HERFEH+

حرفه پلاس، ورود به دنیای حرفه‌ای‌ها

داده کاوی

تعریف داده کاوی:

زمانی که از بحث داده کاوی صحبت می‌شود دو مبحث را باید در نظر گرفت یکی بحث داده یا Data و دیگری بحث اطلاعات یا Information گفته می‌شود که اطلاعات قدرت می‌آورد چون می تواند تصمیمات به هنگام انجام دهد ولی موضوع این است هر چیزی که در گام اول دریافت می‌شود قابل تصمیم‌گیری نیست ما ممکن است داده‌های زیادی دریافت کنیم اما این داده‌ها ممکن است داده‌های درستی نباشند یا حجم این داده‌ها زیاد باشد یا تنوع این داده‌ها زیاد باشد یا مسائل دیگری وجود داشته باشد.
وقتی ما می‌توانیم تصمیم‌گیری درستی برای کسب‌وکار خود بگیریم که این داده‌ها را به‌نوعی کاویده باشیم و آن چیزی که برای کسب‌وکار ما ارزشمند است برای تصمیمات استفاده کنیم به عبارتی ما باید با فیلتر کردن داده‌هایی که به دست ما می‌رسد بخشی که خیلی ارزشمند در جهت منافع سازمانی ما هست را پیدا کنیم و آن‌ها را به اطلاعات تبدیل کنیم و از آن‌ها استفاده کنیم به‌عبارت دیگر داده‌ها یک مبحث خام است و اطلاعات یک مبحث پخته و کامل است که عملا از دیتای اولیه به‌دست‌آمده است.
داده کاوی در حقیقت استخراج اطلاعات از داده‌های اولیه است بنابراین هوشمندی بسیاری می‌خواهد تا بتوانیم از داده‌های اولیه اطلاعات موردنیاز خود را به دست آوریم.
وظایف داده کاوی یا data mining چیست؟
کاویدن و استخراج دانش از منابع عظیم داده برای آشکارسازی اطلاعات گران‌بهای پنهان‌شده در حجم انبوه اطلاعات سطحی. مثلا کجا سرمایه‌گذاری کنیم، یا کجا محصولمان را بفروشیم یا کجا خط تولیدمان را بیشتر کنیم یا کجا وارد بورس شویم یا نشویم. و چشم بسته در زمینه حرفه خود تصمیمات مدیریتی نگیریم.

پایه‌های دانش داده کاوی شامل:

-آمار(مطالعه عددی روابط داده‌ها)
-هوش مصنوعی(هوش انسان مانند نرم‌افزار و یا ماشین)
-یادگیری ماشین(الگوریتم‌هایی برای پیش‌بینی داده‌ها)

تفاوت داده کاوی و تحلیل‌های علم آمار

داده کاوی در حقیقت زیرمجموعه علم آمار است اما تفاوت‌هایی نیز بین آن‌ها وجود دارد که شامل:
-حجم داده‌های مورد تحلیل
-روش مدل‌سازی داده‌ها
-استفاده از هوش مصنوعی
-فراتر رفتن اهمیت داده کاوی از علم آمار
-در اختیار قراردادن امکانات
-امکان استفاده از مدل‌های مدل‌سازی داده‌های ورودی
-دستیابی به اطلاعات پنهان موجود در حجم عظیم داده
-استفاده از شیوه‌های خودکار یادگیری

ویژگی‌ها و فواید داده کاوی یا DM

ویژگی‌های اصلی داده کاوی عبارت‌اند از:

-کشف اتوماتیک الگوها در انجام پروژه‌ها

-پیش‌بینی احتمالی نتایج و خروجی‌ها

-تولید اطلاعات اجرایی و مفید

-متمرکز بر داده‌های بزرگ

-ایجاد روابط به‌صورت اتوماتیک

-استفاده از داده‌های متنوع

-داینامیک بودن

-عدم نیاز به داده‌های صحیح

-ساخت مدل‌های واقعی

-آنالیز کردن داده‌های واقعی

-دوری از اشکالات احتمالی نمونه‌گیری

برخی از فایده‌های داده کاوی:

-تصمیم‌گیری‌های واقع‌بینانه

-کاهش هزینه‌ها

-تکرار تصمیمات سودآور

-بالابردن میزان فروش

-ارتقای کیفی پژوهش‌ها

-ارزیابی میزان ریسک

-حذف ضررهای ناشی از ناآگاهی مدیران

-شفاف کردن فضای حاکم برکسب‌وکار

-تشخیص جرائم و شناسایی سوابق افراد

برخی از محدودیت‌های داده کاوی:

-مشخص نبودن ارزش یا میزان اهمیت الگوها و روابط
-نیازمند بودن به کارشناسان فنی و تحلیل گران کارآزموده
-نیاز به‌صرف هزینه‌های بالا در برخی موارد

نمونه کاربردهای داده کاوی

-استفاده تجاری : مانند الگوی خرید افراد از فروشگاه‌های زنجیره‌ای
-استفاده پزشکی: مانند کشف الگوهای ناشناخته تاثیر داروها بر بیمارهای مختلف و گروه‌های سنی مختلف
-استفاده در بانک داری: سرمایه‌گذاری و بورس مانند شناسایی مشتریان پرخطر یا سودجو

چگونگی عملکرد سیستم و فرآیندهای داده کاوی

گام اول: درک کسب‌وکار
گام دوم: بررسی و درک داده‌ها
گام سوم: آماده‌سازی داده‌ها
گام چهارم: مدل‌سازی
گام پنجم: تست و ارزیابی مدل
گام ششم: توسعه مدل نهایی

سه گام اول در حقیقت گام‌های مقدماتی و سه گام دوم گام‌های اجرایی است.

۱-درک کسب‌وکار: در میان گذاشتن مسئله کسب‌وکار با متخصص داده کاوی و واگذاری تشخیص مسئله کسب‌وکار به متخصص داده کاوی.

۲-بررسی و درک داده‌ها: بررسی داده‌ها با توجه به حجم و کیفیت داده ها و تعدیل مسئله برای واقع‌بینانه شدن پروسه داده کاوی.

شدن پروسه داده کاوی.
۳- آماده‌سازی داده‌ها:یکی کردن انبارهای داده در کسب‌وکار و شناسایی و حذف داده‌های اشتباه همچنین تغییر فرمت داده‌ها با توجه به مسئله تعدیل‌شده در گام دوم فرآیند
۴- مدل‌سازی: انتخاب مدل با توجه به متدهای مختلف شامل:
-مدل‌سازی توصیفی
-مدل‌سازی پیش بینانه
-مدل‌سازی تجویزی
-مدل‌سازی وابستگی
است که باید بر اساس شرایط کسب‌وکار تعریف‌شده باشد.
۵-تست و ارزیابی مدل: تست و ارزیابی مدل انتخاب‌شده و بررسی موثر بودن و میزان تاثیر و در صورت لزوم تکرار این تست و ارزیابی
۶-توسعه مدل نهایی: ارائه راه‌حل‌ها با توجه به مسئله و ارائه نرم افزار شبیه‌سازی کسب‌وکار

مدل استاندارد برای داده کاوی کسب‌وکار

در بحث پروسس مدل دیگری را در موردش گفتگو می‌کنیم به نام مدل کریسپ CRISP-DM که مخفف جمله Cross-Industry Standard Process for Data Mining است که با دو تفکر به موضوع نگاه می‌کند که یکی در مورد سگمنتیشن است که می‌گوید من داده‌های اولیه‌ای که دارم پس‌ازاینکه گام‌ها و فرآیندهای داده کاوی که در بالا ذکر شد را انجام داد همه این مراحل می‌تواند بر اساس دسته‌بندی یا Classification می‌تواند انجام شود و می‌توانیم آن‌ها را به‌صورت خوشه‌بندی یا Clustering کنیم و بعد به سراغ تحلیل قواعد انجمنی یا Association Rules برویم و سیستم‌های مصورسازی یا Visualization را اجرا کنیم البته این‌ها به‌صورت AND , OR است یعنی همه را می‌توانیم داشته باشیم و یا بعضی از مکانیسم‌ها را.

در کنار این مسئله می‌تواند داده کاوی با نظارت ما که به آن یادگیری با ناظر یا Supervised Learning یا یادگیری بدون ناظر یا Unsupervised Learning یا به‌صورت سیستم‌های میانگین یا یادگیری نیمه ناظر Semi- Supervised Learning انجام شود.

معرفی ابزارهای داده کاوی:

ابزارهای داده کاوی خیلی زیاد هستند برخی از نرم‌افزارها بسیار سنگین هستند برخی سبک و اتفاقا کارا هستند ما باید از کدام‌یک از ابزارها استفاده کنیم، آیا می‌توانیم از چند ابزار در کنار هم استفاده کنیم. جواب بله است.

بهتر است که از یک مدل استفاده کنیم اما طبیعتا می‌توانیم از چند مکانیسم استفاده کنیم.

مثلا برخی از داده‌ها که نیاز به خوشه‌بندی است می‌توانیم سراغ نرم‌افزارهایی برویم که کارشان خوشه‌بندی است، برخی با آمار و ارقام کار می‌کنند برخی با خصلت‌ها کار می‌کنند برخی از زبانهای برنامه‌نویسی با برنامه‌نویسی دیگری نوشته شده‌اند

مثلا پایه‌شان پایتون یا C++ باشد

بنابراین بسته به تفکری که پشت این داده کاوی است در کوتاه مدت می‌تواند تاثیرات شگرفی را ایجاد کند.

در اینجا برخی از نرم افزارهایی که بسیار کاربری هستند و قابل دسترسی و به‌راحتی قابل بارگذاری هستند به شما معرفی می کنیم:

Rapid Miner

این نرم‌افزار دارای طیف گسترده‌ای از برنامه‌های کاربری است، می‌تواند برای پروژه‌های تجاری، علم و تحقیق دانشگاه، آموزش و غیره مورداستفاده قرار گیرد. کاربران می‌توانند با استفاده از این ابزار از تکنیک‌های مختلف داده کاوی استفاده کنند.

اگرچه این یک نرم‌افزار جدید نیست، هنوز هم به‌عنوان یکی از بهترین‌های شناخته‌شده است و در مورد قیمت‌گذاری خیلی انعطاف‌پذیر است.

Orange

این ابزار داده کاوی از بهترین نرم‌افزارهای متن بازی است که برای مصورسازی داده روی وب موجود است. این نرم‌افزار به پردازش متن و یادگیری ماشین اختصاص دارد. همچنین دارای برخی از ویژگی‌های پیشرفته برای تجزیه‌وتحلیل است. با اینکه نرم‌افزار ساده‌ای است اما می‌تواند کارهای زیادی برای شرکت شما انجام دهد.

GraphLab Create

این نرم‌افزار، بهینه‌سازی مدل، رگرسیون، مصورسازی یادگیری ماشین، تشخیص آنومالی، ساختار داده‌های مقیاس‌پذیر، تجزیه وتحلیل تصویر، یادگیری داده‌ها را ارائه می‌دهد. این ابزار اجازه می‌دهد تا شما در مورد مجموعه داده‌ها بیشتر بدانید و الگوهای مختلف را با استفاده از تکنیک‌های مختلف پیدا کنید.

با این کار، می‌توانید برنامه‌های پیش‌بینی کننده بسازید. چنین ابزاری برای شرکت‌ها شگفت‌انگیز است زیرا دوره‌های برنامه‌ریزی آن‌ها را کاهش می‌دهد و توسعه پروژه‌ها را تسریع می‌بخشد.

R Studio

آر یکی از ابزارهای داده کاوی رایگان در این لیست است. بسیاری از شرکت‌ها با محاسبات آماری و گرافیک در حال مبارزه هستند. این ابزار به شما این امکان را می‌دهد که یاد بگیرید که این ابزار همه‌چیزهایی است که شما نیاز دارید تا دانش خود را در مورد موضوع افزایش دهید.

محیط نرم‌افزار آر اجازه می‌دهد تا شما محاسبات گرافیکی مختلف انجام دهید. این ابزار برای مدل‌های خطی و غیرخطی، طبقه بندی، خوشه‌بندی و غیره مناسب است. شما می‌توانید با زبان اسکریپتی مختلف به آن دسترسی پیدا کنید.

Weka

هنگامی‌که از طریق منوی ویکا جستجو می‌کنید، به‌سرعت متوجه می‌شوید که این نرم‌افزار بر اساس چندین الگوریتم یادگیری ماشین استوار است. برخی از ویژگی‌های ویکا شامل پردازش، طبقه بندی، رگرسیون، ارتباط، انتخاب ویژگی، آزمایش‌های مختلف و غیره است. ویکا از اکسپلورر استفاده می‌کند.

KNIME

این نرم‌افزار پلتفرمی است که بر سه چیز تمرکز دارد: ادغام، تجزیه‌وتحلیل و گزارش دهی. با استفاده از این ابزار داده کاوی، کارهای زیادی می‌توانید انجام دهید. بیش از ۱۰۰۰ تجزیه‌وتحلیل مختلف وجود دارد که می‌توانید با آن انجام دهید. این ابزار برای تحلیل خود از منابع مختلف استفاده می‌کند. اگر تصمیم به‌اضافه کردن آن به لیست نرم‌افزاری خود داشته باشید قطعا ارزش آن را دارد.

OpenNN

این نرم‌افزار عمدتا بر روی پیاده‌سازی شبکه‌های عصبی تمرکز دارد. این پلتفرم در سی پلاس پلاس ساخته‌شده و مزیت اصلی آن کارایی بالا است. با توجه به مدیریت حافظه و سرعت پردازش بالا، به‌راحتی می‌تواند نرم‌افزارهای رقیب را مغلوب کند.

Apache

نام این نرم‌افزار به معنای برنامه مدیریت اطلاعات بدون ساختار است. به‌عبارت‌دیگر، مانند بسیاری دیگر از ابزارهای داده کاوی، این ابزار یک الگو را در مجموعه داده‌های بزرگ دنبال می‌کند. با آپاچی می‌توانید برنامه‌ها را به اجزای اصلی تقسیم کنید. همچنین می تواند این اجزا را به خدمات شبکه متصل کند. اگرچه این ابزار نسبتا قدیمی است اما به‌طور مرتب بروز می‌شود که بدین معنی است که شما بهترین محصول را خواهید گرفت

CLUTO

همان‌طور که از نامش برمی‌آید، این نرم‌افزار بر خوشه‌بندی به عنوان تکنیک داده کاوی تمرکز دارد. از مجموعه داده‌های مختلف صرف نظر از اندازه آن‌ها برای تجزیه‌وتحلیل خوشه‌ها استفاده می‌کند. این نرم‌افزار شامل هر دو برنامه مستقل و همچنین کتابخانه است که مورداستفاده قرار می‌گیرد که بر مبنای آن برنامه‌های کاربردی می‌توانند به الگوریتم‌های خوشه‌بندی دسترسی داشته باشند. این ابزار از چنین روش برای خوشه‌بندی به‌طور صحیح و کارآمد خلاصه می‌کند.

Anaconda

یک پلتفرم کامل داده کاوی است که با پایتون ساخته‌شده است. مانند بسیاری از این برنامه‌ها، مواردی از قبیل ساده‌سازی جریان های کاری از نقطه اولیه تا استقرار وجود دارد. همچنین به شما اجازه می‌دهد که منابع خود را به‌منظور به دست آوردن بهترین نتایج و الگوها، ادغام کنید. این نرم‌افزار برای تیم‌های بزرگ ساخته‌شده است. شما می‌توانید نتایج و پروژه‌های خود را با کل شرکت به اشتراک بگذارید.

Shogun

این ابزار می‌تواند توسط هرکسی استفاده شود. این نرم‌افزار بر اساس الگوریتم‌های مختلفی است که ما را قادر به یافتن الگوهای مختلف می‌کند. این نرم‌افزار در سی پلاس پلاس ساخته‌شده است و از سایر زبان‌های برنامه‌نویسی نیز پشتیبانی می‌کند. شاید یکی از بهترین موارد در مورد این ابزار رایگان بودن آن است.

TraMineR

یک نرم‌افزار داده کاوی است که برای موارد مختلفی استفاده می شود، با این ابزار شما می‌توانید داده‌ها را تحلیل و مستندسازی کنید.

ROSETTA

شما می‌توانید از آن برای مرور و پیش‌پردازش اطلاعات در فازهای اولیه داده کاوی، برای اعتبار سنجی و تجزیه‌وتحلیل قواعد و الگوهای القاشده استفاده کنید. این نرم‌افزار برای یک برنامه خاص ساخته نشده بلکه باهدف ابزار عام و چندمنظوره ساخته‌شده است.