

برنام خدا

---

داده کاوی با  
**Python**

نویسندگان:

گالیت شمولی، پیتر سی. بروس، پیتر گدک، نایتین آر. پیتل

---

مترجمان:

دکتر مهدی اسماعیلی

دکتر سیدمهدی وحیدی پور

# داده‌کاوی با Python

مترجمان: دکتر مهدی اسماعیلی، دکتر سیدمهدی وحیدی‌پور

ویراستار علمی: دکتر رامین مولاناپور

ناشر: انتشارات آتی‌نگر

ناشر همکار: انتشارات وینا

طراحی جلد و صفحه‌آرایی: همتا بیداریان

چاپ اول، ۱۴۰۰

شمارگان: ۱۰۰۰ نسخه

قیمت: ۹۵۰,۰۰۰ ریال

شابک: ۹۷۸-۶۲۲-۷۵۷۱-۱۷-۲

ISBN: 978-622-7571-17-2

حق چاپ برای انتشارات آتی‌نگر محفوظ است.

نشانی دفتر فروش: خیابان جمالزاده جنوبی، روبه‌روی کوچه رشتچی، پلاک ۱۴۴، واحد ۱

نمابر: ۶۶۵۶۵۳۳۷

تلفن: ۸-۶۶۵۶۵۳۳۶



www.ati-negar.com \* info@ati-negar.com

داده‌کاوی با Python [ گالیت شمولی، پیتر سی. بروس... و دیگران ]؛ مترجمان: دکتر مهدی اسماعیلی، دکتر سیدمهدی وحیدی‌پور

تهران: آتینگر، وینا ۱۴۰۰

۶۰۰ ص.: مصور، جدول، نمودار.

ISBN: 978-622-7571-17-2

فیبا.

یادداشت: نویسندگان گالیت شمولی، پیتر سی. بروس، پیتر گدک، نایتین آر. پیتل.

یادداشت: عنوان اصلی کتاب: Data mining for business analytics : concepts, techniques and applications in Python, c 2020.

یادداشت: کتابنامه: ص. [۵۹۹] - ۶۰۰.

موضوع: ریاضیات بازرگانی -- برنامه‌های کامپیوتری - Business mathematics -- Computer programs

موضوع: کسب و کار -- داده‌پردازی - داده‌کاوی - Data processing -- Business - Data mining

موضوع: پایتون (زبان برنامه‌نویسی کامپیوتر) - Python (Computer program language)

شناسه افزوده: شمولی، گالیت، ۱۹۷۱-م. Shmueli, Galit, 1971

شناسه افزوده: اسماعیلی، مهدی، ۱۳۵۰-، مترجم

شناسه افزوده: وحیدی‌پور، سیدمهدی، ۱۳۵۶-، مترجم

شناسه افزوده: رامین، مولاناپور، ۱۳۵۲-، ویراستار

HF۵۵۴۸/۲

۶۵۰/۰۷۳۷

۷۶۳۵۹۲۶

رده‌بندی کنگره

رده‌بندی دیویی

شماره کتابشناسی ملی

# فهرست مطالب

## پیشگفتار

۹

## فصل ۱: مقدمه

۱۱

- ۱-۱ تحلیل‌شناسی کسب‌وکار چیست؟..... ۱۱
- ۲-۱ داده‌کاوی چیست؟..... ۱۳
- ۳-۱ داده‌کاوی و عبارات مرتبط..... ۱۴
- ۴-۱ داده‌های بزرگ..... ۱۵
- ۵-۱ علوم داده..... ۱۶
- ۶-۱ چرا روش‌های متفاوت زیادی وجود دارند؟..... ۱۷
- ۷-۱ واژه‌شناسی و اصطلاحات..... ۱۷
- ۸-۱ نقشه راه این کتاب..... ۲۰

## فصل ۲: مروری بر فرایند داده‌کاوی

۲۵

- ۱-۲ مقدمه..... ۲۵
- ۲-۲ ایده‌های اصلی در داده‌کاوی..... ۲۶
- ۳-۲ گام‌های داده‌کاوی..... ۲۹
- ۴-۲ گام‌های مقدماتی..... ۳۱
- ۵-۲ قدرت پیش‌بینی و بیش‌برازش..... ۴۵
- ۶-۲ ساخت یک مدل پیش‌بینی..... ۵۲
- ۷-۲ استفاده از Python برای داده‌کاوی روی یک ماشین محلی..... ۵۷
- ۸-۲ خودکارسازی راه‌حل‌های داده‌کاوی..... ۵۸
- ۹-۲ اخلاق در داده‌کاوی..... ۵۹

## فصل ۳: مصورسازی داده‌ها

۷۱

- ۱-۳ مقدمه..... ۷۱
- ۲-۳ معرفی دو مجموعه..... ۷۴
- ۳-۳ نمودارهای پایه‌ای: نمودار میله‌ای، گراف‌های خطی و نمودارهای پراکنشی..... ۷۶
- ۴-۳ مصورسازی چند بُعدی..... ۸۶
- ۵-۳ مصورسازی‌های تخصصی..... ۱۰۲
- ۶-۳ خلاصه: مصورسازی با هدف داده‌کاوی..... ۱۱۰

#### فصل ۴: کاهش ابعاد

۱۱۵	
۱۱۶	۱-۴ مقدمه
۱۱۶	۲-۴ مصیبت ابعاد
۱۱۷	۳-۴ ملاحظات کاربردی
۱۱۸	۴-۴ خلاصه‌های داده‌ها
۱۲۳	۵-۴ تحلیل همبستگی
۱۲۴	۶-۴ کاهش تعداد طبقه‌ها در متغیرهای طبقه‌ای
۱۲۵	۷-۴ تبدیل یک متغیر طبقه‌ای به یک متغیر عددی
۱۲۵	۸-۴ تحلیل مؤلفه‌های اصلی
۱۳۸	۹-۴ کاهش ابعاد با استفاده از مدل‌های رگرسیون
۱۳۹	۱۰-۴ کاهش ابعاد با استفاده از درختان رده‌بندی و رگرسیون

#### فصل ۵: ارزیابی کارایی

۱۴۵	
۱۴۶	۱-۵ مقدمه
۱۴۷	۲-۵ ارزیابی کارایی روش‌های پیش‌بینی
۱۵۳	۳-۵ تشخیص کارایی رده‌بند
۱۶۷	۴-۵ بررسی کارایی رتبه‌بندی
۱۷۲	۵-۵ بیش‌نمونه‌گیری

#### فصل ۶: رگرسیون خطی چندگانه

۱۸۳	
۱۸۴	۱-۶ مقدمه
۱۸۴	۲-۶ مدل‌سازی تبیینی (توضیحی) در مقابل مدل‌سازی پیش‌بینانه
۱۸۶	۳-۶ تخمین معادله رگرسیون و پیش‌بینی
۱۹۲	۴-۶ انتخاب متغیر در رگرسیون خطی

#### فصل ۷: K نزدیک‌ترین همسایه

۲۱۱	
۲۱۱	۱-۷ رده‌بند KNN (متغیر خروجی طبقه‌ای)
۲۲۰	۲-۷ الگوریتم KNN برای خروجی عددی
۲۲۲	۳-۷ مزایا و معایب الگوریتم‌های KNN

#### فصل ۸: رده بیز ساده

۲۲۷	
۲۲۷	۱-۸ مقدمه
۲۲۹	۲-۸ اعمال رده‌بند بیزین به صورت کامل
۲۳۹	۳-۸ مزایا و معایب بیز ساده

## فصل ۹: درختان رده‌بندی و رگرسیون

۲۴۵	۱-۹ مقدمه
۲۴۶	۲-۹ درختان رده‌بندی
۲۴۸	۳-۹ ارزیابی کارایی یک درخت رده‌بندی
۲۵۵	۴-۹ اجتناب از بیش‌برازش
۲۶۱	۵-۹ استخراج قواعد رده‌بندی از درخت‌ها
۲۶۲	۶-۹ درختان رده‌بندی برای بیش از دو رده
۲۶۸	۷-۹ درختان رگرسیون
۲۶۸	۸-۹ بهبود پیش‌بینی: جنگل‌های تصادفی و درختان تقویت شده
۲۷۲	۹-۹ مزایا و معایب یک درخت

## فصل ۱۰: رگرسیون لجستیک

۲۸۳	۱-۱۰ مقدمه
۲۸۴	۲-۱۰ مدل رگرسیون لجستیک
۲۸۵	۳-۱۰ مثال پذیرش وام
۲۸۶	۴-۱۰ ارزیابی کارایی رده‌بندی
۲۹۳	۵-۱۰ تعمیم رگرسیون لجستیک برای شرایطی با بیش از دو رده
۲۹۶	۶-۱۰ مثالی از یک تحلیل کامل: پیش‌بینی پروازهای تأخیری

## فصل ۱۱: شبکه‌های عصبی

۳۱۹	۱-۱۱ مقدمه
۳۲۰	۲-۱۱ مفهوم و ساختار یک شبکه عصبی
۳۲۰	۳-۱۱ برازش یک شبکه برای داده‌ها
۳۲۱	۴-۱۱ ورودی‌هایی که لازم است کاربر تعیین کند
۳۳۴	۵-۱۱ کاوش رابطه میان متغیرها و خروجی
۳۳۶	۶-۱۱ یادگیری ژرف
۳۳۶	۷-۱۱ نقاط ضعف و قوت شبکه‌های عصبی

## فصل ۱۲: تحلیل تشخیصی

۳۴۷	۱-۱۲ مقدمه
۳۴۸	۲-۱۲ فاصله یک رکورد از یک رده
۳۴۹	۳-۱۲ توابع رده‌بندی خطی فیشر
۳۵۱	۴-۱۲ کارایی رده‌بندی در تحلیل تشخیصی
۳۵۶	۵-۱۲ احتمالات پیشین

۳۵۷	۶-۱۲ هزینه‌های نابرابر برای رده‌بندی‌های نادرست
۳۵۸	۷-۱۲ رده‌بندی بیش از دو کلاس
۳۶۲	۸-۱۲ نقاط ضعف و قوت

### فصل ۱۳: روش‌های ترکیبی: مدل‌سازی تلفیقی و uplift

۳۶۷	
۳۶۸	۱-۱۳ مدل‌سازی تلفیقی
۳۷۴	۲-۱۳ مدل‌سازی uplift
۳۸۱	۳-۱۳ خلاصه

### فصل ۱۴: قواعد انجمنی و پالایش مشارکتی

۳۸۵	
۳۸۶	۱-۱۴ قواعد انجمنی
۴۰۱	۲-۱۴ پالایش مشارکتی
۴۱۰	۳-۱۴ خلاصه

### فصل ۱۵: تحلیل خوشه

۴۱۷	
۴۱۸	۱-۱۵ مقدمه
۴۲۲	۲-۱۵ محاسبه فاصله میان دو رکورد
۴۲۶	۳-۱۵ محاسبه فاصله میان دو خوشه
۴۳۰	۴-۱۵ خوشه‌بندی سلسله‌مراتبی (تجمیعی)
۴۳۸	۵-۱۵ خوشه‌بندی غیرسلسله‌مراتبی: الگوریتم k-Means

### فصل ۱۶: سری‌های زمانی

۴۴۷	
۴۴۸	۱-۱۶ مقدمه
۴۴۹	۲-۱۶ مدل‌سازی توصیفی در مقابل پیش‌بینانه
۴۴۹	۳-۱۶ روش‌های رایج پیش‌بینی در کسب‌وکار
۴۵۰	۴-۱۶ مؤلفه‌های سری‌های زمانی
۴۵۵	۵-۱۶ افزایش داده‌ها و ارزیابی کارایی

### فصل ۱۷: پیش‌بینی مبتنی بر رگرسیون

۴۶۳	
۴۶۴	۱-۱۷ مدلی با روند
۴۷۱	۲-۱۷ مدلی با الگوی فصلی
۴۷۴	۳-۱۷ مدلی با روند و الگوی فصلی
۴۷۶	۴-۱۷ خودهمبستگی و مدل ARIMA

## فصل ۱۸: روش‌های هموارسازی

۴۹۵	۱-۱۸ مقدمه
۴۹۶	۲-۱۸ میانگین متحرک
۴۹۶	۳-۱۸ هموارسازی نمایی ساده
۵۰۲	۴-۱۸ هموارسازی نمایی پیشرفته

## فصل ۱۹: تحلیل شبکه اجتماعی

۵۱۹	۱-۱۹ مقدمه
۵۱۹	۲-۱۹ شبکه‌های جهت‌دار در مقابل بی‌جهت
۵۲۱	۳-۱۹ مصورسازی و تحلیل شبکه‌ها
۵۲۲	۴-۱۹ متریک‌هایی برای تحلیل شبکه‌ها
۵۲۶	۵-۱۹ استفاده از متریک‌های شبکه در پیش‌بینی و رده‌بندی
۵۳۲	۶-۱۹ جمع‌آوری داده‌های شبکه‌های اجتماعی با Python
۵۳۷	۷-۱۹ مزایا و معایب

## فصل ۲۰: متن‌کاوی

۵۴۱	۱-۲۰ مقدمه
۵۴۲	۲-۲۰ نمایش جدولی متن: ماتریس عبارت-سند و سید واژگان
۵۴۲	۳-۲۰ سید واژگان در مقابل استخراج معنی در سطح سند
۵۴۳	۴-۲۰ پیش‌پردازش متن
۵۴۴	۵-۲۰ پیاده‌سازی روش‌های داده‌کاوی
۵۵۳	۶-۲۰ مثال: بحث در مورد خودروها و وسایل الکترونیکی
۵۵۳	۷-۲۰ خلاصه

## فصل ۲۱: چند مثال کاربردی

۵۶۱	۱-۲۱ باشگاه کتاب Charles
۵۶۱	۲-۲۱ مجموعه داده‌های German Credit
۵۶۷	۳-۲۱ شرکت کامپیوتری Tayko
۵۷۱	۴-۲۱ ترغیب سیاسی
۵۷۴	۵-۲۱ لغو کردن تاکسی‌ها
۵۷۸	۶-۲۱ بخش‌بندی مشتریان
۵۷۹	۷-۲۱ جذب سرمایه
۵۸۳	۸-۲۱ فروش مکمل
۵۸۵	۹-۲۱ پیش‌بینی درخواست حمل‌ونقل عمومی (سری‌های زمانی)