پژوهشنامهٔ بیمه (صنعت بیمهٔ سابق)/سال‌بیست‌وششم/شمارهٔ ۳/پاییز ۱۳۹۰/شماره‌مسلسل۱۰۳/صفحات ۱۰۳ تا ۱۲۸

# شناسایی تقلب در بیمهٔ اتومبیل بااستفاده از روش‌های داده‌کاوی

مهدی فیروزی [۱]
مرتضی شکوری [۲]
لیلا کاظمی [۳]
سحر زاهدی [۴]

تاریخ دریافت مقاله: ۱۳۹۰/۰۱/۲۷
تاریخ پذیرش مقاله: ۱۳۹۰/۱۰/۱۲

## چکیده

*تقلب‌های بیمه‌ای از مسائل مهم و خسارت‌زا برای شرکت‌های بیمه و بیمه‌گذاران، در تمام رشته‌های بیمه‌ای است. یکی از راه‌های شناسایی تقلب در خسارت‌های اعلام‌شده، استفاده از اطلاعات تقلب‌های کشف‌شده در گذشته است. امروزه روش‌های داده‌کاوی به‌طور گسترده در کشف الگوها در داده‌ها استفاده می‌شوند. استفاده از این روش‌ها می‌تواند در شناسایی خسارت‌های تقلبی در صنعت بیمه مفید باشد. در این مقاله ضمن بررسی روش‌های رایج برای شناسایی تقلب در بیمهٔ اتومبیل از سه روش داده‌کاوی رگرسیون لجستیک، بیز ساده و درخت تصمیم برای پیداکردن الگوهایی استفاده شده است که به شرکت‌های بیمه در شناسایی تقلب‌ها در بیمه اتومبیل کمک می‌کنند. همچنین در یک مطالعهٔ تجربی این روش‌ها بر روی داده‌های واقعی (شامل اطلاعات ۷۲ پروندهٔ خسارت بیمه‌نامه‌های شخص ثالث و بدنهٔ اتومبیل) آزمایش و کارایی هر روش سنجیده شد. روش بیز ساده با دقت ۹۰/۲۸ درصد در شناسایی صحیح جعلی یا غیرجعلی‌بودن پرونده‌های خسارت بهترین کارایی را در مقایسه با دو روش درخت تصمیم با دقت ۸۸/۹ درصد و رگرسیون لجستیک با دقت ۸۶/۱ درصد داشت.*

**واژگان کلیدی:** *شناسایی تقلب، رگرسیون لجستیک، بیز ساده، درخت تصمیم، داده‌کاوی*

۱. کارشناس‌ارشد آمار ریاضی، دانشگاه صنعتی امیرکبیر (Email: Mehdi.firouzi@aut.ac.ir)
۲. کارشناس‌ارشد علوم اکچوئری، دانشگاه علامه طباطبائی (Email:Mrtshakouri@Gmail. com)
۳. کارشناس‌ارشد مدیریت بازرگانی، دانشگاه تهران (نویسنده مسئول) (Email: Leila.kazemi@ut.ac.ir)
۴. کارشناس‌ارشد فناوری اطلاعات، دانشگاه صنعتی امیرکبیر (Email: Sahar.zahedi@aut.ac.ir)

## ۱. مقدمه

کلاهبرداری‌های بیمه‌ای، هرساله خسارت‌های زیادی را به شرکت‌های بیمه تحمیل می‌کند. باوجود پیشرفت‌های فراوان در شناسایی این تقلب‌ها، هزینه‌های ایجادشده برای شرکت‌های بیمه در اثر این کلاهبرداری‌ها در حال افزایش است. تقلب در صنعت بیمه ممکن است در مراحل مختلف و توسط اشخاص مختلفی رخ دهد: بیمه‌گذاران جدید، بیمه‌گذاران فعلی، اشخاص ثالث زیان‌دیده یا متخصصانی که به بیمه‌گذاران خدمات ارائه می‌دهند (Terisa, 2010). متخصصان بیمه‌ای هنوز به یک توافق کلی در مورد تعریف تقلب در بیمه نرسیده‌اند. دریگ و همکارانش[1] چهار اصل را برای تعریف تقلب بیان کرده‌اند: تقلب باید آشکار و عمدی باشد؛ علیهٔ قانون باشد؛ در آن منفعت مالی وجود داشته باشد و اطلاع‌رسانی نادرست اتفاق افتاده باشد. گیل و همکارانش[2]، کلاهبرداری بیمه‌ای را این‌گونه تعریف کرده‌اند: «اعلام عمدی خسارت‌های جعلی، اعلام خسارت بیش از مقدار واقعی آن، یا هر روش دیگر برای به‌دست‌آوردن مبلغی بیش از آنچه که بیمه‌گذار قانوناً مستحق دریافت آن باشد». کلاهبرداری‌های بیمه‌ای در قوانین بسیاری از کشورهای جهان، جرم کیفری تلقی شده و در صورت اثبات، مرتکبین آن علاوه‌بر بازگرداندن وجوه ناشی از کلاهبرداری، محکوم به جریمهٔ مالی و حتی حبس می‌شوند. انواع کلاهبرداری و تقلب در صنعت بیمه بسیار متنوع است و به‌صورت روزانه، حتی بیش از آنچه که ما فکرش را بکنیم در اطراف ما اتفاق می‌افتد. از بزرگ‌ترین انواع کلاهبرداری‌ها در بیمه، ارائهٔ اطلاعات نادرست است. برخی بیمه‌گذاران اطلاعات غیرواقعی به بیمه‌گر می‌دهند، این در حالی است که در صورت ارائهٔ اطلاعات صحیح، بیمه‌گر تصمیمی متفاوت اتخاذ می‌کرد (Wilson, 2003).

1. Derrig et al, 2006
2. Gill et al, 1994

در این مقاله ابتدا مروری بر مراحل کنترل تقلب، در مدل رایج در روند ارزیابی خسارت شده است. در ادامه سه روش رگرسیون لجستیک[1]، بیز ساده[2] و درخت تصمیم[3]، که از ابزارهای مورد استفاده در دانش داده‌کاوی هستند به‌عنوان تکنیک‌هایی برای شناسایی و دسته‌بندی خسارت‌های تقلبی معرفی شده‌اند. سپس در یک تحقیق تجربی بااستفاده از داده‌های واقعی، مدل‌هایی براساس این سه روش ارائه شده است که به کشف تقلب در بیمهٔ اتومبیل کمک می‌کنند. برای این منظور اطلاعات ۷۲ پروندهٔ خسارت بیمهٔ اتومبیل که در ۳۶ مورد از آنها تقلب صورت گرفته مورد استفاده قرار گرفته است. در انتها نتایج به‌کارگیری این سه روش با یکدیگر مقایسه شده‌اند.

## ۲. کلاهبرداری در صنعت بیمه

در سال ۲۰۰۲، مؤسسهٔ تحقیقاتی فرانک به سفارش انجمن بیمه‌گران بریتانیا، تحقیقی با شرکت ۲۰۰۰ نفر انجام داد. هدف اصلی این تحقیق سنجش دیدگاه‌های مردم در خصوص ادعاهای تقلبی در صنعت بیمه بود. هدف دیگری که از طراحی این تحقیق دنبال می‌شد، این بود که تقلب و سوءاستفاده از بیمه را جزو اقدامات خلاف قانون در جامعه مطرح کند. نتایج این تحقیق نشان می‌دهد که بخشی از تقلب و سوءاستفاده در بیمه، ناشی از ناآگاهی و عدم‌شناخت مردم دربارهٔ چیزی است که درست است. بیشتر کسانی که در این تحقیق مورد پرسش قرار گرفته‌اند، دربارهٔ آنچه که رفتار درست تلقی می‌شود، اطلاع دقیقی نداشته‌اند. نتایج این تحقیق نشان داد که:

- اگرچه بیشتر پرونده‌ها و دعاوی بیمه‌ای درست و صحیح است، تقریباً نیمی از پرسش‌شوندگان احتمال تقلبی‌بودن یک ادعا را رد نکرده‌اند.
- احتمال وقوع تقلب بیمه‌ای بیشتر از سایر سوء استفاده‌هاست.

1. Logistic Regression
2. Naive Bayes
3. Decision Tree

– در میان افراد شرکت‌کننده در تحقیق، در خصوص درست یا نادرست‌بودن اقداماتی مانند خرید مال مسروقه یا رانندگی در حال مستی، دیدگاه‌های متفاوتی وجود دارد.

یکی دیگر از یافته‌های تحقیق این بود که ۶٪ از شرکت‌کنندگان در تحقیق، به تقلب در بیمه اذعان کرده بودند که نمونهٔ بارز آن اغراق در خسارت‌های وارد شده بود. ۲٪ از شرکت‌کنندگان نیز به طرح ادعای ساختگی اذعان کرده بودند. به‌دلیل اینکه برخی از افراد ممکن است عملکرد نادرست خود را کتمان کنند، آمار مربوط‌به تقلب ممکن است بیشتر از این باشد (راه‌چمنی، ۱۳۸۵).

کلاهبرداری در بیمهٔ اتومبیل از روش‌های مختلفی صورت می‌گیرد، برخی از شرکت‌ها اغراق در اعلام میزان خسارت و برخی دیگر سایر فعالیت‌های هدفمند، مانند تصادفات ساختگی، اسناد جعلی و ارائهٔ اطلاعات نادرست را به‌عنوان مصادیق تقلب درنظرمی‌گیرند. بعضی از کلاهبرداری‌ها در صنعت بیمه کاملاً آگاهانه و عمدی است. بیمه‌گذار ممکن است موجبات بروز خسارتی را فراهم آورد تا بدین طریق از محل بیمه‌نامهٔ خود منفعتی کسب کند. به‌طورکلی، بیمه‌گذاران در دو موقعیت مرتکب تقلب می‌شوند: مورد اول، شرایطی است که در آن، فرد آگاهانه سعی در ایجاد خسارت یا اغراق در میزان و نوع خسارت دارد؛ به‌عنوان‌مثال، در یک سانحهٔ تصادف ممکن است فرد بیمه‌گذار باتوجه به حق‌بیمه‌ای که برای سالیان متمادی به شرکت بیمه پرداخت نموده است درصدد بهره‌برداری از فرصت برآید و با تجمیع کلیهٔ زیان‌های پیشین با خسارت فعلی سعی در کسب موقعیت مالی بهتر کند. مورد دوم که ممکن است منجربه خسارت‌های جعلی گردد، مواردی است که بیمه‌گذار به صرف داشتن بیمه‌نامه، احتیاط کمتری می‌کند. بدین معنی که گرچه ممکن است شخص قصد ایجاد خسارت یا اغراق در میزان آن را نداشته باشد، با این حال اقدام به انجام فعالیت‌هایی می‌کند که در صورت نداشتن بیمه‌نامه، این فعالیت‌ها را انجام نمی‌داد.

وجود بیمه و تعهد به جبران خسارت باعث می‌شود بیمه‌گذاران ریسک‌هایی را متحمل شوند که در صورت عدم‌وجود بیمه از این ریسک‌ها دوری می‌جستند (Wilson, 2003). تصادفات ساختگی برای دریافت خسارت از محل بیمه‌نامه، ویژگی‌های خاصی دارند از جمله اینکه معمولاً در ساعات پایانی شبانه روز و خارج از نواحی شهری اتفاق می‌افتند. در این‌گونه موارد اغلب، رانندگان جوان و تعداد سرنشینان خودرو به‌نسبت زیاد است، اما کودکان و سالخوردگان حضور ندارند و همچنین نشانه‌های دیگری که در اینجا به آنها اشاره نمی‌کنیم (Subelj et al, 2010). ارائهٔ اطلاعات نادرست دربارهٔ میزان استفاده از خودرو در هنگام صدور بیمه‌نامه، اعلام خلاف واقع خسارت‌های بدنی به شرکت بیمه هنگام تصادف از طریق تلفن، اعلام خلاف واقع به شرکت بیمه دربارهٔ دزدیده‌شدن اتومبیل، اعلام خلاف واقع به شرکت بیمه دربارهٔ صدمه‌دیدن اتومبیل و اینکه در تصادف راننده مقصر فرار کرده است و همچنین اعلام واردشدن صدمات بدنی به کسانی که هنگام تصادف داخل اتومبیل نبوده‌اند نیز از مصادیق دیگر تقلب در بیمه‌های اتومبیل است.

مسئولیت تشخیص خسارت‌های تقلبی در شرکت‌های بیمه برعهده کارشناسان خسارت است که بعضاً افرادی کم‌تجربه بوده و اغلب آموزش‌های لازم را در خصوص کشف تقلب سپری نکرده‌اند (Doig et al, 1999). براساس برآوردی که از طریق مصاحبه با مدیران و خبرگان به‌دست‌آمده‌است نرخ کشف تقلب در حدود ۱۰٪ است که این بدان معنی است که تعداد بسیار زیادی از خسارت‌های تقلبی شناسایی نمی‌شوند (Morley et al, 2006). ازآنجایی‌که روند ارزیابی خسارت‌ها معمولاً به‌صورت دستی انجام می‌گیرد و کمتر از سیستم‌های کامپیوتری استفاده می‌شود، ادعاهای تقلبی معمولاً شناسایی نمی‌شوند. همچنین باتوجه به اینکه همواره شیوه‌های جدیدی در کلاهبرداری‌ها به‌کارگرفته‌می‌شود، روش‌های مورد استفاده در کشف تقلب باید قابلیت کافی برای شناسایی تقلب را داشته باشند. هرچند کشف کاملاً

خودکار کلاهبرداری‌ها در عمل ممکن نیست، اما استفاده از اطلاعات تقلب‌های کشف‌شده در گذشته و بهره‌گیری از تکنیک‌های آماری می‌تواند به کارشناسان خسارت در شناسایی خسارت‌های جعلی کمک کند (Viaene et al, 2007).

## ۳. مروری بر تحقیقات گذشته

از دههٔ ۹۰ تاکنون تحقیقات بسیاری در زمینهٔ کشف خسارت‌های تقلبی در رشتهٔ بیمهٔ اتومبیل انجام شده است. برای مثال می‌توان به تحقیق ویسبرگ و دریگ[1] اشاره کرد. بلهادجی و دیون[2] تحقیقاتی را دراین‌زمینه بااستفاده از داده‌های خسارت بیمهٔ اتومبیل کشور کانادا انجام داده‌اند. همچنین کیومینز و تنیسن[3] تحقیقات مشابهی را در ایالات متحده انجام دادند. محققان دیگری مانند دریگ و استاسزیوسکی[4]، ویزبرگ و دریگ[5] و براکت و همکارانش[6] در مقالات خود تکنیک‌هایی برای شناسایی خسارت‌های تقلبی و دسته‌بندی کلاهبرداری‌ها ارائه دادند. آرتیس و همکارانش[7] عملکرد مدل‌های انتخاب باینری را برای کشف تقلب در بازار بیمهٔ اتومبیل اسپانیا برای سال‌های ۱۹۹۶تا ۱۹۹۳ تجزیه‌وتحلیل کردند. آنها روشی برای اصلاح طبقه‌بندی نوع خسارت معرفی کردند.

باتوجه به تنوع حجم و نوع داده‌ها، روش‌های آماری زیادی برای کشف تقلب‌ها وجود دارند. این روش‌ها می‌توانند با ناظر یا بی‌ناظر باشند. در روش‌های باناظر، نمونه‌هایی از موارد تقلبی و غیرتقلبی موجود است و مدلی ساخته می‌شود که بر اساس آن، تقلبی یا غیرتقلبی بودن نمونه‌های جدید مشخص می‌شود. این روش جهت تشخیص انواع تقلباتی

1. Weisberg, & Derrig, 1993
2. Belhadji, & Dionne, 1997
3. Cummins, & Tennyson, 1992
4. Derrig, & Ostazewski, 1995, 1995
5. Weisberg & Derrig, 1993
6. Brockett et al, 1998
7. Artis et al, 2002

مناسب است که از قبل وجود داشته‌اند. روش‌های بی‌ناظر، به دنبال کشف نمونه‌هایی هستند که کمترین شباهت را با نمونه‌های نرمال دارند (Bolton & Hand, 2002)

واین و ددِن[1] در مقاله‌ای در سال ۲۰۰۴، با به‌کارگیری روش بیز ساده[2]، اقدام به کشف تقلب در داده‌های بیمهٔ اتومبیل کردند. در این مقاله از الگوریتم‌های تقویت‌کننده[3] استفاده شده و نتایج آن با نتایج حاصل از اعمال الگوریتم بیز ساده بدون به‌کارگیری الگوریتم‌های تقویت‌کننده مقایسه شده و ثابت شده است که استفاده از الگوریتم‌های تقویت‌کننده، نتایج دقیق‌تری به دست می‌دهند. بلهادجی و دیون در مقاله خود در سال ۱۹۹۷، ابتدا با تحقیق و تفحص از خبرگان صنعت بیمه اتومبیل، به شناسایی عوامل کلیدی در تقلبات بیمه‌ای پرداختند و سپس با محاسبهٔ احتمال شرطی تقلب برای هر شاخص و به‌کارگیری الگوریتم رگرسیون، مهم‌ترین شاخص‌ها را تعیین کردند. همچنین به کمک الگوریتم رگرسیون، به پیش‌بینی خسارت‌های تقلبی پرداختند. آرتیس و همکارانش نیز در سال ۲۰۰۲، به مقایسهٔ مدل‌های لاجیت چند جمله‌ای[4] و مدل لاجیت چند جمله‌ای تو در تو[5] در شناسایی تقلبات بیمهٔ اتومبیل پرداختند. فوآ و همکارانش[6]، با ترکیب الگوریتم‌های شبکه‌های عصبی پس انتشاری[7]، بیز ساده و درخت تصمیم C4.5 به کشف تقلب در بیمه‌های اتومبیل پرداختند. بروکت و همکارانش[8] در سال ۱۹۹۸، نیز در مقاله خود، ابتدا به کمک الگوریتم تحلیل مؤلفه‌های اصلی[9] به انتخاب ویژگی‌ها پرداختند و سپس با ترکیب

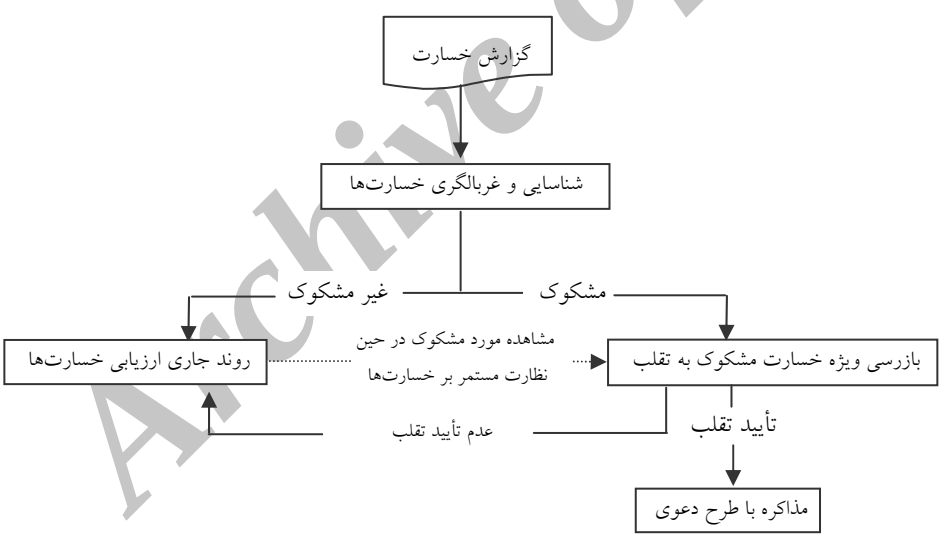
1. Viaene & Dedene, 2004
2. Naive Bayes
3. Boosting Algorithms
4. Multinominal Logit Model (MLM)
5. Nested Multinominal Logit Model (NMLM)
6. Phua et al, 2004
7. Backpropagation Neural Network
8. Brockett et al, 1998
9. Principal Component Analysis (PCA)

الگوریتم‌های خوشه‌بندی و شبکه‌های عصبی BP به کشف تقلبات بیمهٔ اتومبیل پرداختند.

## ۴. روند شناسایی تقلب

مدل رایج (نمودار ۱) برای کنترل تقلب‌ها، شامل مراحل شناسایی و غربال‌گری، تحقیق و بررسی، مذاکره با بیمه‌گذار یا طرح دعوی است که در روند ارزیابی خسارت‌ها اجرا می‌شود. روند ارزیابی خسارت‌ها با رخداد یک حادثه و اعلام گزارش به شرکت بیمه آغاز و با پرداخت یا عدم‌پرداخت خسارت پایان می‌یابد. عواملی چون عدم‌تمایل به ارائهٔ اطلاعات صحیح از نشانه‌های کلاهبرداری است که در صورت اثبات تخلف منجربه عدم‌پرداخت خسارت می‌گردند.

**نمودار۱. مدل کنترل کلاهبرداری و تقلب**




(Viaene et al, 2007)

### ۱-۴. شناسایی و غربال‌گری

در این مرحله خسارت‌های مشکوک به تقلب شناسایی و تفکیک می‌شوند. خسارت‌هایی که از این مرحله گذر می‌کنند، طبق روال معمول و با حداقل هزینه‌های اداری ارزیابی می‌شوند، اما خسارت‌هایی که مشکوک به تقلب‌اند باید در مراحل بعدی مورد ارزیابی دقیق‌تر قرار گیرند که این امر مستلزم صرف زمان، هزینه و نیروی انسانی بیشتر است (Ghezzi, 1983). بدون وجود سیستم‌های هوشمند، بررسی خسارت‌ها تنها براساس اطلاعات موجود در مورد بیمه‌گذار و خسارت وارده ممکن است. اما ازآنجاکه معمولاً جستجوی دستی در پرونده‌ها و موارد مشابه گذشته، بسیار مشکل و زمان‌بر است، کارشناسان خسارت باید براساس اطلاعات بسیار محدود و اغلب با اتکا به تجربیات به تصمیم‌گیری بپردازند (Viaene et al, 2007). با دراختیارداشتن داده‌های مربوط‌به بیمه‌گذاران و خسارت‌ها در چند سال گذشته و شناسایی معیارهای تقلب در هر رشته بیمه‌ای، شرکت‌های بیمه به‌راحتی می‌توانند موارد مشکوک به تقلب را غربال کنند (Derrig et al , 2006). معمولاً داده‌ها از سه منبع قابل دستیابی‌اند:

- گردآوری داده‌ها در مرحلهٔ صدور بیمه‌نامه از طریق فرم‌هایی که توسط بیمه‌گذاران پر می‌شوند، اطلاعاتی در مورد بیمه‌گذاران و اتومبیل بیمه‌شده از قبیل تاریخ تولد، نشانی، نوع اتومبیل، تاریخ اخذ گواهینامهٔ رانندگی، نوع کاربری اتومبیل و ... که غالباً عوامل مؤثر در شناسایی ریسک تحت پوشش و تعیین نرخ مناسب در محاسبهٔ حق‌بیمه است را برای بیمه‌گر فراهم می‌آورد. همچنین این اطلاعات در آینده به همراه جزئیات خسارت در تکمیل پروفایل مشتریان استفاده می‌شود.
- گردآوری داده‌ها در مرحلهٔ ارزیابی خسارت‌ها که توسط کارشناسان مربوطه جهت پرداخت خسارت استفاده می‌شود، داده‌هایی از قبیل زمان، مکان، شرح وقوع و علت

حادثه، شاهدان و مشخصات اتومبیل‌های ثالث (نوع، سال ساخت، سازنده) و ... را در اختیار شرکت بیمه قرار می‌دهند.

- گردآوری داده‌های موجود در پایگاه داده‌هایی که در صنعت اتومبیل اطلاعات مربوط‌به خودروها، مدل‌های آنها و هزینهٔ خرید و تعمیر قطعات مختلف را در اختیار کارشناسان خسارت قرار می‌دهند. به کمک چنین پایگاه‌های داده‌ای، ارزیابان خسارت می‌توانند به سرعت مبالغ قابل پرداخت را محاسبه کنند (Viaene et al, 2007).

**۲-۴. تحقیق و بررسی**

تشخیص تقلبی‌بودن ادعا برعهده ارزیاب خسارت است که وی براساس تجربه، توانایی و خلاقیت خود این فرایند را انجام می‌دهد (Viaene & Dadene, 2004). براساس تحقیق تنیسن و سالساس[1] روش‌های رایج رسیدگی خسارت‌ها عبارت‌اند از: بازدید از محل، بررسی پیشینه، گزارش‌های واحدهای ویژهٔ بازرسی و نظارت بر فعالیت‌های بیمه‌گذار (Derrig, 2002).

**۳-۴. مذاکره با بیمه‌گذار یا طرح دعوی**

اغلب شرکت‌های بیمه ترجیح می‌دهند به همان روش‌های سنتی به بازرسی خسارت‌ها جهت کشف و جلوگیری از تقلب بپردازند، ولی بااین‌حال در برخی از موارد نیز نیاز به دادگاه خواهد بود. اما دعوی قضایی و بازرسی‌های ویژه معمولاً مستلزم صرف هزینه و زمان زیادی است. معمولاً شرکت‌های بیمه به‌دلیل تأثیری که ممکن است طرح دعوی در دادگاه و شکست احتمالی در آن، بر شهرت شرکت در بازار داشته باشد تمایلی به طرح دعوی در دادگاه‌ها ندارند (Viaene et al, 2007).

1. Tennyson & Salsas, 2002

## ۵. فراگیری ماشین و داده‌کاوی

عصر حاضر، عصر اطلاعات است. اطلاعات فراوانی در قالب پایگاه‌های داده ذخیره شده است که تبدیل آنها به دانش مورد نیاز جهت تصمیم‌گیری، نیازمند ابزارهای جدیدی است. روش‌های آماری برای تحلیل داده‌ها بیشتر بر پایهٔ استخراج شاخص‌های کمی استوار است. اگرچه این روش‌ها به‌صورت غیرمستقیم ما را به دانش مورد نیاز جهت تصمیم‌گیری سوق می‌دهند، اما درنهایت تفسیر نتایج آنها نیازمند تحلیل‌های انسانی است. روش‌های نوین تحلیل داده، تفسیر داده‌ها را تسهیل نموده و می‌توانند درک بهتر فرایندها را فراهم کنند. به‌منظور تسهیل فرایند تصمیم‌گیری، سیستم‌های تحلیل داده باید به دانش لازم و قابلیت تصمیم‌گیری براساس داده‌ها تجهیز شوند. جهت دستیابی به این هدف، محققین به ارائه ایده‌های جدیدی از فراگیری ماشین پرداخته‌اند. باتوجه به این ایده‌ها وظیفهٔ فراگیری ماشین، تبدیل داده‌ها (ورودی) به دانش تصمیم‌گیری (خروجی) خواهد بود. همچنین براساس این ایده‌ها، ضرورت پیدایش یک حوزهٔ تحقیقاتی جدید که داده‌کاوی نام گرفته به وجود آمده است (Michalski et al, 1998). داده‌کاوی فرایند کشف الگوها در داده‌هاست. این فرایند باید خودکار یا نیمه خودکار باشد. الگوهای شناسایی‌شده باید معتبر بوده و برای ما مزایایی از جمله مزایای اقتصادی داشته باشند. همچنین داده‌ها باید همواره در قالب کمیت‌های معتبر ارائه شوند (Witten & Frank, 2000). استفاده از مدل‌های ریاضی برای شناسایی تقلب، این امکان را به متخصصین شرکت‌های بیمه می‌دهد که با صرف زمان و هزینهٔ کمتری تشخیص دهند که ادعای خسارت اعلام‌شده ازلحاظ آماری مشکوک به تقلب است یا خیر. در ادامه سه روش رگرسیون لجستیک، بیز ساده و درخت تصمیم که از ابزارهای رایج در داده‌کاوی هستند معرفی و بااستفاده از این روش‌ها مدل‌هایی برای شناسایی و دسته‌بندی خسارت‌های تقلبی بر روی داده‌های واقعی برازش داده خواهد شد.

## ۶. مطالعهٔ تجربی

برای ساختن یک مدل ریاضی، نیاز به داده‌هایی از هر دو دستهٔ ادعاهای جعلی و غیرجعلی داریم. در این بخش بااستفاده از اطلاعات ۷۲ پروندهٔ خسارت بیمهٔ اتومبیل (شامل بیمه‌نامه‌های شخص ثالث و بدنهٔ اتومبیل) که در طی سال‌های ۱۳۸۹ تا ۱۳۸۳ ثبت شده‌اند و به‌کارگیری روش‌های ذکرشده در بخش قبل، مدل‌هایی جهت شناسایی تقلب در بیمهٔ اتومبیل ارائه شده است. نیمی از این پرونده‌ها (۳۶ پرونده)، پرونده‌هایی هستند که وقوع تقلب در آنها توسط کارشناسان ذی‌ربط تشخیص داده شده است و ۳۶ پروندهٔ باقی‌مانده به‌صورت تصادفی از بین پرونده‌های خسارت پرداختی انتخاب شده است (لازم به ذکر است که این امکان وجود دارد که در بین پرونده‌های خسارت پرداختی نیز موارد جعلی وجود داشته باشد که به هر دلیلی شناسایی نشده‌اند). سابقهٔ بیمهٔ بیمه‌گذارانی که پرونده‌های خسارت آنها انتخاب شده است بین ۱ تا ۷ سال متغیر بوده است. این بیمه‌گذاران در طول سابقهٔ بیمه‌شان ۱ تا ۳ ادعای خسارت داشته‌اند. فاصلهٔ زمانی بین وقوع حادثه تا اعلام خسارت به شرکت بیمه از سوی بیمه‌گذار در پرونده‌های خسارت مورد بررسی از ۰ روز تا ۴۹۹ روز بوده است. در بین ۷۲ پروندهٔ خسارت مورد بررسی، ۶۶ مورد دارای کروکی و بقیه فاقد کروکی بوده‌اند. همچنین ۵۰ مورد از پرونده‌های خسارت، مالی و بقیه جانی بوده‌اند. حجم نمونهٔ مورد استفاده برای ساختن مدل کوچک است، اما این تعداد، تمام پرونده‌های خسارت جعلی شناسایی‌شده بوده است. بهتر است از تعدادی پروندهٔ خسارت جعلی دیگر (غیر از پرونده خسارت‌های جعلی مورد استفاده جهت ساخت مدل) برای بررسی صحت مدل‌های ارائه‌شده استفاده گردد. اما باتوجه به کم‌بودن تعداد پرونده‌های جعلی موجود این امکان وجود نداشت.

**۱-۶. متغیرهای مورد استفاده در مدل**

در هریک از سه مدلی که در این بخش برای شناسایی تقلب معرفی خواهند شد، جعلی یا غیرجعلی‌بودن یک پرونده، به‌عنوان متغیر وابسته درنظرگرفته می‌شود. مقدار ۱ برای متغیر وابسته به معنای جعلی‌بودن پرونده خسارت و مقدار ۰ به معنای غیرجعلی بودن آن پرونده است. در این مطالعه، فرایند شناسایی تقلب بااستفاده از شش متغیر مستقل صورت گرفته است. برای انتخاب متغیرها از روش ترکیبی استفاده شده است. روش ترکیبی از ترکیب دو روش پیش‌رونده و پس‌رونده تشکیل شده است. در این روش در هر مرحله از ورود متغیرها به مدل، متغیری که کمترین ارتباط را با متغیر وابسته داشته باشد، حذف و متغیری که بیشترین ارتباط را داشته باشد، انتخاب می‌شود. همچنین برای بهترشدن نتایج از نظرات کارشناسان بیمه اتومبیل نیز بهره گرفته شده است. اولین متغیر مستقل، سابقهٔ بیمه‌ای هریک از بیمه‌گذاران در شرکت بیمه است. این متغیر به این دلیل انتخاب شده است که انتظار می‌رود احتمال ارتکاب تقلب توسط بیمه‌گذارانی که سابقهٔ بیمه‌ای بیشتری در یک شرکت بیمه دارند، کمتر باشد. بنابراین یک رابطهٔ معکوس بین این متغیر و متغیر وابسته وجود دارد. دومین متغیر مستقل، تعداد ادعاهای خسارت بیمه‌گذاران در طول دورهٔ سابقهٔ بیمه است. تعداد ادعاهای خسارت بیشتر توسط یک بیمه‌گذار می‌تواند به این معنا باشد که بیمه‌گذار از بیمه‌نامه به‌منظور مقاصد سودجویانه استفاده کرده باشد. ازاین‌رو بین این متغیر و متغیر وابسته یک رابطهٔ مستقیم وجود خواهد داشت. سومین متغیر مستقل، فاصلهٔ زمانی بین وقوع حادثه تا اعلام خسارت به شرکت بیمه از سوی بیمه‌گذار است. فرض شده است که هرچه این فاصلهٔ زمانی طولانی‌تر باشد، احتمال تقلب افزایش خواهد یافت. بنابراین یک رابطهٔ مستقیم بین این متغیر و متغیر وابسته وجود خواهد داشت. چهارمین متغیر مستقل، وضعیت کروکی خسارت رخ داده است. مقدار ۱ برای این متغیر به معنی نداشتن کروکی و مقدار ۰ به معنی داشتن

کروکی است. این متغیر به‌این‌دلیل انتخاب شده است که با حضور پلیس در صحنهٔ حادثه، شانس تقلب از قبیل صحنه‌سازی کاهش می‌یابد. پنجمین متغیر مستقل، جانی یا مالی‌بودن خسارت است. مقدار ۱ برای این متغیر به معنی مالی‌بودن خسارت و مقدار ۰ به معنی جانی بودن آن است. علت انتخاب این متغیر در مدل آن است که بنا به تجربه کارشناسان، فراوانی تقلب در خسارت‌های مالی نسبت‌به خسارت‌های جانی بیشتر بوده است. ازآنجایی‌که شرکت‌های بیمه در پرداخت خسارت‌های با مبالغ بالا حساسیت بیشتری داشته و بررسی‌های بیشتری نسبت‌به علل وقوع حادثه انجام می‌دهند، این‌گونه به‌نظرمی‌رسد که با افزایش مبلغ خسارت، احتمال تقلب کاهش می‌یابد، بنابراین مبلغ خسارت به‌عنوان ششمین متغیر مستقل درنظرگرفته‌شده‌است.

**۲-۶. روش رگرسیون لجستیک**

زمانی که متغیر وابسته، متغیری کیفی با دو سطح باشد، مدل‌های رگرسیون معمولی قابل استفاده نیستند. در این‌گونه موارد معمولاً از رگرسیون لجستیک استفاده می‌شود. مدل رگرسیون لجستیک به‌این‌صورت تعریف می‌شود:

$$\text{logit}(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + ... + \beta_d X_d$$

که در این مدل $X_1,...,X_d$ متغیرهای مستقل و $p$ احتمال مشاهده مقدار ۱ برای متغیر وابسته به شرط مشاهده مقادیر $x_1,...,x_d$ است. ضرایب رگرسیونی در این حالت با فرض دوجمله‌ای بودن توزیع متغیر وابسته از روش حداکثر درست‌نمایی برآورد می‌شوند. باتوجه به اینکه در این تحقیق متغیر وابسته (وضعیت پروندهٔ خسارت) یک متغیر دو سطحی است، از رگرسیون لجستیک برای تشخیص جعلی یا غیرجعلی بودن پرونده‌های خسارت استفاده شده است. بااستفاده از رگرسیون لجستیک پیش‌رو متغیرهایی که نقش مهم‌تری در تعیین وضعیت پروندهٔ خسارت داشته‌اند، شناسایی و وارد مدل شده‌اند. در گام اول مبلغ کل پروندهٔ خسارت ($X_6$) و مقدار ثابت در مدل قرار گرفته‌اند. در گام‌های دوم و سوم به ترتیب متغیرهای فاصلهٔ زمانی وقوع حادثه تا

اعلام خسارت ($X_3$) و نوع خسارت (جانی یا مالی‌بودن خسارت) ($X_5$) به مدل افزوده شده‌اند. شاخص‌های لگاریتم درستنمایی، مربع R کاکس و سل و مربع R، ناجل کرک در جدول ۱، معنی‌داری مدل را در هریک از گام‌ها نشان می‌دهند. کاهش مقدار لگاریتم درستنمایی در هر گام گویای این مطلب است که متغیر مستقل واردشده باعث بهبود مدل شده است. مقادیر مربع R کاکس وسل و مربع R ناجل کرک درصد تغییرات متغیر وابسته که توسط مدل بیان می‌شود را نشان می‌دهند.

**جدول ۱. شاخص‌های معنی‌داری مدل**

| گام | لگاریتم درست‌نمایی | مربع R کاسل وسل | مربع R ناجل کرک |
|---|---|---|---|
| ۱ | ۶۷/۸۰۹ | ۰/۳۵۹ | ۰/۴۷۸ |
| ۲ | ۵۸/۰۰۰ | ۰/۴۴۱ | ۰/۵۸۷ |
| ۳ | ۴۷/۳۲۳ | ۰/۵۱۸ | ۰/۶۹۰ |

باتوجه به مقادیر p- مقدار ضرایب مدل در گام سوم (پیوست۱)، مدل رگرسیون لجستیک به شکل زیر خواهد بود:

$$\text{logit}(Y) = -۶/۸۳۲ + ۱/۵۶X_3 + ۴/۶۲۶X_5 + ۶/۸۴۵X_4$$

که در آن Y متغیر وضعیت پروندهٔ خسارت است. لازم به ذکر است مقادیر متغیرهای فاصلهٔ زمانی وقوع حادثه تا اعلام خسارت و مبلغ کل پرونده خسارت، قبل از ورود به مدل با تقسیم بر انحراف معیارشان استاندارد شده‌اند. با به‌کارگیری این مدل بر روی داده‌های اولیه، دقت مدل سنجیده شده است. همان‌طور که در جدول ۲، مشاهده می‌شود دقت مدل در شناسایی پرونده‌های خسارت جعلی، ۸۸/۳ درصد و در شناسایی پرونده‌های غیرجعلی ۸۸/۹ درصد بوده است. ضمن اینکه دقت کلی مدل در شناسایی صحیح جعلی یا غیرجعلی‌بودن هر پرونده خسارت برابر ۸۶/۱ درصد است.

**جدول ۲. دقت مدل در شناسایی وضعیت پرونده‌های خسارت بااستفاده از رگرسیون لجستیک**

| مشاهده‌شده | | برآوردشده | | |
|---|---|---|---|---|
| | | وضعیت پرونده | | درصد صحیح |
| | | جعلی | غیرجعلی | |
| وضعیت پرونده | جعلی | ۳۰ | ۶ | ۸۸/۳ |
| | غیرجعلی | ۴ | ۳۲ | ۸۸/۹ |
| کل | | | | ۸۶/۱ |

### ۳-۶. روش بیز ساده

بیز ساده، شکل بسیار مقدماتی از مدل احتمال بیزی است. احتمال رخداد هریک از نتایج نهایی، براساس احتمالات رخداد متغیرهای مستقل به شرط رخداد همان نتیجه به‌دست‌می‌آید. فرض ما بر این است که احتمال رخداد هریک از متغیرهای مستقل به شرط رخداد یک نتیجهٔ نهایی خاص، مستقل از احتمال رخداد سایر متغیرهای مستقل به شرط رخداد همان نتیجه باشد. عملکرد بیز ساده دسته‌کننده[1] بر فرضیات استقلال قوی استوار است. یعنی اینکه احتمال رخداد یک صفت روی احتمال سایر صفت‌ها بی‌تأثیر است (Russell & Norvig, 2004). تئوری بیز امکان محاسبهٔ احتمال پسین را برمبنای احتمالات پیشین فراهم می‌کند. در مدل احتمال بیز اگر $h$ یک پیشامد و D مشاهدات باشد آنگاه خواهیم داشت:

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)}$$

1. Naive Bayes Classifier

که در آن P(h)احتمال رخداد h، P(D) احتمال رخداد D، $P(D|h)$ احتمال رخداد D به شرط رخداد $h$ و P(h|D) احتمال رخداد پیشامد h به شرط رخداد D است. در مواردی که مجموعه‌ای از پیشامدهای H (در اینجا جعلی و غیرجعلی‌بودن پرونده خسارت) وجود داشته باشد و بخواهیم محتمل‌ترین فرضیه را از میان آنان انتخاب کنیم، از فرضیهٔ حداکثر احتمال[1] استفاده می‌شود که رابطه آن به این شکل است:

$$\begin{aligned} h_{MAP} &= \arg\max_{h\in H} P(h|D) \\ &= \arg\max_{h\in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h\in H} P(D|h)P(h) \end{aligned}$$

حال اگر پیشامد $V$ به مجموعه‌ای از متغیرهای مستقل، وابسته باشد، آنگاه می‌توان روابط فرضیه حداکثر احتمال را به شکل زیر بیان کرد:

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j\in V} P(v_j|a_1,...,a_d) \\ &= \arg\max_{v_j\in V} \frac{P(a_1,...,a_d|v_j)P(v_j)}{P(a_1,...,a_d)} \\ &= \arg\max_{v_j\in V} P(a_1,...,a_d|v_j)P(v_j) \end{aligned}$$

که در آن $\alpha_1,...,\alpha_d$ پیشامدهای مستقل، $v$ طبقهٔ مربوطبه پیشامد وابسته و $V$ مجموعهٔ تمامی پیشامدهای وابستهٔ ممکن است. بااستفاده از داده‌های موجود و براساس روش بیز ساده، مدل شناسایی ادعاهای خسارت تقلبی در بیمهٔ اتومبیل مطابق جدول پ-۲ (پیوست پ، جدول پ-۲)، به‌دست‌آمده‌است.

با اعمال این مدل بر روی داده‌های اولیه، دسته‌بندی زیر جهت بررسی دقت مدل به‌دست‌آمده‌است.

1. Maximum A Posteriori hypothesis (MAP)

**جدول ۳. دقت مدل در شناسایی وضعیت پرونده‌های خسارت بااستفاده از مدل بیز ساده**

| مشاهده شده | | برآورد شده | | |
|---|---|---|---|---|
| | | وضعیت پرونده | | درصد صحیح |
| | | جعلی | غیرجعلی | |
| وضعیت پرونده | جعلی | ۳۲ | ۴ | ۸۸/۸۹ |
| | غیرجعلی | ۳ | ۳۳ | ۹۱/۶۷ |
| کل | | | | ۹۰/۲۸ |

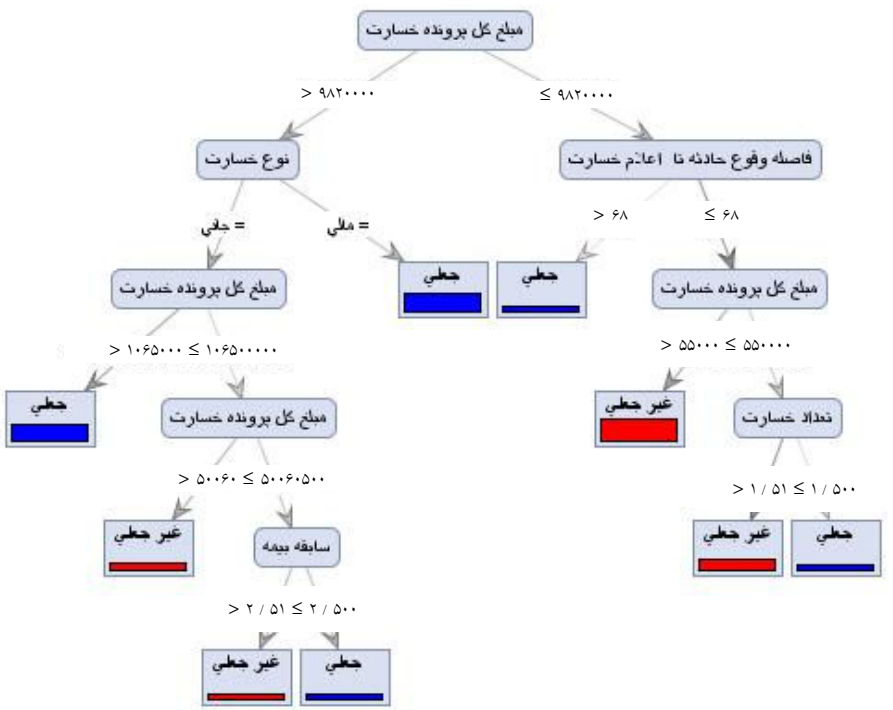
همان‌طور که در جدول ۳ مشاهده می‌شود، دقت مدل در شناسایی پرونده‌های خسارت جعلی، ۸۸/۸۹ درصد و در شناسایی پرونده‌های غیرجعلی ۹۱/۶۷ درصد بوده است. ضمن اینکه دقت مدل در شناسایی صحیح جعلی یا غیرجعلی‌بودن هر پروندهٔ خسارت برابر ۹۰/۲۸ درصد است.

### ۴-۶. درخت تصمیم

درخت تصمیم از ابزارهای داده‌کاوی است که در رده‌بندی داده‌های کیفی استفاده می‌شود. در درخت تصمیم، درخت کلی به‌وسیله خردکردن داده‌ها به گره‌هایی ساخته می‌شود که مقادیری از متغیرها را در خود جای می‌دهند. با ایجاد درخت تصمیم براساس داده‌های پیشین که رده آنها معلوم است، می‌توان داده‌های جدید را دسته‌بندی کرد. درخت تصمیم دارای قابلیت فهم بالا و سرعت مناسب در یادگیری الگو بوده و می‌توان از آن برای کشف تقلب در شرکت‌های بیمه استفاده کرد. هدف از ایجاد درخت تصمیم در این تحقیق، طبقه‌بندی داده‌های خسارت موجود به‌منظور تصمیم‌گیری در مورد جعلی یا غیرجعلی‌بودن پرونده‌های خسارت جدید در بیمهٔ اتومبیل است. معیارهای مختلفی برای تعیین صفتی که خردکردن داده‌ها باید براساس

آن انجام شود، وجود دارد که از آن جمله می‌توان به معیارهای بهرهٔ اطلاعاتی[1]، نسبت بهره[2] و شاخص جینی[3] اشاره کرد. در این تحقیق با کمک نـرم‌افـزار RapidMiner و بااستفاده از معیار بهرهٔ اطلاعاتی، درخت تصمیم‌گیری رسم شـده اسـت (نمـودار ۲). باتوجه به درخت ایجادشده، متغیر مبلغ کل پروندهٔ خسارت به‌عنوان اولین و مهم‌ترین عامل در بررسی جعلی یا غیرجعلی بودن ادعای خسارت درنظرگرفته‌می‌شود.

**نمودار۲. درخت تصمیم مطابق معیار بهره اطلاعاتی**




باتوجه به درخت ایجادشده، متغیر مبلغ کل پروندهٔ خسارت به‌عنوان اولین و مهم‌ترین عامل در بررسی جعلی یا غیرجعلی بودن ادعای خسارت درنظرگرفته

1. Information Gain
2. Gain Ratio
3. Gini Index

می‌شود. این متغیر در ریشهٔ درخت، به دو شاخهٔ مقادیر کمتر و بیشتر از ۹٫۸۲۰٫۰۰۰ ریال تقسیم می‌شود. در سطح دوم درخت، دو پارامتر نوع خسارت و فاصلهٔ زمان وقوع حادثه تا اعلام خسارت قرار دارند که هریک به‌ترتیب براساس مقادیر مالی و جانی و مقادیر کمتر و بیشتر از ۶۸ روز به دو شاخه تقسیم می‌شوند و به‌همین‌ترتیب، درخت حاصل تا ۵ سطح ادامه می‌یابد. براساس درخت حاصل می‌توان قوانینی به‌صورت اگر-آنگاه به شرح زیر استخراج نمود:

– اگر مبلغ کل پروندهٔ خسارت، بزرگ‌تر از ۹٫۸۲۰٫۰۰۰ ریال باشد و نوع خسارت، مالی باشد، مورد خسارت جعلی خواهد بود.

– اگر مبلغ کل پروندهٔ خسارت، کوچک‌تر یا مساوی با ۹٫۸۲۰٫۰۰۰ ریال باشد و فاصلهٔ بین زمان وقوع حادثه تا زمان اعلام وقوع حادثه بیشتر از ۶۸ روز باشد، مورد خسارت جعلی خواهد بود.

– اگر مبلغ کل پروندهٔ خسارت، بزرگ‌تر از ۹٫۸۲۰٫۰۰۰ ریال باشد، نوع خسارت، جانی باشد، مبلغ کل پروندهٔ خسارت، کوچک‌تر مساوی با ۵۰٫۰۶۰٫۰۰۰ ریال باشد و سابقهٔ بیمه‌ای فرد بیمه‌گذار کوچک‌تر یا مساوی با ۲/۵ سال باشد، مورد خسارت جعلی خواهد بود.

– اگر مبلغ کل پروندهٔ خسارت، کوچک‌تر یا مساوی با ۹٫۸۲۰٫۰۰۰ ریال باشد، فاصلهٔ بین زمان وقوع حادثه تا زمان اعلام وقوع حادثه کوچک‌تر یا مساوی با ۶۸ روز باشد، مبلغ کل پروندهٔ خسارت، کوچک‌تر مساوی با ۵۵۰٫۰۰۰ ریال باشد و تعداد خسارت‌های بیمه‌گذار بزرگ‌تر از ۱/۵ باشد، مورد خسارت جعلی نخواهد بود.

– سایر قوانین نیز به‌همین‌ترتیب قابل استخراج است.

باتوجه به دخالت عامل انسانی در پرونده‌های تقلب، بدیهی است که موضوع پرونده‌های جعلی تنها با استناد به نتایج حاصل از مدل‌های کمی، قابل بررسی نیستند. برای کاربردی شدن این قوانین و اعمال در دنیای واقعی، نیاز به بررسی‌های بیشتر و

استفاده از تجارب متخصصان بیمه الزامی است. در ادامه با اعمال این مدل بر روی داده‌های اولیه، نتایج زیر جهت بررسی دقت مدل به‌دست‌آمده است.

**جدول ۴. دقت مدل در شناسایی وضعیت پرونده‌های خسارت بااستفاده از درخت تصمیم‌گیری**

| مشاهده شده | | برآوردشده | | |
|---|---|---|---|---|
| | | وضعیت پرونده | | درصد صحیح |
| | | جعلی | غیرجعلی | |
| وضعیت پرونده | جعلی | ۳۵ | ۱ | ۹۷/۲ |
| | غیرجعلی | ۷ | ۲۹ | ۸۰/۶ |
| کل | | | | ۸۸/۹ |

همان‌طور که در جدول ۴ مشاهده می‌شود، دقت مدل در شناسایی پرونده‌های خسارت جعلی، ۹۷/۲ درصد و در شناسایی پرونده‌های غیرجعلی۸۰/۶ درصد بوده است. ضمن اینکه دقت مدل در شناسایی صحیح جعلی یا غیرجعلی‌بودن هر پروندهٔ خسارت برابر ۸۸/۹ درصد است.

## ۷. نتیجه‌گیری و پیشنهادها

در این مقاله سه روش داده‌کاوی رگرسیون لجستیک، بیز ساده و درخت تصمیم برای ساخت مدل‌هایی جهت شناسایی ادعاهای خسارت تقلبی در بیمهٔ اتومبیل معرفی شدند. در ادامه این روش‌ها بر روی داده‌های واقعی آزمایش و کارایی هر روش سنجیده شد. روش بیز ساده با دقت ۹۰/۲۸ درصد در شناسایی صحیح جعلی یا غیرجعلی‌بودن پرونده‌های خسارت بهترین کارایی را در مقایسه با دو روش درخت تصمیم با دقت کلی ۸۸/۹ درصد و رگرسیون لجستیک با دقت کلی ۸۶/۱ درصد داشت. البته باید به این نکته توجه داشت که در مدل بیز ساده برای تشخیص جعلی یا غیرجعلی‌بودن هر خسارت، شش متغیر و در مدل درخت تصمیم، پنج متغیر حضور دارند. این در حالی است که تصمیم‌گیری در مدل رگرسیون لجستیک برمبنای

سه متغیری است که بیشترین همبستگی را با متغیر وابسته دارند. در نتیجه استفاده از مدل رگرسیون لجستیک نسبت‌به مدل بیز ساده به محاسبات کمتری نیاز دارد. مدل درخت تصمیم باتوجه به شهودی‌بودن آن در زمان‌هایی که نیاز به سرعت در تصمیم‌گیری باشد، قابلیت بهتری دارد. علاوه‌بر متغیرهای استفاده‌شده در این تحقیق، متغیرهای دیگری نیز وجود دارند که به شرط ثبت اطلاعات آنها در خسارت‌های پیشین، می‌توانند با ورود به مدل به تشخیص بهتر تقلب در خسارت‌های جعلی کمک کنند. برای مثال می‌توان به ساعت وقوع حادثه، داخل یا خارج از شهربودن محل حادثه، سن راننده و تعداد سرنشینان خودرو در زمان حادثه اشاره کرد. باتوجه به هزینه‌های هنگفتی که سالانه شرکت‌های بیمه بابت خسارت‌های جعلی متحمل می‌شوند و همچنین مطالعات اندک انجام‌گرفته در زمینهٔ راه‌های کشف و کاهش تقلب، این تحقیق می‌تواند مبنایی علمی جهت کشف تقلب، پیش روی مدیران بیمه و پژوهشگران علاقمند به مطالعه بگذارد. باتوجه به دقت خوبی که این روش‌ها در شناسایی صحیح جعلی یا غیرجعلی‌بودن پرونده‌های خسارت دارند، پیشنهاد می‌گردد از این روش‌ها برای بررسی پرونده‌های جعلی در سایر رشته‌های بیمه‌ای نیز استفاده گردد. علاوه‌براین باتوجه به دخالت عامل انسانی در پدیدهٔ تقلب در صنعت بیمه، بررسی‌های بیشتر و استفاده از نظرات خبرگان و متخصصان بیمه به کاربردی‌تر شدن نتایج این تحقیق خواهد انجامید.

## پیوست‌ها

### پیوست۱. جدول متغیرهای مورد استفاده در مدل

| نام اختصاری متغیر | نوع متغیر | شرح متغیر |
|---|---|---|
| Y | وابسته | وضعیت پرونده (جعلی یا غیرجعلی‌بودن یک پرونده) |
| $X_1$ | مستقل | سابقه بیمه‌ای بیمه‌گذاران |
| $X_2$ | مستقل | تعداد ادعاهای خسارت بیمه‌گذاران در طول دورهٔ سابقهٔ بیمه |
| $X_3$ | مستقل | فاصله زمانی بین وقوع حادثه تا اعلام خسارت |
| $X_4$ | مستقل | وضعیت کروکی خسارت رخ‌داده |
| $X_5$ | مستقل | جانی یا مالی‌بودن خسارت |
| $X_6$ | مستقل | مبلغ خسارت |

### پیوست ۲. ضرایب متغیرهای مدل و مقادیر P مقدار متناظر

ضرایب هریک از متغیرهای واردشده در مدل رگرسیون لجستیک و همچنین مقدار ثابت مدل همراه با مقادیر p مقدار آنها در جدول نشان داده شده است.

| | | B | S.E. | Sig. |
|---|---|---|---|---|
| گام ۱ | $X_6$ | ۳/۴۵۴ | ۱/۱۱۷ | ۰/۰۰۲ |
| | ثابت مدل | -۱/۲۸۴ | ۰/۳۹۱ | ۰/۰۰۱ |
| گام ۲ | $X_3$ | ۱/۴۴۴ | ۰/۶۹۹ | ۰/۰۳۹ |
| | $X_6$ | ۳/۰۵۴ | ۱/۰۷۱ | ۰/۰۰۴ |
| | ثابت مدل | -۱/۸۹۵ | ۰/۴۸۴ | ۰/۰۰۰ |
| گام ۳ | X3 | ۱/۵۶۰ | ۰/۵۴۶ | ۰/۰۰۴ |
| | X5 | ۴/۶۲۶ | ۱/۶۷۴ | ۰/۰۰۶ |
| | X6 | ۶/۸۴۵ | ۱/۸۹۰ | ۰/۰۰۰ |
| | ثابت مدل | -۶/۸۳۲ | ۱/۹۶۷ | ۰/۰۰۱ |

**پیوست۳. احتمالات جعلی و غیرجعلی‌بودن پروندهٔ خسارت‌ها به تفکیک هریک از متغیرها**

| وضعیت پرونده | | مبلغ کل پروندهٔ خسارت* | | | نوع خسارت | | | وضعیت کروکی | | | فاصله زمانی وقوع تا اعلام ** | | | تعداد خسارت | | | سابقهٔ بیمه | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| غیرجعلی | جعلی | | غیرجعلی | جعلی | | غیرجعلی | جعلی | | غیرجعلی | جعلی | | غیرجعلی | جعلی | | غیرجعلی | جعلی | سال | غیرجعلی | جعلی |
| ۰/۵ | ۰/۵ | ۰ | ۰/۰۶ | ۰/۶۹ | جانی (۰) | ۰/۱۷ | ۰/۴۴ | دارد (۰) | ۰/۱۱ | ۰/۰۶ | ۱ | ۰/۹۴ | ۰/۴۷ | ۱ | ۰/۷۸ | ۰/۷۸ | ۱ | ۰/۲۹ | ۰/۳۱ |
| | ۱ | ۱ | ۰/۹۴ | ۰/۳۱ | مالی(۱) | ۰/۸۳ | ۰/۵۶ | ندارد (۱) | ۰/۸۹ | ۰/۹۴ | ۲ | ۰/۰۳ | ۰/۳۱ | ۲ | ۰/۱۷ | ۰/۱۹ | ۲ | ۰/۴۱ | ۰/۴۴ |
| | | | | | | | | | | | ۳ | ۰/۰۰ | ۰/۰۶ | ۳ | ۰/۰۶ | ۰/۰۳ | ۳ | ۰/۲۹ | ۰/۱۹ |
| | | | | | | | | | | | ۴ | ۰/۰۰ | ۰/۰۶ | | | | ۴ | ۰/۰۰ | ۰/۰۳ |
| | | | | | | | | | | | ۵ | ۰/۰۰ | ۰/۰۶ | | | | ۵ | ۰/۰۰ | ۰/۰۳ |
| | | | | | | | | | | | ۶ | ۰/۰۰ | ۰/۰۰ | | | | ۶ | ۰/۰۳ | ۰/۰۰ |
| | | | | | | | | | | | ۷ | ۰/۰۰ | ۰/۰۳ | | | | ۷ | ۰/۰۳ | ۰/۰۰ |
| | | | | | | | | | | | ۸ | ۰/۰۳ | ۰/۰۰ | | | | | | |
| | | | | | | | | | | | ۹ | ۰/۰۰ | ۰/۰۰ | | | | | | |
| | | | | | | | | | | | ۱۰ | ۰/۰۰ | ۰/۰۳ | | | | | | |

* مقدار ۰ برای مقادیر بزرگ‌تر از میانگین

مقدار ۱ برای مقادیر کمتر یا مساوی میانگین

** ۱: کمتر از ۵۰ روز، ۲: ۵۱ تا ۱۰۰ روز، ۳: ۱۰۱ تا ۱۵۰ روز، ...، ۱۰: بیشتر از ۴۵۱ روز

تذکر: بیشترشدن جمع احتمالات از ۱، ناشی از خطای گردکردن است.

## منابع


۱. راه‌چمنی، ابوالقاسم ۱۳۸۵، 'تقلب و کلاهبرداری تهدید همیشگی صنعت بیمه'، *فصلنامه آسیا*، ش۳۸، صص۹-۱۶.

2. Artis, M, Ayuso, M & Guillen, M 2002, 'Detection of automobile insurance fraud with discrete choice models and misclassified claims', *Journal of Risk and Insurance,* pp. 325-40.

3. Belhadji, DB & Dionne, G 1997, 'development of an expert system for the automatic detection of automobile insurance fraud', *Risk Management Chair, HEC-Montreal*.

4. Bolton, RJ & Hand, DJ 2002, 'Statistical fraud detection: a review', *Statistical Science*, vol. 17, no. 3, pp. 235–55.

5. Brockett, PL, Xia, X & Derrig, RA 1998, 'Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud', *The J. of Risk and Insurance,* pp. 245-74.

6. Cummins, JD & Tennyson, S 1992, 'Controlling automobile insurance costs', *Journal of Economic Perspectives*, pp. 95-115.
7. Derrig, RA 2002, 'Insurance fraud', *Journal of Risk and Insurance*, vol. 69, pp. 271-87.
8. Derrig, R, Johnston, D & Sprinkel, E 2006, 'Auto insurance fraud: measurements and efforts to combat it', *Risk Management and Insurance Review*, vol. 9, pp. 109-30.
9. Derrig, RA & Ostazewski, KM 1995, 'Fuzzy techniques of pattern recognition in risk and claim classification', *The J. of Risk and Insurance*, pp. 447-82.
10. Doig, A, Jones, B & Wait, B 1999, 'The insurance industry response to fraud', *Security Journal*, vol. 12, pp.19-30.
11. Ghezzi, SG 1983, 'A private network of social control: insurance investigation units', *Social Problems,* vol. 30, pp.521–30.
12. Gill, KM, Woolley, KA & Gill, M 1994, 'Insurance fraud: the business as a victim', *Crime at Work,* M. Gill (Ed.), Vol. 1, Leicester: Perpetuity Press.
13.Michalski, RS, Bratko, I & Kubat, M 1998, *Machine learning and data mining - methods and applications*, John Wiley & Sons Ltd., 1st ed.
14. Morley, N, Ball, L & Ormerod, T 2006, 'How the detection of insurance fraud succeeds and fails', *Psychology, Crime & Law*, vol. 12, pp.163-80.
15. Phua, C, Alahakoon, D & Lee, V 2004, 'Minority report in fraud detection: classification of skewed data', *Sigkdd Explorations*, vol. 6, no. 1, pp. 50-9.
16. Russel, SJ & Norvig, P 2004, *Artificial intelligence: a modern approach*, Pearson Education International, USA, 2nd ed.
17.Šubelj, L, Furlan, Š & Bajec, M 2010, 'An expert system for detecting automobile insurance fraud using social network analysis', *Expert Systems with Applications: An International Journal*, vol. 38, pp. 1039-52.
18. Tennyson, S & Salsas, P 2002, 'claims auditing in automobile insurance: fraud detection & deterrence objective', *Journal of Risk and insurance,* vol. 69, no. 3, pp. 289-308.
19. Terisa, R 2010, *Improving the defense lines: the future of fraud detection in the insurance industry (with fraud risk models, text mining, and social networks)*, Paper Presented in the SAS Global forum, Washington.
20.Viaene, S, Ayuso, M, Guillen, M, Gheel, DV & Dedene, G 2005, 'Strategies for detecting fraudulent claims in the automobile insurance industry', *European Journal of Operational Research*, vol. 176, pp. 565–83.

21. Viaene, S, Ayuso, M, Guillen, M, Gheel, D & Dedene, G, 2007, 'Strategies for advertising fraudulent claims in the automobile insurance industry', *European Journal of Operational Research,* vol. 176, pp. 565-83.
22. Viaene, S & Dedene, G 2004, 'Insurance fraud: Issues and challenges', *Geneva Papers on Risk and Insurance Issues and Practice*, vol. 29, pp. 313-33.
23.Weisberg, HI & Derrig, RA 1993, 'Quantitative methods for detecting fraudulent aumobile bodily insurance claims', *AIB Cost Containment/Fraud Filing,* pp. 49-82.
24. Wilson, HJ 2003, 'An analytical approach to detecting insurance fraud using logistic regression', *Journal of Finance and Accountancy*, pp. 1-15
25. Witten, I & Frank, E 2000, *Data mining: practical machine learning tools and techniques with Java implementations*, San Francisco, Calif.: Morgen Kaufmann, 2nd ed.