

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/286048668>

# Veri Madenciliği: Kümeleme ve Sınıflama Algoritmaları

Book · October 2011

CITATIONS

10

READS

19,031

1 author:



Ulas Akkucuk

Bogazici University

55 PUBLICATIONS 305 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Call for Chapters: Handbook of Research on Sustainable Supply Chain Management for the Global Economy [View project](#)



Call for Chapters: Recent Developments on Creating Sustainable Value in the Global Economy [View project](#)

---

# VERİ MADENCİLİĞİ

## KÜMELEME VE SINIFLAMA

### ALGORİTMALARI

---

**Yard. Doç. Dr. ULAŞ AKKÜÇÜK**

Ψ  
YALIN YAYINCILIK  
İstanbul - 2011

**VERİ MADENCİLİĞİ: KÜMELEME VE SINIFLAMA ALGORİTMALARI****Yard. Doç. Dr. ULAŞ AKKÜÇÜK**

Boğaziçi Üniversitesi  
İktisadi ve İdari Bilimler Fakültesi  
İşletme Bölümü

YALIN YAYINCILIK  
Ordu Caddesi Özbek Çarşısı No. 25 / 41  
(İstanbul Üniversitesi Rektörlüğü'nün Karşısında)  
34452 - Beyazıt - İSTANBUL  
Tel: (0212) 518 43 63 - (0212) 546 97 54  
Fax: (0212) 518 43 63  
www.yalinkitap.com e-mail: yalinkitap@yahoo.com  
Yayıncı Sertifika No: 16116  
Baskı: Deniz Ofset Matbaa  
Gümüşsuyu Cad. Topkapı Center B Blok No:403  
Topkapı - Zeytinburnu - İstanbul  
Sertifika No: 13901

**ISBN: 978-605-4539-03-1**

Birinci Basım: 2011

© Bu kitabın 5846 sayılı Yasa ile korunan tüm hakları yazarına aittir. İzinsiz olarak herhangi bir şekilde çoğaltılması, basılması, kaynak gösterilmeksizin alıntılar yapılması yasaktır ve anılan Yasa gereği kovuşturulur.

© All rights reserved

No part of this book may be reproduced or stored in a retrieval system, or transmitted in any form or by any means mechanical, electronic, photocopy, magnetic, tape or otherwise, without permission in writing from the writer.

## YAZAR HAKKINDA

**Ulaş AKKÜÇÜK**

Ulaş Akkçük 1975 yılında İstanbul'da doğmuştur. 1993 yılında Robert Kolej, ardından 1997 yılında Bilkent Üniversitesi Endüstri Mühendisliği bölümünü bitirdikten sonra Boğaziçi Üniversitesi'nde MBA yapmış, 2004 yılında ise ABD New Jersey'de bulunan Rutgers University'den İşletme alanında Doktorasını almıştır. 2005 yılından beri Boğaziçi Üniversitesi İktisadi İdari Bilimler Fakültesi İşletme Bölümü'nde Sayısal Yöntemler Anabilim Dalı'nda Yardımcı Doçent olarak görev yapmaktadır. İstatistik, Kalite Yönetimi, Üretim Yönetimi, Uygulamalı Veri Madenciliği, Pazarlama Araştırmalarında Özel Konular, Araştırma Yöntemleri gibi lisans, yüksek lisans ve doktora seviyesinde çeşitli dersler vermiştir.

Dr. Akkçük'ün araştırma ve uzmanlık alanları arasında, pazarlama araştırmalarında sıkça kullanılan sayısal yöntemler (özellikle görselleştirme/boyut azaltma teknikleri), optimizasyon teknikleri ve bu tekniklerin görselleştirme uygulamalarında kullanımı, kalite yönetimi, veri madenciliği, müşteri ilişkileri yönetimi (CRM) ve tedarik zinciri yönetimi sayılabilir. Yayınlanmış akademik çalışmalarında çok değişkenli analiz teknikleri arasında yer alan görselleştirme tekniklerinin hem metodolojik boyutlarıyla ilgilenmiş hem de Sosyal Bilimler alanında uygulamalarını yapmıştır. INFORMS ve Classification Society gibi çeşitli uluslararası bilimsel kuruluşlara üyedir. 2009 yılında International Federation of Classification Societies tarafından verilen Chikio Hayashi ödülüne sahip olmuştur. Çalışmaları Journal of Classification, Boğaziçi Journal, European Journal of Social Sciences, European Journal of Economics Finance and Administrative Sciences gibi bilimsel dergilerde yayınlanmıştır. Ayrıca Joint Statistical Meetings 2006, International Federation of Classification Societies 2009, 2011 ve Classification Society 2006, 2007 yıllık toplantılarında tebliğler sunmuştur. Boğaziçi Journal ve Journal of Multivariate Analysis gibi bilimsel dergilerde hakemlik çalışmaları yapmıştır. Yayınlanmış "Görselleştirme Teknikleri: Pazarlama ve Patent Analizi Örnek Uygulamaları İle" isimli bir de bilimsel kitabı bulunmaktadır.



*Haziran 2011'de aramızdan ayrılan  
doktora tez hocam J. Douglas Carroll'a  
ithaf ediyorum.*



## İÇİNDEKİLER

ŞEKİLLER LİSTESİ .....	11
TABLolar LİSTESİ.....	13
ÖNSÖZ .....	15

## BİRİNCİ BÖLÜM

### VERİ MADENCİLİĞİ NEDİR?

1.1. Veri Madenciliği Tanımları .....	17
1.2. CRISP-DM Modeli .....	17

## İKİNCİ BÖLÜM

### VERİ ÖNİŞLEMESİ

2.1. Yanlış Tasnif Edilmiş ya da Mantıksız Değerler .....	21
2.2. Eksik Değerler .....	23
2.3. Veri Dönüşümleri.....	24
2.4. Aykırı Değerlerin Tespiti .....	25

## ÜÇÜNCÜ BÖLÜM

### AÇINSAYICI VERİ ANALİZİ

3.1. Kutu Bıyık Diyagramı .....	29
3.2. Histogram .....	31
3.3. Pareto Diyagramı .....	32
3.4. Serpilme Diyagramı .....	33
3.5. Çapraz Tablo.....	35
3.6. Diğer Görselleştirme Teknikleri .....	36



## DÖRDÜNCÜ BÖLÜM

## İSTATİSTİKİ TAHMİN TEKNİKLERİ

4.1. Tek Değişkenli İstatistikler.....	37
4.2. Güven Aralıkları .....	42
4.3. Basit Regresyon .....	44
4.4. Çoklu Regresyon .....	47

## BEŞİNCİ BÖLÜM

## k-MEANS ALGORİTMASI

5.1. Uzaklık Fonksiyonları .....	50
5.2. Değişken Ölçekleri ve Nitel Değişkenler .....	52
5.3. Sıradüzensel Yöntemler .....	56
5.4. K-Means Algoritması.....	57
5.5. Küme Sayısının ( $k$ ) Belirlenmesi .....	60

## ALTINCI BÖLÜM

## K EN YAKIN KOMŞU ALGORİTMASI

6.1. En Yakın Komşu Algoritma .....	65
6.2. Ağırlıklandırma Fonksiyonları .....	68
6.3. $K$ seçimi .....	70

## YEDİNCİ BÖLÜM

## KARAR AĞAÇLARI

7.1. Bir Örnek .....	71
7.2. CART Algoritması.....	74
7.3. C 4.5 Algoritması.....	80
7.4. C 5.0 Algoritması.....	82

## SEKİZİNCİ BÖLÜM

## MODEL KARŞILAŞTIRMA YÖNTEMLERİ

8.1. Tahmin ve Kestirim İçin Model Karşılaştırma Yöntemleri .....	83
8.2. Sınıflama Uygulamaları İçin Model Karşılaştırma Yöntemleri .....	84
8.3. Hatalı Sınıflama Türleri.....	85
8.4. Fayda/Maliyet Hesabı ile Karar Verme .....	88
8.5. Kazanım ve Birikimli Kazanım Grafikleri.....	90

## DOKUZUNCU BÖLÜM

## VAKA ANALİZİ: ARAMA A.Ş.

9.1. Açımsayıcı Veri Analizi .....	95
9.2. Diskriminant Analizi .....	103
9.3. Lojistik regresyon .....	107
9.4. Model Karşılaştırması .....	108

## ONUNCU BÖLÜM

## VAKA ANALİZİ: YAZ OKULU

10.1. Açımsayıcı Veri Analizi .....	114
10.2. Çoklu Regresyon.....	114
10.3. Diskriminant Analizi ile “AA” Notu Tahmini.....	117
10.4. Karar Ağaçları C5.0 ile .....	118

KAYNAKÇA A: MAKALELER, KİTAPLAR, BİLDİRİLER.....	125
--	-----

KAYNAKÇA B: AĞ ADRESLERİ .....	127
--------------------------------	-----

EK: ONUNCU BÖLÜM VAKA ANALİZİ İLE İLGİLİ VERİ.....	128
--	-----



## ŞEKİLLER LİSTESİ

Şekil 1.1. CRISP-DM aşamaları (URL13) .....	19
Şekil 3.1. İMKB100 endeksi Ocak 2000-Ekim 2008 arası kutu bıyık diyagramı.....	30
Şekil 3.2. İMKB100 endeksi ve Dow Jones yan yana kutu bıyık diyagramı .	31
Şekil 3.3. Bir istatistik dersinde 35 öğrencinin sınav notları.....	32
Şekil 3.4. Pareto diyagram örneği .....	33
Şekil 3.5. Yatay eksen Dow Jones, dikey eksen Bovespa.....	34
Şekil 3.6. Yatay eksen Bovespa, dikey eksen Jakarta .....	34
Şekil 3.7. Yatay eksen Jakarta, dikey eksen Dow Jones .....	35
Şekil 4.1. 35 sınav sonucu yüzdelik ve Z değerlerinin serpilme diyagramı .	42
Şekil 4.2. Yatay eksen ağırlık dikey eksen benzin tüketimi serpilme diyagramı .....	45
Şekil 4.3. Yatay eksen ağırlık dikey eksen benzin tüketimi serpilme diyagramı üzerinde regresyon eğrisi çizilmiş hali. ....	47
Şekil 5.1. WEKA'dan alınmış golf veri seti.....	50
Şekil 5.2. Kümeleme çıktısı sıradüzensel yapıya bir örnek.....	56
Şekil 5.3. İki boyutlu bir yapıda belirlenmiş beş merkez .....	58
Şekil 5.4. İki boyutlu bir yapıda belirlenmiş ilk beş küme .....	59
Şekil 5.5. İki boyutlu bir yapıda homojen dağılmış nesneler .....	60
Şekil 5.6. Pseudo-F ve r-kare değerleri (Akküçük, 2011b) .....	63
Şekil 6.1. Na ve K değerleri, reçeteye yazılan ilaç (A, B, C, X ve Y ile gösterilmiştir) .....	66
Şekil 6.2. Gözlem 1 yakınlaşmış .....	67
Şekil 6.3. Gözlem 2 yakınlaşmış .....	68
Şekil 7.1. Golf veri seti için karar ağacı (C4.5).....	74
Şekil 8.1. Örnek birikimli kazanım grafiği, dikey eksen olumlu cevap/20.000, diyagonal çizgi rastgele seçimde olacak eğriyi gösterir.....	91

Şekil 8.2. Örnek kazanım grafiği, dikey eksen kazanım, yatay çizgi “1” değerini ifade eder. ....	92
Şekil 9.1. İlk 15 gözlem için örnek veri.....	95
Şekil 9.2. Akşam konuşma sayısı (evecalls), dakikası (evemins) ve ücretinin (evecharge) aralarında serpilme diyagramları.....	97
Şekil 9.3. Gündüz konuşma dakikası histogram .....	100
Şekil 9.4. ROC Diskriminant Analizi (eğri altındaki alan 0,826) .....	109
Şekil 9.5. ROC Lojistik Regresyon Analizi (eğri altındaki alan 0,823) .....	110
Şekil 9.6. İki farklı analiz için lift (kazanım) grafiği .....	111
Şekil 9.7. İki farklı analiz için birikimli kazanım (ROC) grafiği .....	111
Şekil 10.1. See5 Çıktısı, 59 hata 42 doğru tahmin .....	119
Şekil 10.2. See5 Çıktısı, karar ağacı, global budama seçilmeden.....	120
Şekil 10.3. See5 Çıktısı, hata oranları.....	121
Şekil 10.4. See5 Çıktısı, karar ağacı, budama güven seviyesi %75 .....	122
Şekil 10.5. See5 Çıktısı, karar ağacı, hata oranları, budama güven seviyesi %75 .....	123

**TABLolar LİSTESİ**

Tablo 2.1. 2005 Capital 500 listesi “İl” değişkenine göre frekans.....	22
Tablo 2.2. 2005 Capital 500 listesi “Sektör” değişkenine göre frekans .....	23
Tablo 2.3. İMKB100 endeksi Ocak 2000-Ekim 2008 arası şiddetli aykırı değerler ve tarihleri.....	27
Tablo 3.1. 2005 Capital 500 listesi “Sektör” ve “İl” arasında çapraz tablo..	36
Tablo 4.1. DJI-BVSPA-JKSE 2010 haftalık getiri özet istatistikleri.....	39
Tablo 4.2. 45 sınav sonucu yüzdelik ve Z değerleri.....	41
Tablo 4.3. Ağırlık ve yakıt tüketimi basit regresyon sonuçları .....	46
Tablo 4.4. Yakıt tüketimi çoklu regresyon sonuçları.....	48
Tablo 5.1. Banka müşteri verisi.....	53
Tablo 5.2. Banka müşteri verisi yaş ve gelir kullanılarak hesaplanan uzaklık ..	53
Tablo 5.3. Banka müşteri verisi, min-max dönüşümü .....	54
Tablo 5.4. Banka müşteri verisi Z-skor dönüşümü .....	54
Tablo 5.5. Banka müşteri verisi yaş ve gelir kullanılarak hesaplanan uzaklık, min-max dönüşümünden sonra.....	54
Tablo 5.6. Banka müşteri verisi yaş ve gelir kullanılarak hesaplanan uzaklık, Z-skor dönüşümünden sonra .....	55
Tablo 5.7. Banka müşteri verisi yaş ve gelir kullanılarak hesaplanan uzaklık, min-max dönüşümünden sonra ve cinsiyet eklenerek ..	56
Tablo 7.1. Golf veri seti .....	72
Tablo 7.2. Hava değişkeninin değerleri için golf a uygun olan gün sayıları .	73
Tablo 7.3. Rüzgar değişkeninin değerleri için golf a uygun olan gün sayıları...	73
Tablo 7.4. Olası bölünme adayları .....	77
Tablo 7.5. İlk aşamada hesaplanan $\Phi$ değerleri.....	78
Tablo 7.6. İkinci yinelemede olası bölünme adayları .....	79
Tablo 7.7. İkinci aşamada hesaplanan $\Phi$ değerleri .....	79
Tablo 8.1. Örnek hata matrisi .....	86
Tablo 8.2. Örnek hata matrisi .....	86

Tablo 8.3. Örnek kazanım grafiği verisi .....	91
Tablo 9.1. Kayıp müşteri özet bilgileri .....	98
Tablo 9.2. Uluslararası plan özet bilgileri.....	98
Tablo 9.3. Telesekreter planı özet bilgileri.....	98
Tablo 9.4. Çapraz tablo kayıp müşteri ile uluslararası plan.....	99
Tablo 9.5. Çapraz tablo kayıp müşteri ile telesekreter planı .....	99
Tablo 9.6. Gündüz ortalama çağrı sayısı, kayıp ve kayıp olmayan müşterilere göre .....	101
Tablo 9.7. Gece ortalama çağrı sayısı, kayıp ve kayıp olmayan müşterilere göre .....	101
Tablo 9.8. Müşteri hizmetleri arama sayısı, kayıp ve kayıp olmayan müşterilere göre .....	102
Tablo 9.9. Müşteri hizmetleri arama sayısına göre ortalama kayıp müşteri (kayıp müşteri 1-0 olarak yeniden kodlanmıştır) .....	102
Tablo 9.10. Diskriminant fonksiyon katsayıları.....	104
Tablo 9.11. Diskriminant fonksiyon katsayıları standartlaştırılmış halde ..	105
Tablo 9.12. Diskriminant analizi hata matrisi .....	106
Tablo 9.13. Lojistik regresyon analizi hata matrisi.....	107
Tablo 9.14. Lojistik regresyon analizi hata matrisi (eşik olasılık 0,3).....	108
Tablo 10.1. Değişkenler arasındaki korelasyon .....	115
Tablo 10.2. Çoklu Regresyon .....	116
Tablo 10.3. Diskriminant Fonksiyonu .....	118

## ÖNSÖZ

Özel ve kamu kuruluşlarının topladığı veri miktarı arttıkça bu veriden anlamlı sonuçlar çıkartma isteği çeşitli matematiksel yöntemlerin “Veri Madenciliği” adı altında toparlanmasına sebep olmuştur. Veri madenciliği, kitabın başlığına da esin kaynağı olan iki önemli ve işletmeler için faydalı amaç için kullanılabilir. Bunlardan ilki “sınıflama” olarak nitelendirilebilir. Örneğin bir banka kredi başvurularını “krediye uygun” ve “krediye uygun değil” olarak sınıflamak isteyecektir. Diğer önemli amaç ise “kümeleme” olarak nitelenebilir ve aslında ilk amaçla ortak olarak da kullanılabilir. Örneğin aynı banka kredi başvurusu yapan müşterileri bir hedef değişken gözetmeden demografik ve diğer özelliklere göre segmentlere ayırmak isteyebilir. Sadece bir amaca hizmet eden algoritmalar olmakla birlikte hem kümeleme hem sınıflama algoritması olarak kullanılabilen algoritmalar da mevcuttur.

Bu kitabın temel amacı veri madenciliği tekniklerini temel matematiksel prensipleri ve uygulama yöntemleriyle göstermektir. Kitapta ilk iki bölümde veri madenciliğindeki temel kavramlar anlatılacak ve veri önışlemesinden bahsedilecektir. Daha sonraki iki bölümde temel açmsayıcı veri analizi tekniklerinden ve regresyon gibi genel ama veri madenciliğine de uygulanabilecek temel istatistiki tekniklerden söz edilecektir. Daha sonra kitabın başlığında da bahsedilen iki temel veri madenciliği görevi olan “sınıflama” ve “kümeleme” algoritmalarından önemli görülenler anlatılacaktır. En sonunda, ilk olarak hayali bir şirket olan ARAMA A.Ş. üzerinden kayıp müşteri tespit etmekle ilgili bir vaka analizi ardından öğrencilerin özelliklerinden dersten alacakları notu tahmin etmekle ilgili başka bir vaka okuyuculara sunulacaktır.

Yard. Doç. Dr. Ulaş Akküçük  
Eylöl 2011 - Bebek-İstanbul





## 1. BÖLÜM:

### VERİ MADENCİLİĞİ NEDİR?

#### 1.1. Veri Madenciliği Tanımları

Veri madenciliği için birçok farklı tanım bulunmaktadır. Amerikan Pazarlama Birliği (AMA) veri madenciliğini şu şekilde tanımlamıştır (URL1), “Verilerin, yeni ve potansiyel olarak yararlı bilgi bulma amaçlı analiz süreci. Bu süreç bulunması zor örüntülerin ortaya çıkarılması için matematiksel araçların kullanımını içerir.” Gartner tarafından verilen bir diğer tanım ise şöyledir (URL2), “Veri depolarında saklanan büyük miktarda veri üzerinden eleme ile anlamlı korelasyonlar, örüntüler ve trendler keşfetme süreci. Veri madenciliği, örüntü tanıma teknolojilerinin yanı sıra matematiksel ve istatistiksel teknikleri kullanmaktadır.”

Çok farklı tanımlar da yapılabilir ancak yukarıdaki iki tanım veri madenciliği sürecinin olmazsa olmaz üç ana temelini açıkça belirtmektedir:

- Büyük miktarda veri: Örneğin bir cep telefonu şirketinin 34 milyonu aşan abonesi hakkında geçmiş üç yılda tuttuğu demografik özellikleri, fatura detayları ve kullanım alışkanlıklarını içeren veri.
- Potansiyel olarak yararlı bilgi: Örneğin geçmişte başka telefon şirketine geçmiş müşterilerin özelliklerini kullanarak şu anda başka şirkette geçme eğiliminde olan müşterileri önceden belirleyip bu müşterilerin geçiş yapmasını engelleyebilme.
- Matematiksel ve istatistiksel teknikler: Aynı örneğe devam ederek, bu kayıp müşteriyi tahmin etme işlemini gerçekleştirebilmek için bu kitapta da bahsedilecek “lojistik regresyon” gibi bir istatistiksel yöntem kullanma.

#### 1.2. CRISP-DM Modeli

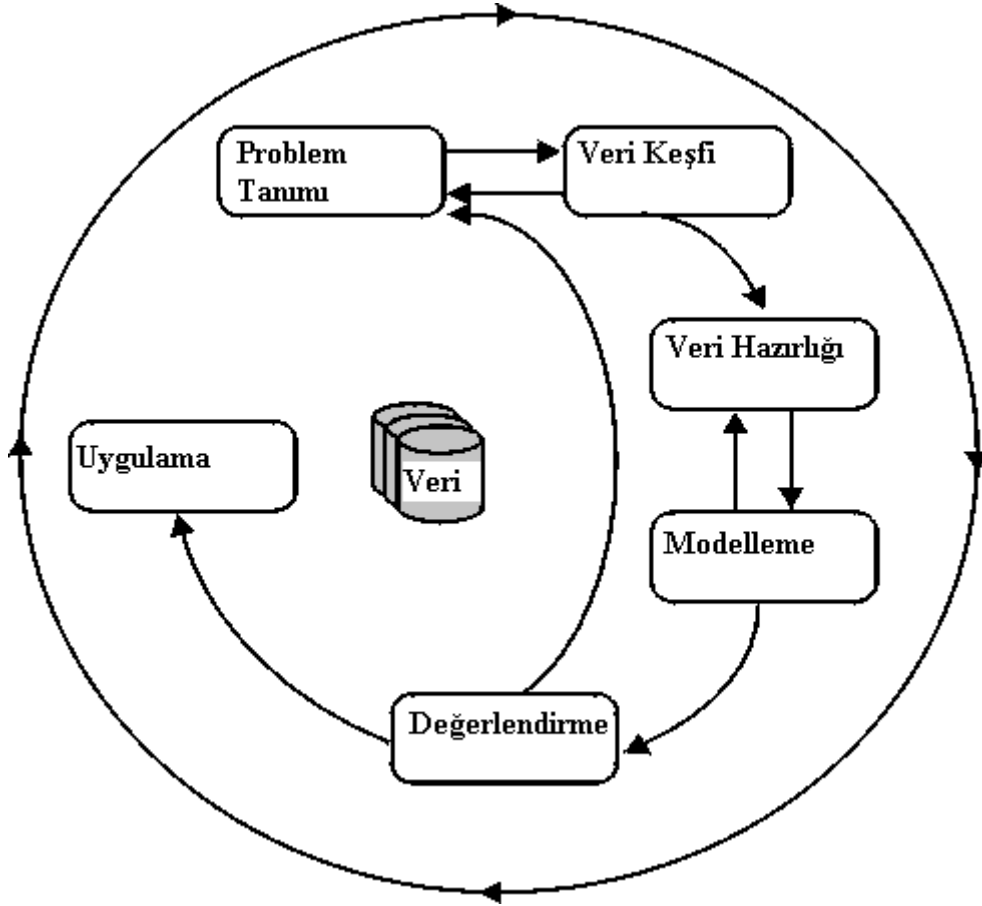
Veri madenciliği için endüstriler arası bir standart model 1996 yılında CRISP-DM adı altında geliştirildi (Larose, 2005). Buna göre bir veri madenciliği projesi, birbirine bağlı altı aşamadan oluşan bir döngü olarak özet-

lenebilir. Bu altı aşama aşağıda özetlenmiş ve Şekil 1.1’de grafik olarak gösterilmiştir.

- **Problem tanımı:** Bir veri madenciliği projesi bir iş sorununun farkına varılması ile başlar. Veri madenciliği uzmanları, iş uzmanları ve alan uzmanları bir iş perspektifinden proje hedefleri ve gereksinimlerini tanımlamak için yakın bir işbirliği içinde çalışır. Projenin amacı, daha sonra bir veri madenciliği problem tanımına çevrilir. Problem tanımı aşamasında, veri madenciliği araçları, henüz gerekli değildir.
- **Veri keşfi:** Alan uzmanları “meta data” üzerinde anlaşmaya varır. Bu uzmanlar veriyi toplar, tarif eder ve inceler. Ayrıca, veri kalitesi sorunlarını belirler. Problem tanımı aşamasında veri madenciliği uzmanları ve iş uzmanlarının sık sık görüş alışverişi hayati önem taşımaktadır. Veri keşif aşamasında geleneksel veri analizi araçları, örneğin betimsel istatistikler veriyi keşfetmek için kullanılır.
- **Veri hazırlama:** Alan uzmanları modelleme süreci için veri modeli oluşturur. Bazı veri madenciliği araçları sadece belli bir formatta veri kabul eder, dolayısıyla, uzmanlar veriyi toplamak, temizlemek ve gerekli formata uydurmak durumundadır. Ayrıca, örneğin, ortalama bir değer gibi yeni hesaplanan değişkenler oluşturulabilir. Veri hazırlama aşamasında, veri, birden çok defa değişikliğe uğratılabilir. Tablolar, kayıtlar ve niteliklerin seçilerek modelleme aracı için veri hazırlanması, bu aşamanın sonucudur.
- **Modelleme:** Veri madenciliği uzmanları, veri madenciliği için geliştirilmiş farklı matematiksel modelleri uygulamalıdır çünkü aynı sorunu çözmeyi amaçlayan birden fazla yöntem olabilir. Veri madenciliği uzmanlarının, her modeli değerlendirmesi gerekir. Modelleme aşamasında, veri hazırlama aşamasında çalışmış alan uzmanları ile sık sık görüşmek gereklidir. Modelleme aşaması ve değerlendirme aşaması birleşik gibidir, birden fazla defa tekrarlanabilirler. En iyi değerler elde edilene kadar birkaç kez modeldeki matematiksel parametreleri değiştirmek gerekebilir. Son modelleme aşaması tamamlandığında, yüksek kaliteli bir model kurulmuştur.
- **Değerlendirme:** Veri madenciliği uzmanları modeli değerlendirir. Model, onların beklentilerini tatmin etmiyorsa, bu modelleme aşamasına geri dönmeyi ve parametreleri en iyi değerleri elde edene ka-

dar değiştirerek modeli yeniden kurmayı gerektirebilir. Nihayet, modelden memnun olduğunda şu sorulara cevap verilmelidir: Model iş hedeflerine ulaşıyor mu? Tüm iş sorunları değerlendirildi mi? Değerlendirme aşamasının sonunda, veri madenciliği uzmanları, veri madenciliği sonuçlarının nasıl kullanılacağına karar verir.

- Uygulama: Veri madenciliği uzmanları, sonuçları diğer uygulamalara aktarırlar. Bu uygulamalar şirket çalışanları tarafından kolayca kullanılabilen veritabanları ya da elektronik çizelge uygulamaları olabilir.



Şekil 1.1. CRISP-DM aşamaları (URL13).



## 2. BÖLÜM:

### VERİ ÖNİŞLEMESİ

Crisp-DM aşamalarında da görüldüğü üzere ham veri, veri madenciliği işlemlerine girmeden önce veri keşfi ve veri hazırlama aşamalarından geçmektedir. Pyle, bir veri madenciliği projesinde veri hazırlamaya ayrılan zamanın toplam zamanın %60'ı, buna karşın modellemeye ayrılan zamanın toplam zamanın %5'i olduğunu iddia etmektedir (Pyle, 1999). Genelde veritabanlarında bulunan ham veri önışlemeye tabii tutulmadan önce çeşitli problemlere sahiptir. Bunların arasında:

- Gereksiz ya da kullanımdan düşmüş alanlar
- Eksik değerler
- Aykırı değerler
- Mantıksız değerler
- Kullanılması düşünülen modele uygun olmayan girdi formatı

#### 2.1. Yanlış Tasnif Edilmiş ya da Mantıksız Değerler

Mantıksız değerler birçok şekilde oluşabilir. Nitel ve nicel değişkenlerde farklı şekillerde mantıksız değerler girilmiş olabilir. Eğer bir veritabanında gelir değişkeni negatif olarak kaydedilmişse ya da sayısal olması gereken yaş değişkeni bir harfle belirtilmişse bu durumda bu kayıta bir aksaklık olduğu düşünülebilir. Bu mantıksız değerlerin düzeltilerek eksik değere dönüştürülmesi bir çözümdür.

Nitel değişkenlerde önemli bir sorun ise veri girişi yapanların aynı değer için farklı şekilde giriş yapmalarıdır. Örneğin kimi kayıta ülke "USA" kimisinde de "US" olarak girilmiş olabilir. Tablo 2.1'de verilen ve Capital 500 listesinde bulunan firmaların faaliyet gösterdiği illeri gösteren frekans tablosu da aynı hataya bir örnek oluşturmaktadır (URL4).

**Tablo 2.1.** 2005 Capital 500 listesi “İl” değişkenine göre frekans

İstanbul	261	Manisa	4	Zonguldak	2
İzmir	46	Mersin	4	Amasya	1
Bursa	30	Ordu	4	Aydın	1
Ankara	27	Antalya	3	Çanakkale	1
Kocaeli	20	Isparta	3	Çankırı	1
Kayseri	16	K.Maraş	3	Edirne	1
Gaziantep	14	Karaman	3	Giresun	1
Denizli	9	Konya	3	İzmit	1
Adana	8	Adıyaman	2	Kahramanmaraş	1
Eskişehir	5	Hatay	2	Karabük	1
Sakarya	5	Kütahya	2	Kırklareli	1
Balıkesir	4	Niğde	2	Mardin	1
Bolu	4	Tokat	2	Trabzon	1

Yukarıda İzmit Kocaeli olarak düzeltildikten sonra Kocaeli frekansı 21 olacaktır. Bir başka örnek yine aynı veritabanından, bu sefer sanayi kuruluşunun faaliyette bulunduğu sektöre bakarak verilebilir. Tablo 2.2’ye bakıldığında elektrik-elektronik, elektronik-elektrik, elektik-elektronik olarak üç farklı, ama aynı sektörü ifade eden kodlama olduğu görülebilir. Bu basit bir kodlama hatası olarak hemen düzeltilebilir. Aynı listenin ileriki senelerde yayınlanan versiyonları ile karşılaştırmalı bir analiz yapılacaksa kodlamanın sabit kalıp kalmadığına da bakılmalıdır. Örneğin 2009 listesinde “telekomünikasyon” ve “telekom” diye iki farklı kod ortaya çıkmıştır (URL5). Bu değişimin de bir hata mı olduğu yoksa mantıklı bir farklılaşmayı ifade ettiği araştırılmalıdır.

**Tablo 2.2.** 2005 Capital 500 listesi “Sektör” değişkenine göre frekans

Tekstil/konfeksiyon	79	İnşaat	7
Gıda	78	Maden	7
Otomotiv	44	Tütün	4
Kimya-ilaç	35	Alkollü içecek	3
Bilişim	30	Gıda-içecek	3
Ticaret-hizmet	26	Lastik	3
Demir-çelik	22	Basın	2
Elektrik-elektronik	21	Elektronik-elektrik	2
Çimento	20	Kuyum	2
Ağaç-Orman	14	Eletrik-elektronik	1
Enerji-petrol	14	Hızlı tüketim	1
Makine	13	Hizmet-ticaret	1
Ambalaj	12	İmalat	1
Metal	12	İnşaat	1
Cam-seramik	10	Kağıt	1
Plastik	10	Metal sanayi	1
Perakende	9	Orman ürünleri	1
Telekomünikasyon	8	Pazarlama	1
		Tekstil	1

## 2.2. Eksik Değerler

Bir kaydın tümüyle eksik değerlerden oluşması çok ender görülür, bu durumda zaten bu kaydı değerlendirme dışında tutmak tek uygun yöntemdir. Ancak çoğu zaman bazı değişkenlerin değerleri eksik, diğerlerinininki tamdır. Bu durumda, kullanılan istatistiki analizin özelliklerine göre çeşitli eksik değer tamamlama yöntemleri uygulanabilir. Bunlar:



- Eksik değeri bir sabit değeriyle yer değiştirme
- Eksik değeri o değişkenin ortalaması ile yer değiştirme
- Eksik değeri rastgele bir dağılımdan türetilen bir değeriyle değiştirme.

Yukarıdaki yöntemler birçok istatistiki pakette opsiyon olarak verilir. Sabit bir değeriyle eksik değerleri yer değiştirme varyansı ve kovaryansı olduğundan daha düşük gösterdiğinden çok fazla eksik değer varsa tercih edilen bir yöntem olmayabilir. Eksik değer tamamlama ile ilgili birçok teorik kaynak ve yazılım bulunmaktadır (Allison, 2002; Rubin, 1976, 1987; Schafer, 1997). Modelleme yapılan program içinde eksik değer tamamlama yapılabildiği gibi farklı bir platformda eksik veri tamamlanarak veri madenciliği programına işlenmiş veri girdi olarak kullanılabilir.

### 2.3. Veri Dönüşümleri

Kullanılan istatistiki model ham veri üzerinde çeşitli dönüşümler yapılmasını gerekli kılabilir. Bunun sebeplerinden biri de değişkenlerin ölçüldüğü birimlerin çok büyük farklılık gösterebilecek olmasıdır. Örneğin bir cep telefonu abonesinin aylık faturasının TL cinsinden değeri ve attığı SMS sayısı çok farklı ortalama ve varyansa sahip olabilir. Bu ölçek farkları kullanılan veri madenciliği algoritmasında bir değişkenin sonuca olması gerektiğinden fazla etki etmesini sağlayabilir. Değişkenleri “tek tip” hale getirmek için birçok standartizasyon yöntemi mevcuttur. En sık kullanılan standartlaştırma yöntemleri *min-max* ve *z-skor* standartlaştırma yöntemleridir.

*Min-max* standartlaştırma yöntemi sadece minimum ve maksimum değere göre minimum değeri 0 maksimum değeri de 1 olacak şekilde veriyi standartize eder. Yani,

$$X^* = \frac{X - \min(X)}{\text{aralık}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Burada *min-max* standart değerlerin verideki uç değerlere karşı hassas olduğunu belirtmek gerekir. Ayrıca veri aşağı yukarı tekdüze bir dağılıma

sahipse *min-max* standart değerleri yaklaşık olarak yüzdeliğe denk gelecektir. Bir örnek vermek gerekirse 2005 Capital 500 listesindeki 500 şirketin cirolarına bakıldığında en yüksek 10.364.153.845 TL ile Petrol Ofisi, en düşük ise 63.082.552 TL ile Rast Gıda olarak görülmektedir. Min-max standardizasyonu sonucu en düşük değer 0 en yüksek ise 1 değerini alacaktır. Hemen birinci şirketten sonra gelen şirket olan Ford Otosan ise 0,53 değeri almaktadır ki en yüksek değer yarısı kadar ciro yaptığını gösterir. Aynı işlem çalışan sayısı değişkenine göre yapıldığında en fazla çalışan sayısına sahip şirketin 7539 adet çalışana, en düşük çalışan sayısına sahip şirketin de 14 çalışana sahip olduğu görülmektedir. Bu durumda örneğin 710 çalışana sahip Pınar Süt 0,09 değerine sahip olacaktır.

Z-skor standartlaştırması ise değişkeni ortalaması 0 ve standart sapması 1 olan bir değişkene dönüştürür. Bunu yapmak için ortalamayı çıkarıp standart sapmaya bölmek yeterlidir. Burada “-“ değerler ortalamadan alttaki gözlemleri “+” değerler ise ortalamadan yüksek gözlemleri gösterir. Ortalamadan fark ise standart sapma biriminden ölçülür. Z değeri şu şekilde hesaplanır:

$$X^* = \frac{X - ortalama(X)}{s \text{ standart sapma}(X)} = \frac{X - \bar{X}_x}{s_x}$$

Bir örnek vermek gerekirse 35 kişilik bir sınıfta sınav ortalamasının 63,09 standart sapmasının ise 13,77 olduğunu düşünelim. Bu durumda en yüksek not olan 94 z değeri olarak 2,24’e denk gelecektir. En düşük değer olan 35 ise -2,04 z değerine denk gelecektir.

#### 2.4. Aykırı Değerlerin Tespiti

Aykırı değerler modele göre sonuçları çok etkileyen bir özelliğe sahip olabilirler. Kimi zaman bu aykırı değerler bir hata sonucu girilmiş değerler de olabilir. Dolayısıyla tespitleri önemlidir. Aykırı değerler grafik yöntemlerle olduğu gibi bir önceki alt bölümde anlatılan dönüşümlerden sonra ortaya çıkan standart değerlere bakarak da anlaşılabilir. Aykırı değerleri tespit etmenin en sık kullanılan yöntemi z değerlerine bakarak +3’ten fazla ya da -3’ten az gözlemleri aykırı değer olarak belirlemektir. Yalnız burada z değeri

lerini bulmak için hesaplanan standart sapma ve ortalama, tespit edilecek aykırı değerleri de içereceğinden bu değerler tarafından olumsuz yönde etkilenebilir. Bu hassasiyet durumu yüzünden aykırı değerlerin tespiti için farklı yöntemler de geliştirilmiştir.

Bu yöntemlerden biri de kartiller arası farkı (*IQR*) kullanarak aykırı değerleri belirler. Kutu ve bıyık diyagramı için de kullanılan 1. ve 3. kartiller arası fark, 1. ve 3. kartile 1,5 ile çarpılarak eklendiğinde ortaya çıkan sonuçlardan daha ekstrem olan değerler aykırı değerler olarak adlandırılır. Yani aşağıdaki formülde verilen alt ve üst değerlerden daha küçük ya da daha büyük değerler aykırı değer olarak sayılabilir:

$$Q1 - 1,5 \times IQR$$

$$Q3 + 1,5 \times IQR$$

Bir başka uygulamada (istatistiki paketlerde sıkça kullanılan) yukarıdaki aralıktan  $1,5IQR$  sonraya kadar bulunan değerlere “orta derecede aykırı değer” daha sonrakilere ise “şiddetli aykırı değer” denmektedir. Yani buna göre şiddetli aykırı değerlerin başlangıç sınırları aşağıdaki gibidir.

$$Q1 - 3 \times IQR$$

$$Q3 + 3 \times IQR$$

Aynı test skorları örneğine dönersek burada 1. Kartil ( $Q1$ ) 50, 2. ise 71,5 olarak ortaya çıkmaktadır. Yani  $IQR$  21,5’tur. Dolayısıyla 103,75’ten fazla test skorları ve de 13,75’ten az test skorları aykırı değer olarak kabul edilebilir. En düşük değer 35 en yüksek değer ise 94 olduğundan (zaten 100’ü geçemeyeceğinden) aykırı değer bu veri setinde yoktur.

Bir başka veri setine bakarak aykırı değerlere bir örnek verebiliriz. 4 Ocak 2000 tarihinden 17 Ekim 2008 tarihine kadar İMKB-100 endeks kapanış değerleri 2092 gözlemi içeren bir veri setidir. Günlük getiri hesaplanırsa 2091 adetlik getiri verisi oluşur. Bu veri incelendiğinde en kötü düşüşün %-18,11 ile 21 Şubat 2001 en büyük yükselişin ise %19,45 ile 5 Aralık 2000’de olduğu görülür.  $Q1$  %-1,39,  $Q3$  ise %1,47’dir. Dolayısıyla  $IQR$

%2,86 olarak hesaplanır. Dolayısıyla -5,68 ile 5,76 arasındaki bölge dışındaki getiriler aykırı değer olarak adlandırılabilir. Ayrıca -9,97 ile 10,05 dışındaki bölgede de şiddetli aykırı değer olarak adlandırılabilir. Bu şekilde incelendiğinde (bir sonraki bölümde sonuçlar grafik olarak da gösterilecektir) negatif bölümde 4 şiddetli olmak üzere 37 aykırı değer, pozitif bölümde ise 10 şiddetli olmak üzere 44 aykırı değer bulunmaktadır. Şiddetli aykırı değerler ve tarihleri Tablo 2.3'te verilmiştir.

**Tablo 2.3.** İMKB100 endeksi Ocak 2000-Ekim 2008 arası şiddetli aykırı değerler ve tarihleri

21.02.2001	-18,11%
19.02.2001	-14,62%
03.03.2003	-12,49%
17.03.2003	-10,57%
17.07.2002	10,13%
05.11.2002	10,17%
06.10.2003	10,62%
18.03.2003	11,58%
04.01.2001	11,61%
30.03.2001	12,05%
07.11.2002	12,52%
27.04.2001	13,53%
06.12.2000	18,64%
05.12.2000	19,45%



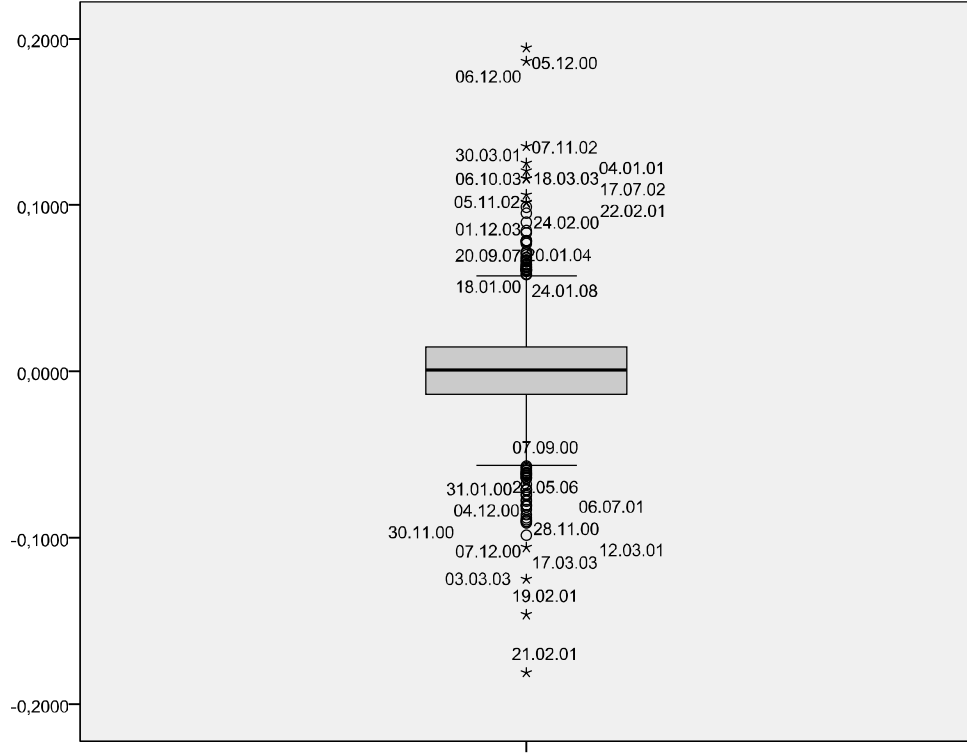
### **3. BÖLÜM:**

## **AÇINSAYICI VERİ ANALİZİ**

Açımsayıcı veri analizi kısaca “veriyi tanımak” için kullanılan çeşitli grafik ve tablo üretme yöntemleri olarak tanımlanabilir. İlk olarak Tukey tarafından “Exploratory Data Analysis” olarak popüler kullanıma kazandırılmıştır (Tukey, 1977). Veri madenciliği açısından düşünülürse, büyük miktarda veri ilk ele alındığında veriyi tanımak için kullanılan ve analiz dışı bırakılabilecek, yeniden kodlanması gerekebilecek ve de kurulan modellerde ne tür ilişkilerin araştırılması gerektiği konusuna ışık tutabilecek bir dizi grafik ve tablo üretme tekniği olarak önemlidir.

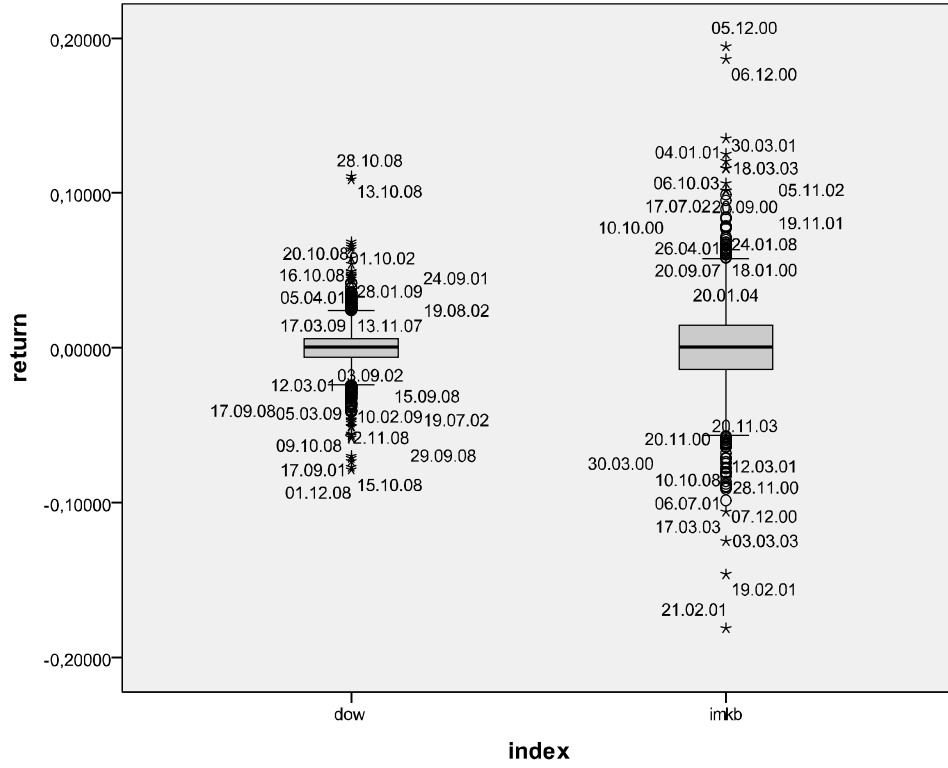
#### **3.1. Kutu Bıyık Diyagramı**

Kutu bıyık diyagramı farklı şekillerde oluşturulabilir. Kimi uygulamada doğrudan 5-Rakam, yani minimum, Q1, medyan, Q3 ve maksimum değerleri bir kutu ve bıyık formatında gösterilir. Farklı uygulamalarda aykırı değerler şiddetlerine göre gösterilebilir. Şekil 3.1, daha önce de kullanılan İMKB100 endeksi verisinin SPSS ile aykırı değerler, oluştukları tarih ile etiketlenmiş şekilde üretilmiş kutu bıyık diyagramını göstermektedir.



**Şekil 3.1.** İMKB100 endeksi Ocak 2000-Ekim 2008 arası kutu bıyık diyagramı

Yukarıdaki gibi bir grafik veri hakkında ilk bakışta bir fikir sahibi olmamızı sağlar. Aynı zamanda farklı değişkenler yan yana olarak da kutu bıyık diyagramında incelenebilir. Örneğin aynı süre aralığında Dow Jones endeksinin getirilerini İMKB getirilerinin yanında kutu bıyık diyagramı olarak göstermek istersek de bunu yapabiliriz. Şekil 3.2 iki grafiği yan yana göstermektedir (grafik SPSS ile yapılmıştır). Bu grafikte gerçekten iki piyasa arasında getiri değişkenliği açısından ne kadar fark olduğu açık bir biçimde görülebilir.

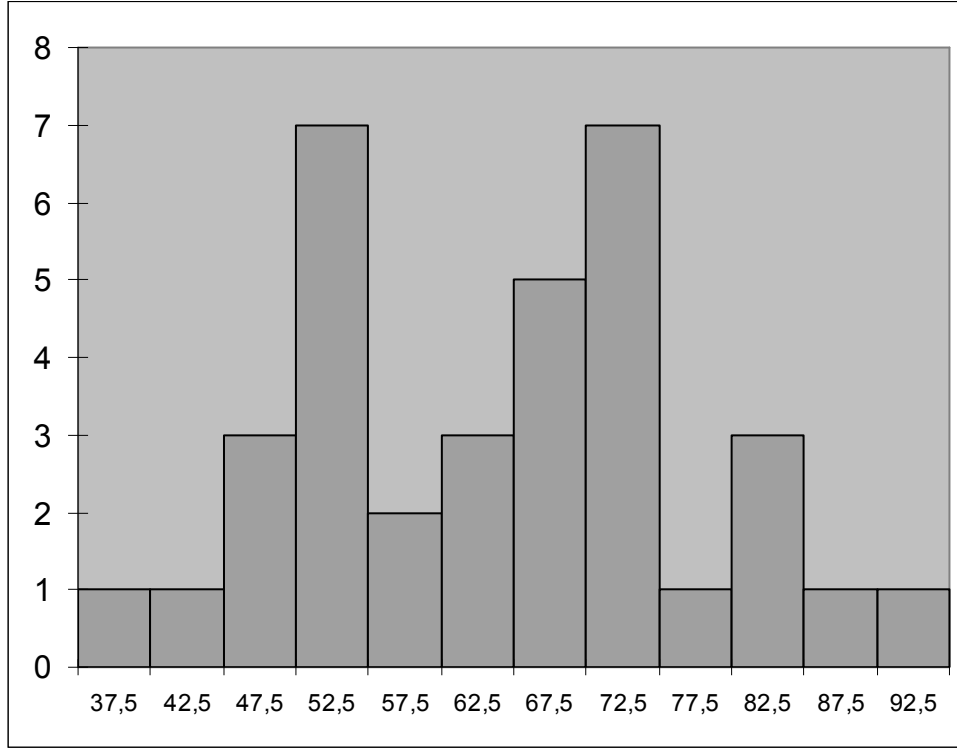


**Şekil 3.2.** İMKB100 endeksi ve Dow Jones yan yana kutu bıyık diyagramı

### 3.2. Histogram

Kutu bıyık diyagramı verinin genel şeklini gösterse de daha detaylı bir grafik için Histogram kullanılabilir. Histogram için nicel verinin çeşitli aralıklara bölünmesi gerekir. Bu aralıklar genellikle eşit uzunlukta seçilir. Grafikteki yatay eksen nicel değişken değerlerini (genellikle aralık ortalaması), dikey eksen ise verilen aralıktaki frekansı verir. Sütunların boyu bu frekansa orantısal olmalıdır. Ancak eşit uzunlukta olmayan versiyonda histogramdaki sütun uzunlukları değil altlarındaki alan ham frekansa orantısal olmalıdır. Daha önce verilen 35 öğrencinin sınav notlarına ilişkin örnekte  $[35,40)$ ,  $[40,45)$ ,  $[45,50)$ ,  $[50,55)$ ,  $[55,60)$ ,  $[60,65)$ ,  $[65,70)$ ,  $[70,75)$ ,  $[75,80)$ ,  $[80,85)$ ,  $[85,90)$  ve  $[90,95)$  aralıkları kullanılarak yapılan Histogram Şekil 3.3'te verilmiştir. Burada yatay eksen etiketleri ilgili aralıktaki ortalama değeri vermektedir.

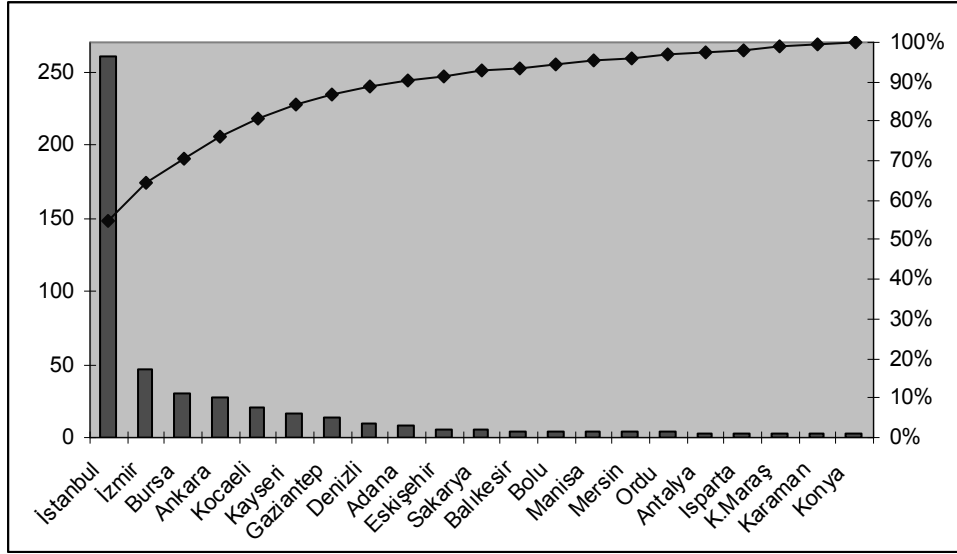




**Şekil 3.3.** Bir istatistik dersinde 35 öğrencinin sınav notları

### 3.3. Pareto Diyagramı

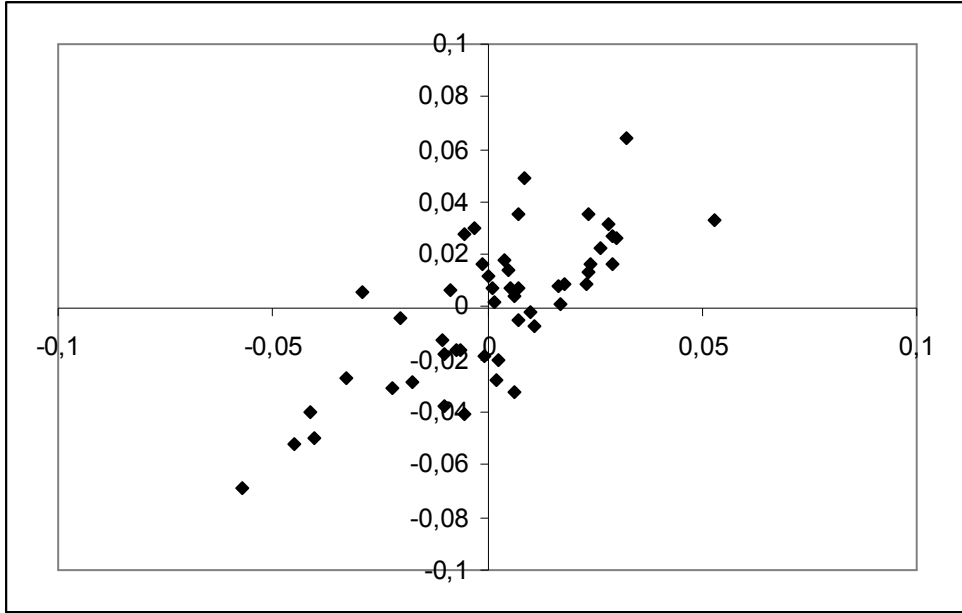
Pareto grafiği nitel veriler için yapılan sütun grafiğinin değişik bir adaptasyonudur ve kalite yönetiminde özellikle uygulama alanı oldukça fazladır. Pareto kuralı ya da 20-80 kuralı olarak da bilinen kural, etkilerin yaklaşık %20'sinin etkilenenlerin yaklaşık %80'ini oluşturduğunu söyler. Örneğin bir kalite araştırmasında tüketici şikayet sayıları “etkilenen”, çeşitli tüketici şikayet sebepleri de “etkiler” olarak adlandırılırsa bu kural işletilebilir. Pareto grafiği ise etkileri yatay eksen, yüksek frekanstan düşüğe, dikey eksen ise (çift dikey eksen) hem frekanslar hem de birikimli frekansları gösterir. Şekil 3.4’te Capital 500 örneğinde verilen sanayi kuruluşlarının illere göre dağılımı gösterilmektedir. En az üç sanayi kuruluşu bulunan 21 ildeki toplam sanayi kuruluşu sayısı olan 477 adet in %80’i ilk beş ilde bulunmaktadır, yani İstanbul, İzmir, Bursa, Ankara ve Kocaeli.



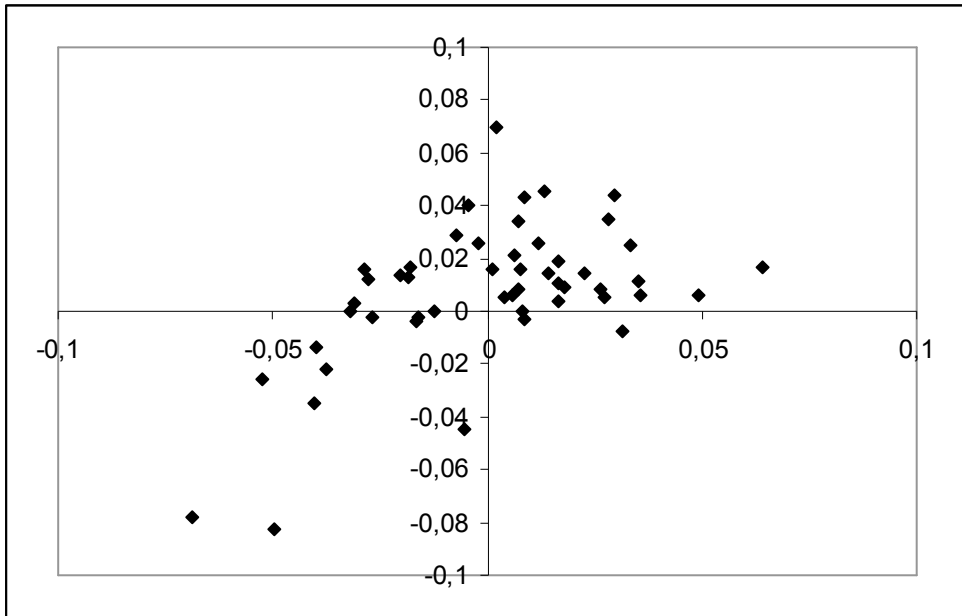
Şekil 3.4. Pareto diyagram örneği

### 3.4. Serpilme Diyagramı

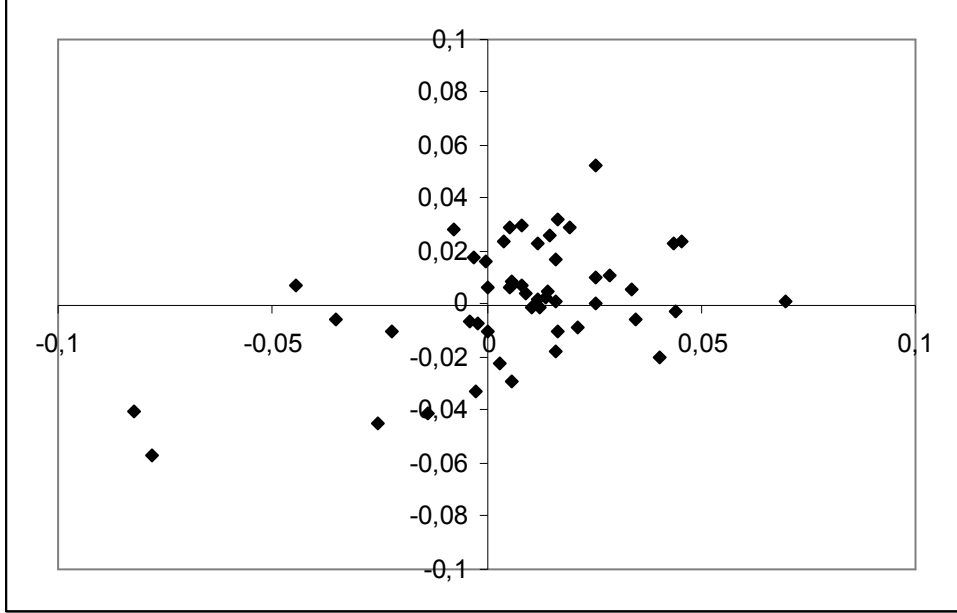
İki nicel değişkenin arasındaki ilişkiyi merak edersek bu değişkenlerin serpilme diyagramını kullanabiliriz. Değişkenler doğal olarak eşleşmiş olmalıdır, örneğin aynı haftalarda iki farklı piyasa endeksindeki hareket gibi. Örneğin 2010 yılında her hafta için Dow Jones, Bovespa ve Jakarta getiri verilerini indirmiş olalım (URL6). Bir haftayı Dow Jones ve Bovespa'da silerek (Jakarta'da tatil olan) veriyi ön hazırlıktan geçirerek serpilme diyagramlarını yapabiliriz. Bu değişkenleri üç farklı serpilme diyagramında göstererek aralarındaki ilişkiyi görmeye çalışabiliriz. Şekil 3.5, 3.6, ve 3.7 olası üç farklı kombinasyonu sunmaktadır. Şekillerden de anlaşılacağı gibi Bovespa-Dow Jones çifti yüksek lineer korelasyona sahip olmakla birlikte, diğer iki çift düşük lineer korelasyona sahip görünmektedir. Bu gözlem altında hesaplanan Pearson korelasyon katsayıları ile de doğrulanabilir. İlk grafikte görülen Dow Jones-Bovespa çifti arasındaki korelasyon 0,76'dır. İkinci grafikteki korelasyon ise 0,55, son grafikteki ise 0,48 olarak hesaplanmıştır. Bu hesaplanan değerler de görsel değerlendirmeler ile örtüşmektedir.



Şekil 3.5. Yatay eksen Dow Jones, düşey eksen Bovespa



Şekil 3.6. Yatay eksen Bovespa, düşey eksen Jakarta



**Şekil 3.7.** Yatay eksen Jakarta, düşey eksen Dow Jones

### 3.5. Çapraz Tablo

Daha önce kullanılan 500 büyük şirket verisine baktığımızda iki kategorik değişkenin arasında ne tür bir ilişki olduğunu merak edersek bir çapraz tablo kullanabiliriz. Örneğin sektör ve il değişkenleri arasında nasıl bir ilişki vardır? Belli sektörler belli illerde toplanmış mıdır? Daha önce bahsedilen düzeltmeleri de yapıp tabloyu kısaltmak amacıyla bazı illeri ve sektörleri de çıkarırsak Tablo 3.1 oluşturulan çapraz tabloyu göstermektedir. Çapraz tabloda herhangi bir ilişki olup olmadığı görüldükten sonra bu ilişkinin şiddetini ölçmek için de çeşitli istatistikler mevcuttur. Ayrıca ki-kare testi ile bir örneklemden toplanan veri ile ilgili iki nitel değişkenin bağımsız olup olmadığı testi yapılabilir. Tablo 3.1’de tablonun çok büyük olmaması için bazı il ve sektörler elimine edilip 402 adet şirket seçilmiştir.

**Tablo 3.1.** 2005 Capital 500 listesi “Sektör” ve “İl” arasında çapraz tablo

Sektör / il	Adana	Ankara	Bursa	Denizli	G.Antep	İst.	İzmir	Kayseri	Kocaeli	Top.
Ağaç-Orman	1					5	3	3	2	14
Ambalaj		2		1	1	6	1			11
Bilişim		1			1	28				30
Cam-seramik						6	1		1	8
Çimento	1	2	1	1		3	3		1	12
Demir-çelik		1	2	1		9	3		2	18
Elektronik-elektrik		4	1	1		15	1	2		24
Enerji-petrol			2			11			1	14
Gıda		2	2		2	34	11	2	2	55
Kimya-ilaç		2				22	6	1	4	35
Maden		3				2				5
Makine		1	2			5			1	9
Metal			1			7		2	2	12
Otomotiv		4	8		1	17	5		4	39
Perakende						8	1			9
Plastik	1				1	4	3			9
Tekstil / konfeksiyon	5	1	10	4	8	31	3	5		67
Telekomünikasyon						8				8
Ticaret-hizmet		2	1			17	2	1		23
Toplam	8	25	30	8	14	238	43	16	20	402

### 3.6. Diğer Görselleştirme Teknikleri

Yukarıda bahsedilen tek değişkenli ya da çift değişkenli görselleştirme yöntemleri yanı sıra çok boyutlu görselleştirme yöntemleri de veri madenciliğinde önemli açmsayıcı yöntemler olarak kullanılabilir. Bu yöntemlerden Çok Boyutlu Ölçekleme ve Uyum Analizi yöntemlerinin uygulamalarla geniş anlatımı Akküçük (2011a) kaynağında bulunabilir. Ayrıca çeşitli başka kaynaklarda da farklı çok boyutlu görselleştirme tekniklerinin algoritmik özellikleri ve karşılaştırmada kullanılan yöntemler ile ilgili detaylar bulunabilir (Akküçük, 2004; 2009, Akküçük ve Carroll, 2006a; 2006b; 2010).

## 4. BÖLÜM:

### İSTATİSTİKİ TAHMİN TEKNİKLERİ

Temel tek değişkenli istatistikler, güven aralıkları, basit regresyon ve çoklu regresyon gibi istatistikçilerin senelerdir uyguladığı yöntemler de birer veri madenciliği unsuru sayılabilir. Bu nedenle bu bölümde sıkça kullanılan tek değişkenli istatistikleri, güven aralıklarını, iki nicel değişken arasındaki ilişkiyi tespit etmek için kullanılabilecek korelasyon ve bu konuyla bağlantılı olarak basit regresyon ve çoklu regresyon konularının üzerinden geçeceğiz.

#### 4.1. Tek Değişkenli İstatistikler

Veriyi tanımak için analizin başında grafik yöntemler yanında sayısal bazı istatistiklerin de hesaplanması faydalı olacaktır. Bu istatistikleri merkezi yatkinlık ölçüleri, değişkenlik ölçüleri ve asimetri ölçüleri olarak gruplasak şu şekilde özetleyebiliriz.

- Merkezi Yatkinlık Ölçüleri:

- Ortalama:  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ .
- Kırılmış ortalama: Yukarıdaki değer aykırı değerler tarafından etkilenebileceğinden en yüksek ve düşük %  $i$  değer çıkarılarak kırılmış ortalama hesaplanabilir, “ $i$ ” burada 5, 10 gibi bir değer olabilir.
- Mod: En sık tekrar eden değer.
- Medyan: Gözlem sayısı çift ise sıralı veride  $n/2$  ile  $(n+1)/2$  gözlem değerlerinin ortalaması, gözlem sayısı tek ise sıralı veride  $(n+1)/2$ ’inci veri. Aynı zamanda 2. Kartil ya da 50. yüzdelik.
- Yüzdelikler ve kartiller: Değişik isimleri olsa da veriyi belirli bir yüzdesi altta belirli bir yüzdesi üstte olmak üzere ikiye ayıran sayı-

lar grubu. Örneğin 25. yüzdelik (aynı zamanda 1. Kartil) veriyi  $\frac{1}{4}$  ve  $\frac{3}{4}$  büyüklüğünde iki parçaya ayırır. Bu konuya bu altbölümün sonunda verinin normal dağılıma ne kadar yakın olduğunun kontrolü ile ilgili bir analiz için tekrar dönülecektir.

- Değişkenlik Ölçüleri:

- Değişim aralığı: en yüksek değer-en düşük değer
- Kartiller arası fark (*IQR*): 3. Kartil – 1. Kartil

- Varyans(evren):  $\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$

- Varyans(örneklem):  $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

- Standart sapma (evren):  $\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$

- Standart sapma (örneklem):  $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

- Değişkenlik katsayısı (ölçme seviyesi oranlı olmalıdır):  $CV = \frac{s}{\bar{x}}$

- Asimetri ve basıklık ölçüleri

- Çarpıklık (asimetri):  $\gamma = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$

- Basıklık:  $\beta = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2}$

$$\circ \text{ Alternatif basıklık formülü: } \beta = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

[bu formülde yukarıdaki formülde bulunan basıklık, aslında standart basıklık değeri kabul edilen standart normal dağılımın basıklık değeri olan “3” çıkarılarak standardize edilir, dolayısıyla (+) değerler leptokurtic (sivri), (–) değerler ise platykurtic (düz) olarak adlandırılabilir. Diğer formülde – değer imkanı yoktur]

Daha önce verilmiş üç farklı piyasa endeksi için hesaplanan temel istatistikler Tablo 4.1’de verilmiştir. Tablo incelendiğinde en fazla ortalama getiri JKSE’de gözlenir. Değişkenlik ölçülerinde ise standart sapmada BVSPA, değişim aralığında ise JKSE en fazladır. Basıklık ölçüsünde JKSE ciddi derecede fazladır ki bu da “lepto-kurtic” yani sivri bir dağılımı işaret eder. Ortada sivri ama dağılımın uçlarında da yoğunluk olan bir dağılım bu şekilde yüksek basıklık değeri verebilir. Haftalık getiriler içerisinde çok daha fazla aykırı değeri JKSE endeksinde bulmayı bekleyebiliriz.

**Tablo 4.1. DJI-BVSPA-JKSE 2010 haftalık getiri özet istatistikleri**

<i>DJI</i>		<i>BVSPA</i>		<i>JKSE</i>	
Ortalama	0,001694	Ortalama	0,000020	Ortalama	0,007355
Standart Hata	0,003113	Standart Hata	0,003921	Standart Hata	0,003832
Medyan	0,003222	Medyan	0,006111	Medyan	0,009702
Mod	#N/A	Mod	#N/A	Mod	#N/A
Standart Sapma	0,022014	Standart Sapma	0,027724	Standart Sapma	0,027093
Varyans	0,000485	Varyans	0,000769	Varyans	0,000734
Basıklık	0,511833	Basıklık	-0,111067	Basıklık	3,397742
Çarpıklık	-0,503970	Çarpıklık	-0,271317	Çarpıklık	-1,211156
Aralık	0,109873	Aralık	0,132900	Aralık	0,151917
Minimum	-0,057063	Minimum	-0,068992	Minimum	-0,082274
Maksimum	0,052811	Maksimum	0,063909	Maksimum	0,069643
Toplam	0,084723	Toplam	0,001001	Toplam	0,367735
Sayı	50	Sayı	50	Sayı	50



Toplanan verinin hangi dağılımdan olduğu analizde önem taşır. Örneğin normal dağılım birçok istatistiki tekniğin varsayımı olarak karşımıza çıktığından verinin normal dağılımdan gelip gelmediği araştırılmak istenebilir. Betimsel istatistikler hesaplandıktan sonra aralarından bazıları bu amaçla kullanılabilir. Örneğin normal dağılımda basıklık ve çarpıklık sıfır ya da sıfıra yakın olmalıdır. Histogram çizilerek grafik olarak da dağılımın simetrik ve çan eğrisi şeklinde olup olmadığı gözlenebilir. Bunların dışında *pp* plot ya da *qq* plot olarak da bilinen normal dağılıma uygunluk grafikleri yapılabilir ve bu grafikten bir korelasyon katsayısı hesaplanarak normal dağılıma uygunluk için bir sayısal ölçü de getirilebilir (*ppcc* – percentile plot correlation coefficient). Bu grafik birkaç şekilde yapılabilir ama öncelikle verideki her gözlemin sıralanarak (küçükten büyüğe) bir yüzdelik değeri belirlenmesi gerekir. Küçükten büyüğe sıralandığında ve bir sıra numarası verildiğinde aşağıdaki gibi yüzdelik hesaplanabilir:

$$yüzdelik = \frac{sıra}{n+1}$$

Farklı bir formülasyonda sıra numaraları yüzdeliğe şu şekilde çevrilebilir:

$$1. \text{ sıra için: } yüzdelik = \frac{0,5}{n}$$

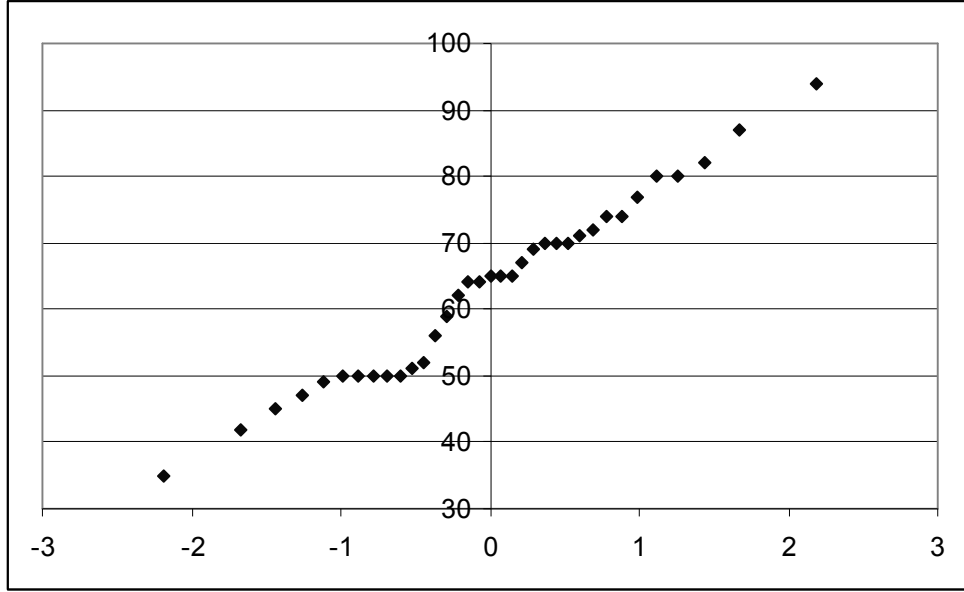
$$2. \text{ 3. ve } n-1 \text{ sırasına kadar: } yüzdelik = \frac{sıra - 0,3175}{n + 0,365}$$

$$\text{En büyük değer yani } n \text{ için: } yüzdelik = 1 - \frac{0,5}{n}$$

Yüzdeliklerin hesaplanmasından sonra bu yüzdeliklere normal dağılımda tekabül eden *Z* değerleri bulunmalıdır. Bu işlem tablo yardımı ile ya da Excel kullanarak “=normsinv(yüzdelik)” fonksiyonu ile yapılabilir. Daha sonra orijinal veri ve bu *Z* değerleri arasında serpilme diyagramı ve korelasyon katsayısı hesaplanabilir. Tablo 4.2 35 sınav sonucu için iki farklı şekilde hesaplanmış yüzdelik ve 2. formüle göre bulunan *Z* değerlerini vermektedir. Şekil 4.1 ise oluşturulan grafiği göstermektedir. Normal dağılımdan ufak bir sapma dışında çok fark yoktur ki bu sapma da histogramda zaten görülmektedir. Korelasyon (*ppcc*) oldukça yüksek, 0,9886 olarak ortaya çıkmaktadır.

**Tablo 4.2. 45 sınav sonucu yüzdeler ve Z değerleri**

Veri	Sıra	Yüzdelik 1	Yüzdelik 2	Z
35	1	2,78%	1,43%	-2,18935
42	2	5,56%	4,76%	-1,66883
45	3	8,33%	7,59%	-1,43354
47	4	11,11%	10,41%	-1,25837
49	5	13,89%	13,24%	-1,11509
50	6	16,67%	16,07%	-0,99166
50	7	19,44%	18,90%	-0,88174
50	8	22,22%	21,72%	-0,78157
50	9	25,00%	24,55%	-0,68868
50	10	27,78%	27,38%	-0,6014
51	11	30,56%	30,21%	-0,51847
52	12	33,33%	33,03%	-0,43897
56	13	36,11%	35,86%	-0,36216
59	14	38,89%	38,69%	-0,28742
62	15	41,67%	41,52%	-0,21426
64	16	44,44%	44,34%	-0,14224
64	17	47,22%	47,17%	-0,07094
65	18	50,00%	50,00%	-1,4E-16
65	19	52,78%	52,83%	0,070938
65	20	55,56%	55,66%	0,142236
67	21	58,33%	58,48%	0,214265
69	22	61,11%	61,31%	0,287424
70	23	63,89%	64,14%	0,362157
70	24	66,67%	66,97%	0,438972
70	25	69,44%	69,79%	0,518473
71	26	72,22%	72,62%	0,601397
72	27	75,00%	75,45%	0,688684
74	28	77,78%	78,28%	0,781567
74	29	80,56%	81,10%	0,881743
77	30	83,33%	83,93%	0,991661
80	31	86,11%	86,76%	1,115095
80	32	88,89%	89,59%	1,258373
82	33	91,67%	92,41%	1,43354
87	34	94,44%	95,24%	1,668833
94	35	97,22%	98,57%	2,18935



**Şekil 4.1. 35 sınav sonucu yüzdeler ve Z değerlerinin serpilme diyagramı**

#### 4.2. Güven Aralıkları

Bazı durumlarda yukarıda anlatılan tek değişkenli istatistikler örneklem yerine tüm evrenden hesaplanmış olabilir. Bazı veri madenciliği görevlerinde örneklem yerine eldeki tüm veri inceleniyor olabilir. Bu durumda elbette güven aralıklarına gerek duyulmaz. Ancak bir örneklem alınmışsa nokta kestirimi için belirlenecek artı eksi bir hata payı araştırma sonuçlarının incelenmesinde faydalı olacaktır. Güven aralıkları yukarıdaki altbölümde anlatılan tek değişkenli istatistiklerin örneklem büyüklüğünü de göz önüne alarak “ne kadar hassas” olduğunu göstermek için kullanılan bir istatistiksel araçtır. Güven aralıkları genelde aşağıdaki formu alırlar:

$$\text{nokta kestirimi} \pm \text{standart hata}$$

Burada nokta kestirimi güven aralığı hesaplanan istatistik için bulunan örneklem istatistiğidir. Standart hata ise matematiksel bir formül, verilen güven seviyesi ve örneklem sayısından bulunabilir. Örneklem dağılımı ma-

tematiksel olarak ifade edilemediği durumlarda yeniden örnekleme (bootstrapping) yoluyla standart hata kestirilebilir. Bir örneklemden bulunan ortalama değer için genel güven aralığı aşağıdaki gibi bulunabilir:

$$\bar{x} \pm t_{\alpha/2; n-1} \frac{s}{\sqrt{n}}$$

Aynı şekilde bir örneklemden ortaya çıkan oran için güven aralığı da şu şekilde bulunur:

$$p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Bir örnek vermek gerekirse bir bankaya kredi başvurusunda bulunanlardan rastgele seçilen 300 kişilik beyaz ve 300 kişilik siyah gruplarda bankalara yaptıkları kredi başvurusunun geri çevrilme oranını incelediğimizi düşünelim. Siyahlarda oran %37 beyazlarda ise 15% olsun<sup>1</sup>. Bu oranlar için %95 güven aralıkları hesaplandığında:

$$0,37 \pm 1,96 \sqrt{\frac{0,37 \times 0,63}{300}} = (\%31,54 - \%42,46)$$

$$0,15 \pm 1,96 \sqrt{\frac{0,15 \times 0,85}{300}} = (\%10,96 - \%19,04)$$

Bu şekilde siyah ve beyaz ret oranları arasındaki farkın rastgele bir fark olmadığı görülür, örneklem dağılımını da hesaba katarak hata payının artı eksi % 5 civarındadır. Burada örneklem sayısının küçülmesi güven aralıklarını da genişletecektir.

<sup>1</sup> Carnegie Mellon Üniversitesi “Data and Story Library” (URL7) içerisinde benzer bir veri seti vardır (URL8).

### 4.3. Basit Regresyon

İncelenen değişken sayısı birden fazla olduğu durumlarda çift değişkenli analiz yöntemleri kullanılabilir. Bu yöntemlerden en sık kullanılanı iki değişken arasında lineer bir ilişki olup olmadığı sorusunu cevaplayan Pearson korelasyon katsayısıdır. Gözlemler ve değişkenlerden oluşan bir veri matrisi  $\mathbf{X}$ 'ten değişken ortalamalarını çıkararak  $\mathbf{X}^*$ , aynı zamanda da standart sapmalara bölerek  $\mathbf{X}^{**}$  oluşturursak varyans-kovaryans matrisi  $\mathbf{S}$  ve korelasyon matrisi  $\mathbf{R}$  aşağıdaki işlemlerle bulunur. Burada  $\mathbf{S}$  matrisinde diyagonalde varyanslar,  $\mathbf{R}$  matrisinde ise diyagonalde 1'ler bulunmaktadır.

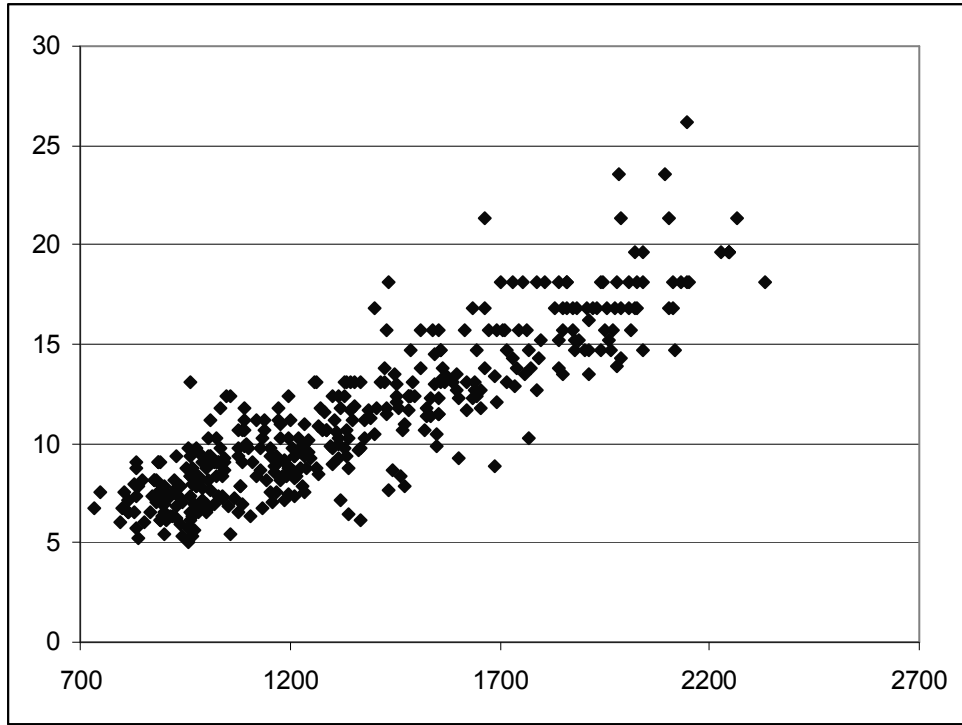
$$\mathbf{S} = \frac{1}{n-1}(\mathbf{X}^*)'(\mathbf{X}^*)$$

$$\mathbf{R} = \frac{1}{n-1}(\mathbf{X}^{**})'(\mathbf{X}^{**})$$

Korelasyon analizini ve basit regresyonu uygulamalı olarak göstermek amacıyla University of California Irvine veritabanından (URL9) indirilen otomobil benzin tüketimi (URL10) verisini kullanacağız. 398 otomobil için toplanan bu veride aşağıdaki değişkenler bulunuyor:

- Benzin tüketimi: Orijinalde “Miles per gallon” (galon başına mil) 100 km.de litre olarak çevrildi.
- Silindir sayısı
- Motor hacmi (kübik inç cinsinden)
- Beygirgücü
- Ağırlık (pound cinsinden kg.'ye çevrildi)
- Hızlanma
- Model senesi (79 ve 82 arasındadır)
- Üretim yeri, nitel değişken (1- Amerika, 2-Avrupa, 3-Japonya)
- Arabanın marka ve modeli (Honda Accord gibi)

İlk değişken yani benzin tüketimi ile arabanın ağırlığını incelemeye çalışırsak serpilme diyagramı yardımcı olabilir. Şekil 4.2 de görülen serpilme diyagramı ilişkinin doğrusal bir ilişki olduğu yolunda izlenim vermektedir. Görüldüğü gibi ağırlık 800 kg ile 2300 kg. Aralığında artarken yakıt tüketimi de neredeyse doğrusal şekilde 100 km’de 5 litreden 25 litrenin üzerine kadar çıkmaktadır.



**Şekil 4.2. Yatay eksen ağırlık düşey eksen benzin tüketimi serpilme diyagramı**

Aradaki ilişkinin sayısal büyüklüğünden bahsedersek korelasyon katsayısı kullanılabilir, bu veri için 0,8852’ye eşittir. Eğer yakıt tüketimi değişkenini ağırlık yoluyla tahmin etmeye çalışırsa basit regresyon kullanabiliriz. Tahmin edilen değişken ya da bağımlı değişken “yakıt tüketimi” ve tahmin eden değişken ya da bağımsız değişken “ağırlık” olmak üzere yapılan regresyon hesaplarından aşağıdaki gibi gösterilebilecek bir tahmin eğrisi ortaya çıkar.

$$\hat{y} = b_0 + b_1x$$

Burada  $b_0$  ile ifade edilen değer regresyon denkleminin  $y$  eksenini kestiği yeri ve  $b_1$  ile ifade edilen değer ağırlıktaki birim artışta yakıt tüketimini ne kadar arttığını göstermektedir. Excel ile bulunan regresyon sonuçları Tablo 4.3'te verilmiştir. Buna göre her kilogram ek ağırlık yakıt tüketimini 100 kilometrede 0,009 litre derecesinde artırmaktadır. Model  $r^2$  istatistiğine bakıldığında  $Y$ 'deki değişkenliğin % 78'inin  $X$  tarafından açıklandığı görülmür. Başka bir deyişle ağırlık bilgisinin yokluğunda, sadece  $Y$  değerlerini kullanarak yaptığımız tahmin hatası (ki en iyi tahmin ortalama  $Y$  olacaktır), ağırlık bilgisini eklediğimizde %78 oranında azalmaktadır. Yani  $X$  kullanarak,  $X$  yokluğundaki tahmin hatasının %22'sini yaparız. Tahminler gözlemler ile aynı düzlemde gösterildiğinde tahmin hatasını görsel olarak da inceleyebiliriz. Excel tarafından üretilen grafik Şekil 4.3'te verilmiştir.

**Tablo 4.3. Ağırlık ve yakıt tüketimi basit regresyon sonuçları**

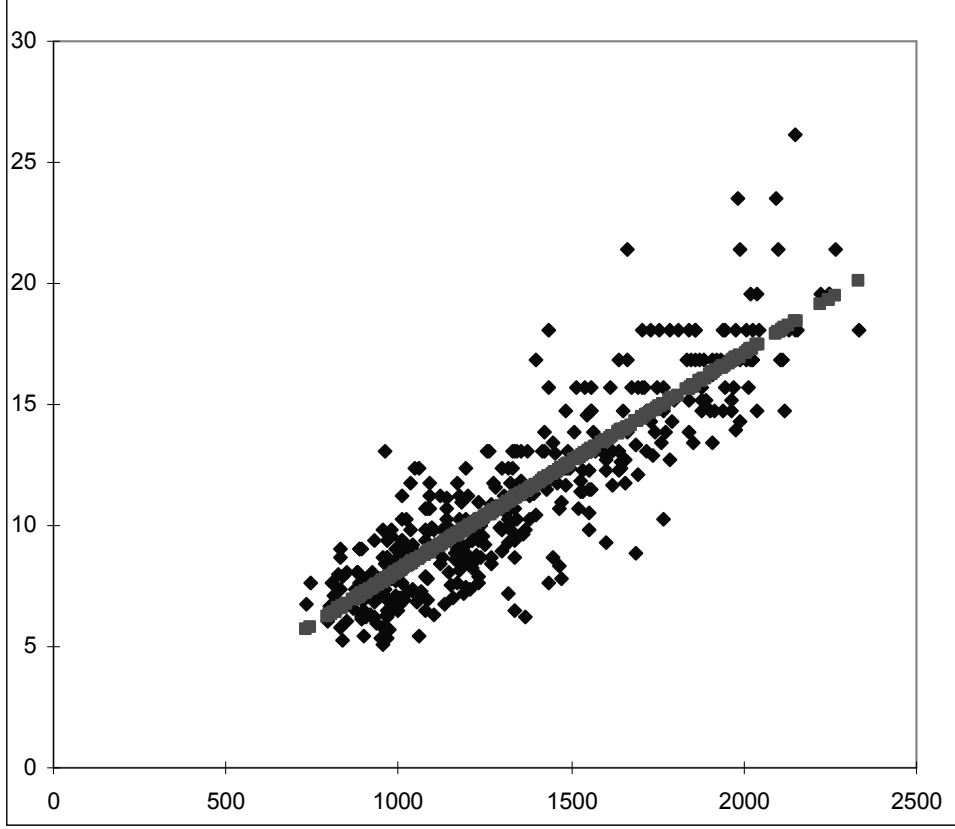
<i>Regresyon İstatistikleri</i>					
Çoklu R	0,88521856				
R Kare	0,783611898				
Düzeltilmiş R Kare	0,783065464				
Standart Hata	1,817486647				
Gözlem Sayısı	398				

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Anlamlılık F</i>
Regresyon	1	4737,020763	4737,020763	1434,045	1,1E-133
Kalıntı	396	1308,090054	3,303257712		
Toplam	397	6045,110817			

	<i>Katsayılar</i>	<i>Standart Hata</i>	<i>t Stat</i>	<i>P-değeri</i>
Sabit	-0,903073931	0,332674642	-2,714586014	0,006926
Ağırlık	0,008992534	0,000237466	37,86878885	1,1E-133



**Şekil 4.3. Yatay eksen ağırlık düşey eksen benzin tüketimi serpilme diyagramı üzerinde regresyon eğrisi çizilmiş hali.**

#### 4.4. Çoklu Regresyon

Bağımlı değişken, birden fazla bağımsız değişken ile açıklanmak istenirse çoklu regresyon uygulanabilir. Bölüm 4.3'te verilen araba örneğine devam edersek ve yakıt tüketimini sadece tek değişkenle değil silindir sayısı, beygir gücü, motor hacmi ve hızlanmayı da ekleyerek 5 değişkenle açıklamaya çalışırsak sonuçta çoklu regresyon yapmış oluruz. Bu örnekte sonuçlar Tablo 4.4'teki gibi çıkar. Tahminin standart hatası iyileşmiş, 100 km.de yapılan tahmin hatası ortalama 1.82'den 1.66'ya düşmüştür. Ancak modelin açıkladığı varyans oranı %78'den % 82'ye yükselerek çok fazla iyileşme göstermemiştir. Düzeltilmiş  $r^2$  ise gözlem sayısının değişken sayısından çok fazla olması dolayısıyla zaten  $r^2$  değerinden çok farklı değildir. Modelde



regresyon katsayıları  $t$  değerleri incelendiğinde motor hacminin 0,05 seviyesinde anlamlı olmayan tek değişken olduğu görülmektedir. Modeli bu şekilde kullanmak yada bazı değişkenleri eksiltmek model kurucunun karar vermesi gereken noktalardır. Tahmin hatasındaki iyileşmenin, fazladan değişken kullanımını haklı çıkarıp çıkarmayacağı da analizcinin cevaplaması gereken bir sorudur.

**Tablo 4.4. Yakıt tüketimi çoklu regresyon sonuçları**

<i>Regresyon İstatistikleri</i>					
Çoklu R	0,906458				
R Kare	0,821666				
Düzeltilmiş R Kare	0,819392				
Standart Hata	1,658349				
Gözlem Sayısı	398				

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Anlamlılık F</i>
Regresyon	5	4967,064	993,4128	361,2253	2,7E-144
Kalıntı	392	1078,047	2,75012		
Toplam	397	6045,111			

	<i>Katsayılar</i>	<i>Standart Hata</i>	<i>t Stat</i>	<i>P-değeri</i>
Sabit	-2,99004	1,030623	-2,90119	0,003927
Silindir	0,356531	0,159182	2,239769	0,025666
Hacim	-0,00093	0,00354	-0,26396	0,791952
Beygircü	0,044503	0,006417	6,935309	1,68E-11
Hızlanma	0,12438	0,04819	2,58101	0,010214
Ağırlık	0,004354	0,000694	6,275834	9,22E-10

## 5. BÖLÜM:

### k-MEANS ALGORİTMASI

*k*-means algoritması sıkça kullanılan bir kümeleme tekniğidir. Örneğin bir banka ya da perakendeci müşterilerini grup içinde birbirlerine yakın özellikler gösteren farklı gruplara ayırmak istiyorsa kullanılabilir. Çok değişkenli analiz yöntemleri içinde kümeleme tekniği bir içbağımlılık modeli olarak adlandırılabilir. Yani belirli bir değişken (ya da bir grup değişken) bağımlı olarak adlandırılmamaktadır. Bir önceki bölümde bahsedilen regresyon yöntemi ise bağımlılık analiz eden bir yöntemdir. Veri madenciliği terminolojisinde ise aynı ayrım “gözetimli öğrenme”(supervised learning) ve “gözetimsiz öğrenme” (unsupervised learning) olarak yapılmaktadır. Bir kümeleme algoritması da gözetimli çalışabilir (örneğin bir sonraki bölümdeki *k*-en yakın komşu algoritması) ancak bu durumda bir hedef değişkenin belirlenmesi ve kümeleme algoritmasının bu hedef değişkeni en yüksek şekilde tahmin edecek şekilde çalışması gerekir.

Gözetimli ve gözetimsiz algoritmaları anlayabilmek için bir örnek vermek faydalı olacaktır. Örneğin Şekil 5.1’de verilen, golf veri setini ele alalım. Burada 14 gözlem ve 3 nitel, 2 nicel olmak üzere 5 değişken bulunmaktadır<sup>2</sup>. Burada en sonda yer alan “play” değişkeni, diğer 4 değişkenin belirlediği hava şartlarının golf oynamaya müsait olup olmadığını gösterir. Bu şekilde hedef değişken kullanarak, verilen bir hava şartı için golf oynamaya müsait olup olmadığını çıkartacak bir algoritma gözetimli öğrenmeye örnektir.

---

<sup>2</sup> Veri URL11’den indirilebilen WEKA veri madenciliği programı içinde yer alan veri dosyasında yer almaktadır, ana menüden “tools” altında “arff viewer” seçilerek bakılabilir.

ARFF-Viewer - C:\Program Files\Weka-3-6\data

File Edit View

weather.arff weather.arff

Relation: weather

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

**Şekil 5.1. WEKA’da alınmış golf veri seti**

Gözetimli ve gözetimsiz öğrenme tekniklerinin farklarını açıkladıktan sonra bir gözetimsiz teknik olan k-means algoritmasının temel özelliklerinden bahsedebiliriz. Öncelikle kümeleme algoritmasının temelinde olan uzaklık/mesafe kavramını anlamak ve nasıl hesaplandığını incelemek faydalı olacaktır.

### 5.1. Uzaklık Fonksiyonları

Temelde kümeleme algoritmaları nesneler arasındaki uzaklıkları hesaplayarak işe başlar. Kümelemeye girecek  $n$  nesne varsa  $n(n-1)/2$  farklı çift için uzaklık hesaplanabilir. Uzaklık bir çeşit benzerlik ölçüsüdür, yani iki

nesne arasında uzaklık fazla ise birbirlerine benzemedikleri sonucu çıkabilir. Çoğu kişi uzaklık denildiğinde sıkça kullanılan düz çizgi “Öklid” mesafesini anlasa da çok farklı uzaklık fonksiyonları da mevcuttur. Temelde bir uzaklık fonksiyonu üç farklı nesne için (a, b, c diyelim) şu özelliklere sahiptir:

- $D(a,b) \geq 0$
- Sadece  $a=b$  koşulu sağlanırsa  $D(a,b)=0$
- $D(a,b)=D(b,a)$
- $D(a,b) \leq D(a,c) + D(c,b)$

Yukarıdaki özelliklere uygun birçok uzaklık fonksiyonu vardır. Birinci özellik uzaklığın eksi değerler almamasını garanti eder. Gerçekten de eğer iki gözlem tıpatıp aynı ise uzaklık “0” olmalıdır, negatif değerler anlamsızdır. Bu da ikinci özellikte belirtilmiştir. Üçüncü özellik ise uzaklığın hesaplanmasında a, b karşılaştırmasının b, a ile aynı sonucu vermesi gerektiğini söyler. Dolayısıyla fonksiyon hesaplama sırasından bağımsız olmalıdır. Son özellik ise üçgen eşitsizliği olarak bilinir, basitçe açıklamak gerekirse İstanbul-Ankara mesafesi artı Ankara-Mersin mesafesi, İstanbul-Mersin mesafesinden kısa olamaz.

En sık kullanılan uzaklık fonksiyonu Öklid mesafesidir. Kullanılan istatistiki paketler çok farklı opsiyonlar da sunabilir. İlki  $a=(x_1, x_2)$  ikincisi ise  $b=(y_1, y_2)$  olarak verilen iki nokta arasındaki Öklid mesafesi şu şekilde hesaplanır:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Bu formülü  $n$  boyuta, bu sefer iki nokta  $a=(x_1, x_2, x_3 \dots x_n)$  ve  $b=(y_1, y_2, y_3 \dots y_n)$  olacak şekilde genellersek ise:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 \dots + (x_n - y_n)^2}$$

Formülünü elde ederiz. Aslında Öklid mesafesi, daha genel bir uzaklık fonksiyonu olan Minkowski- $p$  uzaklığının  $p=2$  olduğu durumdur. Bu genel uzaklık fonksiyonu ise şu formülle gösterilebilir:

$$d = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p \cdots + |x_n - y_n|^p}$$

Minkowski- $p$  formülünde  $p$ 'nin aldığı üç farklı değer ilginçtir. Bunları  $a=(2,3)$   $b=(4,7)$  noktaları üzerinden gösterelim:

- $p=1$ : Şehir blok mesafesi (ya da “Manhattan<sup>3</sup>” mesafesi). Bu durumda  $d(a,b)=2+4=6$  olacaktır. Yani kesişen cadde ve sokaklardan oluşan bir şehirde bir yerden bir yere gitmek için sadece yatay ve dikey hareket edilmesi durumunda bu mesafe ortaya çıkar.
- $p=2$ : Öklid mesafesi, iki nokta arasına düz bir çizgi çekilmesi durumunda geçerlidir. Bu durumda  $d(a,b)=\sqrt{4+16}=4,47$  olacaktır.
- $p=\infty$ : Bu durumda “sup” ya da “dominans” ölçüsü olarak bilinir ve koordinatlardan hangisi en çok farkı verirse mesafe odur, bu durumda  $d(a,b)=4$  olur, çünkü “7-3” “3-2”den büyüktür.

## 5.2. Değişken Ölçekleri ve Nitel Değişkenler

Değişkenlerin ölçü birimleri uzaklıkları etkileyecek boyutta olabilir. Daha önceki bölümlerde ele aldığımız Capital 500 listesini ele alalım. Burada ciro değişkeni ile çalışan sayısı değişkenini kullanarak Öklid mesafesi hesaplırsak, sayısal değer olarak ciro çok daha fazla olduğu için, uzaklıkta cironun etkisi doğal olarak daha fazla olacaktır. Bu durumda Bölüm 2.3'te bahsedilen dönüşümleri yapmak algoritmanın sağlıklı çalışması açısından önemli olacaktır. Elbette bir değişkenin ağırlığının daha fazla olması doğal olarak istenebilir, bu durumda sırf o değişken kullanılarak kümeleme yapılabilir. Bir başka sorun da nitel değişkenlerin mesafe hesaplaması için nasıl kullanılacağıdır. Her iki konuyu da Tablo 5.1'de verilen basit bir örnekle

<sup>3</sup> Aslında alt Manhattan bölgesinde diğer bölgelerin aksine caddeler ve sokaklar birbirine dik değildir, bu sokakların Hollandalılar zamanında ineklerin oluşturduğu patikalar üzerine yapıldığı söylenir.

açıklamaya çalışalım. Tablo 5.1 bir bankanın yaptığı müşteri segmentasyon çalışmasında topladığı müşteri verilerinin bir kısmını içeriyor. Veride beş müşterinin yaş, aylık gelir ve cinsiyet olmak üzere üç değişkenden aldığı değerler görülüyor. Tablo 5.2 ise yaş ve aylık gelir üzerinden (standardizasyon olmadan) hesaplanan Öklid mesafelerini veriyor.

**Tablo 5.1. Banka müşteri verisi**

Müşteri	Yaş	Aylık Gelir	Cinsiyet
A	50	2100	Kadın
B	20	2000	Kadın
C	30	5000	Erkek
D	20	5000	Kadın
E	20	1850	Erkek

**Tablo 5.2. Banka müşteri verisi yaş ve gelir kullanılarak hesaplanan uzaklık**

	A	B	C	D	E
A		104,403	2900,069	2900,155	251,794
B	104,403		3000,017	3000,000	150,000
C	2900,069	3000,017		10,000	3150,016
D	2900,155	3000,000	10,000		3150,000
E	251,794	150,000	3150,016	3150,000	

Öncelikle standardizasyonun etkisini görmek için cinsiyet değişkenini kullanmadan sadece yaş ve gelir üzerinden müşteriler arasındaki mesafeleri hesaplayalım. Örneğin  $D(A,B) = \sqrt{10900} = 104$  olacaktır.  $D(B,E)$  ise  $\sqrt{22500} = 150$  olacaktır. Yani 20 yaşında 2000 TL kazanan bir müşteri, 50 yaşında 2100 TL kazanan bir müşteriye, 20 yaşında 1850 TL kazanan bir müşteriden daha çok benzemektedir! Burada aylık gelir değişkeninin uzaklıkta, yaş değişkeninden çok daha fazla etkili olduğu görülür. Bu şekilde hesaplanan uzaklık Tablo 5.2’de verilmiştir. Bu sorunu düzeltmek üzere daha önce de bahsi geçen min-max ya da Z skor standardizasyonu uygulanabilir. Hangisinin uygulanması gerektiği metoda göre değişir, yapay sinir ağları gibi 0-1 arası girdi gerektiren uygulamalarda min-max tercih edilebilir. Min-max’te 0-1 arası değişim aralığı olacağı, z-skorda ise -3,+3 arası

değişim aralığı olacağı bilinmelidir. Her iki tür standartlaşmış halleriyle Tablo 5.1'deki değerler Tablo 5.3 ve Tablo 5.4'te verilmektedir. Ayrıca farklı yöntemle hesaplanmış Öklid mesafeleri de Tablo 5.5 ve 5.6'da verilmektedir.

**Tablo 5.3. Banka müşteri verisi, min-max dönüşümü**

Müşteri	Yaş	Min-Max	Aylık Gelir	Min-Max
A	50	1,00	2100	0,08
B	20	0,00	2000	0,05
C	30	0,33	5000	1,00
D	20	0,00	5000	1,00
E	20	0,00	1850	0,00

**Tablo 5.4. Banka müşteri verisi Z-skor dönüşümü**

Müşteri	Yaş	Z-Skor	Aylık Gelir	Z-Skor
A	50	1,69	2100	-0,66
B	20	-0,61	2000	-0,72
C	30	0,15	5000	1,09
D	20	-0,61	5000	1,09
E	20	-0,61	1850	-0,81

**Tablo 5.5. Banka müşteri verisi yaş ve gelir kullanılarak hesaplanan uzaklık, min-max dönüşümünden sonra**

	A	B	C	D	E
A		1,001	1,137	1,359	1,003
B	1,001		1,009	,952	,048
C	1,137	1,009		,333	1,054
D	1,359	,952	,333		1,000
E	1,003	,048	1,054	1,000	

**Tablo 5.6. Banka müşteri verisi yaş ve gelir kullanılarak hesaplanan uzaklık, Z-skor dönüşümünden sonra**

	A	B	C	D	E
A		2,302	2,329	2,892	2,306
B	2,302		1,969	1,813	,091
C	2,329	1,969		,767	2,052
D	2,892	1,813	,767		1,904
E	2,306	,091	2,052	1,904	

Standardizasyon sonrası hem min-max hem de Z-skor ile A-B arasının B-E arasına göre oldukça büyük olduğunu gözlemliyoruz. Çünkü B ile E müşteriler aynı yaşta ve 150TL gelir farkına sahipler, dolayısıyla çok yakın müşteriler. Ancak A ile B daha az gelir farkına sahip olmalarına rağmen (50 TL kadar) 30 yaş farkına sahipler. Tablo 5.5 ve 5.6 ya bakıldığında beklendiği gibi z-skorların değişim aralığının daha fazla olması sebebi ile uzaklıkların da değişim aralığının daha fazla olduğudur. Bu fark cinsiyet değişkenini de uzaklık hesabına eklediğimizde daha belirgin ortaya çıkar. Cinsiyet üzerinde uzaklık hesaplama için Öklid mesafe fonksiyonunu kullanırız ancak sadece 1-0 değeri alan “farklı(a,b)” isimli bir fonksiyon kullanarak iki nitel gözlemin farkını sayısallaştırırız. En basit yöntem, bu örnekte, cinsiyet her iki müşteride de aynı ise “0”, farklı ise “1” değerini atamaktır. Min-max standartlaşması ile kullanıldığında oldukça anlamlı olacaktır çünkü aynı min-max’te olduğu gibi en fazla mesafe 1 en az ise 0 olacaktır. Tablo 5.7 bu şekilde hesaplanan uzaklıkları vermektedir. Görüldüğü gibi C ve E olmayan mesafeler aynı kalmış, C-E mesafesi aynı kalmış diğerleri değişmiştir. İlginç olan bu sefer A-B, B-E karşılaştırmasında neredeyse eşit uzaklık çıkmasıdır, aslında B-E arası ( $\approx 1,0011$ ) bu sefer A-B’den ( $\approx 1,0005$ ) çok az farkla daha büyüktür, bu da cinsiyet değişkeninin eklenmesiyle olmuştur.

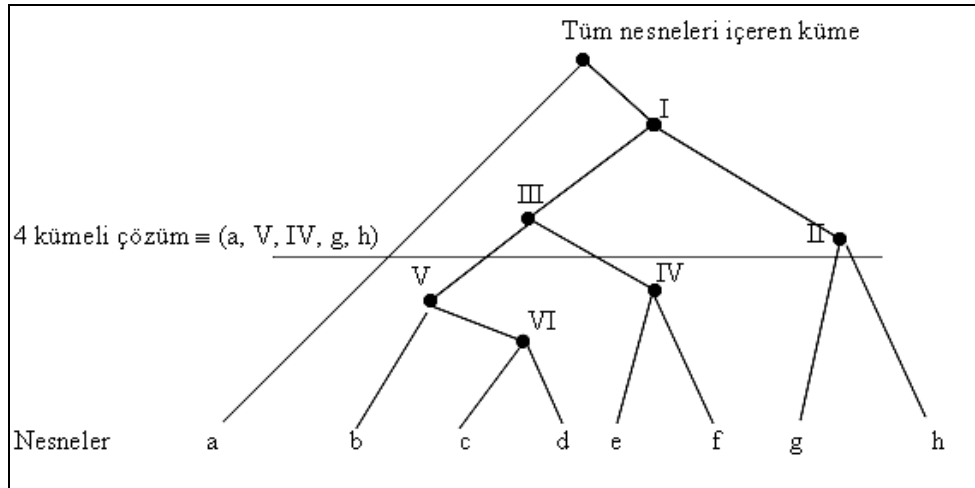


**Tablo 5.7. Banka müşteri verisi yaş ve gelir kullanılarak hesaplanan uzaklık, min-max dönüşümünden sonra ve cinsiyet eklenerek**

	A	B	C	D	E
A	0,000	1,001	1,137	1,359	1,416
B	1,001	0,000	1,009	0,952	1,001
C	1,514	1,421	0,000	1,054	1,054
D	1,359	0,952	0,333	0,000	1,414
E	1,003	0,048	1,054	1,000	0,000

### 5.3. Sıradüzensel Yöntemler

Aslında  $k$ -means algoritması bir sıradüzensel metot değildir ancak bu tür metotlardan da bahsetmek, kümeleme algoritmalarının genel mantığını anlamak açısından önemli olabilir. Sıradüzensel (hierarchical) yöntemlerde iç içe geçmiş bir dizi kümeleme sonucu ortaya çıkar. Az küme sayısına sahip çözümler, çok küme sayısına sahip çözümlerdeki bazı kümelerin (tek nesne de küme olabilir) birleştirilmesi sonucu ortaya çıkar. Bu “birleşme” (ya da ayrılma) çizelgesi ise bir ağaç yapısı içinde gösterilebilir. Bu yapıya bir örnek Şekil 5.2’de verilmiştir. Bu yapıyı herhangi bir noktada keserek  $k$  kümelikli bir çözüm bulabiliriz. Örneğin dört kümelikli çözüm şekilde gösterilmiştir.



**Şekil 5.2. Kümeleme çıktısı sıradüzensel yapıya bir örnek**

Sıradüzensel algoritmalar tüm nesnelerin tek başına oluşturdukları kümeler ile başlayıp en yakın mesafedeki nesneleri birleştirir. Bu ilk adım seçilen opsiyondan bağımsız olarak tüm algoritmalarda aynıdır. Şekilde görülen sekiz nesnede birbirine en yakın olan nesneler c-d olduğu için ilk birleşen de c-d ikilisi olmuştur. Bu durumda yedi kümeli çözüm a, b, VI (yani a ile b'nin ortak kümesi), e, f, g, ve h olabilir. İkinci adımda birleşmiş kümeler ile diğer nesneler ya da kümeler arasındaki mesafeyi tanımlamak için bir opsiyon seçmek gerekir. Genelde bu opsiyon istatistikî paketlerde üç adettir:

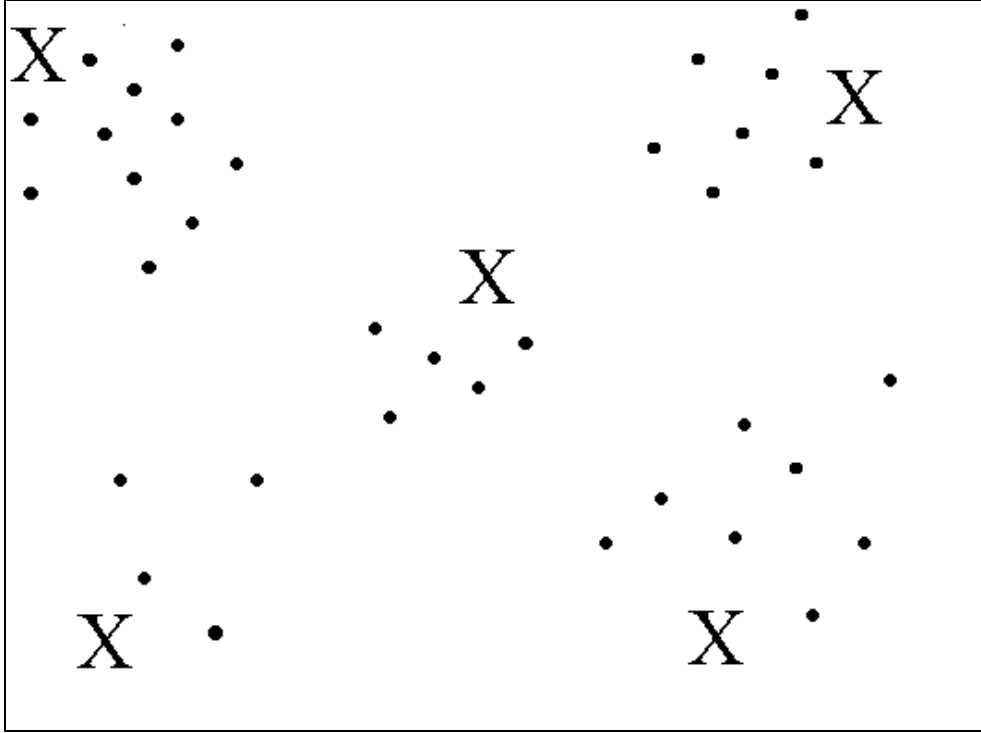
- İki küme arasındaki mesafe bu kümelerin en yakın elemanları arasındaki mesafeye eşittir (single linkage). Yani c,d birleşmesinde oluşan VI kümesi ile b nesnesi arasındaki mesafe b'ye daha yakın olan nesneden hesaplanır.
- İki küme arasındaki mesafe nesneler arasındaki en uzak mesafe temel alınarak hesaplanır (complete linkage). Bu durumda c,d birleşmesinde oluşan VI kümesi ile b nesnesi arasındaki mesafe b'ye daha uzak olan nesneden hesaplanır.
- İki küme arasındaki mesafe nesneler arasındaki ortalama mesafe temel alınarak hesaplanır (average linkage). Bu durumda c,d birleşmesinde oluşan VI kümesi ile b nesnesi arasındaki mesafe b'ye c'nin uzaklığı ile d'nin uzaklığının ortalaması alınarak hesaplanır.

Herhangi bir opsiyonla ikinci aşamada VI kümesine en yakın nesnenin b olduğunu düşünürsek ikinci birleşme VI ile c arasında olur ve V kümesi ortaya çıkar. Bu durumda altı kümeli çözüm elde edilmiş olur. Algoritma bu şekilde tüm nesneleri birleştirerek tek küme haline gelene kadar devam eder.

#### 5.4. K-Means Algoritması

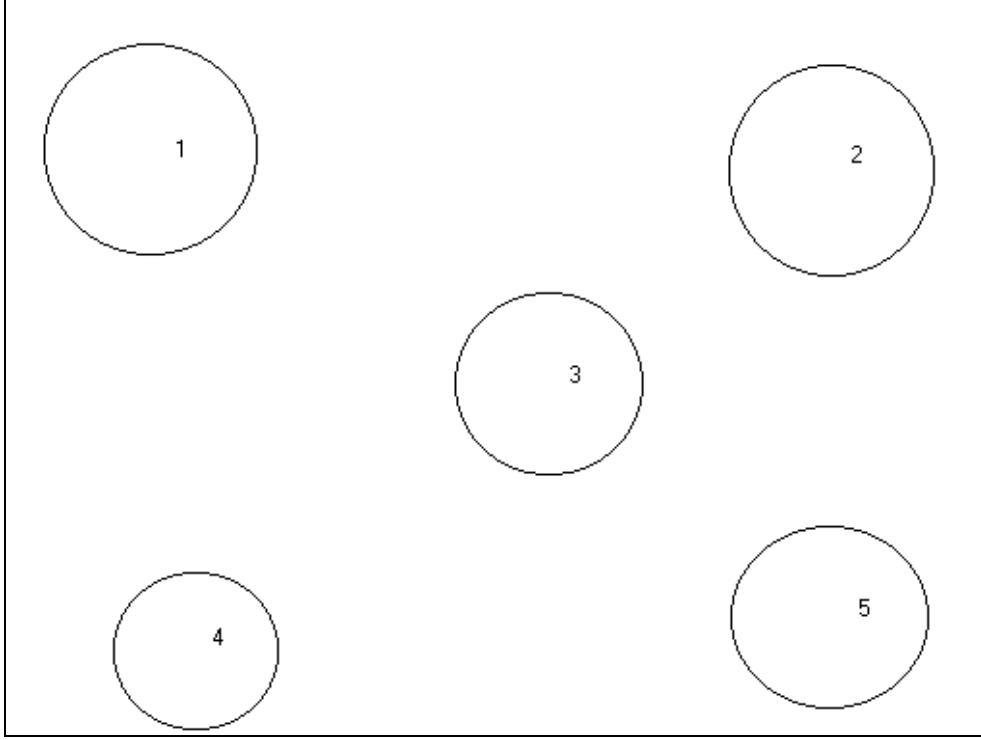
K-means algoritması sıradüzensel bir yapı sağlamaz, sadece istenen sayıda küme üretir. Kümeler birbiriyle ayrıdır yani bir nesne ancak bir kümeye konulabilir. K-means algoritmasını  $k$  sayısını değiştirerek çeşitli kümeleme sonuçları elde edebiliriz. Bu şekilde bu sonuçları karşılaştırarak optimal küme sayısını belirleyebiliriz (5.5'te bahsedilecek). Metodun başlaması için ilk olarak  $k$  merkez (centroid) belirlenmesi gerekir. Bu işlem yazılım tarafından otomatik olarak gerçekleştirilebilir ya da kullanıcı bu başlangıç

merkezini yazılıma bir dosyada verebilir. Şekil 5.3'te bu şekilde belirlenmiş beş merkez görülmektedir. Merkezlerin birbirine çok yakın olmaması mümkün olduğu kadar dağınık olması algoritmanın işleyişi açısından idealdir.



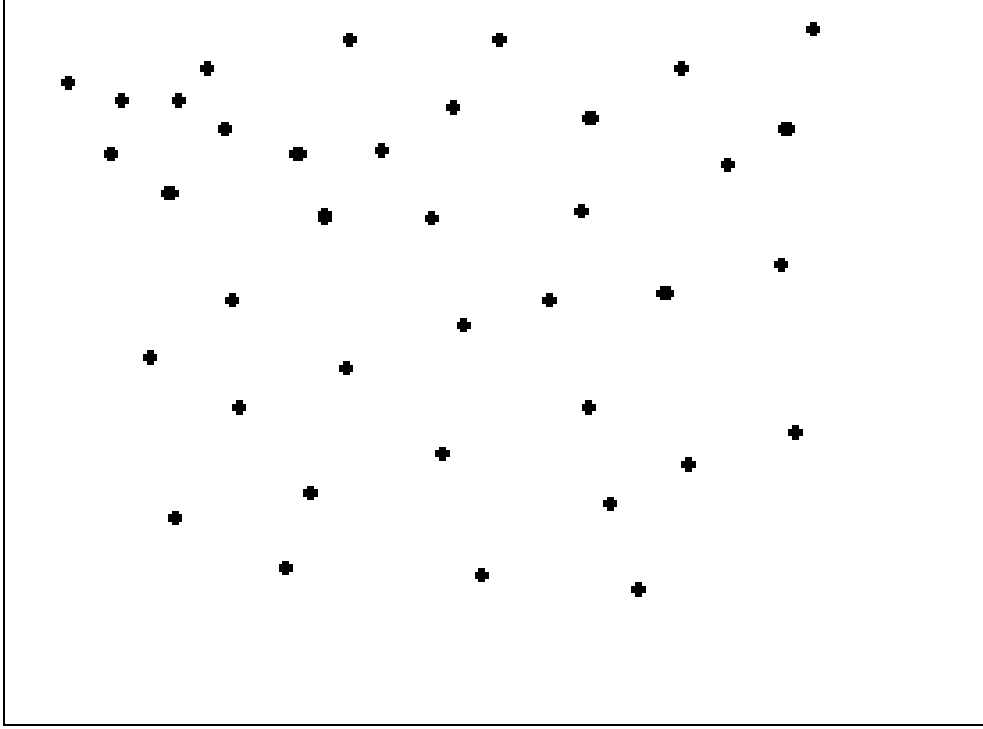
**Şekil 5.3. İki boyutlu bir yapıda belirlenmiş beş merkez**

Bu ilk adımdaki belirlenen merkezler ile diğer noktalar arasındaki mesafeler ikinci adımda hesaplanır. Bu mesafe genelde Öklid mesafesidir ama yazılım uygulamaları farklı opsiyonlar da verebilir. Her nesne en yakında olduğu merkeze atanır ve algoritmanın ikinci adımı da bitmiş olur. Bu ilk etapta bir kümeleme oluşmuştur. Oluşan kümeleri Şekil 5.4'te olduğu gibi gösterebiliriz.



**Şekil 5.4. İki boyutlu bir yapıda belirlenmiş ilk beş küme**

Algoritma şimdi de bu oluşan kümelerin merkezlerini hesaplar. Bu merkezler ilk girilen merkezlerle aynı ya da çok fazla değişim göstermemişse algoritma sonlanır. Bu sonlanma kriteri de yazılımlarda bir seçenek olarak bulunur. Şekil 5.3'te aslında doğal kümelenmenin bulunduğu söylenebilir. Yani iki-boyutta gözle görülen bir öbeklenme ya da gruplanma mevcuttur ve gözle görülen doğal küme sayısı da beştir. Daha gerçekçi bir veri seti Şekil 5.5'teki gibi olabilir. Burada gözle görülebilen bir küme yapısı yoktur. Çok boyutlu veride gözle bu yapıyı da keşfetmek zaten mümkün olmayacaktır. Bu durumda optimal  $k$  seviyesini belirlemek için bazı sayısal yöntemlere başvurulması uygun olacaktır. Alt bölüm 5.5 bu tür yöntemlerden bahsedecektir.



**Şekil 5.5. İki boyutlu bir yapıda homojen dağılmış nesneler**

### 5.5. Küme Sayısının ( $k$ ) Belirlenmesi

Küme sayısının belirlenmesi belki de bu algoritmanın veri madenciliği uygulamalarında kullanılmasında en zor kısımdır. Kesin ve her zaman çalışan bir formül olmamaklar beraber birçok farklı yöntem uygulanabilir. Ayrıca çıktının yorumlanabilir olması da matematiksel yöntemlerin verdiği değerler kadar önemlidir. Değişik  $k$  sayılarını karşılaştırmak için dört farklı metottan bahsedeceğiz:

- Küme sayısı arttıkça ardışık çözümlerin birbiriyle uyum oranı.
- Gruplar arası kareler toplamı ve gruplar içi kareler toplamından faydalanarak hesaplanan Pseudo-F istatistiği.

- Kareler toplamı ile hesaplanan r-kare değerleri.
- Çözümlerde en az eleman içeren küme ile en çok eleman içeren kümelerin içerdiği eleman sayısının toplama oranı ve birbirlerine oranı.

Her istatistiki modelde olduğu gibi kümeleme analizinde de daha basit modeller tercih sebebidir. Dolayısıyla küme sayısını arttırmak tercih edilen bir durum olmamalıdır. Örneğin biri dört biri beş kümeli iki çözümü karşılaştırmak için bu çözümlerin gerçekte ne kadar farklı olduğu sorusunu sorabiliriz. 200 nesneli bir analizde dört kümeli çözümden beş kümeliye geçerken sadece bir nesne tek başına bir küme oluştursa ve diğer tüm nesneler aynı kümelerinde kalsa iki çözüm arasında oldukça fazla benzerlik olduğunu söyleyebiliriz. Belki kabaca benzerlik oranının 199/200 (199 nesne aynı kümelerinde kaldığı için) olduğu söylenebilir. Daha formel bir ölçü “uyuşma oranı” (agreement rate ya da RAND index) adı altında Rand tarafından tanımlanmıştır (Rand, 1971), daha sonra rastlantısal dağılım için düzeltilmiş bir versiyonu da ortaya çıkmıştır (Arabie & Hubert, 1985). Bu indeks her olası nesne çifti için aynı kümede olup olmadıklarına bakmaktadır. Bir çift iki farklı kümeleme sonucunda aynı kümede gruplanmışlar ise uyuşma oranı  $1/(\text{toplam nesne çifti})$  oranında artmaktadır. İndeks %100 uyum gösterirse kümeleme sonuçları tamamen aynıdır. %100’e yakın sonuçlar ise kümeleme sonuçlarının çok da değişmediği izlenimini verir. Bu yöntem farklı algoritmalarından elde edilen sonuçları karşılaştırmakta kullanılabileceği gibi, uygun  $k$  sayısının tespitinde de kullanılabilir. Örneğin Akküçük’ün (2011b, URL12) verdiği bir örnekte beş kümeden altı kümeye geçerken uyuşma oranının %94 olduğu görülmüştür. Bu durumda gerçekten de beş kümeden altı kümeye geçmenin çok anlamı yoktur.

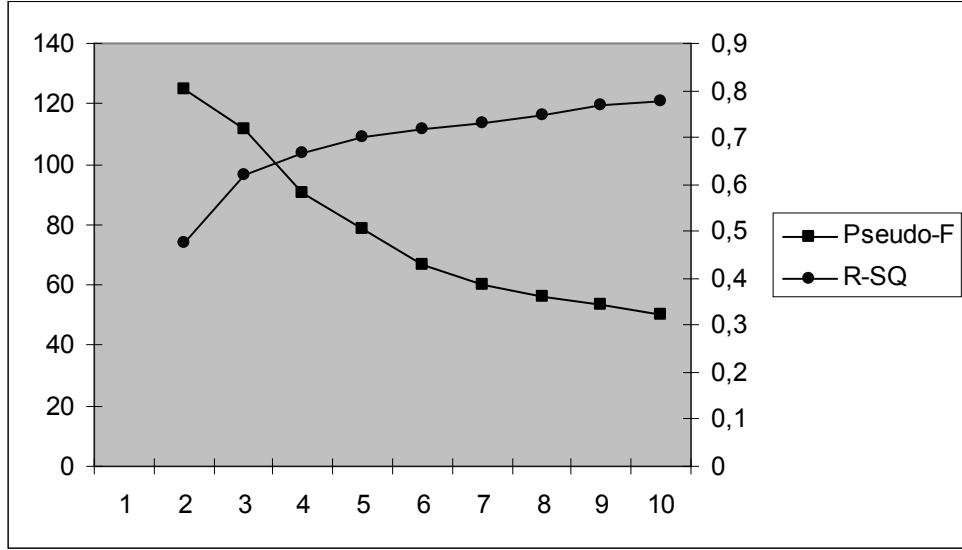
Kareler toplamı ise küme sayısının tespiti konusunda faydalanılabilecek yöntemlerden biridir. Bir küme geometrik merkezi (yani o kümedeki her değişkenin ortalamaları), o küme elemanlarının farklarının karelerinin toplamı “grup içi kareler toplamı - GIKT”ı verir. Tüm nesnelerin genel geometrik merkezden farklarının karelerinin toplamı ise (yani tüm nesneler için hesaplanan değişken ortalamaları) “toplam kareler toplamı – TKT” verir. İkisinin arasındaki fark ise “gruplar arası kareler toplamı – GAKT” olacaktır. GIKT’ı sıfıra indirmek için her nesneyi bir kümeye koymak yetecektir. Bu durumda GAKT da en üst seviyeye çıkacaktır. Ancak amaç tabii tek kümeli ve  $n$  (nesne sayısı kadar) kümeli çözümlerden çok mantıklı bir çö-

züm bulmaktır. Bu durumda Pseudo-F (Calinski & Harabasz, 1974) ve r-kare ölçüleri aşağıdaki gibi hesaplanabilir.

$$\text{Pseudo F} = \frac{\left[ \sum_{i=1}^n (X_i - \bar{X})^2 - \sum_{k=1}^G \sum_{i=1}^{n_k} (X_i - \bar{X}_k)^2 \right] / (G-1)}{\left[ \sum_{k=1}^G \sum_{i=1}^{n_k} (X_i - \bar{X}_k)^2 \right] / (n-G)}$$

$$\begin{aligned} r^2 &= 1 - \frac{\text{GIKT}}{\text{TKT}} \\ &= 1 - \frac{\sum_{k=1}^G \sum_{i=1}^{n_k} (X_i - \bar{X}_k)^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

Akküçük'ün (2011b) verdiği örnekte bu iki değerin grafiği Şekil 5.6'da olduğu gibi görülmektedir. Pseudo-f ve r-kare için yüksek değerler tercih edilir olsa da bu örnekte her iki değer için bir denge noktası bulmak daha mantıklı görünmektedir. Burada dört ya da beş kümeden sonra değerlerin arası açılmaktadır.



**Şekil 5.6. Pseudo-F ve r-kare değerleri (Akküçük, 2011b)**

En son olarak kümelerdeki eleman sayısına bakılarak da bir karar verilebilir. Bazı kümelerde çok fazla eleman bazılarında da çok az eleman varsa bu kümeleme sonuçlarını yorumlamak zor olacaktır. Ayrıca en fazla elemanlı kümenin eleman sayısının, en az elemanlı kümenin eleman sayısına bölünmesi ile ortaya çıkan sayıya bakılarak da bir karar verilebilir. Akküçük (2011b) örneğinde beş kümeden altı kümeye geçerken bu oran 1,43'ten 31'e çıkmaktadır. Yani bu atlama esnasında küme yapıları çok az değişmekte sadece ayrılan bazı nesneler (sayısı bir yada iki olan) kendi başlarına bir küme oluşturmaktadırlar. Bu da optimal küme sayısının diğer kriterlere de bakarak beş olması gerektiğini destekler.





## 6. BÖLÜM:

### ***K* EN YAKIN KOMŞU ALGORİTMASI**

En basit haliyle  $k$  en yakın komşu algoritması bir kümeleme algoritması olarak kullanılabilir. Her nesne kendisine (belirli bir tanımlanmış benzerlik ölçüsüne göre) en yakın  $k$  nesne ile aynı kümeye konur. Böylece nesne sayısı kadar küme ortaya çıkar. Elbette burada ortaya çıkan küme yapısı  $k$ -means ya da sıradüzensel kümeleme ile ortaya çıkan yapıdan farklıdır, zira kümeler birbiriyle örtüşür, yani aynı nesne birden fazla kümede ortaya çıkar. Veri madenciliği anlamında ise yöntem bir gözetimli öğrenme yöntemi olarak kullanılır.  $K$  en yakın komşu algoritması yeni bir gözlemi her zaman saklanan bir veritabanındaki gözlemler ile karşılaştırarak en yakın  $k$  gözlemi belirler. Bu “yakın” gözlemlerin sınıflamalarına göre yeni gözlemi de bir sınıfa koymaya çalışır. Yakın gözlemlerin hepsi aynı sınıfta ise sorun çok daha basittir ama farklı sınıflamalar var ise aradaki mesafeyi de kullanarak bir çeşit ağırlıklı ortalama alınması gerekebilir. Ayrıca  $k$  katsayısının nasıl seçileceği de bu yöntemle ilgili pratik sorunlardan bir tanesidir.

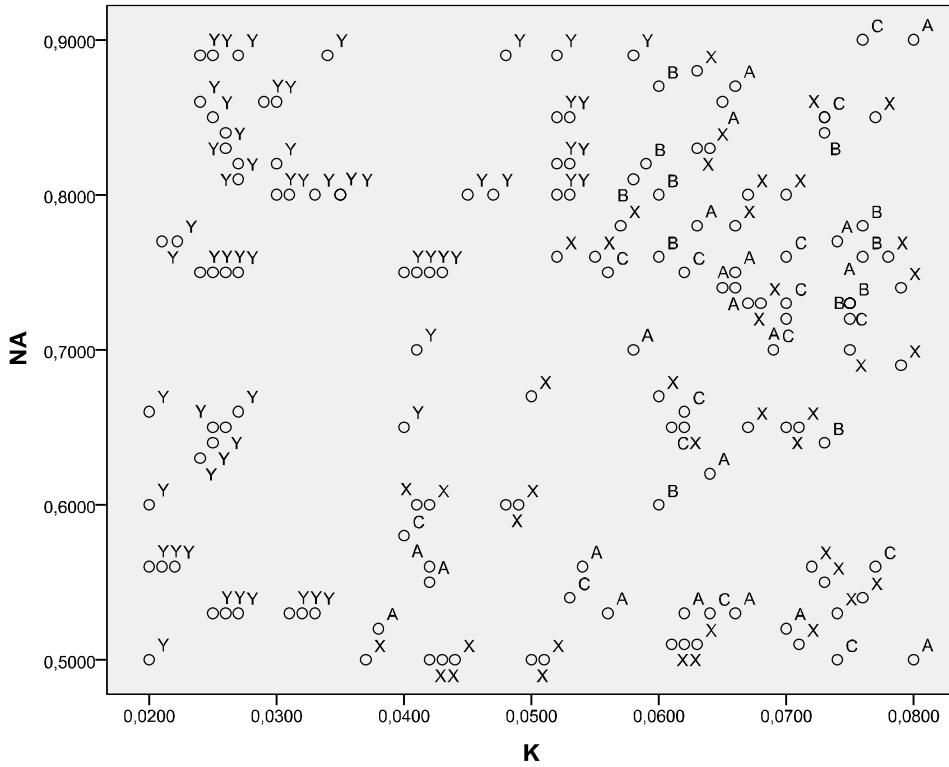
#### **6.1. En Yakın Komşu Algoritması**

$K$  en yakın komşu algoritması bir örnek tabanlı öğrenme (instance based learning) biçimidir, yani yeni gözlemler sınıflanırken hafızada tutulan bir veri seti kullanılır. Bu hafızadaki veri setinde hedef değişkenin alacağı değerler bellidir. Yeni bir gözlemi sınıflamak gerektiğinde algoritmanın en temel şekli şu şekilde işler:

- Öncelikle  $k$  sayısı belirlenmelidir, bu  $k$  sayısı yeni gözlemi sınıflandırmak için kaç adet en yakın gözlemin kullanılacağını belirler.
- Yeni gözleme belirlenen değişkenler üzerinde hesaplanan (elbette gerekli standardizasyon işlemlerinde sonra) mesafeye göre en yakın  $k$  kayıt incelenir. Burada kullanılan mesafe Öklid ya da tercih edilen başka bir mesafe türü olabilir.
- Basit bir oylama yöntemiyle yeni gözlem gereken sınıfa ayrılır. Örneğin  $k$  üç olarak seçilmiş ise kayıtlardaki en yakın iki ya da daha fazla kayıt hangi sınıfa aitse yeni gözlem de o şekilde sınıflanır.

Yöntemi anlamak için çeşitli hastaların sodyum (Na) ve potasyum (K) değerleri ve reçetelenen ilaç tipini gösteren (A, B, C, X, Y olmak üzere) Şekil 6.1'deki veriyi kullanalım. Burada Na değişkeni minimum 0,5 maksimum 0,9 değerlerini almaktadır, K değişkeni ise minimum 0,02 maksimum 0,08 değerlerini almaktadır. İki adet yeni gözlemin hangi ilaçları alması gerektiğine karar vermemiz gerektiğini düşünelim. Bu iki yeni gözlem ve Na, K değerleri şu şekilde verilmiştir:

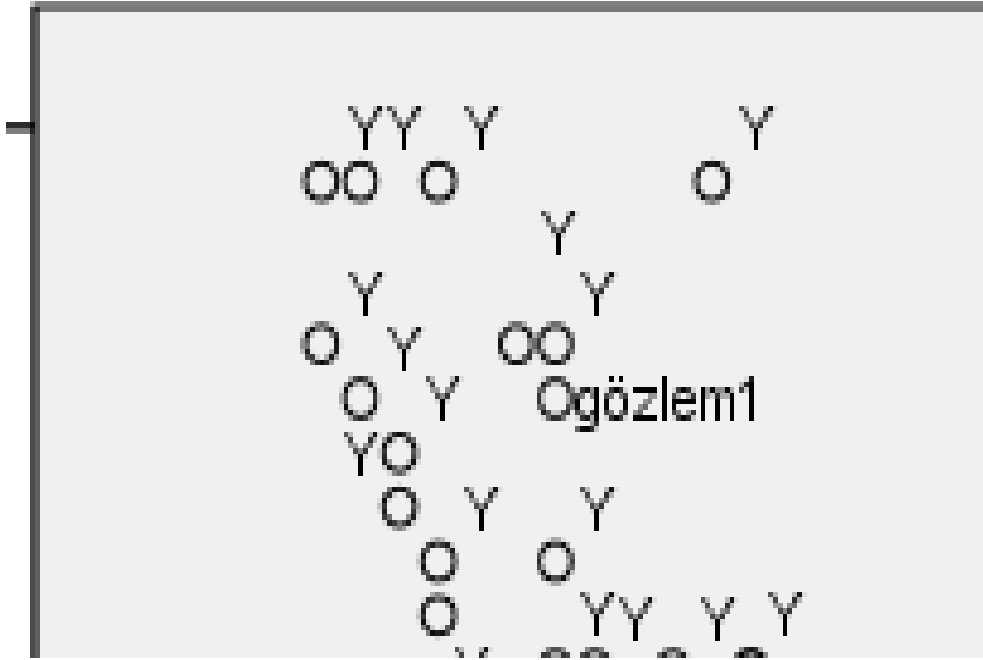
- Gözlem 1: Na=0,85 ve K=0,030
- Gözlem 1 min-max normalizasyonu: Na=0,875 ve K=0,167
- Gözlem 2: Na=0,57 ve K=0,075
- Gözlem 2 min-max normalizasyonu: Na=0,175 ve K=0,917



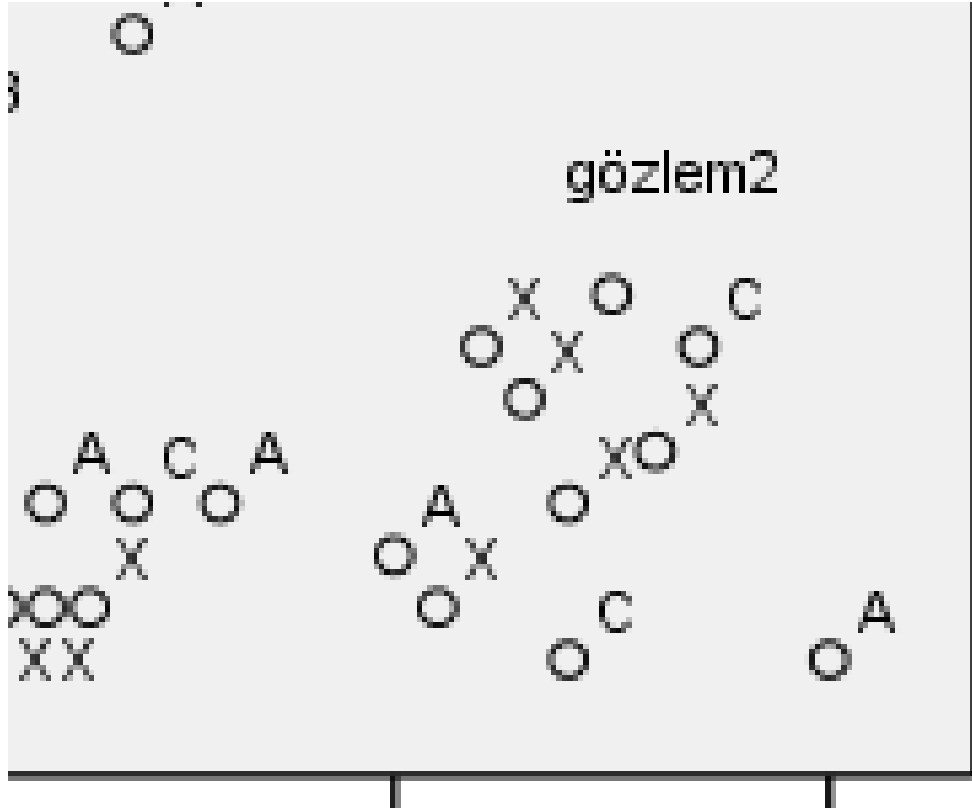
**Şekil 6.1. Na ve K değerleri, reçeteye yazılan ilaç (A, B, C, X ve Y ile gösterilmiştir)<sup>4</sup>**

<sup>4</sup> Daha önce SPSS Clementine ismiyle anılan yazılım içinde yer alan bir veri setinden esinlenilmiştir.

Gözlem 1 ve Gözlem 2'nin yerleştikleri yerler haritaya yerleştirilir ve yaklaştırılırsa Şekil 6.2 ve 6.3 elde edilir. Burada açıkça görülen Gözlem 1'in üç Y oyu aldığı, Gözlem 2'nin ise iki X bir C oyu aldığıdır. Dolayısıyla basit oylamada Gözlem 1 Y ilacı almaya, Gözlem 2 ise X ilacı almaya göre sınıflandırılır.



Şekil 6.2. Gözlem 1 yaklaşmış



Şekil 6.3. Gözlem 2 yaklaşmış

## 6.2. Ağırlıklandırma Fonksiyonları

Daha önceki bölümde basit oylama yoluyla 2. Gözlemi X olarak sınıflamıştık. Ancak Basit oylama her zaman kolay uygulanamayabilir. Örneğin her üç yakın komşu da farklı ilaç olabilir. Ya da sayısal üstünlüğü olmasa da C noktası yeni gözleme diğerlerinde çok daha yakın olabilir. Bu durumun üstesinden gelmek için oylar ağırlıklandırılabilir. Ağırlıklandırma mesafenin karesinin tersine göre yapılır. Örneğimizde C'nin aldığı oy ve iki adet X noktasının aldığı oy şu şekilde hesaplanabilir:

$$oy(C) = \frac{1}{d(yeni, C)^2}$$

$$oy(X) = \frac{1}{d(yeni, x_1)^2} + \frac{1}{d(yeni, x_2)^2}$$

Bu durumda C'nin sayısı bir de olsa ağırlıklandırma sonucu X'i geçme imkanı vardır. Bu durumda Gözlem 2'nin en yakın olduğu üç noktayı ve bu noktaların koordinat değerlerini görmek gerekir:

- Nokta C: Na=0,56 ve K=0,077
- Nokta C normalize: Na=0,15 ve K=0,95
- Nokta X<sub>1</sub>: Na=0,56 ve K=0,072
- Nokta X<sub>1</sub> normalize: Na=0,15 ve K=0,87
- Nokta X<sub>2</sub>: Na=0,55 ve K=0,073
- Nokta X<sub>2</sub> normalize: Na=0,125 ve K=0,88

Normalize değerler kullanarak her üç mesafe de hesaplanırsa:

$$d(yeni, C) = \sqrt{(0,917 - 0,95)^2 + (0,175 - 0,15)^2} = 0,0414$$

$$d(yeni, X_1) = \sqrt{(0,917 - 0,87)^2 + (0,175 - 0,15)^2} = 0,0532$$

$$d(yeni, X_2) = \sqrt{(0,917 - 0,88)^2 + (0,175 - 0,125)^2} = 0,0622$$

Bu durumda C ve X'in aldığı oylar şu şekilde hesaplanabilir:

$$oy(c) = \frac{1}{d^2(yeni, C)} = 583$$

$$oy(x) = \frac{1}{d^2(yeni, X_1)} + \frac{1}{d^2(yeni, X_2)} = 358 + 258 = 611$$

Görüldüğü gibi az farkla da olsa X kazanmaktadır ama C biraz daha yakın olsaydı belki de basit oy sayısı fazla olmasına rağmen X kazanamayacaktı. Gerek basit oylama gerekse ağırlıklı oylama ile yaptığımız sınıflamaya duyduğumuz güveni de ifade edebiliriz. Örneğin basit oylamada 2/3 yani % 67 derecesinde güvenle X sınıflamasını yaptığımızı söyleyebiliriz. Ağırlıklı ise  $611/(583+611) = \% 51$  derecesinde güvenle X sınıflaması yaptığımızdan bahsedebiliriz.

Ağırlıklı oylamanın bir başka kullanımı da yeni gözleme ait başka bir değişkeni (nicel bir değişken) yakın gözlemlerdeki değerleri kullanarak tahmin etmek olabilir. Yukarıda hesaplanan oy oranlarını ağırlık olarak kabul ederek yeni gözlem için tahmin şu formülle yapılabilir:

$$tahmin = \frac{\sum w_i y_i}{\sum w_i}$$

Bu örnekte yakın gözlemlerde sırasıyla 120, 130 ve 140 tansiyon değerlerinin ölçülmüş olduğunu varsayalım. Yeni gözlem için tansiyon değerini tahmin etmemiz gerekirse şu şekilde hesaplayabiliriz:

$$yeni\ tan\ siyon = \frac{583 \times 120 + 358 \times 130 + 258 \times 140}{1194} = 127$$

### 6.3. K seçimi

Aslında ağırlıklandırma fonksiyonu kullanılacaksa  $k$  seçiminin çok önemi kalmaz, çünkü doğal olarak uzaktaki kayıtların sonuç üzerine etkisi çok az olur. Ancak bu hesaplama açısından pahalı bir çözüm olabilir ve gereksiz yere fazla işlem yapılmasına sebep olabilir.  $K$  değerinin çok küçük seçimi (örneğin en küçük değer olan bir) metodun örnek veriyi ezberlemesine yol açar ve genelleştirme zor olur. Büyük  $k$  ise çok fazla düzleştirmeye yol açar ve veriden öğrenilecek özel bazı örüntülerin saklanmasına sebebiyet verir. Belki de en uygun çözüm veriyi iki gruba ayırıp (eğitim ve doğrulama olmak üzere) farklı  $k$  değerleri için eğitim seti kullanılarak yapılan tahminlerin doğrulama setinde verdiği doğru tahmin yüzdesini en fazla veren  $k$  değerini kullanmak olabilir.

## 7. BÖLÜM:

### KARAR AĞAÇLARI

Karar ağaçları da 5. Bölüm’de anlatılan gözetimli öğrenme yöntemlerinden biridir. Birkaç farklı algoritmik uygulaması olsa da, yeni bir gözlemi daha önceden belirlenmiş sınıflara ayırmak için bir karar ağacı kullanır. Bu karar ağacı eğitim verisi ile çalıştırıldığından en doğru tahminleri yapacak şekilde seçilir. Mükemmel tahmin sağlayan ama çok karışık kurallar pratikte uygun olmayabilir, dolayısıyla eğitim verisinde doğru tahmin oranı ile model karmaşıklığını dengeleyecek bir çözüm bulunmalıdır. Bir karar ağacı kök düğümden başlar (en üstte bulunur) ve aşağı doğru ilerleyen dallardan oluşur. Dallar diğer düğümlere ulaşır ve bazı düğümler uç düğümdür (yani kendisinden sonra gelen başka düğüm yoktur). Her düğümden bir karar verilmesi gerekir. Bu nitel bir değişkenin çeşitli seviyeleri olabileceği gibi nicel bir değişken söz konusuysa da eşik değerlerden büyük ya da küçük olmaya göre hangi dala gidileceğine karar verilebilir. Karar ağaçları bir önceki bölümde anlatılan  $k$  en yakın komşu algoritması gibi bir gözetimli öğrenme yöntemidir. Yani bir hedef değişkenin olması gerekir, zaten karar ağacını oluşturan algoritma da bu hedef değişkeni en iyi tahmin edecek kurallar silsileleri oluşturmaya çalışır.

#### 7.1. Bir Örnek

Daha önce de verdiğimiz golf örneği karar ağaçlarını anlatmak için uygun bir örnek olacaktır. Tablo 7.1 golf verisini oluşturan iki nitel ve iki nicel değişkeni göstermektedir. Her gözlem için ayrıca belirtilen koşulların golf oynamaya uygun olup olmadığı verilmektedir. Bu durumda yeni bir koşul söz konusu olduğunda, örneğin “kapalı, rüzgarlı, 65 derece sıcaklıkta (Fahrenheit cinsinden<sup>5</sup>) ve nem oranı 70 olan bir havada golf oynanabilir mi?” sorusu cevaplanmak istenmektedir. 14 gözlemden oluşan veri setimizde bu özelliklere %100 uyan bir vaka yoktur ama yakın vakalar olabilir. Örneğin havanın kapalı olduğunu göz önüne alırsak veri setinden kapalı hava

---

<sup>5</sup> (Fahrenheit-32)/1,8 Celcius cinsinden sıcaklığı verir



koşullarının golfü uygun olup olmadığını kontrol edebiliriz. Gerçekten de dört kapalı hava koşulu olan günde de golfü uygun koşullar oluşmuştur. Bu durumda basit bir kural yaratarak “hava=kapalı” olma durumunda golfü uygundur diyebiliriz.

**Tablo 7.1. Golf veri seti**

Gün	Hava	Sıcaklık	Nem	Rüzgar	Golfü Uygun
1	Güneşli	85	85	Yok	Hayır
2	Güneşli	80	90	Var	Hayır
3	Kapalı	83	78	Yok	Evet
4	Yağmurlu	70	96	Yok	Evet
5	Yağmurlu	68	80	Yok	Evet
6	Yağmurlu	65	70	Var	Hayır
7	Kapalı	64	65	Var	Evet
8	Güneşli	72	95	Yok	Hayır
9	Güneşli	69	70	Yok	Evet
10	Yağmurlu	75	80	Yok	Evet
11	Güneşli	75	70	Var	Evet
12	Kapalı	72	90	Var	Evet
13	Kapalı	81	75	Yok	Evet
14	Yağmurlu	71	80	Var	Hayır

Bu şekilde diğer değişkenleri de inceleyerek oluşabilecek basit kuralları görmek, bu bölümde anlatacağımız algoritmaların da çalışma biçimini anlamak açısından faydalı olacaktır. Hava değişkenini ele alırsak, sadece bu değişkenin üç seviyesinden yola çıkarak basit bir kural dizisi oluşturmak istersek, her seviye için golfü uygun ve uygun olmayan gün sayılarına bakmak yerinde olur. Tablo 7.2 bu sayıları vermektedir. Bu durumda kapalı havalara “uygun”, güneşli havalara “uygun değil” ve en son olarak yağmurlu havalara “uygun” dersek doğru tahmin oranımız  $4 \times 1 + 5 \times 0,6 + 5 \times 0,6 = 10/14$  yani %71 olacaktır.

**Tablo 7.2. Hava değişkeninin değerleri için golfa uygun olan gün sayıları**

Hava	Gözlem Sayısı	Golfa Uygun Oranı	Golfa Uygun Değil Oranı
Kapalı	4	1	0
Güneşli	5	0,4	0,6
Yağmurlu	5	0,6	0,4

Aynı analizi bu sefer de rüzgar durumu için yapacak olursak Tablo 7.3'teki durumla karşılaşırız. Bu durumda rüzgar olduğu zaman “uygun” ya da “uygun değil”(çünkü eşit sayıda uygun ve uygun olmayan gün vardır), rüzgar olmadığı zaman ise “uygun” dersek doğru tahmin oranımız 9/14 yani % 64 olacaktır.

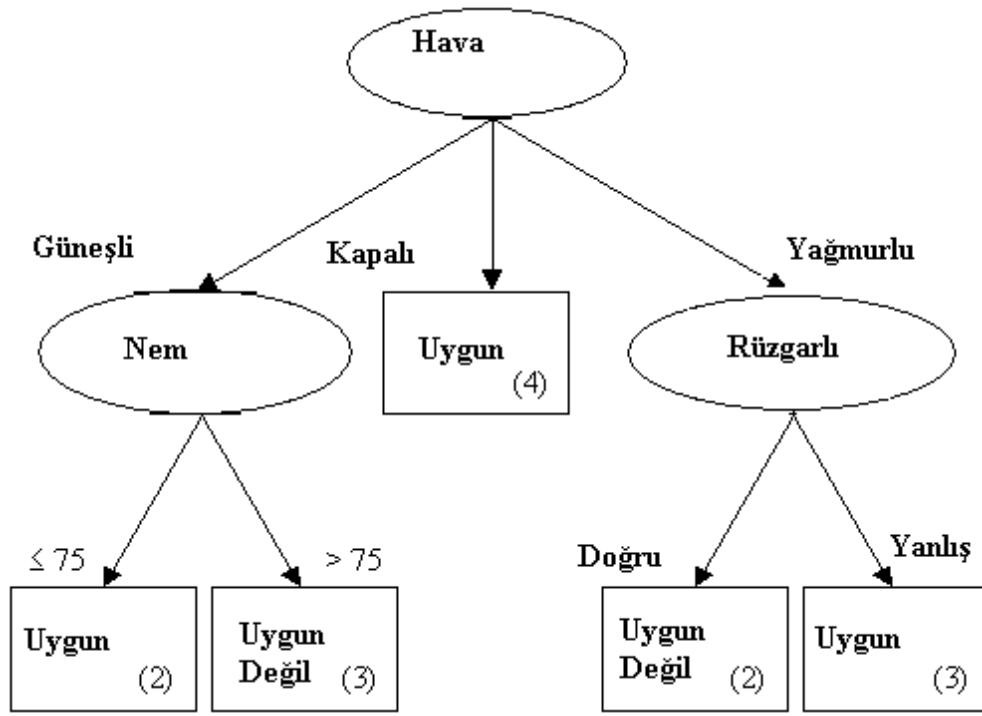
**Tablo 7.3. Rüzgar değişkeninin değerleri için golfa uygun olan gün sayıları**

Rüzgar	Gözlem Sayısı	Golfa Uygun Oranı	Golfa Uygun Değil Oranı
Var	6	0,5	0,5
Yok	8	0,75	0,25

Nicel değişkenlere gelince, bu durumda karar ağacı algoritmaları genelde bir eşik değer tespit ederek bundan fazla ya da eksik olma durumuna göre hangi dala gidileceğine karar verirler. Bu eşik değer elbette doğru tahmin oranını maksimize edecek şekilde yapılır. Veriyi incelediğimizde uygun günlerde ortalama sıcaklığın 73 nemin ise 78,22 olduğu görülür. Aynı şekilde uygun olmayan günlerde ortalama sıcaklık 74,6 nem ise 84'tür. Bu durumda düşük nem ve sıcaklığın golf oynanması için ideal olduğu düşünülebilir.

Elbette tüm bu açınsayıcı analiz teknikleri daha karar ağacı modeline karar vermeden yapılmış olmalıdır ve yazılım kullanılarak bir karar ağacı oluşturulmadan hangi değişkenlerin girdi olarak kullanılacağına karar vermekte kullanılabilir. Bu bölümde anlatacağımız algoritmalarından biri kullanılarak

bir karar ağacı üretildiğinde ise Şekil 7.1'deki gibi bir karar ağacı ortaya çıkar. Bu karar ağacı C4.5 ile üretilmiştir ve doğru tahmin oranı %100'dür. Yuvarlak düğümler değişkenin aldığı değere göre farklı dallara gidilebilecek düğümleri, kare olanlar ise son düğümleri yani hedef değişkenin değerinin tespit edildiği düğümleri göstermektedir.



Şekil 7.1. Golf veri seti için karar ağacı (C4.5)

## 7.2. CART Algoritması

CART algoritması Breiman ve diğerleri tarafından (1984) önerilmiştir. “Classification and Regression Trees” olarak adlandırılan İngilizce ismin baş harflerinden yola çıkarak CART algoritması olarak adlandırılmıştır. CART algoritması kök düğümden başlayarak her düğüm için olası tüm ayırma şekillerini gözden geçirerek bunlardan en iyisini seçer. CART algoritması, yukarıda grafiği verilen C4.5 ile üretilmiş ağacın aksine, her düğümde iki dal üretir. Yani bölünmeler (ya da ayrılmalar) ikilidir, nitel değişken üç seviyeli de olsa üç farklı bölünme şekli de incelenmelidir. Örneğin

yukarıdaki golf veri seti için sadece nitel değişkenler göz önüne alındığında dört farklı bölünme şekli olduğu kolaylıkla anlaşılır. Nicel değişkenler için tüm olasılıklar denenmek istenirse verideki tüm farklı sayılar göz önüne alınmalıdır. Örneğin nem değişkeni için dokuz farklı seviye vardır. Bu durumda sekiz farklı bölünme kuralı getirilebilir. Sıcaklık için ise 11 farklı bölünme kuralı olabilir. Bunların hangilerinin tercih edileceği program tarafından belirlenebilir ya da analizi yapan kişi eşik değerler belirleyerek nicel değişkenleri analiz öncesi gruplara ayırabilir. Nicel değişkenin çok fazla sayıda farklı değeri varsa ikinci yöntem daha anlaşılabilir bir karar ağacı ortaya çıkaracak ve hesapları kolaylaştıracaktır.

Olası farklı bölünme şekillerine karar verildikten sonra algoritma her adayın ne kadar başarılı bir ayırıcı olduğunu hesaplayan bir formül kullanır. Bu performans kriterine  $\Phi$  dersek şu şekilde hesaplanır:

$$\Phi = 2P_L P_R \sum_{j=1}^{\text{sin ifsayisi}} |P(j | t_L) - P(j | t_R)|$$

Burada:

$t_L$  = düğüm t'nin sol alt düğümü

$t_R$  = düğüm t'nin sağ alt düğümü

$$P_L = \frac{t_L \text{ deki kayıt sayısı}}{\text{toplam kayıt sayısı}}$$

$$P_R = \frac{t_R \text{ deki kayıt sayısı}}{\text{toplam kayıt sayısı}}$$

$$P(j | t_L) = \frac{t_L \text{ deki } j \text{ sinifına uyan kayıt sayısı}}{t_L \text{ deki toplam kayıt sayısı}}$$

$$P(j | t_R) = \frac{t_R \text{ deki } j \text{ sinifına uyan kayıt sayısı}}{t_R \text{ deki toplam kayıt sayısı}}$$

Aslında daha önce golf örneği için çok benzer hesaplamalar yapmıştık. Burada aday bölünmelerden biri “rüzgarlı gün” ve “rüzgarsız gün” olabilir.

Bu bölünme için  $\Phi$  değerini hesaplamak istersek aşağıdaki rakamlara ihtiyaç duyarız:

- $t_L$ : Sol alt düğüm rüzgarlı gün olsun
- $t_R$ : Sağ alt düğüm rüzgarsız gün olsun
- $P_L$ : rüzgarlı gün sayısının toplama oranı =  $6/14 = 0,43$
- $P_R$ : rüzgarsız gün sayısının toplama oranı =  $8/14 = 0,57$
- Burada  $2P_L \times P_R$  işlemi yapılırsa  $2(6/14) \times (8/14) = 0,49$  çıkar. Bu işlem olası en yüksek değerini hem sol hem de sağ alt düğümlerin eşit oranda gözlem içermesi ile elde eder. Yani en yüksek değer olan 0,50 ancak her iki alt düğümde de 7 gözlem olması neticesinde olur. Dolayısıyla bölünme kuralının başarısı bir yanda gözlemleri eşit iki parçaya bölmesi ile orantılıdır.
- $P(j | t_L) = \frac{t_L \text{ deki } j \text{ sinifına uyan kayıt sayısı}}{t_L \text{ deki toplam kayıt sayısı}}$ : Bir başka deyişle sol alt düğümde golfa uygun gün bulmanın koşullu olasılığı, ve golfa uygun olmayan gün bulmanın koşullu olasılığı. Bu durumda rüzgarlı günlerde golfa uygun gün sayısı 3 olduğunda bu oran her iki durum için de  $3/6 = 0,50$  olmalıdır. Yani  $P(\text{uygun} | \text{sol düğüm}) = 0,5$ ,  $P(\text{uygun değil} | \text{sol düğüm}) = 0,5$ .
- $P(j | t_R) = \frac{t_R \text{ deki } j \text{ sinifına uyan kayıt sayısı}}{t_R \text{ deki toplam kayıt sayısı}}$ : Bir başka deyişle sağ alt düğümde golfa uygun gün bulmanın koşullu olasılığı, ve golfa uygun olmayan gün bulmanın koşullu olasılığı. Bu durumda rüzgarsız günlerde golfa uygun gün sayısı 6 olduğunda bu oran sırasıyla  $6/8 = 0,75$  ve  $2/8 = 0,25$  olmalıdır. Yani  $P(\text{uygun} | \text{sağ düğüm}) = 0,75$ ,  $P(\text{uygun değil} | \text{sağ düğüm}) = 0,25$ .
- Tüm bu rakamları kullanarak  $\Phi$  hesaplanmak istenirse, formülün toplama işaretiinden sonrasında bulunan  $|0,5 - 0,75| + |0,5 - 0,25|$  işlemi yapılmalıdır, bu da 0,5 çıkacaktır. Bu işlemin teorik en fazla değeri sınıf sayısı kadardır. Yani bizim durumumuzda iki sınıf olduğuna göre teorik en fazla değer 2'dir. Bu değer ise her iki alt düğümün "saf" olması durumunda hesaplanacaktır. Saf düğümde sadece tek bir cins hedef değişken bulunur. Örneğin sol düğümde sadece "uy-

gun” sağ düğümde ise sadece “uygun değil” olsaydı bu şekilde gerçekleşirdi (yani  $|1-0|+|0-1|$  olarak hesaplanırdı).

- En son olarak  $\Phi$  değeri  $0,49 \times 0,5 = 0,2449$  olur.

Bu şekilde tüm olası ayırmalar için  $\Phi$  değeri hesaplanarak en yüksek olan kök düğüm olarak seçilecektir. Bunu yapmadan diğer bölünme olasılıklarına göz atmalıyız. Burada “hava” değişkeni üç seviyeli olduğundan, “rüzgar” değişkeni de iki seviyeli olduğundan dört farklı bölünme vardır. Nicel değişkenleri de eşik değeri belirleyerek iki seviyeli hale getirmek işimizi kolaylaştıracaktır. Nem değişkeni için 78’den küçük eşit ve 78’den büyük kuralı 6-8 bir bölünme getirmektedir. Bu da 80’i seçmekten daha uygun görünmektedir. Sıcaklık için ise 72’den küçük eşit ve 72’den büyük kuralı benzer şekilde uygun görünmektedir. Bu durumda algoritmanın ilk adımında değerlendirilmesi gereken altı farklı bölünme şeklini Tablo 7.4’te özetleyebiliriz.

**Tablo 7.4. Olası bölünme adayları**

Bölünme	Sol Alt Düğüm	Sağ Alt Düğüm
1	Rüzgarlı=doğru	Rüzgarlı=yanlış
2	Hava=kapalı	Hava {güneşli, yağmurlu}
3	Hava=güneşli	Hava {kapalı, yağmurlu}
4	Hava=yağmurlu	Hava {güneşli, kapalı}
5	Nem $\leq$ 78	Nem $>$ 78
6	Sıcaklık $\leq$ 72	Sıcaklık $>$ 72

Bu altı farklı bölünme şekli için hesaplanan  $\Phi$  değerleri ve bu hesaplamaları yaparken kullanılan değerler Tablo 7.5’te verilmiştir. Buna göre ilk düğüm sola hava=”kapalı” seçeneği ile ayrılacak sağa ise hava=”güneşli veya yağmurlu” seçenekleri ile ayrılacaktır. Sola ayrılan düğüm “saf” olduğu için (yani sadece golfu uygun günleri içerdiği için) son düğüm olacaktır.

Tablo 7.5. İlk aşamada hesaplanan  $\Phi$  değerleri

	$P_L$	$P_R$	$2P_LP_R$	$P(j t_L)$	$P(j t_R)$	$\Phi$
1	0,428571429	0,571428571	0,489795918	0,5	0,75	0,244897959
				0,5	0,25	
2	0,285714286	0,714285714	0,408163265	1	0,5	<b>0,408163265</b>
				0	0,5	
3	0,357142857	0,642857143	0,459183673	0,4	0,78	0,346938776
				0,6	0,22	
4	0,357142857	0,642857143	0,459183673	0,6	0,67	0,06122449
				0,4	0,33	
5	0,428571429	0,571428571	0,489795918	0,8	0,5	0,326530612
				0,2	0,5	
6	0,571428571	0,428571429	0,489795918	0,6	0,67	0,040816327
				0,4	0,33	

Algoritma ikinci aşamada olası bölünme adaylarını ve bu adayların  $\Phi$  değerlerini bulacaktır. Kapalı günler (dört adet) eksildiği için halihazırda sadece dört çeşit bölünme olabilir. Bunlar da Tablo 7.6'da gösterilmiştir. Tablo 7.7 hesaplanan  $\Phi$  değerlerini vermektedir. Burada görülmektedir ki 0,4 değeri ile rüzgarın olması ya da olmaması golf a uygunluğu belirleyen en önemli değişken olarak görülmektedir. Bir başka ilginç nokta da  $\Phi$  değerinin 4. düğüm için sıfır çıkmasıdır, burada soldaki 6 ve sağdaki 4 gün içinde eşit oranda golf a uygun gün bulunmaktadır. Yani bir başka deyişle hava sıcaklığının 72 dereceden düşük ya da fazla olduğunu bilmek golf a uygunluk açısından ekstra bir bilgi vermemektedir. İki değişkenin istatistiki olarak bağımsız olduğunu da söyleyebiliriz (tabii bu karar düğümünde tüm veride değil).

**Tablo 7.6. İkinci yinelemede olası bölünme adayları**

Bölünme	Sol Alt Düğüm	Sağ Alt Düğüm
1	Rüzgarlı=doğru	Rüzgarlı=yanlış
2	Hava=güneşli	Hava=yağmurlu
3	Nem $\leq$ 78	Nem $>$ 78
4	Sıcaklık $\leq$ 72	Sıcaklık $>$ 72

**Tablo 7.7. İkinci aşamada hesaplanan  $\Phi$  değerleri**

	$P_L$	$P_R$	$2P_LP_R$	$P(j t_L)$	$P(j t_R)$	$\Phi$
1	0,4	0,6	0,48	0,25	0,666667	0,4
				0,75	0,333333	
2	0,5	0,5	0,5	0,4	0,6	0,2
				0,6	0,4	
3	0,3	0,7	0,42	0,666667	0,428571	0,2
				0,333333	0,571429	
4	0,6	0,4	0,48	0,5	0,5	0
				0,5	0,5	

Bu aşamadan sonra oluşan her ki düğüm de saf değildir. Yani rüzgar=var (sol alt düğüm) seçeneğinde üç uygun değil bir uygun gün vardır. Sağ alt düğümde ise dört uygun iki uygun olmayan gün vardır. Bu durumda yukarıdaki hesaplamaların bu sefer her iki karara düğümü için de tekrarlanması gerekir. Ya da alternatif olarak karar ağacı bu şekilde tutulup, sol alt düğüm “oynama” sağ alt düğüm ise “oyna” şeklinde sonlandırılıp son düğüm haline getirilebilir. Bu durumda doğru tahmin oranı 11/14 yani %79 olacaktır. Ancak her düğüm saflaşana kadar algoritmaya devam edilirse, hava koşullarına göre (yağmurlu, güneşli) bir kere daha bölünme gerçekleşecektir. Bu aşamada iki saf düğüm daha elde edilir: rüzgar var yağmurlu ve rüzgar yok yağmurlu. Bu noktada algoritma durdurulursa doğru tahmin oranı 12/14 yani %86 olacaktır. Saflaşma için nem değişkenini kullanmak yeterli olacak: yani rüzgar var, güneşli ise nem düşükse oyna yüksekse oynama; rüzgar yok güneşli ise nem düşükse oyna yüksekse oynama. Karar ağa-



cının ne kadar büyütüleceği ve seviyesinin ne kadar fazla olacağı kullanıcının isteğine göre değişebilir. Çok detaylı, veriyi “ezberlemiş” bir karar ağacı genel kullanım için çok faydalı olmayabilir. Ayrıca her seviye arttığında tahmin oranı ne kadar artmaktadır? Bu da hangi seviyede karar ağacının “budanması” gerektiği konusunda bize yol gösterebilir.

### 7.3. C 4.5 Algoritması

C4.5 algoritması CART algoritmasının bazı özellikleri geliştirilerek Quinlan tarafından önerilmiştir (1992). CART algoritmasından temel farkı bir düğümden çıkan dal sayısının iki ile sınırlı olmamasıdır. Ayrıca düğüm homojenliğini ölçmek için kullandığı formül de farklıdır.

C4.5 algoritması en iyi bölünmeye karar verirken bilgi kazanımı (information gain) ya da entropi azalması ismi verilen kriteri kullanır. Homojenliği (ya da saflığı) ölçmek için Gini indeksi, sınıflama hatası ve entropi gibi bir takım ölçüler kullanılabilir. Kısaca bunların üstünden geçerek:

- Gini indeks:  $1 - \sum p_j^2$ , en fazla değeri olan “0” değerini herhangi bir sınıfın olasılığı “1” ise alır. Yani “saflık” durumu Gini indeksinin 0 olması durumudur.
- Sınıflama Hatası:  $1 - \max\{p_j\}$ , aynı şekilde tek bir sınıftan oluşan veride sınıflama hatası “0” olacaktır.
- Entropi:  $-\sum p_j \log_2(p_j)$ ,  $\log(0)=0$  olarak tanımlanırsa  $\log(1)=0$  olduğu için herhangi bir sınıfın olasılığı “1” ise entropi değeri “0” olacaktır, yani bu indekste de saflık durumu indeksin en düşük değeri ile ifade edilmektedir. Tam tersi en yüksek değer tüm olasılıkların eşit olması durumunda elde edilir. İki sınıf olması durumunda maksimum entropi “1” olacaktır.

Golf veri setinde hiçbir sınıflama kuralı uygulanmadan önce 9 golfa uygun 5 de uygun olmayan gün olduğunu ele alırsak bu değerler şu şekilde hesaplanabilir.

- Gini indeksi:  $1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0,4592$
- Sınıflama Hatası:  $1 - \left(\frac{9}{14}\right) = \frac{5}{14} = 0,3571$
- Entropi:  $-\left(\frac{9}{14} \log_2 \left(\frac{9}{14}\right) + \frac{5}{14} \log_2 \left(\frac{5}{14}\right)\right) = 0,9403$

C4.5 algoritması kök düğüm için entropiyi ölçtüktan sonra olası bölünmeleri tespit eder (burada daha önce söylendiği gibi ikili bölünme zorunluluğu yoktur) ve bu oluşan grupların ayrı ayrı entropi değerlerini hesaplar. Daha sonra grup büyüklüğü ile ağırlıklı ortalama alınarak bölünmenin entropisi hesaplanır. Aradaki fark da entropi kazanımı olarak adlandırılır. Entropi kazanımı entropi kazanım oranı olarak yüzde ile de ifade edilebilir. Golf örneği için kök düğümden sonra hava koşullarına göre veri üç farklı gruba ayrılabilir, bunlar:

- Kapalı Hava: 4 gözlem: 4 Uygun, 0 Uygun değil
- Güneşli Hava: 5 Gözlem: 2 Uygun, 3 Uygun değil
- Yağmurlu Hava: 5 Gözlem: 3 Uygun, 2 Uygun değil

Her bir grup için entropi hesaplanırsa aşağıdaki sonuçlar ortaya çıkar:

- Kapalı Hava:  $-\frac{4}{4} \log_2 \left(\frac{4}{4}\right) - \frac{0}{4} \log_2 \left(\frac{0}{4}\right) = 0$
- Güneşli Hava:  $-\frac{2}{5} \log_2 \left(\frac{2}{5}\right) - \frac{3}{5} \log_2 \left(\frac{3}{5}\right) = 0,9710$
- Yağmurlu Hava:  $-\frac{3}{5} \log_2 \left(\frac{3}{5}\right) - \frac{2}{5} \log_2 \left(\frac{2}{5}\right) = 0,9710$

Bu bölünme için ağırlıklı ortalama alınarak entropi hesaplanırsa:

$$entropi = \frac{4}{14} 0 + \frac{5}{14} 0,9710 + \frac{5}{14} 0,9710 = 0,6935$$

$$kazanim = 0,9403 - 0,6935 = 0,2467$$

$$kazanimorani = \frac{0,2467}{0,9403} = 0,2624 = \%26,24$$

Dolayısıyla bu bölünme şekli için entropi kazanımı 0,2467 olacaktır. Bu seviye diğer bölünme şekilleriyle karşılaştırılmalı ve en yüksek kazanım seçilerek devam edilmelidir.

#### 7.4. C 5.0 Algoritması

C5.0 (See5 olarak da adlandırılır) algoritması, C4.5 algoritmasının geliştirilmiş halidir. Programın 400 gözlemle sınırlı Windows uyumlu bir versiyonu URL13'ten indirilebilir. URL14'te ise C4.5 ile ayrıntılı bir karşılaştırma yer almaktadır. Temelde tahmin hatasında çok fazla fark olmamakla beraber, bilgisayar süresi ve karar ağaçlarının karmaşıklığı konularında C5.0, C4.5'e üstünlük sağlamaktadır.

## 8. BÖLÜM:

### MODEL KARŞILAŞTIRMA YÖNTEMLERİ

Kitabın başlarında da belirtildiği gibi CRISP-DM modelinde (ya da benzer diğer modellerde) modelleme aşamasından sonra model değerlendirme aşaması gelir. Alternatif modeller (örneğin C4.5 ve C5.0) ile aynı veri seti üzerinde uygulanan algoritmaların ne kadar başarılı olduklarının değerlendirilmesini yapmak için bir takım yöntemler geliştirilmiştir. Bu bölümde bu yöntemleri inceleyeceğiz.

#### 8.1. Tahmin ve Kestirim İçin Model Karşılaştırma Yöntemleri

Regresyon gibi klasik istatistikî yöntemler de veri madenciliği uygulamalarında kullanılabilir. Dolayısıyla bu tür yöntemlerde de farklı modelleri (örneğin farklı bağımsız değişken kullanan çoklu regresyon yöntemleri) karşılaştırmak için kullanılan yöntemleri şöyle özetleyebiliriz:

- Ortalama Karesel Hata (MSE) =  $\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1}$
- Standart Hata =  $\sqrt{MSE}$ , bu yaklaşık ortalama tahmin hatasını değişkenin kendi biriminden verir. Ne kadar düşükse o kadar iyidir.
- Belirleme Katsayısı  $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$ , bağımlı değişkendeki

varyansın ne kadarının bağımsız değişkenler tarafında açıklandığı, bir başka deyişle bağımsız değişkenler olmadan yapılan tahminin (ortalama  $\bar{y}$ ) ne kadar iyileştiğini gösterir.

- Düzeltilmiş Belirleme Katsayısı  $= R_{adj}^2 = 1 - \left[ 1 - R^2 \left( \frac{n-1}{n-p-1} \right) \right]$ , her modelde olduğu gibi regresyonda da fazla değişken kullanımı modeli karmaşık hale getirir ve istenmez. Dolayısıyla belirleme katsayısı ek bir değişkenle sürekli artacağından bu artışın gerekli olup olmadığına karar verirken düzeltilmiş belirleme katsayısı kullanılabilir zira bu katsayı ek bir değişkenle artmak zorunda değildir.

Yukarıdaki ölçüler regresyon modeli için geçerli olsa da farklı bir tahmin yöntemi ile (örneğin bu kitapta anlatılan k-en yakın komşu yöntemi ile kan basıncını tahmin örneği) yapılan tahminlerle regresyon tahminlerini karşılaştırmak için de kullanılabilir. Burada değişken sayısını belirten  $p$  değeri anlam taşımayabilir ama gözlem sayısı değişken sayısına göre yüksek olduğunda  $p$ 'nin formüle katılıp katılmaması anlamlı farklar çıkarmayacaktır.

## 8.2. Sınıflama Uygulamaları İçin Model Karşılaştırma Yöntemleri

Sınıflama probleminde veri madenciliği yeni bir gözlemi bir hedef değişkenin çeşitli seviyelerine göre sınıflamaya çalışır. Model kurulmaya çalışırken eldeki hedef değişkenin seviyeleri bilinen veri setini iki gruba ayırmak faydalı olur:

- Eğitim Seti
- Doğrulama Seti

Eğitim seti ile, örneğin C4.5 algoritması kullanılıyorsa, önce hedef değişkeni eğitim setine göre optimal belirleyen karar silsilesi oluşturulur. Daha sonra karar kurallarını belirlerken hiç kullanılmamış doğrulama seti ile ne kadar başarılı tahmin yapıldığı test edilebilir. Bu işlem farklı modellerin (örneğin C4.5, C5.0 ve k-en yakın komşu) eldeki problem için ne kadar doğru tahminler yaptığını belirlemekte ya da modelde belirlenmesi gereken parametreleri optimize etmekte kullanılabilir. Ayrıca sonuçta modelin rastgele sınıflamadan ne kadar farklı olduğunun da belirlenmesi gerekir. Kuşkusuz

en maliyetsiz teknik yeni yöntemleri rastgele, örneğin zar atarak, sınıflamak olacaktır.

### 8.3. Hatalı Sınıflama Türleri

Hatalı sınıflama türlerini anlamak için tiptan bir örnek verelim. Bir virüsün (örneğin H1N1) varlığı ile ilgili yapılan bir testin iki sonucu olur, hasta ya pozitifdir ya da negatif. İşletmelerde de sınıflama yapılırken bir cep telefonu abonesi gelecek ay içinde başka bir şirkete geçecek ya da geçmeyecek diye sınıflanabilir. Epidemiyoloji uzmanları testin sonucu ve kişinin gerçekten virüs taşıyıp taşımadığı ile ilgili yapılan doğru tahminleri iki gruba ayırırlar ve firmaların ürettiği testleri bu değerleri bulmak için değerlendirirler. Bu iki farklı veri ülkenin uzman epidemiyoloji otoritesi tarafından (örneğin ABD için Center for Disease Control – CDC, URL15’te bir örneğe bakılabilir) yayınlanır. Bu iki tür doğru sınıflama ve olasılıksal anlamları aşağıdaki gibidir:

- Sensitivity: Doğru pozitif yani  $P(\text{test pozitif} \mid \text{virüs var})$ , 1-sensitivity yanlış pozitif oranını verir.
- Specificity: Doğru negatif yani  $P(\text{test negatif} \mid \text{virüs yok})$ , 1-specificity yanlış negatif oranını verir.

Genelde çoğu test yüksek sensitivity verirken bazılarının specificity değerleri düşüktür, dolayısıyla yanlış negatif ihtimalleri yüksektir. Bu da negatif sonuçlara güvenilemeyeceği anlamına gelir. Ülkemizde de korkuya sebep olan H1N1 virüsünde piyasaya çıkan pahalı testler ise ikinci tip doğru tahmini (specificity) yüksek veren testlerdi. Şunu da belirtmek gerekir ki her zaman pozitif veren sahte bir test %100 sensitivity’ye sahip olacaktır, aynı şekilde her zaman negatif veren sahte bir test %100 specificity’ye sahip olacaktır. Hiçbir şey yapmadan her gözlemi aynı sınıfa koyan bir veri madenciliği algoritması da aynı şekilde herhangi bir hata tipi oranının maksimize edebilir. Tabii bu başarılı olduğu anlamına gelmez. Dolayısıyla bir veri madenciliği algoritmasının başarısından söz etmek için farklı tip hata oranlarına bakmak ve maliyet unsurunu işin içine sokmak gerekir.

Bu açıklayıcı örnekten sonra genel olarak sınıflama hatasının nasıl belirlendiğine dair bir örnek verelim. Sınıflama hatalarını (iki tipini) görmek için, sınıflama algoritmasını, sonuçlarını bildiğimiz veride çalıştırdıktan sonra 2x2 bir çapraz tablo yapmamız gerekir. Bu çapraz tablo hata matrisi (confusion matrix) olarak adlandırılır ve Tablo 8.1'deki gibi ortaya çıkar. Tablo 8.2 ise kayıp müşteri tahmin etmek için kullanılan bir veri madenciliği algoritması ile tahmin edilen kayıp müşteri ve gerçekte olan kayıp müşterileri karşılaştıran çapraz tablodur.

**Tablo 8.1. Örnek hata matrisi**

		Tahmin Edilen		Toplam
		1	0	
Gerçekte Olan	1	Doğru pozitif (a)	Yanlış Negatif (b)	P' (a+b)
	0	Yanlış pozitif (c)	Doğru Negatif (d)	N' (c+d)
Toplam		P (a+c)	N (b+d)	a+b+c+d

**Tablo 8.2. Örnek hata matrisi**

	Kayıp Müşteri Olarak Sınıflanmış	Kayıp Müşteri Değil Olarak Sınıflanmış	Toplam
Gerçekte Kayıp Müşteri	2.000	3.000	5.000
Gerçekte Kayıp Müşteri Değil	500	19.500	20.000
Toplam	2.500	22.500	25.000

Öncelikle beş farklı tip hata oranını tanımlayabiliriz:

- Genel Hata Oranı =  $\frac{b+c}{a+b+c+d}$

- Negatif Tahminler İçinde Yanlış Negatif Oranı =  $\frac{b}{b+d} = \frac{b}{N}$
- Pozitif Tahminler İçinde Yanlış Pozitif Oranı =  $\frac{c}{a+c} = \frac{c}{P}$
- Pozitif Olduğu Bilinen Gözlemlerde Doğru Tahmin Oranı (Sensitivity) =  $\frac{a}{a+b}$
- Negatif Olduğu Bilinen Tahminlerde Doğru Tahmin Oranı (Specificity) =  $\frac{d}{c+d}$

Bu üç farklı grubu Tablo 8.2’teki veriye uygularsak sonuçlar şöyle çıkar:

- Genel Hata Oranı =  $\frac{3500}{25000} = 0,14$ , yani model genelde %14 hata yapmaktadır ve doğru tahmin oranı %86’dır. Peki komplike bir matematiksel model kullanmadan her gözlemi “kayıp” olarak sınıflasak hata oranı ne olurdu? Verideki “kayıp” oranı  $5.000/25.000 = \%20$  olduğundan hata oranımız da %20 olurdu. Dolayısıyla hata oranı olarak %14 küçük görünse de rastgele bir algoritmanın bulduğu sonuçlarla da karşılaştırmak yararlı olacaktır. Ayrıca bir sonraki alt bölümde anlatılan maliyet noktasında da genel hata oranından bağımsız çok farklı maliyet yapıları ortaya çıkabilir.
- Negatif Tahminler İçinde Yanlış Negatif Oranı =  $\frac{3000}{22500} = 0,13$ , yani modelin yaptığı negatif tahminlerin %13’ü yanlış, %87’si doğrudur.
- Pozitif Tahminler İçinde Yanlış Pozitif Oranı =  $\frac{500}{2500} = 0,20$ , yani modelin yaptığı pozitif sınıflamaların %20’si yanlış %80’i doğrudur. Buradan modelin yaptığı pozitif tahminler negatiflerden daha az güven duyulacağı sonucu çıkar.



- Pozitif Olduğu Bilinen Gözlemlerde Doğru Tahmin Oranı

$$(\text{Sensitivity}) = \frac{2000}{5000} = 0,40$$

- Negatif Olduğu Bilinen Gözlemlerde Doğru Tahmin Oranı

$$(\text{Specificity}) = \frac{19500}{20000} = 0,975$$

#### 8.4. Fayda/Maliyet Hesabı ile Karar Verme

Bir önceki alt bölümde verilen örnekte tıbbi testte virüs yokken var denilmesi ya da varken yok denilmesi elbette aynı kefiye konulamaz. Yokken var denilmesi durumu müteakip doğrulama testleri yapılarak belki çözülebilir ama ikinci durum elbette çok daha riskli maliyetli bir durum olacaktır. Bu durumun hastaya maliyetini belki ölçmek mümkün olmaz çünkü bilmeden virüs ilerleyecek tedavi mümkün olmayacaktır.

Ancak işletme uygulamalarında hatalı sınıflandırılan gözlemlerin bize maliyetini aşağı yukarı tahmin edebiliriz ve buna göre de sınıflama uygulamamızın belirlediği sınıflama eşik değerlerini kalibre edebiliriz. Daha önce verdiğimiz, ve vaka çalışmasında inceleyeceğimiz kayıp müşteri modelini ele alırsak, bir cep telefonu abonesinin gelecek ay içinde başka bir şirkete geçecek ya da geçmeyecek şekilde “yanlış” sınıflanmasının bize maliyetini araştırmamız gerekir. Elimizde cep telefonu abonesi ile ilgili şu bilgiler olsun:

- Müşteriyi kaybedeceğimiz sinyali alırsak müşteriye gelecek ayki faturasında kullanılmak üzere 30TL vereceğiz. Bu durumda müşterinin çıkma fikrini değiştirip şirkete kalacağınız düşünelim.
- Müşteriyi kaybedersek ise müşterinin hesaplanmış hayat boyu değeri olan 500TL’yi kaybedeceğiz.

Bu durumda bir müşteriyi kaybedeceğimiz halde kaybetmeyecekmiş gibi sınıflarsak (yani yanlış negatif) 500 TL maliyetimiz olacak. Bununla beraber normalde kaybetmeyeceğimiz bir müşteriyi kaybedecekmişiz gibi sı-

nıflarsak (yani yanlış pozitif) kaybımız sadece 30TL olacaktır. Burada açıktır ki yanlış pozitif ve yanlış negatif maliyetleri aynı değildir. Kuşkusuz yanlış negatif maliyeti daha yüksek olmalıdır ve eğer kullandığımız modele bu bilgiyi de dahil ederek parametreleri kalibre edebiliyorsak bunu yapmamız gerekir. Burada veri madenciliği kullanmadan uygulanabilecek iki ekstrem karar söz konusu olabilir. Bunların maliyetleri Tablo 8.2'deki rakamlara göre:

- Hiç bir şey yapma: 5000 müşteri şirketi terk eder, toplam maliyet 2.500.000 TL.
- Herkese 30TL'lik öneriyi götür: Bu durumda maliyet 750.000 TL olur ve kimse şirketi terk etmez.

Veri madenciliği kullanarak ve Tablo 8.2'deki verileri esas alarak ise maliyetler şu şekilde olur:

- 2500 kişi kayıp müşteri olarak sınıflandığı için 30TL'lik paket sunulur bunun maliyeti ise 75.000 TL olur. Diğer yandan bu kişilerin 500 tanesine gereksiz yere paket sunulmuştur, 3.000 kişi ise kayıp müşteri oldukları halde yanlış sınıflandırılmıştır, bu şekilde 3.000 kişiden 1.500.000 TL yaşam boyu değer kaybedilir. Bu durumda toplam maliyet 1.575.000 olur.

Aslında hiçbir şey yapmamaktan daha iyi bir maliyet olmasına rağmen herkese 30TL'lik öneriyi götürmek daha iyi görünmektedir. Bu durum iki şekilde çözülebilir; ya daha iyi tahmin oranları veren bir algoritma kullanılacak (sıfır hata durumunda maliyet sadece 150.000 TL olacaktır) ya da algoritmada yanlış tahmin maliyetleri seçeneği varsa bu kullanılarak toplam hata oranını yükseltmek pahasına yanlış negatif oranını düşürmek olacaktır. Böyle bir opsiyon bazı istatistikî programlarda verilmektedir. Öntanımlı olan seçenek maliyetlerin oranının 1 olduğudur, ancak bu uygulamada bu oran 17 (yaklaşık 500/30) olarak değiştirilerek toplam maliyeti düşüren tahmin sonuçları elde edilebilir.

### 8.5. Kazanım ve Birikimli Kazanım Grafikleri

Kazanım (Lift) kavramını anlamak için Tablo 8.2'deki örnekte kayıp müşteri olarak sınıfladığımız 2500 kişi ile görüştüğümüzü varsayalım. Burada görüştüğümüz kişilerin %80'i gerçekten de kayıp müşteridir. Halbuki bu 2500 kişiyi rastgele bir yöntemle belirleseydik kayıp müşteri oranı 5.000/25.000 yani %20 olacaktı. Bu iki değer oranı kazanım (lift) olarak adlandırılır. Yani:

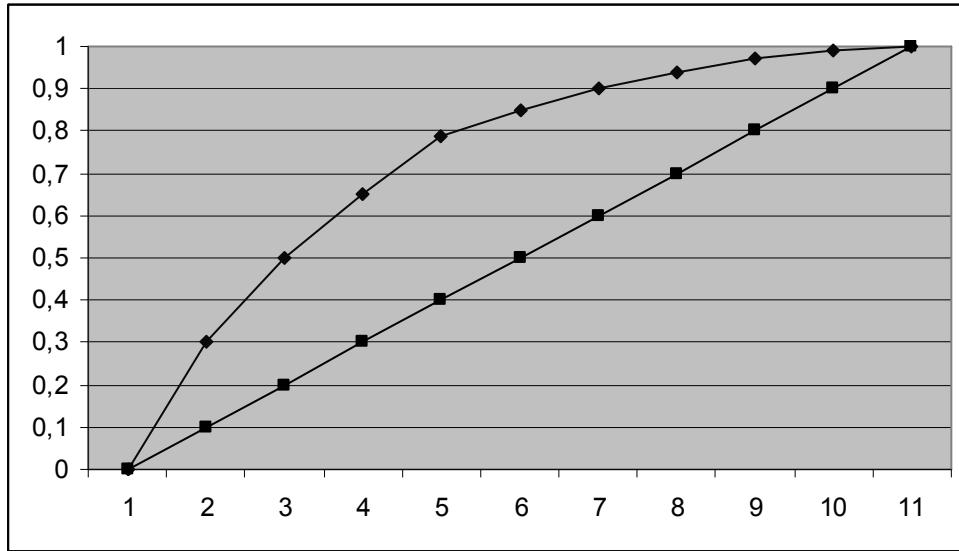
$$\text{Kazanım} = \frac{\frac{a}{a+c}}{\frac{a+b}{a+b+c+d}}$$

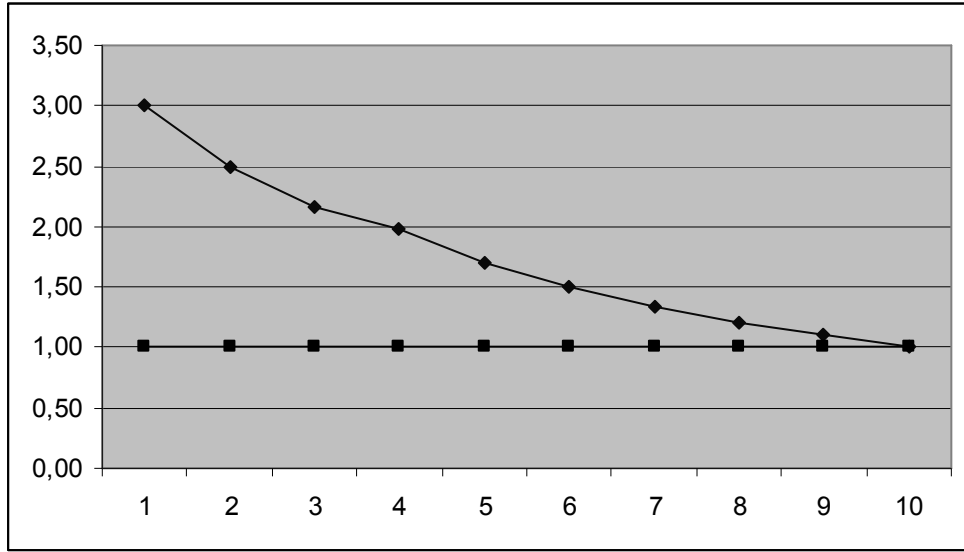
Örneğimizde kazanım  $80/20=4$  olarak bulunur. Bir başka deyişle veri madenciliği algoritması “rastgele belirlenecek 2500” kişiye oranla 4 kat daha fazla etkili bir belirleme/sınıflama yapmıştır. Burada kritik nokta kazanım kavramının örneklem büyüklüğü ile değişmesidir. Yani kayıp müşteri olacağından daha emin olduğunuz 1000 kişiyle temasa geçseydik kazanım kuşkusuz daha yüksek olurdu. Tam tersi temas ettiğimiz kişi sayısını arttırsaydık da kazanım giderek düşer sonunda 25.000 kişinin tümüyle görüşüldüğünde “1” sayısına düşerdi. Kazanım grafiği (lift chart) ve bunun birikimli hali (gain chart) çeşitli veri madenciliği program çıktılarını karşılaştırmakta kullanılan çok önemli bir araçtır. Kazanım grafiğini oluşturmak için gözlemlerin pozitif olma olasılıklarına göre büyükten küçüğe sıralanması gerekir.

Kazanım grafiğini anlamak için URL16'dan alınma şu örneği verelim. Bir doğrudan pazarlama şirketinin katalogla alışveriş amacıyla katalog gönderebileceği 100.000 adres olduğunu düşünelim. Bunların %20'sinin kataloga cevap vereceği (yani satın alacağı) biliniyor olsun. Ayrıca bir veri madenciliği programı adres listesindeki kişileri 10.000'erlik gruplara satın alma olasılıklarına göre yüksekten küçüğe sıralamış olsun. Aynı zamanda her grupta gerçekten kaç kişinin satın alma yapacağı verilmiş olsun. Bu durum Tablo 8.3'teki gibi verildiğinde Şekil 8.1 ve 8.2 de verilen birikimli kazanım ve kazanım grafikleri ortaya çıkar.

**Tablo 8.3. Örnek kazanım grafiği verisi**

İletişime Geçilen Müşteri Sayısı	Olumlu Cevap Sayısı	Olumlu Cevap Oranı	Kazanım	Olumlu Cevap/20.000
10.000	6.000	0,60	3,00	0,30
20.000	10.000	0,50	2,50	0,50
30.000	13.000	0,43	2,17	0,65
40.000	15.800	0,40	1,98	0,79
50.000	17.000	0,34	1,70	0,85
60.000	18.000	0,30	1,50	0,90
70.000	18.800	0,27	1,34	0,94
80.000	19.400	0,24	1,21	0,97
90.000	19.800	0,22	1,10	0,99
100.000	20.000	0,20	1,00	1,00

**Şekil 8.1. Örnek birikimli kazanım grafiği, dikey eksen olumlu cevap/20.000, diyagonal çizgi rastgele seçimde olacak eğriyi gösterir.**



Şekil 8.2. Örnek kazanım grafiği, dikey eksen kazanım, yatay çizgi “1” değerini ifade eder.

## **9. BÖLÜM:**

### **VAKA ANALİZİ - ARAMA A.Ş.**

Kayıp müşteri (İngilizce “churn” ya da “attrition”) bir müşterinin bir şirketin hizmetlerinden memnun olmayıp bir başka şirkete geçmesi için kullanılan bir terimdir. Cep telefonu, kredi kartı gibi ürünlerde başka şirketlerin sürekli cazip fırsat tekliflerine maruz kalan tüketiciler bir şirketten diğerine geçebilirler. Yeni bir müşteri elde etmenin maliyetinin mevcut müşteriyi tutmaktan çok daha pahalı olduğu varsayılırsa, şirketlerin müşterileri kaybetmeden önce belirleyip onları kalmaya ikna etmeye çabalamak istemeleri oldukça doğaldır.

Bu problem veri madenciliğinde de gözetimli öğrenme sınıfına giren bir problemdir. Hedef değişken müşterinin başka şirkete geçip geçmediği, bağımsız değişkenler ise müşteri hakkında kaydı tutulan çeşitli verilerdir. Burada Larose (2005) web sitesinde verdiği (URL17) kayıp müşteri verisini kullanarak çeşitli yöntemleri uygulamalı olarak göstereceğiz. Kayıp müşteri verisi hedef değişken (kayıp müşteri doğru ya da yanlış olmak üzere ikili bir değişken), ve bu değişkeni tahmin etmekte kullanılabilecek 20 farklı değişkenden oluşmaktadır. Bu değişkenler 3.333 kayıt için verilmiştir. Analizden önce değişkenleri tanımlamak gerekir:

1. Eyalet: Müşterinin bulunduğu eyalet
2. Hesap süresi: Kaç aydır hesap aktif
3. Alan kodu
4. Telefon numarası
5. Uluslararası plan: İkili değişken
6. Telesekreter servisi: İkili değişken
7. Telesekretere bırakılan mesaj sayısı
8. Gün içi kullanılan dakika sayısı
9. Gün içi yapılan arama sayısı

10. Toplam günlük ücret
11. Akşam kullanılan dakika sayısı
12. Akşam yapılan arama sayısı
13. Toplam akşam ücret
14. Gece kullanılan dakika sayısı
15. Gece yapılan arama sayısı
16. Toplam gece ücret
17. Uluslararası kullanılan dakika sayısı
18. Uluslararası yapılan arama sayısı
19. Toplam uluslararası ücret
20. Müşteri hizmetlerine yapılan arama sayısı

Verilen URL’den indirilen .txt sonlu dosya rahatlıkla SPSS (PASW) programına aktarılabilir. Burada önemli olan veri aktarılırken ondalık ayraçının orijinal veride “.” olarak kullanıldığını bilmektir. Bu düzeltilmezse sayısal veriler “string” olarak algılanacaktır. Verideki yukarıda sayılan 20 değişken ve hedef değişken olan kayıp müşteri değişkeni (churn? Olarak ifade edilmiştir) ilk 15 gözlem için Şekil 9.1’de verilmiştir.

	State	AccountLength	AreaCode	Phone	IntlPlan	VMailPlan	VMailMessage	DayMins	DayCalls	DayCharge	EveMins	EveCalls	EveCharge
1	KS	128	415	382-4657	no	yes	25	265,100000	110	45,070000	197,400000	99	16,780000
2	OH	107	415	371-7191	no	yes	26	161,600000	123	27,470000	195,500000	103	16,620000
3	NJ	137	415	358-1921	no	no	0	243,400000	114	41,380000	121,200000	110	10,300000
4	OH	84	408	375-9999	yes	no	0	299,400000	71	50,900000	61,900000	88	5,260000
5	OK	75	415	330-8626	yes	no	0	166,700000	113	28,340000	148,300000	122	12,610000
6	AL	118	510	391-8027	yes	no	0	223,400000	98	37,980000	220,600000	101	18,750000
7	MA	121	510	355-9993	no	yes	24	218,200000	88	37,090000	348,500000	108	29,620000
8	MO	147	415	329-9001	yes	no	0	157,000000	79	26,690000	103,100000	94	8,760000
9	LA	117	408	335-4719	no	no	0	184,500000	97	31,370000	351,600000	80	29,890000
10	WV	141	415	330-8173	yes	yes	37	258,600000	84	43,960000	222,000000	111	18,870000
11	IN	65	415	329-8803	no	no	0	129,100000	137	21,950000	228,500000	83	19,420000
12	RI	74	415	344-9403	no	no	0	187,700000	127	31,910000	163,400000	148	13,890000
13	IA	168	408	363-1107	no	no	0	128,800000	96	21,900000	104,900000	71	8,920000
14	MT	95	510	394-8006	no	no	0	156,600000	88	26,620000	247,600000	75	21,050000
15	IA	62	415	366-9238	no	no	0	120,700000	70	20,520000	307,200000	76	26,110000

Şekil 9.1. İlk 15 gözlem için örnek veri

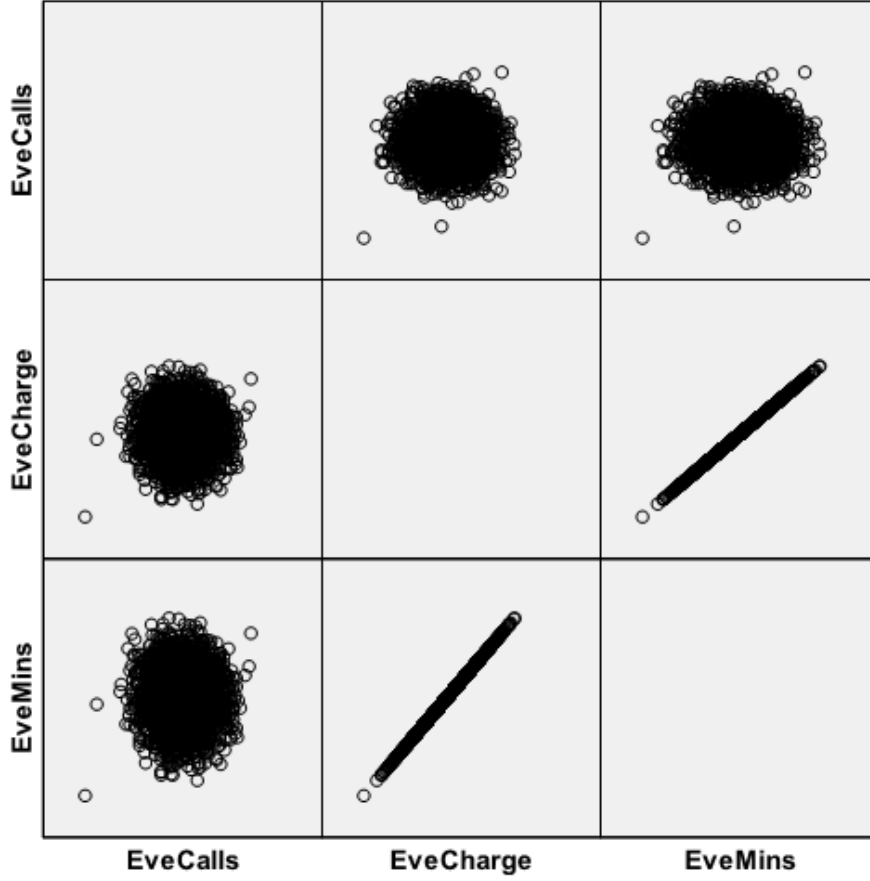
## 9.1. Açımsayıcı Veri Analizi

Analize başlamadan önce değişkenleri teker teker incelemek faydalı olacaktır. Öncelikle eyalet, alan kodu ve telefon numarasıyla başlayalım. Telefon numarası her gözlemin tanımlayıcısı olabilir ancak analizde bir değeri yoktur, dolayısıyla silinebilir. Alan koduna bakıldığında sadece üç farklı alan kodunun gözlemlendiği görülür. Ancak eyalet sayısına bakıldığında 51 farklı değer görülür. Burada ABD’de cep telefonu alan kodlarının yerel alan koduna göre verildiği düşünülürse bir mantıksızlık vardır. Muhtemelen veriyi bağışlayan şirket bazı alanları karıştırarak veri gizliliğini korumaya çalışmıştır. Gerek eyalet gerek de alan kodu analizde kullanılmaya uygun ol-



madıkları için çıkarılabilir. Bu durumda toplam bağımsız (tahmin edici) değişken sayısı 20 değişkenden 17'ye indirilmiştir.

Veriyi oluşturan diğer nitel ve nicel değişkenler incelendiğinde elenecek başka değişkenler de bulunabilir. Örneğin gün içi konuşma sayısı, dakikası ve ücreti üç farklı değişken olarak kaydedilmiştir. Bu akşam, gece ve uluslararası konuşmalar için de geçerlidir. Ancak basit bir gözlem bu değişkenlerden dört tanesinin elememizi sağlar. Şekil 9.2'de olduğu gibi akşam konuşma sayısı, dakikası ve ücretlerinin serpilme diyagramlarını çizdiğimizde konuşma ücreti ile dakikanın korelasyonunun "1" olduğunu görürüz. Aynı gözlemi diğer benzer değişkenler için de yapabiliriz. Aslında her grupta toplam maliyet belirlenmiş dakika ücreti ile konuşma dakikasının çarpımından oluşur. Yaklaşık gündüz 17 cent, akşam 8,5 cent, gece 4,5 cent ve son olarak da uluslararası için 27 cent olarak ücretlendirme olmuştur. Bu eleme opsiyonel değildir, çoklu regresyon gibi bazı analizler bu mükemmel korelasyon durumunda uyarı verecek ve çalışmayacaktır. Karar ağacı, k-en yakın komşu gibi bir modelde ise sorun olmaz ama program fazladan hesaplama yapacağı için sadece zaman kaybetmiş oluruz. Dolayısıyla kullanılacak değişken sayısını 17'den 13'e maliyet ile ilgili değişkenleri elimine ederek indirebiliriz.



**Şekil 9.2. Akşam konuşma sayısı (evecalls), dakikası (evemins) ve ücretinin (evecharge) aralarında serpilme diyagramları**

Veride hedef değişken dahil toplam üç nitel değişken bulunmaktadır. Bunlar: Uluslararası plan varlığı, telesekreter planı varlığı ve hedef değişken olan kayıp müşteri değişkeni. Bunların özelliklerini grafiklerle ve çapraz tablolarla inceleyebiliriz. Tablo 9.1 kayıp müşteri değişkeninin özet bilgilerini vermektedir. Aynı şekilde Tablo 9.2 ve 9.3 diğer iki nitel değişkenin özet bilgilerini vermektedir. Buna göre genel müşteri kaybetme oranı  $483/3.333 = \%14,49$  olarak ortaya çıkmaktadır. Burada ilginç olan telesekreter planı sahipliğinin ve uluslararası plan sahipliğinin kayıp müşteriyi tahmin etmekte kullanılıp kullanılmayacağıdır. Bunun için ise çapraz tablolar işimize yarayabilir. Tablo 9.4'ten görebildiğimiz kadarıyla uluslararası plana sahip müşterilerin  $\%42$ 'si kaybedilmiş ancak sahip olmayanların ise sa-

dece % 11'i kaybedilmiştir. Bu da başka bir şirkete geçme ile uluslararası plana sahip olma arasında bir ilişki olabileceğini gösterir. Tablo 9.5'ten de anlaşılacağı gibi aynı şey telesekreter planı için bu kadar kuvvetle söylene-  
mez ama bu sefer plana sahip müşteriler daha az oranla başka bir şirkete  
geçmişlerdir.

**Tablo 9.1. Kayıp müşteri özet bilgileri**

Churn					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	false.	2850	85,5	85,5	85,5
	true.	483	14,5	14,5	100,0
	Total	3333	100,0	100,0	

**Tablo 9.2. Uluslararası plan özet bilgileri**

IntlPlan					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	3010	90,3	90,3	90,3
	yes	323	9,7	9,7	100,0
	Total	3333	100,0	100,0	

**Tablo 9.3. Telesekreter planı özet bilgileri**

VMailPlan					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	2411	72,3	72,3	72,3
	yes	922	27,7	27,7	100,0
	Total	3333	100,0	100,0	

**Tablo 9.4. Çapraz tablo kayıp müşteri ile uluslararası plan**

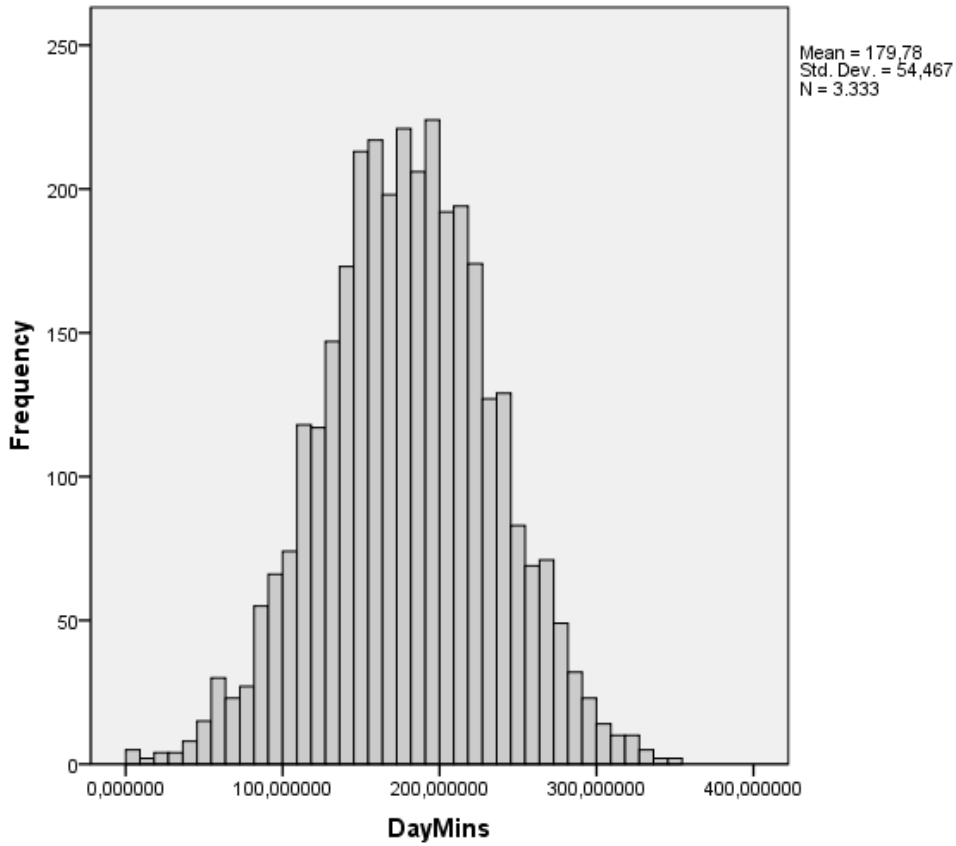
Churn * IntlPlan Crosstabulation					
			IntlPlan		Total
			no	yes	
Churn	false.	Count	2664	186	2850
		% within IntlPlan	88,5%	57,6%	85,5%
	true.	Count	346	137	483
		% within IntlPlan	11,5%	42,4%	14,5%
Total		Count	3010	323	3333
		% within IntlPlan	100,0%	100,0%	100,0%

**Tablo 9.5. Çapraz tablo kayıp müşteri ile telesekreter planı**

Churn * VMailPlan Crosstabulation					
			VMailPlan		Total
			no	yes	
Churn	false.	Count	2008	842	2850
		% within VMailPlan	83,3%	91,3%	85,5%
	true.	Count	403	80	483
		% within VMailPlan	16,7%	8,7%	14,5%
Total		Count	2411	922	3333
		% within VMailPlan	100,0%	100,0%	100,0%

Nicel değişkenler için betimsel istatistikler ve histogram gibi grafik araçlardan yararlanabiliriz. Ayrıca kayıp müşteri ile ilgilerini de kayıp müşterinin iki farklı seviyesi için sürekli değişkenlerin ortalama değerlerini buralar inceleyebiliriz. Birkaç örnek vermek gerekirse gündüz dakikaların histogramı Şekil 9.3'te verilmiştir. Ortalamanın 179 dakika standart sapmanın ise 54 dakika olduğu ve dağılımın oldukça simetrik olduğu gözlemlenebilir. Tablo 9.6 ise iki farklı kayıp müşteri grubuna göre ortalama gündüz yapılan çağrı sayısını vermektedir. Burada kayıp müşterilerde ortalamada

çok az fazla çağrı yapıldığı görülür. Aynı tablo akşam çağrıları için yapıldığında (Tablo 9.7) benzer bir sonuca varılabilir. Tablo 9.8’de ise müşteri hizmetlerine yapılan çağrı sayısı iki gruba göre ortalama değer olarak ifade edildiğinde, diğer şirketlere geçen müşterilerin ortalama 1,5 kat daha fazla müşteri hizmetlerini aradıkları ortaya çıkmaktadır. Bu da veri madenciliği programının keşfetmesi gereken önemli bir ilişkiye işaret olabilir. Aynı tür bir çıkarıma, kayıp müşteri değişkenini yeniden kodlayarak (1-0) bu sefer müşteri hizmetlerini arama sayısına göre ortalama kayıp müşteri sayısına bakarak yapabiliriz. Tablo 9.8 bu çalışmayı göstermektedir. Buna göre örneğin dokuz kere müşteri hizmetlerini arayan müşterilerin hepsi kaybedilmiştir. Bir kez arayanların % 13’ü kayıpkken, beş kez arayanların % 60’tan fazlası kaybedilmiştir. Bu da bir önceki çıkarımı destekleyen ve mantığa da uyan bir ilişkinin varlığını göstermektedir.



**Şekil 9.3. Gündüz konuşma dakikası histogram**

**Tablo 9.6. Gündüz ortalama çağrı sayısı, kayıp ve kayıp olmayan müşterilere göre**

Report					
DayCalls					
Churn		Mean	N	Std. Deviation	Median
dimension1	false.	100,28	2850	19,801	100,00
	true.	101,34	483	21,582	103,00
	Total	100,44	3333	20,069	101,00

**Tablo 9.7. Gece ortalama çağrı sayısı, kayıp ve kayıp olmayan müşterilere göre**

Report					
EveCalls					
Churn		Mean	N	Std. Deviation	Median
dimension1	false.	100,04	2850	19,958	100,00
	true.	100,56	483	19,725	101,00
	Total	100,11	3333	19,923	100,00

**Tablo 9.8. Müşteri hizmetleri arama sayısı, kayıp ve kayıp olmayan müşterilere göre**

Report					
CustServCalls					
Churn		Mean	N	Std. Deviation	Median
dimension1	false.	1,45	2850	1,164	1,00
	true.	2,23	483	1,853	2,00
	Total	1,56	3333	1,315	1,00

**Tablo 9.9. Müşteri hizmetleri arama sayısına göre ortalama kayıp müşteri (kayıp müşteri 1-0 olarak yeniden kodlanmıştır)**

Report					
churn2					
CustServCalls		Mean	N	Std. Deviation	Median
dimension1	0	,1320	697	,33873	,0000
	1	,1033	1181	,30448	,0000
	2	,1146	759	,31878	,0000
	3	,1026	429	,30374	,0000
	4	,4578	166	,49973	,0000
	5	,6061	66	,49237	1,0000
	6	,6364	22	,49237	1,0000
	7	,5556	9	,52705	1,0000
	8	,5000	2	,70711	,5000
	9	1,0000	2	,00000	1,0000
	Total	,1449	3333	,35207	,0000

## 9.2. Diskriminant Analizi

Bu analiz temelde çoklu regresyon analizi ile benzerdir. Aslında hedef değişkenin yeniden kodlanması ile Diskriminant analizi çoklu regresyon analizi ile tam olarak eşdeğer olmaktadır. Örneğimizde “true” ve “false” olarak adlandırılan “churn” değişkeni zaten Diskriminant analizi için 1, 0 olarak yeniden kodlanacaktır. Bu 1, 0 kodlaması aslında isteğe bağlıdır ve sonuçları değiştirmeyen bir kodlamadır. Yani -10, 20 ya da 5, 15 diye de kodlanabilir. Eğer çoklu regresyon analizi için “0”lar  $-1/n_1$  “1”ler ise  $1/n_2$  olarak kodlanırsa (burada  $n_1$  ve  $n_2$  sırasıyla girdi verisindeki 0 ve 1 sayılarını vermektedir) çoklu regresyon sonucunda çıkan tahminler pozitifse “1” negatifse “0” tahmini yapılabilir (Lattin, Carroll ve Green, 2003). Ya da eşik değeri o değil de başka bir rakam yapılarak tahmin maliyetleri arasındaki fark (daha önceki bölümlerde anlatıldığı gibi) gözetilebilir.

Temelde SPSS programı ile Diskriminant analizi uygulamak için bir dizi bağımsız değişken ve iki değerli bir hedef değişkenin belirtilmesi gerekir. Çoklu regresyonda olduğu gibi program en iyi bağımsız değişken kombinasyonunu çok adımlı yöntemlerle bulabileceği gibi tüm değişkenleri kullanarak da tahmin gerçekleştirebilir. Diskriminant Analizi, çıktı olarak bir Diskriminant fonksiyonu verir ayrıca her gözlem için tahminler (0 ve 1 olma olasılıkları da dahil) hesaplanıp programca kaydedilir. Yeni bir gözlem için Diskriminant fonksiyonu kullanılarak tahmin yapılabilir. Kayıp müşteri verisinde 13 bağımsız değişken ile hedef değişken olan “churn” tahmin edildiğinde tablo 9.10’da olduğu gibi bir Diskriminant fonksiyonu çıkar. Burada orijinal verinin ölçeğinin etkisini azaltmak için standartlaştırılmış Diskriminant fonksiyonu katsayıları daha yararlıdır. Bunlar da Tablo 9.11’de verilmiştir. Bu tablo aynı zamanda kayıp müşteri tahmininde değişkenlerin göreceli önemini de bize verir.



**Tablo 9.10. Diskriminant fonksiyon katsayıları**

<b>Canonical Discriminant Function Coefficients</b>	
	Function
	1
AccountLength	,001
IntlPlan	2,225
VMailPlan	-1,009
VMailMessage	,014
DayMins	,009
DayCalls	,002
EveMins	,005
EveCalls	,001
NightMins	,002
NightCalls	,000
IntlMins	,059
IntlCalls	-,060
CustServCalls	,429
(Constant)	-4,597
Unstandardized coefficients	

**Tablo 9.11. Diskriminant fonksiyon katsayıları standartlaştırılmış halde**

Standardized Canonical Discriminant Function Coefficients	
	Function
	1
AccountLength	,030
IntlPlan	,636
VMailPlan	-,449
VMailMessage	,196
DayMins	,497
DayCalls	,048
EveMins	,243
EveCalls	,012
NightMins	,118
NightCalls	,009
IntlMins	,164
IntlCalls	-,148
CustServCalls	,552

Tablo 9.11 daha önce açınsayıcı analizler ile yapılan çeşitli çıkarımları desteklemektedir. Örneğin uluslararası plan sahipleri ve daha çok müşteri hizmetlerini arayan müşteriler şirketi terk etmeye meyilli olmaktadır. Tam tersine telesekreter planı sahipleri de şirkette kalmaya meyilli olmaktadır. Gün içinde çok dakika kullanmanın da aynı şekilde kuvvetli bir şekilde şirketi terk etmeyi etkilediği saptanabilir. Arama sayısının ise etkisinin az olduğu görülmektedir. Belki de arama sayılarını içeren değişkenleri analizden çıkarmak çok daha uygun olabilir. Çeşitli sınıflama hatalarını hesaplamak için Tablo 9.12’de verilen hata matrisini kullanabiliriz.

**Tablo 9.12. Diskriminant analizi hata matrisi**

Classification Results <sup>a</sup>					
churn2			Predicted Group Membership		Total
			,00	1,00	
Original	Count dimension2	,00	2263	587	2850
		1,00	129	354	483
	% dimension2	,00	79,4	20,6	100,0
		1,00	26,7	73,3	100,0
a. 78,5% of original grouped cases correctly classified.					

Tabloya göre çeşitli hatalar şu şekilde hesaplanabilir:

- Genel Doğru Tahmin Oranı:  $(2263+354)/3333 = \%78,51$  gözlem doğru tahmin edilmiştir
- Toplam 941 pozitif sınıflama içinde 354 adet gözlem doğru sınıflanmıştır dolayısıyla pozitif tahminlerde doğruluk oranı  $354/941 = \%37,62$ 'dir.
- Toplam 2392 negatif sınıflama içinde 2263 tanesi doğru sınıflamadır yani negatif tahminlerde doğruluk oranı  $2263/2392 = \%94,61$  olarak gerçekleşmiştir.
- Pozitif olduğu bilinen 483 gözlemin 354'ü doğru sınıflanmıştır dolayısıyla "sensitivity"  $354/483 = \%71,81$  olarak gerçekleşmiştir.
- Negatif olduğu bilinen 2850 gözlemin 2263 adeti doğru sınıflanmıştır dolayısıyla "specificity"  $2263/2850 = \%79,40$  olmuştur.

Burada program çıktısı kullanılarak 941 kişi ile görüşüldüğünü varsayalım. Bunların 354'ü gerçekten kayıp müşteridir. Rastgele seçilen 941 kişide ise kayıp müşteri sayısı (kayıp müşteri oranı  $483/3333=14,50\%$  olduğunda göre) 136 kişi olacaktır.  $354/136=2,60$  ise bu modelin kazanımı (lift) olarak düşünülebilir. Diskriminant analizinde eşik değeri değiştirilerek yukarıdaki hata oranları manipüle edilebilir. Genel hata oranı artar ancak toplam maliyet düşebilir. Ayrıca bağımsız değişkenlerin hepsini kullanmak yerine bir kısmı ayıklanabilir bu da modelin anlaşılabilirliğini arttıracaktır.

### 9.3. Lojistik regresyon

İkili değişkenleri tahmin etmek için istatistikî paketlerde sıkça kullanılan diğer opsiyon da lojistik regresyon opsiyonudur. Diskriminant analizi gibi tüm değişkenler kullanılabileceği gibi en iyi bağımsız değişkenlerin program tarafından seçilmesi de sağlanabilir. Tablo 9.13 bu analizde ortaya çıkan hata matrisini vermektedir.

**Tablo 9.13. Lojistik regresyon analizi hata matrisi**

Classification Table <sup>a</sup>					
Observed			Predicted		
			churn2		Percentage Correct
			,00	1,00	
Step 1	churn2	,00	2771	79	97,2
		1,00	379	104	21,5
	Overall Percentage				86,3
a. The cut value is ,500					

Burada 2875 adet tahmin doğru sınıflandırılmıştır. Yani genel doğru tahmin oranı % 86,26 olarak gerçekleşmiştir. Yalnız lojistik regresyon modeli 183 gibi oldukça az sayıda gözlemi kayıp olarak sınıflamıştır. Bu şekilde % 56,83 ile pozitif tahminler içinde doğru tahmin oranı yükselmiştir. Negatif tahminler içinde doğru tahmin oranı ise % 88 olarak gerçekleşmiştir. Kazanım ise  $104/26 = 4$  olarak hesaplanır. Kazanım daha yüksektir ama tes-

pit edilmeyen ve ulaşlamayan kayıp müşterilerden oluşacak zarar daha büyük olabilir. Sensitivity bu durumda  $104/483=21\%$  ile düşük bir seviyededir. Diskriminant analizinde olduğu gibi eşik değiştirilerek sınıflama sistemi değiştirilebilir ve bu şekilde kazanım ve çeşitli sınıflama hataları da değişebilir. Bu şekilde lojistik regresyon seçeneklerinden eşik olasılık değeri öntanımlı değer olan 0,5'ten 0,3'e düşürülürse Tablo 9.14'teki gibi bir hata matrisi ortaya çıkar.

**Tablo 9.14. Lojistik regresyon analizi hata matrisi (eşik olasılık 0,3)**

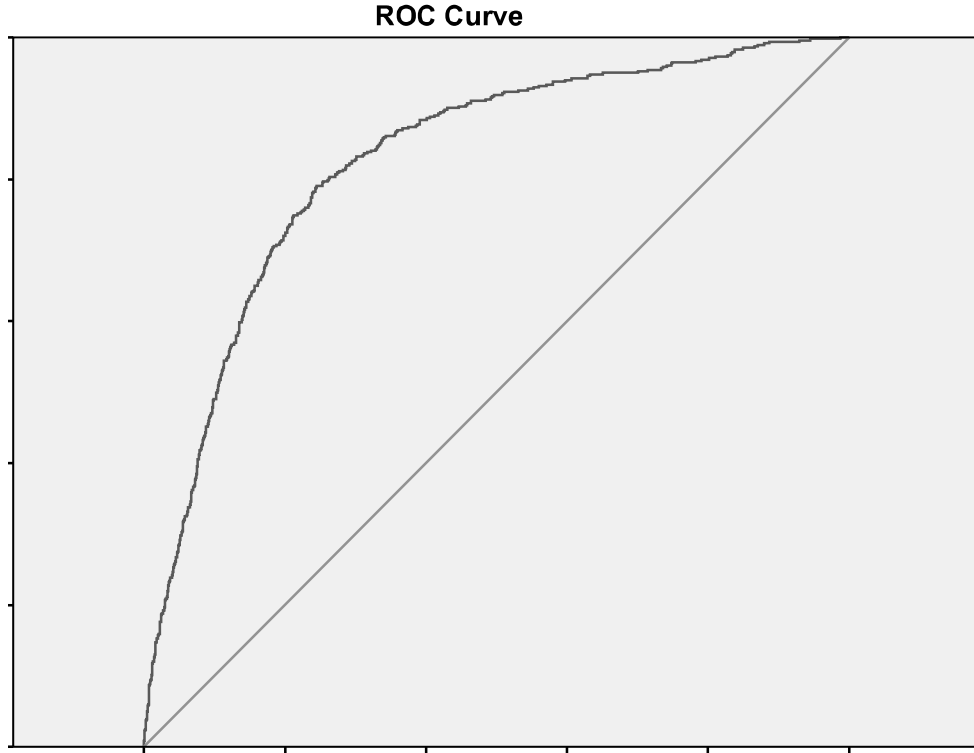
Classification Table <sup>a</sup>					
Observed			Predicted		
			churn2		Percentage Correct
			,00	1,00	
Step 1	churn2	,00	2619	231	91,9
		1,00	254	229	47,4
	Overall Percentage				85,4
a. The cut value is ,300					

Genel hata oranı çok az artmasına rağmen sensitivity, yani kayıp olduğu bilinen müşteriler içinde doğru tespit oranı  $229/483=47\%$ 'ye ulaşmıştır. Burada kazanım hesaplanırsa 460 müşteri ile temasa geçildiği düşünülebilir. 460 müşteri rastgele seçildiğinde 67 kişi kayıp müşteri olacaktır. Bu durumda kazanım  $229/67=3,42$ 'dir. Aslında 4'ten düşük olsa da maliyetler göz önüne alınırsa daha avantajlı bir durum ortaya çıkabilir.

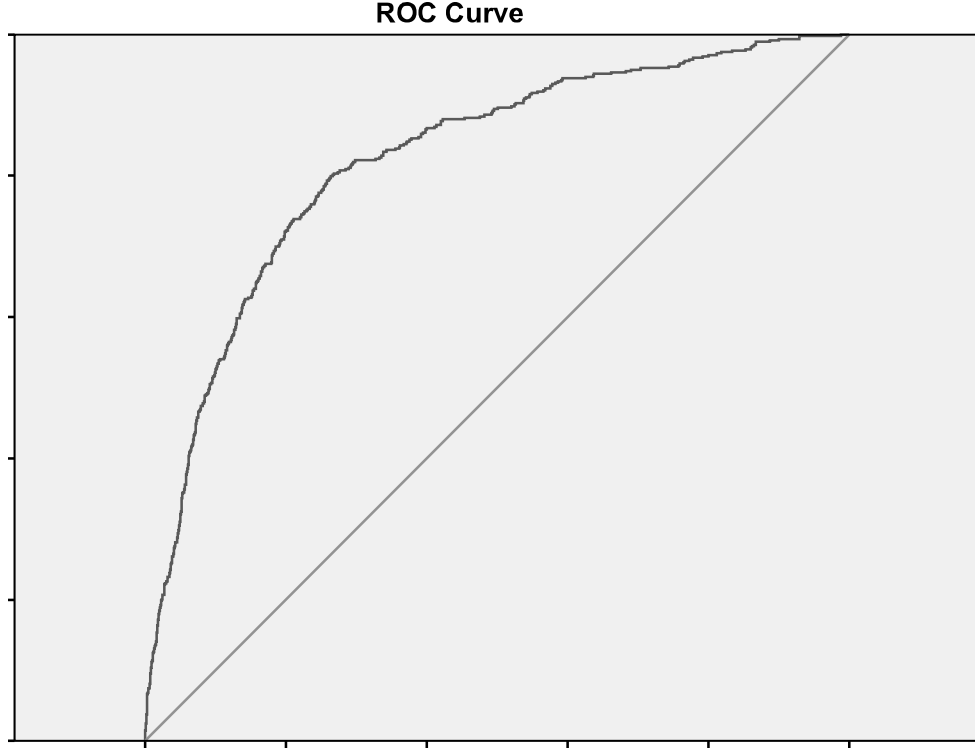
#### 9.4. Model Karşılaştırması

Genel hata oranına bakıldığında (eşik değerler değiştirilmediğinde) lojistik regresyon Diskriminant analizine üstün gözükmektedir. Ancak sensitivity lojistik regresyonda oldukça düşük çıkmıştır. Aslında eşik değerlerin değişebileceği düşünüldüğünde iki metodu karşılaştırmak için birikimli

kazanım grafiklerini (ROC) üst üste görmek ve grafiğin altındaki alanı hesaplamak faydalı olacaktır. SPSS tarafından üretilen ROC grafikleri eğri altındaki alanı da vermektedir. Bu şekilde çok az fark olmakla beraber, genel olarak tüm seviyelerde, Diskriminant analizinin üstün olduğu görülmektedir.

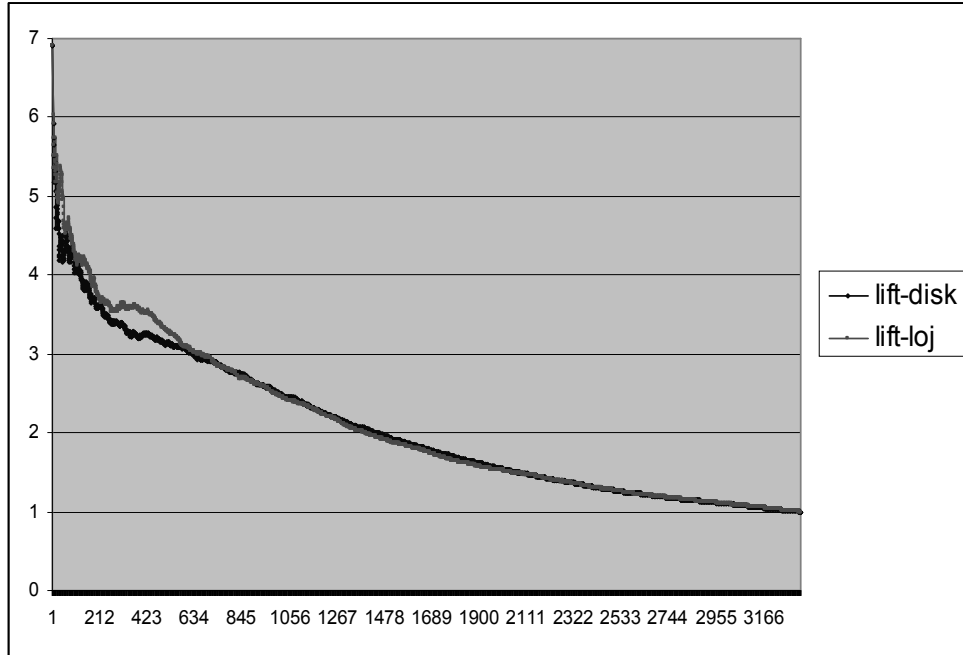


**Şekil 9.4. ROC Diskriminant Analizi (eğri altındaki alan 0,826)**

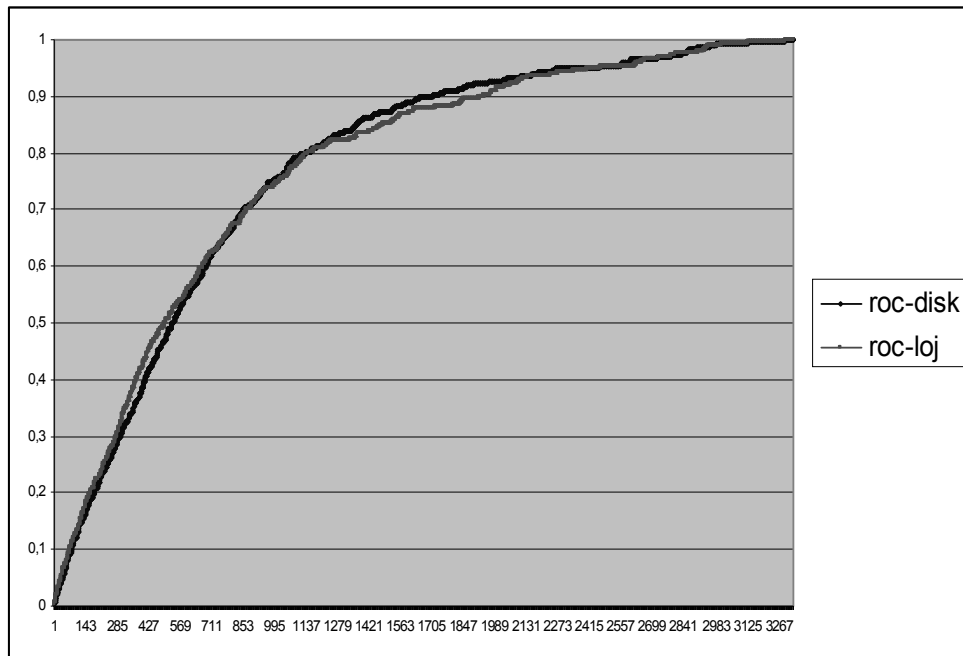


**Şekil 9.5. ROC Lojistik Regresyon Analizi (eğri altındaki alan 0,823)**

Bir başka ilginç nokta da temas edilen müşteri sayısına göre belli noktalarda bir metodun diğerine üstün olabileceğidir. Bu anlamda iki grafiği üst üste getirerek hangi noktalarda bir metodun diğerine üstün olduğu görülebilir. Kazanım ve birikimli kazanım (ROC) grafikleri için tek gereken her gözlem için Diskriminant skorları ve lojistik regresyon analizinden elde edilen olasılıklardır. Bunlar ayrı ayrı büyükten küçüğe sıralanırsa, en yüksek skora sahip gözlemden başlayarak 3.333 sayısına kadar kazanım (lift) ve birikimli kazanım hesaplanabilir. Bu şekilde iki metot üst üste konulduğunda lift ve ROC grafikleri Şekil 9.6 ve 9.7 gibi ortaya çıkmaktadır. Burada aslında 700 gibi bir müşteri sayısına kadar lojistik regresyonun üstün olduğu daha sonra ise Diskriminant analizinin üstün geldiği anlaşılmaktadır.



**Şekil 9.6. İki farklı analiz için lift (kazanım) grafiği**



**Şekil 9.7. İki farklı analiz için birikimli kazanım (ROC) grafiği**





## 10. BÖLÜM:

### VAKA ANALİZİ: YAZ OKULU

Bu bölümde örnek veri olarak bir üniversite yaz okulunda 2006 Yaz döneminden beri verilen bir istatistik dersinden alınan sonuçların öğrencilerin çeşitli özellikleri kullanılarak tahmin edilip edilemeyeceği üzerinde bir örnek verilecektir. Bunun için dersi 2010 Yaz dönemine kadar alan 114 öğrenci ile ilgili veri seti kullanılmıştır. Öğrenci kayıt sisteminde şu veriler bulunmaktadır:

- Öğrenci numarası
- İsim
- Soyadı
- Tekrar durumu (dersi daha önce almış mı)
- Bölüm (Yaz okulu özel öğrenci, A, B, C olarak adlandırılmıştır)
- Sömestr
- Öğrenci statüsü (özel, normal) –özel öğrenciler zaten bölüm değişkeninde yaz okulu olarak belirtildiği için bu değişken aslında fazladır
- Genel statü (sınamalı, tekrar, normal, son sınıf)
- Harf notu (AA, BA, BB, CB, CC, DC, DD, F aynı notların rakamsal ifadesi 4,0 - 3,5 - 3,0 - 2,5 - 2,0 - 1,5 - 1,0 - 0)

Bunların dışında türetilmiş bir değişken olan cinsiyet de kullanılabilir. Öğrenci numarası, isim ve soyadı analizde kullanılmaya değer olmayan, bir önceki analizdeki telefon numarası ve alan koduna çok benzeyen değişkenler olduğundan analiz dışında tutulacaktır. Böylece harf notuna olası etkisi olabilecek altı değişken ayırt edilebilir. Bunların biri (sömestr) dışında hepsi nitel değişkenlerdir. Sömestr bile, “0” değeri sadece özel öğrencilere ayrıldığından yarı nitel yarı nicel bir özelliğe sahiptir. Analizden önce “F” alan dört kişi (muhtemelen kayıt olup gelmeyen) ve tek bir kişi tarafından temsil

edilen bölümler analizden çıkarılmıştır. Bu tek bölümler dahil edilse, örneğin karar ağacı bölüme göre ayırarak bazı gözlemleri mükemmel tahmin etme olanağına sahip olur, “fazla” öğrenme gerçekleşirdi. Böylece 101 gözlem analize dahil edilmiştir.

### 10.1. Açımsayıcı Veri Analizi

Değişkenlerin harf notunu nasıl etkilediğini görmek için öncelikle açımsayıcı veri analizi faydalı olacaktır. Bölümler elendikten sonra yaz okulu dışında isimleri gizlenerek; yaz okulu, A, B ve C olarak ortaya çıkmıştır. Nitel değişkenlerin harf notu ile ilişkisini görmek için çeşitli seviyelerinin harf notu ortalamalarına bakmak faydalı olabilir. Buna göre:

- Kız öğrencilerin not ortalaması 3,36 erkeklerin 3,25 olarak ortaya çıkmıştır
- Dersi tekrar edenlerin not ortalaması 3,18 etmeyenlerin 3,36’dır
- Son sınıf öğrencilerin not ortalaması en yüksektir, bunu takip eden normal, başarısız ve sınamalı statü gruplarıdır
- Departmanlara bakıldığında Yaz Okulu grubu en düşük A grubu en yüksek ve aralarında 0,40 puan kadar fark vardır
- Sömestr durumu 0 ile 16 arasında değişmektedir. Zayıf da olsa sömestr yükseldikçe harf notunun düştüğü görülür. “0” sömestrde olanların hepsi yaz okuluna gelen özel öğrencilerdir. “0”lar çıkarılarak harf notu ile korelasyon hesaplandığında -0,37 çıkmaktadır.

### 10.2. Çoklu Regresyon

Çoklu regresyondan önce çeşitli istatistiki programlar otomatik olarak nitel verilerin kukla değişkenler kullanımı ile, istatistiksel analizlere uygun hale getirebilse de bir kısmı hata mesajı verebilir. Bu nedenle yukarıdaki değişkenleri 1-0 kukla değişkenler ile değiştirmek gerekecektir. Bu şekilde aşağıdaki gibi kukla değişkenler oluşturulabilir:

- Erkek: 1 Erkek 0 Kız
- Tekrar: Dersi tekrarlayan 1, ilk defa alan 0
- Yaz: 1 ise yaz okulu özel öğrenci
- A: 1 ise A departmanı
- B: 1 ise departman B (C departmanı diğerlerinin 0 olduğu duruma denk gelir)
- Sem: Sömestr sayısı
- Normal: 1 ise normal
- Son sınıf: 1 ise son sınıf
- Başarısız: 1 ise başarısız (0 durumu sınamalı duruma denk gelir)

Bu dokuz değişkenin birbirleri arasındaki ilişkileri incelemek için korelasyon tablosuna bakılabilir. Tablo 10.1 bunu vermektedir. Burada gözlemlenebilen ilişkilerden biri sömestr ile normal son sınıf gibi statülerin yüksek korelasyon gösterdiği. Dolayısıyla sömestr değişkeni dururken diğerleri fazla olabilir. Ayrıca sömestr ile yaz okulu yüksek korelasyon göstermektedir, çünkü yaz okulu öğrencileri zaten “0” sömestr gözükmemektedir. Tekrar değişkeni ile B bölümü ve sömestr da yüksek korelasyon göstermektedir. B bölümü öğrencileri genelde dersi tekrar etmektedir, yüksek sömestr sayısındaki öğrencilerin de ders tekrar etmeleri elbette mantıklıdır. Burada normal, son sınıf, başarısız kategorilerini kaldırıp (nitekim bir kısmı çok az gözleme sahiptir) çoklu regresyon yapmak mantıklı görünmektedir.

**Tablo 10.1. Değişkenler arasındaki korelasyon**

	<i>Erkek</i>	<i>Tekrar</i>	<i>Yaz</i>	<i>A</i>	<i>B</i>	<i>Sem</i>	<i>Normal</i>	<i>Son sınıf</i>	<i>Başarısız</i>
Erkek	1,00								
Tekrar	-0,13	1,00							
Yaz	0,19	-0,51	1,00						
A	0,06	-0,17	-0,41	1,00					
B	-0,01	0,67	-0,36	-0,43	1,00				
Sem	-0,24	0,65	-0,83	0,10	0,41	1,00			
Normal	0,22	-0,59	0,76	-0,06	-0,44	-0,87	1,00		
Son sınıf	-0,29	0,50	-0,68	0,00	0,40	0,80	-0,90	1,00	
Başarısız	0,15	0,13	-0,12	0,08	0,10	0,10	-0,16	-0,24	1,00

Çoklu regresyon analizi sonucu Tablo 10.2'deki gibi ortaya çıkmıştır. Beklenenin aksine tekrar değişkeni pozitif katsayıya sahip olmuştur ama bu sömestr ve yaz değişkenleriyle olan yüksek korelasyondan meydana gelmektedir.

**Tablo 10.2. Çoklu Regresyon**

<i>Regresyon İstatistikleri</i>	
Çoklu R	0,48
R Kare	0,23
Düzeltilmiş R Kare	0,18
Standart Hata	0,64
Gözlem Sayısı	101

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Anlamlılık F</i>
Regresyon	6	11,40	1,90	4,65	0,00
Kalıntı	94	38,36	0,41		
Toplam	100	49,76			

	<i>Katsayılar</i>	<i>Standart Hata</i>	<i>t Stat</i>	<i>P-değeri</i>
Sabit	3,99	0,36	11,03	0,00
Erkek	-0,15	0,14	-1,04	0,30
Tekrar	0,15	0,21	0,68	0,50
Yaz	-0,79	0,38	-2,07	0,04
A	0,23	0,24	0,96	0,34
B	-0,44	0,26	-1,70	0,09
Sem	-0,06	0,03	-2,09	0,04

Model elbette geliştirilebilir ancak bu şekliyle notları tahmin modeli standart hata olarak 0,64 vermektedir. Yani ortalama 0,64 puan (yaklaşık bir harf notu) hata yapmaktadır. Aslında modelin çıkardığı sonuç gerçekte bir üst ya da alt harf notuna yuvarlanırsa aşağıdaki gibi bir durum elde edilir:

- 32 kişi tam tahmin
- 41 kişi artı eksi bir harf notu hata
- 26 kişi artı eksi iki harf notu hata
- 2 kişi eksi 3 harf notu hata

Dolayısıyla bu modeli kullanacak bir uzman sistemin yaz okulu ders seçimi yapan öğrencilere (ayrıca belki bizde olmayan genel not ortalaması ve daha önce alınan derslerdeki notları da girdi olarak kullanan daha gelişmiş bir model ile) % 75 ihtimalle en fazla bir harf notu hata yapan bir tavsiyede bulunması oldukça faydalı olabilir.

### 10.3. Diskriminant Analizi ile “AA” Notu Tahmini

Bazı öğrenciler sadece AA alma ihtimalleri ile ilgilenebilirler. Bu durumda Diskriminant analizi ya da lojistik regresyon kullanılabilir. Hedef değişken not AA ise 1 değilse 0 olarak dönüştürülebilir. Bu şekilde yapılan analizde Tablo 10.3'teki gibi bir Diskriminant fonksiyonu çıkmıştır. Bu şekilde yeni bir gözlem için fonksiyon sonucu hesaplanıp 1'den fazla ise AA notu alacağı tahmin edilebilir.

**Tablo 10.3. Diskriminant Fonksiyonu**

<b>Canonical Discriminant Function Coefficients</b>	
	Function
	1
erkek	-,719
tekrar	-,730
A	2,018
B	,053
Sem	-,059
Normal	,101
Sonsınıf	1,094
(Constant)	-,160
Unstandardized coefficients	

Tahmin sonuçları açısından bakıldığında şu genel tablo ortaya çıkar:

- 36 kişi AA olarak tahmin edilmiştir aslında 23'ü gerçekten almıştır, yani pozitif tahminler içinde doğru tahmin oranı % 64
- Genel doğru tahmin oranı ise % 75'tir
- 35 gerçekte AA alan içinde 23'ü doğru tahmin edilmiş dolayısıyla "sensitivity" % 66 olarak gerçekleşmiştir.
- Specificity ise % 80'dir.
- Kazanım 1,84 olarak hesaplanır (23 kişi / rastgele seçilen 36 kişide olabilecek AA sayısı)

#### 10.4. Karar Ağaçları C5.0 ile

Aynı işlemi sınırlı sayıda gözlem için bedava olarak sunulan ve C5.0 algoritmasını uygulayan See5 isimli programla denemek faydalı olacaktır.

Bunun için düz metin şeklinde hazırlanan \*.dat sonlu bir veri dosyası ve değişkenlerin ve hedef değişkenin tanımlandığı \*.names sonlu bir başka dosya gerekmektedir. Aslında bir de doğrulama seti varsa \*.test sonuyla programa girdi olarak kullanılabilir. Temelde öntanımlı opsiyonlarla çalıştırıldığında program Şekil 10.1'deki çıktıyı verir. Sadece departman kullanarak % 41,6'lık bir tahmin doğruluk oranı elde etmektedir. Bu da aslında bir önceki regresyon örneğinde elde edilen %32'den (tam doğru tahmini değerlendirirsek) daha iyi bir sonuçtur.

```

Results for Student
File Edit

See5 [Release 2.08] Tue Sep 27 19:23:58 2011

Class specified by attribute `grade`

Read 101 cases (7 attributes) from Student.data

Decision tree:

department = SUMMER: BA (26/19)
department in {A,C}: AA (47/22)
department = B: BB (28/18)

Evaluation on training data (101 cases):

Decision Tree
-----
Size      Errors
3      59(58.4%)  <<

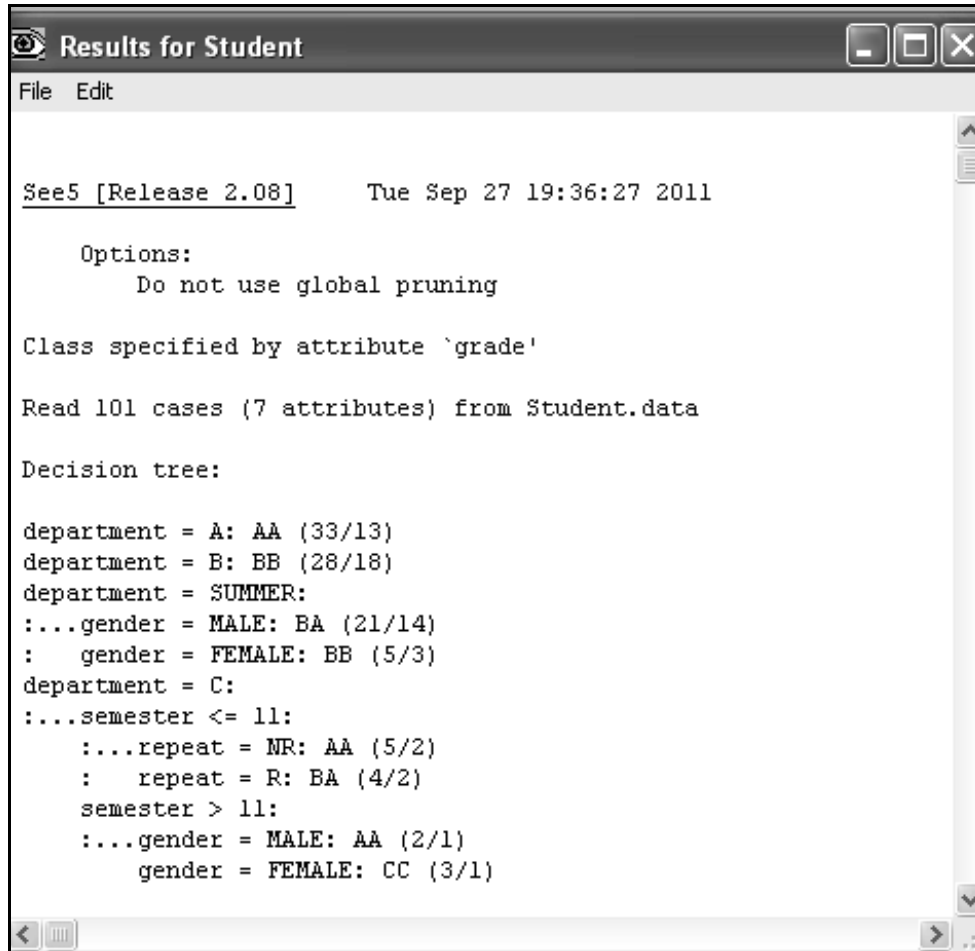
(a)  (b)  (c)  (d)  (e)  <-classified as
----  -
25    5    5           (a): class AA
11    7    4           (b): class BA
6     6   10           (c): class BB
2     3    2           (d): class CB
3     5    7           (e): class CC

```

Şekil 10.1. See5 Çıktısı, 59 hata 42 doğru tahmin



Program öntanımlı opsiyonda “global budama” (global pruning) seçilmiştir. Bu kaldırılarak daha komplike bir ağaç ortaya çıkabilir, örneğin Şekil 10.2’de olduğu gibi. Burada doğru tahmin oranı Şekil 10.3’de görüldüğü gibi % 46,5’a ulaşmış ancak ağaç daha karmaşık ve çok seviyeli bir hale gelmiştir. Tahmin oranındaki ufak iyileşmenin, eklenen karmaşıklığı haklı çıkarıp çıkarmayacağı karar vericinin cevap vermesi gereken bir sorudur. İlk ağacın sadece üç son dalı varken ikincinin ise sekiz adet son dalı vardır. İkinci ağaç karar vermek için dört değişken kullanırken ilki ise sadece bir değişken kullanmaktadır.



```

Results for Student
File Edit

See5 [Release 2.08]      Tue Sep 27 19:36:27 2011

Options:
  Do not use global pruning

Class specified by attribute 'grade'

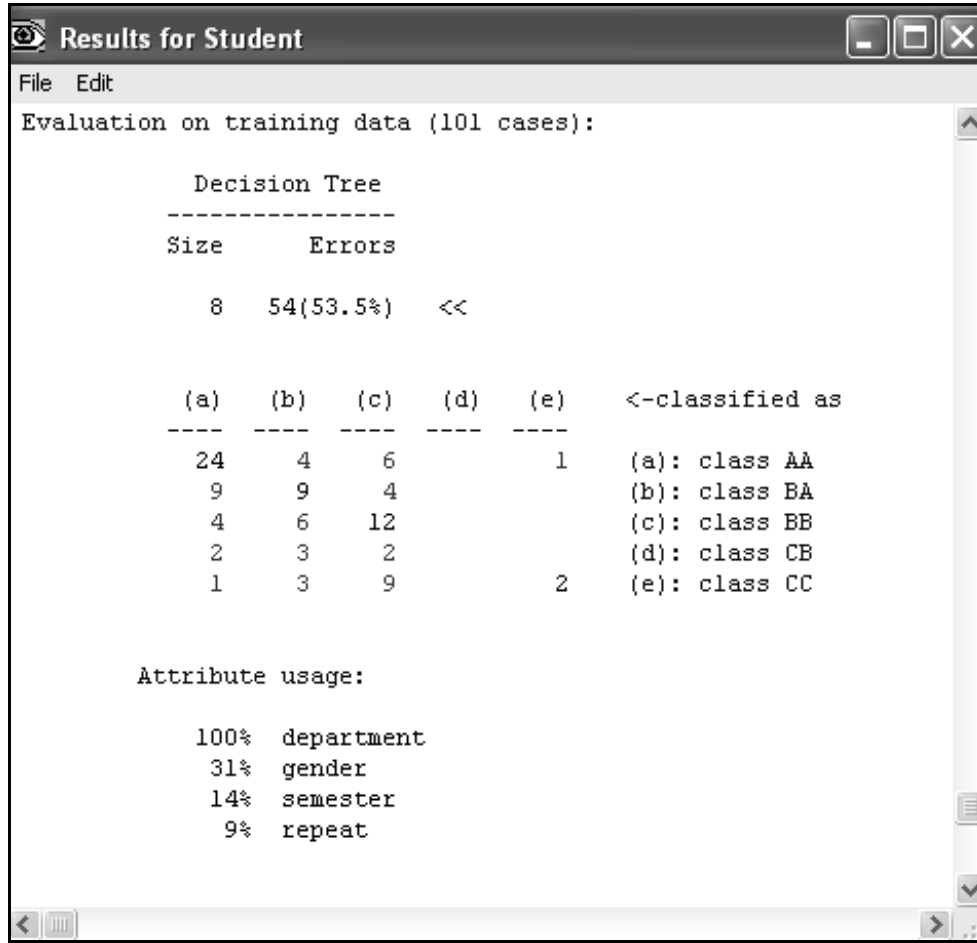
Read 101 cases (7 attributes) from Student.data

Decision tree:

department = A: AA (33/13)
department = B: BB (28/18)
department = SUMMER:
...gender = MALE: BA (21/14)
:   gender = FEMALE: BB (5/3)
department = C:
...semester <= 11:
...repeat = NR: AA (5/2)
:   repeat = R: BA (4/2)
semester > 11:
...gender = MALE: AA (2/1)
      gender = FEMALE: CC (3/1)

```

Şekil 10.2. See5 Çıktısı, karar ağacı, global budama seçilmeden



**Şekil 10.3. See5 Çıktısı, hata oranları**

Bir başka seçenek olan “pruning CF – pruning Confidence Level” (budama güven seviyesi) öntanımlı olarak %25’tir. Bu yükseltilerek daha karmaşık bir ağaç elde edilebilir. Örnek olarak Şekil 10.4 ve 10.5’te bu seviye %75 olarak verilmiştir. Burada hata oranı düşmüş doğru tahmin oranı % 51,5’ kadar ulaşmıştır. Daha yükseltildiğinde sonuç değişmemektedir.

See5 [Release 2.08] Tue Sep 27 19:50:41 2011

Options:

Do not use global pruning  
Pruning confidence level 75%

Class specified by attribute 'grade'

Read 101 cases (7 attributes) from Student.data

Decision tree:

```

department = A: AA (33/13)
department = SUMMER:
:...gender = MALE: BA (21/14)
:  gender = FEMALE: BB (5/3)
department = C:
:...semester <= 11:
:  :...repeat = NR: AA (5/2)
:  :  repeat = R: BA (4/2)
:  semester > 11:
:  :...gender = MALE: AA (2/1)
:  :  gender = FEMALE: CC (3/1)
department = B:
:...general = ONPROBATION: BB (0)
:  general = NORMAL: AA (1)
:  general = REPEATING: BB (2/1)
:  general = SENIOR:
:...semester > 11: CC (6/3)
:  semester <= 11:
:  :...gender = FEMALE: AA (7/4)
:  :  gender = MALE:
:  :...semester > 9: BA (5/3)
:  :  semester <= 9:
:  :  :...semester <= 8: CC (5/2)
:  :  :  semester > 8: BB (2)

```

Şekil 10.4. See5 Çıktısı, karar ağacı, budama güven seviyesi %75

```

Evaluation on training data (101 cases):

      Decision Tree
      -----
Size      Errors

      14    49 (48.5%)    <<

(a)      (b)      (c)      (d)      (e)      <-classified as
-----
      28         5         1                1      (a): class AA
      11        11                12          (b): class BA
       6         7         5                4      (c): class BB
       2         4                1          (d): class CB
       1         3         3                8      (e): class CC

Attribute usage:

100% department
 50% gender
 39% semester
 28% general
  9% repeat

```

**Şekil 10.5. See5 Çıktısı, karar ağacı, hata oranları, budama güven seviyesi %75**

Maksimum karmaşıklıkta bir ağaç elde etmek için ise bir başka opsiyon olan “minimum number of cases” (minimum gözlem sayısı) opsiyonu değiştirilebilir. Öntanımlı opsiyonda bu 2’dir, yani son dallarda en az 2 gözlem olması gerekir. Bu “1” e düşürülürse ise en karmaşık ağaç ortaya çıkar. Bu ağaç 36 son dala sahip ancak hata oranı % 34,7 ile en düşük ağaçtır. Burada gereğinden fazla karmaşıklık ve eğitim verisini fazla öğrenme problemi olduğu elbette çok açıktır. Bu en karmaşık ağaç aşağıda verilmiştir.

```

department = SUMMER:
...gender = MALE: BA (21/14)
: gender = FEMALE: BB (5/3)
department = A:
...general = ONPROBATION: BB (1)
: general = REPEATING:
: ...repeat = NR: AA (1)
: : repeat = R: BA (1)
: general = NORMAL:
: ...semester <= 3:
: : ...gender = MALE: BB (1)
: : : gender = FEMALE: AA (1)
: : semester > 3:
: : ...repeat = R: AA (1)
: : : repeat = NR:
: : : ...gender = MALE: AA (7/3)
: : : : gender = FEMALE: BA (1)
: general = SENIOR:
: ...repeat = NR: AA (12/3)
: : repeat = R:
: : ...gender = MALE:
: : : ...semester <= 11: BA (1)
: : : : semester > 11: AA (3/1)
: : : gender = FEMALE:
: : : ...semester <= 11: AA (2)
: : : : semester > 11: BA (1)
department = C:
...semester > 11:
: ...gender = MALE:
: : ...repeat = NR: AA (1)
: : : repeat = R: CB (1)
: : gender = FEMALE:
: : : ...repeat = NR: CC (2)
: : : : repeat = R: AA (1)
: semester <= 11:
: ...repeat = R:
: : ...gender = MALE: BB (1)
: : : gender = FEMALE: BA (3/1)
: : repeat = NR:
: : ...gender = MALE: BA (1)
: : : gender = FEMALE:
: : : ...semester <= 10: AA (3)
: : : : semester > 10: BB (1)
department = B:
...general = ONPROBATION: BB (0)
: general = NORMAL: AA (1)
: general = REPEATING:
: ...semester <= 10: CC (1)
: : semester > 10: BB (1)
: general = SENIOR:
: ...semester > 11:
: : ...gender = MALE: CC (3/1)
: : : gender = FEMALE:
: : : ...semester <= 13: CC (1)
: : : : semester > 13: BB (2/1)
: semester <= 11:
: ...gender = FEMALE: AA (7/4)
: : gender = MALE:
: : : ...semester > 9: BA (5/3)
: : : : semester <= 9:
: : : : ...repeat = NR: CC (1)
: : : : : repeat = R:
: : : : : ...semester > 8: BB (2)
: : : : : : semester <= 8:
: : : : : : ...semester <= 7: BB (1)
: : : : : : : semester > 7: CC (3/1)

```

## KAYNAKÇA

### A: MAKALELER, KİTAPLAR, BİLDİRİLER

Akküçük, U. (2004). *Nonlinear Mapping: Approaches Based on Optimizing an Index of Continuity and Applying Classical Metric MDS on Revised Distances*. Doktora Tezi, Rutgers University, New Jersey, ABD. (UMI no AAT 3148774, bkz. URL19)

Akküçük, U. (2009). “Bir Çok Boyutlu Ölçekleme Tekniği Olarak Torgersen Ölçekleme Yöntemi ve Temel Bileşenler Analizi ile Karşılaştırılması”, *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 25: 311-322.

Akküçük, U. ve Carroll, J. D. (2006a). “PARAMAP vs. Isomap: A Comparison of Two Nonlinear Mapping Algorithms,” *Journal of Classification*, 23:2, 221-254.

Akküçük, U. ve Carroll, J. D. (2006b). American Statistical Association Joint Statistical Meetings (JSM) dahilinde “Parametric Mapping (PARAMAP): An Approach to Nonlinear Mapping,” 6-10 Ağustos 2006, Seattle, ABD. [CD-ROM olarak yayınlandı, ss. 1980-1986]

Akküçük, U. ve Carroll, J. D. (2010). “Nonlinear Mapping Using a Hybrid of PARAMAP and Isomap Approaches,” Hermann Locarek-Junge ve Claus Weihs (Editörler), *Classification as a Tool for Research*: 371-380. Berlin-Heidelberg-New York: Springer.

Akküçük, U. (2011a). *Görselleştirme Teknikleri: Pazarlama ve Patent Analizi Örnek Uygulamaları İle*. İstanbul: Yalın Yayıncılık.

Akküçük, U. (2011b). “A Study on the Competitive Positions of Countries Using Cluster Analysis and Multidimensional Scaling”, *European Journal of Economics, Finance, and Administrative Sciences*, 37, 17-26.

Allison, P. D. (2002). *Missing Data*. Thousand Oaks: Sage Publications.

- Arabie, P. ve Hubert, L. (1985). “Comparing Partitions”, *Journal of Classification*, 2, 193-218.
- Breiman, L., Friedman, J., Olshen, R. ve Stone, C. (1984). *Classification and Regression Trees*. Boca Raton: Chapman and Hall/CRC Press.
- Calinski, T., ve Harabasz, J. (1974). “A Dendrite Method for Cluster Analysis”, *Communications in Statistics* , 3, 1-27.
- Giudici, P. (2003). *Applied Data Mining: Statistical Methods for Business and Industry*. John Wiley & Sons.
- Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken: John Wiley & Sons.
- Lattin, J., Carroll, J. D. ve Green, P. E. (2003). *Analyzing Multivariate Data*. Pacific Grove: Duxbury.
- Pyle, D. (1999). *Data Preparation for Data Mining*. San Fransisco: Morgan Kaufmann.
- Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning*. San Fransisco: Morgan Kaufmann.
- Rand, W. M. (1971). “Objective Criteria for the Evaluation of Clustering Methods”, *Journal of the American Statistical Association* , 66, 846-850.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, ss. 581-592.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall/CRC.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

**B: AĞ ADRESLERİ**

URL1: American Marketing Association (AMA) pazarlama terimleri sözlüğü

[http://www.marketingpower.com/\\_layouts/Dictionary.aspx?dLetter=D](http://www.marketingpower.com/_layouts/Dictionary.aspx?dLetter=D)

URL2: Gartner bilişim terimleri sözlüğü

[http://www.gartner.com/technology/it-glossary/#3\\_0](http://www.gartner.com/technology/it-glossary/#3_0)

URL3: Crisp-DM adımları

[http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=/com.ibm.im.easy.doc/c\\_dm\\_process.html](http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=/com.ibm.im.easy.doc/c_dm_process.html)

URL4: 2005 Capital 500 firma listesi

<http://www.capital500.net/capital/ca05.htm>

URL5: 2008 Capital 500 firma listesi

<http://www.capital.com.tr/Siralamalar/Html/2009.htm>

URL6: Piyasa getiri verileri

<http://finance.yahoo.com/intlindices>

URL7: Carnegie Mellon Data and Story Library

<http://lib.stat.cmu.edu/DASL/>

URL8: Kredi başvuru geri çevrilme verisi

<http://lib.stat.cmu.edu/DASL/Stories/mortgagerefusals.html>

URL9: University of California at Irvine (UCI) yapay öğrenme veritabanları

<http://archive.ics.uci.edu/ml/>

URL10: Otomobil verisi

<http://archive.ics.uci.edu/ml/datasets/Auto+MPG>

URL11: WEKA yazılımı

<http://www.cs.waikato.ac.nz/ml/weka/>

URL12: Akküçük 2011b makalesi

<http://www.eurojournals.com/EJEFAS.htm>



URL13: Quinlan web sitesi See5/C5.0 programının Windows uyumlu yazılımı

<http://rulequest.com/see5-comparison.html>

URL14: Quinlan web sitesi C5.0 – C4.5 karşılaştırması

<http://rulequest.com/see5-comparison.html>

URL15: H1N1 testleri ile ilgili specificity ve sensitivity verileri

<http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5830a2.htm>

URL16: Kazanım ve Birikimli grafik örnekleri

[http://www2.cs.uregina.ca/~dbd/cs831/notes/lift\\_chart/lift\\_chart.html](http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html)

URL17: Larose (2005) Kitabı yardımcı web sitesinde bulunan kayıp müşteri verisi

<http://www.dataminingconsultant.com/data/churn.txt>

## EK: VAKA İLE İLGİLİ VERİ

Gender	Repeat	Department	Sem	Student Status	General Status	Grade
MALE	NR	SUMMER	0	SPECIAL	NORMAL	AA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	AA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	AA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	AA
FEMALE	NR	SUMMER	0	SPECIAL	NORMAL	AA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BA
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BB
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BB
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BB
MALE	NR	SUMMER	0	SPECIAL	NORMAL	BB
FEMALE	NR	SUMMER	0	SPECIAL	NORMAL	BB
FEMALE	NR	SUMMER	0	SPECIAL	NORMAL	BB
MALE	NR	SUMMER	0	SPECIAL	NORMAL	CB
MALE	NR	SUMMER	0	SPECIAL	NORMAL	CB

MALE	NR	SUMMER	0	SPECIAL	NORMAL	CB
MALE	NR	SUMMER	0	SPECIAL	NORMAL	CC
MALE	NR	SUMMER	0	SPECIAL	NORMAL	CC
FEMALE	NR	SUMMER	0	SPECIAL	NORMAL	CC
FEMALE	NR	SUMMER	0	SPECIAL	NORMAL	CC
MALE	NR	SUMMER	0	SPECIAL	NORMAL	CC
MALE	NR	A	10	UGRAD	SENIOR	AA
MALE	R	A	10	UGRAD	SENIOR	BA
MALE	NR	A	10	UGRAD	SENIOR	BA
MALE	R	A	12	UGRAD	SENIOR	AA
MALE	R	A	12	UGRAD	REPEATING	BA
MALE	R	A	14	UGRAD	SENIOR	AA
FEMALE	R	A	14	UGRAD	SENIOR	BA
MALE	R	A	14	UGRAD	SENIOR	CC
FEMALE	NR	A	3	UGRAD	NORMAL	AA
MALE	NR	A	3	UGRAD	NORMAL	BB
MALE	NR	A	4	UGRAD	NORMAL	AA
MALE	NR	A	4	UGRAD	NORMAL	AA
MALE	NR	A	4	UGRAD	NORMAL	AA
MALE	NR	A	4	UGRAD	NORMAL	AA
MALE	NR	A	4	UGRAD	REPEATING	AA
FEMALE	NR	A	4	UGRAD	NORMAL	BA
MALE	NR	A	4	UGRAD	NORMAL	BA

MALE	NR	A	4	UGRAD	NORMAL	BA
MALE	NR	A	4	UGRAD	NORMAL	CB
FEMALE	R	A	6	UGRAD	NORMAL	AA
MALE	NR	A	6	UGRAD	SENIOR	AA
MALE	NR	A	6	UGRAD	SENIOR	AA
MALE	NR	A	7	UGRAD	SENIOR	AA
MALE	NR	A	7	UGRAD	SENIOR	AA
MALE	NR	A	8	UGRAD	SENIOR	AA
FEMALE	NR	A	8	UGRAD	SENIOR	AA
FEMALE	NR	A	8	UGRAD	SENIOR	AA
FEMALE	R	A	8	UGRAD	SENIOR	AA
MALE	NR	A	8	UGRAD	SENIOR	BA
FEMALE	NR	A	8	UGRAD	SENIOR	BB
FEMALE	R	A	9	UGRAD	SENIOR	AA
FEMALE	NR	A	9	UGRAD	SENIOR	AA
MALE	R	A	9	UGRAD	ONPROBATION	BB
MALE	R	B	10	UGRAD	SENIOR	AA
FEMALE	R	B	10	UGRAD	SENIOR	AA
FEMALE	R	B	10	UGRAD	SENIOR	BA
MALE	R	B	10	UGRAD	SENIOR	BA
MALE	R	B	10	UGRAD	SENIOR	BA
FEMALE	R	B	10	UGRAD	SENIOR	BB
MALE	R	B	10	UGRAD	SENIOR	CB

MALE	R	B	12	UGRAD	SENIOR	BB
MALE	R	B	12	UGRAD	REPEATING	BB
MALE	R	B	12	UGRAD	SENIOR	CC
FEMALE	R	B	14	UGRAD	SENIOR	BB
FEMALE	R	B	14	UGRAD	SENIOR	CB
MALE	R	B	7	UGRAD	SENIOR	BB
MALE	NR	B	7	UGRAD	SENIOR	CC
MALE	R	B	8	UGRAD	NORMAL	AA
FEMALE	R	B	8	UGRAD	SENIOR	AA
FEMALE	R	B	8	UGRAD	SENIOR	BA
MALE	R	B	8	UGRAD	SENIOR	BB
FEMALE	R	B	8	UGRAD	SENIOR	BB
MALE	R	B	8	UGRAD	SENIOR	CC
MALE	R	B	8	UGRAD	SENIOR	CC
MALE	R	B	8	UGRAD	REPEATING	CC
MALE	R	B	9	UGRAD	SENIOR	BB
MALE	R	B	10	UGRAD	SENIOR	BB
FEMALE	R	B	13	UGRAD	SENIOR	CC
MALE	R	B	14	UGRAD	SENIOR	CC
FEMALE	R	B	9	UGRAD	SENIOR	AA
MALE	R	B	9	UGRAD	SENIOR	BB
MALE	R	C	10	UGRAD	SENIOR	BB
FEMALE	NR	C	10	UGRAD	SENIOR	AA

FEMALE	NR	C	10	UGRAD	SENIOR	AA
FEMALE	R	C	10	UGRAD	SENIOR	BA
MALE	NR	C	10	UGRAD	SENIOR	BA
FEMALE	R	C	10	UGRAD	SENIOR	BB
FEMALE	NR	C	11	UGRAD	SENIOR	BB
MALE	NR	C	12	UGRAD	SENIOR	AA
FEMALE	NR	C	12	UGRAD	SENIOR	CC
FEMALE	R	C	14	UGRAD	SENIOR	AA
MALE	R	C	16	UGRAD	SENIOR	CB
FEMALE	NR	C	16	UGRAD	SENIOR	CC
FEMALE	R	C	8	UGRAD	SENIOR	BA
FEMALE	NR	C	9	UGRAD	SENIOR	AA

