

# VERİ MADENCİLİĞİ

## Kavram ve Algoritmaları

Doç. Dr. Gökhan SİLAHTAROĞLU

PAPATYA YAYINCILIK EĞİTİM

İstanbul, Ankara, İzmir, Adana

© PAPATYA YAYINCILIK EĞİTİM  
BİLGİSAYAR SİS. SAN. VE TİC. A.Ş.  
İnönü Cad. Hacıhanım Sok. 10/6, 80090, Gümüşsuyu/İstanbul

Tel : (212) 245 37 40 (Merkez)  
Faks : (212) 245 37 41  
e-mail : bilgi@papatya.gen.tr  
Web : <http://www.papatya.gen.tr>  
<http://www.papatya.info.tr>  
Dağıtım : İstanbul : (212) 527 52 96  
Adana : (322) 432 00 73

### **Veri Madenciliği (Kavram ve Algoritmalar)** - Gökhan SILAHTAROĞLU

2. Basım Mart 2013

Yayın Danışmanı : Dr. Rifat ÇÖLKESEN  
Yayına Hazırlayan : Dr. Cengiz UĞURKAYA (Post-Edu Institute)  
Türk Dili : Necdet AVCI ve Batuhan AVCI  
Üretim : Olcay KAYA ve Ziya ÇÖLKESEN  
Sayfa Düzenleme : Papatya - Kelebek Tasarım  
Kapak Tasarım : Papatya - Kelebek Tasarım  
Basım ve Ciltleme : Altan Basım San. Ltd. Şti. / İstanbul

© Bu kitabın her türlü yayın hakkı **Papatya Yayıncılık Eğitim A.Ş.**'ye aittir.  
Yayinevinden yazılı izin alınmaksızın alıntı yapılamaz, kısmen veya tamamen  
hiçbir şekilde COĞALTILAMAZ, BASILAMAZ, YAYIMLANAMAZ. Kitabın, tamamı veya bir kısmının fotokopi makinesi, ofset gibi teknikle  
çoğaltılmaması, hem çoğaltan hem de bulunduranlar için yasadışı bir davranıştır.

Lütfen kitabımızın fotokopi yöntemiyle çoğaltımasına engel olunuz. Fotokopi  
hırsızlığıdır.

Silahtaroğlu, Gökhan.  
Veri Madenciliği (Kavram ve Algoritmaları) / Gökhan Silahtaroğlu. - İstanbul: Papatya Yayıncılık Eğitim,  
2013  
viii, 300 s.; 24 cm.  
Kaynakça ve dizin var.  
ISBN 978-975-6797-81-5  
1. Veritabanı. 2. Kümeleme. 3. Sınıflandırma. 4. Bilgisayar Algoritma. 5. Veri Ambarı.  
I. Title

*Aileme ve Sevdiklerime...*

# Teşekkür

Veri madenciliği konusuyla ilk tanışmam doktora eğitimim sırasında sayın Prof. Dr. Haldun AKPINAR'la başladım; ondan aldığım ilk veri madenciliği dersi, tez konumu da bu alanda seçmem ve çalışmalarımı bu yöne kaydirmama neden oldu diyebiliyorum. Daha sonra danışman hocam Prof. Dr. Öner ESEN ile yürüttüğüm çalışmalarımındaki literatür taramaları da yıllar sonra böylesi bir kitaba ışık tuttu. Ayrıca doktora çalışmalarımından sonra üniversitemde verdiği veri madenciliği dersleri de bu kitabı yazmama neden diğer bir unsurudur. Kitabın oluşmasında, aslında farkında olmasalar da dolaylı ve dolaysız katkısı olanlara ve kitap ekinde geçen kodların yazımı, derlenmesi ve testinde emeği geçen Sayın *Mehmet ÖZAKAN'a* teşekkürü bir borç bilirim.

Kiabın oldukça genişletilmiş ikinci baskısında okuyucuya daha fazla örneklerle veri madenciliği algoritmaları anlatılmaya çalışılmıştır. Ayrıca, literatürde öne çıkan bazıları çok yeni olan algoritmalar da ikinci baskıkda yerini almıştır.

Kitabı ders kitabı olarak kullanmak isteyen akademisyenler için ders sunumlarında kullanılmak üzere Power Point sunum dosyaları da hazırlanmıştır.

İlk baskıkta tanıtılan ticari ve açık kaynak kodlu veri madenciliği yazılımlarına ek olarak, bu baskıkta IBM Modeler (Clementine 12.0) programı ve KNIME açık kaynak kodlu veri madenciliği yazılımı tanıtılmış ve örnek uygulamalar adım adım anlatılmıştır. Bu uygulamaların daha fazlası PDF dosyası olarak da hazırlanarak, okuyucuya ayrıca bir kaynak olarak da sunulmaktadır.

Gerek PDF gerekse de Power Point sunum dosyalarını akademisyen okuyucularımız [www.tdk.com.tr](http://www.tdk.com.tr) adresinden alabilirler. Kitabın tüm veri madenciliği alnında çalışanlara yararlı olmasını diliyorum.

*Gökhan SİLAHTAROĞLU*

11 Şubat 2013, Sahrayıcedid.

# İçindekiler

<b>Önsöz</b>	<b>7</b>
<b>Bölüm 1. VERİ MADENCİLİĞİ</b>	<b>9</b>
1.1. Veri Madenciliğinin Uygulama Alanları	11
1.2. Veri Ambarları ve OLAP	15
1.3. Veri Madenciliği İçin Verilerin Hazırlanması	19
1.3.1. Verilerin Temizlenmesi	20
1.3.2. Verilerin Yeniden Yapılandırılması	24
1.4. Özet	26
1.5. Sorular	26
<b>Bölüm 2. VERİ MADENCİLİĞİ MODELLERİ</b>	<b>29</b>
2.1. Değer Tahmini Modeli	30
2.2. Bağlantı Analizi	32
2.3. Birlikteşlik Kuralları	33
2.4. Örüntü Tanıma	34
2.5. Ardışık Zaman Örüntüleri	35
2.6. Dolandırıcılık Tespiti	38
2.7. Kümeleme Analizi	39
2.7.1. Kümeleme Analizinin Kullanım Alanları	42
2.8. Özet	43
2.9. Sorular	43
<b>Bölüm 3. SINIFLANDIRMA</b>	<b>45</b>
3.1. Karar Ağaçları	46
3.1.1. ID3	51
3.1.2. C 4.5 ve C 5	56
3.1.3. CART	58
3.1.4. SLIQ	58
3.1.5. SPRINT	59
3.2. İstatistiğe Dayalı Algoritmalar	60
3.2.1. Bayesyen Sınıflandırma	60
3.2.2. Regresyon	63
3.3. Mesafeye Dayalı Sınıflandırma Algoritmaları	65
3.3.1. K-En Yakın Komşu	65
3.4. Yapay Sinir Ağları	66
3.4.1. İleri Sürümlü Yapay Sinir Ağları	71
3.4.2. Hata Geriye Yayma Yöntemi	72
3.5. SEE5 Yazılımı	75
3.6. Özet	80
3.7. Sorular	80

<b>Bölüm 4. BİRLİKTELİK KURALLARI VE İLİŞKİ ANALİZİ</b>	<b>83</b>
4.1. AIS Algoritması	85
4.2. SETM Algoritması	85
4.3. Apriori Algoritması	86
4.4. AprioriTid Algoritması	89
4.5. Diğer Algoritmalar	90
4.6. Weka Yazılımı	91
4.7. Özet	96
4.8. Sorular	97
<b>Bölüm 5. KÜMELEME</b>	<b>99</b>
5.1. Benzerlik ve Uzaklık	100
5.2. Kümeleme Analizinin Sınıflandırılması	106
5.3. Hiyerarşik Yöntemler	107
5.3.1. SLINK Algoritması ve Tek Bağlantı Tekniği	107
5.3.2. CURE Algoritması	109
5.3.3. CHAMELEON Algoritması	110
5.3.4. BIRCH	113
5.4. Bölümlemeli Yöntemler	114
5.4.1. K-Means (K-Ortalama) Algoritması	114
5.4.2. PAM Algoritması	117
5.4.3. CLARA Algoritması	119
5.4.4. CLARANS Algoritması	119
5.5. Yoğunluğa Dayalı Algoritmalar	121
5.5.1. DBSCAN Algoritması	122
5.5.2. OPTICS Algoritması	124
5.5.3. DENCLUE Algoritması	126
5.6. Grid Temelli Algoritmalar	129
5.6.1. STING Algoritması	129
5.6.2. Dalga Kümeleme	131
5.6.3. CLIQUE Algoritması	133
5.7. Genetik Algoritmalar	134
5.8. Özet	138
5.9. Sorular	138
<b>Ek-1. Veri Madenciliği Alanında Geliştirilmiş Programlar</b>	<b>141</b>
<b>Ek-2. Korelasyon Hesaplama</b>	<b>147</b>
<b>Ek-3. Bayes Teoremi</b>	<b>153</b>
<b>Ek-4. En Küçük Kareler ve Normalizasyon</b>	<b>157</b>
<b>Ek-5. Benzerlik Uzaklık</b>	<b>161</b>
<b>Kaynakça</b>	<b>165</b>
<b>Dizin</b>	<b>173</b>

## Önsöz

Veri madenciliği konusu güncel ve eksikliği hissedilen bir alandır. Çünkü veri toplama kaynakları ve bilişim teknolojisi oldukça gelişmiş ve yaygınlaşmıştır. Dolayısıyla kurumların elinde çok ciddi veritabanları oluşmaya başlamıştır. Hemen her disiplin tarafından çeşitli amaçlar için kullanılan veri madenciliği alanındaki gelişmeler, akademisyen ve araştırmacı bilim çevreleri tarafından, uluslararası bilimsel dergiler ve kongreler aracılığıyla takip edilmektedir. Bu eser de güzel Türkçemizde böylesi bir konuda kaynak oluşturmak için hazırlanmıştır.

Bu kitap ile hem Türk okuyucusuna veri madenciliği konusunda yardımcı olmak, hem de yeni başlayanlar ve deneyimli kişilere yol göstermek hedeflenmiştir. Ele alınan tüm teknik ve algoritmalar uluslararası hakemli kongrelerde sunulmuş, basılmış veya yine uluslararası bilimsel dergilerde yayınlanmış eserlerden derlenmiştir. Bu algoritmalar, okuyucunun kolay takip edeceği bir şekilde, bilinen veri madenciliği modelleri arasına konularak sunulmuştur. Kitap içerisinde algoritmaların kaba kodları ve çalışma ilkeleri de anlatılmış olup ayrıca en çok kullanılan ve literatürde adı öne çıkmış algoritmalarla ilişkin çözümü örnekler de sunulmuştur. Bu anlamda, bu alanda yazılım geliştirmek isteyenlere ışık tutulmaya çalışılmıştır. Kitabın bu ikinci baskısında, bu örnekler ve algoritmaların sayısı artırılmış ve kitaba zenginlik kazandırılmıştır.

Kitap içerisinde ikisi ticari, diğeri açık kaynak kodlu olmak üzere üç ayrı yazılım tanıtılmıştır. Bunun yanısıra kitabın ekinde, günümüzde kullanılan veri madenciliği yazılımlarının kısa kopyeleri ve bulunabilecekleri Internet adresleri verilmiştir. Bu yönyle de kendi yazılımını geliştirmek yerine bilinen kavram ve teknikleri başka yazılımlar aracılığıyla uygulamak isteyenlere yol gösterilmeye çalışılmıştır. Kitabın genişletilmiş ikinci baskısında okuyucuya teorik konuların dışında uygulama deneyimi de kazandırmak için çeşitli saha uygulama örnekleri verilmiş ve Clementine programıyla çözüm ve raporları sunulmuştur.

Kitabın ülkemizdeki tüm öğrencilerimize, kendi öğrencilerime, mühendislik, işletme ve diğer bilim dallarında araştırmacılara yararlı olmasını dilerim.

**Doç. Dr. Gökhan SİLAHTAROĞLU**

