ağacı oluşturma algoritması Şekil 5.1' deki gibi olacaktır.

Bu algoritma en temel ağaç yaratma algoritmasıdır. Algoritmanın enine boyuna incelenmesi ve geliştirilmesine yönelik olarak çeşitli çalışmalar vardır. Bunlardan en önemlisi tek değişkenli karar ağaçları için Quinlan' ın 1983' te geliştirdiği ID3 algoritması ve yine ardından Quinlan' ın geliştirdiği C4.5 algoritmasıdır. Çok değişkenli karar ağaçları için ise Breiman' ın geliştirdiği CART algoritması vardır.



Şekil 5.1: Hunt' ın ağaç oluşturma metodu.

5.2.2 ID3 algoritması

Yukarıda akış diyagramı gösterilen HUNT' ın meşhur algortimasındaki en önemli eksiklik özelliklerin rastgele seçilmesidir. Oysa ki bu seçim sırasında bilgi kazancı en yüksek olan özellik dikkate alınırsa oluşturulan ağaç o kadar sade ve anlaşılır olacaktır. Buna yönelik olarak Quinlan entropi kurallarını içeren bilgi teorisini kullanmıştır. Shannon ve Weaver' ın bilgi teorisinde temel olarak kaynak, mesaj ve alıcı vardır. Bu sistemde bilgi, mesaja bakılarak değil de, alıcıya bakılarak elde edilir. Alıcı mümkün olan mesaj uzayı bilgisine ve bu mesajların olasılıklarına sahiptir. Ağaçlardaki bazı düğümler ve bu düğümlerdeki kararlar anlamsız ve gereksiz olabilmektedir. Ancak bu tip düğümler de negatif – pozitif olay balansına sahiptir. İşte bu şekilde sınıflandırma yapılabilmektedir (Kocabaş, 1991).

Örnek olarak, X düğümünde 5 pozitif ve 3 negatif olay olduğunu düşünelim. Bu noktada yapılacak sınıflandırmanın pozitif olasılığı 5/8' dir, negatif olasılığı 3/8' dir. Bu olasılıksal sınıflandırmayı türetme yeteneğinin anlamı şudur: Doğru olarak sınıflandırılmış bir örneğin söylediği mesajın bilgi içeriği artık hesaplanabilir.

Öyle ki bir tablonun sonuçları mesaj olsun ve bu mesajlar iki değere sahip olsunlar. Bu değerlerle birlikte p bilgisi pozitif olasılığını, q negatif olasılığını gösterir. Bu iki değerin toplamı zaten 1 (p+q) olmak zorundadır. Doğru sınıflandırma veren bir mesajın bilgi içeriği;

$$I(p,n) = -p \lg_2 p - q \lg_2 q \qquad \textbf{(5.1)}$$

şeklinde hesaplanır.

Bu formül, genel bilgi içerik formülünün özel bir durumudur. Çünkü özel olarak iki olasılık mevcuttur: pozitif ve negatif.

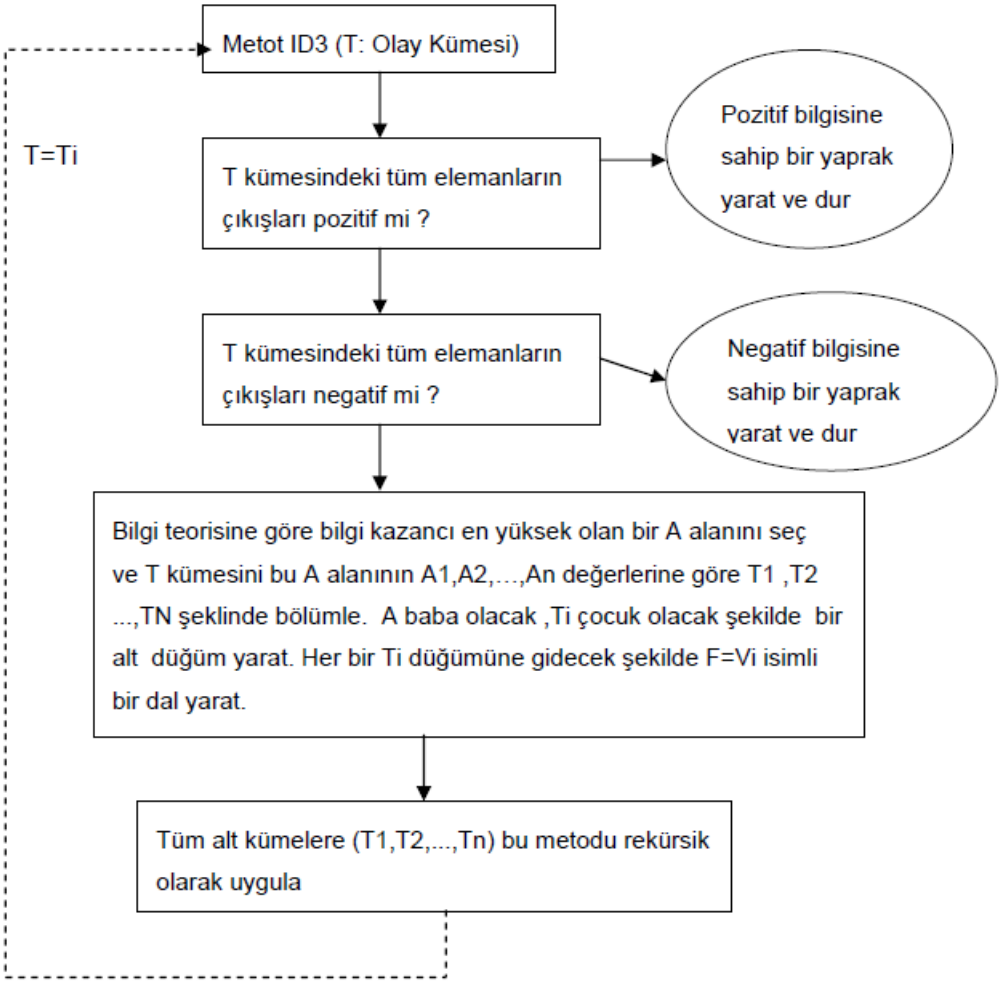
$\{A_1, A_2, \ldots, A_N\}$ değerlerine sahip A özelliği ağacın bölümlenmesi için kullandığında, T kümesi $\{T_1, T_2, \ldots, T_N\}$ şeklinde bölünecektir. Bu bölümleme de T kümesinin A özelliğinin Ai olduğu bölgelere Ti densin. Bu kümedeki pozitif olayların sayısını p_i temsil etsin, negatif olayların sayısı n_i temsil etsin. Bu durumda T_i alt ağacı için beklenen bilgi gereksinimi ise $I(p_i,n_i)$ olur. T ağacı için beklenen bilgi gereksinimi tüm T_i ağaçlarının beklenen bilgi gereksinimlerinin ağırlıklı ortalamalarının toplamı olur ve

$$E(A) = \sum_{i=1}^{n} \frac{p_i+n_i}{p+n} I(p_i+n_i) \qquad \textbf{(5.2)}$$

şeklinde hesaplanır. Dolayısı ile A özelliği üzerinden sağlanan bilgi kazancı

$$\text{Bilgi kazancı}(A) = I(p,n)-E(A) \qquad \textbf{(5.3)}$$

şeklinde ifade edilir. Şekil 5.2' de ID3 algoritmasının aşamaları ifade edilmiştir.



Şekil 5.2 : ID3 algoritmasının akış diyagramı.

Bilgi gereksinimi ve bilgi kazancı ID3 algoritması için iki önemli kavramdır. Belirleyici bir sınıflandırmada bilgi ihtiyacı aslında doğru sınıflandırmayı sağlayan mesajın bilgi içeriğinden başka bir şey değildir. Buna yönelik olarak yaratmak istenilen karar ağaçlarının amacı doğru soruları sormasıdır ve sonunda öyle bir noktaya ulaşılmalı ki, bu noktanın karar için bilgi gereksinimi 0 olsun. Kurulu karar ağacının her seviyesinde geriye kalan bilgi gereksinimi minimize edilir.

5.2.3 C4.5 algoritması

ID3 algortimasında bazı eksiklikler ve sorunlar bulunmaktadır. Bu sorunlar yine Quinlan' ın geliştirdiği C4.5 algoritması ile giderilmiştir. C4.5 algoritması, ID3 algoritmasının bütün özelliklerinin kendine miras alarak oluşturulmuş bir algoritmadır. Yukarıda bahsedilen tüm içeriğin üzerine yeni kavramlar eklenmiştir. Bölünme – Dağılma Bilgisi (Split - Info), özelliklerin kayıp değerleriyle baş edilmesi, sayısal özellik değerlerinin hesaplara katılması bu başlıklardan en önemlileridir.

5.2.3.1 Bölünme – dallanma bilgisi

Bir kategorik özelliğin olası değer çeşitliliği ne kadar yüksek olursa o özelliğin bilgi kazancı gereksiz bir şekilde yüksek çıkar ve bu durum ağacın doğruluğunu kötü bir şekilde etkiler. Bu tip özellikler işe yaramadıkları gibi bilgi kazancı yüksek özelliklerin de önüne geçip veride gizlenmiş kuralların çıkarılmasına engel teşkil ederler. Bu tip gereksiz bilgilerle başa çıkabilmek için Quinlan bölünme bilgisi kavramı ile algoritmasını güncellemiştir. Bu algoritma değer çeşitliliği fazla olan özelliklerin bilgi kazancını azaltarak algoritmanın gereksiz bazı çıkarımlar yapmasını engellemektedir.

Algoritmaya bölünme bilgisi denilen yeni bir kavram eklenmiştir. A bir özellik, A_i bu özelliğin değerleri, T_i A_i özelliğinin bu veride kaç kez tekrarladığı ve T ise ele alınan olay sayısını temsil ediyorsa, bölünme bilgisi

$$\sum_{i=1}^{n} \frac{T_i}{T} * \lg_2 \frac{|T_i|}{|T|} \qquad \textbf{(5.4)}$$

olarak hesaplanır. Bu bölünme bilgisi tüm özelliklerin bilgi kazanç formülüne bölen olarak eklenir ve bu sonuç kazanç oranı olarak ifade edilir. Bu durumda A özelliğinin kazanç oranı,

Kazanç_Oranı = Bilgi Kazancı(A) / Bölünme_Bilgisi(A) **(5.5)**

şeklinde hesap edilir.

5.2.3.2 Sayısal özellikler

Veri kümesinde iki tip veri bulunmaktadır; nominal (kategorik) ve sayısal. Nominal tipi örnek olarak; renk özelliği mavi, kırmızı ve sarı olan, büyüklük özelliği büyük, orta ve küçük olan toplar olarak ifade edilebilir. ID3 algoritması sadece nominal

değerlere sahip veri tipleri ile işlem yapabilmektedir. Ancak bu algoritmanın devamı ve geliştirilmişi olan C4.5 algoritması bu tip veriler için de bir yöntem geliştirmiştir.

İlk bakışta sayısal özellikler ile uğraşmak ve onların bilgi kazancını hesaplamak oldukça zor gelebilir. Yapılması gereken iş sadece bu özelliğin sayısal değerleri arasında uygun eşik değerini bulmaktır. Bu eşik değeri bulunduktan sonra ikili bir bölünme ile veri kümesi bölünebilir; bu eşik değerinden büyük veriler ile bu eşik değerinden küçük veriler olmak üzere.

Böyle bir bölünme işlemi bir çok veri üzerinde denendiğinde olumlu sonuçlar vermiştir. Ancak eksik olan yanı sadece ikili bölme işlemi gerçekleştirmesidir. Oysa bu tip bir bölünme üçlü ya da daha fazla olursa veri yığını içerisine gizlenmiş olan kuralları bulma olasılığı daha çok artar.

5.2.3.3 Kayıp veriler

Bir önceki bölümde gizli bir kabullenme vardır; söz konusu ID3 algortiması eksik olmayan bir veri yığınını dikkate alır. Fakat veride bazı bilgilerin bulunmaması tüm algoritmanın çalışmasını engeller ve yanlış çıkarımlara yol açar. Gerçek uygulamalarda, çeşitli sebeplerden dolayı eksik veriler olabilir. Bu aşamada veri yığınını tamamlamak yerine algoritmaya bazı yetenekli metotlar eklemek gerekir.

5.3 Ağacın Testi

Oluşturulan ağacın performansı o ağaçtaki hata oranı ölçülerek tespit edilir. Ağaçlar, her bir olay için bir sınıf tahmininde bulunurlar; bu tahmin doğru ise söz konusu ağaç başarılıdır, değilse başarısızdır. Hata oranı ise tüm veri kümesindeki hatalı sınıflandırılmış olayların tüm olaylara oranıdır. Bu sonuç o sınıflandırıcının performasına karşılık gelir.

Eğer ağaçta hiçbir budama işlemi ya da bir kestirim yapılmamışsa, genelde ağacın hata oranı %0' a yakındır. Bu nedenle ağacın yaratılmasında kullanılan verilerle gerçekleştirilen hata oranı gelecekteki yeni verilere ilişkin hata oranına pek fazla ışık tutmamaktadır. Dolayısıyla hata oranı tespiti yeni bir veri kümesi üzerinde yapılmalıdır. Ağacı yaratan verilerdeki hata oranı genelde o ağacı yaratırken kullanılan yer değiştirme işlemlerini ifade etmektedir. Bu hata oranı güvenilir

değildir. Önemli olan bu veriden bağımsız bazı veriler üzerinde bu testi uygulamaktır. Bu noktada söz konusu olan iki ayrı veri kümesi mevcuttur.

1 – Eğitim Kümesi: Ağacı oluşturan veri kümesi.

2 – Sağlama Kümesi: Ağacın güvenilirliğini (hata oranını) tespit eden veri kümesi.

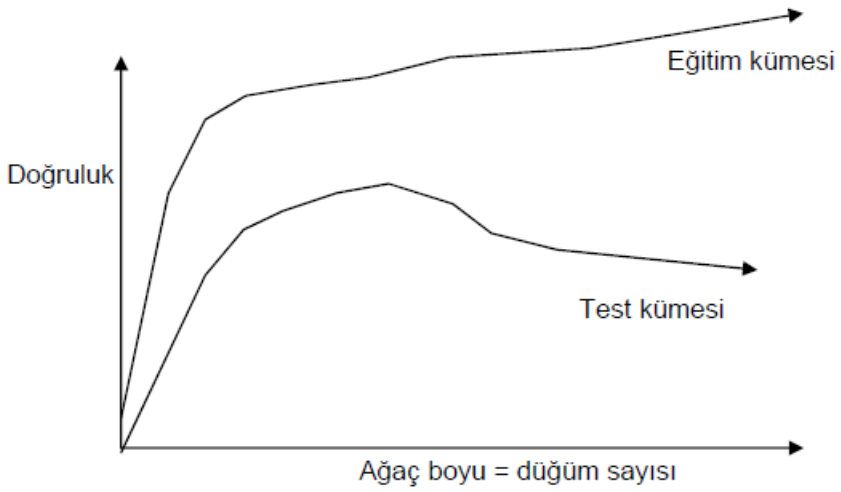
Test veri kümesi tamamen bağımsız olmak zorundadır. Eğer tam bir genelleme yapılmak isteniyorsa bu nokta şüphesiz ki çok önemlidir. Sonuç olarak, bir karar ağacı yaratılmadan önce veri kümesi ikiye ayrılır: eğitici küme ve sağlama kümesi. Bu noktadaki en önemli problem ikiye bölünecek kadar verinin bulunmamasıdır. Veri çok olduğu zaman bu işlem kolaylıkla yapılabilir. Ancak veri yetersiz ise, test için ayrılacak veri kümesini azaltmak gerekecektir.

Test aşamasında çok kritik bir aşama vardır ki o da seçilen bu iki kümenin birbirine tamamen zıt olmamasıdır. Buna yönelik olarak çeşitli yöntemler geliştirilmiştir. Bu yöntemlerden biri olan ikili çapraz sağlama testi (cross validation) veri kümesini iki eşit parçaya böler; önce birini sonra diğerini eğitici küme olarak seçer ve bu işlemi aynı şekilde test kümesi için de yapar. Bu durumda ortaya çıkacak olan ortalama hata oranı daha sağlıklı sonuç verecektir. Ancak bazı çalışmalar bunun da yetersiz kaldığını göstermiştir. Bu amaçla ikili sağlama işleminin yaptığı ikiye bölme yerine ona bölme denenmiştir. Onlu çapraz sağlama işlemi önce tüm veri yığınını on eşit parçaya böler. Ardından her birini sırayla dışarı alır ve geriye kalan dokuz adet veri kümesinden bir eğitim seti oluşturarak karar ağacını oluşturur. Ardından bu dışarıya aldığı veri kümesi ile test eder. Bu işlemi her biri için yapar ve işlemlerden çıkan on adet hata oranının ortalamasını gerçek ortalama olarak kabul eder.

5.4 Ağaç Budama Metotları

Ağaç budama, karar ağaçlarının en önemli konularından biridir. Veri yığınından oluşan ağaç çok büyük olabilir ve hiçbir anlam ifade etmeyebilir. Basit bir veri yığınından ağacın çok büyük çıkmasına şişme (overfitting) denir. Böyle bir sonuç çıkmasının iki nedeni olabilir. Birincisi veri yığını üzerinde gürültü vardır. Gürültü, veri yığınının gereksiz kurallar türetmesine yol açar. Bu durumda ağaçtan çıkarılan kurallar aslında gürültüye aittir. İkinci olasılık ise, seçilen veri kümesinin o olayı temsil edebilme yeteneğinin olmamasıdır.

Ağaç algoritmaları her zaman overfitting oluştururlar. Oluşan ağacın çok büyük olması bu etkiyi arttırır. Şekil 5.3' te görüldüğü gibi ağacın dengeli olabilmesi için belli bir büyüklüğün üzerinde olmalıdır. Ancak bu büyüklük arttıkça test verisinin hata oranı yükselmekte ve ağacın doğruluğu azalmaktadır. Bu noktadaki en önemli işlem ağaç budamasıdır. Şekilde de görüldüğü gibi ağaç belli bir büyüklükten sonra doğruluğunu kaybetmeye başlamaktadır. Şekil 5.3' ten de anlaşılacağı gibi, eğitim kümesini tutarlı kılmak için ağaç şişmeye başlarken, sağlama kümesi tutarsızlaşmaktadır. Dolayısı ile eğitim kümesi için hata oranını yükselten ağaç budama işlemi, aslında sadece eğitim kümesine özgü bazı kuralları temizlemektedir. Bu budama işlemi, sanıldığının aksine sağlama verisini daha çok doğrulamaktadır.



Şekil 5.3 : Eğitim kümesi ile test kümesinin ağaç boyuna göre doğruluk performansı.

Budama, ağacın daha sade görünmesi ve daha rahat yorumlanması için bir avantajdır. İki türlü ağaç budama tekniği bulunmaktadır.

- Ön Budama (Pre - Pruning)
- Sonradan Budama (Post Pruning)

Ön budama işlemi ağaç yaratılırken devreye girer. Ağaç oluşturulurken bir noktada ağacı durdurarak görevi yerine getirir. Örnek olarak, bölünen kümeler özellikleri belli bir eşik değerinin üstünde değilse o noktada ağaç bölümleme işlemi durdurulur ve o kümedeki baskın sınıf yaprak olarak yaratılır. Böylelikle gereksiz alt ağaçların oluşmasının önüne geçilmiş olur. Ön budama çok önemlidir, çünkü çalışma zamanında ağaca müdahele eder ve zamandan kazanır, bu nedenle verimli bir yöntemdir.

Ancak Breiman, bu yöntemin çelişkili performanslar verdiğini savunmuştur. Yöntem ağacı çok önceden durdurup, yeterli olgunluğa gelmesini engellemektedir. Bu tutarsızlık fark edildikten sonra ikinci bir teknik olan Post – Pruning yöntemi kullanılmıştır. Ancak, çok büyük veri yığınlarında verimli ve hızlı çalışma, tutarlı çalışmanın önüne geçmektedir. Bu sebeple ön budama yöntemi tamamen terkedilmiş bir yöntem değildir.

Post – pruning ağaç oluşturulduktan sonra devreye giren bir yöntemdir. Budanmamış T ağacını basitleştirirek T' ağacına çevirir. Budama yöntemi belirli alt ağaçların yerine o alt ağaçlarda en sık bulunan sınıfa sahip bir yaprak koyar. Bu yöntem, aslında eğitim kümesindeki hata oranını arttırır. Ancak yeni bir olayın test edilmesindeki hata oranını düşürür. Aslında budama işlemi hata yaratmaktan ziyade doğruluğu arttırmak olarak tanımlanabilir.

Post – pruning altında üç tip budama metodu bulunmaktadır:

- Alt ağaçları silme ve yaprak oluşturma
- Alt ağaçları yükseltme
- Dal kesme

6. VERİ MADENCİLİĞİNDE KULLANILAN YAZILIMLAR

Piyasadaki güçlü ve kullanışlı veri madenciliği programları incelenmiştir.

6.1 SPSS Clementine

1980' li yıllarda MatLab, MathCad, I-Think ve Stella gibi bilimsel programlama araçlarının kullandığı grafik programlama yaklaşımını kullanan ilk veri madenciliği programıdır. İngiltere' de geliştirilen Clementine, olduça uzun bir süredir kullanılmakta olup, araçları giderek iyileştirilmektedir. Clementine tarihi bilgi içeren veritabanları ve veriler için idealdir (Yurtsever, 2002).

Avantajları;

- Veri madenciliği algoritmaları çok çeşitlidir.
- Bir çok veri madenciliği algoritmasının içine yerleştirilmiş olukça güçlü, en optimal parametreyi araştıran programları mevcuttur.
- Kayıp gözlem tahmini veri kalitesinin kontrolü için araçlar mevcuttur.
- Bir modelleme algortimasının sonuçlarının diğer bir modelleme algoritmasında girdi olarak kullanılabileceği güçlü öğrenme modelleri kurulabilir.
- Ortalama kullanım kolaylığı vardır.

Dezavantajları;

- Diğer programlara göre az sayıda betimsel istatistik ve parametrik çözümlemeler mevcuttur.
- Dahili bazı özelliklerin daha gelişmişleri ayrı ürün olarak satılmaktadır (Yurtsever, 2002).

6.2 STATISTICA Data Miner

STATISTICA Data Miner, kendine has kategoride bir programdır. Veri madenciliği projelerindeki tüm görevleri kolaylaştırmadaki başarısı ve birçok işlemi başarıyla

gerçekleştirmesi açısından eşsizdir. STATISTICA Data Miner, diğer veri madenciliği programlarına göre oldukça fazla araç sunar (Nisbet, 2004).

Avantajları;

- Veri madenciliği için parametrik istatistik ve makine öğrenimi kombinasyonu zengin algoritmaları sunar.
- Diğer programlara göre grafiksel programlama arayüzüyle kullanımı kolaydır.
- Tüm ortak veri madenciliği görevleri için araçlar sunar.
- Model çıktısı için oldukça esnek araçlar mevcuttur.
- Boyut azaltmada kullanılan güçlü araçları mevcuttur.
- Visual Basic dilini temel alan güçlü geliştirme araçları mevcuttur.

Dezavantajları;

- Sinir ağı modellerinin değerlendirilmesi için kullanılan grafikler kolaylıkla elde edilememektedir (Nisbet, 2004).

6.3 Darwin

Darwin, Oracle firmasını geliştirdiği bir veri madenciliği yazılımıdır. Yapay sinir ağları, Bayesian öğrenme, k – en yakın komşu, regresyon ağaçları, kümeleme ve sınıflama gibi farklı algoritmaları destekleyen paralel sunucular için geliştirilmiş bir veri madenciliği yazılımıdır. Darwin; StarTree, CART karar ağacı, StarNet ve StarMatch makine öğrenme algoritmalarının kullanıldığı bazı modülleri de içermektedir (Yurtsever, 2002).

6.4 DBMiner

Kanada da Simon Fraser Üniversitesi Veritabanı Araştırma Laboratuarlarında geliştirilmiş bir veri madenciliği yazılım prototipidir. DBMiner bir araştırma prototipi olmaktan çıkıp ticari bir ürün haline gelmiştir. DBMiner' ı diğer veri madenciliği yazılımlarından ayıran en önemli özellik, OLAP yöntemleriyle veri madenciliği algoritmalarını bütünleştirmesidir. Bu bütünleşme sınucunda çevrimiçi analitik madencilik (Online Analytical Mining, OLAM) adı verilen yeni bir veri madenciliği metodolojisi ortaya çıkmıştır. OLAM metodolojisi ile verilerin çok

boyutlu görünümü elde edilmekte ve etkileşimli bir veri madenciliği ortamı yaratılmaktadır. Yazılım, ilişkisel veritabanlarından veya veri ambarlarından verileri alır ve bu verileri bütünleştirerek çok boyutlu veritabanlarına aktarır.

DBMiner yazılımına yapılan bazı eklemeler ile karmaşık verilerin üstesinden gelebilen, GeoMiner, MultiMediaMiner ve WebLogMiner gibi özelleşmiş veri madenciliği yazılımlarının prototipleri geliştirilmiştir (Yurtsever, 2002).

6.5 Enterprise Miner

SAS firmasının veri madenciliği aracıdır. Enterprise Miner karar ağaçları, yapay sinir ağları, regresyon çözümlemesi, kümeleme, zaman serileri, birliktelik kurallarının bulunması gibi veri madenciliği sorgularını ele alabilmektedir. Grafiksel arayüzü sayesinde kullanım kolaylığı sağlar. Diğer veri madenciliği programlarına oranla istatiksel ve regresyon açısından fazla sayıda araca sahiptir. Algoritma derinliği ve görsel arabirimi güçlü olduğu yönlerindendir.

6.6 Intelligent Miner

Eşleştirme, sınıflama, tahmin modelleme, gruplama, sıralı desen analizi, regresyon analizi gibi fonksiyonları içeren IBM firmasının geliştirdiği bir yazılımdır. Ayrıca yapay sinir ağları algoritmaları, istatistik metotları, veri hazırlama ve görsel gösterim araçları gibi ek özellikler içerir.

6.7 WEKA

Yeni Zelanda Waikito Üniversitesi' nde, nesneye yönelik programlama dillerinden biri olan Java ile geliştirilmiş ve halen yeni sürümleri geliştirilmeye devam eden açık kaynak kodlu bir veri madenciliği yazılımdır. WEKA'nın Java ile geliştirilmiş olması, Linux, Unix, Windows ve Macintosh gibi başlıca işletim sistemi platformlarında kullanılabilirliğini sağlamıştır. Ayrıca WEKA programı internet üzerinden bedelsiz olarak dağıtılmaktadır.

7. UYGULAMA

7.1 Veri Kümesinin Tanımlanması

Müşteri bağlılığının tespit edilmesinde zor olan kısım, müşterinin hangi koşullarda kaybedildiğinin ortaya çıkarılmasıdır. Bu aşamada, hesaba katılması gereken bir çok nitelik olabilir. Güçlü bir model oluşturabilmek için işin amaçlarına ve hedeflerine ters düşmeyecek bir "müşteri bağlılık tanımı" ortaya konmalıdır.

Bu tezde kullanılan veriler için müşteri bağlılık tanımı, aşağıdaki gibi belirlenmiştir:

- Emeklilik tarihinden önce sistemden ayrılan müşteriler

Çizelge 7.1' de analizi yapılan bireysel emeklilik sistemi müşterilerine ait hangi niteliklerin kullanıldığı sıralanmıştır.

SOZ_DURUM olarak belirtilen nitelik, müşterinin yukarıdaki tanıma göre bağlılık bilgisini içerir. Müşteri kaybedilmiş ise bu alan "AYRILMIŞ", kaybedilmemiş ise bu alan "AKTİF" değerini alır. Bir üçüncü durum olan "ARA VERME" statüsü de veride yer almaktadır. Fakat bu durum, tez çalışması sırasında inceleme dışında bırakılmıştır.

Veriyi sağlayan, Türkiye' de sektöründe ilk 5 arasında yer alan emeklilik şirketi tabloda belirtilen niteliklere sahip 35.000 adet kaydı paylaşmıştır. Bu kayıtların 12.981' i kaybedilmiş, 13.877'si kaybedilmemiş müşterilerden oluşmaktadır. Geri kalan 8.142 kayıt ara vermiş durumda olan müşterilere aittir.

Analizde, SAS Enterprise Miner 4.3 yazılımı kullanılmıştır. Kullanılan yöntem ise karar ağaçları algoritmasıdır.

Çizelge 7.1 : Veri kümesinde kullanılan nitelikler ve açıklamaları.

NİTELİK ADI	**AÇIKLAMA**
SOZLESME_NO	Sözleşme numarası
SOZ_KATILIMCI_NO	Sözleşme üzerindeki katılımcı müşteri numarası
SOZ_TEKLIF_TANZIM_TARIHI	Sözleşmeye ait teklifin tanzim tarihi
SOZ_YURURLUK_TARIHI	Sözleşmenin yürürlüğe giriş tarihi
SOZ_SON_DEGISIKLIK_TARIHI	Sözleşmenin son değişiklik tarihi
SOZ_EMEKLILIK_TARIHI	Katılımcının emeklilik tarihi
SOZ_BIREYSEL_KURUMSAL	Sözleşmenin bireysel mi kurumsal mı olduğu bilgisi
SOZ_URUN_NO	Sözleşmenin ürün tipi
SOZ_TECDIT_YILI	Sözleşme tecdit (yenileme) yılı
SOZ_DURUM	Sözleşmenin statüsü (Aktif, Ara Vermiş, Ayrılmış)
SOZ_SON_DEGISIKLIK_KODU	Sözleşme üzerindeki son değişiklik kodu
SOZ_SON_DEGISIKLIK	Sözleşme üzerindeki son değişiklik kodunun açıklaması
SOZ_SON_DEGISIKLIK_GRUP	Sözleşme üzerindeki son değişikliğin bağlı olduğu grup
SOZ_ODEME_SEKLI	Sözleşme ödeme tipi
SOZ_ODEME_SIKLIGI	Sözleşme ödeme dönemi
SOZ_BOLGE_NO	Sözleşmenin bağlı olduğu bölge numarası
SOZ_BOLGE_ADI	Sözleşmenin bağlı olduğu bölge adı
SOZ_ARACI_GRUBU	Sözleşmenin aracı tipi
SOZ_KAYNAK	Sözleşmenin kaynağı
SOZ_ILK_KAYNAK	Sözleşmenin ilk kaynağı
SOZ_KAYNAK_SATIS	Sözleşmenin satış kanalı
SOZ_KAMPANYA	Sözleşme kampanya dahilinde satıldı ise kampanya adı
SOZ_GIRIS_AIDATI_PARA_BIRIMI	Sözleşme giriş aidatı para birimi
SOZ_GIRIS_AIDATI_DOVIZ	Sözleşme giriş aidatı tutarı (döviz cinsinden)
SOZ_CIKIS_AIDATI_DOVIZ	Sözleşme çıkış aidatı tutarı (döviz cinsinden)
SOZ_TOPLAM_KATKI_PAYI	Sözleşme ödenen toplam katkı payı tutarı
SOZ_TAHAKKUK_TOPLAM	Sözleşme tahakkuk eden toplam tutar
SOZ_GIRIS_AIDATI_ODEMESI	Sözleşmeye ödenen giriş aidatı tutarı
SOZ_ISVEREN_KATKI_PAYI_ODEMESI	Sözleşmeye ödenen işveren katkı payı tutarı
SOZ_CALISAN_KATKI_PAYI_ODEMESI	Sözleşmeye ödenen çalışan katkı payı tutarı
SOZ_ARA_FONU_ODEMESI	Sözleşmeye ödenen ara fonu tutarı
SOZ_BASLANGIC_FONU_ODEMESI	Sözleşmeye ödenen başlangıç fonu tutarı
SOZ_KAMPANYA_ODEMESI	Sözleşmeye ödenen kampanya tutarı
SOZ_YATIRIMA_YONLENDIRILEN	Sözleşme yatırıma yönlendirilmiş toplam tutar
CINSIYET	Katılımcının cinsiyeti
MEDENI_DURUM	Katılımcının medeni durumu
EGITIM_DURUMU	Katılımcının eğitim durumu
DOGUM_TARIHI	Katılımcının doğum tarihi
UYRUK	Katılımcının uyruğu
SOSYAL_GUVENLIK_TIPI	Katılımcının bağlı olduğu sosyal güvenlik kurum tipi
CALISMA_SEKTORU	Katılımcının çalışma sektörü
MESLEK	Katılımcının mesleği
CEPTEL_VARMI	Katılımcının cep telefonu bilgisinin olup olmadığı
EPOSTA_VARMI	Katılımcının e-posta bilgisinin olup olmadığı
ILETISIM_ILI	Katılımcının iletişim ili

Çizelge 7.2 : Veri kümesinde kullanılan niteliklerin özellikleri.

NİTELİK ADI	VERİ TİPİ	MIN.	MAX.	ORT.	STD. SAP.
SOZLESME_NO	number	1	35000		
SOZ_KATILIMCI_NO	number	1	34029		
SOZ_TEKLIF_TANZIM_TARIHI	date	31/12/2003	31/12/2009		
SOZ_YURURLUK_TARIHI	date	31/12/2003	27/03/2010		
SOZ_SON_DEGISIKLIK_TARIHI	date	31/12/2003	21/04/2010		
SOZ_EMEKLILIK_TARIHI	date	26/07/2000	14/07/2047		
SOZ_BIREYSEL_KURUMSAL	varchar2				
SOZ_URUN_NO	number	1	901		
SOZ_TECDIT_YILI	number	1	7		
SOZ_DURUM	varchar2				
SOZ_SON_DEGISIKLIK_KODU	number	11	96		
SOZ_SON_DEGISIKLIK	varchar2				
SOZ_SON_DEGISIKLIK_GRUP	varchar2				
SOZ_ODEME_SEKLI	varchar2				
SOZ_ODEME_SIKLIGI	varchar2				
SOZ_BOLGE_NO	number	1	9		
SOZ_BOLGE_ADI	varchar2				
SOZ_ARACI_GRUBU	varchar2				
SOZ_KAYNAK	varchar2				
SOZ_ILK_KAYNAK	varchar2				
SOZ_KAYNAK_SATIS	varchar2				
SOZ_KAMPANYA	varchar2				
SOZ_GIRIS_AIDATI_PARA_BIRIMI	varchar2				
SOZ_GIRIS_AIDATI_DOVIZ	number	0	168,00	76,79	56,70
SOZ_CIKIS_AIDATI_DOVIZ	number	0	342,00	100,65	75,04
SOZ_TOPLAM_KATKI_PAYI	number	10	2.570,00	156,41	168,15
SOZ_TAHAKKUK_TOPLAM	number	0	124.000,00	1.432,70	4.191,20
SOZ_GIRIS_AIDATI_ODEMESI	number	0	227,93	59,11	54,52
SOZ_ISVEREN_KATKI_PAYI_ODM	number	0	30.717,00	259,44	1.532,90
SOZ_CALISAN_KATKI_PAYI_ODM	number	0	69.680,00	2.519,20	4.561,80
SOZ_ARA_FONU_ODEMESI	number	0	24.300,00	39,72	743,53
SOZ_BASLANGIC_FONU_ODM	number	0	100.000,00	167,52	3.329,20
SOZ_KAMPANYA_ODEMESI	number	0	1.369,60	17,08	92.152,00
SOZ_YATIRIMA_YONLENDIRILEN	number	0	124.000,00	2.962,90	6.197,30
CINSIYET	varchar2				
MEDENI_DURUM	varchar2				
EGITIM_DURUMU	varchar2				
DOGUM_TARIHI	date	20/04/1923	14/12/1991		
UYRUK	varchar2				
SOSYAL_GUVENLIK_TIPI	varchar2				
CALISMA_SEKTORU	varchar2				
MESLEK	varchar2				
CEPTEL_VARMI	varchar2				
EPOSTA_VARMI	varchar2				
ILETISIM_ILI	varchar2				

7.2 Veri Temizleme ve Dönüştürme

Veriyi dönüştürmek için aşağıdaki işlemler yapılmıştır:

DOGUM_TARIHI: Bu nitelik, müşterinin yaşının hesaplamasında kullanılmıştır. Yeni hesaplanan nitelik YAS olarak adlandırılmıştır.

SOZ_TEKLIF_TANZIM_TARIHI : Bu nitelikten teklif tanzim tarihi bilgisinin sadece yıl kısmı alınarak TEKLIF_YIL niteliği oluşturulmuştur.

SOZ_YURURLUK_TARIHI : Bu nitelikten sözleşme yürürlük tarihi bilgisinin sadece yıl kısmı alınarak YURURLUK_YIL niteliği oluşturulmuştur.

SOZ_SON_DEGISIKLIK_TARIHI : Bu nitelikten sözleşme son değişiklik tarihi bilgisinin sadece yıl kısmı alınarak SON_DEG_YIL niteliği oluşturulmuştur.

SOZ_EMEKLILIK_TARIHI : Bu nitelikten katılımcı emeklilik tarihi bilgisinin sadece yıl kısmı alınarak EMEKLILIK_YIL niteliği oluşturulmuştur.

SOZ_ISVEREN_KATKI_PAYI_ODEMESI,

SOZ_CALISAN_KATKI_PAYI_ODEMESI, SOZ_ARA_FONU_ODEMESI,

SOZ_BASLANGIC_FONU_ODEMESI ve SOZ_KAMPANYA_ODEMESI nitelikleri toplanarak TOPLAM_TAHSILAT niteliği oluşturulmuştur.

TOPLAM_TAHSILAT niteliğinden, SOZ_YATIRIMA_YONLENDIRILEN niteliği çıkartılarak KESINTILER niteliği oluşturulmuştur.

7.3 Veri Modelleme ve Karar Ağacı Algoritmasının Uygulanması

Veri temizleme ve dönüşümünden sonraki adım modelleme adımıdır. Bu çalışmada model kurma aşamasında SAS Enterprise Miner 4.3 programı kullanılmıştır.

Bu tezde kullanılan veriler, excel formatında hazırlanmış olup SAS EM' a bu formatta yüklenmiştir.

Kurulacak modeli belirlemek için, farklı algoritmalar ile veri üzerinde çalıştırılarak doğrulukları karşılaştırılmalıdır. Çapraz geçerleme (cross validation) ya da yüzde ayırma (percentage split) seçeneklerini kullanarak kurulan modellerin doğruluğunu ölçmek mümkündür. Çapraz geçerleme önceki bölümlerde ifade edildiği gibi veri kümesini n eş büyüklükte (program default olarak 10 katman kullanmaktadır) katmana ayırarak; her iterasyonda n-1 parçayı modelin öğrenilmesinde 1 parçayı ise

modelin testinde kullanan bir yöntemdir. Yüzde ayırma yönteminde ise veri kümesinin belirli bir yüzdesi (program default olarak %66 kullanmaktadır) modelin öğrenilmesine kalan kısmı ise modelin testine ayrılmaktadır.

Tez çalışmasında, çapraz geçerleme yöntemi kullanılmıştır. Kullanılacak eğitim ve test kümesinin büyüklüklerinin belirlenmesi için Çizelge 7.3' te detayı belirtilen eğitim ve test kümesi oranları ile karar ağaçları oluşturularak doğruluk oranları hesaplanmıştır. Çalışmada, verinin %40'ı eğitim kümesi ve geri kalan %60' ı test kümesi olacak şekilde bölümlenmesine karar verilmiştir. Doğruluk oranın yüksek, yaprak sayısının küçük olması göz önünde bulundurulmuştur.

Çizelge 7.3 : Eğitim ve test kümesi büyüklük değişimine göre doğruluk oranları.

Eğitim Kümesi	Test Kümesi	Test Kümesi Doğruluk Oranı	Yaprak Sayısı
%90	%10	0,8918	24
%80	%10	0,9012	26
%70	%10	0,9035	22
%60	%10	0,9078	21
%50	%10	0,9154	21
%40	%10	0,929	13
%30	%10	0,9546	11
%20	%10	0,9635	7
%10	%10	0,9628	2
%5	%10	0,9836	3

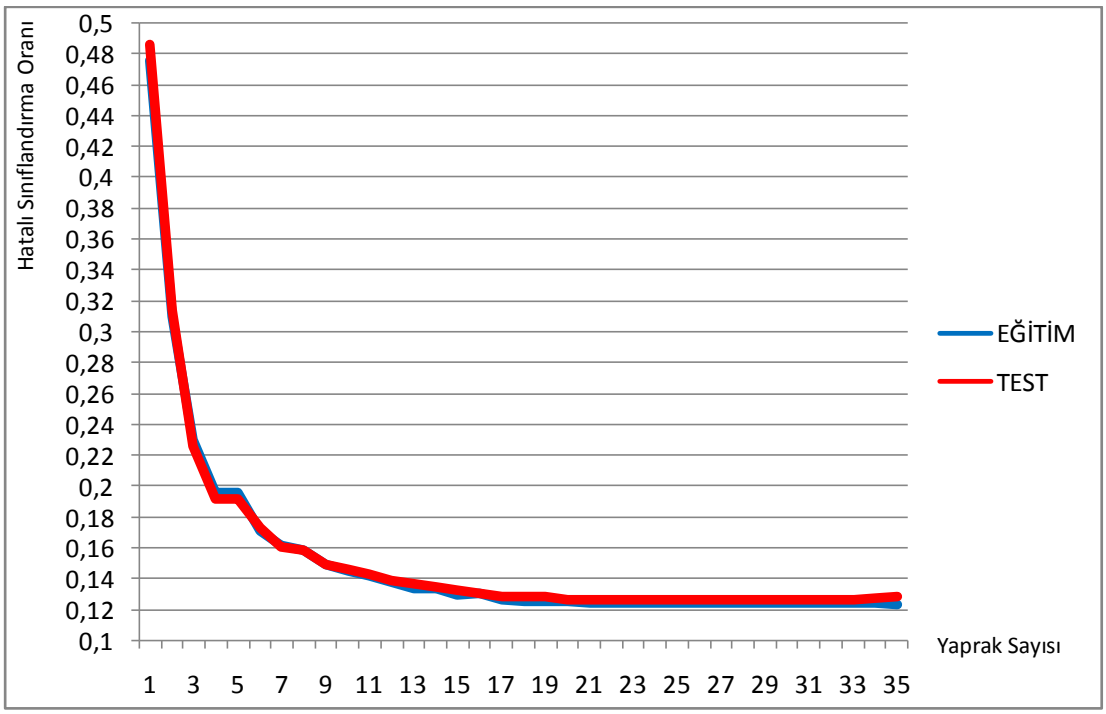
Tez çalışmasında amaç, ayrılma eğilimi gösteren müşteri karakteristiklerini tanımlamak ve bu müşterileri elde tutmaya yönelik programlar ve pazarlama politikaları geliştirmektir. Ayrılma eğilimi taşıyan müşteri profilini belirlemek için sınıflandırma teknikleri kullanılmıştır. "SOZ_DURUM" niteliği, hedef değişken (target variable) olarak belirlenmiştir. 3 farklı sınıflandırma algoritması kullanılarak, kurulan modelin güvenilirliği araştırılmıştır.

İlk sınıflandırma işleminde, Chi-Square dallanma yöntemini kullanan karar ağacı algoritması oluşturulmuştur. Karar ağacı, Çizelge 7.4' te belirtilen 9 farklı düğüme sahiptir. Çizelgedeki, düğüm sayısı alanı ilgili niteliğin ağaçta kaç kez düğüm olarak kullandığını belirtmektedir. TOPLAM_TAHSILAT niteliği, oluşturulan ağaçta 6 kez kullanılmıştır. Bu niteliğin önem derecesi, hem eğitim hem de test setinde 1 olarak hesaplanmıştır. Bu niteliğin ardından YURURLUK_YIL niteliği karar ağacından belirleyici olarak kullanılmıştır.

Çizelge 7.4 : Chi-Square dallanma yöntemi kullanan karar ağacında düğümler.

NİTELİK ADI	DÜĞÜM SAYISI	EĞİTİM	TEST
TOPLAM_TAHSILAT	6	1,00	1,00
YURURLUK_YIL	3	0,84	0,79
SOZ_TOPLAM_KATKI_PAYI	4	0,33	0,34
KESINTILER	2	0,31	0,26
SOZ_BIREYSEL_KURUMSAL	1	0,27	0,26
TEKLIF_YIL	1	0,22	0,20
SOZ_GIRIS_AIDATI_ODEMESI	1	0,15	0,11
CALISMA_SEKTORU	1	0,13	0,07
MESLEK	1	0,11	0,09

Şekil 7.1, Chi-Square dallanma yöntemini kullanan karar ağacının hatalı sınıflandırma oranının yaprak sayısına değişimini göstermektedir. Şekilde görüldüğü gibi ağacın dengeli olabilmesi için belli bir büyüklüğün üzerinde olmalıdır. Ancak bu büyüklük arttıkça test verisinin hata oranı yükselmekte ve ağacın doğruluğu azalmaktadır. Bu noktadaki en önemli işlem ağaç budamasıdır. Şekilde de görüldüğü gibi ağaç belli bir büyüklükten sonra doğruluğunu kaybetmeye başlamaktadır. Şekil 7.1' den de anlaşılacağı gibi, eğitim kümesini tutarlı kılmak için ağaç gereğinden çok büyümeye başlarken, sağlama kümesi tutarsızlaşmaktadır. Chi-Square dallanma yöntemini ile oluşturulan karar ağacında 21 yaprak bulunmaktadır.



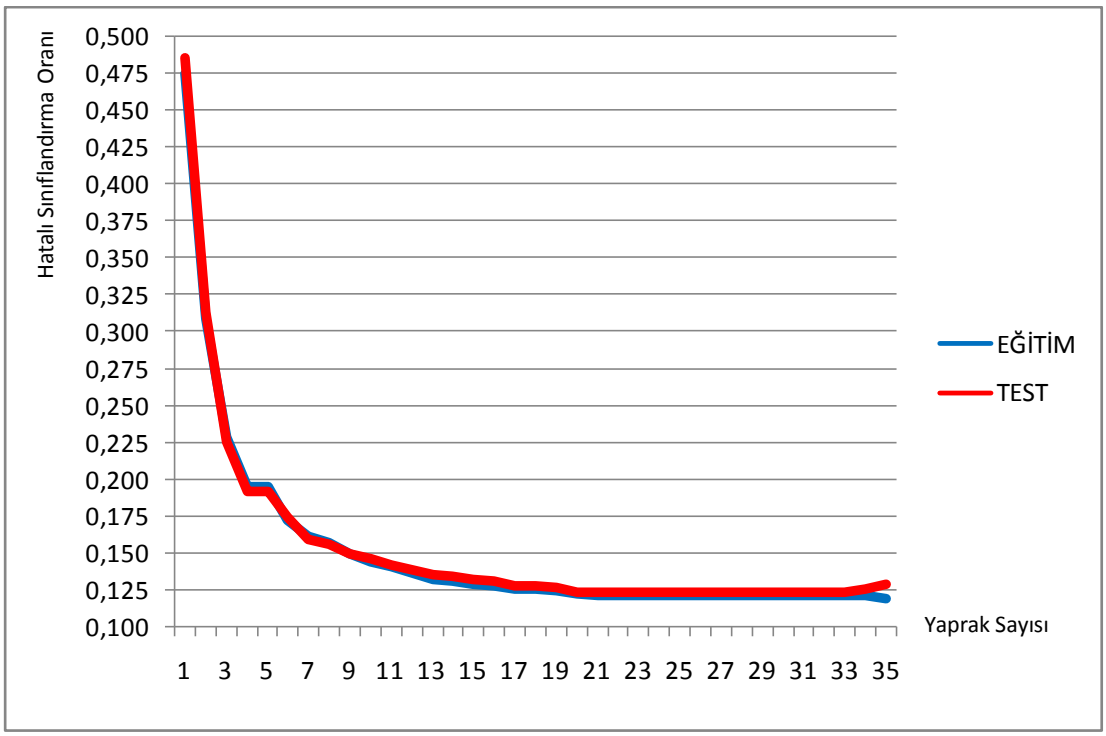
Şekil 7.1 : Chi-Square hatalı sınıflandırma oranı – yaprak sayısı dağılımı.

İkinci sınıflandırma işleminde, entropy tabanlı dallanma yöntemini kullanan karar ağacı algoritması oluşturulmuştur. Karar ağacı, Çizelge 7.5' te belirtilen 10 farklı düğüme sahiptir. TOPLAM_TAHSILAT niteliği, oluşturulan ağaçta 6 kez kullanılmıştır. Bu niteliğin önem derecesi, hem eğitim hem de test setinde 1 olarak hesaplanmıştır. Bu niteliğin ardından YURURLUK_YIL niteliği karar ağacından belirleyici olarak kullanılmıştır. Chi-Square dallanma yöntemi ile oluşturulan karar ağacından farklı olarak, entropy tabanlı oluşturulan karar ağacında SOZ_KAYNAK, SOZ_CIKIS_AIDATI_DOVIZ, SOZ_GIRIS_AIDATI_DOVIZ nitelikleri belirleyici olarak kullanılmıştır.

Çizelge 7.5 : Entropy dallanma yöntemi kullanan karar ağacında düğümler.

NİTELİK ADI	**DÜĞÜM SAYISI**	**EĞİTİM**	**TEST**
TOPLAM_TAHSILAT	6	1,00	1,00
YURURLUK_YIL	3	0,84	0,80
KESINTILER	1	0,28	0,24
SOZ_KAYNAK	1	0,27	0,26
TEKLIF_YIL	2	0,26	0,26
SOZ_TOPLAM_KATKI_PAYI	2	0,26	0,27
SOZ_CIKIS_AIDATI_DOVIZ	2	0,20	0,18
SOZ_GIRIS_AIDATI_DOVIZ	1	0,15	0,15
SOZ_GIRIS_AIDATI_ODEMESI	1	0,15	0,11
MESLEK	1	0,11	0,09

Şekil 7.2, entropy tabanlı dallanma yöntemini kullanan karar ağacının hatalı sınıflandırma oranının yaprak sayısına değişimini göstermektedir. Şekilde görüldüğü gibi ağaç belli bir büyüklükten sonra doğruluğunu kaybetmeye başlamaktadır. Şekil 7.2' den de anlaşılacağı gibi, eğitim kümesini tutarlı kılmak için ağaç gereğinden fazla büyümeye başlarken, sağlama kümesi tutarsızlaşmaktadır. Entropy dallanma yöntemini ile oluşturulan karar ağacında 21 yaprak bulunmaktadır.



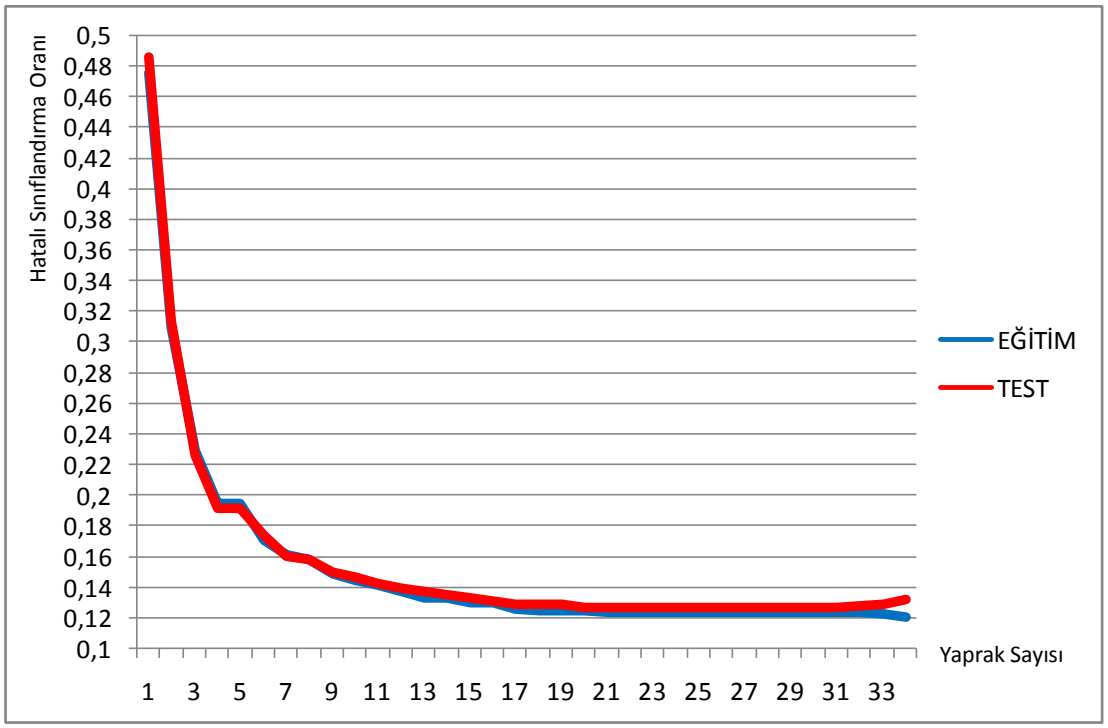
Şekil 7.2 : Entropy hatalı sınıflandırma oranı – yaprak sayısı dağılımı.

Son olarak sınıflandırma işleminde, gini index dallanma yöntemini kullanan karar ağacı algoritması oluşturulmuştur. Karar ağacı, Çizelge 7.6' da belirtilen 9 farklı düğüme sahiptir. TOPLAM_TAHSILAT niteliği, oluşturulan ağaçta 6 kez kullanılmıştır. Bu niteliğin önem derecesi, hem eğitim hem de test setinde 1 olarak hesaplanmıştır. Bu niteliğin ardından YURURLUK_YIL niteliği karar ağacından belirleyici olarak kullanılmıştır. Entropy dallanma yönteimine göre oluşturulan karar ağacından farklı olarak, gini index dallanma yöntemini kullanan karar ağacında CALISMA_SEKTORU, MESLEK nitelikleri belirleyici olarak kullanılmıştır.

Çizelge 7.6 : Gini index dallanma yöntemi kullanan karar ağacında düğümler.

NİTELİK ADI	DÜĞÜM SAYISI	EĞİTİM	TEST
TOPLAM_TAHSILAT	6	1,00	1,00
YURURLUK_YIL	3	0,84	0,79
SOZ_TOPLAM_KATKI_PAYI	4	0,33	0,34
KESINTILER	2	0,31	0,26
SOZ_KAYNAK	1	0,27	0,26
TEKLIF_YIL	1	0,22	0,20
SOZ_GIRIS_AIDATI_ODEMESI	1	0,15	0,11
CALISMA_SEKTORU	1	0,13	0,07
MESLEK	1	0,11	0,09

Şekil 7.3, gini index dallanma yöntemini kullanan karar ağacının hatalı sınıflandırma oranının yaprak sayısına değişimini göstermektedir. Şekilde görüldüğü gibi ağaç belli bir büyüklükten sonra doğruluğunu kaybetmeye başlamaktadır. Şekil 7.3' ten de anlaşılacağı gibi, eğitim kümesini tutarlı kılmak için ağaç gereğinden fazla büyümeye başlarken, sağlama kümesi tutarsızlaşmaktadır. Entropy dallanma yöntemini ile oluşturulan karar ağacında 21 yaprak bulunmaktadır.



Şekil 7.3 : Gini index hatalı sınıflandırma oranı – yaprak sayısı dağılımı.

Confusion matrix: Yakınsaklık matrisi olarak da adlandırılır. Doğru olarak sınıflandırılan örneklerin sayısı bu matrisin diyagonal elemanlarının toplamına eşittir.

Çizelge 7.7' de testleri yapılmış 3 karar ağacı algoritmasının yakınsaklık matrisi bulunmaktadır. Ayrıca, çizelgede eğitim ve test kümeleri için ağacın doğruluk oranları hesaplanmıştır. En iyi sonucu veren algoritmanın entropy dallanama yöntemini kullanan olduğu görülmektedir.

Çizelge 7.7 : Karar ağacı yakınsaklık matrisi.

	EĞİTİM KÜMESİ			DOĞRULUK ORANI	TEST KÜMESİ			DOĞRULUK ORANI
chi-square	1	0	Toplam		1	0	Toplam	
1	5288	351	5639		7674	564	8238	
0	983	4121	5104		1528	6349	7877	
Toplam	6271	4472	10743	0,8758	9202	6913	16115	0,8702

entropy	1	0	Toplam		1	0	Toplam	
1	5300	339	5639		7722	516	8238	
0	967	4137	5104		1513	6364	7877	
Toplam	6267	4476	10743	0,8784	9235	6880	16115	0,8741

gini index	1	0	Toplam		1	0	Toplam	
1	5288	351	5639		7681	557	8238	
0	983	4121	5104		1535	6342	7877	
Toplam	6271	4472	10743	0,8758	9216	6899	16115	0,8702

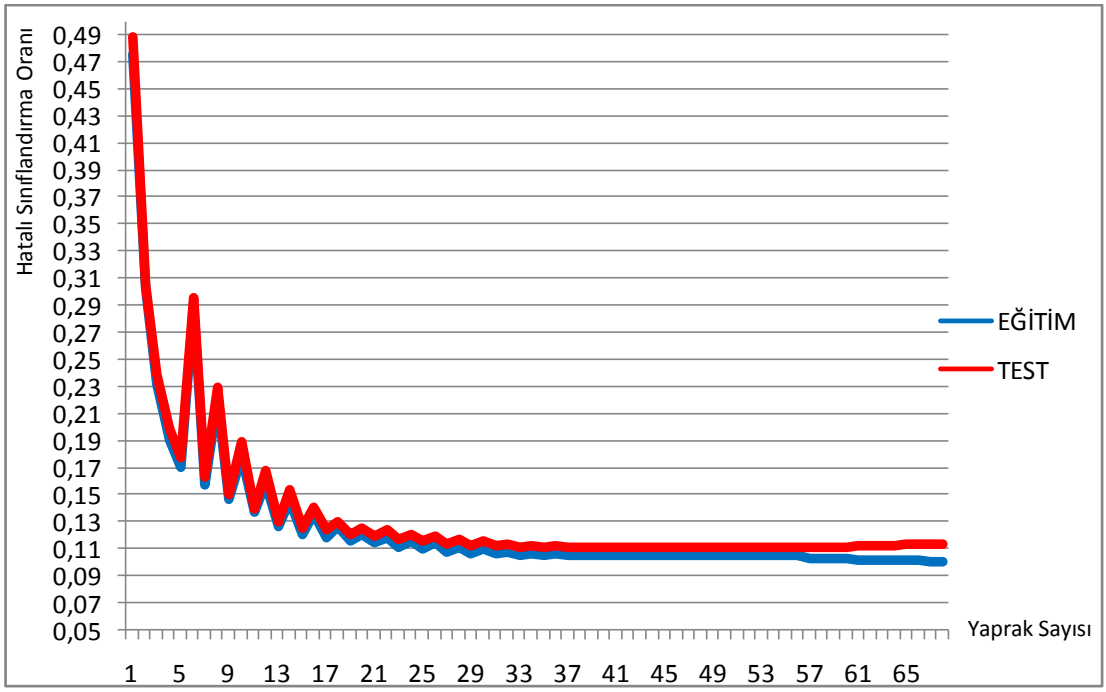
Karşılaştırması yapılan 3 farklı karar ağacı algoritmasının arasından en iyi sonucu verenin entropy dallanma yöntemini kullanan olduğunun görülmesinin ardından ağacın doğruluğunu daha yüksek seviyeye taşıyabilmek için farklı bir çalışma yapılmıştır. Şu ana kadar oluşturulan karar ağaçlarda, dallanma kriteri 2 olarak kabul edilmiştir. Dallanma kriteri 3 ve 4 olan entropy dallanma yöntemini kullanan karar ağaçları da oluşturularak sonuçları incelenmiştir.

Entropy-3 tabanlı karar ağacı, Çizelge 7.8' de belirtilen 8 farklı düğüme sahiptir. TOPLAM_TAHSILAT niteliği, oluşturulan ağaçta 5 kez kullanılmıştır. Bu niteliğin önem derecesi, hem eğitim hem de test setinde 1 olarak hesaplanmıştır. Bu niteliğin ardından TEKLIF_YIL niteliği karar ağacından belirleyici olarak kullanılmıştır. Entropy-2 tabanlı oluşturulan karar ağacından farklı olarak, entropy-3 tabanlı oluşturulan karar ağacında CALISMA_SEKTORU niteliği belirleyici olarak kullanılmıştır.

Çizelge 7.8 : Entropy-3 dallanma yöntemi kullanan karar ağacında düğümler.

NİTELİK ADI	DÜĞÜM SAYISI	EĞİTİM	TEST
TOPLAM_TAHSILAT	5	1,00	1,00
TEKLIF_YIL	1	0,78	0,78
SOZ_CIKIS_AIDATI_DOVIZ	7	0,48	0,49
SOZ_GIRIS_AIDATI_DOVIZ	1	0,29	0,29
SOZ_GIRIS_AIDATI_ODEMESI	1	0,26	0,27
CALISMA_SEKTORU	1	0,19	0,17
SOZ_TOPLAM_KATKI_PAYI	1	0,16	0,16
KESINTILER	1	0,13	0,10

Şekil 7.4, entropy-3 tabanlı dallanma yöntemini kullanan karar ağacının hatalı sınıflandırma oranının yaprak sayısına değişimini göstermektedir. Şekilde görüldüğü gibi ağaç belli bir büyüklükten sonra doğruluğunu kaybetmeye başlamaktadır. Şekil 7.4' ten de anlaşılacağı gibi, eğitim kümesini tutarlı kılmak için ağaç gereğinden fazla büyümeye başlarken, sağlama kümesi tutarsızlaşmaktadır. Entropy-3 tabanlı hesaplama yöntemini ile oluşturulan karar ağacında 37 yaprak bulunmaktadır.



Şekil 7.4 : Entropy-3 hatalı sınıflandırma oranı – yaprak sayısı dağılımı.

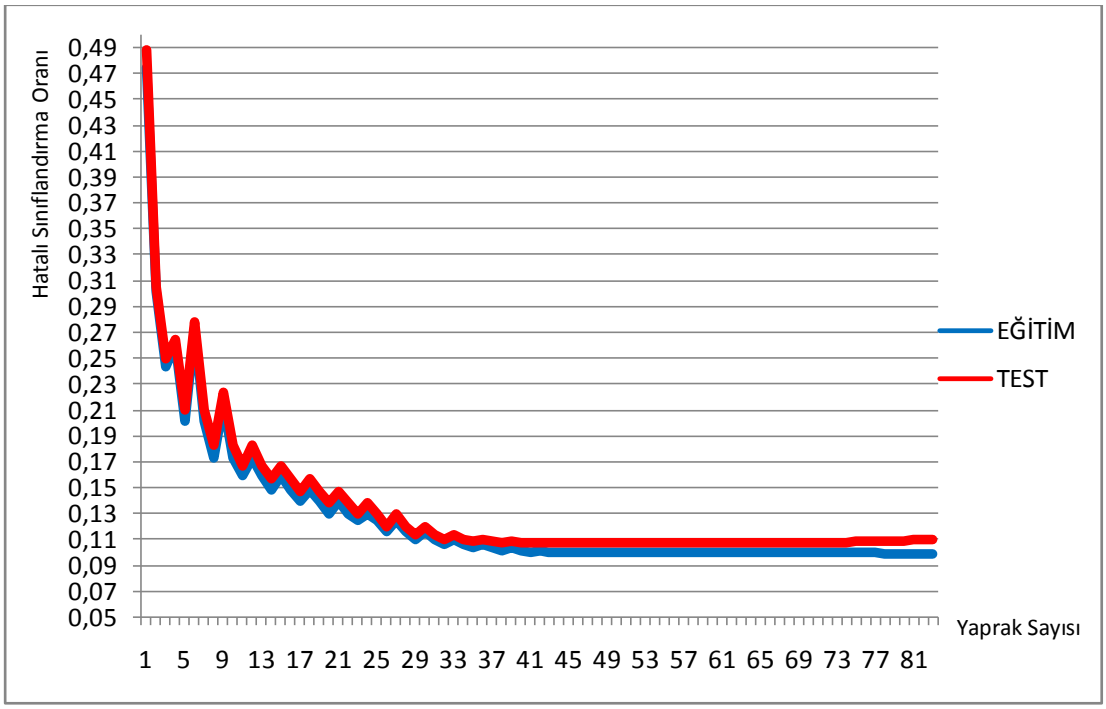
Entropy-4 dallanma yöntemini kullanan karar ağacı, Çizelge 7.9' da belirtilen 7 farklı düğüme sahiptir. TOPLAM_TAHSILAT niteliği, oluşturulan ağaçta 4 kez kullanılmıştır. Bu niteliğin önem derecesi, hem eğitim hem de test setinde 1 olarak hesaplanmıştır. Bu niteliğin ardından TEKLIF_YIL niteliği karar ağacından

belirleyici olarak kullanılmıştır. Entropy-3 dallanma yöntemi ile oluşturulan karar ağacından farklı olarak, entropy-4 dallanma yöntemi ile oluşturulan karar ağacında EMEKLILIK_YIL niteliği belirleyici olarak kullanılmıştır.

Çizelge 7.9 : Entropy-4 dallanma yöntemi kullanan karar ağacında düğümler.

NİTELİK ADI	DÜĞÜM SAYISI	EĞİTİM	TEST
TOPLAM_TAHSILAT	4	1,00	1,00
TEKLIF_YIL	1	0,80	0,81
SOZ_CIKIS_AIDATI_DOVIZ	6	0,50	0,51
SOZ_GIRIS_AIDATI_ODEMESI	1	0,26	0,29
KESINTILER	1	0,26	0,26
SOZ_TOPLAM_KATKI_PAYI	1	0,22	0,20
EMEKLILIK_YIL	1	0,10	0,06

Şekil 7.5, entropy-4 dallanma yöntemini kullanan karar ağacının hatalı sınıflandırma oranının yaprak sayısına değişimini göstermektedir. Şekilde görüldüğü gibi ağaç belli bir büyüklükten sonra doğruluğunu kaybetmeye başlamaktadır. Şekil 7.5' ten de anlaşılacağı gibi, eğitim kümesini tutarlı kılmak için ağaç gereğinden fazla büyümeye başlarken, sağlama kümesi tutarsızlaşmaktadır. Entropy-4 dallanma yöntemi ile oluşturulan karar ağacında 46 yaprak bulunmaktadır.



Şekil 7.5 : Entropy-4 hatalı sınıflandırma oranı – yaprak sayısı dağılımı.

Çizelge 7.10' da testleri yapılmış 3 karar ağacı algoritmasının yakınsaklık matrisi bulunmaktadır. Ayrıca, çizelgede eğitim ve test kümeleri için ağacın doğruluk

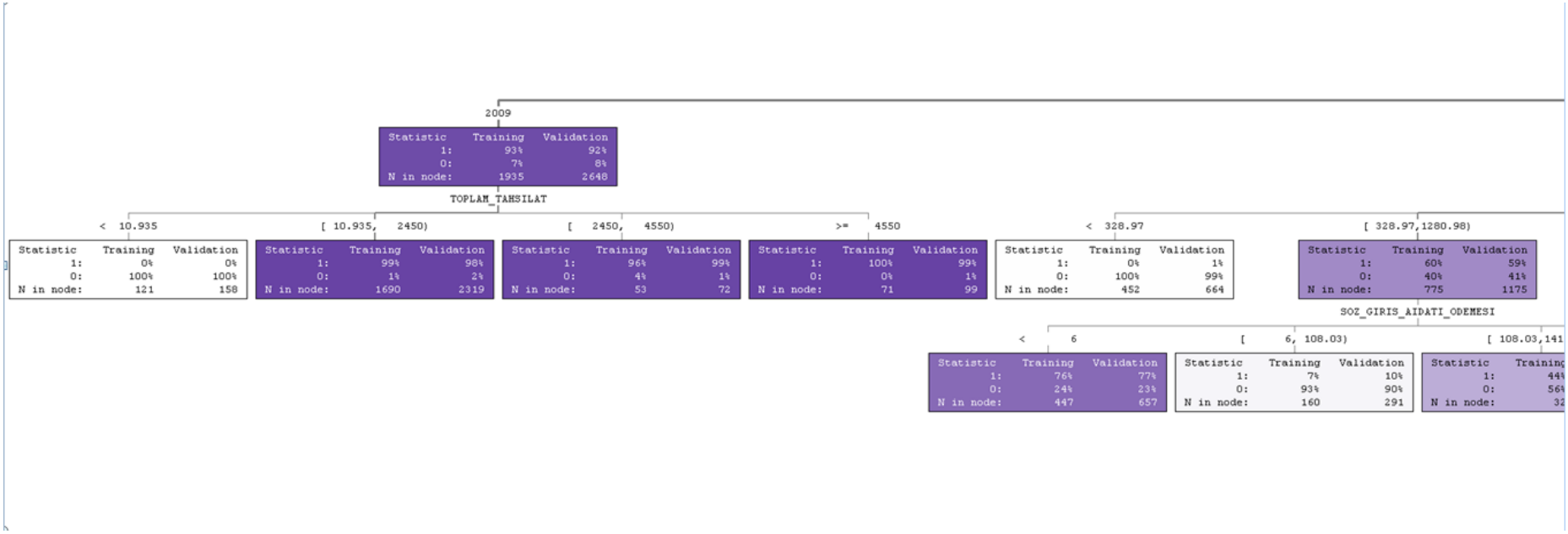
oranları hesaplanmıştır. En iyi sonucu veren algoritmanın, entropy-4 dallanma yöntemini kullanan olduğu görülmektedir.

Çizelge 7.10 : Entropy tabanlı karar ağaçlarının yakınsaklık matrisi.

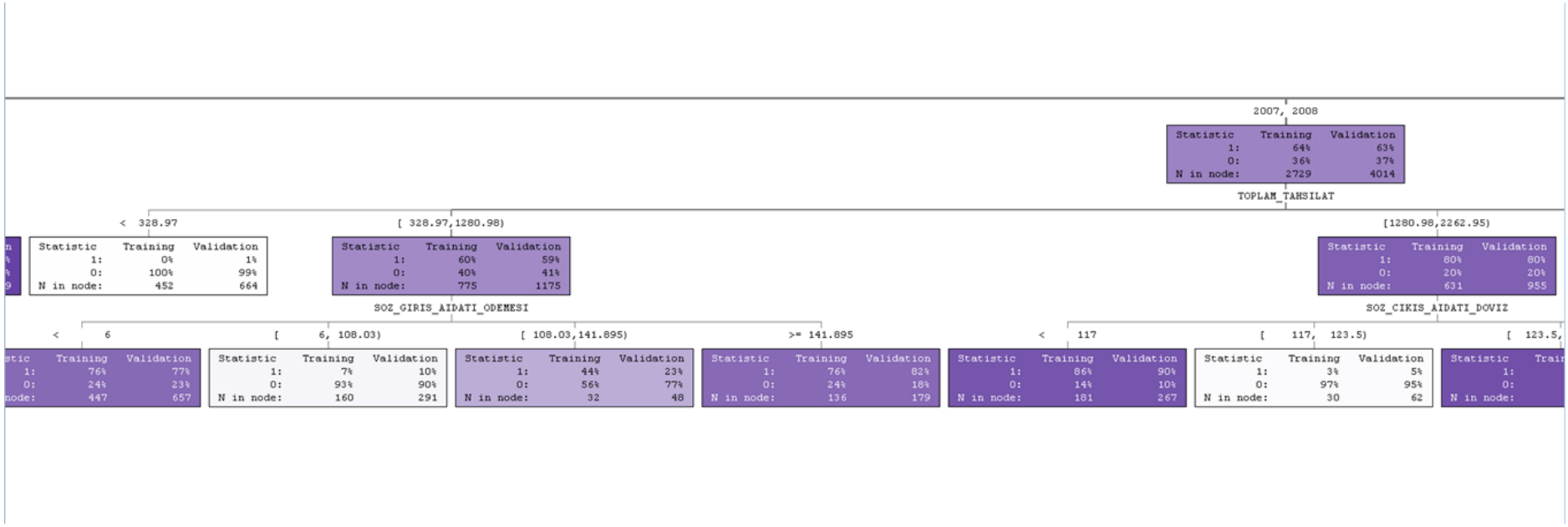
	EĞİTİM KÜMESİ			DOĞRULUK ORANI	TEST KÜMESİ			DOĞRULUK ORANI
entropy-2	1	0	Toplam		1	0	Toplam	
1	5300	339	5639		7722	516	8238	
0	967	4137	5104		1513	6364	7877	
Toplam	6267	4476	10743	0,8784	9235	6880	16115	0,8741

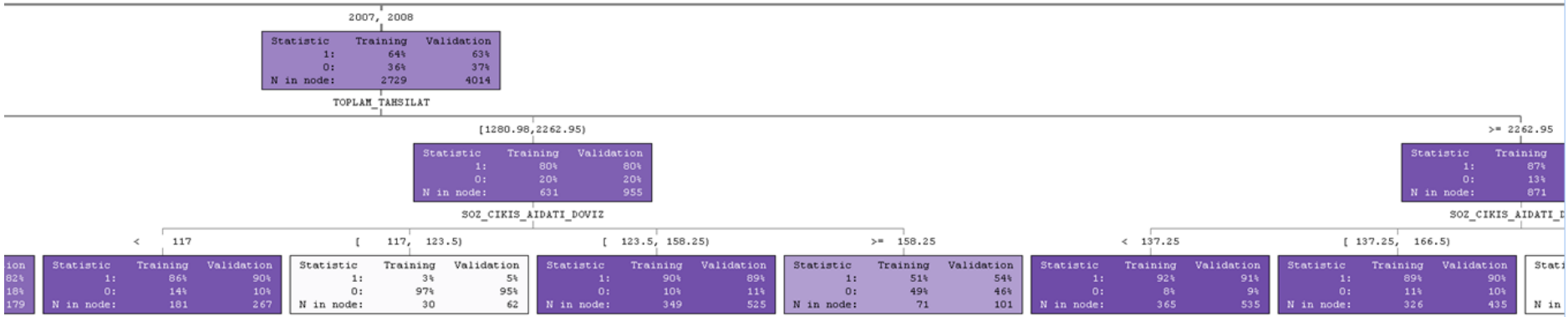
entropy-3	1	0	Toplam		1	0	Toplam	
1	5322	317	5639		7670	568	8238	
0	820	4284	5104		1220	6657	7877	
Toplam	6142	4601	10743	0,8942	8890	7225	16115	0,8890

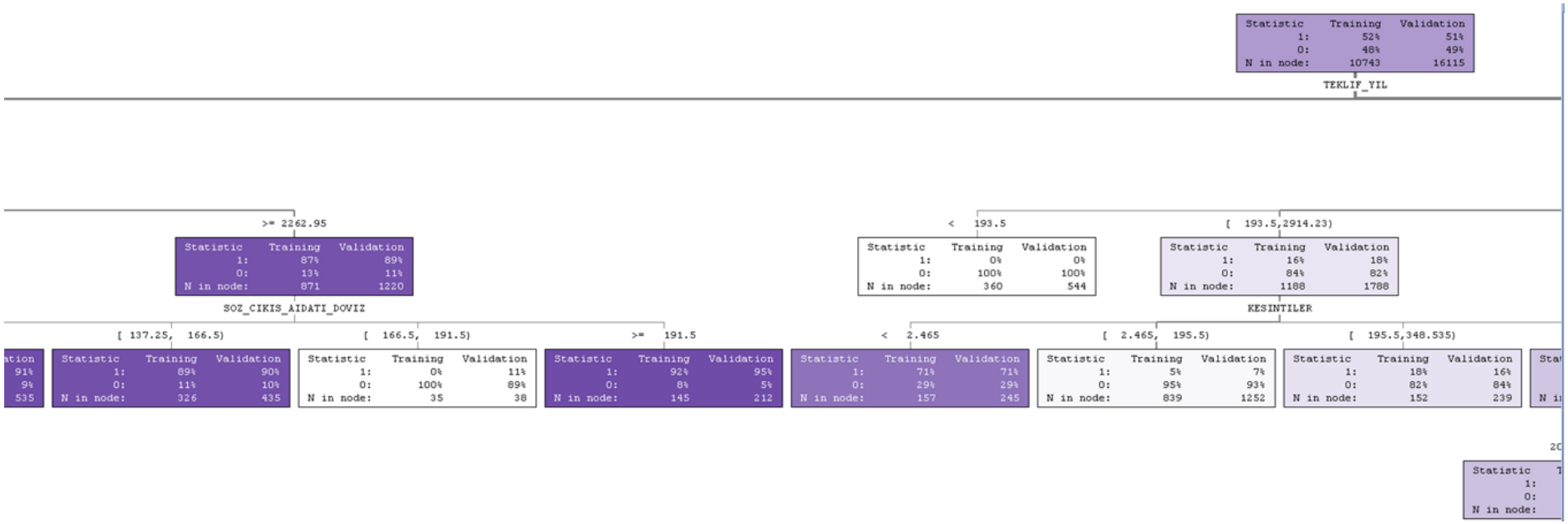
entropy-4	1	0	Toplam		1	0	Toplam	
1	5259	380	5639		7610	628	8238	
0	706	4398	5104		1107	6770	7877	
Toplam	5965	4778	10743	0,8989	8717	7398	16115	0,8923

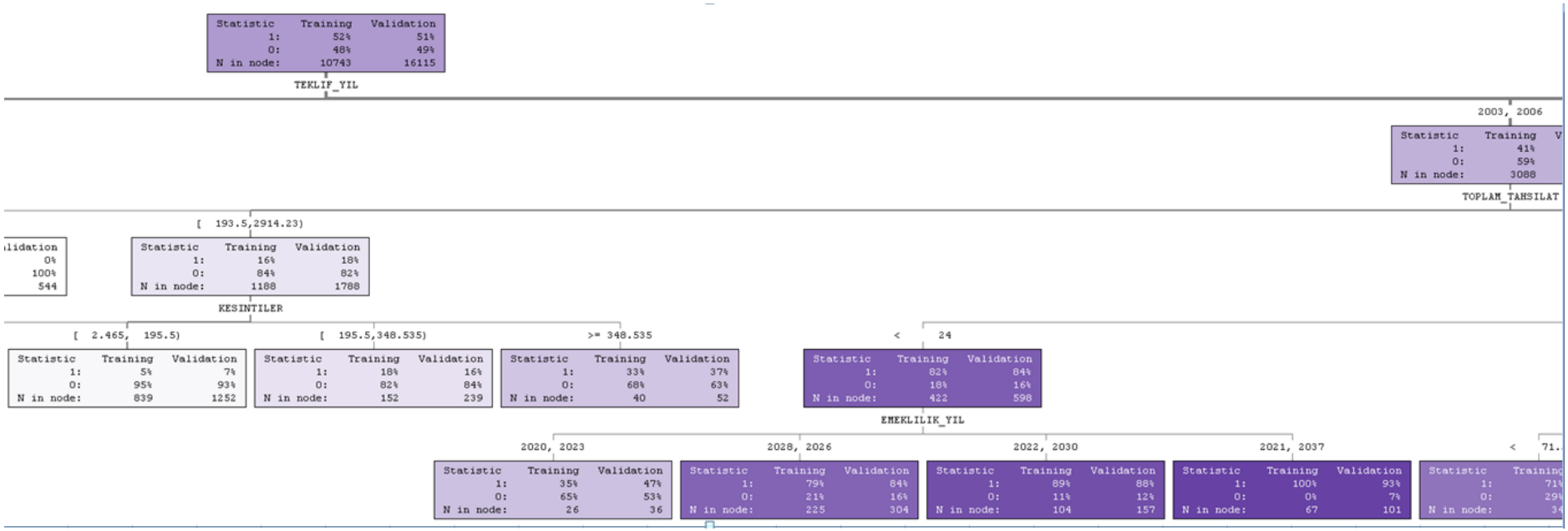


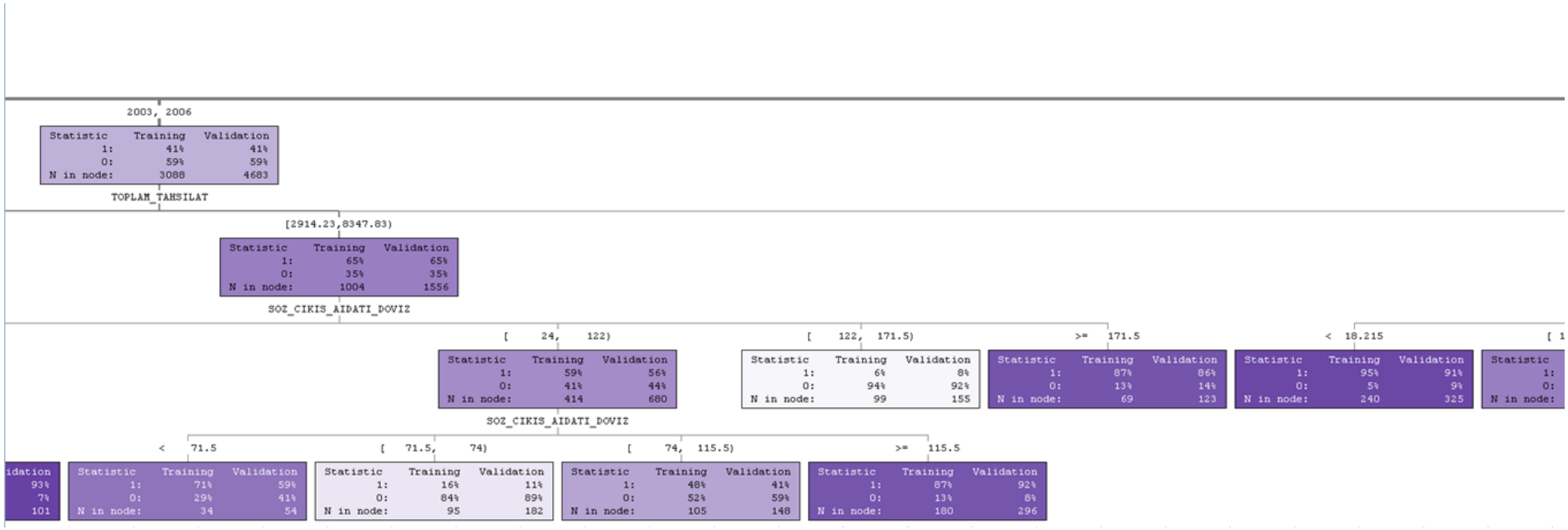
Şekil 7.6 : Entropy-4 dallanma yöntemini ile oluşturulan karar ağacı.

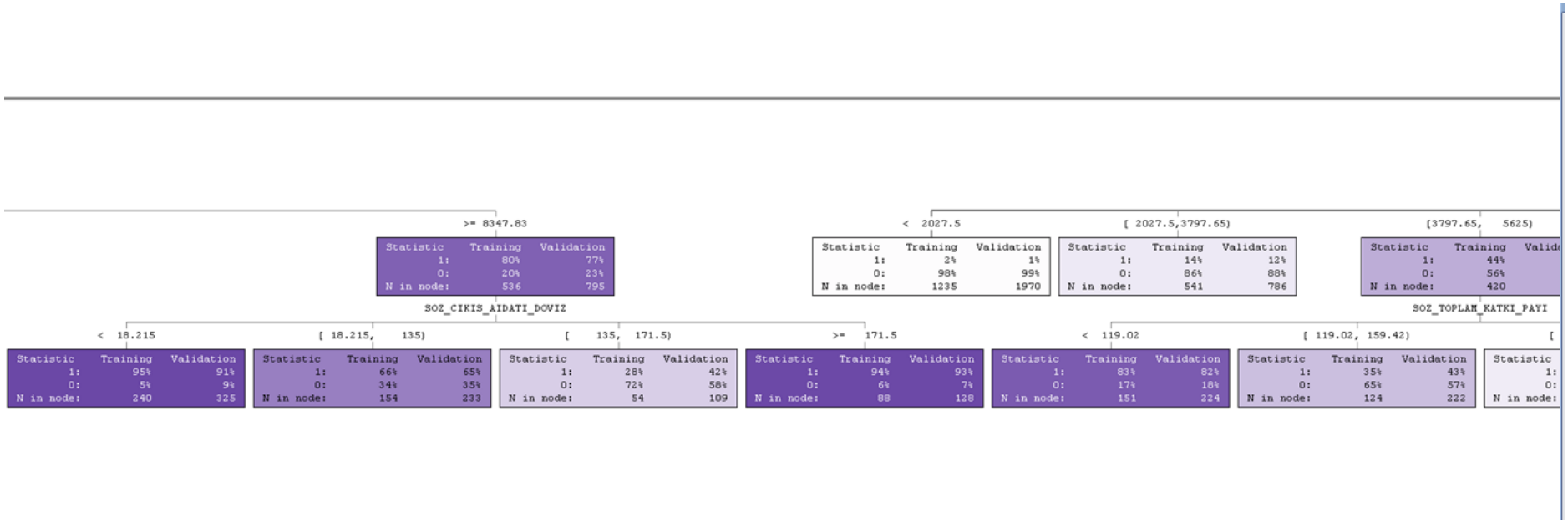

2007, 2008
Statistic Training Validation
1: 64% 63%
0: 36% 37%
N in node: 2729 4014
TOPLAM_TAHSILAT
< 328.97
Statistic Training Validation
1: 0% 1%
0: 100% 99%
N in node: 452 664
[328.97,1280.98)
Statistic Training Validation
1: 60% 59%
0: 40% 41%
N in node: 775 1175
[1280.98,2262.95)
Statistic Training Validation
1: 80% 80%
0: 20% 20%
N in node: 631 955
SOZ_GIRIS_AIDATI_ODEMESI
SOZ_CIKIS_AIDATI_DOVIZ
< 6
Statistic Training Validation
1: 76% 77%
0: 24% 23%
N in node: 447 657
[6, 108.03)
Statistic Training Validation
1: 7% 10%
0: 93% 90%
N in node: 160 291
[108.03,141.895)
Statistic Training Validation
1: 44% 23%
0: 56% 77%
N in node: 32 48
>= 141.895
Statistic Training Validation
1: 76% 82%
0: 24% 18%
N in node: 136 179
< 117
Statistic Training Validation
1: 86% 90%
0: 14% 10%
N in node: 181 267
[117, 123.5)
Statistic Training Validation
1: 3% 5%
0: 97% 95%
N in node: 30 62
[123.5,

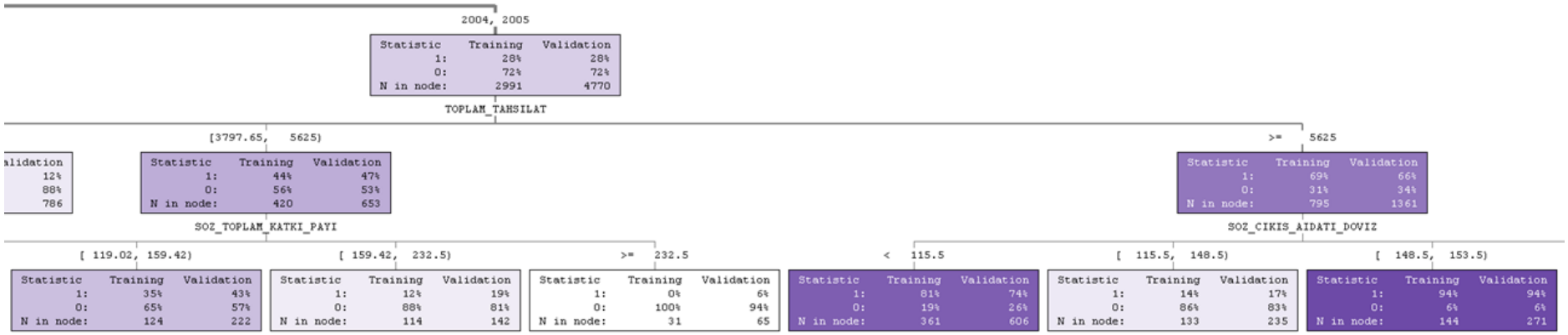

2007, 2008
Statistic Training Validation
1: 64% 63%
0: 36% 37%
N in node: 2729 4014
TOPLAM_TAHSILAT
[1280.98,2262.95)
Statistic Training Validation
1: 80% 80%
0: 20% 20%
N in node: 631 955
>= 2262.95
Statistic Training
1: 87%
0: 13%
N in node: 871
SOZ_CIKIS_AIDATI_DOVIZ
< 117
[117, 123.5)
[123.5, 158.25)
>= 158.25
< 137.25
[137.25, 166.5)
Statistic Training Validation
1: 86% 90%
0: 14% 10%
N in node: 181 267
Statistic Training Validation
1: 3% 5%
0: 97% 95%
N in node: 30 62
Statistic Training Validation
1: 90% 89%
0: 10% 11%
N in node: 349 525
Statistic Training Validation
1: 51% 54%
0: 49% 46%
N in node: 71 101
Statistic Training Validation
1: 92% 91%
0: 8% 9%
N in node: 365 535
Statistic Training Validation
1: 89% 90%
0: 11% 10%
N in node: 326 435


Statistic Training Validation
1: 52% 51%
0: 48% 49%
N in node: 10743 16115
TEKLIF_YIL
>= 2262.95
Statistic Training Validation
1: 87% 89%
0: 13% 11%
N in node: 871 1220
SOZ_CIKIS_AIDATI_DOVIZ
[137.25, 166.5)
Statistic Training Validation
1: 89% 90%
0: 11% 10%
N in node: 326 435
[166.5, 191.5)
Statistic Training Validation
1: 0% 11%
0: 100% 89%
N in node: 35 38
>= 191.5
Statistic Training Validation
1: 92% 95%
0: 8% 5%
N in node: 145 212
< 193.5
Statistic Training Validation
1: 0% 0%
0: 100% 100%
N in node: 360 544
[193.5,2914.23)
Statistic Training Validation
1: 16% 18%
0: 84% 82%
N in node: 1188 1788
KESINTILER
< 2.465
Statistic Training Validation
1: 71% 71%
0: 29% 29%
N in node: 157 245
[2.465, 195.5)
Statistic Training Validation
1: 5% 7%
0: 95% 93%
N in node: 839 1252
[195.5,348.535)
Statistic Training Validation
1: 18% 16%
0: 82% 84%


Statistic Training Validation
1: 52% 51%
0: 48% 49%
N in node: 10743 16115
TEKLIF_YIL
2003, 2006
Statistic Training V
1: 41%
0: 59%
N in node: 3088
TOPLAM_TAHSILAT
[193.5,2914.23)
Statistic Training Validation
1: 16% 18%
0: 84% 82%
N in node: 1188 1788
KESINTILER
[2.465, 195.5)
Statistic Training Validation
1: 5% 7%
0: 95% 93%
N in node: 839 1252
[195.5,348.535)
Statistic Training Validation
1: 18% 16%
0: 82% 84%
N in node: 152 239
>= 348.535
Statistic Training Validation
1: 33% 37%
0: 68% 63%
N in node: 40 52
< 24
Statistic Training Validation
1: 82% 84%
0: 18% 16%
N in node: 422 598
EMEKLILIK_YIL
2020, 2023
Statistic Training Validation
1: 35% 47%
0: 65% 53%
N in node: 26 36
2028, 2026
Statistic Training Validation
1: 79% 84%
0: 21% 16%
N in node: 225 304
2022, 2030
Statistic Training Validation
1: 89% 88%
0: 11% 12%
N in node: 104 157
2021, 2037
Statistic Training Validation
1: 100% 93%
0: 0% 7%
N in node: 67 101
< 71.
Statistic Training
1: 71%
0: 29%
N in node: 34


2003, 2006
Statistic Training Validation
1: 41% 41%
0: 59% 59%
N in node: 3088 4683
TOPLAM_TAHSILAT
[2914.23,8347.83)
Statistic Training Validation
1: 65% 65%
0: 35% 35%
N in node: 1004 1556
SOZ_CIKIS_AIDATI_DOVIZ
[24, 122)
Statistic Training Validation
1: 59% 56%
0: 41% 44%
N in node: 414 680
[122, 171.5)
Statistic Training Validation
1: 6% 8%
0: 94% 92%
N in node: 99 155
>= 171.5
Statistic Training Validation
1: 87% 86%
0: 13% 14%
N in node: 69 123
< 18.215
Statistic Training Validation
1: 95% 91%
0: 5% 9%
N in node: 240 325
SOZ_CIKIS_AIDATI_DOVIZ
< 71.5
Statistic Training Validation
1: 71% 59%
0: 29% 41%
N in node: 34 54
[71.5, 74)
Statistic Training Validation
1: 16% 11%
0: 84% 89%
N in node: 95 182
[74, 115.5)
Statistic Training Validation
1: 48% 41%
0: 52% 59%
N in node: 105 148
>= 115.5
Statistic Training Validation
1: 87% 92%
0: 13% 8%
N in node: 180 296


>= 8347.83
Statistic Training Validation
1: 80% 77%
0: 20% 23%
N in node: 536 795
SOZ_CIKIS_AIDATI_DOVIZ
< 18.215
Statistic Training Validation
1: 95% 91%
0: 5% 9%
N in node: 240 325
[18.215, 135)
Statistic Training Validation
1: 66% 65%
0: 34% 35%
N in node: 154 233
[135, 171.5)
Statistic Training Validation
1: 28% 42%
0: 72% 58%
N in node: 54 109
>= 171.5
Statistic Training Validation
1: 94% 93%
0: 6% 7%
N in node: 88 128
< 2027.5
Statistic Training Validation
1: 2% 1%
0: 98% 99%
N in node: 1235 1970
[2027.5,3797.65)
Statistic Training Validation
1: 14% 12%
0: 86% 88%
N in node: 541 786
[3797.65, 5625)
Statistic Training
1: 44%
0: 56%
N in node: 420
SOZ_TOPLAM_KATKI_PAYI
< 119.02
Statistic Training Validation
1: 83% 82%
0: 17% 18%
N in node: 151 224
[119.02, 159.42)
Statistic Training Validation
1: 35% 43%
0: 65% 57%
N in node: 124 222


2004, 2005
Statistic Training Validation
1: 28% 28%
0: 72% 72%
N in node: 2991 4770
TOPLAM_TAHSILAT
[3797.65, 5625)
Statistic Training Validation
1: 44% 47%
0: 56% 53%
N in node: 420 653
>= 5625
Statistic Training Validation
1: 69% 66%
0: 31% 34%
N in node: 795 1361
SOZ_TOPLAM_KATKI_PAYI
SOZ_CIKIS_AIDATI_DOVIZ
[119.02, 159.42)
Statistic Training Validation
1: 35% 43%
0: 65% 57%
N in node: 124 222
[159.42, 232.5)
Statistic Training Validation
1: 12% 19%
0: 88% 81%
N in node: 114 142
>= 232.5
Statistic Training Validation
1: 0% 6%
0: 100% 94%
N in node: 31 65
< 115.5
Statistic Training Validation
1: 81% 74%
0: 19% 26%
N in node: 361 606
[115.5, 148.5)
Statistic Training Validation
1: 14% 17%
0: 86% 83%
N in node: 133 235
[148.5, 153.5)
Statistic Training Validation
1: 94% 94%
0: 6% 6%
N in node: 144 271

8. SONUÇ VE TARTIŞMA

Veri hacmi arttıkça işlenmesi ve istenilen bilgilere ulaşmak oldukça güçleşecektir. Veri madenciliği bu kısımda devreye girer; büyük miktarda veri içinden gelecek ile ilgili tahmin yapmamızı sağlayacak bağıntı ve örüntülerin bulunmasını sağlar.

Veri madenciliğinin en yaygın uygulama alanlarından biri pazarlamadır. Şirketler, veri madenciliğini kampanya yönetimi, müşteri segmentasyonu, müşteri edinme, mevcut müşterileri elde tutma, müşteri yaşam ömrü analizi, müşteri ilişkileri yönetimi, pazar sepeti analizleri ve kredi derecelendirme gibi pek çok alanda kullanmaktadır.

Bu tez çalışmasında, veri madenciliği teknikleri kullanılarak bireysel emeklilik sahibi müşterilerin sigorta şirketinden ayrılarak rakip şirketlerin müşterisi olmayı tercih eden davranışlarını ortaya koyan modeller geliştirilmiştir. Verinin temizlenmesi ve istenilen formata dönüştürülmesi işlemlerinin ardından; SAS Enterprise Miner yardımı ile 3 farklı karar ağacı algoritması ile müşterinin ayrılarak rakip şirketi tercih etme (müşteri bağlılık) davranışı modellenmiştir.

Yapılan karar ağacı uygulamasında denenen kurallardan sonra şu sonuçlar çıkarılmıştır:

- Eğitim ve test kümesinin büyütülmesi, verinin özelliklerine bağlı olarak çıkan sınıflandırmanın ne kadar verimli olduğunu etkiler.
- Chi-Square, entropy ve gini index dallanma yöntemleri ile oluşturulan karar ağaçları içinde en yüksek doğruluk oranını entropy tabanlı 21 yapraklı karar ağacı vermiştir.
- Entropy hesaplama yöntemini kullanarak 3 farklı dallanma kriteri ile oluşturulan ağaçların testi yapılmıştır. Bu analiz sonucunda, en yüksek doğruluk oranını dallanma kriteri 4 olan 46 yapraklı ağacın verdiği görülmüştür.

- Entropy-4 tabanlı ağacın detayı incelendiğinde,
  TOPLAM_TAHSILAT < 10.94 ve TEKLIF_YIL =2009 olan,
  TOPLAM_TAHSILAT < 193.5 ve TEKLIF_YIL 2003 ya da 2006 olan,
  166.5<=SOZ_CIKIS_AIDATI_DOVIZ<191.5 ve
  2262.96 <= TOPLAM_TAHSILAT ve TEKLIF_YIL 2007 ya da 2008 olan,
  232.5<= SOZ_TOPLAM_KATKI_PAYI ve
  797.65 <= TOPLAM_TAHSILAT <5625 ve TEKLIF_YIL 2004 ya da 2005 kritelerine uyan müşterilerin bağlılıkların en az olduğu saptanmıştır.

Bu çalışma aşağıda belirtilen işlemlerin yapılması ile genişletilebilir:

- Karar ağaçları yerine, farklı bir veri madenciliği yöntemi kullanılarak ortaya çıkan sonuçlar karşılaştırılabilir.
- Hangi modelin böyle bir veri kümesi için uygun olup olmayacağı tartışılabilir.
- Ara verme durumundaki, müşteriler ayrıca incelenerek bunların ayrılmaya olan eğilimleri detaylandırılabilir. Böylelikle henüz müşteri kaybedilmeden gereken aksiyon alınabilir.
- Ayrılan müşteriler üzerinde farklı bir veri madenciliği yöntemi olan Support Vector Machine analizi yapılarak, bu müşterilere ait ortak özellikler belirlenebilir. Elde edilecek sonuçlar, karar ağacı algoritması ile karşılaştırılarak ortaya konmuş modelin doğruluğu incelenebilir.

KAYNAKLAR

Akalın, M., 2004: CRM ve Sigortacılık Sektöründe Bir Uygulama, Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

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

Akbulut S., Aydoğan E., Gencer C., 2007: Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Bölünmesi, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Bayram, E., 2001: Customer segmentation and churn modeling in wireless communication, Yüksek Lisans Tezi, Boğaziçi Üniveristesi Fen Bilimleri Enstitüsü, İstanbul, 1-5.

Chen, M., Chiu, A., Chang, H., 2005: Mining changes in customer behavior in retail marketing, Expert Systems with Applications, 28 (4): 773-781.

Çoban, S., 2004: Toplam Kalite Yönetimi Perspektifinde İçsel Pazarlama Anlayışı, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Sayı: 22.

Fayyad, U., 1997: Data Ming and Knowledge Discovery in Databases: Implications for Scientific Databases, Ninth Internation Conferance on Scientific and Statistical Database Management.

Gazi, V. E., 2007: Veri Madenciliğinde Duyarlılık, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Griffin, J., 1997: Customer Loyalty-How To Earn It, How to Keep It, Jossey Boss Publishers, San Fransisco.

Günaydın, Y., 2005: Konaklama İşletmelerinde müşteri ilişkileri yönetimi, memnuniyeti ve sadakati, Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü, İzmir.

Gündaş, K., 2003: İşletmelerde Müşteri İlişkileri Yönetimi ve Bir Uygulama, Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü, İzmir.

Güzel, M., 2001: Müşteri İlişkileri Yönetimi (MİY) ve Türkiye Değerlendirmesi, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Hançer, M., 2003: Konaklama Endüstrisinde Müsteri Sadakati: Anlam, Önem, Etki ve Sonuçlar, Seyahat ve Turizm Araştırmaları Dergisi, Cilt: 3, No:1-2.

Hsieh N., 2004: An integrated data mining and behavioral scoring model for analyzing bank customers, Expert Systems with Applications, 27 (4): 623-633.

Hwang H., Jung T., Suh E., 2004: An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry, Expert System with Applications, 26: 181-188.

Jonker J., Piersma N., Poel D., 2003: Joint optimization of customer segmentation and marketing policy to maximize long term profitability, Working Paper of Faculty of Economics and Business Administration Ghent University, 3: 214-239.

Kalder Grubu, 2000: Müşteri Memnuniyeti Yönetimi, Kalder Yayınları, No:31, Rota Yayınevi, İstanbul.

Kayaalp K., 2007: Asenkron Motorlarda Veri Madenciliği ile Hata Tespiti, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Isparta.

Kırım, A., 2001: Strateji ve Birebir Pazarlama CRM, Sistem Yayıncılık, İstanbul.

Kocabaş, Ş., 1991: A Review of Learning. The Knowledge Enginering Review, Vol.6 No:3, 195-222.

Kotler, P., 2000: Pazarlama Yönetimi (Çev.N. Muallimoğlu), Millenium Baskı, Beta Yayıncılık, İstanbul.

Larose, D. T., 2005: Discovering Knowledge in Data: An Introduction to Data Mining, John and Wiley Sons Incorporated, USA.

Law, M., Lau, T., Wong, Y. H., 2003: From Customer Relationship Management to Customer-Managed Relationship: Unraveling The Paradox With a Co-Creative Perspective, Marketing Intelligence&Planning, Vol 21, No 1.

Liu D., Shih Y., 2005: Integrating AHP and data mining for product recommendation based on customer lifetime value, Information & Management, 42 (3): 387-400.

Madran, C., Canbolat, C., 2006: Müşteri İlişkileri Yönetimi ile Müsteri Sadakati Arasındaki İlişki, 2. Ulusal Pazarlama Kongresi Bildiri Kitabı, Dokuz Eylül Üniversitesi Deniz İşletmeciliği ve Yönetimi Yüksekokulu, 1-2-3 Kasım 2006, İzmir.

Odabaşı, Y., 2005: Müşteri İlişkileri Yönetimi, Sistem Yayıncılık, 5.Basım, İstanbul.

Odabaşı, Y., Gülfidan, B. 2003: Tüketici Davranışı, Mediacat Yayınları, 3.Basım, İstanbul.

Oliver, R.L., 1999: Whence Customer Loyalty?, Journal of Marketing, Vol:63, Special Issue.

Reicheld, F.F., Sasser W.E., 1990: Zero Defections: Quality Comes To Services, Harward Business Review, September-October.

Rosenfield, R.J., 1999: Whatever Happened to Relationship Marketing?Nine Big Mistakes, Direct Marketing, Vol 62, Issue 2.

Ryals, L., Knox, S., 2001: Cross-Functional Issues in the Implementation of Relationship Marketing Through Customer Relationship Management, European Management Journal, Vol. 19, No. 5.

Sever H., Oğuz B., 2002: Veri Tabanlarında Bilgi Keşifine Formel Bir Yaklaşım Kısım I: Eşleştirme Sorgulamaları ve Algoritmalar, Bilgi Dünyası, 3(2), S: 173-204, Ankara.

Soyaslan, M., 2006: Müşteri İlişkileri Yönetimi ve Türkiye'deki Oteller Üzerine Bir Arastırma, Yüksek Lisans Tezi, Uludag Üniversitesi Sosyal Bilimler Enstitüsü, Bursa.

Taşkın, E., 2000: Müşteri İlişkileri Egitimi, Papatya Yayıncılık, 2.basım, İstanbul.

Tek, B.Ö., 1999: Pazarlama İlkeleri, Global Yönetimsel Yaklaşım, Türkiye Uygulamaları, Beta Yayıncılık, İstanbul.

Timur, M.N., Sarıyer, N., 2004: Kayseri'deki Otomobil Bayilerinde Müşteri Tatmin Aracı Olarak Sikayet Toplama Yöntemlerine İlişkin Bir Uygulama, Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Sayı : 17.

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

Tütüncü, Ö., Dogan, Ö., 2003: Müsteri Tatmini Kapsamında Ögrenci Memnuniyetinin Ölçülmesi ve Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Uygulaması, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Cilt 5, Sayı 4.

Vahaplar A., İnceoğlu M. M., 2001: Veri Madenciliği ve Elektronik Ticaret, VII. Tükiye' de Internet Konferansı.

Yurtsever, U., 2002: Veri Madenciliği ve Uygulaması, Yüksek Lisans Tezi, Sakarya Üniversitesi Sosyal Bilimler Enstitüsü, Sakarya.

Url-1 <www.tech-strategy.org/Manuel/CRM-Nedir-Ne-Degildir.pdf>, alındığı tarih 17.12.2009.

ÖZGEÇMİŞ

Ad Soyad	**:** Yıldız Koçtürk
Doğum Yeri ve Tarihi	**:** Almanya, 31/01/1980
Adres	**:** Yıldıztabya Cad. Tunalı Apt. No:10/4 GaziOsmanPaşa/İSTANBUL
Lisans Üniversite	**:** Yıldız Teknik Üniversitesi