Veri Bilimi

John D. Kelleher
Brendan Tierney

ÇEVİRİ: ONUR ÖZTÜRK

3. BASKI

Tellekt


Can Yayınları
Bu tanıtım kopyası
Can okurları için
özel olarak hazırlandı.
Can Yayınları

VERİ BİLİMİ

Tellekt_19

Veri Bilimi

Çeviri: Onur Öztürk

Data Science (MIT Press Essential Knowledge series)
İlk basım (çeviriye kaynak alınan basım): MIT Press, 2018

Bu kitap ilk kez 2019 yılında QNB Finansbank işbirliğiyle yapılan Ufuk Açan Yayınlar dizisinde yayımlanmıştır.

1. baskı: 2020
3. basım: Aralık 2022, İstanbul
Bu kitabın 3. baskısı 1000 adet yapılmıştır.

Yayına hazırlayan: Didem Bayındır
Düzelti: Melis Oflas
Mizanpaj: Bahar Kuru Yerek

Kapak Tasarımı ve Uygulama: Bora Başkan
İç Kapak Görseli: Elina Krima

Baskı ve cilt: Matsis Matbaa Hizmetleri San. ve Tic. Ltd. Şti.
Tevfikbey Mahallesi Dr. Ali Demir Caddesi No: 51 Giriş Depo Bölümü
Küçükçekmece-İstanbul
Sertifika No: 40421

ISBN 978-625-7118-08-8

Tellekt
tellekt.com • bilgi@tellekt.com

Maslak Mah. Eski Büyükdere Cad. İz Plaza Giz, No: 9/25 Sarıyer / İstanbul
Telefon: (0212) 252 56 75 / 252 59 88 / 252 59 89 Faks: (0212) 252 72 33
Sertifika No: 43514
Tellekt, Can Sanat Yayınları Yapım ve Dağıtım Ticaret ve Sanayi A.Ş.'nin markasıdır.

twitter.com/tellekt • facebook.com/tellekt • instagram.com/tellekt

VERİ BİLİMİ

JOHN D. KELLEHER
&
BRENDAN TIERNEY

ÇEVİRİ:
ONUR ÖZTÜRK

Tellekt

JOHN D. KELLEHER, Technological University Dublin'deki Bilgi, İletişim ve Eğlence Araştırma Enstitüsü'nün akademik lideridir. Uzmanlık alanları arasında yapay zekâ, veri analizi, yapay öğrenme, uzamsal biliş ve metin analitiği gibi konular bulunmaktadır. Brendan Tierney'le birlikte yazdıkları *Veli Bilimi* dışında *Deep Learning* [Derin Öğrenme] adlı bir başka kitabı daha vardır.

BRENDAN TIERNEY, Dublin Institute of Technology'de veri madenciliği ve gelişmiş veritabanları üzerine ders vermektedir. Yirmi yılı aşkın süredir veri madenciliği, veri depolama ve veritabanı tasarımı alanındaki çalışmalarının ve farklı ülkelerde yürüttüğü projelerin yanı sıra çeşitli teknoloji dergilerinde makaleler yazan Brendan, *UKOUG Oracle Scene* dergisinin editörlüğünü yürütmektedir.

ONUR ÖZTÜRK, 1985 yılında Ankara'da doğdu. Ankara Fen Lisesi mezunudur. Lisans derecesini ODTÜ Elektrik ve Elektronik Mühendisliği bölümünden, yüksek lisans derecesini ise aynı üniversitenin mimarlık fakültesinden aldı. Bilim, felsefe, teknoloji ve matematik alanlarında çeşitli süreli yayınlar için içerik üretmekte ve çeviriler yapmaktadır.

İÇİNDEKİLER

TEŞEKKÜR

John ve Brendan, Paul McElroy ile Brian Leahy'ye ilk taslakları okuyup yorumda bulundukları için teşekkür ediyor. Ayrıca, taslak üzerine detaylı ve faydalı geribildirimler sunan iki anonim hakeme ve MIT Press çalışanlarına destek ve yönlendirmeleri için teşekkür ediyor.

John, ailesi ile dostlarına, bu kitabın hazırlığı süresince verdikleri destek ve cesaret için teşekkür ediyor ve kitabı, sevgisi ve dostluğunun nişanesi olarak babası John Bernard Kelleher'e adıyor.

Brendan, (dördüncü) bir kitap daha yazarken, esas işlerine yetişmeye çalışır ve seyahat ederken sundukları sürekli destek için Grace, Daniel ve Eleanor'a teşekkür ediyor.

ÖNSÖZ

Veri biliminin hedefi, alınacak kararları hacimli veri kümelerinden elde edilen içgörülere dayandırarak karar alma süreçlerini geliştirmektir. Bir faaliyet alanı olarak veri bilimi, hacimli veri kümelerinden, kolayca saptanamayacak faydalı örüntüler çıkarmaya yarayan bir dizi ilkeyi, sorun tanımını, algoritmayı ve süreci içerir. Veri madenciliği ve yapay öğrenme yakından ilişkili olsa da, veri biliminin etkinlik alanı bunlarla sınırlı değildir. Günümüzde veri bilimi modern toplumların hemen her bileşeninde karar alma süreçlerini yönlendiriyor. Veri biliminin gündelik yaşamınızı hangi yollardan etkileyebileceğine dair örneklerden bazıları, çevrimiçi olarak size hangi reklamların sunulacağına; hangi filmler, kitaplar ve arkadaş bağlantılarının tavsiye edileceğine; hangi e-postaların istenmeyen postalar kutunuza düşeceğine; cep telefonu hattınızı değiştirdiğinizde hangi teklifleri alacağınıza; sağlık sigortanızın size ne kadara patlayacağına; muhitinizdeki trafik lambalarında hangi ışıkların ne zaman yanacağına; ihtiyaç duyabileceğiniz ilaçların nasıl tasarlanacağına ve şehrinizde polisin hangi lokasyonlarda suç takibine odaklanacağına benzer kararların verilmesidir.

Toplumlarımızda veri biliminin kullanımındaki artışı güdüleyen etkenler, büyük veri [*big data*] ve sosyal medyanın ortaya çıkışı, bilgi işlem [*computing*] gücündeki hızlanma, bilgisayar belleğinin maliyetindeki muazzam düşüş ve derin öğrenme gibi daha güçlü veri

analizi ve veri modelleme yöntemlerinin gelişimidir. Bu faktörler bir arada şu anlama gelir: Kuruluşlar için veri toplamak, depolamak ve işlemek hiçbir zaman bugünkü kadar kolay olmamıştır. Söz konusu teknik inovasyonlar ve veri biliminin daha geniş ölçekte uygulanışı aynı zamanda şu anlama da gelir: Veri kullanımına ve bireysel mahremiyete ilişkin etik güçlüklerin aşılması hiçbir zaman bugünkü kadar aciliyet kazanmamıştır. Bu kitabın hedefi, okuyucuya veri bilim alanına dair ilkeli bir anlayış kazandıracak derinlikle, alanın en önemli unsurlarını ele alan bir giriş sunmaktır.

Birinci bölüm, veri bilimi alanını tanıtıyor ve onun nasıl gelişip evrimleştiğine dair tarihsel bir özet sunuyor. Aynı zamanda, veri biliminin bugün niçin önemli olduğunu ve benimsenmesinde itici güç olan faktörlerden bazılarını inceliyor. Bölüm, veri bilimiyle ilişkili mitlerden bazılarının incelenmesi ve çürütülmesiyle sona eriyor. İkinci bölüm, veriyle ilişkili temel kavramları tanıtarak, bir veri bilimi projesindeki standart aşamaları tarif ediyor: işin anlaşılması, verinin anlaşılması, verinin hazırlanması, modelleme, değerlendirme ve konuşlandırma. Üçüncü bölüm, hem veri altyapısına hem de birden fazla kaynaktan veri bütünleştirmenin ve büyük verinin ortaya çıkardığı güçlüklere odaklanıyor. Tipik bir veri altyapısının güçlük yaratabilen yönlerinden biri, veritabanındaki ve veri ambarındaki verinin sıklıkla veri analizi için kullanılan sunuculardan farklı sunucularda bulunmasıdır. Bunun sonucu olarak hacimli veri kümeleriyle çalışıldığında, veritabanı ya da veri ambarının bulunduğu sunucular ile veri analizi ve yapay öğrenme için kullanılan sunucular arasında veriyi aktarmak şaşırtıcı derecede uzun zaman alabiliyor. Üçüncü bölüm, kuruluşların kullanımına yönelik standart bir veri bilimi altyapısının tarifi ve bir veri altyapısı içinde hacimli veri kümelerini bir yerden bir yere taşımanın zorluğuyla baş etmek için geliştirilen çözümlerden bazılarının açıklanmasıyla başlıyor. Yapay öğrenmenin veritabanı içinde kullanımı, veri depolamak ve işlemek için Hadoop'un kullanımı ve geleneksel veritabanı programları ile Hadoop benzeri çözümleri sorunsuzca kaynaştıran karma veritabanı sistemlerinin geliştirilmesi bu çözümler arasında. Bölüm, bir kuruluşun farklı bileşenlerinden gelen veri kümelerinin, yapay öğrenmeyle

uyumlu, birleşik bir gösterim oluşturacak şekilde bütünleştirilmesine ilişkin bazı güçlüklere dikkat çekerek sona eriyor. Dördüncü bölüm, yapay öğrenme alanını tanıtıp sinir ağları, derin öğrenme ve karar ağacı modelleri gibi en popüler yapay öğrenme algoritmaları ve modellerinden bazılarını açıklıyor. Beşinci bölüm, bir dizi iş sorununu inceleyip bunların yapay öğrenme çözümleri tarafından nasıl çözülebileceğini betimleyerek yapay öğrenme uzmanlığını gerçek dünyanın sorunlarıyla ilişkilendirmeye odaklanıyor. Altıncı bölüm, veri biliminin yaratabileceği etik sorunları, veri mevzuatlarındaki yeni gelişmeleri ve veri bilimi sürecinde bireylerin mahremiyetini korumaya yönelik yeni bilgi işlem yaklaşımlarından bazılarını inceliyor. Son olarak yedinci bölüm, yakın gelecekte veri biliminin kayda değer ölçüde etkileyeceği bazı alanları tanıtıyor ve bir veri bilimi projesinin başarılı olup olmayacağını belirlemede önem taşıyan ilkelerden bazılarını açıklıyor.

1

VERİ BİLİMİ NEDİR?

Veri bilimi, hacimli veri kümelerinden kolayca saptanamayacak faydalı örüntüler çıkarmaya yarayan bir dizi ilkeyi, sorun tanımını, algoritmayı ve süreci içerir. Veri biliminin çoğu unsuru, yapay öğrenme ve veri madenciliği gibi veri bilimiyle bağlantılı alanlarda geliştirilmiştir. Hatta *veri bilimi*, *yapay öğrenme* ve *veri madenciliği* terimleri sıklıkla birbirinin yerine kullanılır. Bu disiplinlerin ortaklığı, verinin analizi yoluyla karar almayı geliştirmeye odaklanmalarıdır. Gelgelelim, veri bilimi bu diğer alanlara borçlu da olsa, onun etkinlik alanı daha geniştir. Yapay öğrenme (YÖ), veriden örüntüler çıkarmak amacıyla algoritmaların tasarlanmasına ve değerlendirilmesine odaklanır. Veri madenciliği genel olarak yapılandırılmış verinin analiziyle uğraşır ve çoğunlukla ticari uygulamaların önem kazandığı bir alandır. Veri bilimiyse tüm bunları dikkate almakla kalmaz, yapılandırılmamış sosyal medya ve web verilerinin yakalanması [*data capturing*], temizlenmesi ve dönüştürülmesi gibi diğer güçlüklerle; yapılandırılmamış hacimli veri kümelerinin depolanması ve işlenmesi için büyük veri teknolojilerinin kullanımıyla; ayrıca veri etiği ve mevzuatına dair sorularla da meşguldür.

Veri bilimini kullanarak elimizdeki veriden farklı tiplerde örüntüler çıkarabiliriz. Örneğin benzer davranış ve beğeniler sergileyen müşteri gruplarını tanımlamamıza yardım edecek örüntüler çıkarmak isteyebiliriz. İş jargonunda, bu işlem *müşteri segmentasyonu* olarak bilinir ve veri bilimi terminolojisinde buna *öbekleme* [*clustering*] denir. Bunun yerine, çoğunlukla birlikte satın alınan ürünleri tanımlayan bir örüntü çıkarmak da isteyebiliriz, ki bu işleme de *birliktelik kuralları madenciliği* [*association rule mining*] denir. Veya düzmece sigorta talepleri gibi şüpheli ya da anormal olayları tanımlayan örüntüler çıkarmak isteyebiliriz, ki bu işleme *anomali* ya da *uçdeğer tespiti* denir. Son olarak da bir şeyleri sınıflandırmamıza yardımcı olan örüntüleri belirlemek isteyebiliriz. Örneğin, bir e-posta veri kümesinden çıkarılmış sınıflandırma örüntüsünün nasıl olabileceğini gösteren bir kural şöyledir: *Eğer bir e-posta "kolayca para kazan" ifadesini içeriyorsa, muhtemelen istenmeyen bir e-postadır.* Bu tip sınıflandırma kurallarının belirlenmesine *tahmin* denir. *Tahmin* kelimesi tuhaf bir seçim gibi görünebilir, çünkü kural gelecekte ne olacağına dair tahminde bulunmaz: E-posta halihazırda ya istenmeyen bir postadır ya da değildir. Öyleyse en iyisi, tahmin örüntülerini, geleceğe dair bir tahmin değil, bir niteliğin bilinmeyen bir değerine dair tahmin olarak düşünmektir. Bu örnekte, e-posta sınıflandırma niteliğinin "istenmeyen" değerini alıp almayacağına dair bir tahminde bulunuyoruz.

Her ne kadar veri bilimini farklı tip örüntüler çıkarmak için kullanabilsek de, bu örüntülerin hem bariz olmamasını hem de yararlı olmasını isteriz. Önceki paragrafta verilen e-posta sınıflandırma kuralı örneği o kadar basit ve barizdir ki, eğer bir veri bilimi işleminden çıkarılan tek kural olsaydı, hayal kırıklığına uğrardık. Örneğin, bu e-posta sınıflandırma kuralı, bir e-postanın tek bir niteliğini kontrol ediyor: E-posta "kolayca para kazan" ifadesini içeriyor mu? Eğer bir insan uzman, bir örüntüyü kolaylıkla kendi zihninde yaratabiliyorsa, onu "keşfetmek" için veri bilimini kullanmak, genellikle harcanacak zamana ve çabaya değmez. Veri bilimi genellikle çok sayıda veri örneğimiz olduğunda ve örüntüler insanların kendi başlarına keşfedip çıkaramayacağı kadar karmaşık olduğunda yararlı hale

gelir. Bunun için bir insan uzmanın kolayca kontrol edebileceğinden fazla sayılabilecek miktarda veri örneğini bir alt sınır olarak alabiliriz. Örüntülerin karmaşıklığını da, yine insan yeteneklerine kıyasla tanımlayabiliriz. Biz insanlar, bir, iki ya da üç niteliği (sıklıkla *özellikler* ya da *değişkenler* biçiminde de ifade edilir) kontrol eden kuralları tanımlamada epey iyiyiz; ancak üç nitelikten yukarıya çıktığımızda, onlar arasındaki etkileşimlerle başa çıkmakta zorlanmaya başlayabiliriz. Oysa veri bilimi, çoğunlukla onlarca, yüzlerce, binlerce, hatta aşırı durumlarda milyonlarca nitelik arasındaki örüntüleri bulmak istediğimiz bağlamlarda uygulanır.

Veri bilimini kullanarak çıkardığımız örüntülerin işimize yarayabilmesi için, bu örüntülerin mutlaka sorunumuzun çözümüne yardımcı olacak bir şeyler yapmamıza olanak tanıyan içgörüleri bize kazandırması gerekir. *Eyleme geçirilebilir içgörü* [*actionable insight*] ifadesi bu bağlamda bazen, çıkarılan örüntülerden ne beklediğimizi ifade etmek için kullanılır. *İçgörü* terimi, örüntünün bize, sorunumuza ilişkin bariz olmayan, geçerli bilgiyi vermesi gerektiğini vurgular. *Eyleme geçirilebilir* terimi ise, edindiğimiz içgörünün de öyle veya böyle kullanabileceğimiz bir şey olması gerektiğini vurgular. Örneğin, müşteri *kaybı* [*churn*] –yani çok sayıda müşterinin başka şirketlere kayması– sorununu çözmeye çalışan bir cep telefonu şirketi için çalıştığımızı hayal edelim. Veri biliminin bu sorunu gidermek için uygulamaya koyulmasına bir örnek, önceki müşterilere ilişkin verilerden, kaybetme riskimiz olan mevcut müşterileri saptamamızı sağlayacak örüntüleri çıkarmak, ardından da bu müşterilerle temas kurup onları bizimle kalmaya ikna etmektir. Kaybedilme olasılığı yüksek müşterileri saptamamıza imkân veren bir örüntü ancak (*a*) bize müşterilerimizi kaybetmeden önce onlarla temasa geçme vakti tanıyacak süre içinde saptamayı yapabiliyorsa ve (*b*) şirketimiz müşterilerle temasa geçecek bir ekip kurabiliyorsa yararlıdır. Örüntülerin bize kazandırdığı içgörü üzerine şirketimizin eyleme geçebilmesi için bunların her ikisi de gereklidir.

Veri Biliminin Kısa Tarihi

Veri bilimi teriminin 1990'lı yıllara dayanan özgün bir tarihi var. Gelgelelim faydalandığı alanlar daha uzun bir geçmişe sahip. Bu daha uzun tarihsel sürecin bileşenlerinden biri veri toplamanın, diğeri ise veri analizinin tarihidir. Kitabın bu kısımda, söz konusu dallardaki başlıca gelişmeleri inceleyecek ve onların nasıl ve neden veri bilimi alanında birleştiklerini açıklayacağız. Ortaya çıkan önemli teknik inovasyonları tanıtıp adlandırırken, bu incelemede ister istemez yeni bir terminolojiye de giriş yapacağız. Bahsi geçen her yeni terimin anlamına dair kısa bir açıklama sunduktan sonra, ilerleyen bölümlerde bu terimlerin çoğuna geri dönüp daha detaylı açıklamalarda bulunacağız. İncelememiz veri toplamanın tarihiyle başlayıp, veri analizinin tarihiyle devam edecek ve veri biliminin gelişimini ele alarak sonlanacak.

Veri Biriktirmenin Tarihi

Verileri kaydetmenin en eski yöntemleri, günlerin geçişini çubuklara çentik atarak işaretlemek veya gün dönümlerinde güneşin doğuş zamanını belirlemek için zemine direk dikmek olabilir. Fakat yazının gelişmesiyle, dünyadaki deneyimlerimizi ve olayları kaydetme becerimiz, toplayabildiğimiz veri miktarını da muazzam ölçüde artırdı. Yazının en eski biçimi Mezopotamya'da MÖ 3200 civarında gelişti ve ticari kayıtları tutmak için kullanıldı. Bu tip kayıt tutma, *işlemsel veri* [*transactional data*] olarak bilinen verileri kaydeder. İşlemsel veri bir maddenin satışı, bir faturanın düzenlenmesi, malların teslimi, kredi kartı ödemesi, sigorta tazminat talepleri vb. olayların bilgisini içerir. Demografik veri gibi *işlemsel olmayan veri*lerin de uzun bir tarihi vardır. Bilinen en eski nüfus sayımları, MÖ 3000 yıllarında firavunlar dönemi Mısır'ında gerçekleştirilmiştir. Eski devletlerin kapsamlı veri toplama operasyonlarına bu denli çaba ve kaynak ayırmasının nedeni, bu devletlerin vergilerini ve ordularını büyütmeye ihtiyaç duymasıydı; ki bu da Benjamin Franklin'in yaşamda yalnızca iki şeyin kesin olduğuna dair iddiasını kanıtlıyor: ölüm ve vergiler.

Son 150 yılda elektronik sensörün geliştirilmesi, verinin dijitalleştirilmesi ve bilgisayarın icadı, toplanan ve depolanan veri miktarındaki devasa artışa katkıda bulundu. Veri toplama ve depolamadaki kilometre taşlarından biri, 1970'te, Edgar F. Codd'un *ilişkisel veri modeli*ni [*relational data model*] açıkladığı bir makale yayımlaması oldu. Bu model, (o sırada) verilerin nasıl depolandığını, dizinlendiğini ve veritabanlarından nasıl çekildiğini ortaya koyması açısından devrimciydi. Model, kullanıcıların istedikleri veriyi tanımladığı basit sorgular yardımıyla veritabanlarından veri çıkarabilmesini sağlıyordu. Bu sorgular, kullanıcının veri yapısına ya da verilerin nerede depolandığına kafa yormasını gerektirmeksizin, istenen veriye ulaşılabilmesini sağlıyordu. Codd'un makalesi modern veritabanlarının temelini atmış ve veritabanı sorgulama tanımları için uluslararası bir standart olan *Structured Query Language*'ın (SQL)* gelişimine zemin hazırlamıştı. İlişkisel veritabanları, verileri her örneğin bir satırda, her niteliğin bir sütunda gösterildiği tablolarda depolar. Bu yapı veri depolama için idealdir, çünkü veri doğal niteliklerine ayrıştırılabilir.

Veritabanları, *işlemsel* veya *operasyonel* olarak adlandırılan verinin –yani bir şirketin günlük operasyonları tarafından üretilen veri türünün– yapılandırılmış biçimini depolamak ve çekmekte kullanılan doğal teknolojidir. Gelgelelim, şirketler büyüyüp daha otomatize hale geldikçe, bu şirketlerin farklı birimlerince üretilen veri miktarı ve çeşitliliği de çarpıcı ölçüde artmıştır. 1990'larda, şirketler mahşeri miktarlarda veri biriktirdikleri halde, bu verileri analiz etmede sürekli zorlukla karşılaştıklarını fark ettiler. Sorunun bir yönü, verilerin çoğunlukla kuruluşların bünyesinde, çok sayıda farklı veritabanında depolanmasıydı. Diğer bir zorluksa, veritabanlarının veriyi depolamak ve geri çekmek için optimize edilmiş olmasıydı, ki bu etkinlikler SEÇ, EKLE, GÜNCELLE ve SİL gibi basit işlemlerin yoğun kullanımıyla yapılıyordu. Bu şirketler kendi verilerini analiz etmek için ayrık veritabanlarından verileri bir araya getirip birleştirebilen ve daha karmaşık analitik veri işlemlerini kolaylaştırabilen

* Yapılandırılmış Sorgu Dili. (Ç.N.)


Can Yayınları
Bu tanıtım kopyası
Can okurları için
özel olarak hazırlandı.
Can Yayınları

Bugün internette gördüğümüz ilanlar, bizlere önerilen kitap ve filmler, istenmeyen e-postalar, hatta sağlık sigortamızın primi veri bilimiyle belirleniyor. Veri toplamak, depolamak ve işlemek hiç bu kadar kolay olmamıştı; bu da veri biliminin büyük bir hızla gelişmesini açıklıyor.

Veri Bilimi alanın tarihçesiyle açılıyor; temel veri kavramlarını, veri bilimi projelerinin aşamalarını, veri altyapısını, çoklu kaynaklardan veri elde etmenin zorluklarını ve bu konudaki uzmanlığın gerçek dünyada karşılaştığımız sorunlara nasıl bir yaklaşım getireceğini ele alıyor. Konuyu ahlaki ve hukuki boyutları açısından da gözden geçiren *Veri Bilimi*, bu bilim dalını başarıyla kullanma yollarını ve gelecekte nasıl bir etki alanı yaratacağını merak edenlere ideal bir kaynak sunuyor.

Tellekt

www.tellekt.com

ISBN 978-625-7118-08-8

9 786257 118088

Bilim | Teknoloji