ارائه چهارچوب برای پیش‌بینی سطح خسارت مشتریان بیمه بدنه اتومبیل با استفاده از راهکار داده‌کاوی

نویسندگان: دکتر فرامرز فتح نژاد- سید محمود ایزدپرست

- مدیرعامل شرکت بیمه دی
- کارشناس‌ارشد مدیریت فناوری اطلاعات، دانشگاه پیام نور تهران

چکیده

امروزه مشتریان به عامل بسیار مهم و حیاتی در هدایت سرمایه‌گذاران، تولیدکنندگان و حتی محققان و نوآوران مبدل گشته‌اند. به‌همین‌دلیل سازمان‌ها نیاز دارند مشتریان خود را بشناسند و برای آنان برنامه‌ریزی کنند. تاکنون برای شناسایی مشتریان و به‌خصوص شناسایی رفتار آنها روش‌هایی استفاده شده که از جمله می‌توان روش‌های آماری را معرفی کرد که البته محدودیت‌ها و مشکلاتی دارد. در این پژوهش با استفاده از روش‌های داده‌کاوی، چهارچوبی را برای شناسایی مشتریان (با تمرکز بر مشتریان بیمه بدنه اتومبیل) ارائه می‌کنیم. به‌منظور کاهش خطاها و محدودیت‌ها از دو روش درخت تصمیم و خوشه‌بندی، به‌صورت مکمل استفاده کرده‌ایم. ابتدا با استفاده از تکنیک خوشه‌بندی، مشتریان را براساس ویژگی‌هایشان خوشه‌بندی کرده و سطح خسارت هرکدام از این خوشه‌ها را محاسبه کردیم. سپس مشتریان آتی را براساس ویژگی‌هایشان و تکنیک درخت تصمیم در یکی از این خوشه‌ها دسته‌بندی نمودیم. دسته‌ای که مشتری در آن قرار گرفته معرف سطح خطرپذیری اوست. با استفاده از این معیار و نوع بیمه‌نامه مشتری می‌توان میزان خسارت او را پیش‌بینی کرد. البته برای تعیین صحت نتایج، مدل‌ها را ارزیابی کرده و با یکدیگر مقایسه کردیم. شایان ذکر است تکنیک درخت تصمیم برای این منظور نتایج بهتری را حاصل نمود ولی تکنیک خوشه‌بندی نیز تفکیک خوبی میان مشتریان ایجاد می‌کند.

واژگان کلیدی: داده‌کاوی، بیمه، دسته‌بندی، درخت تصمیم، خوشه‌بندی، خسارت

مقدمه

امروزه گسترش روزافزون فناوری اطلاعات و حجم بالای داده‌های[1] ذخیره‌شده از مشتریان، ابزار بسیار مناسبی برای پژوهشگران است تا از آن برای به‌دست آوردن راهکارهای مناسب برای پیشرفت صنایع استفاده کنند. در این پژوهش قصد داریم با تمرکز بر حوزه بیمه، با کاربرد عملی داده‌کاوی در بیمه، بررسی کنیم که چگونه می‌توان مشتریان بیمه بدنه اتومبیل را دسته‌بندی نموده و براساس ویژگی‌های آنان میزان ریسک یا خطرپذیری مشتریان را برآورد کرد. برای این منظور از روش‌های داده‌کاوی[2]، جهت دستیابی به قوانین تصمیم‌گیری و ایجاد مدل برای پیش‌بینی خطرپذیری مشتریان در صنعت بیمه استفاده کرده‌ایم. در این مقاله ابتدا روش‌های پیشین که برای این منظور استفاده شده است را بررسی می‌کنیم و سپس به مفهوم داده‌کاوی می‌پردازیم. پس از آن روش پیشنهادی را مطرح کرده و در ادامه چگونگی پیاده‌سازی آن را بیان می‌نماییم. درنهایت به ارزیابی مدل ایجادشده و تحلیل نتایج به‌دست‌آمده خواهیم پرداخت.

۱. اهداف و سؤالات پژوهش

از مهم‌ترین ارکان صنعت بیمه، مشتریان آن می‌باشند که با عضویت خود در بیمه، پشتوانه و تأمین‌کننده منابع مالی آن به‌شمار می‌آیند. به‌همین‌دلیل شناسایی مشتریان و نیز پیش‌بینی سطح خسارت آنها، از این جهت که عامل اصلی سودآوری در بیمه‌اند از اهمیت بسیار بالایی برخوردار است (فلاح، ۱۳۷۹). درحال‌حاضر تعرفه بیمه بدنه اتومبیل براساس ویژگی‌های وسیله نقلیه است و شخص بیمه‌کننده و سبک زندگی او در تعرفه بیمه هیچ نقشی ندارد (رستمی، ۱۳۷۸). درحالی‌که علاوه‌بر اتومبیل، یکی از مهم‌ترین عوامل تأثیرگذار در میزان خسارات، سبک زندگی بیمه‌گذار یعنی ویژگی‌های شخصی فرد، کار، محل زندگی و ... است که به شناسایی میزان ریسک مشتریان کمک می‌کند و در واقع خسارات مشتریان هم در گرو همین ریسک است (حسین‌زاده و الهی، ۱۳۸۶).

1. Data
2. Data Mining

از مهم‌ترین ارکان صنعت بیمه، مشتریان آن می‌باشند که با عضویت خود در بیمه، پشتوانه و تأمین‌کننده منابع مالی آن به‌شمار می‌آیند

– سؤال ۱: آیا ویژگی‌های دموگرافیک مشتریان بیمه بدنه اتومبیل در سطح خطرپذیری و خسارت آنان تأثیرگذار است؟

– سوال ۲: اولویت اثرگذاری این ویژگی‌های دموگرافیک، در سطح خطرپذیری و خسارت مشتریان به چه ترتیبی است؟

هدف اصلی این پژوهش پاسخ‌گویی به سؤالات مطرح شده است که طی آن ابتدا مشتریان بیمه را خوشه‌بندی کرده و سپس سطح خسارت هر خوشه را محاسبه می‌نماییم. با استفاده از نتایج خوشه‌بندی، شاخص‌های تأثیرگذار در خطرپذیری مشتریان را مشخص کرده و با استفاده از آن خطرپذیری مشتریان آتی را پیش‌بینی می‌کنیم. نهایتاً با استفاده از سطح خطرپذیری و نوع بیمه‌نامه فرد می‌توان میزان خسارت او را پیش‌بینی کرد.

۲. پیشینه پژوهش

تا اینجا توضیحاتی در مورد مشکلاتی بیان کردیم که در صنعت بدنه اتومبیل وجود دارد که البته در طول سال‌هایی که از عمر بیمه می‌گذرد راه‌حل‌هایی برای آن ارائه شده که هرکدام از این راه‌حل‌ها کاستی‌هایی دارد که هنوز نتوانسته مشکلات موجود را به‌طور کامل حل نماید و یا راه‌حل مناسبی برای آن ارائه کند. در ادامه به بیان بعضی از این راه‌حل‌ها می‌پردازیم.

۲-۱. روش‌های آماری

قبل از پیدایش داده‌کاوی، برای تحلیل داده‌ها و نتیجه‌گیری از آنها از روش‌های گوناگونی استفاده می‌نمودند که از جمله آن می‌توان روش‌های آماری را نام برد که به دلیل سادگی، محبوبیت بالایی دارند. البته روش‌های آماری مانند رگرسیون و دیمانسیون نیز برای تحلیل داده‌ها، قابلیت‌های بسیاری دارند و در بسیاری از موارد نتایج خوبی می‌توان از آنها حاصل نمود ولی این روش‌ها نیز ضعف‌ها و محدودیت‌هایی دارند که در ادامه به تشریح آن خواهیم پرداخت (Kantardzic, 2002).

روش‌های آماری مرسوم، مزیت‌هایی دارند که در ادامه آمده است (Kantardzic, 2002):

– ساده می‌باشند یعنی برای استفاده از آنها نیاز به دانش

بالای آماری و ریاضی نیست؛

– قابل فهم‌اند یعنی نتایج حاصل از آنها، روشن و گویا است و نیاز به تفسیر خاصی ندارد؛

– زمانی که تعداد ابعاد[1] داده‌ها کم است بسیار خوب جواب می‌دهند و نتایج مناسبی حاصل می‌نمایند؛

– سریع بوده و همچنین نرم‌افزارهای قدرتمندی برای آنها وجود دارد که کار با آنها را ساده‌تر می‌کند.

از جمله معایب این روش عبارت‌اند از (Kantardzic, 2002):

– زمانی که تعداد ابعاد داده‌ها بالا باشد از قابلیت‌های آن کاسته می‌شود و حتی در بسیاری موارد در تحلیل داده‌ها ناتوان می‌ماند.

– روش‌های آماری معمولاً براساس یک قاعده خاصی می‌باشند و هر روشی برای نوع خاصی از مسئله کاربرد دارد و معمولاً این روش‌ها از انعطاف‌پذیری پایینی برخوردارند.

از جمله پژوهش‌هایی که از این روش استفاده نموده‌اند می‌توان به (رستمی، ۱۳۷۸) اشاره کرد. در این پژوهش پژوهشگر، تعدادی از مشخصه‌های تأثیرگذار بر میزان ریسک و خطرپذیری مشتریان بیمه را درنظرگرفته و تأثیر هریک را بررسی کرده است. البته باتوجه به اینکه فقط از روش‌های آماری و قوانین احتمالات استفاده کرده، نتایج دقیقی استخراج نشده و فقط میزان اثرگذاری هریک از مشخصه‌ها به‌صورت منفرد و مستقل محاسبه شده است، در حالی که در عالم واقع این گونه نیست و مشخصه‌های درنظرگرفته‌شده علاوه‌بر اینکه بر مشخصه هدف یعنی خطرپذیری مشتری تأثیر گذارند بر یکدیگر نیز اثر می‌گذارند. زیرا این مشخصه‌ها به‌صورت کامل از یکدیگر مستقل نیستند و باید همگی این مشخصه‌ها به‌صورت هم‌زمان و همراه با یکدیگر درنظرگرفته‌شوند تا نتیجه مطلوب حاصل شود. روش‌های آماری هنگامی که تعداد ابعاد داده‌ها زیاد می‌شوند دچار محدودیت‌هایی شده و در برخی موارد ناتوان می‌مانند و نمی‌توانند نتیجه مناسب را ایجاد کنند.

۲-۲. یادگیری ماشین

روش‌های فراگیری ماشین از جمله روش‌های داده‌کاوی است که از گذشته برای پردازش داده‌ها و همچنین داده‌های

1. Dimention

بیان‌کننده مشخصه‌ها یا صفت‌های اشیاء است.

بیمه استفاده می‌شدند. روش‌های یادگیری ماشینی عموماً رویه‌های خودکاری هستند که براساس عملیاتی منطقی یا دودویی طراحی شده‌اند. تاکنون توجه زیادی معطوف رویکرد درخت‌های تصمیم و الگوریتم‌های استنتاج قواعد شده است که در این رویکردها نتایج رده‌بندی به‌صورت یکسری تصمیمات منطقی (قواعد) اگر- آنگاه است. این مدل‌ها، نظام یا حالتی را از نمونه‌ها یاد می‌گیرند که آنها را قادر می‌سازد تا با مسائل پیچیده‌تر با داده‌های کافی به طرزی مقبول رویارو شوند. تکنیک‌های دیگر مانند الگوریتم‌های ژنتیک و رویه‌های منطقی استقرا درحال‌حاضر به‌سرعت در حال توسعه هستند و به‌نظرمی‌رسد که این نوع الگوریتم‌ها امکان رویارویی با گونه‌های عمومی‌تر داده را فراهم کنند. مثلاً مواردی که در آن تعداد و نوع مشخصه‌ها ممکن است تغییر کند یا درجایی‌که لایه‌های اضافی از یادگیری مورد نیاز باشند و به بقیه یادگیری‌ها اضافه شوند. همچنین اگر این موارد با ساختاری سلسله مراتبی از مشخصه‌ها، رده‌ها و ... توأم باشند، ارزش این رویکردها بیشتر روشن می‌شود (Raquel, 2007).

هدف فراگیری ماشین تهیه عبارات ساده است به‌طوری‌که به‌راحتی توسط انسان فهمیده شوند. برای این منظور تکنیک‌های آن باید بتوانند منطق استدلال انسان را برای کمک به فرآیند تصمیم تقلید کنند. همانند رویکردهای آماری، دانش زمینه ممکن است در توسعه این راهکارها مورد استفاده قرار گیرد، اما عملیات اصلی بدون دخالت انسان انجام می‌گیرد. این روش‌ها بسیار مفیدند و در طول زمان پیشرفت کرده و قدرت زیادی پیدا کرده‌اند. همچنین باتوجه به اینکه قواعد تولیدشده توسط آن به منطق انسان بسیار شبیه است امروزه بسیار فراگیر شده و در زمینه‌های گوناگونی به‌کارمی‌رود.

همان‌طور که بیان شد روش‌های ارائه‌شده در بالا برای حل مشکلات بیمه بدنه اتومبیل مورد استفاده قرار می‌گرفتند که البته هر کدام کاستی‌هایی دارند. از جمله ضعف در کار با داده‌هایی با حجم بالا، تعداد ابعاد زیاد داده‌ها و نیز اینکه معمولاً روش‌های آماری، انعطاف‌پذیری کمی دارند و نمی‌توان آنها را برای هر نوع مسئله‌ای به‌کاربرد یا نتایج به‌دست‌آمده را تعمیم داد. به‌همین‌دلیل روش‌های داده‌کاوی آمده‌اند تا علاوه‌بر حل این مشکلات، انعطاف‌پذیری بالاتر و قابلیت‌های بیشتری را ارائه کنند. روش‌های داده‌کاوی از

ترکیب و تجمیع روش‌هـای علوم مختلفی مثل آمار، هوش مصنوعی، بصری‌سازی، سیسـتم‌های پایگاه داده، یادگیری ماشـینی، الگوریتـم و ... ایجاد می‌شـود که با اسـتفاده از قابلیت‌های مختلفی کـه در هریک از این علوم وجود دارد سعی کرده تا محدودیت‌هایی که وجود دارد را تا حد ممکن برطرف سازد (Han & Kamber, 2006). هر یک از علوم نامبرده شـده نقاط قوت و ضعفی دارند کـه در داده‌کاوی سعی شـده تا این نقاط قوت در کنار یکدیگر قرار گیرند و ضعف‌های موجود به حداقل برسند.

۳. فعالیت‌های مرتبط با پژوهش

پس از بررسـی و مطالعه مقالات مربوط‌بـه کاربردهای داده‌کاوی، تقسـیم‌بندی به صورت نمـودار ۱ از حوزه‌های داده‌کاوی در بیمـه به‌دسـت‌آمد کـه در ادامه شـرح داده می‌شود:

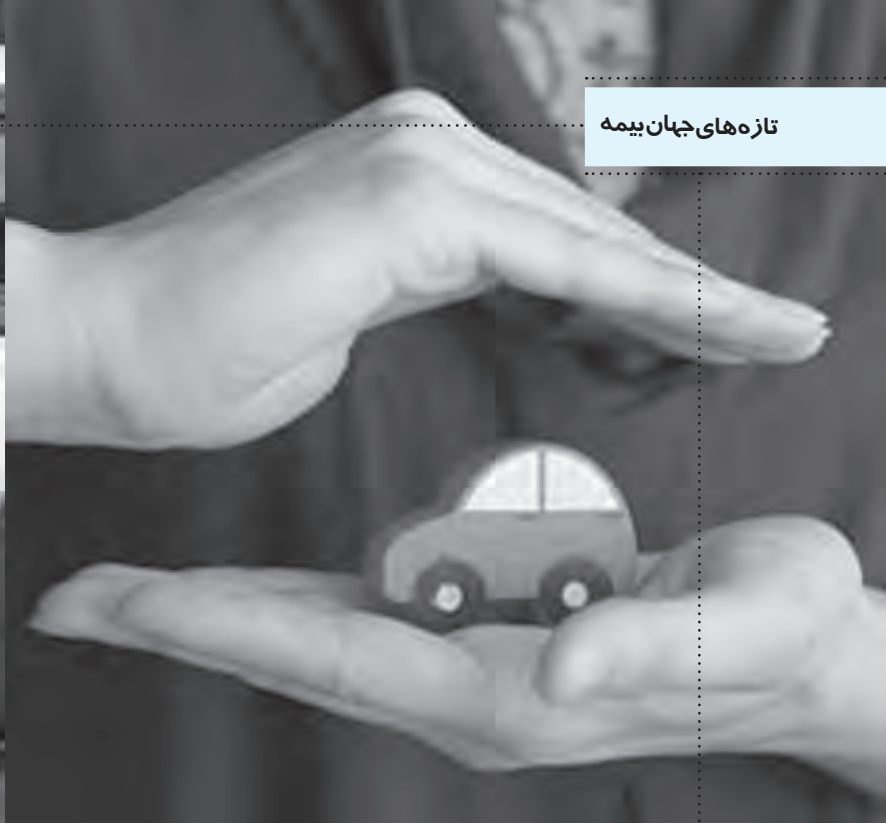
نمودار ۱. دسته‌بندی تحقیقات در زمینه داده‌کاوی

- دسته اول

تحقیقات سـعی کرده‌اند با استفاده از داده‌کاوی ریسک مشـتریان را برآورد نمایند. ریسک مشتریان براساس مقدار خسـارت و احتمال وقوع آن برای افراد تعیین می‌شـود. از جمله روش‌های تعیین خسـارت، روش‌هایی اسـت که در (Lin, 2009) و (Smith, 2000) ارائه شـده است. در این مقالات از طریق روش‌های رگرسـیونی و شبکه‌های عصبی میزان خسارت پیش‌بینی شده است.

- دسته دوم

مقالات مربوط می‌شـود به کشف و کنترل تخلفات، که سـهم عمده‌ای در کاهش هزینه‌های شرکت‌های بیمه دارد. بنابراین لازم است از ابزارهایی برای تسهیل، تسریع و افزایش دقت این فرآیند استفاده شود. داده‌کاوی از ابزارهایی است



که می‌تواند برای این منظور به‌کاررود (Morley, 2006) . به‌طورکلی در همه روش‌های ارائه‌شده برای کشف تخلفات به‌دنبال یافتن ویژگی‌های تخلفات و متخلفین می‌باشند. در ادامه برخی از روش‌های ارائه‌شده برای کشف تخلفات به‌خصوص تخلفات بیمه‌ای از ابعاد مختلف، انواع داده به‌کاررفته، تکنیک‌ها و ... بررسی می‌شود. برای کشف تخلفات از روش‌های بسیاری استفاده می‌شود. به‌طورکلی روش‌های استفاده‌شده را می‌توان به سه رده تقسیم کرد:

- روش‌های رده‌بندی: که از جمله آن می‌توان به مقالات (Morley, 2006) و (Dalkilic, 2009) اشاره کرد.
- روش‌های خوشه‌بندی: که از جمله آن می‌توان به مقالات (Castro, 2000) و (Kuo, 2006) اشاره کرد.
- ترکیب روش‌های رده‌بندی و خوشه‌بندی: که از جمله آن می‌توان به مقالات (Sumathi, 2006) و (Chann, 2010) اشاره کرد.

- دسته سوم

تحقیقات، که از تکنیک‌های داده‌کاوی برای گروه‌بندی مشتریان و تحلیل الگوهای رفتاری مشتریان در حوزه مدیریت ارتباط با مشتری استفاده شده است. در این مقالات بیشتر از روش‌های خوشه‌بندی و رده‌بندی استفاده می‌گردد. به‌گونه‌ای که با استفاده از روش‌های خوشه‌بندی، ابتدا گروه‌های لازم ایجاد می‌گردد و سپس با استفاده از روش‌های رده‌بندی هر کدام از مشتریان براساس ویژگی‌های‌شان در هر یک از این گروه‌ها قرار می‌گیرند. از جمله مقالاتی که در این زمینه منتشر شده است می‌توان به (Saha, 2009) و (Das, 2009) اشاره کرد.

۴. فرضیات پژوهش

- **فرضیه ۱:** ویژگی‌های دموگرافیک مشتریان بیمه بدنه اتومبیل در سطح خطرپذیری و خسارت مشتریان تأثیرگذار است.

- **فرضیه ۲:** با استفاده از روش‌های داده‌کاوی می‌توان ویژگی‌های اثرگذار روی سطح خطرپذیری و خسارت مشتریان را استخراج کرده و اولویت‌بندی کرد.

۵. مفاهیم داده‌کاوی

از زمانی که علم آمار به‌وجودآمد دانشمندان، نیاز به کشف خصوصیات داده‌ها را احساس کرده بودند. با استفاده از آن آمار و روش‌ها در آن زمان خصوصیات داده‌ها از قبیل پراکندگی و تمرکز آنها بررسی می‌شد (Chen, 2006). زمانی که می‌خواهیم به بررسی تأثیر تعداد کمی از عوامل بر روی هدف بپردازیم معمولاً روش‌های آماری مناسب‌اند، ولی زمانی که تعداد این عوامل زیاد می‌شود دیگر این روش‌ها کارایی مناسبی ندارند و حتی در مواردی ناتوان

می‌باشند. مثلاً در تحلیل داده‌های سبک زندگی افراد[1] به‌دلیل اینکه این داده‌ها ابعاد بسیار زیادی دارند کمتر از روش‌های آماری استفاده می‌شود. دانشمندان برای رفع این مشکل تصمیم گرفتند که از سرعت بالای کامپیوترها استفاده نمایند، همین امر سبب شد که روش‌های ابتکاری دیگری علاوه بر روش‌های آماری مثل شبکه‌های عصبی و الگوریتم ژنتیک ایجاد شود (Lee, 2006).

داده‌کاوی عبارت است از «استخراج اطلاعات و دانش و کشف الگوهای پنهان از یک پایگاه داده‌های بسیار بزرگ». که این الگوها و دانش‌ها معمولاً مستتر در داده می‌باشند (Chan & Lewis, 2002).

از داده‌کاوی می‌توان برای انجام کارهایی مثل دسته‌بندی، پیش بینی[2]، تخمین[3] و خوشه بندی[4] داده‌ها استفاده کرد. برای انجام این کارها تکنیک‌هایی توسعه یافته‌اند که باتوجه به پیشرفت کامپیوترها و این علم همه روزه برتعداد و کیفیت این تکنیک‌ها افزوده می‌شود. تعدادی از معروف‌ترین این تکنیک‌ها عبارت‌اند از: الگوریتم‌های خوشه‌بندی[5]، شبکه‌های عصبی، الگوریتم ژنتیک[6]، نزدیک‌ترین همسایگی[7] و درخت تصمیم‌گیری.

۶. روش پیشنهادی پژوهش

در این پژوهش می‌خواهیم با استفاده از تکنیک‌های داده‌کاوی، چهارچوبی برای پیش‌بینی سطح خطرپذیری و خسارت مشتریان ارائه کنیم. برای این منظور از دو تکنیک خوشه‌بندی و درخت تصمیم استفاده کرده‌ایم. ابتدا بااستفاده از روش خوشه‌بندی، مشتریان براساس ویژگی‌هایشان در خوشه‌هایی تفکیک می‌شوند. سپس میانگین سطح خسارت در هریک از این خوشه‌ها را محاسبه می‌کنیم. حال مشتریان آتی باتوجه به اینکه به کدامیک از این خوشه‌ها شبیه‌تر هستند در یکی از این خوشه‌ها قرار می‌گیرند تا سطح خطرپذیری آنها براساس خوشه‌ای که در آن قرار گرفته‌اند، مشخص شود. همچنین با استفاده از تکنیک درخت تصمیم با استفاده از داده‌ها، درختی را که براساس قوانین «اگر-آنگاه» است، ایجاد می‌کنیم و سپس مشتریان جدید را از نود ریشه وارد می‌کنیم تا به نود برگ رسیده و سطح خطرپذیری آنها مشخص شود. با داشتن سطح خطرپذیری و نوع بیمه‌نامه مشتری می‌توان سطح خسارت مشتری را پیش‌بینی کرد. درنهایت هر دو این مدل‌ها را ارزیابی کرده‌ایم. روش‌های ارزیابی به‌کاررفته در این پژوهش به دو دسته ارزیابی درونی و بیرونی تقسیم می‌شود. در ارزیابی درونی صحت مدل

1. Demographic
2. Prediction
3. Estimation
4. Clustering
5. Cluster Detection Algorithm

6. Genetic Algorithm
7. Nearest Neighboring

ایجادشده را بررسی کرده‌ایم که برای این منظور از روش پهنه سایه روشن[1] و از داده‌های آموزش و آزمایش بهره برده‌ایم و در ارزیابی بیرونی به مقایسه نتایج به‌دست‌آمده از مدل با نظرات کارشناسان خبره بیمه بدنه و تحلیل آن پرداخته‌ایم.

۷. جامعه آماری و نمونه آماری

جامعه آماری مورد استفاده در این پژوهش داده‌های مربوط‌به یک میلیون مشتری بیمه بدنه اتومبیل، از پانزده شرکت ارائه‌کننده خدمات بیمه بدنه[2] است که طی پنج سال اخیر جمع‌آوری شده و بالغ بر پنج میلیون رکورد است (جدول ۱). این داده‌ها در پایگاه داده بیمه مرکزی ذخیره شده که شامل دو دسته اطلاعات است: اطلاعات فردی مشتریان و اطلاعات مربوط‌به خسارات مشتریان. البته تعداد کل مشتریانی که سابقه خسارت دارند، حدود هفتاد هزار مشتری است که در جدولی مجزا آمده است.

باتوجه به اینکه تعداد مشتریان حدود یک میلیون نفر است در برخی عملیات‌ها مانند تحلیل اکتشافی به دلیل محدودیت‌های موجود نمونه‌ای از داده‌ها را به روش نمونه‌گیری کاهشی[3] ایجاد نمودیم.

جدول ۱. درصد مشتریان تحت پوشش شرکت‌های بیمه‌ای

شرکت بیمه	درصد تعداد مشتریان تحت پوشش
دانا	٪۹/۴
ایران	٪۳۳/۷
البرز	٪۱۲/۳
آسیا	٪۱۴/۲
سایر شرکت‌ها (یازده شرکت)	٪۳۰/۴

1. Silhouette
۲. شرکت‌های ایران، آسیا، البرز، پارسیان، کارآفرین، پاسارگاد، دانا، دی، رازی، سامان، سینا، معلم، ملت، نوین و توسعه
3. Down Sampling

۸. تکنیک‌های داده کاوی

دسته‌بندی و خوشه‌بندی از جمله تکنیک‌های بسیار پرکاربردی است که توسط متخصصان آمار و محققان فراگیری ماشینی به‌کارمی‌رود. ارائه یک تعریف دقیق از این دو روش، دشوار بوده اما مطابق با تعریف کلی، تکنیک دسته‌بندی و خوشه‌بندی، جداسازی یا قراردادن اجزا یا اشیا در تعدای از کلاس‌هاست که در خوشه‌بندی این کلاس‌ها از قبل وجود ندارد و طی فرآیند و باتوجه به ویژگی‌های اشیا ایجاد می‌شوند ولی در دسته‌بندی، این کلاس‌ها وجود داشته و اشیا براساس ویژگی‌هایشان در این کلاس‌ها قرار می‌گیرند (Tan & Steinbach, 2006).

- درخت تصمیم

درخت تصمیم، روشی معروف برای دسته‌بندی است که نتایج آن در یک نمودار جریان[4] شبیه ساختار درخت ارائه شده است، که هر گره[5] نشانگر یک تست بر روی ارزش مشخصه و هر شاخه، خروجی هر تست را نمایش می‌دهد، برگ‌های درخت نیز نمایانگر کلاس‌ها هستند. به‌طور عادی، پیچیدگی یک درخت تصمیم با افزایش تعداد مشخصه‌ها افزایش می‌یابد. اگرچه در بعضی از شرایط دیده شده است که تنها تعداد کمی از مشخصه‌ها می‌توانند کلاسی را که هر شیء به آن تعلق دارد، تعیین کنند و بقیه مشخصه‌ها کم و یا بی‌تأثیرند (Tan & Steinbach, 2006).

در ساخت درخت‌های تصمیم معمولاً داده‌ها را به دو دسته تقسیم می‌کنند:

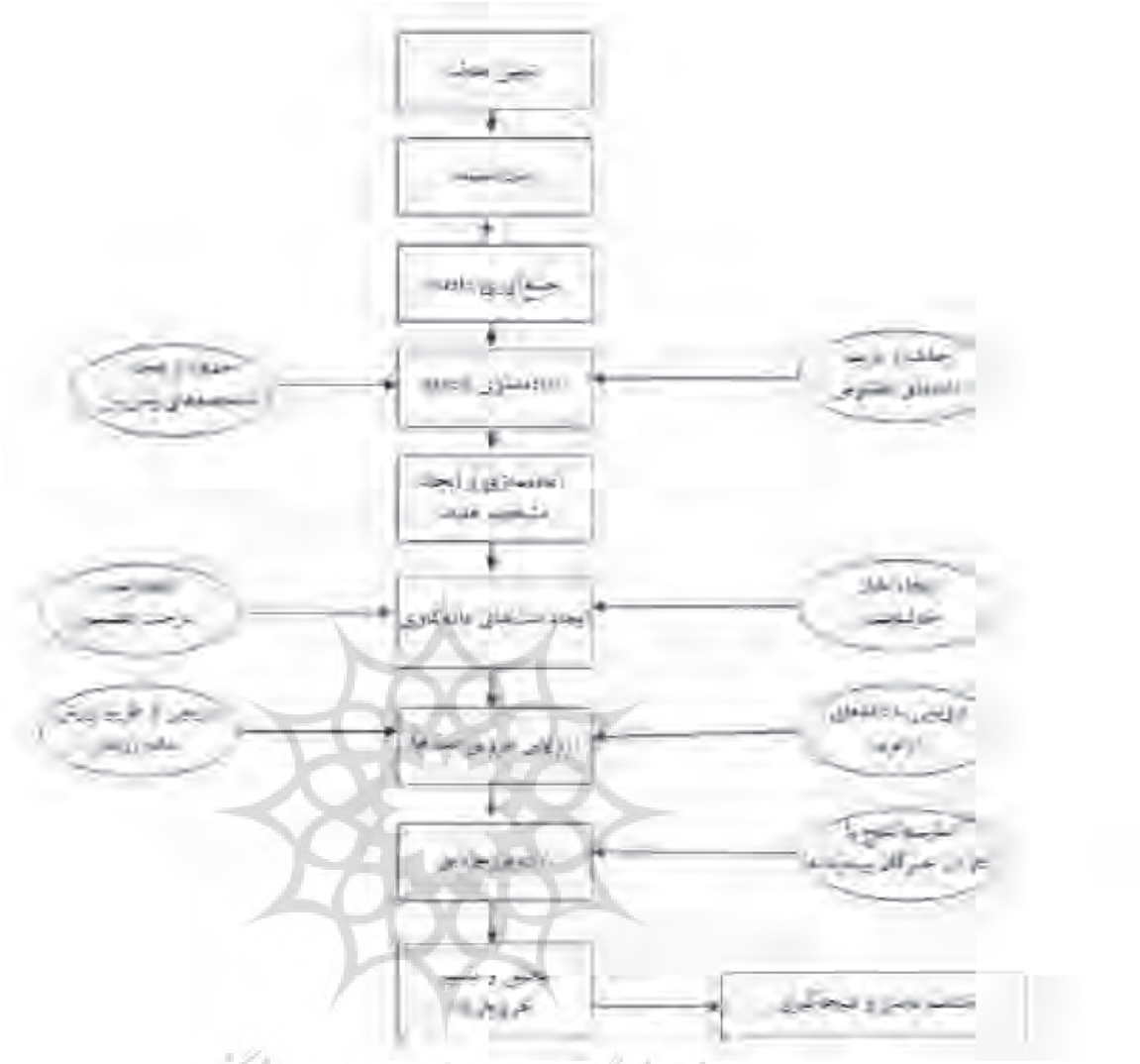
- داده‌های آموزشی[6] که برای ساخت مدل مورد استفاده قرار می‌گیرند.

- داده‌های تست[7] که برای تست و ارزیابی مدل ساخته‌شده کاربرد دارند. کیفیت داده‌های آموزشی اغلب نقش مهمی در تعیین کیفیت درخت تصمیم دارد. درصورتی‌که آموزش سیستم زیاد شود -یعنی داده‌هایی که برای آموزش و ساخت مدل به‌کارمی‌رود درصد زیادی از داده‌ها باشد- دچار حالتی به‌نام «آموزش بیش از حد مدل»[8]

درخت تصمیم، روشی معروف برای دسته‌بندی است که نتایج آن در یک نمودار جریان شبیه ساختار درخت ارائه شده است

4. Flow chart
5. Node
6. Train Data
7. Test Data
8. Over Fitting

نمودار ۲. مراحل تحقیق



خواهیم شـد که به‌دلیل وجود موارد غیرعادی در داده‌های آموزشی، خطا تولید می‌کند (Chan & Lewis, 2002).

- خوشه و خوشه‌بندی

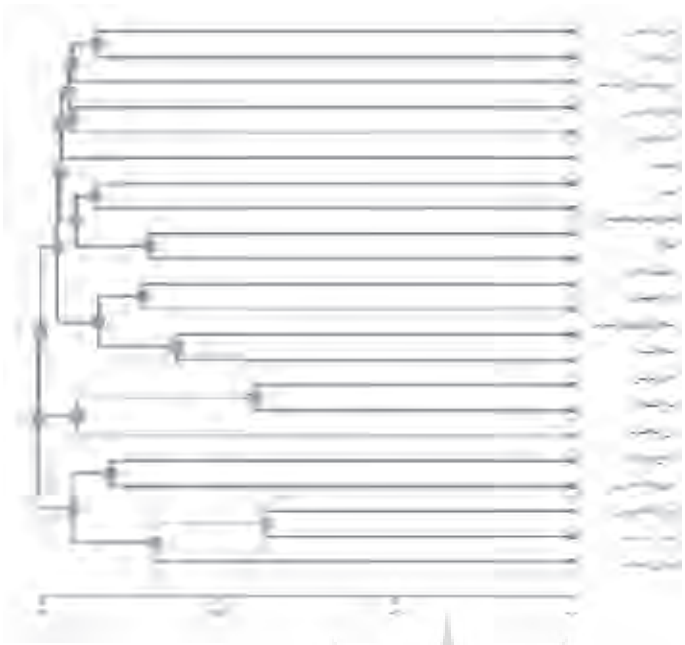
خوشـه‌بندی در واقع یـک عملیات غیرنظارتی اسـت. این عملیات هنگامی اسـتفاده می‌شـود که ما به‌دنبال یافتن گروه‌هایـی از داده‌های مشـابه باشـیم، بـدون اینکه از قبل پیش‌بینـی در مـورد شـباهت‌های موجود داشـته باشـیم. خوشـه‌بندی معمولاً هنگامی اسـتفاده می‌شـود که به‌دنبال یافتن گروه‌هایی از مشتریان هستیم که قبلاً شناخته نشده‌اند؛ یعنی گروه‌ها از قبل وجود ندارند و در فرآیند خوشـه‌بندی ایجاد می‌شـوند. برای مثال می‌توان شباهت‌های مشتریان در اسـتفاده از تلفن همـراه را به‌منظور گروه‌بندی مشـتریان و تشخیص خدمت جدید جستجو کرد (Borgelt, 2008).

۱-۸. پیش‌پردازش[1] داده‌ها

داده‌های مورد استفاده در این پژوهش، داده‌های مشتریان بیمه بدنه اتومبیل اسـت. این داده‌ها، اغتشـاش بسیار زیادی دارند و شـاید علت اصلی آن این است که این داده‌ها فقط برای ذخیره سـوابق مشـتریان کاربـرد دارد و اصولاً برای کاربردهای داده‌کاوی جمع‌آوری نشده است. داده‌های بیمه که بـرای این پژوهش مورد اسـتفاده قرار گرفتـه نیز دارای مشکلات بسیاری از جمله مقادیر گمشده، داده‌های متناقض و پراکنده اسـت. بااین‌وجود برای اینکه این داده‌ها به‌گونه‌ای تبدیل شـود که الگوریتم‌هـا و تکنیک‌های داده‌کاوی قادر به اعمـال روی آنها باشـند از روش‌های مختلفی اسـتفاده

1. Preprocess

نمودار۳. نمایش گرافیکی ماتریس همبستگی



نموده‌ایم که عبارت‌اند از:

- پاکسازی داده‌ها
- تبدیل و یکپارچه‌کردن داده‌ها
- کاهش بعد داده (Han & Kamber, 2006).

۹. مراحل تحقیق

با توضیحات ارائه شده می‌توان روند کار را بدین صورت تعریف کـرد. در ابتدا باید هدف اصلـی از این عملیات را مشـخص کنیم که می‌خواهیم چه نتایجی را به‌دست‌آوریم. در نتیجه براساس این هدف، مشخصه‌هایی را که می‌خواهیم از آنهـا برای این منظور اسـتفاده کنیم، مشـخص می‌کنیم. سـپس برای اینکه این داده‌ها برای انجام عملیات داده‌کاوی مناسـب شـوند مجموعه عملیات پیش‌پردازش را روی آن انجام می‌دهیم. پس از آن مدل‌سـازی نمـوده و به تحلیل و ارزیابی آن می‌پردازیم (نمودار ۲).

۱۰. مشخصه‌های تأثیرگذار در خطرپذیری مشتریان

برای انتخاب مشـخصه‌های[1] مناسب در این پژوهش ابتدا به بررسـی هرکدام از مشـخصه‌های موجـود پرداختیم. در این مرحله تعدادی از مشـخصه‌ها مانند (نام، نام خانوادگی، شـماره شناسـنامه و ...) که اطلاعات شـخصی افراد بوده مربوط‌به ویژگی‌های رفتاری آنها نمی‌شـود، مشخصه‌هایی مانند (شـماره شاسی، شـماره پلاک و ...) که مشخصه‌های یکتای اتومبیل اسـت و برخی مشخصه‌های مربوط‌به جدول خسارت مثل (شماره پرونده، وضعیت پرونده و ...) که یکتا بوده و نسـبت به متغیر هدف تأثیرگذار نیسـتند را باتوجه به مصاحبـات و مشـاوره‌هایی که با کارشناسـان بیمه صورت گرفت، حذف کردیم. در واقع این متغیرها، تعداد حالت‌های بسیار زیادی دارند که علاوه‌بر آنکه نمی‌توانند عامل مناسبی برای تفکیک اشـیا باشند، ممکن است روی مدل تأثیر منفی گذاشـته و نتایج را مغشوش نمایند. مشخصه‌های باقی مانده را بررسـی کردیم تا ببینیم تغییرات هر کدام از آنها نسبت‌به متغیر هدف چگونه است. در واقع بررسی کردیم که آیا آن مشخصه نسبت‌به متغیر هدف تغییر معناداری دارد یا خیر؟ و اینکه این تغییر به چه صورتی اسـت (مسـتقیم یا معکوس). برای این منظور از ماتریس همبستگی[2]، نمودار هیستوگرام[3] و نمودار جعبه‌ای[4] استفاده نمودیم.

نمـودار هیسـتوگرام نشـان‌دهنده توزیـع متغیـر اسـت و اینکـه مقادیـر بـه چـه شـکلی پراکنـده شـده‌اند و

1. Future Selection

2. Correlation
3. Histogram Chart
4. BoxPlot

نمودار جعبه‌ای نشان‌دهنده این مطلب است که داده‌ها در چه بازه‌ای بیشترین تجمع را دارند و به کمک آنها می‌توان داده‌های پرت را شناسایی نمود (Han & Kamber, 2006). برای این کار ابتدا با استفاده از نرم‌افزار R، یک نمونه از داده‌ها را وارد یک صفحه گسترده نمودیم و سپس این داده‌ها را در نرم‌افزار بارگذاری نمودیم. پس از آن با استفاده از دستورات برنامه‌نویسی این نرم‌افزار، نمودارهای کلیه مشخصه‌ها را رسم کردیم تا توزیع هر کدام از آنها را مشاهده کنیم.

درنهایت باتوجه به اینکه هدف این تحقیق دسته‌بندی بیمه‌گذاران و یافتن روابطی پنهان در میان داده‌ها برای پیش‌بینی‌های آتی است، مشخصه‌های زیر را به‌عنوان مشخصه‌های اصلی که بر میزان خسارات بیمه‌گذاران در بیمه بدنه اتومبیل تأثیرگذارند، انتخاب کردیم:

- نوع استفاده اتومبیل (آموزشی، مسافربری، بارکش، آتش‌نشانی و شخصی)؛
- سال ساخت اتومبیل؛
- تعداد سالی که مشتری خسارت نداشته است؛
- مقدار تخفیفات مشتری؛
- نوع اتومبیل (اتوبوس، بارکش، کشاورزی، سواری و موتور)؛
- ظرفیت اتومبیل؛
- سن مشتری؛
- تاریخ صدور گواهینامه مشتری؛
- شغل مشتری؛
- جنسیت مشتری؛
- وضعیت تأهل مشتری؛
- شهر محل زندگی مشتری؛
- وضعیت بیمه‌نامه (انفرادی، گروهی)؛
- سطح تحصیلی مشتری؛
- نوع پلاک اتومبیل (دولتی، شخصی)؛
- تعداد سیلندر اتومبیل؛
- شهر محل صدور شناسنامه مشتری؛
- کد شهر پلاک اتومبیل؛
- نوع مالکیت (حقوقی، خصوصی و دولتی)؛
- نوع تیپ اتومبیل.

که این مشخصه‌ها را می‌توان به سه دسته کلی تقسیم کرد:

- داده‌های دموگرافی بیمه‌گذاران؛
- داده‌های مربوط‌به ویژگی‌های اتومبیل؛
- داده‌های خسارات.

در نمودار ۳ سلسله مراتب همبستگی بین این مشخصه‌ها آمده است. نمودار سلسله مراتبی در واقع نشانگر رابطه بین متغیرهاست. بدین معنا که تا چه اندازه‌ای متغیرها به یکدیگر وابسته‌اند و کدام متغیرها وابستگی بیشتری با یکدیگر دارند. به‌عبارت‌دیگر نشان‌دهنده این مطلب است که مقدار یک متغیر را از روی کدام متغیر می‌توان با دقت بیشتری به‌دست‌آورد (Tan & Steinbach, 2006).

همان‌طور که در نمودار مشاهده می‌کنید مشخصه‌ها، همبستگی بالایی با یکدیگر ندارند و همچنین مشخصه سطح خسارت که متغیر هدف است با مشخصه‌های دیگر همبستگی مشخصی ندارد. درصورتی که متغیر هدف، با مشخصه دیگری همبستگی بالایی داشته باشد باید آن مشخصه را حذف کنیم زیرا در صورت وجود آن مشخصه پیش‌بینی به‌دست‌آمده فقط ناشی از آن مشخصه خواهد شد و تأثیر سایر مشخصه‌ها بر مشخصه هدف نشان داده نخواهد شد.

۱۱. مشخصه هدف

در این پژوهش ما به دنبال پیش‌بینی سطح خطرپذیری و درنهایت خسارت مشتریان هستیم و مشخصه مبلغ کل خسارت به خوبی نشان‌دهنده این مسئله است. ولی باید این نکته را درنظر گرفت که مبلغ خسارت همبستگی بالایی با نوع بیمه‌نامه و مبلغ تعهد بیمه‌نامه دارد و این مشخصه (مبلغ بیمه‌نامه) از ویژگی‌های دموگرافیک شخص نیست و مربوط‌به نوع وسیله نقلیه و مبلغ وسیله نقلیه اوست. به‌همین‌دلیل برای حذف آن و نرمال‌سازی سطوح خطرپذیری، مبلغ خسارت را بر مبلغ تعهد بیمه‌نامه تقسیم کردیم تا نتایج نرمال شده و بتوانیم مشخصه سطح خطرپذیری مشتری را گسسته سازیم. این مشخصه را سطح خسارت[1] می‌نامیم. در جدول ۲ حدود این سطوح مشخص شده است.

1. Level

جدول ۲. سطوح خسارت تعریف‌شده بر پایه مبلغ خسارت پرداختی بیمه به مبلغ تعرفه بیمه‌نامه

سطح خسارت	حد پایین	حد بالا
۱	کمتر از یک برابر	یک برابر
۲	یک برابر	دو برابر
۳	دو برابر	پنج برابر
۴	پنج برابر	ده برابر
۵	ده برابر	بیشتر از ده برابر

۱۲. نرم‌افزارهای داده‌کاوی به‌کاررفته

برای انجام عملیات داده‌کاوی نرم‌افزارهای زیادی موجود است که می‌توان برحسب نیاز از آنها استفاده کرد. از جمله این نرم‌افزارها:

- Microsoft SQLServer که قابلیت کار با حجم بالای داده را دارد ولی امکانات آن هنوز محدود است.
- Rapid Miner که نرم‌افزار قدرتمندی برای ساخت مدل‌های زیادی از داده‌کاوی است.
- R در واقع نرم‌افزار آماری است که برای انجام عملیات تحلیل اکتشافی بسیار مناسب است.

در این پژوهش باتوجه به اینکه حجم داده‌ها بسیار بالاست برای مدل‌سازی از نرم‌افزار Microsoft SQLServer استفاده نموده‌ایم. در واقع مجموعه عملیات پیش‌پردازش را توسط خود این نرم‌افزار انجام می‌دهیم و برای انجام عملیات داده‌کاوی و مدل‌سازی از قسمتی از این نرم‌افزار با نام SQLServer Analysis Services استفاده می‌کنیم. در واقع این نرم‌افزار برای پشتیبانی از OLAP[1] است که هر دو هدف داده‌کاوی[2] و انبار داده[3] را پشتیبانی می‌کند. بدین‌صورت که می‌توانیم یک انبار داده ایجاد کرده و سپس روش‌ها و مدل‌های داده‌کاوی را روی این انبار اعمال نماییم. البته برای عملیات تحلیل اکتشافی باتوجه به محدودیت‌های این نرم‌افزار از نرم‌افزارهای R و Rapid Miner استفاده کرده‌ایم.

1. Online Analysis Services
2. Data Mining
3. Data Warehouse

۱۳. پیاده‌سازی مدل‌های داده‌کاوی

همانطور که قبلاً بیان شد در این پژوهش از دو روش خوشه‌بندی و درخت تصمیم استفاده نموده‌ایم که در ادامه به بیان چگونگی مدل‌سازی آن خواهیم پرداخت.

- خوشه‌بندی

در این پژوهش ما از الگوریتم خوشه‌بندی K-Means استفاده می‌کنیم که شرکت ماکروسافت[4] ارائه کرده و در نرم‌افزار SQLServer نیز موجود است. همان‌طور که پیشتر ذکر کردیم این الگوریتم با وجود مزایای زیادی که دارد معایبی نیز دارد که در اینجا برای حل هر یک از این معایب این الگوریتم، راهکارهایی ارائه شده که در ادامه به بیان آن می‌پردازیم. این الگوریتم برای حل مشکل تعداد خوشه بهینه از یکسری روش‌های هیوریستیک[5] استفاده می‌کند به‌گونه‌ای که داده‌ها را با مقادیر مختلف k خوشه‌بندی می‌کند و سعی می‌کند تا با استفاده از روش‌های بهینه‌سازی مثل الگوریتم ژنتیک[6] و... مقدار بهینه k را به‌دست‌آورد البته برای اینکه محاسبه کنیم که خوشه‌های به‌دست‌آمده، کیفیت مناسبی دارد یا نه و اینکه تفکیک خوبی صورت گرفته می‌توان از روش‌هایی مثل پهنه سایه روشن استفاده نمود. این روش‌ها براساس محاسبه تفاوت بین خوشه‌ها و نیز همسایگی هر عنصر می‌توانند تعداد خوشه‌های مناسب را تشخیص دهند. در مورد ایراد دومی که بر الگوریتم K-Means وارد می‌کنند نیز تمهیداتی را درنظرگرفته‌ایم. در این پژوهش برای اینکه با این مشکل مواجه نشویم ابتدا سعی کرده‌ایم تا با استفاده از روش‌های دیگر مثل خوشه‌بندی سلسله مراتبی، نقاط مناسب مرکز خوشه را تشخیص دهیم. البته باتوجه به محدودیت‌های این روش و هزینه بالای آن از نظر زمان و حافظه، از نمونه‌ای از داده‌ها برای این الگوریتم استفاده نموده‌ایم. همچنین در روش خوشه‌بندی Microsoft نیز نقاط ابتدایی خوشه به‌صورت

4. Microsoft
5. Heuristics
6. Genetic Algorithm

جدول ۳. ماتریس مقادیر واقعی و پیش‌بینی‌شده سطح خطرپذیری

	سطح ۱ واقعی	سطح ۲ واقعی	سطح ۳ واقعی	سطح ۴ واقعی	سطح ۵ واقعی
سطح ۱ پیش‌بینی شده	۶۹۵۷	۱۵۳۲	۹۱۶	۵۱۸	۳۸۲
سطح ۲ پیش‌بینی شده	۲۳	۲۲۱۵	۶۳۹	۴۵۷	۲۴۴
سطح ۳ پیش‌بینی شده	۱۴	۷۱۲	۱۳۲۹	۳۵۷	۲۲۴
سطح ۴ پیش‌بینی شده	۵	۵۷۳	۴۴۳	۹۲۱	۱۶۷
سطح ۵ پیش‌بینی شده	۱	۲۹۳	۲۷۴	۲۲۶	۵۷۸

کاملاً تصادفی نمی‌باشد بلکه از یکسری روش‌های هیوریستیک در مورد پراکندگی داده‌ها، برای تعیین مراکز خوشه‌ها استفاده می‌شود (Robert, 2001).

- درخت تصمیم

درخت تصمیم، ابزاری برای پیش‌بینی است که در اینجا می‌خواهیم برای پیش‌بینی سطح خسارت از آن استفاده کنیم. برای ایجاد درخت تصمیم نیز از الگوریتم درخت تصمیم Microsoft استفاده می‌نماییم. در این الگوریتم مانند سایر روش‌های درخت تصمیم یکسری از مشخصه‌ها را به عنوان ورودی و یک متغیر را به‌عنوان مشخصه هدف تعریف می‌کنیم. یعنی از مشخصه‌های ورودی برای پیش‌بینی مشخصه هدف استفاده می‌کنیم. روش کار این الگوریتم به‌این‌صورت است که مشخصه‌های ورودی را براساس میزان تأثیرگذاری‌شان بر متغیر هدف درنظر گرفته و به آنها اولویت می‌دهد و نهایتاً بر طبق این اولویت‌بندی، درخت پیش‌بینی را ایجاد می‌کند (Microsoft, 2010).

۱۴. تحلیل مدل

۱-۱۴. تحلیل خوشه‌بندی

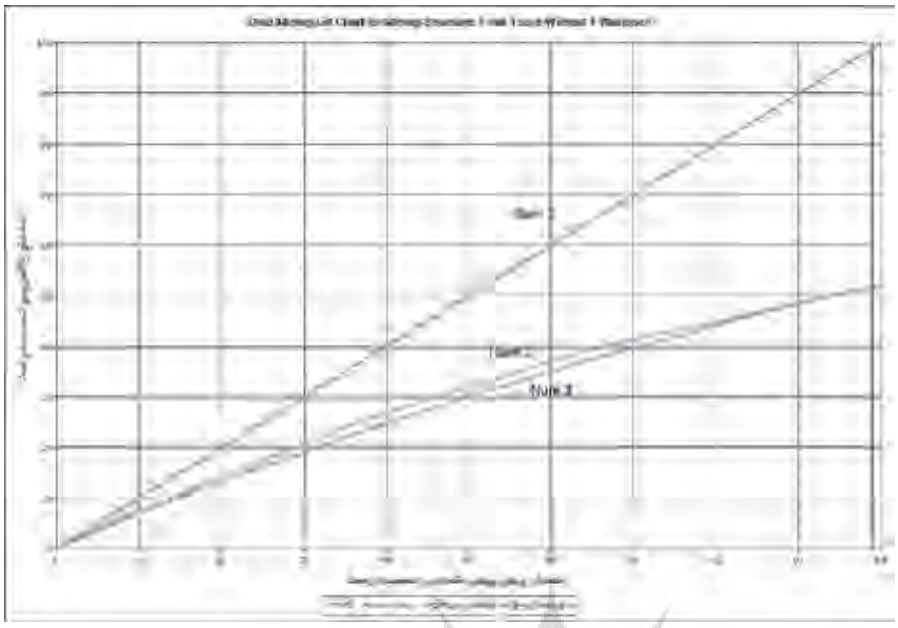
با استفاده از مشخصه‌ها و تنظیم پارامترهای الگوریتم خوشه‌بندی، این تکنیک را نیز روی داده‌ها اعمال کردیم. نمودار ۴ نشان‌دهنده خوشه‌های ایجادشده است. خطوطی که بین خوشه‌ها قرار دارد و نیز شدت رنگ آن نشان‌دهنده نزدیکی مراکز این خوشه‌ها با یکدیگر است. محل قرارگرفتن خوشه‌ها در شکل براساس الگوریتم MDS است. این الگوریتم در (Tan & Steinbach, 2006) بیان شده و مشخص‌کننده فاصله بین خوشه‌هاست. همان‌طور که مشاهده می‌کنید خوشه هشتم نسبت به سایر خوشه‌ها دورتر قرار گرفته و نیز با هیچ‌کدام ارتباط نزدیکی ندارد که نشان‌دهنده استقلال و دور بودن مرکز آن از سایر خوشه‌هاست. به‌عبارت‌دیگر در صورت تلفیق آن با سایر خوشه‌ها کیفیت خوشه‌بندی بسیار خراب خواهد شد.

نمودار ۴. شمای گرافیکی خوشه‌های ایجادشده

۲-۱۴. تحلیل درخت تصمیم

برای ساخت درخت از بیست مشخصه که به‌عنوان مشخصه ورودی تعریف کرده بودیم، استفاده نمودیم (این مشخصه‌ها به‌صورت منفرد همبستگی پایینی با مشخصه هدف دارند ولی زمانی که در کنار یکدیگر بررسی شوند می‌توان با استفاده از آن مشخصه هدف را پیش‌بینی نمود). همچنین پارامترهای لازم در الگوریتم درخت‌تصمیم را باتوجه به مطالبی که پیشتر بیان شد، تنظیم کردیم تا بتوانیم درخت مورد نظر را ایجاد کنیم. برای ساخت درخت ابتدا با استفاده از ۷۰٪ داده‌ها آن

نمودار ۵. نمودار ترفیع ایجادشده برای ارزیابی مدل درخت‌تصمیم و خوشه‌بندی



را ایجاد کردیم و سپس مدل ایجادشده را با ۳۰٪ باقی‌مانده داده‌ها ارزیابی کردیم. جدول ۳ نشانگر پیش‌بینی‌هایی است که توسط این مدل صورت گرفته است.

پس از ایجاد مدل، آن را مورد ارزیابی قرار دادیم. صحت مدل روی داده‌های تست ۶۰٪ به‌دست‌آمد. برای بررسی صحت نتیجه، آن را با درصد کلاس غالب می‌سنجیم. در این مدل کلاس غالب دارای ۳۵٪ داده‌هاست؛ یعنی بدون هیچ مدلی و به‌صورت تصادفی می‌توانیم تا دقت ۳۵٪ را پیش‌بینی کنیم. در اینجا دقت ۶۰٪ دقت بسیار خوبی است.

۱۵. ارزیابی نتایج

نتایج به‌دست‌آمده را به دو صورت ارزیابی می‌کنیم. یکی ارزیابی درونی که مدل‌های ایجادشده را با استفاده از روش‌های آزمون داده‌کاوی ارزیابی می‌کنیم و دیگری ارزیابی بیرونی که نتایج به‌دست‌آمده را با استفاده از تکنیک پرسش‌نامه با نظرات کارشناسان خبره مقایسه می‌کنیم.

۱-۱۵. ارزیابی درونی

ارزیابی درونی برای تأیید این مسئله تنظیم شده‌اند که آیا پارامترهای مدل به‌صورت مناسب، برای هدف در نظر گرفته شده‌اند. برای ارزیابی درونی از دو روش استفاده می‌کنیم.

- **ارزیابی با داده‌های آموزش و آزمایش**

در این روش برای اینکه ببینیم مدل ایجادشده تا چه حدی قابل اعتماد بوده و می‌تواند متغیر هدف را پیش‌بینی نماید داده‌ها را به دو دسته آموزش و تست تقسیم می‌کنیم. روش کار بدین‌صورت است که ابتدا مدل را با استفاده از داده‌های آموزشی ایجاد می‌کنیم و سپس مدل ساخته‌شده را با داده‌های تست آزمون می‌کنیم تا صحت عملکرد آن را به‌دست‌آوریم. این روش برای آزمون مدل‌های پیش‌بین بسیار مناسب است (Han & Kamber, 2006).

در این پژوهش داده‌ها را به دو دسته ۷۰٪ برای آموزش و ۳۰٪ برای تست تقسیم نمودیم. سپس با استفاده از داده‌های آموزش مدل درخت‌تصمیم و خوشه‌بندی را ایجاد کردیم. پس از آن ۳۰٪ داده‌های تست را وارد مدل کردیم از مدل برای پیش‌بینی مشخصه هدف یعنی سطح خطرپذیری مشتری استفاده نمودیم.

برای نمایش نحوه عملکرد مدل و پیش‌بینی آن از نمودار ترفیع[1] استفاده می‌کنیم. این نمودار بدین صورت است که در یک نمودار به‌عنوان حالت ایده‌آل (شکل نمودار در صورتی که کلیه حالت‌ها صحیح پیش‌بینی شده باشد) رسم می‌شود و همچنین نمودار دیگری که بیان‌کننده حالت مدل ما میباشد را در همان گراف رسم می‌کند (Han & Kamber, 2006).

1. Lift Chart

در نتیجه با مقایسه این دو نمودار می‌توان دریافت مدل در چه قسمت‌هایی توانسته پیش‌بینی خوبی داشته باشد و یا اینکه در کدام قسمت‌ها ضعیف عمل کرده است. همچنین پارامتری به‌نام امتیاز[1] در این نمودار داریم که نشان‌دهنده درصدی از مواردی است که به‌صورت صحیح پیش‌بینی شده است. که این پارمتر برای درخت‌تصمیم برابر ۶۰ و برای خوشه‌بندی ۵۹ به‌دست‌آمد.

نمودار ۵ نشان‌دهنده ترفیع مدل ایجاد شده است. همانطور که در نمودار مشاهده می‌کنید سه شاخص رسم شده است که شاخص ۱ نشان دهنده نمودار ایده‌آل، شاخص ۲ درخت‌تصمیم و شاخص ۳ برای خوشه‌بندی است. در اینجا ما از هر دو روش خوشه‌بندی و درخت‌تصمیم برای پیش‌بینی سطح خطرپذیری مشتریان مقادیر استفاده نموده‌ایم که همانطور که در نمودار می‌بینید این دو تکنیک تقریباً به یک اندازه توانسته‌اند پیش‌بینی صورت دهند (درخت‌تصمیم ۶۰٪ و خوشه‌بندی ۵۹٪) که باتوجه به آن می‌توان دریافت تکنیک درخت‌تصمیم برای پیش‌بینی سطح خطرپذیری مشتریان دقت بیشتری دارد و برای این منظور مناسب‌تر است.

۲-۱۵. ارزیابی بیرونی

در ارزیابی بیرونی به‌دنبال مقایسه نتایج به‌دست‌آمده با ارزیابی‌های ذهنی کارشناسان خبره بیمه بدنه اتومبیل هستیم. برای این منظور از روش پرسش‌نامه استفاده نمودیم تا به‌وسیله آن ذهنیاتی را که کارشناسان خبره نسبت‌به پارامترهای تأثیرگذار بر خطرپذیری و ریسک مشتریان دارند، استخراج کرده و با نتایج حاصل از عملیات داده‌کاوی و مدل ایجادشده مقایسه نماییم. این پرسش‌نامه شامل مشخصه‌هایی است که برای مدل‌سازی از آنها استفاده کردیم. براساس مشخصه‌هایی که برای ایجاد مدل از آنها استفاده کردیم، پرسش‌نامه را ایجاد می‌نماییم. به‌عبارت‌دیگر سؤالات پرسش‌نامه بدین‌صورت است که اولویت هر کدام از این مشخصه‌ها توسط کارشناسان هم تعیین می‌گردد. کارشناسان به هرکدام از این مشخصه‌ها امتیازی از یک تا پنج می‌دهند که این امتیاز، تعیین‌کننده اولویت و اهمیت آن مشخصه، در خطرپذیری مشتری و سطح خسارت اوست. پرسش‌نامه ایجادشده شامل ۲۰ سؤال، بر گرفته از مشخصه‌های اصلی تأثیرگذار در خطرپذیری مشتریان است.

به‌دلیل اینکه شرکت‌های بیمه‌ای مختلفی وجود دارد برای پرکردن پرسش‌نامه‌ها از چهار شرکت بزرگ بیمه‌ای (آسیا، ایران، دانا و البرز که ۷۰٪ مشتریان بیمه بدنه اتومبیل را شامل می‌شوند) استفاده نمودیم. تعداد ۲۲ پرسش‌نامه توسط کارشناسان این شرکت‌ها پر گردید. سپس با استفاده از نرم‌افزار صفحه گسترده[2] امتیازاتی را که کارشناسان به هرکدام از این مشخصه‌ها داده‌اند، ذخیره کرده و مجموع امتیاز برای هر مشخصه را محاسبه کردیم. نهایتاً براساس این امتیازات، مشخصه‌ها را اولویت‌بندی نمودیم. به‌گونه‌ای که مشخصه‌ای که دارای امتیاز بالاتری است، اولویت بالاتری نیز دارد.

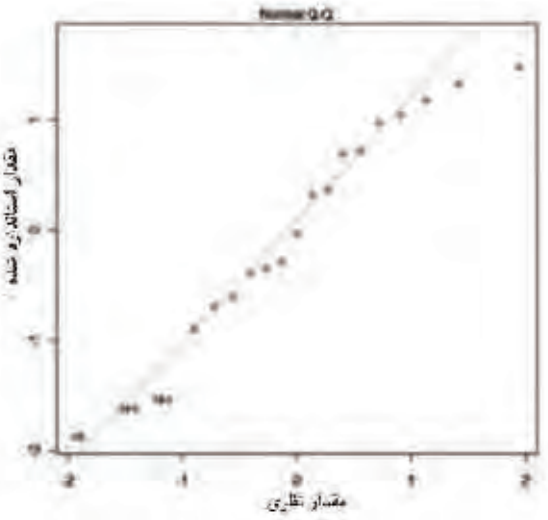
- تحلیل نتایج نظرات کارشناسان

پس از اولویت‌بندی مشخصه‌ها براساس نظرات کارشناسان، حال نتایج حاصل از مدل‌ها را با آن مقایسه می‌کنیم تا ببینیم این نتایج تا چه حدی هم‌راستا می‌باشند. برای این منظور نمودار اولویت‌های به‌دست‌آمده از مدل را برحسب اولویت‌های به‌دست‌آمده از نظرسنجی خبرگان رسم می‌کنیم. هرچه نمودار ایجادشده به خط نیمساز ناحیه اول نزدیک‌تر باشد بدین معناست که نتایج به‌دست‌آمده از دو روش به یکدیگر نزدیک‌تر است. نمودار ۶ نشان‌دهنده این امر است. در این نمودار نقاط، نشان‌دهنده نسبت نتایج به‌دست‌آمده از پرسش‌نامه به نتایج حاصل از مدل است و نیز خط رسم‌شده، خط رگرسیونی است که براساس این نقاط به‌دست‌آمده‌است. خط رگرسیون، خطی است که مجموع فواصل آن از نقاط ایجادشده حداقل باشد. به‌عبارت‌دیگر هرچه نقاط به این خط نزدیک‌تر باشند به معنای آن است که این نتایج به‌دست‌آمده بیشتر به یکدیگر نزدیک می‌باشند. مقدار ضریب همبستگی محاسبه‌شده برابر ۸۲٪ است که به معنای همبستگی بالا میان نتایج به‌دست‌آمده از مدل و پرسش‌نامه است.

1. Score

2. Spread Sheet

نمودار۶. نمودار همبستگی و خط رگرسیون نتایج حاصل از مدل و نتایج حاصل از پرسش‌نامه



نمودار ۶ به‌صورت خط مستقیم نیست؛ به این معنا که نتایج به‌صورت دقیق یکسان نمی‌باشند. ولی این مطلب قابل توجیح است. زیرا اولویت‌هایی که در ذهن کارشناسان است کاملاً دقیق نیست، زیرا براساس ذهنیات و تجربیات آنان بوده و پایه علمی قوی ندارد. به‌همین‌دلیل اختلاف‌هایی بین نتایج وجود دارد، ولی نتایج به‌دست‌آمده از دو مدل تقریباً هم‌راستا هستند. ضریب همبستگی برای این دو برابر ۰/۸۲ است که این مطلب می‌تواند به‌عنوان تأییدی بر نتایج حاصل‌شده باشد، این دو نتیجه به‌دست‌آمده یکدیگر را تأیید می‌کنند.

۱۶. آزمون فرضیه‌ها

در ابتدا دو فرضیه بیان کردیم که در طول این پژوهش سعی بر اثبات آن داشتیم. براساس این دو فرضیه و اهداف درنظرگرفته‌شده، ساختاری تعریف نموده و به مدل‌سازی برای آن پرداختیم و درنهایت مدل‌های ایجادشده را ارزیابی و تحلیل کردیم تا به نتایج لازم برای اثبات این فرضیه‌ها دست یابیم که در ادامه به بیان آن می‌پردازیم.

– **فرضیه ۱:** ویژگی‌های دموگرافیک مشتریان بیمه بدنه اتومبیل در سطح خطرپذیری و خسارت مشتریان تأثیرگذار است.

همانطور که از نتایج به‌دست‌آمده از مدل ایجادشده دیدیم، توانستیم مشتریان را براساس ویژگی‌های دموگرافیک آنها خوشه‌بندی کنیم. در این مدل سطح خسارت مشتریان را به پنج رده تقسیم کردیم که مشتریان براساس شباهتشان به مشتریان دیگر در خوشه‌ها، در یکی از این رده‌ها قرار گرفتند. هر کدام از این رده‌ها، بازه‌ای از خسارت دارند که این بازه‌ها را از پر خسارت به کم خسارت تقسیم‌بندی کردیم.

آنها را در یکی از این خوشه‌ها قرار دادیم و براساس خوشه‌ای که مشتری در آن قرار گرفته بود، میزان سطح خطرپذیری و خسارت او را پیش‌بینی کردیم.

– **فرضیه ۲:** با استفاده از روش‌های داده‌کاوی می‌توان ویژگی‌های اثرگذار روی سطح خطرپذیری و خسارت مشتریان را استخراج کرده و اولویت‌بندی نمود.

در این پژوهش ما با استفاده از روش‌های خوشه‌بندی، مشتریان را براساس ویژگی‌هایشان در خوشه‌هایی قرار دادیم. برای خوشه‌بندی از مشخصه‌هایی که تمایز قابل توجهی میان مشتریان ایجاد می‌کنند استفاده کردیم که این مشخصه‌ها با استفاده از نظر کارشناسان خبره بیمه و عملیات تحلیلی اکتشافی که انجام دادیم، به‌دست‌آمد. سپس با استفاده از روش‌های رده‌بندی مانند درخت‌تصمیم، الویت تأثیرگذاری این مشخصه‌ها در سطح خطرپذیری مشتریان را به‌دست آوردیم.

۱۷. دستاوردهای پژوهش

از دستاوردهای مهم این پژوهش، نتایج حاصل از عملیات داده‌کاوی بر روی داده‌های بیمه بدنه اتومبیل است. براساس نتایج این مرحله مشاهده گردید که علاوه‌بر مشخصه‌های ظاهری اتومبیل مثل تعداد سیلندر، ظرفیت، سال ساخت و نوع کاربری، مشخصه‌های رفتاری مشتری یا همان ویژگی‌های دموگرافیک آنها نیز در پیش‌بینی سطح خسارت مشتریان بیمه تأثیر عمده‌ای دارد. این نتیجه به‌خصوص در مورد سطح خسارت مشتری اهمیت ویژه‌ای در این صنعت دارد، چرا که مصوبه بیمه مرکزی برای نرخ حق‌بیمه بدنه اتومبیل دقیقاً براساس همین مشخصه‌هاست. این نکته نشان‌دهنده آن است که پارامترهای مورد استفاده برای نرخ گذاری بیمه نیازمند بازنگری و توجه بیشتر به سوابق و رفتار خرید و خسارت مشتری است. در واقع تعیین

تعرفه بیمه که هم اکنون فقط براساس ویژگی‌های اتومبیل است، نقص‌های بسیار زیادی دارد. البته این مسئله بدین معنا نیست که ویژگی‌های اتومبیل در تعیین تعرفه بیمه تأثیرگذار نمی‌باشد بلکه بدین معناست که این ویژگی‌ها برای این امر ناقص بوده و به تنهایی نمی‌تواند نتیجه مناسب را داشته باشد. زیرا در یک تصادف راننده اتومبیل و ویژگی‌های او نیز نقش بسیار مهمی دارند.

نتایج حاصل از این پژوهش هم برای بیمه مرکزی و هم برای شرکت‌های بیمه می‌تواند سودمند باشد. بیمه مرکزی می‌تواند از آن برای تعیین تعرفه بیمه‌نامه استفاده کند و شرکت‌های بیمه می‌توانند از آن برای شناسایی مشتریانشان و نیز سفارشی‌سازی محصولات برای مشتریان استفاده نمایند که این امر کمک بسیاری به سودآوری شرکت‌ها خواهد کرد.

۱۸. پیشنهادات آتی

در این پژوهش ما توانستیم مشخصه‌هایی را دریابیم که در سطح خطرپذیری مشتریان دخیل هستند و ترتیبی برای اولویت‌بندی تأثیر آنها، در سطح خطرپذیر به‌دست‌آوردیم. البته برای تعیین تعرفه بیمه‌نامه علاوه‌بر مشخص‌شدن مشخصه‌ها و اولویت آنها، میزان وزن و تأثیرگذاری هریک از این مشخصه‌ها در خطرپذیری مشتریان نیز بسیار مهم است که به‌دست‌آوردن وزن هریک از این مشخصه‌ها در مبلغ تعرفه بیمه‌نامه و خطرپذیری مشتریان نیز می‌تواند موضوعی برای پژوهش باشد.

داده‌های موجود در پایگاه داده بیمه، نقص‌های بسیار زیادی دارد که این مسئله ناشی از آن است که این داده‌ها برای تحلیل و داده‌کاوی جمع‌آوری نمی شوند بلکه برای ایجاد پرونده‌ها و سوابق مشتریان هستند. در نتیجه باعث کاهش کیفیت نتایج و حتی ایجاد نتایج اشتباه می‌گردند. لذا تبیین قوانینی برای درج این مشخصه‌ها در پایگاه داده توسط نمایندگی‌های بیمه می‌تواند کمک شایانی در امر داده‌کاوی باشد تا بتوان نتایج واقعی‌تری از آنها استخراج کرد.

بررسی کاربردها و ارائه استراتژی‌های مناسب برای مدیریت ارتباط با مشتری و مدیریت خسارات مشتری براساس نتایج چهارچوب ارائه‌شده، از مسائل راه‌گشا در صنعت بیمه اتومبیل است که هم شرکت‌های بیمه برای حفظ یا افزایش مشتریانشان و هم سازمان بیمه مرکزی برای تعیین تعرفه مناسب، بنابر میزان ریسک مشتری، می‌توانند از آن استفاده کنند.

درصورتی‌که داده‌های مربوط‌به جرائم رانندگی و میزان ریسک مشتریان بیمه در دسترس باشد می‌تواند کمک شایانی به شناسایی مشتریان و در نتیجه پیش‌بینی دقیق‌تر میزان خسارت آنها کند. داده‌های مربوط‌به جرائم رانندگی در اختیار پلیس راهور می‌باشد. درصورتی‌که این داده‌ها در سوابق مشتریان بیمه موجود باشد یعنی این دو بانک اطلاعاتی در پیوند با یکدیگر باشند می‌توان نتایج بسیار دقیق‌تری از شناسایی رفتار مشتریان و خطرپذیری آنها به‌دست‌آورد.

منابع

۱. حسین‌زاده، لیلا و الهی، شعبان۱۳۸۶ ، *دسته‌بندی مشتریان هدف در صنعت بیمه با استفاده از داده‌کاوی*، پایان‌نامه، دانشگاه تربیت مدرس.

۲. رستمی، حمیدرضا ۱۳۷۸، 'نقش عوامل ایجادکننده ریسک در قیمت‌گذاری بیمه بدنه اتومبیل،' *فصلنامه صنعت بیمه*، ش ۴۸، صص ۹۰-۸۵.

۳. فلاح، زهرا ۱۳۷۹، *بررسی عوامل عمده اثرگذار بر تقاضای بیمه بدنه اتومبیل و برآورد الگوی مناسب*، پایان‌نامه دانشگاه آزاد اسلامی.

4. Borgelt, C 2008, 'Accelerating fuzzy clustering', *Information Sciences*, vol.179. no.23.

5. Castro, E 2000, *Automatic clustering via boundary extraction for mining massive point-data sets*, In Proceedings of the Fifth International Conference on Geo-Computation.

6. Chan, C & Lewis, B 2002, 'A basic primer on data mining', *Information Systems Management*, vol.19, no.4, pp.56-60.

7. Chann, P 2010, 'Data hirarchical

algorithm, Proc. of the 19th ACM Symposium on Applied Computing.

17. Microsoft 2010, Microsoft SQLServer 2008, Viewed 7 Feb 2010 <http://msdn.microsoft.com>.

18. Morley, B 2006, *How the detection of insurance fraud, Psychology, Crime & Law*, pp.163-180.

19. Raquel, F 2007, 'Modelling of insurers' rating determinants', An application of machine learning techniques and statistical models, *European Journal of Operational* Research, vol.183, no.3, pp.1488-512.

20. Robert, A 2001, 'SQL Server Analysis Services', *Designing SQL Server 2000 Databases*, pp.453-98.

21. Saha, S 2009, 'A new point symmetry based fuzzy genetic clustering technique for automatic evolution of clusters, 'Information Sciences, vol.51, pp.532-41.

22. Smith, K 2000, 'An analysis of customer retention and insurance claim patterns using data mining: a case study' *Journal of the Operational Research Society*, pp.532-41.

23. Sumathi,S 2006, *Data mining for insurance*, Berlin, Springer.

24. Tan, PN & Steinbach, M 2006, *Intruduction To data mining*, United State Of America, Addison Wesly.

clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry', *Knowledge-Based Systems*, vol.23, no.8, pp.800-8.

8. Chen, S 2006, 'A KanoCKM model for customer knowledge discovery', *Total Quality Management*, Vol.1, No.5, pp.589-608.

9. Dalkilic, T 2009, *Neural networks approach for determining total claim amounts in insurance*, Insurance: Mathematics and Economics .

10. Das, S 2009, *Utilization of self-organizing map and fuzzy clustering for site characterization using piezocone data*, Computers and Geotechnics.

11. Han, J & Kamber, M 2006, *Data mining: concepts and techniques*, San Morgan, Francisco Kaufman.

12. Kantardzic, M 2002, *Data mining: concepts, models, methods, and algorithms*, Wiley-IEEE Press.

13. Kuo, R 2006, *Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan*, Expert Systems with Applications.

14. Lee, S 2006, 'Decision tree approaches for zeroinflated count data', J*ournal of Applied Statistics*, vol.33, no.8, pp.853-65.

15. Lin, C 2009, U*sing neural networks as a support tool in the decision making for insurance industry*, Expert Systems With Applications.

16. Lu, F 2004, *A fast genetic k-means*