
بسم الله الرحمن الرحيم

دانشکده مهندسی صنایع و مدیریت

گروه مدیریت

پایان‌نامه جهت اخذ مدرک کارشناسی ارشد مدیریت اجرایی MBA

مدیریت ریزش کارکنان با استفاده از تکنیک‌های داده‌کاوی در شرکت نفت و گاز پارس

علیرضا ابراهیمی

استاد راهنما

دکتر بزرگمهر اشرفی

بهمن ۱۳۹۲

ماحصل آموخته‌هایم را تقدیم می‌کنم به آنان که مهر آسمانی شان آرام‌بخش آلام زمینی‌ام است. به استوارترین تکیه‌گاهم، دستان پرمهر پدرم، به مادرم که تار مویی از او بپای من سیاه نماند، به همسرم که الگوی صبر، مهربانی و گذشت است.

تشکر و قدردانی

سپاسگذار کسانی هستم که سرآغاز تولد من هستند، از یکی زاده می شوم و از دیگری جاودانه.

درود فراوان خدمت پدر و مادر عزیز، دلسوز و فداکارم که پیوسته جرعه نوش جام تعلیم و تربیت، فضیلت و انسانیت آنها بوده ام و با سپاس فراوان از همسرم که حس تعهد و مسئولیت را در زندگیمان تلالویی خدایی داده است.

با امتنان بیکران از مساعدت های بی شائبه ی استاد فرهیخته ام جناب آقای دکتر بزرگمهر اشرفی که با راهنمایی های دلسوزانه ی خود راهگشای اینجانب بوده اند.

برای همه ی این عزیزان آرزوی سربلندی و روزگاری سبز دارم.

تعهدنامه

اینجانب علیرضا ابراهیمی دانشجوی دوره کارشناسی ارشد رشته MBA دانشکده صنایع و مدیریت دانشگاه صنعتی شاهرود نویسنده پایان نامه مدیریت ریزش کارکنان با استفاده از تکنیک-های داده‌کاوی در شرکت نفت و گاز پارس تحت راهنمایی دکتر بزرگمهر اشرفی متعهد می‌شوم:

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام دانشگاه صنعتی شاهرود به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تاثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.

در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری و اصول انسانی رعایت شده است.

چکیده

نیروی انسانی، عنصر و سرمایه اصلی شرکت‌ها و سازمان‌ها می‌باشد. تعدد شرکت‌هایی که به دنبال نیروی انسانی خبره و کارآمد هستند، موجب شده تا کارکنان براساس انتظارات خود از شغل، شرکت مورد نظرشان را انتخاب کرده و در آن مشغول به کار شوند. افزایش قدرت انتخاب افراد سبب شده تا آنها به محض نارضایتی از شرایط حاکم بر سازمان، به راحتی سازمان را ترک کرده و به سازمان دیگری جهت کار مراجعه کنند. لذا برای شرکت‌ها انتخاب افرادی با تعهد سازمانی بیشتر اهمیت ویژه‌ای پیدا کرده است تا بدین طریق بتوانند هزینه‌های ناشی از ریزش کارکنان را کاهش دهند. لذا مسئله‌ی مدیریت ریزش کارکنان در بسیاری از شرکت‌ها مطرح شده است. در این تحقیق سعی شد تا با استفاده از داده‌های جمع‌آوری شده از شرکت نفت و گاز پارس و با به کارگیری تکنیک‌های داده‌کاوی الگوهایی جهت پیش‌بینی ریزش و عدم ریزش کارکنان استخراج شود تا این شرکت بتوانند برمبنای آن تصمیم به جذب نیروهای متعهد به سازمان بگیرد.

واژگان کلیدی: ریزش کارکنان، تعهد سازمانی، داده‌کاوی، تکنیک نظریه راف‌ست

فصل اول

مقدمه

۱-۱. مقدمه

نیروی انسانی، عنصر و سرمایه اصلی شرکت‌ها و سازمان‌ها می‌باشد. تعدد شرکت‌هایی که به دنبال نیروی انسانی خبره و کارآمد هستند، موجب شده تا کارکنان براساس انتظارات خود از شغل، شرکت مورد نظرشان را انتخاب کرده و در آن مشغول به کار شوند. افزایش قدرت انتخاب افراد سبب شده تا به آنها به محض نارضایتی از شرایط حاکم بر سازمان، به راحتی سازمان را ترک کرده و به سازمان دیگری جهت کار مراجعه کنند. در واقع کارکنان ممکن است به دلایل مختلف بک سازمان را ترک کنند، به چنین رخدادی که در سازمان اجتناب ناپذیر است، ریزش کارکنان می-گویند. ریزش کارکنان باعث اختلال و صرف هزینه و وقت در سازمان ها می شود به همین دلیل به یک دغدغه مهم برای هر سازمانی به ویژه سازمان های خدماتی و سازمان های فرافن[۱] تبدیل شده است. از این رو مدیران شرکت‌ها به دنبال راهکارهایی برای حفظ و نگهداری کارکنان کنونی خود هستند. برای حفظ کارکنان ابتدا نیاز به شناسایی کارکنانی ست که ریزش آنها محتمل است. بدین منظور مدیران و کارشناسان شرکت‌ها نیاز به روش‌هایی برای شناسایی کارکنان دارای خاصیت ریزش خواهند داشت تا بتوانند با شناسایی آنها و اعمال استراتژی‌های مناسب مدیریتی از ریزش آنها جلوگیری کنند.

در پژوهش پیش‌رو سعی شده تا با بهره‌گیری از تکنیک راف ست[۲] الگوهایی از داده‌های گذشته استخراج شود که بر مبنای آنها بتوان کارکنانی که ریزش آنها محتمل‌تر است را شناسایی و هنگام گزینش کارکنان این معیار را مدنظر قرار داد.

۲-۱. عنوان تحقیق

جهت پاسخ به سئوالات، این پژوهش با عنوان زیر معرفی شده است:

مدیریت ریزش کارکنان با استفاده از تکنیک‌های داده‌کاوی در شرکت نفت و گاز پارس

---

۱ high- tech
۲ rough set

۳-۱. بیان مساله تحقیق

مدیریت استراتژیک، سطوح استراتژی را در سه سطح عملیاتی، تجاری و شرکت بیان می‌کند. یکی از عواملی که استراتژی‌های سطح عملیاتی را تحت‌الشعاع قرار می‌دهد، منابع انسانی است. عامل مذکور در راستای تحقق اهداف شرکت‌ها نقش کلیدی ایفا می‌کند و در صورتی که به درستی گزینش شود و کادر قوی‌ای از نیروی انسانی گرد هم آیند، به سرمایه‌ها اصلی سازمان تبدیل می-شوند.

دلایل متعددی در یک سازمان، کارکنان را که سرمایه‌های اصلی در یک سازمان هستند، وادار به ترک سازمان می‌کند. لذا بررسی این عوامل و شناسایی الگوهایی از روند ترک کارکنان در سازمان چالشی‌ست که کمتر شرکتی از موج آن در امان مانده است.

بنا به گفته‌ی معاون توسعه منابع انسانی و مدیریت وزیر نفت "کار اقماری، تامین نشدن بموقع نیروی انسانی به دلیل فرآیند طولانی جذب، علاقه نداشتن نیروی غیر بومی به کار در مناطق کمترتوسعه یافته و توسعه نیافته، جاذبه کاذب ایجاد شده در کشورهای همسایه و کمبود نیروی متخصص به دلیل استمرار نداشتن در سیستم جذب نیرو یکی از مشکلات این سازمان است. از جمله راهکارهای برون رفت از این مشکلات به کارگیری نیروی‌های بومی، کاهش سن بازنشستگی به دلیل وجود فرسودگی‌های بالای شغلی و بازنگری در حقوق و مزایا است.

اگر سعی شود تا انتخاب کارکنان مطابق با روش‌های علمی صورت پذیرد، می‌توان بخش اعظمی از مشکل ریزش کارکنان را حل کرد. لذا در این پژوهش با بررسی صورت گرفته در زمینه‌ی ریزش-های کارکنان در شرکت نفت و گاز پارس ، این شرکت بعنوان مطالعه‌ی موردی انتخاب شده و داده‌های مورد نیاز جهت کشف الگوهای ریزش کارکنان از این شرکت استخراج شده است.

۴-۱. ضرورت و اهداف پایان نامه

روش‌های داده‌کاوی از جمله ابزارهای قدرتمند در تحلیل اطلاعات و استخراج دانش محسوب می‌شود. از طرفی، استخراج دانش از داده‌های خام، همیشه موضوعی چالش برانگیز در میان متخصصین علوم مختلف بوده است. از این رو داده‌کاوی نتایج بسیار مفیدی را در اختیار متخصصان قرار داده است و کارشناسان علوم مختلف می‌توانند با بهره گرفتن از دیدگاه‌ها و روش‌های نوین داده کاوی دقت تجزیه و تحلیل اطلاعات خود را افزایش دهند. این مسئله انگیزه‌ای شد تا در این پایان‌نامه بر گوشه‌ای از قابلیت‌های داده‌کاوی تمرکز کرده و میزان قدرت آنها را نمایش دهیم.

یکی از زمینه‌هایی که داده‌کاوی در آن به ایفای نقش می‌پردازد، مدیریت منابع انسانی است که امروزه از آن به عنوان یکی از دغدغه‌های مدیران شرکت‌ها و سازمان‌ها یاد می‌شود. این مسئله به ویژه در سازمان‌هایی که نیروی انسانی چرخ‌های گرداننده‌ی آن هستند و حفظ آنها قطعا موجب افزایش سرعت پیشرفت و ترقی خواهد شد و در مقابل، از دست دادن آنها بسیار ضررآفرین بوده، نمود جدی‌تری دارد. از این رو در مطالعات مختلفی، از مدل‌های داده‌کاوی به منظور شناسایی کارکنانی که ریزش آنها محتمل است، استفاده شده است. در این پایان نامه قصد داریم از مدل نظریه راف ست برای شناسایی کارکنانی که ریزش آنها محتمل است، استفاده کنیم.

۱-۵. اهداف تحقیق

هدف اصلی این تحقیق استخراج الگوهایی کارآ از اطلاعات تاریخی‌ست تا برمبنای آن بتوان ریزش و یا عدم ریزش کارکنان را پیش‌بینی کرد.

۱-۶. سئوالات تحقیق

با توجه به مباحث فوق و هدف تحقیق، می‌توان سئوالات تحقیق را به صورت زیر طرح کرد:

۱- آیا می‌توان بابکارگیری تکنیک‌های داده‌کاوی الگویی کارآ برای پیش‌بینی ریزش کارکنان استخراج کرد؟

۲- آیا می‌توان با استفاده از داده‌های تاریخی موجود در شرکت، الگویی کارآ برای پیش‌بینی ریزش کارکنان استخراج کرد؟

۱-۷. جنبه‌های جدید و نوآوری تحقیق

در این پژوهش سعی شده، با بهره‌گیری از رویکردهای داده‌کاوی، با جامع‌نگری بیشتری به مقوله‌ی گزینش کارکنان که سرمایه‌های اصلی در یک سازمان هستند، پرداخته شود و بابکارگیری تکنیک راف‌ست و داده‌های تاریخی شرکت الگوهایی جدید جهت تصمیم‌گیری گزینش کارکنان ارائه شود.

۱-۸. کاربردهای حاصل از نتایج تحقیق

از آنجا که منابع انسانی یکی از سرمایه‌های حیاتی شرکت‌ها محسوب می‌شود و مدیریت آنها بسیاری از شرکت‌ها را به چالش کشیده است و به دغدغه‌ی اصلی بسیاری از مدیران تبدیل شده است، از جمله کاربردهای این پژوهش می‌توان به تصمیم‌گیری در مورد گزینش و یا عدم گزینش کارکنان در سازمان اشاره کرد.

۱-۹. روش‌ها و ابزار تجزیه و تحلیل داده‌ها

در این تحقیق، اطلاعات مربوط به کارکنان استخراج شده است و سپس از طریق پیاده‌سازی الگوریتم‌های داده‌کاوی به شناسایی الگوهای مشترک در مورد کارکنان پرداخته شده است و در نهایت براساس دانش به دست آمده می‌توان توصیه‌های لازم و مناسب در مورد گزینش و یا عدم گزینش کارکنان ارائه کرد.

۱-۱۰. ساختار پایان نامه

بخش‌های مختلف پایان نامه مطابق ذیل تنظیم شده است:

فصل دوم اختصاص به داده‌کاوی و نحوه کاربرد آن در پیش‌بینی ریزش کارکنان دارد. در فصل سوم روش شناسی تحقیق، مدل نظریه راف ست، معرفی و در پایان فصل معیارهای ارزیابی مدل

ذکر شده‌اند. در فصل چهارم ابتدا به توضیح داده‌های مورد استفاده در این پایان نامه که مربوط به اطلاعات کارکنان شرکت نفت و گاز پارس ست، پرداخته و سپس مدل مذکور بر روی داده‌ها برازش شده و نتایج مورد ارزیابی قرار می‌گیرند.

فصل دوم

ادبیات موضوع و پیشینه تحقیق

۱-۲. مقدمه

امروزه بسیاری از سازمان‌ها افراد را، خواه به صورت کارشناسان فنی، کارشناسان متخصص در ارائه‌ی خدمات به مشتریان و مدیران و ... ، مزیت رقابتی خود می‌دانند. در عصر حاضر منابع انسانی به طور متناقض می‌توانند عاملی برای موفقیت یا شکست برای همه‌ی سازمان‌ها باشند. یکی از عواملی که در تاثیرگذاری کارکنان در سازمان تاثیر دارد، مدیریت آنها می‌باشد. همین مسئله باعث پیدایش شاخه‌ای در مدیریت با عنوان مدیریت منابع انسانی شده است. مدیریت منابع انسانی عبارت است از فرآیند کار کردن با افراد، به طوری که این افراد و سازمان‌شان به توانمندی کاملی دست یابند، حتی زمانی که تغییر، نیاز به کسب مهارت‌های جدید، تقبل مسئولیت‌های جدید و شکل جدیدی از روابط را ملزم باشد در واقع مدیریت منابع انسانی استفاده از نیروی انسانی در جهت اهداف سازمان است و شامل فعالیت‌هایی نظیر کارمندیابی و جذب، آموزش، حقوق و دستمزد و روابط سازمانی می‌شود. در منابع انسانی مدرن، فعالیت‌هایی نظیر برنامه‌ریزی کارراهه شغلی، جبران خدمت و مزایا، جانشین پروری، مدیریت استعدادها، نگرش کارکنان، تقویت کار تیمی و استراتژی منابع انسانی نیز وارد شده‌اند.

مدیریت منابع انسانی باید برای سازمان‌ها ارزش‌آفرینی کند. این ارزش‌آفرینی هم برای ذی‌نفعان داخلی و هم ذی‌نفعان خارجی است. ذینفعان داخلی کارکنان و سازمان هستند در حالی که ذی‌نفعان خارجی مشتریان سازمان، سرمایه‌گذاران و همچنین جامعه است.

اما متخصصین منابع انسانی با تمرکز بر سه راه حل می‌توانند باعث ارزش‌آفرینی شوند. اولین راه حل تمرکز بر مدیریت استعداد است که به سطح فردی بر می‌گردد.نیروی کار متعهد، قابل و نتیجه بخش از این طریق پرورش می‌یابد. دومین راه حل تمرکز بر سازمان است که با ارتقا فرهنگ سازمانی و توسعه سازمانی حاصل می شود. ترکیب این دو به راه حل سوم که رهبری است می‌رسد.

یکی از مهمترین شاخصهای مدیریت منابع انسانی شاخص نرخ ریزش منابع انسانی[۳] است. مبحث حفظ و نگهداری نیروی انسانی و جذب و استخدام مستقیما تحت تاثیر این شاخص قابل مدیریت می باشند اما این شاخص چیست و چگونه اندازه گیری می شود:

در یک تقسیم بندی ساده دو نوع خروج کارکنان از سازمان رخ می دهد: خروج داوطلبانه و یا خروج غیر داوطلبانه. همانگونه که اسامی آنها نشان می دهد، خروج داوطلبانه با میل و اختیار کارکنان صورت می گیرد در حالی که خروج غیرداوطلبانه جبرا بر کارکنان تحمیل می شود حال چه از طریق اخراج و یا تعدیل نیرو و یا رسیدن به سن بازنشستگی. اما خروج داوطلبانه نیز قابل تقسیم بندی به زیر شاخه های دیگری است.

از لحاظ ساختاری، چارچوب مدیریت منابع انسانی را می‌توان به دو شاخه عملیاتی و تحلیلی طبقه بندی کرد(برسون و همکارن، ۲۰۰۰). شاخه عملیاتی آن شامل مهارت‌ها و دانش فنی است و شاخه تحلیلی آن به تحلیل خصوصیات و رفتار کارکنان برای محافظت از استراتژی‌های مدیریتی اشاره دارد. این تجزیه و تحلیل بر روی اطلاعات حاصله از کارکنان در شاخه عملیاتی انجام می-شود. بعبارت دیگر شاخه تحلیلی مدیریت منابع انسانی، بر انبار داده سازمان‌ها و کشف دانش از آنها تکیه دارد.

به طور کلی مدیریت ارتباط منابع انسانی به سه مرحله تقسیم می‌شود:

۱. شناسایی کارکنان

چرخه مدیریت ارتباط با کارکنان، با شناسایی کارکنان آغاز می‌شود. در این مرحله، افرادی که احتمال می‌رود که برای شرکت سود آور باشند، شناسایی شوند. این مرحله، شامل تحلیل کارکنانی که شرکت را ترک کرده و به سمت شرکت‌های رقیب سوق پیدا کرده‌اند می‌باشد. دو عنصر مهم در شناسایی کارکنان، تحلیل فرد و تقسیم‌بندی کارکنان است. تحلیل فرد، با تکیه بر خصوصیات افراد

[3] Turnover rate

به دنبال دسته‌های سودآوری از آنها می‌گردد، در حالیکه در تقسیم‌بندی کارکنان، مجموعه کامل کارکنان به زیرمجموعه‌های مجزا از هم به گونه‌ای تقسیم می‌شوند که دارای بیشترین شباهت و همگونی باشند.

۲. جذب کارکنان

مرحله‌ی دوم مدیریت منابع انسانی، دنباله‌رو مرحله اول است. بعد از شناسایی دسته افراد بالقوه، سازمان‌ها می‌توانند برای جذب آنها تلاش کرده و آنها را مورد هدف قرار دهند.

۳. نگهداری کارکنان

برای شناسایی و جذب نیرو، قطعا هزینه‌هایی صورت گرفته که با ترک نیرو این هزینه به هدر می-رود. به همین دلیل نگهداری افراد جذب شده یکی از مهم‌ترین مسائل نگران‌کننده برای شرکت-هاست. رضایت‌مندی کارکنان، مهم‌ترین شرط در نگهداری کارکنان است که از طریق برآورده شدن انتظاراتش صورت می‌گیرد.

موضوع این پایان‌نامه مرتبط با مرحله حفظ و نگهداری کارکنان است که تحت عنوان " مدیریت ریزش کارکنان" در زیر به طور خلاصه توضیح داده شده است.

۲-۲. تعریف گسترده‌ی اصطلاحات

مدیریت

تعریف اول : مدیریت عبارت ا ست از هما هنگ کردن منا بع انسا نی ومادی برای نیل به هدف سازمان

منابع انسانی

درمیان عوامل متعدد در مدیریت منابع انسانی از بالاترین اهمیت برخوردار است زیراکه کا رآمدی دیگر عوامل نیز به وضعیت و عملکرد انسان و رفتارهای او مرتبط است و برهمین مبناست که در بازگشت سرمایه بعنوان یکی از باارزش‌ترین شاخص‌های بهره وری است و سازمان مرهون عامل حیاتی تر تحت عنوان نیروی انسانی است.

مدیریت منابع انسانی

مدیریت منابع انسانی عبارت است از فرآیند کار کردن با افراد، به طوری که این افراد و سازمان‌شان به توانمندی کاملی دست یابند، حتی زمانی که تغییر، نیاز به کسب مهارت‌های جدید، تقبل مسئولیت‌های جدید و شکل جدیدی از روابط را ملزم باشد در واقع مدیریت منابع انسانی استفاده از نیروی انسانی در جهت اهداف سازمان است و شامل فعالیت‌هایی نظیر کارمندیابی و جذب، آموزش، حقوق و دستمزد و روابط سازمانی می‌شود. در منابع انسانی مدرن، فعالیت‌هایی نظیر برنامه‌ریزی کارراهه شغلی، جبران خدمت و مزایا، جانشین پروری، مدیریت استعدادها، نگرش کارکنان، تقویت کار تیمی و استراتژی منابع انسانی نیز وارد شده‌اند.

ریزش کارکنان

ریزش کارکنان که رویگردانی کارکنان نیز نامیده می‌شود، اصطلاحی‌ست که در مواقع ترک یک فرد از سازمان مورد استفاده قرار می‌گیرد. به طور کلی کارکنان ریزش شده را براساس علت ریزش به دو دسته تقسیم می‌کنند که عبارتند از:

۱- ریزش داوطلبانه

- ریزش تصادفی: زمانی رخ می‌دهد که به دلایل غیر عمدی از جمله تغییر محل زندگی و ... فرد دیگر قادر به همکاری با سازمان نیست.
- ریزش عمدی: زمانی رخ می‌دهد که فرد به دلیل مزایای بهتر، حقوق بالاتر، پست شغلی و موقعیت اجتماعی بهتر و ... سازمان را ترک گفته و با سازمان دیگری همکاری می‌کند.

۲- ریزش غیر داوطلبانه: در این نوع ریزش سازمان به دلایلی از جمله ناکارآمدی و کارشکنی فرد در سازمان و ... فرد در سازمان شناسایی و اخراج می‌شود.

اولین نوع تقسیم بندی بر اساس کارکرد خروج است برخی از خروجها برای سازمان مخل هستند و برخی اینگونه نیستند .آنچه که معمولا در سازمانها و نزد مدیران منابع انسانی حائز اهمیت است خروجهای مخل برای سازمان می باشند همانند از دست دادن نیروهایی انسانی کارآمد که با رفتن

آنها سازمان نیاز به جانشین یابی می باشد. اما این تقسیم بندی باز نیز می تواند ادامه یابد از میان خروجهای داوطلبانه مخل به سازمان برخی اجتناب پذیر و برخی اجتناب ناپذیر می باشند. مثالهای خروجهای اجتناب ناپذیر می تواند به انتقال محل سکونت کارمند و یا بیماری و مانند اینها اشاره نمود. گرچه همه خروجها قابل بررسی هستند ولی آنچه که باید در درجه اول بررسی گردد خروجهای داوطلبانه مخل برای سازمان اجتناب پذیر می باشند. این نوع خروجها نشان دهنده ضعف در سیستم نگهداری انسانی و یا جذب و استخدام می باشد.

اما برای محاسبه نرخ ریزش نیروی انسانی ، می توان این نرخ را بر مبنای ماهانه و سالانه محاسبه نمود. برای محاسبه ماهانه آن می توان تعداد نیروهای انسانی خروجی را به میانگین کل نیروهای انسانی در آن ماه تقسیم کرد .میانگین نیروی انسانی در ماه نیز با گرفتن میانگین تعداد نفرات اول ماه و تعداد نفرات آخر ماه در واحد مورد بررسی حاصل می شود. هنگام محاسبه بر مبنای سالانه تمامی موارد فوق را در بازه یکسال انجام می دهیم.

اما دانستن نرخ ریزش نیروی انسانی به تنهایی کمکی به شناخت بهتر از سازمان نمی کند. این شاخص زمانی مفید است که به عنوان ابزاری برای مدیریت اقدامات منابع انسانی در دو حوزه یاد شده و یا سایر اقدامات استفاده شود .حالت ایده آل آن است که این شاخص نه تنها به عنوان شاخصی نسبی در تحلیل اقدامات ما استفاده شود بلکه با بهینه یابی در بازار به مقایسه نرخ ریزش نیروی انسانی خود با سایر سازمانها در صنعت خود بپردازد. به عنوان مثال نرخ ریزش نیروی انسانی در رستورانهای تهران قابل مقایسه با نرخ ریزش نیروی انسانی در مراکز تماس اپراتورهای تلفن همراه نمی باشد چون گرچه هر دو نرخ ممکن است بالا باشد ولی تابع عوامل مختلفند. مثلا رستورانهای تهران تحت تاثیرات فصلی دچار ریزش نیروی انسانی می گردند در حالی که مراکز تماس عموما و به طور متوسط دارای نرخ ریزش بالایی هستند.

مجموعه شاخص‌ها باید در کنار هم قرار گیرند تا تفسیری صحیح بتوان از آن ارائه کرد. نکته‌ی حائز اهمیت این است که بسیاری از سازمانهای بزرگ و کوچک ما این شاخص کلیدی را مورد اغفال قرار می دهند و به صورت منظم نسبت به اندازه گیری و ثبت آن اقدام نمی کنند.

این در حالیست که جذب یک فرد در سازمان هزینه‌های زیادی را به سازمان متحمل می‌کند. چو[۴] و همکاران (۲۰۰۷)، هزینه جذب یک کارمند جدید را دو برابر بیشتر از هزینه نگهداری کارمندان موجود برآورد کردند. بنابراین از بین حالات ذکر ذکر شده برای ریزش کارکنان، ترک عمدی برای مدیران شرکت‌ها حائز اهمیت است. اگر مدیران قادر به شناسایی کارکنان با احتمال ریزش بالا باشند، می‌توانند از همان ابتدا در جذب کارکنان تامل بیشتری کنند. مدیریت ریزش کارکنان لزوما روی تمامی کارکنان با احتمال ریزش اعمال نمی‌شود و برای سازمان‌ها جلوگیری از ریزش کارکنان کارا و خبره در اولویت بالاتری قرار دارد. جلوگیری از ریزش کارکنان، باعث کاهش نرخ ریزش کارکنان و افزایش اعتبار شرکت خواهد شد.

مروری بر مطالعات قبلی طی سال‌های ۲۰۰۰ تا ۲۰۰۶ نشان داده است که حدود ۶۰ درصد مقاله-های مرتبط با موضوع مدیریت ریزش کارکنان، در زمینه نگهداری کارکنان بوده است. بنابراین نگهداری کارکنان در سال‌های اخیر بیشتر مورد توجه بوده و محققان را بر آن داشته تا راهکارهایی را برای حفظ کارکنان خبره ارائه دهند.

همانطور که بیان شد، در شاخه تحلیلی مدیریت منابع انسانی و ریزش کارکنان، کسب دانش از انبار داده‌های کارکنان است. در مدیریت ریزش کارکنان، دانش مورد نظر برای شناسایی کارکنانی بکار می‌رود که خصوصیات ریزش در آنها وجود دارد. مطالعات نشان داده‌اند که ابزارهای داده‌کاوی به منظور کشف دانش از پایگاه‌های داده کارکنان، بسیار مفید عمل کرده‌اند و مقاله‌ها و کتاب‌های متعددی در این زمینه به چاپ رسیده‌اند. در بخش بعدی به معرفی داده‌کاوی و نحوه کاربرد آن در استخراج دانش از پایگاه داده کارکنان می‌پردازیم.

---

[۴] chu

۳-۲. مروری بر مطالعات قبلی انجام شده در زمینه ریزش کارکنان

ریزش کارکنان پدیده‌ای ست که بسیاری از محققان به آن پرداخته‌اند. اما هیچ دلیل استانداردی برای ترک افراد در سازمان وجود ندارد[۱]. ریزش کارکنان چرخش نیروی کار بین شرکت‌ها و مشاغل در بازار کار است[۲]. اصطلاح ریزش طبق تعریفی که پرایس(۱۹۷۷) ارائه کرده است نسبت تعداد کارکنانی که در بازه‌ی زمانی مشخصی سازمان را ترک کرده‌اند به تعداد کارکنانی که در آن بازه در آن سازمان کار می‌کنند. غالبا مدیران از ریزش با عنوان کل فرآیند ریزش و پر کردن پست‌های خالی در سازمان یاد می‌کنند. مادامی‌که پستی در سازمان خالی می‌شود، داوطلبانه و یا غیر داوطلبانه باید شخصی برای آن پست نهاده شده و آموزش ببیند. این چرخه‌ی جایگزینی را چرخه‌ی ریزش می‌گویند[۳].

مدل‌های مرتبط با ریزش داوطلبانه، با تمرکز روی جنبه‌های تصمیماتی ریزش کارکنان، واگرایی از تفکرات گذشته را نشان می‌دهد[۴]. تحقیقات صورت گرفته علل ریزش کارکنان را به دو دسته تقسیم کرده است که ذیلا آورده شده‌اند.

۳-۲-۱. فاکتورهای مرتبط با شغل

محققان بسیاری[۵] تلاش کرده‌اند تا پاسخی برای انگیزه‌های افراد در ترک سازمان بیابند. تاکنون به دلیل تنوعی که بین کارکنان وجود داشته، نتایج به‌دست آمده از تحقیقات ناسازگار است[۵]. بنابراین دلایل متفاوتی برای ترک سازمان توسط افراد وجود دارد. دلایلی از جمله استرس شغلی، عوامل استرس‌زا در شغل، نبود تعهد سازمانی و نارضایتی شغلی می‌توانند سبب ترک سازمان شوند[۶]. بدیهی‌ست که تمامی دلایل فوق و سایر دلایل شخصی هستند. فاکتورهای دیگری مثل نمایندگی شخصی، عدم داشتن حس قدرت، کنترل شخصی نیز می‌تواند بر ریزش کارکنان تاثیرگذار باشند. فیرث[۶] و مانو[۷] علل ترک سازمان توسط افراد را عوامل مالی و اقتصادی می‌دانند. آنها بابکارگیری مدل‌های اقتصادی نشان دادند که علت ترک کارکنان اقتصادی است و از این نتایج

۵Bluedorn, Kalliath, Beck, Kramer, … .
۶ Firth
۷ manu

برای پیش‌بینی ریزش استفاده کردند[۷]. سازمان‌های بزرگتر موقعیت مناسب‌تری را برای رشد کارکنان، حقوق و مزایای مناسب‌تری را فراهم کنند ولذا اطمینان بیشتری از تعهد کارکنان به سازمان حاصل می‌کنند. تروور[۸] تعدد ریزش‌ها را نارضایتی شغلی می‌داند و از این فاکتور برای پیش‌بینی ریزش کارکنان استفاده کرده است. مسئله‌ی دیگری با عنوان ابهام در شعل وجود دارد و از تفاوت بین انتظارات دیگران از کارکنان و فهم آنها از شغل شان ناشی می‌شود. این باعث عدم اطمینان کارکنان از شغلشان می‌شود.

اطلاعات ناکافی از نحوه‌ی انجام کار، چگونگی برآوردن انتظارات، ابهامات موجود در روش‌های ارزیابی عملکرد، فشارهای کاری و عدم رضایت شغلی ممکن است سبب شود تا افراد تعهد کمتری را نسبت به سازمان داشته باشند و در نهایت باعث ایجاد تمایل در افراد به ترک سازمان می-شود[۸]. اگر نقش کارکنان در سازمان به وضوح مشخص نشده باشد، میزان ریزش کارکنان افزایش می‌یابد. فاکتورهای از قبیل مرگ و یا عدم صلاحیت کارکنان وجود دارند که خارج از کنترل‌های مدیریتی هستند.

۲-۳-۲. فاکتورهای مرتبط با سازمان

بی‌ثباتی سازمانی یکی از دلایل عمده در نرخ بالای ریزش کارکنان است. شواهد نشان می‌دهند، در صورتی که محیط کاری باثبات باشد کارکنان تمایل دارند که در سازمان بمانند. در سازمان‌هایی که نرخ ناکارآمدی بالاست، نرخ ریزش کارکنان نیز بالاست[۹].

تحمیل رویکردهای کمی برای مدیریت کارکنان سبب نارضایتی کارکنان و در نهایت خروج آنها از سازمان می‌شود. مادامی‌که کارکنان در پروسه‌ی تصمیم‌گیری در سازمان شرکت داشته باشند، ندرتا سازمان را ترک می‌کنند. کارکنان باید نسبت یه همه‌ی مسائلی که شرایط کاری را تحت تاثیر قرار می‌دهند، آگاهی داشته باشند. ولی در سازمان‌ها بدلیل عدم به اشتراک‌گذاری اطلاعات، حوزه‌ی کاری کارکنان محدود ست[۱۰]. کاستلی[۹] اشاره کرده است که نرخ بالای ریزش کارکنان

8. Trevor
9. costly

می‌تواند به دلیل ضعف سیاست‌های شخصی، سیاست‌های استخدام، ضعف نظارتی، رویه‌های کاری ضعیف و نبود انگیزه اتفاق بیفتد. تمامی عوامل فوق در قالب ضعف‌های مدیریتی می‌توانند باعث نرخ بالای ریزش کارکنان شوند[۱۱].

گریفث و همکارانش متغیرهایی مرتبط به پرداخت‌ها را در ریزش‌ها اثربخش می‌دانند. آنها همچنین در تحقیقات خود نشان دادند که بین پرداخت‌ها، عملکرد و ریزش رابطه‌ی مستقیم وجود دارد. بدین طریق که وقتی کارکنان در ازای عملکرد بالا پاداش‌های متناسب دریافت نمی‌کنند، سازمان را ترک می‌کنند. اگر مشاغل، مشوق‌های مالی کافی داشته باشند، احتمال ریزش کارکنان کاهش می‌یابد. همچنین عوامل دیگری از جمله استیل مدیریتی، عدم آگاهی از سازمان، عدم مزایای رقابتی در سازمان و ... سبب ریزش کارکنان می‌شود[۱۲].

۳-۳-۲. اثرات ناشی از ریزش کارکنان

از دید سازمانی خروج کارکنان از سازمان، هزینه‌بر است. خروج‌های سازمانی داوطلبانه به مثابه خروج سرمایه‌های انسانی از سازمان است و فرآیند جایگزینی در سازمان هزینه‌هایی را برای سازمان در پی دارد. این هزینه‌ها شامل جستجو برای کارکنان مناسب، انتخاب بین گزینه‌ها، آموزش‌های رسمی و غیر رسمی کارکنان تا زمانی که وی به سطح عملکرد مناسب برسند[۱۳]. علاوه بر این هزینه‌های جایگزینی، خروجی‌ها ممکن است تحت تاثیر برخی از این جایگزینی‌ها قرار بگیرند و هزینه‌های پنهانی را به سازمان تحمیل کنند. بسیاری از محققان بیان کرده‌اند، نرخ بالای ریزش کارکنان در صورتی که به درستی مدیریت نشوند، تاثیرات بسیاری را بر سودآوری سازمان دارند[۱۴].

خروج از سازمان هزینه‌های پنهان و نامشهود زیادی را بر سازمان تحمیل می‌کند و این هزینه‌های نامشهود ناشی از کارکنان جدید، همکاران کارکنان جدید، کارکنانی که مرتبط به افراد جداشده از سازمان هستند و ... است. تمامی موارد فوق سودآوری سازمان را تحت‌الشعاع قرار می‌دهند[۱۵]. بعبارت دیگر، ریزش کارکنان روی سرویش دهی به

ان و رضایت آنها اثر می‌گذارد[۱۶]. کاترین بیان کرد که ریزش کارکنان اثراتی از جمله از دست رفتن سودآوری، فروش، مدیریت زمان و ... می‌شود.

تخمین‌های محققان نشان می‌دهد استخدام و آموزش افراد جایگزین هزینه‌ای حدود ۵۰ درصد حقوق سالانه فرد را به خود اختصاص می‌دهد. ولی هزینه‌ها به همین میزان خلاصه نمی‌شود. زمانی که یک کارمند سازمانی را ترک می‌کند، فرض می‌شود بهره‌وری به دلیل اینکه منحنی آموزش مسنلزم درک شعلی و سازمانی است، کاهش می‌یابد. علاوه بر این، فقدان سرمایه‌های انتزاعی بر هزینه‌های تحمیلی می‌افزاید، چرا که نه تنها سازمان‌ها سرمایه‌ی انسانی خود را از دست می‌دهند، بلکه فرصت بالقوه‌ای برای رقبا بدست می‌آید تا این سرمایه‌ها را جذب کنند[۱۷]. بنابراین اگر ریزش کارکنان به درستی مدیریت نشود، ممکن است تاثیرات منفی از حیث هزینه-های پرسنلی در کوتاه مدت و میزان نقدینگی شرکت در بلند مدت داشته باشد[۱۸].

علاوه بر تحقیقات فوق، تحقیقاتی در زمینه‌ی استراتژی‌هایی جهت کاهش ریزش کارکنان انجام شده است که ذیلا به آن پرداخته شده است.

برای مقابله با ریزش‌های کارکنان و اثرات منفی ناشی از آنها، چندین استراتژی وجود دارد. این سیاست‌ها و استراتژی‌ها می‌توانند نسبت به استخدام، انتخاب، آموزش و القا، طراحی شغل و ... اتخاذ شوند. این سیاست‌ها باید دقیقا با مشکل پیش آمده هم‌خوانی داشته باشند و متناسب با آن انتخاب شوند[۱۸].

جدول ۱. دفعات استفاده از مدل های آماری و داده کاوی در مقاله ها

	سال چاپ							تعداد کل
	۲۰۰۳	۲۰۰۴	۲۰۰۵	۲۰۰۶	۲۰۰۷	۲۰۰۸	۲۰۰۹	
Neural Networks		۲	۳	۳	۲	۲	۳	۱۵
Decision Tree			۲	۱	۲	۳	۵	۱۳
Logistic Regression			۲	۲	۱	۳	۵	۱۳
Random Forests			۲			۲	۳	۷
Support Vector Machine						۳	۴	۷
Survival Analysis		۱			۱	۱		۳
Bayesian Network			۱			۲		۳
Self Organizing Maps		۱				۱		۲
AdaCost							۱	۱
Gradient Boosting Machine							۱	۱
Linear Discriminant Analysis						۱		۱
AdaBoost						۱		۱
Rough Set Theory						۱		۱
K-Nearest Neighbor					۱			۱
K-Means				۱				۱
Tailor-Butina				۱				۱
Time Series				۱				۱
ROCK			۱					۱
Regression Forests			۱					۱
Linear Regression			۱					۱
Association Rules		۱						۱
Sequence Discovery	۱							۱

فصل سوم

روش‌شناسی تحقیق

۱-۳. مقدمه

تصمیم‌گیری جوهر اصلی مدیریت است و عمل تصمیم‌گیری در واقع دشوارترین و در بعضی مواقع خطرناک‌ترین کار هر مدیر می‌تواند تلقی شود. یک مدیر با یک تصمیم‌گیری نادرست ممکن است صدمات جبران‌ناپذیری را بر پیکره سازمان خود وارد آورد.

درواقع مدیران باید بررسی کنند که هدف اصلی چیست و چگونه می‌توانند به آن دست یابند و از طرفی هدف پیشرو تا چه اندازه‌ای برای آنها مهم وقابل توجه است. مساله تصمیم‌گیری سخت‌تر میشود، زمانیکه سازمان دارای اهداف مختلف با اولویتهای متفاوت دارد و نیز خروجی‌های متفاوت از تصمیم‌گیری‌های مدیریتی انتظار می‌رود. شرط اولیه دستیابی به اهداف سازمان داشتن نیروی انسانی و کادر قوی است. در شرایطی که شرکت‌ها و سایر رقبا سعی در افزایش سودآوری دارند، عدم توجه به نیروی انسانی کارآمد، می‌تواند باعث عقب‌ماندگی و به مرور خروج از عرصه‌ی رقابت گردد. لذا اولین مسئله‌ی پیش روی مدیران داشتن نیروی انسانی کارآمد است که به مثابه سرمایه-های یک سازمان است.

از آنجا که هدف اصلی این پژوهش عبارتست از ‘‘مدیریت ریزش کارکنان با استفاده از تکنیک‌های داده‌کاوی’’ ، در این پژوهش از تکنیک راف‌ست برای داده‌کاوی جهت شناسایی افراد مستعد ریزش در مراحل اولیه‌ی جذب است تا از هزینه‌ها مشهود و نامشهود ریزش کارکنان در سازمان استفاده شود.

در این فصل به معرفی تکنیک‌های داده‌کاوی و نظریه راف ست، که از جمله متداول‌ترین تکنیک-های داده‌کاوی در پیش‌بینی ریزش کارکنان است، پرداخته شده است.

لازم به ذکر است که در این مدل، مجموعه متغیرهای توضیحی را با $\{X_1,X_2,...,X_N\}$، متغیر پاسخ را با Y، مقدار مشاهده شده متغیرهای توضیحی را با مجموعه $\{x_1,x_{2,...,}x_n\}$ و مقدار مشاهده متغیر پاسخ را با y نشان می‌دهیم.

۲-۳. نوع و روش تحقیق

براساس هدف پژوهش‌ها به پژوهش‌های بنیادی و کاربردی تقسیم می شوند.

پژوهش بنیادی :پژوهشی است که به کشف ماهیت اشیاء پدیده‌ها و روابط بین متغیرها، اصول، قوانین و ساخت یا آزمایش تئوری‌ها و نظریه‌ها می‌پردازد و به توسعه مرزهای دانش رشته علمی کمک می‌نماید. هدف اساسی این نوع پژوهش تبیین روابط بین پدیده ها، آزمون نظریه ها و افزودن به دانش موجود در یک زمینه خاص است.

پژوهش کاربردی :پژوهشی است که با استفاده از نتایج تحقیقات بنیادی به منظور بهبود و به کمال رساندن رفتارها، روش‌ها، ابزارها، وسایل، تولیدات، ساختارها و الگوهای مورد استفاده جوامع انسانی انجام می‌شود. هدف تحقیق کاربردی توسعه دانش کاربردی در یک زمینه خاص است.

تحقیق پیش‌رو به لحاظ هدف تحقیق از نوع تحقیقات کاربردی است.

پژوهش‌ها براساس نحوه‌ی گردآوری داده‌ها به پژوهش‌های توصیفی و آزمایشی تقسیم می شوند.

پژوهش توصیفی یا غیر آزمایشی شامل ۵ دسته است: پیمایشی، همبستگی، پس رویدادی، اقدام پژوهی، بررسی موردی

پژوهش آزمایشی به دو دسته تقسیم می شود: تحقیق تمام آزمایشی و تحقیق نیمه آزمایشی

این تحقیق براساس نحوه‌ی گردآوری داده‌ها از نوع تحقیقات توصیفی- پیمایشی است.

۳-۳. جامعه‌ی آماری

با مطالعات صورت گرفته پیرامون ریزش کارکنان، مشخص شد یکی از شرکت‌هایی که این مسئله آن را به چالش کشیده است، شرکت نفت و گاز پارس است. بنابر تحقیقات صورت گرفته در این شرکت افراد زیادی پس از مدتی کار تصمیم به ترک سازمان می‌گیرند. علت این امر را می‌توان در دلایلی از جمله کار اقماری، تامین نشدن بموقع نیروی انسانی به دلیل فرآیند طولانی جذب، علاقه نداشتن نیروی غیر بومی به کار در مناطق کمترتوسعه یافته و توسعه نیافته، جاذبه کاذب ایجاد شده در کشورهای همسایه و کمبود نیروی متخصص به دلیل استمرار نداشتن در سیستم جذب نیرو از دیگر مشکلات این سازمان در کشور است. این در حالیست که هزینه‌ی جذب و آموزش کارکنان در این شرکت به دلیل تخصصی بودن ماهیت کار بسیار بالاست. لذا خروج کارکنان از سازمان هزینه‌های مشهود و نامشهود بسیاری را بر سازمان متحمل می‌کند. در راستای کاهش نرخ ریزش در این شرکت اقداماتی از جمله به کارگیری نیروی‌های بومی، کاهش سن بازنشستگی به دلیل وجود فرسودگی‌های بالای شغلی و بازنگری در حقوق و مزایا به کار گرفته شده است. ولی جهت ریشه‌کن شدن این پدیده‌ی منفی در سازمان، باید از ابتدای امر کارکنانی که مستعد ریزش هستند شناسایی شده و از استخدام آنها ممانعت شود. لذا در این تحقیق با کسب اطلاعات از شرکت نفت و گاز پارس و بکارگیری تکنیک‌های داده‌کاوی سعی شده تا با ارائه‌ی الگوهایی، ریزش کارکنان پیش‌بینی شود.

۴-۳. نمونه و روش نمونه‌گیری

از بعد اقتصادی و زمانی، مطالعه‌ی کامل جوامع بزرگ، و یا حتّی متوسط امکانپذیر و مقرون به صرفه نیست و ناچار به نمونه‌گیری از جوامع، انجام پژوهش بر روی نمونه، و تعمیم نتایج نمونه به جامعه هستیم. این امر در صورتی بدون خطا است که نمونه نماینده‌ی مناسب جامعه باشد.

لذا در این تحقیق اطلاعات مربوط به ۵۸۴ نفر از کارکنان شرکت نفت و گاز پارس جهت اجرای مسئله جمع‌آوری شده است.

۵-۳. کشف دانش و نقش داده‌کاوی در آن

۱-۵-۳. مروری بر داده کاوی

به تازگی داده‌کاوی موضوع بسیاری از مقالات کسب‌وکار و مجلات و نرم‌افزارها شده است. با این حال، تنها چند سال پیش، فقط چند نفر واژه داده‌کاوی را شنیده بودند. هرچند داده‌کاوی شکل تکامل یافته‌ی رشته‌ای با سابقه‌ی طولانی است اما این واژه به خودی خود به تازگی و در دهه ۹۰ معرفی شده است.

داده‌کاوی پایگاه‌ها و مجموعه‌های حجیم داده‌ها را در پی کشف و استخراج دانش مورد تحلیل و کندوکاوی ماشینی قرار می‌دهد. این‌گونه مطالعات و کاوش‌ها را می‌توان همان امتداد و استمرار دانش کهن و همه‌جا گیر آمار دانست. تفاوت عمده در مقیاس، وسعت و گوناگونی زمینه‌ها و کاربردها، و نیز ابعاد و اندازه‌های داده‌های امروزی است که شیوه‌های ماشینی مربوط به یادگیری، مدل‌سازی و تعلم را طلب می‌کند.

اصطلاح داده‌کاوی همانطور که از ترجمه‌ی آن به معنی داده‌کاوی مشخص است به مفهوم استخراج اطلاعات نهان و یا الگوها و روابط مشخص در حجم زیادی از داده‌ها به یک یا چند بانک اطلاعاتی بزرگ است. بسیاری از شرکت‌ها و موسسات دارای حجم انبوهی از اطلاعات هستند. تکنیک‌های داده‌کاوی به طور تاریخی به گونه‌ای گسترش یافته‌اند که به سادگی می‌توان آنها را با ابزارهای نرم‌افزاری امروزی و موجود دراین موسسات تطبیق داده و از اطلاعات جمع‌آوری شده فعلی بهترین بهره را برد. در صورتی که سیستم‌های داده‌کاوی بر روی سیستمهای سرویس دهنده نصب شده باشد و دسترسی به بانک‌های اطلاعاتی بزرگ فراهم باشد، به کمک چنین سیستم‌هایی می‌توان به سوالاتی از قبیل :کدامیک از مشتریان ممکن است خریدار کدامیک از محصولات آینده شرکت باشد (چرا،در کدام مقطع زمانی)و بسیاری از موارد مشابه پاسخ داد.

در تعریفی، داده کاوی را یک رشته میان رشته ای دانسته و آنرا حاصل تلاش و همکاری مجموعه ای از رشته های آمار، هوش مصنوعی، یادگیری ماشین، پایگاه داده، بازیابی اطلاعات[۱۰] ، تشخیص الگو[۱۱] و تکنیک های تصویرسازی داده ها[۱۲] در طی یک پروسه دانسته اند (Kamber,2001 & Han).

همچنین در تعریفی دیگر از داده کاوی، آنرا فرایندی نامیده اند که طی آن الگوهای جالبی که به صورت آشکار جرئی از داده ها نیستند کشف می شوند (Frank & Witten). از این الگوها می توان به منظور پیش بینی استفاده نمود.

ریشه های داده کاوی از طریق دو مسیر به نیاکان خود بر می گردد. آمار کلاسیک یکی از این مسیر هاست. بدون آمار، داده کاوی وجود نخواهد داشت، آمار پایه اکثر فناوری هایی است که داده کاوی بر مبنای آنها ساخته شده است. آمار کلاسیک در بستر مفاهیمی مانند تحلیل رگرسیون، توزیع استاندارد، واریانس استاندارد، تجزیه و تحلیل تفکیکی[۱۳]، تجزیه و تحلیل خوشه ای و فاصله اطمینان[۱۴] و ... به مطالعه داده ها و روابط داده ها می پردازد. این مفاهیم و روش ها بلوک های تشکیل دهنده ساختمان بسیار پیشرفته تری هستند که با تجزیه و تحلیل آماری کلاسیک نقش مهمی ایفا می کند و می توان آمار کلاسیک را یکی از پایه های اصلی و بنیادین داده کاوی دانست.

هوش مصنوعی یکی دیگر از نیاکان داده کاوی است. هوش مصنوعی که اساس آن تکنیک های اکتشافی[۱۵] است و در برابر تکنیک های آماری قرار دارد، تلاشی است برای اعمال روشی مانند روش پردازش تفکر انسان بر مشکلات آماری از بین متدهای هوش مصنوعی، تکنیک های یادگیری ماشین در زمینه داده کاوی بیشتر مورد استفاده قرار گرفتند.

---

۱۰ Information Retrieval
۱۱ Pattern Recognition
۱۲ Data Visualization
۱۳ Discriminate Analysis
۱۴ Confidence Intervals
۱۵ Heuristc

در آنها تکنیک های اکتشافی هوش مصنوعی با تکنیک های پیشرفته آمار ترکیب می شوند. در یادگیری ماشین تلاش می شود تا به برنامه های کامپیوتری امکان داده شود از داده های خود جهت آموزش استفاده کنند و اصطلاحا آموزش[۱۶] داده شوند، مانند برنامه هایی که بر اساس کیفیت داده های ورودی تصمیمات مختلفی می گیرند.

داده کاوی، در بسیاری از زمینه ها اساسا اقتباس روش های یادگیری ماشین و اعمال آن ها به یک برنامه کسب و کار است. به عنوان بهترین توصیف، داده کاوی را می توان ترکیب پیشرفت های اخیر و قدیمی آمار، هوش مصنوعی و یادگیری ماشیت دانست. این تکنیک ها بصورت تجمعی مورد استفاده قرار می گیرند تا به مطالعه داده ها بپردازند تا روندها و الگوهای ناشناخته قبلی را کشف نمایند. داده کاوی در زمینه هایی مورد اقبال قرار گرفته است که می بایست داده های زیادی مورد تجزیه و تحلیل قرار گیرند تا روندهای خاصی کشف شوند و این روندها به هیچ روش یا ابزار دیگری قابل کشف شدن نباشند.

به طور عام می توان داده کاوی را فرآیند استخراج الگوها ار داده ها دانست. همچنان که داده ها به طور پیوسته در حال گرد آوری هستند، با نرخ دو برابر شدن حجم داده ها هر سه سال یکبار (Lyman& Hal R, 2003) و یا به برآوردی دیگر هر نه ماه یکبار (Fayyad, g & Uthurusamy, 2003) ، داده کاوی به طور فزاینده ای در حال تبدیل شدن به یک ابزار مهم برای تبدیل این داده ها به اطلاعات است. در تحقیقی که در سال ۲۰۰۵ ارایه گردید برآورد شد که حدود ۶ بیلیون دلار در فعایت های متن کاوی[۱۷] و داده کاوی سرمایه گذاری انجام گردد (Ebecken, 2005 & Zanasi, Brebbia). داده کاوی معمولا در طیف گسترده ای از شیوه های پروفایلینگ مانند بازاریابی، نظارت، ردیابی و کشف تقلب و اکتشافات علمی مورد استفاده قرار می گیرد.

---

[۱۶] Learn
[۱۷] Text Mining

تکنیک های داده کاوی را بر روی بازه گسترده ای از انواع مجموعه های داده ها نظیر پایگاه داده ای، انبار داده ای، داده های جغرافیایی، داده های مالتی مدیا، اینترنت، وب، داده های متنی و ... قابل اعمال است. شاید عبارت کشف دانش برای کل فرآیند استخراج الگو از داده ها مناسب تر باشد ولی واژه داده کاوی انتخاب شده و مورد اقبال قرار گرفته است (Andrassoya & Paralic, 1999).

زمانی که به داده کاوی به عنوان ابزاری برای کشف الگوهای موجود در مجموعه داده ای[۱۸] می نگریم، توجه به سه نکته بسیار مهم حائز اهمیت است. اولین نکته این است که با استفاده از داده ها و نمونه هایی[۱۹] که در دامنه (محدوده مساله مورد بررسی) حضور ندارد ممکن است نتایجی را تولید کند که از آن دامنه نباشند و دومین نکته این است که در صورتیکه داده ها موجود در دامنه در قالب مشخص و قابل کاوشی ارایه شده نباشند داده کاوی قادر نخواهد بود که اقدام به کشف الگو از آن داده ها نماید. این نکته به این معنی است که برای انجام یک داده کاوی موفق می بایست داده های مناسبی ارایه شود. داده کاوی نوعی ابزار است و مانند هر ابزار دیگری می باست برای نتیجه بخش بودن مواد اولیه مناسبی آماده شود.بدین منظور کاربر می بایست به عنوان اولین قدم از فرآیند داده کاوی مجموعه ای از داده های نماینده دامنه جمع آوری کند.سومین نکته مهم در فرآیند داده کاوی این است که کشف یک الگوی خاص در مجموعه ای خاص از داده ها لزوما به این معنی نیست که این الگو به طور عام نماینده کل جمعیت آماری مورد بررسی است.

بعضی از کاربردهای داده کاوی را می توان در کاربردهای معمول تجاری (مثل تحلیل و مدیریت بازار، تحلیل سبد بازار، پیش بینی قیمت نفت، بازاریابی هدف، فهم رفتار مشتری و تحلیل و مدیریت ریسک)، مدیریت و کشف فریب (کشف فریب تلفنی، کشف فریب های بیمه ای اتومبیل، کشف حقه های کارت اعتباری، کشف تراکنش های مشکوک مالی و پول شویی)،متن کاوی

[۱۸] Training Set
[۱۹] Sample

(خلاصه سازی، یافتن متون مشابه و کلمات کلیدی، پالایش نامه های الکترونیکی، گروه های خبری و غیره)، پزشکی (کشف ارتباط علامت و بیماری، تحلیل آرایه های DNA ، تصاویر پزشکی)، وب کاوی (پیشنهاد صفحات مرتبط، بهبود ماشین های جستجوگر یا شخصی سازی حرکت در وب سایت) و یافتن روندهای فرهنگی سیاسی در وب، تحلیل شبکه های اجتماعی وب (وبلاگها، ویکی ها) ، آنالیر ترافیک وب، تشخیص نفوذی به شبکه، متن کاوی، بیوانفورماتیک، سیستم پیشنهاد دهنده برای آموزش مجازی و کاربردهای بسیار دیگری در شاخه های مختلف مهندسی دانست. البته داده¬کاوی هر کاری را انجام نمی دهد و هر کار آماری را داده کاوی نمی نامند. برای داده کاوی شناخت و تحلیل داده ها مورد نیاز است، به طوری که بتوان روابط و الگوهای بین داده ها را با کمک افراد خبره پیدا کرد.

امروزه، بیشترین کاربرد داده کاوی در بانکها، مراکز صنعتی و کارخانجات بزرگ، مراکز درمانی و بیمارستانها، مراکز تحقیقاتی، بازاریابی هوشمند و بسیاری از موارد دیگر می باشد.

داده کاوی پل ارتباطی میان علم آمار ، علم کامپیوتر ، هوش مصنوعی ، الگوشناسی ، فراگیری ماشین و بازنمایی بصری داده می باشد. داده کاوی فرآیندی پیچیده جهت شناسایی الگوها و مدل های صحیح، جدید و به صورت بالقوه مفید، در حجم وسیعی از داده می باشد، به طریقی که این الگوها و مدلها برای انسانها قابل درک باشند. داده کاوی به صورت یک محصول قابل خریداری نمی باشد، بلکه یک رشته علمی و فرآیندی است که بایستی به صورت یک پروژه پیاده سازی شود.

کاوش داد ه ها به معنی کنکاش داده های موجود در پایگاه داده و انجام تحلیل های مختلف بر روی آن به منظور استخراج اطلاعات است.

داده کاوی فرایندی تحلیلی است که برای کاوش داده ها (معمولا حجم عظیمی از داده ها - در زمینه های کسب وکار و بازار) صورت می‌گیرد و یافته‌هابابه‌کارگیری الگوهایی،احراز اعتبار می‌شوند . هدف اصلی داده کاوی پیش بینی است. و به صورت دقیق تر میتوان گفت:

"کاوش داده ها شناسایی الگوهای صحیح، بدیع، سودمند و قابل درک از داده های موجود در یک پایگاه داده است که با استفاده از پرداز شهای معمول قابل دستیابی نیستند"

۶-۳. آشنایی با ادبیات موضوع

۱-۶-۳. داده کاوی

استفاده گسترده از بارکد برای تولیدات تجاری، به خدمت گرفتن کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع‌آوری داده، از اسکن کردن متون و تصاویر تا سیستم‌های سنجش از راه دور ماهواره‌ای، در این تغییرات نقش مهمی دارند.

به طور کلی استفاده همگانی از وب و اینترنت به عنوان یک سیستم اطلاع‌رسانی جهانی ما را مواجه با حجم زیادی از داده و اطلاعات می‌کند. این رشته انفجاری در داده‌های ذخیره شده، نیاز مبرم وجود تکنولوژی‌های جدید و ابزارهای خودکاری را ایجاد کرده که به صورت هوشمند به انسان، یاری می‌رسانند تا این حجم زیاد داده را به صورت اطلاعات و دانش تبدیل کند. داده کاوی به عنوان یک راه حل برای این مسائل مطرح می‌باشد. در یک تعریف غیررسمی داده کاوی فرایندی است، خودکار برای استخراج الگوهایی که دانش را بازنمایی می‌کنند، که این دانش به صورت قسمتی در پایگاه داده‌های عظیم انباره داده‌ها و دیگر مخازن بزرگ اطلاعات، ذخیره شده است. داده کاوی به طور همزمان از چندین رشته علمی بهره‌ها می‌برد. نظیر تکنولوژی ایجاد پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه‌های عصبی، آمار، شناسایی الگو، سیستم‌های مبتنی بر دانش، حصول دانش، بازیابی اطلاعات، محاسبات سرعت بالا و بازنمایی بصری داده . داده کاوی در اواخر دهه ۱۹۸۰ پدیدار گشته، در دهه ۱۹۹۰ گام‌های بلندی در این شاخه از علم برداشته

شده است. و انتظار می‌رود در این قرن به رشد و پیشرفت خود ادامه دهد. واژه‌های داده کاوی و کشف دانش در پایگاه داده اغلب به صورت مترادف یکدیگر مورد استفاده قرار می‌گیرد. کشف دانش در پایگاه داده فرایند شناسایی درست، ساده مفید و در نهایت الگوها و مدل‌های قابل فهم در داده‌ها می‌باشد. داده کاوی، مرحله‌ای از فرایند کشف دانش می‌باشد و شامل الگوریتم‌های مخصوص داده کاوی است. به طوری که تحت محدودیت‌های موثر محاسباتی قابل قبول، الگوها و یا مدل‌ها را در داده کشف می‌کند. به بیان ساده‌تر، داده کاوی به فرایند استخراج دانش ناشناخته، درست و بالقوه مفید از داده اطلاق می‌شود. تعریف دیگر اینست که داده کاوی گونه‌ای از تکنیک‌ها برای شنسایی اطلاعات و یا دانش تصمیم‌گیری از قطعات داده می‌باشد، به نحوی که با استخراج آنها، در حوزه‌های تصمیم‌گیری، پیش‌بینی، پیش‌گویی و تخمین مورد استفاده قرار گیرند. داده‌ها اغلب حجیم، اما بدون ارزش می‌باشند. داده به تنهایی قابل استفاده نیست. بلکه دانش نهفته در داده‌ها قابل استفاده می‌باشد. به این دلیل اغلب به داده کاوی، تحلیل داده‌ای ثانویه گفته می‌شود.

چه چیزی سبب پیدایش داده کاوی شده است؟

اصلی‌ترین دلیلی که باعث شده داده کاوی کانون توجهات در صنعت اطلاعات قرار بگیرد، مساله در دسترس بودن حجم وسیعی از داده‌ها و نیاز شدید به اینکه از این داده‌ها اطلاعات و دانش سودمند استخراج کنیم. اطلاعات و دانش بدست آمده در کاربردهای وسیعی از مدیریت کسب و کار کنترل تولید و تحلیل بازار تا طراحی مهندس و تحقیقاتی علمی مورداستفاده قرار می‌گیرد. داده کاوی را می‌توان حاصل سیر تکاملی طبیعی تکنولوژی اطلاعات دانست، که این سیر تکاملی ناشی از یک سیر تکاملی در صنعت پایگاه داده می‌باشد، نظیر عملیات جمع‌آوری داده‌ها و ایجاد پایگاه داده، مدیریت داده و تحلیل و فهم داده.

تکامل تکنولوژی پایگاه داده و استفاده فراوان آن در کاربردهای مختلف سبب جمع‌آوری حجم فراوانی داده شده است. این داده‌های فراوان باعث ایجاد نیاز برای ابزارهای قدرتمند برای تحلیل

داده‌ها گشته زیرا در حال حاضر به لحاظ داده ثروتمند هستیم ولی دچار کمبود اطلاعات می‌باشیم.

ابزارهای داده کاوی داده‌ها را آنالیز می‌کنند و الگوهای داده‌های را کشف می‌کنند که می‌توان از آن در کاربردهایی نظیر تعیین استراتژی برای کسب و کار، پایگاه دانش و تحقیقات علمی و پزشکی، استفاده کرد. شکاف موجود بین داده‌ها و اطلاعات سبب ایجاد نیاز برای ابزارهای داده کاوی شده است تا داده‌های بی‌ارزش را به دانشی ارزشمند تبدیل کنیم. به طور ساده داده کاوی به معنای استخراج یا مدل برداشتن از مقدار زیادی داده خام است، البته این نام‌گذاری برای این فرایند تا حدی نامناسب است، زیرا به طور مثال عملیات معدن کاری برای استخراج طلا از صخره و ماسه را طلاکاوی می‌نامیم. نه ماسه کاوی یا صخره کاوی، بنابراین بهتر بود به این فرایند نامی شبیه به استخراج دانش از داده می‌دادیم که متاسفانه بسیار طولانی است. دانش کاوی به عنوان یک عبارت کوتاه‌تر به عنوان جایگزین نمی‌تواند بیانگر تاکید و اهمیت بر معدن کاری زیاد داده باشد. معدن‌کاری عبارتی است که بلافاصله انسان را به یاد فرایندی می‌اندازد که به دنبال یافتن مجموعه کوچکی از قطعات ارزشمند از حجم بسیار زیادی از مواد خام هستیم. با توجه به مطالب عنوان شده، با اینکه این فرایند تا حدی دارای نامگذاری نقص است ولی این نامگذاری یعنی داده کاوی بسیار عمومیت پیدا کرده است. البته اساسمی دیگری برای این فرایند پیشنهاد شده است که بعضاً بسیار متفاوت با واژه داده کاوی است. نظیر استخراج دانش از پایگاه داده‌ها، استخراج دانش، آنالیز داده و الگو، باستان‌شناسی داده، و لایروپی داده‌ها.

۲-۶-۳. مراحل کشف دانش

کشف دانش دارای مراحل تکراری زیر است:

۱. پاکسازی داده‌ها (از بین بردن نویز و ناسازگاری داده‌ها)

۲. یکپارچه‌سازی داده‌ها (چندین منبع داده ترکیب می‌شوند)

۳. انتخاب داده‌ها (داده‌ها مرتبط با آنالیز داده بازیابی می‌شوند)

۴. تبدیل کردن داده‌ها (تبدیل داده‌ها به فرمی که مناسب برای داده کاوی باشد مثل خلاصه‌سازی و همسان‌سازی

۵. داده کاوی (فرایند اصلی که روال‌های هوشمند برای استخراج الگوها از داده‌ها به کار گرفته می‌شود)

۶. ارزیابی الگو (برای مشخص کردن الگوهای صحیح و موردنظر به وسیله معیارهای اندازه‌گیری)

۷. ارائه‌ی دانش (یعنی نمایش بصری، تکنیک‌های بازنمایی برای ارائه دانش کشف شده به کاربر استفاده می‌شود).

در هر مرحله داده کاوی باید به کار بر با پایگاه دانش تعامل داشته باشد. الگوهای کشف شده به کاربر ارائه می‌شوند و در صورت خواست او به عنوان دانش به پایگاه دانش اضافه می‌شوند. توجه شود که بر طبق این دیدگاه داده کاوی تنها یک مرحله از کل فرایند است. البته به عنوان تک مرحله اساسی که الگوهای منحنی را آشکار می‌سازد. با توجه به مطالب عنوان شده، در اینجا تعریفی از داده کاوی ارائه می‌دهیم:

داده کاوی عبارتست از فرایند یافتن دانش از مقادیر عظیم داده‌های ذخیره شده در پایگاه داده انبار داده و یا دیگر مخازن اطلاعات.

براساس این دیدگاه یک سیستم داده کاوی به طور نمونه دارای اجزای اصلی زیر است:

این اجزاء اصلی شامل:

۱- پایگاه داده، انبار داده یا دیگر مخازن اطلاعات: که از مجموعه‌ای از پایگاه داده‌ها و انبار داده‌ها، صفحات گسترده یا دیگر انواع مخازن اطلاعات، پاکساز ی داده‌ها و تکنیک‌های یکپارچه‌سازی روی این داده‌ها انجام می‌شود.

۲- سرویس‌دهنده پایگاه داده یا انبار داده، که مسئول بازیابی داده‌های مرتبط براساس نوع درخواست داده کاوی کاربر می‌باشد.

۳- پایگاه دانش: این پایگاه از دانش زمینه تشکیل شده تا به جستجو کمک کند، یا برای ارزیابی الگوهای یافته شده از آن استفاده می‌شود.

۴- موتور داده کاوی: این موتور جزء اصلی از سیستم داده کاوی است و به طور ایده‌آل شامل مجموعه‌ای از پیمانه‌هایی نظیر توصیف، تداعی، کلاسبندی، آنالیز خوشه‌ها و آنالیز تکامل و انحراف است.

۵- پیمانه ارزیابی الگو: این جزء معیارهای جذابیت

۶- واسط کاربر گرافیکی

۳-۶-۳. جایگاه داده کاوی در میان علوم مختلف

کاربردهای معمول تجاری: از قبیل تحلیل و مدیریت بازار، تحلیل سبد بازار، بازاریابی هدف، فهم رفتار مشتری، تحلیل و مدیریت ریسک

مدیریت و کشف ضریب: کشف ضریب تلفنی، کشف ضریب‌های بیمه‌ای و اتومبیل، کشف حقه‌های کارت اعتباری، کشف تراکنش‌های مشکوک مالی (پونستویی)

متن کاوی: پالایش متن (نامه‌های الکترونیکی و گروه‌های خبری و غیره)

پزشکی: کشف ارتباط علامت و بیماری، تحلیل آرایه‌های DNA (تصاویر پزشکی)

ورزش: آمارهای ورزشی

وب کاوی: پیشنهاد صفحات مرتبط بهبود ماشین‌های جستجوگر یا شخصی سازی حرکت در وب‌سایت

۴-۶-۳. داده کاوی چه کارهایی نمی‌تواند انجام دهد؟

داده کاوی فقط یک ابزار است و نه یک عصای جادویی داده کاوی به این معنی نیست که شما راحت به کناری بنشینید و ابزارهای داده کاوی همه کار را انجام دهد.

داده کاوی نیاز به شناخت داده‌ها و ابزارهای تحلیل و افراد خبره در این زمینه‌ها را از بین نمی‌برد. داده کاوی فقط به تحلیلگران برای پیدا کردن الگوها و روابط بین داده‌ها کمک می‌کند و در این مورد نیز روابطی که یافته می‌شود باید به وسیله داده‌های واقعی دوباره بررسی و تست گردد.

۵-۶-۳. داده کاوی و انبار داده‌ها

معمولاً داده‌هایی که در داده کاوی مورد استفاده قرار می‌گیرد از یک انبار داده استخراج می‌گردند و یک پایگاه داده یا مرکز داده‌ای ویژه برای داده کاوی قرار می‌گیرند.

اگر داده‌های انتخابی جزیی از انبار داده‌ها باشد بسیار مفید است چون بسیاری از اعمالی که برای ساختن انبار داده‌ها انجا می‌گیرد با اعمال مقدماتی داده کاوی مشترک است و در نتیجه نیاز به انجام مجدد این اعمال وجود ندارد.از جمله این اعمال پاکسازی داده‌ها می‌باشد. پایگاه داده مربوط به داده کاوی می‌تواند جزیی از سیستم انبار داده‌ها باشد و یا می‌تواند یک پایگاه داده جدا باشد.

ولی با این حال وجود انبار داده‌ها برای انجام داده کاوی شرط لازم نیست و بدون آن هم اگر داده‌ها در یک و یا چندین پایگاه داده باشند می‌توان داده کاوی را انجام دهیم و بدین منظور فقط کافیست داده‌ها را در یک پایگاه داده جمع‌آوری کنیم و اعمال جامعیت داده‌ها و پاکسازی داده‌ها را روی آن انجام دهیم. این پایگاه داده جدید مثل یک مرکز داده‌ای عمل می‌کنند.

۶-۶-۳. سابقه و تاریخچه داده کاوی

داده کاوی[۲۰] به بهره گیری از ابزارهای تجزیه و تحلیل داده‌ها به منظور کشف الگوها و روابط معتبری که تا کنون ناشناخته بوده‌اند اطلاق می‌شود. این ابزارها ممکن است مدلهای آماری،الگوریتم‌های ریاضی و روش‌های یاد گیرنده[۲۱] باشند. داده کاوی منحصر به گردآوری و مدیریت داده‌ها نبوده و تجزیه و تحلیل اطلاعات و پیش بینی را نیز شامل می‌شود.

علاوه بر پیشرفت ابزارهای مدیریت داده، افزایش قابلیت دسترسی به داده و کاهش نرخ نگهداری داده نقش ایفا می‌کند. در طول چند سال کذشته افزایش سریع جمع آوری و نگه داری حجم اطلاعات وجود داشته‌است. با پیشنهادهای برخی از ناظران مبنی بر آنکه کمیت داده‌های دنیا به طور تخمینی هر ساله دوبرابر می‌گردد. در همین زمان هزینه ذخیره سازی داده‌ها بطور قابل توجهی از دلار برای هر مگابایت به پنی برای مگابایت کاهش پیدا کرده‌است. مطابقا قدرت محاسبه‌ها در هر ۱۸ – ۲۴ ماه به دوبرابر ارتقاء پیدا کرده‌است این در حالی است که هزینه قدرت محاسبه رو به کاهش است. داده کاو به طور معمول در دو حوزه خصوصی و عمومی افزایش پیدا کرده‌است. سازمانها داده کاوی را به عنوان ابزاری برای بازدید اطلاعات مشتریان کاهش تقلب و اتلاف و کمک به تحقیقات پزشکی استفاده می‌کنند. با اینهمه ازدیاد داده کاوی به طبع بعضی از پیاده سازی و پیامد اشتباه را هم دارد.اینها شامل نگرانی‌هایی در مورد کیفیت داده‌ای که تحلیل

[۲۰] Data Mining
[۲۱] Machine Laming Method

می‌گردد، توانایی کار گروهی پایگاههای داده و نرم‌افزارها بین ارگانها و تخطی‌های بالقوه به حریم شخصی می‌باشد.همچنین ملاحظاتی در مورد محدودیتهایی در داده کاوی در ارگان‌ها که کارشان تاثیر بر امنیت دارد، نادیده گرفته می‌شود.

در حالیکه محصولات داده کاوی ابزارهای قدرتمندی می‌باشند، اما در نوع کاربردی کافی نیستند.برای کسب موفقیت، داده کاوی نیازمند تحلیل گران حرفه‌ای و متخصصان ماهری می‌باشد که بتوانند ترکیب خروجی بوجود آمده را تحلیل و تفسیر نمایند.در نتیجه محدودیتهای داده کاوی مربوط به داده اولیه یا افراد است تا اینکه مربوط به تکنولوژی باشد.

اگرچه داده کاوی به الگوهای مشخص و روابط آنها کمک می‌کند، اما برای کاربر اهمیت و ارزش این الگوها را بیان نمی‌کند.تصمیماتی از این قبیل بر عهده خود کاربر است.برای نمونه در ارزیابی صحت داده کاوی، برنامه کاربردی در تشخیص مظنونان تروریست طراحی شده که ممکن است این مدل به کمک اطلاعات موجود در مورد تروریستهای شناخته شده، آزمایش شود.با اینهمه در حالیکه ممکن است اطلاعات شخص بطور معین دوباره تصدیق گردد، که این مورد به این منظور نیست که برنامه مظنونی را که رفتارش به طور خاص از مدل اصلی منحرف شده را تشخیص بدهد.

از سال ۱۹۵۰ رایانه‌ها در تحلیل و ذخیره‌سازی داده‌ها به کار گرفته شدند. پس از حدود ۲۰ سال حجم داده‌ها دو برابر شد و پس از آن تقریبا هر دو سال یک‌بار همزمان با پیشرفت فناوری اطلاعات، حجم داده‌ها هم به دو برابر افزایش یافت. این پیشرفت آن قدر زیاد بود که تعداد رکورد-های برخی از پایگاه داده‌ها به چند صد میلیارد رسید. پدیده شبکه جهانی وب، استفاده گسترد ه از بارکد برای تولیدات تجاری، به خدمت گرفتن کامپیوتر در کسب و کار، خدمات الکترونیکی دولتی و پیشرفت در وسایل جمع‌آوری داد ه ، انفجاری را در مجموعه‌های اطلاعاتی سازمان‌ها و موسسات ایجاد کرد ه است. حجم زیاد اطلاعات، مدیران این مجموعه ها را در تحلیل و یافتن اطلاعات مفید

دچار چالش کرده است. داده‌کاوی، ابزار مناسب را برای تجزیه و تحلیل اطلاعات و کشف و استخراج روابط پنهان در مجموعه‌های داده‌ای سنگین فراهم می‌کند.

داده کاوی، فرآیند کشف الگوهای پنهان، جالب توجه، غیر منتظره و با ارزش از داخل مجموعه وسیعی از داده-هاست و فعالیتی در ارتباط با تحلیل دقیق داده های سنگین بی ساختار است که علم آمار ناتوان از تحلیل آنهاست. بعضی مواقع دانش کشف شده توسط داده کاوی عجیب به نظر می رسد؛ مثلا ارتباط افراد دارای کارت اعتباری و جنسیت با داشتن دفترچه تامین اجتماعی یا سن، جنسیت و درآمد اشخاص با پیش بینی خوش حسابی او در بازپرداخت اقساط وام. داده کاوی از علو می مانند یادگیری ماشین، هوش مصنوعی، آمار، پایگاه داده و شناسایی الگو به طور همزمان بهره گرفته و در حوزه های تصمیم گیری، پیش بینی، و تخمین مورد استفاده قرار می گیرد.

واژه کشف دانش در پایگاه داده ها[۲۲] در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، گسترده، سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است. این لغت به بیان دیگر به همه شیوه هایی اشاره دارد که هدف آنها پی بردن به ارتباط و نظم بین اطلاعات قابل مشاهد است. لغت KDD برای توصیف همه مراحل استخراج اطلاعات از پایگاه داده و نیز بیان اهداف کارهای اولیه کاربرد قوانین تصمیم گیری است. این واژه به طور رسمی اولین بار توسط Usama Fayaad در اولین کنفرانس بین المللی داده کاوی و کشف دانش که در سال ۱۹۹۵ در مونترال برگزار شده بود، معرفی شد که به بیان ارتباط تکنیکهای آنالیز در چندین مرحله با هدف استخراج دانش های ناشناخته قبلی از داده های در دسترس می پرداخت. داده هایی که ارتباط منظم و پراهمیت آنها قبلاً به نظر نمی رسید. کم کم واژه داده کاوی جای خود را پیدا کرد و مترادفی برای همه مراحل استخراج دانش شد. هر چند که داده کاوی مرحله ای از KDD است، اما در کل KDD فرآیند یافتن اطلاعات و الگوهای مفید از داده را گویند و داده کاوی بهره گیری از الگوریتم هایی برای یافتن اطلاعات مفید در فرآیند KDD است.

[۲۲] Knowledge Discovery in Database (KDD)

فرآیند KDD عبارت است از:

۱) پاک سازی و یکپارچه سازی داده(پیش پردازش داده)

۲) ایجاد یک انبار داده مشترک برای تمام منابع

۳) داده کاوی

۴) بصری سازی نتایج تولید شده

که مرحله پیش پردازش غالباً یکی از مراحل زمان بر و در عین حال بسیار مهم در کسب نتیجه مطلوب است.

در تعاریف قبلی جنبه بسیار مهمی که همان هدف نهایی داده کاوی است حذف شد ه است. هدف نهایی داده کاوی به دست آوردن نتایجی است که می تواند منافع کاری داشته باشد.

داده کاوی کاربرد سطح بالای فنون و ابزار به کار برده شده برای معرفی و تحلیل داده های تصمیم گیرندگان است. اصطلاح داده کاوی را متخصصین آمار، تحلیلگران داده ها و انجمن سیستم های اطلاعات مدیریت به کار برده اند در حالی که پژوهشگران یادگیری ماشین و هوش مصنوعی بیشتر از KDD استفاده می کنند. از نقطه نظر محققان، داده کاوی یک نظم نسبتاً جدید است که به طور عمد ه از میان مطالعاتی که به منظور نظم بخشیدن به برخی از فعالیتها همچون تخمین زدن، بازاریابی و سرشماری و آمار انجام گرفته، توسعه یافته است. اید ه ای که مبنای داده کاوی است یک فرآیند با اهمیت از شناخت الگوهای بالقوه مفید، تازه و درنهایت قابل درک در داده هاست. کشف دانش در پایگاه داده ها برای کشف اطلاعات مفید از مجموعه بزرگ داده هاست. دانش کشف شده می تواند قاعده ای باشد که با کمک آن ویژگی های داده ها، الگوهایی که به طور متناسب رخ می دهند، خوشه بندی موضوع های درون پایگاه داده ها و غیره را توصیف کند.

یک کاربر سیستم KDD به منظور انتخاب زیر مجموعه صحیحی از داده ها باید درک بالایی از قلمرو داده ها، رده مناسبی از الگوها و معیار خوبی برای الگوهای جالب داشته باشد. بنابراین سیستم KDD باید ابزارهایی با اثر تعاملی داشته باشد نه سیستم های تجزیه و تحلیل خودکار.

پژوهش جدی روی موضوع داده کاوی از اوایل دهه ۹۰ شروع شد. پژوهش ها و مطالعه های زیادی در این زمینه صورت گرفته؛ همچنین سمینارها، دوره های آموزشی و کنفرانس هایی نیز برگزار شد ه است. نتایج پایه های نظری داده کاوی در تعدادی از مقاله های پژوهشی آورده شد ه است. سال ۱۹۹۵ با استفاد ه از داده کاوی، انباره های داده بانک های آمریکا را بررسی کرده و بیان کردند که چگونه این سیستم ها برای بانک های آمریکا قدرت رقابت بیشتری ایجاد می کنند. در این سال انجمن داده کاوی همزمان با اولین کنفرانس بین المللی «کشف دانش و داده کاوی» شروع به کار و یک سازمان علمی به نام ACM- SIGKDD را تاسیس کرد. سال ۱۹۹۶ دیدگاهی از داده کاوی به عنوان «پرس و جو کننده از پایگاه های استنتاجی» پیشنهاد شد و فیاض و شاپیرو پیشرفت های کشف دانش و داده کاوی را اعلام کردند. همان سال دیدگا ه اقتصاد سنجی روی داده کاوی و عملکرد داده کاوی به عنوان یک مسأله بهینه ارائه و کنفرانس های ناحیه ای و بین المللی در مورد داده کاوی برگزار شد که از جمله می توان به کنفرانس آسیا و اقیانوسیه درباره کشف دانش و داده کاوی اشاره کرد. سال ۲۰۰۰ بحث های مقایسه ای بین آمار و داده کاوی و نیز استفاده از وب در کاوش داده ها و کاربردهای آن ارائه شد. سال ۲۰۰۲ «داده کاوی ساختارهای پیوند برای مدل رفتار مصرف کننده» عرضه شد.

از لحاظ تاریخی، توسعه داده کاوی را در طول زمان می توان به مراحل زیر تقسیم کرد:

مرحله اولیه: گردآوری وایجاد پایگاه اطلاعاتی (تا دهه ۱۹۶۰)

مرحله دوم: نظام های مدیریتی مبنی بر پایگاه اطلاعاتی (دهه۱۹۷۰ واوایل دهه ۱۹۸۰)

مرحله سوم: نظام های پایگاه اطلاعاتی پیشرفته (اواسط دهه ۱۹۸۰ تازمان حاضر)

مرحله چهارم: انبارش اطلاعات و داده کاوی (اواخر دهه ۱۹۸۰ تا به امروز)

مرحله پنجم: نظام پایگاه اطلاعاتی مبنی برشبکه (دهه ۱۹۹۰ تا کنون)

مرحله ششم: نسل نوین نظام های اطلاعاتی یکپارچه شده (از ۲۰۰۰ به بعد)

بدین ترتیب فعالیتی که از دهه ۱۹۶۰ شروع شده بود، در دهه ۱۹۹۰ گام های بلندی برداشت و انتظار می رود در این قرن به رشد و بالندگی خود ادامه دهد.

از هنگامی که رایانه در تحلیل و ذخیره‌سازی داده‌ها به کار رفت پس از حدود ۲۰ سال، حجم داده‌ها در پایگاه داده‌ها دو برابر شد، ولی پس از گذشت دو دهه و همزمان با پیشرفت فن‌آوری اطلاعات[۲۳] هر دو سال یکبار حجم داده‌ها، دو برابر شد. همچنین تعداد پایگاه داده‌ها با سرعت بیشتری رشد نمود.حال با وجود سیستم‌های یکپارچه اطلاعاتی سیستم‌های یکپارچه بانکی و تجارت الکترونیک، لحظه به لحظه به حجم داده‌ها در پایگاه داده‌های مربوط اضافه شده است.

انسان ها فرآیند استخراج الگوها از داده ها را برای قرن ها بصورت "دستی" انجام داده اند، اما با افزایش حجم اطلاعات در عصر مدرن ، لزوم استفاده از شیوه های خودکار امری کاملا بدیهی به نظر می رسد. متد های قدیمی و اولیه ای که در تشخیص الگوها در داده ها مورد استفاده قرار می گرفته اندعبارتند از قضیه بیز (دهه ۱۷۰۰) و تجزیه و تحلیل رگرسیون (دهه ۱۸۰۰) تکثیر سریع ، دسترسی آسان و قدرت و ظرفیت های روزافزون تکنولوژی های رایانه ای، حجم ذخیره سازی آنها را افزایش داده و این امکان را فراهم می آورد تا بتوان حجم بسیار زیادی از مجموعه های داده ای را ذخیره سازی کرد. با گذشت زمان همانطور که مجموعه داده ها در بعد اندازه و پیچیدگی رشد کرده اند، تجزیه و تحلیل مستقیم دستی[۲۴] داده ها به طور فزاینده با پردازش های غیرمستقیم و خودکار داده ها تکمیل (و نه جایگزین) شده است.این فرآیند به کمک سایر اکتشافات در علوم رایانه از قبیل شبکه های عصبی، خوشه بندی[۲۵]، الگوریتم های ژنتیکی (دهه ۱۹۵۰)، درخت

---

[۲۳] Information Technology
[۲۴] Manual
[۲۵] Clustering

تصمیم گیری (دهه ۱۹۶۰) و ماشین های بردار پشتیبانی[۲۶] (دهه ۱۹۸۰) انجام شده است. در تعریفی از داده کاوی فرآیند اعمال این روش ها به داده ها با هدف کشف الگوه های پنهان معرفی شده است (Kantradzic, 2003) داده کاوی برای سال های زیادی در کسب و کار های مختلف، دانشمندات و دولت ها استفاده شده است تا حجم انبوهی از داده ها را نظیر سوابق و تاریخچه سفر مسافران خطوط هوایی ، داده های سرشماری، داده های تولید شده توسط اسکنر سوپرمارکت ها و ... غربال کرده تا گزارش های متنوعی را تولید کنند. با این حال می بایست توجه شود که اینگونه گزارش ها همیشه نمی تواند به عنوان داده کاوی تلقی شود.

در جهت تعریف استاندارد های مربوط به داده کاوی تلاش هایی انجام گرفته است، برای مثال می توان به European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0) اشاره کرد که در سال ۱۹۹۹ انجام پذیرفت و یا Java Data Mining Standard (JDM 1.0) که در سال ۲۰۰۴ انجام شد. لازم به ذکر است که این دو استاندارد هایی هستند که در حال تکامل اند و نسخه های بعدی و کاملتر در دست توسعه است. مستقل از تلاش های انجام گرفته برای این استاندارد ها ، سیستم های نرم افزاری با منابع باز[۲۷] مانند RapidMiner، Weka، Knime و R project که آزادانه در دسترس هستند به یک استاندارد غیر رسمی برای تعریف فرآیند های داده کاوی تبدیل شده اند.بسیاری از این سیستم ها قادر به دریافت و ارسال[۲۸] مدل ها به PMML (Predictive Model Markup Language) هستند. PMML یک روش استاندارد برای نمایش مدل های داده کاوی ارایه می کند به طوری که این مدل ها بتواند در برنامه های مختلف آماری بصورت اشتراکی مورد استفاده قرار گیرند. PMML یک زبان مبتنی بر XML است که توسط گروه داده کاوی[۲۹] (DMG) که یک گروه مستقل و متشکل از بسیاری شرکت های داده کاوی است، توسعه داده شده و نسخه ۴.۰ آن در ژوئن ۲۰۰۹ منتشر گردید

---

[۲۶] Support Vector Machine

[۲۷] Open-Source

[۲۸] import & export

[۲۹] Data Mining Group

.(Guazzelli, lin,& (PMLL project page, Williams, 2009))

۷-۳. داده‌کاوی و نحوه کاربرد آن در ریزش کارکنان

داده‌کاوی پل ارتباطی میان علوم کامپیوتر، هوش مصنوعی[۳۰]، الگوشناسی[۳۱]، یادگیری ماشینی[۳۲] و مجسم‌سازی[۳۳] داده‌هاست که فرآیندهایی را جهت شناسایی الگوها و مدل‌هایی مفید، در حجم وسیعی از داده‌ها بکار می‌گیرد. بعبارت دیگر داده‌کاوی فرآیندی است که با استفاده از روش‌های هوشمند، دانش را از مجموعه‌ای از داده‌ها استخراج می‌کند. این روش، امروزه کاربردهای وسیعی در زمینه در حوزه‌های مختلف از جمله صنعت ارتباطات، بیمه، بانکداری، علوم زیستی و پزشکی پیدا کرده است. سیستم‌های داده‌کاوی تقریبا از اوایل دهه ۹۰ میلادی مورد توجه قرار گرفته‌اند. علت این امر نیز آن بود که تا آن زمان، سازمان‌ها بیشتر در پی ایجاد سیستم‌های عملیاتی کامپیوتری بودند که به وسیله آن‌ها بتوانندداده‌های موجود را سازماندهی کنند. پس از ایجاد این سیستم‌ها، روزانه حجم زیادی از اطلاعات جمع‌آوری می‌شد که تحلیل آنها از عهده انسان خارج بود. بررسی محققان نشان داده است که حجم داده‌هادر جهان، در هر ۲۰ ماه حدودا دو برابر می-شود. در یک تحقیق که بر روی گروه‌های تجاری بسیار بزرگ، صورت گرفت مشخص گردید که ۱۹ درصد از این گروه‌ها دارای پایگاه داده‌هایی با حجم بیش از ۵۰ گیگا بایت می‌باشند و ۵۹ درصد از آنها نیز انتظار می‌رود در آینده‌ای نزدیک در چنین سطحی از اطلاعات قرار گیرند. به

۳۰ Artificial Intelligence

۳۱ pattern Recognition

۳۲ Machine Learning

۳۳ Visualization

همین دلیل، نیاز به روش‌هایی بود که از میان انبوه داده‌ها، اطلاعات مفید یا دانش استخراج شود. بدین ترتیب داده‌کاوی در مسیر ایجاد و رشد قرار گرفت. در سال ۱۹۸۹ و ۱۹۹۱، کارگاه‌های کشف دانش از پیاگاه‌های داده توسط پیانتسکی[۳۴] و همکاران برگزار گردید. اصطلاح داده‌کاوی برای اولین بار توسط فیاد و همکاران در اولین کنفرانس بین المللی کشف "دانش و داده‌کاوی" در سال ۱۹۹۵ مطرح شد. از سال ۱۹۹۵، داده‌کاوی به صورت جدی وارد مباحث آمار شد و در سال ۱۹۹۶ اولین مجله "کشف دانش از پایگاه داده‌ها" منتشر شد. در حال حاضر، داده‌کاوی مهم‌ترین فن‌آوری جهت بهره‌برداری موثر از داده‌های حجیم بوده و اهمیت آن رو به فزونی است.

از منظرهای مختلفی می‌توان به موضوع الگویابی در داده‌ها نظاره کرد که هر یک از آنها دارای مبانی و روش‌های خاصی است. شکل ۲-۲ شامل فهرستی از عناوین کلی مربوطه است که در زیر به اختصار شرح داده شده‌اند.

پیوند[۳۵]

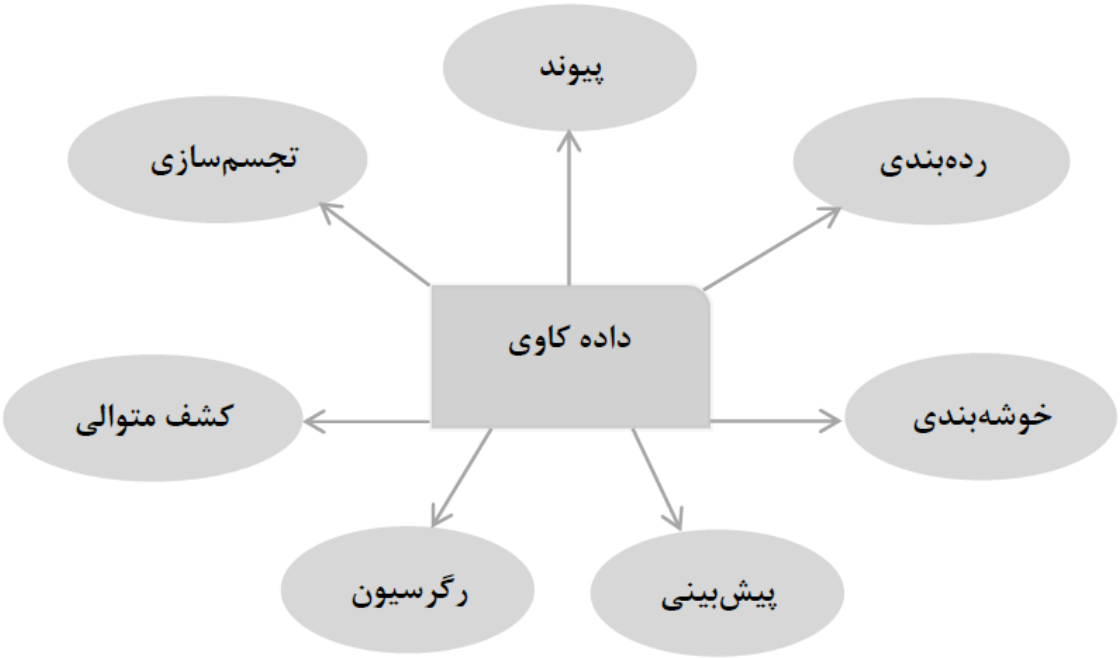
هدف یافتن ارتباط میان مجموعه‌ای از متغیرهاست. این ارتباط معمولا به صورت قوانین " اگر-آنگاه" از مجموعه داده‌ها استخراج می‌شود. به عنوان مثال می‌توان به شناسایی ارتباط میان اقلام یک فروشگاه بر اساس اطلاعات موجود در فاکتورهای فروش اشاره کرد. برنامه‌های تحلیل سبد بازار و باهم فروشی از جمله مثال‌هایی هستند که مدل‌های پیوند برای آنها استفاده می‌شود.

(دسته بندی خوشه‌های داده کاوی)

---

۳٤ Pianetsky

۳۵ Association



رده‌بندی[۳۶]

روش‌های رده‌بندی، با استفاده از اطلاعات موجود در مشاهدات، هریک از آنها را در رده‌های از پیش تعیین شده‌ای قرار می‌دهند (احمد، ۲۰۰۴). تحلیل ریزش مشتری از جمله مواردی ست که تکنیک‌های رده‌بندی در آن کاربرد دارد.

خوشه‌بندی[۳۷]

روش‌های خوشه‌بندی، مشاهداتی که دارای بیشترین شباهت هستند شناسایی کرده و درون یک خوشه قرار می‌دهند(احمد۲۰۰۴، کریبر و همکاران، ۲۰۰۳). تفاوت رده‌بندی و خوشه‌بندی در این است که در رده‌بندی، فرض می‌شود متغیر پاسخ دارای چند سطح بوده و هدف آن است که سطح مربوط به مشاهدات، براساس متغیرهای توضیحی، پیش‌بینی شود، اما در خوشه‌بندی وجود متغیر پاسخ موضوعیت نداشته و هدف ، تشخیص مشاهداتی است که به گونه‌ای، مشابه هستند.

پیش‌بینی[۳۸]

---

۳۶ Clasification

۳۷ clustering

۳۸ Forcasting

در پیش‌بینی، مقدار یک متغیر براساس یک الگوی از پیش تعریف شده‌ای، پیش‌بینی می‌شود. به عنوان مثال می‌توان به سری‌های زمانی اشاره کرد که در آن رفتار یک متغیر در طول زمان بررسی شده و رفتار آینده آن پیش‌بینی می‌شود.

رگرسیون[۳۹]

یک روش آماری برای شناسایی رابطه‌ی بین مجموعه متغیرهای توضیحی و متغیر پاسخ است. دراین روش، برای شناسایی این رابطه، از تصویر متغیر پاسخ در فضای ایجاد شده توسط مجموعه متغیرهای توضیحی استفاده می‌شود

کشف متوالی

برسون و همکاران(۲۰۰۰)، کشف متوالی را شناسایی روابط و الگوها در طول زمان تعریف کرده است. هدف از کشف متوالی، مدل‌سازی یک فرآیند در طول زمان به منظور شناسایی انحراف و روند در آن است.

تجسم‌سازی[۴۰]

تجسم‌سازی، نمایش داده‌ها به نحوی است که کاربر می‌تواند الگوها و روابط پیچیده در داده‌ها را مشاهده کند(شاو و همکاران، ۲۰۰۱)، روش‌های بصری‌سازی به موازات روش‌های دیگر داده‌کاوی برای درک الگوها و روابط پیچیده درون داده‌ها مورد استفاده قرار می‌گیرد.

هریک از موارد فوق به نحوی در مدیریت ریزش کارکنان مورد استفاده قرار می‌گیرند. به عنوان مثال، منظور از استفاده از روش‌های خوه‌بندی در مرحله شناسایی کارکنان این است که کارشناسان امر با استفاده از این روش‌ها، کارکنان را براساس اطلاعاتشان خوشه‌بندی کرده و آن خوشه‌هایی که از نظر آنها بهترین است شناسایی می‌کنند و برمبنای آنها عمل می‌کنند.

---

۳۹ Regression
۴۰ Visualization

برای پیش‌بینی ریزش کارکنان، روش‌های رده‌بندی داده‌کاوی به منظور استخراج دانش مورد نظر مناسب می‌باشند. در این پایان نامه قصد داریم با استفاده از ابزار رده‌بندی، کارکنان با ریزش محتمل را شناسایی کنیم. بدین منظور ابتدا مروری بر مطالعات قبلی خواهیم کرد.

۲-۷-۳. نظریه راف ست

نظریه راف ست در اوایل سال ۱۹۸۰ میلادی توسط پروفسور زدیسلاوپاولاک[۴۱] معرفی شد. این روش،الگوها و قوانین نهفته در مجموعه داده‌ها را به صورت قوانین اگر... آنگاه استخراج کرده و با استفاده از این قوانین، عمل رده‌بندی را برای یک مشاهده جدید انجام می‌دهد. همچنین این نظریه یک ابزار قدرتمند برای استدلال در موارد ابهام و نایقینی است و روش‌هایی را برای زدودن و کاستن اطلاعات و دانش نامربوط یا مازاد برنیاز از پایگاه‌های داده ارائه می‌دهد(پاولاک، ۱۹۸۲). در نهایت، نظریه راف ست با کاهش اطلاعات زائد، مجموعه‌ای از قواعد تلخیص شده پرمعنا را از داده-ها استخراج می‌کند که کار تصمیم‌گیرنده را بسیار ساده می‌کنند. لذا با توجه به رشد سریع حجم اطلاعات، نظریه راف ست نقش بسیار موثری را در سیستم‌های پشتیبان تصمیم‌گیری ایفا می-کند(زیارکو[۴۲]، ۱۹۹۳).

سرفصل مطالبی که برای نظریه راف ست باید به آنها پرداخته شود، به قرار زیر است:

۱. سیستم‌های اطلاعاتی و سیستم‌های تصمیم

۲. الگوریتم‌های تصمیم‌گیری

۳. دقت و کیفیت رده‌بندی الگوریتم تصمیم‌گیری

- روابط شباهت
- تقریب مجموعه‌ها
- دقت و کیفیت تقریب

---

۴۱ Zdzislaw Pawlak

۴۲ Ziarko

- دقت و کیفیت رده‌بندی

۴. کاهش متغیرها و هسته

۵. رده‌بندی یک مشاهده جدید

- معیارهای محافظ تصمیم
- عمل تصمیم‌گیری

۶. گسسته‌سازی

۱-۲-۷-۳. سیستم‌های اطلاعاتی و سیستم‌های تصمیم

در نظریه راف ست، سیستم‌های اطلاعاتی به جداولی اطلاق می‌شود که در سطرهای آن افراد یا اشیا و در ستون‌های آن متغیرها قرار دارند و به صورت رابطه زیر نمایش داده می‌شود.

$S=(U,C)$

که در آن U یک مجموعه ناتهی از اشیا یا افراد و C یک مجموعه ناتهی از متغیرهای شرطی(توضیحی) است که از نوع رده می‌باشند.

سیستم‌های اطلاعاتی به صورت $S=(U, C \cup D)$ را یک سیستم تصمیم گویند که در آن D $(D\neq C)$، مجموعه متغیرهای تصمیم(پاسخ) از نوع رده هستند(سورج[۴۳]، ۲۰۰۴). در مسئله مورد بررسی در این پایان نامه، مجموعه متغیرهای پاسخ دارای یک عضو به صورت $D=\{d\}$ است که دارای دو رده می‌باشد.

به ازای هر $a\in \{C \cup D\}$، مجموعه مقادیری که متغیر a اختیار می‌کند را دامنه a نامیده و با V_a نشان داده می‌شود. همچنین مقدار متغیر a برای فرد x *را با* $a(x)$ نمایش داده می‌شود.

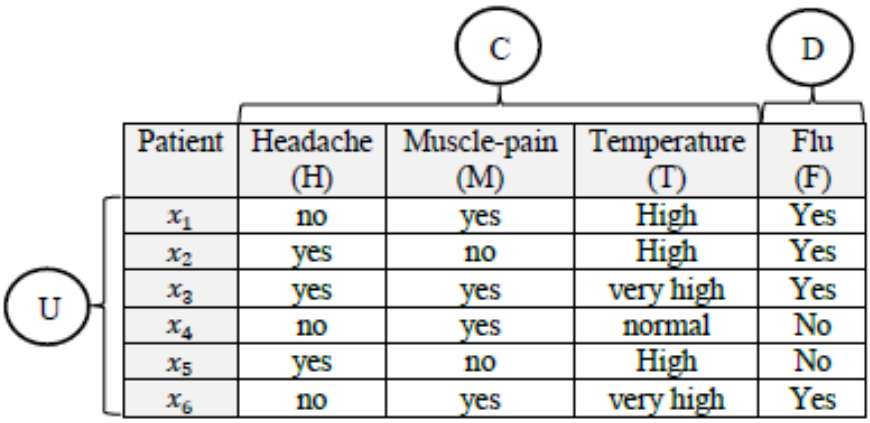
مثال ۳-۱

---

۴۳ Suraj

فرض کنید جدول تصمیمی به صورت جدول زیر داریم که حاوی اطلاعات مربوط به ۶ فرد است.

(جد.ل تصمیم)



Patient	Headache (H)	Muscle-pain (M)	Temperature (T)	Flu (F)
x_1	no	yes	High	Yes
x_2	yes	no	High	Yes
x_3	yes	yes	very high	Yes
x_4	no	yes	normal	No
x_5	yes	no	High	No
x_6	no	yes	very high	Yes

مجموعه‌های C,D,U به ترتیب مجموعه متغیرهای شرطی، متغیر پاسخ و کل افراد جامعه هستند. دامنه متغیرها عبارتند از:

$V_H=\{yes,no\}$

$V_M=\{yes,no\}$

$V_T=\{normal,high,very\ high\}$

$V_F=\{yes,no\}$

به عنوان مثال، مقدار متغیر H برای فرد x_a برابر با yes است که آن را به صورت $H(x_a)=yes$ نشان می‌دهند.

۲-۲-۷-۳. الگوریتم‌های تصمیم‌گیری

فرض کنید در سیستم تصمیم $S=(U, C \cup D)$، مجموعه‌های C و D به ترتیب دارای n و m عضئ به صورت $C=\{c_1,c_2,\ldots,c_n\}$ و $D=\{d_1,d_2,\ldots,d_m\}$ باشند، آنگاه برای هر $x \in U$، مجموعه مقادیر متغیرهای شرطی و تصمیم به صورت

$d_1(x),d_2(x),\ldots,d_m(x),c_1(x),c_2(x),\ldots,c_n(x)$ وجود خواهند داشت به طوریکه براساس آن‌ها قوانین تصمیم به صورت رابطه زیر حاصل می‌شود.

$$x\rightarrow_x D \qquad 1\text{-}3$$

که به آن قانون تصمیم ایجاد شده توسط فرد x در S گویند (پاولاک،۲۰۰۲). در خصوص مثال قبل با توجه به اینکه $n=3$ و $m=1$ است داریم:

$C=\{c_1,c_2,c_a\}$, $c_1=H$, $c_2=M$, $c_3=T$

$D=\{d_1\}$, $d_1=F$

برای فرد x_1 مقادیر متغیرهای شرطی و تصمیم عبارتنداز:

$c_1(x_1)=$ no, $c_2(x_2)=$ yes, $c_a(x_1)=$ high, $d_1(x_1)=$ yes

بنابراین قانون ایجاد شده توسط فرد x_1 عبارت است از:

$c_1(x_1)=$ no, $c_2(x_1)=$ yes, $c_a(x_1)=$ high$\rightarrow$ $d_1(x_1)=$ yes

یا بعبارت دیگر:

(1) H=no & M=yes & T=high $\rightarrow_{x1}$ F=yes

به همین صورت برای سایر افراد جامعه (U) نیز می‌توان قوانین تصمیم را به صورت زیر استخراج کرد:

(2) H=no & M=no & T=high $\rightarrow_{x2}$ F=yes

(3) H=yes & M=yes & T=very high $\rightarrow_{x3}$ F=yes

(4) H=no & M=yes & T=normal $\rightarrow_{x4}$ F=no

(5) H=yes & M=no & T=high $\rightarrow_{x5}$ F=no

(6) H=no & M=yes & T=very high $\rightarrow_{x6}$ F=yes

به مجموعه کلیه قوانین تصمیم استخراج شده از یک جدول تصمیم، در اصطلاح الگوریتم تصمیم-گیری گویند. گاهی اوقات ممکن است دو قانون به ازای زیر مجموعه‌ای از متغیرهای شرطی دقیقا مقادیر مشابهی را دارا باشند اما مقدار متغیر پاسخ آنها متفاوت باشند. در این صورت حالت عدم قطعیت در مورد این قانون‌ها پیش می‌آید. یعنی برای مشاهده جدیدی که از این قانون‌ها پیروی می‌کند، نمی‌توان بطور حتمی تصمیم گرفت که مقدار متغیر پاسخ آن چیست. به الگوریتم تصمیم‌گیری که تمامی قانون‌های قطعی باشد، الگوریتم تصمیم‌گیری سازگار، و در غیر اینصورت ناسازگار گویند.

حال در اینجا سئوال‌های زیر مطرح می‌شود:

۱. آیا می‌توان الگوریتم تصمیم‌گیری استخراج شده را به نجوی که دقت رده‌بندی آن کاهش نیابد، ساده‌تر و خلاصه‌تر کرد؟

۲. آیا با استفاده از الگوریتم تصمیم‌گیری استخراج شده، می‌توان عمل رده‌بندی را برای هر مشاهده جدید انجام داد؟

۳. زمانی که متغیرهای شرطی از نوع پیوسته(غیر رده) باشند، چگونه می‌توان الگوریتم تصمیم-گیری را استخراج نمود؟

نظریه راف ست برای هریک از سئوالات مطرح شده راه‌حلی ارائه می‌کند که در زیربخش‌های بعد به آنها می‌پردازیم.

۳-۲-۷-۳. دقت و کیفیت رده‌بندی الگوریتم تصمیم‌گیری

در نظریه راف ست به منظور ساده‌سازی جدول تصمیم و یا الگوریتم تصمیم‌گیری، ابتدا میزان قدرت رده‌بندی الگوریتم تصمیم‌گیری اولیه را با استفاده از معیارهایی محاسبه می‌کنند. سپس با استفاده از روشی که در زیربخش بعدی معرفی می‌شود، الگوریتم تصمیم‌گیری را به گونه‌ای ساده می‌کنند که دقت رده‌بندی الگوریتم جدید با دقت رده‌بندی الگوریتم اولیه برابر باشد. در این زیر بخش معیارهایی را جهت سنجش دقت رده‌بندی یک الگوریتم تصمیم‌گیری معرفی خواهیم کرد. بدین منظور ابتدا نیاز به ذکر بعضی از تعاریف پایه است که در زیر به آن می‌پردازیم.

روابط شباهت با هم ارزی

در جدول اطلاعات ممکن است خصوصیات بعضی از افراد به ازای زیرمجموعه‌ای از متغیرهای شرطی یکسان باشند که در اصطلاح به آنها افراد شبیه[۴۴] گفته می‌شود. در مثال قبل افراد x_2 و x_5 به ازای تمامی متغیرهای شرطی مقادیر یکسانی را دارا هستند. روابط شباهت، روابطی هستند که براساس آنها افراد شبیه از لحاظ زیرمجموعه‌ای از متغیرها، درون یک مجموعه قرار می‌گیرند. در سیستم‌های اطلاعاتی S=(U,C)، به ازای هر $B \subseteq C$ یک رابطه شباهت به صورت زیر تعریف می‌شود:

$$IND_s(B)=\{(x,x') \in u \times u | \forall a \in B, a(x) = a(x')\} \quad ۲-۳$$

که به آن در اصطلاح رابطه B– شبیه گویند و دارای خواص زیر است:

- خاصیت بازتابی: به ازای هر $x \in u$ ، $(x,x) \in IND_s(B)$
- خاصیت تقارنی: اگر $(x, x') \in IND_s(B)$ باشد، آنگاه $(x', x) \in IND_s(B)$
- خاصیت تعدی: اگر $(x, y) \in IND_s(B)$ و $(y, z) \in IND_s(B)$، آنگاه $(x, z) \in IND_s(B)$

[۴۴] Indiscernible

بنابراین اگر $(x, x') \in IND_S(B)$، آنگاه x و x' به ازای تمام متغیرهای درون مجموعه B، مقادیر یکسانی را دارا می‌باشند. بدیهی است $IND_s(B)$ یک رابطه هم‌ارزی است. خانواده‌ای از همه کلاس‌های هم‌ارزی $IND_s(B)$، یک افراز ایجاد شده توسط B است که آن را با نماد U/ $IND_s(B)$ با یه اختصار با U/B نشان می‌دهند. همچنین بلوکی از افراز U/B که شامل x است را با نماد $[x]_B$ نشان می‌دهند. زیر نویس s در رابطه هم‌ارزی، زمانی که سیستم اطلاعاتی برای تمامی روابط هم‌ارزی مشترک باشد، حذف می‌شود(سورج، ۲۰۰۴).

تعریف۱-۳. به بلوک‌های افراز U/B در اصطلاح یک مجموعه ابتدایی می‌گویند.

تقریب مجموعه‌ها

فرض کنید S=(U,C) یک سیستم اطلاعات است و فرض کنید $B \subseteq C$ و $x \subseteq U$. آنگاه تقریب پایین مجموعه x براساس مجموعه متغیر B، به صورت زیر تعریف می‌شود و شامل تمام افرادی است که به طور قطع براساس مجموعه متغیرهای B درون مجموعه x قرار می‌گیرند.

$\underline{B}(x|[x]_B \subseteq X)$ ۳-۳

تقریب بالای مجموعه x براساس مجموعه متغیر B، به صورت رابطه زیر تعریف می‌شود و شامل تمام افرادی است که ممکن است براساس مجموعه متغیرهای B، عضو مجموعه X باشند.

$$\bar{B}(X) = \{x|[x]_B \cap x \neq \phi\}$$ ۴ -۳

به اختلاف میان تقریب بالا و پایین مجموعه x براساس مجموعه متغیر B ناحیه مرزی گویند که به صورت زیر تعریف می‌شود و شامل مجموعه افرادی است که براساس مجموعه متغیرهایB نمی‌توان به طور قطع در مورد تعلق آنها به مجموعه x تصمیم‌گیری کرد.

$$BN_s(x) = \bar{B}(x) - \underline{B}(x)$$ ۵-۳

مجموعه U-$\underline{B}(x)$، شامل تمام افرادی است که به طور قطع، براساس مجموعه متغیر B، درون مجموعه X قرار نمی‌گیرند. اگر $BN_s(x)$ تهی باشد، مجموعه X، براساس مجموعه متغیر B دقیق است و در غیر اینصورت یک مجموعه مبهم است(سورج،۲۰۰۴).

دقت و کیفیت تقریب

دقیق یا مبهم بودن مجموعه $X \subseteq U$ براساس مجموعه متغیر $B \subseteq C$، را می‌توان با استفاده از معیارهای دقت و کیفیت تقریب که به ترتیب در روابط زیر آمده‌اند، بررسی کرد.

$$\alpha_\beta(x) = \frac{|\underline{B}(x)|}{|\overline{B}(x)|} \qquad ۶\text{-}۳$$

$$\rho_B(X) = \frac{|\underline{B}(x)|}{|x|} \qquad ۷\text{-}۳$$

که در آنها |.| تعداد اعضا مجموعه مربوطه است. با توجه به اینکه $\underline{B(X)} \subseteq X \subseteq \overline{B(X)}$ آنگاه $0 \leq \alpha_B(X) \leq \rho_B(X) \leq 1$.

اگر $\alpha_B(X) = 1$ آنگاه X براساس B دقیق و در غیر اینصورت مبهم است. همچنین $\rho_B(X) = 0$ اگر و تنها اگر $\alpha_B(X) = 0$، و $\rho_B(X) = 1$ اگر و تنها اگر $\alpha_B(X) = 1$.

دقت و کیفیت رده‌بندی

فرض کنید $S = (U, C \cup D)$ یک سیستم تصمیم و G=$\{X_1, X_2, \ldots, X_n\}$ مجموعه‌ای از زیرمجموعه‌های جامعه باشد که در آن $x_i \cap x_j = \emptyset$ و $U = \cup_{i=1}^{n} x_i$ هستند، آنگاه G یک افراز جامعه U است و هر کدام از آن x_i ها یک رده از این افراز محسوب می‌شوند.

به ازای هر $B \subseteq C$ تقریب بالا و پایین مجموعه G به ترتیب به صورت زیر تعریف می‌شوند:

۸-۳

$$\underline{B}(G) = \{\underline{B}(X_1), \underline{B}(X_2), \dots, \underline{B}(X_n)\}$$

$$\overline{B}(G) = \{\overline{B}(X_1), \overline{B}(X_2), \dots, \overline{B}(X_n)\}$$

آنگاه دقت و کیفیت رده‌بندی به ترتیب به صورت روابط زیر تعریف می‌شوند

۹-۳

$$\mu_B(G) = \frac{\sum_{i=1}^{n}|\underline{B}(X_i)|}{\sum_{i=1}^{n}|\overline{B}(X_i)|}$$

$$\partial_B(G) = \frac{\sum_{i=1}^{n}|\underline{B}(X_i)|}{|U|}$$

۴- ۲-۷-۳. کاهش‌ها و هسته متغیرها

در یک سیستم تصمیم، ممکن است وجود بعضی از متغیرها غیرضروری بوده و باعث بیهوده جدول و به دنبال آن پیچیدگی الگوریتم تصمیم‌گیری استخراج شده از آن شود. بنابراین یافتن متغیرهایی غیرضروری و حذف آنها از جداول تصمیم بسیار مهم و حائز اهمیت است. از این رو در این بخش، روش کاهش متغیرها را با استفاده از نظریه راف معرفی می‌کنیم.

تعریف ۲-۳. فرض کنید S=(U,C) یک سیستم تصمیم *بوده و* $B \subseteq C$ و $a \in B$ گوییم a متغیر غیر ضروری در B است اگر $IND_s(B)=IND_s(B-\{a\})$، در غیر اینصورت متغیر a در B ضروری است.

تعریف ۳-۳. مجموعه B مستقل است اگر همه متغیرهای آن ضروری باشند.

تعریف۴-۳. هر زیر مجموعه B' از B را یک کاهش مجموعه B نامند، اگر B' مستقل و $IND_s(B')=IND_s(B)$. بعبارت دیگر کاهش‌ها، به زیرمجموعه‌هایی از متغیرها اطلاق می‌شود که قادر هستند افراز یکسانی همانند مجموعه اصلی متغیرها، روی افراد جامعه ایجاد کنند. کاهش مجموعه متغیر B را با Red(B) نشان می‌دهند. متغیرهایی که متعلق به کاهش‌ها نیستند، غیرضروری محسوب می‌شوند. بدیهی است یک مجموعه متغیر می‌تواند کاهش‌های بسیاری داشته باشد.

تعریف۵-۳. به مجموعه تمام متغیرهای ضروری در B که $B \subseteq C$، هسته B می‌گویند و آن را با نماد Core(B) به صورت زیر نشان می‌دهند.

$$Core(B) = \cap Red(B) \qquad ۱۰-۳$$

که در آن $Red(B)$ مجموعه همه کاهش‌های B است. از آنجاییکه هسته اشتراک همه کاهش-هاست، از این‌رو مهم‌ترین زیرمجموعه از متغیرها است و حذف هر یک از اعضا آن قطعا در توان رده‌بتدی الگوریتم تصمیم‌گیری موثر خواهد بود.

بنابراین بدین طریق می‌توان یک الگوریتم تصمیم‌گیری را با حفظ قدرت رده‌بندی ساده کرد.

۵-۲-۷-۳. نحوه رده‌بندی یک مشاهده

هدف اصلی از معرفی نظریه راف ست، انجام عمل رده‌بندی برای یک مشاهده جدید است. پس از استخراج الگوریتم تصمیم‌گیری و ساده کردن آن نوبت به استفاده از آن برای رده‌بندی یک مشاهده جدید می‌رسد. با توجه به اینکه ممکن است مشخصات یک مشاهده جدید با چند قانون مختلف مطابقت داشته باشد، از این رو برای انتخاب بهترین قانون برای مشاهده و انجام عمل رده-بندی، نیاز به معرفی معیارهایی بمنظور سنجش میزان اهمیت هر قانون خواهیم داشت. بنابراین در ادامه، ابتدا معیارهای محافظ تصمیم‌گیری را معرفی کرده و سپس با استفاده از آنها عمل رده-بندی را برای یک مشاهده جدید انجام می‌دهیم.

معیارهای محافظ تصمیم

در سیستم‌های تصمیم‌گیری، معیارهایی تعریف می‌شوند که براساس آنها می‌توان به میزان قطعیت و اهمیت هر قانون پی برد. در ادامه برخی از این معیارهای مهم را معرفی می‌کنیم.

تعریف۶-۳. معیار پشتیبان قانون، عبارت است از تعداد دفعات تکرار یک قانون که به صورت رابطه زیر تعریف می‌شود.

$$Support_x(C,D)= |[x]_c \cap [x]_D| \quad (۱۱-۳)$$

تعریف ۷-۳. معیار توان قانون، به نسبت تعداد دفعات تکرار یک قانون به تعداد کل قانون‌ها گفته می‌شود که عبارت است از:

$$\sigma_x(C,D) = \frac{support_x(C,D)}{|U|} \quad (۱۲-۳)$$

تعریف ۸-۳. معیار قطعیت قانون عبارت است از:

$$cer_x(C,D) = \frac{support_x(C,D)}{|[x]_c|} = \frac{\sigma_x(C,D)}{\pi([x]_c)}, \pi([x]_c) = \frac{\pi([x]_c}{|U|} \quad (۱۳-۳)$$

اگر $cer_x(C,D) = 1$ باشد، آنگاه تصمیم مربوطه یک تصمیم قطعی و در غیر اینصورت غیرقطعی میباشد(پاولاک،۲۰۰۲)

تعریف ۹-۳. معیار پوشش قانون تصمیم، عبارت است از:

$$cov_x(C,D) = \frac{support_x(C,D)}{|[x]_D|} = \frac{\sigma_x(C,D)}{\pi([x]_D)}, \pi([x]_D) = \frac{|[x]_D|}{|U|} \quad (۱۴-۳)$$

قانونی که معیارهای محافظ تصمیم بزرگتری داشته باشد، اعتبار بیشتری دارد.

رده‌بندی مشاهده جدید

حال اگر یک مشاهده جدید داشته باشیم و بخواهیم رده متغیر پاسخ مربوط به آن را پیش‌بینی کنیم، ممکن است یکی از حالت‌های زیر برای آن اتفاق بیافتد:

۱. مشخصات متغیرهای شرطی مشاهده جدید، دقیقا با یکی از قانون‌های قطعی موجود در الگوریتم تصمیم‌گیری مطابقت داشته باشد.

۲. مشخصات متغیرهای شرطی مشاهده جدید، با یکی از قانون‌های غیرقطعی موجود در الگوریتم تصمیم‌گیری مطابقت داشته باشد.

۳. مشخصات متغیرهای شرطی مشاهده جدید، با هیچ یک از قانون‌های موجود در الگوریتم تصمیم‌گیری مطابقت نداشته باشد.

۴. مشخصات متغیرهای شرطی مشاهده جدید، با بیش از یک قانون موجود در الگوریتم تصمیم-گیری مطابقت داشته باشد.

برای حالت اول، پیش‌بینی به آسانی انجام می‌شود. برای حالت دوم قانون موجود غیرقطعی است، بنابراین این قانون به همراه چند قانون دیگر دارای مقادیر متغیرهای شرطی یکسان و مقدار متغیر تصمیم متفاوت هستند. از این رو تصمیم‌گیرنده از میان این قانون‌ها، قانونی را که از لحاظ معیارهای توان و پوشش دارای اعتبار بیشتری باشد به عنوان قانون پیش‌بینی برای مشاهده جدید انتخاب می‌کند. برای حالت چهارم همه قانون‌هایی که با مشخصات مشاهده جدید مطابقت دارند، می‌توانند به تصمیم‌گیرنده پیشنهاد شوند. اگر همه قانون‌ها رده مشابهی را انتخاب کنند، آنگاه هیچ ابهامی وجود نخواهد داشت، در غیر اینصورت معیارهای محافظ تصمیم، تعیین‌کننده بهترین قانون خواهند بود و مشابه حالت دوم، عمل پیش‌بینی صورت می‌پذیرد. تصمیم‌گیری برای حالت سوم نسبت به حالت‌های دیگر مشکل‌تر است. برای این حالت مجموعه‌ای از قوانین تصمیم که به مشخصات مشاهده جدید شبیه‌تر و نزدیک‌تر باشد به تصمیم‌گیرنده پیشنهاد می‌شوند. بدین منظور اسلوینسکی و همکاران روشی را تحت عنوان رابطه نزذیکی ارزشمند پیشنهاد کردند که با استفاده از آن مجموعه‌ای از قوانین که بیشترین شباهت را به مشخصات فرد مورد نظر دارند به تصمیم-گیرنده پیشنهاد می‌شود.

رابطه نزدیکی ارزشمند

فرض کنید x یک مشاهده و r یک قانون باشد. رابطه نزدیکی ارزشمند، به میزان نزدیکی یا شباهت مشاهده x به قانون r گفته می‌شود که با $(x\ S\ r)$ نشان داده می‌شود. از آنجاییکه ممکن است قانون r نسبت به یک مشاهده دارای شرط‌های کمتری باشد، بنابراین میزان نزدیکی قانون r به مشاهده x براساس متغیرهای موجود در قانون r مورد بررسی قرار می‌گیرد. اگر $A= \{a_1,a_2,$

$\{a_1, \ldots, a_m\}$ مجموعه متغیرهای شرطی موجود در قانون r باشد و $a_1^x, a_2^x, \ldots, a_m^x$ مقادیر این متغیرها برای مشاهده x باشند، آنگاه رابطه نزدیکی ارزشمند برای متغیر i ام به صورت رابطه زیر محاسبه می‌شود:

۱۵-۳

$$a_i^x \, S_i a_i^r \Leftrightarrow g_i(a_i^x) \geq g_i(a_i^r)$$

که در آن g_i یک تابع حقیقی مقدار است. با تعریف رابطه فوق مجموعه متغیرهای شرطی به دو زیر مجموعه جدا از هم با نام‌های اتحاد موافق و اتحاد ناسازگار تقسیم می‌شوند که به ترتیب به صورت روابط زیر تعریف می‌شوند:

$$C(x\ S\ r) = \{a_i \in A;\ g_i(a_i^x) \geq g_i(a_i^r)\} \qquad ۱۶\text{-}۳$$

$$D(x\ S\ r) = \{a_i \in A; g_i(a_i^x) < g_i(a_i^r)\} \qquad ۱۷\text{-}۳$$

به عبارت دیگر اتحاد موافق شامل متغیرهایی است که نزدیکی مشاهده x و قانون رابطه r را تایید می‌کنند و در مقابل، اتحاد ناسازگار شامل متغیرهایی است که نزدیکی مشاهدهx و قانون r را تایید نمی‌کنند. برای تایید یا رد گزاره‌ی $(x\ S\ r)$ دو شرط زیر باید برقرار شود:

۱. شرط توافق: $C(x\ S\ r)$ به قدر کافی با اهمیت باشد.

برای ارزیابی شرط توافق از شاخص توافق که به صورت رابطه زیر تعریف شده استفاده می‌شود:

$$c(x,r) = \frac{\sum_{a_i \in C(x\ S\ r)} k_i}{\sum_{a_i \in A} k_i} \quad , \quad c(x,r) \in [0,1] \qquad ۱۸\text{-}۳$$

که در آن k_i یک وزن مثبت برای متغیر a_i است. بعبارت دیگر $c(x,r)$ بیانگر اهمیت نسبی اتحاد $C(x\ S\ r)$ در مجموعه A است. با در نظر گرفتن یک مقدار آستانه برای شاخص توافق میزان اهمیت مورد نظر در شرط توافق مورد بررسی قرار می‌گیرد.

۲. شرط ناسازگاری: برای هر $a_i \in D(x\ S\ r)$ میزان ناسازگاری مشاهده x با قانون r به اندازه کافی بزرگتر باشد.

برای اعمال شرط دوم نیاز به معرفی یک مقدار آستانه v_i برای هر متغیر a_i داریم. به عبارت دیگر باید رابطه زیر برقرار باشد:

$$g_i(a_i^x) - g_i(a_i^r) < v_i \quad , \quad a_i \in D(x\ S\ r) \qquad ۱۹\text{-}۳$$

بتابراین به طور کلی برای تایید وجود رابطه ارزشمندی بین مشاهده x و قانون r داریم:

$$x\ S\ r \Leftrightarrow \begin{cases} c(x,r) \geq s \\ \text{و} \\ g_i(a_i^x) - g_i(a_i^r) < v_i \quad , \quad a_i \in D(x\ S\ r) \end{cases} \qquad ۲۰\text{-}۳$$

مقادیر مجهول در رابطه فوق باید توسط تصمیم‌گیرنده تعیین شوند. به همین صورت برای تمامی قانون‌های موجود، وجود رابطه نزدیکی ارزشمند را با مشاهده x بررسی می‌کنیم. در نهایت شاهد مجموعه‌ای از قوانین خواهیم بود که وجود شباهت بین آنها و مشاهده x مورد تایید قرار گرفته و تصمیم‌گیرنده می‌تواند براساس معیارهای محافظ تصمیم، یکی از این قانون‌ها را برای پیش‌بینی رده مشاهده جدید انتخاب کند.

۶-۲-۷-۳. گسسته‌سازی

نظریه راف‌ست تنهاقابلیت کار کردن با متغیرهای گسسته را دارد. اما در جداول و سیستم‌های تصمیم، ممکن است بعضی از متغیرها از نوع پیوسته باشند. لذا برای کار کردن با نظریه راف‌ست، لازم است این متغیرها به متغیر گسسته تبدیل شوند. به عمل تبدیل یک نتغیر پیوسته به گسسته، در اصطلاح عمل گسسته‌سازی گویند. گسسته‌سازی یک تکنیک پیش‌پردازش است که طی آن حدود تغییرات یک متغیر به زیر فاصله‌هایی تقسیم می‌شود که بتوانند بهترین رده‌بندی را برای اشیا یا افراد انجام دهند. فاصله‌های بلند موجب از دست دادن اطلاعات و به وجود آمدن اغتشاش داده‌ها در داده‌ها خواهند شد، اما برای رده‌بندی یک مشاهده جدید مناسب‌اند. فاصله‌های کوتاه اغشاش موجود در داده‌ها را کم می‌کنند، اما قدرت رده‌بندی یک مشاهده جدید را کاهش می-دهند. بنابراین انتخاب یک فاصله مناسب امری حائز اهمیت است. روش‌های مختلفی برای انجام عمل گسسته‌سازی وجود دارد که یکی از آنها روش مبتنی بر الگوریتم استدلال بولی است.

روش مبتنی بر استدلال بولی *(BRA)*

فرض کنید $S = (U, C \cup D, V, f)$ صورت دیگری از نمایش یک سیستم باشد، که در آن C= $\{a_1, \ldots, a_k\}$ ، U=$\{x_1, \ldots, x_k\}$ و D=$\{d\}$. مجموعه $v = U_{a \in c} v_a$ مجموعه مقادیر متغیرها است که در آن $v_a = [l_a, r_a) \subset R$ حدود تغییرات متغیر a است. برای هر $a \in c$ تابع $f = U \to v_a$ را خواهیم داشت که بیانگر مقدار متغیر a برای فرد x است.

تعریف. هر جفت (a,b) به عنوان عضوی از یک افراز روی v_a تعریف می‌شود که در آن $a \in c$ و $b \in R$ است. این جفت در اصطلاح یک برش روی v_a نامیده می‌شود.

فرض کنید $B_a = \{(a, b_1^a), (a, b_2^a), \ldots, (a, b_{k_a}^a)\}$، مجموعه برش‌های ممکن برای $a \in c$ باشد که در آن $l_a = [b_0^a, b_1^a] \cup [b_1^a, b_2^a] \cup \ldots \cup [b_{k_a}^a, b_{k_a+1}^a]$ هستند. این مجموعه برش‌ها، افرازی روی مجموعه v_a به صورت زیر ایجاد می‌کند:

$$B_a = [b_0^a, b_1^a] \cup [b_1^a, b_2^a] \cup \ldots \cup [b_{k_a}^a, b_{k_a+1}^a] \qquad ۲۱\text{-}۳$$

بنابراین مجموعه کل افرازهای ایجاد شده توسط متغیرهای شرطی عبارت است از:

$$B = \cup_{a \in C} B_a \qquad ۲۲\text{-}۳$$

مجموعه B، سیستم تصمیم اصلی را به سیستم تصمیم گسسته $S^B = (U, C \cup D, V^B, f^B)$ تبدیل می‌کند به طوری که:

$$\forall x \in U, a \in C, i \in \{0,1, \ldots, k_a\} \;\; ; \;\; f^B(x_a) = i \Leftrightarrow f(x_a) \in [b_i^a, b_{i+1}^a]$$

مجموعه برش‌های متفاوت، سیستم‌های گسسته متفاوت تولید می‌کنند. هدف اصلی در گسسته-سازی، انتخاب کوچکترین زیرمجموعه از برش‌ها است که بتوانند افراد جامعه U را مانند مجموعه اصلی برش‌ها از هم ممیزی کند.

حال باید نقاط ممکن برای انجام برش مشخص شوند. فرض کنید برای هر $a \in c$، دنباله $V_1^a < V_2^a < \cdots < V_{n_a}^a$ مجموعه مقادیر مرتب شده‌ای باشد که متغیرa می‌تواند اختیار کند، آنگاه متغیرهای بولی به صورت زیر تعریف می‌شوند:

۲۳-۳

$$\forall\, a \in C\ ,\ 1 \le l \le n_{a-1} :\ p_l^a = [V_l^a, V_{l+1}^a)$$

و مجموعه همه متغیرهای بولی جامعه را به صورت زیر نشان می‌دهند:

$$BV(U) = \{p_1^{a_1}, \dots, p_{n_{a-1}}^{a_1}, \dots, p_1^{a_k}, \dots, p_{n_{a-1}}^{a_k}\} \qquad ۲۴\text{-}۳$$

مجموعه همه برش‌های ممکن روی a را میتوان به صورت رابطه زیر تعریف نمود:

$$B_a = \left\{\left(a, \frac{V_1^a + V_2^a}{2}\right), \left(a, \frac{V_2^a + V_3^a}{2}\right), \dots, \left(a, \frac{V_{n_a-1}^a + V_{n_a}^a}{2}\right)\right\} \qquad ۲۵\text{-}۳$$

در گام بعدی باید زیر مجموعه‌ای از این برش‌ها به گونه‌ای انتخاب شود که بتواند تمامی افراد جامعه را از هم ممیز کند. برای پیدا کردن این برش‌ها از روشی مبتنی بر استدلال بولی استفاده شده است. این روش جدول S^* را از روی جدول S طی مراحل زیر می‌سازد.

۱. ستون‌های جدول S^* را مقادیر p_i^a ها، و سطرهای آن را جفت متغیرهایی با رده تصمیم متفاوت تشکیل می‌دهند. اگر جفت متغیر i ام توسط برش j ام به درستی از هم تمیز شوند، آنگاه درایه S_{ij}^* مقدار یک و در غیر این صورت مقدار صفر را می‌گیرد.

۲. تابع تمایز را برای تمام جفت افرادی که در سطرهای جدول قرار دارند طبق رابطه زیر به دست می‌آید.

$$\forall\ k = 1, \dots, n_a\ \ ,\ \ \psi(x_i, x_j) = \vee_{a \in C}\, \{p_k^a | p_k^a = 1\} \qquad ۲۶\text{-}۳$$

که در آن $\vee$ عملگر انفصال نامیده می‌شود. رابطه فوق به این معنی است که برای ممیز کردن جفت فرد (x_i, x_j) حداقل به یک برش در فاصله‌هایی که متغیر بولی آنها مقدار یک را گرفته، نیاز است.

۳. در پایان برای یافتن کوچکترین زیر مجموعه از متغیرهای بولی، تابع بولی را به صورت رابطه تعریف می‌کنند:

$$\Phi^U = \wedge \{\psi(x_i, x_j) : d(x_i) \neq d(x_j)\} \qquad ۲۷\text{-}۳$$

که در آن منظور از $d(x_i)$ مقدار متغیر پاسخ مربوط به متغیر x_i است. بااستفاده از قوانین جبر بولی، مجموعه‌های موجود در Φ^U تا حد امکان ساده شده و از میان مجموعه‌های باقیمانده، کوچکترین مجموعه به عنوان مجموعه برش‌های مناسب، انتخاب می‌شوند.

فصل چهارم

پیاده‌سازی مدل و استخراج نتایج

۴-۱. مقدمه

تجزیه و تحلیل داده‌ها فرآیند چند مرحله‌ای است که طی آن داده‌هایی که از طریق بکارگیری ابزارهای جمع‌آوری در نمونه آماری فراهم آمده‌اند، خلاصه، کدبندی، دسته‌بندی و در نهایت پردازش می‌شوند تا زمینه برقراری انواع تحلیل‌ها و ارتباط‌ها بین داده‌ها به منظور آزمون فرضیه‌ها فراهم آید. در این فرآیند داده‌ها هم از لحاظ مفهومی و هم از جنبه تجربی پالایش می‌شوند.

تجزیه و تحلیل داده‌ها برای بررسی صحت و سقم فرضیات برای هر نوع تحقیق از اهمیت خاصی برخوردار است. امروزه در بیشتر تحقیقاتی که متکی بر اطلاعات جمع‌آوری شده از موضوع مورد بررسی تحقیق می‌باشد، تجزیه و تحلیل اطلاعات از اصلی‌ترین و مهم‌ترین بخش‌های تحقیق محسوب می‌شود. در این بخش ساختار اجرایی و پیاده‌سازی روش تحقیق به تفصیل بیان شده است.

۴-۲. انتخاب جامعه‌ی هدف

همانطور که در فصل قبل مطرح شد یکی از شرکت‌هایی که این مسئله آن را به چالش کشیده است، شرکت نفت و گاز پارس است. بنابر تحقیقات صورت گرفته در این شرکت افراد زیادی پس از مدتی کار تصمیم به ترک سازمان می‌گیرند. علت این امر را می‌توان در دلایلی از جمله کار اقماری، تامین نشدن بموقع نیروی انسانی به دلیل فرآیند طولانی جذب، علاقه نداشتن نیروی غیر بومی به کار در مناطق کمترتوسعه یافته و توسعه نیافته، جاذبه کاذب ایجاد شده در کشورهای همسایه و کمبود نیروی متخصص به دلیل استمرار نداشتن در سیستم جذب نیرو از دیگر مشکلات صنعت حفاری در کشور است. این در حالیست که هزینه‌ی جذب و آموزش کارکنان در این شرکت به دلیل تخصصی بودن ماهیت کار بسیار بالاست. لذا خروج کارکنان از سازمان هزینه-های مشهود و نامشهود بسیاری را بر سازمان متحمل می‌کند. جهت ریشه‌کن شدن این پدیده‌ی منفی در سازمان، باید از ابتدای امر کارکنانی که مستعد ریزش هستند شناسایی شده و از

استخدام آنها ممانعت شود. لذا در این تحقیق با کسب اطلاعات از شرکت نفت و گاز پارس در مورد کارکنان استخدام شده و بکارگیری تکنیک‌های داده‌کاوی سعی شده تا با ارائه‌ی الگوهایی، ریزش کارکنان پیش‌بینی شود.

۴-۳. انتخاب نمونه‌ی آماری

داده‌های گردآوری شده از شرکت نفت و گاز پارس حاوی اطلاعاتی از ۵۸۴ نفر از کارکنان شرکت است که طبق تحقیقات به دست آمده در وفاداری و متقابلا ریزش کارکنان تاثیرگذار بوده‌اند.

۴-۴. تعیین پارامترهای ارزیابی کارکنان

تکنیک‌ها باید مجهز به پارامترهای کارآ و قابل تنظیم باشند، تا تعادل مطلوب بین جواب‌های بدست آمده و میزان محاسبات را برقرار نماید. همچنین تعیین پارامترهایی مناسب، دستیابی به پاسخ‌های بهینه‌تر را ممکن می‌سازد. بدیهی‌ست که پارامترها برای همه‌ی مسائل ثابت نمی‌باشند و باید پارامترهای سازگار با هر مسئله را برای تکنیک‌ها یافت.

در جدول زیر پارامترهای ارزیابی شرکت‌ها که مطابق با تحقیقات به دست آمده در وفاداری و عدم وفاداری کارکنان تاثیر دارد، نمایش داده شده است. از بین ۱۲ متغیر مشخص شده، ۱۱ متغیر آن توضیحی و یک متغیر آن به عنوان متغیر پاسخ در نظر گرفته شده است. مشخصات این متغیرها در جدول زیر بیان شده‌اند.

جدول ۲. متغیرها، نقش و نوع آنها

نام متغیر	نقش متغیر	نوع متغیر	نام متغیر	نقش متغیر	نوع متغیر
سن "Age"	توضیحی	عددی	اختیارات شغلی "*Authority*"	توضیحی	اسمی
جنسیت "Sex"	توضیحی	اسمی	حقوق و مزایا "*wage*"	توضیحی	عددی
وضعیت تاهل "Marit-status"	توضیحی	اسمی	سطح استرس شغل "*Stress*"	توضیحی	عددی
سطح تحصیلات "Educ"	توضیحی	اسمی	سابقه کار مرتبط و یا غیر مرتبط "*experience*"	توضیحی	اسمی
بومی و یا غیربومی بودن "native-expatriate"	توضیحی	اسمی	کار در مناطق کمتر توسعه یافته و یا شهرهای بزرگ "*Location*"	توضیحی	اسمی
اقماری یا روزکار بودن "work-stat"	توضیحی	اسمی	ریزش فرد "*Churn*"	پاسخ	اسمی

متغیر پاسخ، مقادیر ۱ و صفر را متناسب با ریزش و عدم‌ریزش کارکنان اختیار می‌کند. در مطالعات داده‌کاوی به مظور سنجش دقت مدل‌ها، معمولا مجموعه اصلی داده‌ها را به دو زیرمجموعه مجزا تقسیم نموده و یکی از آنها را به عنوان مجموعه داده مدل‌ساز و دیگری را به عنوان مجموعه داده آزمون در نظر می‌گیرند. در این مطالعه ۷۰٪ از داده‌های اصلی به عنوان داده‌های مدل‌ساز و ۳۰٪ مابقی به عنوان داده‌های آزمون در نظر گرفته شده است. طبق این تقسیم‌بندی، مجموعه داده‌های مدل‌ساز شامل ۴۷۵ مشاهده است که در ۹۰/۹درصد آنها یا به عبارت دیگر ۳۷۱مورد، متغیر پاسخ مقدار صفر و در ۹/۱ درصد مابقی یعنی ۳۷ مورد مقدار ۱ را اختیار کرده است و می‌توان گفت نسبت ریزش به عدم ریزش کارکنان ۱ به ۱۰ می‌باشد. هنگامی‌که یکی از سطوح متغیر پاسخ نسبت به سطوح دیگر فراوانی بسیار بیشتری داشته باشد، حالت عدم تعادل یا عدم بالانس پیش

می‌آید. در بروز چنین مواردی، الگوریتم‌های یادگیری و روش‌های آماری، عملکرد خوبی نداشته و نتایج قابل اعتمادی ارائه نمی‌دهند. برای حل این مشکل، روش‌های مختلفی پیشنهاد شده که با افزایش یا کاهش حجم داده‌ها، نسبت مشاهدات با رده‌های مختلف را به میزان قابل قبولی تعدیل می‌کنند. یکی از این روش‌ها، روش بیش‌نمونه‌گیری است. در این روش با بکارگیری الگوریتم‌های مختلف، فراوانی رده‌ای از متغیر پاسخ که در اقلیت است، افزایش می‌یابد. یکی از این الگوریتم‌ها، الگوریتم اسموت است که در زیر شرح داده شده است.

۱-۷-۳. الگوریتم اسموت

الگوریتم اسموت توسط چائولا و همکارانش معرفی شد. این الگوریتم با تولید مشاهدات ساختگی از رده اقلیت عمل بیش‌نمونه‌گیری انجام می‌دهد. فرض کنید در اینجا نیز مشابه فصل قبل $\{x_1,x_2, \dots , x_N\}$ مجموعه متغیرهای توضیحی، Y متغیر پاسخ، $\{x_1,x_2, \dots , x_N\}$ مقدار مشاهده متغیر پاسخ، $o_i=\{x_1(o_i), x_2(o_i), \dots, x_N(o_i)\}$ مجموعه مقادیر متغیرهای توضیحی می‌توانند از نوع متغیرهای گسسته و پیوسته باشند و متغیر پاسخ به صورت گسسته می‌باشند که دارای k رده است. اگر فرض کنیم رده l در اقلیت باشد و تعداد T مشاهده از n مشاهده دارای رده l باشند، آنگاه نحوه تولید مشاهدات مصنوعی به این صورت است که ابتدا باید هر یک از T مشاهده، k نزدیک‌ترین همسایگی از همان رده پیدا کنیم. سپس اگر بخواهیم $M\%$ بیش نمونه‌گیری انجام دهیم (یعنی تعداد مشاهدات با رده اقلیت را به اندازه $M\%$ افزایش دهیم) آنگاه باید برای هر مشاهده، از میان k تزدیک‌ترین همسایگی، M همسایگی را به طور تصادفی انتخاب کنیم. به عنوان مثال، اگر ۶۰ مشاهده از رده اقلیت داشته باشیم و بخواهیم تعداد مشاهدات این رده را دو برابر کنیم (یعنی بخواهیم بیش‌نمونه‌گیری ۲۰۰٪ انجام دهیم)، آنگاه باید ۲ همسایگی از میان k نزدیک‌ترین همسایگی برای هر یک از این مشاهدات به طور تصادفی انتخاب کنیم. حال اگر مشاهده ساختگی تولید شده توسط مشاهده iام و نزدیک‌ترین همسایگی jام را با o_{ij}^* نشان می-دهیم، آنگاه این مشاهده با استفاده از رابطه زیر تولید می‌شود:

$$\vec{o}^{*}_{i,j} = \vec{o}_i + \lambda(\vec{o}_j - \vec{o}_i) \quad , \quad i = 1,2,\dots,T \quad , \quad j \in \{1,2,\dots,k\}$$

یا بعبارت دیگر:

$$\begin{pmatrix} x_1(o^{*}_{i,j}) \\ x_2(o^{*}_{i,j}) \\ \cdot \\ \cdot \\ \cdot \\ x_N(o^{*}_{i,j}) \end{pmatrix} = \begin{pmatrix} x_1(o_i) \\ x_2(o_i) \\ \cdot \\ \cdot \\ \cdot \\ x_N(o_i) \end{pmatrix} + \lambda \begin{pmatrix} x_1(o_j) - x_1(o_i) \\ x_2(o_j) - x_2(o_i) \\ \cdot \\ \cdot \\ \cdot \\ x_N(o_j) - x_N(o_i) \end{pmatrix}$$

که در آن λ مقداری بین صفر و یک می‌باشد. بعنوان مثال اگر دو متغیر توضیحی داشته باشیم و $o_i=(6,4)$ مقادیر متغیرهای توضیحی مربوط به مشاهده‌ای از رده اقلیت باشد و فرض کنیم $o_i=(4,3)$ یکی از k نزدیک‌ترین همسایگی برای مشاهده o_i باشد، آنگاه با فرض $\lambda = 0.2$ مشاهده ساختگی o^{*}_{ij} به صورت زیر ساخته می‌شود:

$$\vec{o}^{*}_{i,j} = \begin{pmatrix} 6 \\ 4 \end{pmatrix} + 0.2 * \begin{pmatrix} 4-6 \\ 3-4 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix} + \begin{pmatrix} -0.4 \\ -0.2 \end{pmatrix} = \begin{pmatrix} 5.6 \\ 3.8 \end{pmatrix}$$

مقادیر λ و k باید توسط کاربر مشخص شوند و k حداقل باید برابر با M باشد. همچنین $M \geq 100$ می‌باشد.

حال با استفاده از الگوریتم اسموت، قصد داریم عمل بیش‌نمونه‌گیری را به ازای ۱۰۰٪ الی ۸۰۰٪ روی داده‌های مدل‌ساز انجام دهیم. برای انجام این کار از نرم‌افزار وکا که یک نرم‌افزار داده‌کاوی است، استفاده کردیم که نتایج آن به قرار جدول زیر است. از این پس به منظور سهولت در بیان، مجموعه داده‌های مدل‌سازی که عمل بیش‌نمونه‌گیری ۱۰۰٪ الی ۸۰۰٪ روی آنها انجام گرفته است را به ترتیب با نام‌های اسموت ۱۰۰٪، ۲۰۰٪، ۳۰۰٪، ۴۰۰٪ الی ۸۰۰٪ مورد استفاده قرار می‌دهیم.

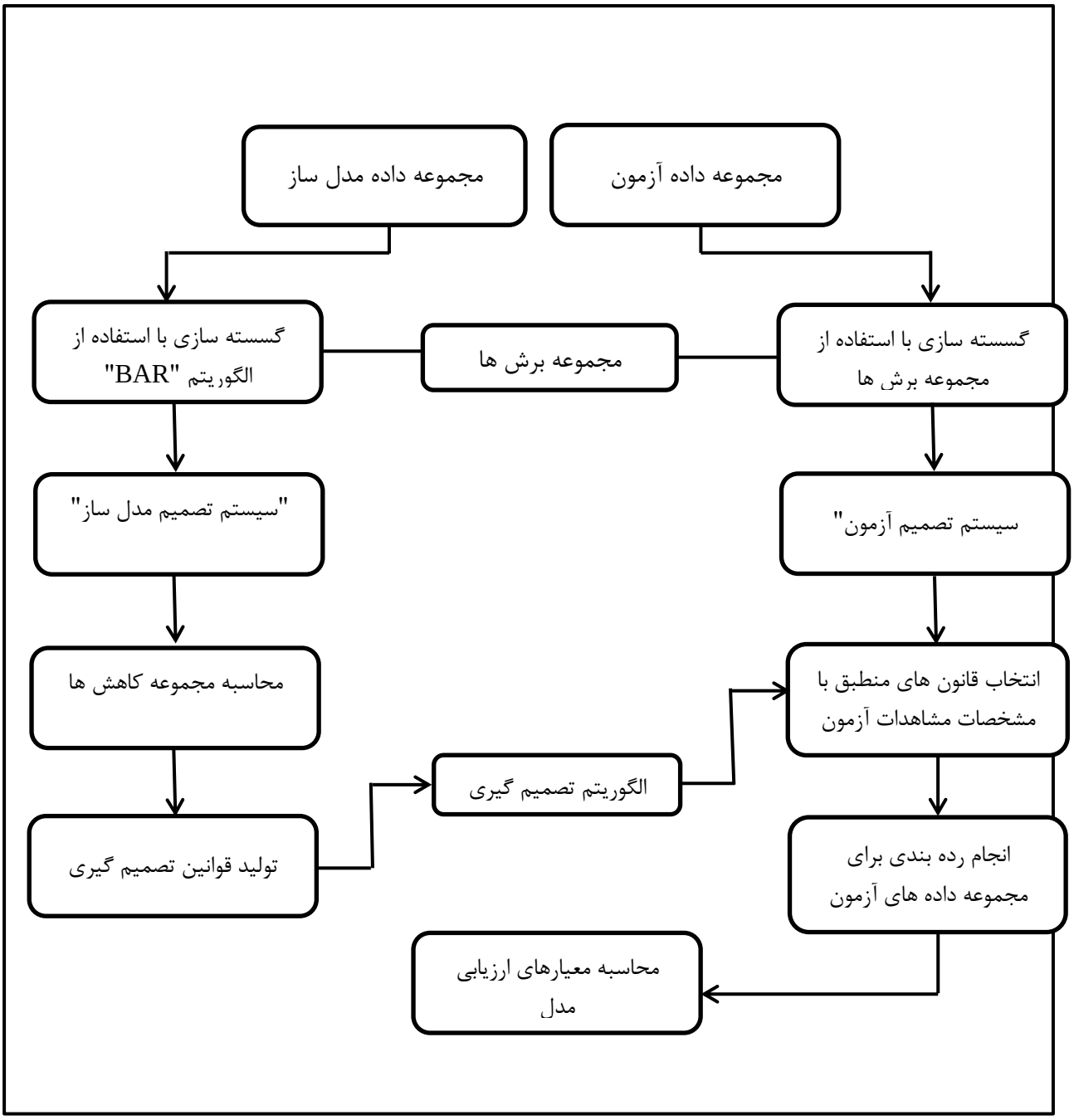
رده یک (اقلیت)	رده صفر (اکثریت)	تعداد مشاهدات

		داده مدل ساز
۳۷	۳۷۱	مجموعه داده اصلی مدل‌ساز
۷۴	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۱۰۰
۱۱۱	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۲۰۰
۱۴۸	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۳۰۰
۱۸۵	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۴۰۰
۲۲۲	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۴۰۰
۲۵۹	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۵۰۰
۲۹۶	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۶۰۰
۳۳۳	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۷۰۰
۳۷۰	۳۷۱	بیش‌نمونه‌گیری با اسموت ٪۸۰۰

علت اینکه بیش‌نمونه‌گیری را تا ٪۸۰۰ ادامه دادیم این است که با انجام اسموت ٪۸۰۰ نسبت مشاهدات با رده صفر و یک تقریبا برابر یک می‌شود و افزایش بیش از این مقدار الگوریتم اسموت موجب بیشتر شدن تعداد مشاهدات با رده یک از تعداد مشاهدات با رده صفر می‌شود.

۵-۴. نتایج حاصل از پیاده‌سازی الگوریتم نظریه راف ست

برای پیاده سازی مدل نظریه مجموعه مبهم بر مجموعه داده مدل ساز و استخراج نتایج، از نرم افزار "روزتا[۴۵]" استفاده کرده ایم که برای انجام این کار باید دنباله ای از مراحل طی شوند که در شکل زیر نمایش داده شده است.



رسم توضیحی ۱. فرآیند پیاده سازی مدل نظریه مجموعه مبهم

۴۵ ROSETTA

این نمودار بیانگر این است که برای پیاده سازی یک مدل نظریه مجموعه مبهم، ابتدا باید مجموعه داده مدل ساز و آزمون به شکل یک جدول یا سیستم تصمیم گیری درآیند.از آنجایی که بعضی از متغیرهای مورد استفاده در پیش بینی ریزش کارکنان از نوع کمی هستند، باید گسسته سازی را برای مجموعه داده مدل ساز انجام دهیم. سپس با استفاده از برش های ایجاد شده طی گسسته سازی روی مجموعه داده های مدل ساز، عمل گسسته سازی را برای مجموعه داده آزمون انجام دهیم. با انجام این دو عمل، سیستم تصمیم مدل ساز و سیستم تصمیم آزمون تولید می شوند. در مرحله بعد مجموعه کاهش را برای جدول تصمیم مدل ساز پیدا می کنیم. با استفاده از این کاهش ها، قانون های تصمیم گیری و به دنبال آن الگوریتم تصمیم گیری استخراج می شود. سپس از میان قانون های تصمیم گیری موجود، آن دسته از قانون هایی که با مشخصات مشاهدات موجود در مجموعه داده ی آزمون مطابقت دارند جدا شده و سایر قانون ها حذف می شوند.در نهایت عمل رده بندی، انجام شده و معیارهای ارزیابی مدل محاسبه می شوند. بدیهی است برای پیاده سازی مدل برهرکدام از مجموعه داده های مدل ساز (مجموعه داده اصلی، اسموت ۱۰۰٪، اسموت ۲۰۰٪، اسموت ۳۰۰٪، اسموت ۴۰۰٪، اسموت ۵۰۰٪،اسموت۶۰۰٪،اسموت۷۰۰٪،اسموت۸۰۰٪)، باید فرآیند پیاده سازی مدل نظریه مجموعه مبهم را یک بار تکرار کنیم. در نهایت پس از برازش مدل روی تمام مجموعه داده های مدل ساز، نتایج به صورت نمودارهای ۱تا ۴ خلاصه شده اند. در هر یک از این نمودارها، مقادیر معیارهای ارزیابی مدل به ازای مجموعه داده های مختلف قرار دارند و بهترین مدل با دایره روی نقطه متناظر با آن مشخص شده است. این مراحل به در ادامه به طور کامل شرح داده شده است.

در ابتدا لازم است که داده های خام را به اطلاعاتی تبدیل کنیم که برای نرم افزار قابل خواندن باشد. پس از وارد کردن اطلاعات، آنها را به دو مجموعه مدل ساز و آزمون تبدیل می کنیم که با توجه به اینکه این دو مجموعه به نسبت ۷۰ به ۳۰ تقسیم بندی شده اند، ۴۰۵ مورد در مجموعه داده مدل

ساز و ۱۷۵ مورد در مجموعه داده آزمون قرار میگیرند که این مرحله در شکل ۱ نشان داده شده است.

شکل۱. تولید مجموعه داده آزمون و مدل ساز

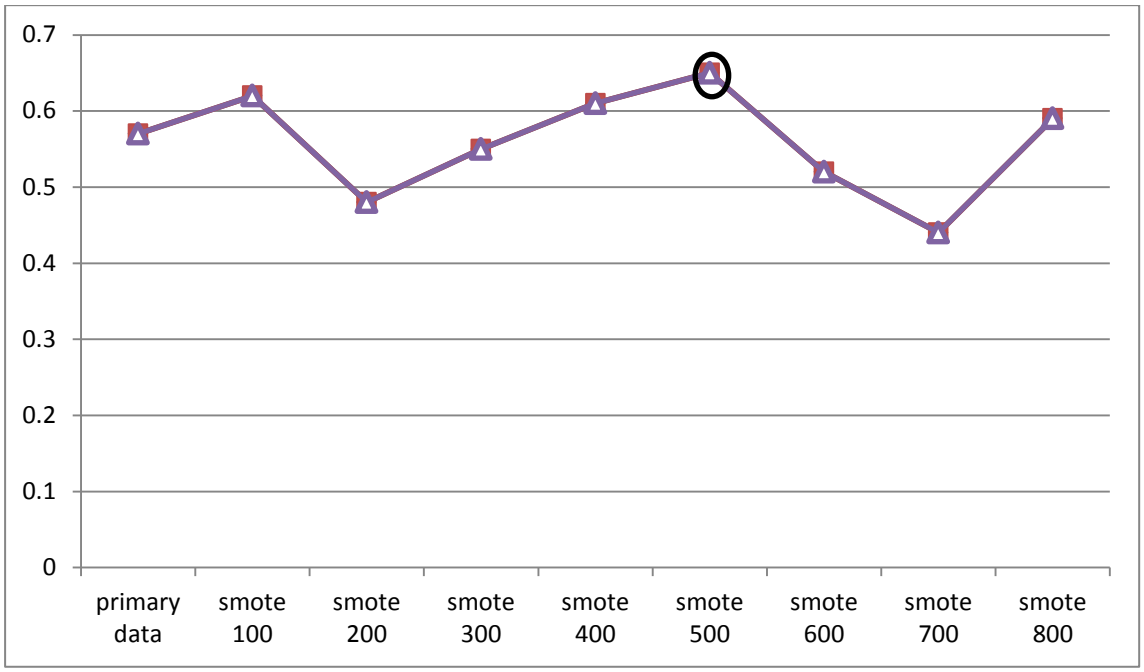
با توجه به کم بودن نسبت ریزش به عدم ریزش و برای افزایش دقت عملکرد ابزار داده کاوی عمل بیش نمونه گیری با کمک الگوریتم اسموت و نرم افزار وکا به میزان ۸۰۰٪ انجام میگیرد. پس از انجام این مرحله و با توجه به اینکه ابزار مورد استفاده تنها با داده های گسسته کار می کند، گسسته سازی را با استفاده از الگوریتم بولی(BAR) بر روی داده های مدل ساز انجام می دهیم که بخشی از خروجی در جدول ۳ آورده شده است.

	age	sex	marit-status	education	native	work-status	authority	wage	stress	experience	location	churn
1	[54, *)	male	married	BA	no	30day	yes	high	no	yes	tehran	0
2	[30, 34)	male	married	BA	yes	15day	yes	high	yes	no	asaluye	0
3	[27, 29)	male	married	BA	yes	15day	yes	med	no	no	asaluye	1
4	[*, 27)	male	single	BA	no	15day	yes	med	no	no	asaluye	0
5	[45, 48)	male	married	BA	no	30day	no	high	no	no	asaluye	0
6	[51, 52)	male	married	UN-BA	no	15day	yes	high	no	yes	asaluye	0
7	[*, 27)	male	single	UN-BA	yes	30day	no	low	yes	no	asaluye	0
8	[*, 27)	male	single	BA	yes	15day	yes	med	no	no	asaluye	0
9	[53, 54)	male	married	BA	no	15day	no	high	no	no	asaluye	1
10	[39, 40)	male	married	BA	yes	15day	yes	high	yes	no	asaluye	0
11	[30, 34)	male	married	MA	no	15day	yes	high	no	no	asaluye	1
12	[51, 52)	male	married	MA	no	30day	yes	high	no	yes	tehran	1
13	[27, 29)	male	single	BA	yes	15day	no	med	no	no	asaluye	0
14	[35, 39)	male	married	BA	no	30day	yes	high	no	no	asaluye	0
15	[35, 39)	female	married	BA	no	15day	no	high	no	no	asaluye	1
16	[40, 41)	male	married	MA	no	15day	yes	high	no	yes	asaluye	1
17	[30, 34)	male	married	BA	yes	15day	yes	med	no	no	tehran	0
18	[30, 34)	male	married	BA	no	15day	yes	high	no	no	asaluye	0
19	[*, 27)	male	single	BA	no	15day	no	low	no	yes	asaluye	0
20	[41, 45)	male	married	BA	yes	15day	yes	med	yes	no	asaluye	0
21	[45, 48)	male	married	MA	no	15day	yes	high	no	yes	asaluye	0
22	[54, *)	male	married	UN-BA	no	15day	no	high	no	no	asaluye	0
23	[39, 40)	male	married	MA	no	15day	yes	high	yes	no	asaluye	0
24	[41, 45)	male	married	MA	no	30day	yes	high	no	no	tehran	0
25	[45, 48)	male	married	BA	no	15day	yes	high	no	no	asaluye	0
26	[54, *)	male	married	MA	yes	15day	yes	high	no	no	asaluye	0
27	[45, 48)	male	married	DOC	no	15day	yes	high	no	no	asaluye	1
28	[30, 34)	female	married	MA	no	15day	no	high	no	no	asaluye	1
29	[30, 34)	male	married	BA	yes	15day	yes	med	no	yes	asaluye	0
30	[27, 29)	male	single	BA	yes	15day	yes	med	no	no	asaluye	0
31	[27, 29)	male	married	UN-BA	yes	15day	yes	low	no	no	asaluye	0
32	[30, 34)	male	married	MA	no	15day	no	med	no	no	asaluye	1
33	[35, 39)	male	married	BA	no	30day	yes	high	no	no	tehran	0
34	[30, 34)	male	married	BA	yes	15day	yes	med	no	no	asaluye	0
35	[35, 39)	male	married	BA	yes	30day	yes	med	no	no	tehran	0

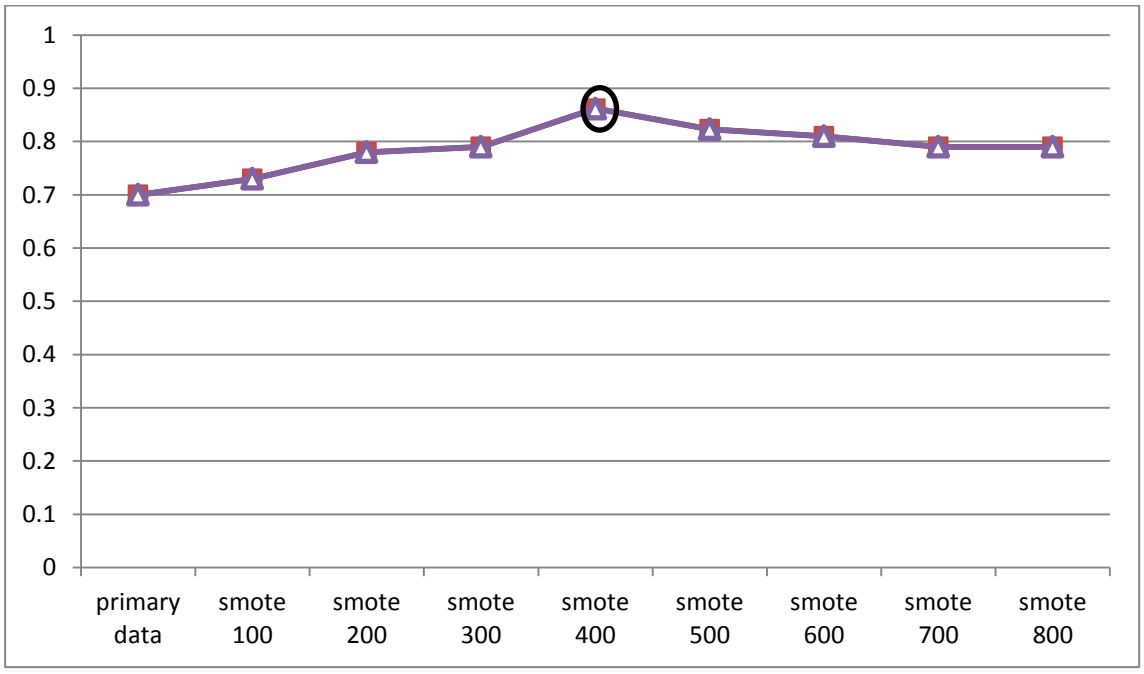
جدول ۳. گسسته سازی مجموعه داده مدل ساز

سپس با استفاده از برش های ایجاد شده طی گسسته سازی روی مجموعه داده های مدل ساز، عمل گسسته سازی را برای مجموعه داده آزمون انجام می دهیم. با انجام این دو عمل، سیستم تصمیم مدل ساز و سیستم تصمیم آزمون تولید می شوند.

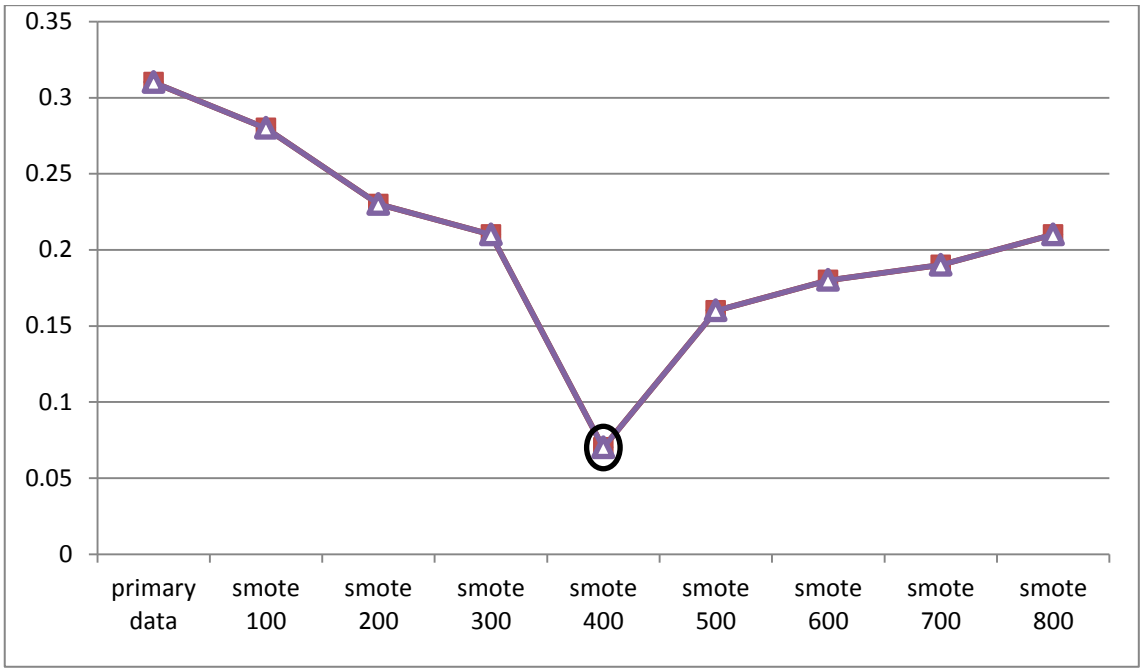
در نهایت پس از برازش مدل روی روی تمام مجموعه داده های مدل ساز، نتایج به صورت به صورت نمودارهای ۱تا ۴ خلاصه شده اند. در هریک از این نمودارها، مقادیر معیارهای ارزیابی مدل به ازای مجموعه داده های مختلف قرار دارند و بهترین مدل با دایره روی نقطه متناظر با آن مشخص شده است.



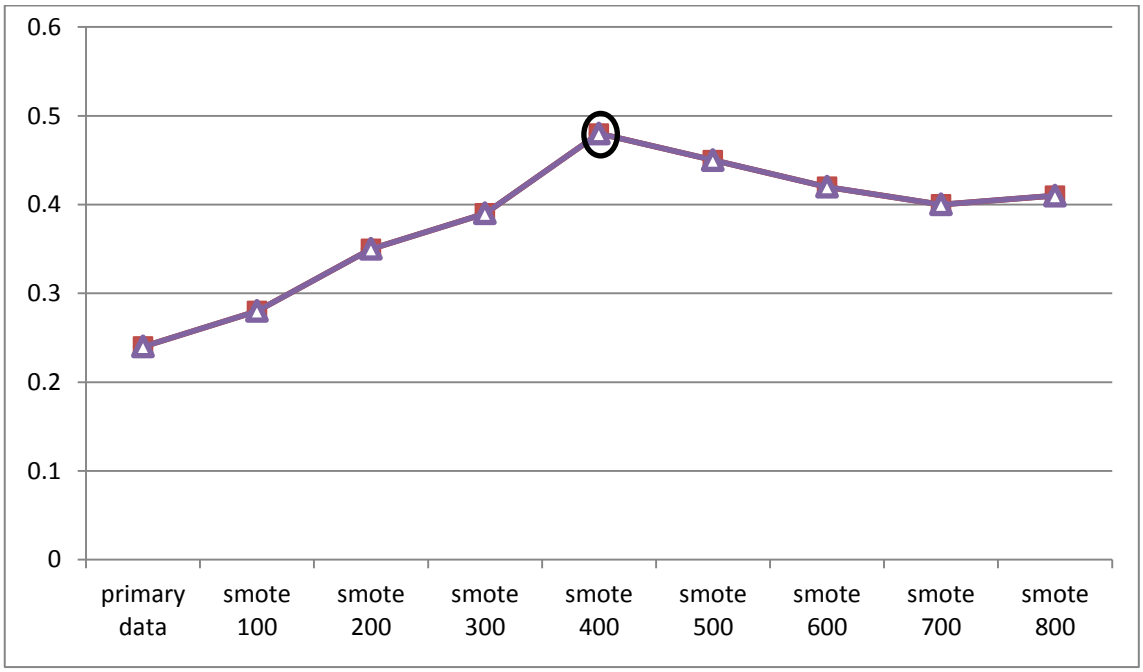
نمودار ۱.معیار ارزیابی مدل ***Recall***



نمودار ۲. معیار ارزیابی مدل ***Accuracy***



نمودار ۳. معیار ارزیابی مدل FPrate



نمودار ٤. معیار ارزیابی مدل *Precision*

با توجه به نمودارهای ۱ تا ۴ ، مقادیر بهینه معیارهای *Accuracy*، *Recall*، *Precision* و *Fprate*، به ترتیب به ازای میزان پیش نمونه گیری ۴۰۰٪، ۵۰۰٪، ۴۰۰٪ و ۴۰۰٪ به دست آمده است. حال بسته به اینکه برای کارشناسان مربوطه کدام رده از متغیر پاسخ مهم تر باشد می توان به صورت زیر میزان پیش نمونه گیری مناسب را پیشنهاد داد:

۱. اگر هزینه رده بندی اشتباه رده ۱(ریزش کارکنان) و رده صفر(عدم ریزش کارکنان) یکسان باشد، آنگاه مجموعه داده اسموت ۴۰۰٪ به دلیل دارا بودن بیشترین مقادیر معیارهای *Accurracy* و *Precision* به عنوان بهترین مجموعه پیشنهاد می شود.مدل برازش داده شده بر روی این مجموعه داده دارای *Accuracy* برابر با ۸۶/۲٪ می باشد.

۲. اگر هزینه رده بندی اشتباه رده ۱ (ریزش کارکنان) بیشتر از هزینه رده بندی اشتباه رده صفر (عدم ریزش کارکنان) باشد، آنگاه مجموعه داده اسموت ۵۰۰٪ به دلیل دارا بودن بیشترین مقدار *Recall*، به عنوان بهترین مجموعه داده پیشنهاد می شود. مدل برازش داده شده روی این مجموعه داده دارای *Accuracy* برابر با ۸۲/۳٪ می باشد.

در نهایت مجموعه اسموت ۴۰۰٪ و مجموعه اسموت ۵۰۰٪، به ترتیب به عنوان دو مورد پیشنهادی اول و دوم به کارشناسان ارائه می شود. مدل برازش داده شده بر این دو مجموعه داده به ترتیب از لحاظ معیار *Accuracy* دارای مقادیر ۸۸/۶٪ و ۸۲/۸٪ می باشند.

از آنجایی که در پیش بینی ریزش کارکنان مورد دوم پیشنهادی برای انتخاب کارشناسان محتمل تر است، لذا نتایج مربوط به برازش مدل نظریه راف ست را این مجموعه مورد بررسی قرار دادیم و پس از گسسته سازی این مجموعه و گسسته سازی مجموعه داده آزمون با استفاده از برش های ایجاد شده، سیستم تصمیم مدل ساز و سیستم تصمیم آزمون ساخته شد. در مرحله بعد مجموعه کاهش را برای جدول تصمیم مدل ساز پیدا کردیم که تنها کاهش صورت گرفته به صورت زیر به دست آمد.

Core(c)=Red(c)= {age, sex, marit-status, education, native, work-status, authority, wage, stress, experience}

با استفاده از این کاهش ها، قانون های تصمیم گیری و به دنبال آن الگوریتم تصمیم گیری استخراج می شود. بر این اساس ۲۱۴ قانون تصمیم گیری تولید شد که در زیر سه مورد از آن آورده شده است:

1. *age([30, 34)) AND sex(female) AND marit-status(married) AND education(MA) AND native(no) AND work-status(15day) AND authority(no) AND wage(high) AND stress(no) AND experience(no) => churn(1)*
2. *age([27, 29)) AND sex(male) AND marit-status(single) AND education(BA) AND native(yes) AND work-status(15day) AND authority(yes) AND wage(med) AND stress(no) AND experience(no) => churn(0)*
3. *age([41, 45)) AND sex(male) AND marit-status(married) AND education(BA) AND native(yes) AND work-status(15day) AND authority(yes) AND wage(med) AND stress(yes) AND experience(no) => churn(0)*

با استفاده از این قانون های تولید شده عمل رده بندی را برای مجموعه داده آزمون انجام دادیم که در نتیجه به مدلی با دقت ۷۲٪ از لحاظ معیار *Recall* رسیدیم.

فصل پنجم

نتیجه‌گیری و پیشنهادات

۵-۱. مقدمه

مدیران باید بررسی کنند که هدف اصلی چیست و چگونه می‌توانند به آن دست یابند و از طرفی هدف پیشرو تا چه اندازه‌ای برای آنها مهم وقابل توجه است. مساله تصمیم‌گیری سخت‌تر میشود، زمانیکه سازمان دارای اهداف مختلف با اولویتهای متفاوت دارد و نیز خروجی‌های متفاوت از تصمیم‌گیری‌های مدیریتی انتظار می‌رود. شرط اولیه دستیابی به اهداف سازمان داشتن نیروی انسانی و کادر قوی است. در شرایطی که شرکت‌ها و سایر رقبا سعی در افزایش سودآوری دارند، عدم توجه به نیروی انسانی کارآمد، می‌تواند باعث عقب‌ماندگی . به مرور خروج از عرصه‌ی رقابت گردد. لذا اولین مسئله‌ی پیش روی مدیران داشتن نیروی انسانی کارآمد است که به مثابه سرمایه-های یک سازمان است.

از آنجا که هدف اصلی این پژوهش عبارتست از ‘‘مدیریت ریزش کارکنان با استفاده از تکنیک‌های داده‌کاوی’’ ، در این پژوهش از تکنیک راف‌ست برای داده‌کاوی جهت شناسایی افراد مستعد ریزش در مراحل اولیه‌ی جذب است تا از هزینه‌ها مشهود و نامشهود ریزش کارکنان در سازمان استفاده شود.

در این فصل به معرفی تکنیک‌های داده‌کاوی و نظریه راف ست، که از جمله متداول‌ترین تکنیک-های داده‌کاوی در پیش‌بینی ریزش کارکنان است، پرداخته شده است.

۵-۲. نتیجه‌گیری

با توجه به یافته‌های تحقیق، کاربرد تکنیک‌های داده‌کاوی به ویژه تکنیک نظریه‌ی راف ست توصیه می‌شود.

در این پژوهش با استفاده از تکنیک نظریه راف ست و با استخراج اطلاعات از پایگاه داده‌های استخدامی شرکت نفت و گاز پارس ، تلاش شد تا براساس داده‌های تاریخی الگوهایی که برمبنای آن بتوان کارکنان مستعد ریزش را شناسایی کرد، شناسایی شوند تا مدیران منابع انسانی با استناد

به این الگوها بتوانند تصمیمات مناسب‌تری را در استخدام‌ها اتخاذ کنند و بدین طریق هزینه‌های مشهود و نامشهودی را که ریزش کارکنان بر سازمان تحمیل می‌کند، حذف و یا کنترل کنند.

نتایج حاصل از پیاده‌سازی تکنیک نظریه راف ست روی داده‌های جمع‌آوری شده از شرکت نفت و گاز پارس نشان می‌دهد، در صورتی که مدیران از الگوهای بدست آمده که چند مورد آن در زیر آورده شده است برای استخدام افراد استفاده کنند، تا حد زیادی می‌توانند کارکنانی را به کار بگیرند که تعهد سازمانی آنها بیشتر و احتمال ریزش آنها کمتر است.

چند مورد ازالگوهای بدست آمده:

1. *age([30, 34)) AND sex(female) AND marit-status(married) AND education(MA) AND native(no) AND work-status(15day) AND authority(no) AND wage(high) AND stress(no) AND experience(no) => churn(1)*
2. *age([27, 29)) AND sex(male) AND marit-status(single) AND education(BA) AND native(yes) AND work-status(15day) AND authority(yes) AND wage(med) AND stress(no) AND experience(no) => churn(0)*
3. *age([41, 45)) AND sex(male) AND marit-status(married) AND education(BA) AND native(yes) AND work-status(15day) AND authority(yes) AND wage(med) AND stress(yes) AND experience(no) => churn(0)*

۳-۵. کاربردهای تحقیق

استفاده از این تکنیک و دیگر ابزار داده کاوی جهت تجزیه و تحلیل اطلاعات موجود در پایگاه داده های سازمان ها بالاخص سازمان های خدماتی و سازمان های فرافن که به دلیل ریزش کارکنان متحمل ضرر و زیان زیادی می‌شوند، توصیه می‌شود.

۳-۵.پیشنهادات

به منظور انجام تحقیقات آینده در زمینه مدیریت ریزش کارکنان می توان موارد زیر را انجام داد:

۱.استفاده از دیگر الگوریتم های بیش نمونه گیری برای متعادل سازی داده ها

۲.استفاده از سایر مدل ها و تکنیک های داده کاوی جهت پیش بینی ریزش کارکنان

۳.استفاده از چندین تکنیک داده کاوی و مقایسه عملکرد و نتایج آنها

منابع و مآخذ

1. Abassi SM, Hollman KW (2000). “Turnover: the real bottom line”, Public Personnel Management, 2 (3) :333-342.
2. Basta N, Johnson E (1989). “ChEs are back in high demand”, Chem. Eng. 96 (8): 22-29.
3. Bluedorn AC (1982). “A unified model of turnover from organizations”, Hum. Relat. 35: 135-153.
4. Brooke PP, Russell DW, Price JL (1988). “Discuss validation of measures of job satisfaction, job involvement and organizational commitment”, J. Appl. Psychol. 73 (2) : 139-145
5. Kalliath TJ, Beck A (2001). “Is the path to burnout and turnover paved by a lack of supervisory support: a structural equations test”, New Zealand J. Psychol. 30: 72-78.
6. Locke E (1976). “The nature and causes of job satisfaction”,in Dunnette. MD (Eds). Handbook of Industrial and Organizational Psychology, Rand McNally, Chicago, IL, pp. 1297-1349.
7, Mano Rita –Negrin, Shay S Tzafrir (2004). “Job search modes and Turnover” Career development international. (5): 442-446
8. Catherine M Gustafson (2002). “staff turnover: Retention”. International j. contemp. Hosp. manage. 14 (3) : 106-110.
9. Phillips DJ (1990). “The price tag on turnover”, Pers. J. pp. 58-61. Porter LW, Steers RM, Mowday RT, Boulian PV (1974). “Organizational commitment, job satisfaction, and turnover among psychiatric technicians”, J. Appl. Psychol. 59: 603-609.
10. Tor Guinmaraes JE Owen (1997). “Assessing employee turnover intentions before and after TQM” International J. Qual. Reliability manage. 14 (1): 46-63.
11. Griffeth RW, Hom PW, Gaertner S (2000). “A meta-analysis of antecedents and correlates of employee turnover: update, moderator tests, and research implications for the next millennium”, J. Manage. 26 (3): 463-88.
12. Schervish PG (1983). The structural Determinants of unemployment, Vulnerability and power in market relations, academic press, New York, NY. Pp. 71-112.
13. Hackman, JR, Oldham GR (1975). “Development of the job diagnostic survey” J. Appl. Psychol. 60: 159-70.
14. Zuber A (2001). “A career in food service cons: high turnover”, Nations Restaurant News, 35 (21):147-148.
15. Wasmuth WJ, Davis SW (1983). “Managing employee turnover: why employees leave”, The Cornell HRA Quarterly, pp. 11-18.
16. Peters L, Bhagat R, O’Connor EJ (1981). “An examination of the independent and joint contribution of organizational commitment and job satisfaction on employee intention to quit”, Group Org. Studies, 6: 73-82.
17. DeMicco FJ, Giridharan J (1987). “Managing employee turnover in the hospitality industry”, FIU Hosp. Rev. pp.26-32.

18. Morrell K, Loan-Clarke J, Wilkinson A (2001). “Unweaving leaving: the use of models in the management of employee turnover”, Int. J. Manage. Rev. 3 (3): 219-144.

Abstract

Organizations invest a lot on their employees in terms of induction and training, developing, maintaining and retaining them in their organization. Therefore, managers at all costs must minimize employee's turnover. Although, there is no standard framework for understanding the employees
turnover process as a whole. So employee turnover problem has engaged companies.so we tried to find rules from historical data gathered form Pars oil and gas company using rough set theory to predict the employee turnover. Using these rules helps managers to select engaged employees efficiently.

Key words: Employee churn, Organizational commitment,data mining, rough set theory,Rosetta.

Shahrood University of technology

Faculty of Industrial Engineering and Management

Applying Data Mining to Employee Churn Management in Pars Oil and Gas Company

Alireza Ebrahimi

Supervisor:

Dr. Bozorgmehr Ashrafi

Date: January 2014