Journal of New Achievements in Electrical, Computer and Technology
Vol 2, No 1, Mar-Jun 2022,(107-126)
WWW.JNAECT.IR

مجله دستاوردهای نوین در برق،کامپیوتر و فناوری
دوره ۲، شماره ۱، بهار ۱۴۰۱ (۱۲۶-۱۰۷)

پیشنهاد روش نوینی برای پیش بینی و تشخیص کلاهبرداری مشتریان و بیمه شدگان در شرکتهای بیمه به منظور کاهش میزان خسارات مالی وارده به این شرکت‌ها

علی پاسبان اسدآبادی[۱](نویسنده مسئول)، زهرا اسماعیلی[۲]، علی اسماعیلی[۳]

[۱]دانشجوی کارشناسی ارشد ،گروه مهندسی کامپیوتر مهندسی نرم افزار، دانشگاه آزاد اسلامی، واحد الکترونیک ،تهران، ایران.

[۲]فوق لیسانس مدیریت دولتی گرایش مدیریت رفتار سازمانی دانشکده مدیریت و اقتصاد دانشگاه آزاد اسلامی واحد علوم و تحقیقات - تهران – ایران.

[۳]دانشجوی دکتری تخصصی رشته فناوری اطلاعات گرایش شبکه و رایانش دانشکده فنی مهندسی دانشگاه آزاد اسلامی واحد تهران شمال . تهران – ایران.

Alipasban1370@gmail.com , marjanoxygen@gmail.com , ali_e_24@yahoo.com

چکیده

متاسفانه مشاهده می شود که به خصوص در کشور ایران، تمامی سازمانهای بیمه با خطر ورشکستگی مواجه هستند، برای پاسخ به علت این امر، به تقلب ها و کلاهبرداری های مختلفی که در بیمه بدنه ماشینها، بیمه ساختمان، بیمه آتش سوزی و ... می توان نظری انداخت که واقعا این کلاهبرداری ها هزینه های سنگینی را برای سازمانهای بیمه تحمیل می کند. در این تحقیق بر روی یکی از بزرگترین معضلات کنونی در صنعت بیمه پرداختیم. یکی دوسالی از ایجاد نسخه های الکترونیکی در ایران می گذرد ولی با حجم بالایی از استفاده متقلبانه از دفترچه های بیمه افراد دیگر به منظور خرید داروهای خاص و یا استفاده از خدمات پاراکلینیکی و ... مواجه هستیم که این امر هم باعث افزایش فشارهای مالی بر سازمانهای بیمه شده است. در این تحقیق از تکنیکهای داده کاوی برای ارائه روش نوینی به منظور کشف موارد تقلب در نسخه های الکترونیکی به منظور دریافت داروهای خاص با دفترچه بیمه افراد دیگر استفاده نمودیم. نتایج بدست آمده نشان داد که استفاده از قوانین انجمنی می تواند با دقت بالا رابطه بین رفتارهای متقلبانه و ویژگیهای دیگر شخص متقلب را کشف نماید و استفاده از جنگلهای تصادفی در مقایسه با روشها و الگوریتمهای دیگر، با دقت بالاتری می تواند به شناسایی این موارد خطاکار یا ناهنجار منتهی شود.

کلمات کلیدی: پیش بینی و تشخیص، کلاهبرداری، مشتریان و بیمه شدگان، شرکتهای بیمه، خسارات مالی، داده کاوی.

مقدمه

از آنجا که در دهه اخیر به علل مختلف (اعم از تغییرات اقلیمی و آب و هوایی، پایین آمدن کیفیت غذاهای تولیدشده، تغییر سبک زندگی مردم و روی آوردن آنها به زندگی ماشینی و بی تحرک، ...) بر تعداد بیماریها در جوامع مختلف افزوده شده و برخی از بیماریها همانند کرونا تمام مردم دنیا را به شدت گرفتار نموده و داروهای معالجه برخی از این بیماریها اغلب بسیار گرانقیمت هستند لذا، در برخی موارد این بالابودن قیمت داروها می تواند عاملی برای سوء استفاده جمعی از مردم باشد. لذا، در حال حاضر، کلاهبرداری با نسخه یکی از انواع تقلب در مراقبت های بهداشتی است که در اغلب کشورها (منجمله ایران) مرسوم بوده و بار سنگینی را بر دوش سازمان تامین اجتماعی و شرکت های بیمه خصوصی قرار می دهد. این نوع تقلب، با استفاده از تجویز بیش از حد دارو صورت می گیرد. این تقلب نه تنها صدمات مالی بسیار زیادی می تواند برای شرکتهای بیمه و تامین اجتماعی وارد نماید بلکه برای محققان پزشکی و فراپزشکی می تواند باعث ایجاد مشکل در مطالعه و آنالیز ویژگی های بیماران بر اساس داروهای تجویز شده، گردد. تصور جامعه مبنی بر اینکه تقلب در تجویز داروهای غیرضروری، یک جرم بدون قربانی است، باعث شده تا ارتکاب آنها ساده گرفته شده و به کرات شاهد آن باشیم و زنجیره کلاهبرداری بین شرکت های دارویی، پزشکان، داروخانه ها و بیماران، تقویت شود.

با در نظر گرفتن اینکه همه روزه حجم عظیمی از داده های مرتبط با سیستم مراقبت های بهداشتی تولید شده و به ادرات و سازمانهای بیمه ارسال می شوند، لزوم ایجاد سیستمی برای تشخیص انواع مختلف تقلب های صورت گرفته شده در نسخه های پزشکی بسیار مشهود می باشد. از آنجا که تقریبا نیمی از هزینه های شرکت های بیمه صرف مصارف دارویی می شود لذا، می توان متوجه شد که هزینه های نسخه های تقلبی برای سازمانهای بیمه و تامین اجتماعی بسیار سنگین است. بنابراین هر سیستمی که بتواند نسخه هایی را که احتمال کلاهبرداری خاصی را که توسط کاربر انجام می شود را، پیدا کند، می تواند کمک شایانی به این سازمانهای بیمه بنماید. این سیستم خودکار باید به گونه ای باشد که با حداقل مقدار منفی کاذب عمل کند، یعنی حداقل مقدار نسخه های تقلبی بدون تشخیص باقی بماند.

این مقاله در شش بخش تنظیم شده است. در بخش دوم پیشینه تحقیقات انجام شده توسط محققان مختلف که مرتبط به این تحقیق است را، مورد بررسی قرار خواهیم داد. در بخش سوم درباره تکنیکهای داده کاوی توضیحاتی ارائه می کنیم. در بخش چهارم روش پیشنهادی را توضیح خواهیم داد و در بخش پنجم نحوه پیاده سازی و ارزیای روش پیشنهادی را توضیح خواهیم داد و نهایتا در بخش ششم نتیجه گیری را خواهیم آورد و کارهای آتی را بیان خواهیم کرد.

۱- پیشینه تحقیق

تجزیه و تحلیل مطالبات یکی از جنبه های مهم تجزیه و تحلیل پیش بینی کننده در صنعت بیمه است زیرا تقریباً ۸۰ درصد درآمد حاصل از حق بیمه برای مطالبات شرکت های بیمه هزینه می شود [۱]. بنابراین، تجزیه و تحلیل کامل مطالبات برای بهبود جریان هزینه ها ضروری است. با تجزیه و تحلیل داده های بیمه ، روابط بین عوامل مختلف (متغیرها) مشاهده شده و تابعی از مدل پیش بینی ها به دست می آید [۲]. از این پیش بینی ها می توان برای تصمیم گیری استفاده کرد. جدا از داده های ساختاریافته مورد استفاده شرکت های بیمه، گستره عظیمی از داده های بدون ساختار وجود دارد که اطلاعات حیاتی را ارائه می دهد [3, 4]. می توان از آن داده ها برای دسته بندی دسته های مختلف ذینفعان و محاسبه پرداخت و پردازش مورد انتظار برای این دسته ها استفاده کرد [5]. برخی از شاخص های کلیدی عملکرد (KPI) در مورد خسارت های بیمه مانند زمان چرخه خسارت، رضایت مشتری ، تشخیص تقلب، بازیابی خسارت و هزینه های رسیدگی به خسارت وجود دارد [6]. مشاهده شده است که از حجم عظیمی از داده هایی که یک شرکت بیمه به آن دسترسی دارد ، تنها از ۱۰ تا ۱۵ درصد از این داده ها استفاده می کند [5. 6] .تکنیکهای یادگیری ماشین می تواند به افزایش استفاده از داده ها با در نظر گرفتن KPIادعا شده توسط استفاده کننده ها، کمک کند و بسیاری از فرایندهای معمول را برای کاهش زمان چرخه ادعاها، افزایش رضایت مشتری ، مبارزه با تقلب ، بهینه سازی بازیابی ادعاها و کاهش هزینه های رسیدگی به ادعا اتخاذ نماید.

Melih Kirlidog و همکارش [7] روش های داده کاوی مانند تشخیص ناهنجاری، خوشه بندی و طبقه بندی که می توانند با موفقیت ناهنجاری ها یا فاصله های زیاد را در مجموعه وسیعی از داده ها تشخیص دهند را، مورد بررسی قرار دادند. این امر، می تواند برای صنعت بیمه که با مطالبات کلاهبرداری مشکل دارد بسیار مفید باشد. پس از تشخیص ادعاهای ناهنجار، باید چندین تجزیه و تحلیل در مورد آنها انجام شود تا تحقیقات کاملی انجام شود. وظیفه اصلی در این تجزیه و تحلیلها محدود کردن هدف برای تشخیص کلاهبرداری است. اگرچه اکثر الگوهای کلاهبرداری معمولاً توسط کارشناسان بیمه شناخته می شود، اما چنین تحقیقاتی می تواند برخی الگوهای جدید و ناشناخته را نیز آشکار کند. Noorhannah Boodhun و همکارش [8] راهکارهایی برای افزایش ارزیابی ریسک در شرکت های بیمه عمر با استفاده از تجزیه و تحلیل پیش بینی ارائه نمودند. دنیای واقعی برای انجام تجزیه و تحلیل از مجموعه داده ای با بیش از صد ویژگی استفاده شده است. در این تحقیق، الگوریتم های یادگیری ماشین، یعنی رگرسیون خطی چندگانه، شبکه عصبی مصنوعی، طبقه بندی کننده های REPTree و Random Treeبرای پیش بینی سطح ریسک متقاضیان بر روی مجموعه داده اجرا شد. یافته های این پژوهش

نشان داد که الگوریتم REPTree بالاترین عملکرد را با کمترین مقدار متوسط خطای مطلق (MAE) ۱.۵۲۸۵ و کمترین مقدار خطای میانگین مربع (RMSE) ۲.۰۲۷ برای روش CFS نشان داد ، در حالی که رگرسیون خطی چندگانه بهترین عملکرد را برای PCA با کمترین مقادیر MAE و RMSE به ترتیب ۱.۶۳۹۶ و ۲.۰۶۵۹ ، در مقایسه با سایر مدلها دارد. Finkelstein سعی کرد از اطلاعات نامتقارن استفاده کرده و سعی کند ویژگی های فردی را شناسایی کند که با ریسک مرتبط است و با تقاضای بیمه ارتباط دارد [9]، اما هنوز توسط شرکت های بیمه از این اطلاعات استفاده نمی شود. نتایج بدست آمده در این تحقیق، نشان می دهد که مقررات بیمه می تواند نقش مهمی در تعیین عملکرد قیمت گذاری داشته باشد. Rahman از تکنیک های انتخاب ویژگی برای طبقه بندی صحیح داده ها استفاده کرد و ثابت نمود که تکنیک های طبقه بندی در طبقه بندی مشتریان با توجه به ویژگی های آنها بسیار مفید است. Kang الگوریتم هایی برای انتخاب ویژگی جدید برای تجزیه و تحلیل داده های کلی را ارائه دادند [10] . آنها برای پاسخ مداوم بر مدلهای رگرسیون خطی تمرکز کردند، آنها نشان دادند که بسط متغیر پاسخ غیر پیوسته با رگرسیون لجستیک در الگوریتم آنها امکان پذیر است. Seema Rawat و همکارانش [11] بر روی روشهای شناسایی، عوامل معنی دار و تعیین کننده برای تشکیل پرونده و پذیرش ادعاهای مطرح شده توسط بیمه شده ها در ادارات بیمه تحقیق کردند. آنها روش نوینی ارائه دادند که از روشهای یادگیری ماشین استفاده کرده و بر اساس تجزیه و تحلیل داده های اکتشافی (EDA) و تکنیک های انتخاب ویژگی است. همچنین، الگوریتم های یادگیری ماشین را بر روی مجموعه داده ها اعمال کردند و با استفاده از معیارهای عملکرد، نتایج بدست آمده را مورد ارزیابی قرار دادند.

۲- مفاهیم کلیدی

۱-۳-سازمانهای ارائه دهنده خدمات بیمه در ایران

انواع مختلفی از خدمات برای اشخاص در سازمانهای مختلف بیمه ارائه می شود لذا بیمه های اشخاص به سه گروه کلی تقسیم می شوند که عبارتند از [۱].

- بیمه اشخاص (درمان)
- بیمه اشخاص (عمر)
- بیمه اشخاص (حوادث)
- بیمه تکمیلی

۲-۳-شرکتهای ارائه دهنده خدمات بیمه در ایران

در حال حاضر، ۲۵ شرکت بیمه در ایران فعال هستند که تنها یک شرکت رسما دولتی به شمار می‌رود. بیمه‌های باسابقه که در جریان انقلاب اسلامی به نفع دولت مصادره شده بودند اکثرا در سال‌های دهه ۸۰ مجددا در جریان خصوصی‌سازی، خصوصی شدند. تعرفه های تعیین شده برای ارائه بیمه نامه های مختلف از طرف سازمان بیمه مرکزی و بر اساس نوع خدمات ارائه شده توسط هر سازمان می باشد.

۳-۳-رشد صنعت بیمه در ایران و جهان

بیمه در زبان فرانسه Assurance و در زبان انگلیسی Insurance اطلاق می شود. می توان گفت که هر دو این معانی که از ریشه لاتینی secures به معنای اطمینان گرفته شده است، اما معادل آن در پارسی را می توان برگرفته از "بیم" که همان عدم اطمینان خاطر از حصول نتیجه مطلوب می باشد دانست [12].

اولین فعالیتهای مربوط به بیمه برای اولین بار در سال ۱۵۵۲ میلادی در شهر فلورانس ایتالیا آغاز شد ولی اولین و مهمترین اجتماع بیمه گران در سده هفدهم توسط شرکت لویدز لندن تشکیل شد.

در ایران، در سال ۱۳۱۰ قانون نظامنامه ثبت شرکتها به تصویب رسید و متعاقب آن در همین سال اولین شرکت بیمه با نام "میهن ما" شروع به کار کرد. در همین زمان بسیاری از شرکتهای بیمه خارجی (مانند: ایگل استار، یورکشایر، رویال، ویکتوریا، ناسیونال سویس، فنیکس، اتحادالوطنی) در ایران نمایندگی ایجاد کردند. شرکت سهامی بیمه ایران توسط دولت در شانزدهم شهریور ۱۳۱۴ با سرمایه ۲۰ میلیون ریال تاسیس شد. در سال ۱۳۲۹ اولین شرکت بیمه خصوصی با نام "بیمه شرق" در ایران تاسیس شد. بیمه مرکزی ایران در سال ۱۳۵۰ برای سامان بخشیدن به فعالیتهای شرکتهای بیمه و ایجاد ارتباط بین فعالیتهای آنها و ایجاد امکان برای نظارت دولت بر رعایت ضوابط و قوانین وضع شده برای حفظ حقوق بیمه گر و بیمه شونده تاسیس شد. در حال حاضر در حدود ۲۵ شرکت بیمه داخلی در ایران در حال فعالیت می باشند [13]. صنعت بیمه در ایران که دارای سهم ۰.۰۹ درصدی از کل حق بیمه های تولیدی جهان است دارای رتبه چهل و ششم در صنعت بیمه جهان می باشد در حالیکه رتبه کشور ترکیه در صنعت بیمه سی و چهار می باشد. همچنین در ایران، سرانه حق بیمه ۵۰ دلار است در حالیکه سرانه حق بیمه در جهان ۶۰۸ دلار می باشد لذا ایران از لحاظ سرانه حق بیمه دارای رتبه ۷۶ در کل جهان می باشد. مقایسه سهم بیمه های زندگی و غیر زندگی در ایران با میانگین جهانی را نشان می دهد که، سهم بیمه های زندگی (بیمه عمر و پس انداز) در ایران بسیار پایین از نرخ میانگین جهانی می باشد. در حالیکه سهم بیمه های غیرزندگی (بیمه شرکتها، محصولات دریایی، خودروها و...) در ایران بسیار بالاتر از نرخ میانگین جهانی است [13].

نرخ رشد واقعی سالانه پرداخت های ناخالص مطالبات مشتریان (بیمه گذاران) در بخش بیمه عمر در سال ۲۰۱۹ نشان می دهد که متاسفانه کشور ما ایران در این جدول قرار ندارد یعنی نرخ رشد سالانه پرداختهای ناخالص مطالبات در بخش بیمه عمر در ایران بسیار ناچیز و در حد بسیار منفی باید باشد. بیشترین نرخ رشد واقعی سالانه پرداخت های ناخالص مطالبات در بخش بیمه عمر متعلق به روسیه و فرانسه و کمترین میزان نرخ رشد در بین این چند کشور مطالعه شده مربوط به مصر و برزیل است. همچنین برخی کشورها نرخ رشد منفی دارند که بیشترین نرخ منفی محاسبه شده در بین این کشورها، برای کشور انگلستان بوده است. متاسفانه کشور ایران در این مطالعه مدنظر قرار نگرفته است یعنی احتمالا نرخ رشد بسیار منفی را در کشور ایران در این زمینه شاهد بوده ایم لذا در این لیست گنجانده نشده است. نرخ‌های رشد واقعی سالانه پرداخت‌های ناخالص خسارت در بخش غیرزندگی، در سال ۲۰۱۹ نشان می دهد که، بیشترین نرخ رشد واقعی سالانه پرداخت های ناخالص خسارت در بخشهای غیرزندگی متعلق به لوگزامبورک و ایرلند و کمترین نرخ نیوزیلند و استرالیا می باشد. السالوادور بیشترین درصد منفی را در این شکل دارا می باشد و متاسفانه به نظر میرسد درصد منفی رشد ایران بسیار پایین تر از این کشورهاست.

۳- روش پیشنهادی

در این تحقیق برای شناسایی موارد تقلب و یا کلاهبرداری در نسخه های الکترونیکی از تکنیکهای یادگیری ماشین به شرح زیر استفاده خواهیم کرد.

- **در مرحله اول**- تمام بیمه شدگان را با استفاده از تکنیکهای رده بندی به گروههایی شامل (دارای بیماری خاص، افراد مسن، نوزادان و کودکان، افراد های ریسک و افراد سالم) طبقه بندی خواهیم کرد.

  - گروه دارای بیماری خاص: شامل تمام بیمارانی است که دارای بیماری های علاج ناپذیر یا خاص هستند مانند: سرطان، ام اس، تالاسمی، فشارخون، پارکینسون، آلزایمر، برخی از بیماریهای عصبی یا عقب ماندگی های ذهنی کودکان، بیماریهای قلبی و.... این بیماران داروهای خاصی را هر ماه بایستی از هلال احمر بایستی دریافت کنند که اغلب دوز خاصی باید مصرف کنند و اغلب داروهای خارجی آنها با ارز دولتی وارد می شود و هزینه آنها در داروخانه های هلال احمر بسیار کمتر از بازار آزاد است و این داروها بیشترین سود را برای کلاهبرداران دارند تا بتوانند با جعل نسخه این داروها را دریافت کنند چون اغلب این داروها با شماره کارت ملی به بیماران تحویل داده می شود.

- افراد مسن، اغلب این افراد دارای بیماری های خاصی مانند فشار خون، چربی بالا و... هستند که به علت کهولت سن کاملا طبیعی می باشد.
- نوزادان و کودکان: برخی از داروهای مخصوص نوزادان و کودکان با ارز دولتی وارد کشور می شوند به خصوص داروهای مخصوص بیماریهای مادرزادی و یا بیماریهای حاد کودکان که تقلب در این داروها هم می تواند باعث ضرر و زیان زیادی برای بیمه ها شود.
- افراد های ریسک: افرادی که در خانواده درجه اول آنها به خصوص پدر، مادر، خواهر و برادر سابقه ابتلا به بیماریهای صعب العلاج وجود دارد.
- افراد سالم: این افراد هیچ بیماری خاصی ندارند.

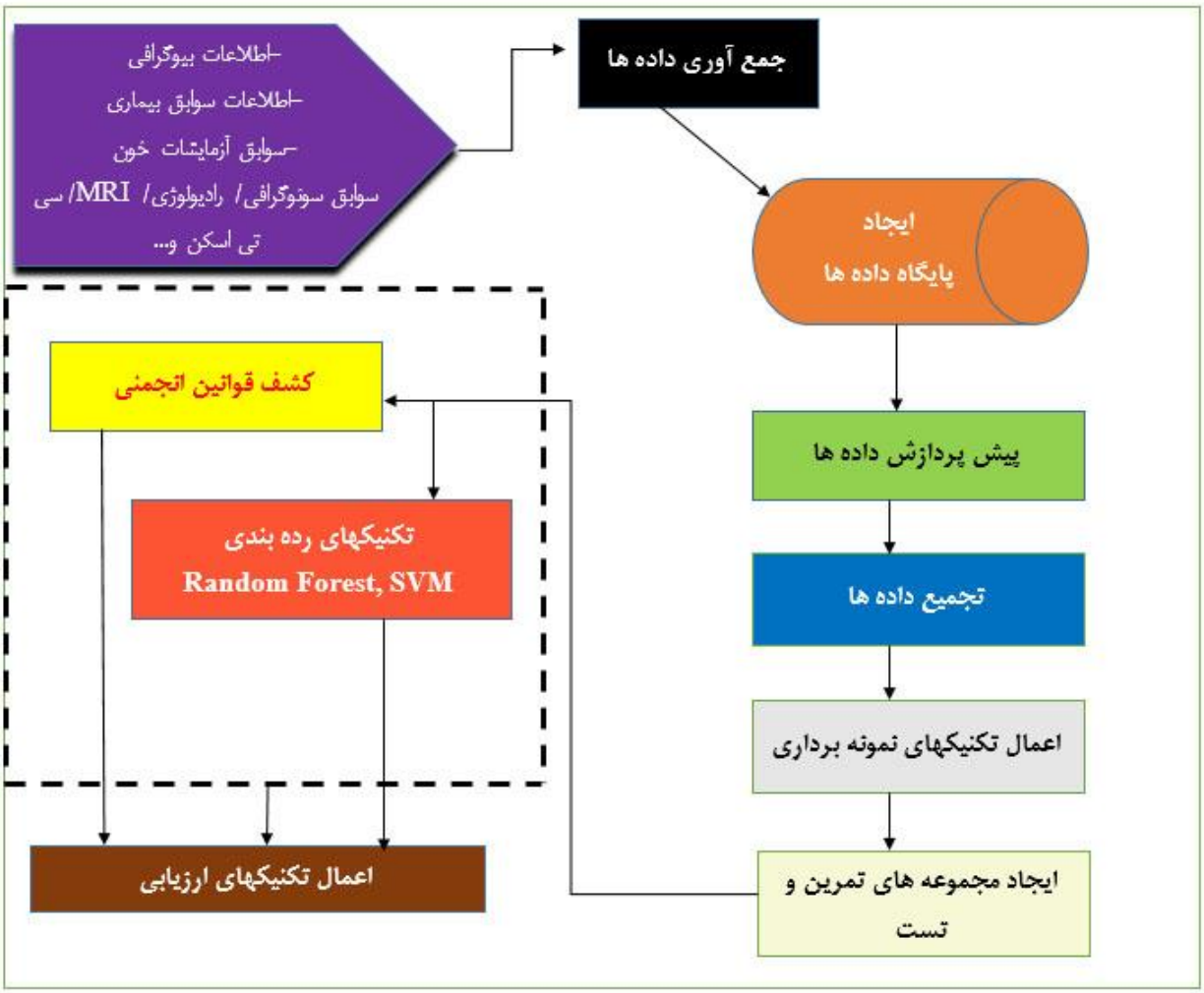
- **در مرحله دوم،** در هر گروه، بیماران را بر اساس درجه بیماری و داروهای دریافتی و میزان پیشرفت بیماری و خصلت های دیگر خوشه بندی خواهیم کرد.
- **در مرحله سوم،** به پایش داروهای تجویز شده برای بیماران در هر خوشه پرداخته و با استفاده از تکنیکهای کشف آنومالی اقدام به شناسایی موارد پرت یا آنومالی خواهیم کرد که این موارد با احتمال بسیار بالایی همان موارد تقلب را تشکیل خواهند داد.

شکل ۱ شمای کلی روش پیشنهادی در این تحقیق را نشان می دهد.

مراحل انجام شده در شکل ۱ عبارتند از:

- جمع آوری داده ها: داده های مورد نیاز در این پژوهش از منابع مختلف (سابقه انجام آزمایشات خون، ادرار، سونوگرافی،رادیولوژی، MRI، CTScan و....) و سابقه استفاده از داروهای خاص، بیماریهای زمینه ای و....
- ذخیره در پایگاه داده ها: تمام داده های جمع آوری شده از منابع مختلف در پایگاه داده ها ذخیره می شوند.
- اعمال پیش پردازش داده ها، داده های جمع آوری شده ممکنست حاوی برخی از فیلدهای خالی بوده یا دارای برخی داده های نامرتبط باشند که این رکوردهای ناقص در این مرحله حذف خواهند شد.
- تجمیع داده ها: در این مرحله، داده های جمع آوری شده از منابع مختلف باهم تجمیع شده و سپس اقدام به انتخاب بهینه ترین ویژگیها از بین کل ویژگیهای جمع آوری شده می نماییم.

- اعمال تکنیکهای نمونه برداری: تکنیکهای نمونه برداری مختلفی وجود دارد که پرکاربردترین آنها عبارتند از: روش نمونه برداری ساده تصادفی (بدون جایگزینی و با جایگزینی)، روش نمونه برداری سیستماتیک، روش نمونه برداری Stratified، روش نمونه برداری خوشه ای.



شکل ۱-شمای کلی روش پیشنهادی در این تحقیق

- ایجاد مجموعه تمرین و مجموعه تست: با اعمال تکنیکهای نمونه برداری، می توانیم مجموعه تمرین و مجموعه تست را ایجاد کنیم. مجموعه تمرین برای ایجاد الگو و خوشه بندی به کار خواهند رفت و مجموعه تست برای ارزیابی الگوهای ایجاد شده مورد استفاده قرار خواهند گرفت.
- اعمال تکنیکهای رده بندی: در این تحقیق از تکنیکهای رده بندی درخت تصمیم ، AdaBoost Regressor و Gradient Boosting برای ایجاد الگو برای هر بیمه شده استفاده خواهیم کرد با استفاده از این الگو می توان پیش بینی کرد که نیازهای دارویی هر شخص چه مواردی می باشند.

- ایجاد خوشه بندی: برای هر دسته از بیمه شدگان که با استفاده از تکنیکهای رده بندی شناسایی شدند، تکنیکهای خوشه بندی سلسله مراتبی استفاده خواهیم کرد. در هر گروه از بیماران بر اساس شدت و پیشرفت بیماری خوشه های مختلفی ایجاد خواهیم کرد زیرا این بیماران بر اساس شدت بیماری به داروهای مختلف با دوزهای متفاوت نیاز دارند.
- استفاده از تکنیکهای کشف آنومالی: شناسایی داده های پرت و ناهنجار یکی از راههای شناسایی تقلب ها یا کلاهبرداریهای انجام شده می باشد. با استفاده از روشهای مبتنی بر چگالی اقدام به شناسایی موارد آنومالی یا داده های پرت در هر خوشه از کاربران خواهیم کرد.
- اعمال تکنیکهای ارزیابی الگوها یا تکنیکهای شناسایی آنومالی ها: با اعمال معیارهای مختلف از قبیل F-measure, Recall, Precision, Accuracy اقدام به ارزیابی الگوها یا روشهای اعمال شده و پیش بینی های انجام شده در این تحقیق خواهیم کرد.

۴- پیاده سازی و ارزیابی نتایج

مراحل انجام این پیاده سازی به شرح زیر است.

۵-۱-مرحله اول-جمع آوری داده ها

یکی از بزرگترین چالشها در انجام تحقیقات علمی پژوهشی در حوزه درمان بدست آوردن داده های واقعی از بیمارستانها و بیمه ها می باشد چون به علت احترام به حریم خصوصی بیماران و بنا به برخی قوانین و سیاستهای شرکتهای بیمه و یا بیمارستانها، در اختیار قرار دادن پرونده های پزشکی بیماران از نظر قانونی یا اخلاقی از طرف سازمانها و شرکتهای بیمه مجاز نمی باشد. لذا در این تحقیق، با برخی محدودیت ها در جمع آوری داده مواجه شدیم لذا از یکی از شعب بیمه با حذف مشخصات شخصی از قبیل شماره سریال بیمه، نام و نام خانوادگی، آدرس محل کار، آدرس محل زندگی و سایر مشخصات احتمالی که منجر به شناسایی شخص بیمه شده گردد، داده ها را جمع آوری نمودیم.

۵-۲-پیش پردازش و تبدیل داده ها

داده های جمع آوری دارای برخی از فیلدهای خالی یا داده های نامرتبط و یا فرمت نامشخص هستند که بایستی در این مرحله اقدام به اعمال تکنیکهای پیش پردازش برای حذف یا تصحیح رکوردهای دارای داده های نامرتبط، ایجاد فرمت مناسب برای فیلدها و فایل مورد نظر و در نهایت نرمالسازی بنماییم.

کل رکوردهای داده ای جمع آوری شده در این پایان نامه، ۱۰۰۰ بیمه شده تامین اجتماعی است که در حدود ۵۰ رکورد دارای داده های نامرتبط یا فیلدهای ناقص بودند که در این مرحله حذف شدند و تنها ۹۵۰ رکورد برای آنالیز بیشتر انتخاب شدند.

در ادامه این مرحله با اعمال دو تکنیک نمونه برداری Stratified و نمونه برداری ساده تصادفی بدون جایگزینی اقدام به ایجاد مجموعه های تمرین و تست خواهیم کرد. در نمونه برداری Stratified بر اساس بیماریهای ارثی یا زمینه ای سعی داریم تا بیمه گذاران را انتخاب کنیم و سپس از بین آن گروههای انتخاب شده سعی داریم با اعمال روش نمونه برداری ساده تصادفی بدون جایگذاری اقدام به انتخاب نمونه ها نماییم.

بعد از انتخاب نمونه ها، اقدام به انتخاب ویژگیهای مناسب از بین کل ویژگیهای جمع آوری شده برای هر بیماری می کنیم. چون عمده هدف ما شناسایی کلاهبرداری در نسخه های الکترونیکی است مواردی مانند هزینه ترحیم، زایمان، هزینه های دندانپزشکی، هزینه های چشم پزشکی را از ویژگیهای جمع آوری شده، مدنظر قرار نمی دهیم.

شناسه خاص، جنسیت، وضعیت تاهل، میزان تحصیلات، میزان درآمد ماهیانه، تعداد فرزندان، تعداد افراد تحت تکفل، ابتلا به بیماری خاص، مصرف داروی خاص بیماریهای ناعلاج، هزینه های دارویی، هزینه های پاراکلینیکی، هزینه های بستری در بیمارستان، سابقه بیماری خانوادگی.

۱۲ ویژگی از بین ۲۰ ویژگی جمع آوری شده، برای تجزیه و تحلیل و کشف تقلب ها انتخاب شد. بعد از انتخاب ویژگیها اقدام به ایجاد تبدیلات لازم در داده ها می نماییم تا آماده کار در نرم افزارهای داده کاوی (در این تحقیق از نرم افزار Rapid miner) استفاده خواهیم کرد، باشند. در نهایت از ۹۵۰ رکورد کلی جمع آوری شده پیش پردازش شده، ۷۵۰ رکورد به عنوان مجموعه تمرین و ۲۰۰ رکورد به عنوان مجموعه تست انتخاب می شوند.

۳-۵-تجزیه و تحلیل اولیه داده ها

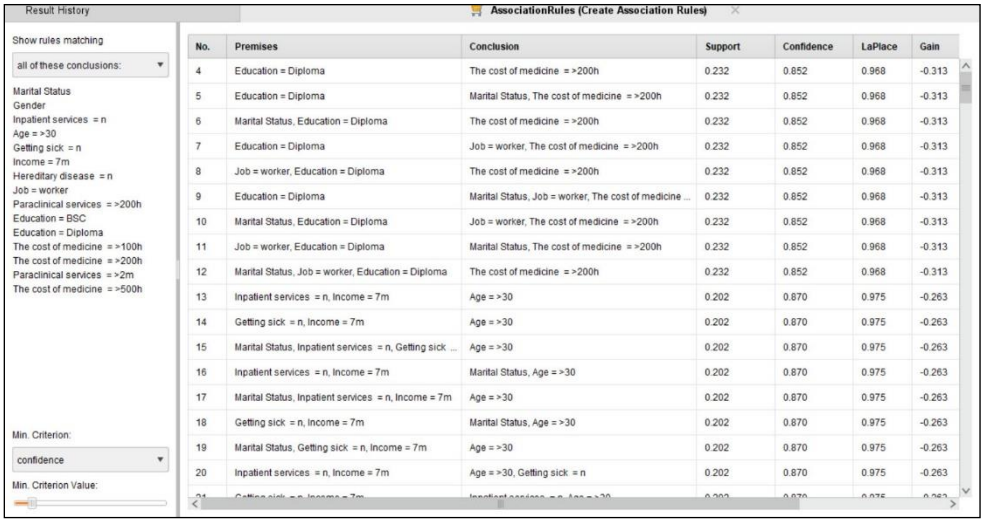
از ابعاد مختلف تجزیه و تحلیل اولیه داده ها را انجام دادیم.

- فراوانی جنسیتی بیمه شدگان را در مجموعه داده های جمع آوری شده نشان می دهد که۵۷ درصد از بیمه گذاران تحت مطالعه آقا و ۴۳ درصد خانم می باشند.
- پراکندگی بیمه گذاران از نظر سنی، نشان می دهد که ۵۷ درصد بیمه گذاران دارای سن بین ۳۰ تا ۴۰ هستند. اقلیت بیمه گذاران تحت مطالعه دارای سن بالای ۵۰ و کمتر از ۶۰ سال (۳ درصد) هستند و ۶۰ سال به بالا (۲ درصد) هستند.

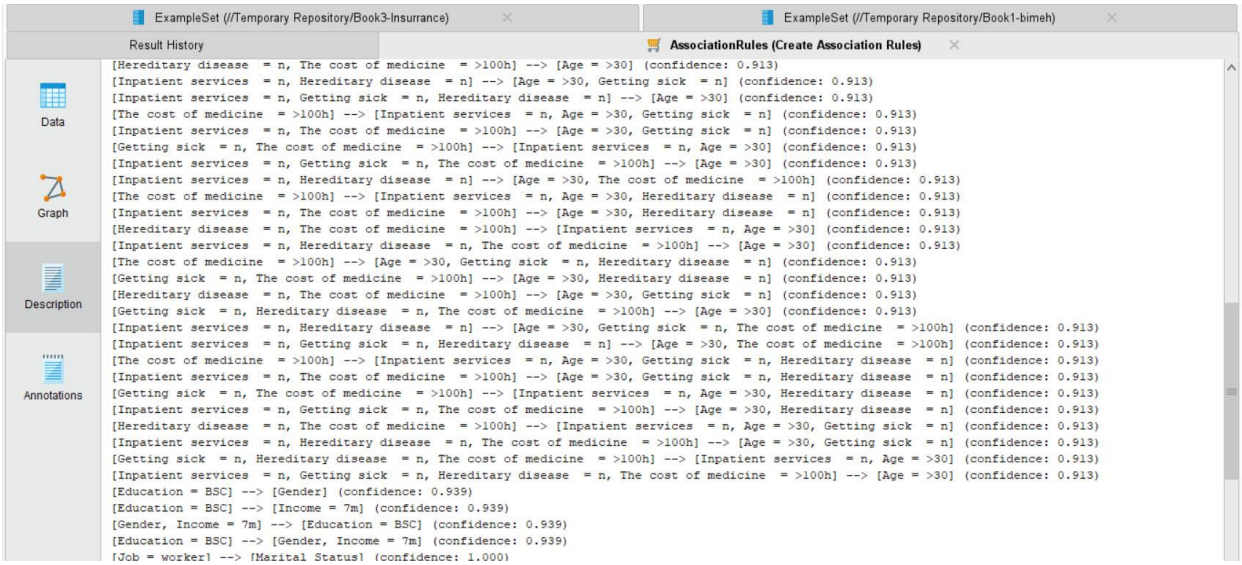
- پراکندگی بیمه گذاران از لحاظ سطح تحصیلات نشان می دهد که، تنها ۳ درصد از بیمه گذاران دارای مدرک دیپلم یا پایین تر از آن هستند و ۳۷ درصد دارای مدرک لیسانس می باشند.
- پراکندگی بیمه گذاران تحت مطالعه از لحاظ وضعیت تاهل نشان می دهد که، ۷۵ درصد از بیمه گذاران تحت مطالعه متاهل و ۲۵ درصد از آنها مجرد می باشند.
- پراکندگی بیمه گذاران بر اساس بیماری لاعلاج که نیاز به مصرف مداوم دارو دارند را، نشان می دهد که، ۴۹ درصد از افراد تحت مطالعه خودشان یا خانواده تحت پوشش و تکفل آنها دارای هیچ نوع بیماری ارثی یا بیماری لاعلاجی که نیاز به استفاده مداوم از دارو باشد، نیستند. در عین حال می بینیم که شایع ترین بیماری ارثی در بین این بیمه گذاران و خانواده های آنها بیماری فشار خون است (٪۳۱)، بیماری سرطان و ام اس هم هر کدام (٪۲) در بین بیمه گذاران و یا افراد خانواده آنها وجود دارد. بیماریهای عصبی مانند دیابت، آلزایمر و پارکینسون در زمره بیماریهای دیگر در نظر گرفته شد که کماکان افراد از داروهای مخصوصی تا آخر عمر خود مجبور هستند تا، استفاده کنند.
- پراکندگی بیمه گذاران بر اساس استفاده از داروهای خاص که برخی از آنها از طریق ارزهای دولتی وارد کشور می شوند و تفاوت بسیار زیادی بین قیمت آنها در داروخانه و بازار سیاه وجود دارد، نشان می دهد که، برخی از داروهای مخصوص بیماریهای سرطان، ام اس، آلزایمر، پارکینسون، فشارخون، دیابت، افسردگی و.... تنها با استفاده از نسخه های پزشک متخصص و به صورت جیره بندی شده (بر حسب دوز استفاده بیماران) به صورت ماهیانه به بیماران از طرف داروخانه های دولتی و یا داروخانه های هلال احمر، با قیمت دولتی عرضه می شوند. تهیه این داروها در بازار سیاه بدون نسخه برای بیماران یا مقدور نیست یا دارای قیمت بسیار بالایی می باشد. لذا شناسایی دقیق بیماران و عرضه صحیح داروهای آنها از اهمیت خاصی هم برای دولت، وزارت بهداشت و هم بیمه های تکمیلی، برخوردار می باشند. ۲۸ درصد از بیمه گذاران یا خانواده تحت تکفل آنها از این نوع داروها استفاده می کنند.

۴-۵-اعمال تکنیکهای داده کاوی

ابتدا قوانین انجمنی را بر روی داده های موجود اعمال نمودیم و نتایج زیر (شکل ۲ و ۳) بدست آمد.



شکل ۲-نمایش عددی روابط ایجاد شده بین متغیرهای مختلف



شکل ۳-کشف قوانین انجمنی

برای توضیح قوانین انجمنی ایجاد شده چند خط از شکل ۳ را توضیح می دهیم.

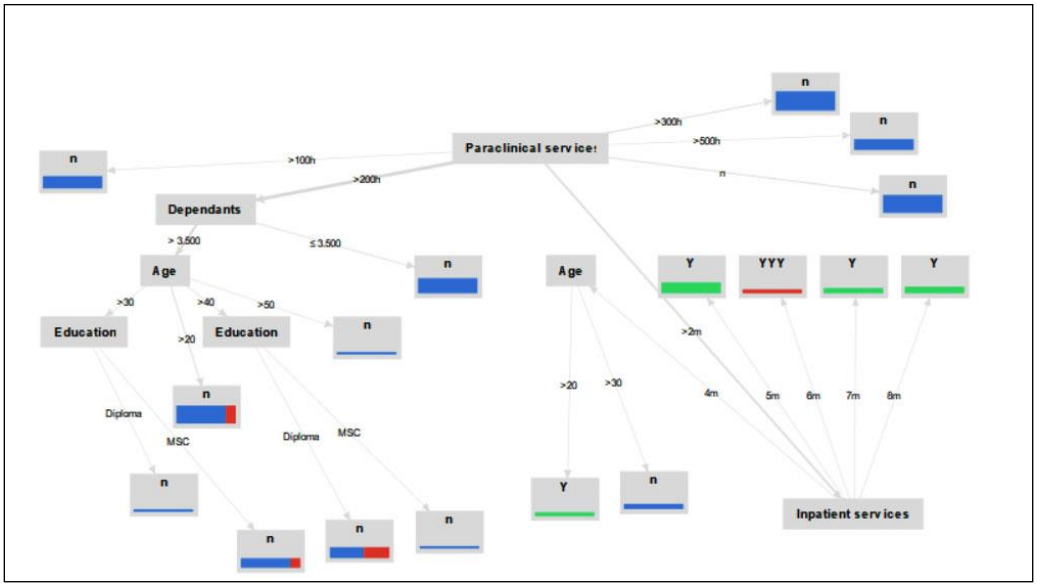
- با اطمینان تقریبا ۸۰.۷ درصدی می توان گفت افرادی که خدمات بستری در بیمارستان استفاده نکرده باشد، اغلب افراد متاهل می باشند.
- با اطمینان ۸۱ درصد می توان گفت افرادی که دارای سن بین ۳۰ الی ۴۰ سال دارند و سابقه بیماری زمینه ای ندارند آنگاه این افراد از سرویسهای بستری استفاده نخواهند کرد.

همانطور که مشاهده می شود قوانین زیادی ایجاد شده است که در شکل ۲ قسمتی از خروجی نشان دهنده میزان Support و Confidence قوانین ایجاد شده را مشاهده می کنیم. چند قانون مهم که می تواند به ما در شناسایی کلاهبرداران کمک نماید عبارتند از:

- با درصد پشتیبانی ۸۰ درصد و درصد Confidence برابر با ۷۰ درصد می توان گفت، افرادی که اقدام به استفاده از نسخه های جعلی در راستای استفاده از داروهای بیماران خاص می نمایند، دارای مدرک فوق لیسانس، مجرد، مذکر و اغلب شغلهای آزاد یا مدیران کسب و کارها هستند.
- با درصد پشتیبانی ۷۰ درصد و درصد اطمینان ۶۵ درصد، داروهایی که مخصوص بیماران ام اس می باشد، بیشترین نرخ کلاهبرداری را دارا می باشند.
- با درصد پشتیبانی ۶۵ درصد و درصد اطمینان ۶۰ درصد، داروهای اعصاب (مانند مرفین، دیازپام، فلوکستین) از جمله داروهایی هستند که در داروخانه ها با جعل نسخه و به صورت آزاد فروخته می شوند در حالیکه وجود نسخه پزشکی برای ارائه این داروها در چندسال اخیر اجباری می باشد.
- با درصد پشتیبانی و درصد اطمینان ۶۰ درصد می توان گفت بین مدرک تحصیلی و مصرف داروهای اعصاب رابطه مستقیمی وجود دارد. افراد دارای مدارک فوق لیسانس و دکترا بیشترین استفاده کنندگان از داروهای ضدافسردگی را تشکیل می دهند.
- با درصد پشتیبانی و اطمینان ۹۰ درصد رابطه مستقیم بین جنسیت و ابتلا به بیماری های ام اس و بیماری افسردگی وجود دارد. به طوریکه اغلب بیماران را جنس مونث اغلب تحصیل کرده (لیسانس و بالاتر) تشکیل می دهند.
- و.....

۵-۵-اعمال تکنیکهای کشف الگو

برای کشف آنومالی در این تحقیق از الگوریتم جنگل تصادفی Random Forest و الگوریتم ماشین بردار پشتیبان (SVM) استفاده نمودیم. یکی از درختهای ایجاد شده در شکل ۴ قابل مشاهده می باشد.



شکل ۴--یکی از درختهای ایجاد شده برای شناسایی موارد استفاده غیرمجاز از داروهای خاص

قابل ذکر است که برچسب های موجود (Y , n, YY) به معنای:

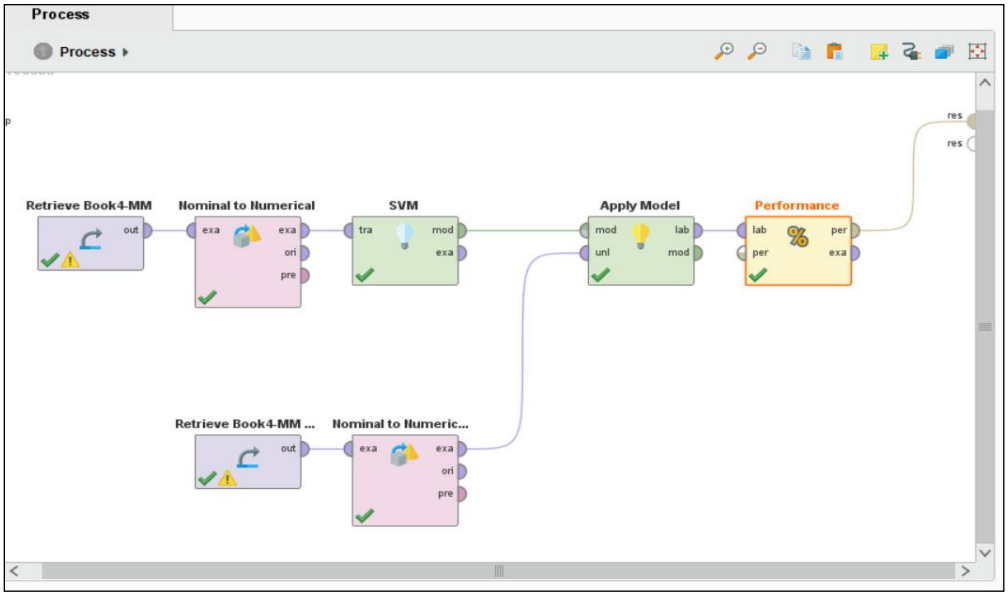
Y نشانگر بیمارانی است که استفاده مجاز و با مجوز پزشک یا مرکز بهداشت از داروهای خاص دارند.

n : نشانگر بیمه گذارانی است که هیچ استفاده ای از داروهای خاص نمی کنند.

YY : بیانگر افرادی است که به صورت غیرمجاز از دفترچه تامین اجتماعی آنها اقدام به خرید داروهای خاص با نسخه الکترونیکی شده است. اغلب این افراد اصلا بیماری خاصی ندارند.

تنها گروه خلافکاری که در این شکل مشاهده می شود شامل کسانی است که پرونده پزشکی بستری در بیمارستان دارند و بیشتر از ۲ میلیون برای آنها هزینه پاراکلینیکی و بیشتر از ۶ میلیون هزینه بستری در بیمارستان به آنها از بیمه پرداخت شده است. ضمن بررسی که انجام دادیم متاسفانه به این نتیجه رسیدیم که این امر بارها دیده شده که متاسفانه برخی از کادر درمانی یا پزشکان با استفاده از پرونده ها و مدارک بیمارانی که در بیمارستان بستری هستند اقدام به تهیه داروهای خاص به اسم آنها برای اهداف مختلف می نمایند. حتی موردی را در یکی از بیمارستانهای دولتی مشاهده کردیم که بیمار برای گرفتن جواب آزمایشات و نوار قلب و اکوی قلب خود به بیمارستان مراجعه کرده بود ولی در پرونده وی هیچ گونه مدرکی موجود نبود و با تمام پیگیری ها انجام شده در نهایت هیچ کسی پاسخگوی بیمار نبود و ایشان ناچارا قادر به اخذ مبلغ بیمه تکمیلی نشده بودند و در نهایت متوجه شده بودند که با اسم ایشان چندین نسخه الکترونیکی صادر شده ولی هیچ کسی پاسخگوی ایشان نبود.

الگوریتم SVM هم بر روی داده های تمرین اعمال شد (شکل ۵).



شکل ۴-۵-اعمال الگوریتم SVM

۶-۵-ارزیابی نتایج

برای ارزیابی دقت شناسایی موارد خلافکار توسط روشهای مورد استفاده در این تحقیق، از معیارهای دقت، صحت، ماتریس پریشانی استفاده خواهیم کرد.

حداکثر مقدار Precision برابر با یک و یا ۱۰۰ درصد و حداقل مقدار آن صفر است شامل مواردی است که مدل ایجاد شده غلط پیش بینی کرده است که به آن False Positive می‌گوییم. هر چقدر نسبت پیش بینی‌های درست یا True Positive بیشتر باشد مقدار Precision کمتر خواهد شد. از رابطه ۱ می توان مقدار آن را محاسبه کرد.

رابطه ۱ محاسبه Precision

$$Precision = \frac{TP}{TP + FP}$$

حداکثر مقدار Recall برابر با یک و یا ۱۰۰ درصد و حداقل مقدار آن صفر است و شامل مواردی است که انتظار داشته ایم درست پیش بینی شود ولی درست پیش بینی نشده اند که به آن False Negative می‌گوییم وقتی نسبت پیش بینی‌های درست یا True Positive بیشتر باشد مقدار Recall کمتر خواهد شد. با استفاده از رابطه ۲ محاسبه می شود.

رابطه ۲ محاسبه Recall

$$Recall = \frac{TP}{TP + FN}$$

معیار accuracy یا همان صحت برابر است با تعداد مواردی که درست پیش بینی کرده ایم و آن را True Positive یا (TP) می‌نامیم تقسیم بر تعداد کل پیش بینی‌هایی که انجام شده‌است. با رابطه ۳ محاسبه می شود.

رابطه ۳ محاسبه Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

زمانی که می‌خواهید معیار ارزیابی الگوهای ایجاد شده، میانگینی از دو مورد قبلی یعنی همان Recall یا Precision باشد می‌توان از میانگین هارمونیک این دو معیار استفاده کرد که به آن معیار F-score می‌گویند. از رابطه ۴ محاسبه می شود.

رابطه ۴ محاسبه F-score

$$F1 = \frac{2}{\frac{1}{Recall} \cdot \frac{1}{Precision}} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

جدول ۱ نتایج بدست آمده از ارزیابی مدلهای ایجاد شده توسط الگوریتم درخت تصمیم و SVM را نشان می دهد.

جدول ۱–ارزیابی مدلهای استفاده در این تحقیق

الگوریتم استفاده شده	Accuracy	Precision	Recall	F-Score
جنگل تصادفی	0.98881	0.97458	1	0.98712
SVM	0.96642	0.93222	0.99099	0.9607

همانطور که مشاهده می کنیم از نظر تمام معیارهای مورد نظر ما، الگوریتم جنگل تصادفی کارایی بهتری را در شناسایی موارد خلافکار از خود نشان می دهد.

۷-۵-مقایسه عملکرد روشهای مورد استفاده با روشهای دیگر

در این قسمت ما الگوریتم های دیگر کشف الگو شامل (درخت تصمیم، naïve Bayes، KNN و رگرسیون لجستیک) را بر روی داده های موجود اعمال نمودیم. نتایج بدست آمده از ارزیابی مدل ایجاد شده توسط این الگوریتم ها در مقایسه با الگوریتم های به کار رفته در این پژوهش در جدول ۲ نشان داده شده است.

جدول ۲-مقایسه کارایی الگوریتمهای استفاده شده با روشهای دیگر

الگوریتم استفاده شده	Accuracy	Precision	Recall	F-Score
جنگل تصادفی	0.98881	0.97458	1	0.98712
SVM	0.96642	۰.۹۶۶۳۹	0.99099	0.9607
درخت تصمیم	۰.۹۶۳۳۲	۰.۹۳۲۲۲	۰.۹۵۶۸۹	۰.۹۶۲۳۴
KNN	۰.۷۱۲۶۸	۰.۵۷۱۴۸۸	۰.۷۳۳۴۰	۰.۶۳۸۴۹
Naïve Bayes	۰.۸۴۶۳۸	۰.۸۰۶۷۲	۰.۸۳۴۸۷	۰.۸۲۰۵۱
رگرسیون لجستیک	۰.۶۵۹۳۲	۰.۷۷۱۴۲	۰.۸۳۵۰۵	۰.۸۰۱۹۸

همانطور که در جدول ۲ مشاهده می شود بیشترین عملکرد در بین تمام الگوریتم های اعمال شده متعلق به جنگل تصادفی و کمترین عملکرد مربوط به الگوریتم KNN می باشد. در نهایت می توان با استفاده از دو الگوریتم رده بندی استفاده شده در این پژوهش و همچنین با استفاده از قوانین انجمنی بدست آمده اقدام به شناسایی موارد خطا یا در واقع خلافکارانی که با جعل نسخه ها یا مدارک پزشکی اقدام به گرفتن داروهای خاص از بیمارستانها به نرخ دولتی می کنند را، شناسایی نموده و در اسرع وقت اقدام به بلوکه کردن آن نسخه نمود.

۵- نتیجه گیری و کارهای آتی

شرکتهای بیمه یکی ازمهمترین شرکتهای حیاتی در هر کشور محسوب می شوند. تقریبا اغلب افراد در دنیا در یکی از سازمانها یا شرکتهای بیمه عضو هستند و ماهیانه هزینه ای را برای این عضویت پرداخت می کنند. این شرکتها و سازمانهای بیمه خدمات مختلفی برای اعضای خود ارائه می کنند که در قبال آن مبالغی از آنها ماهیانه یا سالیانه اخذ می کنند. یکی از مهمترین مشکلات که برای این شرکتها و سازمانهای بیمه رخ می دهد تقلب یا

کلاهبرداری به منظور اخذ مبلغ خسارت یا ضرر و زیان از این سازمانها می باشد. از آنجا که این مبالغ ضرر و زیان در برخی مواقع (از قبیل بیمه آتش سوزی، بیمه بدنه ماشین و...) بسیار زیاد است می تواند بار مالی زیادی برای این شرکتها و سازمانهای بیمه وارد نماید لذا شناسایی این تقلب ها و کلاهبرداری ها می تواند در جلوگیری از خسارات مالی تحمیل شده به آنها نقش مهمی داشته باشد. در این تحقیق از تکنیکهای داده کاوی شامل الگوریتم جنگل تصادفی، SVM برای شناسایی مدل و شناسایی آنومالی ها استفاده شده و از تکنیکهای کشف قوانین انجمنی برای کشف روابط بین متغیرهای مختلف استفاده شده است. ارزیابی تکنیکهای استفاده شده نشان می دهد که، کارایی بسیار خوبی در شناسایی موارد تقلب دارا می باشند.

تعداد نمونه های بدست آمده از سازمان بیمه تامین اجتماعی هر چقدر بیشتر باشد می توانیم روشها و مدلهای بدست آمده را بر روی آنها اعمال نموده و نسبت به افزایش دقت روشهای به کار رفته اعتماد بیشتری داشته باشیم. در آینده انتظار می رود با افزایش استفاده از نسخه های الکترونیکی یکی از نیازها و چالشهای این سازمان های بیمه، شناسایی زودهنگام جعل یا سوء استفاده از دفترچه های بیمه افراد مختلف برای خرید داروهای خاص یا پرداخت هزینه های پاراکلینیکی و ... باشد لذا شاید بتوان امید داشت که قوانینی برای در اختیار گذاشتن داده های واقعی بیشتری از طرف این سازمانها در اختیار پژوهشگران، اتخاذ شود تا راحتتر و با دقت بالاتر بتوانیم روشی جامع برای شناسایی تقلب ها در این سازمان اتخاذ نماییم.

منابع

1. Pappas, I.O. and A.G. Woodside, *Fuzzy-set Qualitative Comparative Analysis (fsQCA): Guidelines for research practice in Information Systems and marketing.* International Journal of Information Management, 2021. **58**: p. 102310.
2. Petropoulos, A., et al., Predicting bank insolvencies using machine learning techniques. International Journal of Forecasting, 2020. 36(3): p. 1092-1113.
3. Pourhabibi, T., et al., *Fraud detection: A systematic literature review of graph-based anomaly detection approaches.* Decision Support Systems, 2020. **133**: p. 113303.
4. Waring, J., C. Lindvall, and R. Umeton, *Automated machine learning: Review of the state-of-the-art and opportunities for healthcare.* Artificial Intelligence in Medicine, 2020. **104**: p. 101822.
5. Pramanik, M.I., et al., *Healthcare informatics and analytics in big data.* Expert Systems with Applications, 2020. **152**: p. 113388.
6. Richter, A.N. and T.M. Khoshgoftaar, *A review of statistical and machine learning methods for modeling cancer risk using structured clinical data.* Artificial intelligence in medicine, 2018. **90**: p. 1-14.
7. Kirlidog, M. and C. Asuk, *A fraud detection approach with data mining in health insurance.* Procedia-Social and Behavioral Sciences, 2012. **62**: p. 989-994.
8. Boodhun, N. and M. Jayabalan, *Risk prediction in life insurance industry using supervised learning algorithms.* Complex & Intelligent Systems, 2018. **4**(2): p. 145-154.
9. Finkelstein, A. and J. Poterba, *Testing for asymmetric information using "unused observables" in insurance markets: Evidence from the UK annuity market.* Journal of Risk and Insurance, 2014. **81**(4): p. 709-734.
10. Kang, S. and J. Song, *Feature selection for continuous aggregate response and its application to auto insurance data.* Expert Systems with Applications, 2018. **93**: p. 104-117.
11. Rawat, S., et al., *Application of machine learning and data visualization techniques for decision support in the insurance sector.* International Journal of Information Management Data Insights, 2021. **1**(2): p. 100012.
12. Yang, C., et al., *Estimation of Prevalence of Kidney Disease Treated With Dialysis in China: A Study of Insurance Claims Data.* American Journal of Kidney Diseases, 2021. **77**(6): p. 889-897. e1.

13. Palanisamy, V. and R. Thirunavukarasu, *Implications of big data analytics in developing healthcare frameworks–A review.* Journal of King Saud University-Computer and Information Sciences, 2019. **31**(4): p. 415-425.

Proposing a New Method for Predicting and Detecting Fraud of Customers and Policyholders in Insurance Companies in order to Reduce the Amount of Financial Losses to These Companies

Ali pasban asadabadi[1](Corresponding author), Zahra Ismaeili[2], Ali Ismaeili[3]

[1]Computer Eng Student, Electronic Branch, Islamic Azad University, tehran, Iran
Alipasban1370@gmail.com
[2]Master of Public Administration, Department of Organizational Behavior Management, Faculty of Management and Economics, Islamic Azad University, Science and Research Branch , Tehran , Iran
marjanoxygen@gmail.com
[3]PhD student in Information Technology, majoring in Network and Computing, Faculty of Engineering, Islamic Azad University, North Tehran Branch. Tehran Iran
ali_e_24@yahoo.com

Abstract

Unfortunately, it is observed that especially in Iran, all insurance organizations are at risk of bankruptcy, to respond to this, to various frauds and scams in car body insurance, building insurance, fire insurance and ... It can be argued that these scams really impose heavy costs on insurance companies. In this study, we addressed one of the biggest current problems in the insurance industry. One or two years have passed since the creation of electronic prescriptions in Iran, but we are facing a high volume of fraudulent use of other people's insurance booklets in order to buy certain drugs or use paraclinical services, etc., which also increases the financial pressures on Insured organizations. In this study, we used data mining techniques to provide a new method to detect cases of fraud in electronic versions in order to receive certain drugs with other people's insurance booklets. The results showed that the use of association rules can accurately detect the relationship between fraudulent behaviors and other characteristics of the fraudster, and the use of random forests compared to other methods and algorithms, can more accurately identify these erroneous cases or lead to abnormal.