مقاله پژوهشی

متغیر جایگزین، رویکردی نوین در جهت افزایش کیفیت کشف تقلبات بیمه‌های اتومبیل با استفاده از الگوریتم‌های با نظارت

فربد خانی‌زاده
عضو هیئت علمی پژوهشکده بیمه، تهران، ایران. khanizadeh@irc.ac.ir

مریم اثنی‌عشری*
(نویسنده مسئول)، عضو هیئت علمی پژوهشکده بیمه، تهران، ایران. esnaashari@irc.ac.ir

فرزان خامسیان
عضو هیئت علمی پژوهشکده بیمه، تهران، ایران. khamesian@irc.ac.ir

آزاده بهادر
راهبر میز بیمه‌های اتومبیل، پژوهشکده بیمه، تهران، ایران. bahador@irc.ac.ir

چکیده: در سال‌های اخیر تمایل صنعت بیمه به تجهیز شرکت‌ها به سیستم‌های کشف تقلب افزایش یافته است. با توجه به هزینه زیادی که اینگونه پرونده‌ها به صنعت وارد می‌کند، الگوریتم‌های کشف و شناسایی تقلب می‌بایست بخش جدایی‌ناپذیری از شرکت‌های بیمه باشند. لیکن مشکل اساسی، کیفیت خروجی سیستم‌های کشف تقلب است. از طرفی الگوریتم‌های با نظارت نسبت به الگوریتم‌های بدون نظارت دقت بالاتری دارند. از طرف دیگر در حوزه کشف تقلب، داده‌های برچسب‌گذاری شده محدودند و بنابراین به‌کارگیری الگوریتم‌های با نظارت، دقت و کیفیت آنها با چالش مواجه می‌شود. در این مقاله برای رفع این چالش، با استفاده از رویکرد "متغیر جایگزین"، از متغیر دیگری که مقادیر آن در دسترس بوده و شاخص مناسبی برای پرونده‌های مشکوک می‌باشد استفاده شده است. این رویکرد، باعث بهبود کارایی و کیفیت سیستم شده و به شرکت‌های بیمه این امکان را می‌دهد که با اطمینان بیشتر و خطای کمتر نسبت به پرونده‌های مشکوک اقدام کنند.

واژگان کلیدی: الگوریتم با نظارت، متغیر جایگزین، کشف تقلب، بیمه خودرو.

۱- مقدمه

بیمه به‌عنوان یک عنصر اساسی در استراتژی‌های مدیریت ریسک افراد، گروه‌های اجتماعی و مشاغل شناخته می‌شود. لذا حفظ شرایط پایدار مالی برای شرکت‌های بیمه جهت ارائه تعهدات بیمه‌ای مناسب به بیمه‌گذاران در شرایط بحران و وقوع حادثه بسیار حائز اهمیت است. یکی از عوامل مؤثر در تضعیف شرایط پایدار مالی شرکت‌ها، موارد کلاه‌برداری و تقلبات بیمه‌ای است. کلاهبرداری بیمه، تلاشی برای بهره‌برداری از قرارداد بیمه است. به عبارت دیگر، منظور از بیمه محافظت در برابر خطرات و ریسک‌های احتمالی است، و نه ابزاری برای ثروتمند کردن بیمه‌گذاران. اگرچه تقلبات بیمه‌ای توسط صادرکننده بیمه‌نامه نیز اتفاق می‌افتد، لیکن اکثر موارد مربوط به پرونده‌های خسارت و تلاش بیمه‌گذاران برای دریافت پول بیشتر می‌باشد. در بیمه‌های اموال و مسئولیت، یکی از رشته‌هایی که تعداد زیاد پرونده‌های تقلب را به خود اختصاص می‌دهد، حوزه بیمه‌های اتومبیل می‌باشد. عدم مقابله با این پدیده توسط شرکت‌های بیمه

Corresponding author: esnaashari@irc.ac.ir

تاریخ دریافت: ۱۴۰۰/۱۱/۱۳ تاریخ پذیرش: ۱۴۰۱/۰۵/۱۵
دوره ۱۱/ شماره ۴
صفحات ۴۱۳-۴۲۸

در بلند مدت، هزینه سنگینی را به بیمه‌گران وارد می‌کند. شایان ذکر است که شرکت‌ها برای جبران ضررهای مالی ایجاد شده به ناچار نرخ حق بیمه را افزایش داده که این عمل منجر به فشار مالی بر بیمه‌گذاران و متعاقباً نارضایتی آنها می‌گردد. شاید بتوان گفت، اولین قدم در راه مبارزه با کلاهبرداری و تقلبات بیمه کشف این پرونده‌ها است. لیکن فرهنگ مبارزه با تقلب امری فراتر از یک سیستم شناسایی و تشخیص خودکار پرونده‌های تقلب است. در واقع فرهنگ مبارزه با تقلب مستلزم ارتباط ساختاریافته بین بخش‌های مختلف ازجمله شرکت‌های بیمه، نهاد ناظر و نهاد قانون‌گذاری و همچنین مشارکت مدیریت ارشد، آموزش‌های آگاهی از تقلب و استانداردهای عملکرد همسو برای کارکنان مرتبط با پرونده‌های خسارت و بیمه‌گری است.

موضوع مهم دیگر در حوزه تقلب، نگرانی شرکت‌های بیمه از موارد و پرونده‌هایی است که به اشتباه به‌عنوان تقلب شناسایی می‌شوند (مثبت کاذب). در واقع با توجه به آنکه بیمه از جمله خدمات و محصولاتی است که می‌بایست در زمان بحران و وقوع خسارت رضایت مشتریان را جلب کند، برچسب تقلب به اشتباه به افرادی که پرونده خسارت آنها مشکلی ندارد می‌تواند باعث نارضایتی بیمه‌گذار و چه بسا عدم تمدید بیمه‌نامه با شرکت مزبور شود. در همین راستا ترجیح بیمه‌گران بر استفاده از آن دسته از سیستم‌های کشف تقلب است که خروجی آنها از کیفیت و دقت بالاتری برخوردار است.

در این مقاله به دلیل اهمیت این موضوع و رایج شدن تقلب در رشته بیمه‌های اتومبیل و همچنین الگوریتم‌های داده‌کاوی در شناسایی پرونده‌های تقلب، از الگوریتم‌های با نظارت یادگیری ماشین در جهت کشف پرونده‌های مشکوک به تقلب در رشته بیمه‌های شخص ثالث اتومبیل استفاده شده است. شایان ذکر است که مشکل اساسی در استفاده از الگوریتم‌های با نظارت در حوزه شناسایی تقلبات، تعداد کم اینگونه پرونده‌ها نسبت به کل پرونده‌های خسارت است. در همین راستا یکی از ابتکارات این مقاله استفاده از متغیر هدف جایگزین، جهت افزایش دقت مدل و ارتقاء کیفیت خروجی سیستم می‌باشد که در ادامه به آن خواهیم پرداخت.

۲- پیشینه پژوهش

مثلث کلاهبرداری توسط جرم‌شناس دونالد کریسی در دهه ۱۹۵۰ ایجاد شد. فرضیه وی بر اساس مصاحبه با زندانیان بوده است. کریسی در این تئوری مطرح می‌نماید که سه عنصر عقلانیت، فشار و فرصت، عناصر تأثیرگذار برای ارتکاب کلاهبرداری توسط یک فرد هستند و هنگامی که هر سه ضلع مثلث در زندگی یک فرد وجود داشته باشد بسیار محتمل است که او مرتکب تقلب شود (یا هم اکنون در حال انجام تقلب باشد). شرکت‌های بیمه قادر به کنترل عقلانیت یا فشار نیستند؛ آنچه آنها می‌توانند کنترل کنند مولفه فرصت است که این امر با استفاده از روندهای نظارتی و کنترلی مناسب و یا نصب فناوری‌های کشف تقلب میسر می‌شود. شناسایی متغیرهای تأثیرگذار در کشف تقلب نیز در مقالات و پژوهش‌های متعددی مورد بررسی و مطالعه قرار گرفته است. شیوه‌ها و روش‌های کشف پرونده‌های مشکوک را نیز می‌توان به دو دسته کلی روش‌های آماری، الگوریتم‌های یادگیری ماشین تقسیم کرد. در رویکرد نخست از روش‌های کلاسیک آماری و توابع توزیع در بررسی پرونده‌ها استفاده می‌شود [۴] و[۱۴]. هنگام استفاده از داده‌کاوی و الگوریتم‌های یادگیری ماشین از دو شیوه‌ی با نظارت و بدون نظارت می‌توان بهره برد. با توجه به ماهیت داده‌های مربوط به تقلب، استفاده از الگوریتم‌های بدون نظارت از عمومیت بیشتری برخوردار است [۳]، [۱۰]، [۱۳]، [۱]، [۱۱]. به‌طور خاص الگوریتم‌های بدون نظارتی که در تحقیقات استفاده شده‌اند عبارتند از: رتبه بندی اسپکترال [۱۵ و ۲۸]؛ شناسایی نمونه‌های نامتعارف [۱۶]، [۶] و [۲]، خوشه‌بندی [۱۷]، [۲۰] و [۲۷]. در برخی از مقالات با دسترسی به داده‌های برچسب‌گذاری شده این امکان فراهم گردید که از الگوریتم‌های با نظارت استفاده شود. برای مثال زو و همکاران[1] (۲۰۱۲) و سهین و همکاران[2] (۲۰۱۳) از درخت تصمیم و دیپا و داناپال[3] (۲۰۱۲) و گیامفی و عبدولای[4] (۲۰۱۸) در

1 .Zou, et al

2 Sahin, et al.

3 Dheepa & Dhanapal

4. Gyamfi & Abdulai

پژوهش‌های خود از ماشین بردار پشتیبان استفاده نمودند. به علاوه، سهین و دومان (۲۰۱۱)، روشین و همکاران[5] (۲۰۱۷) و چن و لای[6] (۲۰۲۱) در مطالعات خود الگوریتم‌های رگرسیون لوجستیک و شبکه‌های عصبی را استفاده نمودند.

به‌طور کلی تقلبات را می‌توان بر اساس سه معیار مهم «منبع ایجاد»، «شدت تقلب» و «زمان وقوع تقلب» طبقه‌بندی نمود. بر همین اساس تقلبات از نظر منبع ایجاد آن به دو دسته: داخلی و خارجی؛ از نظر زمان وقوع، به چهار دسته: زمان بیمه‌گری و صدور بیمه‌نامه، زمان وقوع حادثه، هنگام پذیرش در بیمارستان و هنگام ارائه مدارک قضایی؛ و از نظر شدت تقلب به دو دسته: نرم و سخت تقسیم می‌شوند [۸] و [۲۵]. بر اساس این مطالعات و بسیاری از تحقیق‌های دیگر در زمینه کشف تقلب می‌توان گفت، متغیرهای اثرگذار بر ادعاهای خسارت تقلبی در هفت متغیر اصلی شامل مشخصات زیان‌دیدگان/ مصدومان، مشخصات رانندگان مقصر حادثه، مشخصات بیمه‌گذار، مشخصات خودرو، مشخصات بیمه‌نامه، مشخصات حادثه و مشخصات درمان خلاصه می‌شوند [۵]، [۱۸]، [۲۳]، [۲۴]، [۲۶].

۳- روش پژوهش

همانطور که قبلاً اشاره شد یکی از چالش‌های موجود در برخورد با بررسی داده‌های تقلب، عدم وجود داده‌های برچسب‌گذاری شده می‌باشد. این امر باعث می‌شود که امکان استفاده از الگوریتم‌های با نظارت یادگیری ماشین امکان‌پذیر نباشد. در این پژوهش برای غلبه بر این مشکل، متغیر دیگری (متغیر هدف جایگزین) انتخاب شده و از نتایج آن به‌عنوان راهنمایی برای تشخیص متغیرهای مشکوک استفاده شده است. در همین راستا در این مقاله متغیر مدت زمان بیمه‌نامه به‌عنوان شاخصی برای تشخیص تقلب در بیمه‌های خودرو انتخاب شده است. همانطور که می‌دانیم بیمه‌نامه‌های شخص ثالث به طور معمول به صورت یک ساله صادر می‌شوند و بر اساس نظرات کارشناسان بیمه، هر چه مدت زمان بیمه‌نامه کمتر باشد نیاز به بررسی بیشتر دارد. در داده‌هایی که در اختیار تیم نویسندگان قرار گرفت برای برخی مشاهدات موجود، زمان شروع و پایان بیمه‌نامه یکسان صادر شده است. این امر به طور طبیعی از موارد مشکوک بوده و در این مقاله به‌عنوان شاخصی برای برچسب‌گذاری داده‌ها استفاده شده است.

۳-۱- داده‌های پژوهش

متغیرها و داده‌های استفاده شده در این مقاله بر اساس ۱- نظرسنجی از خبرگان صنعت؛ ۲- بررسی تحقیقات مشابه و ۳- محدودیت‌های دسترسی به پایگاه داده استخراج شده‌اند. مهمترین شاخصه‌های شناسایی‌شده در تقلبات بر اساس نظرات خبرگان، شامل ۱۰ شاخصه کلی با عناوین مشخصات حادثه، مشخصات مصدومان/ زیان‌دیدگان، مشخصات بیمه‌نامه، مشخصات رانندگان مقصر حادثه، تناسب و تطابق، شرایط پس از وقوع خسارت، مشخصات کروکی، مشخصات خودرو، مشخصات مدارک پزشکی و اصالت مدارک است. همچنین شاخصه‌های شناسایی‌شده در تقلبات بر اساس مطالعات مشابه، شامل ۷ شاخصه کلی می‌باشد. در نهایت بر اساس مهمترین شاخصه‌های شناسایی شده از این دو منبع و متغیرهای ثبت‌شده در پایگاه داده شرکت‌های بیمه، ۱۵ ویژگی تأثیرگذار به شرح جدول ۱، انتخاب شد. در این مقاله همچنین یک متغیر هدف جایگزین با نام "مدت زمان بیمه‌نامه"، به‌عنوان متغیر وابسته استفاده شده است. جامعه آماری نیز شامل پنجاه هزار نمونه از داده‌های مربوط به خسارت رشته بیمه شخص ثالث است.

جهت افزایش کارایی مدل‌های یادگیری ماشین نیاز است که بر روی داده‌ها پیش پردازش‌های لازم انجام گیرد. شایان ذکر است مراحل پیش پردازش ثابت نبوده و بر اساس مجموعه داده‌های در اختیار تعیین می‌شوند. برای پیش‌پردازش داده‌های این پژوهش، اقدامات به شرح زیر انجام شد.

استخراج بخش‌های قابل استفاده

در مواجهه با برخی متغیرها، تمامی اطلاعات وارد شده قابل استفاده نبوده و زمانی برای استخراج بخش‌های قابل استفاده در انجام تحقیق صرف شد.

[5] Rushin, et al.

[6] Chen, J. L., & Lai

تلفیق ویژگی‌ها

برخی از متغیرها و ویژگی‌ها به تنهایی اطلاعات مؤثری جهت کشف پرونده‌های تقلب در اختیار قرار نمی‌دادند و جهت دستیابی به شاخص‌های تأثیرگذار، از تلفیق آن با متغیر یا متغیرهای دیگر استفاده کردیم. برای مثال از تفاضل تاریخ شروع و پایان بیمه‌نامه، متغیر مدت بیمه‌نامه به‌دست آمد.

جدول۱- متغیرهای مستقل در مجموعه داده

نام متغیر	کد انگلیسی متغیر
سن خودرو	PrdDte
محدوده وقوع حادثه	InCty
نوع وسیله نقلیه	MapCarTypCod
سن مقصر حادثه	Age
سن زیان‌دیده	AgeLoser
نوع گواهی‌نامه	IsLcnsFit
نوع کاربری	UsgCod
جنسیت زیان‌دیده	IsMale
جنسیت راننده مقصر حادثه	CusMaleCod
فاصله حادثه تا اعلام خسارت	Days
ساعت وقوع حادثه	Hour
نوع گروه‌بندی وسیله نقلیه	CarGrpCod
نوع خسارت زیان‌دیده	ThrLosTyp
کد نوع حادثه	AcdTypCod
کد استان	CtyNam

حذف داده‌های نویز (نوفه)

منظور از داده‌های نویز در این تحقیق متغیرهایی هستند که برای آنها در نمونه‌های مختلف مقادیر متفاوت و متناقض آورده شده است. به‌عنوان مثال، در برخی موارد خودرو پژو، یک بار به‌عنوان سواری و یک بار به‌عنوان بارکش ثبت شده بود که تا حد امکان، اینگونه اطلاعات دارای نویز از داده‌ها حذف شد.

یکسان‌سازی و ارائه کدهای عددی به تمام متغیرها

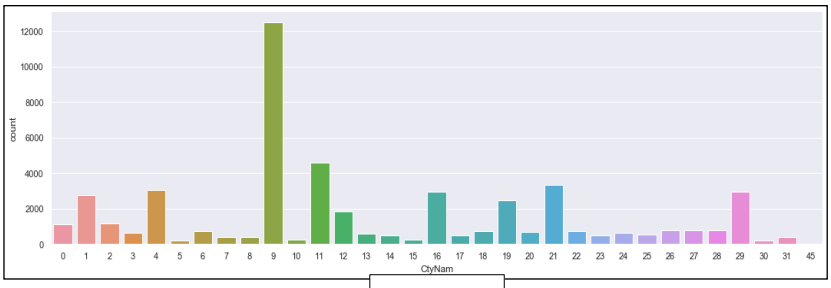
در این مرحله ابتدا ماهیت و نوع تمام متغیرها از منظر عددی یا اسمی بودن مشخص شد و سپس متغیرهی اسمی کدگذاری شدند. از میان متغیرهای مستقل موجود، سن مقصر حادثه، سن زیان‌دیده، ساعت وقوع حادثه، فاصله زمانی وقوع حادثه تا اعلام خسارت، عددی و سایر متغیرها، اسمی هستند.

گروه‌بندی متغیرهای مستقل عددی

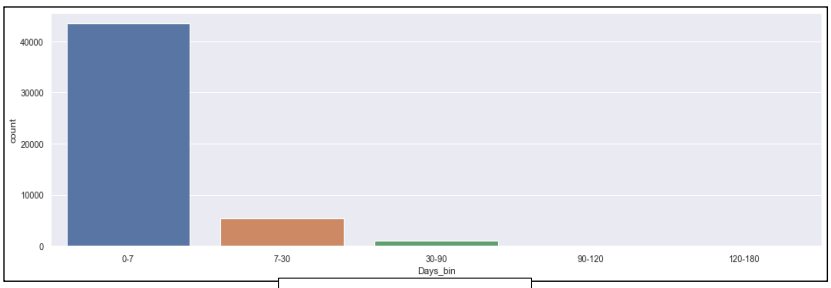
برای یکسان‌سازی متغیرها و استفاده بهینه از الگوریتم‌های یادگیری ماشین، متغیرهای عددی به گروه و دسته‌های مختلف تقسیم شدند. از این طریق داده‌های پیوسته‌ای که بررسی آن دشوار بوده به شکل گسسته درآمده و درک آنها آسان‌تر شد. به‌عنوان نمونه، متغیر پیوسته‌ای مانند سن به‌صورت دسته‌هایی با طول ۱۰ سال در نظر گرفته شد.

۴- آمار توصیفی

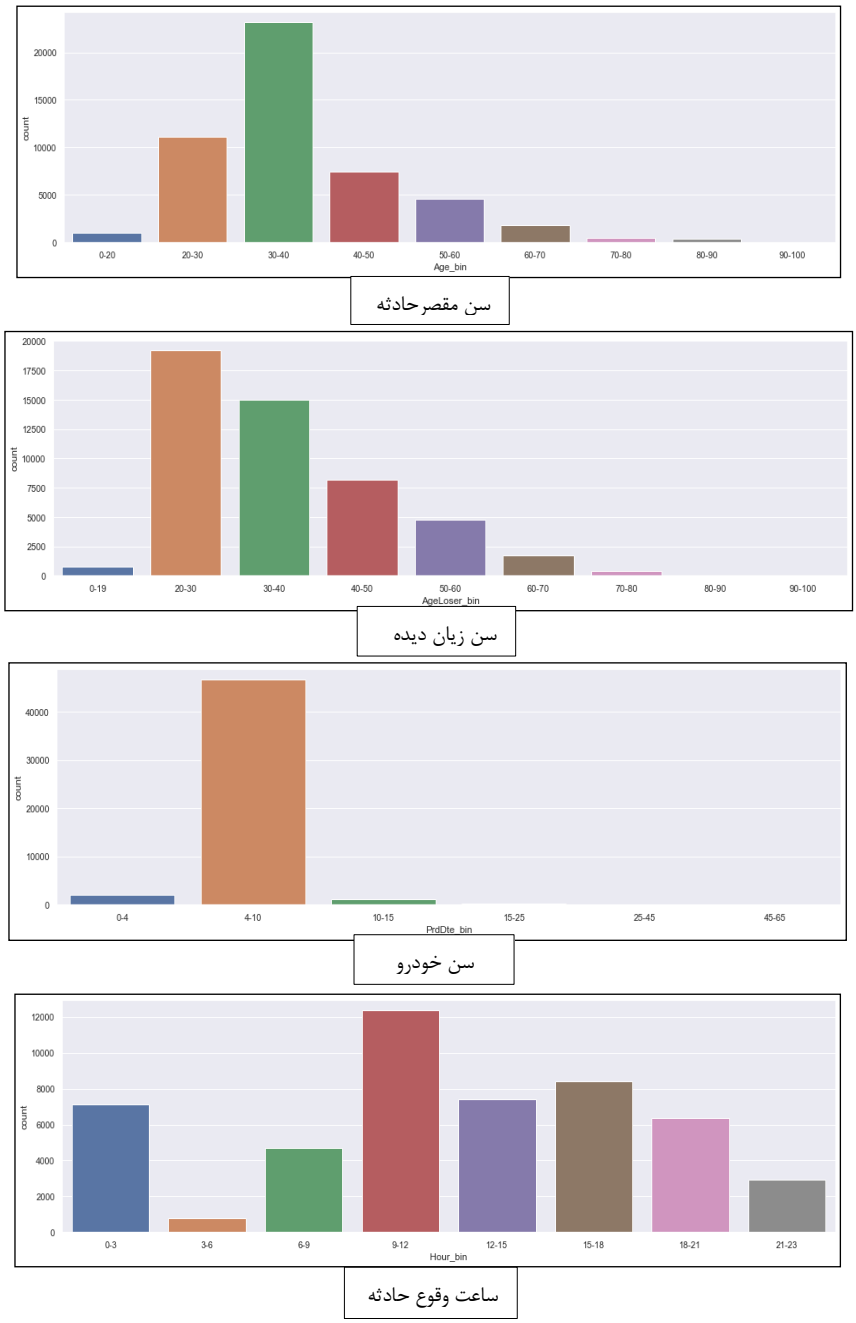
از آمار توصیفی به منظور سازمان‌دهی، خلاصه کردن و توصیف اطلاعات استفاده می‌شود. معمولاً سازمان‌دهی داده‌ها قبل از آنالیز آنها، می‌تواند منجر به آشکار شدن نکات پنهان بسیاری شود. به همین منظور در این بخش نیز، به توصیف متغیرها پرداخته شده است.



کد استان



فاصله حادثه تا اعلام خسارت



شکل ۱- آمار توصیفی مربوط به متغیرهای استفاده شده در تحلیل

در این پژوهش گروه‌بندی متغیرهای گوناگون، به شرح ذیل است:

- نوع کاربری: (۱- سواری آژانس، تاکسی، کرایه و مسافرکش شخصی درون شهری؛ ۲- سواری کرایه و مسافرکش شخصی برون شهری؛ ۳- حمل مواد منفجره؛ ۴-حمل مواد سوختی؛ ۵- تعلیم رانندگی؛ ۶- مسابقات رانندگی؛ ۷- وسایل نقلیه مخصوص جابجایی کارکنان بیمه گذار، دانش آموزان، دانشجویان؛ ۸- شخصی؛ ۹- آمبولانس؛ ۱۰- وسیله نقلیه ویژه حمل خون؛ ۱۱- وسیله نقلیه حمل وسایل رادیولوژی؛ ۱۲- آتش نشانی؛ ۱۳- سایر)

• نام نوع وسیله نقلیه: ۱- سواری؛ ۲- موتورسیکلت؛ ۳- بارکش؛ ۴- اتوکار؛ ۵- سایر (مانند کشاورزی و راه سازی).

• جنسیت زیاندیده/راننده مقصر حادثه: (۰- زن؛ ۱- مرد)

• سن زیان‌دیده/مقصر حادثه: (۰- ۰ تا ۲۰ سال؛ ۱- ۲۰ تا ۳۰ سال؛ ۲- ۳۰ تا ۴۰ سال؛ ۳- ۴۰ تا ۵۰ سال؛ ۴-۵۰ تا ۶۰ سال؛ ۵- ۶۰ تا ۷۰ سال؛ ۶- ۷۰ تا ۸۰ سال؛ ۷- ۸۰ تا ۹۰ سال؛ ۸- ۹۰ تا ۱۰۰ سال)

• فاصله حادثه تا اعلام خسارت: (۰- ۰ تا ۷ روز؛ ۱- ۱۲۰ تا ۱۸۰ روز؛ ۲- ۳۰ تا ۹۰ روز؛۳- ۷ تا ۳۰ روز؛ ۴- ۹۰ تا ۱۲۰ روز)

• ساعت وقوع حادثه: (۰-ساعت ۰ تا ۳؛ ۱-ساعت ۱۲ تا ۱۵؛ ۲- ساعت ۱۵ تا ۱۸؛ ۳- ساعت۱۸ تا ۲۱؛ ۴- ساعت ۲۱ تا ۲۳؛ ۵- ساعت۳ تا ۶؛ ۶- ساعت ۶ تا ۹؛ ۷- ساعت ۹ تا ۱۲)

• سن خودرو: (۰-بین۰ تا ۴؛ ۱- بین ۱۰ تا ۱۵؛ ۲- بین ۱۵ تا ۲۵؛ ۳- بین ۲۵ تا ۴۵؛ ۴- بین ۴ تا ۱۰؛ ۵- بین ۴۵ تا ۶۵)

• نوع گروه‌بندی وسیله نقلیه: (۱- سواری کمتر از چهار سیلندر؛ ۲- سواری چهار سیلندر؛ ۳- سواری چهار سیلندر (پیکان، پراید، سمند)؛ ۴- سواری بیش از چهار سیلندر؛ ۵- وسیله نقلیه عمومی با ظرفیت هفت تا پانزده نفر(استیشن، ون)؛ ۶- وسیله نقلیه عمومی با ظرفیت شانزده تا بیست و شش نفر(مینی بوس)؛ ۷- وسیله نقلیه عمومی با ظرفیت بیست و هفت نفر وبیشتر (اتوبوس)؛ ۸- وسیله نقلیه بارکش با ظرفیت تا یک تن؛ ۹- وسیله نقلیه بارکش با ظرفیت بیش از یک تن تا سه تن؛ ۱۰- وسیله نقلیه بارکش با ظرفیت بیش از سه تن تا پنج تن؛ ۱۱- وسیله نقلیه بارکش با ظرفیت بیش از پنج تن تا ده تن؛ ۱۲- وسیله نقلیه بارکش با ظرفیت بیش از ده تن تا بیست تن؛ ۱۳- وسیله نقلیه بارکش با ظرفیت بیش از بیست تن؛ ۱۴- وسایل نقلیه ویژه حمل زباله وخیابان پاک‌کن‌ها؛ ۱۵- موتورسیکلت گازی؛
۱۶- موتورسیکلت دنده‌ای یک سیلندر؛ ۱۷- موتورسیکلت دنده‌ای دو سیلندر و به بالا؛ ۱۸- موتورسیکلت دنده‌ای دارای سه چرخ، ۱۹- سایر)

• نوع خسارت زیان‌دیده: (۱- افراد خارج از خودرو؛ ۲- راننده غیر مقصر؛۳- راننده مقصر؛ ۴- سرنشین خودرو غیر مقصر؛ ۵- سرنشین خودرو مقصر؛ ۶- مالی)

• نوع گواهی‌نامه: (۰- مناسب نیست؛ ۱- مناسب است)

• کد نوع حادثه: (۱- برخورد وسیله نقلیه با وسیله نقلیه دیگر؛ ۲- برخورد وسیله نقلیه با انسان؛ ۳- برخورد وسیله نقلیه با اشیاء و یا حیوانات؛ ۴- حادثه برای وسیله نقلیه به علت واژگونی؛ ۵- حادثه برای وسیله نقلیه به علت بلایای طبیعی؛ ۶- حادثه برای سرنشین خودرو (غیر از راننده) به‌دلیل تصادف یا واژگونی)

• کد محل وقوع حادثه: (۰-خارج از شهر؛ ۱- داخل شهر)

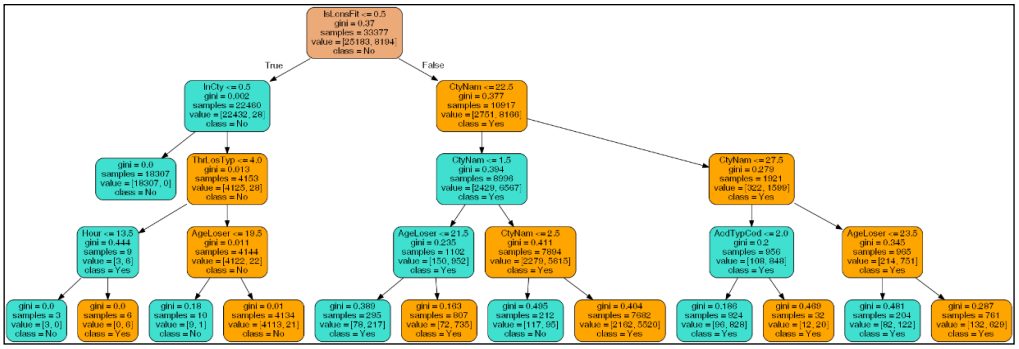
• کد استان: (۰- نامشخص؛ ۱- آذربایجان شرقی؛ ۲- آذربایجان غربی؛ ۳- اردبیل؛ ۴- اصفهان؛ ۵- ایلام؛ ۶- کرمانشاه؛ ۷- بوشهر؛ ۸- کهکیلویه و بویراحمد؛ ۹- تهران؛ ۱۰- چهارمحال و بختیاری؛ ۱۱- خراسان رضوی؛ ۱۲- خوزستان؛ ۱۳- زنجان؛ ۱۴- سمنان؛ ۱۵- سیستان و بلوچستان؛ ۱۶- فارس؛ ۱۷- کردستان؛ ۱۸- کرمان؛ ۱۹- گیلان؛ ۲۰- لرستان؛ ۲۱- مازندران؛ ۲۲- مرکزی؛ ۲۳- هرمزگان؛ ۲۴- همدان؛ ۲۵- یزد؛ ۲۶- قم؛۲۷- قزوین؛ ۲۸- گلستان؛ ۲۹- البرز؛ ۳۰- خراسان جنوبی؛ ۳۱- خراسان شمالی؛ ۴۵- خارج از ایران)

همانطور که در شکل ۱ نمایش داده شده است، همانگونه که در شکل ۱ مشاهده می‌شود، بیشترین فراوانی کلاس در هر متغیر به قرار ذیل است.

- محدوده شهری: خارج از شهر
- نوع حادثه: برخورد وسیله نقلیه با وسیله نقلیه دیگر
- جنسیت راننده مقصر حادثه: مرد
- نوع گواهی‌نامه: نامناسب
- نوع خسارت زیان‌دیده: راننده غیر مقصر
- نوع گروه‌بندی وسیله نقلیه: سواری چهار سیلندر
- جنسیت زیان‌دیده: مرد
- نوع وسیله نقلیه: سواری
- نوع کاربری: شخصی
- استان: تهران
- فاصله حادثه تا اعلام خسارت: تا ۷ روز
- سن مقصر حادثه: بین ۳۰ تا ۴۰ سال

- سن زیان‌دیده: بین ۲۰ تا ۳۰ سال
- سن خودرو: بین ۴ تا ۱۰ سال
- ساعت وقوع حادثه: بین ۹ تا ۱۲

در بخش بعد ابتدا به تجزیه و تحلیل داده‌ها با استفاده از مدل‌های مختلف یادگیری ماشین خواهیم پرداخت و سپس مدلی را ارائه می‌دهیم که کارایی و کیفیت سیستم کشف تقلب‌ها را تا حد زیادی افزایش می‌دهد. در شکل‌های ۲ و ۳ نمودار درخت تصمیم با عمق ۴ و مدل‌های بردار پشتیبان با کرنل خطی و RBF رسم شده است.



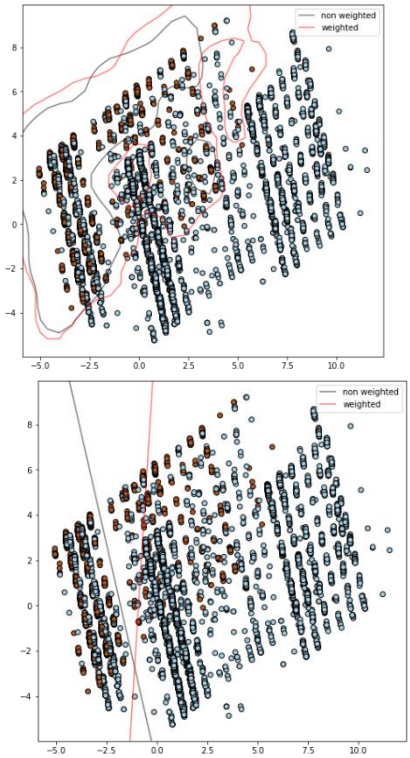
شکل ۲: درخت تصمیم با عمق ۴

۵- بررسی نتایج مدل

همانگونه که در بخش‌های قبل بیان گردید، پس از بررسی متغیر «مدت زمان بیمه‌نامه»، نمونه‌هایی که تاریخ شروع و پایان بیمه‌نامه یکسان می‌باشد به‌عنوان نمونه‌های مشکوک در نظر گرفته شد و سایر نمونه‌ها مربوط به پرونده‌های نرمال و سالم می‌باشد. لذا با توجه به داده‌های پژوهش که قبلا معرفی شد، یک مجموعه داده در اختیار خواهیم داشت به همراه ۱۵ متغیر مستقل و یک متغیر وابسته دودویی. با توجه به ساختار این مجموعه داده، برای مدل‌سازی از مدل‌های طبقه‌بندی لجستیک، درخت تصمیم و ماشین بردار پشتیبان استفاده شد که نتایج آن در جدول ۲ قابل مشاهده است.

جدول۲: کارایی مدل

مدل	صحت	دقت	یادآوری	امتیاز F1
لوجستیک	٪۹۲	٪۷۵	٪۹۹	٪۸۶
درخت تصمیم	٪۸۹	٪۷۸	٪۷۶	٪۷۷
ماشین بردار پشتیبان (خطی)	٪۹۲	٪۷۵	٪۹۹	٪۸۶
ماشین بردار پشتیبان (چندجمله‌ای)	٪۹۰	٪۷۲	٪۹۹	٪۸۳
ماشین بردار پشتیبان (تابع شعاعی)	٪۹۱	٪۷۳	٪۹۹	٪۸۴



شکل ۳- مرز جداکننده داده‌ها در فضای دوبعدی PCA با کرنل خطی و RBF

در جدول ۲ معیارهای مختلف کارایی برای مدل‌ها ارائه شده که روابط محاسبات آنها به شرح ذیل است:

$$\text{صحت} = \frac{\text{منفی واقعی+مثبت واقعی}}{\text{کل نمونه}}$$

و

$$\text{دقت} = \frac{\text{مثبت واقعی}}{\text{مثبت کاذب+مثبت واقعی}}$$

و

$$\text{بازخوانی} = \frac{\text{مثبت واقعی}}{\text{منفی کاذب+مثبت واقعی}}.$$

اگر به مخرج دو کسر آخر توجه کنیم، برای رابطه دقت[7] تمام مقادیری که مثبت پیش‌بینی شده‌اند را خواهیم داشت و در مورد رابطه بازخوانی[8]، تمام مقادیر مثبت واقعی به‌دست می‌آید. بنابراین روابط بالا را می‌توان به صورت ذیل بازنویسی کرد:

$$\text{دقت} = \frac{\text{مثبت واقعی}}{\text{کل مثبت های پیش بینی شده}}$$

$$\text{بازخوانی} = \frac{\text{مثبت واقعی}}{\text{کل مثبت های واقعی}}$$

همان‌طور که از روابط بالا مشخص است، معیار دقت زمانی استفاده می‌شود که تشخیص تعداد مواردی که واقعاً مثبت هستند از تعداد کل مواردی که مثبت پیش‌بینی شده‌اند، برای محقق اهمیت دارد. به‌عبارت دیگر، معیار دقت مواردی که صحیح پیش‌بینی شده‌اند را در اختیار محقق یا کاربر قرار می‌دهد و کاربرد اصلی آن زمانی است که هزینه نتایج مثبت‌های کاذب (FP) بالا باشد. از طرف دیگر معیار بازخوانی، بیانگر تعداد نمونه‌هایی است که از بین کل نمونه‌های واقعی به درستی مثبت پیش‌بینی شده‌اند. به طور مشابه این معیار زمانی استفاده می‌شود که هزینه نتایج منفی‌های کاذب بالا باشد. می‌توان گفت، رابطه دقت میزان دقت مدل را در موارد پیش‌بینی شده توسط مدل ارزیابی کرده و معیار فراخوانی بیانگر کارایی مدل در شناسایی موارد مثبت می باشد. در نهایت معیار F1 که از رابطه

$$F1 = 2 * \frac{\text{بازخوانی*دقت}}{\text{بازخوانی+دقت}}$$

به‌دست می‌آید، بین دو معیار صحت و بازخوانی نوعی تعادل برقرار می‌کند. همانطور که در جدول ۲ قابل مشاهده است میانگین معیارهای کارایی برای هیچکدام از الگوریتم‌های استفاده شده کمتر از ۸۰ درصد نمی باشد. این امر یک دستاورد بسیار مناسب در زمینه کارایی آن دسته از سیستم‌های هوشمند کشف تقلب در صنعت بیمه است که از الگوریتم‌های با نظارت بهره می‌برند. در حقیقت شیوه‌ی جدید ارائه شده در تحقیق حاضر باعث افزایش نرخ کشف پرونده‌های کلاهبرداری در صنعت بیمه می‌شود.

۶- بحث و نتیجه‌گیری

در این مقاله از الگوریتم‌های با نظارت جهت کشف تقلبات مشکوک استفاده گردید و تلاش شد با مقایسه نتایج حاصل از به‌کارگیری متداول‌ترین مدل‌های با نظارت در حوزه یادگیری ماشین، بهترین مدل از نظر عملکرد و کاراترین مدل جهت کشف پرونده‌های تقلب شناسایی و پیشنهاد شود. طبیعت نامتوازن داده‌های تقلب، کار را برای استفاده از الگوریتم‌های یادگیری ماشین دشوار می‌سازد. در واقع الگوریتم‌های با نظارت برای تحلیل بر روی داده‌های متوازن طراحی شده‌اند و این خصوصیت داده‌های تقلب باعث کاهش اطمینان به نتایج مدل می‌شود. ابتکار صورت‌گرفته در این مقاله به‌منظور رفع مشکل مطرح شده و ارتقا کیفیت سیستم کشف تقلب، استفاده از متغیر جایگزین است. در همین راستا با توجه به آنکه داده‌های مزبور از قبل بر اساس تقلبی بودن یا نبودن پرونده‌ها برچسب‌گذاری نشده بودند، این امر توسط تیم نویسندگان و با استفاده از متغیر جایگزین (مدت بیمه‌نامه) به‌عنوان شاخصی برای تشخیص پرونده‌های تقلبی انجام گرفت. پس از برچسب‌گذاری داده‌ها، الگوریتم‌های طبقه‌بندی لجستیک، درخت تصمیم و ماشین بردار پشتیبان بر روی مجموعه داده اعمال گردید که از آن میان مدل لجستیک و بردار پشتیبان خطی با دقت ۸۶٪ بهترین کارایی را از خود نشان دادند. شایان ذکر است الگوریتم‌ها و رویکرد استفاده شده در مقاله می‌تواند اساس سیستم‌های کشف تقلب خودکار قرار گیرد. در حقیقت از این طریق داده‌های جمع‌آوری‌شده سالانه به‌عنوان ورودی به سیستم داده می‌شود و پرونده‌های مشکوک برای تحقیقات بیشتر شناسایی می‌شوند و ظرف چندین سال داده‌های برچسب‌گذاری شده بیشتر و دقیق‌تری در اختیار سیستم و متخصصان قرار می‌گیرد تا با حساسیت و دقت بالاتری به تشخیص پرونده‌های متقلبانه بپردازند و به این ترتیب کیفیت کشف تقلبات بیمه‌های اتومبیل افزایش خواهد یافت.

[7]. Precision
[8]. Recall

۷- منابع

[1] Agaskar, V., Babariya, M., Chandran, S., & Giri, N. (2017). Unsupervised learning for credit card fraud detection. International Research Journal of Engineering and Technology (IRJET), 4(3), 2343-2346.

[2] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31.

[3] Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. Credit scoring and credit control VII, 235-255.

[4] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical science, 17(3), 235-255.

[5] Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A., & Alpert, M. (2002). Fraud classification using principal component analysis of RIDITs. Journal of Risk and insurance, 69(3), 341-371.

[6] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[7] Chen, J. L., & Lai, K. L. (2021). Deep convolution neural network model for credit-card fraud detection and alert. J. Artif. Intell., 3(02), 101-112.

[8] Derrig, R. A. (2002). Insurance fraud. Journal of Risk and Insurance, 69(3), 271-287.

[9] Dheepa, V., & Dhanapal, R. (2012). Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft computing, 2(4), 391-397.

[10] Domingues, R. (2015). Machine Learning for Unsupervised Fraud Detection.

[11] Gomes, C., Jin, Z., & Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. Journal of Risk and Insurance, 88(3), 591-624.

[12] Gyamfi, N. K., & Abdulai, J. D. (2018, November). Bank fraud detection using support vector machine. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 37-41). IEEE.

[13] Lepoivre, M. R., Avanzini, C. O., Bignon, G., Legendre, L., & Piwele, A. K. (2016). Credit card fraud detection with unsupervised algorithms. Journal of advances in information technology, 7(1).

[14] 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(3), 275-287.

[15] Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2(1), 58-75.

[16] Noble, C. C., Cook, D. J. (2003, August). Graph-based anomaly detection. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 631-636).

[17] Peng, Y., Kou, G., Sabatka, A., Chen, Z., Khazanchi, D., & Shi, Y. (2006, October). Application of clustering methods to health insurance fraud detection. In 2006 International Conference on Service Systems and Service Management (Vol. 1, pp. 116-120). IEEE.

[18] Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. Acm sigkdd explorations newsletter, 6(1), 50-59.

[19] Rushin, G., Stancil, C., Sun, M., Adams, S., & Beling, P. (2017, April). Horse race analysis in credit card fraud—deep learning, logistic regression, and Gradient Boosted Tree. In 2017 systems and information engineering design symposium (SIEDS) (pp. 117-121). IEEE.

[20] Sabau, A. S. (2012). Survey of clustering based financial fraud detection research. Informatica Economica, 16(1), 110.

[21] Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, 40(15), 5916-5923.

[22] Sahin, Y., Duman, E. (2011, June). Detecting credit card fraud by ANN and logistic regression. In 2011 International Symposium on Innovations in Intelligent Systems and Applications (pp. 315-319). IEEE.

[23] Subudhi, S., Panigrahi, S. (2018). Detection of automobile insurance fraud using feature selection and data mining techniques. International Journal of Rough Sets and Data Analysis, 5(3), 1-20.

[24] Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., & Dedene, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. European Journal of Operational Research, 176(1), 565-583.

[25] Viaene, S., Dedene, G. (2004). Insurance fraud: issues and challenges. The Geneva Papers on Risk and Insurance-Issues and Practice, 29(2), 313-333.

[26] Viaene, S., Derrig, R. A., Baesens, B., & Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. Journal of Risk and Insurance, 69(3), 373-421.

[27] Yaram, S. (2016, August). Machine learning algorithms for document clustering and fraud detection. In 2016 International Conference on Data Science and Engineering (ICDSE) (pp. 1-6). IEEE.

[28] Yuan, S., Wu, X., Li, J., & Lu, A. (2017, November). Spectrum-based deep neural networks for fraud detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 2419-2422).

[29] Zou, K., Sun, W., Yu, H., & Liu, F. (2012, March). ID3 decision tree in fraud detection application. In 2012 International Conference on Computer Science and Electronics Engineering (Vol. 3, pp. 399-402). IEEE.

Target replacement, a new approach to increase the performance of fraud detection system in auto insurance utilizing supervising learning

Farbod Khanizadeh
Assistant professor, Insurance research center. Tehran. Iran. khanizadeh@irc.ac.ir

Maryam Esna-ashari[9*]
Assistant professor, Insurance research center. Tehran. Iran. esnaashari@irc.ac.ir

Farzan Khamesian
Assistant professor, Insurance research center. Tehran. Iran. khamesian@irc.ac.ir

Azadeh Bahador
Assistant professor, Insurance research center. Tehran. Iran. bahador@irc.ac.ir

Abstract: Recent years, the insurance industry has been experiencing an increase in equipping insurance companies with fraud detection systems. Furthermore due to the significant cost imposed on the insurance industry by the rise in such claims, the role of data mining techniques in detecting fraudulent claims has become widespread. However an essential issue with such systems is the quality of their outputs. On one hand, supervised algorithms are more accurate comparing to unsupervised counterparts. On the other hand, as data labeled fraud is really limited, the efficiency of supervised algorithms is severely challenged. Within this regard, a novel approach is introduced as "alternative feature" to overcome the challenge. Basically, alternative feature is a variable whose values are available and can be considered a suitable indicator to detect suspicious cases. This approach improves the efficiency of the system and allows experts and insurance companies to investigate suspicious cases with more confidence and less error.

Keywords: supervised learning, target replacement, fraud detection, auto insurance

Introduction

In recent years, the insurance industry has shown willing to equip companies with fraud detection systems. In fact, due to the large cost imposed on the

[9] Corresponding author: esnaashari@irc.ac.ir

industry, fraud detection and identification algorithms should be part of all modern financial systems. In this regard, the role of data mining methods in discovering fraud cases has become more significant. However there is a fundamental problem. On the one hand, supervised algorithms are more accurate than unsupervised ones. On the other hand, due to the nature of fraud detection, the labeled data is very limited, and this makes the use of supervised algorithms and their accuracy a big challenge.

In this article, due to the importance of detecting fraudulent claims in car insurance, we have investigated the issue through machine learning methods. When using data mining and machine learning algorithms, two supervised and unsupervised methods can be utilized. Due to the nature of fraud datasets, the use of unsupervised algorithms is more common ((Bolton and Hand, 2001) and (Dominguez, 2015)). In some cases in which labeled data are accessible, it is possible to use supervised algorithms. For example, Zhou et al. (2012), Gyamfi and Abdoulai (2018) used support vector machine in their research. In this article, we have also used supervised machine learning algorithms to detect suspicious claims in third-party car insurance. Moreover, in order to solve the challenge of the small number of detected fraud cases, we proposed the notion of "variable replacement", through which another variable whose values are available and is a suitable indicator for suspicious cases is considered as the target variable. The approach presented in the article allows experts and insurance companies to choose the right supervised algorithm and to act more confidently and accurately on suspicious cases.

research method: One of the challenges in dealing with fraudulent claims is the lack of labeled targets. This makes it impossible to use supervised algorithms. To overcome this issue, the research provides a different variable (replacement target variable) whose results were used as a guide to detect suspicious features. The aforementioned alternative variable is the duration of the insurance policy as an indicator to detect frauds in car insurance. In the following, according to the structure of the data set, logistic regression, decision tree and support vector machine have been used.

Research data: In this article, 15 independent variables and an alternative target variable named "insurance policy duration" are used. The statistical population includes fifty thousand examples third party car insurance claims. In order to increase the efficiency of machine learning models, necessary pre-processing has been done on the data. It is worth mentioning that the pre-processing steps are not fixed and are determined based on the available data set. In this regard, for the pre-processing the following steps were taken: extraction of usable parts, integration of features, removing noise data, assigning numerical codes to all variables, grouping of numerical independent variables. After examining the variable "duration of the insurance policy", the samples with the same start and end date are considered as suspicious samples while others are normal cases. Data and

Findings and conclusions: The initiative taken in this article to solve the problem of data set imbalance was the use of "variable replacement". In this regard, considering that the aforementioned data were not labeled based on whether the cases were fraudulent or not, this was done by the team of authors using an alternative variable (policy duration) as an indicator to detect fraudulent cases. After labeling the data, logistic models, decision tree and support vector machine were applied on the data, of which logistic model and linear support vector showed the best performance with 86% accuracy. The results can be seen in Table 1.

Table 1. Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Logestic	92%	75%	99%	86%
Decision Tree	89%	78%	76%	77%
SVM (linear)	92%	75%	99%	86%
SVM (poli)	90%	72%	99%	83%
SVM (rbf)	91%	73%	99%	84%

It is worth mentioning that the algorithms and approach used in the article can be used as the basis of automatic fraud detection systems. In fact, in this way, the data collected annually is given as input to the system, and suspicious cases

are identified for further investigation, and within several years, more and more accurate labeled data will be provided to the system and experts to detect fraudulent cases with higher sensitivity and accuracy.

References:

[1] Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. Credit scoring and credit control VII, 235-255.

[2] Domingues, R. (2015). Machine Learning for Unsupervised Fraud Detection.

[3] Zou, K., Sun, W., Yu, H., & Liu, F. (2012, March). ID3 decision tree in fraud detection application. In 2012 International Conference on Computer Science and Electronics Engineering (Vol. 3, pp. 399-402). IEEE.

[4] Gyamfi, N. K., & Abdulai, J. D. (2018, November). Bank fraud detection using support vector machine. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 37-41). IEEE.