

داده‌های کاربردی با R

مولفان:

محمد مرادی

مجید جوادی

سهیلا مهرمولایی



نشر دانشگاهی کیان
Kian Publication

سرشناسه

عنوان و نام پدیدآور

مشخصات نشر

مشخصات ظاهری

شابک

وضعیت فهرست نویسی

موضوع

موضوع

شناسه افزوده

شناسه افزوده

زده بندی کنگره

زده بندی دیویی

شماره کتابشناسی ملی

مرادی، محمد، ۱۳۶۶.

داده‌کاوی کاربردی R / مولفان محمد مرادی، مجید جوادی، سهیلا مهرمولایی.

تهران: انتشارات دانشگاهی کیان، ۱۳۹۶.

۳۶۷ ص.: مصور.

۹۷۸-۶۰۰-۳۰۷-۱۶۸-۱

فیبا.

داده کاوی. Data mining.

داده کاوی — نرم افزار. Data mining — Software.

جوادی، مجید، ۱۳۶۶.

مهرمولایی، سهیلا، ۱۳۶۲.

۱۳۹۶ ۷۶/۹/۵۲م ۱۳۹۶ QA

۰۰۶/۳۱۲

۴۶۴۰ ۲۳۷



نشر دانشگاهی کیان
Kian Publication

انتشارات دانشگاهی کیان

نام کتاب : داده کاوی کاربردی با R

مولفان : محمد مرادی - مجید جوادی - سهیلا مهرمولایی

ناظر فنی : علی محمودی

ویراستار : لیلا رفیعی

صفحه آرا : مرضیه امانت

طراح جلد : شیلان هوشیاری

چاپ اول : ۱۳۹۶

تیراژ : ۱۰۰۰

چاپ : ستاره سبز

صحافی : نمونه

قیمت : ۲۵۰۰۰ تومان

شابک : ۹۷۸-۶۰۰-۳۰۷-۱۶۸-۱

ISBN : 978-600-307-168-1



خرید اینترنتی آسان از:

www.kianpub.com

بر اساس قانون حقوق مولفان و مصنفان، کلیه حقوق چاپ و نشر این کتاب به طور انحصاری به نشر دانشگاهی کیان تعلق دارد و هرگونه استفاده و برداشت از محتوای این اثر به هر شکلی اعم از چاپ، کپی، اسکن، لوح فشرده، نشر الکترونیک و اینترنتی یا به صورت هرگونه فایل رایانه‌ای، بدون مجوز رسمی ناشر ممنوع و حرام شرعی است و پیگرد قانونی دارد.



kianpublication

برای دریافت اخبار و اطلاعات
مفید و شرکت در قرعه کشی، ما
را در این شبکه‌ها دنبال کنید.

سخن ناشر

«سپس، به کاتبان و نویسندگان بنگر و بهترین آنها را بر کارهای خود بگمار...»

کاتبان و نویسندگانی برگزین که قدر خود را بشناسند، چون کسی که به قدر خود شناخت ندارد، دیگران را هم نمی‌شناسد.»

اگرچه نوشتن و باختن زکات علم از توصیه‌های اکید بزرگان و گواه بر کرامت اهل دانش است؛ اما امروزه پرداختن به اجیزه‌ها و اهداف نوشتن بیش‌تر جلوه می‌کند. بی‌شک این‌که چه کسی می‌نویسد مهم نیست؛ اما این‌که چرا و به چه منظورهایی می‌نویسد، درخور تأمل است.

ما معتقدیم که چاپ روزانه کتاب‌های به اصطلاح «زرد» که خالی از هرگونه نوآوری و بی‌توجه به استانداردهای چاپ کتاب و نیازهایی محلی است، حاصل تفکر بازاری مستولی بر جامعه‌ی نشر است. بی‌پرده آن‌که عنوان پر زرق و برق و ستاره‌ساز دادن مضمون‌های نو با هدف فروش بالا و طویل‌کردن سیاهه‌ی سابقه‌ی علمی، نمی‌تواند دلیل مبرمی برای چاپ و نشر کتابی باشد که خواننده‌ی مشتاق با صرف هزینه‌های نه چندان کم آن را تهیه می‌کند؛ به امید آن‌که پس از آن بیاموزد.

باید پذیرفت که انگیزه‌ی نوشتن کم از محتوای نوشته نیست و بین این دو رابطه‌ای مستقیم برقرار است. اگر انگیزه‌ی نوشتن، تولید دانش باشد، بی‌شک نویسنده از قلم به محتوای کم‌عمق پرهیز می‌کند و اگر دغدغه‌ی دانش و فرهنگ زخم‌خورده در میان باشد، ناشر تنها به عنوان پرطمپا، قفسه نمی‌کند.

و چقدر امروزه، فرهنگ و دانش این مرز و بوم که گرفتار آفت بی‌انگیزگی و رخم هوس است، نیازمند ناشران و نویسندگانی است که نیت‌شان کمک به رشد دانش و ارتقای فرهنگ جامعه است و به راستی که التیامی بر این درد نیست؛ مگر نویسندگانی که قدر خود و دیگران را می‌دانند و خوب می‌دانند که کتاب، ابزار سودجویی‌های مغرضانه نیست و می‌کوشند تا خود را از هرگونه عطش نام و رسم و ثروت بی‌همی‌کنند.

ما در انتشارات دانشگاهی کیان خود را بر بی‌ای از عیب و خطا نمی‌دانیم؛ اما همواره بیش از پیش می‌کوشیم تا در راستای تولید علم و نشر کتاب‌های پرمحتوا، دست نویسندگانی که انگیزه‌ی پاک دارند را بفشاریم و در کنارشان باشیم و از خداوند متعال می‌خواهیم که در این مسیر صعب و پرخطر در سایه‌ی لطف و عنایت خود از آن‌چه به عهده‌ی ما نهاده شده، سربلند و پیروز بر آییم.

انتشارات دانشگاهی کیان

مقدمه مولفان

در حالی که بیش از دو دهه از مطرح شدن اصول و مفاهیم داده‌کاوی می‌گذرد، به جرات می‌توان گفت که اکنون در دوران شکوفایی و اوج کاربردهای آن هستیم. در واقع، در سال‌های اخیر و در روبه‌روی با حجم بسیار زیادی از داده‌ها در حوزه‌های مختلف کاری، استفاده از داده‌کاوی به عنوان یکی از الزامات فرایندهای علمی، پژوهشی و تجاری مطرح شده است. بر این اساس، به ندرت می‌توان فرایندها و پروژه‌های مختلف پژوهشی و تجاری را یافت که از ابزارهای پرشمار داده‌کاوی بهره‌ای نجسته باشند.

در پی این اقبال عمومی به استفاده از روش‌های داده‌کاوی، یکی از مسایل و چالش‌های مهم، البته از دیدگاه نشر و کارشناسان علوم رایانه، چگونگی پیاده‌سازی الگوریتم‌ها و روش‌های داده‌کاوی و اعمال آن بر مجموعه‌های داده‌ای بوده و می‌باشد.

اگرچه بسیاری از زبان‌های برنامه‌نویسی و سیستم‌های بانک‌های اطلاعاتی، امکانات و تسهیلاتی را برای انجام فرایندهای داده‌کاوی فراهم می‌نمایند، اما به کارگیری ابزارهای اختصاصی که به صورت قابل قبولی از گام‌های مختلف فرایند داده‌کاوی پشتیبانی کند، یکی از خواسته‌های منطقی جامعه کاربری می‌باشد.

بر همین اساس، ابزارها و زبان‌های برنامه‌نویسی و محاسبات معرفی شده‌اند، ولی آنچه در طی سال‌های اخیر توجهات را به صورت گسترده‌ای به خود جلب کرده است، زبان R می‌باشد. آنچه این زبان برنامه‌نویسی را از سایر رقبایش متمایز می‌سازد، قابلیت‌های تعبیه شده در آن به عنوان زبانی برای انجام محاسبات و تحلیل‌های آماری است. از این رو، می‌توان این زبان برنامه‌نویسی را یک ابزار ایده‌آل برای انجام فرایندهای تحلیل داده و به صورت مشخص داده‌کاوی دانست. شاهد این ادعا نیز آمارهای منتشر شده از سوی منابع معتبر است که میزان اقبال بالای متخصصان به این زبان را تایید می‌کند.

با توجه به ویژگی‌ها و اهمیت این زبان، در این کتاب بر آن شدیم که مروری کلی بر چگونگی انجام فرایندهای تعامل با داده با استفاده از R داشته باشیم و دریچه‌ای نو بر روی کارشناسان و پژوهشگران عرصه تحلیل داده بگشاییم. در واقع هدف اصلی کتاب حاضر این است که قابلیت‌هایی بیشتر و انعطاف‌پذیرتر از ابزارها و بسته‌های نرم‌افزاری مربوط به داده‌کاوی که استفاده از آن‌ها در میان دانشجویان و پژوهشگران رایج است

را به مخاطبان ارایه کند و این مهم از طریق یادگیری زبان R و افزونه‌های مربوط به آن محقق می‌شود.

البته لازم به ذکر است که بررسی تمامی جوانب و نکات مربوط به استفاده از این زبان برای داده‌کاوی، احتیاج به چندین جلد کتاب دارد؛ از این رو در این کتاب تلاش بر این بوده است که بتوانیم ضمن پرداختن به مباحث متنوع و در عین حال مهم، امکان فراگیری ساده و به‌دور از پیچیدگی را نیز فراهم آوریم.

آنچه پیش‌نیاز بهره‌برداری کامل از این کتاب، آشنایی اولیه با مفاهیم زبان R و نیز اصول داده‌کاوی می‌باشد، اما به‌منظور تسهیل فرایند آموزش، مقدمه‌ای بر مفاهیم ذکر شده در این کتاب ابتدایی ارایه شده است.

در انتهای فصل تشکر از مدیریت و دست‌اندرکاران نشر دانشگاهی کیان که در طی مراحل تکمیل این کتاب، نهایت همکاری را با مولفان داشتند، از خوانندگان گرامی درخواست می‌کنیم تا نظرات پیشنهادات و انتقادات خود درباره کتاب حاضر را از طریق پست الکترونیکی RDataMining'@gmail.com با ما در میان بگذارند.

فصل اول: مقدمه‌ای بر زبان R

۱۵	۱-۱. معرفی R
۱۶	۲-۱. نصب و راه‌اندازی
۲۷	۳-۱. کنسول R
۲۸	۴-۱. متغیرها
۲۸	۵-۱. توابع
۲۹	۶-۱. اشیا
۳۰	۷-۱. بردارها
۳۲	۸-۱. برداری‌سازی
۳۴	۹-۱. فاکتورها
۳۷	۱۰-۱. ایجاد دنباله‌ها
۳۹	۱۱-۱. Sub-Setting
۴۱	۱۲-۱. ماتریس‌ها و آرایه‌ها
۴۳	۱۳-۱. لیست‌ها (فهرست‌ها)
۴۵	۱۴-۱. قالب‌های داده‌ای
۴۹	۱۵-۱. خواندن داده‌ها

فصل دوم: مفاهیم اولیه‌ی داده‌کاوی

۵۳	۱-۲. انگیزه‌های استفاده از داده‌کاوی
۵۴	۲-۲. مقدمه‌ای بر داده‌کاوی
۶۰	۳-۲. داده‌کاوی از چند دیدگاه مختلف
۶۶	۴-۲. محدودیت‌های داده‌کاوی

فصل سوم: شروع کار با Rattle و داده‌ها

۶۹	۱-۳. نصب Rattle
۷۷	۲-۳. کار با داده‌ها
۸۳	۳-۳. تعامل با داده‌ها با استفاده از R

فصل چهارم: بارگذاری داده‌ها

۹۰	۱-۴ داده‌های CSV
۹۵	۲-۴ داده‌های ARFF
۹۷	۳-۴ داده‌هایی با منبع ODBC
۹۹	۴-۴ مجموعه‌های داده‌ای R، منابع داده‌ای دیگر
۱۰۲	۵-۴ RData
۱۰۳	۶-۴ جابجانه
۱۰۴	۷-۴ رینه‌های مشترک

فصل پنجم: پودن داده‌ها

۱۱۰	۱-۵ خلاصه‌سازی
۱۱۸	۲-۵ بازنمایی بصیرت زیج
۱۳۸	۳-۵ تحلیل همبستگی

فصل ششم: گرافیک‌های تعاملی

۱۴۶	۱-۶ بسته Latticist
۱۴۸	۲-۶ بسته GGobi

فصل هفتم: تبدیل و انتقال داده‌ها

۱۵۸	۱-۷ مسایل مربوط به داده‌ها
۱۶۱	۲-۷ تبدیل (انتقال) داده‌ها
۱۶۲	۳-۷ مقیاس‌دهی مجدد داده‌ها
۱۶۸	۴-۷ Imputation
۱۷۱	۵-۷ Recoding
۱۷۴	۶-۷ پاک‌سازی

فصل هشتم: خوشه‌بندی داده‌ها

۱۷۶	۱-۸ بازنمایی دانش
-----	-------------------

۱۷۷	۲-۸. جست و جوی اکتشافی
۱۷۸	۳-۸. معیارها
۱۸۱	۴-۸. مثال
۱۸۷	۵-۸. نکات پایانی

فصل نهم: روش، قواعد انجمنی

۱۹۰	۱-۹. بازنمایی دانش
۱۹۰	۲-۹. جست و جوی اکتشافی
۱۹۲	۳-۹. معیارها
۱۹۳	۴-۹. مثال
۱۹۷	۵-۹. ایجاد مدل با استفاده از روش

فصل دهم: طبقه بندی داده ها با روش اشریف ردلر پشتیبان

۲۰۳	۱-۱۰. بازنمایی دانش
۲۰۶	۲-۱۰. الگوریتم
۲۰۹	۳-۱۰. مثال
۲۱۱	۴-۱۰. مدل سازی با استفاده از R
۲۱۲	۵-۱۰. پارامترهای تنظیم

فصل یازدهم: استقرار

۲۱۵	۱-۱۱. استقرار در زبان برنامه نویسی R
۲۱۸	۲-۱۱. تبدیل به PMML

فصل دوازدهم: تحلیل و کاوش شبکه های اجتماعی

۲۲۲	۱-۱۲. فرایند کلی کاوش شبکه های اجتماعی
۲۲۶	۲-۱۲. کاوش نظرات و بررسی الگوها در توییتز
۲۳۴	۳-۱۲. تحلیل نظرات

فصل سیزدهم: تحلیل و پیشگویی شاخص‌های قیمت منزل مسکونی

۲۶۰	۱-۱۳. واردکردن داده‌های HPI
۲۶۱	۲-۱۳. پوشش داده‌های HPI
۲۷۱	۳-۱۳. مولفه‌های فصلی و روندی مربوط به HPI
۲۷۳	۴-۱۳. پیش‌بینی (وضعیت آینده) HPI
۲۷۵	۵-۱۳. قیمت تخمین زده شده‌ی یک ملک

صا چهاردهم: پیش‌بینی پاسخ مشتری و بهینه‌سازی سود

۲۷۸	۱-۱۴. داده‌های KDD Cup 1998
۲۸۷	۲-۱۴. پوشش داده‌ها
۲۹۹	۳-۱۴. ارزیابی مدل
۳۰۳	۴-۱۴. انتخاب بهتر: رخت
۳۰۵	۵-۱۴. نمره‌دهی (امتیاز هی)

فصل پانزدهم: مدل‌سازی پیش‌بینی داده‌های بزرگ با حافظه محدود

۳۱۰	۱-۱۵. مقدمه
۳۱۰	۲-۱۵. شیوه انجام فرایند
۳۱۱	۳-۱۵. متغیرها و داده‌ها
۳۱۲	۴-۱۵. جنگل تصادفی
۳۱۴	۵-۱۵. مساله حافظه
۳۱۵	۶-۱۵. یادگیری مدل‌ها بر روی داده‌های نمونه
۳۱۷	۷-۱۵. ایجاد مدل با متغیرهای انتخابی
۳۲۴	۸-۱۵. نمره‌دهی
۳۳۱	۹-۱۵. چاپ قواعد به دست آمده
۳۴۳	پیوست ۱: آشنایی با RStudio
۳۵۳	پیوست ۲: اتصال R به محیط MS SQL Server
۳۶۷	منابع