فصلنامه مدل‌های کمی مدیریت، نشریهٔ علمی، سال دوم، شمارهٔ دوم، بهار۱٤۰٤ صص ۱-

ارایه مدل یادگیری ماشین در تشخیص تقلب‌های واقع شده در بیمه درمان

پویان سلطانیه [۱]، کیومرث پیرزاده کرسفی [۲]

۱ مدیریت صنعتی، مدیریت، رجا، قزوین، ایران، Pooian.soltany@gmail.com

۲فناوری اطلاعات، فناوری اطلاعات، غیاث الدین جمشید کاشانی، قزوین، ایران

تاریخ دریافت مقاله :۱٤۰٤/۰۲/۲۲

تاریخ پذیرش مقاله : **۱٤۰٤/۰۵/۱۲**

چکیده

امروزه تقلب یکی از مشکلات عمده صنعت بیمه است که کلان داده و یادگیری ماشین در تلاش برای حل آن هستند. این تحقیق به بررسی تقلب‌های بیمه درمان با استفاده از الگوریتم‌های یادگیری ماشین و داده‌کاوی می‌پردازد. این رویکرد نوآورانه می‌تواند دقت تشخیص تقلب را افزایش دهد و در کاهش هزینه‌های سلامت مؤثر باشد. ما روش نظارت شده ای را اتخاذ کردیم که برای ادعای داده های خودرو یک شرکت بیمه ناشناس اعمال می شود. هدف ما ارائه رویکردی است که ارتباط نتایج هوش مصنوعی را بهبود می بخشد. این مطالعه نشان می دهد که الگوریتم جنگل تصادفی بهترین عملکرد را بر اساس معیار K-measure و بهترین امتیاز را با توجه به ارزیابی RMSE دارد. جنگل تصادفی همچنین بهترین نسبت بین یادآوری و صحت را دارد، سهم تقلب‌های کشف شده در بین کل تقلب‌ها ۲۳٫۸٪ است.

کلمات کلیدی

یادگیری ماشینی، تشخیص تقلب بیمه ای، یادگیری تحت نظارت، جنگل تصادفی.

ارایه مدل یادگیری ماشین در تشخیص تقلب‌های واقع شده در بیمه درمان

۱.مقدمه

در دنیای امروز معماری سازمانی یکی از پرکاربردترین راهکارهای سازمان برای همسویی استراتژیک فناوری اطلاعات با کسب و کار سازمان و ابزار قدرتمندی برای ساماندهی به سیستم های اطلاعاتی پیچیده می باشد. در کسب و کار بیمه معماری داده یکی از وجوه اساسی معماری در کنار معماری کسب و کار، معماری اپلیکیشن و معماری تکنولوژی است و استفاده از تحلیل داده در راستای گسترش ضریب نفوذ بیمه، پیشگیری و کشف تقلب بسیار کاربردی می باشد. بیمه درمانی محدوده وسیعی از برنامه های بیمه ای است که هزینه های ناشی از بیماری ها، تصادفات و ناتوانی ها را تحت پوشش قرار می دهد. متاسفانه با توجه به گسترش نظام بیمه در کشور ایران، رویکرد تقلب و سوء استفاده از انواع بیمه ها نیز گسترش یافته است، و هر ساله زیانه ای بسیاری را متوجه شرکت های بیمه ای می کند. رویکرد پیشنهادی استفاده از روش های داده کاوی جهت استخراج دانش از انبوه داده ها و کشف تقلب است. در ایران اکثر تحقیقات انجام شده در این زمینه متعلق به رشته اتومبیل بوده و رشته درمان تاحدودی مورد غفلت قرار گرفته است. در همین راستا هدف پژوهش حاضر ارایه مدل یادگیری ماشین در تشخیص تقلب‌های واقع شده در بیمه درمان بوده که طی فصل حاضر به کلیات پژوهش پرداخته شده است.

تقلب و کلاهبرداری بیمه درمان در حال افزایش است، تقلب از بسیاری جهات اثرات و پیامدهای منفی بر سازمان می‌گذارد. علیرغم وجود پیشرفت‌های فراوان در شناسایی این تقلب‌ها، هزینه و ضرر و زیان زیادی در اثر این کلاهبرداری‌ها، به این شرکتها وارد می‌نماید. تقلب در بیمه ممکن است در مراحل مختلف و توسط اشخاص مختلفی رخ می‌دهد از جمله بیمه‌گذاران جدید، بیمه‌گذاران فعلی، اشخاص ثالث زیان دیده یا متخصصانی که به بیمه‌گذاران خدمات ارائه می‌دهند.

هزینه بهداشت و درمان با توجه به جمعیت، اقتصاد، جامعه، و تغییرات قانون به سرعت در حال افزایش است(دهیب[1] ۲۰۲۰). این افزایش در هزینه های بهداشت و درمان بر دولت و سیستم‌های بیمه سالم خصوصی تاثیر می گذارد. رفتارهای متقلبانه ی ارائه دهندگان بهداشت و درمان و بیماران با تحمیل هزینه های غیر ضروری به مشکلی جدی برای سیستم های بیمه تبدیل شده است (احمدی‌نژاد[2] ۲۰۱۶). بنابراین شرکت های بیمه روش هایی را برای تشخیص تقلب ایجاد می کنند که عمدتا برگرفته از تجارب خبرگان بوده و کمتر به روشهای مبتنی بر تحلیل داده متکی است با توجه به حجم اسناد و انبوه بیمه شدگان و ارائه دهندگان خدمت، کشف تقلب[3]به صورت سنتی غیر ممکن است(عسکر زاده، تارخ ، ۱۳۹۶). به علاوه، در گذشته، پیامدهای مالی تقلب بیمهای در حدی نبود که ارزش بررسی و تلاش برای یافتن راه‌حلهای ممکن را داشته باشد، اما به‌تازگی این وضع دگرگون شده و شرکتهای بیمه‌ای اهمیت مطالعه و ضرورت بررسی بیشتر عوامل تقلب بیمه‌ای را درک کرده‌اند. امروزه با توجه به در صنایع مختلف، استفاده از تکنیکهای هوش مصنوعی، مانند شبکه عصبی مصنوعی[4]،یادگیری ماشین و الگوریتم ژنتیک (عبادی و دیگران[5]،۲۰۱۷) به‌دلیل توانمندی بالایی که در مدل‌کردن مسائل پیچیده دارند، به ابزار رایج در کشف تقلب تبدیل شده‌اند (تقوی فرد و جعفری، ۱۳۹۴) و با ظهور فناوریهای مختلف، فرصتهایی برای کشف بهتر تقلب را فراهم ساخته است. در این بین، نگرش داده‌کاوی به عنوان روشی مؤثر و کارامد مطرح می‌شود. با تکیه بر داده‌کاوی و تکنیک‌های یادگیری ماشین در زمان کوتاه و با دقت بالا (لی[6]و دیگران، ۲۰۱۸) امکان شناسایی الگوهای رفتاری را در بین حجم زیادی از داده فراهم می‌آورد که این امر امکان شناسایی و کشف الگوهای مشکوک و یا حتی تقلب را میسر می‌سازد (لیو[7]و واسارلی، ۲۰۱۳). تکنیک‌های داده‌کاوی در بیمه درمان شامل برخی موارد مانند اطلاعات اولیه سیستم بیمه و ، تقلبی بودن رفتارها و تحلیل ویژگی‌های آن می‌شود (راوته و آنارادا،[8]۲۰۱۵) . در روش‌های یادگیری ماشین، الگوریتم‌هایی که براساس

---

[1] Dhieb
[2] Ahmadinejad
[3] Fraud detection
[4] Artificial Neural Networks (ANN)
[5] Abadi et al
[6] LI, et al
[7] Liu,Vasarhelyi
[8] Rawte, V., & Anuradha

داده‌ها توانایی یادگیری و پیش‌بینی دارند، ایجاد و مورد بررسی قرار می‌گیرد. چنین الگوریتم‌هایی صرفا از دستورهای برنامه پیروی نمی‌کنند بلکه از طریق مدلسازی داده‌های ورودی نمونه، پیشبینی یا تصمیم‌گیری می‌کنند (ملکی و دیگران، ۱۴۰۰، صفحه ۲۶۹).

۱-۱.اهداف

روش‌های ماشین بردار پشتیبان[9]، درخت تصمیم، دسته‌بند بیز[10]و روش‌های مبتنی بر یادگیری عمیق[12]روش‌های یادگیری ماشینی هستند که برای تشخیص خودکار تقلب و کلاهبرداری به کار گرفته شده‌اند (شیرگیو و دیگران [13]، ۲۰۱۹). هر چند تحقیقات زیادی در حوزه استفاده از روش‌های داده‌کاوی برای تشخیص تقلب در حوزه بیمه در ایران انجام گرفته است اما به موضوع تشخیص تقلب در بیمه درمان با روش‌های داده‌کاوی بسیار کمتر پرداخته شده و تاحدودی مورد غفلت قرار گرفته است. بنابراین، پژوهش حاضر به‌دنبال آن است با توجه به اهمیت موضوع تشخیص تقلب در بیمه درمان، به ارائه یک مدل پیش‌بینی‌کننده به کمک روش‌های یادگیری ماشین بپیردازد و در پایان محقق به دنبال پرسش به این پاسخ می‌باشد که آیا می‌توان مدلی ارائه داد که بتواند رفتارهای متقلبانه بیمه درمان را با دقت مطلوب پیش‌بینی کرد.

۲_۱. اهمیت و ضرورت پژوهش

تقلب تأثیرات بسیاری چه بصورت مستقیم و چه غیرمستقیم بر روی بیمه‌گران و بیمه‌گذاران دارد(عبد اللهی[14]۲۰۱۶). تقلبات بیمه‌ای می‌تواند منجر به افزایش مخاطرات اجتماعی و مخاطرات اخلاقی از طریق رفتارهای بیمه‌گذاران و افراد درگیر شود که در درازمدت کسری بودجه و بار مالی سنگینی را به صنعت بیمه تحمیل می‌کند(گوپتا[15]۲۰۲۱). تقلب بیمه‌گذاران به عنوان یک تهدید جدی برای رشد(هانکوک[16]۲۰۲۱)، توسعه و سودآوری صنعت بیمه مطرح می شود که با زیر سوال بردن اصل حسن نیت، فضای اعتماد و شفافیت را از بین خواهد برد. با بروز تقلب در صنعت بیمه ممکن است ارائه خدمات به جامعه هدف محدود شده و رضایتمندی بیمه‌شدگان و ذی‌نفعان بیمه کاهش یابد.

با توجه به سالنامه آماری بیمه مرکزی در سال ۱۳۹۷، مجموع حق بیمه‌های تولیدی در کل صنعت بیمه (دولتی و غیردولتی) حدوداً ۴۵۴۴۳ میلیارد تومان و مجموع کل خسارات پرداختی در کلیه رشته‌ها تقریباً برابر با ۲۸۸۶۸ میلیارد تومان بوده است. اگر فرض شود که تنها ۱۰ درصد خسارات پرداختی مربوط به ادعاهای متقلبانه باشد، با اعمال این مبلغ بر خسارات کشور، مجموع تقلبات بیمه‌ای در کل صنعت حدود ۲۸۸۶ میلیارد تومان می‌شود که مبلغ قابل توجهی است و این آمارها لزوم تحلیل و بررسی تقلب را در جهت کاهش هزینه های ناشی از وقوع آن آشکار می‌سازد(بهرامی نسب، ۱۴۰۱).

یکی از رشته‌های بیمه ای که در آن تقلب زیادی به روش‌های گوناگونی صورت می‌پذیرد، رشته درمان بویژه در بخش غیردولتی است. از نظر خبرگان حوزه بیمه درمان تکمیلی، مهم‌ترین پیامد تقلب بر صنعت بیمه جهان «غیرواقعی بودن حدود ۲۰ الی ۳۰ درصد خسارت‌ها» عنوان شده است. طبق برآورد خوشبینانه کارشناسان داخلی، تقلب در رشته درمان برای صنعت بیمه ایران نیز بین ۱۰ تا ۱۵ درصد برآورد شده است. با نگاهی به میزان خسارات پرداختی شرکت‌های بیمه در رشته درمان در بخش غیردولتی (تقریباً ۷۵۲۵ میلیارد تومان) در سال ۱۳۹۷، حجم خسارات پرداختی شرکت‌های بیمه غیردولتی از محل تقلب صورت‌گرفته در حدود ۷۵۲ تا ۱۱۲۸میلیارد تومان برآورد می‌گردد. تعداد خسارات پرداختی این بخش همچنین در سال ۱۳۹۷، ۴۲۲۹۶۵۷۴ مورد

[9] Support vector machines(SVM)
[10] Decision Tree
[11] Naive Bayes classifier
[12] Deep learning
[13] Shirgave et al
[14] Abdallah
[15] Gupta
[16] Hancock

بوده است که تنها نسبت به سال قبل از آن چیزی حدود ۸۹ درصد رشد داشته است(حسین زاده، ۱۴۰۱). این ارقام تا حدودی می تواند اهمیت توجه و حجم تقلبات را نمایان می‌کند.

فرضیات پژوهش

۱.الگوریتم جنگل تصادفی در شناسایی تقلب‌های بیمه درمان دقت بیشتری نسبت به سایر مدل‌ها دارد.

۲.استفاده از معیار K-Measure به جای F-Measure دقت و کارایی مدل‌های تشخیص تقلب بیمه درمان را افزایش می‌دهد.

۳.پیشینه پژوهش

از سال ۲۰۱۰ تحقیقات گسترده ای پیرامون این موضوع انجام شده است که طی اخرین تحقیقات که جز پایه های اساسی این تحقیق به شمار می رود می توان به تحقیقات پژوهشگر مستقل شاشانک آگراوال[۱۷]اشاره نمود که مقاله ای با عنوان"رویکرد هوشمند یادگیری ماشین برای شناسایی تقلب در بیمه پزشکی[۱۸] چاپ رسانده است.

پژوهشی دیگر نیز با عنوان"سیستم تشخیص کلاهبرداری و سواستفاده در پزشکی به وسیله یادگیری ماشین[۱۹]در دانشگاه چینگ دائو مورد بررسی قرار گرفته که یکی ازموارد مورد مطالعه در این مقاله نیز به شمار می آید.

Anoun و Maizate (2024) در مقاله‌ای با عنوان "معیاری برای تشخیص تقلب در بیمه سلامت با استفاده از تکنیک‌های یادگیری ماشین" که در مجله بین‌المللی هوش مصنوعی IAES منتشر شد، به بررسی کاربرد تکنیک‌های یادگیری ماشین در تشخیص تقلب در بیمه سلامت پرداختند. آنها با توجه به مشکل کمبود مجموعه داده‌های واقعی و برچسب‌های تقلب قابل اعتماد، از داده‌های واقعی و دقیق ادعاهای بیمه سلامت استفاده کردند تا شش مورد از رایج‌ترین الگوریتم‌های طبقه‌بندی نظارت شده، از جمله شبکه‌های عصبی را با هم مقایسه کنند. این مطالعه تحت هدایت کارشناسان بیمه انجام شد که قوانین استنتاج برچسب تقلب را ارائه کردند و نتایج را بررسی کردند.نتایج این تحقیق نشان داد که طبقه‌بندی نظارت شده می‌تواند به طور مؤثر برای تشخیص تقلب در بیمه سلامت استفاده شود و دقت تشخیص را با ضریب ۴,۲ بهبود بخشد (۸۴٪ فراخوانی برای نرخ مثبت ۲۰٪).

جدول ۱:لگوریتم تجزیه و تحلیل عملکرد مقایسه ای برای تشخیص تقلب در بیمه

تحقیقات	دقت
ژو[۲۰] و همکاران (۲۰۱۰)	۸۸,۷۰
سانادرکومار[۲۱] و همکاران (۲۰۱۵)	۵۸,۹۲
سانادرکومار و راوی[۲۲](۲۰۱۵)	۶۰,۳۱
نیان و همکاران (۲۰۱۶)	۷۶,۴۳

[1] Shashank Agarwal [7]

[1] An Intelligent Machine Learning Approach for Fraud Detection in Medical Claim Insurance: A Comprehensive Study [8]

[1] Medical Fraud and Abuse Detection System Based on Machine Learning [9]

[2] Xue [0]

[2] Sundarkumar [1]

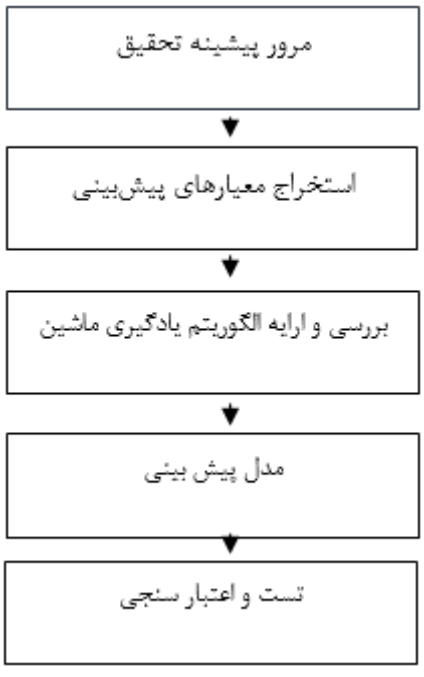
[2] Sundarkumar and Ravi [2]

شارمیلا و پنیگراهی[3](۲۰۱۷)	۸۷,۰۲
کونگای ژانک و همکارام(۲۰۲۰)	۸۰,۰۲
شاشانک آگراوال(۲۰۲۳)	۸۸
درخت تصمیم (پژوهش حاضر)	۸۹,۵۷

٤. روش پژوهش

این پژوهش به توسعه یک مدل پیش‌بینی‌کننده برای شناسایی تقلب در بیمه‌های درمان با بهره‌گیری از روش‌های داده‌کاوی و یادگیری ماشین می‌پردازد. این رویکرد نوآورانه، بهبود دقت در تشخیص تقلب را هدف قرار می‌دهد و قابلیت پیاده‌سازی در صنایع مختلف را دارد. با توجه به عدم توجه کافی به این مسئله در ایران و هزینه‌های بالای ناشی از تقلب، استفاده از الگوریتم‌های یادگیری ماشین ضروری به نظر می‌رسد. از موفقیت‌های قابل ذکر می‌توان به صرفه‌جویی‌های چشمگیر در کشورهایی مانند استرالیا اشاره کرد. در نهایت، این تحقیق ضرورت بررسی تقلب‌های بیمه را تأکید می‌کند. در ابتدا ، بررسی ادبیات تحقیق و سپس جمع آوری داده، پیش‌پردازش، آماده‌سازی داده‌ها، مشورت با کارشناسان خبرگان و تحلیل داده صورت می‌پذیرد. همچنین پیاده سازی با کمک روش‌های یادگیری ماشین توسط نرم‌افزار برنامه‌نویسی پایتون به همراه کتابخانه‌های داده کاوی از جمله keras , sckitlearn , TensorFlow انجام می‌گیرد. در تحقیق، الگوریتم‌های طبقه‌بندی مورد استفاده شامل جنگل تصادفی، تصمیم درخت، ماشین بردار پشتیبان، **K**–نزدیک همسایه، Adaboost، رگرسیون خطی، بیز ساده و پرسپترون چند خطی برای تشخیص تقلب می باشند.

در این پژوهش تکنیک‌های مختلف مورد ارزیابی و آزمایش روی مجموعه داده‌های بیمه خودکار از یک مخزن قابل اعتماد انجام خواهد شد تا بهترین طبقه‌بندی‌کننده را برای سیستم تشخیص تقلب پیدا گردد. علاوه بر این، سیستم در جنبه‌های دقت، یادآوری و امتیاز F1 برای همه الگوریتم‌ها تحلیل خواهد شد.



[2] Sharmila and Panigrahi [3]

۵.یافته ها

در این مطالعه، یک تحلیل مقایسه ای بین مدل های طبقه بندی مختلف با توجه به دو معیار عملکرد F-Measure و K-Measure ارائه شده است. شاخص RMSE نیز محاسبه و با نتیجه دو اندازه گیری مقایسه می شود. در بخش "انتخاب ویژگی ها" تعداد ویژگی ها را از ۳۶ به ۱۹ کاهش داده شد تا فقط ویژگی های مربوطه حفظ شود. همانطور که در جدول ۱ ارائه شده است، قبل از کاهش می‌توان مشاهده نمود که Adaboost.M1 بهترین امتیاز طبقه بندی (٪۹۶٫۶۸) را بر اساس معیار K-Measure نشان می دهد، اما درخت تصمیم (٪۲۴٫۷) بر اساس معیار F-Measure پیشتاز است.

جدول ۱: تحلیل عملکرد مقایسه ای قبل از انتخاب ویژگی در تعادل مجدد مجموعه داده

طبقه‌بندی کننده	یادآوری	صحت	دقت	F-measure	K-Measure
RandomForest	٪۱۲٫۶۷	٪۷۰٫۹۶	٪۷۶	٪۱۵٫۷۲	٪۹۶٫۶۸
AdaBoostM1	٪۱۳٫۳۲	٪۷۲٫۲۸	٪۷۸	٪۱۶٫۶۲	٪۹۶٫۴۴
Linear Regression	٪۳۴٫۷۸	٪۶۵٫۹۰	٪۷۳	٪۲۱٫۷۹	٪۹۲٫۷۸
MLP	٪۶۳٫۸۱	٪۶۵٫۲۶	٪۷۳	٪۲۴٫۴۰	٪۷۹٫۰۵
Decision Tree	٪۷۵٫۸۳	٪۶۴٫۹۰	٪۷۹	٪۲۴٫۷۹	٪۷۴٫۱۲
KNN	٪۷۶٫۴۴	٪۶۳٫۵۷	٪۷۷	٪۲۲٫۹۱	٪۷۱٫۹۵
Naïve Bayes	٪۸۴٫۶۱	٪۶۲٫۲۷	٪۷۳	٪۲۱٫۴۴	٪۶۶٫۵۱
SVM	٪۹۱٫۷۶	٪۶۲٫۶۴	٪۷۳	٪۲۲٫۲۲	٪۶۵٫۲۴

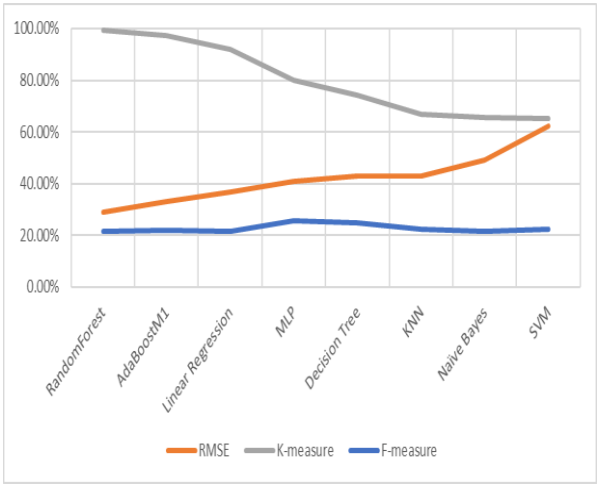
با این حال، پس از کاهش ویژگی‌ها، نتایج F-Measure و K-measure به‌طور مشهودی بهبود می‌یابند که در جدول ۱ نشان داده شده است. طبق معیار K-Measure، جنگل تصادفی بهترین امتیاز را با ٪۹۹٫۴ (+٪۴) حفظ می‌کند. در مقابل F-Measure، درخت تصمیم تقریبا همان امتیاز ٪۲۴٫۷ را حفظ می کند و رتبه اول را به MLP با ٪۲۵٫۶ (+٪۱٫۲) می‌دهد.

جدول ۲: تحلیل عملکرد

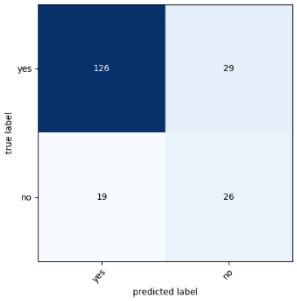
طبقه‌بندی کننده	یادآوری	صحت	دقت	F-measure	K-Measure
RandomForest	٪۲۸,۸۳	٪۶۹,۶۶	٪۸۹,۵۷	٪۲۱,۵۲	٪۹۹,۴۶
AdaBoostM1	٪۲۸,۳۸	٪۶۷,۹۲	٪۸۵,۳۷	٪۲۱,۹۴	٪۹۷,۳۴
Linear Regression	٪۳۵,۶۴	٪۶۵,۵۱	٪۷۶,۳۱	٪۲۱,۶۰	٪۹۱,۸۹
MLP	٪۶۵,۱۲	٪۶۶,۱۰	٪۷۹,۲۸	٪۲۵,۶۶	٪۷۹,۹۲
Decision Tree	٪۷۵,۸۳	٪۶۴,۹۰	٪۸۲,۴۴	٪۲۴,۷۹	٪۷۴,۱۲
KNN	٪۸۸,۵۲	٪۶۲,۹۳	٪۸۰,۱۹	٪۲۲,۵۶	٪۶۶,۶۹
Naïve Bayes	٪۸۷,۲۱	٪۶۲,۳۰	٪۸۱,۲۴	٪۲۱,۵۵	٪۶۵,۸۱
SVM	٪۹۱,۷۶	٪۶۲,۶۴	٪۸۰,۶۷	٪۲۲,۲۲	٪۶۵,۲۴

مقایسه‌ای پس از انتخاب ویژگی در تعادل مجدد مجموعه‌داده ها مسلماً، این دو رتبه متفاوت یا حتی مخالف هستند، در واقع جنگل تصادفی بر اساس معیار F-Measure در رتبه آخر قرار دارد، در حالی که بر اساس معیار K-Measure در رتبه اول قرار دارد. به منظور قضاوت در مورد ارتباط این دو معیار، شاخص RMSE همانطور که در شکل ۲رائه شده است، معرفی کردیم، نتیجه به دست آمده یک همبستگی رتبه قوی بین K-measure و RMSE را نشان می دهد. در واقع کمترین مقدار مربوط به جنگل تصادفی با ۲۹,۸ درصد است. با این حال شبکه عصبی چندلایه با ۴۱,۲۶ درصد در رتبه پنجم قرار دارد.

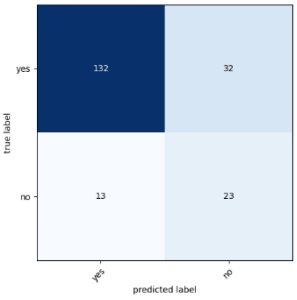
شکل ۱ :نمایش همبستگی بین K-Measure، F-Measure و RMSE



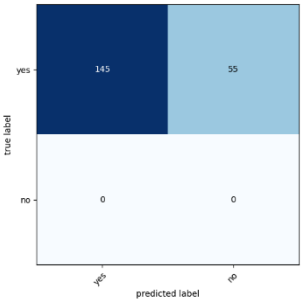
همچنین در ادامه با توجه به یافته‌های پژوهش و معیارهای ارزیابی برای هر هشت الگوریتم به دست آمده، در شکل‌های۲ الی ۹ ماتریس‌های درهم‌تنیدگی نمایش داده شده است.



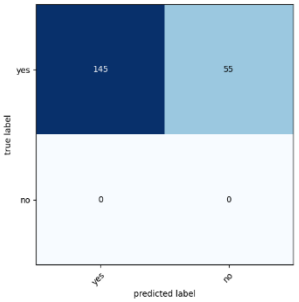
شکل ۲:ماتریس درهم‌تنیدگی الگوریتم جنگل تصادفی



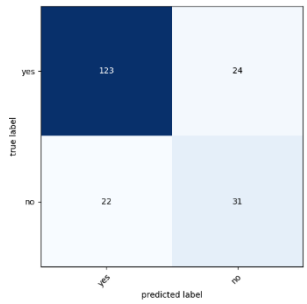
شکل ۳ :ماتریس درهم‌تنیدگی الگوریتم AdaBoostM1



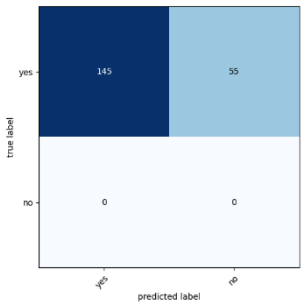
شکل ۴ :ماتریس درهم‌تنیدگی الگوریتم رگرسیون خطی



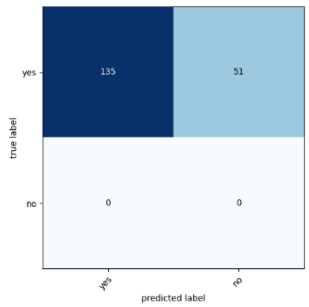
شکل ۵ :ماتریس درهم‌تنیدگی الگوریتم شبکه عصبی چندلایه



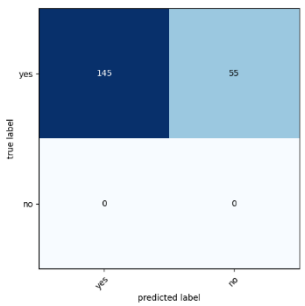
شکل ۶: ماتریس درهم‌تنیدگی الگوریتم درخت تصمیم



شکل ۷ :ماتریس درهم‌تنیدگی الگوریتم KNN



شکل ۸ :ماتریس درهم‌تنیدگی الگوریتم Naïve Bayes



شکل ۹ :ماتریس درهم‌تنیدگی الگوریتم SVM

با توجه به ماتریس‌های درهم‌تنیدگی ارائه شده، می‌توان به این نتیجه رسید که ملکرد الگوریتم‌ها در تشخیص تقلب در بیمه درمان:

۱. جنگل تصادفی: ماتریس درهم‌تنیدگی نشان می‌دهد که این الگوریتم در تشخیص تقلب در بیمه درمان به خوبی عمل کرده است. (شکل ۲)

۲. AdaBoostM1: ماتریس درهم‌تنیدگی نشان می‌دهد که این الگوریتم نیز در تشخیص تقلب در بیمه درمان به خوبی عمل کرده است. (شکل۳)

۳. رگرسیون خطی: ماتریس درهم‌تنیدگی نشان می‌دهد که این الگوریتم در تشخیص تقلب در بیمه درمان به نسبت دیگر الگوریتم‌ها kémتر عمل کرده است. (شکل ۴)

۴. شبکه عصبی چندلایه: ماتریس درهم‌تنیدگی نشان می‌دهد که این الگوریتم در تشخیص تقلب در بیمه درمان به خوبی عمل کرده است. (شکل ۵)

۵. درخت تصمیم: ماتریس درهم‌تنیدگی نشان می‌دهد که این الگوریتم در تشخیص تقلب در بیمه درمان به نسبت دیگر الگوریتم‌ها متوسط عمل کرده است. (شکل ۶)

۶. KNN: ماتریس درهم‌تنیدگی نشان می‌دهد که این الگوریتم در تشخیص تقلب در بیمه درمان به نسبت دیگر الگوریتم‌ها کمتر عمل کرده است. (شکل ۷)

۷. Naïve Bayes: ماتریس درهم‌تنیدگی نشان می‌دهد که این الگوریتم در تشخیص تقلب در بیمه درمان به نسبت دیگر الگوریتم‌ها کمتر عمل کرده است. (شکل ۸)

۸. SVM: ماتریس درهم‌تنیدگی نشان می‌دهد که این الگوریتم در تشخیص تقلب در بیمه درمان به خوبی عمل کرده است. (شکل ۹)

در نتیجه با توجه به ماتریس‌های درهم‌تنیدگی ارائه شده، می‌توان گفت که الگوریتم‌های جنگل تصادفی، AdaBoostM1، شبکه عصبی چندلایه، و SVM در تشخیص تقلب در بیمه درمان به خوبی عمل کرده‌اند. در حالی که الگوریتم‌های رگرسیون خطی، درخت تصمیم، KNN، و Naïve Bayes به نسبت دیگر الگوریتم‌ها عملکرد کمتری داشته‌اند.

٦.بحث و نتیجه گیری

در این کار، مقایسه ای از هشت الگوریتم پرکاربرد یادگیری ماشین ارائه شده است و عملکرد آنها با دو روش ارزیابی، مقایسه شده است تا مشخص شود که کدام اگوریتم مناسب ترین روش است، تا در داده های دنیای واقعی برای پیش بینی تقلب اعمال می شود. این مطالعه نشان می دهد که الگوریتم جنگل تصادفی بهترین عملکرد را با ارزیابی K-measure و بهترین امتیاز را با RMSE دارد. جنگل تصادفی همچنین بهترین نسبت بین یادآوری و صحت را دارد، سهم تقلب‌های کشف شده در بین کل تقلب‌ها ۲۳٫۸٪ است. از سوی دیگر، صحت ۶۹٫۷ درصدی را داریم که بالاترین امتیاز را بین مدل ها دارد، یعنی تقریبا در ۰٫۷ مواردی که ادعایی به عنوان تقلب پیش بینی شده است، مدل صحیح است. علاوه بر این، اگر مقدار صحت را با نتایج بدست آمده توسط تحقیقات قبلی ذکر شده در جدول ۵–۱ آمده است، مقایسه کنیم، بالاترین امتیاز را با ۸۹٫۵۷ نسبت به جنگل تصادفی کسب کرده ایم. علاوه بر این، می‌توان نتیجه گرفت که روش ارزیابی عملکرد برای یک مدل طبقه‌بندی با استفاده از فرمول F-Measure قدیمی است، این روش باید ارتقا یابد یا با روش K-Measure جایگزین شود. همچنین پیشنهاد می گردد ترکیب یادگیری ماشین ، بلاک چین و هوش مصنوعی برای تشخیص کلاهبرداری در بیمه درمان در آینده به عنوان رویه نو برای تشخیص دقیق تر و برخط استفاده گردد. ترکیب این سه فناوری می‌تواند به مؤسسات بیمه کمک کند تا به طور مؤثرتری کلاهبرداری‌ها را شناسایی و پیشگیری کنند. استفاده از یادگیری ماشین برای تجزیه و تحلیل داده، بلاک‌چین برای ذخیره‌سازی امن و هوش مصنوعی برای تسریع روندهای تشخیصی، می‌تواند به کاهش هزینه‌ها و حفاظت از منافع مشتریان منجر شود.

۷.منابع فارسی و انگلیسی

ابراهیمی، علی اکبر و ظهرابی واحد، محمد،۱۳۹۸،راه های کشف تقلب در بیمه(بخش دیه) با استفاده از الگوریتم k_mens،سومین کنفرانس بین المللی مهندسی برق،مهندسی مکانیک، کامپیوتر و علوم مهندسی.
اعتمادی، ح.، و زلقی، ح. (۱۳۹۲). کاربرد رگرسیون لجستیک در شناسایی گزارشگری مالی متقلبانه. دانش حسابرسی, ۳(۵۱), .
بهرامی نسب، بهاره و محقق، عارفه،۱۴۰۱،زنجیره بلوکی، پاسخگویی و کشف تقلب در شرکت های بیمه،چهارمین همایش ملی تحقیقات میان رشته ای در مدیریت و علوم انسانی،تهران
جهانی فر فاطمه ، راد عباس و بهرامی سامانی ، احسان.;۱۴۰۰. پیش بینی تقلب در صنایع خدماتی با استفاده از الگوریتم های داده کاوی(مورد مطالعه: شرکت بیمه کوثر) پایان نامه کارشناسی ارشد ، دانشگاه شهید بهشتی.
حسین زاده، سحر و رسولی، آمنه،۱۴۰۱،تشخیص تقلب بیمه مراقبتهای بهداشتی با استفاده از بلاک چین،هفدهمین کنفرانس علوم و مهندسی کامپیوتر و فناوری اطلاعات،بابل.
حسینی، علی و رضایی، عباسعلی،۱۳۹۷،کشف تقلب و راهکارهای مقابله با آن در سازمانهای بیمه ای با استفاده از داده کاوی (مطالعه موردی : سازمان تامین اجتماعی)،سومین کنفرانس ملی فناوری در مهندسی برق و کامپیوتر،سمنان.
ربطی، میناسادات،۱۳۹۶،بهبود دقت رده بندهای تشخیص تقلب روی مجموعه داده نامتوازن بیمه با استفاده از نمونه برداری ترکیبی،اولین کنفرانس ملی کاربردسازی هوش تجاری (راهکارها و چالش ها)،تهران.
روح الامین، مهدی و پایگذار، حمید و خلیلی درمنی، محمد،۱۳۹۸،تشخیص تقلب در صنعت بیمه با ترکیب درخت تصمیم و الگوریتم ژنتیک،سومین کنفرانس ملی مهندسی برق و کامپیوتر،تهران.
عبدالرزاق نژاد مجید، خرد مهدی. استفاده از خوشه بندی BIRCH و الگوریتم بهینه سازی واکنش شیمیایی جهت کشف تقلب در حوزه سلامت.
عسکرزاده، حسن و تارخ، محمدجعفر،۱۳۹۶،کشف تقلب در بیمه سلامت بر اساس رویکرد داده کاوی،کنفرانس بین المللی پژوهش های نوین در مدیریت ، اقتصاد ، توانمندی صنعت جهانگردی در توسعه،مشهد.
علی پور، آسیه و قائمی، رضا،۱۳۹۵،مروری بر روش های کشف تقلب در بیمه های درمانی با استفاده از تکنیک های داده کاوی،دومین همایش ملی کاربرد سیستم های هوشمند (محاسبات نرم) در علوم و صنایع،قوچان.

فیروزی، مهدی و شکوری، مرتضی و کاظمی، لیلا و زاهدی، سحر،۱۳۹۰،شناسایی تقلب در بیمه اتومبیل با استفاده از روش های داده کاوی، فصلنامه پژوهشنامه بیمه، دوره: ۲۶، شماره: ۳
فیق دوست، احمد و فریدی ماسوله، مرضیه و باقری، احمد،۱۳۹۹،تشخیص تقلب بیمه با استفاده از الگوریتم خوشه بندی و شبکه عصبی،اولین کنفرانس ملی بهینه سازی سیستمهای تولیدی و خدماتی،رودسر
ملکی کاکلر، ح.، و بحری ثالث، ج.، و جبارزاده کنگرلویی، س.، و آشتاب، ع. (۱۴۰۰). کارایی مدل های آماری و الگوهای یادگیری ماشین در پیش بینی گزارشگری مالی متقلبانه. اقتصاد مالی (اقتصاد مالی و توسعه), ۱۵(۵۴), ۲۶۷-۲۹۲.
یوسفی کیا، فایزه و باقری، علیرضا،۱۳۹۷،طراحی نرم افزار کشف تقلب در بیمه اتومبیل با استفاده از روش بیز ساده،چهارمین کنفرانس ملی دستاوردهای نوین در برق و کامپیوتر و صنایع، اسفراین.
Abadi, M., Erlingsson, U., Goodfellow, I., McMahan, H. B., Mironov, I., Papernot, N., ... & Zhang, L. (2017, August). On the protection of private information in machine learning systems: Two recent approches. In 2017 IEEE 30th Computer Security Foundations Symposium (CSF) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/8049647

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90-113.

Ahmadinejad, H., Norouzi, A., Ahmadi, A., & Yousefi, A. (2016). Distance based model to detect healthcare insurance fraud within unsupervised database. Indian Journal of Science and Technology, 9(43).

Bodaghi, A., & Teimourpour, B. (2018). Automobile insurance fraud detection using social network analysis. Applications of Data Management and Analysis: Case Studies in Social Networks and Beyond, 11-16. https://link.springer.com/chapter/10.1007/978-3-319-95810-1_2

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. IEEE Access, 8, 58546-58558.

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

Gupta, R. Y., Mudigonda, S. S., & Baruah, P. K. (2021). A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage schemes. International Journal of Engineering Trends and Technology, 69(3), 96-102.

Gupta, R. Y., Mudigonda, S. S., Baruah, P. K., & Kandala, P. K. (2021). Markov model with machine learning integration for fraud detection in health insurance. arXiv preprint arXiv:2102.10978.

Hancock, J. T., & Khoshgoftaar, T. M. (2021). Gradient boosted decision tree algorithms for medicare fraud detection. SN Computer Science, 2(4), 1-12.

Indarto, S. L., & Ghozali, I. (2016). Fraud diamond: Detection analysis on the fraudulent financial reporting. Risk Governance & Control: Financial Markets & Institutions, 6(4), 116-123.

Itri, B., Mohamed, Y., Mohammed, Q., & Omar, B. (2019, October). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS) (pp. 1-4). IEEE.

Johnson, J. M., & Khoshgoftaar, T. M. (2021). Medical provider embeddings for healthcare fraud detection. SN Computer Science, 2(4), 1-15.

Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Applied Soft Computing, 36, 283-299.

Li, Y., Yan, C., Liu, W., & Li, M. (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Applied Soft Computing, 70, 1000-1009.

Liu, Q., & Vasarhelyi, M. (2013, November). Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. In 29th world continuous auditing and reporting symposium (29WCARS), Brisbane, Australia.

Majhi, S. K. (2021). Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection. Evolutionary intelligence, 14(1), 35-46.

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.

Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., & Reynkens, T. (2022). Social network analytics for supervised fraud detection in insurance. Risk Analysis, 42(8), 1872-1890.

Pandey, P., Saroliya, A., & Kumar, R. (2018). Analyses and detection of health insurance fraud using data mining and predictive modeling techniques. In Soft Computing: Theories and Applications (pp. 41-49). Springer, Singapore.

Rawte, V., & Anuradha, G. (2015, January). Fraud detection in health insurance using data mining techniques. In 2015 International Conference on Communication, Information & Computing Technology (ICCICT) (pp. 1-5). IEEE.

Rawte, V., & Anuradha, G. (2015, January). Fraud detection in health insurance using data mining techniques. In 2015 International Conference on Communication, Information & Computing Technology (ICCICT) (pp. 1-5). IEEE.

Saldamli, G., Reddy, V., Bojja, K. S., Gururaja, M. K., Doddaveerappa, Y., & Tawalbeh, L. (2020, April). Health care insurance fraud detection using blockchain. In 2020 Seventh international conference on software defined systems (SDS) (pp. 145-152). IEEE.

Shirgave, S., Awati, C., More, R., & Patil, S. (2019). A Review On Credit Card Fraud Detection Using Machine Learning. Int. J. Sci. Technol. Res, 8, 1217-1220.

Subudhi, S., & Panigrahi, S. (2020). Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. Journal of King Saud University-Computer and Information Sciences, 32(5), 568-575.

Agarwal, S. (2023). An Intelligent Machine Learning Approach for Fraud Detection in Medical Claim Insurance: A Comprehensive Study. Scholars Journal of Engineering and Technology, 11(9), 191-200.

Zhang, C., Xiao, X., & Wu, C. (2020). Medical fraud and abuse detection system based on machine learning. International journal of environmental research and public health, 17(19), 7265.Cherkaoui, O., Anoun, H., & Maizate, A. (2024). A benchmark of health insurance fraud detection using machine learning techniques. IAES International Journal of Artificial Intelligence (IJ-AI)

A Machine Learning Model for Detecting Fraud in Health Insurance

Pouyan Soltaniyeh[1], Kiumars Pirzadeh Karsafi[2]

[1] Industrial Management, Management, Raja University, Qazvin, Iran, Pooian.soltany@gmail.com
[2] Information Technology, Information Technology, Ghiasuddin Jamshid Kashani University, Qazvin, Iran

Abstract

Fraud has become one of the major challenges in the insurance industry, where big data and machine learning are actively employed to address this issue. This study investigates health insurance fraud using machine learning algorithms and data mining techniques. The proposed innovative approach enhances fraud detection accuracy and contributes to reducing healthcare costs. We adopted a supervised learning method applied to automobile claim data from an anonymous insurance company. Our objective is to present an approach that improves the interpretability of artificial intelligence outcomes. The findings demonstrate that the Random Forest algorithm delivers the best performance based on the K-measure criterion and achieves the lowest RMSE score. Moreover, Random Forest provides the most balanced trade-off between recall and precision, with a detection rate of 23.8% of fraudulent cases among the total identified frauds.

Keywords

Machine Learning; Insurance Fraud Detection; Supervised Learning; Random Forest