



تحلیل داده های آماری

دکتر محمد ستاری

استادیار گروه مدیریت و فناوری اطلاعات سلامت دانشگاه علوم پزشکی اصفهان



مقدمه (داده کاوی)

از هنگامی که رایانه در تحلیل و ذخیره سازی داده ها بکار رفت (۱۹۵۰) پس از حدود ۲۱ سال حجم داده ها در پایگاه داده ها دو برابر شد. ولی پس از گذشت دو دهه و همزمان با پیشرفت فن آوری اطلاعات، هر دو سال یکبار حجم داده ها دو برابر شده و همچنین تعداد پایگاه داده ها با سرعت بیشتری در حال رشد است

حال با ورود سیستم های یکپارچه اطلاعاتی، لحظه به لحظه به حجم داده ها در پایگاه داده های مربوط اضافه شده و باعث به وجود آمدن انبارهای عظیمی از داده ها شده است.

سازمان ها از لحاظ داده غنی از لحاظ دانش ضعیف



تعریف داده کاوی

- ▶ فلسفه ی داده کاوی این است که آینده بسیار به گذشته شبیه است.
- ▶ داده کاوی به شما کمک می کند تا رفتار کسب و کار خود در گذشته را دقیق بشناسید و بر اساس آن آینده را با تقریب بالایی پیش بینی کنید.



تعریف داده کاوی

▶ به مجموعه‌ای از روش‌های قابل اعمال بر پایگاه داده‌های بزرگ و پیچیده به منظور کشف الگوهای پنهان و جالب توجه نهفته در میان داده‌ها، داده کاوی گفته می‌شود.

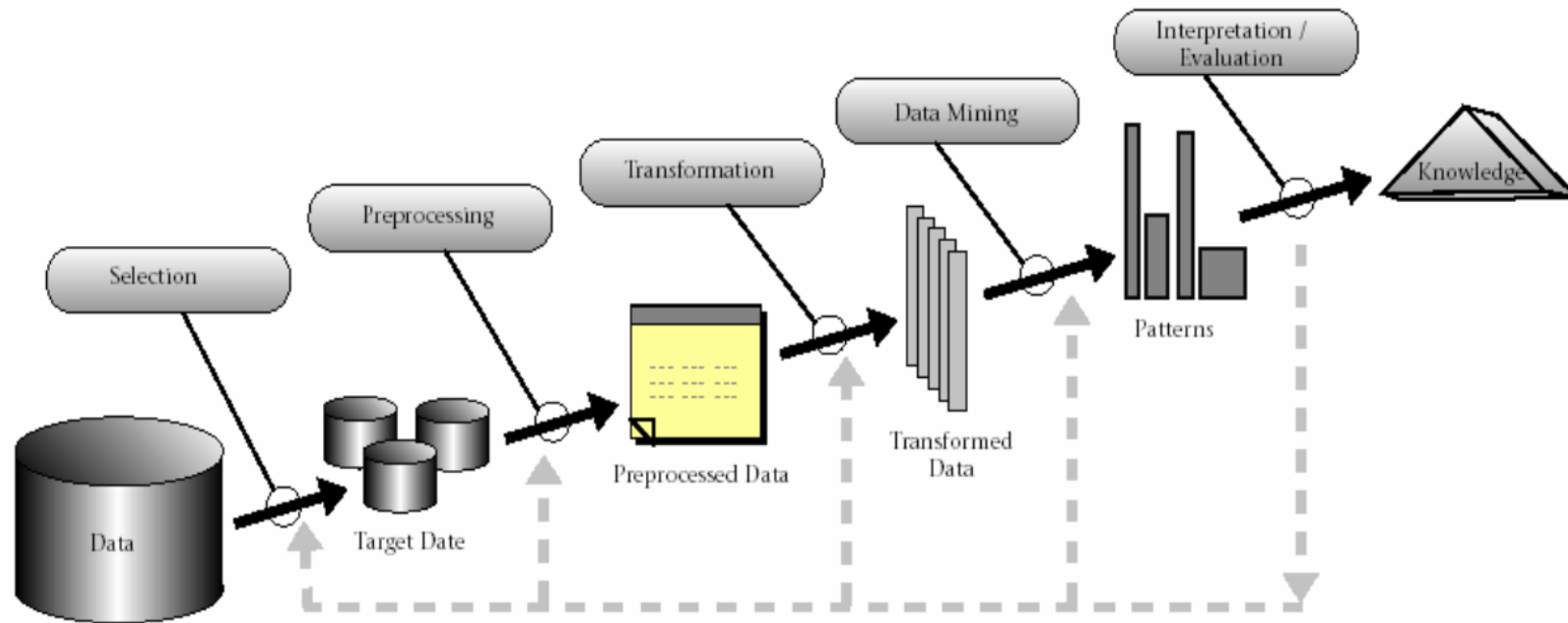




مطالعات داده کاوی

- ▶ مقطعی : این مطالعات در یک مقطع زمانی مشخص انجام می شوند و داده ها از یک جمعیت یا نمونه در همان زمان جمع آوری می گردد.
- ▶ گذشته نگر مثلا اطلاعات بیماران قلبی در ده سال اخیر

مراحل تبدیل داده به دانش





مشکلات قابل حل توسط داده کاوی

- ▶ مسئله ای پیچیده و ناساخت یافته و یا نیمه ساخت یافته
- ▶ داده‌های مرتبط وجود داشته باشند و به آنها دسترسی داشته باشیم
- ▶ داده ها در یکجا مجتمع شده و انباره داده ها ایجاد شود.
- ▶ توانایی کامپیوترها امکان استفاده از نرم افزارهای مرتبط با داده کاوی را به ما بدهند
- ▶ مدیران نیاز به استفاده از دانش استخراج شده از داده ها را حس کرده باشند



چرا داده کاوی

(unstructured

۹۰ درصد اطلاعات دنیای دیجیتال را داده‌های بدون ساختار (data) تشکیل می‌دهد.

منجر به تصمیمات واقع بینانه می‌شود.

سبب تکرار تصمیمات سودآور رخ داده در گذشته می‌گردد.



اهمیت داده کاوی در حوزه سلامت

امروزه صنعت بهداشت و درمان مقادیر زیادی داده پیچیده در مورد منابع بیمارستانی ، تشخیص بیماری ها ، پرونده های الکترونیکی بیماران ، تجهیزات پزشکی و غیره تولید می کند.

مقادیر زیادی از داده ها یک منبع کلیدی برای پردازش و تجزیه و تحلیل برای استخراج دانش است که پشتیبانی از صرفه جویی در هزینه و تصمیم گیری را امکان پذیر می کند.

کاربرد داده کاوی در حوزه سلامت

(۱) تشخیص، درمان بیماری

(۲) مدیریت منابع پزشکی





تشخیص، درمان بیماری

- ▶ تشخیص زودهنگام بیماری ها: با تحلیل داده های پزشکی، تصاویر پزشکی مانند ماموگرافی، سیگنالهای بیولوژیکی و سوابق بالینی بیماران، داده کاوی به شناسایی بیماری ها (مانند سرطان پستان، دیابت، بیماری قلبی و سرطان پوست) در مراحل اولیه کمک می کند. این تشخیص زودهنگام موجب افزایش موفقیت درمان و کاهش عوارض می شود
- ▶ پیش بینی و پیشگیری: الگوریتم های داده کاوی قادر به پیشبینی احتمال ابتلا به بیماری ها بر اساس ویژگیهای بالینی و رفتاری بیماران هستند، که امکان اتخاذ اقدامات پیشگیرانه را فراهم می آورد



تشخیص، درمان بیماری

- ▶ شخصی سازی درمان: با توجه به داده های ژنتیکی، بالینی و رفتاری، داده کاوی می تواند به طراحی پروتکل های درمانی اختصاصی برای هر بیمار کمک کند و عوارض جانبی داروها را کاهش دهد.
- ▶ تشخیص هوشمند ناهنجاری ها: داده کاوی قادر است الگوهای غیرعادی یا ناهنجار در داده های بیماران را شناسایی کند، مانند الگوهای متفاوت در بیماران کبدی، که می تواند به درمان به موقع بیماریهای خطرناک کمک کند.
- ▶ مدیریت و بهینه سازی سلامت: داده کاوی در بهینه سازی عملکرد بیمارستانها، مدیریت تجهیزات پزشکی و پایش سلامت عمومی بکار می رود



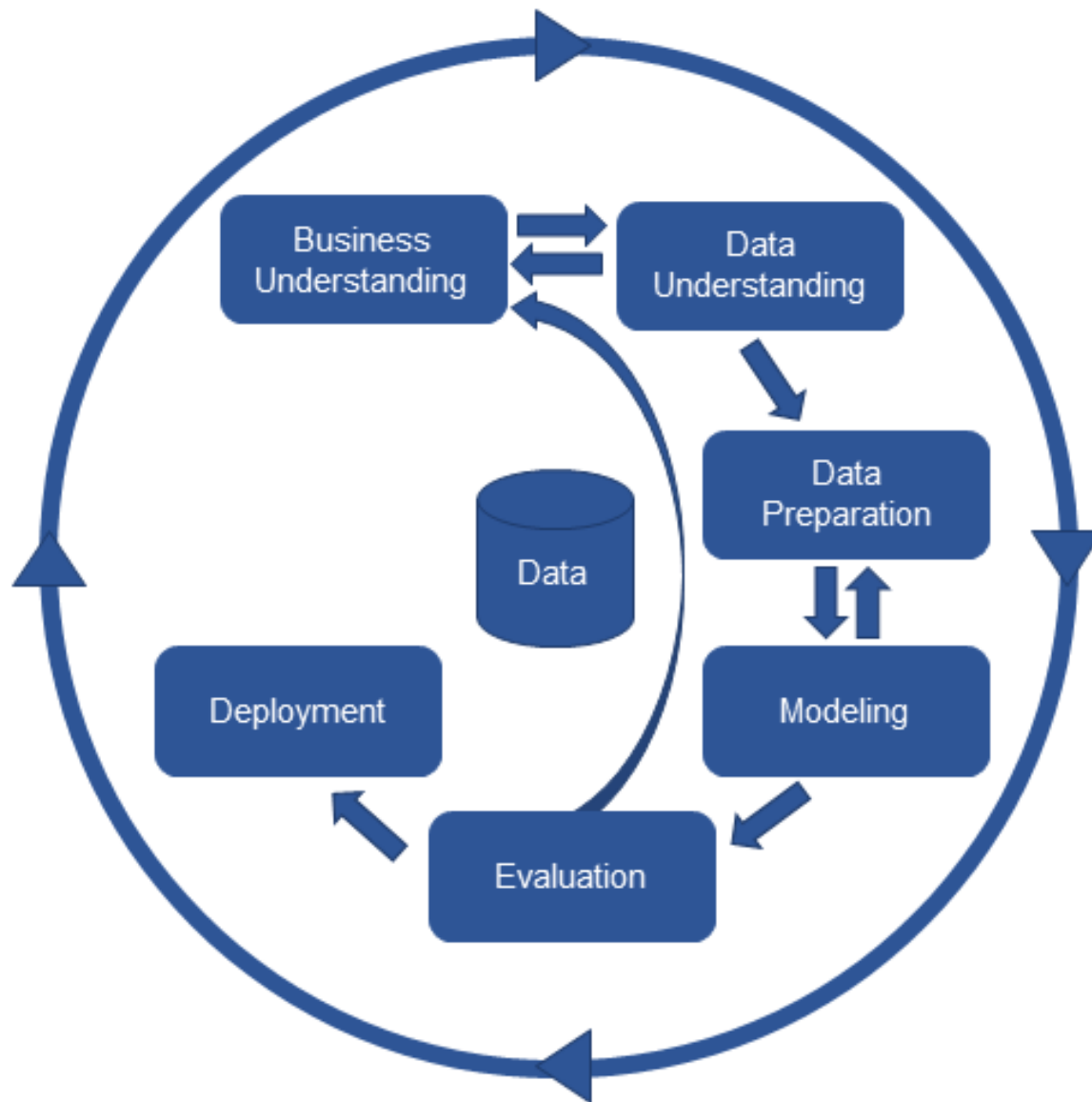
مدیریت منابع پزشکی

- ▶ بهینه‌سازی عملکرد بیمارستان: با تحلیل داده‌های مراجعه بیماران، زمان انتظار، فیلدهای خالی تختها و منابع انسانی، میتوان ظرفیتهای را به نحوی تنظیم کرد که کمترین هدررفت و بیشترین بهره‌وری حاصل شود.
- ▶ پیشبینی نیازهای منابع: مدل‌های پیشبینی داده‌های توانایی تخمین تقاضا برای داروها، تجهیزات پزشکی و نیروی انسانی در بازه‌های زمانی آینده را دارند که موجب برنامه‌ریزی دقیقتر و کاهش کمبود یا اضافی منابع میشود.
- ▶ شناسایی بیماران با ریسک بالا: داده‌های می‌تواند بیماران را بر اساس داده‌های پزشکی به گروه‌هایی تقسیم کند و بیمارانی که نیاز به مراقبتهای ویژه یا سریعتر دارند را شناسایی کند تا بهتر تخصیص منابع صورت گیرد.



مدیریت منابع پزشکی

- ▶ کاهش هزینه ها و بهبود خدمات: با تحلیل رفتار و الگوهای درمانی، داده کاوی به طراحی روش های درمانی مؤثرتر و شخصی سازی شده کمک میکند که هم کیفیت خدمات را افزایش و هم هزینه های مراقبتی را کاهش می دهد.
- ▶ مدیریت ارتباط با بیمار: تحلیل داده های تعاملی بیماران با سیستم های درمانی، مراکز تماس و بخش صدور صورتحساب به بهبود خدمات پس از درمان و رضایت بیماران کمک میکند.



16 روش های جمع آوری اطلاعات





مشکلات جمع آوری اطلاعات در پزشکی

► ۱. چالش یکپارچه سازی و استخراج داده ها

داده های مراقبت های بهداشتی از منابع با فرمت های مختلف مانند داده های ساختاریافته، کاغذ، فیلم، چند رسانه ای، تصاویر دیجیتال و غیره به دست می آیند که یکپارچه سازی و استخراج داده ها را یک چالش واقعی می سازد.

► ۲. داده های دائماً در حال تغییر:

بیماران و پزشکان ممکن است نقل مکان کنند، نام و حرفه خود را تغییر دهند، بازنشسته شوند. سازمان ها همچنین جابه جا می شوند، داروها و درمان جدید معرفی می شوند.

مشکلات جمع آوری اطلاعات در پزشکی



۳. حریم خصوصی و مقررات امنیتی: حفظ اعتماد بیماران، پایه و اساس ساخت یک سیستم پزشکی است.

۴. انتظارات بیمار: صنعت بهداشت و درمان باید درک درستی از نیازهای متغیر بیماران و ترجیحات آنها داشته باشد و سپس راه حل هایی همسو با روش زندگی آنها ارائه شوند.

۵. عدم وجود فرآیندهای تضمین کیفیت: چنین فرآیندی معمولاً وجود ندارد. کیفیت داده ها اغلب به شخصی بستگی دارد که داده ها را وارد می نماید



پیش پردازش داده

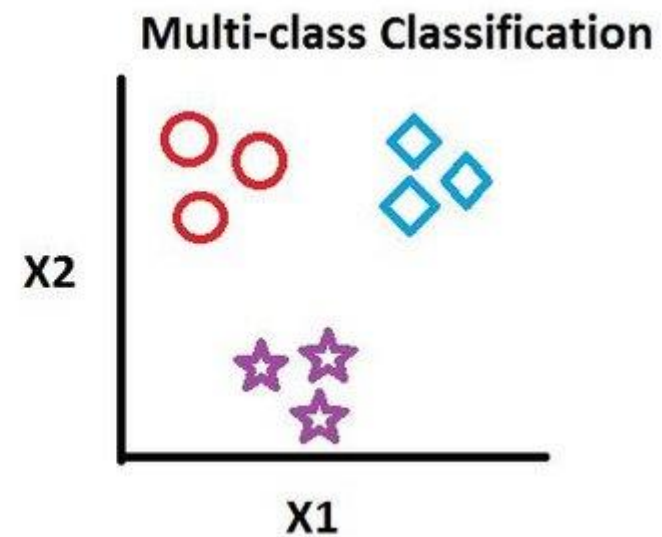
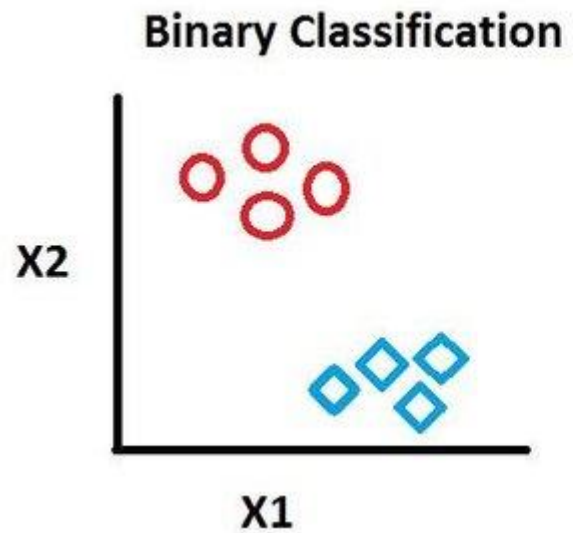
- انتخاب کلاس های هدف
- تکنیک های پاکسازی داده یا Data cleaning
- تکنیک های یکپارچگی داده یا Data integration
- تکنیک های کاهش داده یا Data reduction
- تکنیک های تبدیل داده یا Data transformations



انتخاب کلاس های هدف

Binary Class

Multi Class





پاکسازی داده

مدیریت داده های از دست رفته

مدیریت داده های پرت





پاکسازی داده - مدیریت داده های از دست رفته

حذف یک ردیف خاص - در این روش، یک ردیفی حذف می شود که بیش از ۷۵ درصد از مقادیر آن، نامشخص است یا وارد نشده است.

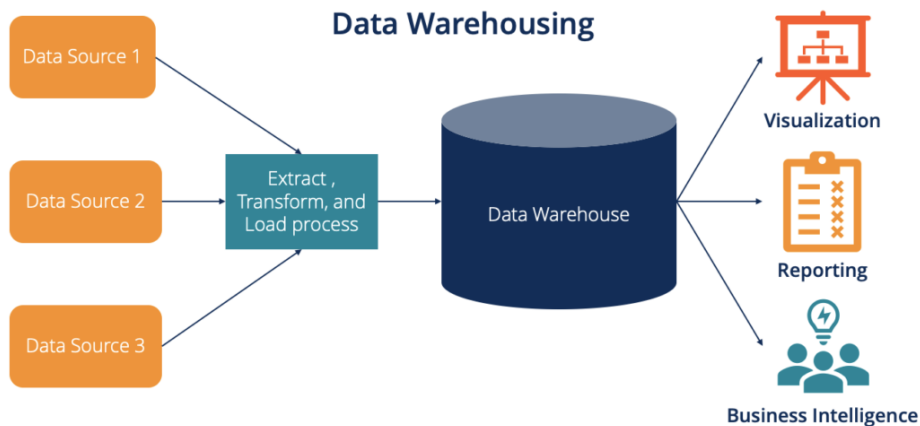
محاسبه میانگین - این روش برای ویژگی‌هایی که داده‌های عددی دارند مانند سن، حقوق، سال و غیره مفید است. در اینجا می‌توانید میانگین و میانه یک ستون حاوی مقادیر نامشخص محاسبه و جایگزین نمایید.



یکپارچگی داده

بسیاری از موارد ممکن است داده ها در فایل ها و منابع مختلف نگهداری شوند و در این صورت نیاز است تا داده ها پیش از اجرای تکنیک های داده کاوی با یکدیگر یکپارچه شوند.

استفاده از انباره داده توصیه می شود





تبدیل داده

فعالیت های مانند نرمال سازی داده ها و گسسته سازی داده

نرمال سازی داده ها : در مواقعی که فاصله ی اعداد زیاد باشد می توان با نرمال سازی محدوده ی کوتاه تری را انتخاب نمود نرمال سازی سن به بازه ی صفر تا یک

گسسته سازی داده ها : مثلا تبدیل مقادیر عددی فشارخون به سه سطح نرمال، پایین، بالا



کاهش ابعاد داده

دو رویکرد اصلی در کاهش ابعاد وجود دارد:

انتخاب ویژگی Feature Selection

در این روش، زیرمجموعه ای از ویژگیهای اصلی و مرتبط با مسئله انتخاب می شوند و ویژگی های غیرمفید یا تکراری حذف می شوند. این کار به کمک روش هایی مانند فیلتر (مثلاً حذف ویژگی های با واریانس کم یا همبستگی زیاد) انجام می شود. این روش ها کمک می کنند تا ویژگی های پراهمیت حفظ شوند و ابعاد کاهش یابد بدون اینکه ترکیب یا تغییر شکل داده ها انجام شود.



کاهش ابعاد داده

دو رویکرد اصلی در کاهش ابعاد وجود دارد:

استخراج ویژگی (Feature Extraction)

در این روش، ویژگیهای اصلی جدیدی که ترکیبی از ویژگی های اولیه هستند، ساخته می شوند. تکنیکهایی مانند تحلیل مولفههای اصلی (PCA) نمونه های رایج این روش به شمار می روند. این روش بیشتر برای داده های با ابعاد بالا کاربرد دارد و به قابل مصورسازی و تحلیل بهتر داده کمک می کند.



مشخصات یک مجموعه ی داده ای برای استفاده در داده کاوی

- (۱) داشتن حداقل تعداد رکوردهای لازم
- (۲) داشتن حداقل تعداد صفات
- (۳) میزان **missing value** خیلی زیاد نباشد
- (۴) صفت هدف تا حد امکان **missing value** نداشته باشد



مشکلات داده کاوی در حوزه سلامت

▶ حجم بسیار عظیم و ناهمگن داده ها: داده‌های پزشکی از منابع مختلفی مانند آزمایشها، پرونده های الکترونیکی سلامت، مکالمات پزشک و بیمار، دستگاه های پزشکی و ... جمع آوری می شوند که تنوع و حجم زیاد آنها، پردازش و یکپارچه سازی را دشوار می کند

▶ کیفیت پایین داده ها: داده های پزشکی ممکن است ناقص، نادرست، متناقض یا ناهمگن باشند که دقت مدل های داده کاوی را کاهش می دهد و مانع موفقیت تحلیل ها می شود.



مشکلات داده کاوی در حوزه سلامت

- ▶ حریم خصوصی و امنیت اطلاعات: داده های پزشکی شامل اطلاعات حساس بیماران هستند که حفظ حریم خصوصی و امنیت آنها به شدت اهمیت دارد. مدیریت دسترسی، رمزنگاری و سیاست های امنیتی از چالشهای اصلی در این زمینه اند.
- ▶ خطا و عدم قطعیت در داده ها و نتایج: تشخیص ها و درمان ها در پزشکی ممکن است اشتباه یا ناقص باشند و این موضوع موجب می شود که تحلیل داده ها نیازمند در نظر گرفتن خطاهای احتمالی و حساسیت ها باشد.
- ▶ ناهمگونی و استانداردهای مختلف داده ها: وجود فرمت ها و استانداردهای متنوع در سیستمهای اطلاعات بهداشتی، ترکیب داده ها را پیچیده می کند و توسعه الگوریتم های موثری را دشوار می سازد



انواع داده

Numerical ➤ عددی

Ordinal ➤ ترتیبی

Nominal ➤ اسمی



ابزارهای داده کاوی

- (۱) اکسل
- (۲) نرم افزار R
- (۳) نرم افزار ریپیدماینر
- (۴) نرم افزار وکا
- (۵) زبان برنامه نویسی پایتون



اکسل به عنوان ابزار داده کاوی

- ▶ نرم افزار اکسل یکی از مشهورترین و پرکاربردترین ابزارهای صفحه گسترده است که به دلیل قابلیت های گسترده اش، در حوزه داده کاوی نیز مورد استفاده قرار می گیرد. اکسل ابزاری ساده، در دسترس و کاربردی برای انجام تحلیل های اولیه داده ها و استخراج الگوهای نهفته در آن ها محسوب می شود.
- ▶ اکسل به کمک امکانات مختلف از جمله فرمول نویسی، جداول محوری، نمودارها، و افزونه های اختصاصی قابلیت انجام محاسبات آماری و تحلیل داده ها را دارد. این امکانات کمک می کنند تا بتوان به صورت پایه ای داده ها را پاک سازی، مرتب سازی، گروه بندی و تحلیل کرد.



اکسل به عنوان ابزار داده کاوی

- ▶ در زمینه داده کاوی، اکسل قابلیت‌هایی ذیل را دارد:
- ▶ تعریف مدل‌های تحلیلی و پیش‌بینی
- ▶ انجام تحلیل‌های توصیفی و آماری (میانگین، واریانس، آزمون‌های آماری)
- ▶ تشخیص الگوهای مهم در داده‌ها مانند تشخیص دسته‌بندی و تحلیل سبد خرید
- ▶ امکان استفاده از فرمول‌های پیچیده و توابع تخصصی مانند VLOOKUP
- ▶ استفاده از افزونه‌ها برای تحلیل پیشرفته‌تر را دارد که باعث می‌شود بتوان از آن برای داده کاوی در کسب و کارها و پروژه‌های مختلف استفاده کرد.



اکسل به عنوان ابزار داده کاوی

بنابراین، اکسل به عنوان یک ابزار داده کاوی پایه، به ویژه برای تحلیل داده های ساختاریافته و در حجم متوسط، استفاده می شود و با افزونه های موجود و ابزارهای پیشرفته تر می تواند تا حد قابل توجهی فراتر از محاسبات ساده عمل کند.



نقاط قوت اکسل در داده کاوی

- ▶ سهولت استفاده و دسترسی گسترده: اکسل به دلیل سادگی کار با آن و نصب پیش فرض روی اکثر سیستم‌های ویندوزی، برای کاربران با سطوح مختلف تجربه، بسیار قابل دسترس است.
- ▶ انعطاف پذیری بالا: امکان تطبیق اکسل با نیازهای خاص پروژه یا کسب و کار و قابلیت ساخت مدل‌های تحلیلی مختلف با فرمول نویسی پیچیده.
- ▶ ابزارهای تحلیل داده: جداول محوری، نمودارها، توابع آماری و محاسباتی مختلف، فرمت بندی شرطی برای تشخیص الگوها و پاکسازی اولیه داده‌ها.



نقاط قوت اکسل در داده کاوی

- ▶ امکان ارتباط و یکپارچگی با سایر نرم افزارها: توانایی اتصال به نرم افزارهای دیگر مایکروسافت آفیس و حتی پایگاه های داده SQL
- ▶ پیشرفته تر شدن با افزونه ها: افزونه هایی مثل Power Query، Power Pivot و قابلیت ماکرو و برنامه نویسی VBA برای خودکارسازی و تحلیل های پیچیده تر.
- ▶ گزارش گیری و مصورسازی: امکانات قوی برای ساخت نمودارهای مختلف و گزارش های گرافیکی قابل فهم برای ارائه نتایج.



نقاط ضعف اکسل در داده کاوی

- ▶ محدودیت در حجم داده‌ها: اکسل برای داده‌های حجیم و بسیار بزرگ مناسب نیست و پردازش داده‌های بزرگ با تاخیر و خطا مواجه می‌شود.
- ▶ عدم تخصصی بودن: اکسل ابزار تخصصی داده‌کاوی نیست و برای تحلیل‌های پیچیده‌تر باید به نرم‌افزارهای تخصصی دیگر مراجعه کرد.



نقاط ضعف اکسل در داده کاوی

- ▶ احتمال خطاهای انسانی: خطاهای فرمول نویسی و نگهداری داده‌ها در اکسل می‌تواند منجر به اشتباهات و نتایج نادرست شود که به خصوص در پروژه‌های مهم، ریسک بالایی دارد.
- ▶ پیچیدگی در کار با داده‌های نامنظم و فاقد ساختار: اکسل عمدتاً برای داده‌های ساختار یافته مناسب است و داده‌های نامنظم را به خوبی پشتیبانی نمی‌کند.



انواع تکنیک های داده کاوی

(۱) قواعد انجمنی :

تولید قواعد از الگوهای پرتکرار

(۲) خوشه بندی :

تشخیص مجموعه ای از دسته ها برای توصیف داده مثلا تشخیص گروه های مختلف در یک شبکه
یادگیری بدون نظارت

(۳) کلاسه بندی :

قرار دادن داده ها در یک دسته بندی از پیش تعریف شده
یادگیری با نظارت



انواع تکنیک های داده کاوی - قواعد انجمنی

Basket	Product 1	Product 2	Product 3
1	Milk	Cheese	
2	Milk	Apples	Cheese
3	Apples	Banana	
4	Milk	Cheese	
5	Apples	Banana	
6	Milk	Cheese	Banana
7	Milk	Cheese	
8	Cheese	Banana	
9	Cheese	Milk	

Itemset

{Milk,Cheese,Apples,Banana}

k-itemset : زیرمجموعه ی k عنصری از Itemset

فرض کنید زیر مجموعه ی مورد نظر : {Milk,Cheese}

فرکانس تکرار : در شش سبد خرید Milk,Cheese با هم رخ داده اند پس فرکانس تکرار آنها برابر با ۶ است (absolutesupport)

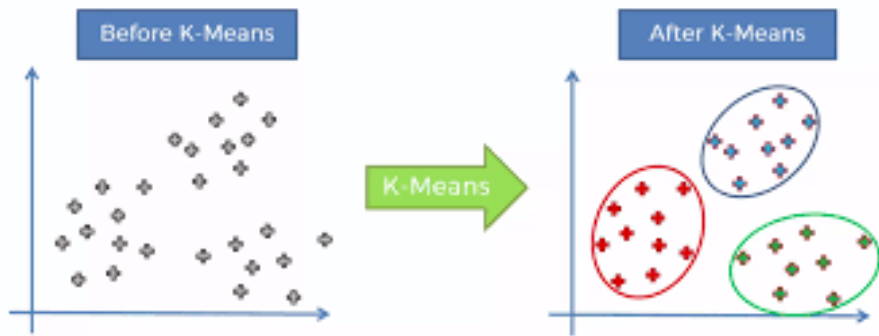
فرکانس تکرار نسبی : از ۹ سبد خرید، در ۶ سبد خرید Milk,Cheese با هم رخ داده اند پس فرکانس تکرار نسبی برابر با ۶۶ درصد است (relativesupport)

آستانه کمترین تکرار : برابر با ۵۰ درصد است (min_sup)

پس Milk,Cheese یک مجموعه ی پرتکرار است



انواع تکنیک های داده کاوی - خوشه بندی



خوشه بندی k-means

- (۱) k نقطه را صورت تصادفی به عنوان مرکز خوشه در نظر گرفته می شود برای این مثال می خواهیم سه خوشه داشته باشیم $K=3$
- (۲) فاصله هر نقطه با مرکز خوشه محاسبه شده و هر نقطه به نزدیک ترین مرکز خوشه اختصاص می یابد
- (۳) به مرحله ۱ بازگردید ، وقتی تغییری ایجاد نشد، متوقف شوید.



انواع تکنیک های داده کاوی - K - نزدیکترین همسایه

- ▶ الگوریتم k-نزدیکترین همسایه k-Nearest Neighbors یا KNN یک روش ساده و مشهور در یادگیری ماشین با ناظر است که برای مسائل طبقه‌بندی و رگرسیون استفاده می‌شود.
- ▶ یکی از مهم‌ترین کاربردها استفاده از الگوریتم KNN، دسته‌بندی مشتریان است. فروشگاه‌های اینترنتی به وسیله این الگوریتم، مشتریان خود را دسته‌بندی می‌کنند و محصولات متناسب با سبد خرید را به آنها پیشنهاد می‌دهند..



انواع تکنیک های داده کاوی - خوشه بندی - K - نزدیکترین همسایه

► نحوه کار:

- (۱) ابتدا مقدار K تعداد نزدیکترین همسایه ها تعیین می شود.
- (۲) سپس فاصله نمونه آزمایشی با هر نمونه آموزشی محاسبه می شود (فواصل مانند اقلیدسی).
- (۳) پس از مرتب سازی نمونه ها بر اساس فاصله، K نمونه نزدیکترین انتخاب می شوند.
- (۴) در مرحله آخر، بر اساس کلاس های K همسایه نزدیک، کلاس نمونه آزمایشی تعیین می شود. در طبقه بندی، این کار با استفاده از رأی اکثریت انجام می شود (یعنی کلاسی که بیشترین تعداد همسایه ها را دارد).



انواع تکنیک های داده کاوی - رگرسیون

رگرسیون یک تکنیک آماری و مدل سازی پیش بینی است که رابطه میان یک متغیر وابسته (هدف) و یک یا چند متغیر مستقل (ویژگی ها) را بررسی می کند. هدف آن یافتن مدل یا معادله ای است که بتواند مقدار متغیر وابسته را بر اساس مقادیر متغیرهای مستقل پیش بینی کند.

به طور ساده تر، رگرسیون به دنبال یافتن خط یا تابعی است که بهترین توضیح را برای داده ها ارائه دهد و بتوان از آن برای پیش بینی مقدار متغیر وابسته استفاده کرد. این رابطه معمولاً به شکل خطی در نظر گرفته می شود که به آن رگرسیون خطی گفته می شود، اما انواع دیگری مثل رگرسیون چند جمله ای و لجستیک هم وجود دارد.



انواع تکنیک های داده کاوی - رگرسیون خطی

▶ رگرسیون خطی یک روش آماری و مدل سازی است که رابطه ی خطی بین یک متغیر وابسته (هدف) و یک یا چند متغیر مستقل (ویژگی) را مدل می کند. هدف آن یافتن یک معادله ی خطی است که بهترین تناسب را با داده های موجود داشته و بتواند مقدار متغیر وابسته را بر اساس مقادیر متغیرهای مستقل پیش بینی کند..



انواع تکنیک های داده کاوی - رگرسیون خطی

- ▶ رگرسیون خطی ساده Simple Linear Regression: زمانی به یک الگوریتم رگرسیون خطی ساده گفته می شود که برای پیش بینی مقدار متغیر عددی وابسته از یک متغیر مستقل واحد استفاده شود.
- ▶ «رگرسیون خطی چندگانه» Multiple Linear regression: اگر برای پیش بینی مقدار متغیر وابسته عددی بیش از یک متغیر مستقل استفاده شود، به این مدل، رگرسیون خطی چندگانه گفته می گویند.



انواع تکنیک های داده کاوی - رگرسیون خطی

▶ مثال از رگرسیون خطی ساده:

فرض کنید معادله رگرسیون ما به این شکل باشد $y = 10x + 200$

در این معادله:

y قیمت خانه است

x متراژ خانه است

ضریب ۲ نشان می‌دهد که با افزایش هر متر مربع متراژ، قیمت خانه به طور متوسط ۱۰ واحد (مثلاً ۱۰ میلیون تومان) افزایش می‌یابد.

۲۰۰ قیمت پایه خانه.



انواع تکنیک های داده کاوی - آنالیز افتراق خطی

▶ آنالیز افتراق خطی که به انگلیسی **Linear Discriminant Analysis** یا **LDA** نامیده می شود، یک روش آماری و الگوریتم یادگیری ماشین برای دسته بندی و کاهش ابعاد داده ها است. این روش با هدف یافتن ترکیب خطی از ویژگی های ورودی به گونه ای که بیشترین تمایز (افتراق) را میان کلاس های مختلف داده ها ایجاد کند، به کار می رود.

▶ به طور دقیق تر، آنالیز افتراق خطی سعی می کند نسبت پراکندگی بین گروه ها (بین کلاس ها) را به پراکندگی درون گروه ها (داخل هر کلاس) بیشینه کند. این کار با یافتن محور یا ترکیب خطی میسر می شود که داده ها را به شکلی تبدیل کند که تفکیک کلاس ها حداکثر شود.



انواع تکنیک های داده کاوی - آنالیز افتراق خطی

تعیین اینکه آیا مصرف‌کنندگان جدید خریدی انجام خواهند داد یا خیر، آیا آنها به یک برند خاص وفادار خواهند بود یا خیر، آیا رویکرد فروش نرخ موفقیت ضعیف، متوسط یا قوی خواهد داشت یا نه، یا این که خریداران جدید کدام دسته را خواهند داشت. علاوه بر این، نشان می‌دهد که کدام یک از پیش‌بینی‌کننده‌ها متمایزترین هستند (بالاترین وزن تمایز)



ارزیابی تکنیک ها – 10-fold cross validation

قسمت اول، مجموعه‌ی داده‌ای آموزشی است که ۹۰ درصد مجموعه‌ی داده‌ای اصلی را تشکیل می‌دهد

قسمت دوم، مجموعه‌ی داده‌ای آزمایشی است که ۱۰ درصد مجموعه‌ی داده‌ای اصلی را تشکیل می‌دهد.

تقسیم به این دو مجموعه، ۱۰ بار انجام می‌شود و هر بار یک گروه به صورت تصادفی به عنوان مجموعه‌ی آزمایشی و بقیه‌ی گروه‌ها به عنوان مجموعه‌ی آموزشی در نظر گرفته می‌شوند. از مجموعه‌ی داده‌ای آموزشی برای تولید الگوها و از مجموعه‌ی داده‌ای آزمایشی برای ارزیابی صحت تکنیک‌های مورد استفاده در نظر گرفته می‌شود.



پارامترهای ارزیابی عملکرد یک کلاسه بند - تعاریف براساس کلاسه بندی دودویی

Accuracy

نسبت نمونه‌های درست پیش بینی شده به کل نمونه‌های پیش بینی شده

Sensitivity

حساسیت نسبتی از نمونه‌های مثبت است که مدل آن‌ها را به درستی به عنوان نمونه مثبت تشخیص داده است.

Specificity

نسبتی از نمونه‌های منفی است که مدل آن‌ها را به درستی به عنوان نمونه منفی تشخیص داده است



پارامترهای ارزیابی عملکرد یک کلاسه بند - تعاریف براساس کلاسه بندی دودویی

AUC

ROC : منحنی‌های **ROC** منحنی‌های دو بعدی هستند که در آنها **Sensitivity** در محور **Y** و **1-Specificity** در محور **X** رسم می‌شوند.

وقتی مقدار ناحیه‌ی تحت منحنی **ROC** برابر ۱ باشد، دسته‌بندی می‌تواند بین تمام نقاط کلاس مثبت و منفی به طور صحیح تمایز قائل شود. اما، اگر مقدار ناحیه‌ی تحت منحنی **ROC** برابر ۰ باشد، دسته‌بندی همه منفی‌ها را مثبت و همه مثبت‌ها را منفی پیش‌بینی می‌کند.

AUC سطح ناحیه‌ی تحت منحنی **ROC** را نشان می‌دهد.



معرفی کلان داده

کلان داده اصطلاحی است که برای حجم زیادی از داده‌ها استفاده می‌شود که خیلی سریع و پیچیده هستند و پردازش آنها با استفاده از روش‌های سنتی، مشکل است.

کلان‌داده از ترابایت شروع شود و می‌تواند از پتابایت هم عبور نماید



معرفی کلان داده

کلان داده در سه نوع می تواند باشد:

- (۱) بدون ساختار مانند داده های متنی
- (۲) نیمه ساخت یافته: داده های نیمه ساختاریافته شامل هر دو داده ساختاریافته و بدون ساختار است.
- (۳) ساخت یافته مانند داده های جدولی

کلان داده



کلان داده یا سه V :

حجم یا Volume :

اولین ویژگی بیگ دیتا حجم و مقدار آن است که با کلمه **Volume** شناخته می شود. سازمان ها و نهادهای گوناگون با جمع آوری اطلاعات از منابع متنوع و متفاوت اقدام به ایجاد یک پایگاه داده می کنند. این اطلاعات می تواند شامل مواردی مثل اطلاعات مربوط به حجم معاملات و تبادلات تجاری، اطلاعات دریافت شده از شبکه های اجتماعی یا اطلاعات ماشینی و طبقه بندی شده دیگر باشد حجم زیاد داده ها، چالش هایی را ایجاد می نماید که نیاز به سیستم های جدید برای مدیریت این نوع داده ها مطرح می شود

کلان داده



کلان داده یا سه V :

سرعت یا Velocity

دومین ویژگی مهم کلان داده بحث سرعت فرآیندهای آن است. مبادلات اطلاعات باید با سرعت بسیار بالایی انجام شود
تولید داده‌ها با سرعتی بی سابقه انجام می‌شود و باید به موقع پردازش شود.

کلان داده



کلان داده یا سه V :

تنوع یا Variety :

فرمت و نوع اطلاعات با یکدیگر تفاوت دارند. این اطلاعات می توانند به صورت متنی، تصویری، صوتی یا حتی کدهای صفر و یک باشند! این تنوع باعث شده تا دسته بندیهای متعددی برای جمع آوری و ذخیره سازی اطلاعات در نظر گرفته شود.

نقش داده کاوی در تحلیل کلان داده



نقش داده کاوی در تحلیل کلان داده بسیار کلیدی و بنیادی است؛ به طوری که داده کاوی را می‌توان فرایند تحلیل و استخراج الگوها، روابط و بینش‌های پنهان از مجموعه‌های بزرگ و پیچیده داده (کلان داده) تعریف کرد.

داده کاوی به عنوان فرآیندی پیشرفته در تحلیل کلان داده، با استفاده از الگوریتم‌ها و مدل‌های آماری، یادگیری ماشین و تکنیک‌های متنوع، به کشف الگوهای پنهان و روابط مهم در این داده‌ها کمک می‌کند. این الگوها و دانش استخراج شده می‌تواند در پیش‌بینی روندها، تصمیم‌گیری‌های بهتر تجاری و بهینه‌سازی فرآیندها نقش داشته باشد.

در واقع، داده کاوی به سازمان‌ها کمک می‌کند تا حجم عظیمی از داده‌های خام کلان داده را به اطلاعات معنی‌دار و قابل استفاده تبدیل کنند که برای بهبود استراتژی‌ها، افزایش کارایی، شناخت رفتار مشتریان، پیش‌بینی بازار و شناسایی ریسک‌ها حیاتی است.

آینده داده کاوی



▶ انتظار می رود تا سالهای آینده داده کاوی نقش کلیدی در بهبود نتایج درمانی، افزایش بهره وری سیستمهای بهداشتی، توسعه درمان های شخصی سازی شده و ارتقای سلامت عمومی ایفا کند.

▶ آینده داده کاوی با سیستمهای هوشمند و بلادرنگ و افزایش میزان خودکارسازی فرآیند مشخص می شود. این آینده با پیشرفت در هوش مصنوعی، یادگیری ماشین و رشد نمایی داده ها شکل خواهد گرفت.

.

.



مهمترین بخش هایی که در آینده داده کاوی در آنها موثر است

- ▶ تشخیص زودهنگام و پیش بینی بیماریها: داده کاوی امکان شناسایی الگوهای مخفی در داده های سلامت را فراهم کرده که منجر به تشخیص سریعتر بیماری ها، به ویژه بیماریهای خاموش مانند سرطان و دیابت می شود. همچنین مدل های پیش بینی می توانند احتمال بازگشت بیماری یا عوارض جانبی را تعیین کنند.
- ▶ شخصی سازی درمان: با بهره گیری از داده های ژنتیکی، بالینی و رفتاری بیماران، پروتکلهای درمانی متناسب با هر فرد طراحی میشود که این امر منجر به افزایش اثربخشی درمان و کاهش عوارض می گردد.

مهمترین بخش هایی که در آینده داده کاوی در آنها موثر است



▶ یکپارچه سازی داده ها و فناوری های نوین: با پیشرفت های فناوری مانند بلاکچین، پرونده های سلامت دیجیتال با امنیت و شفافیت بالا مدیریت می شوند و دسترسی سریع و مطمئن به داده ها برای تصمیم گیری های درمانی فراهم می گردد.

▶ افزایش بهره وری و بهینه سازی منابع: تحلیل داده ها به بهینه سازی عملکرد بیمارستانها، مدیریت بهتر منابع انسانی و تجهیزات کمک می کند و هزینه های مراقبتی را کاهش می دهد