کشف تقلب در بیمه سلامت بر اساس رویکرد داده کاوی

حسن عسکرزاده
دانشجوی دکتری دانشگاه خواجه نصیرالدین طوسی
محمد جعفر تارخ
استاد دانشگاه خواجه نصیرالدین طوسی

چکیده

هزینه بهداشت و درمان با توجه به جمعیت، اقتصاد، جامعه، و تغییرات قانون به سرعت در حال افزایش است. این افزایش در هزینه های بهداشت و درمان بر دولت و سیستم های بیمه سلامت خصوصی تاثیر می گذارد. رفتارهای متقلبانه ی ارائه دهندگان بهداشت و درمان و بیماران با تحمیل هزینه های غیر ضروری به مشکلی جدی برای سیستم های بیمه تبدیل شده است. بنابراین شرکت های بیمه روش هایی را برای تشخیص تقلب ایجاد می کنند که عمدتا برگرفته از تجارب خبرگان بوده و کمتر به روش های مبتنی بر تحلیل داده متکی است . صنعت بیمه و در راس آن بیمه سلامت با هزینه ای بالغ بر ۵۰ هزار میلیارد تومان یکی از کلیدی ترین هزینه های تحت مدیریت و نظارت دولت است که با تخمینی معادل ۳ تا ۱۰ درصد یعنی ۱۵ تا ۵۰ هزار میلیارد ریال تقلب مواجه است . با توجه به حجم اسناد و انبوه بیمه شدگان و ارائه دهندگان خدمت ، کشف تقلب به صورت سنتی غیر ممکن است .در این پژوهش با بررسیپژوهش های قبلی در ایران و سایر نقاط جهان و بر اساس ۱۵۰،۰۰۰،۰۰۰ رکورد اطلاعات بیمه سلامت بیمه شدگان استان تهران ضمن اجرای فرآیند فراخوانی , پاکسازی و اعمال داده ها در یک انباره داده ای و با بهره گیری ازالگوریتم های غیر نظارتی , مدلها و ابزارهای دادهکاوی , نسبت به تهیه لیست کوتاه جهت شناسایی و کشف تقلب در بیمه سلامت استان تهران در حوزه فعالیت پزشکان و ارجاع متقلبانه پزشک و داروخانه و پزشک و آزمایشگاه اقدام گردید. نتایج حاصل شامل ارایه لیست کوتاه شامل ۱۴۵ پزشک از ۷۱,۵۴۵ پزشک در ۳ خوشه که حدود ۳۸٪ هزینه ها بر اساس نسخه آنان ایجاد شده است گردید.

واژگان کلیدی: دادکاوی , کشف تقلب ، بیمه سلامت ، خوشه بندی

مقدمه

سهم قالب هزینه صنعت بیمه سلامت در ایران توسط دو نهاد اصلی سازمان بیمه سلامت ایران و سازمان تامین اجتماعیکه هر دو توسط دولت مدیریت میشوند پرداخت میشود. برآورد به دست آمده از قانون بودجه سال ۱۳۹۶ حاکی از ۱۴,۳۹۳ میلیارد تومان هزینه در حوزه وزارت بهداشت است که به طور مشخص ۸,۹۵۸ میلیارد تومان به صورت مستقیم توسط سازمان بیمه سلامت ایران هزینه خواهد شد. به این اعداد باید پرداختهای مستقیم مردم در قالب فرانشیز اضافه گردد. با توجه به اینکه سالانه ۳ ٪ تا ۱۰ ٪از هزینه های بیمه سلامت به صورت متقلبانه دریافت میشود(Li et al , 2008) و عدد مورد نظر به ۱۰ ٪ نزدیکتر است (Sparrow , 1998) این عدد برای سال ۹۶ به ۸۹۵ میلیارد تومان بالغ خواهد شد . با توجه به حجم اسناد قابل رسیدگی و کند بودن روش های مبتنی بر الگوهای ذهنی افراد خبره و همچنین کمبود منابع انسانی در صورتیکه بتوان بر اساس روشهای مبتنی بر تحلیل داده ها ، نسبت به کشف داده های تقلبی اقدام کرد حجم بیشتری از هزینه های غیر قابل پرداخت را در زمان کوتاهتری از سبد هزینه سلامت حذف میگردد.

بیان مسئله :

سارمانهای بیمه گر پس از دریافت اسناد هزینه از ارایه دهندگان خدمات سلامت اعم از مراکز درمانی ، پزشکان، داروخانه ها ، آزمایشگاهها فرآیند بررسی هزینه ها و تطبیق آنها با معیارها و جداول هزینه شده توسط آنرا را که به آن رسیدگی به اسناد میگویند را آغاز مینمایند . این رسیدگی هم اکنون به صورت دستی و توسط افراد خبره در سازمان صورت میگیرد . با توجه به انسانی بود فعالیت فوق, محدودیتهایی نظیر خطای انسانی , کمبود نیروی انسانی خبره , محدودیتهای زمانی فعالیت انسانی , عدم کیفیت یکسان در رسیدگی، احتمال وجود تعاملات انسانی ارزیاب و ارزیابی شونده و سایر موارد بر رسیدگی تاثیر گذار است. حجم زیاد پرونده ها نیز بر مشکل افزوده و احتمال کشف موفق تقلب های پیچیده را کاهش میدهد. استفاده از روش های تحلیل داده های بزرگ نظیر داده کاوی به ذینفعان کمک میکند تا بتوانند ضمن تعمیم و بهره برداری از الگوهای شناخته شده جهت بکارگیری الگوریتم های همراه با ناظر نسبت به کشف الگوهای ناشناخته از طریق به کار گیری الگوریتم ها و مدلهای نظارت نشده بپردازند.

اهمیت موضوع :

کاهش ۱۰ ٪ هزینه های سلامت از طریق حذف اسناد تقلبی میتواند منجر به افزایشی به همین میزان در کیفیت و کمیت خدمات سلامت به بیمه شدگان باشد. ایجاد یک انباره داده حاصل از فرآیند فراخوانی، پالایش و بارگذاری داده ضمن استنادپذیر کردن داده‌های موجود در پایگاه‌داده‌ای سازمانهای بیمه‌گر و ایجاد بستر داشبورد برای برپایی سامانه های هوش تجاری امکان تجزیه تحلیل و بهرمندی از روشهای داده کاوی برای کشف تقلب را نیز فراهم می‌کند .

ادبیات و پیشینه :

تعریف سند سازی، تقلب، و سوء استفاده از بیمه

تقلب و سوء استفاده ،که به موضوع بزرگی در راستای توسعه سیستم‌های اطلاعاتی تبدیل شده است، در حال مختل کردن صنایع زیادی است. صنایع بهداشت و درمان و مخابرات، مانند صنعت بانکداری، از تقلب و سوء استفاده مکرر رنج می برد. البته مردم زیادی تقلب را با سوء استفاده اشتباه می گیرند؛ این واژه ها نمی توانند با هم ترکیب شوند. تقلب به عنوان یک فریب

عمدی یا ارائه اطلاعات نادرست تعریف می‌شود که توسط شخصی که می‌داند این فریب یا ارائه نادرست اطلاعات ممکن است سود غیر مجازی برای او یا شخص دیگری داشته باشد انجام می گیرد (راهنمای تقلب بهداشت و درمان امریکا، ۱۹۹۱). به طور مختصر، تقلب گفته‌ای غلط است که عمدا برای رسیدن به چیزی غیر منصفانه و غیرقانونی، رواج داده شده است. درحالیکه سوء استفاده به عنوان رفتاری متناقض و نامناسب با هدفی غیر قانونی تعریف می شود بدون اینکه لزوما عواقب قانونی داشته باشد.

سند سازی توسط ارائه دهندگان خدمات بهداشتی و درمانی

هشتاد درصد هزینه بهداشت و درمان مربوط به تصمیم پزشکان درباره خدماتی است که بیماران نیاز دارند. بنابراین، تقلب و سوء استفاده رخ داده توسط پزشکان می تواند خیلی قابل توجه باشد(Wynia et al ,2000). البته دلایل و انگیزه هایی وجود دارد که چرا پزشکان قانون مربوط به تقلب و سوء استفاده را زیر پا می‌گذارند. دیدی که پزشکان از از فعالیت خود به عنوان کسب و کار دارند، میتواند نقشی حیاتی در ارتکاب به تقلب یا سوء استفاده ایفا کند. برای مثال، هزینه صدور صورتحساب میتواند انگیزه بزرگی برای پزشکانی باشد که خودشان را به عنوان یک فروشنده می بینند. پزشکان می توانند اقدامات غیر ضروری برای افزایش هزینه ها انجام دهند. اگرچه این روش ها بر سابقهپزشکی بیمار اثر میگذارد و آن را تحریف می کند و ممکن است منجر به درمان اشتباه در آینده شود (Price and Norris,2009). از طرف دیگر، پزشکان ممکن است در شرایط دشواری بین انتخاب تعهد حرفه ای در مقابل بیماران یا قوانین پوشش مندرج در قرارداد شان قرار گیرند. برای مثال، برخی پزشکان ممکن است در شرایط بیمار اغراق کنند یا درخواست آزمایشی را بکنند که نشان دهد این دارو یا درمان برای بیمار ضروری است، تا در بدست آوردن پوشش اضافه به آنها کمک کنند .(Wynia et al ,2000)

شاخص های بالقوه سند سازی، تقلب و سوء استفاده در بیمه

راه های بیشماری برای تقلب و سوء استفاده وجود دارد. همچنین ارتباطی قوی بین سند سازی، تقلب و سوء استفاده در بیمه وجود دارد. بیشتر دلایلی که یک صورتحساب در بیمه رد میشود، این است که شاخص های مشکوک دارد. در این شرایط، بیمه‌گر از ارائه کننده خدمات سلامت یا بیمه شده می خواهد تا اطلاعات ارائه شده را تایید کند. بنابراین، تعیین و طبقه بندی دقیق این پارامتر ها حیاتی است. انواعتقلب های شناخته شده در جدول ۱ آمده است . در ادامه به تشریح هر کدام از انواع تقلب خواهیم پرداخت

جدول ۱ انواع تقلب در بیمه سلامت

ردیف	انواع تقلب
۱	کدگذاری اشتباه خدمات درمانی
۲	صدور مجدد صورتحساب
۳	تجزیه یک فعالیت ترکیبی با کد واحد به فعالیتهای جزئی تر
۴	صورتحساب مواردیکه تحت پوشش نیستند
۵	ارایه خدمات غیر ضروری
۶	عدم تطبیق تشخیص و درمان
۷	ارایه خدمات بیش از ظرفیت
۸	ارجاع منفعت طلبانه

کدگذاری اشتباه فعالیت ها , می تواند سرنخ هایی از تقلب و سوء استفاده داشته باشد. کدگذاری فعالیت ها زمانی رخ می دهد که ارائه کنندگان خدمات بهداشتی و درمانی از کدی استفاده می کنند که گران تر از خدمات بهداشت و درمان، تست ها، یا آیتم هایی است که واقعا برای بیمار انجام شده است. برای مثال، کد ۹۹۲۱۱ برای یک مشکل پزشکی ساده و یک ویزیت کوتاه است که ۲۰ دلار هزینه دارد، در حالیکه کد ۹۹۲۱۵ نشان دهنده یک مشکل پیچیده و ویزیتی طولانی با هزینه ۱۴۰ دلار است. در نتیجه، چک کردن خطاهای صورتحساب مربوط به کدگذاری فعالیت ها برای کاهش هزینه بهداشت و درمان و جلوگیری از تقلب و سوء استفاده، حیاتی است (Mukherjee ,2012). از طرفی دیگر، بسیاری از پزشکان معتقدند که دقت در کدگذاری درست در صورت حسابیه اندازه ویزیت بیمار زمان می برد و آنرا بهانه ای برای عدم دقت و بروز اشتباه می‌دانند. در ایران از سال ۱۳۸۴ اقداماتی در خصوص یکسان سازی نرخ خدمات درمانی شکل گرفته که نتیجه آن تولد کتاب ارزش نسبی خدمات و مراقبت‌های سلامت است که بر اساس فرآیندی با همین هدف از کشور امریکا اقتباس شده است . هر چند در این کتاب هدف کدینگ واحد پیگیری نمی‌شود ولی از نتایج مشخص آن رویکرد یکسان سازی کدینک و کاهش این گونه از تقلبها میباشد.صدور مجدد صورتحساب، که به صدور دوباره صورتحساب برای یک فعالیت در یک زمان با تغییراتی کوچک گفته میشود، مانند تاریخ، هم می‌تواند یک اشتباه ساده باشد، هم می‌تواند یک سوء استفاده باشد. در هر صورت، ارزش بررسی مجدد و حذف را دارد.تجزیه یک فعالیت ترکیبی با کد واحد به فعالیتهای جزئی تر به چندین کد جزئی تر، روشی دیگر برای افزایش هزینه و بدست آوردن منفعت غیر مجاز است. درمان ها یا آزمایش هایی وجود دارند که شامل بیش از یک خدمت است. وقتی این خدمات با هم انجام شوند، تامین کننده خدمات بهداشتی و درمانی نیاز به استفاده از کد های مشخصی دارد که دو خدمت یا بیشتر را گروه بندی کند. اگر تامین کننده خدمات بهداشتی و درمانی از این کدهای صورتحساب مشخص، برای تمام خدمات اختصاص یافته استفاده نکند و به صورت مجزا آنها را صورتحساب کند، ممکن است پولی بیشتر از خدماتی که واقعاً انجام داده دریافت کند. برای مثال، تست کامل خون شامل آزمایش های زیادی مانند اندازه گیری آنزیم ها و مواد معدنی مختلف است. زمانی که این آزمایش ها جداگانه صورتحساب شود، نرخ پرداخت ممکن است دو برابر شود.ارایه صورتحساب برای مواردی که تحت پوشش بیمه نیست به جای موارد تحت پوشش نیز یکی از فعالیت های سوء استفاده گرانه و دلیلی برای سند سازی است که مکرر دیده می شود. زیرا تامین کنندگان خدمات بهداشتی و درمانی موظف هستند بهترین مراقبت ممکن را پیشنهاد بدهند، بعضی اوقات ممکن است به خاطر سلامت بیمارشان، مواردی که تحت پوشش نیستند را به جای موارد تحت پوشش صورتحساب کنند. پزشکان اغلب قوانین بازپرداخت را دستکاری می کنند تا به بیمارانشان کمک کنند تا برای خدمات ضروری در طرح درمان، پوشش لازم را بگیرند (Wynia et al ,2000).انجام خدماتی که برای رفاه بیمار ضروری نیست، به عنوان مواردی که از نظر پزشکی ضروری نیست در نظر گرفته می شود. بیمه‌گر پوشش را فقط برای تشخیص و درمان خدمات قانونی، منطقی و ضروری از نظر پزشکی، فراهم می کند. صورتحسابها یا صورتحسابهای بیمه که شامل خدمات غیر ضروری است ممکن است منجر به رد صورتحساب شود یا نیاز به تحقیق داشته باشد که بفهمیمآیا تقلب یا سوء استفاده است یا خیر (Robbins & Anderson , 2011).زمانی که یک طرح درمان که نیازمند شرایط پیش نیاز است برای بیماری به کار برده می شود که شرایط پیش نیاز را ندارد،یک نشانهقرمز می تواند رفتار متقلبانه یا سوء استفاده گرانه بالقوه را نشان دهد. گذشته از شرایط پیش نیاز، یک عدم تطابق بین تشخیص و طرح درمان می تواند نشانه یک رفتار مشکوک باشد. برای مثال، تشخیصی که نیاز به داروی خاص برای بیمار ندارد ممکن است نشان دهنده تقلب یا سوء استفاده بالقوه باشد.

نسبت برخورد غیر معمول با بیمار، پارامتر دیگری برای تخمین ریسک تقلب و سوء استفاده است. برای مثال، اگر پزشکی هر روز تعداد زیادی از بیماران را ببیند که بیشتر از میزانی است که او می توانسته بپذیرد، اثبات کننده ی مراقبت ضعیف او از بیمارانش یا ارتکاب به تقلب باشد. یک طرح درمان ناکافی که به پزشکی که بیمارانی بیشتر از حد توانش را می‌بیند اختصاص یافته است، بینشی نسبت به رفتار پزشک می دهد. علاوه بر این، بیمارستان هایی که تعداد پزشکانی که استخدام کرده اند را بیشتر از تعداد واقعی گزارش می دهند، تقلب کرده اند، زیرا ارائه اطلاعات نادرست نیز تقلب است.ارجاع منفعت طلبانه ، معرفی بیماران به پزشکی خاص یا ارایه دهنده خدمات بهداشتی و درمانی خاص است. برای مثال، اگر یک پزشک منفعتی شخصی از یک کلینیک داشته باشد، نمی تواند هیچ بیماری را به آن کلینیک ارجاع دهد. در بعضی از کشورها از جمله امریکا قانونی برای مقابله با این امر وجود دارد . در ایران اشتراک منافع پزشکان با داروخانه ها و آزمایشگاهها و بیمارستانها به تناسب قرارداد سازمانهای بیمه گر ممکن است با جرایمی همراه باشد.به صورت خلاصه، قرارداد مقابله با ارجاع منفعت طلبانه زمانی نقض می شود که ارائه دهنده خدمات بهداشتی و درمانی بیماران را به جایی که ارتباط مالی با آن دارد ارجاع دهد. این معرفی ها توسط قوانین یا قراردادهای ضد ارجاع منفعت طلبانه ممنوع شده اند و در صورت رخ دادن تقلب محسوب می شوند.

چالش در تشخیص سندسازی، تقلب و سوء استفاده در بیمه بهداشت و درمان

اگرچه تشخیص سندسازی و تقلب در بیمه حیاتی و به شدت مورد نیاز است، چالش ها و محدودیت های زیادی هستند که این کار را سخت می‌کنند. اول،تشخیص تقلب و سوء استفادهاز بیمه سلامت نیازمند کارشناسانی است که از دانش پزشکی در سطح بالایی برخوردار باشند (Yang & Hwang ,2006). بیشتر شرکت های بیمه از روش هایی استفاده می کنند که برای تشخیص فعالیت های متقلبانه یا سوء استفاده گرانه بالقوه ، نیازمند نیروی انسانی جهت ارزیابی مدارک است. این روش ها که مبتنی بر دانش افراد خبره است، نیاز به کارکنان خبره ای دارد که به اندازه کافی در دسترس نیستند. به علاوه، تکنیک های تشخیص دستی تقلب، به تلاش، زمان و تخصص انسانی زیادی نیاز دارد که منجر به تاخیر در اثبات یا رد صورتحساب می شود. علاوه بر این، آزمایشات و تشخیص های دستی بسیار هزینه بر هستند.با استفاده از تکنیک های اتوماتیک هر مورد با قوانین ساده ای که برای تست استفاده می شوند، کشف می شود. اگرچه، تشخیص تقلب و سوء استفاده مستلزم بررسی متغیرها و ابهامات زیادی است. این ابهامات به فناوری اطلاعات دقیق و جامعی نیاز دارد تا بتواند صحت صورتحساب را آزمون کند (Shin et al , 2012). با اینکه صورتحسابهای پزشکی و مستنداتی که الکترونیکی ارائه شده اند کشفتقلب را ساده تر می کنند، اما چالش های دیگری نیز وجود دارد. برای مثال، ارائه دهندگان خدمات بهداشتی و درمانی و بیمارستان ها انتظار دارند شرکت های بیمه به صورتحسابهای ارائه شده از سوی آنها پاسخی سریع بدهند. حقیقت این است که سرعت عمل در پردازش صورتحساب در شرکت های بیمه احتمال اشتباه را بالا می‌برد، و باعث می شود برخی صورتحسابهایمتقلبانه کشف نشوند. چالش دیگر تشخیص تقلب این است که در روش ها و نظارت های کنونی، داده‌هایی که نیازمند تحلیل هستند ، می توانند نسبت به هر تغییری حساس باشند. این نوسانات و بی ثباتی در سیستم بیمه سلامت، مانع از بررسی صورتحسابهای بیمه میشود. از این رو، کشف رفتارهای مشکوک در بهداشت و درمان نیازمند تکنیک های انطباقی است (Li et al , 2008).پس از ارایه ادبیات موضوع تحقیق ،مفاهیم فنی مرتبط با این پژوهش به صورت اجمالی مرور می گردد که شامل موارد زیر است:

یادگیری ماشین :

به عنوان یکی از شاخه‌های وسیع و پرکاربرد هوش مصنوعی، یادگیری ماشین(Machine learning)به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌هایی می‌پردازد که بر اساس آنها رایانه‌ها و سامانه‌ها توانایی تعلّم و یادگیری پیدا می‌کنند.یادگیری ماشین به دو نوع با نظارت یا بدون نظارت تقسیم بندی میشود .

یادگیری با نظارت : به نوعی از یادگیری توسط ماشین (رایانه) گفته میشود که الگوهایی که ماشین باید به استناد آن داده ها و روابط بین آنان را بیاموزد وجود دارد و ماشین سعی میکند برای کشف رابطه های پنهان در داده ها , از نمونه های موجود که به آن داده های برچسب دار میگویند بهره برداری نماید.

یادگیری بی نظارت (بدون نظارت، در مقابل یادگیری بانظارت): یکی از انواع یادگیری در یادگیری ماشینی است. اگر یادگیری بر روی داده‌های بدون برچسب و برای یافتن الگوهای پنهان در این داده‌ها انجام شود، یادگیری، بدون نظارت خواهد بود.

از انواع یادگیری بدون نظارت می‌توان به خوشه‌بندی، مدل پنهان مارکوف و برخی شبکه‌های عصبی مصنوعی اشاره کرد.

خوشه بندی یا آنالیز خوشه (Clustering) :

در آمار و یادگیری ماشینی، یکی از شاخه های یادگیری بی‌نظارت می‌باشد و فرآیندی است که در طی آن، نمونه‌ها به دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند تقسیم می‌شوند که به این دسته ها خوشه گفته میشود. بنابراین خوشه مجموعه ای از اشیاء می‌باشد که در آن اشیاء با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های دیگر غیر مشابه می‌باشند.مسایل خوشه‌بندی عموما به دو شکل مطرح شود: (۱) یک ماتریس بی‌شباهتی داده می‌شود یا (۲) یک ماتریس که هر سطر آن یک شیء را توصیف می‌کند. خروجی الگوریتم می‌تواند به دو صورت باشد: (۱) گروه‌بندی اشیا به مجموعه‌های مجزا یا (۲) خوشه‌بندی سلسله مراتبی که یک درخت برای تقسیم‌بندی اشیا پیدا می‌کند. الگوریتم‌های نوع اول سریعتر هستند. از الگوریتم‌های مشهور برای خوشه‌بندی می‌توان به k-means اشاره کرد.

k-means clustering :

روش میانگین k در عین سادگی یک روش بسیار کاربردی و پایه چند روش دیگر مثل خوشه بندی فازی و Segment-wise distributional clustering Algorithmاست. روش کار به این صورت است که ابتدا به تعداد دلخواه نقاط به عنوان مرکز خوشه در نظر گرفته می شود. سپس با بررسی هر داده ، آن را به نزدیک ترین مرکز خوشه نسبت می‌دهیم. پس از اتمام این کار با گرفتن میانگین در هر خوشه می توانیم مراکز خوشه و به دنبال آن خوشه های جدید ایجاد کنیم. (با تکرار مراحل قبل) از جمله مشکلات این روش این است که بهینگی آن وابسته به انتخاب اولیه مراکز بوده و بنابراین بهینه نیست. مشکلات دیگر آن تعیین تعداد خوشه ها و صفر شدن خوشه ها می باشد.

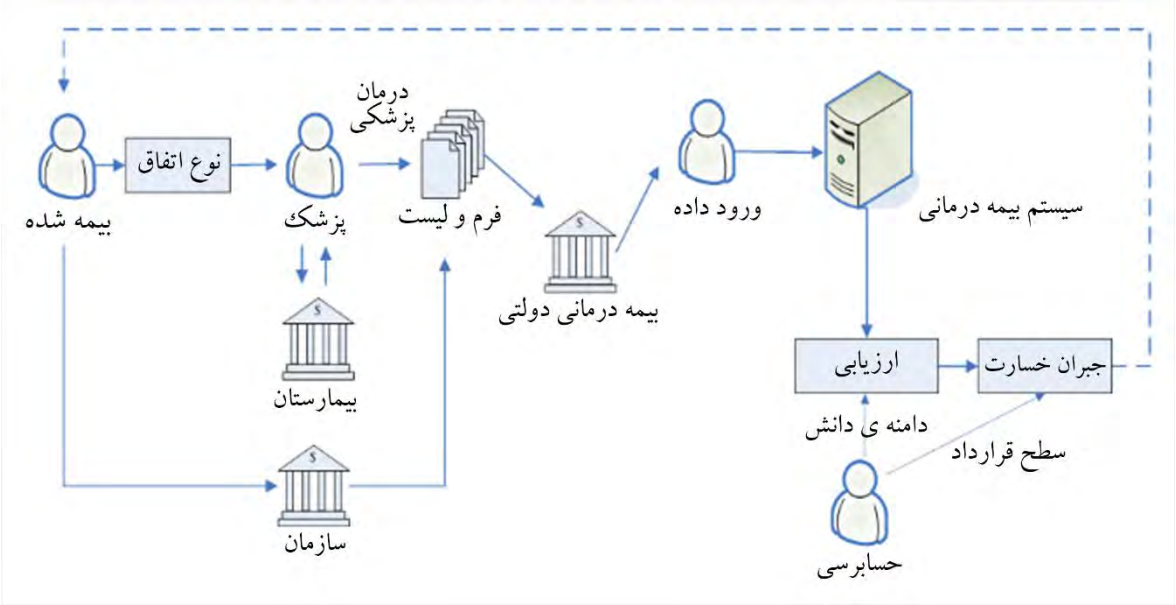
نرمالسازی گوسی یا Z-Score :

برای انجام عمل خوشه بندی روی بردارهایی که اعداد هر بعد آن در بازه های متفاوتی قرار دارند عمل نرمالسازی صورت میگیرد . روشهای مختلفی برای نرمالسازی وجود دارد که یکی از پر کاربردترین آنها روش گوسی یا Z-Score است . در این توزیع داده های اصلی در فاصله ۰ و ۱ نگاشت میشوند.

مروری بر پژوهش های پیشین با استفاده از خوشه بندی

سیستم های بیمه سلامت

انجمن بیمه سلامت آمریکا، بیمه سلامت را به عنوان پوششی علیه ریسک هزینه های درمانی به علت بیماری یا آسیب دیدگی تعریف می کند. این پوشش می تواند توسط بعضی سازمان های مرکزی، برای مثال شرکت های خصوصی یا دولتی، ارائه شود.منبع این پوشش در بسیاری از کشور ها صرف نظر از سیستم های بهداشت و درمان شان، متفاوت است. بررسی سالیانه انجام شده توسط صندوق مشترک المنافع، سیستم های بهداشت و درمان استرالیا، نیوزیلند، بریتانیا، آلمان، کانادا و ایالات متحده را مقایسه می کند. این بررسی تاکید می کند که ایالات متحده تنها کشور بدون پوشش بیمه سلامت سراسری است. اداره آمار ایالات متحده بیان می کند که ۳۱ درصد از آمریکایی ها طرح بیمه سلامت عمومی دارند، در حالی که ۵۵ درصد از آنها پوشش خود را از طریق کارفرمایانشان می گیرد. اگرچه، تحت پوشش بودن تضمین نمی کند که شخص بیمه شده هیچ هزینه پزشکی پرداخت نکند. میزانی که بیمه شده باید بپردازد قبل از اینکه بیمه گر برای یک ویزیت یا خدمت خاص بپردازد، پرداخت مشترک نامیده می شود. جدای از پرداخت مشترک، ممکن است خدماتی باشد که بیمه گر بر اساس حق بیمه ای که بیمه شونده می خرد، بازپرداخت می کند. مانند خدماتی که به عنوان بیمه تکمیلی شناخته می شوند که در آن درصد بیشتری از هزینه ها در قبال دریافت حق بیمه بیشتر پرداخت میشود. روند گردش اسناد در سیستم بیمه در شکل ۱ نشانداده شده است.

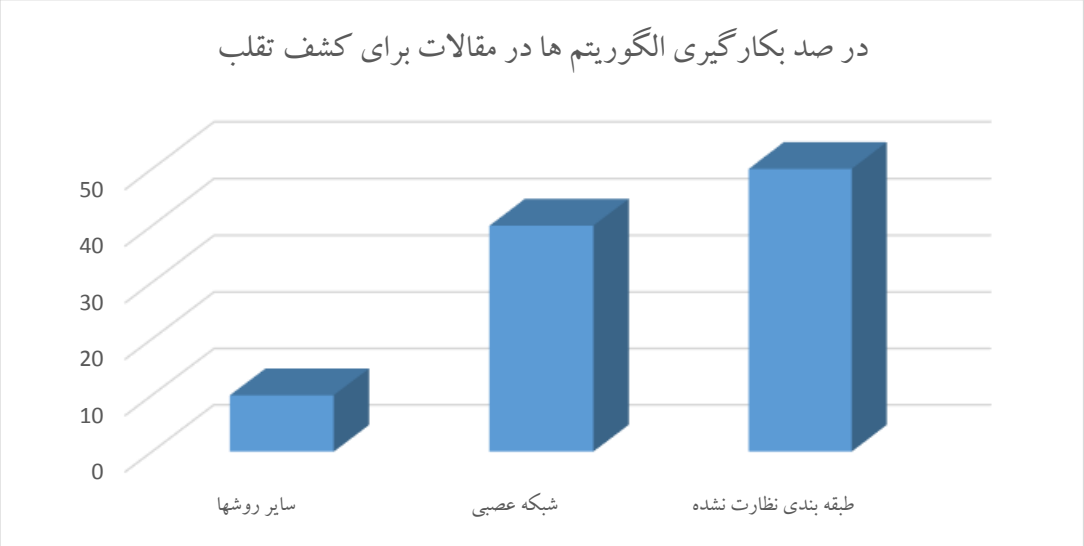


شکل۱ نمودار کلی گردش اطلاعات در سامانه های بیمه سلامت

تقریبا در هر سیستم بیمه سلامت، بیماران با پرداخت حق بیمه، پوشش سلامت می خرند. و هنگام مراجعه به ارایه دهندگان خدمات بهداشتی و درمانی ، پرداخت مشترکشان یا همان فرانشیز را انجام می‌دهند و خدمات دریافت می کنند. ارائه دهندگان، خدماتی را که به بیمار ارائه داده اند ثبت کرده و برای شرکت بیمه می‌فرستند. شرکت‌های بیمه فرم‌های صورتحساب را تحلیل می‌کنند و در خصوص مبلغی که باید به ارائه کنندگان بپردازند تصمیم می‌گیرند. این مبلغ به موارد عدم پوشش بیمه ای، الزامات پزشکی خدمات، و دقت فرم صورتحساب بستگی دارد. شرکت های بیمه دستورالعمل هایی به مراکز درمانی ارسال می کنند که اعلام می‌کند کدامیک از خدمات پزشکی تحت پوشش بوده و نحوه پرداخت و میزان تعیین شده که بیمار باید بپردازد را توضیح می دهد.

تشخیص سند سازی، تقلب و سوء استفاده در بیمه سلامت

یکی از بزرگترین چالش های پیش روی شرکت های بیمه این است که فرم های صورتحساب نیازمند تحلیل هستند و باید در زمان محدودی تصمیم بگیرند کدام موارد باید بازپرداخت شوند. متاسفانه، تمام فرم های صورتحساب شامل اطلاعات صحیح نیستند، و عدم صحت فرم های صورتحساب هزینه بهداشت و درمان را افزایش می دهد. این اشتباهات میتواند خطاهای سهوی باشد، یا یک روش عمدی برای فریب دادن شرکت های بیمه. بنابراین، بسیاری از شرکت های بیمه به یک سیستم غربالگری بدون دخالت انسان برای بررسی فرم های صورتحساب نیاز دارند. این سیستم می تواند تصمیم بگیرد کدام صورتحسابها باید دقیق تر بررسی شوند. این سیستم های تشخیص اولیه برای شکار ناهنجاری ها و بالا بردن پرچم قرمز با استفاده از روش های جدید مانند داده کاوی و روش های آماری معمولی طراحی شده اند نمودار ۱ درصد بکارگیری الگوریتم ها جهت کشف تقلب را نشان میدهد.



نمودار ۱ : : درصد روش های استفاده شده برای تشخیص تقلب در بهداشت و درمان

بازبینی ادبیات نشان می دهد که شبکه های عصبی و الگوریتم های طبقه بندی نظارت نشده (ماشین های پشتیبانی بردار، درخت های تصمیم، الگوریتم K-نزدیک ترین) برای مشکلات تشخیص تقلب در بهداشت و درمان بسیار استفاده می شوند.

تحلیل خوشه نیز به اندازه سایر ابزار های داده کاوی موثر است. پنگ و همکاران (Peng et al ,2006)دو روش خوشه ای را پیشنهاد می دهند. انگیزه اولیه پشت این تحقیق، طراحی سیستمی است که بتواند تقلب را در پایگاه داده های بزرگ کشف کند. آنها خوشه ها را با استفاده از پنجاه و سه ویژگی، مانند میزان صورتحساب، خدمات دریافت شده، و جمعیت شناسی بیمار، تولید میکنند. لیو و همکاران (Li et al , 2008) یک سیستم را با به کار گیری رگرسیون منطقی درختهای طبقه بندی و یک شبکه عصبی برای کشف تقلب و سوء استفاده در خدمات دیابتی، بهبود بخشیدند. داده های برچسب دار شامل ۹ متغیر مرتبط با هزینه، مانند هزینه تشخیص و دارو در روز می شود. هر سه الگوریتم با دقت بالایی قادر به تشخیص بیمارستان های متقلب و سوء استفاده گر هستند.

هی و همکاران (He et al , 1998)یک الگوریتم K-نزدیکترین و یک الگوریتم ژنتیک را با استفاده از بیست و هشت ویژگی برای ارائه کنندگان خوشه، ترکیب کردند. وزن ها در الگوریتم K-نزدیکترین برای تعیین نزدیکترین شیوه های همسایگی با استفاده از الگوریتم های ژنتیک به صورت بهینه ای تنظیم شده اند. قانون بیزین و قانون اکثریت , پزشکان را با مشخصات کم خطر و پر خطر تعیین می کنند.

یوزی و گدی(Uzi and Gadi , 1999) از یک الگوریتم خوشه بندی با یک تابع فاصله و توزیع های احتمال برای تعیین الگوری عمومی داده استفاده می کنند. انحراف از الگوی عمومی اگر بزرگتر از آستانه از پیش تعریف شده باشند نشان دهنده فعالیت های متقلبانه است.

ترنتن و همکاران(Trenten et al , 2013)مدلی برای کشف تقلب در حوزه کشف صورتحساب صوریمراکز درمانی یا بیماران ارایه نمودند که بر اساس هر رکورد خدماتی , صورتحسابها طبقه بندی و سپس در هر طبقه با خوشه بندی موارد غیر متعارف را استخراج و بر اساس نمونه های تخلف کشف شده و درصد تعلق آنها به هر خوشه , تشخیص تقلب با درصد تعیین شده داده میشد .آنها معیارهایی شامل مبلغ صورتحساب , کد خدمت ارایه شده , تاریخ ارایه خدمت , داروی تجویز شده , کد بیمه شده , کد مرکز بهداشت و درمان را برای تحلیل به عنوان ورودی دریافت و آنهارا در هفت طبقه گروه بندی کرده و در هر گروه آنالیز متفاوتی صورت میدهند .جودکی وهمکاران(Joudaki et al , 2011) به استناد اطلاعات اسناد پزشکی سازمان تامین اجتماعی استان لرستان و بر اساس اطلاعات نسخ دارویی طی ۵ گام نسبت با شناسایی پزشکان متقلب اقدام کرده اند . ایشان برای سنجش میزان توانمندی ویزیت پزشکان با توجه به عدم وجود معیار واحد . ۴۰ پزشکی که در هر دو بحش خصوصی و دولتی کار میکردند را بررسی و میزان کار در هر دو را معیار قرارداده .یک دیتاست از پزشکان شامل ۱۶۴ پزشک عمومی و 474897نسخه دارویی تهیه و رکوردهایی که داده های ناشناس زیادی داشتند از دیتا ست حذف شده اند. در روش آنان روشهای آماری برای پر کردن داده های مفقود استفاده نشده است .برای شناسایی رفتار تقلبی پزشکان ،۱۵ مصاحبه با افراد مختلف صورت گرفته که ۸ نفر آنها ارزیاب بیمه ، ۵ نفر مدیرملی و استانی و ۲ نفر پزشکبوده اند و راه های تقلب پزشکان را بررسی کرده اند. نمونه استنتاج منطقی ذکر شده این است که پزشک برگه سفید از دفترچه بیمار بدون اطلاع او برداشته و دارویی روی آن مینویسد و با تبانی با داروخانه هزینه نسخه تقلب دریافت شده را با داروخانه دریافت میکنند. چون ارزیابهای نسخه، به نسخ بالای ۴ دارو حساس هستند ، آنها ۳ یا کمتر دارو در یک نسخه تقلبی قرار میدهند.13 شاخص به دست آمده از تحلیل منطقی در بحث تخلف عبارتست از: درصد بیمارانی که بیش از یکبار در ماه ویزیت شده اند، میانگین اقلام دارو در یک نسخه ، میانگین هزینه نسخه دارویی پزشک (مشترک با تقلب) ،نسبت ۵ گرانترین نسخه آنتی بیوتیک به نسخ همه پزشکان ، نسبت تعداد نسخ تزریقی به تعداد نسخ همه نسخ پزشکان،نسبت هزینه نسخ تزریقی به جمع هزینه نسخ پزشکان ، نسبت تعداد نسخ حاوی آنتی بیوتیکبه کل نسخ پزشکان ، نسبت تعداد نسخ حاوی آنتی بیوتیک تزریقی به کل نسخ پزشکان ، نسبت تعداد نسخ حاوی کورتن تزریقی به کل نسخ پزشکان و در موضوع تقلب شاخص ها عبارت بودند از: درصد بیماران تکراری ،درصد بیماران – داروخانه های تکراری ،درصد بیماران – داروخانه های تکراری در ماه ،میانگین هزینه نسخه دارویی پزشک (مشترک با تخلف) ، نسبت ارجاع به داروخانه های گران قیمت سپس برای هرکدام از گروهها هم میانگین و هم انحراف معیار محاسبه شده است. در آخرین گام توسط روش خوشه بندی۹۲٪ صورتحسابها که مربوط به ۱۱ ماه است جدا شدند ، برای هر پزشک مقادیر نشانه ها یا شاخص ها را محاسبه واستفاده از Z Score نرمال سازی شد سپس بر اساس hierarchical clustering methodعمل کلاسترینگ انجام و پزشکان به دو گروه عادی و مظنون و در دو ویژگی تخلف و تقلب تقسیم بندی شدند. بر اساس معیار فاصله Euclidian distance measures ، تعداد بهینه کلاسترها با استفاده از شاخص اعتباریmaximum value of the silhouette coefficientمحاسبه گردید . در نتیجه ۱۳ شناسه یا نشانگر شناسایی شد که ۲ نشانگر بر اساس هزینه هاست ،۴ نشانگر بر اساس تعدد و الگوهای ویزیت پزشکان است .۷ نشانگر بر اساس الگوی نسخه تعیین شده است . تحقیق فوق نشان میدهد ۴ نشانگر معرف تقلب و ۸ نشانگر معرف تخلف هستند کهنشانگر متوسط هزینه دارو در نسخه در هر دو مشترک است و اعداد بزرگتر در هر نشانگر ، احتمال بالاتری از تقلب و تخلف را نشان میدهد

اهداف و فرضیه های پژوهش :

هدف از پژوهش ارایه لیستی کوتاه شده از مجموعه ۷۱,۵۴۵ پزشکی است که در شبکه درمان استان تهران بالقوه فعال هست و کاهش آن به مجموعه ای با حداکثر ۱ ٪ جهت بررسی سریعتر و دقیقتر ارزیابان و ناظران سازمان بیمه سلامت است به گونه ای که در آن با احتمال بیشتری کشف تقلب میسر گردد . بدیهی است خروجی ارزیابان ورودی لازم جهت به کارگیری الگوریتم های با نظارت را در پژوهش های بعدی فراهم خواهد کرد.

فرضهای پژوهش عمدتا بر صحت داده های تاریخ دار و صحت کدهای دارویی به کار رفته استوار است. با توجه به انجام فرآیند پالایش داده ها ، این فرض نادقیق نیست . زیرا در موارد زیادی خطا در ورود اطلاعات تاریخ و کد دارو باعث عدم رعایت الگوهای از پیش تعریف شده مانند قید ماه در فاصله ۱ تا ۱۲ و روز از ۱ تا ۳۱ همچنین تغییر کدینگ دارو به داروهایی که اطلاعات آن در پایگاه داده ای دارو یافت نخواهد شد می‌گردد.

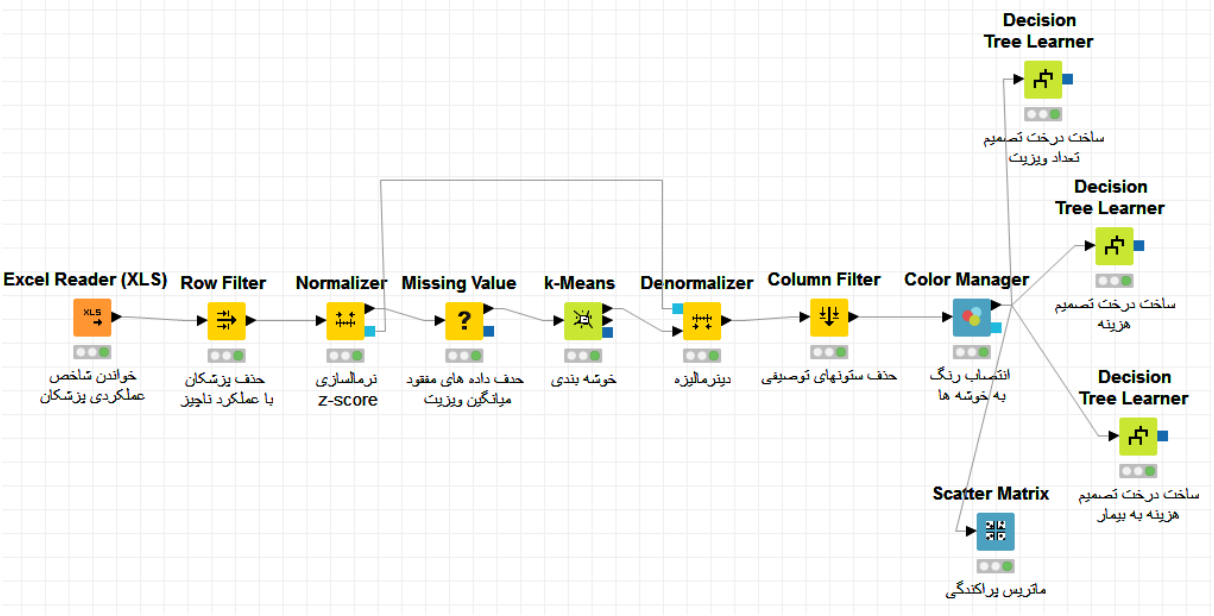
روش تحقیق

تحقیق فوق بر اساس محاسبه شاخص های خبرگانی روی اطلاعات نسخ کلیه پزشکانی است که حداقل یک خدمت به بیمه شده گانی که در سال ۹۵ مراجعه نموده اند ارایه کرده اند. در پژوهش حاضر هدف تهیه لیست پزشکان با بیشترین احتمال تقلب است در فرآیند این تحقیق ابتدا داده های نسخ پزشکان استان تهران از پایگاه داده اوراکل متعلق به سازمان بیمه سلامت استخراج شده است .این اطلاعات شامل فیلدهای اطلاعاتی شماره نظام پزشکی پزشک , کد بیمه شده , تاریخ ویزیت , تاریخ دریافت نسخه , کد داروخانه ، ریز اطلاعات نسخه شامل کد دارو ، تعداد دارو و حاوی ۱۵۰,۰۰۰,۰۰۰ رکورد اطلاعات ریز نسخ , ۷۵,۴۵۴ رکورد اطلاعات پزشک و نزدیک به ۳,۶ میلیون رکورد بیمه شده ای است که در سال ۹۵ با سازمان بیمه سلامت در استان تهران تعامل داشته اند . سپس فرآیند پالایش داده ها شامل الگوهای غلط تاریخی , الگوهای غلط تقدم و تاخر تاریخ ویزیت و نسخه , کدینگ اشتباه داروها , کدهای نظام پزشکی و بیمه شده غیر معتبر، کدینگ داروخانه های غیر معتبر در محیط کلیک ویو (QlikView) صورت گرفته و داده های مربوط به سال ۱۳۹۵ استان تهران فیلتر گردید. باتوجه به اینکه مدل کشف تقلب در داده های بزرگ ، بخصوص در مواردیکه داده های نظارتی وجود ندارد باید مشتمل بر دو بخش تحلیلی و پردازشی باشد. بخش تحلیلی عموما مبتنی بر اطلاعات ادراکی خبرگان است . خبرگان شاخص هایی را برای غیر متعارف بودن به صورت تجربی و ادراکی تعیین مینمایند که میتوان آنرا از پایگاه داده واکشی و دیتا ستی از موارد غیر متعارف به دست آورد.در این پژوهش نقطه نظر خبرگان در قالب ۹ شاخص شناسایی گردیده است.

عموما کشف تقلب به صورت سنتی معمولاز دو روش صورت میگرد . ۱- بررسیاتفاقی در اسناد و محاسبه شاخص های تجربی یا ادراکی ۲ – گزارشات مردمی که در نهایت اگر تکرار شونده باشد خود تبدیل به یک الگوی تجربی میگردد.

با توجه به حجم اطلاعات تلفیق دو روش نیز کارآمد است . به اینصورت که به جای تحلیل صدها میلیون رکورد اطلاعات ، میتوان از رویکردهای تحلیلی استفاده کرده و عملیات کاهش داده ها را بدون از دست دادن داده های کلیدی صورت داد. به عنوان نمونه به جای پردازش ۱۵۰ میلیون رکورد اطلاعات ارجاعات پزشکی که در آن پزشک , بیمار , مرکز درمانی , اقلام نسخ و آزمایش قراردارد میتوان با پردازش اولیه بر اساس موجودیتهای اصلی مثلا بیمه شدگان استان تهران ، رکوردها را به ۳,۵

میلیون رکورد تقلیل داد یا با تمرکز بر موجودیت پزشک و داروخانه تعداد رکوردها را به ۵۰۰,۰۰۰ تقلیل داد . تمرکز بر پزشکان , رکوردها را به ۸۰,۰۰۰ رکورد تقلیل داده و در صورت تمرکز بر داروهای مصرفی رکوردها به کمتر از ۱۰,۰۰۰ رکورد کاهش یافته و در صورت تمرکز بر داروخانه ها نیز این دیتاست به حدود ۲۰۰۰ رکورد خواهد رسید . بدیهی است هر دیتا ست میتواند حاوی بخشی از اطلاعات عملیاتی باشد که بر اساس نوع تقلب محتمل ، از تجمع یا گروه بندی سایر موجودیتهای اصلی محاسبه میگردد. در این روش عملیات پیش پردازش داده ها نسبتا زمانبر است اماعملیات محاسباتی خوشه بندی که هزینه محاسباتی و حافظه ای زیادی دارد بسیار سریعتر انجام خواهد شد .ابتدا شاخص های عملکردی پزشکان استان تهران در سال ۹۵ به تفکیک هر پزشک شاملجمع هزینه نسخ تجویز شده توسط هر پزشک ،تعداد نسخ تجویز شده توسط هر پزشک،تعداد عناوین دارویی تجویز شده، تعداد ویزیت در سال، تعداد بیمار در سال (اختلاف این فیلد با فیلد تعداد ویزیت در سال نشاندهنده این است که چند بیمار بیش از یکبار در یکسال توسط یک پزشک ویزیت شده باشد.) ، روزهایی که پزشک در سال فعال بوده است، میانگین ویزیت هر پزشک در روز، سرانه هزینه دارو و آزمایش به ازای هر ویزیت، سرانه هزینه دارو و آزمایش تجویز شده به ازای هر بیمار از پایگاه داده استخراج گردید سپس بر اساس مدل نشانداده شده در شکل ۳ فرآیند خوشه بندی با الگوریتم k_means و نرمالسازی z ، پزشکان به ۵ خوشه عملکردی تقسیم شدند تا بتوان بر اساس غیر متعارف ترین رفتار،نسبت به شناسایی و ارزیابی پزشکان اقدام نمود.نرم افزار مورد استفاده جهت مدلسازی ، نرم افزار متن باز knime میباشد . مزیت استفاده از این نرم افزار در مقابل spss modeler متعلق به IBM یا RapidMiner متن باز بودن و عدم محدودیت در حجم داده های مورد پردازش است . ضمن اینکه برای تغییر در یک المان در مدل نیازی به محاسبه بسیار زمانبر همه مدل نیست . در واقع ویژگی منحصر به فرد Knime نگهداری داده های پردازش شده از هر گره در مدل است که تا وقتی مبنای ورودی ها و تنظیمات آن تغییری نکند ، دوباره محاسبه نخواهد شد . این ویژگی بخصوص در کار با پردازش حجم زیاد داده ها مزیت بزرگی در سرعت بخشیدن به تغییر مدل و مشاهده نتایج به دست میدهد.برای انجام عمل خوشه بندی و برای تسهیل در امر پردازش داده ها ، ابتدا گزارش عملکردی پزشکان به تفکیک ۷۱,۵۴۵ پزشک از مجموع ۱۵۰,۰۰۰,۰۰۰ رکورد اطلاعات سال ۹۵ استان تهران در محیط پایگاه داده ای اوراکل بازیابی و در محیط qlikview به عنوان انباره داده ای هوش تجاری مورد پیش پردازش قرار گرفت. پس از اعمال فرآیند پاکسازی داده های پرت با الگوی حفظ داده های مشکوک به تقلب (حذف داده های پرت کم ارزش) ، خروجی جهت پردازش و مدلسازی به فرمت فایل اکسل ذخیره گردید.فایل اکسل به دست آمده به عنوان فایل ورودی مدل جهت پردازش به نرم افزار knime سپرده شد. سپس فرآیند خوشه بندی در نرم افزار فوق پیاده سازی گردید . شکل ۲ نمونه مدل ایجاد شده در محیط نرم افزار Knime میباشد.



شکل۲ : مدل خوشه بندی K_MEANS با =K۵ و نرمالسازی Z

توضیح مدل:

در گام اول اطلاعات عملکردی پزشکان که در فایل اکسل ذخیره گردیده بود، خوانده و پزشکانی که عملکردی ریالی بسیارکمی داشته اند حذف میشوند . سپس اطلاعات ۹ گانه ذکر شده با توزیع گوسی (Z-Score) نرمالسازی شدند . سپس برای جلوگیری از بروز خطا در عمل خوشه بندی مقادیر مفقود از میانگین ویزیت حذف و با الگوریتم k_means با تعداد 5 خوشه، عمل خوشه بندی صورت گرفت. سپس داده ها مجددا به حالت اصلی و غیر نرمالتبدیل شدند تا نتایج به صورت اعداد در بازه های واقعی نمایش داده شوند. سپس برای ایجاد وضوح بهتر به هر خوشه رنگی اختصاص داده شد.پس از آن با حذف ستونهای توصیفی ، به ۳ روش متفاوت شامل شاخص هزینه به بیمار ، هزینه کل و تعداد ویزیت ، درخت تصمیم ساخته شد تا از بین آنان بهترین و ساده ترین درخت برای پیاده سازی نرم افزار انتخاب گردد. همچنین خروجی نمایش پراکندگی خوشه ها نیز در این مدل تهیه شده است.

در این پژوهش بر اساس نظر خبرگانویژگی ۹ گانه به ازای همه ۷۱،۵۴۵ پزشکی که در سال ۹۵ حداقل یک رکورد اطلاعاتی داشتند محاسبه گردید .

یافته ها

یافته هایحاصل از این پژوهش به صورت ۳ خوشه غیر نرمال و ۲ خوشه نرمال براساس آنچه در جداول ۳ و ۴ ارایه گردیده به شرح ذیل تحلیل میگردد.

جدول ۲ : نتایج عددی خوشه بندی در یک نگاه کلی

خوشه	تعداد پزشک	تعداد نسخه	تعداد بیمار	جمع نسخه	روزهای فعال	سرانه هزینه پزشک	سرانه هزینه بیمار	سرانه هزینه روزانه
۰	6	5,852	2,683	179,584,226,414	819	29,930,704,402	66,934,113	219,272,560
۱	92	63,142	32,273	600,446,874,004	12,635	6,526,596,457	18,605,239	47,522,507
۲	4,993	2,101,144	1,708,146	878,651,575,501	484,789	175,976,682	514,389	1,812,441
۳	137	914	599	38,113,966,152	737	278,204,132	63,629,326	51,715,015
۴	66,317	967,801	834,281	435,333,084,962	631,764	6,564,427	521,806	689,075
جمع	71,545	3,138,853	2,577,982	2,132,129,727,033	1,130,744	29,801,240	827,054	1,885,599

جدول ۳ نتایج درصدی سهم هر بخش از خوشه بندی در یک دید کلی

خوشه	نظام پزشکی	تعداد بیمار	جمع نسخه	روزهای فعال
0	0/01%	0/10%	8/42%	0/07%
1	0/13%	1/25%	28/16%	1/12%
2	6/98%	66/26%	41/21%	42/87%
3	0/19%	0/02%	1/79%	0/07%
4	92/69%	32/36%	20/42%	55/87%
جمع	100/00%	100/00%	100/00%	100/00%

تحلیل و استنتاج از نتایج

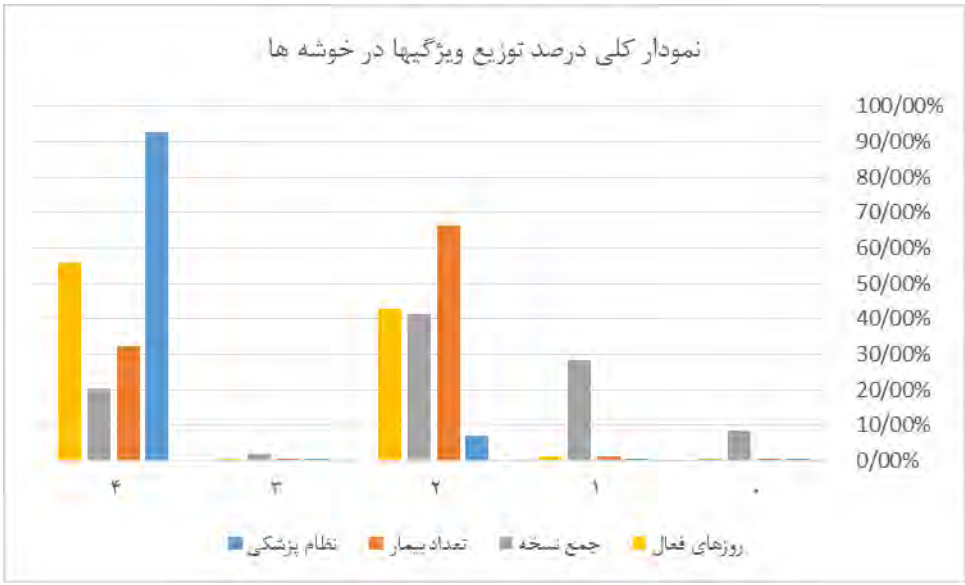
خوشه صفر: چنانکه از این جدول مشخص است ۶ پزشک شناسایی شده در خوشه ۰ رفتاری کاملا غیر نرمال دارند زیرا علیرغم اینکه کمتر از ۱ صدم درصد از پزشکان را شامل میشوند اما ۸,۴۳ درصد از هزینه های سازمان در استان تهران را تعیین میکنند. سرانه هزینه ایجاد شده نزدیک ۳۰ میلیارد ریال به ازای هر پزشک در سال ۹۵ است . اینکه این هزینه در مورد تنها یک دهم درصد از بیماران پرداخت شده جای بررسی به عنوان کیس بازرسی دارد.این پزشکان در هر روز فعالیت خود هزینه ای به طور متوسط 219,272,560 ریال برای دارو و آزمایش بیماران خود تجویز نموده اند. ممکن است این گروه شامل پزشکان معتمد وویژه در تعامل با بیماران ویژه و خاص باشند.

خوشه یک : این خوشه شامل ۹۲ پزشک با رفتار غیر نرمال هستند . هر چند سرانه هزینه به پزشک و بیمار در این خوشه به مراتب کمتر از خوشه قبل است اما هنوز با ویزیت ۱,۲۵ درصد از بیماران هزینه ای بالغ بر ۲۸ درصد را ایجاد کرده اند.در واقع ۱,۲۵ درصد از بیمارانی که ۲۸ درصد هزینه ها را ایجاد کرده اند توسط این ۹۲ پزشک معاینه شده اند.

خوشه دو : تعداد ۴,۹۹۲ پزشک شناسایی شده در این خوشه که کمتر از ۷ درصد از پزشکان هستند ۶۶,۲۶ درصد از بیماران را معاینه و این معاینه ۴۱,۲۱ درصد از هزینه های استان تهران را شامل شده است که به نظر خوشه ای پزشکان بسیار فعال است .

خوشه ۳ : این خوشه با ۱۳۷ پزشک ، کمترین بیمار را ویزیت کرده و کمترین میزان فعالیت را در جمع پزشکان داشته اند اما به دلیل داشتن فاصله نسبتا زیاد از متوسط سرانه هزینه به عنوان خوشه غیر نرمال شناسایی میشوند. علیرغم اینکه بیماران ویزیت شده توسط این پزشکان حدود ۰,۰۳ درصد را شامل میشود ولی هنوز ۱۰ برابر هزینه میانگین و ۱,۷۹ درصد از هزینه کل را شامل میشوند.

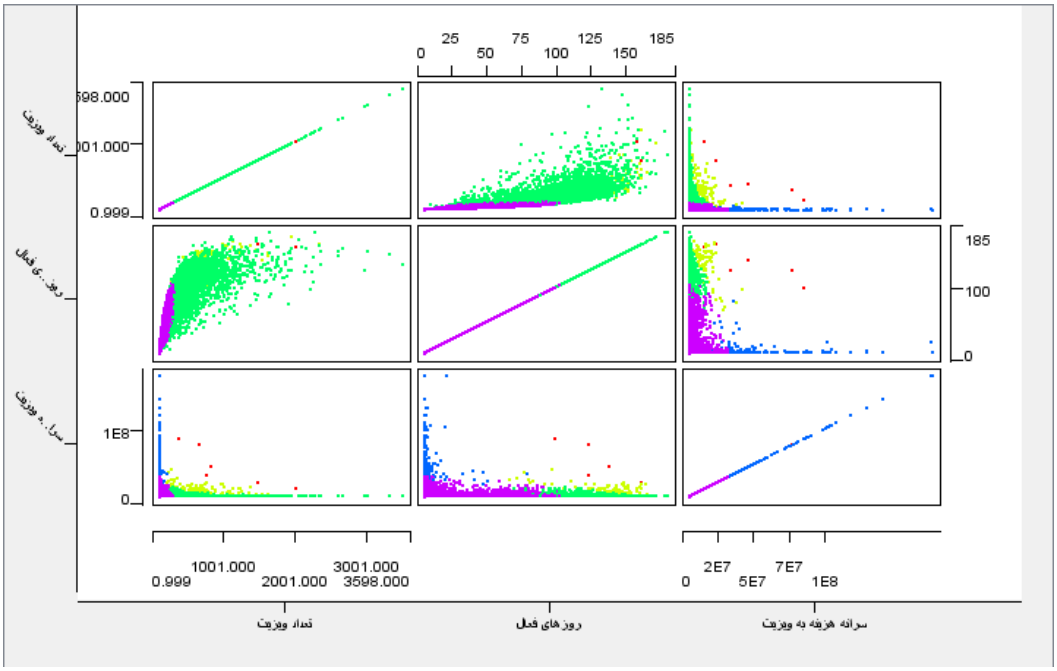
خوشه ۴: پزشکان این خوشه ۹۲,۶۹ درصد از کل پزشکان را شامل میشوند که ۳۲,۳۶ درصد از بیماران را ویزیت کرده اند . روزهای فعالیت این پزشکان ۵۵,۸۷ درصد از کل فعالیت پزشکان را شامل میشود. ۲۰,۴۲ درصد از هزینه های سازمان توسط بیماران ویزیت شده توسط این پزشکان هزینه شده است .این خوشه نیز به نظر رفتاری نرمال دارد.در انتها اطلاعات ویزیت های صورت گرفته و داروهای تجویز شده توسط پزشکان ۳ خوشه غیر نرمال بر اساس شماره نظام پزشکی از پایگاه داده ای اوراکل استخراج و در اختیار معاونت بیمه ایسازمان بیمه سلامت قرار گرفت تا به جای بررسی ۱۵۰ میلیون رکورد اطلاعاتی یا ۷۱,۵۴۵ پزشک تنها اطلاعات ۲۳۵ پزشک با بیشترین احتمال مورد رسیدگی قرار گیرد. در بررسی اولیه موارد متعددی که رفتاری مغایر دستورالعمل های موجود اعم از تخلف یا تقلب باشد شناسایی شد اما به دلیل محرمانه تلقی شدن این اطلاعات,فیدبک لازم به پژوهشگر منتقل نشده است.این یکی از بزرگترین مشکلات پژوهش در حوزه کشف تقلب و تخلف است که توسط کارگزاران به عنوان تهدید علیه نیروهای خبره و از طرفی نگرانی از افشا یا کشف روابط افراد درون سازمان با افراد متخلف یا متقلب تلقی می‌گردد.



نمودار ۲ : نمودار ستونی خوشه ها و شاخص های کلیدی در یک نگاه

چنانکه از نمودار ۲ مشخص است خوشه های ۰ و ۱ و ۳ کاملا رفتاری غیر طبیعی دارند . خوشه ۲ هم نشان میدهد که ۷ درصد پزشکان حدود ۶۶ درصد بیماران را ویزیت میکنند.

نمودارپراکندگی خوشه های پنجگانه در نمودار نشانداده شده است . این نمودار بر اساس سه ویژگی سرانه هزینه هر پزشک ، تعداد ویزیت و روزهای فعالیت پزشک تهیه شده است. نقاط بنفش و سبز وضعیت نرمالی را نشان میدهند.



نمودار ۳ : پراکندگی داده ها بر اساس تعداد ویزیت؛ سرانه ویزیت و روزهای فعال

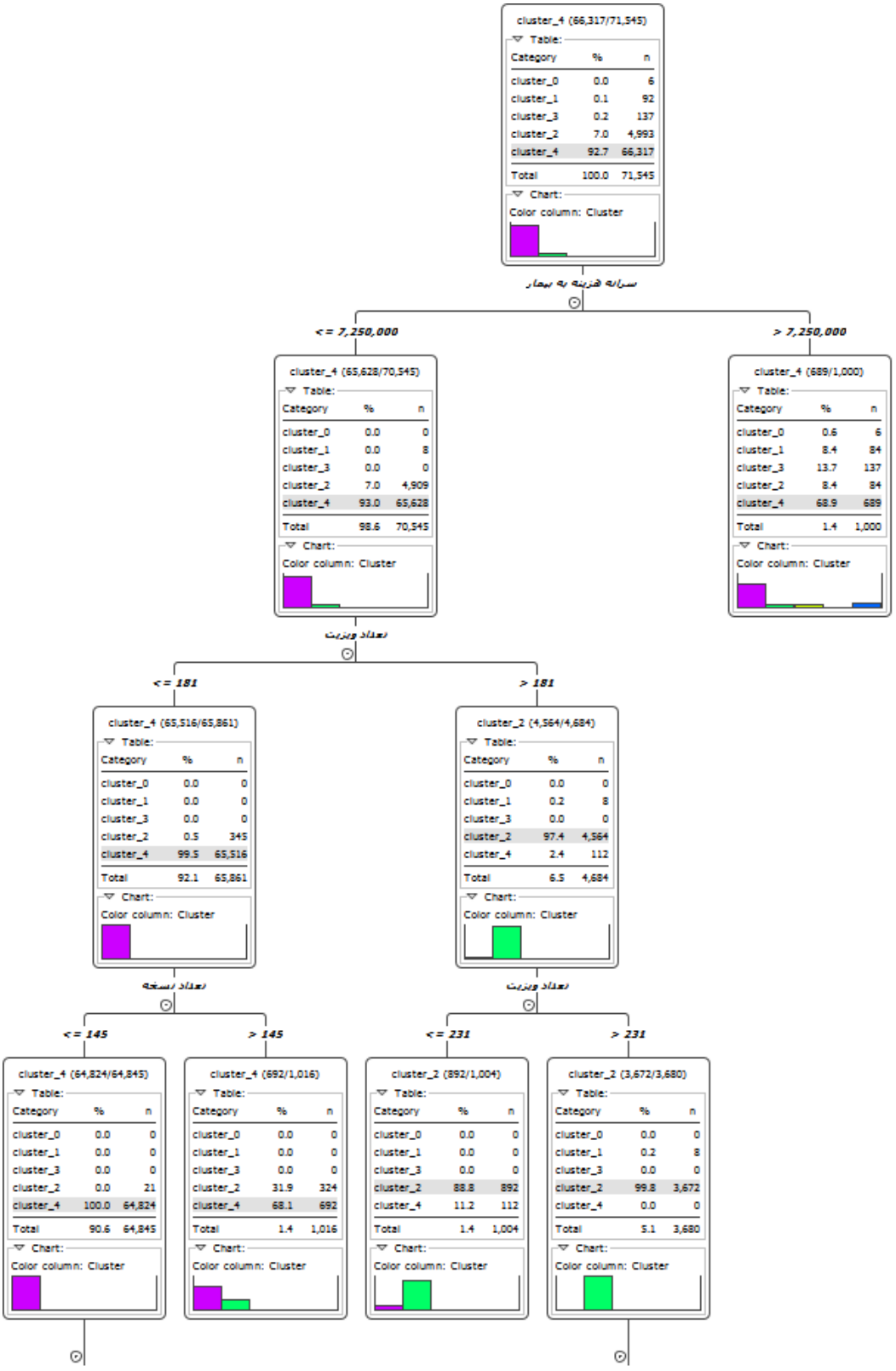
در پایان جهت پیاده سازی مدل تهیه شده در محیط نرم افزار ، درخت تصمیم مرتبط که میتواند برای تولید قواعد منطقی برای شناسایی هر کدام از خوشه ها تهیه شده است . شکل ۳ وضعیت درخت تصمیم بر اساس سرانه هزینه به بیماربه عنوان گره ریشه را نشان میدهد . شکل ۴ نمایش دهنده درخت تصمیم بر اساس گره ریشه هزینه کل و شکل ۵ درخت تصمیم را در حالتیکه گره ریشه بر اساس تعداد ویزیت است را نمایش میدهد

بحث و نتیجه‌گیری

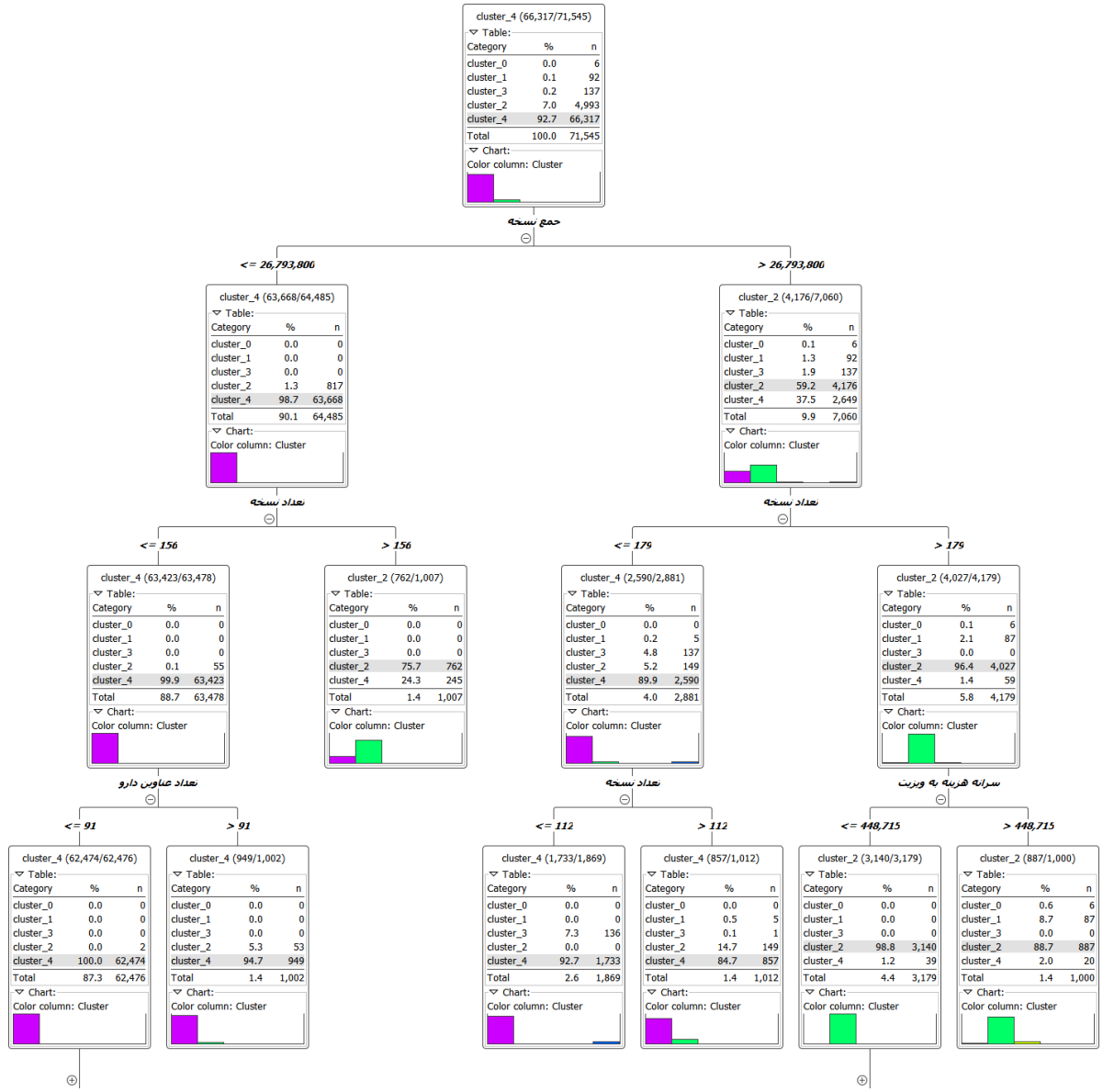
نتایج این پژوهش نشان میدهد ،با بهره برداری از روشهای مبتنی بر داده کاوی میتوانننسبت به ارزیابی حجم زیادی از داده ها در زمانی کوتاه و با صرف منابع انسانی محدود اقدام نمود . متاسفانه هنوز نتایج بررسی خوشه ها به صورت محرمانه رسیدگی میشود و نتایج آن در اختیار پژوهشگر قرار نمیگیرد . در صورت ارایه این نتایج میتوان نسبت به ایجاد مدلهایی مبتنی بر الگوریتم های با نظارت ، سرعت و صحت مدل را ارتقاء داد.

تقدیر و تشکر

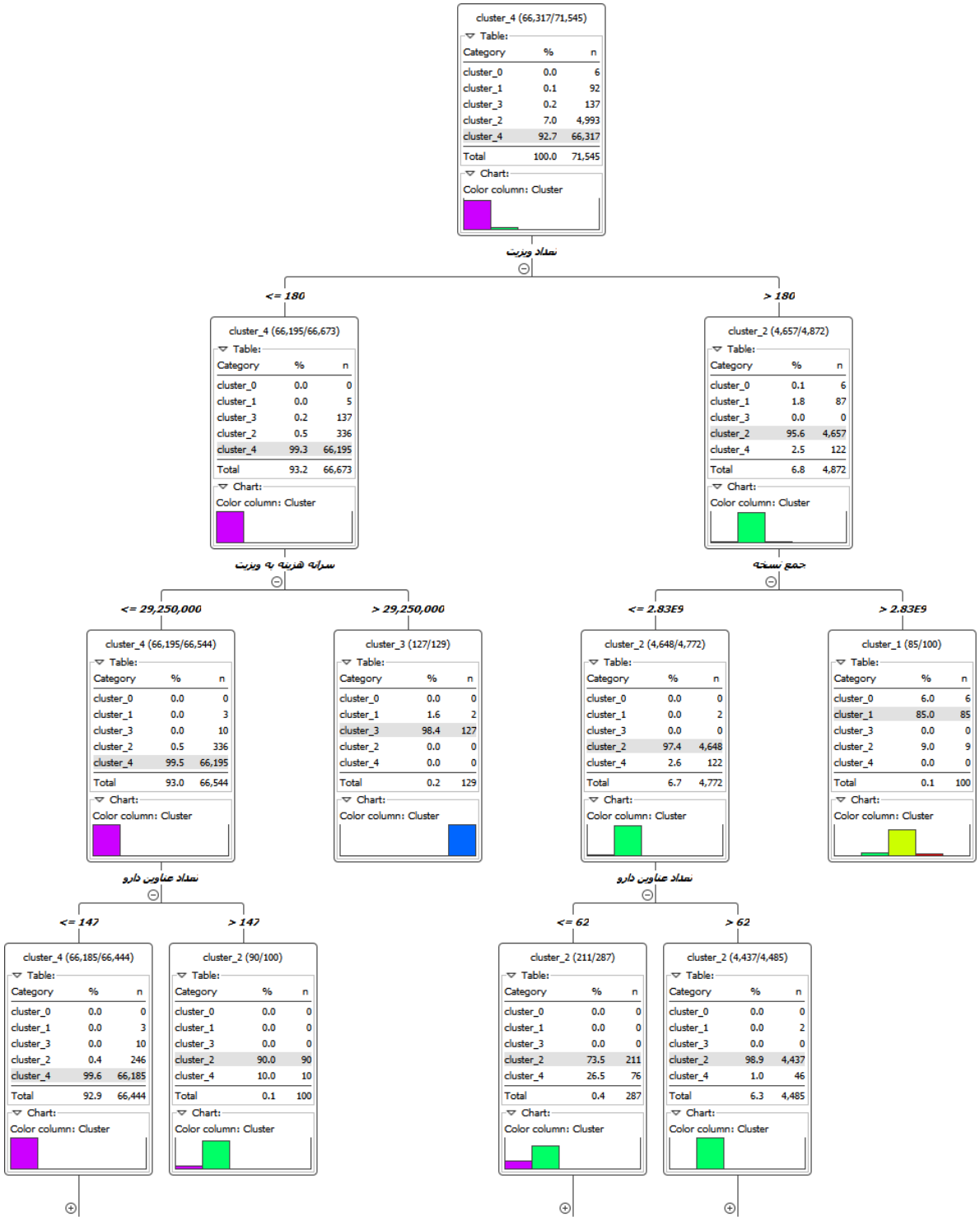
این مقاله با حمایت سازمان بیمه سلامت ایران در قالب پایان نامه سفارشی تهیه گردیده است .به دلیل محدودیتهای موجود در ارتباط محقق با بازرسان و تاخیر در تعاملات نظر خبرگانی این پژوهش با حمایت جمعی از کارشناسان رسیدگی اسناد پزشکی سایر سازمانهای بیمه گذار تهیه گردیده است که بدینوسیله از آنان قدردانی میگردد.



شکل۳ : درخت تصمیم مدل تهیه شده بر اساس سرانه هزینه به بیمار



شکل۴ : درخت تصمیم بر اساس هزینه کل



شکل۵ : درخت تصمیم بر اساس گره ریشه ی تعداد نسخه

منابع:

He, H., Wang, J., Graco, W., & Hawkins, S. (1998). Application of Neural Networks to Detection of Medical Fraud. Expert Systems with Applications, 13(4), 329-336.

Hossein Joudaki,Arash Rashidian,· Behrouz Minaei-Bidgoli,Mahmood Mahmoodi,Bijan Geraili, Mahdi Nasiri,and Mohammad Arab (2011). Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study, Int J Health Policy Manag. 2016: 165–172. doi:10.15171. PMCID: PMC4770922

Li, Jing, Kuei-Ying Huang, Jionghua Jin, and Jianjun Shi. "A survey on statistical methods for health care fraud detection." Health Care Management Science 3, no. 11 (2008): 275-287.

Li, J., Huang, K.-Y., Jin, J., & Shi, J. (2008). A Survey on Statistical Methods for Health Care Fraud Detection. Health Care Management Science, 11, 275-287.

Peng, Y., Kou, G., Sabatka, A., Chen, Z., Khazanchil, D., & Shi, Y. (2006). Application of Clustering Methods to Health Insurance Fraud Detection. International Conference on Service Systems and Service Management, 116-120. Troyes, France: IEEE.

Robbins, D. B., & Anderson, A. (2011, October 10). Too Much Care? Stepped Up Medical Necessity fraud Litigation Against Hospitals . Retrieved September 15, 2012, from Washington Healthcare News: www.wahcnews.com.

Shin, H., Park, H., lee, J., & Jhee, W. C. (2012). A Scoring Model to Detect Abusive Billing Patterns in Health Insurance Claims. Expert Systems with Applications, 39(1), 7441-7450.

Sparrow, M. (1998, December). Fraud Control in the Health Care Industry*: Assessing the State of the Art*. Retrieved 08 28, 2012, from National Criminal Justice Reference service: https://www.ncjrs.gov.

Uzi, M., & Gadi, P. (1999). Unsupervised Profiling for Identifying Superimposed Fraud.Principles of Data Mining and Knowledge Discovery Lecture Notes in Computer Science, 1704, 251-261.

Wynia, M., Cummins, D., VanGeest, J., & Wilson, I. (2000, April 12). Physician Manipluation of Reimbursement Rules for Patients. Journal of American Medical Association (JAMA), 283(14), 1858-1865.

Yang, W.-S., & Hwang, S.-Y. (2006). A Process Mining Framework for the Detection of Healthcare Fraud and Abuse. Expert Systems with Applications, 31(1), 56-68.