محبوب و رو به رشد: شغل وسوسه‌انگیزی به نام دانشمند داده‌ها - دیجیاتو

هدا عربشاهی | چهارشنبه، ۱۵ اردیبهشت ۱۴۰۰

به‌نظر می‌رسد حوزه علم داده‌ها هر روز بیش‌از گذشته بزرگ‌تر و محبوب‌تر می‌شود. براساس جست‌وجوهای لینکدین، علم داده‌ها یکی از روبه‌رشدترین حوزه‌های کاری در سال 2017 بوده و در سال 2020 وب‌سایت Glassdoor فعالیت در بخش علم داده‌ها را به‌عنوان یکی از سه حرفه برتر در ایالات متحده طبقه‌بندی کرده است. اما علم داده‌ها دقیقا چه شاخه‌ای از علم را در برمی‌گیرد که چرا در سال‌های اخیر با محبوبیت فزاینده‌ای همراه شده است؟

«پتر نائور»، متخصص انفورماتیک دانمارکی نخسیتن‌بار در سال 1974 اصطلاح «علم داده‌ها» را در کتابش با عنوان «بررسی اجمالی روش‌های رایانه‌ای» به‌عنوان انقلاب داده‌شناسی (دیتالوژی) به‌کار برد. در این تعریف اولیه، نائور علم داده‌ها را صرفاً به‌عنوان رشته‌ای مرتبط با مدیریت و دستکاری داده‌ها همان‌طورکه به‌نظر می‌رسند، معرفی می‌کند و تاکید کمی بر امکان استخراج اطلاعات ارزشمند از خود داده‌ها دارد.

اما ویلیام کلیولند با آغاز قرن جدید در سال 2001 موجودیت علم داده‌ها را به‌عنوان رشته‌ای مستقل و نه به‌عنوان زیرشاخه‌ای از انفورماتیک و علم آمار به‌رسمیت شناخت و نشان داد که این علم می‌تواند در 6 حوزه تخصصی مختلف شامل پژوهش‌های چندرشته‌ای، الگوها، پردازش داده‌ها، آموزش، ارزیابی ابزارها و نظریه خلاصه شود.

با ظهور کلان‌داده‌ها و استقبال از ایده «مقدار داده‌ای»، مفهوم علم داده‌ها تکامل یافت و به‌این‌ترتیب به علمی کل‌نگر تبدیل شد که اصل بنیادین آن فقط مدیریت داده نیست بلکه ارزیابی وسیع‌تر مقدار ناهمگنی از داده‌های برآمده از منابع مختلفی است که پایگاه داده‌ها، تحلیل، حسگرها، وب و غیره را شامل می‌شود.

بنابراین، امروزه علم داده‌ها را باید به‌عنوان رشته‌ای در نظر گرفت که علوم رایانه، آمار و ریاضیات را در بر می‌گیرد. نتایج پژوهشی که سال 2018 از سوی دانشگاه پلی‌تکنیک میلان در ایتالیا برپایه تحلیل مشاغل عرضه شده روی شبکه اجتماعی لینکدین انجام شد، نشان می‌دهد که بیشترین مشاغلی که از سوی شرکت‌ها نیاز به آنها عرضه شده مربوط به بخش علم داده‌ها بوده است. این مطالعه دست‌کم سه نوع شغل را در این بخش شناسایی کرده که دانشمند داده‌ها، مهندس داده‌ها و تحلیلگر داده‌ها را شامل می‌شود.

تعریف علم داده‌ها

اگر بخواهیم به هرنوع ابزار یا نمونه‌ای از علم داده‌ها اشاره کنیم، باید اول بتوانیم تعریفی دقیق از این دانش را ارائه دهیم. اما ارائه تعریفی که بتواند مفهوم علم داده‌ها را به‌درستی بیان کند کمی پیچیده است. زیرا این اصطلاح در شیوه‌های مختلف تحقیق و تحلیل به‌کار می‌رود. بنابراین، بهتر است پیش‌از هرچیز این سوال را مطرح کنیم که خود اصطلاح «علم» به چه معنی است؟

علم مطالعه سیستماتیک دنیای مادی و طبیعی از طریق مشاهده و تجربه با هدف ارتقای درک بشر از فرآیندهای طبیعی است. به‌این‌ترتیب، «مشاهده» و «درک» دو واژه مهم در تعریف مفهوم علم هستند. اگر علم داده‌ها را به‌عنوان فرآیندی برای درک جهان از طریق الگوهایی که در داده‌ها وجود دارند درنظر بگیریم، پس وظیفه دانشمند داده‌ها تبدیل داده‌ها و تحلیل آنها و همچنین استخراج الگوها از داده‌های تحلیل شده است.

به‌بیانی دیگر، داده‌ها به دانشمند داده‌ها عرضه می‌شود و او از مجموعه‌ای از ابزارها و تکنیک‌های مختلف استفاده می‌کند تا داده‌ها را پیش‌پردازش و آنها را برای تحلیل آماده کند. پس از انجام این کار، داده‌ها برای رسیدن به الگوهای معنادار تحلیل می‌شوند.

نقش دانشمند داده‌ها شبیه به نقش یک دانشمند سنتی است. هر دو برای حمایت یا رد فرضیه‌هایی درباره چگونگی عملکرد جهان، به تحلیل داده‌ها مشغول هستند و هر دو در تلاشند برای بهتر کردن درک ما از جهان به الگوهای داده‌ها معنا بخشند. دانشمندان داده‌ها از همان شیوه‌های علمی دانشمندان سنتی استفاده می‌کنند.

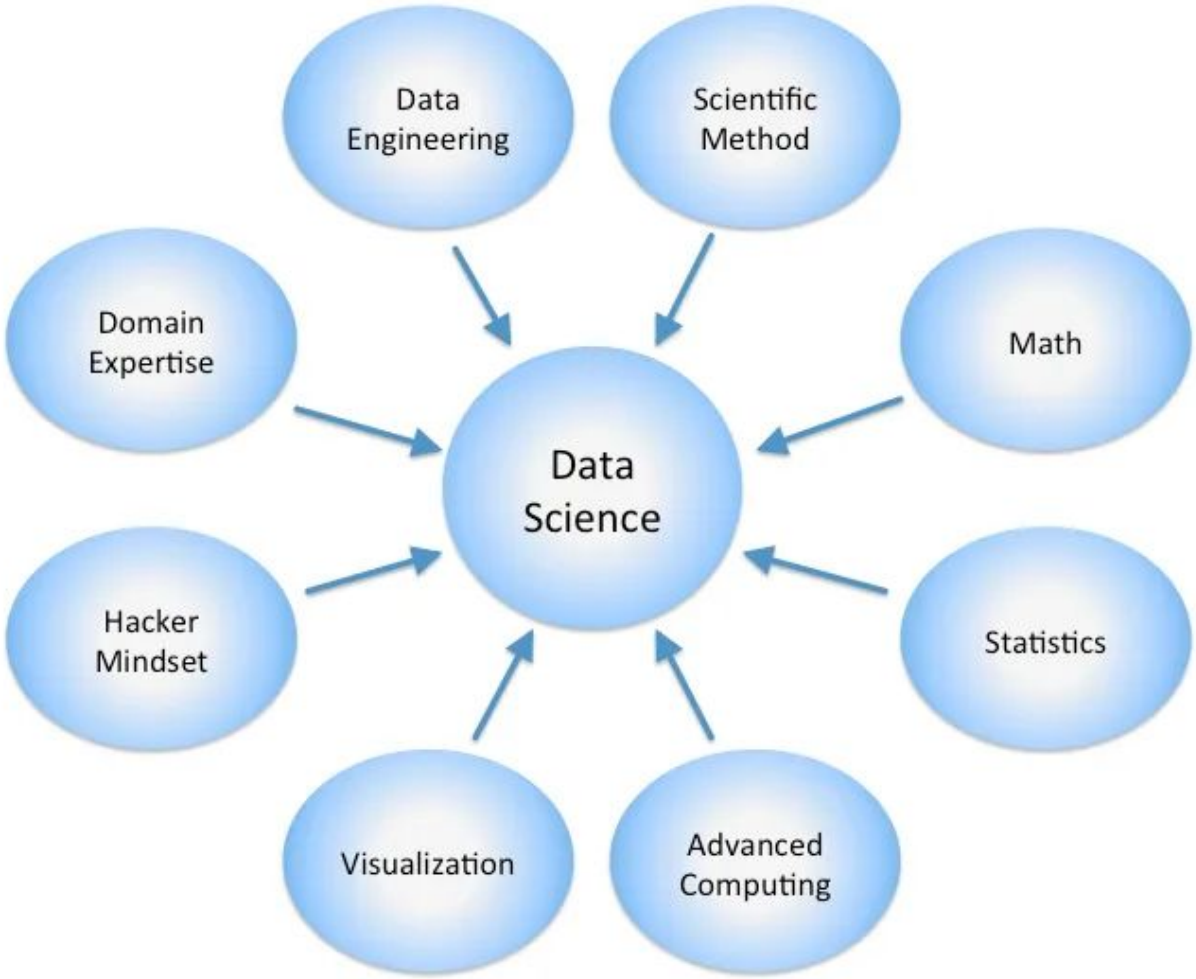
دانشمند داده‌ها با جمع‌آوری مشاهداتی که روی برخی پدیده‌هایی که می‌خواهد مطالعه کند، کارش را آغاز می‌کند. سپس، فرضیه‌ای را درباره پدیده مورد سوال مطرح می‌کند و سعی می‌کند داده‌هایی را پیدا کند که به طرق مختلف فرضیه‌اش را رد کنند. درصورتی‌که فرضیه توسط این داده‌ها نقض نشود، دانشمند قادر خواهد بود نظریه یا الگویی را درباره چگونگی عملکرد پدیده ارائه دهد.

این نظریه یا الگو بازهم می‌تواند آزمایش شود و دانشمند داده‌ها همچنان می‌تواند ببیند که آیا نظریه‌اش با دیگر داده‌های مشابه قابل ارزیابی است یا خیر. اگر یک الگو به‌حد کافی محکم باشد و طی سایر آزمایش‌ها رد نشود، می‌تواند برای پیش‌بینی اتفاقات آینده آن پدیده خاص مورد استفاده قرار گیرد.

اما نکته‌ای که درباره دانشمندان داده‌ها حائز اهمیت است، این است که به‌طور کلی این دانشمندان داده‌های مورد نیازشان را از طریق تجربه جمع‌آوری نمی‌کنند و معمولا برای کشف متغیرهای مخدوش‌کننده‌ای که می‌توانند با فرضیه‌ای خاص تداخل داشته باشند، آزمایش‌ها را با گروه‌های کنترل و کارآزمایی‌های دو سر کور طراحی نمی‌کنند.

بخش وسیعی از داده‌هایی که دانشمند داده‌ها تحلیل می‌کند، آنهایی هستند که از طریق مطالعات

و سیستم‌های مشاهداتی به‌دست آمده‌اند و درست در اینجاست که کار دانشمند داده‌ها از کار دانشمند سنتی که همواره به انجام آزمایش‌های بیشتر تمایل دارد، متفاوت می‌شود. از این‌رو، دانشمند داده‌ها می‌تواند نوعی آزمایش را تحت عنوان آزمایش A / B انجام دهد که در این آزمایش، برای دیدن چگونگی تغییر الگوهای داده‌ای، در سامانه‌ای که داده‌ها را جمع‌آوری می‌کند به‌عمد، تغییراتی ایجاد شده باشد.



سوای تکنیک‌ها و ابزارهای مورد استفاده، علم داده‌ها در نهایت قصد دارد تا با درک معنای داده‌هایی که از طریق مشاهده و آزمایش به‌دست می‌آیند، درک ما را از جهان بهبود بخشد. علم داده‌ها فرآیند استفاده از الگوریتم‌ها، اصول آماری و ابزارها و ماشین‌های مختلف برای استخراج بینش از داده‌ها است. این بینش‌ها به ما کمک می‌کنند الگوهای جهان پیرامون خودمان را درک کنیم.

وظیفه دانشمند داده‌ها چیست؟

همان‌طور که مشاهده شد هر فعالیتی که شامل تحلیل داده‌ها به روش علمی باشد را می‌توان علم داده‌ها نامید و این همان بخشی است که ارائه تعریفی درست برای علم داده‌ها را بسیار دشوار می‌کند.

برای روشن کردن موضوع، پیش از هرچیز بهتر است بعضی از فعالیت‌هایی را که دانشمند داده‌ها به‌طور روزانه انجام می‌دهد، بررسی کنیم: درطول روز، ممکن است از دانشمند داده‌ها خواسته شود که الگویی را برای بایگانی کردن و بازیابی داده‌ها طراحی کند، خطوطی را برای داده‌های ETL (استخراج، تبدیل، بارگذاری) ایجاد کند و داده‌ها را دوباره پاک‌سازی کند، از روش‌های آماری استفاده کند، امکان مشاهده داده‌ها را فراهم کند، هوش‌مصنوعی را پیاده‌سازی کند و الگوریتم‌های یادگیری خودکار و توصیه‌هایی برای اقدامات داده‌محور را ارائه دهد.

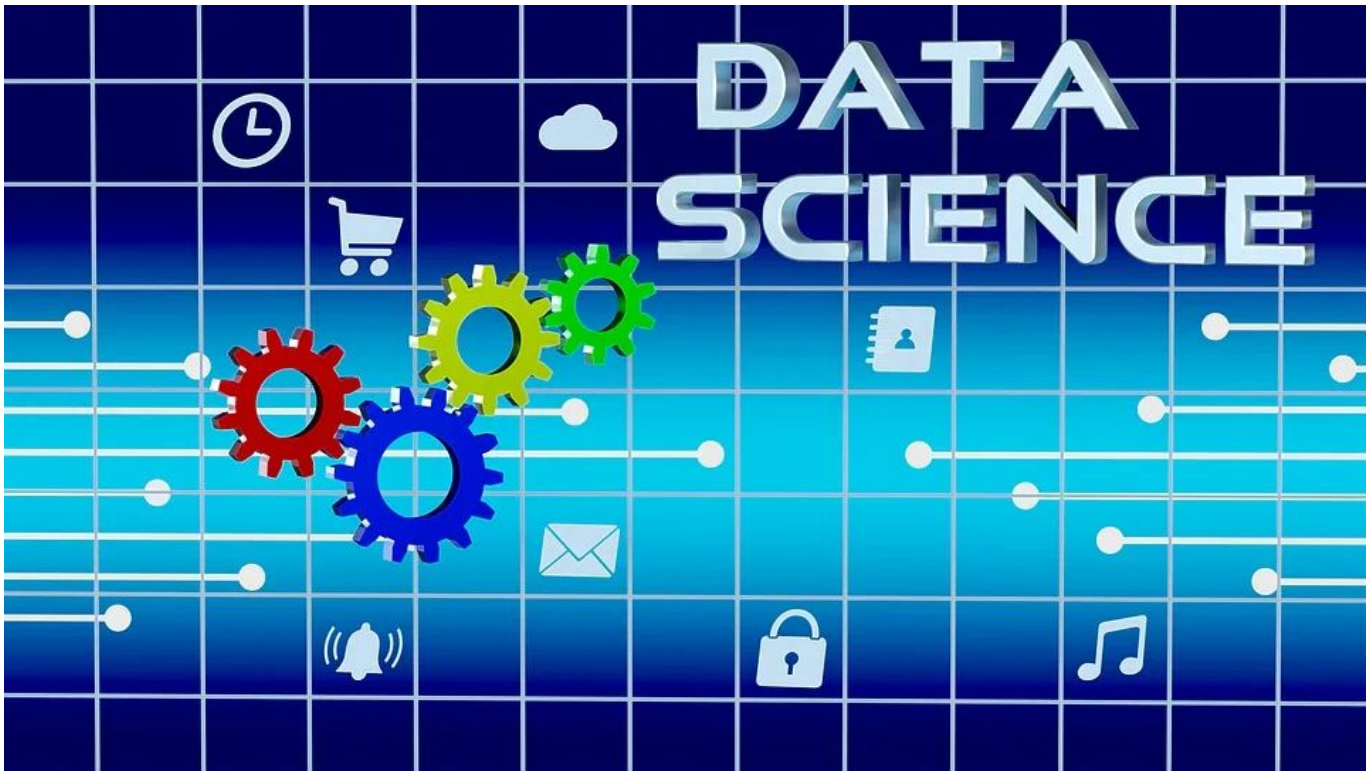
بایگانی، بازیابی، ای‌تی‌ال و پاک‌سازی داده‌ها

ممکن است از دانشمند داده‌ها خواسته شود تا با نصب سخت‌افزارها و نرم‌افزارها، فناوری‌های لازم برای ذخیره و بازیابی اطلاعات را مدیریت کند. مسئول این بخش را می‌توان «مهندس داده‌ها» نامید. با این وجود، بعضی از شرکت‌ها ترجیح می‌دهند کل این مسئولیت‌ها به‌عهده دانشمند داده‌ها باشد.

همان‌طور که پیشتر گفته شد، دانشمند داده‌ها همچنین ممکن است نیاز داشته باشد که خطوطی را برای داده‌های ETL ایجاد کند. داده‌ها به‌ندرت همان‌طور که دانشمند داده‌ها به آنها نیاز دارد، قالب‌بندی می‌شوند. در واقع، داده‌ها باید به صورت خام از منبع داده دریافت شوند و سپس به فرمت‌های قابل استفاده و پیش‌پردازش شده تبدیل شوند. مواردی چون استانداردسازی داده‌ها، حذف افزونگی‌ها و حذف داده‌های خراب از جمله کارهایی هستند که باید برای تبدیل کردن داده‌های خام به فرمت‌های قابل استفاده انجام داد.

شیوه‌های آماری

برای تبدیل کردن داده‌ها، استفاده از آمار ضروری است. در واقع، از شیوه‌های آماری برای استخراج الگوهای مورد نیاز از مجموعه داده‌ها استفاده می‌شود. از این‌رو، دانشمند داده‌ها باید درک درستی از مفاهیم آماری داشته باشد.



این دانشمند باید بتواند از طریق بررسی متغیرهای مغشوش، همبستگی قابل‌توجهی را از همبستگی‌های جعلی تشخیص دهد و همچنین برای اینکه بتواند تعیین کند که در مجموعه داده‌ها کدام ویژگی‌ها برای الگوی مورد نیازش کاربردی است، باید با ابزارهای مناسب این کار به‌خوبی آشنا باشد و باید بداند در الگوهای آماری چه زمانی باید از رویکرد رِگرسیون (تحلیل وایازشی) و چه‌زمانی از رویکرد طبقه‌بندی استفاده کند و چه‌زمانی باید نگران میانگین نمونه باشد. به بیانی ساده، دانشمند داده‌ها بدون این مهارت‌های اساسی دانشمند نخواهد بود.

نمایش داده‌ها

یکی از بخش‌های حیاتی کار دانشمند داده‌ها این است که یافته‌های خودش را به دیگران منتقل کند و اگر نتواند به‌طور موثری کشف‌هایش را به دیگران معرفی کند، نتایج بررسی‌هایش از حیث اهمیت خارج خواهند شد.

از سوی دیگر، دانشمند داده‌ها باید راوی بسیار خوبی باشد. بدین‌معنی که بتواند نماهایی را تولید کند و از طریق آنها ارتباط معنایی نکات مرتبط به‌هم را روی مجموعه داده‌ها و الگوهایی که کشف کرده، نشان دهد. ابزارهای مختلف و متنوعی برای به تصویر کشیدن و به نمایش گذاشتن داده‌ها وجود دارد که با استفاده از آنها می‌توان داده‌ها را برای اهداف اولیه (تحلیل اکتشافی داده‌ها) به‌معرض دید گذاشت و نتایجی را که برپایه الگوها به‌دست آمده‌اند به تصویر کشید.

توصیه‌ها و اهداف سازمانی

دانشمند داده‌ها همچنین باید در مورد نیازها، اهداف و فعالیت‌های سازمان یا کسب‌وکاری که در

خدمت آنها است درک واضحی داشته باشد، از محدودیت‌هایی که آنها اعمال می‌کنند و فرضیاتی که مقامات راس سازمان ارائه می‌دهند، آگاه باشد و بداند که باید چه نوع متغیرها و ویژگی‌هایی را تحلیل کند. به‌این‌ترتیب، می‌تواند الگوهایی را که در رسیدن به اهداف و برنامه‌های آن سازمان و شرکت خاص موثرند، بررسی کند.

یادگیری خودکار و هوش‌مصنوعی

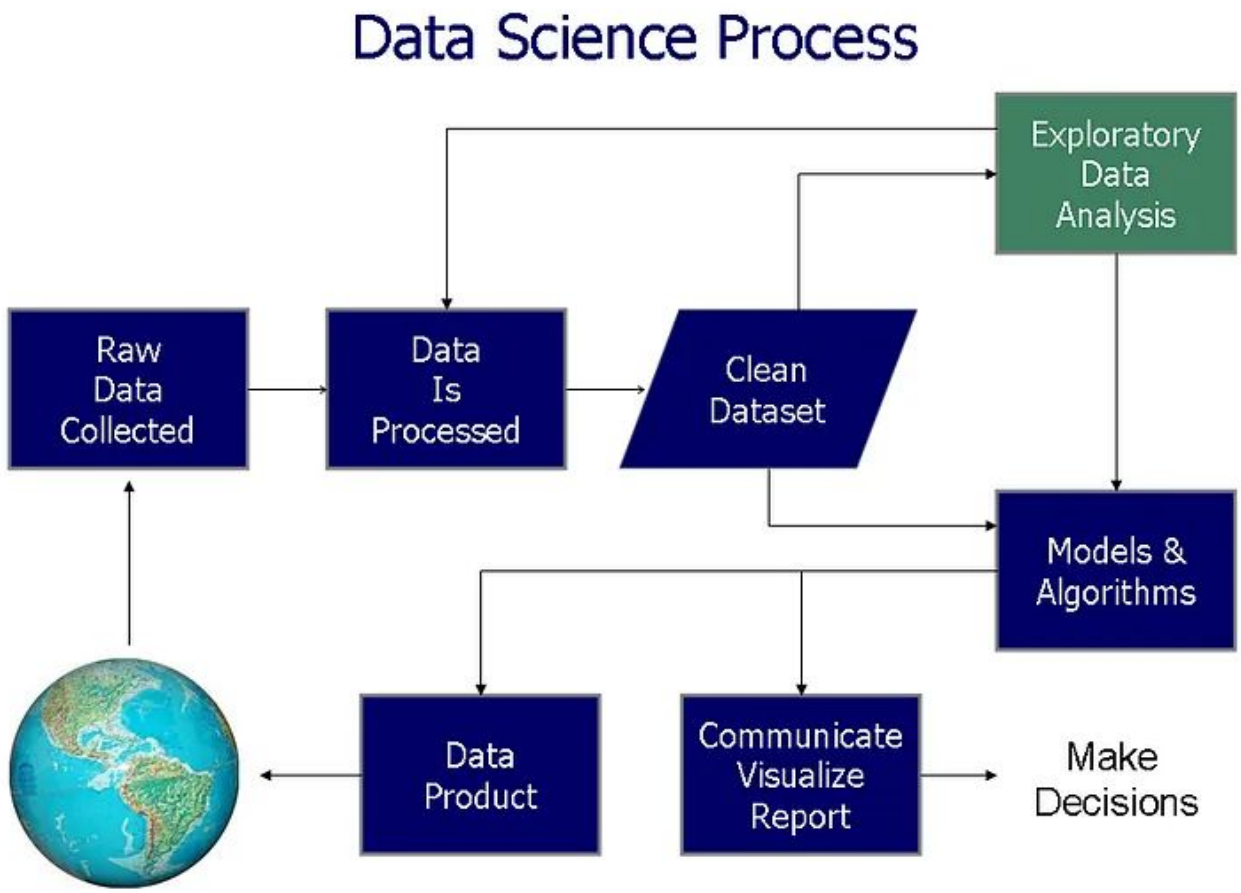
الگوریتم‌ها و الگوهای ماشین یادگیری و هوش مصنوعی از جمله ابزارهایی به‌شمار می‌روند که دانشمند داده‌ها باید از آنها برای تحلیل، شناسایی الگوهای داخل داده‌ها و یافتن ارتباط میان متغیرها و پیش‌بینی رویدادهای آینده استفاده کند.

علم داده‌های سنتی دربرابر علم کلان‌داده‌ها

از زمانی‌که شیوه‌های جمع‌آوری داده‌ها پیچیده‌تر و پایگاه‌های داده‌ها بزرگ‌تر شده‌اند، بین علم داده‌های سنتی و علم کلان‌داده‌ها تفاوت‌هایی نمایان شده است.

تحلیل داده‌های سنتی و علم داده‌ها از طریق شیوه تحلیل توصیفی و اکتشافی و با هدف یافتن الگوها و تحلیل نتایج عملکرد طرح انجام می‌شود. روش‌های سنتی تحلیل داده‌ها اغلب فقط بر داده‌های گذشته و داده‌های فعلی متمرکز هستند و تحلیل‌گر اغلب با داده‌هایی سروکار دارد که از قبل پاک‌سازی و استانداردسازی شده‌اند.

این در حالی است که دانشمند کلان‌داده‌ها اغلب با داده‌های پیچیده و پاک‌سازی نشده سروکار دارد. تحلیل داده‌های پیشرفته‌تر و تکنیک‌های جدیدتر علم داده‌ها می‌تواند برای پیش‌بینی رفتار آینده استفاده شود. اما این کار اغلب با کلان‌داده‌ها انجام می‌شود زیرا الگوهای پیش‌بینی کننده معمولا به داده‌های زیادی احتیاج دارند تا بتوان آنها را به روشی قابل اعتماد ساخت.



ابزارهای مورد استفاده در علم داده‌ها

ابزارهای رایج برای علم داده‌ها سامانه‌هایی برای بایگانی‌سازی داده‌ها، اجرای تحلیل اکتشافی داده‌ها (EDA)، الگوهای داده‌ها، اجرای ETL (استخراج، تبدیل، بارگذاری) و نمایش داده‌ها را شامل می‌شود.

بسترهایی چون مایکروسافت آژور، سرویس‌های وب آمازون و گوگل کلود تمام ابزارهای لازم را برای کمک به دانشمند داده‌ها در بایگانی‌سازی، تبدیل، تحلیل و الگوسازی داده‌ها عرضه می‌کنند. به‌علاوه، ابزارهای مستقلی چون Airflow (زیرساخت داده‌ها) و Tableau (نمایش و تحلیل داده‌ها) برای علم داده‌ها وجود دارند.

همچنین بسترها و ماژول‌هایی چون PyTorch، TensorFlow و Azure Machine-learning studio الگوریتم‌های ماشین یادگیری و هوش مصنوعی را که برای الگوسازی داده‌ها استفاده می‌شوند، عرضه می‌کنند.

نمونه‌ای برای درک بهتر علم داده‌ها

علم داده‌ها در همه زمینه‌ها از تحویل مواد غذایی تا ورزش، ترافیک و سلامت کاربرد دارد. یکی از نمونه‌های بارز آن در حوزه تحویل غذا، سرویس Uber Eats (معادل اسنپ فود) است.

Uber Eats باید غذای مردم را در کمترین زمان در وضعیتی که هنوز گرم و تازه است، تحویل دهد. به‌منظور نیل به این هدف، دانشمند داده‌های شرکت اوبر باید از الگوی آماری استفاده کند که جنبه‌هایی از جمله فاصله رستوران‌ها تا محل تحویل، ایام تعطیل، زمان مورد نیاز برای تهیه غذا و حتی شرایط آب‌وهوایی را در نظر بگیرد. با تحلیل این داده‌ها می‌توان زمان تحویل غذا را به بهترین شکل بهینه‌سازی کرد.

دیجیاتو