

МЕЃУНАРОДЕН УНИВЕРЗИТЕТ ВИЗИОН
UNIVERSITETI NDËRKOMBËTAR VIZION



INTERNATIONAL VISION UNIVERSITY
ULUSLARARASI VİZYON ÜNİVERSİTESİ

Teoriden Pratiğe Veri Madenciliği

Aybeyan SELİM

Kuzey Makedonya – Gostivar, 2024

Teoriden Pratięe Veri Madencilięi

Yazar

Prof. Dr. Aybeyan SELİM

Yayımcı

Uluslararası VİZYON Üniversitesi Gostivar, Kuzey Makedonya
International VISION University – Gostivar, North Macedonia

Editör

Prof. Dr. Fadıl HOCA – Uluslararası Vizyon Üniversitesi Kurucu Rektörü

Denetleyenler

Prof. Dr. Muzafer SARAÇEVİÇ – Uluslararası Novi Pazar Üniversitesi,
Bilgisayar Bilimleri Bölümü Öğretim Üyesi

Doç. Dr. Fehmi SKENDER – Uluslararası Vizyon Üniversitesi, Mühendislik
ve Mimarlık Fakültesi, Bilgisayar Bilimleri Bölümü Öğretim Üyesi

Bilgisayar Tasarımı

Doç. Dr. İlker ALİ

Lektör

Dr. Mesut UĞURLU

Basımevi

Print Factory DOOEL Skopje

Yayın yeri ve yılı: Gostivar, 2024

Tiraj : 500

Önsöz

Bu kitap, Uluslararası Vizyon Üniversitesi Mühendislik ve Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü'nde verilen Veri Madenciliği dersi için bir ders kitabı olarak hazırlanmıştır. Günümüz dünyasında veri, en değerli kaynaklardan biri haline gelmiş ve doğru analiz edildiğinde büyük fırsatlar sunmaktadır. Bu kitabın amacı, öğrencilere veri madenciliğinin temel kavramlarını, yöntemlerini ve uygulamalarını öğretmek ve bu alanda yetkinlik kazanmalarına katkı sağlamaktır.

Kitapta, veri madenciliği süreci, veri ön işleme teknikleri, farklı modelleme yöntemleri ve değerlendirme metrikleri detaylı bir şekilde ele alınmıştır. Teorik bilgilerin yanı sıra, pratik örneklerle desteklenen uygulamalı bölümler sayesinde öğrenciler, öğrendiklerini pekiştirme ve gerçek dünyadaki sorunlara çözüm bulma imkanı bulacaklardır.

Bu kitabın hazırlanmasında desteklerini esirgemeyen Uluslararası Vizyon Üniversitesiyönetimine içtenlikle teşekkür ederim. Ayrıca, kitabın metninin iyileştirilmesi konusunda değerli önerileri ile katkı sağlayan Prof. Dr. Muzafer Saraçevîç ve Doç. Dr. Fehmi Skender'e özel teşekkürlerimi sunarım. Bu süreçte bana her daim destek olan sevgili eşim Semra Hanım'a ve oğullarım Erdem ile Eren'e en derin teşekkürlerimi sunarım. Onların sabrı ve sevgisi, bu kitabın hazırlanmasında en büyük motivasyon kaynağım olmuştur.

Prof. Dr. Aybeyan SELİM
Uluslararası Vizyon Üniversitesi
Mühendislik ve Mimarlık Fakültesi Dekanı

12. 09. 2024
Gostivar

Bilgi ve Verinin Gücü Üzerine Sözler

"Eğer bilgi edinmek istersen, insanlarla tartış ve fikirlerini paylaş; bilgi paylaşımıyla artar."

– El- Harezmi

"Bilgi, sadece öğrenmekle değil, paylaşmakla değer kazanır."

– Fârâbî

"Veri, yeni petroldür."

– Clive Humby

"Büyük veri olmadan, kör ve sağır bir şekilde otoyolun ortasındasınız."

– Geoffrey Moore

"Tanrı'ya güveniriz, diğer herkes veri getirmelidir."

– W. Edwards Deming



İçindekiler

Önsöz.....	III
1. Veri Madenciliğine Giriş.....	1
1.1 Veri Madenciliği Nedir?	1
1.2 Veri Madenciliği Görevleri.....	3
1.3 Veri Madenciliği Uygulamaları.....	5
1.4 Veri Madenciliği Süreci.....	6
1.5 Veri Madenciliği Araçları ve Teknolojileri.....	7
1.6 Veri Madenciliği Sürecinde Dikkat Edilmesi Gereken Hususlar	9
1.7 Veri Madenciliği ve İş Zekası	10
1.8 Veri Madenciliği ve Büyük Veri	10
1.9 Veri Madenciliği Projelerinin Yönetimi.....	11
1.10 Veri Madenciliği Teknikleri ve Algoritmaları	12
1.11 Veri Madenciliği ve Makine Öğrenmesi	13
1.12 Veri Madenciliği ve İş Analitiği	14
1.13 Veri Madenciliği ve İş Zekası Arasındaki İlişki	15
1.14 Veri Madenciliği ve Büyük Veri Analitiği.....	15
1.15 Veri Madenciliği ve Yapay Zeka	16
1.16 Veri Madenciliği ve İş Zekası Araçları.....	17
1.17 Veri Madenciliği ve İş Zekası Stratejileri	17
1.18 Veri Madenciliği ve Gelecek Yönelimler	18
1.19 Veri Madenciliği ve Etik	19
1.20 Bilgi ve Verinin Gücü Üzerine Sözler	20
2. Neden Python Programlama Dili?	21
2.1 Python'da Temel Veri Türleri	23
2.2 Bileşik Veri Türleri	30
2.3 Kontrol Akışı Komutları	33

2.4 Kullanıcı Tanımlı Fonksiyonlar.....	36
2.5 Dosyadan Veri Okuma/Yazdırma	37
2.6 Veri Madenciliğinin Etkileyici Örnekleri ve Olayları	38
3. Veri.....	39
3.1. Veri Türleri.....	40
3.2 Nitelikler ve Ölçekler	40
3.3 İstatistiksel Analiz	44
3.4 Veri Seti Türleri	45
3. 5 Veri Kümesinin Genel Özellikleri	46
3. 6 Benzerlik Ölçüleri	50
3.7 Numerik Verilerde Uzaklık.....	52
3.8 Benzerlik ölçümleri	54
3.9 Nominal Veriler İçin Mesafe Ölçümleri	60
3.10 Ordinal Veriler İçin Uzaklık Hesaplaması.....	62
4. Veri Madenciliği için Python - Numpy Kütüphanesi.....	65
4.1 ndarray oluşturma (NumPy dizisi oluşturma).....	65
4.2 Eleman Bazında İşlemler	69
4.3 İndeksleme ve Dilimleme	70
4.4 Numpy Kütüphanesinde Aritmetik ve İstatistiksel Fonksiyonlar	74
4.5 Numpy kütüphanesinde lineer cebir işlemleri	78
5. Veri Madenciliği için Python - Pandas Kütüphanesi	81
5.1 Series Nesnesi	82
5.2 Veri Çerçevesi (Data Frame)	90
5.3 Pandas Kütüphanesinde Bazı Önemli Fonksiyonlar	103
6. Veri Madenciliği için Python - Matplotlib Kütüphanesi.....	109
6.1 Matplotlib Çoklu Grafikleri	113
6.2 Grafiğe Farklı Elemanlar Ekleme.....	117

6.3 Grafiklerin Çizilmesi	122
7. Makine Öğrenmesine Giriş	131
7.1 Makine Öğrenmesinde Temel Kavramlar	131
7.2 Regresyon Analizinde Performans Ölçütleri	134
7.3 Makine Öğrenmesi Projeleri İçin Uygulanabilecek Adımlar	136
7.4 Makine Öğrenmesi Projeleri İçin Başarısızlık Nedenleri	136
7.4 Scikit-Learn Kütüphanesi Hakkında	138
7.5 Veri Madenciliğinin Etkileyici Örnekleri ve Olayları	142
8. Veri Temizleme	143
8.1 Aykırı Değerler	144
8.2 Eksik Değerler	148
8.3 Veri Manipülasyonu Fonksiyonları	161
9. Veri Ön İşleme ve Kodlama Teknikleri	173
9.1 Pratik Uygulama ve Kodlama Teknikleri	177
9.2 Etiket Kodlama (Label Encoding)	177
9.3 Özellik Eşleme (Feature Mapping)	180
9.4 Tek Sıcak Kodlama (One Hot Encoding)	182
9.5 İkili Kodlama (Binary Encoding)	185
9.6 Hedef Kodlama (Target Encoding)	190
9.7 Çarpıklık (Skewness) ve Normal Dağılım	192
9.8 Karekök ve Küpkök Dönüşümleri	194
9.9 Logaritmik Dönüşüm	201
9.10 Box-Cox Dönüşümü	205
9.11 Veri Bölme ve Özellik Ölçeklendirme	210
10. Lineer Regresyon Modeli	221
10.1 Lineer Regresyonun Türleri	222
10.2 Lineer Regresyonun Sınırlamaları	222

10.3 Katsayıların Tahmini	223
10. 4 En Küçük Kareler Metodu (Least Squares Method)	224
10. 5 Sapmalar ile Tahmin.....	224
10.6 Determinasyon Katsayısı (R^2).....	225
10.7 Basit Lineer Regresyonun Python Programlama Dilinde Uygulanması	226
10.8 Çoklu Regresyon Modeli	231
10.9 Veri Madenciliğinin Etkileyici Örnekleri ve Olayları	234
11. Lojistik Regresyon Modeli	235
11.1 Logit Fonksiyonu	239
11.2 Sigmoid Fonksiyonu.....	241
11. 3 Basit Lojistik Regresyonda Bir Bağımsız Değişkenin Olasılığı	243
11.4 Python Programlama Dilinde Örnek Lojistik Regresyon Çalışması	245
12. Sınıflandırma Algoritmaları	251
12.1 Karar Ağaçları	251
12.2 Karar Ağacı Oluşturma İçin Temel Algoritma.....	254
12.3 Öznitelik Testi Koşullarını İfade Etme Yöntemleri	257
12.4 Öznitelik Testi Koşulu Seçme Ölçümleri	258
12.5 Veri Setine Karar Ağacı Algoritmasını Python'da Uygulama	261
13. Kümeleme Algoritmaları	267
13.1 Kümeleme Analizi	267
13.2 Farklı Kümeleme Türleri	268
13.3 K-means (K-ortalamlar) Kümeleme Algoritması	270
13.4 K-means Algoritmasının Python'da Uygulanması	272
13.5 Bilgi ve Verinin Gücü Üzerine Sözler.....	276
14. Kendimizi Sınayalım	277
15. Kaynakça.....	279

Bilgi ve Verinin Gücü Üzerine Sözler

"Veriyi anlamadan harekete geçmek, karanlık bir odada ışığı açmadan yol bulmaya çalışmaktır."

– Cahit Arf

"Geleceği şekillendirmek isteyen, veriyi anlamalı ve doğru soruları sormalıdır."

– Aziz Sancar

"Bilim, aklımızı aydınlatırken veri, yolumuzu belirler."

– Ali Kuşçu

"Bilgiye ulaşmak kadar, onu doğru kullanmak da bir erdemdir."

– İbn-i Sina





15. Kaynakça

- [1] C. Aggarwal, *Data Mining: The Textbook*. Springer, 2009.
- [2] Desrosiers and G. Karypis, "A comprehensive survey of neighborhood-based recommendation methods," in *Recommender Systems Handbook*, pp. 107–144, 2011.
- [3] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.
- [4] J. H. Faghmous, A. Banerjee, S. Shekhar, M. Steinbach, V. Kumar, A. R. Ganguly, and N. Samatova, "Theory-guided data science for climate change," *Computer*, vol. 47, no. 11, pp. 74–78, 2014.
- [5] J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. MIT Press, 2001.
- [6] J. Kleinberg, J. Ludwig, and S. Mullainathan, "A Guide to Solving Social Problems with Machine Learning," *Harvard Business Review*, Dec. 2016.
- [7] Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data," *IEEE Transactions on Knowledge and Data Engineering*, 2017.
- [8] M. H. Dunham, *Data Mining: Introductory and Advanced Topics*. Prentice Hall, 2006.
- [9] M. Tomov and C. Velkoska, "Contribution of the quality costs to sustainable development," *Production Engineering Archives*, vol. 28, no. 2, pp. 164–171, 2022, doi: 10.30657/pea.2022.28.19.
- [10] Nuredin and M. Nuredin, *Farklı Alanlarda Etik*, 2023.
- [11] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proc. of the 6th Intl. Conf. on Knowledge Discovery and Data Mining*, Boston, MA, 2000, pp. 71–80, ACM Press.
- [12] P. S. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. Pearson Education, 2019.
- [13] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

-
- [14] R. L. Grossman, M. F. Hornick, and G. Meyer, "Data mining standards initiatives," *Communications of the ACM*, vol. 45, no. 8, pp. 59–61, 2002.
- [15] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, Inc., New York, 2nd ed., 2001.
- [16] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, San Francisco, CA, 2003.
- [17] S. Papadimitriou, A. Brockwell, and C. Faloutsos, "Adaptive, Unsupervised Stream Mining," *VLDB Journal*, vol. 13, no. 3, pp. 222–239, 2004.
- [18] Selim and I. Ali, "Dijital Okuryazarlık Becerilerinin Kuzey Makedonya'da Yükseköğretim Üzerindeki Etkisi," in *International Congress of Economics & Business*, pp. 81–91, 2022.
- [19] Selim, F. Skender, and I. Ali, "Systematic Review Of Big Data, Digital Transformation Areas And Industry 4.0 Trends In 2021," *International Scientific Journal Vision*, vol. 6, no. 2, pp. 27–41, 2021.
- [20] Selim, I. Ali, and B. Ristevski, "University Information System's Impact on Academic Performance: A Comprehensive Logistic Regression Analysis with Principal Component Analysis and Performance Metrics," *TEM Journal*, vol. 13, no. 2, 2024.
- [21] Selim, I. Ali, M. Saracevic, and B. Ristevski, "Application of the digital twin model in higher education," *Multimedia Tools and Applications*, pp. 1–18, 2024.
- [22] Selimi and A. Üseini, "Yenilikçi eğitim ile dijital yetkinlik ve girişimcilik becerilerinin geliştirilmesi–Kuzey Makedonya örneği," in *ICEB'19-International Congress of Economics and Business*, 2019.
- [23] Skender and V. Manevska, "Data Visualization Tools-Preview and Comparison," *Journal of Emerging Computer Technologies*, vol. 2, no. 1, pp. 30–35, 2022.
- [24] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, Prediction*. Springer, 2nd ed., 2009.
- [25] T. Hey, S. Tansley, and K. M. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery, Volume 1*. Microsoft Research, Redmond, WA, 2009.
- [26] Velkoska and M. Tomov, "Understanding and application of quality costs in automotive manufacturing companies in North

- Macedonia: empirical study," *International Journal of Quality & Reliability Management*, vol. 40, no. 6, pp. 1463–1484, 2023, doi: 10.1108/IJQRM-01-2022-0032.
- [27] Velkoska and M. Tomov, "Visualization of the Process of Tracking Quality Using Quality Costs: An Empirical Study," *TEM Journal*, vol. 11, no. 2, pp. 971–980, 2022, doi: 10.18421/TEM112-61.
- [28] Velkoska, "Integration of Juran's trilogy, Deming's quality cycle and DMAIC methodology in the development of management with quality cost methodology," *Vision International Refereed Scientific Journal*, vol. 7, no. 2, pp. 109–123, 2022, doi: 10.55843/ivisum227
- [29] X. Amatriain and J. M. Pujol, "Data mining methods for recommender systems," in *Recommender Systems Handbook*, pp. 227–262. Springer, 2015.