فصلنامه تأمین اجتماعی
● سال بیست و یکم ● شماره اول ● بهار ۱۴۰۴

شناسایی هوشمند تقلب در دستمزدهای اعلامی در سال‌های پایانی بیمه‌پردازی با بهره‌گیری از الگوریتم‌های یادگیری ماشین بدون نظارت[۱]

امیرسالار ربیعی[۲]، غلامرضا رضایی[۳]، داود قراخانی[۴]

چکیده

هدف: تقلب درگزارش دستمزد از چالش‌های مهم نظام‌های بیمه‌ای است که با افزایش حجم داده‌ها و گسترش سامانه‌های غیرحضوری، شناسایی آن با روش‌های سنتی دشوارتر شده است. این پژوهش با هدف ارزیابی توان الگوریتم‌های یادگیری ماشین بدون نظارت در کشف ناهنجاری‌های مرتبط با تقلب دستمزدی و ارائه رویکردی خودکار برای تقویت فرایندهای نظارتی سازمان تأمین اجتماعی انجام شده است.

روش: این پژوهش ازنظر هدف، کاربردی و از نظر ماهیت داده‌ها، توصیفی-تحلیلی است. سه الگوریتم جنگل ایزوله، الگوریتم خوشه‌بندی مبتنی بر چگالی (DBSCAN) و ماشین بردار پشتیبان (One-Class SVM) روی مجموعه‌ای شامل ۲۶۲۵۸ رکورد ماهانه دستمزد ۴۷۰ بیمه‌شده طی سال‌های ۱۳۹۸ تا ۱۴۰۲ در تشخیص تقلب برای شناسایی دستمزدهای نامتعارف استفاده شده است. تحلیل‌ها با تمرکز بر الگوهای ناهنجار در سطوح فردی و بین‌فردی انجام و کارایی روش‌ها بر اساس منطق تشخیص و سازگاری با رفتار واقعی داده‌ها ارزیابی شد.

یافته‌ها: نتایج نشان داد الگوریتم جنگل ایزوله دقیق‌ترین و پایدارترین عملکرد را داشته و ناهنجاری‌ها را با توزیع منطقی‌تر نسبت به سایر روش‌ها شناسایی کرده است. الگوریتم خوشه‌بندی مبتنی‌بر چگالی در داده‌های پراکنده دچار حذف بیش از حد شده و ماشین بردار پشتیبان، حساسیت بالایی همراه با نرخ هشدار کاذب بیشتر نشان داده است.

نتیجه‌گیری: یادگیری ماشین بدون نظارت، توانایی مؤثری در شناسایی خودکار رفتارهای مشکوک دستمزدی دارد و استفاده از جنگل ایزوله می‌تواند راهکاری مقیاس‌پذیر و قابل‌اعتماد برای کاهش

۱- این مقاله برگرفته از رساله دکتری با عنوان «شناسایی ناهنجاری در دستمزد بیمه‌شدگان تأمین اجتماعی» می‌باشد.

۲- دانشجوی دکتری مدیریت صنعتی دانشگاه آزاد اسلامی استان قزوین.
۳- دکتری اقتصاد، استادیار دانشگاه آزاد اسلامی استان قزوین. (نویسنده مسئول). q.r.rezaei@gmail.com
۴-دکتری مدیریت، استادیار گروه مدیریت دانشگاه آزاد اسلامی استان قزوین.

ریسک تقلب در نظام‌های بیمه‌ای فراهم کند. پیشنهاد می‌شود این الگوریتم به‌عنوان هسته یک سامانه هشدار هوشمند در ساختارهای نظارتی مورد استفاده قرار گیرد.

واژه‌های کلیدی: تقلب بیمه‌ای، جنگل ایزوله، خوشه‌بندی مبتنی بر چگالی، ماشین بردار پشتیبان تک کلاسه، سازمان تأمین اجتماعی.

۱. مقدمه

در دهه‌های اخیر، رشد روزافزون داده‌های سازمانی و توسعه سامانه‌های دیجیتال در حوزه‌های بیمه‌ای، منجربه تولید حجم عظیمی از داده‌های مرتبط با دستمزد، سوابق بیمه و بازنشستگی بیمه‌شدگان شده است. در چنین محیط‌های داده‌محور، تشخیص ناهنجاری‌ها[۱] به‌عنوان یکی از مسائل کلیدی در تحلیل داده و کشف دانش مطرح است. ناهنجاری‌ها در داده‌های بیمه‌ای می‌توانند نشانگر اشتباهات انسانی، خطاهای سیستمی یا رفتارهای غیرعادی و حتی تقلب‌های احتمالی در گزارش‌دهی دستمزد توسط کارفرمایان باشند. این مسئله در نظام بیمه‌ای کشور، به‌ویژه در سازمان تأمین اجتماعی، اهمیت ویژه‌ای دارد؛ چراکه دستمزدهای ارسالی مبنای محاسبه حق‌بیمه، مزایای کوتاه‌مدت و بلندمدت، و درنهایت میزان مستمری بازنشستگی افراد قرار می‌گیرد. ایران در مقایسه با کشورهای دیگر دنیا یکی از بالاترین نرخ‌های جایگزینی[۲] را در صندوق‌های بازنشستگی دارد؛ به‌عبارت‌دیگر، در برخی موارد زمانی‌که فردی بازنشسته می‌شود، تقریباً همان حقوق مشمول کسور زمان اشتغال را دریافت می‌کند. یکی از دلایل اصلی نرخ جایگزینی بالا در ایران، نحوه محاسبه حقوق بازنشستگی افراد است. در شرایطی که در اغلب نظام‌های بازنشستگی دنیا متوسط بهترین ۵ سال سنوات خدمت فرد یا تمام سال‌های بیمه‌پردازی فرد مبنای محاسبه قرار می‌گیرد، در ایران دو سال آخر خدمت فرد مبنای محاسبه است.[۳] نرخ جایگزینی یکی از شاخص‌های تعیین‌کننده میزان سخاوتمندی نظام‌های مستمری است. میانگین این نسبت برای صندوق‌های بیمه ایران در سال ۱۳۹۸ حدود ۸۳ درصد است که از تقسیم اولین مستمری دریافتی به آخرین دستمزد مشمول کسر حق‌بیمه افراد محاسبه می‌شود. این سخاوتمندی باعث طمع عده‌ای برای دریافت مستمری بیشتر شده است. نظام بیمه‌ای تأمین اجتماعی ایران از نوع مزایای تعریف‌شده[۴] است و سازوکار آن مبتنی‌بر نظام بدون ذخیره (PAYG[۵]) است به‌طوری‌که مبنای اصلی تأمین و پرداخت مستمری‌ها از محل حق‌بیمه‌های افراد شاغل (کسورات) است که در این نظام پرداخت مستمری بر اساس متوسط دستمزد دو سال آخر بیمه‌پردازی تقسیم‌بر سنوات پرداخت حق‌بیمه محاسبه می‌شود[۶] (نعیمی و همکاران، ۱۳۹۰:۳۲۹).

در نظام تأمین اجتماعی وفق ماده ۲۸ و ۳۹، کارفرمایان مسئول تنظیم و ارسال لیست حق‌بیمه به سازمان می‌باشند و در حال حاضر بیش از ۱۰ میلیون بیمه‌شده اجباری در سازمان بیمه‌پرداز می‌باشند.

1- Anomaly Detection.

۲- مقرری بازنشستگی در مقایسه با آخرین حقوق و دستمزد مبنای بیمه‌پردازی.

۳- ماده ۷۷ قانون ت ا ا.

4- Defined Benefit (DB).

5- Pay as you go.

۶- ماده ۷۷- میزان مستمری بازنشستگی عبارت است از یک‌سی‌ام متوسط مزد یا حقوق بیمه‌شده ضرب‌در سنوات پرداخت حق‌بیمه مشروط بر اینکه از سی‌وپنج سی‌ام متوسط مزد با حقوق تجاوز ننماید.

در پلتفرم جاری سازمان اگر کارفرمایی اقدام به افزایش نامتعارف دستمزد بیمه‌شده یا بیمه‌شدگان خود به‌قصد بهره‌برداری بیشتر از مزایای بازنشستگی بنماید راهکار مکانیزه‌ای برای کنترل این تقلب در سازمان وجود ندارد. از طرفی وفق ماده ۳۹ قانون تأمین اجتماعی[۱] فرصتی شش ماهه جهت کنترل لیست‌های ارسالی کارفرما به سازمان داده شده است و باتوجه‌به حجم بالای رکوردهای ماهیانه لیست‌ها، امکان کنترل و نظارت انسانی را غیرممکن کرده است ادعاهای دستمزد دروغین می‌توانند منجربه پرداخت نادرست یا سود مالی برای بیمه‌شدگان متقلب شود و درنهایت هزینه‌های بلندمدت سازمان را به‌طور نامتعادلی افزایش و نرخ جایگزینی غیرواقعی برای متقلبین ایجاد نماید.

تقلب یک مسئله مهم برای سازمان‌های بیمه‌گر است. هزینه‌های هنگفت تقلب نه‌تنها شرکت‌های بیمه، بلکه بیمه‌شدگان صادق را نیز از طریق تأثیر آن‌ها بر حق‌بیمه تحت‌تأثیر قرار می‌دهد (ویان[۲] و همکاران، ۲۰۰۷)، فراتر از صرفه‌جویی در هزینه مرتبط با شناسایی ادعاهای تقلب، تشخیص مؤثر تقلب بیمه نیز به‌عنوان یک عامل بازدارنده عمل می‌کند که برای بازار بیمه حیاتی است (تنسون[۳]، ۲۰۰۲)؛ بنابراین، از دیدگاه شرکت‌های بیمه و بیمه‌گذاران صادق، بررسی این موضوع که کدام روش می‌تواند به بهترین نحو تقلب در ادعاهای بیمه را تشخیص دهد، مهم است.

بررسی ادبیات نظری در این موضوع نشان می‌دهد که اگرچه استفاده از الگوریتم‌های بدون نظارت در حوزه‌های مالی و بیمه‌ای به طور گسترده‌ای موردتوجه قرار گرفته، کاربرد آن‌ها در تحلیل سری‌های زمانی دستمزد بیمه‌شدگان در سال‌های پایانی خدمت کمتر مورد بررسی قرارگرفته است شکاف پژوهشی موجود، فرصتی ارزشمند برای ارائه مدلی نوآورانه فراهم می‌کند. پژوهش حاضر با تمرکز بر الگوریتم جنگل‌های ایزوله[۴] و مقایسهٔ آن با سایر الگوریتم‌های مشابه، درصدد است چهارچوبی برای شناسایی دستمزدهای غیرعادی ارائه دهد و راهکاری عملی برای سازمان تأمین اجتماعی در مدیریت ریسک تقلب دستمزدی و کاهش نرخ جایگزینی فراهم آورد که هدف، ارائه رویکردی هوشمند و مقیاس‌پذیر برای تشخیص ناهنجاری در داده‌های دستمزد و تلاش برای ارتقای دقت در فرایند نظارت و جلوگیری از تحمیل بار مالی غیرواقعی بر سازمان و بهبود شفافیت و عدالت بیمه‌ای باشد. سال‌های اخیر ادغام تکنیک‌های هوش مصنوعی قابل‌تبیین[۵] با مدل‌های تشخیص ناهنجاری بسیار موردتوجه قرار گرفته است (کارلتی و همکاران[۶]، ۲۰۲۱). نوآوری پژوهش ما نیز در همین راستاست در این مقاله تلاش شده است مدل جنگل ایزوله آموزش‌دیده روی داده‌های دستمزد کارگران را با ابزارهای تفسیرپذیری همراه

۱- سازمان حداکثر ظرف شش ماه از تاریخ دریافت صورت مزد اسناد و مدارک کارفرما را مورد رسیدگی قرار داده و...
2- Viaene.
3- Tennyson.
4- Isolation Forest.
5- Explainable AI.
6- Carletti et al.

کنیم تا خروجی مدل برای تصمیم‌گیران قابل‌فهم‌تر شود. به‌طور مشخص، با استخراج شاخص‌های اهمیت ویژگی[1] و الگوهای مؤثر بر تصمیم مدل برای ناهنجار شناختن یک رکورد، مدلی ارائه شده است که قابل توضیح برای انسان باشد. این رویکرد در حقیقت پلی بین دقت الگوریتمی و شفافیت موردنیاز در سازمان تأمین اجتماعی است.

بر این اساس، باتوجه‌به شکاف پژوهشی موجود در این زمینه، این تحقیق در پی پاسخ به سؤال‌های زیر است:

۱- کدام یک از الگوریتم‌های یادگیری ماشین بدون نظارت، در تشخیص ناهنجاری در دستمزد بیمه‌شدگان، از منظر شاخص‌های پایداری و نرخ هشدار (باتوجه‌به محدودیت نبود داده‌های برچسب‌دار) عملکرد بهتری دارد؟

۲- الگوریتم‌های یادگیری ماشین بدون نظارت، با تأکید بر جنگل ایزوله، تا چه حد توانایی شناسایی الگوهای ناهنجار در روند دستمزد بیمه‌شدگان در سال‌های پایانی خدمت را دارند؟

۳- کدام متغیرها و ویژگی‌های زمانی (نظیر شیب جهش دستمزد، انحراف از روند تاریخی یا تغییر رفتار نسبت به هم‌صنفان) بیشترین تأثیر را در ناهنجار قلمداد شدن یک رکورد ایفا می‌کنند؟

۲. مروری بر ادبیات روش‌های شناسایی تقلب در بیمه

تقلب یک مسئله اجتماعی مهم است که توجه بسیاری از محققان دانشگاهی را به خود جلب کرده است. می‌توان آن را به‌عنوان به‌دست آوردن چیزی باارزش یا اجتناب از یک تعهد از طریق فریب تعریف کرد (گرابوسکی و دافیلد[2]، ۲۰۰۱). تقلب بیمه توسط گیل و همکاران[3](۲۰۰۵) به‌عنوان ارائه آگاهانه یک ادعای ساختگی، بزرگ‌نمایی یک ادعا یا افزودن موارد اضافی به یک ادعا، یا به هر نحوی نادرست بودن ادعا، باهدف به‌دست آوردن چیزی بیش از حق قانونی تعریف شده است. از کل حجم هزار میلیارد دلاری صنعت بیمه در جهان، حدود ۲۵درصد از فعالیت‌ها در بعضی حوزه‌ها، مثل بیمه درمانی تا ۴۰درصد دارای تقلب بوده است (کومار و همکاران[4]، ۲۰۱۰). به‌طورکلی، تقلب ناشی از تعامل دو عامل است: انگیزه یک مجرم برای کلاهبرداری و فرصتی برای انجام این کار (ویانی و ددنی[5]، ۲۰۰۴). انگیزه تقلب را می‌توان به‌عنوان مثال به انرژی مجرمانه، اعتبار و طمع و وضعیت (اقتصادی) مجرم تقسیم کرد (گرابوسکی و دافیلد، ۲۰۰۱). فرصت تقلب را می‌توان ترکیبی از یک هدف مناسب و عدم وجود مکانیسم‌های محافظتی مؤثر در نظر گرفت، شرکت‌های بیمه به‌ویژه در پردازش دعاوی خود،

1- Feature Importance.
2- Grabosky & Duffield.
3- Gill.
4- Kumar et al.
5- Viaene & Dedene.

مستعد تقلب هستند (ویان ودنی، ۲۰۰۴) شرکت‌های بیمه‌گر روش‌های مختلفی برای مبارزه با تقلب به کار می‌گیرند که در ادامه به شرح مختصری از آنها پرداخته می شود.

۲-۱. روش‌های سنتی

در این روش‌ها شرکت‌های بیمه‌گر، استراتژی‌های حسابرسی را اجرا می‌کنند که بر اساس آن‌ها تصمیم می‌گیرند که آیا یک ادعا باید بررسی شود یا خیر. ازآنجاکه بررسی مشروعیت یک ادعا، هزینه‌بر و زمان‌بر است، این فرایند نوعی تأیید هزینه بر سازمان بیمه‌گر است (پیکارد[1]، ۱۹۹۶). فرایند رسیدگی به ادعا در شرکت‌های بیمه معمولاً با غربالگری اولیه ادعا آغاز می‌شود، این غربالگری تعیین می‌کند که آیا ادعا به‌طور معمول پردازش می‌شود یا به یک واحد بازرسی ویژه تحویل داده شود. واحد بازرسی ویژه شامل کارشناسانی است که قبل از تصمیم‌گیری در مورد پرداخت ادعا یا آغاز مذاکرات یا اقدامات قانونی، ادعا را به طور مفصل بررسی می‌کنند (ویان و همکاران[2]، ۲۰۰۷). تقلب بیمه معمولاً به‌عنوان یک بازی کم‌خطر و پربازده در نظرگرفته می‌شود (دریگ[3]، ۲۰۰۲) این مشکل همچنین اهمیت یک غربالگری اولیه کارآمد را برجسته می‌کند که می‌تواند مواردی با احتمال بالای تقلب را شناسایی کند تا واحد بازرسی ویژه بتواند منابع ارزشمند خود را به این موارد اختصاص دهد.

۲-۲. روش‌های تشخیص تقلب مبتنی‌بر یادگیری ماشین

داده‌کاوی به شناسایی و جلوگیری از تقلب بیمه کمک می‌کند و می‌تواند تشخیص ناهنجاری، خوشه‌بندی و طبقه‌بندی ادعاهای جعلی بیمه را امکان‌پذیر کند (تورنتون و همکاران[4]، ۲۰۱۴). یادگیری ماشین[5] را می‌توان به‌عنوان فرایندی تعریف کرد که در آن «یک رایانه برخی داده‌ها را مشاهده می‌کند، مدلی را بر اساس داده‌ها می‌سازد و از این مدل هم به‌عنوان فرضیه‌ای در مورد جهان و هم به‌عنوان نرم‌افزاری که می‌تواند مشکلات را حل‌کند، استفاده می‌کند» (راسل و نورویگ[6]، ۲۰۲۰) یادگیری ماشین اساساً کاربردی از تکنیک‌های هوش مصنوعی است تا سیستم‌ها بتوانند خودشان یاد بگیرند. این بدان معناست که سیستم به طور خودکار یاد می‌گیرد، بداهه‌سازی می‌کند و از طریق تجربه سازگار می‌شود، بدون اینکه برای انجام یک عملیات خاص برنامه‌ریزی شده باشد. (چاترای و همکاران[7]، ۲۰۲۰) یادگیری ماشین به‌ویژه برای مدیریت موقعیت‌های پیچیده مناسب است زیرا می‌تواند با مطالعه و تجزیه‌وتحلیل

1- pikard.
2- Viaene et al.
3- Derrig.
4-Thornton et al.
5- Machine learning.
6- Russell & Norvig.
7- Khatri et al.

خودکار حجم زیادی از داده‌ها، تصمیم‌گیری کند و به رفتارهای غیرمعمول بسیار سریع‌تر از انسان پاسخ دهد که ازنظر تشخیص زودهنگام یک مزیت قابل توجه است. بسته به مجموعه داده‌های موجود، پژوهشگران می‌توانند بین یادگیری نظارت شده و یادگیری بدون نظارت و نیمه نظارت شده و یا ترکیب آن‌ها را انتخاب کنند. اولی برای حجم زیادی از داده‌های برچسب‌گذاری شده مناسب است، درحالی‌که دومی برای داده‌های فاقد برچسب‌گذاری به‌خوبی کار می‌کند (وانگ و همکاران[1]، ۲۰۲۰). تشخیص سنتی تقلب در بیمه به‌شدت به حسابرسی و بازرسی ماهرانه متکی بود (نیان و همکاران[2]، ۲۰۱۶) به دلیل پیشرفت‌های تکنولوژیکی و عملیات تجاری در مقیاس بزرگ، اتخاذ چنین روش‌های مرسومی، کار تشخیص تقلب را غیرعملی می‌کند (کمپ[3]، ۲۰۱۰) این حجم بالای هزینه‌های سربار، سازمان‌ها را به استفاده از فنّاوری‌های جدید در عرصه کشف تقلب راغب‌تر کرده است.

الگوریتم‌های یادگیری ماشین مورد استفاده برای تشخیص تقلب را می‌توان در درجه اول به سه دسته مهم زیر تقسیم کرد: مدل‌های تحت نظارت[4]، بدون نظارت[5] و نیمه‌نظارتی[6].

۲-۲-۱. یادگیری تحت نظارت[7]

یادگیری تحت نظارت یکی از رویکردهای اصلی در یادگیری ماشین است که هدف آن برازش یک مدل با استفاده از داده‌های برچسب‌دار به‌منظور پیش‌بینی یا طبقه‌بندی نمونه‌های جدید است این نوع یادگیری به‌عنوان یادگیری پیش‌بینی‌کننده شناخته می‌شود (چاترای و همکاران، ۲۰۲۰) زیرا برچسب کلاسِ نمونه‌های جدید را بر اساس اطلاعات مربوط به نمونه‌های مشابه در داده‌های آموزشی پیش‌بینی می‌کند. در این رویکرد، برای هر مشاهده، مقدار واقعی متغیر وابسته مشخص است و مدل می‌کوشد رابطه‌ای بهینه میان ویژگی‌های ورودی (متغیرهای مستقل) و خروجی (برچسب هدف) بیابد (بیشاپ[8]، ۲۰۰۶) یادگیری تحت نظارت به‌عنوان تلاش برای استنتاج تابعی تعریف می‌شود که به بهترین شکل داده‌های ورودی داده شده را به یک خروجی داده شده نگاشت می‌کند. مجموعه‌ای از داده‌های آموزشی در طول مرحله آموزش مدل به‌عنوان ورودی سیستم ارائه می‌شود. هر ورودی با یک مقدار خروجی مطلوب برچسب‌گذاری می‌شود که اساساً بر مدل نظارت دارد برای توسعه یک مدل تحت نظارت برای تشخیص تقلب، ابتدا باید اطلاعات گذشته در مورد وقوع واقعی تقلب و

1- N. Wang et al.
2- Nian et al.
3- Kemp.
4- Supervised.
5- Unsupervised.
6- Semi-supervised.
7- Supervised Learning.
8- Bishop.

غیر تقلب را به دست آوریم. بر اساس دقت برچسب‌های خروجی، مدل تلاش می‌کند تا پارامترها را برای شناسایی بهتر موارد تقلب و غیر تقلب بهینه کند. بااین‌حال، یکی از چالش‌های اساسی در استفاده از روش‌های تحت نظارت در داده‌های بیمه‌ای، فقدان داده‌های برچسب‌دار معتبر است. در بسیاری از موارد، ناهنجاری‌های واقعی شناسایی یا تأیید نشده‌اند و همین امر باعث می‌شود روش‌های یادگیری بدون نظارت مناسب‌تر باشند. بااین‌وجود، در صورت وجود داده‌های برچسب‌دار، استفاده از یادگیری تحت نظارت می‌تواند دقت تشخیص تقلب را به طور چشمگیری افزایش دهد. بسیاری از مطالعات که از یادگیری ماشین برای تشخیص کلاهبرداری بیمه استفاده می‌کنند، بر روی روش‌هایی از حوزه یادگیری تحت نظارت تمرکز دارند (برای مثال جانسون و خوش‌گفتار[1]، ۲۰۱۹ و وانگ و همکاران[2]، ۲۰۱۸) الوان[3] و همکاران (۲۰۲۲) نشان می‌دهند که چگونه ترکیب تکنیک‌های یادگیری ماشین با روش‌های موجود برای تشخیص کلاهبرداری می‌تواند یافتن کلاهبرداری را آسان‌تر کند. به طور خاص آن‌ها اثربخشی چندین تکنیک داده‌کاوی، ازجمله درخت تصمیم، ماشین بردار پشتیبان، نزدیک‌ترین همسایه K و مدل پنهان مارکوف، را در تشخیص کلاهبرداری کارت اعتباری بررسی می‌کنند. یافته‌ها، پتانسیل یک رویکرد ترکیبی را که این روش‌ها را برای افزایش تشخیص کلاهبرداری ادغام می‌کند، برجسته می‌کند (الوان و همکاران، ۲۰۲۲).

۲-۲-۲. یادگیری بدون نظارت[4]

سازمان‌های بیمه‌گر، در عصری فعالیت می‌کنند که ترکیب فنّاوری و شیوه‌های عملکرد مجرمان، بسیار پویا شده است، به‌طوری‌که نمی‌توان از یک الگوریتم واحد برای تشخیص کلاهبرداری استفاده کرد. دشواری یادگیری این ویژگی‌های رفتاری پویا، ازنظر پیچیدگی به‌صورت تصاعدی افزایش می‌یابد و یادگیری نظارت شده در عمل نمی‌تواند چنین متغیرهای نهفته پیچیده‌ای را در داده‌ها بیاموزد. این محدودیت‌ها مانع استفاده از مدل‌سازی نظارت شده در تشخیص کلاهبرداری بیمه شده است و درنتیجه فرصتی برای یک الگوی مدل‌سازی جایگزین که یادگیری بدون نظارت است، فراهم می‌کند. ادبیات نوظهوری در مورد توسعه روش‌های یادگیری بدون نظارت در تشخیص کلاهبرداری بیمه وجود دارد (اکین و همکاران[5]، ۲۰۱۸ و ظفری و اکین[6]، ۲۰۱۸) یادگیری بدون نظارت شاخه‌ای از یادگیری ماشین است که در آن داده‌ها فاقد برچسب مشخص هستند و الگوریتم‌ها تلاش می‌کنند الگوها،

1- Johnson & Khoshgoftaar.
2- X .Wang et al.
3- Alwan et al.
4- Unsupervised Learnin.
5- Ekin et al.
6- Zafari & Ekin.

ساختارهای پنهان و ناهنجاری‌ها را در داده‌ها شناسایی کنند (گودفلوی و همکاران[1]، ۲۰۱۶) یادگیری بدون نظارت رویکردی داده‌محور و اکتشافی است که به سازمان‌ها این امکان را می‌دهد تا بدون اتکا به داده‌های برچسب‌دار، الگوهای پنهان را استخراج کرده و رفتارهای غیرعادی را در داده‌های پیچیده شناسایی کنند؛ بنابراین، می‌تواند نمایش ویژگی داده‌ها را در غیاب هرگونه سوگیری یا گمراهی ناشی از داده‌های خروجی ذهنی، بهتر تعیین‌کند. برخلاف مدل‌های تحت نظارت، ما برای توسعه یک مدل نیازی به دانش قبلی از برچسب‌های خروجی نداریم (نیو و همکاران[2]، ۲۰۱۹).

در یادگیری بدون نظارت، مدل‌ها بر اساس شباهت، فاصله یا چگالی داده‌ها عمل می‌کنند و از روابط درونی میان مشاهدات برای خوشه‌بندی یا تشخیص نقاط غیرعادی بهره می‌گیرند. یادگیری بدون نظارت شامل دسته‌ای از روش‌های تحلیلی است که تلاش می‌کنند تابعی را استنباط کنندکه به بهترین شکل نمایش داده‌های ورودی را در غیاب هرگونه نظارتی مشخص کند. مدل‌های بدون نظارت مبتنی‌بر شبکه‌های عصبی با چندین لایه «پنهان» که به‌عنوان پرسپترون‌های چندلایه[3] (MLP) نیز شناخته می‌شوند، اغلب به‌عنوان یادگیری عمیق بدون نظارت شناخته می‌شوند (هستی و همکاران[4]، ۲۰۰۹ وگودفلو و همکاران ، ۲۰۱۶) بااین‌حال، هنگام تلاش برای نظارت بر تقلب بیمه، متغیر پیش‌بینی‌کننده مورد نظر در محیط‌های کسب‌وکار واقعی وجود ندارد. این امر مستلزم تأیید دستی به‌صورت موردی است. علاوه‌براین، کلاهبرداری بیمه پدیده‌ای پیچیده است که باوجود تحقیقات کامل شامل زمان و تلاش قابل‌توجه، تأیید نتایج آن با قطعیت مطلق تقریباً غیرممکن است. بنابراین، در معرض خطاهای ارتکاب، حذف و طبقه‌بندی نادرست است که می‌تواند عملکرد مدل را مختل کند (آرتیس و همکاران[5]، ۲۰۰۲).

استریپلینگ و همکاران[6] (۲۰۱۸) از جنگل‌های ایزوله - یک روش تشخیص ناهنجاری بدون نظارت که در بخش ۳-۱ با جزئیات بیشتر معرفی خواهد شد- برای تولید ویژگی‌هایی برای تشخیص تقلب در جبران خسارت کارگران استفاده می‌کنند.گومز و همکاران[7] (۲۰۲۱) رویکردی را برای تشخیص تقلب در حوزه‌های مختلف مانند مطالبات بیمه و پرداخت‌های کارت اعتباری مبتنی‌بر یادگیری عمیق بدون نظارت پیشنهاد می‌کنند. رویکرد آن‌ها همچنین امکان شناسایی مهم‌ترین متغیرها برای این‌کار را فراهم می‌کند (گومز و همکاران،۲۰۲۱). دووال و همکاران (۲۰۲۳) پیش‌تر از یادگیری بدون نظارت، ازجمله جنگل‌های ایزوله، برای استخراج پروفایل‌های ناهنجاری رفتار رانندگی استفاده

1- Goodfellow et al.
2- Niu et al.
3- Perceptron Multi Layer.
4- Hastie et al.
5- Artís et al.
6- Stripling et al.
7- Gomes et al.

می‌کنند و یک رابطه پیش‌بینی‌کننده با احتمال مطالبات بیمه خودرو شناسایی می‌کنند (دووال و همکاران[1]،۲۰۲۳).

با بهره‌گیری از قدرت تکنیک‌های تحلیلی پیشرفته، سازمان تأمین اجتماعی می‌تواند الگوها، روابط، رفتارها و ادعاهای مشکوکی که ممکن است نشان‌دهنده تقلب باشد، شناسایی کند. این قابلیت‌های تحلیلی پیشرفته، سازمان را قادر می‌سازد تا فعالیت‌های متقلبانه را به‌طور مؤثرتری شناسایی و از آن جلوگیری کند علاوه‌براین، با اصلاح و بهبود مستمر الگوریتم‌های تشخیص تقلب بر اساس ورودی‌های داده‌های جدید و روندهای تقلب در حال ظهور، سازمان می‌تواند از طرح‌های کلاهبرداری در حال تکامل جلوتر بماند و استراتژی‌های خود را برای مبارزه مؤثر با تقلب در یک‌چشم انداز پویا و پیچیده تطبیق دهد (ترزیی و همکاران[2]،۲۰۲۱).

۳-۲-۲. روش نیمه نظارتی

یادگیری نیمه‌نظارتی یکی از رویکردهای نوین و میانی در حوزهٔ یادگیری ماشین است که میان دو روش کلاسیک یادگیری تحت نظارت و یادگیری بدون نظارت قرار می‌گیرد. این رویکرد زمانی به‌کارگرفته می‌شود که تنها بخشی از داده‌ها دارای برچسب[3] باشند و بخش عمده‌ای از داده‌ها بدون برچسب باقی مانده باشند. در چنین شرایطی، الگوریتم از ترکیب هر دو نوع داده برای بهبود دقت مدل و استفاده بهینه از اطلاعات نهفته در داده‌های بدون برچسب بهره می‌گیرند (هادی و شوونکر[4]، ۲۰۱۳). الگوریتم‌های نیمه‌نظارتی به دو دستهٔ اصلی تقسیم می‌شوند: ۱- روش‌های مبتنی بر مدل[5]، ۲- روش‌های مبتنی‌بر گراف[6] یادگیری نیمه‌نظارتی از منظر کارایی محاسباتی و هزینه‌های داده‌سازی، نسبت به روش‌های صرفاً تحت نظارت برتری دارد، زیرا نیاز به فرایند برچسب‌گذاری گسترده را کاهش می‌دهد.

۳. روش‌شناسی پژوهش

تحقیقات قابل‌توجهی دررابطه‌با تشخیص تقلب در بیمه عمومی درگذشته وجود داشته است که بر تکنیک‌های داده‌کاوی و یادگیری ماشین تمرکز دارند (بندیک و همکاران[7]، ۲۰۲۲). محققان عمدتاً بر بیمه‌های تجاری (خودرو) و بیمه‌های درمانی (بیمه سلامت) تمرکز کرده‌اند. در این پژوهش برای اولین‌بار به بررسی وجود تقلب‌های نرم در دستمزدهای ارسالی کارفرمایان در سال‌های پایانی

1- Duval et al.
2- Terzi et al.
3- Label.
4- Hady & Schwenker.
5- Model-based.
6- Graph-based Methods.
7- Benedek et al.

بیمه‌پردازی پرداخته شده است. پژوهش حاضر ازنظر هدف، کاربردی و ازنظر ماهیت داده‌ها، توصیفی– تحلیلی و مبتنی‌بر داده‌های واقعی سازمان تأمین اجتماعی است. ازآنجاکه داده‌های مورداستفاده فاقد برچسب بوده و ماهیت ناهنجاری‌ها از پیش مشخص نیست در این پژوهش از پتانسیل سه الگوریتم یادگیری بدون نظارت پرکاربرد در تشخیص ناهنجاری شامل جنگل ایزوله، خوشه‌بندی مبتنی‌بر چگالی[1] و ماشین بردار پشتیبان تک کلاسه[2] در تشخیص تقلب برای شناسایی دستمزدهای نامتعارف استفاده شده است.

۳-۱. الگوریتم جنگل‌های ایزوله[3]

بیشتر رویکردهای رایج در تشخیص ناهنجاری، ابتدا الگوی رفتار عادی (پروفایل/مدل نرمال) را بر اساس داده‌های عادی می‌آموزند و سپس نمونه‌هایی را که انحراف معناداری از این الگو دارند، به‌عنوان ناهنجاری شناسایی می‌کنند. در این زمینه می‌توان به روش‌های آماری (وانگ و همکاران، ۲۰۱۹)، خوشه‌بندی (ابی و همکاران، ۲۰۰۶) و طبقه‌بندی در صورت در دسترس بودن برچسب‌ها (اسمایتی و همکاران،۲۰۲۰) اشاره کرد. درواقع در مدل‌های اشاره شده ابتدا رفتارهای نرمال، بررسی و شناسایی می‌شود و پس از آن، مواردی که از رفتار طبیعی انحراف و فاصله دارند به‌عنوان رفتار ناهنجار انتخاب می‌شوند. این رویکرد شاید برای مجموعه داده‌هایی با نمونه‌های محدود مشکلی ایجاد نکند؛ اما برای مجموعه داده‌های بزرگ، هم ازلحاظ اختصاص حافظه و هم زمان محاسبه، باعث ایجاد یک چالش بزرگ می‌شود. در رویکرد جنگل ایزوله این مشکل از طریق ایزوله کردن نقاط نمونه، قابل حل است. در الگوریتم جنگل ایزوله که یک مدل تجمعی[4] است بر اساس مسیری که منجر به ایزوله شدن یک مشاهده از نمونه می‌شود؛ نقطه قرمز را به‌عنوان ناهنجاری یا یک رفتار عادی در نظر می‌گیرد همان‌طورکه در شکل شماره ۱ قابل مشاهده است دو نقطه قرمز و آبی از دو مسیر متفاوت ایزوله شده‌اند و بر اساس الگوریتم جنگل ایزوله، نقاط ناهنجار به ریشه نزدیک‌تر هستند. بر همین اساس الگوریتم مزبور نقطه قرمز را به‌عنوان یک نمونه ناهنجار در نظر می‌گیرد . ازلحاظ ریاضی به هر یک از مشاهدات مجموعه داده‌ها یک امتیاز ناهنجاری اختصاص داده که از رابطه زیر محاسبه می‌شود:

$$S(x.n) = ۲^{\frac{-E(h(x))}{c(n)}} \quad (۱$$

1- -DBSCAN.
2--One class SVM.
3- -Isolation Forest.
4- -cumulative.

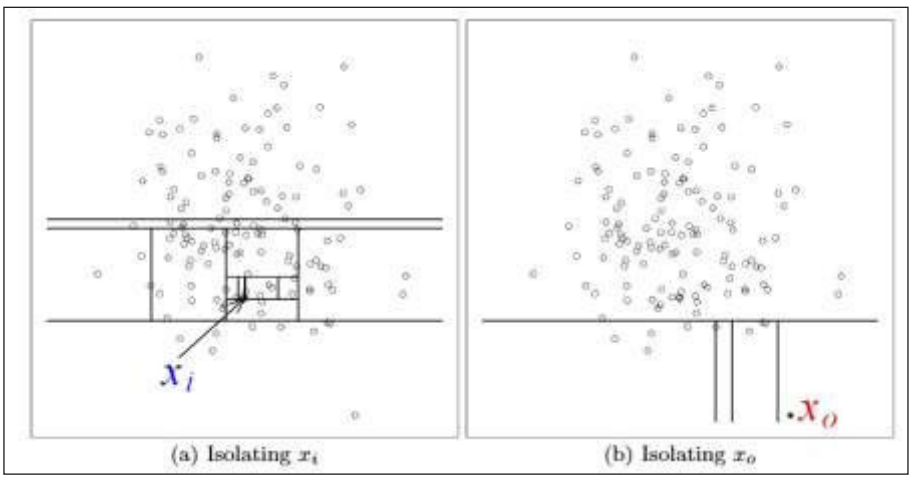
n تعداد نمونه و c(n)، میانگین طول مسیرهای ناموفق در درخت جستجوی دودویی بوده و از رابطه زیر محاسبه می‌شود:

$$C(n)=2H(n-1)-2\frac{(n-1)}{n} \quad (2$$

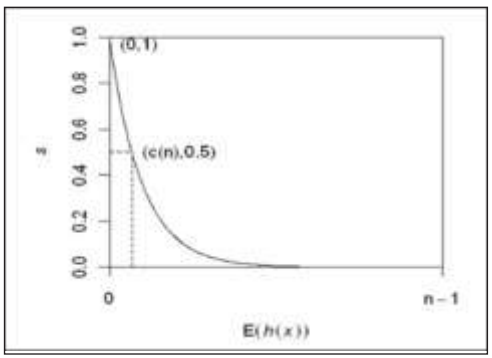
در معادله بالا H(i) عدد هارمونیک بوده که می‌توان آن را با ثابت اویلر به‌صورت زیر تخمین زد:

$$Ln(i)+ 0,5772156649 \quad (3$$

و h(x) طول مسیر منتهی به نقطه x است.



شکل ۱. افرازبندی جنگل ایزوله



شکل ۲. رابطه بین h(x) و s

۳-۲. الگوریتم خوشه‌بندی مبتنی‌بر چگالی[1]

الگوریتم خوشه‌بندی مبتنی‌بر چگالی توسط ایستر[2] و همکارانش در سال ۱۹۹۶ پیشنهاد شد و برای خوشه‌بندی داده‌های با اشکال دلخواه در حضور اختلال در پایگاه‌های داده با ابعاد بالا، چه مکانی و چه غیر مکانی، طراحی‌شده است. ایده اصلی الگوریتم خوشه‌بندی مبتنی‌بر چگالی این است که برای هر شیء از یک خوشه، همسایگی با شعاع معین[3] باید حداقل تعداد کمی از اشیاء[4] را در بربگیرد، به این معنی که کاردینالیتی همسایگی باید از یک آستانه بیشتر باشد. همسایگی یک نقطه دلخواه p به‌صورت زیر تعریف می‌شود:

$$N_{Eps} = \{q \in D / dist \in (p.q) < Eps)\} \quad (۱$$

در اینجا، D پایگاه داده اشیاء است. اگر θ همسایگی‌های یک نقطه P حداقل شامل تعداد کمینه‌ای از نقاط باشند، آنگاه این نقطه، نقطه مرکزی نامیده می‌شود. نقطه مرکزی به‌صورت زیر تعریف می‌شود:

$$N_{Eps}(P) > MinPts \quad (۲$$

در اینجا Eps و MinPts پارامترهای مشخص‌شده توسط کاربر هستند که به‌ترتیب به معنای شعاع همسایگی و حداقل تعداد نقاط در همسایگی یک نقطه مرکزی هستند. اگر این شرط برقرار نباشد، این نقطه به‌عنوان نقطه غیرمرکزی در نظر گرفته می‌شود.

الگوریتم خوشه‌بندی مبتنی‌بر چگالی (DBSACN) با بررسی همسایگی هر شیء در مجموعه داده‌ها، خوشه‌ها را جستجو می‌کند. اگر همسایگی یک شیء p شامل بیش از MinPts باشد، یک خوشه جدید با p به‌عنوان شیء اصلی ایجاد می‌شود. سپس به‌صورت تکراری اشیاء قابل دسترسی مستقیم به چگالی را از این اشیاء اصلی جمع‌آوری می‌کند که ممکن است شامل ادغام یک خوشه قابل دسترسی چگالی جدید باشد. این فرایند زمانی پایان می‌یابد که هیچ شیء جدیدی نتواند به هیچ خوشه‌ای اضافه شود (زانگ[5]، ۲۰۱۳).

1- DBSCAN.
2- Ester.
3- Eps.
4- MinPts.
5- Zhang.

۳-۳. ماشین بردار پشتیبان تک کلاسه[1]

یکی از تکنیک‌های یادگیری بدون نظارت، ماشین بردار پشتیبان تک کلاسه (OCSVM) است که برای تشخیص داده‌های پرت و تشکیل یک فرایند یادگیری افزایشی استفاده می‌شود. کاربرد آن در تشخیص ناهنجاری به طور گسترده در سراسر جهان مانند تشخیص داده‌های پرت، تشخیص داده‌های جدید و بسیاری موارد دیگر استفاده می‌شود. ماشین بردار پشتیبان تک کلاسه (OCSVM) به‌عنوان یک یادگیرنده تک کلاسه از SVM اصلاح شده است که سعی می‌کند یک ابر کره را در میان نمونه‌های کلاس‌های عادی پیدا کند. این مدل داده‌های جدید را به‌صورت عادی یا غیرعادی طبقه‌بندی می‌کند، تمام مشاهدات داخل ابر کره عادی هستند و مشاهدات خارج از ابر کره غیرعادی هستند. یک مزیت قابل‌توجه ماشین‌های بردار پشتیبان (SVM) قابلیت آن‌ها در تولید یک مرز تصمیم‌گیری غیرخطی با تبدیل داده‌ها از طریق یک نگاشت غیرخطی به یک فضای ویژگی با ابعاد بالاتر F است. در این فضای ویژگی، می‌توان کلاس‌ها را با یک ابر صفحه جدا کرد، حتی اگر یک مرز خطی در فضای ورودی اصلی I امکان‌پذیر نباشد. این فرایند منجر به یک منحنی غیرخطی در فضای ورودی می‌شود، زمانی‌که ابر صفحه به عقب تصویر می‌رود. با استفاده از یک هسته چندجمله‌ای برای تصویرسازی، تمام نقاط به بعد سوم منتقل می‌شوند و می‌توان از یک ابر صفحه برای جداسازی استفاده کرد. هنگامی‌که محل تقاطع صفحه با فضا به فضای دوبعدی تصویر می‌شود، منجربه یک مرز دایره‌ای می‌شود.

۳-۴. مقایسه روش‌های تشخیص ناهنجاری و انتخاب الگوریتم جنگل‌های ایزوله

برای تشخیص موارد غیرعادی در داده‌های پژوهش سه الگوریتم جنگل ایزوله، ماشین بردار پشتیبان تک کلاسه و الگوریتم خوشه‌بندی مبتنی بر چگالی آزمایش شدند. بر اساس مبانی نظری و شواهد موجود در متون علمی، الگوریتم جنگل ایزوله عملکرد بهتری روی این نوع داده نشان داد. در ادامه، به دلایل برتری جنگل ایزوله در مقایسه با دو روش دیگر می‌پردازیم:

۳-۴-۱. کارایی و مقیاس‌پذیری

الگوریتم جنگل ایزوله به‌خاطر پیچیدگی زمانی خطی خود نسبت به تعداد داده‌ها تقریباً بسیار مقیاس‌پذیر است. این روش با استفاده از مجموعه‌ای از درخت‌های تصمیم تصادفی، نقاط دور افتاده را با عمق کم در درخت‌ها جداسازی می‌کند و از این طریق امتیاز ناهنجاری به هر نمونه اختصاص می‌دهد. تحقیقات نشان داده که جنگل ایزوله به طور کارآمدی در داده‌های با ابعاد بالا نیز قابل استفاده است و دقت بالایی در شناسایی موارد نادر دارد. به‌علاوه، این روش برای انواع مختلف ناهنجاری (نقطه‌ای،

1- One class SVM.

تجمعی، زمینه‌ای) انعطاف‌پذیر و مؤثر گزارش شده است (هاوون و لبو[1]، ۲۰۲۵). این نقاط قوت به‌ویژه در داده‌های دستمزد بیمه‌شدگان که حجم بالا و ویژگی‌های محدودی (۳ تا ۴ متغیر) دارند اهمیت دارد، چراکه مدل می‌تواند به‌سرعت آموزش ببیند و ناهنجاری‌ها را حتی در صورت چندبعدی بودن داده‌ها شناسایی کند.

۳-۴-۲. حساسیت کمتر به ابر پارامترها و عدم نیاز به برآورد دقیق نرخ آلودگی

الگوریتم جنگل ایزوله نسبت به تنظیمات ابرپارامتر کمتر حساس گزارش شده است. برخلاف الگوریتم خوشه‌بندی مبتنی بر چگالی که نیازمند تنظیم دقیق پارامترهایی مانند epsilon (شعاع همسایگی) و حداقل نقاط برای تشکیل خوشه است، جنگل ایزوله به‌صورت پیش‌فرض با تعداد درخت‌ها و اندازه زیر نمونه‌ها کار می‌کند و وابستگی کمتری به تنظیم دستی دارد. همچنین ماشین بردار پشتیبان احتیاج به تنظیم پارامتر nu نسبت موارد خارج از کلاس و انتخاب کرنل مناسب دارد که در عمل بسیار حساس و چالش‌برانگیز است. منابع علمی تأکید می‌کنند که SVM یک کلاسه به حضور داده‌های پرت بسیار حساس است و عملکرد مطلوبی در سناریوهای کاملاً بدون نظارت که مجموعه آموزش آلوده به موارد ناهنجار ناشناخته است، ندارد (هاوون و لیو، ۲۰۲۵) درواقع، ماشین بردار پشتیبان بیشتر برای تشخیص نوظهوری[2] کاربرد دارد که در آن مدل بر روی داده‌هایی کاملاً پاک آموزش می‌بیند؛ اما در مسئله ما داده آموزش، خود ممکن است شامل دستمزدهای نامتعارف باشند. در مقابل جنگل ایزوله می‌تواند بدون نیاز به دانستن دقیق درصد نرخ آلودگی[3] در ابتدا، یک نمره ناهنجاری برای هر رکورد تولید کند؛ سپس کاربر یا تحلیلگر می‌تواند بر اساس توزیع این نمرات، آستانه‌ای مناسب برای تفکیک موارد مشکوک تعیین نماید . به همین دلیل، جنگل ایزوله در مواجه با نرخ آلودگی ناشناخته، انعطاف‌پذیری بیشتری دارد و نیازمند فرض اولیه کمتری درباره درصد موارد تقلب است. این ویژگی یک مزیت مهم در داده‌های دستمزد بیمه‌شدگان است که نرخ دقیق تخلفات از پیش معلوم نیست.

۴. داده‌های تحقیق و ابزارهای تحلیل

این پژوهش باهدف شناسایی ناهنجاری در سری‌های زمانی دستمزد بیمه‌شدگان سازمان تأمین اجتماعی طراحی شده است. ناهنجاری‌ها در این زمینه اغلب به‌صورت افزایش غیرمتعارف دستمزد در دوره‌های منتهی به بازنشستگی مشاهده می‌شوند که می‌تواند بر میزان مستمری و عدالت بیمه‌ای اثرگذار باشد. بدین منظور این پژوهش برای اولین بار در کشور با داده‌های دستمزدی واقعی از تعدادی از

1- Haowen &lu.
2- novelty detection.
2- Pollution rate.

بیمه‌شدگان اجباری[1] استان اصفهان انجام گردید داده‌ها شامل آیتم‌های دستمزد، کارکرد ماهیانه و شماره کارگاه در طول فاصله زمانی فروردین ۱۳۹۸ لغایت اسفند ۱۴۰۲ در ۴۴ شعبه استان اصفهان است. باتوجه‌به رویکرد گذشته‌نگر پژوهش داده‌ها از بیمه‌شدگانی که در سال ۱۴۰۲ بازنشسته شده‌اند انتخاب گردید و از طرفی باتوجه‌به محرمانه بودن اطلاعات بیمه‌ای افراد و کارگاه‌ها[2] و قوانین جاری سازمان نسبت به اخذ مجوزهای مربوطه برای استخراج داده‌ها با استفاده از کدهای sql از پایگاه داده سازمان تأمین اجتماعی اقدام گردید. در مرحله پیش‌پردازش، ابتدا داده‌های تکراری و فاقد کارکرد حذف شدند. سپس به‌منظور هم‌مقیاس‌سازی متغیرها، از روش نرمال‌سازی Min-Max استفاده گردید. علاوه‌براین، ویژگی‌های داده نظیر دستمزد روزانه، نسبت دستمزد روزانه به حداقل دستمزد و درصد نسبت دستمزد روزانه به حداکثر دستمزد سالیانه اعلامی شورای دستمزد کشور محاسبه شد تا داده‌ها جهت اعمال سه الگوریتم یادگیری ماشین مربوطه آماده گردد.

روش پیشنهادی بر اساس دو ویژگی نسبت دستمزد به حداقل و نسبت دستمزد بیمه‌شده به حداکثر روزانه قانونی طراحی‌شده است. اول اینکه طبق قوانین کشور در ابتدای هرسال شورای‌عالی کار دستمزد حداقل و حداکثر روزانه را برای بیمه‌شدگان مشمول قانون کار و تأمین اجتماعی و دوم، نرخ افزایش مزایای جانبی را در ابتدای هرسال اعلام می‌نماید که کارفرمایان موظف به تبعیت از دستمزدها و مزایای اعلامی هستند.

۴-۱. تحلیل توصیفی داده‌ها

تحلیل توصیفی امکان شناسایی الگوها، روندها و روابط در داده‌ها را فراهم می‌کندکه به نتیجه‌گیری‌های مهم و تصمیم‌گیری‌های آگاهانه کمک می‌نماید. این پژوهش برای اولین‌بار با استفاده از داده‌های واقعی و گذشته‌نگرِ مربوط به پرداخت‌های حق‌بیمه انجام شد که شامل سری زمانی پرداخت پنج‌سالهٔ ۴۷۰ مستمری‌بگیر است. مجموع داده‌ها شامل ۲۶۲۵۸ رکورد ماهانهٔ حق‌بیمه در بازهٔ زمانی فروردین ۱۳۹۸ تا اسفند ۱۴۰۲[3] از ۳۳ کلاس کد مختلف می‌باشد.

داده‌های واقعی اغلب با مشکلات متعددی همچون داده‌های از دست رفته، دورافتاده و غیرعادی مواجه هستند (فان و همکاران، ۲۰۲۱). بنابراین، داده‌های واقعی هرگز به خوبی آن‌گونه که انتظار داریم، نیستند و از مسائل گوناگونی رنج می‌برند که می‌توانند بر تفسیر و پردازش سیستم‌ها و مدل‌های ساخته‌شده بر اساس آن داده‌ها تأثیر بگذارند. برای نمونه، سیستم‌های تصمیم‌گیری به شدت تحت تأثیر

۱- بیمه‌شدگانی‌که دارای کارفرما بوده و حق‌بیمه ایشان از طریق لیست بیمه به سازمان ارسال می‌گردد.
۲- باتوجه‌به محرمانه بودن داده‌ها از آوردن اسم اصناف خودداری شده است و به صورت کد کلاسه معرفی شده‌اند.
۳- باتوجه‌به محرمانگی داده‌های بیمه‌پردازی بیمه‌شدگان امکان نمایش داده‌های استخراجی نبود.

کیفیت داده‌ها قرار می‌گیرند. فساد یا آسیب داده‌ها می‌تواند ناشی از عوامل متعددی از جمله خطاهای انسانی و ذاتی باشد (میچاو و فینک،۲۰۲۱).

جدول شماره ۱ مصوبات شورای عالی کار را در طول سال‌های پژوهش که شامل حداقل دستمزد روزانه و حداقل و حداکثر دستمزد ماهیانه است، نشان می‌دهد.

جدول شماره ۱. حداقل و حداکثر دستمزد قانونی (به ریال)

سال	دستمزد روزانه	حداقل دستمزد ماهیانه	حداکثر دستمزد ماهیانه
۱۳۹۸	۵۰۵۶۲۷	۱۵۱۶۸۸۱۰	۱۰۶۱۸۱۶۷۰
۱۳۹۹	۶۳۶۸۰۹	۱۹۱۰۴۲۷۰	۱۳۳۷۲۹۸۹۰
۱۴۰۰	۸۸۵۱۶۵	۲۶۵۵۴۹۵۰	۱۸۵۸۸۴۶۵۰
۱۴۰۱	۱۳۹۳۲۵۰	۴۱۷۹۷۵۰۰	۲۹۲۵۸۲۵۰۰
۱۴۰۲	۱۷۶۹۴۲۸	۵۳۰۸۲۸۴۰	۳۷۱۵۷۹۸۸۰

در سازمان برای دسته‌بندی فعالیت‌های اقتصادی و شناسایی اصناف از طرح کدینگ اطلاعات کارگاهی استفاده می‌شود که هر صنف و زیر فعالیت‌های آن دارای یک کد اختصاصی هستند جدول شماره ۲ اطلاعات صنفی و تعداد بیمه‌شدگان را در داده‌های پژوهش نشان می‌دهد.

جدول شماره ۲. کد اصناف و تعداد بیمه‌شده در هر صنف

ردیف	کد کلاسه صنف	تعداد بیمه‌شده (نفر)	ردیف	کد کلاسه صنف	تعداد بیمه‌شده (نفر)
۱	۲۳۳	۱۰۱	۱۸	۲۳۲	۹
۲	۸۱۳	۳۸	۱۹	۵۲۱	۸
۳	۲۳۱	۳۷	۲۰	۳۸۳	۸
۴	۸۲۱	۳۲	۲۱	۴۰۹	۸
۵	۸۲۴	۲۰	۲۲	۱۴۴	۷
۶	۳۶۳	۱۹	۲۳	۷۳۲	۷
۷	۹۹۹	۱۴	۲۴	۳۶۲	۷
۸	۴۱۰	۱۳	۲۵	۷۱۲	۶
۹	۸۵۱	۱۲	۲۶	۶۶۷	۶
۱۰	۸۲۶	۱۲	۲۷	۳۷۱	۶
۱۱	۸۲۲	۱۱	۲۸	۲۱۲	۶
۱۲	۳۲۱	۱۰	۲۹	۲۳۷	۶
۱۳	۲۱۶	۱۰	۳۰	۸۴۱	۶

ردیف	کد کلاسه صنف	تعداد بیمه‌شده (نفر)	ردیف	کد کلاسه صنف	تعداد بیمه‌شده (نفر)
۱۴	۱۰	۹	۳۱	۳۴۲	۵
۱۵	۱۷۳	۹	۳۲	۵۲۲	۵
۱۶	۳۵۴	۹	۳۳	۲۵۴	۵
۱۷	۸۱۴	۹			

مهندسی ویژگی[1] یکی از مراحل کلیدی در فرایند داده‌کاوی و یادگیری ماشین است که نقش تعیین‌کننده‌ای در بهبود عملکرد مدل‌های پیش‌بینی دارد (کوهن و جانسون[2]، ۲۰۱۹). در این راستا ویژگی‌های[3] ذیل را تعریف و محاسبه کردیم.

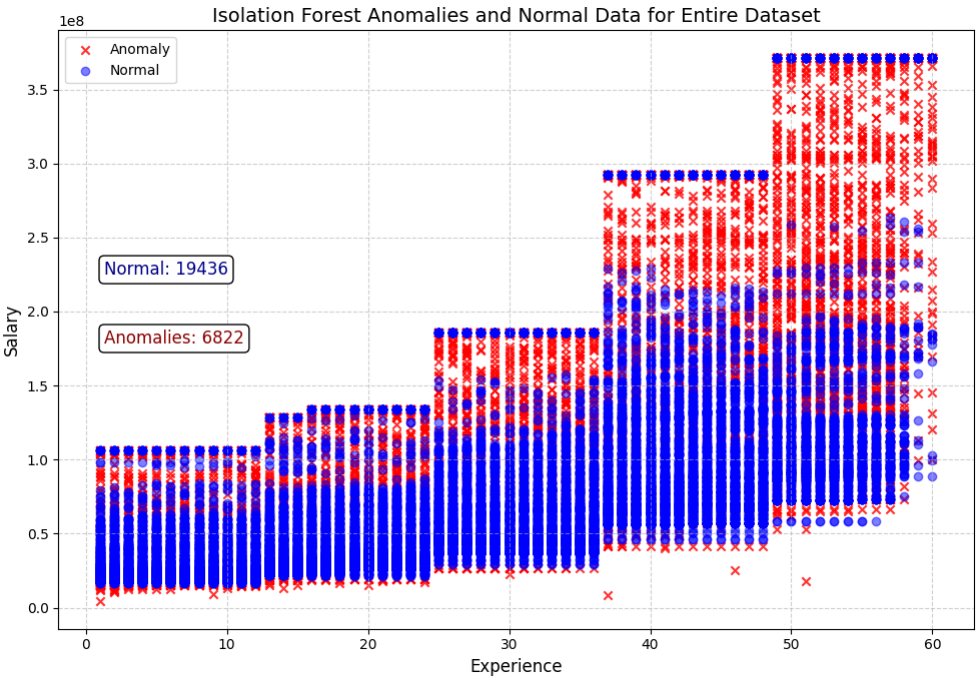
salary- per- day متوسط دستمزد روزانه (تقسیم دستمزد ماهیانه بر کارکرد ماهیانه)

salary-ratio نسبت حقوق دریافتی نسبت‌به حداقل دستمزد قانونی (تقسیم دستمزد ماهیانه بیمه‌شده به حداقل ماهیانه قانون کار) (جدول ۱)

max- ratio نسبت دستمزد روزانه به حداکثر روزانه قانونی (دستمزد ماهیانه بیمه‌شده تقسیم‌بر حداکثر دستمزد ماهیانه قانونی ضربدر ۱۰۰) (جدول ۱)

این ویژگی‌ها به‌دلیل ارتباط مستقیم با تحلیل دستمزد و امکان مقایسه بین بیمه‌شدگان انتخاب شدند. آماده‌سازی داده‌ها در برنامه پایتون با کتابخانه‌های numpy و pandas انجام شد.

1- Feature Engineering.
2- Kuhn & Johnson.
3- Features.



شکل ۳. خروجی الگوریتم جنگل ایزوله روی کل داده‌ها

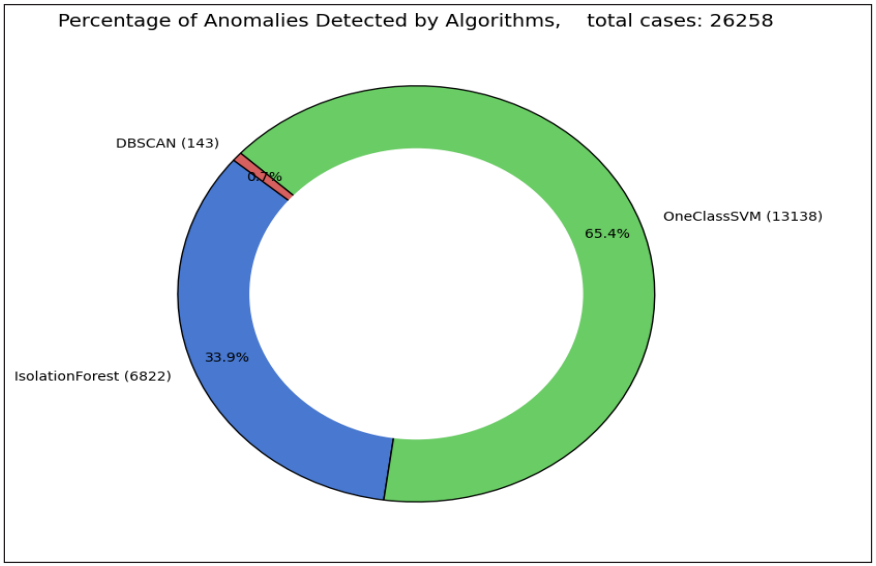
پس از آماده‌سازی داده‌ها، الگوریتم جنگل ایزوله بر روی کل مجموعه داده به‌صورت یکپارچه اجرا شد تا دید کلی از توزیع دستمزدها به دست آید (شکل ۳). هدف بررسی این فرضیه است که آیا با یک مدل واحد می‌توان تمام داده‌های نرمال را از داده‌های ناهنجار[1] تفکیک کرد. همان‌طورکه در شکل ۳ مشاهده می‌شود، پس از اجرای الگوریتم، داده‌های نرمال (آبی) و ناهنجار (قرمز) هم‌پوشانی و درهم‌تنیدگی بسیار شدیدی دارند. این عدم تفکیک‌پذیری به روشنی نشان می‌دهد که الگوهای دستمزد بیمه‌شدگان در میان همه آن‌ها یکسان نیست؛ زیرا داده‌ها شامل اصناف متفاوتی هستند که هر یک شرایط کاری و دستمزد و مزایای مختص به صنعت خود را دارند.

نتیجه‌گیری کلیدی از این مرحله آن است که تحلیل یکپارچه داده‌ها منجربه شناسایی نادرست ناهنجاری‌ها می‌شود؛ بنابراین استراتژی اصلی پژوهش بر تفکیک داده‌ها بر اساس اصناف مجزا و اجرای الگوریتم برای هر صنف به‌طور جداگانه بنیان نهاده شد تا الگوهای خاص هر گروه شغلی به دقت تحلیل شده و ناهنجاری‌ها به شکل قابل‌اعتمادی شناسایی گردند.

1- Anomaly.

۴-۲. ارزیابی مقایسه‌ای و انتخاب الگوریتم بهینه

با اجرای دو الگوریتم خوشه‌بندی مبتنی‌بر چگالی و ماشین بردار پشتیبان تک کلاسه روی داده‌های پژوهش مشخص شد که الگوریتم DBSCAN در تشخیص موارد تک گردایه[1] چندان خوب عمل نکرد مگر آنکه epsilon را کوچک انتخاب می‌کردیم که در آن صورت بخش عمده‌ای از داده‌های عادی نیز پرت شناسایی می‌شدند. این رفتار قابل انتظار است، چراکه DBSCAN بر مبنای چگالی خوشه‌ها کار می‌کند و اگر ناهنجاری‌ها به‌صورت نقاط بدون خوشه و نسبتاً دور از بقیه نباشند یا چگالی داده نرمال یکنواخت نباشد، تنظیم یک epsilon ثابت دشوار است. ازسوی‌دیگر، One-class SVM نیز در داده پژوهش که دارای ابعاد کمی بود، دچار overfit روی نقاط پرت شد و مرز تصمیم آن نتوانست به‌درستی تمامی موارد عادی را پوشش دهد درنتیجه موارد عادی متعددی اشتباهاً پرت شناسایی شدند. این مشاهده نیز مطابق گزارش‌های موجود است که ماشین بردار تک کلاسه در نبود داده‌های برچسب خورده و در حضور توزیع‌های پیچیده، پایدار عمل نمی‌کند. (شکل ۴)



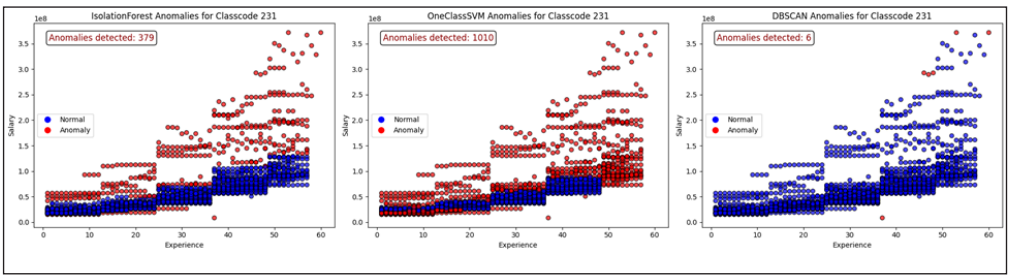
شکل ٤. خروجی سه الگوریتم بر روی داده‌های پژوهش

به‌طورکلی، مطالعات جامع نشان می‌دهند هر یک از این الگوریتم‌ها نقاط قوت خاص خود را دارند اما اگر به‌تنهایی استفاده شوند، از حیث حساسیت به تنظیمات و مقیاس‌پذیری ممکن است دچار مشکل شوند. در نقطه مقابل، جنگل ایزوله به‌عنوان رویکرد پیش‌فرض مؤثرتر برای بسیاری از انواع داده مطرح

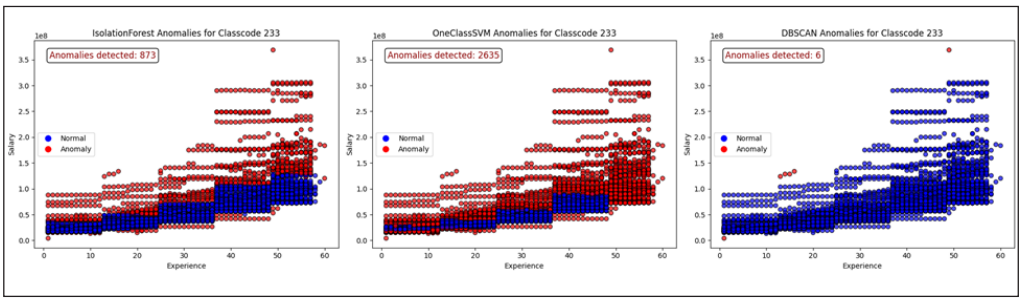
1- anomalies point.

شده است، چراکه بدون نظارت خاصی و با حداقل تنظیمات می‌تواند طیف متنوعی از ناهنجاری‌ها را آشکار سازد. براین‌اساس و باتوجه‌به نتایج به‌دست‌آمده، جنگل ایزوله برای مسئله تشخیص دستمزدهای نامتعارف گزینه مناسب‌تری ارزیابی شد.

ازآنجایی‌که داده‌های پژوهش فاقد برچسب مشخصی برای تفکیک داده‌های عادی از غیرعادی است در این پژوهش از یادگیری غیر نظارتی استفاده شد ارزیابی مدل‌ها به‌جای معیارهای دقت‌سنجی عددی، بر اساس تحلیل بصری و منطق خروجی انجام گرفت. هدف، یافتن الگوریتمی بود که الگوی اصلی داده‌ها را به بهترین شکل شناسایی کرده و انحرافات معنادار از آن الگو را به‌عنوان ناهنجاری معرفی کند. با اعمال سه الگوریتم روی سه کلاس کد ۲۳۳ و ۲۳۱ و ۸۱۳ با بیشترین بیمه‌شده خروجی نمودارهای ۵، ۶، ۷ به دست آمد با بررسی خروجی الگوریتم‌ها مشخص می‌گردد که الگوریتم جنگل ایزوله نسبت به دو الگوریتم one -class SVM, DBSCAN بهتر عمل کرده و ناهنجاری‌های دستمزدی بیمه‌شدگان را منطقی‌تر شناسایی کرده است. برای مثال، در صنف ۲۳۳ الگوریتم ماشین بردار تک کلاسه با شناسایی ۲۶۳۵ ناهنجاری، بیش از حد حساس[1] عمل کرده است. همان‌طورکه مشخص است، بسیاری از نقاطی که در روند طبیعی دستمزد قرار دارند نیز به‌عنوان ناهنجار برچسب خورده‌اندکه این امر منجر به نرخ بالای هشدارهای غلط[2] شده است.

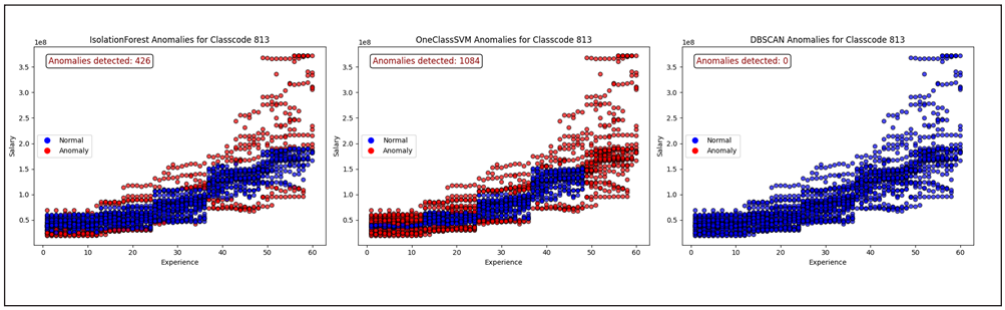


شکل ۵. خروجی سه الگوریتم روی داده‌های صنف ۲۳۱



شکل ۶. خروجی سه الگوریتم روی داده‌های صنف ۲۳۳

1- Overly Sensitive.
2- False Positives.



شکل ۷. خروجی سه الگوریتم روی‌داده‌های صنف ۸۱۳

الگوریتم خوشه‌بندی مبتنی‌بر چگالی، با شناسایی تنها ۶ ناهنجاری، بیش‌ازحد محافظه‌کار[1] بوده است. این مدل تقریباً تمام داده‌ها را به‌عنوان یک خوشه واحد در نظرگرفته و تنها نقاط بسیار دورافتاده را ناهنجار تشخیص داده و درنتیجه، بسیاری از موارد مشکوک را نادیده می‌گیرد.

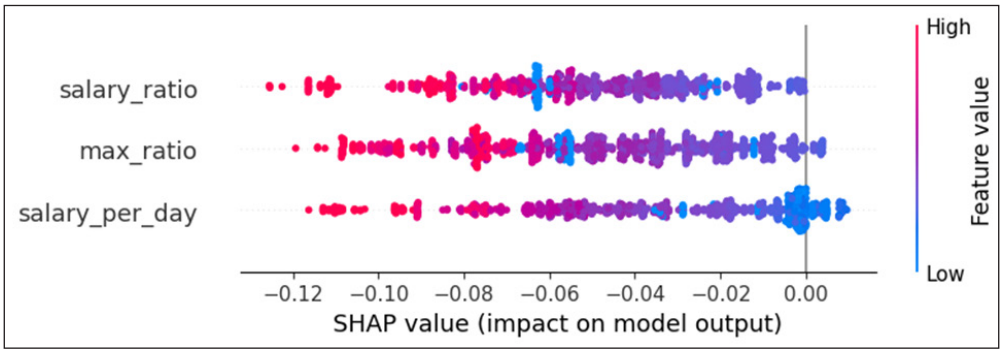
الگوریتم جنگل ایزوله با شناسایی ۸۷۳ ناهنجاری، متعادل‌ترین و منطقی‌ترین عملکرد را از خود به نمایش گذاشت. جنگل ایزوله به‌خوبی توانسته ساختار پلکانی داده‌ها را درک کرده و نقاطی را که به‌طور واضح بالاتر یا پایین‌تر از الگوی دستمزدی متناسب با سابقه کار قرار دارند، به‌عنوان ناهنجاری شناسایی کند.

با مقایسه نتایج و خروجی الگوریتم‌ها، الگوریتم جنگل ایزوله به‌دلیل ارائه خروجی قابل‌تفسیر، متعادل و همسو با منطق کسب‌وکار، به‌عنوان الگوریتم نهایی و بهینه برای شناسایی دستمزدهای نامتعارف در این پژوهش انتخاب گردید.

۴-۳. تفسیرپذیری مدل با SHAP

شناسایی عوامل کلیدی در تشخیص ناهنجاری برای اطمینان از اینکه مدل جنگل ایزوله بر اساس معیارهای منطقی تصمیم‌گیری می‌کند و صرفاً یک «جعبه سیاه» نیست، از ابزار قدرتمند SHAP[2] استفاده شد. این تکنیک به ما نشان می‌دهد که هر یک از ویژگی‌های داده چقدر در ناهنجار یا نرمال تشخیص دادن یک دستمزد تأثیر داشته‌اند. این برنامه تأثیر هر یک از متغیرها را در ناهنجاری بررسی و نشان می‌دهد (شکل ۸).

1- Overly Conservative.
2- Shapley Additive exPlanations.



شکل ۸. خروجی برنامه SHAP

محور عمودی (نمودار ۸) فهرستی از ویژگی‌ها یا عواملی را نشان می‌دهدکه مدل برای تصمیم‌گیری استفاده کرده است. این لیست بر اساس اهمیت مرتب شده است، محور افقی (SHAP value) میزان و جهت تأثیر یک ویژگی را نشان می‌دهد. نقطه صفر(خط عمودی وسط) نقطه بی‌اثر یا مبنا است رنگ هر نقطه، مقدار واقعی آن ویژگی را نشان می‌دهد.

قرمز: یعنی مقدار آن ویژگی برای آن داده بالا[1] بوده است.

آبی: یعنی مقدار آن ویژگی پایین[2] بوده است.

حرکت به سمت چپ (اعداد منفی) هرچه یک نقطه به سمت چپ‌تر باشد، یعنی آن ویژگی با آن مقدار خاص، تصمیم مدل را به سمت ناهنجار بودن سوق داده است .حرکت به سمت راست (اعداد مثبت) یعنی آن ویژگی تصمیم را به سمت نرمال بودن سوق می‌دهد. در این نمودار، تقریباً تمام تأثیرات قوی در سمت چپ (منفی) قرار دارند که منطقی است؛ چون ما به دنبال دلایل ناهنجار بودن هستیم.

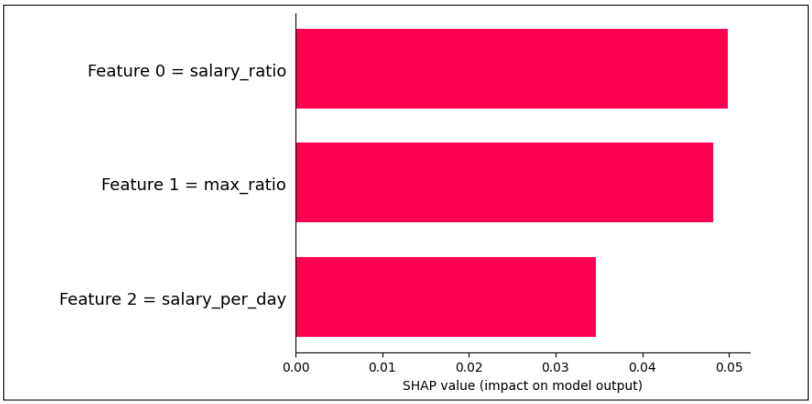
salary_ratio (نسبت دستمزد روزانه بیمه‌شده به حداقل دستمزد روزانه قانونی) در بالاترین جایگاه قرار دارد، پس مهم‌ترین و تأثیرگذارترین عامل است.)

max_ratio نسبت دستمزد روزانه بیمه‌شده به حداکثر دستمزد قانونی) دومین عامل مهم است.

salary_per_day دستمزد روزانه سومین عامل مهم است.

نتیجه‌گرفته می‌شود که مدل، بیش از هر چیز به salary_ratio برای تشخیص ناهنجاری توجه می‌کند (شکل ۹).

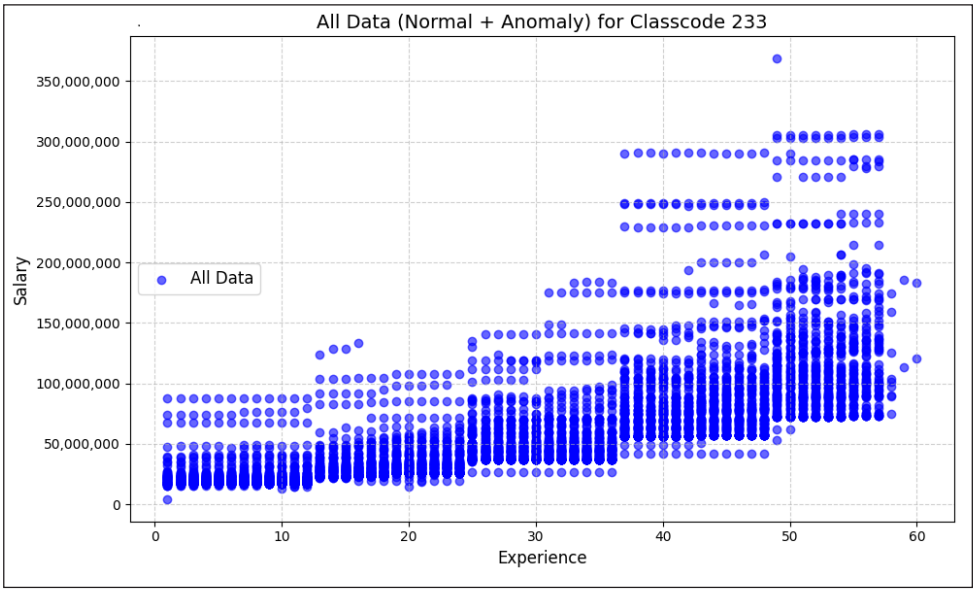
1- High.
2- Low.



شکل ۹. مقدار خروجی برنامه SHAP

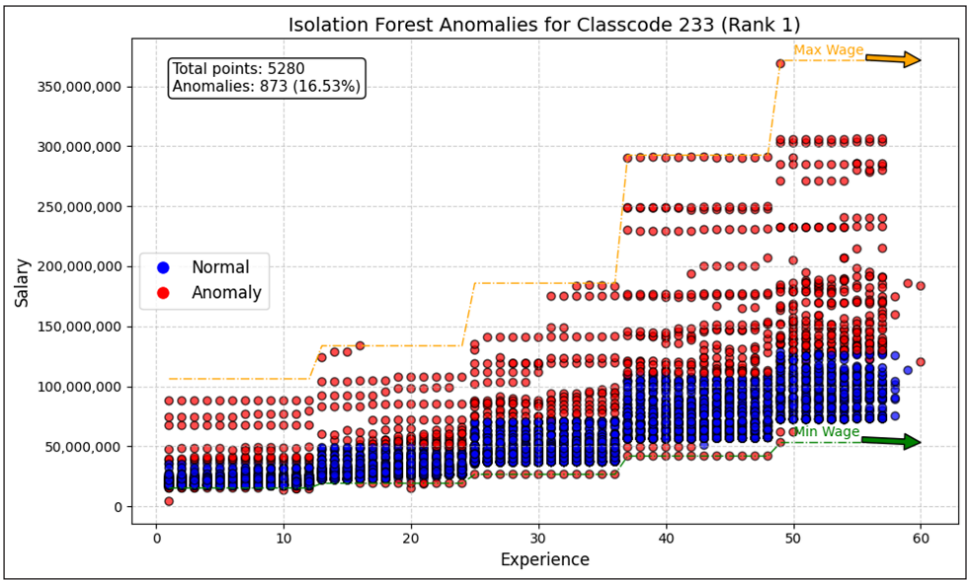
۴-۴. تحلیل خروجی‌ها در صنف منتخب

باتوجه‌به اینکه داده‌های پژوهش برای ۳۳ صنف است و در بخش قبلی توضیح داده شد که استراتژی اصلی پژوهش بر تفکیک داده‌ها بر اساس اصناف مجزا و اجرای الگوریتم برای هر صنف جداگانه بنا نهاده شد تا الگوهای خاص هر گروه شغلی به‌دقت تحلیل شود و ناهنجاری‌ها شناسایی شوند. با این توضیحات، تحلیل همه اصناف در این مقاله به علت محدودیت حجم مقدور نیست، لذا موردی، تحلیل مربوط به صنف با کد کلاسه ۲۳۳ که بیشترین فراوانی را دارد انجام می‌گیرد. نمودار کلی بیمه‌شدگان صنف ۲۳۳ در نمودار شماره ۱۰ رسم شده است. نمودار ۱۰، توزیع خام داده‌های دستمزد در برابر سابقه کار را منحصراً برای افراد صنف ۲۳۳ نمایش می‌دهد. برخلاف نمودار درهم‌تنیده شکل ۴، در اینجا یک الگوی منطقی و ساختاریافته به‌وضوح قابل مشاهده است که در آن، با افزایش سابقه کار، دستمزد نیز به‌صورت پلکانی افزایش می‌یابد. این ساختار منظم، یک مبنای بسیار قابل اعتماد برای الگوریتم فراهم می‌کند تا بتواند رفتار «نرمال» دستمزد را برای این گروه شغلی خاص بیاموزد و در مراحل بعد، هر داده‌ای را که از این الگوی مشخص پیروی نمی‌کند، به‌عنوان یک ناهنجاری بالقوه شناسایی نماید.



شکل ۱۰. اسکاتر پلات داده‌های صنف ۲۳۳

نمودار شماره ۱۰ اسکاتر پلاتر داده‌های صنف ۲۳۳ را در بازه ۶۰ ماهه با مشخص کردن نرمال و غیر نرمال هر نقطه نشان می‌دهد.



شکل ۱۱. اسکاتر پلات ناهنجاری‌های صنف ۲۳۳

۴-۵. پیاده‌سازی نهایی و تعریف کریدور حقوق نرمال (صنف منتخب)

در این مرحله، الگوریتم منتخب جنگل ایزوله بر روی داده‌های تفکیک‌شدهٔ صنف ۲۳۳ پیاده‌سازی شد تا مرز بین حقوق‌های متعارف و نامتعارف به‌صورت دقیق مشخص شود.

همان‌طورکه در نمودار نهایی (شکل ۱۲) مشاهده می‌شود، مدل با موفقیت یک «کریدور حقوق نرمال» را بر اساس سابقه کار تعریف کرده است. از مجموع ۵۲۸۰ رکورد در این صنف، الگوریتم توانست ۸۷۳ مورد (معادل ۱۶٫۵۳درصد) را به‌عنوان ناهنجاری شناسایی کند.

خطوط مرزی[1] این خطوط که به‌صورت بصری در نمودار مشخص شده‌اند، حد بالا و پایین دستمزد متعارف را برای هر سطح از سابقه کار نشان می‌دهند. این مرزها توسط قوانین پرداخت بیمه و بر اساس ساختار دستمزدی بیمه‌شدگان همین صنف آموخته شده‌اند.

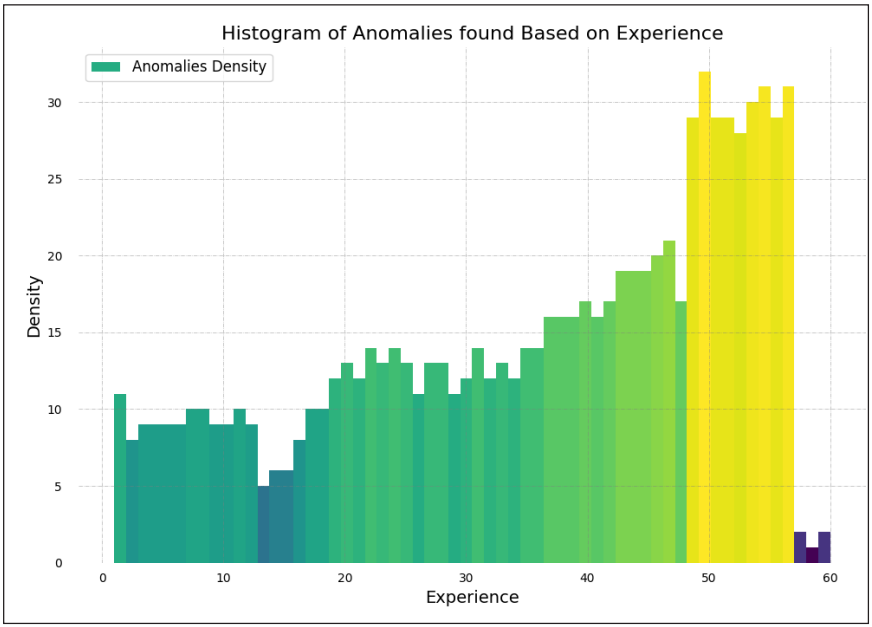
نقاط نرمال (آبی): بیمه‌شدگانی که دستمزدشان درون این کریدور قرار می‌گیرد و از الگوی کلی پیروی می‌کنند.

نقاط ناهنجار (قرمز): بیمه‌شدگانی که دستمزدشان به‌طور قابل‌توجهی بالاتر یا پایین‌تر از این کریدور است و به‌عنوان موارد مشکوک و نیازمند بررسی بیشتر شناسایی شده‌اند. این خروجی به‌وضوح موفقیت استراتژی پروژه را نشان می‌دهد؛ با جداسازی هر صنف، مدل توانست الگوی خاص دستمزدی آن صنف را بیاموزد و به شکل مؤثری محدوده‌های نرمال و ناهنجار را برای شناسایی دقیق تخلفات احتمالی دستمزد تعریف کند.

۴-۶. تحلیل توزیع ناهنجاری‌ها بر اساس سابقه

پس از شناسایی ناهنجاری‌ها توسط مدل جنگل ایزوله، برای اعتبارسنجی و درک عمیق‌تر نتایج، توزیع فراوانی این ناهنجاری‌ها بر اساس سابقه کار تحلیل شد. هدف این بود که ببینیم آیا الگوی کشف شده با انتظارات منطقی و تجربی همخوانی دارد یا خیر.

1- Min/Max Wage.



شکل ۱۲. هیستوگرام ناهنجاری و سن بیمه‌پردازی

نمودار هیستوگرام بالا تراکم تعداد ناهنجاری‌ها را در بازه‌های مختلف سابقه کار به تصویر می‌کشد. همان‌طورکه به‌وضوح دیده می‌شود، یک‌روند صعودی در تعداد تخلفات با افزایش سابقه کار وجود دارد، اما نکته کلیدی، جهش و تراکم بسیار بالای ناهنجاری‌ها در سال‌های پایانی خدمت (حدود ۴۸ تا ۶۰) است که با رنگ زرد مشخص شده است. این الگو یک یافته بسیار مهم است که صحت عملکرد الگوریتم را تأیید می‌کند. در عمل و بر اساس منطق کسب‌وکار، انتظار می‌رود که بیشترین تخلفات یا ثبت دستمزدهای نامتعارف، در سال‌های پایانی خدمت بیمه‌شدگان و نزدیک به شرایط بازنشستگی رخ داده باشد.

۵. نتیجه‌گیری و پیشنهادات

پژوهش حاضر با هدف شناسایی ناهنجاری‌های دستمزدی، به ویژه افزایش‌های غیرمتعارف در سال‌های منتهی به بازنشستگی، بر روی داده‌های واقعی ۴۷۰ بیمه‌شده بازنشسته از استان اصفهان (شامل ۲۶۲۵۸ رکورد ماهانه) انجام شد. در این راستا، سه الگوریتم یادگیری ماشین بدون نظارت – جنگل ایزوله، DBSCAN و ماشین بردار پشتیبان تک‌کلاسه – پس از انجام مراحل پیش‌پردازش، نرمال‌سازی و تفکیک داده‌ها بر اساس کد صنف (به‌دلیل ناهمگونی الگوهای دستمزدی در تحلیل یکپارچه) مورد

ارزیابی مقایسه‌ای قرار گرفتند. یافته‌ها نشان داد که الگوریتم خوشه‌بندی مبتنی‌بر چگالی (DBSCAN) با شناسایی تنها ۶ ناهنجاری در صنف ۲۳۳، عملکردی بسیار محافظه‌کارانه داشته و عملاً قادر به کشف موارد مشکوک نیست؛ در مقابل، الگوریتم ماشین بردار پشتیبان (OneClass SVM) با شناسایی ۲۶۳۵ ناهنجاری، دچار بیش‌برازش (overfitting) شده و هشدارهای غلط بالایی تولید می‌کند. در این میان، الگوریتم جنگل ایزوله با شناسایی ۸۷۳ ناهنجاری (معادل ۱۶٫۵۳درصد از داده‌های این صنف)، متعادل‌ترین و منطقی‌ترین عملکرد را ارائه داد و موفق به ترسیم «کریدور دستمزد نرمال» بر اساس سابقه‌کار شد. تحلیل توزیع ناهنجاری‌ها نیز تأیید کرد که فراوانی موارد مشکوک در سال‌های پایانی سابقه بیمه‌پردازی به‌طور معناداری بیشتر از سایر دوره‌هاست که با منطق هدف پژوهش (تشخیص افزایش‌های فرصت‌طلبانه در آستانه بازنشستگی) همخوانی کامل دارد. بر این اساس، الگوریتم جنگل ایزوله به عنوان رویکرد بهینه برای شناسایی هوشمند دستمزدهای نامتعارف در سازمان تأمین اجتماعی انتخاب شد.

با وجود نتایج امیدوارکننده، پیاده‌سازی عملیاتی این چهارچوب با محدودیت‌هایی همراه است. نخست آنکه اثربخشی مدل به شدت به در دسترس بودن داده‌های تاریخی کافی و باکیفیت وابسته است؛ داده‌هایی که باید تنوع الگوهای ناهنجاری و تقلب را پوشش دهند تا امکان یادگیری جامع ویژگی‌های تفکیک‌کننده فراهم آید. دوم، با آنکه روش پیشنهادی توانایی پردازش تراکنش‌های حجیم را در چهارچوب محدودیت‌های فنی سازمان نشان داده است، اما برای کاربرد در زمان واقعی (realtime) نیاز به بهینه‌سازی بیشتر از نظر پیچیدگی محاسباتی و تأخیر پردازش وجود دارد. سوم، قابلیت اطمینان و انطباق با مقررات جاری (از جمله قوانین حریم خصوصی و محرمانگی اطلاعات بیمه‌ای) از ملزومات جدی هرگونه استقرار عملیاتی است که در پژوهش حاضر صرفاً در سطح داده‌های گذشته‌نگر و با اخذ مجوزهای لازم بررسی شده است.

در راستای توسعه آتی و کاربردی‌سازی نتایج، پیشنهادهای زیر ارائه می‌شود:

۱. ایجاد زیرساخت متمرکز تبادل اطلاعات: سازمان تأمین اجتماعی با عقد تفاهم‌نامه‌های بین‌دستگاهی و تجمیع منابع اطلاعاتی پراکنده (مانند پرداخت‌های بانکی کارفرمایان، اظهارنامه‌های مالیاتی کارگاه‌ها، و سوابق حقوق و دستمزد در سامانه‌های دیگر) می‌تواند یک مرکز تبادل اطلاعات گسترده، متمرکز و در عین حال قانونمند ایجاد کند. چنین زیرساختی امکان پیوند و اشتراک‌گذاری داده‌های مرتبط را در چهارچوب محدودیت‌های قانونی فراهم می‌آورد و دقت تشخیص تقلب را به طور چشمگیری افزایش می‌دهد.

۲. توسعه سیستم هوشمند تطبیق‌پذیر (Adaptive AI): ترکیب موفقیت‌آمیز یادگیری بدون‌نظارت با انتخاب ویژگی‌های مبتنی‌بر دانش بیمه‌ای (مانند نسبت دستمزد به حداقل و حداکثر قانونی) پایه‌ای

مستحکم برای طراحی سیستمی فراهم می‌کند که بتواند به طور پیوسته با الگوهای نوظهور تقلب سازگار شود. توصیه می‌شود معماری مدل به گونه‌ای طراحی شود که قابلیت به‌روزرسانی برخط (online learning) و بازآموزی دوره‌ای را داشته باشد.

۳. آگاهی‌سازی و فرهنگ‌سازی عمومی: سازمان باید با همکاری رسانه‌های جمعی، تشکل‌های کارگری و کارفرمایی، برنامه‌های آگاهی‌بخشی گسترده‌ای را اجرا کند تا پیام «عدم سودآوری تقلب» و «هزینه‌های سنگین کشف جرایم» به مجرمان بالقوه منتقل شود. هم‌زمان، توانمندسازی بیمه‌شدگان قانون‌مدار برای گزارش‌دهی موارد مشکوک و ایجاد فضای عمومی «عدم تحمل نسبت به تقلب بیمه‌ای» می‌تواند به‌عنوان یک سازوکار بازدارنده اجتماعی عمل کند.

۴. تقویت ابزارهای تحلیلی درون‌سازمانی: در کوتاه‌مدت، سازمان می‌تواند با استقرار داشبوردهای نظارت بر کریدور دستمزد نرمال (بر اساس مدل جنگل ایزوله) به تفکیک هر صنف و شعبه، بازرسان خود را به یک ابزار هشدار سریع مجهز کند. این اقدام نیازمند آموزش نیروی انسانی در زمینه تفسیر خروجی‌های مدل و ادغام آن با فرایندهای بازرسی میدانی است.

به‌طور کلی، این پژوهش نشان می‌دهد که با وجود محدودیت‌های ذاتی داده‌های واقعی (نقص، دورافتادگی، عدم برچسب)، رویکرد یادگیری بدون نظارت مبتنی‌بر جنگل ایزوله می‌تواند به عنوان یک ابزار تحلیلی کارآمد در خدمت عدالت بیمه‌ای و جلوگیری از افزایش‌های فرصت‌طلبانه دستمزد در آستانه بازنشستگی قرار گیرد.

منابع

۱. نعیمی، عمران؛ جوان جعفری، محمدرضا؛ فدایی جویباری، حمید؛ قاسمی، محسن؛ رضوانی مفرد، احمد؛ رشوند بوکانی، مهدی؛ جعفری، زهرا؛ طباطبایی حصاری، نسرین. (۱۳۸۹).قانون تأمین اجتماعی در نظم حقوقی کنونی، انتشارات جنگل.

2. Abe, N., Zadrozny, B., & Langford, J. (2006). «Outlier detection by active learning». Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 504–509. https://doi.org/10.1145/1150402.1150459

3. Alwan, R. H., Hamad, M. M., & Dawood, O. A. (2022). «A comprehensive survey of fraud detection methods in credit card based on data mining techniques». AIP Conference Proceedings, 2400(1), 020006. https://pubs.aip.org/aip/acp/article-abstract/2400/1/020006/2821439

4. Artís, M., Ayuso, M., & Guillén, M. (2002). «Detection of automobile insurance fraud with discrete choice models and misclassified claims». Journal of Risk and Insurance, 69(3), 325–340.

5. Bishop, C. M. (2006). «Pattern recognition and machine learning». Springer.

6. Carletti, M., Terzi, M., & Susto, G. A. (2021). «Interpretable Anomaly Detection with DIFFI: Depth-based Isolation Forest Feature Importance» (No. arXiv:2007.11117). arXiv. https://doi.org/10.48550/arXiv.2007.11117

7. Derrig, R. (2002). «Insurance fraud. Journal of Risk and Insurance», 69(3), 271–287.

8. Duval, F., Boucher, J. -P., & Pigeon, M. (2023). «Enhancing claim classification with feature extraction from anomaly-detection-derived routine and peculiarity profiles». Journal of Risk and Insurance.

9. Ekin, T., Ieva, F., Ruggeri, F., & Soyer, R. (2018). «Statistical medical fraud assessment: Exposition to an emerging field. International Statistical Review», 86(3), 379–402.

10. Ekin, T., Lakomski, G., & Musal, R. M. (2019). «An unsupervised bayesian hierarchical method for medical fraud assessment. Statistical Analysis and Data Mining»: The ASA Data Science Journal, 12, 116–124.

11. Gill, K., Woolley, A., & Gill, M. (2005). «Insurance fraud: The business as a victim? In Crime At Work: Studies in Security and Crime Prevention»: Vol. I (pp. 73–82). Springer.

12. Gomes, C., Jin, Z., & Yang, H. (2021).« Insurance fraud detection with unsupervised deep learning. Journal of Risk and Insurance», 88(3), 591–624.

13. Goodfellow, I., Bengio, Y., & Courville, A. (2016). «Deep learning». MIT press.

14. Grabosky, P. N., & Duffield, G. M. (2001).« Red flags of fraud. In Trends and issues in crime and criminal justice »(Vol. 200). Australian Institute of Criminology.

15. Hady, M. F. A., & Schwenker, F. (2013). «Semi-supervised Learning». In M. Bianchini, M. Maggini, & L. C. Jain (Eds.), Handbook on Neural Information Processing (Vol. 49, pp. 215–239). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-«-36657-4_7

16. Hastie, T., Tibshirani, R., & Friedman, J. (2009). «Unsupervised learning. The elements of statistical learning: Data mining», inference, and prediction (pp. 485–585).

17. Johnson, J. M., & Khoshgoftaar, T. M. (2019).« Medicare fraud detection using neural networks». Journal of Big Data, 6(1), 1–35.

18. Kemp, G. (2010). «Fighting public sector fraud in the 21st century». Computer Fraud & Security, 2010(11), 16–18.

19. Kuhn, M., & Johnson, K. (2019). «Feature engineering and selection: A practical approach for predictive models». Chapman and Hall/CRC. https://www.taylorfrancis.com/books/mono/10.1201/9781315108230/feature-engineering-selection-max-kuhn-kjell-johnson

20. Kumar, M., Ghani, R., & Mei, Z. S. (2010). «Data mining to predict and prevent errors in health insurance claims processing». Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 65–74.

21. Lu, Haowen. (2025). «Evaluating the Performance of SVM, Isolation Forest, and DB-SCAN for Anomaly Detection». ITM Web Conf., 70, 04012. https://doi.org/10.1051/itm-conf/20257004012

22. 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.

23. Niu, X., Wang, L., & Yang, X. (2019). «A comparison study of credit card fraud detection: Supervised versus unsupervised».

24. Picard, P. (1996).« Auditing claims in the insurance market with fraud: The credibility issue». Journal of Public Economics, 63(1), 27–56.

25. Russell, S., & Norvig, P. (2020). «Artificial intelligence: A modern approach».-New. Pearson.

26. Smiti, A. (2020). «A critical overview of outlier detection methods». Computer Science Review, 38, 100306.

27. Stripling, E., Baesens, B., Chizi, B., & Broucke, S. (2018).« Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud». Decision Support Systems, 111, 13–26.

28. Tennyson, S. (2002). «Insurance experience and consumers´ attitudes toward insurance fraud». Journal of Insurance Regulation, 21(2), 35–55.

29. Thornton, D., van Capelleveen, G., Poel, M., van Hillegersberg, J., & Mueller, R. M. (2014). «Outlier-based health insurance fraud detection for us medicaid data». Special Session on Information Systems Security, 2, 684–694. https://www.scitepress.org/PublishedPapers/2014/49861/

30. Viaene, S., Ayuso, M., Guillen, M., Gheel, D., & Dedene, G. (2007). «Strategies for detecting fraudulent claims in the automobile insurance industry». European Journal of Operational Research, 176(1), 565–583.

31. Viaene, S., & Dedene, G. (2004). «Insurance fraud: Issues and challenges». In Geneva Papers on Risk and Insurance. Issues and Practice (pp. 313–333).

32. Wang, N., Liu, Y., Liu, Z., & Huang, X. (2020). «Application of artificial intelligence and big data in modern financial management». 2020 International Conference on Artificial Intelligence and Education (ICAIE, 85-87,.

33. Wang, X., Wu, H., & Yi, Z. (2018). «Research on bank anti-fraud model based on K-means and hidden Markov model». 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC, 780–784.

34. Zafari, B., & Ekin, T. (2019). «Topic modelling for medical prescription fraud and abuse detection». Journal of the Royal Statistical Society Series C: Applied Statistics, 68(3), 751–769.