فصلنامه تأمین اجتماعی
• سال چهاردهم • شماره اول • زمستان ۱۳۹۷

کشف تقلب و راهکارهای مقابله با آن در سازمان‌های بیمه‌ای با استفاده از داده‌کاوی (مطالعه موردی: سازمان تأمین اجتماعی)[۱]

علی حسینی[۲]، عباسعلی رضائی[۳]

تاریخ دریافت مقاله: ۱۳۹۷/۰۷/۲۰
تاریخ پذیریش مقاله: ۹۷/۱۰/۱۰

چکیده

هدف: در سال‌های اخیر، توسعه فناوری‌های جدید راه‌های زیادی برای متقلبان و مجرمان باز کرده است که بتوانند مرتکب تقلب شوند. ایجاد یک سیستم اطلاعاتی جدید، علاوه بر تمامی مزایا و منافعی که دارد، ممکن است فرصت‌های بیشتری را برای ارتکاب تقلب در اختیار مجرمان قرار دهد و این تحقیق با هدف پیش‌بینی رفتارهای متقلبانه بیمه‌شدگان تأمین اجتماعی و به‌منظور کشف تقلب در سه حوزه هزینه‌های (دفترچه‌های درمانی، پرداخت کمک هزینه مرخصی زایمان- بیمه بیکاری و دریافت مستمری بازنشستگی و فوت) سامان یافته است.

روش: تحقیق حاضر با استفاده از سه الگوریتم داده‌کاوی (درخت تصمیم، شبکه‌های عصبی مصنوعی و نزدیک‌ترین همسایه KNN) انجام شده است.

یافته‌ها: شناسایی نقاط بحرانی بروز تقلب در نخستین گام انجام شد. ضمن شناسایی بخش‌های تقلب‌خیز سازمان تأمین اجتماعی از طریق تکمیل پرسشنامه، مصاحبه و بررسی پرونده‌های قضایی و اولویت‌بندی‌ها مشخص و داده‌های متناظر از بانک اطلاعات مرتبط استخراج گردید. داده‌های اولیه پالایش و متغییرهای کلیدی شناسایی شد. تمامی حالت‌های ممکن از دو وضعیت موجود «تقلب» و سالم، برای تهیه جداول داده‌های آزمون و تست در قالب فایل‌های اکسل طراحی و جهت شناخت الگوی رفتاری بیمه‌شدگان سازمان تأمین اجتماعی در راستای سوءاستفاده از منابع در اختیار الگوریتم‌های داده‌کاوی قرار گرفت. نتایج پیاده‌سازی الگوریتم‌های شبکه‌های عصبی مصنوعی، درخت تصمیم و نزدیک‌ترین همسایه طراحی، در قالب ۳ آزمایش مجزا توسط نرم‌افزار مطلب شبیه‌سازی و به‌ترین الگوریتم جهت پیاده‌سازی کدهای SQL نهایی شناسایی گردید.

نتیجه: شناسایی نقاط مستعد تقلب به‌دلیل آگاهی کارفرمایان و بیمه‌شدگان از قوانین و مقررات بیم‌های با حداقل سابقه ممکن وگریز از چارچوب‌های بازرسی کارگاهی؛ به‌جهت استفاده حداکثری از خدمات ارائه شده سازمان تأمین اجتماعی تقریباً در سه حوزه مورد بحث انجام می‌پذیرد و در سایر موارد به‌دلیل انجام پروسه‌های کنترل سیستمی و انسانی در سال‌های اخیر تا حدود زیادی قابل پیشگیری است؛ هرچند باید این نکته را در نظر داشت که نبود رویهٔ ثابت اداری باعث گردیده است که در اکثر شعبه‌ها به‌صورت سلیق‌های رفتار شود

واژگان کلیدی: تقلب، بیمه، داده‌کاوی، بیمه اجتماعی، سازمان تأمین اجتماعی.

۱- این مقاله از پایان نامه تحت حمایت موسسه عالی پژوهش تامین اجتماعی استخراج شده است.

۲- دانشجوی ارشد نرم‌افزار کامپیوتر دانشگاه پیام نور واحد بین الملل قشم، a.hosseini242@gmail.com
۳- استادیار، گروه مهندسي كامپيوتر و فناوري اطلاعات، دانشگاه پیام نور

مقدمه

صنعت بیمه در ذات خود با ریسک‌های تجاری همراه است، ازآنجاکه در این صنعت، عملیات بیمه‌گری بر اساس احتمالات و آمار و ارقام است، مدیریت ریسک نقش مهمی دارد. شرکت‌های بیمه‌گر علاوه بر راهبریِ فرآیند قیمت‌گذاری، مدیریت دارایی‌ها، بدهی‌ها و بازاریابی خود؛ به دنبال آن هستند که به جای تدابیر ساده‌انگارانه، استراتژی‌های خود را بر مبنای ریسک جامع طراحی کرده و به‌کاربندند. ریسک جامع یک شرکت بیمه، تک‌تک فعالیت‌های آن شرکت را شامل می‌شود. هر یک از این فعالیت‌ها بر جریان نقدی مثبت و منفی یک شرکت بیمه، اثرگذار است و در واقع نیل به سودآوری بهینه در کل شرکت، جز با مدیریت جامع ریسک امکان‌پذیر نخواهد بود. (رشیدی، ۱۳۸۶)

یکی از کاربردهای مهم داده‌کاوی، کشف تقلب در دامنهٔ گسترده‌ای از داده‌هاست. داده‌کاوی با کشف روابط پنهان میان داده‌ها و با یادگیری مستمر از تقلب‌هایی که در سیستم رخ داده است، پتانسیل‌های متقلبانه را نمایان می‌کند که - در ساختارهای کنترلی مدیریت ریسک تقلب- شامل مراحل پیشگیری و کشف است. کشف تقلب در بازارهای مختلف اهمیت ویژه‌ای دارد. از کل حجم هزارمیلیارد دلاریِ صنعت بیمه در جهان، حدود ۲۵ درصد از فعالیت‌ها (در بعضی حوزه‌ها، مثل بیمه‌درمانی تا ۴۰ درصد) تقلب بوده است. (Mohit & Kumar 2011) این حجم بالای هزینه‌های سربار، سازمان‌ها را به استفاده از فناوری‌های جدید در عرصه کشف تقلب راغب‌تر کرده است. از طرفی سازمان‌هایی که از پایگاه‌های دادهٔ توزیع‌شده استفاده می‌کنند، به علت خاصیت توزیع‌شده‌گی اطلاعات، به افزونگیِ زیاد -جهت سرعت‌بخشیدن به تراکنش‌ها- دچار می‌شوند که این موضوع احتمال تقلب را بالاتر می‌برد. این تحقیق می‌کوشد، پارامترهای زمینه‌ساز بروز تقلب سازمانی را شناسایی کند. تقلب در سازمان تأمین اجتماعی گاهی به‌دلیل ضعف قوانین موجود یا تخلف کارکنان آن و یا سوء استفاده بیمه شدگان در راستای دستیابی به اهدافی شوم صورت می‌گیرد.

مفهوم کلاهبرداری و تقلب در صنعت بیمه

کلاهبرداری و تقلب، ریسکی جدي را به تمام بخش‌های مالی تحمیل می‌کند. در بخش بیمه، بیمه‌گران و بیمه‌گذاران، هردو، هزینه‌هایی را متحمل می‌شوند. زیان‌هایی که از طریق کلاهبرداری‌ها ایجاد می‌شوند، بر منافع بیمه‌گران و به طور بالقوه، بر ثبات مالی آنها اثر می‌گذارد. برای جبران این زیان‌ها، بیمه‌گران حق بیمه‌ها را افزایش می‌دهند و این موضوع به بالاتر رفتن هزینه‌های بیمه‌گذاران می‌انجامد. در واقع کلاهبرداری اطمینان سهام‌دار و مصرف‌کننده را کاهش می‌دهد. همچنین حسن شهرت تک‌تک بیمه‌گران، بخش بیمه و به‌طور بالقوه، ثبات اقتصادی را تحت تاثیر قرار می‌دهد. ارقام تقلب‌های بیمه‌ای تکان‌دهنده هستند. طبق مطالعات انجمن بیمه انگلستان، هزینه‌های جعلی در صنعت بیمه در بریتانیا بیش از یک میلیارد پوند در سال است. همچنین حجم بازار جرایم بیمهٔ صنعت

مخابرات دنیا بالغ بر ۲۱۰ میلیارد دلار برآورد می‌شود و همان‌طورکه پیش‌تر گفتیم: حدود ۲۵ درصد از کل حجم هزارمیلیارد دلاریِ صنعت بیمه در جهان تقلب بوده است. هزینه‌های بیمهٔ سلامت در سال‌های اخیر در دنیا افزایش نگران‌کننده‌ای داشته است. پژوهش‌ها نشان می‌دهند که بیشتر این هزینه‌ها به خطاهای تشخیص ادعا و تقلب در مراحل پرداخت بیمه برمی‌گردند.

بیمه یک رابطه مبتنی بر قرارداد است که بر اساس آن، بیمه‌گر با بیمه‌گذار توافق می‌کند، در قبال پرداخت حق بیمه از طرف بیمه‌گذار، به نمایندگی از او خسارت یک یا چند حادثه تعریف‌شده را، که ممکن است در آینده رخ دهد -پس از این‌که شخص زیان‌دیده ادعای خسارت نمود- بپردازد. (رشیدی ۱۳۸۶)

ادغام روزافزون بازارهای مالی و عدد روبه رشدِ بیمه‌گران بین‌المللی فعال، کلاهبرداری و پیامدهای بالقوه آن را به موضوع

مهمی در سطح جهانی بدل کرده است. امروزه معظل کلاهبرداری‌ها برای انجمن بین المللی ناظران بیمه اولویت بالایی دارد. این انجمن در خصوص زیان‌های ناشی از کلاهبرداری‌های بیمه ای، درصدد شناسایی نقاط آسیب‌پذیر سازمان‌های بیمه‌گر است. (خاکسار-خیابانی، ۱۳۹۱)

کارشناسان، شیوه‌های متنوع و متداول تقلب در صنعت بیمه را این‌طور دسته‌بندی می‌کنند:

۱. تقلب در بیمه‌های کارگری: کارگرانی که با صحنه‌سازی در حادثه‌ای صدمه می‌بینند تا به‌ازای روزهایی که در محل کار حاضر نمی‌شوند، از شرکت بیمه خسارت دریافت کنند: (اغراق در جراحات وارده در حوادث ناشی از کار؛ اعلام حوادث خارج از محل کار به عنوان حادثه محل کار.)

۲. تقلب در بیمه‌های کارفرمایان: گزارش حقوق کارگران به میزان کمتر از حد معمول برای پرداخت بیمه کمتر؛ اعلام کارگران به عنوان پیمانکاران مستقل؛ اعلام عنوان شغلی اشتباه برای پرداخت کمتر حق بیمه؛ عدم پرداخت بیمه؛ تبانی کارفرمایان و واگذارنده‌های پیمانی.

۳. تقلب در بیمه‌های درمانی: فرستادن صورت حسابهایی بیش از هزینه‌های واقعی خدمات درمانی انجام‌شده به بیمه‌گذار، توسط پزشک؛ ارایه خدمات درمانی و پاراکلینیکی، مانند جراحی و آزمایش؛ تصویربرداری‌های پزشکی غیرضروری با هدف دریافت مبالغ گزاف از شرکت‌های بیمه.

۴. تقلب در اجرای روند اخذ مفاصا حساب اتمام پروژه از سازمان تأمین اجتماعی (تعیین ضریب، گزارش صورت وضعیت‌های صوری، محاسبه مبالغ پرداختی، و ... .) به‌دلیل تک‌نرخی نبودن مبنای محاسبات حق بیمه پیمانی.

کمیته اروپایی بیمه‌ها، سال ۱۹۹۶ در گزارشی با عنوان «راهنمای ضدکلاهبرداری بیمه‌ای اروپا» اعلام کرد که از نظر میانگین هزینه، بیش‌ترین کلاهبرداری و تقلب در دو بخشِ بیمه وسایل نقلیه موتوری

و غرامت‌های بیمه کارگری صورت گرفته است.

کاربرد داده کاوی در فرآیند کشف تقلب

نگای و همکاران(۲۰۱۰) کاربرد روش‌های داده‌کاوی در کشف تقلب‌های مالی را بررسی کردند. آنها ۴۹ مقاله چاپ‌شده در مجلات معتبر را مورد مطالعه قرار دادند. در مجموعه تحقیق‌های مورد بررسی، از شش طبقه از وظایف/کاربردهای داده‌کاوی برای کشف تقلب‌های مالی استفاده شده است. این موارد عبارت هستند از: طبقه‌بندی، رگرسیون، خوشه‌بندی، پیش‌بینی، کشف داده‌های پرت و تصویرسازی. هر یک از این شش طبقه، با مجموع‌های از رویکردهای الگوریتمی حمایت می‌شود که به‌دنبال استخراج ارتباط‌هایی مرتبط، از داده‌ها هستند. این رویکردها در نوع مسائلی که قادر به حل آنها هستند، با یکدیگر تفاوت دارند.

طبقه‌بندی، مدلی را می‌سازد و از آن برای پیش‌بینی عنوان طبقات اشیای ناشناخته استفاده می‌کند تا بین اشیای متعلق به طبقات مختلف، تمایز ایجاد کند. این عنوان‌های طبقاتی از قبل تعریف شده‌اند، ولی متمایز و مرتب نشده‌اند. طبقه‌بندی و پیش‌بینی عبارت است از فرآیند شناسایی مجموع‌های از ویژگی‌ها و مدل‌های مشترک که طبقات یا مفاهیم داده‌ها را توصیف و از هم متمایز می‌کنند. روش‌های معمول طبقه‌بندی شاملِ : شبکه‌های عصبی، شبکه‌های بیز ساده، درختان تصمیم و ماشین‌های بُردار پشتیبان، می‌شود. این‌دست وظایف طبقه‌بندی در کشف تقلب‌های کارت اعتباری، بیمه سلامت، بیمه خودرو، تقلب‌های شرکتی و دیگر انواع تقلب استفاده می‌شوند. طبقه‌بندی یکی از رایج‌ترین مدل‌های یادگیری در کاربرد داده‌کاوی برای کشف تقلب‌های مالی است .طبقه‌بندی، فرآیندی متشکل از دو مرحله می‌باشد: در گام اول، با استفاده از نمون‌های آموزشی، یک مدل آموزش داده می‌شود. یکی از صفات، یعنی صفت عنوان طبقه، حاوی مقادیری است که نشان‌دهنده طبقه از پیش تعریف‌شده‌ای است که هر ردیف به آن تعلق دارد. ذکر این نکته لازم است که این گام به‌عنوان یادگیری نظارت‌شده نیز معروف است. در گام دوم، در مدل تلاش می‌شود، اشیایی که به نمونه آموزشی تعلق ندارند، طبقه‌بندی شوند و یک نمونه آزمون (تایید) تشکیل دهند.

پرکاربردترین روش‌های کشف تقلب‌های مالی عبارتند از: یک- مدل‌های رگرسیون لجستیک (رایج‌ترین روش)؛ دو- شبکه‌های عصبی؛ سه- نزدیک‌ترین همسایه؛ چهار- شبکه استنباط بیزین و درختان تصمیم. هر ۴ روش، همگی در گروه «طبقه‌بندی» قرار می‌گیرند. این روش‌ها برای مشکلات ذاتیِ کشف و طبقه‌بندی داده‌های متقلبانه راه‌حل‌های مهمی پیشنهاد می‌دهند. روش‌هایی که مبنای اصلی تحقیق صورت گرفته در این مقاله هستند.

فراگیری ماشین و داده‌کاوی

عصر حاضر، عصر اطلاعات است . اطلاعات فراوانی در پایگاه‌های‌داده ذخیره می‌شود که تبدیل آنها به دانش مورد نیاز برای تصمیم‌گیری، به ابزارهای جدیدی احتیاج دارد. روش‌های آماری در تحلیل داده‌ها بیشتر بر پایه استخراج شاخص های کمّی استوار است. اگرچه این روش‌ها غیر مستقیم ما را به دانش مورد نیاز در تصمیم‌گیری سوق می‌دهند؛ در نهایت تفسیر نتایج آنها نیازمند تحلیل‌های انسانی است. روش‌های نوین تحلیل داده، تفسیر داده‌ها را آسان‌تر کرده و می‌توانند درک بهتر فرایندها را فراهم سازند . برای تسهیل فرآیند تصمیم‌گیری، سیستم‌های تحلیل داده باید به دانش لازم و قابل تصمیم‌گیری بر اساس داده‌ها تجهیز شوند . جهت دستیابی به این هدف، محققان ایده‌های جدیدی از فراگیری ماشین ارایه داده‌اند. با توجه به این ایده‌ها وظیفه فراگیری ماشین، تبدیل داده‌ها (ورودی) به دانش تصمیم‌گیری (خروجی) است. همچنین بر مبنای این ایده‌ها، حوزه تحقیقاتی جدیدی به نام «داده‌کاوی» به وجود آمده است (Michalski et al ,1998) داده‌کاوی، فرآیند کشف الگوها در داده‌هاست. این فرآیند باید خودکار یا نیمه‌خودکار باشد . الگوهای شناسایی شده باید معتبر بوده و برای ما مزایایی، ازجمله مزایای اقتصادی داشته باشند. همچنین داده‌ها باید همواره در قالب کمیت های معتبر ارائه شوند. (Witten & Frank,2000)

استفاده از مدل‌های ریاضی برای شناسایی تقلب، به متخصصان سازمان‌های بیمه‌گر این امکان را می‌دهد که با صرف زمان و هزینه کمتری تشخیص دهند که ادعای خدمات درخواستی از لحاظ آماری، مشکوک به تقلب هست یا خیر .

در این تحقیق با استفاده از سه روشِ شبکه‌های عصبی، درخت تصمیم و نزدیک‌ترین همسایه که از ابزارهای رایج در داده‌کاوی هستند، مدل‌هایی برای شناسایی و دسته‌بندی خسارات‌های تقلبی بر داده‌های واقعی برازش داده می‌شود.

شبکه‌های عصبی

شبکه‌های عصبی مصنوعی که معمولاً « شبکه‌های عصبی» نامیده می‌شوند، یک الگوی ریاضی مبنی بر سیستم زیستی است. سیستم‌های عصبی، الگوریتمی برای بهینه‌سازی و یادگیری آزادانه بر اساس مفاهیم الهام گرفته از تحقیق در ماهیت مغز هستند. مغز با استفاده از قابلیتی شناخته‌شده به عنوان نورون اجزاءساختاری خود را سازماندهی می‌کند، در نتیجه محاسبات معینی را بسیار سریع‌تر از کامپیوتر دیجیتال انجام می‌دهد. در حالت کلی، شبکهٔ عصبی ماشینی است که طراحی شده تا روشی مشابه با کاری را مدل‌سازی کند که مغز برای انجام وظایف خاص یا عملکرد قابل توجه انجام می دهد. این شبکه یک پردازنده توزیع‌شده موازی و بزرگ، متشکل از واحدهای پردازش ساده است که تمایلی طبیعی به ذخیره‌سازی دانش تجربی و ایجاد دسترسی به آن دارد.

آیا شبیه شدن کامپیوترها به مغز انسان اتفاق خارق‌العاده‌ای نخواهد بود؟ این همان بزنگاهی است که شبکه‌های عصبی وارد عرصه می‌شوند.

اما این شبکهٔ عصبی، مغز نیست. باید این نکته مهم را در نظر داشت که شبکه‌های عصبی عموماً شبیه‌سازهای نرم‌افزاری هستند که با برنامه‌نویسی برای کامپیوترهای بسیار ساده و پیش‌پاافتاده راه می‌افتند و با روش‌های قبلی خود و با استفاده از ترانزیستورها و دروازه‌های منطقی خود کار می‌کنند تا مانند میلیاردها سلول مغزی متصل و موازی عمل کنند. هیچ‌کس تا به حال حتی کسی تلاش هم نکرده است تا کامپیوتری بسازد که با ترانزیستورهایی با ساختار موازی، مثل مغز انسان کار کند. به‌عبارت دیگر، تفاوت شبکهٔ عصبی با مغز، مانند تفاوت مدل کامپیوتری آب‌وهوا با ابر، برف، و هوای آفتابی واقعی است. شبیه‌سازی کامپیوتر، تنها، مجموعه‌ای از متغیرهای جبری و معادلات ریاضی است که آن‌ها را به‌هم متصل می‌کند. (اعداد ذخیره‌شده در جعبه‌هایی که مقدار آن‌ها دائما درحال تغییر است). در واقع این شبیه‌سازی‌ها برای کامپیوترها هیچ معنایی ندارد و تنها برای افرادی که برنامه آن‌ها را می‌نویسند بامعنا است.

شبکه‌های عصبی که به این صورت (شبیه‌سازی و برنامه‌نویسی) به وجود می‌آیند، شبکهٔ عصبی مصنوعی (ANN) نامیده می‌شوند تا از شبکه‌های عصبی حقیقی (مجموعه سلول‌های مغزی متصل) متمایز شوند. شاید با اصطلاحات دیگری، مانند ماشین‌های اتصال، پردازنده‌های توزیع‌شدهٔ موازی، ماشین‌های تفکر و ... نیز مواجه شده باشید، اما در این مقاله تنها، اصطلاح شبکهٔ عصبی را به‌کار می‌بریم که منظور همان شبکهٔ عصبی مصنوعی است.

شبکه عصبی به شکل گسترده‌ای در طبقه‌بندی و خوشه‌بندی به‌کار رفته و پس از رگرسیون، پر کاربردترین روش داده‌کاوی در کشف تقلب است.(Yue et al, ۲۰۰۷) در گام نخست، شبکه با مجموعه‌ای از داده‌های زوجی برای ترسیم ورودی‌ها و خروجی‌ها آموزش داده می‌شود. سپس وزن ارتباطات بین نرون‌ها تثبیت شده و از شبکه برای تعیین طبقه‌بندی مجموعه جدیدی از داده‌ها استفاده می‌گردد. (فوآ و همکاران، ۲۰۰۵)

یادگیری درخت تصمیم

ساختار درخت تصمیم در یادگیری ماشین، مدلی پیش‌بینی‌کننده است که با مشاهده دقیق یک پدیده، دربارهٔ مقدار هدف آن نتیجه‌گیری می‌کند . تکنیک یادگیری ماشین برای استنتاج یک درخت تصمیم از داده‌ها، یادگیری درخت تصمیم نامیده می‌شود که یکی از رایج‌ترین روش‌های داده‌کاوی است.

هر گره داخلی، متناظر یک متغیر و هر کمان به یک فرزند، نمایانگر یک مقدار ممکن برای آن متغیر است. یک گره برگ، با داشتن مقادیر متغیرها که با مسیری از ریشه درخت تا آن گره برگ

بازنمایی می‌شود، مقدار پیش‌بینی شدهٔ متغیر هدف را نشان می‌دهد. یک درخت تصمیم ساختاری را نشان می‌دهد که در آن برگ‌ها نشان‌دهندهٔ دسته‌بندی و شاخه‌ها ترکیبات فصلی صفاتی منتج به این دسته‌بندی‌ها را بازنمایی می‌کنند. یادگیری یک درخت می‌تواند با تفکیک‌کردن یک مجموعه منبع به زیرمجموعه‌هایی براساس یک تست مقدار صفت انجام شود. این فرآیند به شکل بازگشتی در هر زیرمجموعهٔ حاصل از تفکیک تکرار می‌شود. عمل بازگشت زمانی کامل می‌شود که تفکیک بیشتر، سودمند نباشد یا این‌که بتوان یک دسته‌بندی را برای همهٔ نمونه‌های موجود در زیرمجموعهٔ به‌دست‌آمده اعمال کرد.

درختان تصمیم قادرند از روابط موجود در یک مجموعهٔ داده‌ای، توصیف‌های قابل درکی برای انسان تولید کنند. همچنین می‌توانند برای وظایف دسته‌بندی و پیش‌بینی به‌کار روند. این تکنیک به شکلی گسترده در زمینه‌های مختلفی همچون تشخیص بیماری، دسته‌بندی گیاهان و استراتژی‌های بازاریابی مشتری به‌کار رفته است.

K نزدیک‌ترین همسایه K Nearest Neighbor (KNN)

روش K (نزدیک‌ترین همسایه): یک گروه شامل K رکورد را ـاز مجموعه رکوردهای آموزشی که نزدیک‌ترین رکوردها به رکورد آزمایشی باشندـ انتخاب می‌کند و بر اساس برتری رده یا برچسب مربوط به آن‌ها، دربارهٔ دستهٔ رکورد آزمایشی مزبور تصمیم‌گیری می‌نماید. به عبارت ساده‌تر، این روش، ردهای را انتخاب می‌کند که در همسایگی انتخاب‌شده بیش‌ترین تعداد رکورد منتسب به آن دسته باشند. بنابراین ردهای که از همه رده‌ها بیشتر در بین K (نزدیک‌ترین همسایه) مشاهده شود، به‌عنوان رده رکورد جدید در نظر گرفته می‌شود. ایده اصلی روش KNN این است که اگر موجودی مثل اردک راه برود و مثل اردک quack quack کند، پس حتما یک اردک است.

استفاده از الگوریتمKNN نیازمند تعیین سه موضوع می‌باشد:

۱) باید یک مجموعه رکورد داشته باشیم؛

۲) یک معیار محاسبهٔ شباهت نیز باید داشته باشیم؛

۳) همچنین مقدار K نیز باید مشخص شود تا بتوان بر اساس آن عمل نمود. برای موضوعات دسته‌بندی دودویی معمولا در نظر گرفتن مقادیر فرد برای K بهتر است، زیرا امکان پیروز شدن یکی از دو دسته را افزایش می‌دهد. برای رده‌بندی چند ردهای باید عددK را بزرگ‌تر از تعداد رده‌ها و نیز متفاوت از عدد تعداد رده‌ها (زوج یا فرد بودن) در نظر گرفت؛ یعنی اگر تعداد رده‌ها زوج باشد، باید K نهایی را فرد در نظر گرفت و بالعکس.

در رده‌بندی‌های KNN برای دسته‌بندی کردن یک رکورد با دسته نامشخص به صورت زیر

عمل می‌شود:

اول- فاصله رکورد جدید از همه رکوردهای آموزشی محاسبه می‌شود.

دوم- نزدیک‌ترین همسایه‌ها مشخص می‌شوند.

ج- از برچسب دسته K (نزدیک‌ترین همسایه)، برای پیش‌بینی دسته رکورد جدید استفاده می‌شود. به این صورت که بین K رکورد رأی‌گیری می‌شود و دسته‌ای که بیش‌ترین تعداد دفعات دیده‌شدن را در بین این K رکورد داراست، به عنوان دسته رکورد جدید در نظر گرفته خواهد شد.

انتخاب مقدار K در این روش دسته‌بندی بسیار مهم و کلیدی است. اگر مقدار K خیلی کوچک انتخاب شود، الگوریتم به نویز حساس می‌شود. در واقع نویزها در نزدیکی آن رکورد، ممکن است ایجاد اشتباه کنند. همچنین اگر مقدار K خیلی بزرگ انتخاب شود، ممکن است در میان نزدیک‌ترین همسایه‌ها، رکوردهایی از دسته‌های دیگر نیز قرار بگیرند.

وقتی K عدد بزرگی انتخاب شود، به خطای دسته‌بند در دسته‌بندی رکورد ورودی می‌انجامد. یکی از ایده‌هایی که برای حل این مشکل ارائه شده، تعریف فاکتور وزن است. این فاکتور وزنی برابر ۱/d۲ را در نظر می‌گیرد که مقدارd بیانگر فاصله هر رکورد تا رکورد ورودی می‌باشد. به این ترتیب، فاصله‌ها برای الگوریتم اهمیت پیدا می‌کنند و این وزن‌دهی سبب می‌شود که به رکوردهایی که نزدیک‌تر به رکورد ورودی هستند، اهمیت بیش‌تری داده شود.

ساختار اجرایی تحقیق

برای ساختن یک مدل از ماشین یادگیری و دستیابی به نتایج قابل اطمینان، به داده‌هایی از هر دو دسته مورد ادعای جعلی و غیرجعلی نیاز داریم. در این بخش داده‌ای از عملکرد روش‌های تحقیق، شامل (پرسشنامه، نظرات نخبگان و کارشناسان و کارکنان سازمان تأمین اجتماعی و همچنین بررسی پرونده‌های تخلف سال‌های ۱۳۹۰ تا ۱۳۹۶)، در قالب ۲۲۰ مورد تقلب صورت گرفته در مقابل ۲۷۸ پرونده موارد سالم در سه حوزهٔ خدماتی «مرخصی زایمان/ بیمه بیکاری»، «صدور/ تمدید دفترچه» و «مستمری بازماندگان» به‌دست آمد که طبق نتایج اولیه تحقیق، بحرانی‌ترین نقاط بروز تقلب در سازمان تأمین اجتماعی شناخته شدند.

سوابق کلی بیمه‌شدگان، نوع کارگاه محل اشتغال، سوابق آخرین کارگاه، نوع بیمه، نرخ حق بیمه و مبلغ پرداختی به سازمان تأمین اجتماعی در سال‌های آخر بیمه‌پردازی یا دریافت خدمات، همچنین وضعیت بازرسی کارگاهی، ازجمله مهم‌ترین متغیرهایی بودند که پس از کاهش متغیرهای بلااستفاده در تحقیق، در ارزیابی دقیق روش‌های پیشنهادی به‌کار برده شدند. (جدول ۱شماره)

نمونه های انتخابی به‌دست آمده از بررسی پرونده‌های تخلف در سالهای ۱۳۹۰ تا ۱۳۹۶ سازمان

تأمین اجتماعی در سه حوزه فوق، نتایج حاصل از نقطه نظرهای کارشناسان و کارکنان سازمان تأمین اجتماعی و ... ، پس از اجماع کلی در دو گروه «متقلب» و «سالم» براساس جدول ۵ ۲- طبقه‌بندی شدند و جهت انتخاب به‌ترین الگوریتم داده‌کاوی در اختیار ماشین یادگیری قرار گرفتند.

جدول شماره ۱) تعاریف متغیرهای تحقیق

اسامی متغیرها
جنسیت
نوع فعالیت آخرین کارگاه (پیمانکاری / کمک دولت / صنفی)
نرح حق بیمه آخرین کارگاه (۲۷ درصد کامل / کمک دولت)
نوع بیمه (اجباری / اختیاری / مشاغل آزاد و...)
میزان سابقه در آخرین کارگاه
حداقل تعداد حضور بیمه شده در بازرسی کارگاهی
تعداد تغییرات نرخ حق بیمه
میزان کل سابقه پرداختی

جدول شماره ۲) مجموعه داده‌های استخراج‌شده نهایی تحقیق

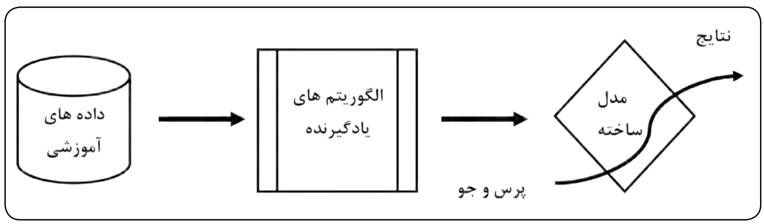
نوع خدمت گرفته شده	نمونه های سالم	نمونه های متقلب	تعداد نمونه ها
مرخصی زایمان / بیمه بیکاری	۷۰	۱۳۷	۲۰۷
صدور / تمدید دفترچه	۱۲۸	۱۴۴	۲۷۲
مستمری بازماندگان	۱۵۲	۷۶	۲۲۸
جمع	۲۷۸	۲۲۰	۵۰۷

طراحی روش پیشنهادی

طبقه‌بندی(Classification) از زیرشاخه‌های اصلی داده‌کاوی و یادگیری ماشین است. با استفاده از طبقه‌بندی کارهای زیادی انجام می‌گیرد. ازجمله: متمایزکردن هوشمندِ بیمه‌شدگان متقلب از سالم؛ شناسایی اشیا مختلف در یک تصویر؛ یافتن مشتری‌های ناراضی قبل از خروج آن‌ها از یک شرکت و رسیدگی به مسأله‌شان؛ خواندن پلاک اتومبیل‌ها با دقت بالا؛ ساختن ماشین‌های خودران(بدون راننده) و ... . به شرط آن‌که الگوهای طراحی ماشین با دقت فراوان انجام گرفته باشد، چون در واقع باید یک ناظر مشخص شود که ستون آخر داده‌های ما را پر کند: افراد متقلب :۱ و افراد سالم: ۰

یادگیری‌ماشین، زیرمجموعه‌های از هوش‌مصنوعی است. با استفاده از تکنیک‌های یادگیری‌ماشین، کامپیوتر، الگوهای موجود در داده‌ها(اطلاعات پردازش‌شده) را می‌آموزد و می‌تواند از آن استفاده کند. باید توجه داشته باشید که در این تکنیک‌ها، یادگیریْ در یک سیستمِ کامپیوتری، بدون برنامه‌نویسیِ صریح(Explicit Programming) انجام می‌شود.

شکل شماره۱) طرح فرایند یادگیری ماشین



فرآیند یادگیری‌ماشین را، می‌توان به‌صورت بالا مدل کرد؛ شکل (۱). داده‌های آموزشی به الگوریتم‌های یادگیری‌ماشین تزریق می‌شوند. این الگوریتم‌ها، وظیفهٔ یادگیری و واکشیِ الگوهای (Patterns)مختلف را در داده‌ها برعهده دارند. بعد از به‌دست آوردنِ الگوها توسط الگوریتم‌ها، یک مدل (Model) ساخته می‌شود. این مدل(Model) می‌تواند در حافظه ذخیره شود. بعد از ذخیرهٔ مدل، سیستم، به تواناییِ پیش‌بینیِ رفتار مجهز می‌گردد. این مدل، می‌تواند خروجیِ پیش‌بینی ـمتقلب یا سالم بودن فرد مورد نظرـ را برگرداند.

در فرآیندهای یادگیری‌ماشین، داده‌ها اهمیت بسیاری دارند؛ یعنی هر چقدر هم که الگوریتم‌های مختلفِ یادگیری‌ماشین، قوی و جامع طراحی شوند، اگر داده‌های خوبی به سیستم وارد نشود(مثلاً داده‌های غلط یا داده‌های ناکافی به سیستم تزریق شوند)، سیستم، پاسخی غیرِدقیق و ناصحیح ارائه می‌دهد. عملکردی که در این پژوهش سعی شده است به‌ترین نحو و با بیش‌ترین دقت در مرحله استخراج داده‌های اولیه انجام گیرد تا نتیجه نهایی و الگوی شناسایی‌شده از بالاترین ضریب اطمینان برخوردار باشد.

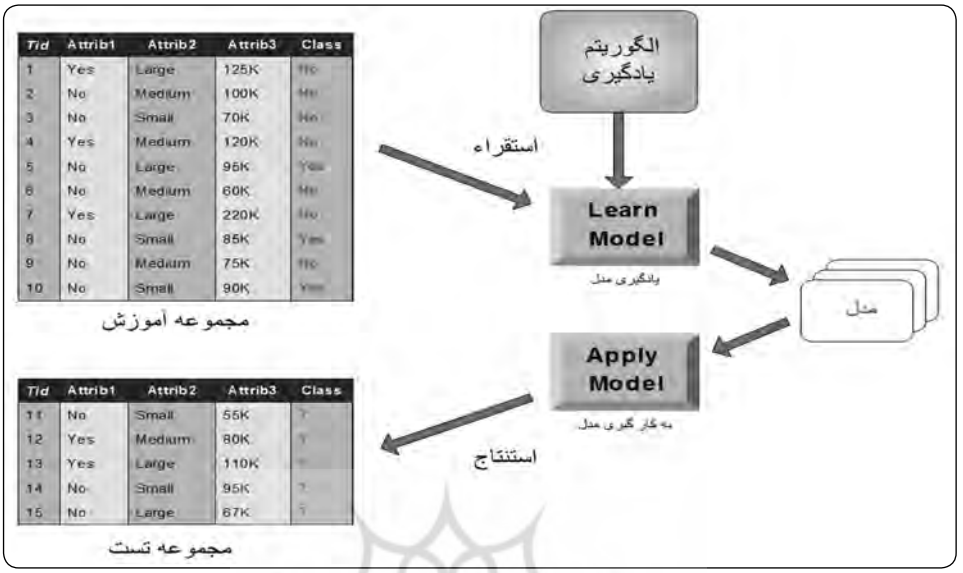
در کل می‌توان روش ارائه شده را به مراحل زیر تقسیم‌بندی کرد:

۱) جمع‌آوری، آماده‌سازی و پیش‌پردازش داده‌ها؛

۲) کاهش بُعد فضای ویژگی (تعداد صفات) با استفاده از الگوریتم اجزای اساسی یا PCA؛

۳) به‌کارگیری طبقه‌بندی درختِ تصمیم.

شکل شماره۲) دیاگرام روش پیشنهادی



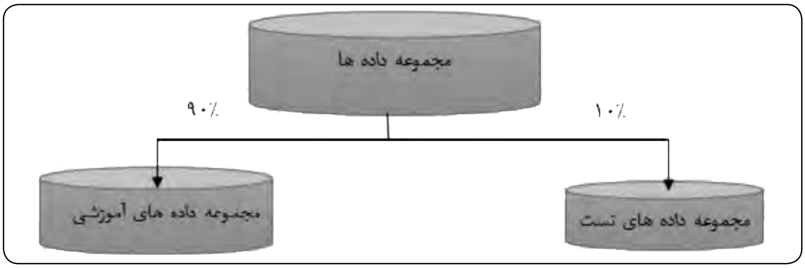
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

فرآیند یادگیری از داده‌ها مهم‌ترین مرحلهٔ داده‌کاوی است که با استفاده از تکنیک‌های مختلف، توصیف، فرمول‌بندی و اجرا می‌شود. روش یادگیری، الگوریتمی است که یک وابستگی مجهول بین ورودی و خروجی سیستم را ـ از مجموعه داده‌های در دسترس ـ برآورد می‌کند، که از آن می‌توان برای پیش‌بینی خروجی‌های آتی سیستم از مقادیر ورودی معلوم استفاده نمود. (شکل شماره۲)

در این مرحله با استفاده از تکنیک‌های داده‌کاوی، داده‌ها مورد کاوش قرار می‌گیرد، دانش نهفته در آنها استخراج می‌شود و الگوسازی صورت می‌پذیرد. سپس ابزار داده‌کاوی، نتایج و الگوهای به‌دست‌آمده را بررسی و نتایج مفید را مشخص می‌کند . در واقع، ابزار داده‌کاوی در درجه اول از توالی ارتباطات برای کشف یک الگو بهره می‌گیرد و در نهایت اطلاعات به‌دست‌آمده را دسته‌بندی می‌کند تا به الگوی خاصی برسد. الگوریتم‌های پیش‌بینی‌کننده، جهت یادگیری الگوها و قوانین موجود در نمونه‌ها، به نمونه‌های آموزشی نیاز دارند تا مدل حاکم بر آنها را تولید و سپس در پیش‌بینی کلاس نمونه‌های جدید، آن را آزمایش کنند، بنابراین مجموعه داده‌ها را به دو بخش نمونه‌های آموزشی و نمونه‌های آزمایشی تقسیم می‌کنیم. برای این منظور ۹۰ درصد از نمونه‌ها را برای مجموعه‌داده آموزشی و ۱۰ درصد باقیمانده را برای مجموعه‌داده آزمایشی (تست) در نظر می‌گیریم.

شکل شماره۳) نحوه اجرای الگوریتم دسته‌بندی classification

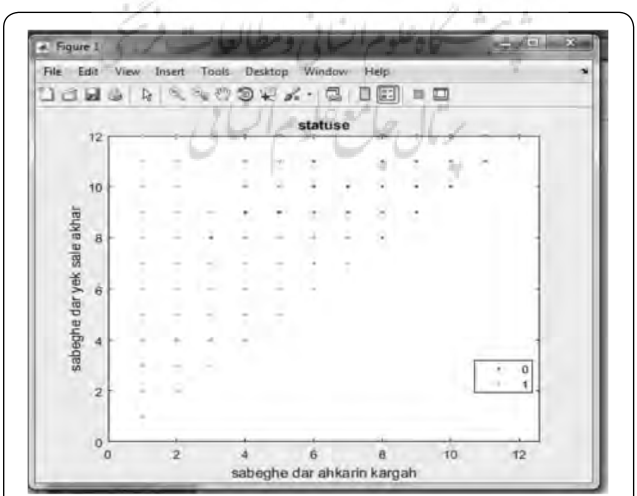


در این مقاله از یکی از متداول‌ترین الگوریتم‌های دسته‌بندی داده‌ها، یعنی درخت تصمیم استفاده شده است. ابتدا داده‌های آموزشی در اختیار درخت قرار داده می‌شود تا دانش حاکم بر داده‌ها را مدل‌سازی کند و در ادامه از این یادگیری برای پیش‌بینی کلاس داده‌های تست استفاده می‌کنیم.

این مجموعه‌داده شامل اطلاعات گسسته (مانند ویژگی جنسیت) است، پس باید این‌گونه آرایه‌های اسمی را به آرایه‌های منطقی تبدیل کنیم. آرایه منطقی در مسیر ویژگی‌ها به‌صورت صفر و یک قرار داده می‌شود. در شکل ۵ برای تجسم این تبدیل، داده‌ها را به‌شکل دو کلاس متمایز نشان دادیم تا پراکندگی آنها را مشاهده کنیم. همان‌طورکه می‌دانیم تعداد ویژگی‌ها در مجموعه‌داده، ۹ بوده است که برای رسم، آنها را به دو بعد کاهش دادیم.

داده‌ها پس از پالایش نهایی، برای بررسی در دو بعد دسته‌بندی گردید: عدد صفر بیانگر افراد سالم و عدد ۱ بیانگر افراد متقلب. (شکل ۴)

شکل شماره٤) توزیع اطلاعات و داده‌ها



استفاده از الگوریتم دسته‌بندی درخت تصمیم:

درخت‌های تصمیم‌گیری شامل بسیاری از متغیرهای ورودی هستند که ممکن است بر طبقه‌بندی مختلف الگوها تأثیر بگذارند. این متغیرهای ورودی اغلب به‌نام صفات خوانده می‌شوند. یک درخت، شامل شاخه‌ها و گره‌ها است. شاخه، نشان‌دهندهٔ نتیجه یک آزمون برای طبقه‌بندی یک الگو (بر اساس یک آزمون)، با استفاده از یکی از ویژگی‌ها است. گره یا برگ در پایان، نشان‌دهندهٔ نهایی انتخاب کلاس برای یک الگو می‌باشد.

الگوریتمی کلی برای ساخت یک درخت تصمیم‌گیری:

۱. ایجاد یک گره ریشه و اختصاص تمام داده‌های آموزشی به آن؛

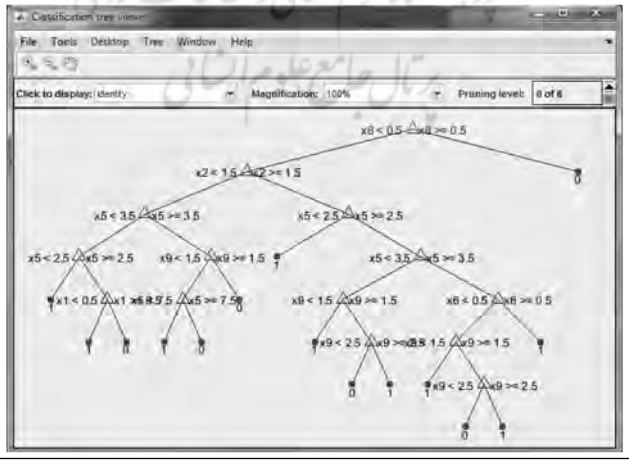
۲. انتخاب به‌ترین ویژگی تقسیم؛

۳. افزودن یک شاخه به گره ریشه، به‌ازای هر مقدار از قسمت‌ها. افراز داده‌ها به زیرمجموعه‌های مجزا در امتداد خطوط تقسیم خاص و حالت‌دادن به شاخه‌ها؛

۴. تکرار مراحل ۲ و ۳ برای هر گره برگ تا رسیدن به معیارهای توقف (برای مثال، گرهای که توسط برچسب یک کلاس ساده، محدود شده است).

بسیاری از الگوریتم‌های مختلف برای ایجاد درخت‌های تصمیم‌گیری ارائه شده‌اند. این الگوریتم‌ها در درجهٔ اول از نظر راهی که برای برآورد و تقسیم ویژگی انتخاب می‌کنند (و مقادیر تقسیم آن)، متفاوت هستند. منظور از تقسیم صفات (تقسیم ویژگی تنها همان یک‌بار و یا چند بار)، تعداد انشعابات در هر گره (باینری در مقابل سه‌تایی)، معیارهای توقف و هرس درخت (پیش‌هرس در مقابل تعویق هرس) است. (شکل۸)

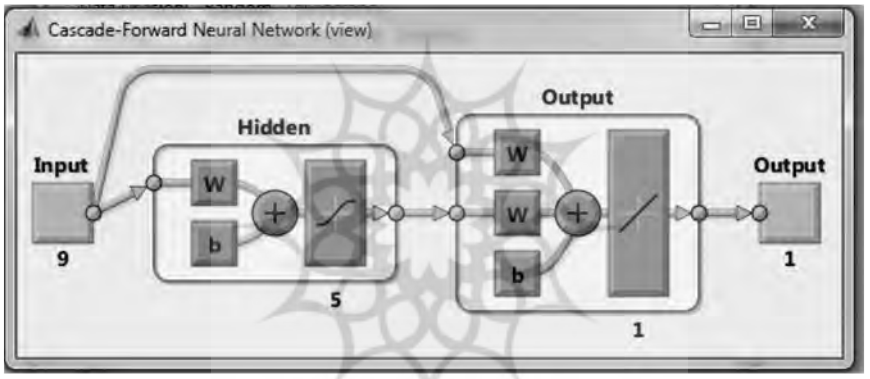
شکل شماره۵) عملکرد اجرایی الگوریتم درخت تصمیم



استفاده از الگوریتم دسته‌بندی شبکه‌های عصبی مصنوعی:

اطلاعات به دو طریق در شبکهٔ عصبی جریان دارند: ۱) زمانی‌که در حال یادگیری است؛ ۲) بعد از این‌که عمل یادگیری انجام شد. در این زمان‌ها الگوهای یادگیری به‌وسیلهٔ واحدهای ورودی وارد شبکه می‌شوند و لایه‌های واحدهای مخفی را برانگیخته می‌کنند و این لایه‌ها به واحدهای خروجی می‌رسند. این طراحی رایج را شبکه عصبی پیش‌خور می‌نامیم. همهٔ واحدها همیشه شلیک نمی‌شوند. هر واحدی اطلاعات ورودی را از واحدهای سمت چپ خود دریافت می‌کند و ورودی‌ها در وزن اتصالات مربوطبه خود ضرب می‌شوند. هرواحدی تمامی ورودی‌هایی را که دریافت می‌کند به این طریق جمع می‌زند و (در ساده‌ترین نوع شبکه) اگر جمع بیش‌از یک مقدار آستانه مشخص شد، این واحد شلیک می‌کند و واحدهای متصل به‌خود را (که در سمت راست هستند) راه می‌اندازد. (شکل ۶)

شکل شماره ۶) عملکرد اجرایی الگوریتم شبکه های عصبی پیش‌رو



برای یادگیری یک شبکه عصبی، به بازخورد نیاز داریم. همان‌طورکه به کودکان گفته می‌شود چه چیزی درست است و چه چیزی غلط. درواقع در این روش ما از باخورد استفاده می‌کنیم. به داده‌های ورودی (۹ متغیر) در هر مرحله بر اساس مقدار داده‌ای وروری و خروجی ـ که بیانگر وضعیت فرد بیمه شده است ـ وزنی اختصاص داده می‌شود و همین اولویت بندی در مرحله بعد، خود به‌عنوان یکی از ورودی‌ها عمل می‌کند و مبنای تصمیم‌گیری صحیح‌تر برای تجسم الگوی دقیق‌تر می‌گردد. (شکل ۶-۶) مانند زمانی‌که می‌خواستیم برای اولین‌بار بازی بولینگ را یاد بگیریم: وقتی توپ سنگینی برمی‌داریم و آن را پرتاب می‌کنیم، مغز ما به‌سرعت چگونگی حرکت توپ و مسیر آن را مشاهده می‌کند و میزان دقت‌مان را بررسی می‌کند ؛ دفعهٔ بعدی که دوباره نوبت ما رسید، مغزمان اشتباهات بار قبلی خود را به‌یاد می‌آورد و حرکت‌ش را باتوجه به آن اشتباهات اصلاح می‌کند، پس امیدوار می‌شویم که این‌بار توپ را بهتر از قبل پرتاب کنیم؛ بنابراین در این مرحله برای مقایسه نتیجه قبلی با نتیجه دلخواه‌مان از «بازخورد» استفاده می‌کنیم. به‌عبارت دیگر، این بازخورد تفاوت‌ها را مشخص

می‌کند و در دستور کار خود برای دفعه بعدی تغییراتی ایجاد می‌کند : باشدت بیشتر پرتاب‌کردن؛ کمی به‌سمت چپ پرتاب‌کردن؛ دیرتر رهاکردن، و ... . هرچه تفاوت بین نتایج حقیقی و نتایج دلخواه بیشتر و بزرگ‌تر باشد، تغییرات نیز بیشتر خواهد شد.

یادگیری با استفاده از یک روند بازخوردی یا پس‌انتشار انجام می‌گیرد؛ نتیجه‌ای‌که در دیاگرام نهایی نرم‌افزار متلب قابل مشاهده است. (شکل ۶-۶) این عمل عبارت است از مقایسه خروجی تولیدی یک شبکه با خروجی دلخواهی که انتظار داریم. از تفاوت بین این دو خروجی، برای تغییر و اصلاح وزن‌های اتصالات بین واحدهای شبکه استفاده می‌شود، با این تفاوت که این روش برعکس است، یعنی از واحدهای خروجی به‌سمت واحدهای مخفی و سپس از آنجا به‌سمت واحدهای ورودی می‌رویم. پس‌انتشار با کاهش تفاوت بین خروجی واقعی و خروجی دلخواه، تاحدی‌که این دو خروجی یکسان شوند، جلو می‌رود تا شبکهٔ عصبی، دقیقاً همان‌طوری‌که باید و انتظار می‌رود، ترسیم شود.

استفاده از الگوریتم نزدیک‌ترین همسایه KNN :

«نزدیک‌ترین همسایه» شاید ساده‌ترین الگوریتمْ در بحثِ طبقه‌بندی باشد. ما مجموعه‌ای از بیمه‌شدگان داریم با ۸ ویژگی کلیدی که بر اساس مقادیر عددی ۱ تا ۳ مقداردهی شده‌اند. هدف، شناسایی افرادی است که با توجه به ویژگی‌های تعریف‌شده در گروه افراد متقلب و یا سالم قرار می‌گیرند. در واقع، سیستم باید افرادی را انتخاب کند که احتمال متقلب‌بودن آن‌ها زیاد است.

داده‌ها در ۱۰ مرحله و با ۸ ویژگی در اختیار الگوریتم قرار می‌گیرند و هر بار ۷ گره ارزیابی می‌شوند. در پایان، بر اساس نتایج الگوی ترسیمی از افراد متقلب و سالم تبیین می‌گردد. داده‌های هر فیلد با نزدیک‌ترین داده، مقایسه می‌شوند و الگوی نهایی بر اساس مقدار KNN اختصاص‌داده‌شده ـ که مقدار ۰ یا ۱ برای هر داده ورودی است ـ ، ترسیم می‌گردد. پس در هر ارزیابی، ۷ داده با هم مقایسه شده و پارامترهای افراد سالم و متقلب اولویت‌بندی می‌گردد.

معیارهای اعتبار سنجی

برای ارزیابی روش پیشنهادی از معیار اعتبارسنجی «فیشر» استفاده شده است.

$$\text{F-Meature} = 2\ \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

این معیار که سال ۱۹۷۴ «ون ریجسبرگن» آن را ارائه داده، ترکیبی از معیارهای دقت و بازیافت است که ازجمله معیارهای معمول ارزیابی ـ به‌ویژه در مواقع کار با مجموعه‌های غیرمتعادل ـ می‌باشد. معیار فیشر مطابق فرمول زیر میانگین توافقی دقت و بازیافت بوده که به‌ترین مقدار این معیار ۱ و بدترین آن ۰ است.

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

در فرمول بالا، Precision به‌معنی دقت بوده و در حقیقت به‌معنای نزدیک بودن مقادیر اندازه‌گیری به همدیگر است، خواه این مقادیر واقعیت را نشان دهند یا خیر. دقت به‌صورت رابطه زیر تعریف می‌شود:

TP: تعداد افراد متقلب که سیستم به‌درستی آن‌ها را متقلب معرفی کرده است.

FP: تعداد افراد سالمی که سیستم به‌اشتباه آن‌ها را متقلب معرفی کرده است.

همچنین Recall به معنی بازیافت است که مطابق با فرمول زیر محاسبه می‌شود که برابر است:

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

FN: تعداد افراد متقلب که سیستم به‌اشتباه آن‌ها را سالم معرفی کرده است.

TN: تعداد افراد سالمی که سیستم به‌درستی آن‌ها را سالم معرفی کرده است.

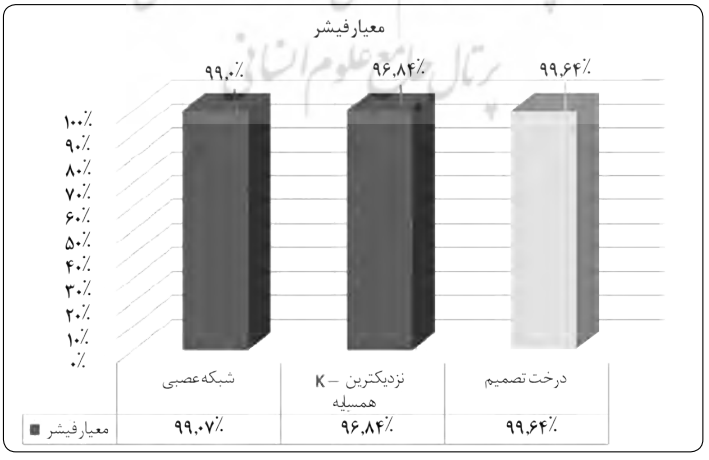
معیار صحت Accuracy با استفاده از برچسب کلاس‌ها، میزان دقت الگوریتم در انتساب کلاس‌ها به نمونه‌ها را نشان می‌دهد. می‌دهد.

$$\text{Accuracy} = \frac{TP + TN}{N}$$

جمع‌بندی

در این مقاله سه روشِ داده‌کاوی: درخت تصمیم، نزدیک‌ترین همسایه و شبکه‌های عصبی برای ساخت مدل‌هایی جهت شناسایی موارد تقلب در سازمان‌های بیمه‌گر معرفی شدند. در ادامه، این روش‌ها روی داده‌های واقعی آزمایش شدند و کارایی هر روش سنجیده شد. روش درخت تصمیم با دقت ۹۹,۶۴ درصد در شناسایی صحیح موارد تقلب به‌ترین کارایی را در مقایسه با دو روش دیگر داشت. همچنین باید به این نکته توجه داشت که متغیرهای به‌کار رفته در هر سه روش یکسان بوده است.

شکل شماره۷) نمودار نتایج آزمایش الگوریتم‌های داده‌کاوی برای تجمیع داده‌های استخراج‌شده



همان‌طورکه نتایج فوق نشان می‌دهد، الگوریتم درخت تصمیم در این مجموعه داده‌ها با توجه به معیار فیشر کارآیی بهتری دارد و عملکرد بالای ۹۹ درصد را نشان می‌دهد.

بررسی نتایج

این تحقیق با هدف پیش‌بینی رفتارهای متقلبانه بیمه‌شدگان تأمین اجتماعی و به‌منظور کشف تقلب در سه حوزه هزینه‌ای (دفترچه‌های درمانی، پرداخت کمک هزینه مرخصی زایمان- بیمه بیکاری و دریافت مستمری بازنشستگی و فوت) با استفاده از سه الگوریتم داده‌کاوی (درخت تصمیم، شبکه‌های عصبی مصنوعی و نزدیک‌ترین همسایه KNN) انجام شد. درنهایت، نتایج نشان داد که الگوریتم درخت تصمیم با ضریب اطمینان نزدیک به ۱۰۰ درصد و با توجه به معیار فیشر نسبت به سایر الگوریتم‌های مورد آزمایش برتری دارد. در زمینهٔ بررسی پیش‌بینی رفتار هم، تحقیقات مختلفی صورت گرفته است که نتایج متفاوتی را گزارش می‌دهد. همچنین در این میان، مقاله‌ای به بررسی برخی برنامه‌های کاربردی در مقیاس‌های بزرگ پرداخته است که به ما نشان می‌دهد، درخت‌های تصمیم، پیش‌بین‌های اهمیت و سیستم‌های GIS در پیش‌بینی نقاط بحرانی، بازدهی بالای ۹۵ درصدی دارند. [۸] در تحقیق دیگری، از سیستم تبدیل داده به دانش استفاده شده است که جهت پیش‌بینی مدل‌های بیمه‌ای اتومبیل به‌کار می‌رود. طبق نتایج این تحقیق، شاخص‌های یافت شده توسط درخت تصمیم‌گیری داده به دانش عبارتند از: انگیزه؛ رویکرد؛ روش‌ها و ابزار[۵]. نتایج به‌دست آمده از تحقیقی با عنوان «کشف تقلب در نسخ دارویی» به کمک روش‌های داده‌کاوی و ترکیب اطلاعات، از الگوریتم درخت تصمیم استفاده نمود، هر چند نتیجه نهایی به عددی حدود ۸۷ درصد رسید[۶].

بررسی ۴۹ مقاله منتشر شده در مجلات معتبر دربارهٔ داده‌کاوی و کشف تقلب، نشان دادند که از شش طبقه از وظایف/کاربردهای داده‌کاوی برای کشف تقلب‌های مالی استفاده شده است که این موارد عبارتند از: طبقه‌بندی، رگرسیون، خوشه‌بندی، پیش‌بینی، کشف داده‌های پِرت و همچنین تصویرسازی که در بخش ۳ به آن اشاره شد. آنها با استفاده از روش‌های معمول طبقه‌بندی، شامل شبکه‌های عصبی، شبکه‌های بیز ساده، درختان تصمیم و ماشین‌های بردار پشتیبان، وظایف طبقه‌بندی را درکشف تقلب‌های کارت‌های بیمه اعتباری، سلامت و بیمه خودرو بررسی کردند. همچنین رایج‌ترین روش‌های تکنیک خوشه‌بندی ازجمله: نزدیک‌ترین همسایه، روش بیز ساده و ... را آزمودند و با استفاده از تکنیک‌های هوش مصنوعی ، الگوهایی متناسب با هدف پیش‌بینی عمر دوره بازنشستگی بیمه‌شدگان به‌دست آوردند[۳]. کارشناسان در پژوهشی از داده‌های جمع‌آوری‌شدهٔ برخی از بازنشستگان متوفی سازمان تأمین اجتماعی در سطح کشور استفاده کرده‌اند. پس از تحلیل نتایج مشخص گردید که در این داده‌ها مدل حاصل از الگوریتم درخت تصمیم C۵,۰ با روش Boosting با روش کلاسه‌کننده‌های چندتایی از دقت بالاتری برخوردار است و قواعد نسبتاً مطلوبی را ارائه می‌دهد.

با به‌کارگیری الگوریتم‌های داده‌کاوی CART و شبکه عصبی به بررسی عوامل مؤثر در میزان مشروطی دانشجویان دانشگاه پیام نور استان قم پرداختند. آنها بعد از ایجاد مدل توسط الگوریتم‌ها به‌صورت تنها، مدل‌ها را در نرم‌افزار clementine با استفاده از گره Ensemble با یکدیگر ادغام نموده و مدل ترکیبی را ساختند. در نهایت مدل‌های به‌دست آمده را ارزیابی کردند. از آنجایی‌که تعداد فیلدها در بانک اطلاعاتی زیاد بود، با استفاده از گره feature Selection، با انتخاب هدف، فیلدهایی را که در هدف تأثیر کمتری داشتند، حذف کردند که تا حدودی از پیچیدگی مدل بکاهند.

بررسی نتایج بیانگر این نکته است که الگوریتم‌های استفاده شده در این پژوهش به‌دلیل انتخاب هوشمندانه نقاط بحرانی تقلب و شناخت متغیرهای کلیدی حاصل از پالایش پرسشنامه‌ها و داده‌های استخراج شده در مرحله اول تحقیق و طراحی فایل اکسل، شاملِ پیش‌بینی همهٔ امکان‌های بروز تقلب، به ثبت نتیجه‌ای نزدیک به ۱۰۰ درصد در مراحل آزمایش‌ها منجر شد. (عکس سایر پژوهش‌های صورت گرفتهٔ قبلی). در مراحل و طراحی کدهای SQL واحد اجرا، که با اطلاعات حقیقی پایگاه داده سازمان تأمین اجتماعی انجام گرفت، در نتیجه ۹۰ درصد موارد بروز تقلب شناسایی گردید.

از طرفی، پیاده‌سازی الگوریتم‌های استفاده‌شده در این پژوهش و طراحی ماشین یادگیری در نرم‌افزار متلب۲۰۱۶ نسخه، در کنارِ استاندارد بودن توابع حوزه داده‌کاوی این نرم‌افزار، نکته‌ای بارزی به‌شمار می‌رود. در حالی‌که سایر پژوهش‌های مشابه این ویژگی را نداشته اند. (عمدتاً از نرم‌افزار کلمنتاین استفاده شده است). به‌همین دلیل، نتیجه نهایی از ضریب اطمینان بالاتری برخوردار بود.

همان‌طورکه در بخش قبلی اشاره کردیم، مهم‌ترین حوزه‌های تقلب در سازمان بیمه‌ای تأمین اجتماعی که در بلند مدت هزینه‌های گزافی را چه در بخش بیمه‌ای و چه در بخش درمانی به این سازمان تحمیل می‌کنند، اولویت‌بندی شدند و با تکنیک طبقه‌بندی و استفاده از الگوریتم‌های یادگیری ماشین، موارد تقلب در ۹ پارامتر و ۳ بخش خدمات‌رسانی (دفترچه‌های درمانی، پرداخت کمک هزینه مرخصی زایمان و بیمه بیکاری و نیز پرداخت مستمری غیرِ کاری) شناسایی و نتایج حاصل از آزمون یادگیری روی داده‌های واقعی آزمایش شدند.

نتایج حاصل از آزمایش‌ها و بررسی متغیرهای کلیدی که در طراحی ماشین یادگیری فرآیند داده‌کاوی استفاده شد، بیانگر این نکته بود که شناسایی نقاط مستعد تقلب به‌دلیل آگاهی کارفرمایان و بیمه‌شدگان از قوانین و مقررات بیمه‌ای با حداقل سابقه ممکن وگریز از چارچوب‌های بازرسی کارگاهی؛ به‌جهت استفاده حداکثری از خدمات ارائه شده سازمان تأمین اجتماعی تقریباً در سه حوزه مورد بحث انجام می‌پذیرد و در سایر موارد به‌دلیل انجام پروسه‌های کنترل سیستمی و انسانی در سال‌های اخیر تا حدود زیادی قابل پیشگیری است؛ هرچند باید این نکته را در نظر داشت که نبود رویهٔ ثابت اداری باعث گردیده است که در اکثر شعبه‌ها به‌صورت سلیقه‌ای رفتار شود که در ادامه به

بعضی از آنها اشاره می‌شود.

◈ بیمه‌شدگانی با حداقل سابقه پرداخت حق بیمه (۲ الی ۳ ماه در سال) بخش زیادی از هزینه‌های درمانی خود را از سازمان تأمین اجتماعی دریافت می‌کنند، زیرا الگوریتم پیشگیرانه‌ای که ارائه خدمات در سامانه متمرکز را به داشتن حداقل سوابق متوالی ۳ ماه و یا وجود بازرسی کارگاهی منوط نماید، وجود ندارد.

◈ بیمه‌شدگانی که به‌دلیل نقص در قانون مرخصی زایمان (ماده ۶۷)، با دارا بودن حداقل ۲ ماه سابقه در یک‌سال آخر منتهی به تاریخ زایمان، از ۶ الی ۹ ماه مرخصی استفاده می‌کنند. (عمدتاً در کارگاه‌های کمک دولت و پیمانکاری‌ها که از نظر قانونی ماهیت بازرسی ندارند.)

◈ بستگان متوفیانی که با اطلاع از نواقص ماده ۷۴ قانون (برقراری حداقل سابقه یکسال در ۱۰ سال به‌شرط وجود سابقه ۳ ماه در یکسال آخر منتهی به بیمه‌پردازی)، در کارگاه‌های پیمانکاری یا کمک دولت، نواقص سابقه‌ای متوفی را برطرف نموده و تا پایان عمر بازماندگان، مستمری آنها را دریافت می‌کنند.

◈ بیمه‌شدگان شاغلی که به طمع دریافت مقرری بیمه بیکاری، با پرونده‌سازی‌های سوری و تعامل ساختگی با کارفرمایان، از ضعف قانونی بیمهٔ بیکاری (داشتن حداقل یکسال سابقه متوالی در آخرین کارگاه) استفاده می‌کنند و با وجود داشتن کار و شغل، به‌صورت ظاهراً غیرارادی از کار بیکار شده و به دریافت مقرری بیمه بیکاری اقدام می‌نمایند و این رویه را به دفعات در طول دوره بیمه‌پردازی انجام می‌دهند.

◈ متقاضیان بیمه‌پردازی که با استفاده از خلأهای قانونی، از خدمات و تسهیلات یارانه‌ای دولت (کارگران ساختمانی، باربران، صیادان، قالی‌بافان و ...) استفاده می‌کنند؛ چراکه سامانه متمرکز کنترلی فی مابین سایر نهادها و سازمان‌ها (اداره کار، اصناف، وزارت امور خارجه، سازمان بنادر و کشتیرانی، سازمان فنی و حرفه‌ای) وجود ندارد و عمده بار مسئولیت نظارت و کنترل متقاضیان بردوش سازمان تأمین اجتماعی ـ به عنوان متولی نهایی ارائه خدمات ـ قرار گرفته است.

پیشنهادات

مسلم است که انجام اقدامات پیشگیرانه برای شناسایی موارد متعدد تقلب و مقابله با آنها، بدون ایجاد تغییرات اساسی در قوانین و بخشنامه‌های موجود و اصلاح نرم‌افزارهای مورد استفاده در شعب بیمه‌ای میسر نخواهد شد؛ هرچند در سال‌های اخیر در این زمینه پیشرفت‌های اساسی صورت گرفته است.

◈ متاسفانه افزایش سالیانه بدهی‌های دولت و تغییر ساختار تشکیلاتی سازمان در سال‌های اخیر، باعث شده است که تأمین اجتماعی ـ که در ابتدا به‌عنوان یک سازمان مستقل مردمی تاسیس

گردید ـ ، امروز به‌دلیل برخورداری از منابع به‌ظاهر عظیم مالی، جولانگاه دولتمردانی گردد که خود را به فراموشی زده‌اند و منابع مالی سازمانی را که حاصل دسترنج کارگران زحمت‌کش این مرز و بوم است، به نفع منافع ملی به تاراج می‌برند؛ بدون این‌که خود را موظف به پاسخگویی بدانند. در حالی‌که بر اساس قانون اساسی کشور، دولت موظف است سهم خود را از حق بیمه‌های پرداختی بیمه‌شدگان مشمول طرح‌های حمایتی و کارگاه‌های کمک دولت، به سازمان تأمین اجتماعی پرداخت نماید. امری‌که در سایر کشورها به‌دلیل تزریق به‌هنگام منابع مالی دولتی بدیهی به‌نظر می‌رسد، اما تا امروز برخلاف پیش‌بینی‌های بودجه سالیانه تقدیمی به مجلس شورای اسلامی، به‌دلیل ضعف دستگاه‌های نظارتی محقق نشده است. بنابراین کاهش منابع مالی و افزایش هزینه‌های این سازمان در بخش درمان و صندوق بازنشستگی باعث شده که در سال‌های اخیر، به‌جای ارائه خدمات رفاهی به کارفرمایان، بیمه‌شدگان و بازنشستگان، تمام ظرفیت‌های انسانی و مالی، صرف وصول مطالبات و بدهی‌های معوقهٔ دولتی و غیردولتی و رسیدگی به تخلفات گردد.

◈ انتقال خدمات ارائه‌شده در سطح شعب، از طریق سامانه‌های یکپارچه سراسری و راه‌اندازی بانک اطلاعات متمرکزی که ضمن تعامل با سامانه اداره ثبت احوال، وضعیت بیمه شده‌ها و افراد تبعی را به‌صورت آنلاین شناسایی کند. سال ۱۳۹۵ سامانه یکپارچه صدور/ تمدید دفترچه‌های درمانی راه‌اندازی گردید، اما جهت جلوگیری از هرگونه سوءاستفاده، بهتر است که سایر خدمات (پرداخت هزینه‌های کوتاه مدت و بلند مدت و ...) به‌صورت سامانه‌های متمرکز انجام گیرد، برای‌مثال در حال حاضر رسیدگی به وضعیت تأهل و فوتی افراد تبعی، جهت صدور و یا تمدید دفترچه‌های درمانی و پرداخت‌های مستمری بگیران، هر ۶ ماه یکبار توسط کارکنان انجام می‌شود که به‌دلیل زمان‌بَر بودن، با دقت کافی همراه نیست.

◈ متأسفانه در سال‌های اخیر، با افزایش روزافزون بیماری‌ها و تغییر قوانین وضع شده در خصوص صدور مجوزهای قانونی، دسترسی بیماران به داروهای خاص درمانی و افزایش سرسام‌آور قیمت اقلام داروئی در دوران تحریم زمینه‌های بروز تقلب را تا حد بسیار زیادی در میان بیمه‌شدگان و زیرمجموعه‌های سازمان نظام پزشکی کشور افزایش داده است. پزشکان و داروسازانی هستند که به‌بهانه‌های واهی، ازجمله عدم تسویه‌حساب به‌موقع مطالبات معوقه سازمان تأمین اجتماعی، در نوبت‌های متوالی بدون اطلاع بیمار، از داروهای کمیاب و گرانقیمت چک لیست‌های صوری تنظیم می‌نمایند و در نوبت‌های مختلف مبلغ‌های گزافی را از سازمان تأمین اجتماعی دریافت می‌کنند. همچنین شرکت‌های داروسازی متخلف که با پرداخت مبالغ هنگفت به پزشکان اقدام به تجویز داروهای غیرضروری می‌نمایند، هم در بخش تأمین دارو مشکل‌سازی کرده و هم به سبب اتمام دفترچه در دوره زمانی ۶ ماهه هزینه‌های چاپ دفترچه را به مجموعه هزینه‌ها اضافه نموده است.

به‌علاوه، اجتناب داروخانه‌ها از تحویل‌دادن نسخه کامل به بیمار و دریافت کل مبلغ داروهای تجویزشده از سازمان تأمین اجتماعی، یکی از نمونه‌های بارز تقلب به‌شمار می‌رود. از طرفی امکان صدور، تجدید و تمدید دفترچه‌های درمانی با حداقل سابقه یک روزه در سامانه نرم‌افزاری فراهم شده است، در صورتی‌که این مدت زمان باید به حداقل ۳ ماه تغییر کند؛ مگر در مواردی که نام بیمه‌شده در بازرسی کارگاهی ثبت شده باشد.

◈ در سال‌های اخیر تلاش بسیار زیادی برای حذف دفترچه‌های درمانی انجام گرفته ولی تا امروز به‌دلیل نبود زیر ساخت‌های لازم در سطح کشور این کار به مرحلهٔ اجرایی نرسیده است. از شواهد امر این‌طور به‌نظر می‌آید که در عمل، این برنامه با تفکرات کنونی اجرایی نخواهد شد و هزینه‌های گزافی را به سازمان تأمین اجتماعی تحمیل خواهد کرد. البته تجهیز شعبه سازمان به دستگاه‌های اسکن انگشت و اسکنرهای پیشرفته در سال ۱۳۹۵ اقدامی مؤثر در این زمینه بود، اما این برنامه تا زمان نگارش این تحقیق به مرحله اجرا نزدیک نشده است. بهتر نیست به‌جای درگیرکردن نیروی انسانی شاغل در شعب و صرف هزینه گزافی که خرید تجهیزات و زیرساخت‌های لازم به سازمان تحمیل می‌کند ، با کمی تغییر از بسترهای آماده‌ای استفاده نماییم که در سال‌های اخیر آزمایش شده و به شرایط ایده‌آل رسیده‌اند؟ می‌توانیم به‌جای صدور کارت‌های درمانی از کارت‌های الکترونیک ملی که توسط اداره ثبت احوال کشور صادر می‌شود، به‌عنوان کارت‌های درمانی و پرونده سلامت استفاده کنیم. به‌صورت آنلاین وضعیت بیمه شده‌های اصلی و تبعی را بررسی کنیم. همچنین در زمان ارائه خدمات درمانی و بازرسی کارگاهی از اثر انگشت‌های اسکن‌شده برای شناسایی بیمه‌شدگان اصلی و تبعی استفاده نماییم و با ارائه یک نرم‌افزار جانبی روی دستگاه‌های خودپرداز سیستم بانکی ATM و POS فروشگاهی، عملیات اعتبار دفترچه‌های درمانی را از طریق این سامانه‌ها انجام دهیم. امکاناتی که با وجود در دسترس بودن، می‌تواند حجم مراجعات را به شعب تأمین اجتماعی کاهش دهد.

◈ انتقال روند کنونی تأمین اعتبار دفترچه‌های درمانی ـ که با مراجعه مستقیم بیمه‌شدگان به شعب تأمین اجتماعی، شعب اقماری و کارگزاری‌ها انجام می‌گیرد ـ به سامانه پیامکی #۱۴۲* علاوه بر کاهش چشمگیر تقلب‌های صورت گرفته در روند کنونی تأمین اعتبار (جعل مهر شعبه و جعل تاریخ اعتبار) می‌تواند از حجم مراجعات ارباب رجوع کم کند و تا حد بسیار زیادی در کاهش هزینه‌های تحمیلی به سازمان تأمین اجتماعی مؤثر باشد.

◈ مهلت دو ماههٔ ارسال لیست حق بیمه برای کارگاه‌های کمک دولت، زمینه‌ساز حجم زیادی از تخلفات بیمه‌ای شده است. خاصه در موارد منجر به فوت بیمه‌شدگانی که سابقه لازم را جهت احراز شرایط ماده ۷۵ قانون تأمین اجتماعی ندارند. برای مثال، کارگاه‌های کمک دولت تا آخرین روز بهمن ماه، از نظر سیستمی برای ارسال لیست آذر ماه سال جاری فرصت دارند، اما بهتر است که ضمن

اصلاح بخشنامه‌های موجود، فرصت ارسال لیست را به‌صورت ماهیانه تعریف نمود و تنها، مهلت پرداخت حق بیمه را دو ماهه در نظر گرفت. با انجام این تغییر، فرصت ارائه لیست خلاف واقع تا حد بسیار زیادی از کارگاه‌های متخلف کمک دولت گرفته می‌شود.

◇ در سال‌های اخیر ارائه لیست‌های معوق کارگاهی، زمینه بروز تقلب‌های گسترده را فراهم نموده است. به‌ویژه در کارگاه‌های پیمانکاری که به بهانه‌های واهی، ازجمله: نبود بودجه، عدم ارسال مدارک بیمه‌شدگان، آماده نشدن صورت وضعیت‌های پیمان، مرخصی کادر واحد حسابداری، فراموشی و ... تخلف صورت می‌گیرد. هرچند از نظر قانونی ارائه لیست‌های معوق بر اساس یک چارچوب تعریف‌شده، انجام می‌شود، اما این کار در شعب تأمین اجتماعی به‌شکل سلیقه‌ای صورت می‌گیرد. در صورتی‌که ارائه لیست معوق کارگاه‌های پیمانکاری تنها به‌شرط مشابهت با لیست ماه قبل و همراه با مدارک مستدل پرداختی و تأییدیه واحد بازرسی کارگاهی، باید به‌عنوان عاملی بازدارنده در واحدهای اجرایی بیمه‌ای انجام شود.

◇ نقص موجود در ماده ۶۷ قانون تأمین اجتماعی (پرداخت کمک هزینه‌های زایمان با حداقل ۲ ماه سابقه حق بیمه در یکسال منتهی‌به تاریخ زایمان)، در سال‌های اخیر هزینهٔ گزافی را به سازمان تأمین اجتماعی تحمیل کرده است؛ در حالی‌که این مدت زمان سابقه در سایر کشورها بین ۶ تا ۷ ماه متغیر است. این مورد بیشتر در کارگاه‌های پیمانکاری دیده می‌شود، با علم به این‌که اصولاً این کارگاه‌ها ماهیت بازرسی نداشته و پیمانکاران برای ارائه لیست بیمه‌شدگان زن منع قانونی ندارند. هرچند به‌دلیل نبود دستورالعملی واحد، این کار در اکثر شعب تأمین اجتماعی به‌صورت سلیقه‌ای اجرا می‌شود. اغلب پیمانکاران عقیده دارند که با توجه به پرداخت بخشی از مبالغ قرادادهای منعقد شده توسط تأمین اجتماعی (۷ و الی ۱۵ درصد از کل قرارداد) در انت‌های قرارداد، لیست کارکرد بیمه‌شدگان زن در قالب قرارداد و تحت عنوان کد پیمانکاری به سازمان تأمین اجتماعی ارائه می‌گردد؛ در صورتی‌که شرکت‌های پیمانکاری موظف هستند اسامی این بیمه‌شدگان را تحت عنوان « دفتر شرکت پیمانکاری» با اختصاص کد کارگاهی اصناف که ماهیت بازرسی دارند، به سازمان تأمین اجتماعی ارائه دهد. ممنوع شدن ارائه لیست بیمه‌شدگان زن در قراردادهای پیمانکاری ـ مگر با تأییدیه واحد بازرسی و ارائه مدارک مستند قانونی ـ می‌تواند راهکاری بازدارنده باشد.

◇ نواقص موجود در طراحی سامانه بازرسی کارگاهی که خود از نقص بخشنامه‌های سال‌های اخیر سازمان تأمین اجتماعی نشئت گرفته، باعث شده است که واحد بازرسی در حاشیه قرار گیرد. واحدی‌که می‌تواند به‌عنوان یک عامل بازدانده قوی در برخورد با کارفرمایان و بیمه‌شدگان متقلب نقش اساسی را ایفا کند. انجام بازرسی کارگاهی حداقل یکبار و حداکثر دو بار در سال برای کارگاه‌های صنفی در صورت نداشتن تغییرات لیست کارگاهی، خود زمینه بروز تقلب و افزایش بدهی و نارضایتی کارفرمایان را تا حد بسیاری

افزایش می‌دهد، هرچند این نارضایتی در اکثر مواقع به‌حق بوده و خسارت‌های مالی زیادی را به کارفرمایان و سازمان تأمین اجتماعی تحمیل می‌کند. (هزینه‌های محاسبه، چاپ اعلامیه‌های بدهی، ابلاغ و جلسات هیأت). همچنین عدم راه‌اندازی نرم‌افزارهای لازم جهت استفاده از اسکن تصویر و اثر انگشت بیمه‌شدگان ـ با وجود تجهیز شعب به سخت افزارهای لازم ـ از مهم‌ترین عواملی است که زمینه بروز تقلب در روند اجرای بازرسی کارگاهی را فراهم می‌نماید. ارائه مشخصات شناسنامه‌ای غیر واقع و جایگزینی سایر افراد به‌جای بیمه‌شدگان غایب از مهم‌ترین چالش‌های بازرسان سازمان تأمین اجتماعی به‌شمار می‌رود. همچنین عدم ارتباط مؤثر سامانه بازرسی کارگاهی با خدمات ارائه شده کوتاه مدت و بلند مدت سازمان تأمین اجتماعی (دفترچه‌های درمانی، مرخصی زایمان، بیمه بیکاری، استراحت پزشکی، پرداخت مستمری و...) در سیستم نرم‌افزاری بیمه‌ای، نتایج حاصل از بازرسی کارگاهی را بی‌ثمر کرده است. برای‌مثال فرد بیمه‌شده در بازرسی کارگاهی ثبت نشده است، اما می‌تواند با ارائه لیست کارکرد، از خدمات دفترچه‌های درمانی و کمک‌های کوتاه مدت استفاده کند. از طرفی، نبود کنترل‌های نظارتی و نبود سیستم تعاملی میان واحد بازرسی و باجه دریافت در زمان ارائه لیست‌های حق بیمه ماهیانه، چه به‌صورت دستی و چه از طریق سامانه متمرکز لیست‌های اینترنتی سامانه Samt.tamin.i، باعث شده است که در سال‌های اخیر کارفرمایان به‌آسانی، لیست‌های تقلبی و غیر واقع ارایه دهند.

◈ سامانه ارسال لیست‌های اینترنتی ـ به آدرس Samt.tamin.ir ـ که با هدف کاهش مراجعات کارفرمایان و سهولت کار راه‌اندازی شد، به‌دلیل رعایت نکردن نکات امنیتی، خود به بستری مناسب جهت انجام اعمال خلاف قانون تبدیل شده است. همچنین به‌علت کاهش موانع نظارتی و کنترلی تا حد بسیار زیادی موارد تقلب را افزایش داده است، چراکه این سامانه فقط اطلاعات هویتی بیمه شده‌ها (نام و نام خانوادگی، شماره شناسنامه، شماره ملی) و وضعیت کارگاه را در زمان دریافت لیست، بررسی می‌نماید. ازاین‌رو بهتر است در فاصلهٔ ارسال لیست تا زمان تأیید نهایی، یک مرحله تأیید اولیه توسط کارکنان سازمان انجام گیرد تا موارد کنترلی، از جمله: (بازرسی کارگاهی، جنسیت، سن، مبالغ دستمزد و....) ملاحظه شود و در صورت عدم رعایت موازین قانونی با ذکر توضیحات توسط کاربر مربوطه به کارفرمایان اعلام گردد. همچنین از آنجاکه اکثر کارفرمایان به‌دلیل مشغله‌های کاری از دفترهای خدماتی اقدام‌به ارسال لیست حق بیمه می‌نمایند، بهتر است جهت جلوگیری از سوء استفاده‌های احتمالیِ افراد سودجویی که اقدام به ثبت کارکرد افرادی خارج از مجموعه کارگاه می‌نمایند، در زمان ارسال لیست یک کد تائید حاوی تعداد نفرات و مجموع عددی حق بیمه برای کارفرما ارسال شود و پس از ثبت کد ارسالی، داده‌های فایل‌های اطلاعاتی توسط سامانه دریافت گردد.

◈ عدم تناسب میان پرداخت‌های کنونی مبالغ حق بیمه (۱۲٪- ۱۴٪- ۱۸٪) بدون ارائه خدمات درمانی، بیمه‌های آزاد و اختیاری با توجه به کاهش سطح درآمد خانوارهای ایرانی در سال‌های اخیر باعث شده

است که زمینه تقلب در بخش‌های کمک دولت و پیمانکاری، رو به افزایش گذارد. بهتر است سازمان تأمین اجتماعی به‌جای صرف هزینه‌های میلیاردی سالیانه بازرسی‌های کارگاه‌های صنفی، ابلاغ بدهی‌ها و برگزاری جلسات هیأت‌های رسیدگی به بدهی‌های کارگاهی، با افزایش طرح‌های حمایتی دولتی و سازمانی، افزایش خدمات رفاهی و جبرانِ بخشی از حق بیمه سهم بیمه‌شده به‌جای تمرکز بر بیمه‌شدگان کارگاهی به‌سمت بیمه‌های آزاد و اختیاری گام نهد. اتفاقی که در اکثر کشورهای پیشرفته جهان با تزریق به‌هنگام منابع دولتی انجام گرفته و آزمون خود را با موفقیت سپری نموده است.

◇ در سال‌های اخیر وجود ضرائب متعدد محاسبه پیمان قراردادهای پیمانکاری، خود زمینه‌ساز بالاترین ضریب بروز تخلف سازمان یافته در میان کارکنان و واگذارنده‌های پیمان شده است. تغییر ضرائب محاسبه پیمان و گزارش‌های سوری صورت وضعیت اعلان درصدهای حسن انجام کار به‌صورت مکانیکی و دستی توسط واگذارنده‌های پیمان و ارائه مبالغ غیر واقعی کارکرد نهایی پیمان، به‌دلیل متغییر بودن ضرائب تعریف شده در قانون تأمین اجتماعی زمینه بیش‌ترین خسارت‌های مالی و بروز تقلب را از طریق رانت‌های اداری، پرداخت رشوه و ... فراهم نموده است. پیمانکارانی هستند که به‌همین واسطه و با پوشش مبالغ پرداختی نهایی به تهیه و تنظیم لیست‌های غیرواقع از بیمه‌شدگان شاغل در کارگاه اقدام می‌نمایند، در حالی‌که طبق قانون، بازرسان سازمان تأمین اجتماعی حق بازرسی از این کارگاه‌ها را ندارند. بنابراین، تعریف یکسان ضریب محاسبه کلیه قراردادهای پیمانکاری باید به‌عنوان عاملی اثرگذار در کاهش تخلفات سازمانی، در اولویت کاری مدیران ارشد سازمان قرار گیرد.

◇ می‌توان گفت تا زمانی‌که فعالیت‌های کارگاه‌های صنفی و پیمانکاری در بخش‌های مالی و بیمه‌ای در سطح کشور به‌صورت سیستماتیک و شفاف توسط نهادهای مرتبط قابل‌کنترل نباشد، زمینه‌های بروز تقلب در تمامی حوزه‌ها اجتناب‌ناپذیر است. عاملی‌که در بسیاری از کشورهای پیشرفته دنیا، از راه سامانه‌های متمرکز از سالها پیش، تحت کنترل کامل نهادهای نظارتی قرار گرفته و کارفرمایان خود را ملزم به رعایت آن می‌دانند. اختصاص کد منحصربه‌فرد به هر کارفرما جهت فعالیت‌های اقتصادی در یک سامانه مشترک تحت نظارت مستقیم وزارتخانه (کار و رفاه اجتماعی، اقتصاد و دارایی)، ازجمله اقداماتی است که می‌تواند در زمینه شفاف‌سازی فعالیت‌های اقتصادی انجام گیرد، به نحوی‌که کلیه فعالیت‌های مرتبط با کارکنان شاغل (پرداخت حقوق، بیمه و مالیات، مزایا و ...) در یک سامانه متمرکز مشترک و تحت یک کد منحصربه فرد اقتصادی برای هر کارفرما انجام گیرد و بر اساس فعالیت‌های انجام گرفته، به هر کارفرما امتیاز مثبت و یا منفی اختصاص داده شود؛ برای مثال: به ارسال به‌موقع لیست حق بیمه کارکنان و پرداخت حقوق و مزایا امتیاز مثبت و به تاخیر در پرداخت‌ها امتیاز منفی داده شود و همهٔ خدمات ارائه‌شده به کارفرمایان (پرداخت تسهیلات، واگذاری قراردادهای پیمانکاری، اعمال تخفیف در پرداخت‌های سالانه همچون مالیات، بیمه و ....) به استناد امتیازهای کسب شده از

این سامانه لحاظ گردد. به‌علاوه کنترل آنلاین شاغلان و بیکاران توسط هر نهاد نظارتی امکان‌پذیر باشد.

◈ قانون کنونی پرداخت بیمه بیکاری با حداقل یک‌سال سابقه پرداخت حق بیمه در آخرین کارگاه ـ به‌شرط نبود کارگاه پیمانکاری با قرارداد ثابت ـ خود زمینهٔ سوءاستفاده حداکثری بیمه‌شدگان و کارفرمایان را فراهم نموده است. در صورتی‌که سازمان تأمین اجتماعی می‌تواند با اختصاص حداقل تخفیف در پرداخت‌های ماهیانه به کارفرمایان و تغییر این مدت زمان از ۱۲ ماه به ۲۴ ماه، تا حد بسیار زیادی هزینه‌های پرداختی و زمینه‌های بروز تقلب را کاهش دهد. از سویی تعامل و همکاری سازمان تأمین اجتماعی به‌عنوان یک سازمان بزرگ بیمه‌گذار با سازمان فنی و حرفه‌ای و وزارت کار در برگزاری دوره‌های تخصصی و لزوم بهره‌گیری نیروی انسانی از تخصص‌های مورد نیاز صنعت کشور باعث خواهد شد که تأمین اجتماعی به‌جای پرداخت ماهیانه مبالغ کلان مقرری بیمه بیکاری، خود زمینه‌ساز اشتغال فراگیر جویندگان کار باشد و در دراز مدت با اصلاح قوانین حمایتی از کارفرمایان متخصص، زمینه برگشت سرمایه را از طریق بیمه‌پردازی همین متخصصانِ آموزش‌دیده فراهم نماید. کارفرمایان متخصصی که با تعریف قوانین حمایتی سازمان تأمین اجتماعی در جهت بیمه‌پردازی خود و کارگران شاغل در مجموعهٔ تحت مدیریت‌شان اقدام خواهند نمود. امری‌که در سال‌های اخیر به فراموشی سپرده شده است. زیرا سازمان تأمین اجتماعی تمام انرژی و توان خود را به کنترل و نظارت مستمر بر نیروی کار و وصول مطالبات معوق معطوف کرده است، به‌جای این‌که با صرف کم‌ترین هزینه‌ها ضمن حمایت از شخص کارفرما زمینه افزایش بیمه‌شدگان کارگاهی را فراهم کند. (اعطای تخفیفات سهم کارفرما به بیمه‌شدگان متخصص شاغل در کارگاه)

◈ استفاده از استارتاپ‌های موبایلی، ابزاری پر کاربرد و عاملی مؤثر در کاهش تقلب‌ها و افزایش سطح رضایت‌مندی در حوزه خدمات سازمان تأمین اجتماعی است. پیاده‌سازی خدمات حضوری و غیرحضوری تأمین اجتماعی در قالب نرم‌افزار استارتاپ و دریافت مبالغی هرچند اندک در قبال خدماتی همچون: امکان ارسال لیست‌های کارگاهی، صدور برگ پرداخت، واریز حق بیمه، مشاهده بدهی‌های کارگاهی، استعلام سابقه، ثبت شکایات، درخواست بازرسی کارگاهی، تأمین اعتبار دفترچه‌های درمانی، سیستم نوبت‌دهی آنلاین و ...، می‌تواند تا حد زیادی ضمن کاستن از حجم مراجعات به شعبه‌ها و کاهش موارد تقلب کارفرمایان، در هزینه‌های جاری نیر صرفه‌جویی نماید.

◈ ناممکن بودن وصول آنی و کنترل برگ پرداخت‌های واریز شده توسط کارفرمایان (جز بانک رفاه و تجارت) در سیستم نرم‌افزاری صبا ـ به‌ویژه در روزهای پایانی هر ماه ـ موجبات جعل مکرر پرفراژ و مهر بانک توسط افراد سودجو گردیده است. بهتر است که امکان وصول آنی برگ پرداخت‌های واریز شده برای تمامی بانک‌های طرف قرارداد در سیستم نرم‌افزاری صبا گنجانده شود.

منابع

- حسینی نسب، مرضیه؛ مشیری، بهزاد؛ رهگذر، مسعود و دیناروند، رسول(۱۳۸۷)، کشف تقلب در نسخ دارویی به کمک روش‌های داده‌کاوی و ترکیب اطلاعات. تهران انتشارات دبیرخانه دائمی کنفرانس داده‌کاوی ایران.
- خاکسار خیابانی، نسیم(۱۳۹۱)، پیشگیری، شناسایی و مقابله با کلاهبرداری در بیمه، پژوهشکده بیمه، اداره کتابخانه‌ها، اسناد علمی و نشریات -گزارش موردی، خرداد و تیر ۱۳۹۱.
- خاکسار خیابانی، نسیم(۱۳۹۱)، پیشگیری، شناسایی و مقابله با کلاهبرداری در بیمه، پژوهشکده بیمه، اداره کتابخانه‌ها، اسناد علمی و نشریات -گزارش موردی، خرداد و تیر ۱۳۹۱.
- قشقایی، محمدحسین(۱۳۹۶)، مجموعه قوانین و مقررات کار و تأمین اجتماعی، تهران انتشارات مذاکره.
- شفیع‌آبادی، محمدحسین (۱۳۹۶)، داده‌کاوی و پیاده‌سازی در متلب، تهران مؤسسه فرهنگی هنری دیباگران.
- ملائی، منیژه. پارسا، سودابه. (۱۳۹۵)، پیش‌بینی رفتار مشتریان با استفاده از تکنیک شبکه‌های عصبی مصنوعی، ماهنامه شباک (شبکه اطلاعات کنفرانس‌های کشور)، ۲(۳): ۱۱-۱۵ .۱۲۴.
- A. Hering ,L. Nisi ,O. Martius , M. Kunz , U. Germann , First published:(2016), Spatial and temporal distribution of hailstorms in the Alpine region: a long-term, high resolution, radarbased analysis`
- Albashrawi, M Lowell - Journal of Data Science, (2016) - researchgate.net ,data mining techniques were used to detect fraud across different financial applications .
- Chen, Ming-Syan, Jiawei Han, J. and Philip, S. Yu., (2016), Data mining: an overview from a database perspective, Knowledge and data Engineering, IEEE Transactions, 8(6), pp 866- 883.
- Ganesh Sundarkumar, G. and Vadlamani R., (2015), A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance, Engineering Applications of Artificial Intelligence, 37, PP. 368–377.
- Gepp, A. Wilson,J.Kumar, K.(2012) A comparative analysis of decision tree vis-à-vis other computational data mining technique in automotive insurance fraud detection. Journal of data Scince 10(2012) p.p 537-561
- Kumar, SP Yadav, S Kumar International , (2015),Journal of Industrial and Systems Engineering 8 (2), 135-156.
- Makridakis, 2017, The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. Futures; Article in Press.