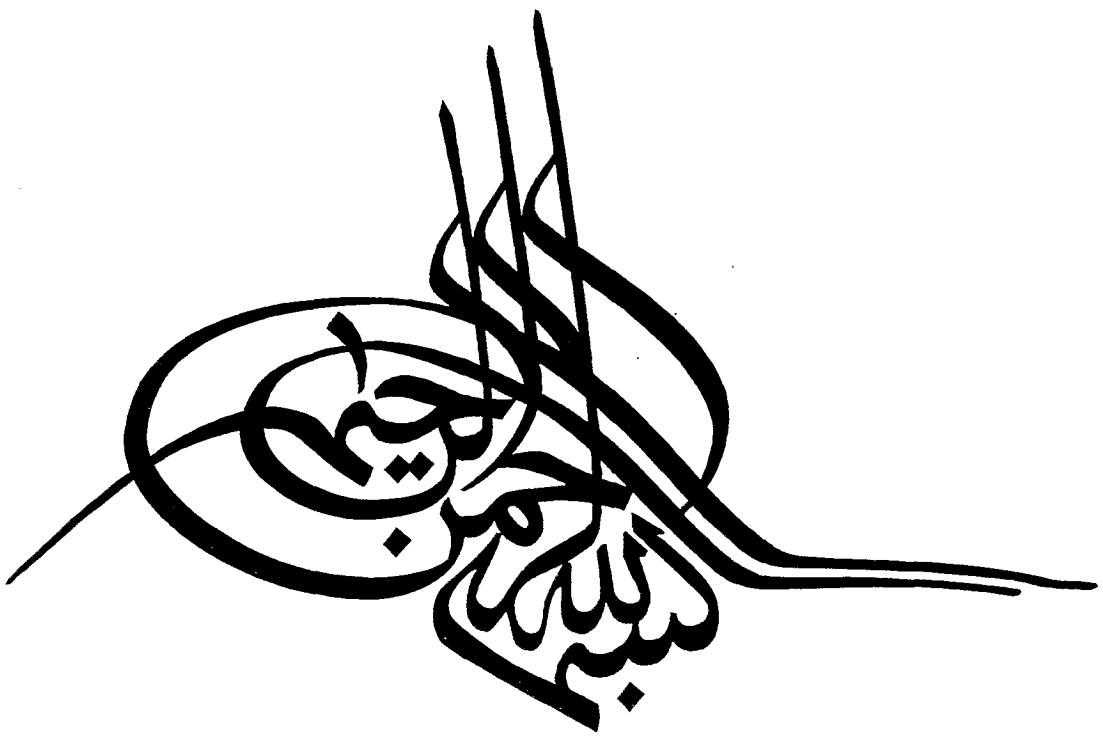

وزارت علوم، تحقیقات و فناوری

دانشگاه تفرش

دانشکده مهندسی صنایع

پایان نامه کارشناسی ارشد

کاربرد داده‌کاوی در مدیریت منابع انسانی

استاد راهنما :

آقای دکتر بهزاد اشجری

استاد مشاور:

آقای دکتر حمیدرضا گلمکانی

دانشجو :

محمود محمدی

۱۳۸۷

تاریخ: ۱۳۸۸/۲۰/...
شماره: ۶۵۴۹/۲۱۵.....
پیوست: .....

دانشگاه تفرش

مدیریت تحصیلات تکمیلی

صورتجلسه دفاعیه پایان نامه کارشناسی ارشد

نام و نام خانوادگی: محمود محمدی شماره دانشجویی: ۸۵۴۱۲۱۰۰۷ گروه: مهندسی صنایع

رشته تحصیلی/گرایش: مهندسی صنایع / مدیریت سیستم و بهره وری

عنوان پروژه: کاربرد داده کاوی در مدیریت منابع انسانی

تعداد واحد: ۶ تاریخ تصویب: ۸۶/۷/۲۴ تاریخ دفاع: ۸۷/۱۱/۳۰

نمره نهایی: به عدد: ۱۷/- به حروف: هفده

هیات داوران	نام و نام خانوادگی	رتبه علمی	امضاء
استاد راهنما	دکتر بهزاد اشجری	استادیار	
استاد مشاور	دکتر حمید رضا گلمکانی	استادیار	
داور داخلی	دکتر ابراهیم شریفی	استادیار	
داور خارجی	دکتر سید مصطفی ترابی	پژوهشگر	
نماینده تحصیلات تکمیلی	حمید رضا کلانتری	-	

مدیرگروه: دکتر بهزاد اشجری

امضاء:

تاریخ:

مهر: دانشگاه تفرش دانشکده مهندسی صنایع و ساخت

مدیر تحصیلات تکمیلی دانشگاه: دکتر حمید رضا صبا

امضاء: صبا

تاریخ: ۸۸/۲/۱۴

مهر: دانشگاه تفرش اداره کل تحصیلات تکمیلی

دانشگاه تفرش اداره کل تحصیلات تکمیلی

تقدیم به آزادگان

در اینجا برخود لازم میدانم خداوند باری تعالی را حمد و سپاس گویم و از راهنمایی‌ها و حمایت‌های استاد راهنمای ارجمند جناب آقای دکتر اشجری مراتب کمال امتنان را به‌جای آورم .

از آنجاییکه تحصیل در کنار کار بدون همراهی همکاران و مسؤولین مافوق میسر نیست لذا در همین مجال از همکاران و مسؤولین سازمان مدیریت و برنامه‌ریزی سابق استان زنجان و شرکت برق منطقه‌ای زنجان سپاسگزارم .

همچنین از تمامی بزرگوارانی که مرا در این مسیر حامی بودند به ویژه مادر عزیزم قدردانی می‌نمایم .

چکیده

یک درک عمیق از دانش پنهان در داده‌های منابع انسانی برای موقعیت رقابتی سازمان وتصمیم‌سازی سازمان حیاتی است. تحلیل الگوها و روابط داده‌های منابع انسانی کاملاً نادر و کمیاب است. اغلب داده‌های منابع انسانی صرفاً برای اهداف گزارش‌مانند و پاسخ به یک سری پرس‌وجوها جمع‌آوری و نگهداری می‌شوند . لازم است که سیستمهای منابع انسانی بر روی داده‌های کمیت‌پذیر بیشتر تمرکز کنند. ما نشان خواهیم داد چگونه داده‌کاوی الگوهای مفیدی از مجموعه‌های داده کشف و استخراج می‌کند ودر نتیجه قابلیتهای داده‌کاوی منجر به افزایش کارآیی و در نهایت تحکیم مزیت رقابتی می‌شود . این پایان نامه بر روی تشریک مساعی داده‌کاوی و مدیریت منابع انسانی در سازمان بحث وبررسی انجام می‌دهد.

در ابتدا مروری بر داده‌کاوی و تکنیکهای آن خواهیم داشت . سپس اشاره‌ای به تفاوت مابین داده‌کاوی و مطالعات معمول داده‌کاوی می‌کنیم . ارزش اطلاعات منابع انسانی درجهت حفظ موقعیت رقابتی بنگاهها و تصمیم‌گیری سازمانی لحاظ شده است . نمونه کاربردی از تکنیکهای داده‌کاوی (تکنیک درخت تصمیم) در سیستم مدیریت منابع انسانی و تصمیم‌گیری مدیران با مثال تشریح خواهد شد .

کلید واژه ها: داده کاوی، سیستم مدیریت منابع انسانی، تصمیم گیری، الگو ، درخت تصمیم

فهرست مطالب

پیشگفتار

برای رقابت کردن در محیط امروزی سازمانها می‌بایست قادر باشند تا بطور هوشمندانه‌ای از دانشهایی که هم اکنون در بطنشان وجود دارد استفاده کنند . امروزه بنگاهها دریافته‌اند که مهمترین منابع سازمانشان همان مردمان آنها هستند. آنها تشخیص داده‌اند که مهارت ، قابلیت انطباق ، دانش و وفاداری کارکنان است که بنگاههای برتر را از بقیه متمایز می‌کند[۲]. افزایش‌نمایی اطلاعات که مرهون ثبت الکترونیکی داده‌هاست وذخیره‌سازی آنها در انبار های داده بزرگ، نیازی راخلق کرده است که درآن تحلیل حجم وسیعی از داده های تولید شده بوسیله‌ی سازمانهای معاصر مورد نظر است و درپی آن بنگاه می بایست بتواند با سرعت عکس‌العمل نشان دهد.

داده‌های منابع انسانی که در دسترس سازمانها می باشند دارای پتانسیلی جهت کمک به پشتیبانی فرآیندهای تصمیم‌گیری هستند. شناسایی اطلاعات مفید در میان پایگاه داده وسیع منابع انسانی که منجر به اتوماسیون فرآیندهای مرتبط با منابع انسانی شود، یکی از چالشهای این زمینه کاری است.

داده‌کاوی[۱] که بطور عامیانه کشف دانش درداده‌های زیاد نامیده می‌شود، بنگاهها و سازمانها را قادر می‌سازد تا بوسیله جمع‌آوری، ذخیره، تحلیل و دسترسی به داده‌های هماهنگ تصمیمات حساب‌شده‌ای بگیرند. در این مسیر از ابزارهای متنوعی مانند ابزارهای پرس‌وجو و گزارش‌گیری، ابزارهای آنالیزگر و تحلیل‌کننده و ابزارهای پشتیبانی‌کننده تصمیم استفاده می‌کنند.

داده‌کاوی بعنوان تحلیل‌گر داده‌های پایگاه داده‌های خیلی بزرگ، ابزار مفیدی برای حرفه‌ای‌های منابع انسانی است. داده‌کاوی شامل استخراج دانش بر مبنای الگوهای برگرفته از داده‌های پایگاه داده است. سازمانهایی که هزاران نفر را استخدام می‌کنند و در پی اطلاعات مستخدمینشان هستند ممکن است الگوهای اطلاعاتی ارزشمندی بیابند که فراهم‌کننده چشم انداز مناسبی جهت نگهداری وجبران خدمات کارکنانشان باشد.

اطلاعات مورد نیاز برای مدیریت منابع انسانی در بسیاری از سازمانها اغلب یا وجود ندارد یا در دسترس نیست، خوشبختانه به برکت داده‌کاوی ، مدیران ارشد اجرایی می توانند دارائی سازمان_کارکنان_را مدیریت نمایند.

تفکر و نقش سیستمهای مدیریت منابع انسانی (HRMS) و در پی آن سیستمهای اطلاعاتی منابع انسانی به موازات توسعه تکنولوژیکی در حال افزایش است . شرکتهای پیشرو این مهم را دریافته‌اند که در دنیای "بمیر یا دانش بیآفرین "، نقش سیستمهای اطلاعاتی منابع انسانی فقط به سیستم پرداخت حقوق و دستمزد محدود نمی‌شود. اغلب سیستمهای اطلاعاتی منابع انسانی زمینه‌های ذخیره‌سازی داده ، فرآیند تعامل (اتوماسیون) و سیستمهای اطلاعاتی مدیریتی (MIS) (جهت تبدیل داده های خام به داده‌های معنی دار) را فراهم می آورند[۱].

در دنیای رقابتی امروز فقدان یک کارمند یا عدم ایفای صحیح نقش عده‌ای ازکارکنان می تواند تفاوت و مزه موفقیت وشکست را آشکار کند. با این تفاسیر دیگر جای تعجب نیست که مدیران ارشد به اطلاعات بیشترو بیشتر از

---

[1] :DataMining

بخش منابع انسانی سازمانشان نیاز داشته باشند. آنها طالب داده- داده‌هایی محکم از توزیع ورود و خروج، استخدام، مقایسه بین حقوق، مزایا و پاداشها ، قومیت و جنسیت _بوده‌اند و هستند. آنها می‌خواهند تا روندها ، تغییرات محتمل در محیط کاری و حتی محیط بیرونی را بدانند.

این‌نوشتار سعی در آن دارد به تکنیکهایی از داده‌کاوی مراجعه کند که داده‌های منابع انسانی را به اطلاعات مفید تبدیل می‌کنند. هدف ما اثبات توانایی داده‌کاوی درارتقاء کیفیت فرآیند تصمیم‌گیری در سیستم مدیریت منابع انسانی است.

در فصل اول ضمن تعریف داده‌کاوی به کلیاتی از آن اشاره خواهیم کرد. در فصل دوم مروری اجمالی به تکنیکهای داده‌کاوی خواهیم داشت و اهم تکنیکها در مثال کاربردی تشریح خواهند شد. در فصل سوم تکنیک درخت تصمیم مورد بررسی قرار خواهد گرفت . در فصل چهارم جنبه‌هایی از مدیریت منابع انسانی که مرتبط با داده‌کاوی است مورد بحث قرار خواهد گرفت و در فصل پنجم تکنیک درخت تصمیم به یک مثال کاربردی اعمال خواهد شد .

انگیزه و محرک اصلی از انتخاب این موضوع برای پایان‌نامه این است‌که تکنیکها و کاربردهای داده‌کاوی تا کنون زیاد مورد توجه مدیران منابع انسانی قرار نگرفته است و با مراجعه به پایان‌نامه‌های کارشناسی ارشد و مقالات ارائه شده دراولین و دومین کنفرانس داده‌کاوی ایران در دوسال اخیر متوجه می شویم که این موضوع در داخل کشور و با کمی احتیاط در دنیا نسبتاً دست نخورده و بکر است.

فصل اول

داده کاوی چیست ؟

از هنگامی که رایانه در تحلیل و ذخیره سازی داده‌ها به کار رفت (۱۹۵۰) حدود ۲۰ سال حجم داده‌ها بصورت خطی افزایش پیدا می‌کرد ولی پس از گذشت دو دهه و همزمان با پیشرفت فن آوری اطلاعات (IT) هر سال یکبار حجم داده‌ها، دو برابر شد. همچنین تعداد پایگاه داده‌ها با سرعت بیشتری رشد نمود .

این در حالی است که تعداد متخصصین تحلیل داده‌ها وآمارشناسان با این سرعت رشد نکرد .حتی اگر چنین اتفاقی می‌افتاد بسیاری از پایگاه داده‌ها چنان گسترش یافته بودند که شامل چند صدمیلیون یا چند صد میلیارد رکورد ثبت شده می‌شدند و امکان تحلیل و استخراج اطلاعات با روش های معمول آماری از دل انبوه داده‌ها مستلزم چند روز کار با رایانه‌های موجود بود.

حال با وجود سیستم‌های یکپارچه اطلاعاتی ، سیستم‌های یکپارچه بانکی و تجارت الکترونیک ، لحظه به لحظه به حجم داده‌ها در پایگاه داده‌های مربوط اضافه شده و باعث به وجود آمدن انبارهای (توده‌های) عظیمی از داده‌ها شده به طوری که ضرورت و کشف و استخراج سریع و دقیق دانش از این پایگاه داده‌ها را بیش از پیش نمایان کرده است . (چنانکه در عصر حاضر گفته می‌شود ،اطلاعات طلاست.)

هم اکنون در هر کشور ، سازمانها ، شرکتها و ... برای امور بازرگانی ، پرسنلی ، آموزشی ، آماری و ... پایگاه داده‌هایی ایجاد یا خریداری شده است . بطوریکه این پایگاه داده‌ها برای مدیران ، برنامه ریزان ، پژوهشگران و ... جهت تصمیم‌گیری‌های راهبردی ، تهیه گزارشهای مختلف ، توصیف وضعیت جاری می‌تواند مفید باشد .داده‌کاوی یا استخراج وکشف سریع و دقیق اطلاعات باارزش و پنهان از این پایگاه داده‌ها از جمله اموری است که هر کشور ، سازمان و شرکتی به منظور توسعه علمی ، فنی و اقتصادی خود به آن نیاز دارد .

متدهای آنالیز قدیمی که اغلب شامل شبکه‌های عصبی و آمار توصیفی آنها هستند ؛کند ، گران و بسیار موضوعی‌اند. بعنوان مثال اگر یک متخصص منابع انسانی مایل به تحلیل هزینه تغییر در یک زمینه از منابع انسانی باشد ، ممکن است مجبور شود داده‌هایش را از چندین منبع متفاوت استخراج کند ؛ منابعی از قبیل : رکوردهای مالی ، گزارشهای خاتمه خدمت ، رکوردهای پرسنل . سپس آنها را دوباره مخلوط و ارزیابی کند . این فرآیند فرصتهای زیادی را برای خطا ایجاد می‌کند. به محض اینکه سایز پایگاه داده‌ها رشد می‌کند، ابزارهای قدیمی بیشتر غیرقابل اعمال می‌شوند .داده‌کاوی فرآیند استخراج اطلاعات از مجموعه داده‌های واقعاً بزرگ با استفاده از الگوریتم ها و تکنیک های بر گرفته از زمینه های آماری ، فراگیری ماشینی[1] و مدیریت پایگاه داده می باشد.[۳]

[1] : Machine learning

داده‌کاوی در بسیاری از نواحی عملیاتی مانند مالی و بازاریابی مورد استفاده قرار گرفته است .کاربرد منابع انسانی هنوز یک فرصت دست نخورده برای اعمال تکنیکهای داده‌کاوی می باشد . سیستمهای اطلاعاتی منابع انسانی[۲] نوعاً حجم عظیمی از داده‌ها را (که مورد نیاز داده‌کاوی است) نگه می دارد . این نکته قابل ذکر است که مانعی در برخورد داده‌کاوی با پایگاه داده‌های کوچک وجود ندارد .

سابقه داده‌کاوی

داده‌کاوی و کشف دانش در پایگاه داده‌ها از جمله موضوع‌هایی هستند که همزمان با ایجاد و استفاده از پایگاه داده‌ها در اوایل دهه ۸۰ میلادی برای جستجوی دانش در داده‌ها شکل گرفت .شاید بتوان لوول (۱۹۸۳) را اولین شخصی دانست که گزارشی تحت عنوان شبیه سازی فعالیت داده‌کاوی در مورد داده‌کاوی ارائه نمود .همزمان با او پژوهشگران و متخصصان علوم رایانه ، آمار ، هوش مصنوعی ، فراگیری ماشینی و ... نیز به پژوهش در این زمینه و زمینه های مرتبط پرداختند . پژوهش جدی روی موضوع داده‌کاوی از اوایل دهه ۹۰ میلادی شروع شد . پژوهش ها و مطالعات زیادی در این زمینه صورت گرفته ، همچنین سمینارها ، دوره های آموزشی و کنفرانس‌هایی نیز برگزار شده است .در برخی از این مقالات پایه های نظری داده‌کاوی آورده شده است. بعنوان مثال در سال ۱۹۹۱، Piatetsky & Shapiro استقلال آماری قاعده‌ها در داده‌کاوی برای بانکهای اطلاعاتی بررسی نموده‌اند .در سال ۱۹۹۵ هافمن و نش استفاده از داده‌کاوی و انبار داده‌ها برای بانکهای آمریکا را بررسی نمود وبیان کردند چگونه این سیستمها برای بانکهای آمریکا قدرت رقابتی بیشتری ایجاد می‌کنند .

در مقاله دیگری توسط چت فیلد مشکلات ایجاد شده توسط داده‌کاوی مورد بحث قرار گرفت و همچنین مقاله ای تحت عنوان مدلهای خطی غیر دقیق داده‌کاوی و استنباط آماری ارائه نمود .

هندری نیز دیدگاه اقتصادسنجی روی داده‌کاوی را تهیه کرد .در این سال انجمن داده‌کاوی همزمان با اولین کنفرانس بین‌المللی "کشف دانش و داده‌کاوی" شروع به کار کرد این کنفرانس پیرو و محصول چهار دوره آموشی بین المللی در پایگاههای داده در سال های ۱۹۸۹تا ۱۹۹۴بود . انجمن مذکور ، یک سازمان علمی به نام ACM-SIGKDD را ایجاد نمود .در سال۱۹۹۷، Mannila خلاصه‌ای از مطالعه روی اساس داده‌کاوی ارائه نمود . در سال ۱۹۹۸ Hand مقاله ای تحت عنوان داده‌کاوی: آمار یا بیشتر ؟ را ارائه نمود .در این سال نیز کنفرانسهای بین المللی و ناحیه ای در مورد داده‌کاوی برگزار شد که ازجمله می توان به کنفرانس آسیا و اقیانوسیه درباره کشف دانش و داده‌کاوی اشاره کرد [۵] . در ایران نیزطی سالهای ۱۳۸۶ ، ۱۳۸۷ اولین و دومین کنفرانس داده کاوی در دانشگاه صنعتی امیرکبیر برگزار شدند .

مفهوم داده‌کاوی

از لحاظ لغت شناسی داده‌کاوی فرآیند کشف الگوهای سودمند در داده‌هاست . الگو در فرهنگ لغات و بستر ، شکل گیری یک طبیعت یا احتمال است .

داده‌کاوی یک مرحله از فرآیند اکتشاف دانش مشتمل بر الگوریتمها (متد های) ویژه است که در راستای اهدافی قابل قبول الگوهای (مدلهای) خاصی روی داده‌هایمان تولید می‌کند[۷] . فرآیند اکتشاف دانش ،فرآیند به کار گیری متدها و الگوریتم‌های داده‌کاوی جهت تشخیص دانش ساری و جاری در چارچوب پایگاه داده با اعمال پیش پردازش یا تبدیل ها مورد نیاز است [۷] .

---

[2] : Human Resource Information Systems

امروزه واژه داده‌کاوی بیان کننده تکنولوژی است که تکنیکهای آماری را در تعامل با فرمولهای ریاضی جهت یافتن روابط مهم ما بین متغیر هادر داده‌های تاریخی به خدمت می‌گیرد [۶] .

داده‌کاوی به بررسی ، تجزیه و تحلیل مقادیر عظیمی از داده‌ها به منظور کشف الگوها و قوانین معنی‌دار اطلاق می‌شود .داده‌کاوی عمدتاً با ساختن مدلها مرتبط است یک مدل اساساً به الگوریتم یا مجموعه ای از قوانین گفته می‌شود که مجموعه ورودی ها را (معمولا به شکل زمینه هایی در پایگاه داده‌های شرکت) با هدف یا مقصد خاصی مرتبط می نماید.[۴]

در بسیاری از متون عبارت داده‌کاوی همتراز و گاهاً مترادف با یکی از عبارت های استخراج دانش ، برداشت اطلاعات ، وارسی داده‌ها وحتی لایروبی کردن داده‌هاست که در حقیقت کشف دانش در پایگاه داده‌ها (Knowledge Discovery of Database[۳]) را توصیف می کند، بنابر این ایده ای که مبنای داده‌کاوی است یک فرآیند با اهمیت از شناخت الگو های بالقوه مفید ، تازه و در نهایت قابل درک در داده‌هاست . واژه کشف دانش در پایگاه داده‌ها در اوایل دهه ۸۰ میلادی در مراجعه به مفهوم کلی گسترده و سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است اصطلاح داده‌کاوی را آماری ها ، تحلیل گران داده‌ها و انجمن سیستم های اطلاعات مدیریت به کار برده‌اند در حالی که پژوهشگران فراگیری ماشین و هوش‌مصنوعی از KDD بیشتر استفاده می‌کنند.

سه تعریف معتبر دیگر داده‌کاوی عبارتند از :

"داده‌کاوی یا به تعبیر دیگر کشف دانش در پایگاه داده‌ها استخراج غیر بدیهی اطلاعات بالقوه مفید از روی داده‌هایی است که قبلا ناشناخته مانده اند. "[۸]

"داده‌کاوی در حقیقت کشف ساختارهای جالب توجه ، غیر منتظره و با ارزش از داخل مجموعه وسیعی از داده‌ها می باشد و فعالیتی است که اساساً با آمار و تحلیل دقیق داده‌ها منطبق است ".[۹]

داده‌کاوی فرآیند کشف رابطه ها ، الگو ها و روند های جدید معنی داریست که به بررسی حجم وسیعی از اطلاعات ذخیره شده در انبار های داده با فن‌آوری های تشخیص الگو (مانند ریاضی و آمار) می پردازد .[۱۰]

با این تعاریف می توان گفت ؛ ایده اصلی داده‌کاوی عبارت است از :داده‌های قدیمی حاوی اطلاعات هستند که در آینده مفید واقع می‌شوند .حجم بیش از اندازه داده‌ها ؛ تعداد زیاد رکورد ها ($۱۰^{۸}$ الی $۱۰^{۱۲}$ بایت) و داده‌های چند بعدی (آرایه های ۱۰۰ تا ۱۰۰۰۰ بعدی) یک مواجهه بدیع است .چگونه می توان میلیونها رکورد ، دهها یا صدها فیلد را برای یافتن الگوها کاوش کرد ؟ این مسائل باعث می شد تنها قسمت کوچکی (نوعا ۵ الی ۱۰درصد) داده‌های جمع آوری شده آنالیز شوند .

در اینصورت است که ممکن است داده‌هایی هرگز آنالیز نشوند داده‌هایی که با هزینه گران جمع آوری شده اند. در اینجاست که نیاز به داده‌کاوی احساس می‌شود.

داده‌کاوی در برابر تحلیلهای سنتی داده‌ها :

جا دارد در این قسمت رابطه و اختلاف بین داده‌کاوی و فعالیت های آشنا تری از قبیل گزارشات ، پرس و جو ها و مطالعات آماری روتین بکار رفته در یک پایگاه داده مورد بررسی قرار گیرد .پرس وجو ها و گزارشها (گزارشها معمولا به‌دنبال پرس و جو ها دیده می‌شوند .) سوالات ساختار یافته ای در پایگاه داده‌ها هستند .بعنوان مثال اگر سازمانی بخواهداطلاعات دو گروه متفاوت

[3] : KDD

از کارکنانش را با هم مقایسه کند ،پرس و جو ها مورد نیاز خواهند بود تا کارکنان منتسب به هر گروه را برای تحلیلهای بیشتر استخراج کنند .بنابراین پرس وجو ها نوعاً یک قدم پایه و اصلی در استخراج داده از پایگاه داده هستند . [۳]

سیستم های داده‌کاوی وسایلی هستند که به استفاده کننده اجازه می‌دهند تا داده‌ها را از پایگاه داده جمع آوری نموده ، تحلیل کنند و آنالیز ها را به صورت مختلف مثل گزارشها و گرافها ارائه دهند . این سیستم ها ممکن است از بسته‌های مدل‌سازی آماری در تحلیل داده‌ها جهت تولید اطلاعات برای پشتیبانی تصمیمات سازمان بهره برند .این سیستم ها می توانند اطلاعات سایر نواحی عملیاتی موسسه را با هم ترکیب نمایند .بعنوان مثال ممکن است کاربری یک مدل آماری برای تحلیل داده‌های زمان دریافت در خواست مشتری و زمان توزیع بسازد . اطلاعات بدست آمده از تحلیل قسمت خدمات مشتری و بازاریابی ممکن است در پشتیبانی تصمیمات مدیریتی منابع انسانی از طریق برآورد شاخص عملکرد مستخدم به کار آید .تا زمانی که ابزارهای این سیستم ها (سیستم های پشتیبانی کننده تصمیمات) اطلاعات مفیدی را برای تصمیم سازان در سازمان فراهم می آورند ، طرحهای بسیارخاصی می‌بایست تدارک دیده شوند .

توانایی ایجاد یک مدل پیشگویی کننده موفق به داده‌های گذشته بستگی دارد . داده‌کاوی طراحی شده است تا از موفقیت‌ها و شکست‌های گذشته بیاموزد و خواهد توانست آنچه که بعداً اتفاق خواهد افتاد را پیشگویی کند. ممکن است برخی فکر کنند چرا از داده‌کاوی در سازمان ها استفاده کنیم وقتی که روشها و تحلیل های آماری مهیا و آماده هستند ؟

آمار شاخه ای از علم ریاضی است که به جمع‌آوری توضیح و تفسیر داده‌ها می پردازد. این مبحث به گونه‌ایست که روزانه کاربرد زیادی دارد. این علم در مقایسه با داده‌کاوی قدمت بیشتری دارد و جزء روشهای کلاسیک داده‌کاوی محسوب می‌شود . وجه اشتراک تکنیک های آماری و داده‌کاوی بیشتر در تخمین و پیش بینی است. اگر تخمین و پیش بینی را جزء وظایف داده‌کاوی بدانیم ،تحلیل های آماری داده کاوی را بیش از یک قرن است که اجرا می‌کنند .به عقیده عده‌ای داده‌کاوی ابتدا از آمار و تحلیل های آماری شروع شد . می‌توان تحلیل های آماری از قبیل فاصله اطمینان ،رگرسیون و . . . را مقدمه و پیش زمینه داده‌کاوی دانست که داده‌کاوی به تدریج در زمینه های دیگر و متدهای دیگر رشد وتوسعه پیدا کرد .پس در واقع متد های آماری جزء روشهای قدیمی و کلاسیک داده‌کاوی محسوب می‌شوند .

در برخی موقعیت ها اینگونه بحث می‌شود که با تعریف دقیق ، آمار یا تکنیک های آماری جزء داده‌کاوی محسوب نمی‌شوند ، با این حال تکنیک های آماری توسط داده‌کاوی بکار برده می‌شوند و برای کشف موضوعات و ساختن مدل های پیشگویی کننده مورد استفاده قرار می‌گیرند. همانگونه که واضح و مشخص است با گذشت زمان علم نیز پیشرفت می‌کند ، هر چه به جلوتر می‌رویم روشهای جدید تر و بهتر مورد استفاده قرار می‌گیرد ،علم امروز نسبت به دیروز جدیدتر است. روشهای جدید علمی در پی کشف محدودیت های روشهای قدیمی ایجاد می‌شود و از آنجایی که روشهای آماری جزء روشهای قدیمی داده‌کاوی محسوب می‌شوند ، از این قاعده کلی که دارای محدودیت هستند مستثنی نیستند. داشتن فرض اولیه در مورد داده‌ها ، یکی از این موارد است .

در اینجا به تشریح بیشتر تفاوت‌های بین مباحث و متد های آماری و دیگر متد های داده‌کاوی می‌پردازیم . تکنیک های داده‌کاوی و تکنیک‌های آماری در مباحثی چون تعریف مقدار هدف برای پیش‌گویی داده‌های تمیز (Clean Data)خوب عمل می‌کنند و برای انواع مسائلی مانند کلاس‌بندی و کشف استفاده می‌شوند . بنابراین تفاوت این دو چیست ؟ چرا ما آنچنان که علاقمند به کار بردن روشهای داده‌کاوی هستیم ، علاقمند روشهای آماری نیستیم ؟

برای جواب این سوال چندین دلیل وجود دارد ؛ اول اینکه روشهای کلاسیک داده‌کاوی از قبیل شبکه‌های عصبی ، تکنیک نزدیک ترین همسایه برای داده‌های واقعی قوی‌تر عمل می‌کنند و استفاده از آنها برای کاربرانی که تجربه کمتری دارند راحت تر است و بهتر می‌توانند از آن استفاده کنند . دلیل دیگر آن است که این روشها با داده‌های بیشتر بهتر می‌توانند عمل کنند . در جایی دیگر اینگونه بیان شده است که داده‌های جمع آوری شده برای داده‌کاوی نوعاً خیلی از فرضهای قدیمی آماری را در نظر

نمی‌گیرند ، از قبیل اینکه مشخصه‌ها باید مستقل باشند ،توزیع داده‌ها باید معین شوند ، داده‌ها بیشترین همپوشانی در فضا و زمان را داشته باشند در حالیکه اغلب داده‌ها همپوشانی دارند. تخلف از هر کدام از فرضها می‌تواند مشکلاتی را در روشهای آماری ایجاد کند و این در حالیست که یک کاربر (تصمیم گیرنده) در جستجوی نتیجه داده‌های جمع آوری شده و مجموعه‌ای از مشاهدات چند بعدی است بدون توجه به اینکه چگونه جمع آوری شده‌اند.

در جایی پایه و اساس داده‌کاوی به دو مقوله آمار و هوش مصنوعی تقسیم شده است که روش‌های هوش مصنوعی به عنوان روش‌های فراگیری ماشینی در نظر گرفته می‌شوند. تفاوت اساسی بین روش‌های فراگیری ماشینی و آمار بر مبنای فرضهایشان است به عنوان یک قانون کلی فرض تکنیک های آماری بر این اساس است که توزیع داده‌ها مشخص است، در بیشتر موارد فرض براین است که توزیع نرمال است و در نهایت درستی یا نادرستی نتایج نهایی به درستی فرض اولیه وابسته است . در مقابل روش‌های فراگیری ماشینی از هیچ فرض در مورد داده‌ها استفاده نمی‌کنند و همین مورد باعث بروز تفاوت‌هایی بین این دو روش می‌شود .

ذکر این نکته ضروری است که بسیاری از روشهای فراگیری ماشین برای ساخت مدل از حداقل چند استنتاج آماری استفاده می‌کندکه این مساله به طور خاص در شبکه عصبی دیده می‌شود.

به طور کلی روشهای آماری ، روشهای قدیمی‌تری هستند که به حالت‌های احتمالی مربوط می‌شوند .داده‌کاوی جایگاه جدیدتری دارد که به هوش مصنوعی ، فراگیری ماشینی ، سیستم‌های اطلاعات مدیریت (MIS) ومتدلوژی پایگاه داده مربوط می‌شود . روش‌های آماری بیشتر زمانی که تعداد داده‌ها کم است و اطلاعات بیشتری می‌توان بدست آورد استفاده می‌شوند به عبارت دیگر این روشها با مجموعه داده‌های کوچکتر سروکار دارند .همچنین به کاربران ابزارهای بیشتری برای امتحان کردن داده‌ها با دقت بیشتر می‌دهد ، بر خلاف روشهایی از قبیل شبکه عصبی که فرآیند مبهمی دارند .

روش‌های آماری چون پایه ریاضی دارند نتایج دقیق‌تری نسبت به دیگر روشهای داده‌کاوی ارائه می‌دهند ولی استفاده از روابط ریاضی نیازمند داشتن اطلاعات بیشتری در مورد داده‌هاست . مزیت دیگر روشهای آماری در تعبیر و تفسیر داده‌هاست . هر چند روشهای آماری به خاطر داشتن ساختار ریاضی تفسیر سخت‌تری دارند ولی دقت نتیجه‌گیری و تعبیر خروجی‌ها در این روش بهتر است .به طور کلی روشهای آماری زمانی که تفسیر داده‌ها توسط روشهای دیگر مشکل است ، بسیار مفید هستند .

روشهای آماری بر مبنای فرض هستند و یک فرضیه شکل گرفته را احساس می‌کنند و این فرضیه را در مقابل داده‌ها مورد آزمون قرار می‌دهند. داده‌کاوی به طور متضادی بر مبنای اکتشاف است و فرضیات را به طور خودکار از داده‌ها استخراج می‌کند. دلیل دیگری که باعث شده است که تکنیک‌های داده‌کاوی در دنیای واقعی قدرتمندتر ظاهر شود داده‌های زیاد و کاربران با مهارت کمتر است [۲]. البته از آزمون های آماری در ارزیابی نتایج داده‌کاوی نیز استفاده می‌شود.

داده‌کاوی می‌تواند به سوالات تحلیلی پاسخ دهد، سوالاتی مانند: عناوین کارآیی کارکنان چیست ؟چه عاملی یا مخلوطی از عوامل مستقیما کارکنان را تحت تاثیر قرار می‌دهند؟ بهترین کارکنان کدام‌ها هستند؟ کدام کارکنان مایلند سازمان را ترک کنند و کدام مایلند ارتقا یابند؟

تحلیل روندها ، پیش گویی و پیش بینی نیازمند داده‌های تاریخی است که بدست آوردن آنها بدون یک پایگاه داده جامع بسیار سخت است .گذشته تک تک کارکنان می توانند قسمتی از تحلیل کارآیی باشد و به تدوین روابط کمک می‌کند و از همه مهمتر در هدایت سیستم مدیریت منابع انسانی از حالت سرپرستی به مدیریت استراتژیک کمک می‌کند.

فصل دوم

کارکردهای داده‌کاوی

زمان آن رسیده است که به داده‌کاوی به عنوان یک فرآیند تکنیکی نگریسته شود. با این نگرش تغییری در طرح کلی مساله ایجاد نمی‌شود اما نقطه اتکا عوض می‌شود . در این فصل به طبقه بندی داده‌کاوی از لحاظ کارکرد و هدف پرداخته‌خواهد شد. اینکه متناسب با مساله ما و هدفی که قصد رسیدن به آن را داریم می‌بایست از چه نوع داده‌کاوی و با چه تکنیکی استفاده کرد . و بدانیم کدام اقدام تکنیکی ما را به پیاده‌سازی مدل رهنمون خواهد ساخت .

داده‌کاوی روش یادگیری از داده‌های گذشته برای اتخاذ تصمیمات بهتری در آینده است . اما از دو نتیجه نامطلوب در این فرآیند یادگیری اجتناب کرد :

۱-یادگیری چیزهایی که درست نیستند .

۲-یادگیری چیزهایی که درست هستند اما مفید نیستند .

یادگیری چیزهای نادرست خطرناک‌تر از یادگیری چیزهای بی‌فایده است چرا که اتخاذ تصمیم‌های مهم کاری و یا تجاری می تواند بر اساس اطلاعات نادرست بنا شود . نتایج داده‌کاوی معمولاً قابل اعتماد به نظر می‌رسند چون به شیوه ظاهراً علمی و بر پایه داده‌های واقعی به دست آمده اند . این ظاهر قابل اعتماد می‌تواند گمراه‌کننده باشد . ممکن است خود داده‌ها نادرست یا نامربوط به مساله باشند و یا روش داده‌کاوی اتخاذ شده متناسب با مساله و هدف آن نباشد .

برخی اوقات الگوهای کشف شده درست هستند اما به لحاظ میزان اثر بخشی نتایج و یا پر هزینه بودن تکنیک مورد استفاده ، مفید واقع نمی‌شوند . در اینجاست که استفاده از یک تکنیک و الگوریتم ساده به حل مساله جنبه اقتصادی می‌دهد . چالش پیش روی داده کاوان اینست که از چه تکنیکی برای رسیدن به الگو بهره برند و بدانند کدام الگوها پیش‌بینی کننده هستند و آیا الگو های منتج از لحاظ هزینه فایده به صرفه می باشند ؟

در فصل قبل مشاهده کردیم که داده‌کاوی در زمینه‌های بسیار زیاد کاربرد دارد اما جدا از طبقه‌بندی کارکردهای داده‌کاوی به لحاظ زمینه کاری ، می توان پرسید داده‌کاوی به لحاظ ماهیت کاری چگونه طبقه بندی می‌شود . بسیاری از مسائل اطراف یک داده کاو را میتوان در قالب یکی از شش عمل زیر گنجاند :

- دسته‌بندی Classification
- تخمین Estimation

- پیش‌بینی — Prediction
- دسته‌بندی شباهت — Association Rules
- خوشه‌بندی — Clustering
- توصیف و نمایه سازی — Profiling

داده‌کاوی از لحاظ هدف در دو نوع ظاهر می شود:

۱- هدایت شده[1] : برخی زمینه‌های هدف خاص را بیان یا دسته‌بندی می کنند .
۲- غیرهدایت شده[2] : یافتن الگوها و تشابهات بین گروه‌هایی از اطلاعات بدون استفاده از زمینه خاص یا مجموعه‌ای ازدانستنی‌های از پیش تعیین شده .

سه مورد دسته‌بندی ، تخمین و پیش‌بینی جزء داده‌کاوی هدایت‌شده هستند که هدف آنها یافتن ارزش متغیر خاص است. دسته‌بندی شباهت و خوشه‌بندی جزء داده‌کاوی غیرهدایت شده هستند که در آن هدف ، یافتن ساختاری در داده‌ها بدون توجه به یک متغیر هدف خاص است . نمایه‌سازی، عملی توصیفی است که میتواند هم هدایت‌شده و هم غیرهدایت‌شده باشد .

دسته‌بندی

دسته‌بندی یکی از معمولترین کارکردهای داده‌کاوی است و در عین حال یکی از واجبات بشر است . ما برای شناخت و برقراری رابطه درباره دنیا به طور مداوم دسته‌بندی ، قسمت‌بندی و درجه‌بندی می‌کنیم . موجودات زنده را به گیاهان ، جانوران و موجودات میکروسکوپیک ، مواد را به عناصر و انسان‌ها را به نژادها تقسیم می‌کنیم .

دسته‌بندی شامل بررسی ویژگی‌های یک مورد جدید و تخصیص آن به یکی از مجموعه‌های از قبل تعیین شده می‌باشد . این عمل شامل ساختن مدلی است که بتوان از آن برای دسته‌بندی کردن داده‌های دسته‌بندی نشده استفاده نمود .

مثالهایی برای دسته‌بندی :

- دسته‌بندی متقاضیان وام و اعتبار به کم خطر ، متوسط و پرخطر
- انتخاب محتویات برای نشان دادن در یک وب
- تشخیص مدعیان حق بیمه که دریافت حق بیمه شامل آنها نمی‌شود
- تشخیص کارکنانی که به همکاری با سازمان ادامه نخواهند داد
- تشخیص مشتریان با ریسک بالا که قصد بیمه کردن اتومبیل خود را دارند

از جمله تکنیکهای دسته‌بندی : درخت‌های تصمیم‌گیری ،نزدیکترین همسایه ، شبکه‌های عصبی و تحلیل پیوند که در شرایط خاص عمل دسته‌بندی را انجام می‌دهد .

---

[1] : Directed
[2] : Undirected

تخمین:

تخمین با نتایج مجزا که با ارقام پیوسته نشان داده شده‌اند سروکار دارد . در تخمین داده‌های ورودی داده می‌شود و به رقمی در خروجی چون در آمد یا تراز کارت اعتباری ختم می‌شود .

فرض کنید شرکت مخابرات قصد دارد فضای تبلیغاتی صورت حساب ماهیانه خود را به یک تولیدکننده کفش کوهنوردی بفروشد . تولید کننده کفش می بایست دریافت کنندگان صورتحساب را به دو دسته کوهنورد و غیرکوهنورد تقسیم کند .روش دیگر ساخت مدلی است که به دریافت کنندگان صورتحساب امتیاز کوهنوردی بدهد .این امتیاز رقمی بین صفر تا یک خواهد بود که نشانگر احتمال تخمین زده شده برای کوهنورد شدن فرد است .برتری روش تخمین به دسته‌بندی امکان مرتب نمودن بر اساس امتیاز است. که امکان مانور بیشتری به فرد تصمیم گیر می‌دهد .

مثالهایی دیگرازتخمین :

- تخمین تعداد فرزندان در یک خانواده
- تخمین درآمد کل یک خانواده
- تخمین عمر یک مشتری
- تخمین احتمال پاسخ فردی خاص به یک پیشنهاد بیمه عمر

تکنیکهای رگرسیون و شبکه‌های عصبی برای تخمین مناسبند. تجزیه بقا نیز وقتی برای تخمین مناسب است که هدف تخمین زمان یک واقعه مانند توصیف یک مشتری است.

پیش‌بینی

پیش‌بینی مانند دسته‌بندی یا تخمین است با این تفاوت که اطلاعات مطابق برخی رفتارهای پیش‌بینی‌شده آینده یا ارقام تخمین زده شده دسته‌بندی می شوند. درعمل پیش‌بینی ، تنها روش برای بررسی صحت عمل ،انتظار دیدن آینده است .

پیش‌بینی مبنای علمی کمتری نسبت به روش دسته‌بندی دارد و مانند دسته‌بندی از قبل کلاسها به طور کامل مشخص نمی‌باشد ولی از آنجا که متغیر پیش‌بینی‌شونده مشخص است جزء داده‌کاوی هدایت شده قرار می گیرند .

هر یک از تکنیک‌های استفاده شده در دسته‌بندی و تخمین را می توان برای استفاده در پیش‌بینی تطبیق داد ، البته متغیری که باید پیش‌بینی شود معلوم است و داده‌های پیشین برای آن وجود دارد. از داده‌های پیشین برای تهیه یک مدل که بیانگر رفتار مشاهده شده کنونی است استفاده می‌شود، وقتی این مدل برای ورودی‌های کنونی بکار رفت نتیجه کار پیش‌بینی رفتار آینده خواهد بود .

دسته‌بندی شباهت یا قوانین وابستگی

عمل دسته‌بندی شباهت برای تعیین چیزهایی است که با هم جور هستند. مثال معمول آن تحلیل سبد بازار است. در تحلیل سبد بازار کالاها و خدماتی که با هم خریداری می شوند، تعیین شده و مورد تحلیل قرار می‌گیرند. فروشگاههای زنجیره‌ای و خرده‌فروشی‌ها می‌توانند از دسته‌بندی شباهت برای تعیین چیدمان کالاها در قفسه فروشگاهها و یا در دفترچه‌های تبلیغاتی و معرفی محصولات فروشگاه استفاده نمایند . تا اقلامی که با هم خریداری می‌شوند همیشه موجود بوده و در کنار هم چیده شوند (و یا حتی در صورت لزوم دور از هم چیده شوند!)

از دسته‌بندی شباهت می توان برای تعیین شرایط فروش متقابل و همزمان و همچنین برای طراحی بسته‌بندی‌های جذاب و یا دسته‌بندی محصولات یا خدمات استفاده کرد.

دسته‌بندی شباهت یک روش ساده برای ایجاد قوانین از داده‌هاست .

جهت دسته‌بندی شباهت از قوانین وابستگی استفاده می‌کنیم .

خوشه‌بندی

به عمل تقسیم جمعیت نا همگن به تعدادی از زیر مجموعه‌ها یا خوشه‌های همگن ،خوشه‌بندی گفته می‌شود .نقطه تمایز خوشه‌بندی از دسته‌بندی اینست که خوشه‌بندی به دسته‌های از پیش تعیین شده تکیه ندارد . در دسته‌بندی بر اساس یک مدل هر کدام از داده‌ها به دسته ای از پیش تعیین شده اختصاص می یابد . این دسته‌ها از طریق یافته‌های پژوهشهای پیشین تعیین گردیده‌اند .

در خوشه‌بندی هیچ دسته از پیش تعیین شده‌ای وجود ندارد و داده‌ها صرفاً براساس تشابه گروه‌بندی می‌شوند و عناوین هر گروه نیز توسط کاربر تعیین می گردد.

خوشه‌بندی معمولاً به عنوان پیش درآمدی برای انواع دیگری از داده‌کاوی یا مدلسازی به کار می‌رود. به عنوان مثال خوشه‌بندی ممکن است اولین گام در تلاش برای تقسیم‌بندی بازار باشد.

برای ایجاد قانونی که در همه موارد کاربرد داشته‌باشد و به این سوال پاسخ دهد که مشتریان به چه نوع تبلیغاتی به بهترین نحو پاسخ می دهند ، نخست باید مشتریان را به خوشه‌هایی با عادتهای مشابه تقسیم نمود سپس پرسید که چه نوع تبلیغاتی برای هر خوشه به بهترین نحو عمل می کند.

توصیف و نمایه سازی

گاهی اوقات هدف داده‌کاوی تنها توصیف آن چیزی است که در یک پایگاه داده پیچیده در جریان است . نتایج نمایه‌سازی درک ما را از مردم ، محصولات یا فرآیندهایی که داده‌ها را در مرحله اول تولید کرده‌اند افزایش می‌دهد. توصیف خوب رفتار ، اغلب توضیح خوبی هم به همراه دارد. حداقل یک توصیف خوب نشان می‌دهد کی می‌توان انتظار یک توضیح مناسب را داشت.