

به نام خدا

مقدمه‌ای بر داده‌کاوی

مفاهیم، روش‌ها، الگوریتم‌ها و کاربردها

مؤلفین :

حسین عباسیان محمدی - علی نیک‌سرشت

پیشگفتار :

دکتر سعید جرفی

دکترای تخصصی مدیریت فناوری اطلاعات، مدرس دانشگاه و مدیر کل در سیستم بانکی

دکتر صدیقه غفرانی

دانشیار گروه مهندسی برق - مخابرات دانشگاه آزاد اسلامی واحد تهران جنوب

سرشناسه	:	عباسیان محمدی، حسین، ۱۳۷۳-
عنوان و نام پدیدآور	:	مقدمه‌ای بر داده‌کاوی مفاهیم، روش‌ها، الگوریتم‌ها و کاربردها/حسین عباسیان محمدی، علی نیک‌سرشت؛ با پیشگفتاری از سعید جرفی، صدیقه غفرانی.
مشخصات نشر	:	تهران: ادبستان، ۱۴۰۰.
مشخصات ظاهری	:	۳۵۶ ص.: مصور، جدول، نمودار.
شابک	:	978-600-196-147-2
وضعیت فهرست نویسی	:	فیپا
موضوع	:	داده‌کاوی -- Data mining
شناسه افزوده	:	نیک‌سرشت، علی، ۱۳۷۲- مترجم
شناسه افزوده	:	جرفی، سعید، غفرانی، صدیقه، مقدمه‌نویس
رده بندی کنگره	:	QAY۶/۹
رده بندی دیویی	:	۰۰۶/۳۱۲
شماره کتابشناسی ملی	:	۸۴۹۵۸۰۱
اطلاعات رکورد کتابشناسی	:	فیپا

نام کتاب	:	مقدمه‌ای بر داده‌کاوی (مفاهیم، روش‌ها، الگوریتم‌ها و کاربردها)
مؤلف	:	حسین عباسیان محمدی، علی نیک‌سرشت
ناشر	:	ادبستان
چاپ و صحافی	:	دیجیتال سام
نوبت چاپ	:	اول - ۱۴۰۰
تیراژ	:	۱۰۰ نسخه
قیمت	:	۱۰۰۰۰۰ تومان

دفتر مرکزی کتاب آیلار

انقلاب - خیابان شهید منیری جاوید (اردیبهشت) - خیابان شهدای ژاندارمری شرقی - شماره ۱۴۶
(ساختمان آیلار) تلفن: ۶۶۴۰۱۲۵۵ - دورنگار: ۶۶۴۹۴۴۳۱

فروشگاه شماره ۱ (کتاب آیلار)

انقلاب - روبروی دبیرخانه دانشگاه تهران - بازارچه کتاب تلفن: ۶۶۴۱۱۸۶۵ - ۶۶۹۵۶۴۳۳

فروشگاه شماره ۳ (کتاب آیلار)

خیابان انقلاب - روبروی دانشگاه تهران - پاساژ فروزنده - همکف - پلاک ۳۲۱ تلفن: ۶۶۹۶۳۵۹۶ - ۹۷

شابک: ۹۷۸-۶۰۰-۱۹۶-۱۴۷-۲ ISBN:978-600-196-145-8

شابک دوره: ۹۷۸-۶۰۰-۱۹۶-۱۴۶-۵ ISBN:978-600-196-146-5

فهرست مطالب

۷	مقدمه
۱۱	پیشگفتار اول
۱۳	پیشگفتار دوم
۱۹	فصل ۱: اصول داده‌کاوی: دیدگاهی اولیه
۱۹	۱.۱ علم‌داده، تحلیل و جستجوی داده و کشف دانش از پایگاه‌های داده
۲۰	۱.۱.۱ علم‌داده و تحلیل آن
۲۰	۱.۱.۲ داده‌کاوی
۲۱	۱.۱.۳ مقایسه علم‌داده با کشف دانش در پایگاه‌های داده
۲۲	۱.۲ کامپیوترها قادر به آموختن چه چیزهایی هستند؟
۲۲	۱.۲.۱ سه دیدگاه مفهومی
۲۶	۱.۲.۴ خوشه‌بندی نظارت‌نشده
۳۰	۱.۳ آیا داده‌کاوی برای مشکل من، مناسب است؟
۳۰	۱.۳.۱ داده‌کاوی یا داده‌جویی؟
۳۲	۱.۴ داده‌کاوی یا مهندسی دانش؟
۳۳	۱.۵ روش نزدیک‌ترین همسایه
۳۴	۱.۶ یک مدل فرایند برای داده‌کاوی
۳۵	۱.۶.۱ تهیه داده‌ها
۳۷	۱.۶.۲ پیش‌پردازش داده‌ها
۳۷	۱.۶.۳ کاوش داده‌ها
۳۸	۱.۶.۴ تفسیر نتایج
۳۸	۱.۶.۵ کاربرد نتیجه
۳۸	۱.۷ داده‌کاوی، کلان داده‌ها و رایانش ابری
۳۹	۱.۷.۱ هَدوپ
۳۹	۱.۷.۲ رایانش ابری
۴۰	۱.۸ اخلاقیات در داده‌کاوی
۴۱	۱.۹ ارزش ذاتی و روی گردانی (ترک) مشتری
۴۵	فصل ۲: نگاهی نزدیک‌تر به داده‌کاوی
۴۵	۲.۱ استراتژی‌های داده‌کاوی
۴۷	۲.۱.۱ طبقه‌بندی
۴۷	۲.۱.۲ تخمین
۴۸	۲.۱.۳ پیش‌بینی

۴۹	۲.۱.۴ خوشه‌بندی نظارت‌نشده
۵۰	۲.۱.۵ تجزیه و تحلیل سبد بازار
۵۰	۲.۲ تکنیک‌های داده‌کاوی نظارت‌شده
۵۰	۲.۲.۱ مثالی بر اساس پایگاه داده تفریع کارت اعتباری
۵۳	۲.۲.۲ تکنیک‌های مبتنی بر قاعده
۵۶	۲.۲.۳ شبکه‌های عصبی
۵۹	۲.۲.۴ رگرسیون آماری
۵۹	۲.۳ قوانین وابستگی
۶۱	۲.۴ تکنیک‌های خوشه‌بندی
۶۳	۲.۵ ارزیابی عملکرد
۶۳	۲.۵.۱ ارزیابی مدل‌های یادگیرنده نظارت‌شده
۶۶	۲.۵.۲ تحلیل خطای دودسته‌ای
۶۷	۲.۵.۳ ارزیابی خروجی عددی
۶۸	۲.۵.۵ ارزیابی مدل نظارت‌نشده

فصل ۳: روش‌های پایه در داده‌کاوی

۷۱	۳.۱ درخت‌های تصمیم‌گیری
۷۱	۳.۱.۱ الگوریتمی برای ایجاد درختان تصمیم‌گیری
۷۸	۳.۱.۲ درخت‌های تصمیم‌گیری
۸۲	۳.۱.۳ قوانین درخت تصمیم‌گیری
۸۳	۳.۱.۴ روش‌های دیگر برای ساخت درختان تصمیم‌گیری
۸۴	۳.۱.۵ ملاحظات عمومی
۸۴	۳.۲ الگوریتم کی میانگین
۸۵	۳.۲.۱ مثالی با استفاده از الگوریتم کی میانگین
۸۹	۳.۲.۲ ملاحظات عمومی
۹۱	۳.۳ یادگیری ژنتیک
۹۲	۳.۳.۱ الگوریتم‌های ژنتیک و یادگیری نظارت‌شده
۹۸	۳.۳.۲ ملاحظات عمومی
۹۸	۳.۴ انتخاب روش مناسب برای داده‌کاوی

فصل ۴: روش‌های یادگیری از داده در داده‌کاوی

۱۰۱	۴.۱ ماشین یادگیری
۱۰۸	۴.۲ نظریه یادگیری آماری
۱۱۶	۴.۳ وظایف یادگیری متداول
۱۲۳	۴.۴ ماشین‌های بردار پشتیبان
۱۴۰	۴.۵ ماشین‌های بردار پشتیبان نیمه نظارت‌شده (S ₂ VM)
۱۴۶	۴.۶ انتخاب مدل
۱۵۱	۴.۷ تخمین مدل
۱۶۳	۴.۸ طبقه‌بندی داده‌های نامتعادل

۱۶۸	۴.۹ صحت ۹۰٪ ... حالا چه؟
۱۶۸	۴.۹.۱ مورد مطالعاتی ۱: تشخیص کلاهبرداری در بیمه
۱۷۱	۴.۹.۲ مورد مطالعاتی ۲: بهبود مراقبت از قلب

فصل ۵: شبکه‌های عصبی مصنوعی

۱۷۷	۵.۱ مدل یک نورون مصنوعی
۱۷۹	۵.۲ معماری‌های شبکه‌های عصبی مصنوعی
۱۸۳	۵.۳ فرایند یادگیری
۱۸۸	۵.۴ وظایف یادگیری با استفاده از این‌ها
۱۸۸	۵.۴.۱ نسبت‌دهی الگو
۱۸۸	۵.۴.۲ شناسایی الگو
۱۸۹	۵.۴.۳ تخمین تابع
۱۸۹	۵.۴.۴ کنترل
۱۹۱	۵.۴.۵ پیش‌بینی
۱۹۲	۵.۵ پرسپترون‌های چندلایه
۲۰۲	۵.۶ شبکه‌های رقابتی و یادگیری رقابتی
۲۰۷	۵.۷ نگاشت‌های خود سازمان‌ده
۲۱۴	۵.۸ یادگیری عمیق
۲۲۲	۵.۹ شبکه‌های عصبی پیچشی (کانولوشنی) (CNNها)

فصل ۶: تجزیه و تحلیل خوشه‌بندی

۲۲۹	۶.۱ مفاهیم خوشه‌بندی
۲۳۳	۶.۲ معیارهای شباهت
۲۴۲	۶.۳ خوشه‌بندی سلسله‌مراتبی تجمیعی
۲۴۷	۶.۴ خوشه‌بندی افرازشونده
۲۵۱	۶.۵ خوشه‌بندی صعودی
۲۵۴	۶.۶ الگوریتم DBSCAN
۲۵۹	۶.۷ اعتبارسنجی خوشه‌بندی

فصل ۷: الگوریتم‌های ژنتیک

۲۶۸	۷.۱ اصول الگوریتم‌های ژنتیک
۲۷۱	۷.۲ بهینه‌سازی با استفاده از الگوریتم‌های ژنتیک
۲۷۲	۷.۲.۱ طرح کدگذاری و مقداردهی اولیه
۲۷۳	۷.۲.۲ ارزیابی برازش
۲۷۴	۷.۲.۳ انتخاب
۲۷۵	۷.۲.۴ تقاطع
۲۷۶	۷.۲.۵ جهش
۲۷۹	۷.۳ یک تصویر ساده از یک الگوریتم ژنتیک
۲۷۹	۷.۳.۱ نمایش
۲۸۰	۷.۳.۲ جمعیت اولیه

ساقی به نور باده برافروز جام ما

مطرب بگو که کار جهان شد به کام ما

ما در پیاله عکس رخ یار دیده‌ایم

ای بی‌خبر ز لذت شرب مدام ما

هرگز نمیرد آن که دلش زنده شد به عشق

ثبت است بر جریده عالم دوام ما

در دورانی که شاهد انقلاب صنعتی چهارم هستیم، مشتریان دیگر مانند قدیم ناآگاه نبوده بلکه، در بسیاری از موارد مشتری حتی نکات و اطلاعاتی را در ذهن دارد که کسب و کارها به دلیل مشغله‌های روزافزون خود از آن غفلت دارند. انسان موجودی است عقلی، اجتماعی، ارتباطی، آزمند، آرزومند، نیازمند، تحول‌گرا، نوگرا و کمال طلب که این ذات پیچیده‌ی انسان در کسب و کاری از موارد محیطی دیگر شرایط را برای کسب و کارها مبهم، نامطمئن و چالش‌برانگیز می‌کند. لذا توجه به مشتری (که بزرگ‌ترین مسئله‌ی پیشروی کسب و کارها در عصر حاضر است)، تمرکز و تمایزی که باید برای او قائل شویم و شناخت و معرفی که باید از او در تمامی مراحل درگیر شدنش با کسب و کارمان به دست آوریم، در قرن و انقلاب صنعتی حاضر بیش‌ازپیش حائز اهمیت بوده و موفقیت کسب و کارها وابسته به آن است. مشتری درواقع قلب تپنده‌ی بازار بوده که مایه‌ی حیات و دوام کسب و کارها می‌باشد. از طرفی در دنیایی که مباحثی از قبیل هوش مصنوعی، فضای ابری، کلان داده‌ها، چابکی، دیجیتال مارکتینگ، علوم داده، اینترنت اشیا، واقعیت مجازی و غیره مطرح است، دیگر مدیران قادر به پاسخگویی به نیازهای بازار از راه‌های سنتی نیستند و باید از طرق پیشرفته راه‌حل‌جویی کرده و بتوانند پاسخ‌خواسته‌ها و نیازهای مشتریان را با کمترین ریسک و خطا، در کمترین مدت‌زمان و با بیشترین بهره‌وری ارائه کنند. جدای از این، روش‌های قدیمی حتی در زمان خود نیز قادر به پاسخگویی مطلوبی نبودند و وقت‌گیر، هزینه‌بر و نادقیق بوده و عموماً امور بر اساس سعی و خطا حل و فصل می‌شدند. در زمره‌ی مهم‌ترین کارهایی که یک مدیر مسئولیت برنامه‌ریزی و پاسخگویی آن را دارد است، انتخاب استراتژی بازاریابی پویا، بادوام و سازگار با عوامل تکنولوژیکی و محیطی بخصوص در شرایط فن‌آورانه است؛ بدین ترتیب یک مدیر زمانی که باید یک محصول، بازار، قیمت یا روش ترفیع جدیدی را

برنامه‌ریزی، انتخاب و پیاده‌سازی کند و یا با مسائلی از قبیل بخش‌بندی، هدف‌گذاری و جایگاه‌یابی طرف است، با سؤالاتی اساسی روبه‌رو خواهد شد که بایستی با تحقیقات بازار، تصمیم‌گیری، خلاقیت و به‌طور کلی استراتژی بازاریابی درست به این ارکان اساسی کسب‌وکار پاسخگو باشد و این مهم به‌جز با اطلاعات، ادراکات، استدلال‌ات و استنتاجات صحیحی که در خدمت مدیر باشد، به‌دست نخواهد آمد. از آنجاکه مسبب شادی و غم یا پیروزی و شکست هر فرد یا هر کسب‌وکاری خارج از انتخاب‌ها، اتفاق‌ها و ارتباط‌ها نیست؛ لذا یک مدیر باید بتواند با انتخاب و تصمیم‌گیری بجا، به‌موقع، به‌اندازه و برازنده و سپس به‌کارگیری و استفاده‌ی درست از منابع و کنترل شرایط (کنش‌گرایی)؛ به موفقیت، پایداری، برتری، رضایت و حمایت اجتماعی هرچه بیشتر دست پیدا کند. از سوی دیگر، در دنیای پررقابت امروز که هر لحظه ستاره‌ای درخشیده و قصد عیان شدن دارد، بایستی به ترتیب با دیده شدن، شنیده شدن و پدیده شدن خود را به وادی برگزیده شدن رسانده تا بتوانیم درنهایت علاوه بر پیروزی در این میدان نبرد، به یک ارزش‌آفرینی و ارزش‌افزایی دوسویه، پویا و بادوام برای کسب‌وکار خود و مشتریان برسیم.

حال چیزی که در دنیای امروز این امر را راحت‌تر کرده و به نیازهای مبرم ذکرشده پاسخ داده و اطلاعات را با هزینه و زمان کمتر و با دقت بیشتر در اختیار مدیران قرار داده است، مبحث هوشمندسازی است. که آقای جوی پاول گیلفرد^۱ آن را یکی از مهم‌ترین متشکل‌های اثرپذیری، اثرگذاری و یادگیری می‌نامد. یکی از روش‌هایی که هوشمندی را برای ما به ارمغان آورده است، مباحث هوش مصنوعی و داده‌کاوی است. داده‌کاوی یک فناوری نسبتاً جدید و جالب است که امروزه در اکثر سازمان‌ها مرسوم شده است. دکتر آرنو پنزیاس^۲ (برنده‌ی جایزه‌ی نوبل و دانشمند ارشد پیشین در آزمایشگاه‌های بل) در ژانویه‌ی سال ۱۹۹۹ میلادی در مصاحبه‌ای با مجله‌ی «کامپیوتر و ورلد»^۳ اظهار داشت که «در آینده‌ی نزدیک، کاوش داده از داخل پایگاه داده‌های سازمان‌ها به بخش مهمی از عملکرد شرکت‌ها تبدیل خواهد شد». دکتر پنزیاس در پاسخ به این سؤال که "در آینده چه کاربردی از علم‌داده در میان شرکت‌ها محبوب می‌شود؟"، پاسخ داد: "داده‌کاوی" و سپس افزود: "در آینده اهمیت داده‌کاوی افزایش می‌یابد و شرکت‌ها تمام اطلاعات مربوط به مشتریان‌شان را حفظ خواهند کرد زیرا در آن زمان این اطلاعات به چیز ارزشمندی تبدیل می‌شوند و اگر کسی این کار را انجام ندهد، از تجارت و رقابت حذف خواهد شد." توماس داونپورت^۴ نیز در مقاله‌ی خود در مجله‌ی کسب‌وکار هاروارد (در سال ۲۰۰۶) اظهار داشت که تصمیم‌گیری تحلیلی، آخرین سلاح

^۱ J. P. Guilford

^۲ Arno Penzias

^۳ Computer World

^۴ Thomas Davenport

استراتژیک شرکت‌ها خواهد بود. او در این مقاله، شرکت‌هایی مثل آمازون، کیتال وان، ماریوت اینترنشنال و غیره را مثال می‌زند که چگونه برای درک بهتر مشتریان و بهینه‌سازی زنجیره‌ی تأمین گسترده‌ی خود از علم تحلیل و واکاوی استفاده می‌کنند تا بتوانند بازده سرمایه‌گذاری خود را به حداکثر برسانند و همچنین بهترین خدمات را به مشتریان خود ارائه دهند. دستیابی یک شرکت به این سطح از موفقیت، به‌شدت به شناخت کامل آن شرکت از مشتریان، فروشندگان، فرایند کسب و کار و زنجیره‌ی تأمین گسترده‌اش بستگی دارد.

با تحلیل داده‌های جمع‌آوری‌شده توسط شرکت‌ها می‌توان به بخش عمده‌ای از «درک مشتری» رسید. در چند سال اخیر، هزینه‌ی ذخیره و پردازش داده‌ها به‌شدت کاهش‌یافته، در نتیجه میزان داده‌هایی که به‌صورت الکترونیکی ذخیره می‌شوند با سرعت زیادی افزایش‌یافته است. ایجاد پایگاه داده‌های بزرگ، این امکان را به وجود آورده که بتوان داده‌های ذخیره‌شده در درون آن‌ها را مورد تجزیه و تحلیل قرار داد. اصطلاح داده‌کاوی در ابتدا برای توصیف فرایندی استفاده می‌شد که از طریق آن، الگوهای ناشناخته‌ای که در داده‌ها وجود داشت را کشف می‌کردند. از آن زمان تاکنون برخی از شرکت‌های نرم‌افزاری، مرزهای این اصطلاح را گسترش داده‌اند و اگرچه تحلیل داده را در این اصطلاح جای داده‌اند تا بتوانند از محبوبیت برجسته داده‌کاوی استفاده کرده و فروش خود را افزایش دهند. داده‌کاوی دارای ابعاد بسیار گوناگونی از جمله مهندسی، مدیریت، پزشکی، سیاست و علوم اجتماعی و غیره می‌باشد که بایستی کوشید و از جبهه امروزی و فردای دنیای خودمان یعنی داده‌کاوی و در نگاهی کلان‌تر هوش مصنوعی جا نمانده و بتوانیم در این میدان نبرد رقابت پیروز بیرون آییم. جا دارد این نکته را نیز یادآور شویم که آنچه به تمام نتوان دریافت، همه‌اش را نتوان گذاشت.

با توجه به مطالب ذکرشده و به‌طور کلی حوزه‌ی مهم هوش مصنوعی و داده‌کاوی که کمک شایانی در شناخت و درک هرچه بیشتر مهم‌ترین دارایی هر کسب و کار یعنی مشتریان کرده و موجب می‌شود تا بافهم دقیق‌تر، سریع‌تر و درست‌تر از محیط، رقبا، نقاط تماس با مشتریان (نقشه سفر مشتری) و به‌طور کلی تمامی ارکان حیاتی کسب و کارمان، به تصمیم‌گیری بهتر و صحیح‌تر و در نهایت موفقیت بیشتر و بادوام‌تر نائل آییم، و همچنین کمک این حوزه به علوم بی‌شمار دیگر، بر آن شدیم تا کتاب حاضر که برگرفته از فصول برتر و پرمایه‌تر ۳ کتاب مرجع، شاخص، پر فروش و جدید (هر کدام دریکی از ارکان داده‌کاوی) است را به‌علاوه دانش مختصرمان^۲ در این حوزه به‌صورت تألیف، ترجمه و تدوین درآوریم تا در اختیار دوستداران علم و مخصوصاً حوزه علوم

۱- در انتهای کتاب لیست این کتب آورده شده است.

۲- چیزی که میدانیم، قطره‌ای است از اقیانوسی که چیزی از آن نمی‌دانیم.

کامپیوتر (هوش مصنوعی، داده‌کاوی، علم‌داده و تجزیه و تحلیل) قرار دهیم. سعی و هدف ما در انجام این کتاب این بود تا با نگاهی دقیق‌تر مبحث داده‌کاوی در جهان را برای مخاطبان و مطالعه‌کنندگان گرامی فراهم آوریم تا دیدی هرچه بیشتر در مورد تحولات این حوزه از تکنولوژی پیشرو پیدا نمایند و از طرفی جبرانی باشد بر کتب حاضر این حوزه در کشور خوبمان که اغلب یا جامع نبوده و مباحث را از پایه بیان نمی‌کنند و یا نسخه قدیمی از کتب مرجعی می‌باشند که دیگر جایگاهی کاربردی نداشته و در این عصر که مخصوصاً در حوزه علوم کامپیوتر همه‌چیز به‌مثابه سرعت نور در حال تغییر و تحول و دگرگونی است، مشتمل بر اطلاعات و دانش بروز نیستند.

امیدواریم کتاب حاضر بتواند به سهم خود در ارتقای دانش حوزه‌ی ذکرشده در میان مهندسان، مدیران، پزشکان، تحلیلگران، دانشجویان، فعالان همه‌ی عرصه‌های مرتبط (احتمالاً همه علوم چرا که همگی کم‌وبیش با داده سروکار دارند) و به‌طور کلی علاقه‌مندان مفید و مؤثر واقع شود.

در خاتمه گفتنی است که برای اینکه پیشگفتار این اثر رنگ و بویی علمی‌تر گرفته و ارزش کتاب صدچندان شود از دو تن از اعضای هیئت علمی مطرح و صاحب‌نظر گرامی از دانشگاه‌های به نام و یا دارای مناصب بالا، درخواست نوشتن پیشگفتار را نمودیم که ایشان نیز لطف نموده و ما را یاری نمودند. از این‌رو پیشگفتار این عزیزان به ترتیب حروف الفبا در دنباله‌ی این مقدمه آورده شده است.

ضمناً از آنجا که مرجع اصلی ارزیابی هر اثر، مخاطبان می‌باشند، از خوانندگان عزیز تقاضا داریم نظرات و سؤالات خود را در مورد کتاب حاضر از طریق نشانی‌های ایمیلی که در ذیل آورده شده‌اند، با ما در میان بگذارند. باشد که همگی ذره‌ذره‌ی ظلمات وجود را به روشنایی تبدیل کنیم.

نشانی‌های ایمیلی:

۱- حسین عباسیان محمدی

محقق، مشاور و مدرس حوزه مهندسی برق (مخابرات) و هوش مصنوعی

h.abbasian.m@aut.ac.ir

۲- علی نیک‌سرشت

محقق، مشاور و مدرس حوزه‌ی مهندسی برق، کامپیوتر (هوش مصنوعی و علم‌داده) و مدیریت

AL.Nikseresht@mail.sbu.ac.ir