

ابزارهای علم داده

برای متخصصان داده

مریم مقدمی

دانشجوی دکتری علم اطلاعات و دانش‌شناسی گرایش مدیریت دانش، دانشگاه تهران

دکتر عزیز هدایتی خوشمهر

هیات علمی دانشگاه شهید مدنی آذربایجان

زمستان ۱۴۰۰



سرناسه	: مقدمی وایقان، مریم، ۱۳۶۸-
عنوان و نام پدیدآور	: ابزارهای علم داده برای متخصصان داده / مریم مقدمی، عزیز هدایتی خوشمهر
مشخصات نشر	: تهران: چاپار: اساطیر پارسی، ۱۴۰۰.
مشخصات ظاهری	: ۱۶۶ ص.
شابک	: ۹۷۸-۶۰۰-۲۷۶-۲۰۶-۱
وضعیت فهرست‌نویسی	: فیپا.
یادداشت	: کتابنامه.
موضوع	: داده‌کاوی -- نرم‌افزار
موضوع	: Data mining -- Software
موضوع	: داده‌کاوی
موضوع	: Data mining
موضوع	: زبان‌های برنامه‌نویسی کامپیوتر
موضوع	: Programming languages (Electronic computers)
شناسه افزوده	: هدایتی خوشمهر، عزیز، ۱۳۵۰ -
رده‌بندی کنگره	: QAV6/9
رده‌بندی دیویی	: ۰۰۶/۳۱۲
شماره کتابشناسی ملی	: ۸۷۴۸۹۰۷



تهران: ۸۸۸۹۹۶۸۰ - ۰۹۱۲۱۹۶۰۲۱۴

ابزارهای علم داده برای متخصصان داده

مریم مقدمی و عزیز هدایتی خوشمهر

چاپ اول: ۱۴۰۰ شمارگان: ۵۰۰

صفحه‌آرایی: ندا هوشمند

شابک ۹۷۸-۶۰۰-۲۷۶-۲۰۶-۱ ISBN 978-600-276-206-1

حق هرگونه چاپ و انتشار محفوظ است.

قیمت: ۴۵۰۰۰ تومان

فهرست

پیشگفتار.....	۱۱
هدف از تالیف کتاب.....	۱۳
چه کسانی از این کتاب می‌توانند استفاده کنند؟.....	۱۴
فصل اول. مقدمه.....	۱۵
تفاوت یادگیری ماشین و علم داده.....	۱۷
علم داده چیست؟.....	۱۹
استخراج الگوهای معنی‌دار.....	۱۹
ترکیبی از آمار، یادگیری ماشین و محاسبات.....	۱۹
الگوریتم‌های یادگیری.....	۲۰
زمینه‌های مرتبط.....	۲۱
مزایای هوش تجاری.....	۲۳
طبقه‌بندی علم داده.....	۲۴
الگوریتم‌های علم داده.....	۲۶
چارچوب فرایند علم داده.....	۲۷
ابزارهای علم داده.....	۳۰
نقشه راه این کتاب.....	۳۰
فصل دوم. ابزارهای علم داده.....	۳۱
مسیر تبدیل شدن به دانشمند داده.....	۳۱
برای دیتا ساینست شدن (دانشمند داده) به چه مهارت‌هایی نیاز دارید؟.....	۳۳

۳۷	ابزارهای لازم برای تبدیل شدن به یک دانشمند داده.....
۵۶	اس کیو ال ^۴
۶۰	زبان برنامه‌نویسی آر.....
۶۱	طراحی این زبان.....
۶۶	زبان سی پلاس پلاس.....
۷۰	زبان برنامه‌نویسی جاوا.....
۷۲	جاوا یا پایتون: کدام یک را دانشمندان داده باید بیاموزند؟.....
۷۴	زبان C.....
۷۸	جاوا اسکریپت.....
۸۰	زبان برنامه‌نویسی متلب.....
۸۴	زبان برنامه‌نویسی Bash.....
۸۴	زبان برنامه‌نویسی جولیا.....
۸۸	زبان Swift.....
۹۳	معایب زبان برنامه‌نویسی سوئیفت.....
	فصل سوم. ابزارهای علم داده بر اساس سهولت استفاده، محبوبیت، شهرت
۹۵	
۹۵	مقدمه.....
۹۵	Algorithms. io.....
۹۶	Apache Giraph.....
۹۷	Apache Hadoop.....
۹۸	Apache HBase.....
۹۹	Apache Hive.....
۱۰۰	Apache Kafka.....
۱۰۱	Apache Mahout.....
۱۰۲	Apache Mesos.....
۱۰۳	Apache Pig.....
۱۰۴	Apache Spark.....
۱۰۵	Apache Storm.....

۱۰۶	BigML
۱۰۸	Bokeh
۱۰۹	Cascading
۱۱۰	Clojure
۱۱۱	D3.js
۱۱۲	DataRobot
۱۱۴	DataRPM
۱۱۵	Excel
۱۱۶	Feature Labs
۱۱۶	ForecastThis
۱۱۷	Fusion Tables
۱۱۸	Gawk
۱۱۹	GraphLab Create
۱۲۰	IPython
۱۲۱	KNIME Analytics Platform
۱۲۲	Matplotlib
۱۲۳	MLBase
۱۲۴	MySQL
۱۲۵	Narrative Science
۱۲۶	NetworkX
۱۲۷	NumPy
۱۲۸	Open Refine
۱۲۹	Pandas
۱۳۰	RapidMiner
۱۳۱	Redis
۱۳۲	Scala
۱۳۳	SciPy
۱۳۴	TIBCO

پیشگفتار

روزی نیست که خبری از علم داده، یادگیری ماشین یا هوش مصنوعی در رسانه‌ها نباشد. بررسی این اخبار دو نکته را برجسته می‌کند: علم داده و یادگیری ماشین، از علم داده بعنوان جذاب‌ترین فرصت شغلی قرن بیست و یکم یاد شده است. در دنیای امروز داده‌های خام از نظر ارزش اقتصادی با نفت خام مقایسه می‌شوند. امروزه اگر سازمان‌ها به‌ویژه آن‌هایی که کسب و کار دیجیتال دارند از قدرت داده‌ها حتی در یک بازه زمانی کوتاه غافل شوند به راحتی جایگاه خود را در بازار رقابت از دست خواهند داد. در حقیقت در دنیایی که بیش از پیش در حال تبدیل شدن به یک محیط دیجیتال است، سازمان‌ها هر روز با حجم انبوهی از داده‌های ساختاریافته و غیرساختاریافته در مقیاس ژتابایت و یوتابایت دست و پنجه نرم می‌کنند. پیشرفت پیوسته فناوری، امکان ذخیره‌سازی مقرون‌به‌صرفه‌تر و هوشمندانه‌تر داده‌های حیاتی را در اختیار ما قرار می‌دهد و این سبب شده است که جمع‌آوری و ذخیره‌سازی داده‌ها توجه بیشتری داشته باشد. علم داده به سازمان‌ها کمک می‌کند تا درکی صحیح و همه‌جانبه از مشتریان خود، بازار و کسب و کار بدست آورند. در دنیایی که با داده‌ها احاطه شده است، استخراج اطلاعات مفید و کارآمد از این داده‌ها، به معنی برتری در

بازار رقابت است و متخصصان علم داده، در استخراج این اطلاعات به سازمان‌ها و شرکت‌ها کمک می‌کنند. در حال حاضر صنعت به متخصصان داده کارآموده‌ای که مهارت‌شان مورد تأیید باشد نیاز فوق‌العاده‌ای دارد. این متخصصان از جمله پردرآمدترین شاغلان صنعت فناوری اطلاعات هستند. بنابر ادعای فوربس^۱ شاغلین حوزه علم داده در آمریکا با حقوق متوسط سالانه ۲۵۷ هزار دلار بهترین درآمد را در این کشور از آن خود کرده‌اند. علاوه بر این با توجه به نیازمندی‌های زیاد و روبه‌رشد، مؤسسه مک کینزی^۲ پیش‌بینی کرده است که در سال‌های آینده نیاز زیادی به متخصصان علم داده خواهیم داشت بطوری که میان میزان تقاضا برای به خدمت گرفتن متخصصان داده و توانایی پاسخگویی به این نیاز، شکافی پنجاه درصدی را تجربه خواهیم کرد. این انگیزه خوبی است تا بیشتر به این موضوع بپردازیم که علم داده چیست و ابزارهای آن کدامند؟ اکنون می‌توانیم مفهوم علم داده را بهتر درک کنیم. علم داده مهارت‌های زیادی را به هم گره می‌زند؛ از آمار و ریاضیات گرفته تا دانش حوزه تجارت. این علم به سازمان کمک می‌کند تا هزینه‌ها را کاهش دهد، به بازارهای جدید فکر کند، طیف مخاطبانش را گسترش دهد، بازاریابی و تبلیغات خود را سر و سامان دهد و محصولات و خدمات جدیدی عرضه کند و این تنها بخشی از فهرست بلندبالای مزایای بهره‌گیری از علم داده در دنیای امروز است. علم داده مجموعه‌ای در هم تنیده از مهارت‌ها و تخصص‌هاست. شاید به جرأت بتوان گفت که این علم یک هنر است. متخصص علم داده باید آنقدر در حرفه‌اش به مهارت رسیده باشد که بر اساس نیاز مشتری، با داده کار کند و آنچه سبب برتری مشتری در بازار رقابت می‌شود را از داده‌ها بیرون بکشد. شاید این حرفه در نگاه نخست، حرفه‌ای پول‌ساز به نظر برسد و از این جهت جذاب باشد اما باید توجه داشت که یک

متخصص علم داده باید احاطه کاملی نه تنها به حرفه خودش بلکه به حیطه کاری مشتریان خود نیز داشته باشد. هم اکنون افراد زیادی با استفاده از دوره‌های آموزشی رایگان پولی در حال فراگیری مهارت‌های علم داده جهت تبدیل شدن به یک دانشمند داده هستند. وظیفه اصلی دانشمند داده استخراج معانی مختلف و کاربردی از انواع داده است (ساختاریافته، نیمه‌ساختاریافته، ساختاریافته) داده‌ای که به‌طور بی‌وقفه وارد بانک‌های اطلاعاتی سازمان می‌شود. استخراج داده از بانک‌های اطلاعاتی، فراهم کردن داده برای تجزیه و تحلیل‌های گوناگون، طراحی و ساخت مدل‌های آماری و تألیف گزارش‌هایی به زبان ساده و قابل فهم از داده‌های مصورسازی شده (گزارش‌هایی که در اختیار مسئولان سازمان قرار می‌گیرد)، از مهم‌ترین وظایف دانشمند داده هستند. دیگر وظایف دانشمند داده به قرار زیر است:

- گردآوری داده
- آماده‌سازی داده
- تجزیه و تحلیل داده‌های اکتشافی (به اختصار EDA) مرحله اول در تحلیل هرگونه داده است و شامل بررسی اشتباهات و فعالیت‌های این‌چنینی می‌شود
- ارزیابی و تفسیر نتایج حاصل شده از تجزیه و تحلیل داده‌های اکتشافی
- طراحی و ساخت مدل‌های آماری؛
- امتحان کردن و گسترش این مدل‌ها؛
- بهینه‌سازی مدل‌ها.

هدف از تألیف کتاب

به دلیل نبود کتابی در حوزه مسیر تبدیل شدن به دانشمند داده و

ابزارهای علم داده در ایران سبب گردید تا کتاب مزبور در حوزه ابزارهای علم داده تدوین و در اختیار علاقه‌مندان علم داده قرار داده شود. در فصل نخست این کتاب علم داده و مفاهیم و شاخه‌های مرتبط با آن معرفی شده است. در فصل دو، سه، چهار و پنج نیز ابزارهای لازم برای متخصصان و دانشمندان علم داده معرفی شده‌اند. مولف این کتاب معتقد است برای تبدیل شدن به یک متخصص و دانشمند داده لازم است تا با ابزارهای ذکر شده در این کتاب آشنایی کاملی وجود داشته باشد. این کتاب به دنبال آن است تا دانش لازم در حوزه ابزارهای علم داده را در اختیار متخصصان داده قرار دهد.

چه کسانی از این کتاب می‌توانند استفاده کنند؟

متخصصان، دانشمندان، تحلیل‌گران، مهندسان، مدیران، روایت‌گران و کتابداران داده و نیز تمام علاقه‌مندان به حوزه علم داده که می‌خواهند وارد کسب و کارهای داده محور شوند می‌توانند از این کتاب به نحو مقتضی استفاده کنند.

با احترام

مریم مقدمی

خرداد ۱۴۰۰