

مجموعه کتاب‌های مطالعات داده‌ها، اطلاعات و دانش - ۱

دانش داده‌ها، هوش مصنوعی و کسب و کارهای مبتنی بر داده

تالیف و گردآوری:

هادی فرزاد

نشر پنداره - ۱۴۰۳

سرشناسه	:	فرزاد، هادی، ۱۳۶۱ -
عنوان و نام پدیدآور	:	دانش داده‌ها، هوش مصنوعی و کسب و کارهای مبتنی بر داده/تالیف و گردآوری هادی فرزاد؛ ویراستار الهام فرزاد
مشخصات نشر	:	تهران: پنداره، ۱۴۰۳.
مشخصات ظاهری	:	۱۶۶ص.
فروست	:	مجموعه کتاب‌های مطالعات داده‌ها، اطلاعات و دانش؛ ۱.
وضعیت فهرست نویسی	:	فیبا
موضوع	:	هوش مصنوعی
		Artificial intelligence
		داده‌پردازی
		Electronic data processing
رده بندی کنگره	:	Q۳۳۵
رده بندی دیویی	:	۳/۰۰۶
شماره کتابشناسی ملی	:	۹۹۵۵۴۸۴
اطلاعات رکورد کتابشناسی	:	فیبا

عنوان کتاب : دانش داده‌ها، هوش مصنوعی و کسب و کارهای مبتنی بر داده

ناشر : نشر پنداره ۶۶۵۲۱۰۱۷ - ۰۲۱

تالیف و گردآوری : هادی فرزاد

ویراستار : الهام فرزاد

نوبت چاپ : چاپ نخست، ۱۴۰۳

شمارگان : ۵۰۰ نسخه

تعداد صفحات : ۱۶۶ صفحه

شابک : ۹۷۸-۶۲۲-۹۳۸۰۶-۰-۴

قیمت : ۳۲۰.۰۰۰ تومان

نشانی اینترنتی : www.pebook.ir

مجموعه‌ی «مطالعات داده‌ها، اطلاعات و دانش» نشر پنداره، بخشی تخصصی ترجمه، تالیف و نشر کتاب است که با هدف ارائه مطالب به روز در زمینه‌ی پیشرفت‌ها و فن‌آوری‌های نوین در زمینه‌های داده‌کاوی و داده‌پردازی فعالیت می‌کند تا تئوری، تحقیق، توسعه و کاربردهای کلان‌داده‌ها را در زمینه‌های مهندسی، علوم رایانه، فیزیک، اقتصاد و علوم زیستی پوشش دهد. کتاب‌های این مجموعه به تجزیه و تحلیل مجموعه‌ی داده‌های کلان مستقل، پیچیده یا توزیع شده، همچنین داده‌های دیجیتالی استخراج شده توسط حسگرها یا دیگر ابزارهای فیزیکی مانند شبیه‌سازها، شبکه‌های عصبی، تراکنش‌های اینترنتی و ... اشاره دارد.

مجموعه کتاب‌های «مطالعات داده‌ها، اطلاعات و دانش» شامل یادداشت‌ها، تک‌نگاره‌ها و منابع مرتبط با کلان‌داده است که حوزه‌های هوش محاسباتی^۱، شبکه‌های عصبی^۲، محاسبات تکاملی^۳، سیستم‌های فازی^۴، هوش مصنوعی^۵، داده‌کاوی^۶، آمار و تحقیقات عملیاتی مدرن^۷، سیستم‌های خودسازمان‌دهنده^۸ و ... را در بر می‌گیرد.

امید است مجموعه کتاب‌های **مطالعات داده‌ها، اطلاعات و دانش** برای علاقه‌مندان به حوزه‌های فوق، حاوی محتوای مفید باشد.

Big Data ^۱
computational intelligence ^۲
neural networks ^۳
evolutionary computation ^۴
fuzzy systems ^۵
artificial intelligence ^۶
data mining ^۷
modern statistics and Operations research ^۸
self-organizing systems ^۹

فهرست مطالب

۷	پیشگفتار
۹	فصل نخست:
۹	مقدمه‌ای بر داده‌ها
۱۰	۱-۱ راهنمای تعریف استراتژی مبتنی بر کلان‌داده
۱۱	۱-۲ تصورات نادرست درباره کلان‌داده‌ها
۱۵	۱-۳ جمع‌بندی و نتیجه‌گیری فصل
۱۷	فصل دوم:
۱۷	مدیریت کلان‌داده؛ چگونه استراتژی داده‌ها را ایجاد، اجرا و سازمان‌دهی کنیم؟
۱۷	۲-۱ رویکرد ناب کلان‌داده
۲۰	۲-۲ ساختار توسعه مراحل داده
۲۴	۲-۳ تجزیه و تحلیل داده‌ها
۲۹	فصل سوم:
۲۹	مقدمه‌ای بر هوش مصنوعی
۳۰	۳-۱ تعریف پایه‌ای و طبقه‌بندی هوش مصنوعی
۳۴	۳-۲ تاریخچه‌ای کوتاه از هوش مصنوعی
۳۹	۳-۳ چرا امروزه هوش مصنوعی مهم است؟
۴۵	فصل چهارم:
۴۵	نقشه دانش هوش مصنوعی: طبقه‌بندی فناوری‌های هوش مصنوعی
۵۵	فصل پنجم:
۵۵	پیشرفت‌های هوش مصنوعی
۵۶	۵-۱ یادگیری ماشین
۶۳	۵-۲ پیشرفت‌های علوم عصبی
۶۸	۵-۳ سخت‌افزار و تراشه‌ها
۷۱	فصل ششم:
۷۱	مدل‌های کسب و کار مبتنی بر داده و هوش مصنوعی
۷۷	۶-۱ شرکت‌های دانشگاه‌محور
۷۸	۶-۲ داده به عنوان خدمت
۷۸	۶-۳ مدل به عنوان خدمت

۷۹	۴-۶ ربات به عنوان خدمت
۸۳	فصل هفتم:
۸۳	کاربردها: هوش مصنوعی در خدمات مالی و صنعت بیمه
۸۳	۷-۱ هوش مصنوعی و خدمات مالی
۸۵	۷-۱-۱ هوش مصنوعی: تحول‌ساز جدید مالی
۸۷	۷-۱-۲ تقسیم‌بندی هوش مصنوعی در فین‌تک
۹۰	۷-۲ هوش مصنوعی و صنعت بیمه
۹۲	۷-۲-۱ چگونه هوش مصنوعی می‌تواند به صنعت بیمه کمک کند؟
۹۴	۷-۲-۲ نوآوری در صنعت بیمه
۹۷	فصل هشتم:
۹۷	هوش مصنوعی و بلاک‌چین
۹۷	۸-۱ مقدمه‌ای غیر تخصصی درباره بلاکچین
۱۰۰	۸-۲ عرضه اولیه سکه‌ها
۱۰۲	۸-۳ چگونه هوش مصنوعی می‌تواند بلاکچین را تغییر دهد؟
۱۰۴	۸-۴ چگونه بلاک‌چین می‌تواند هوش مصنوعی را تغییر دهد؟
۱۰۷	۸-۵ نتیجه‌گیری
۱۰۹	فصل نهم:
۱۰۹	نقش‌ها و عناوین شغلی جدید مرتبط با هوش مصنوعی و دانش داده‌ها
۱۰۹	۹-۱ استخدام رهبران جدید برای انقلاب داده‌ها
۱۱۰	۹-۲ مدیر ارشد داده
۱۱۳	۹-۳ مدیر ارشد هوش مصنوعی
۱۱۵	۹-۴ مدیر ارشد رباتیک
۱۱۷	فصل دهم:
۱۱۷	هوش مصنوعی و مباحث اخلاقی
۱۱۷	۱۰-۱ چگونه ماشین‌هایی با رفتارهای اخلاق‌محور طراحی کنیم
۱۱۸	۱۰-۲ داده‌ها و تعصبات
۱۲۱	۱۰-۳ مسئولیت‌پذیری و اعتماد
۱۲۶	۱۰-۴ استفاده از هوش مصنوعی و مسئله کنترل
۱۲۸	۱۰-۵ ایمنی هوش مصنوعی و خطرات فاجعه‌بار
۱۳۰	۱۰-۶ نتیجه‌گیری
۱۳۱	فصل یازدهم:

- هوش مصنوعی و مالکیت فکری ۱۳۱
- ۱۱-۱ چرا استارت‌آپ‌ها نوآوری‌ها و اختراعات را ثبت می‌کنند؟ ۱۳۱
- ۱۱-۲ مزایای ثبت اختراع برای محصول مرتبط با هوش مصنوعی و داده ۱۳۴
- ۱۱-۳ دلایل عدم پیگیری حفاظت از اختراع ۱۳۷
- ۱۱-۴ چشم‌انداز پتنت‌ها ۱۴۱
- ۱۱-۵ نتیجه‌گیری ۱۴۳
- فصل دوازدهم: ۱۴۵
- سرمایه‌گذاری خطرپذیر در حوزه دانش داده‌ها و هوش مصنوعی ۱۴۵
- ۱۲-۱ دلیل وجودی ۱۴۵
- ۱۲-۲ پژوهش‌های پیشین ۱۴۷
- ۱۲-۲-۱ ویژگی‌های فردی و تیمی ۱۴۷
- ۱۲-۲-۲ ملاحظات مالی ۱۵۰
- ۱۲-۲-۳ ویژگی‌های کسب‌وکار ۱۵۲
- ۱۲-۲-۴ دانش صنعت ۱۵۳
- ۱۲-۲-۵ بررسی یک دیدگاه بیرونی: از بی‌خانمان‌ها تا موفق‌های برجسته ۱۵۴
- ۱۲-۳ نتیجه‌گیری ۱۵۴
- فصل سیزدهم: ۱۵۷
- شتاب‌دهنده‌ها و مراکز رشد هوش مصنوعی و دانش داده‌ها ۱۵۷
- ۱۳-۱ تعریف شتاب‌دهنده و مرکز رشد ۱۵۷
- ۱۳-۲ آیا سرمایه‌گذاری در حوزه داده و هوش مصنوعی ارزش دارد؟ ۱۵۹
- ۱۳.۲.۱ دیدگاه کارآفرین: پیوستن یا نپیوستن ۱۵۹
- ۱۳.۲.۲ دیدگاه سرمایه‌گذار: ماندن یا رفتن ۱۶۰
- ۱۳.۲.۳ شاخص‌های ارزیابی شتاب‌دهنده‌ها ۱۶۲
- ۱۳-۳ نتیجه‌گیری پایانی ۱۶۵

پیشگفتار

در این کتاب، قصد داریم مقدمه‌ای بر کلان‌داده، هوش مصنوعی و دانش داده^۱ و راهکارهای کلاسیک راه‌اندازی کسب و کارهای مبتنی بر داده برای افراد علاقه‌مند به این حوزه‌ها داشته باشیم. این کتاب، نه یک منبع فنی است و نه سندی راهبردی. کتابی که مطالعه خواهید کرد، مجموعه‌ای از درس‌هایی است که می‌تواند شما را به دیدگاهی نو درباره داده‌ها راهنمایی کند. از این نظر، این کتاب، متنی آرگانیک نیست که الزاماً از آغاز تا پایان کتاب به صورت متوالی خوانده شود، هرچند بهتر است که سه تا چهار فصل نخست به صورت متوالی خوانده شود تا چشم‌اندازی کلی از فناوری‌های مرتبط با داده‌ها ایجاد شود و سپس حسب نیاز، بخش‌های دیگر با اولویت خواننده مطالعه شود.

تعریف و کاربردهای هوش مصنوعی که امروزه موضوعی داغ و قابل توجه است، در این کتاب توضیح داده شده است. برای حفظ جامعیت و کاربرد کتاب، تلاش کرده‌ایم بین توضیح مفاهیم، ابزارها و روش‌های استفاده، برنامه‌ها یا روندهای بالقوه استفاده از هوش مصنوعی، تعادل برقرار باشد، به این امید که خواننده بتواند مفاهیم کلی هوش مصنوعی و داده‌ها را درک کرده و با خلاقیت خود، کاربردهای هوش مصنوعی، کلان‌داده و داده‌کاوی را برای کسب و کار خود بیابد.

ساختار کتاب به شکلی پیش خواهد رفت که در فصل‌های نخست به تعاریف اجمالی و زمینه‌های تجاری هوش مصنوعی و کلان‌داده پرداخته و در ادامه، در مورد مسائل و موضوعات اساسی‌تر که توسط هوش مصنوعی به طور معناداری در حال تغییر است بحث خواهیم کرد.

همچنین در این کتاب، مطالبی برای محققان حوزه داده، برای مشاوران راه‌اندازی کسب و کار و همچنین سرمایه‌گذاران وجود دارد.

در پایان، دلیل اصلی تالیف این کتاب، ایجاد تحول در نحوه‌ی تفکر درباره‌ی راهبرد و کاربرد فناوری اطلاعات در جهت راه‌اندازی کسب و کارهای مبتنی بر داده، بهبود زندگی علاقه‌مندان به دانش هوش مصنوعی و داده‌ها است که منجر به خلق پروژه‌هایی شود که تاثیری مثبت در جامعه ایجاد نماید.

هادی فرزاد

تهران؛ زمستان ۱۴۰۳

فصل نخست:

مقدمه‌ای بر داده‌ها

روش‌های گوناگونی برای تعریفی کامل از کلان‌داده^{۱۱} وجود دارد و احتمالاً به همین دلیل است که هنوز به تعریفی جامع و کامل برای **کلان‌داده** نرسیده‌ایم. برخی مجموعه‌ای از داده‌های بزرگ‌تر از یک آستانه‌ی خاص، برای مثال یک ترابایت^{۱۲}، را کلان‌داده توصیف می‌کنند. عده‌ای برای درک ساده‌تر کلان‌داده توسط کاربران عادی، تعریف جالبی به این نحو ارائه می‌کنند که «کلان‌داده، حجمی از داده‌هاست که به علت بزرگ بودنشان، ابزارهای تحلیلی متداول مانند اکسل^{۱۳} قادر به انجام عملیات روی آن‌ها نیستند!»

تعریف دیگری، کلان‌داده را به عنوان منعکس‌کننده‌ی ویژگی‌های تنوع، سرعت و حجم بیان می‌کند. با این حال، تعریفی وجود دارد که به نظر می‌رسد پدیده‌ی کلان‌داده را بهتر توضیح می‌دهد:

«تحلیل کلان‌داده رویکردی نوآورانه است که از فناوری‌ها و فرآیندهای مختلف برای استخراج داده‌های با ارزش از میان انبوهی از داده‌ها برای هر کاربرد دلخواه استفاده می‌کند که در نهایت به خوراکی برای سیستم‌های پایگاه داده تبدیل خواهد شد.»

Big Data^{۱۱}

^{۱۲} TB ترابایت واحد اندازه‌گیری داده؛ هر ترابایت معادل ۱۰۲۴ گیگابایت است.

^{۱۳} Microsoft Excel

در چند سال اخیر، مطالب دانشگاهی در مورد کلان‌داده‌ها به طور قابل توجهی رشد کرده است. به نحوی که یافتن کاربردهای گوناگون کلان‌داده در هر زمینه‌ی تحقیقاتی امکان‌پذیر است. برای مثال، برنامه‌های کاربردی مختلفی از کاربرد کلان‌داده در حوزه‌ی مراقبت‌های پزشکی و بهداشتی، زیست‌شناسی، پروژه‌های دولتی و خدمات عمومی، بازارهای مالی، کنترل انرژی، تشخیص ناهنجاری، پیش‌بینی جرم، مدیریت ریسک و ... ایجاد شده است.

صرف نظر از نوع و ماهیت کسب و کار، کلان‌داده‌ها می‌توانند تاثیر قدرتمندی در پیشبرد اهداف کسب و کارها داشته باشند. تحقیق‌های بسیاری در زمینه‌ی تاثیر استفاده از کلان‌داده بر پروژه‌های تجاری انجام شده است که نتیجه تحقیقات قابل توجه است: «در کسب و کارهایی که از کلان‌داده‌ها استفاده شده است، بین ۵ الی ۶ درصد رشد و بازدهی بیشتر نسبت به کسب و کارهای مشابه بدون استفاده از کلان‌داده مشاهده شده است.»

۱-۱ راهنمای تعریف استراتژی مبتنی بر کلان‌داده

بر مبنای این رویکرد، محققان استفاده از کلان‌داده در مسائل سازمانی و اجرایی متمرکز کرده‌اند. بر همین اساس پنج راهنما برای تعریف استراتژی مبتنی بر کلان‌داده مطرح می‌شود:

۱. قرار دادن افراد در قلب ابتکارات مبتنی بر کلان‌داده.

۲. استفاده از جریان اطلاعات برای خلق ارزش.

۳. افزودن متخصصان علوم رفتاری به تیم.

۴. تمرکز بر یادگیری مستمر.

۵. توجه به مشکلات تجاری به جای مشکلات تکنولوژیکی.

پس از در نظر گرفتن ۵ مورد راهنمایی فوق، در صورتی که شرکت، سازمان یا کسب و کار قصد به کارگیری کلان‌داده در فرآیندهای خود داشته باشد، بایستی داده‌هایی با سه ویژگی اساسی انتخاب شود:

- انتخاب داده‌های مناسب با کسب و کار
- تمرکز بر بزرگ‌ترین محرک عملکرد برای بهینه‌سازی کسب و کار
- تغییر قابلیت‌های شرکت یا سازمان برای استفاده از کلان‌داده

۲-۱ تصورات نادرست درباره کلان‌داده‌ها

داده‌ها به سرعت در حال تبدیل شدن به شکل جدیدی از دارایی سرمایه‌ای و منبعی مبتکرانه از ارزش هستند. باید در نظر داشت که با این که هدایت کلان‌داده در استراتژی کارآمد برای مدیریت کسب و کار ضروری است، اما این موضوع نیز نباید فراموش شود که استراتژی مبتنی بر کلان‌داده ممکن است به دلیل ویژگی‌های ساختاری شرکت، سازمان یا کسب و کار، برای تمام مشاغل ارزشمند نباشد. این تفکر که کلان‌داده‌ها می‌توانند تحول ساختاری قابل توجهی در کسب و کارها ایجاد نمایند، کاملاً صحیح نیست. از این رو برای ایجاد چارچوبی مناسب برای تعریف استراتژی مبتنی بر کلان‌داده، چند تصور نادرست بایستی تبیین شود:

تصور نادرست ۱: داده‌های بیشتر، به معنای دقت بالاتر است!

همه داده‌ها با کیفیت و با ارزش نیستند. آلوده کردن یک مجموعه با حجم بالایی از داده‌های گوناگون و غیر ضروری می‌تواند هدف و محصول نهایی را به خطر بیندازد. درست مانند انتقال خون، در صورتی که یک گروه خونی ناسازگار به بدن تزریق شود، می‌تواند برای کل بدن فاجعه‌بار باشد. همچنین همواره خطر تطبیق و یا تکرار بیش از حد داده‌ها در مدل کسب و کار وجود دارد. در تمام کاربردهای کلان‌داده، بهتر است از تلاش برای رسیدن به کمال اجتناب کنید. چرا که برای رسیدن به کمال استفاده از کلان‌داده، تعداد متغیرها و پردازش‌ها افزایش می‌یابد. متغیرهای زیاد، پیچیدگی مدل افزایش داده و منجر به کاهش دقت یا کارایی می‌شوند. داده‌های بیشتر، حتی اگر به دقت بالاتر بینجامد، اما همواره منجر به هزینه‌ی بالاتر خواهد شد. هزینه‌های مرتبط با کلان‌داده از هزینه بالاتر ذخیره‌سازی و نگهداری مدل تا هزینه پردازش

بیشتر برای فراخوانی تصاویر و نتایج گزارش‌ها و همچنین زمان پردازش است و مشخص است که هزینه‌های بالاتر زمان پردازش و ذخیره‌سازی می‌تواند تمام فرآیند مبتنی بر کلان‌داده را تحت الشعاع قرار دهد.

خلق ارزش واقعی نه در حجم بالای داده‌ها، بلکه در استفاده هوشمندانه از داده‌هاست. استفاده هوشمندانه به این معنا که پیش از ذخیره داده‌ها، مجموعه‌های کوچک‌تر را تجزیه و تحلیل کرده و پیش از ذخیره‌سازی همه داده‌ها، پردازش اولیه برای استخراج داده‌های با ارزش انجام شود. توجه داشته باشید که داده‌ها را صرفاً به خاطر داده بودن ذخیره و پردازش نکنید! چرا که داده‌های بی‌ربط ذخیره شده می‌تواند به نوبت و خطا در ساختار مدل منجر شود!

به قانون ۸۰/۲۰ پارتو^۴ پایبند باشید: «۸۰٪ از پدیده را احتمالاً بتوان با ۲۰٪ از داده‌های متعلق به آن توضیح داد»

تصور نادرست ۲: اگر می‌خواهید فرآیند مبتنی کلان‌داده انجام دهید، باید گام نخست را بزرگ بردارید!

بهترین روش پیش از سرمایه‌گذاری هنگفت در فناوری اطلاعات و زیرساخت‌های کلان‌داده، آن است که تعداد کمی از مشکلات یا فرآیندهای کسب و کار را با کلان‌داده پیش ببرید. در صورت خلق ارزش با استفاده از کلان‌داده‌ها، فرآیندهای دیگری را به صورت گام به گام با استراتژی‌های مرتبط با کلان‌داده اجرا کنید.

تصور نادرست ۳: داده‌ها عینی و واقعی هستند!

پیش از همه باید داده‌ها دسته‌بندی و آماده‌سازی شوند تا معنای «هدفی» که وابسته به آن‌هاست مشخص شود. این موضوع شاید کمی بحث برانگیز، انتزاعی یا گنگ باشد، اما داده‌ها را می‌توان

^۴ Pareto Rule

به عنوان عینی در زمانی که حقایقی را از پدیده‌های گوناگون طبیعی جمع‌آوری می‌کند، یا ذهنی در صورتی که ساختار انسانی یا اجتماعی را منعکس می‌کند تلقی کرد.

برخی داده‌ها می‌توانند واقعی باشند، یعنی بدون توجه به اینکه چه کسی و به چه دلیلی به آن‌ها نگاه می‌کند، دارای ارزش و مفهوم یکسان باشند. در مقابل برخی از داده‌های انتزاعی‌تر، ارزش خود را از تأثیری که بر جامعه تعریف می‌کنند به دست می‌آورند. تفکر بیشتر در مورد ماهیت داده‌ها بسیار ضروری است چرا که داده‌های انتزاعی می‌توانند قربانی عدم شناخت ارزش‌گذاری واقعی شده و یا به شکوفایی و نوآوری منتهی شوند. بنابراین، تفسیر داده‌ها باید بر مبنای ارزش آن برای کسب و کار انجام شود.

هر دو نوع از داده‌های تعریف شده می‌تواند به بینش‌های متفاوتی ختم شود چرا که داده‌ها فاقد عینیت کامل و الگوپذیری قطعی هستند. به همین علت، نمی‌توان هر تغییری را تبدیل به کمیته قطعی کرد.

همچنین نباید فراموش کرد که طیف گسترده‌ای از خطاهای ادراکی یا سوگیری‌های رفتاری که ممکن است عینیت تحلیل را از بین ببرد، می‌تواند بر تصمیم‌گیرندگان مهم فرآیند تأثیر گذارند. برخی از رایج‌ترین گونه‌های سوگیری‌های رفتاری برای مدیران عبارتند از:

- ارتباط‌پنداری یا آپوفنیا^{۱۵} و توجه به ارتباط یا الگوهایی که شاید اصلاً وجود نداشته باشند!
- مغالطه روایی^{۱۶} و نادیده گرفتن برخی از داده‌ها و نشانه‌های تأثیر گذار به نفع مدل کسب و کار
- سوگیری تاییدی^{۱۷} در کلان‌داده‌ها برای پیشبرد اهداف شخصی تصمیم‌گیرنده
- سوگیری انتخاب^{۱۸} و استفاده از داده‌های شناخته شده‌تر

^{۱۵} Apophenia

^{۱۶} Narrative Fallacy

^{۱۷} Confirmation bias

^{۱۸} Selection bias

پیشنهاد می‌شود پیش از آغاز مدل‌سازی استراتژی کلان‌داده، در مورد خطاهای ادراکی فوق مطالعه بیشتری شود.

مثال جالبی از سوگیری‌های رفتاری منجر به خطاهای فاحش، «اثر هاتاوی»^{۱۹} نام دارد. زمانی که بازیگری به نام «آن هاتاوی»^{۲۰} در یک برنامه خبری حاضر شد، ناگهان ارزش سهام شرکت برکشایر-هاتاوی^{۲۱} متعلق به وارن بافت^{۲۲} به شدت رشد کرد. در حالی که تنها شباهت نام خانوادگی با یک شرکت منجر به خطای ادراکی قابل توجهی تبدیل شده بود.

تصور نادرست ۴: داده‌ها تمام حقایق را بیان می‌کنند.

در گام اول، در صورتی که سوال مناسبی مطرح نشود، داده‌ها به خودی خود کاملاً بی‌معنی هستند. با اقتباس از آنچه که «دیپ تات»^{۲۳} در «راهنمای مسافران مجانی کهکشانی»^{۲۴} می‌گوید، داده‌های بزرگ می‌توانند پاسخ نهایی را به زندگی، جهان و همه چیز ارائه دهند، به محض اینکه سوال درست پرسیده شود!

اینجاست که شعور و قضاوت انسان مطرح می‌شود. طرح سوال درست و تفسیر نتایج آن، تنها در صلاحیت مغز انسان است.

طرح سوال درست برای رسیدن به حقیقتی درست برای تعریف استراتژی کلان‌داده حیاتی است. رویکردهای مبتنی بر کشف یک داده تصادفی و پیاده‌سازی بر مبنای حجم عظیمی از داده‌ها که به اصطلاح «اجازه دهید داده‌ها صحبت کنند»^{۲۵} بسیار ناکارآمد، هزینه‌بر و مخرب ایجاد ارزش

^{۱۹} Hathaway's Effect

^{۲۰} Anne Hathaway

^{۲۱} Berkshire Hathaway

^{۲۲} Warren Buffett

^{۲۳} Deep Thought

^{۲۴} The hitchhikers guide to the galaxy یک کمدی رادیویی نوشته شده توسط داگلاس آدامز است که به شکل کتاب، کمیک استریپ،

مجموعه تلویزیونی و حتی بازی ویدیویی منتشر شده است.

^{۲۵} Let the data speak

است. برای پیشگیری از هرگونه خلق هزینه‌ی زاید، فرآیند هوشمند کشف داده‌ها در کنار تجزیه و تحلیل داده‌های استخراج شده به ایجاد ارزش منجر می‌شود.

۱-۳ جمع‌بندی و نتیجه‌گیری فصل

یکی از دلایل اصلی ناکارآمدی داده‌کاوی و رویکردهای کلان‌داده، این است که داده‌کاوی بدون تعریف فرآیندی منطقی انجام می‌شود که این موضوع به اشتباهات رایجی مانند مثبت کاذب^{۲۶}، تاثیر روابط غیرواقعی، سوگیری‌ها، خطاهای شناختی، انتخاب همبستگی‌های غیرواقعی، تعریف و تحلیل متغیرهای اشتباه یا انتخاب مدل نامناسب ختم می‌شود. برای جلوگیری از بروز خطاهای فوق، بایستی توجه ویژه‌ای روی مدل‌سازی اولیه، تعریف سوال درست و یافتن همبستگی‌ها و الگوهای مرتبط داشت.

^{۲۶} False Positive

فصل دوم:

مدیریت کلان‌داده؛ چگونه استراتژی داده‌ها را ایجاد، اجرا و سازمان‌دهی کنیم؟

همان‌گونه در فصل گذشته مطرح شد، استفاده از کلان‌داده در استراتژی‌های مرتبط با مدیریت و رشد کسب و کارها، ضروری به نظر می‌رسد. با این حال، دستیابی به اجماع در مورد چگونگی، رویکرد، میزان و نحوه تحلیل کلان‌داده‌ها کار ساده‌ای نبوده است. به همین دلیل، دورنیک و هندری^{۲۷} رویکردی ناب برای مسائل مرتبط با کلان‌داده‌ها ارائه کرده‌اند که از مدل‌های دیگر، کارآمدتر است.

۱-۲ رویکرد ناب کلان‌داده

این رویکرد زمان، هزینه‌های مرتبط با جمع‌آوری داده‌ها، تجزیه و تحلیل، توسعه فناوری و سنجش بازخورد را کاهش می‌دهد. تعریف این چارچوب، از بروز تضادها و موقعیت‌های اجتناب‌ناپذیر هزینه‌بر پیشگیری می‌کند. شکل ۱-۲ گام‌های کلیدی رویکرد ناب کلان‌داده را نمایش می‌دهد.

توجه به این نکته ضروری است که پیش از آغاز به طراحی رویکرد، تمام فرآیندهای کسب و کار شناسایی شده و پس از آن چارچوب کلان داده تعریف شود. این رویکرد، شامل چهار مرحله متوالی دارای حلقه‌های بازخورد برای انواع داده‌هاست. انواع داده‌هایی که در این رویکرد مد نظر قرار می‌گیرند را می‌توان به سه دسته‌ی در حال سکون و در حال حرکت تقسیم کرد.

- «داده‌های در حال سکون»^{۲۸} داده‌هایی ایستا^{۲۹}، غیرفعال^{۳۰} و ذخیره شده در پایگاه داده^{۳۱} در نظر گرفته می‌شوند.
 - «داده‌های در حال حرکت»^{۳۲} داده‌هایی هستند که به طور مداوم به روز شده و مرتب در پایگاه داده به روز^{۳۳} می‌شوند.
 - «داده‌های در حال استفاده»^{۳۴} داده‌هایی هستند که برای محاسبات، بازخورد، اعتبارسنجی و ... در حافظه موقت قرار می‌گیرند.
- در رویکرد ناب کلان‌داده، توالی مراحل مدل‌سازی بسیار مهم است. در هر مرحله، اعتبارسنجی^{۳۵} تعبیه شده و فرآیند با اجرای مقیاس‌پذیری^{۳۶} و اندازه‌گیری^{۳۷} به پایان می‌رسد.

Data at Rest ^{۲۸}

Static ^{۲۹}

Inactively ^{۳۰}

Database ^{۳۱}

Data at Motion ^{۳۲}

Update ^{۳۳}

Data In Use ^{۳۴}

Validation ^{۳۵}

Scalability ^{۳۶}

Measurement ^{۳۷}



شکل ۱-۲: رویکرد ناب کلان‌داده

مکانیسم بازخورد باید از ایجاد رکوردهای غیرضروری جلوگیری کرده و خروجی مناسبی برای خواسته‌های تجاری فرآیند کسب و کار ارائه کند.

داده‌ها باید به طور مداوم از منابع گوناگون جمع‌آوری شده و با سامانه و سکو^{۳۸}های فرآیند یکپارچه شوند. در ادامه استاندارد برای ارزیابی گزارش‌دهی صحت عملکرد و اطلاعات خروجی ایجاد شود.

در پایان ارزش داده‌های استخراج شده توسط مدیر داده بررسی خواهد شد. این رویکرد، مجموعه داده‌های ناکارآمد را حذف کرده و به مجموعه‌ای داده‌ها در لایه‌های مختلفی از ارزش و اهمیت می‌انجامد.

باید توجه داشت که حتی در صورتی که رویکرد ناب کلان‌داده در شرکت‌ها و سازمان‌ها مورد استفاده قرار گیرد، باز هم ممکن است این رویکرد، پاسخگوی نیازهای کسب و کار نباشد. به همین دلیل، در کنار کلیات رویکرد، تعریف چارچوبی اختصاصی برای هر کسب و کار ضروری است تا با ردیابی مسیر داده‌ها و شناسایی موانع و نقاط اهرمی داخلی، گام‌های رویکرد با موفقیت پیش رود.

۲-۲ ساختار توسعه مراحل داده

«ساختار توسعه مراحل داده^{۳۹}» یک مدل مناسب که برای خلق نقشه راهی در جهت اجرای استراتژی داده‌ها و خلق ارزش ایجاد شده است. از ساختار توسعه مراحل داده می‌توان برای ارزیابی وضعیت فعلی کسب و کار و درک گام‌های آتی برای ارتقای قابلیت‌های کلان‌داده‌ها استفاده کرد.

جدول ۱-۲ ماتریس چهار در چهار ساختار توسعه مراحل داده است که در آن، مراحل تکامل مدل به صورت طرح اولیه، سفارشی، تجاری و علمی نمایش داده شده است. در این جدول، معیارهای اصلی فرهنگ سازمان^{۴۰}، داده، فناوری و نبوغ^{۴۱} و منابع انسانی در نظر گرفته شده‌اند. در ردیف آخر جدول نیز، ملاحظات نهایی شامل موارد تاثیرگذار در امور مالی استراتژی کسب و کار بیان شده است.

مرحله نخست این ساختار، در مورد افزایش آگاهی است. درک این موضوع که علم داده می‌تواند منفعی برای کسب و کار داشته باشد. در این مرحله که فاقد ساختار رسمی و اداری بوده، نیازی به تهیه فناوری خاص وجود ندارد. در مرحله شناخت، داده‌های اولیه به مدیریت و افراد تاثیرگذار انتقال داده می‌شود، بدون اینکه تاثیری در فرآیندهای جاری کسب و کار داشته باشد. با این حال، پروژه‌های موفق، نتیجه آگاهی اولیه در مورد فرآیند پیش رو است. به همین دلیل، توجه به مرحله نخست می‌تواند به شناخت دقیقی از مشکلات و طرح سوالات مناسب برای مدل سازی یاری رساند. این مرحله می‌تواند به شدت زمان‌بر باشد، چرا که عملکرد و خروجی پروژه‌های مبتنی بر داده، دارای متغیرهای زیاد، احتمالات و ناپایداری نتایج هستند، لذا طول کشیدن بیش از حد این مرحله، به معنای ناموفق بودن ساختار توسعه نیست.

مرحله دوم به نوعی تقویت‌کننده آگاهی‌های مرحله‌ی پیشین و در واقع دوره‌ی اکتشاف است. در این مرحله، پذیرفته‌ایم که کلان‌داده‌ها دارای ارزش هستند. اما برای رسیدن به ارزش، نیاز به فناوری و زیرساخت^{۴۲}‌های جدید و حتی تغییر نگرش کلان در سطح مدیران ارشد وجود دارد تا بازیگران جدیدی وارد فرآیند تجزیه و تحلیل داده‌ها شوند. از آنجا که در این مرحله، مشارکت مدیریت هنوز محدود است، برنامه‌های کاربردی بالقوه به یک بخش یا یک عملکرد

Culture^{۴۰}Talent^{۴۱}infrastructures^{۴۲}

خاص واگذار می‌شوند. روش‌ها مورد استفاده، اگر چه پیشرفته‌تر از فاز نخست هستند، اما هنوز به پختگی کافی نرسیده و قابلیت تکرار و استفاده دوباره در فرآیندهای مختلف را ندارند. در ادامه، فاز سوم فرآیندی استانداردتر، بهینه‌تر و قابلیت تکرار را ارائه می‌کند. دسترسی به داده‌ها بسیار گسترده‌تر شده، ابزارها در خط مقدم قرار گرفته و روال استخدام و جذب منابع انسانی تنظیم شده است. به دلیل حمایت سطح بالای مدیران، پروژه از تخصیص منظم بودجه بهره‌مند می‌شود.

مرحله چهارم، به تحول کسب و کار می‌پردازد. در این مرحله، هر عملکرد مبتنی بر داده است. پروژه با روش‌های چابک^{۴۳} هدایت می‌شود و با پشتیبانی کامل مدیران ارشد، پروژه دارای مجموعه‌ای از عملیات به هم مرتبط شده که ممکن است به ایجاد مرکزی برای تعالی، تقویت و تحقیق، آموزش و توسعه فناوری‌های مربوط به داده منجر شود. بودجه قابل توجه برای زیرساخت‌های فناورانه و معماری پیشرفته و بهینه شده پایگاه داده، تاثیر قابل توجهی بر جریان درآمدها خواهد داشت. در این مرحله توجه بیشتری به نوع ذخیره سازی داده معطوف خواهد شد تا زبان‌های چندگانه پشتیبانی شوند. مقایس پذیری بالا و داده‌های متمرکز، به شرکت این اجازه را خواهد داد تا بدون هزینه‌ی اضافی، خروجی‌های مختلف برای سکوها^{۴۴}ی گوناگون ایجاد شده و پایداری همیشگی داده‌های مهم را تضمین کند.

علمی	تجاری	سفارشی	طرح اولیه	
پشتیبانی کامل مدیران - کسب و کار مبتنی بر داده	- مورد حمایت مدیران - تجزیه و تحلیل برای شناسایی مسائل کلان - همسویی با کسب و کار	- مورد توجه و پشتیبانی مدیران میانی - تحلیل برای درک مشکلات - تعریف دپارتمان	- بدون نیاز به حمایت مدیران ارشد - تجزیه و تحلیل به عنوان دارایی IT	فرهنگ سازمانی

^{۴۳} Agile methodologies

^{۴۴} Platforms

	<ul style="list-style-type: none"> - انتقال اطلاعات - بدون بودجه - تحلیل توصیفی 	<ul style="list-style-type: none"> - تامین مالی پروژه - تعریف روش کار - تجزیه و تحلیل دقیق 	<ul style="list-style-type: none"> - بودجه انحصاری - داده‌های پیشرفته - تجزیه و تحلیل انحصاری 	<ul style="list-style-type: none"> - سرمایه‌گذاری قابل توجه زیرساختی و فناوری - کشف داده پیشرفته - تجزیه و تحلیل خودکار
داده	<ul style="list-style-type: none"> - بدون نیاز به زیرساخت داده - داده‌های پراکنده - اطلاعات زیاد و حتی تکراری 	<ul style="list-style-type: none"> - داده‌های همگن - نقاط اهرمی تعریف شده - شکاف واضح بین انواع داده 	<ul style="list-style-type: none"> - داده‌های مجازی - داده‌های داخلی و خارجی - داده‌های ساختار یافته - مدیریت آسان داده‌ها 	<ul style="list-style-type: none"> - داده‌های بی‌کران - پشتیبانی از هر نوع داده - منابع متنوع داده - چرخه حیات اطلاعات
فناوری	<ul style="list-style-type: none"> - بدون حاکمیت داده - بدون فناوری‌های پیشگام - سرمایه محدود 	<ul style="list-style-type: none"> - پایگاه داده SQL - رابطه‌ای یکپارچه - بهبود در معماری داده‌ها - تعریف داده‌های شاخص - زبان‌های اسکریپت و برنامه‌نویسی 	<ul style="list-style-type: none"> - فناوری‌های پیشگام - یکپارچه سازی با تمام زبان‌های برنامه نویسی سازمان - ابزارهای بصری سازی 	<ul style="list-style-type: none"> - مجموعه داده‌های متمرکز - فضای ذخیره ابری - برنامه‌های موبایل - APIها، اینترنت اشیا، ابزارهای یادگیری ماشین
نبوغ و منابع انسانی	<ul style="list-style-type: none"> - نیروی انسانی غیرمتمرکز - تیم کوچک با مهارت‌های تحلیلی اندک 	<ul style="list-style-type: none"> - ترکیبی از چندین متخصص داده تمام وقت و پاره وقت - تیم تحلیل داده - همکاری استراتژیک برای افزایش قابلیت‌ها 	<ul style="list-style-type: none"> - استخدام نیروی متخصص - تیم مدیریت داده - بخش فناوری اطلاعات کامل - پشتیبانی از تیم فناوری به عنوان تیم داده 	<ul style="list-style-type: none"> - مرکز آموزش عالی - کارشناسان و متخصصان - آموزش و یادگیری مستمر - حضور فعال در اکوسیستم داده
تاثیر	<ul style="list-style-type: none"> - بدون بازگشت سرمایه 	<ul style="list-style-type: none"> - درآمد متوسط برای توجیه سرمایه‌گذاران 	<ul style="list-style-type: none"> - درآمد قابل توجه 	<ul style="list-style-type: none"> - مدل کسب و کار متحول شده (اقیانوس آبی)

جدول ۱-۲: ساختار توسعه مراحل داده

در تمام مراحل، نیاز مبرم به مدیریت ابرداده وجود دارد تا با اتخاذ سیاست‌های بررسی دقیق، از کیفیت، تجزیه و تحلیل و امنیت داده‌ها محافظت شود. داده‌ها باید به درستی ذخیره شوند، از طریق مناسب‌ترین ابزار مورد مطالعه قرار بگیرند و با ایجاد و پیگیری چرخه حیات اطلاعات، بایگانی کارآمد و به روز داده‌ها و همچنین بررسی و بهینه‌سازی مداوم پایگاه داده در نظر گرفته شود.

۳-۲ تجزیه و تحلیل داده‌ها

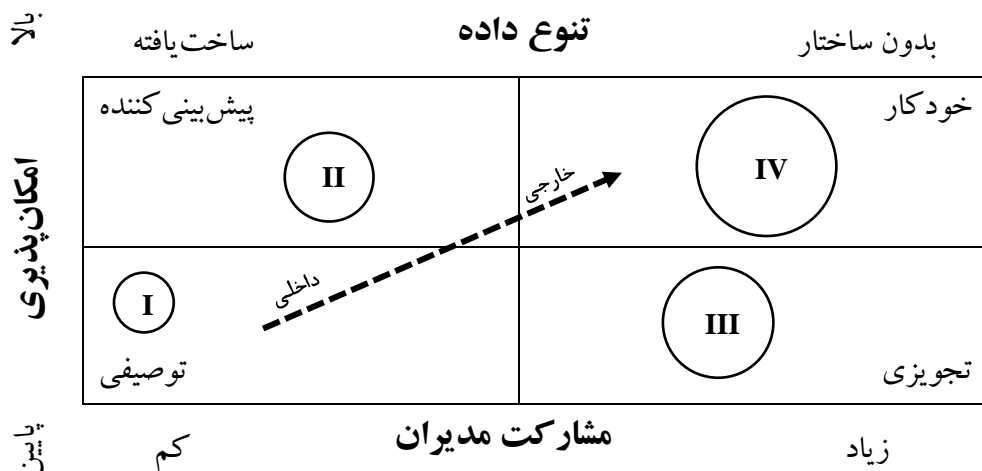
با توجه به نیازمندی‌های هر پروژه، تجزیه و تحلیل‌های گوناگونی بر روی داده‌ها انجام می‌شود. در شکل ۲-۲، رابطه بین پشتیبانی مدیریت از عملکرد تجزیه و تحلیل و پیچیدگی مهارت‌های مورد نیاز برای ایجاد مزیت رقابتی در کسب و کارهای مبتنی بر داده نمایش داده شده است. محور افقی، میزان تعهد مدیریت را مشخص می‌کند و محور عمودی، امکان سنجی پروژه را در نظر می‌گیرد که در اینجا، امکان سنجی به عنوان نسبت پیچیدگی پروژه و قابلیت‌های مورد نیاز برای تکمیل در نظر گرفته شده است. تقاطع بین امکان سنجی و مشارکت مدیریت، ماتریس بلوغ^{۴۵} کلان داده را ایجاد می‌کند که مربوط به چهار نوع تجزیه و تحلیل است. در هر بخش، دایره‌ای نشان‌دهنده تاثیر تجزیه تحلیل بر روی کسب و کار می‌باشد. هرچه دایره بزرگتر باشد، نرخ بازگشت سرمایه گذاری^{۴۶} بالاتر خواهد بود.

همچنین در محور افقی، تنوع داده‌ها از لحاظ نوع ساختار، شامل داده‌های نیمه ساختار یافته و یا بدون ساختار در نظر گرفته شده است.

^{۴۵} Maturity

^{۴۶} ROI – Return On Investment

خط مورب نیز نوع داده‌های مورد استفاده شامل داده‌های روی بستر شبکه‌های داخلی و داده‌های عمومی را مشخص می‌کند.



شکل ۲-۲: ماتریس بلوغ کلان‌داده

تجزیه و تحلیل توصیفی، مربوط است به آنچه که اتفاق افتاده است. تجزیه و تحلیل پیش‌بینی‌کننده، در مورد سناریوهای آینده است. گاهی تجزیه و تحلیل پیش‌بینی‌کننده با تجزیه و تحلیل تشخیصی تقویت می‌شود که علل پدیده‌ای خاص را بررسی کند.

تجزیه و تحلیل تجویزی، توصیه‌هایی برای پیشبرد اهداف پروژه پیشنهاد می‌دهد و در نهایت، تجزیه و تحلیل خودکار، به صورت سیستمی انجام می‌شوند.

پس از آنکه احتمالات و مقاس‌های گوناگون پروژه شناسایی شدند، می‌توان از آن‌ها برای درک اینکه پروژه در کدام بخش قرار می‌گیرد استفاده کرد. همچنین با بررسی بیشتر، نحوه انتقال از سطحی به سطح دیگر امکان‌پذیر خواهد بود.

برای اینکه به آرامی از مرحله نخست به مرحله سفارشی حرکت کنیم، ضروری است تا در مراحل گوناگون، با انجام آزمایش‌ها و دریافت بازخوردها، تاثیر استفاده از رویکردهای مبتنی