

Eđitimde Veri Madenciliđi ve Bilgisayar Uygulamaları

Editörler: Murat KAYRI · Hikmet ŐEVGİN



Editörler: Prof. Dr. Murat KAYRI - Dr. Öğr. Üyesi Hikmet ŞEVGİN

EĞİTİMDE VERİ MADENCİLİĞİ VE BİLGİSAYAR UYGULAMALARI

ISBN 978-625-6810-73-0

Kitap içeriğinin tüm sorumluluğu yazarlarına aittir.

© 2023, PEGEM AKADEMİ

Bu kitabın basım, yayım ve satış hakları Pegem Akademi Yay. Eğt. Dan. Hizm. Tic. AŞ'ye aittir. Anılan kuruluşun izni alınmadan kitabın tümü ya da bölümleri, kapak tasarımı; mekanik, elektronik, fotokopi, manyetik kayıt ya da başka yöntemlerle çoğaltılamaz, basılamaz ve dağıtılamaz. Bu kitap, T.C. Kültür ve Turizm Bakanlığı bandrolü ile satılmaktadır. Okuyucularımızın bandrolü olmayan kitaplar hakkında yayınevimize bilgi vermesini ve bandrolsüz yayınları satın almamasını diliyoruz.

Pegem Akademi Yayıncılık, 1998 yılından bugüne uluslararası düzeyde düzenli faaliyet yürüten **uluslararası akademik bir yayinevidir**. Yayımladığı kitaplar; Yükseköğretim Kurulunca tanınan yükseköğretim kurumlarının kataloglarında yer almaktadır. Dünyadaki en büyük çevrimiçi kamu erişim kataloğu olan **WorldCat** ve ayrıca Türkiye'de kurulan **Turcademy.com** tarafından yayınları taranmaktadır, indekslenmektedir. Aynı alanda farklı yazarlara ait 1000'in üzerinde yayını bulunmaktadır. Pegem Akademi Yayınları ile ilgili detaylı bilgilere <http://pegem.net> adresinden ulaşılabilmektedir.

I. Baskı: Eylül 2023, Ankara

Yayın-Proje: Şehriban Türüldür
Dizgi-Grafik Tasarım: Tuğba Kaplan
Kapak Tasarımı: Pegem Akademi

Baskı: Sonçağ Yayıncılık Matbaacılık Reklam San Tic. Ltd. Şti.
İstanbul Cad. İstanbul Çarşısı 48/48 İskitler/Ankara
Tel: (0312) 341 36 67

Yayıncı Sertifika No: 51818
Matbaa Sertifika No: 47865

İletişim

Macun Mah. 204. Cad. No: 141/A-33 Yenimahalle/ANKARA
Yayınevi: 0312 430 67 50
Dağıtım: 0312 434 54 24
Hazırlık Kursları: 0312 419 05 60
İnternet: www.pegem.net
E-ileti: pegem@pegem.net
WhatsApp Hattı: 0538 594 92 40

ÖN SÖZ

İşlenmemiş bilgi olarak tanımlanan veri, araştırma süreçlerinde keşfedilen bilginin hammaddesi olarak değerlendirilmektedir. Verinin büyüklüğü arttıkça bilgi keşfi daha karmaşık hale gelebilmektedir. Bu durumda, hacim ve değişken çeşitliliği artan veri kümeleri arasında anlamlı ilişki kurmaya yönelik çözümler zorlaşmaktadır. Bilindiği üzere, veri kaynaklarından elde edilen ve hacimce büyük olan veri kümeleri “büyük veri (*big data*)” olarak isimlendirilmektedir. Büyük veri, çoğunlukla veri tipleri açısından çeşitlilik ve karmaşık yapı sergilemektedir. Yaşadığımız bu yüzyılda, neredeyse her sektörde büyük verilerin analizi üzerinden karar süreçlerinin şekillendiğini görmekteyiz. Bununla birlikte, toplum bilimi araştırmalarında, sıklıkla toplumsal eğilimler (davranış kalıpları, harcama biçimleri, siyasi/sosyal eğilimler gibi) büyük verilerin analizi ile ortaya konmaktadır. Tüm bu gerekçelerden kaynaklı olarak, büyük verilere uygulanacak olan istatistiksel yöntemlerin araştırmacılar tarafından doğru şekliyle biliniyor olması gerekmektedir.

Veri madenciliği, hacimce büyük olan veri kümeleri arasında neden-sonuç ilişkisini keşfetme ve eş deyişle bilgiyi madenleme olarak ifade edilebilir. Aynı zamanda, veri madenciliği mevcut verilerden hareketle geleceği tahmin (prediction) etme sanatı olarak düşünülebilir. Literatürde, büyük verileri çözümlemeye yönelik birçok veri madenciliği teknikleri geliştirilmiştir. Bu yönüyle, araştırmalarda; Regresyon, Sınıflandırma (Classification), Kümeleme (Clustering), Karar Ağaçları (Decision Trees), Örüntüyü İzleme (Tracking Patterns) gibi veri madenciliği teknikleri kullanılmaktadır. Veri madenciliğinde bahsi edilen bu tekniklerin (algoritmalar) bünyesinde makine öğrenmesi yöntemleri etkin bir şekilde kullanılmaktadır. Makine öğrenmesi, büyük verinin bir kısmı üzerinde kendini eğiterek veri setinin davranış kalıplarını ve veriler arasındaki ilişkiyi çözümlmeyi (öğrenmeyi) hedefleyen bir kavram olarak karşımıza çıkmaktadır. Bu öğrenme sürecinden sonra büyük verilere ait tahminleme (prediction) ya da tanımlama (description) modelleri kurulmaktadır.

Elinizdeki bu kitap yükseköğretimin birçok alanında kullanılmaya uygun olarak hazırlanmıştır. Veri madenciliğinin istifade edileceği eğitim alanlarının birinde kullanılabileceği gibi; mühendislik, sağlık, sosyal gibi alanlarda da bu kitaptan konforlu bir şekilde istifade edilebilecektir. Araştırmalarda en çok ihtiyaç duyulan veri madenciliği yöntemleri dikkate alınarak bu kitap yapılandırıldı. Bununla birlikte, kitapta ele alınan veri madenciliği yöntemleri sadece teorik boyutu ile ele alınmamış, araştırmacılara örneklik teşkil etmesi açısından gerçekçi veya hi-

potetik veri kmeleri zerinde uygulaması da yapılmıř olup analiz ıktıları zenle yorumlanmıřtır.

Eđitimde Veri Madenciliđi Yntemleri ve Bilgisayar Uygulamaları bařlıklı kitabımız 10 blmden oluřmaktadır. Her blm yazarımız, okuyucuya/arařtırmacıya sunduđu blmde derin bir uzmanlıđa sahiptir; bu ynyle kitabımızın katma deđerinin yksek olduđuna ve sahaya yarar sađlayacađına inanmaktayız. Gerek deđerlendirmenin sahibi ise elbette okuyucularımız olacaktır. Kitabın đrencilerimize, arařtırmacılarımıza ve eđitmenlerimize yarar sađlaması dileđiyle.

Tm Yazarlar Adına

Prof. Dr. Murat KAYRI

Editr

BÖLÜMLER VE YAZARLARI

Editörler: Prof. Dr. Murat KAYRI - Dr. Öğr. Üyesi Hikmet ŞEVGİN

1. Bölüm: Eğitimde Veri Madenciliği

Prof. Dr. Murat KAYRI, Van Yüzüncü Yıl Üniversitesi

ORCID No: 0000-0002-5933-6444

2. Bölüm: Veri Hazırlama ve Ön İşleme Süreci

Dr. Öğr. Üyesi Hikmet ŞEVGİN, Van Yüzüncü Yıl Üniversitesi

ORCID No: 0000-0002-9727-5865

3. Bölüm: Kümeleme Analizi

Dr. Öğr. Üyesi Fuat ELKONCA, Muş Alparslan Üniversitesi

ORCID No: 0000-0002-2733-8891

4. Bölüm: Lojistik Regresyon Analizi

Doç. Dr. Mehmet ŞATA, Van Yüzüncü Yıl Üniversitesi

ORCID No: 0000-0003-2683-4997

Doç. Dr. Ayfer SAYIN, Gazi Üniversitesi

ORCID No: 0000-0003-1357-5674

5. Bölüm: Karar Ağaçları

Dr. Öğr. Üyesi Görkem CEYHAN, Muş Alparslan Üniversitesi

ORCID No: 0000-0001-9342-6876

6. Bölüm: Çok Değişkenli Uyarlanabilir Regresyon Uzanımları

(Multivariate Adaptive Regression Splines-Mars)

Dr. Öğr. Üyesi Hikmet ŞEVGİN, Van Yüzüncü Yıl Üniversitesi

ORCID No: 0000-0002-9727-5865

7. Bölüm: Ensemble Yöntemler

Dr. Öğr. Üyesi Özlem BEZEK GÜRE, Batman Üniversitesi

ORCID No: 0000-0002-5272-4639

8. Bölüm: Yapay Sinir Ağları

Dr. Öğr. Üyesi Özlem BEZEK GÜRE, Batman Üniversitesi

ORCID No: 0000-0002-5272-4639

9. Bölüm: Roc Analizi (Receiver Operating Characteristic Analysis)

Dr. Öğr. Üyesi Mahmut Sami KOYUNCU, Afyon Kocatepe Üniversitesi

ORCID No: 0000-0002-6651-4851

10. Bölüm: Destek Vektör Makineleri

Dr. Öğr. Üyesi Erdal EKER, Muş Alparslan Üniversitesi

ORCID No: 0000-0002-5470-8384

İÇİNDEKİLER

Ön Söz.....	iii
Bölümler ve Yazarları.....	v

1. BÖLÜM EĞİTİMDE VERİ MADENCİLİĞİ

Veri Madenciliğine Giriş	3
Veri Madenciliğinin Tarihçesi	6
Veri Madenciliği Uygulama Alanları.....	7
Eğitsel Veri Madenciliği.....	8
Kaynakça.....	10

2. BÖLÜM VERİ HAZIRLAMA VE ÖN İŞLEME SÜRECİ

Giriş.....	15
Veri Temizleme	16
Kayıp Veriler	16
Uç Değerler	20
Tutarsız Verilerin Düzeltilmesi.....	20
Tekrarlı Verilerin İşlenmesi	20
Veri Bütünleştirme	20
Veri Seçme	21
Veri Dönüştürme	22
Kaynakça.....	23

3. BÖLÜM KÜMELEME ANALİZİ

Giriş.....	27
Kavramsal Çerçeve.....	28
Faktör Analizi ve Diskriminant Analizi ile Karşılaştırılması	30
Yakınlık (Proximity) Ölçüleri.....	31
Kümeleme Analizi Yöntemleri	34
Varsayımlar, Örneklem Büyüklüğü, Veri Temizleme ve Geçerlik	39
SPSS'te Uygulama ve Raporlama.....	41
Veri Seti	41

Aşamalı Kümeleme Analizi SPSS Uygulaması	42
Aşamalı Olmayan Kümeleme Analizi (K-Ortalamalar) SPSS Uygulaması	50
İki Aşamalı Kümeleme Analizi SPSS Uygulaması	56
Sonuç	60
Kaynakça	61

4. BÖLÜM

LOJİSTİK REGRESYON ANALİZİ

Giriş	68
Lojistik Regresyon Analizi Nedir?	68
Lojistik Regresyon Analizinin Amacı ve Kullanım Alanları	70
Açıklayıcı Bir Örnek	72
Lojistik Regresyon Analizinin Diğer Yöntemlerin Karşılaştırılması	73
Lojistik Regresyon Analizinin Aşamaları	75
Lojistik Regresyon Analizinin Cevap Aradığı Sorular	78
Lojistik Regresyon Analizinin Varsayımları	79
Lojistik Regresyon Analizinde Model Katsayılarının Hesaplanması	81
Odds oranı ve Logit değeri	81
Logistik regresyon katsayısı	83
Wald İstatistiği	83
Exp (B) katsayısı	84
Lojistik Regresyon Analizinde Modelin Değerlendirilmesi	84
-2 Log Likelihood Katsayısı	85
Hosmer ve Lemeshow testi	85
Sınıflandırma Tablosu	86
R ve R ² Katsayıları	86
AIC (Akaike Bilgi Kriteri) ve BIC (Bayesian Bilgi Kriteri) Değerleri	87
Lojistik Regresyon Analizinde Sonuçların Raporlanması	87
Lojistik Regresyon Analizinde Kestirim Yöntemleri	88
En Çok Olabilirlik Yöntemi (Maximum Likelihood Estimation (ML))	88
Sıralı En Küçük Kareler Kestirim Yöntemi (Ordinary least squares estimation (OLS))	88
Newton-Raphson Yöntemi	89
Fisher Scoring Yöntemi	89
Stochastic Gradient Descent Yöntemi	90
Bayesian Inference Yöntemi	90
Lojistik Regresyon Analizinin Türleri	90
İkili (Binary) Lojistik Regresyon	90

Çok kategorili (Multinomial) Lojistik Regresyon.....	91
Sıralı (Ordinal) Lojistik Regresyon.....	91
İkili Lojistik Regresyon Analizi ve SPSS Uygulaması.....	91
İkili LR Analiz Çıktılarının APA Stiline Göre Raporlanması.....	94
Çok Kategorili Lojistik Regresyon Analiz Çıktılarının APA Stiline Göre Raporlanması	95
Sıralı Lojistik Regresyon Analizi ve SPSS Uygulaması.....	97
Sıralı LR Analiz Çıktılarının APA Stiline Göre Raporlanması.....	98
Sonuç.....	99
Kaynakça	100

5. BÖLÜM KARAR AĞAÇLARI

Karar Ağaçları.....	105
Karar Ağaçlarının İşleyiş Süreci	110
Karar Ağacı Algoritmaları.....	113
ID3	113
C4.5 (J48)	115
QUEST (Quick, Unbiased, Efficient Statistical Tree)	117
CHAID (Chi-square Automatic Interaction Detection).....	118
CART/C&RT (Classification and Regression Trees).....	121
Diğer Karar Ağacı algoritmaları.....	123
Karar Ağaçlarına Yönelik Örnek Uygulamalar.....	124
QUEST Algoritması Örnek Uygulama.....	125
CHAID Algoritması Örnek Uygulama	134
CART Algoritması Örnek Uygulama	138
Sonuç.....	141
Kaynakça.....	142

6. BÖLÜM ÇOK DEĞİŞKENLİ UYARLANABİLİR REGRESYON UZANIMLARI (MULTIVARIATE ADAPTIVE REGRESSION SPLINES-MARS)

Giriş.....	147
MARS-Çok Değişkenli Uyarlamalı Regresyon Uzanımları.....	147
Temel Fonksiyonların Oluşum Süreci	149
Uygulama ve Raporlama	155

Etkileşim Sayısı Bir Olanlar İçin (One Interaction) Kurulan MARS Modeli.....	159
Etkileşim Sayısı İki Olanlar İçin (Two İnteraction) Kurulan MARS Modeli	162
Etkileşim Sayısı Üç Olanlar İçin (Three İnteraction) Kurulan MARS Modeli....	165
Sonuç.....	168
Kaynakça.....	170

7. BÖLÜM ENSEMBLE YÖNTEMLER

Giriş	173
Boosting	173
Adaboost	174
Bagging	181
Bagging Algoritması	182
Uygulama ve Raporlama	183
Random Forest.....	185
Random Forest Algoritması.....	186
Uygulama ve Raporlama	186
Sonuç	189
Kaynakça.....	190

8. BÖLÜM YAPAY SİNİR AĞLARI

Giriş	195
Yapay Sinir Ağları Kullanım Alanları	195
Biyolojik Sinir Hücresi.....	196
Yapay Sinir Hücresi	196
Yapay Sinir Ağları Yapısı	197
Yapay Sinir Ağlarının Sınıflandırılması.....	197
Geri Beslemeli Ağlar	197
İleri Beslemeli Ağlar	199
Sonuç	219
Kaynakça.....	220

9. BÖLÜM

ROC ANALİZİ

(RECEIVER OPERATING CHARACTERISTIC ANALYSIS)

ROC Analizi.....	225
Atama Tablosu/ Kontenjans Tablosu (Contingency Table)/Hata Matrisi (Confusion).....	226
ROC Analizi Örnek Araştırma Durumu	227
ROC Eğrisi	229
ROC Analizinin Değerlendirilmesi	230
ROC Eğrisi Altında Kalan Alan (Area Under the ROC Curve -AUC))	231
Gini İndeksi (Katsayısı)	231
Youden İndeks (YI).....	233
ROC Analizi SPSS Uygulaması	233
Test Performansını Değerlendirmek ve Kesme Puanı Belirlemek İçin ROC Analizini Kullanma	233
İki Grup Arasındaki Farkın Anlamlılığını Test Etme.....	241
Birden Fazla Testin/Yöntemin/Modelin Sınıflama Performansını Karşılaştırma İçin ROC Analizini Kullanma.....	243
Kaynakça.....	250

10. BÖLÜM

DESTEK VEKTÖR MAKİNELERİ

Giriş.....	253
Destek Vektör Makineleri.....	255
Destek Vektör Makinalarının Optimizasyonu	258
Destek Vektör Makinelerinin Matematiksel Modellemesi	258
Materyal ve Yöntem	262
MATLAB Program Paketinin İçerdiği DVM Modeli.....	262
DeneySEL Sonuçlar	265
Sonuç.....	273
Kaynakça.....	274

Editörler ve Yazarlar Hakkında.....	277
--	------------

