به نام خدا

داده‌کاوی

مفاهیم، روش‌ها، الگوریتم‌ها و کاربردها

با رویکرد آماری

(به همراه مثال‌های عملی در انتهای کتاب)

www.ketab.ir

مؤلفین :

حسین عباسیان محمدی- علی نیک‌سرشت

پیشگفتار :

پرفسور حمیدرضا امین‌داور

عضو هیئت‌علمی دانشکده مهندسی برق دانشگاه امیرکبیر (پلی‌تکنیک تهران)

دکتر یاسر نوروزی

عضو هیئت‌علمی دانشکده مهندسی برق دانشگاه امیرکبیر (پلی‌تکنیک تهران)

سرشناسه	:	عباسیان محمدی، حسین، ۱۳۷۳-
عنوان و نام پدیدآور	:	داده‌کاوی: مفاهیم، روش‌ها، الگوریتم‌ها و کاربردها با رویکرد آماری (بهمراه مثال‌های عملی در انتهای کتاب)/حسین عباسیان‌محمدی، علی نیک‌سرشت ؛ با پیشگفتاری از حمیدرضا امین‌داور، یاسر نوروزی.
مشخصات نشر	:	تهران: ادبستان، ۱۴۰۰.
مشخصات ظاهری	:	۲۵۶ ص.: مصور، جدول، نمودار.
شابک	:	978-600-196-145-8
وضعیت فهرست نویسی	:	فیپا
یادداشت	:	واژه‌نامه.
موضوع	:	داده‌کاوی -- روش‌های آماری Statistical methods -- Data mining
شناسه افزوده	:	نیک‌سرشت، علی، ۱۳۷۲- Nik Seresht, Ali
شناسه افزوده	:	امین‌داور، حمیدرضا، نوروزی، یاسر، مقدمه‌نویس
رده بندی کنگره	:	QA۷۶/۹
رده بندی دیویی	:	۰۰۶/۳۱۲
شماره کتابشناسی ملی	:	۸۴۹۵۷۹۴
اطلاعات رکورد کتابشناسی	:	فیپا

نام کتاب	:	**داده کاوی (مفاهیم، روش‌ها، الگوریتم‌ها و کاربردها)**
مؤلف	:	حسین عباسیان محمدی و علی نیک‌سرشت
ناشر	:	ادبستان
چاپ و صحافی	:	دیجیتال سام
نوبت چاپ	:	اول- ۱۴۰۰
تیراژ	:	۱۰۰ نسخه
قیمت	:	**۷۵۰۰۰ 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 : ۹۷-۶۶۹۶۳۵۹۶

شابک : ۸-۱۴۵-۱۹۶-۶۰۰-۹۷۸ ISBN:978-600-196-145-8

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

این اثر مشمول قانون حمایت از حقوق مؤلفان، مصنفان و هنرمندان مصوب ۱۳۴۸ است، هر کس تمام یا قسمتی از این اثر را بدون اجازه ناشر، نشر یا پخش یا عرضه کند تحت پیگرد قانونی قرار خواهد گرفت.

www.ketab.ir

فهرست مطالب

مقدمه ۶

پیشگفتار اول ۹

پیشگفتار دوم ۱۹

فصل اول: اصول داده‌کاوی: دیدگاهی اولیه ۲۱

علم داده، تحلیل و جستجوی داده و کشف دانش از پایگاه‌های داده ۲۱

کامپیوترها قادر به آموختن چه چیزهایی هستند؟ ۲۳

آیا داده‌کاوی برای مشکل من مناسب است؟ ۳۱

داده‌کاوی یا مهندسی دانش؟ ۳۳

روش نزدیک‌ترین همسایه ۳۴

یک مدل فرایند برای داده‌کاوی ۳۵

داده‌کاوی، کلان‌داده‌ها و رایانش ابری ۳۸

اخلاقیات در داده‌کاوی ۴۰

ارزش ذاتی و روی‌گردانی(ترک) مشتری ۴۱

ارزیابی عملکرد ۴۳

روش‌های یادگیری از داده در داده‌کاوی ۴۹

فصل دوم: روش‌های آماری ۷۰

علم آمار ۷۰

انواع کاربردهای آمار ۷۰

عناصر بنیادی آمار ۷۱

آمار توصیفی و تحلیلی ۷۳

آمار توصیفی: معیارهای مکان و پراکندگی ۷۵

www.ketab.ir

معیارهای شکل توزیع و موقعیت نسبی ۸۵
معیارهای ارتباط بین دو متغیر ۸۸
آشنایی مقدماتی با احتمال(آزمایش‌های تصادفی، قوانین شمارش و نسبت‌دهی احتمالات) ۹۳
فصل سوم: توزیع‌های احتمال و تحلیل رگرسیون ۱۰۴
متغیرهای تصادفی ۱۰۴
تحلیل رگرسیون ۱۱۲
فصل چهارم: داده‌کاوی پیشرفته ۱۴۲
داده‌کاوی زمانی ۱۴۲
داده‌کاوی فضایی ۱۶۰
همبستگی به معنای علیت نیست! ۱۶۶
یادگیری تقویتی ۱۷۴
فصل پنجم: مدل‌های پیش‌بینی‌کننده سری‌های زمانی ۱۸۱
کنترل فرایند آماری ۱۸۱
نمودارهای کنترل ۱۸۲
پیش‌بینی سری‌های زمانی ۱۸۷
صحت پیش‌بینی ۱۹۵
میانگین‌های متحرک و هموارسازی نمایی ۲۰۰
هموارسازی نمایی ۲۰۴
مدل‌های خودهمبسته ۲۰۸
علیت سری‌های زمانی ۲۱۱

www.ketab.ir

فصل ششم: مثال‌های کاربردی عملی با نرم‌افزار ۲۱۳
نرم‌افزارهای مورد نیاز و بحث ۲۱۳
کاربرد هر یک از نرم‌افزارها و دلیل نصب هر کدام ۲۱۳
روند پیاده‌سازی(شفاف‌سازی و آموزش با حل سه مطالعه موردی) ۲۱۵
برخی از لغات و اصطلاحات به‌کاررفته در کتاب ۲۵۰
فهرست منابع ۲۵۶

www.ketab.ir

مقدمه

یا رب مددی کز خودی خود برهم — از بد برهم وز بدی خود برهم
در هستی خود، مرا ز خود بیخود کن — «تا از خودی و بیخودی خود برهم

در دورانی که شاهد انقلاب صنعتی چهارم هستیم، مشتریان دیگر مانند قدیم ناآگاه نبوده، بلکه در بسیاری از موارد، حتی نکات و اطلاعاتی را در ذهن دارند که کسب‌وکارها به‌دلیل مشغله‌های روزافزون خود از آن غافل‌اند. انسان موجودی است اقتصادی، اجتماعی، ارتباطی، آزمند، آرزومند، نیازمند، تحول‌گرا، نوگرا و کمال‌طلب که این ذات پیچیده‌ی او در کنار بسیاری از موارد محیطی دیگر، شرایط را برای کسب‌وکارها مبهم، نامطمئن و چالش‌برانگیز می‌کند. لذا در عصر حاضر و با توجه به انقلاب صنعتی جدید، توجه به مشتری (که بزرگترین مسئله پیش‌روی کسب‌وکارهاست)، تمرکز و تمایزی که باید برای او قائل شویم و شناخت و معرفتی که باید از او در تمامی مراحل درگیر شدنش با کسب‌وکارمان به‌دست آوریم، بیش‌ازپیش اهمیت می‌یابد، تاجایی که موفقیت کسب‌وکارها وابسته به آن است. در واقع، مشتری قلب تپنده‌ی بازار و مایه‌ی حیات و دوام کسب‌وکارها است. از طرفی، در دنیایی که مباحثی از قبیل هوش مصنوعی، فضای ابری، کلان‌داده‌ها، دیجیتال‌مارکتینگ، چابکی، علوم داده، اینترنت اشیا، واقعیت مجازی و غیره مطرح است، مدیران دیگر قادر به پاسخگویی به نیازهای بازار از راه‌های سنتی نیستند و باید از طرق پیشرفته، راه‌حل‌هایی بیابند تا بتوانند پاسخ خواسته‌ها و نیازهای مشتریانشان را با کمترین ریسک و خطا، در کمترین زمان و با بیشترین بهره‌وری ارائه دهند. جدای از این، روش‌های قدیمی حتی در زمان خود نیز قادر به پاسخگویی مطلوبی نبودند و وقت‌گیر، هزینه‌بر و نادقیق بوده و عموماً امور براساس سعی و خطا حل‌وفصل می‌شدند. در زمره‌ی مهم‌ترین کارهایی که یک مدیر مسئولیت برنامه‌ریزی و پاسخگویی آن را دارد، انتخاب استراتژی بازاریابی پویا، بادوام و سازگار با عوامل تکنولوژیکی و محیطی، به‌خصوص در شرایط فناورانه 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 به‌طورکلی استراتژی بازاریابی درست به این ارکان اساسی کسب‌وکار پاسخ دهد و این مهم به‌جز با اطلاعات، ادراکات، استدلال‌ها و استنتاج‌های صحیحی که در خدمت مدیر قرار می‌گیرد، به‌دست نخواهد آمد. از آنجا که 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 به‌جا، به‌موقع، به‌اندازه و برازنده و سپس به‌کارگیری و استفاده درست از منابع و کنترل شرایط (کنش‌گرایی)، به موفقیت، پایداری، برتری، رضایت و حمایت اجتماعی هرچه بیشتر دست یابد. از سوی دیگر، در دنیای پررقابت امروز که هرلحظه ستاره‌ای درخشیده و قصد عیان شدن دارد، بایستی به ترتیب با دیده شدن، شنیده شدن و پدیده شدن خود را به وادی برگزیده شدن برسانیم تا درنهایت بتوانیم علاوه‌بر پیروزی در این میدان نبرد، به یک ارزش‌آفرینی و ارزش‌افزایی دوسویه، پویا و بادوام برای کسب‌وکار خود و مشتریان برسیم.

www.ketab.ir

حال چیزی که در دنیای امروز این امر را راحت‌تر کرده و به نیازهای مبرم ذکر شده پاسخ داده و اطلاعات را با هزینه و زمان کمتر و با دقت و صحت بیشتر در اختیار مدیران قرار داده است، مبحث هوشمندسازی است. که آقای جوی پاول گیلفرد[1] آن را چرخه‌ای متشکل از 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[2] (برنده‌ی جایزه‌ی نوبل و دانشمند ارشد پیشین در آزمایشگاه‌های بل) در ژانویه‌ی سال ۱۹۹۹ میلادی در مصاحبه‌ای با مجله‌ی «کامپیوتر وُرلد[3]» اظهار داشت: «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[4] نیز در مقاله‌ی خود در مجله‌ی کسب‌وکار هاروارد (در سال ۲۰۰۶) اظهار داشت که تصمیم‌گیری تحلیلی، آخرین سلاح استراتژیک شرکت‌ها خواهد بود. او در این مقاله، شرکت‌هایی مثل آمازون، کپیتال وان، ماریوت اینترنشنال و غیره را مثال می‌زند که چگونه برای درک بهتر مشتریان و بهینه‌سازی زنجیره‌ی تأمین گسترده‌ی خود از علم تحلیل و داده‌کاوی 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 کسب‌وکار 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 و درنتیجه، میزان داده‌هایی که به‌صورت الکترونیکی ذخیره می‌شوند با سرعت زیادی افزایش یافته است. ایجاد پایگاه داده‌های بزرگ، این امکان را به وجود آورده که بتوان داده‌های ذخیره‌شده در درون آن‌ها را مورد تجزیه‌وتحلیل 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 اکثر مدل‌های تحلیل داده را در این اصطلاح جای داده‌اند تا بتوانند از محبوبیت برچسب داده‌کاوی استفاده کرده و فروش خود را افزایش دهند. داده‌کاوی دارای ابعاد بسیار گوناگونی ازجمله مهندسی، مدیریت، پزشکی، سیاست، علوم اجتماعی و غیره است. در این میان، باید بکوشیم تا از جبهه امروز و فردای دنیای خودمان یعنی داده‌کاوی و در نگاهی کلان‌تر، هوش مصنوعی جا نمانیم و از این میدان رقابت، پیروز و سربلند بیرون آییم.

باتوجه به مطالب ذکر شده و همچنین، حوزه مهم هوش مصنوعی و داده‌کاوی که کمک شایانی در شناخت و درک هرچه بیشتر مهم‌ترین دارایی هر کسب‌وکار، یعنی مشتریان کرده و موجب می‌شود تا با فهم دقیق‌تر، سریع‌تر و درست‌تر از محیط،

1 J. P. Guilford
2 Arno Penzias
3 Computer World
4 Thomas Davenport

www.ketab.ir

رقبا، نقاط تماس با مشتریان(نقشه سفر مشتری) و به‌طور کلی تمامی ارکان حیاتی کسب‌وکارمان، با تصمیم‌گیری بهتر و صحیح‌تر به 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 پنج کتاب مرجع، شاخص، پرفروش و جدید[1](هرکدام در یکی از ارکان داده‌کاوی) به‌علاوه کتاب "مقدمه‌ای بر داده‌کاوی" از ناشر و مؤلفان همین کتاب است را به‌همراه دانش مختصرمان[2] در این حوزه به‌صورت تألیف، ترجمه و تدوین درآوریم و در اختیار دوست داران علم و مخصوصاً حوزه علوم کامپیوتر(هوش مصنوعی، داده‌کاوی، علم‌داده و تجزیه‌وتحلیل) قرار دهیم. سعی و هدف ما در انجام این مهم، این بوده تا با نگاهی دقیق‌تر مبحث داده‌کاوی در جهان را برای مخاطبان گرامی فراهم آوریم تا دیدی بیشتر در مورد تحولات این حوزه از تکنولوژی پیشرو پیدا نمایند و از طرفی جبرانی باشد بر کتب حاضر این حوزه در کشور خودمان که اغلب یا جامع نبوده و مباحث را از پایه بیان نمی‌کنند و یا نسخه قدیمی از کتب مرجعی هستند که دیگر کاربردی ندارند، به‌خصوص در این عصر که در حوزه علوم کامپیوتر، همه‌چیز به‌مثابه سرعت نور درحال تغییر و تحول و دگرگونی است.

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

در اینجا لازم است از اساتید ارجمند و صاحبنظر، استاد حمیدرضا امین‌داور و استاد یاسر نوروزی که برای غنای بیشتر این اثر با نگارش دو پیشگفتار ما را یاری نمودند، تشکر کنیم. درضمن، از آنجا که مرجع اصلی ارزیابی هر اثر، مخاطبان هستند، 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 درمیان بگذارند. باشد که همگی ذره‌ذره‌ی ظلمات وجود را به روشنایی تبدیل کنیم.

نشانی‌های ایمیلی:

حسین عباسیان‌محمدی، محقق، مشاور و مدرس حوزه مهندسی برق و کامپیوتر (هوش مصنوعی و علم داده):
h.abbasian.m@aut.ac.ir
علی نیک‌سرشت، محقق، مشاور و مدرس حوزه‌ی مهندسی برق، کامپیوتر (هوش مصنوعی و علم داده) و مدیریت:
AL.Nikseresht@mail.sbu.ac.ir

www.ketab.ir

---

۱ در انتهای کتاب لیست این کتب آورده شده است.

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