

Dr. Haydar EKELİK

**KARAR AĞAÇLARI İLE
TOPLULUK ÖĞRENME ALGORİTMALARI
(RASTGELE ORMAN)
TEORİ VE UYGULAMASI**



FİLİZ KİTABEVİ

İSTANBUL

2025

Nisan 2025

ISBN: 978-625-5957-50-4

BASKI - CİLT

Filiz Kitabevi Basım Yayın Dağıtım
Petrol Ürünleri Sanayi ve Tic. Ltd. Şti.
Matbaacı Sertifika Numarası: 77398

YAYINLAYAN



Filiz Kitabevi Basım Yayın Dağıtım
Petrol Ürünleri Sanayi ve Tic. Ltd. Şti.
Cemal Yener Tosyalı Cad. No: 57/A (113)
Vefa-Fatih/İSTANBUL
Tel: (212) 527 0 718
(850)441 0 359
Faks: (212) 519 20 71

www.filizkitabevi.com

bilgi@filizkitabevi.com

Yayıncı Sertifika Numarası: 77398

Copyright© Bu kitabın Türkiye'deki yayın hakları Filiz Kitabevi Basım Yayın Dağıtım Petrol Ürünleri San. ve Tic. Ltd. Şti.'ne aittir. Her hakkı saklıdır. Hiçbir bölümü ve paragrafı kısmen veya tamamen ya da özet halinde, fotokopi, faksimile veya başka herhangi bir şekilde çoğaltılamaz, dağıtılamaz. Normal ölçüyü aşan iktibaslar yapılamaz. Normal ve kanunî iktibaslarda kaynak gösterilmesi zorunludur.

Önsöz

Kitabın konusu, veri madenciliği ve modelleri hakkında bilgi vererek bu alanda sıkça kullanılan yeniden örnekleme yöntemleri ile karar ağaçları ve karar ormanları ile ilgili teorik bilgi ve uygulama adımlarından oluşmaktadır. İlk karar ağacı algoritmalarından olan ID3, C4.5, CART (Classification and Regression Tree) algoritmaları ve bölünme kriterleri ile topluluk öğrenme algoritmaları olan artırma (boosting), torbalama (bagging) ve rastgele orman (random forest) algoritmalarından bahsedilmiştir. Ayrıca rastgele orman algoritması için örnek bir veri kümesi üzerinde R programında nasıl uygulanacağı anlatılmıştır. Kitap dört bölümden oluşmakta olup birinci bölümde veri madenciliği ve modelleri açıklanmıştır. İkinci bölümde yeniden örnekleme yöntemlerinden bahsedilmiştir. Üçüncü bölümde ilk karar ağacı algoritmalarından olan ID3, C4.5 ve CART (Classification and Regression Tree) algoritması ve bölünme kriterleri uygulamalı olarak anlatılmıştır. Dördüncü bölümde ise topluluk öğrenme yöntemlerinden artırma (boosting), torbalama (bagging) ve rastgele orman (random forest) algoritmaları anlatılmış ve örnek bir veri kümesi üzerinde rastgele orman yönteminin uygulama adımları R programında gösterilmiştir.

Bu kitabın, veri madenciliği ve modelleri, karar ağaçları algoritmaları ve topluluk öğrenme yöntemleri ile ilgilenen istatistik, veri bilimi, işletme, ekonometri, iktisat gibi konularda çalışan araştırmacılara, lisans ve lisansüstü öğrencilerine fayda sağlaması beklenmektedir. Yazarın Marmara Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı İstatistik Yüksek Lisans Programı kapsamında yazmış olduğu “İkili Değişkenli Verilerde Boole Cebri Yardımıyla Bir Sınıflandırma Algoritmasının Oluşturulması Ve Dijital Reklamcılık Üzerine Uygulanması “Dijital Reklam Verilerinden Yararlanarak Potansiyel Konut Alıcılarının Rastgele Orman Yöntemiyle Sınıflandırılması” başlıklı yüksek lisans tezinden türetilmiştir. Bu aşamada bilgisini, ilgisini ve zamanını benimle paylaşan değerli hocam Prof. Dr. Dilek Altaş’a sonsuz te-

şekkürlerimi sunarım. Her çalışmamada bilgisini ve desteğini esirgemeyip, akademik hayatıma yön veren saygıdeğer hocam Prof. Dr. Mustafa Tekin'e sonsuz teşekkürlerimi sunarım. Her zaman yanımda olup, desteğini esirgemyen ve hayatımı anlamlı kılan sevgili eşim ve meslektaşım Arş. Gör. Rabia Kapıcıoğlu'na da ayrıca müteşekkirim. Ayrıca Filiz Kitabevi çalışanlarına kitabın hazırlanma sürecinde verdikleri destek ve emek için de teşekkür ederim. Kitabın okuyucuya faydalı olmasını temmeni ederim.

Haydar Ekelik

İstanbul, 2025

İçindekiler

Önsöz.....	III
İçindekiler.....	V

1. BÖLÜM

Veri Madenciliği

1.1. Veri Madenciliği Tanımı ve Veri Tabanlarında Bilgi Keşfi.....	1
1.2. Veri Madenciliğinin Uygulandığı Veri Tabanları	4
1.2.1. İlişkisel Veri Tabanları (Relational Databases).....	4
1.2.3. İşlemsel Veri Tabanları (Transactional Databases).....	5
1.2.4. Gelişmiş Veri Tabanı Sistemleri ve Uygulamaları (Advanced Database Systems And Applications).....	5
1.3. Veri Madenciliği Modelleri	5
1.3.1. Sınıflama ve Regresyon Modelleri.....	6
1.3.2. Kümeleme Modelleri.....	7
1.3.3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler (İlişkilendirme Analizi).....	8

2. BÖLÜM

Yeniden Örnekleme Metotları

2.1. Çapraz Doğrulama (Cross-Validation).....	9
2.1.1. Doğrulama Kümesi Yaklaşımı (Validation Set Approach)	10
2.1.2. Tek-Çıkışlı Çapraz Doğrulama (Leave-One-Out Cross-Validation).....	11
2.1.3. K-Katmanlı Çapraz Doğrulama (K-Fold Cross-Validation)	12
2.2. Bootstrap	13

3. BÖLÜM

Karar Ağaçları

3.1. Karar Ağacı Algoritmaları.....	18
3.1.1. ID3 Algoritması (Iterative Dichotomiser 3)	18
3.1.2. C4.5 Algoritması	19
3.1.3. Sınıflandırma ve Regresyon Ağaçları (Clasification and Regression Tree - CART)	20
3.2. Sınıflandırma Ağaçları Bölünme Kriterleri	23
3.2.1. Entropi (Entropy)	23
3.2.2 Bilgi Kazancı (Information Gain)	25
3.2.3 Bilgi Kazanç Oranı (Information Gain Ratio)	25
3.2.4 Gini Endeksi (Gini Index)	26
3.3. Budama Kriterleri	35
3.4. Durdurma Kriteri	35
3.5. Karar Ağaçlarında Değerlendirme.....	36
3.5.1 Karmaşıklık Matrisi (Confusion Matrix).....	36
3.5.2 ROC Eğrisi (Receiver Operating Characteristic Curve).....	38
3.5.3. Eğri Altında Kalan Alan (Area Under Curve - AUC)	39
3.6. Karar ağaçlarının Avantajları ve Dezavantajları	39

4. BÖLÜM

Karar Ormanları (Topluluk Öğrenme Algoritmaları)

4.1. Bağımlı Yöntemler	41
4.1.1. Arttırma (Boosting)	41
4.2. Bağımsız Yöntemler	43
4.2.1. Torbalama (Bagging)	44
4.2.1.1 Torba Dışı Hata Tahmini (Out-of-Bag (OOB) Error Estimation)	45
4.2.1.2 Değişken Önemlilik Ölçümleri (Variable Importance Measures)	45
4.2.2 Rastgele Orman (Random Forest)	46
Kaynakça	59