پژوهشنامهٔ بیمه (صنعت بیمهٔ سابق)/سال‌بیست‌وششم/شمارهٔ ۴/زمستان ۱۳۹۰/شماره‌مسلسل ۱۰۴/صفحات ۱۰۷تا ۱۲۹

طبقه‌بندی میزان ریسک بیمه‌گذاران بیمهٔ بدنهٔ خودرو با استفاده از الگوریتم‌های داده‌کاوی

(مورد مطالعه: یک شرکت بیمه)

نسترن حاجی حیدری[۱]
سامرند خالهء[۲]
احمد فراهی[۳]

تاریخ دریافت مقاله: ۱۳۹۰/۱۰/۲۰
تاریخ پذیرش مقاله: ۱۳۹۰/۱۲/۰۸

چکیده

شرکت‌های بیمه به‌عنوان یکی از نهادهای مالی تأثیرگذار در اقتصاد و جامعه لازم است به ابزارهای تحلیل ریسک قدرتمندی دسترسی داشته باشند تا بتوانند ریسک دریافتی را به خوبی مدیریت کنند. نتایج حاصل از تحلیل ریسک می‌تواند از طرق مختلفی نظیر قیمت‌گذاری محصولات و بازاریابی و شناسایی مشتریان هدف برای مدیریت ریسک اعمال شود. به‌همین‌دلیل استفاده از ابزار داده‌کاوی می‌تواند در سنجش و پیش‌بینی ریسک بیمه‌گذاران، باتوجه به این تجارب، بسیار راه‌گشا باشد.

هدف این مقاله طبقه‌بندی بیمه‌گذاران بیمهٔ بدنهٔ شرکت بیمه به لحاظ ریسک دریافت یا عدم‌دریافت خسارت طی دورهٔ بیمه است. در ابتدا داده‌های مورد نیاز مشتریان طی یک دورهٔ مشخص، جمع‌آوری شده و سپس فرایند پیش‌پردازش داده‌ها و شناسایی متغیرهای الگوریتم‌های مختلف طبقه‌بندی، روی داده‌های نهایی اعمال شده و نتایج این الگوریتم‌ها به لحاظ صحت پیش‌بینی با یکدیگر مقایسه شده است. درنهایت الگوریتم C5 که بالاترین میزان صحت را در پیش‌بینی ریسک مشتریان دارد به‌عنوان الگوریتم پیشنهادی به شرکت بیمهٔ مورد مطالعه ارائه گردیده است.

واژگان کلیدی: *بیمهٔ بدنهٔ خودرو، داده‌کاوی، طبقه‌بندی ریسک و بیمه‌گذاران*

---

۱. استادیار دانشگاه تهران (نویسنده مسئول) (Email: nhheidari@ut.ac.ir)
۲. کارشناس‌ارشد مدیریت فناوری اطلاعات، دانشگاه پیام نور تهران (Email: samrand.khaleie@gmail.com)
۳. استادیار دانشگاه پیام نور تهران (Email: afarahi@pnu.ac.ir)

۱. مقدمه

پیچیدگی محیطی، شدت رقابت، رواج تکنولوژی‌های نو و پیشرفته، توسعهٔ فناوری اطلاعات و ارتباطات، شیوه‌های نوین عرضهٔ کالاها و خدمات، مسائل زیست محیطی و جهت‌گیری سازمان‌ها از دارایی‌های مشهود به نامشهود و... از عوامل عمده‌ای است که موجب شده است سازمان‌ها و بنگاه‌های اقتصادی در دوران حیات خود با ریسک‌های بسیار متعدد و خطرات زیاد و حتی پیش‌بینی‌نشده مواجه شوند. به همین جهت به‌منظور کاهش ریسک و جبران زیان‌های ناشی از آن، امروزه در ادبیات علمی انواع مدیریت ریسک نظیر مدیریت ریسک بنگاه، مدیریت ریسک کسب‌وکار و مدیریت ریسک استراتژیک مطرح شده و هریک جایگاه خاصی دارند.

بدیهی است هر سازمان باتوجه به ماهیت کار خود، ریسک‌های گوناگونی را تجربه می‌کند و در شرایط متحول امروز، اساساً موفقیت هر بنگاه به تسلط آن بر ریسک‌ها و نوع مدیریتی است که بر انواع ریسک‌ها اعمال می‌کند. مدیریت ریسک، زمانی معنا و مفهوم می‌یابد که شرایط با احتمال متحمل‌شدن زیان و عدم‌اطمینان مواجه شود. این نوع مدیریت شامل حوزه‌های گسترده‌ای است که مسائل مالی، عملیاتی، تجاری، استراتژیک و حوزهٔ وسیع‌تری به نام حوادث خطرآفرین را دربرمی‌گیرد. در مجموع مدیریت ریسک، فرایند سنجش یا ارزیابی ریسک و سپس طرح استراتژی‌هایی برای ادارهٔ ریسک است. متفکران، چهار استراتژی متداول برای مدیریت ریسک برشمرده‌اند: انتقال ریسک (قبول ریسک توسط بخش دیگر)، اجتناب از ریسک (عدم‌انجام فعالیتی که موجب ریسک شود)، کاهش ریسک (شیوه‌هایی که موجب کاهش شدت زیان شود) و پذیرش ریسک (قبول زیان در هنگام وقوع). اما نکته اصلی در بنگاه‌های اقتصادی ما این است که نگاه استراتژیک در مورد شناسایی مدیریت ریسک در آنها وجود ندارد؛ به‌طوری‌که جایگاه خاصی برای این نوع مدیریت در سازمان‌ها و بنگاه‌های ما تعیین نشده است. بدیهی است باتوجه به شرایط

پیچیده و رقابتی کسب‌وکار در عصر امروز، مدیریت ریسک بیش از گذشته اهمیت خود را بازیافته و مدیران برای بقای بنگاه‌های خود و کاهش زیان، ناگزیرند به آن روی آورده و متعهد به اجرای آن باشند. ریسک دلیل وجود بیمه است و بدون ریسک در واقع بیمه مفهوم خود را از دست می‌دهد. کار بیمه‌گری با ریسک و ریسک‌پذیری و کاهش ریسک و محاسبهٔ ریسک سروکار دارد. شرکت‌های بیمه می‌توانند براساس مقدار ریسک هر بیمه‌گذار، مقدار حق‌بیمه را تعیین کنند. بنابراین ابزارهای پیش‌بینی‌کنندهٔ داده‌کاوی، وسیلهٔ مناسبی برای تحلیل ریسک در بیمه خواهد بود. علاوه‌براین از طریق روش‌های خوشه‌بندی یا دسته‌بندی می‌توان در مورد مشتریان هدف و استراتژی‌های بازاریابی برای گروه‌های مختلف تصمیم‌گیری کرد.

در صنعت بیمه، داده‌کاوی می‌تواند به شرکت‌ها جهت کسب مزیت تجاری کمک کند. به‌طورمثال با به‌کارگیری تکنیک‌های داده‌کاوی، شرکت‌ها می‌توانند با استفاده از داده‌ها در مورد الگوهای خرید و رفتار مشتری، به کسب دانش پرداخته و همچنین درک خود را از کسب‌وکار برای کمک به کاهش تقلب، ارتقا بیمه‌گری و بالابردن مدیریت ریسک (Gayle, 1999) افزایش دهند. صنعت بیمه برای تحلیل ریسک وابسته به تجارب پرداخت خسارت است. به‌همین‌خاطر تکنیک‌های داده‌کاوی به علت متکی‌بودن به داده می‌تواند کاربرد زیادی در صنعت بیمه داشته باشد. شرکت‌های بیمه، داده‌های زیادی راجع‌به مشتریان خود ذخیره و نگهداری می‌کنند و این در حالی است که در کشف دانش یا ارزش نهفته در این داده‌ها کم‌توان هستند. با انجام داده‌کاوی و تحلیل اطلاعات موجود در پایگاه‌های داده و سیستم‌های اطلاعاتی، شرکت‌های بیمه قادر خواهند بود ضمن تشخیص بهتر رفتار مشتریان، مشتریان موجود را حفظ و مشتریان بالقوه و بازارهای هدف را شناسایی و فعالیت‌های بازاریابی و تبلیغاتی خود را در آن بخش متمرکز کنند و مدیریت بهینه‌ای را در ارتباط و تعامل با مشتری داشته باشند.

پژوهش حاضر درصدد است با استفاده از داده‌های مربوط‌به بیمه‌نامه‌های بدنهٔ اتومبیل یک شرکت بیمه طی سال‌های ۱۳۸۹-۱۳۸۸، ابتدا مشخصه‌های تأثیرگذار در ریسک بیمه‌گذاران بیمه بدنهٔ اتومبیل را در پایگاه داده موجود تعیین نموده سپس با استفاده از الگوریتم‌های داده‌کاوی مدلی ارائه دهد تا با به‌کارگیری آن بتوان میزان ریسک بیمه‌گذاران آتی را (به لحاظ ریسک داشتن خسارت یا عدم‌خسارت) پیش‌بینی کرد. این مدل می‌تواند در سیاست‌گذاری‌های آتی شرکت بیمه به‌کارگرفته‌شود. به‌عنوان‌مثال شرکت‌های بیمه با استفاده از نتایج داده‌کاوی می‌توانند در میزان حق‌بیمهٔ دریافتی از بیمه‌گذاران مختلف تعدیل ایجاد کنند و با ایجاد سیستم نرخ‌گذاری مبتنی‌بر ریسک بیمه‌گذاران، میزان رضایت بیمه‌گذاران را افزایش داده و از طرفی سودآوری خود را ارتقا دهند. برای ساخت مدل پیش‌بینی لازم است که ابتدا تکنیک مدل‌سازی انتخاب شود که در این پژوهش ۶ تکنیک (درخت تصمیم[1]، شبکه‌های عصبی[2]، شبکه‌های بیزین[3]، ماشین بردار پشتیبان[4]، رگرسیون لجستیک[5] و تحلیل تمایزی[6]) انتخاب گردیده است. ازآنجاکه برخی الگوریتم‌ها صرفاً با داده‌های عددی و برخی صرفاً با داده‌های غیرعددی قابل پیاده‌سازی‌اند و باتوجه به اینکه ساختار پایگاه داده پژوهش شامل داده‌های عددی و داده‌های غیرعددی به‌صورت تواماً است، الگوریتم‌های انتخابی با این نوع داده‌ها هم‌خوانی داشته و قابل اجراست. همچنین برای انتخاب مدلی که بالاترین میزان دقت را در پیش‌بینی داشته باشد، روش‌های مذکور با استفاده از نرم‌افزار Weka از لحاظ درجه صحت دسته‌بندی گردیده و در نهایت الگوریتم C5 که دارای بالاترین درجه صحت است، استفاده شده است.

1. Decision Tree
2. Neural Network
3. Baysian Network
4. Support Vector Machine
5. Logistic Regression
6. Discriminant Analysis

در این مقاله ابتدا مبانی نظری تحقیق تشریح می‌گردد که شامل مباحث داده‌کاوی، ریسک در بیمه و کاربردهای داده‌کاوی در بیمه است. سپس به روش اجرایی و داده‌های مورد استفاده در تحقیق پرداخته می‌شود و نهایتاً فرایند مدل‌سازی صورت گرفته و نتایج بیان می‌گردند.

۲. مروری بر ادبیات تحقیق

۲-۱. انواع بیمه

بیمه یک ابزار مهم مدیریت ریسک است که می‌تواند به چهار طبقه تقسیم‌بندی شود:

- **اجتماعی در مقابل خصوصی:** بیمه‌های اجتماعی توسط دولت تأمین می‌شوند و چند ویژگی دارند اول، مشارکت در این بیمه‌ها اجباری است و تأمین مالی آن توسط دولت انجام می‌گیرد. دوم، امنیت وجود درآمد را برای ریسک‌های شناخته‌شده‌ای چون بیکاری و بازنشستگی تأمین می‌کند، نهایتاً بر دارایی اجتماعی تأکید می‌کند که وجه تمایز این بیمه از بیمه‌های خصوصی است.

- **زندگی در مقابل غیرزندگی:** شرکت‌هایی که بیمه را برای حمایت از اموال می‌فروشند غیرزندگی و آنهایی را که بیمه‌نامه‌هایی برای حمایت از جان اشخاص صادر می‌کنند، شرکت‌های بیمهٔ زندگی می‌نامند.

- **شخصی در مقابل تجاری:** ما می‌توانیم بیمه را براساس گروه هدف خریداران بیمه طبقه‌بندی کنیم؛ مانند بیمهٔ شخصی که برای مصرف‌کنندگان انفرادی تهیه می‌شود که می‌تواند شامل بیمه‌های زندگی انفرادی، صاحبان خانه و خودرو باشد. بیمهٔ تجاری شامل بیمه‌ای است که برای سازمان‌ها طراحی می‌شود، مانند بیمهٔ اموال تجاری، بیمهٔ مسئولیت عمومی و ... .

- **مستقیم در مقابل اتکایی:** بیمه‌ای که به عموم فروخته می‌شود، بیمهٔ مستقیم نام دارد. بیمهٔ اتکایی، قراردادی بین یک بیمه‌گر مستقیم با یک بیمه‌گر دیگر است (Skipper & Kwon, 2007).

باتوجه به طبقه‌بندی عنوان‌شده، ریسک بیمهٔ بدنهٔ اتومبیل خصوصی، غیرزندگی و شخصی است. شرکت‌های بیمه ممکن است این ریسک دریافتی را جهت انتقال ریسک به شرکت‌های دیگر، واگذار کنند یعنی اتکایی کنند.

۲-۲. داده‌کاوی و دسته‌بندی

داده‌کاوی، اقتباس یا استخراج دانش از مجموعه‌ای از داده‌هاست (Edward & Mishkin, 1995). به‌بیان‌دیگر، داده‌کاوی فرایندی است که با استفاده از تکنیک‌های هوشمند، دانش را از مجموعه‌ای از داده‌ها استخراج می‌کند. دانش استخراج‌شده در قالب مدل‌ها، الگوها یا قواعد ارائه می‌شود. این الگوها، مدل‌ها و قواعد اشکال مختلفی برای ارائهٔ دانش استخراج‌شده، هستند. این دانش می‌تواند ملاک تصمیم‌گیری‌های آتی، عملکردهای بعدی یا تغییرات لازم در سیستم قرار گیرند (مروج، ۱۳۸۳). برطبق تعریف مؤسسه سیستم تحلیل آماری[1] در سال ۱۹۹۸ داده‌کاوی، فرایند انتخاب، اکتشاف، مدل‌سازی و شفاف‌سازی الگوهایی مفید و ناشناخته در حجم زیادی از داده است (Koh & Low, 2004). داده‌کاوی، فرایندی است که مراحل متعددی دارد. با اجرای این مراحل می‌توان به یک الگوی دانشی در میان مجموعهٔ کثیری از داده‌ها دست یافت (Lee & Siau, 2001). فایاد و همکارانش[2] مراحل داده‌کاوی را به شرح زیر ارائه کرده‌اند:

- بازیابی اطلاعات از یک بانک اطلاعاتی؛
- انتخاب یک زیربخش مرتبط برای انجام داده‌کاوی؛
- تصمیم‌گیری در مورد سیستم‌های نمونه‌گیری مناسب و تمیزکردن داده‌های موجود در بانک اطلاعاتی؛
- استفاده از یک روش مناسب برای پیش‌پردازش داده‌ها؛

1. Statistical Analysis System (SAS)
2. Fayyad et al, 1996

– ایجاد مدلی از داده‌های پیش‌پردازش‌شده.

فرایند داده‌کاوی به دو قسمت عمده تقسیم می‌شود (Kantardzic, 2003):

– داده‌کاوی توصیفی؛

– داده‌کاوی پیشگویی‌کننده.

داده‌کاوی توصیفی به توصیف روابط الگوها و مدل‌های پنهان در حجم زیادی از داده می‌پردازد. درصورتی‌که داده‌کاوی پیشگویی‌کننده به کشف الگوها و روابط ناشناخته در میان انبوه داده‌ها اشاره دارد. فرایند داده‌کاوی دارای ابزارهایی است که می‌توان آنها را در این دو قسمت قرار داد. بدین‌ترتیب که فعالیت‌های طبقه‌بندی، رگرسیون، تحلیل سری‌های زمانی و تخمین در حوزهٔ داده‌کاوی پیشگویی‌کننده‌اند و فعالیت‌های خوشه‌بندی، خلاصه‌سازی، کشف توالی و قوانین وابستگی در بخش داده‌کاوی توصیفی قرار دارند. داده‌کاوی، کاربردهای فراوانی دارد که در این پژوهش به کاربرد آن در مدیریت و تحلیل ریسک در صنعت بیمه پرداخته می‌شود.

۳-۲. روش‌ها و تکنیک‌های داده‌کاوی

برحسب اینکه در فرایند داده‌کاوی، استنتاج چه نوع دانشی از مجموعهٔ آموزشی مورد نظر است، از روش‌های مختلف داده‌کاوی می‌توان بهره جست. این روش‌ها از نظر شیوه یادگیری به دو دسته اصلی تقسیم می‌شوند:

– الگوریتم‌های یادگیری با نظارت[1]؛

– الگوریتم‌های یادگیری بدون نظارت[2].

دو هدف اصلی داده‌کاوی، پیشگویی و توصیف است (France, 2003). داده‌کاوی پیشگویی‌کننده، مدلی را از سیستم ارائه می‌دهد که توسط مجموعه‌ای از داده‌های مشخص پیش‌بینی می‌گردد. هدف کلی آن طبقه‌بندی، پیش‌بینی و تخمین داده‌هاست.

1. Supervised Learning
2. Unsupervised Learning

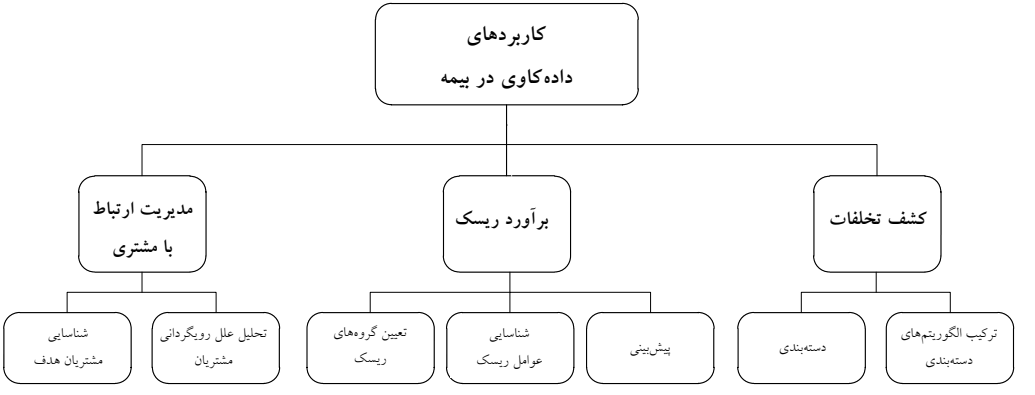
داده‌کاوی توصیفی، اطلاعات جدید و غیربدیهی را براساس مجموعه‌ای از داده‌های موجود ارائه می‌دهد و هدف کلی آن درک و شناخت سیستم‌های تجزیه‌وتحلیل‌شده، با استفاده از الگوها و روابط موجود است.

باتوجه به نتیجه مورد نیاز که پیش‌بینی و دسته‌بندی ریسک مشتریان است در این پژوهش از مدل داده‌کاوی پیش‌بینی‌کننده استفاده شده است.

۲-۴. کاربردهای داده‌کاوی در صنعت بیمه

داده‌کاوی می‌تواند در صنعت بیمه به شرکت‌ها جهت کسب مزیت تجاری کمک کند. به‌طورمثال با به‌کارگیری تکنیک‌های داده‌کاوی، شرکت‌ها می‌توانند با استفاده از داده‌ها در مورد الگوهای خرید مشتری و رفتار مشتری، به کشف دانش بپردازند. همچنین داده‌کاوی در درک بیشتر از کسب‌وکار برای کمک به کاهش تقلب، ارتقای بیمه‌گری و بالابردن مدیریت ریسک، ابزارهای مناسب و مؤثری ارائه می‌کند (Yeo et al, 2001). با مطالعه پیشینهٔ موضوع، کاربردهای داده‌کاوی در صنعت بیمه را می‌توان به سه دسته اصلی تقسیم‌بندی کرد (نمودار۱).

نمودار ۱. کاربردهای داده‌کاوی در صنعت بیمه



(چوبدار،۱۳۸۷)

۲-۵. اهمیت تحلیل ریسک در بیمه و قیمت‌گذاری

روش‌های مرسوم برای تعیین ریسک یا خسارت در شرکت‌های بیمه استفاده از روش‌های اکچوئری است، در این روش‌ها، بیمه‌گذاران به گروه‌های مختلف (مانند سن، جنسیت، مسافت

رانندگی و...) تقسیم می‌شوند و برای هر گروه از طریق داده‌های خسارتی گذشته، مقدار و احتمال خسارت تخمین زده می‌شود (Bigus, 1996). متدولوژی داده‌کاوی غالباً می‌تواند مدل‌های اکچوئری موجود را از طریق یافتن متغیرهای تأثیرگذار اضافی از طریق شناسایی تعاملات و روابط غیرخطی ارتقا دهد (Bigus, 1996). مهم‌ترین سؤال در نرخ‌گذاری این است که فاکتورهای ریسک یا متغیرهایی که برای پیش‌بینی احتمال ادعای خسارت و اندازهٔ یک خسارت مهم‌اند، کدامند. بنابراین یکی از نتایج این پژوهش تعیین اهمیت فاکتورهای مؤثر بر ریسک مشتریان است که به نرخ‌گذاری در صنعت بیمه کمک شایانی می‌کند.

۶-۲. پژوهش‌های مرتبط با به‌کارگیری الگوریتم‌های داده‌کاوی در صنعت بیمه

خلاصه‌ای از برخی پژوهش‌های صورت‌گرفته در زمینهٔ این پژوهش در جدول ۱ ارائه شده است:

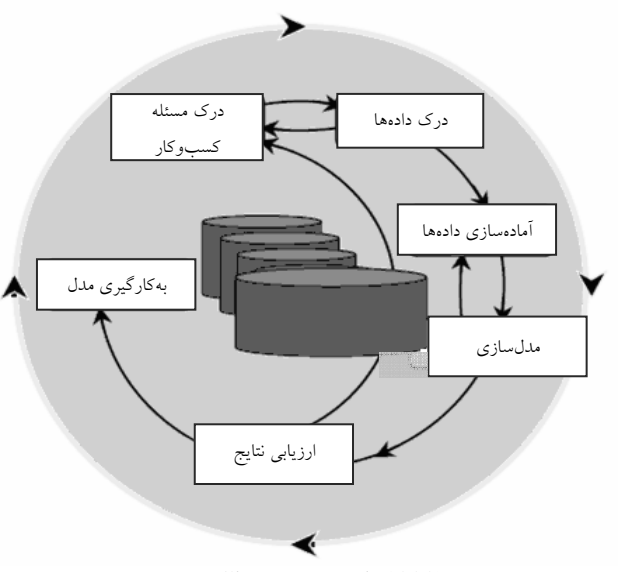
جدول ۱. پژوهش‌های کاربردی داده‌کاوی در صنعت بیمه

مدل استفاده‌شده در داده‌های مربوط‌به بیمه	حوزه استفاده‌شده	هدف نهایی	نام پژوهش
استفاده از K-means clustering	بیمه خودرو	یافتن ویژگی‌های مشترک بیمه‌گذاران	(Goe, 2003)
استفاده از درخت تصمیم CHIAD	بیمه خودرو	یافتن اهمیت فاکتورهای ریسک گوناگون	(Goe, 2003)
استفاده از K-means clustering	بیمه خودرو	یافتن ویژگی‌های مشترک بیمه‌گذاران	(Yeo et al, 2001)
استفاده از درخت تصمیم و شبکه‌های عصبی	بیمه عمر، بیمه حوادث انفرادی، بیمه حوادث خانواده	دسته‌بندی مشتریان هدف	(حسین‌زاده، ۱۳۸۶)
استفاده از درخت تصمیم CHIAD	بیمه خودرو	شناسایی مشتریان آتی بیمهٔ بدنه اتومبیل/ پیش‌بینی مشتریان دارای خسارت/ پیش‌بینی سطح خسارت بیمه‌گذاران بیمهٔ بدنه	(چوبدار، ۱۳۸۷)
استفاده از الگوریتم‌های درخت تصمیم و شبکه عصبی و رگرسیون لجستیک	بیمه خودرو	پیش‌بینی ریسک مشتریان بیمهٔ بدنه اتومبیل	(عنبری، ۱۳۸۹)

۳. روش پژوهش

فرایند این پژوهشِ کاربردی، با کار با داده‌ها آغاز می‌شود و سعی بر آن است تا به کشف الگوهای پنهان داده‌ها پرداخته و الگوریتمی که از بالاترین صحت برخوردار است برای استفاده در امور پیش‌بینی بیمه معرفی گردد. باتوجه به ماهیت پژوهش که استفاده از داده‌کاوی جهت طبقه‌بندی بیمه‌گذاران بیمه بدنهٔ اتومبیل است، این پژوهش از نوع داده‌محور است. پایهٔ اصلی آن بر کشف دانش از پایگاه دادهٔ شرکت بیمهٔ مورد مطالعه نهاده شده است. ازاین‌رو استاندارد جهانی فرایند استاندارد داده‌کاوی در صنعت[1] جهت انجام فرایند پژوهش استفاده شده است که این مراحل شامل درک مسئله کسب‌وکار، درک داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی نتایج و به‌کارگیری مدل است. نمودار۲ مراحل این استاندارد را در قالب یک فرایند نشان می‌دهد.

نمودار ۲. مراحل مدل فرایندی داده‌کاوی براساس استاندارد CRISP-DM



(Chapman et al, 1999)

1. Cross-Industry Standard Process for Data Mining (CRISP-DM)

۱-۳. مراحل اجرای پژوهش

– جمع‌آوری داده از پایگاه داده بیمه‌گذاران فعلی بانک اطلاعاتی بیمهٔ خودرو و پالایش داده‌ها و تعیین شاخص‌هایی برای تعریف طبقات ریسک بیمه‌گذاران؛

– تقسیم آنها به دو دستهٔ داده‌های آزمایشی و داده‌های آموزشی؛

– استخراج الگوها با استفاده از داده‌های آزمایشی با تکنیک‌های مختلف طبقه‌بندی و مقایسهٔ نتایج با استفاده از روش‌های دیگر؛

– اعتبارسنجی مدل با استفاده از مجموعه داده‌های آزمایشی و معرفی بهترین الگوریتم؛

– ارائهٔ الگوی کشف‌شده از طبقه‌بندی بیمه‌گذاران.

۲-۳. داده‌های پژوهش

داده‌های این پژوهش شامل داده‌های مربوط‌به بیمه‌نامه‌های بیمه‌گذاران بیمهٔ بدنهٔ اتومبیل است که بخشی از آنها دچار حادثه شده و خسارت دیده‌اند. جامعهٔ آماری متشکل از ۱۳۷۶۸ دادهٔ مربوط‌به بیمه‌نامه‌های بیمهٔ بدنهٔ اتومبیل است که طی سال‌های ۱۳۸۹–۱۳۸۸ توسط شرکت بیمهٔ مورد مطالعه صادر شده است که شامل بیمه‌نامه‌های خسارتی و غیرخسارتی است. در بین این حجم داده، تمام داده‌ها از کیفیت لازم برخوردار نبودند و صلاحیت ورود به مدل نهایی را نداشتند. معیوب‌بودن داده‌ها از دو لحاظ مورد بررسی قرار گرفته، اول از لحاظ خطاهای اندازه‌گیری بررسی شده است. بدین معنی که بعضی از فیلدها دارای مقادیر نامناسب بودند؛ به‌عنوان‌مثال در بعضی از آنها سن بیمه‌گذار اعداد نامتعارفی (اعداد یک رقمی یا سه رقمی) وارد شده بود؛ درنهایت رکوردها از لحاظ وجود داده‌های نامرتبط مورد بررسی قرار گرفته است؛ بدین‌صورت‌که در بعضی از فیلدها مقادیر نامرتبط وارد شده بود؛ به‌عنوان‌مثال برای نوع خودرو مقادیر عددی اختصاص یافته بود. برهمین‌اساس تا حد امکان سعی شد تا رکوردهای معیوب طی مصاحبه با متصدیان بیمه یا باتوجه به فیلدهای مرتبط دیگر اصلاح شود، ولی در نهایت بعضی از داده‌ها که قابلیت اصلاح نداشتند، حذف

گردیدند. بنابراین داده‌های نهایی به ۱۲۴۵۵ داده کاهش پیدا کرد. داده‌هایی که فیلدهای اصلی آنها گمشده یا اشتباه ثبت شده بود، حذف گردیدند. همچنین در ادامهٔ فرایند آماده‌سازی، برای پاک‌سازی و پیش‌پردازش داده‌ها، دو عملیات مهم کاهش داده و اعمال تغییرات در شکل داده‌ها بر روی پایگاه داده رابطه‌ای صورت گرفت. پاک‌سازی داده‌ها در دو بخش اصلی اصلاح اشتباهات کاربر و یک شکل‌کردن داده‌ها انجام شد. از طرفی فیلدهایی که ارزش اطلاعاتی مناسبی نداشتند از پایگاه داده حذف گردیدند.

خصیصه‌های مشتریان (متغیرهای مستقل پژوهش) و طبقهٔ مشتری (متغیر وابسته پژوهش)، فیلدهای این پایگاه داده را تشکیل می‌دهند. عناوین این فیلدها در پایگاه داده نهایی و پالایش‌شده، شامل ۱۱ خصیصه مشتریان و ۱ فیلد طبقه است (جدول۲).

جدول ۲. عناوین فیلدهای استفاده‌شده در مدل داده‌کاوی

نام فیلد	سن خودرو	سیستم خودرو	نوع خودرو	ارزش خودرو
نام اختصاری	CarAge	CarSystem	CarKind	CarValue
نام فیلد	نوع بیمه‌گذار	سن بیمه‌گذار	نوع شهر	مبلغ حق‌بیمه
نام اختصاری	Customerkind	CustomerAge	CityKind	Price
نام فیلد	کاربری خودرو	پوشش اضافی قطعات	نوع پرداخت حق‌بیمه	طبقه بیمه‌گذار
نام اختصاری	CarUsage	ExtraCover	PaymentKind	Class

داده‌های اولیه قبل از مدل‌سازی، پیش‌پردازش شده‌اند تا از کیفیت قابل قبولی برخوردار باشند. پیش‌پردازش هم در خصوص رکوردهای ناقص صورت گرفته است که به کاهش بسیار زیاد داده‌های نهایی و کاهش تعداد فیلدها انجامید. علت کاهش فیلدها، داشتن داده‌های گمشدهٔ بسیار و بی‌ارتباط‌بودن برخی از فیلدها مانند کد ملی، رنگ خودرو و ... است.

۳-۳. مدل‌سازی داده‌ها با استفاده از الگوریتم‌های طبقه‌بندی

برای ساخت مدل لازم است که ابتدا تکنیک مدل‌سازی انتخاب شود که در این پژوهش۶ تکنیک (درخت تصمیم، شبکه‌های عصبی، شبکه‌های بیزین، ماشین بردار پشتیبان، رگرسیون لجستیک و تحلیل تمایزی) انتخاب گردیده است. ازآنجاکه برخی الگوریتم‌ها صرفاً با داده‌های عددی و برخی صرفاً با داده‌های غیرعددی قابل پیاده‌سازی‌اند و باتوجه به اینکه ساختار پایگاه دادهٔ پژوهش شامل داده‌های عددی و داده‌های غیرعددی به‌صورت توأم است، الگوریتم‌های انتخابی با این نوع داده‌ها هم‌خوانی داشته و قابل اجراست. همچنین ازآنجاکه ابزار مورد استفاده در این پژوهش نرم‌افزار کلمنتاین[1] ۱۲ است، بنابراین کلیهٔ الگوریتم‌های طبقه‌بندی قابل اجرا برای پژوهش ما این ۹ الگوریتم است.

۱-۳-۳. شبکهٔ عصبی

شبکه‌های عصبی با تشخیص نحوهٔ فعالیت و محاسبات مغز انسان برای اولین بار در ۵۰ سال قبل توسط رزنبلت[2] مطرح شد. شبکهٔ عصبی، روشی برای تقریب توابع، یافتن تابع مدل‌ساز خروجی برحسب ورودی و استخراج الگو است که بر پایه اتصال بهم پیوسته چندین واحد پردازشی به نام نرون، ساخته می‌شود که این نرون‌ها در لایه‌های مشخصی آرایش یافته‌اند. نرون‌های هر لایه دارای ارتباطات وزن‌داری با نرون‌های لایه‌های قبل و بعد خود است. هر شبکه حداقل دارای یک لایهٔ ورودی و یک لایهٔ خروجی و در صورت نیاز تعدادی لایهٔ پنهان است. شبکه‌های عصبی مصنوعی با پردازش داده‌های آموزشی، دانش یا قانون نهفته در داده‌ها را به ساختار شبکه منتقل می‌کنند که به این عمل یادگیری می‌گویند.

1. Clementine
2. Rosenblatt

۲-۳-۳. شبکهٔ بیزین

باتوجه به توانایی شبکه‌های بیزی در زمینهٔ مدل‌سازی شبکه‌ها در سال‌های اخیر به استفاده از آنها توجه زیادی شده است. مزایای این شبکه عبارت‌اند از: توانایی کار با تعداد زیادی متغیر؛ توانایی مدل‌کردن انواع ارتباطات خطی، غیرخطی و تصادفی؛ توانایی کار با مقادیر گمشده؛ درنظرگرفتن متغیرهای پنهان که اندازه‌گیری از آنها وجود ندارد؛ توانایی کار با داده‌های نویزی به جهت داشتن مبنای احتمالاتی قوی.

شبکهٔ بیزی، یک مدل گرافیکی احتمالاتی است که ارتباطات بین مجموعه‌ای از متغیرهای تصادفی را کد می‌کند و از دو جزء اصلی تشکیل شده است. یک گراف بدون حلقه جهت‌دار[1] که ساختار ارتباطات بین متغیرها را در شبکه نشان می‌دهد و توزیع احتمال شرطی[2] که توزیع احتمال هر گرهٔ شبکه را به شرط والدینش تعریف می‌کند.

۳-۳-۳. درخت تصمیم

الگوریتم ایجاد یک درخت تصمیم، یک الگوریتم حریصانه است که در یک فرایند چرخشی از بالا به پایین، درخت تصمیم را ایجاد می‌کند. این الگوریتم یکی از شناخته‌شده‌ترین روش‌های ایجاد درخت تصمیم است که به نام ID3[3] معروف است و اساس بقیه الگوریتم‌های درخت تصمیم است.

۴-۳-۳. ماشین بردار پشتیبان

ماشین بردار پشتیبان، تکنیک به‌نسبت جدیدی در حوزهٔ داده‌کاوی است که اولین بار وپنیک[4] آن را مطرح کرد و در بسیاری از مسائل طبقه‌بندی از جمله طبقه‌بندی متن و ... به‌طور موفقیت‌آمیزی به‌کاررفته‌است. استفاده از ماشین بردار پشتیبان برای اعتبارسنجی،

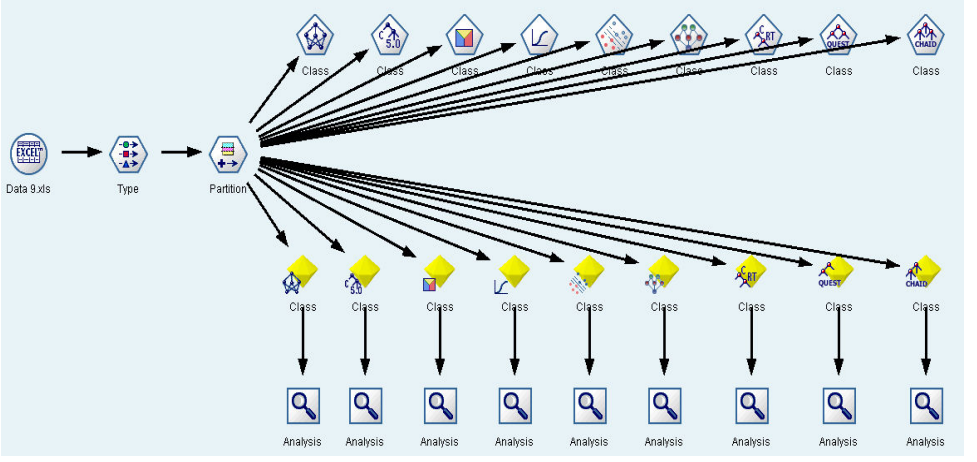
1. Directed Acyclic Graph (DAG)
2. Conditional Probability Distribution (CPD)
3. Iterative Dichotomiser 3 (Examples, Target, Attributes)
4. Vapnik

موضوع نسبتاً جدیدی است و تحقیقات اندکی دراین‌زمینه صورت گرفته است (Bellotti & Crook, 2008).

ماشین بردار پشتیبان، یک طبقه‌بندی‌کنندهٔ دو تایی است که با استفاده از نگاشت داده‌ها از فضای ورودی اصلی به فضایی با بُعد بالاتر برای جداسازی آنها عمل می‌کند. این مدل، ابرصفحه‌ای را جستجو می‌کند که فاصله‌اش با داده‌های دو طبقه ماکزیمم است. در این روش سعی بر آن است تا جهت به‌دست‌آوردن مرز طبقه‌ها، سیستمی با ظرفیت کمینه یا به بیان بهتر سیستمی با حداقل پیچیدگی پیاده‌سازی شود. در نتیجه ماشین بردار پشتیبان می‌تواند با استفاده از داده‌های آموزشی کمتر نسبت‌به روش‌های رقیب، مرزهای سیستم را با دقت مناسبی تخمین بزند، بدون آنکه تعمیم‌پذیری سیستم را مخدوش کند(منیری، ۱۳۸۵).

برای اعمال تکنیک‌های مذکور در پژوهش حاضر از نرم‌افزار کلمنتاین ۱۲ استفاده شده است (نمودار ۳).

نمودار ۳. مسیر[1] ساخته‌شده بااستفاده از نرم‌افزار جهت اجرای الگوریتم داده‌کاوی



1. Stream

ازآنجاکه روش ارائه‌شده در هر پژوهشی باید به لحاظ اعتبار، مورد سنجش قرار گیرد، بنابراین در این پژوهش نیز با عنایت به اینکه روش پژوهش از نوع «داده محور» است، روش اعتبارسنجی به‌این‌صورت است که داده‌ها به دو مجموعه داده‌های آموزشی[1] و داده‌های آزمایشی[2] تقسیم می‌شوند؛ هدف این است که با تعداد داده‌های آموزشی، الگوریتم انتخابی، دانشی را حاصل می‌کند، ولی اینکه نتایج به‌دست‌آمده تا چه میزان دارای اعتبار هستند باید توسط نتایج داده‌های جدید و قدرت پیش‌بینی الگوریتم در مورد داده‌هایی که تاکنون با آن مواجه نبوده، آزمون شوند. از این جهت داده‌های آزمایشی به‌عنوان داده‌های ناظر به الگوریتم داده می‌شوند و نتایج به‌دست‌آمده می‌توانند میزان صحت مدل را ارزیابی کنند.

تقسیم داده‌ها توسط نرم‌افزار مورد استفاده جهت داده‌کاوی و به‌صورت تصادفی صورت می‌گیرد. ازلحاظ تعداد داده‌های هر مجموعه، باید گفت همیشه تعداد داده‌های آموزشی بیشتر از داده‌های آزمایشی درنظرگرفته‌می‌شود. در این پژوهش تعداد مجموعه‌های آموزش ۷۰درصد از کل داده‌ها بوده (۸۶۸۳) و ۳۰درصد باقی‌مانده (۳۷۷۲) به عنوان داده‌های آزمایش درنظرگرفته‌شدند.

برای ارزیابی صحت مدل در طبقه‌بندی مشتریان، از مدل‌های طبقه‌بندی[3] استفاده شده است (جدول ۳).

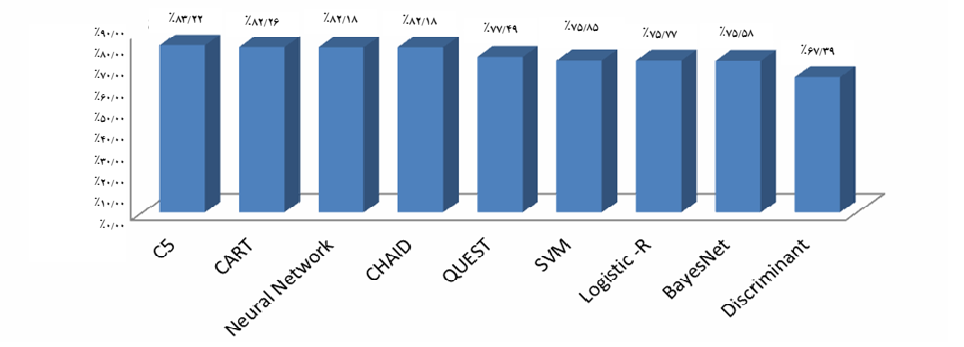
1. Training Data Set
2. Test Data Set
3. Classifier

جدول ۳. مقایسه نتایج صحت الگوریتم‌های طبقه‌بندی

نوع الگوریتم	میزان صحت پیش‌بینی
Neural Network	٪۸۲/۱۸
C5	٪۸۳/۲۲
Discriminant	٪۶۷/۳۹
Logistic –R	٪۷۵/۷۷
SVM	٪۷۵/۸۵
Bayes Net	٪۷۵/۵۸
CART	٪۸۲/۲۶
QUEST	٪۷۷/۴۹
CHAID	٪۸۲/۱۸

نتایج فوق را می‌توان در نمودار ۴ ترسیم کرد.

نمودار ۴. مقایسهٔ تطبیقی صحت الگوریتم‌های طبقه‌بندی



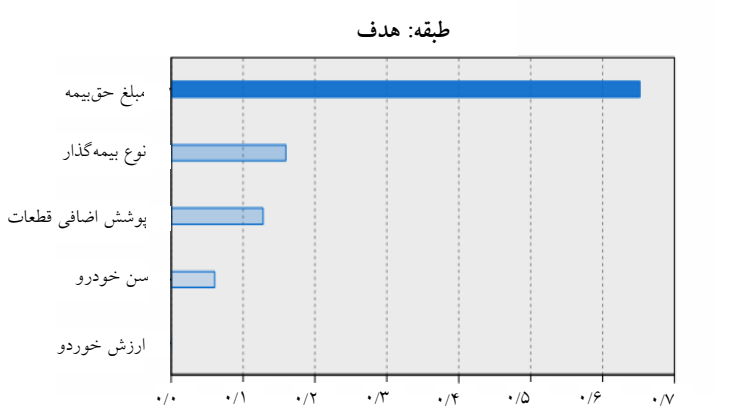
باتوجه به نمودار ۴ که اولویت‌بندی الگوریتم‌ها را به لحاظ دقت طبقه‌بندی نشان می‌دهد، می‌توان بیان کرد که الگوریتم C5 صحت بالاتری را در تفکیک مشتریان خسارتی و غیرخسارتی ارائه می‌دهد. ماتریس متقابل این الگوریتم در جدول ۴ قابل مشاهده است.

جدول۴. ماتریس متقاطع درخت تصمیم C5

مدل / واقعیت	طبقه‌بندی‌شده در طبقهٔ غیرخسارتی	طبقه‌بندی‌شده در طبقهٔ خسارتی
طبقهٔ غیرخسارتی	۹۶۴	۱۱۳۸
طبقهٔ خسارتی	۳۰۴	۶۲۷۷

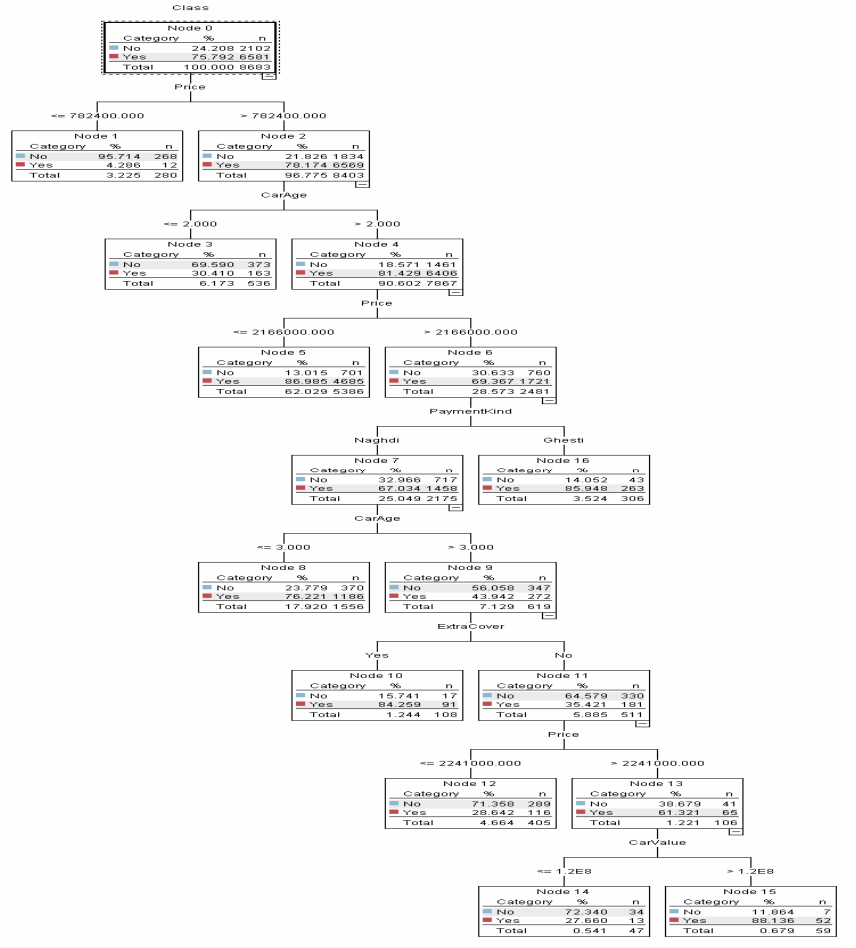
نظر به اینکه صحت الگوریتم از سایر الگوریتم‌ها بیشتر شده است، بنابراین اهمیت متغیرها (خصیصه‌های بیمه‌گذاران) از لحاظ این الگوریتم جهت تفکیک بیمه‌گذاران خسارتی و غیرخسارتی به‌عنوان یک دانش سازمانی مناسب می‌تواند مورد استفاده بیمه‌گران قرار گیرد (نمودار ۵).

نمودار ۵. اهمیت متغیرها از لحاظ الگوریتم C5



همچنین درخت تصمیم به‌دست‌آمده نیز به‌عنوان دانش استخراج‌شده از داده‌ها برای پیش‌بینی وضعیت طبقهٔ مشتریان جدید مورد استفاده قرار می‌گیرد (نمودار ۶).

نمودار ۶. درخت تصمیم C5



تفسیر درخت فوق بدین صورت است:

از ریشه تا هر برگ یک قانون، نمایش داده شده است. تفسیر قوانین به‌دست‌آمده به‌این‌صورت است:

- اگر مبلغ حق‌بیمه کوچک‌تر یا مساوی ۷۸۲،۴۰۰ باشد مشتری کم‌ریسک است.
- اگر مبلغ حق‌بیمه بزرگ‌تر از ۷۸۲،۴۰۰ باشد آنگاه باید سن خودرو درنظرگرفته‌شود:

– اگر سن خودرو کوچک‌تر و مساوی ۲ سال باشد مشتری کم‌ریسک است.
– اگر سن خودرو بزرگ‌تر از ۲ سال باشد آنگاه باید مبلغ حق‌بیمه درنظرگرفته‌شود:
– اگر مبلغ حق‌بیمه کمتر یا مساوی ۲۱۶،۶۰۰ باشد مشتری ریسک بالایی دارد.
– اگر مبلغ حق‌بیمه بیشتر از ۲۱۶،۶۰۰ باشد آنگاه باید نحوه پرداخت حق‌بیمه درنظرگرفته‌شود:
– اگر نحوه پرداخت حق‌بیمه قسطی باشد مشتری ریسک بالایی دارد.
– اگر نحوه پرداخت حق بیمه نقدی باشد آنگاه باید سن خودرو درنظرگرفته‌شود:
– اگر سن خودرو کوچک‌تر و مساوی ۳ سال باشد مشتری ریسک بالایی دارد.
– اگر سن خودرو بزرگ‌تر از ۳ سال باشد آنگاه باید پوشش‌های اضافی بیمه خودرو درنظرگرفته‌شود:
– اگر بیمه‌گذار پوشش‌های اضافی برای بیمهٔ خودرو درخواست داده باشد مشتری ریسک بالایی دارد.
– اگر بیمه‌گذار پوشش‌های اضافی برای بیمهٔ خودرو درخواست نداده باشد آنگاه باید مبلغ حق‌بیمه درنظرگرفته‌شود:
– اگر مبلغ حق‌بیمه کمتر یا مساوی ۲،۲۴۱،۰۰۰ ریال باشد مشتری ریسک بالایی دارد.
– اگر مبلغ حق‌بیمه بیشتر از ۲،۲۴۱،۰۰۰ ریال باشد آنگاه باید ارزش خودرو درنظرگرفته‌شود:
– اگر ارزش خودرو کمتر یا مساوی ۱۲۱،۰۰۰،۰۰۰ ریال باشد مشتری ریسک پایینی دارد.
– اگر ارزش خودرو بیشتر از ۱۲۱،۰۰۰،۰۰۰ ریال باشد مشتری ریسک بالایی دارد.

۴. نتیجه‌گیری

نتایج پژوهش نشان داد که درخت تصمیم C5 نسبت‌به سایر الگوریتم‌های مورد استفاده برای طبقه‌بندی بیمه‌گذاران، صحت بالاتری داشته است. این الگوریتم، مشتریان را با دقت ۷۶/۴درصد در دو دسته از پیش تعیین‌شده بیمه‌گذاران

خسارت‌دیده (خسارتی) و خسارت‌ندیده (غیرخسارتی) طبقه‌بندی کرد که نشان از قوت این الگوریتم در بین سایر الگوریتم‌های اجراشده در این پژوهش است؛ بنابراین دانش استخراج‌شده از این درخت به‌عنوان مورد اعتمادترین دانش داده‌های مورد بررسی است. ازاین‌رو با استفاده از این پژوهش، توان پیش‌بینی‌کنندگی ریسک بیمه‌گذاران این خصیصه‌ها به‌طور عملی مورد آزمون قرار می‌گیرد و می‌توان مقایسهٔ ضمنی نیز بین روش‌های اکچوئری تعیین نرخ و روش‌های داده‌کاوی داشته باشیم. از نقاط قوت مدل ارائه‌شده در این پژوهش، صحت نسبتاً بالای این مدل در مقایسه با مدل‌های ارائه‌شده است. قوانین حاصل از این درخت تصمیم، الگوی پنهان درون پایگاه دادهٔ شرکت بیمه تلقی می‌گردد و می‌توان با استفاده از این الگو، مشتریان خسارتی و غیرخسارتی را پیش‌بینی کرد و سیاست‌گذاری‌هایی را در خصوص هر دو گروه از مشتریان اعمال کرد. این سیاست‌گذاری‌ها می‌تواند در جهت کاهش ریسک بیمه‌گذاران باشد که یکی از آنها اعمال تخفیفات برای مشتریان غیرخسارتی و اعمال جریمه برای مشتریان خسارتی با افزایش نرخ بیمه و اجرای محدودیت‌هایی برای آنهاست. بنابراین با استفاده از نتایج این پژوهش شرکت‌های بیمه می‌توانند نقش این خصیصه‌ها را در پیش‌بینی طبقات ریسک بیمه‌گذاران در یابند و از آن جهت اعمال تخفیفات و جریمه‌ها استفاده کنند.

۵. پیشنهادها برای پژوهش‌های آتی

یافته‌های این پژوهش می‌تواند راهنمایی برای انجام پژوهش‌های آتی باشد:

– استفاده از تکنیک قوانین وابستگی (در حالت با نظارت) برای پیش‌بینی ریسک اعتباری مشتریان و استخراج قوانین مفید.

– خوشه‌بندی مشتریان در گروه‌های مختلف و سپس طبقه‌بندی مشتریان در هر خوشه برای کسب دانش در شناسایی وضعیت اعتباری مشتریان در هر گروه مشابه از مشتریان.

منابع

۱. آقابیگی، ژینا و رضایی، سعید ۱۳۸۶، 'اعتبارسنجی مشتریان اعتباری بانک ملی براساس تکنیک‌های داده‌کاوی (رگرسیون لجستیک)'، *مجموعه مقالات اولین کنفرانس داده‌کاوی ایران*، دانشگاه صنعتی امیرکبیر، صص ۷-۶.

۲. چوبدار، سروناز ۱۳۸۷، *طراحی چهارچوبی برای پیش‌بینی مشتریان آتی بیمه بدنهٔ اتومبیل بر پایه داده‌کاوی*، پایان‌نامه کارشناسی‌ارشد، دانشگاه تربیت مدرس.

۳. حسین‌زاده، لیلا ۱۳۸۶، *دسته‌بندی مشتریان هدف در صنعت بیمه با استفاده از داده‌کاوی*، پایان‌نامه کارشناسی‌ارشد، دانشگاه تربیت مدرس.

۴. عنبری، الهام ۱۳۸۹، *طبقه‌بندی ریسک بیمه‌گذاران بیمهٔ بدنهٔ اتومبیل با استفاده از داده‌کاوی*، پایان‌نامه کارشناسی‌ارشد، دانشگاه شهید بهشتی.

۵. غضنفری، مهدی، علیزاده، سمیه و تیمورپور، بابک ۱۳۸۷، *داده‌کاوی و کشف دانش*، انتشارات دانشگاه علم و صنعت، چ۱، صص ۶۶-۳۵۷.

۶. مروج، مصطفی ۱۳۸۳، *افزودن قابلیت داده‌کاوی فازی به بانک‌های اطلاعاتی رابطه‌ای*، پایان‌نامه کارشناسی‌ارشد، دانشگاه صنعتی امیرکبیر.

۷. منیری، آرش ۱۳۸۵، *استفاده از ماشین بردار پشتیبان در بازشناسی کلمات گسسته فارسی*، پایان‌نامه کارشناسی‌ارشد، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران.

8. Bellotti, T & Crook, J 2008, *Support vector machines for credit scoring and discovery of significant features*, Expert Systems with Applications, pp. 102-9.

9. Bigus, JP 1996, *Data mining with neural networks*, Mc Graw-Hill Publi cation, New York, USA.

10. Chapman, p Clinton, J, Kerber, R, Khabaza, T, Reinartz, T, Shearer, C & Wirth R 1999, *CRISP-DM 1.0: Step-by-step data mining guide*, Viewed 22 October 2011<http://www.crisp-dm.org/CRISPwP-0800.pdf>.

11. Edward, FR & Mishkin, FS 1995, *The decline of traditional banking: implication for financial stability and regulatory policy*, Federal Reserve Bank of New York Policy Review, pp. 27-45.

12. Fayyad, UM, Piatetsky-Shapiro, G & Smyth, P (Editors)1996, Advances in knowledge discovery and Data mining, AAA Press/MITPress, Menlo Park, CA.

13. France, K 2003, *Credit scoring process from a knowledge management prospective*, Budapest University of Technology And Economics, pp. 96-108.
14. Gayle, S 1999, *Data mining in insurance industry*, SAS Institute Inc.
15. Goa, L 2003, *Applying DM in property/casualty insurance*, University of Central Florida.
16. Kantardzic, M 2003, *Data mining: concepts, models, methods, and algorithms*, 1st ed, Wiley.
17. Koh, HC & Low, Ck 2004, 'Going concern prediction using data mining techniques', *Managerial Auditing Journal*, vol. 19, no. 3, pp.462 – 76.
18. Lee, SJ & Siau, K 2001, 'A review of data mining techniques', *Industrial Management & Data Systems*, vol. 101, no. 1, pp.41 – 6.
19. Sanford, G 1999, *Data mining in insurance industry*, SAS Institute Inc.
20. Skipper, HD & Kwon, WJ 2007, *Risk management and insurance: perspective in a globall economy*, Wiley, New York, USA, vol. 1, p. 23-25.
21. Yeo, AC, *Smith*, KA, Willis, RJ & Brooks, M 2001, *Modeling the effect of premium changes on motor insurance customer retention rates using neural networks*, Computational Science - ICCS 2001: International Conference, San Francisco, CA, USA,May 28-30, 2001, Proceedings, Part II.