فصلنامه علمی پژوهشی
دانش حسابداری و حسابرسی مدیریت
دوره ۱۷/ شماره ۱ (پیاپی ۶۵)/ بهار ۱۴۰۷
صفحه ۰۱ تا ۱۵

کشف تقلب در صورت‌های مالی شرکت‌های پذیرفته شده در بورس تهران با استفاده از داده‌کاوی

احمد خاکساران
دانشجوی دکترا، گروه حسابداری، واحد سنندج، دانشگاه آزاد اسلامی، سنندج، ایران
ahmadkhaksaran33@gmail.com

امید محمودی خوشرو
استادیار، گروه حسابداری، واحد سنندج، دانشگاه آزاد اسلامی، سنندج، ایران (نویسنده مسئول)
omid.khoshroo@gmail.com

سید کریم قادرزاده
استادیار، گروه حسابداری، دانشگاه پیام نور، تهران ، ایران.
K.ghaderzadeh@pnu.ac.ir

پیمان امینی
استادیار، گروه حسابداری، دانشکده علوم انسانی و اجتماعی، دانشگاه کردستان
p.amini@uok.ac.ir

تاریخ دریافت: ۱۴۰۳/۰۸/۲۶ تاریخ پذیرش: ۱۴۰۳/۰۹/۱۱

چکیده

تقلب در صورت‌های مالی نه تنها باعث زیان‌های هنگفت برای سرمایه‌گذاران می‌شود، بلکه باعث ایجاد بحران بی‌اعتمادی به حسابداری و صورت های مالی شرکت‌ها می‌شود. علاوه بر این، وضعیت مالی شرکت را در یک چرخه معیوب قرار می دهد که توسعه پایدار بلندمدت کل اقتصاد جامعه را کاهش می دهد. بنابراین، شناسایی صورت‌های مالی تقلبی نقش مهمی در افزایش پایداری اقتصاد اجتماعی ما ایفا می‌کند. در این پژوهش هدف کشف تقلب در صورت‌های مالی شرکت‌های پذیرفته شده در بورس تهران با استفاده از داده‌کاوی است. نمونه پژوهش ۲۰۰شرکت بورسی در بازه ۱۱ ساله ۱۳۹۲-۱۴۰۲ است. معیارهای ارزیابی برای چهار مدل درخت تصمیم، شبکه عصبی، نزدیک‌ترین همسایه و قاعده بیزی بدست آمد که نتایج نشان می دهد مقادیر امتیاز F قاعده بیز بسیار پایین است و تنها عملکرد این مدل مورد تایید نیست، در حالی که، تمام معیارهای ارزیابی سه مدل دیگر مؤثر هستند یعنی درخت تصمیم(DT) ، نزدیک‌ترین همسایه(KNN) و شبکه عصبی (MLP) پیش‌بینی خوبی در تشخیص تقلب شرکت‌های بورسی دارد. تکنیک داده‌کاوی موجود می‌تواند تا حد زیادی یک سیستم پشتیبانی تصمیم برای مدیران در تشخیص تقلب در صورت‌های مالی فراهم کند.

واژه‌های کلیدی: تقلب در صورت‌های مالی، بورس اوراق بهادار تهران، داده‌کاوی.

۱- مقدمه

تقلب در صورت‌های مالی [1]یک نگرانی مهم در دنیای امروز است که مخرب ترین شکل مالی کلاه‌برداری شغلی را نشان می‌دهد. این امر بر طیف گسترده‌ای از ذینفعان، از جمله خود سازمان، سرمایه گذاران، اعتباردهندگان، تأمین‌کنندگان، کارکنان و عموم مردم تأثیر منفی می‌گذارد. با توجه به اهمیت حیاتی آن، این موضوع در دهه‌های گذشته کانون تحقیقات علمی بوده است(کرکاس و همکاران[2]،۲۰۲۴). در سطح جهانی، هزینه تقلب صورت های مالی در طول سال‌ها افزایش یافته است. مستقیم‌ترین قربانیان سرمایه گذاران و سرمایه‌دارانی هستند که به بهانه‌های واهی به شرکت‌ها وجه پرداخت می‌کنند. اما هزینه‌های بیشتری برای جامعه در سطح کلان وجود دارد، مانند از دست دادن اعتماد به سیستم‌های مالی، حتی اگر تأثیر اقتصادی در سطح شرکت بسیار متغیر باشد. از آنجایی که تقلب صورت های مالی به سرمایه گذاران، ذینفعان و جامعه آسیب زیادی وارد می‌کند، مطالعات زیادی انجام‌شده است اندرو و رابین[3](۲۰۲۲). به گفته انجمن بازرسان خبره تقلب، تقلب در صورت‌های مالی کمترین میزان تقلب را نشان می‌دهد اما بیشترین ضرر مالی را به همراه دارد (دو توا[4]،۲۰۲۳). شناسایی تقلب در صورت‌های مالی در چند دهه گذشته با تأکید بر ناهنجاری‌های حسابداری به‌طور کلی و به‌طور خاص بر تقلب در صورت‌های مالی مورد توجه قرار گرفته است. در حالی که تحقیقات اولیه از تکنیک‌های آماری یا سنتی استفاده می‌کرد که هم زمان بر و هم پرهزینه هستند، اخیراً با ظهور داده‌های بزرگ و ML تمرکز بیشتر شده است(محمدی و همکاران[5]، ۲۰۲۰). ثابت شده است که روش‌های حسابرسی دارای کاستی‌های زیادی برای شناسایی گزارشگری مالی متقلبانه هستند. در یک سازمان، مدیر کسی است که از نظر اخلاقی در قبال کشف داده‌های مالی متقلبانه مسئول و پاسخگو است. اما درواقع اکثر تقلب‌های صورت‌های مالی با آگاهی یا رضایت مدیریت انجام می‌شود. هرگونه شکست در تشخیص صورت‌های مالی تحریف شده می‌تواند به شدت به اعتبار حرفه حسابرسی آسیب برساند. در حال حاضر، شیوه‌های حسابرسی باید به موقع انجام شود تا با افزایش تعداد و وقوع موارد تقلب صورت‌های مالی مقابله شود. تکنیک‌های جدید مانند داده‌کاوی ادعا می‌کنند که دارای قابلیت‌های طبقه‌بندی و پیش‌بینی پیشرفته‌ای است و می‌تواند برای تسهیل نقش حسابرسان ازنظر انجام موفقیت‌آمیز وظیفه کشف تقلب مورد استفاده قرار گیرد (لین و همکاران[6]، ۲۰۱۵). نسبت‌های مالی را می‌توان برای بررسی ویژگی‌های کلیدی تقلب‌های مالی بر اساس ارزش‌های برگرفته از صورت‌های مالی استخراج کرد و می‌تواند با کمی کردن بسیاری از جنبه‌های کسب‌وکار، کشف مواد تقلبی ارائه شده در صورت‌های مالی را تجمیع کند. به عنوان بخشی جدایی ناپذیر از تجزیه‌وتحلیل صورت‌های مالی، این نسبت‌ها می‌توانند وضعیت مطالبات و بدهی‌های بد را نشان دهند، اینکه آیا کسب‌وکار دارای بدهی یا موجودی مازاد است، آیا هزینه‌های عملیاتی بالا است و آیا دارایی‌های شرکت به درستی استفاده می‌شود. ایجاد درآمد، نسبت‌های نقدینگی، ایمنی، سودآوری و کارایی شاخص‌های مهم نسبت‌های مالی هستند. با توجه به پیشرفت تکنیک‌های داده‌کاوی، رویکردهای پیچیده‌ای برای کشف دانش برای استخراج اطلاعات ناشناخته قبلی از داده‌ها مورد نیاز است.

در این زمینه، مطالعه حاضر از اهمیت عملی برای مشاغل و حسابرسان برخوردار است، زیرا بازار جهانی شاهد افزایش تقلب در حسابداری مالی است که میلیاردها دلار در سال برای کسب‌وکارها هزینه دارد.

آمار دقیقی از میزان تقلب در شرکت‌های بورس تهران در دست نیست اما بر اساس پژوهش علیخانی و همکاران(۱۳۹۹) ٪۲۱ شرکت‌های مورد بررسی متقلب و ۸۸ درصد غیرمتقلب بودند، در پژوهش رضایی و همکاران(۱۳۹۹) این آمار برای شرکت‌های متقلب ۷۸ درصد و غیرمتلقب ۴۲ درصد بود در نهایت نیز در پژوهش ممنون و همکاران از ۹۶۰ شرکت مورد بررسی ۶۳.۸ درصد شرکت‌ها غیر متقلب و ۳۶.۱۵ درصد متقلب بودند. همان طور که این پژوهش‌ها نشان می‌دهد نسبت تقلب در صورت‌های مالی پژوهش های مختلف، فرق دارد.

اکثر آثار موجود در ادبیات به‌عنوان مثال، اندرو و رابین[7](۲۰۲۲)؛ ماسدو و ویرا[8](۲۰۲۲)؛ وادوا و کومار[9](۲۰۲۰)؛ السینگاوی[10](۲۰۲۱) از تحلیل رگرسیون سنتی استفاده کرده‌اند. در سال‌های اخیر، تعدادی از کارشناسان و محققین سعی کرده‌اند از روش‌های یادگیری ماشین (ML) و داده‌کاوی برای انجام تحقیقات در این زمینه به عنوان راهی برای کاهش خطاهای تشخیص استفاده کنند. در حالی که بسیاری از استراتژی‌های تجاری مبتنی بر دقت صورت‌های مالی هستند، منابع کافی برای تجزیه‌وتحلیل همه آن‌ها به‌طور کامل وجود ندارد. از آنجایی که حسابرسان در چندین نمونه مقصر FSF

[1] FSF
[2] Kirkos, et al.
[3] Andrew, C.; Robin
[4] du Toit, E
[5] Mohammadi, M
[6] C. Lin
[7] Andrew, C.; Robin
[8] Macedo and Vieira
[9] Wadhwa, A.V.K.; Kumar, S
[10] Alsinglawi and Almari

شناخته‌شده‌اند، مدل‌های کشف تقلب مالی معتبر باید به روشی آسان برای حسابرسان، سرمایه‌گذاران، تنظیم‌کننده‌ها و سایر ذینفعان ارائه شوند. به دلیل اتکای ذاتی به مفروضات توزیعی محدود، تکنیک‌های پارامتریک مانند LR فاقد کاربرد کلی است که رویکردهای ناپارامتریک ممکن است ارائه کنند(شرایبر و همکاران[1]، ۲۰۱۸). از سوی دیگر، تحقیقات موجود در مورد روش‌های کمی کشف تقلب مالی، عمدتاً بر صنعت بانکداری و خدمات مالی مبتنی بر کشف تقلب بیمه و کارت اعتباری متمرکزند. ادبیات علمی و آکادمیک فعلی در حال حاضر به دنبال قوانین یا طبقه‌بندی‌های بیشتر از داده‌های قبلی برای دستیابی به هدف پیش‌بینی یا تشخیص است. یادگیری ماشینی (ML) با مقدار مناسب داده می‌تواند نتایج دقیق‌تری را در مقایسه با رویکرد سنتی پیش‌بینی و طبقه‌بندی کند.

نقش داده‌کاوی در کشف تقلب

داده‌کاوی فرآیندی است که در آن از تکنیک‌های مختلف برای استخراج الگوهای داده معتبر استفاده می‌شود. این مرحله شامل انتخاب مناسب‌ترین روش داده‌کاوی (مانند طبقه‌بندی، رگرسیون، خوشه‌بندی یا همبستگی) و انتخاب الگوریتم متعلق به یکی از خانواده‌های قبلی است. در نهایت، الگوریتم انتخاب‌شده برای حل مسئله با تنظیم پارامترهای اصلی و روش‌های اعتبار سنجی استفاده می‌شود. به‌طور کلی روش‌های داده‌کاوی به دو نوع روش‌های پیش‌بینی و توصیفی (مثلاً خوشه‌بندی) تقسیم می‌شوند و روش اول به دو بخش روش‌های آماری و روش‌های نمادین تقسیم می‌شوند. روش‌های آماری با نمایش دانش از طریق مدل‌های ریاضی با محاسبات آن‌ها شناخته می‌شوند. روش‌های نمادین دانش را با استفاده از نمادها و تداعی‌ها نمایش می‌دهند که در نهایت مدل‌های قابل تفسیر بیشتری تولید کند. مدل‌های رگرسیون، شبکه‌های عصبی مصنوعی، پشتیبانی ماشین‌های برداری، مدل‌های بیزی و ... در دسته روش‌های آماری قرار می‌گیرند. مدل‌های رگرسیون از جمله مدل‌های کلاسیک هستند که نیاز به یک کلاس از مدل‌سازی معادلات دارند. رگرسیون‌های خطی، درجه دوم و لجستیک از شناخته‌شده‌ترین مدل‌ها در داده‌کاوی هستند. آنها الزامات اساسی را بر داده‌ها تحمیل می‌کنند و سعی می‌کنند از همه ویژگی‌های آن، چه مفید و چه غیرمفید استفاده کنند. یکی از قوی‌ترین مدل‌های ریاضی، شبکه‌های عصبی مصنوعی (ANN) است که تقریباً برای تمام وظایف داده‌کاوی مناسب است. فرمول‌بندی‌های مختلفی از شبکه‌های عصبی وجود دارد. رایج‌ترین آنها عبارت‌اند از یادگیری کوانتیزاسیون برداری، و شبکه‌های عصبی پیشخور چند لایه، شبکه‌های تابع پایه شعاعی، و ادراک چند لایه. شبکه‌های عصبی بر اساس تعریف نورون‌ها هستند، که بخش‌های اتمی هستند که ورودی‌های خود را برای خروجی با توجه به یک تابع فعال‌سازی جمع‌آوری می‌کنند. این مدل‌ها در صورتی که شبکه‌ها به‌درستی پیکربندی شوند، عموماً بهتر از مدل‌های دیگر هستند، اما به دلیل ساختار پیچیده‌شان چندان محبوب نیستند. ماشین‌های بردار پشتیبان مبتنی بر تئوری یادگیری هستند و زمانی که داده‌ها به‌صورت خطی قابل‌تفکیک هستند، بسیار خوب کار می‌کنند. برخلاف مدل‌های رگرسیون، این مدل‌ها معمولاً نیازی به ایجاد برهمکنش بین متغیرها ندارند و مانند شبکه‌های عصبی مصنوعی، نسبت به نویز و نقاط پرت مقاوم هستند. روش‌های پیش‌بینی عمدتاً به یادگیری نظارت‌شده نسبت داده می‌شود. روش‌های نظارت‌شده روابط بین صفات ورودی و یک ویژگی هدف را در ساختاری که ما آن را "مدل" می‌نامیم، آشکار می‌کند(امیدی، ۲۰۱۹).

مشکلات رگرسیون و طبقه‌بندی هر دو در این دسته هستند. یک سناریوی نظارت‌شده به‌گونه‌ای کار می‌کند که یک مدل ابتدا به مجموعه داده‌های آموزشی برازش می‌شود و سپس برای پیش‌بینی موارد مشاهده نشده استفاده می‌شود. در اینجا، هدف اصلی نگاشت ورودی‌ها به خروجی است که مقادیر صحیح آن توسط سرپرست تعیین می‌شود. از سوی دیگر، چنین ناظری در یادگیری بدون نظارت وجود ندارد و فقط داده‌های ورودی در دسترس است.

بنابراین، هدف یافتن مقررات، بی‌نظمی‌ها، شباهت‌ها، روابط و همبستگی‌ها در ورودی‌ها است. یکی از مزایای یادگیری بدون نظارت این است که برخلاف یادگیری تحت نظارت، می‌توان مدل‌های پیچیده‌تری را یاد گرفت. از آنجایی که در یادگیری نظارت‌شده، هدف یافتن رابطه بین دو مجموعه مشاهدات است، با افزایش تعداد مراحل، دشواری فرآیند یادگیری به دلیل هزینه‌های محاسباتی بالا به‌طور تصاعدی افزایش می‌یابد و بنابراین نمی‌توان این مدل‌ها را عمیقاً آموخت. دو مشکل معروف که به یادگیری بدون نظارت تعلق دارند، خوشه‌بندی و قوانین انجمن هستند. داده‌کاوی به‌طور گسترده در بسیاری از حوزه‌ها کاربرد دارد. در زیر خلاصه‌ای از کاربردهای داده‌کاوی آورده شده است:

(i) کاربردهای داده‌کاوی در امور مالی: پول‌شویی و کشف سایر جرائم مالی، طبقه‌بندی مشتریان و بازاریابی هدف، پرداخت وام، پیش‌بینی و تحلیل اعتبار مشتری

[1] Schreiber et al.

(ii) کاربرد داده‌کاوی در صنعت خرده‌فروشی: تجزیه و تحلیل، اثربخشی کمپین فروش، حفظ مشتری و تحلیل وفاداری مشتری، توصیه محصول و ارجاع متقابل اقلام.

(iii) داده‌کاوی برای صنعت مخابرات: تجزیه‌وتحلیل الگوی تقلبی و شناسایی الگوهای غیرمعمول و خدمات مخابراتی سیار. مطالعات متعددی به دنبال کشف تقلب با استفاده از رویکردهای نوآورانه بوده است.

بسیاری از تحقیقات در ادبیات قبلی تشخیص FSF را به عنوان یک مسئله طبقه‌بندی باینری، گاهی به عنوان یک مسئله چند کلاسه و گاهی اوقات به عنوان یک مشکل خوشه‌بندی فرموله کرده‌اند. محققان هر دو تجزیه‌وتحلیل کمی و کیفی FSF را انجام داده‌اند. متن‌کاوی به‌طور گسترده برای تحقیقات کیفی استفاده‌شده است. تکنیک‌های یادگیری ماشینی(ML) در مراحل اولیه، تحقیقات عمدتاً شامل رویکردهای شبکه عصبی (NN)، LR، DT، SVM، تجزیه‌وتحلیل متمایز (DA)، و شبکه بیزی است(الفیض و فاتی[1]، ۲۰۲۲؛ استرلسینا و همکاران[2]، ۲۰۲۳، کومار و همکاران[3]، ۲۰۲۲).

تقلب در صورت‌های مالی (FSF) یک نگرانی جهانی است. FSF به عنوان حذفیات عمده یا ارائه نادرست در صورت‌های مالی که به دلیل عدم گزارش عمدی داده‌های مالی مطابق با استانداردهای حسابداری پذیرفته شده است، مشخص می‌شود. FSF می‌تواند تأثیرات قابل‌توجهی بر ذینفعان شرکت‌های متقلب و شرکت‌های غیر متقلب در صورت عدم شناسایی و پیشگیری به‌موقع داشته باشد (خدر و همکاران[4]، ۲۰۲۱). متأسفانه تشخیص FSF آسان نیست. علاوه بر این، حتی در صورت کشف، آسیب قابل‌توجهی به‌طور کلی قبلاً رخ‌داده است. درنتیجه، تنظیم‌کننده‌ها، حسابرسان و سرمایه‌گذاران از تکنیک‌های کارآمدتر و مؤثرتری که می‌توانند FSF را شناسایی کنند، سود زیادی خواهند برد. انجمن بازرسان خبره تقلب (ACFE) بیان می‌کند که تقلب در صورت‌های مالی عبارت است از ارائه نادرست عمدی وضعیت مالی یک شرکت با تحریف یا حذف عمدی مبالغ یا افشا در صورت‌های مالی برای گمراه کردن استفاده‌کنندگان از صورت‌های مالی. به گفته مرکز کیفیت حسابرسی (CAQ)، افراد یا شرکت‌ها به دلایل مختلفی از جمله مزایای پولی، نیاز به تحقق اهداف مالی کوتاه‌مدت یا سرپوش گذاشتن بر اخبار ناگوار درگیر دست‌کاری صورت‌های مالی هستند. مصرف‌کنندگان خارجی و داخلی صورت‌های مالی دائماً صورت‌های مالی را زیر سؤال می‌برند و نهادهای نظارتی نمی‌توانند با اطمینان بگویند که صورت‌های مالی معتبر است و مطابق با دستورات قانونی و اخلاقی رویه‌های حسابداران و حسابرسان تهیه‌شده است(کیلکوا و استاداروا[5]، ۲۰۱۶؛ دیبک و همکاران[6]، ۲۰۲۲). درنتیجه، کشف تقلب یا فریب به منظور اطمینان از صحت صورت‌های مالی مهم است.

پیشینه پژوهش

رحمانیان و همکاران (۱۴۰۲)تأثیر کاهش مکانیزم نظارتی بر رابطه میان حق‌الزحمه غیرعادی حسابرسان و تقلب در صورت‌های مالی را موردمطالعه قراردادند. هدف مطالعه بررسی تأثیر کاهش مکانیزم نظارتی بر رابطه میان حق‌الزحمه غیرعادی حسابرسان و تقلب در صورت‌های مالی است. پژوهش کاربردی و از بعد روش‌شناسی، همبستگی از نوع علّی (پس رویدادی) می‌باشد. جامعه آماری کلیه شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران بوده و با استفاده از روش نمونه‌گیری حذف سیستماتیک، ۹۰ شرکت به‌عنوان نمونه پژوهش انتخاب‌شده و در دوره زمانی ۱۰ ساله بین سال‌های ۱۳۹۱ الی ۱۴۰۰ موردبررسی قرار گرفتند. نهایتاً با استفاده از رگرسیون لجستیک و نرم‌افزار ایویوز آزمون فرضیه‌های نهایی پژوهش انجام‌شده است. نتایج حاصل از آزمون فرضیه‌های پژوهش نشان می دهد که میان حق‌الزحمه غیرعادی حسابرس و تقلب در صورت‌های مالی رابطه مستقیم وجود دارد و همچنین کاهش مکانیزم نظارتی(ضعف کنترل داخلی) این رابطه را تشدید می‌کند.نتایج نشان می دهد که میان حق‌الزحمه غیرعادی حسابرس و تقلب در صورت‌های مالی رابطه مستقیم وجود دارد و همچنین کاهش مکانیزم نظارتی(ضعف کنترل داخلی) این رابطه را تشدید می‌کند.

کرکاس و همکاران[7](۲۰۲۴) مرزهای تشخیص تقلب در صورت‌های مالی با مدل‌های زبان بزرگ را بررسی کردند. مطالعات پیشین از مجموعه متنوعی از تکنیک‌های آماری، یادگیری ماشینی و یادگیری عمیق برای تجزیه‌وتحلیل داده‌های حسابی و متنی استفاده کرده‌اند. بااین‌حال، استفاده از مدل‌های زبان بزرگ (LLM) برای شناسایی تقلب در صورت‌های مالی تا حد زیادی ناشناخته باقی‌مانده است. کار علمی اخیر بر پتانسیل LLM ها در تحلیل‌های مختلف زبانی و وظایف حسابرسی تأکید می‌کند. در مطالعه، توانایی یک LLM، به‌ویژه ChatGPT-4 را

[1] Alfaiz, N.S.; Fati
[2] Strelcenia, E
[3] Kumar, S
[4] El-Bannany
[5] Kulikova, L.; Satdarova, D.
[6] Deebak, B et al.
[7] Kirkos, et al.

برای شناسایی صورت‌های مالی تقلبی با تجزیه‌وتحلیل بخش‌های متنی استخراج‌شده از گزارش‌های سالانه بررسی می‌کند. به‌ویژه برنامه‌های مدیرعامل به سهامداران و بخش‌هایی که در مورد ریسک‌های شرکت بحث می‌کنند تمرکز می‌کند. به‌جای استفاده از تنظیم دقیق پیچیده، از تکنیک‌های مهندسی سریع استفاده می‌کند که می‌توانند به‌راحتی توسط حسابرسان بدون تخصص یادگیری ماشین پیاده‌سازی شوند. یافته‌ها امیدوارکننده هستند و پتانسیل LLM را برای کشف تقلب صورت‌های مالی نشان می‌دهند. با گنجاندن بازخورد انسانی در نتایج طبقه‌بندی مدل، به امتیاز عملکرد ۶۷ درصد در مقیاس‌های حساسیت، ویژگی و F-Measure دست یافتند. با رشد مداوم در اندازه و قابلیت‌های مدل، همراه با تحقیقات فعال در مهندسی سریع، به‌زودی پیشرفت‌های بیشتری را در عملکرد پیش‌بینی می‌کنیم. تحقیق می‌تواند برای حسابرسان، وام‌دهندگان، سرمایه گذاران و نهادهای نظارتی ارزشمند باشد.

دو توا[1](۲۰۲۳) پرچم‌های قرمز تقلب در صورت‌های مالی را در یک مطالعه موردی مورد بررسی قرارداد. نویسنده پیش‌ازاین ویژگی‌های تقلب در صورت‌های مالی را بررسی کرده و وجود ۱۶ شاخص تقلب را مشخص کرده است. هدف از مطالعه تعیین اینکه آیا سرمایه گذاران و سایر ذینفعان می‌توانند با استفاده از این ویژگی‌ها تقلب در صورت‌های مالی را در تجزیه‌وتحلیل گزارش سالانه یک شرکت شناسایی کنند. مطالعه یک مورد تقلب در صورت‌های مالی را با استفاده از تکنیک‌های مشابهی که قبلاً اعمال می‌شد، از جمله تحلیل افقی، عمودی و نسبتی تحلیل می‌کند. یافته‌ها چندین ویژگی تقلب و چند مورد دیگر که قبلاً رایج نبوده‌اند را نشان می‌دهند. بنابراین تقلب در صورت‌های مالی بین موارد متفاوت است. همچنین شناسایی شاخص‌های تقلب پس از عمل آسان‌تر است.

لی و همکاران[2](۲۰۲۳) تجزیه‌وتحلیل متنی و کشف تقلب مالی: شواهد از شرکت‌های تولیدی چینی را ارائه کردند. با افزایش پیچیدگی دست‌کاری صورت‌های مالی، تنها با تکیه بر داده‌های مالی کمی ممکن است تقلب مالی به‌طور مؤثر شناسایی نشود. در حالی که تجزیه‌وتحلیل متن می‌تواند بینش بیشتری ارائه دهد، تحقیقات کمی در مورد ابعاد چندگانه آن انجام‌شده است. با استفاده از ۵۷۹ شرکت تولیدی چینی در سال ۲۰۲۰، خوانایی، آینده‌نگر، شباهت، درجه تطبیق و شاخص‌های احساسات مثبت و منفی را از ساختار زبان متنی، کیفیت و بیان متون بحث و تحلیل مدیریت، در ترکیب با شاخص‌های مالی انتخاب شدند. برای کشف تقلب مالی یافته‌های آنها نشان می‌دهد که شرکت‌های متقلب در گزارش‌های مالی خود بیش از حد محتاط هستند، احساسات مثبت کمتری را ابراز می‌کنند و با افزایش پیچیدگی گزارش‌های سالانه خود و استفاده از قیدهای درجه بیشتر برای اصلاح اطلاعات آینده‌نگر، کلاه‌برداری مالی را پنهان می‌کنند. مطالعه همچنین اهمیت در نظر گرفتن بیان زبان متنی را در کشف تقلب مالی در شرکت‌های دولتی و غیردولتی نشان می‌دهد.

لوکانان و شرما[3](۲۰۲۳) پیش‌بینی تقلب با استفاده از یادگیری ماشین: مورد مشاوران سرمایه‌گذاری در کانادا را ارائه کردند.نتایج حاصل از چهار مدل یادگیری ماشین نشان می‌دهد که در تمام ویژگی‌ها، مقدار پول سرمایه‌گذاری شده و اینکه آیا مجرم از یک شرکت سرمایه‌گذاری متعلق به بانک بوده است، پیش‌بینی بالایی از کلاه‌برداری ازنظر ضریب استاندارد هستند. مدیران شعب و رگولاتورها باید به پرتفوی‌هایی که به عنوان نشانه‌ای از تقلب احتمالی به‌طور مداوم متحمل زیان می‌شوند، توجه دقیق داشته باشند. این یافته‌ها به‌ویژه به تنظیم‌کننده‌هایی مرتبط است که به دنبال تکنیک‌های جدید و مؤثر کشف تقلب هستند و درعین‌حال وضوح بیشتری را برای خودتنظیمی بازارهای مالی کانادا فراهم می‌کنند.

خدر و همکاران[4](۲۰۲۳) یک مدل پیش‌بینی قدرتمند برای تقلب در صورت‌های مالی بر اساس تکنیک یادگیری گروه XGBoostبهینه ارائه کردند. هدف مطالعه ایجاد مدل تشخیص تقلب در صورت‌های مالی (FSF) بهتر با استفاده از داده‌های صورت‌های مالی در دسترس عموم شرکت‌ها در منطقه مناست. آنها یک مدل FSF را با استفاده از یک تکنیک گروهی قدرتمند، الگوریتم XGBoost (تقویت گرادیان فوق‌العاده) توسعه دادند، که به شناسایی تقلب در مجموعه‌ای از شرکت‌های نمونه برگرفته از منطقه خاورمیانه و شمال آفریقا (MENA) کمک می‌کند. مسئله عدم تعادل کلاس در مجموعه داده با استفاده از الگوریتم روش نمونه‌برداری بیش از حد اقلیت مصنوعی (SMOTE) مورد بررسی قرار می‌گیرد. آنها از تکنیک‌های مختلف یادگیری ماشین در پایتون برای پیش‌بینی FSF استفاده کردند و یافته‌های تجربی آنها نشان داد که الگوریتم XGBoost از دیگر الگوریتم‌های مطالعه، یعنی رگرسیون لجستیک(LR) ، درخت تصمیم(DT) ، جنگل تصادفی(RF)ماشین بردار پشتیبانی (SVM) بهتر عمل کرده است. سپس الگوریتم XGBoost را برای به دست آوردن

[1] du Toit, E
[2] Li
[3] Lokanan, Kush Sharma
[4] Khedr

بهترین نتیجه با دقت نهایی ۹۶.۰۵ درصد در تشخیص FSF بهینه به کاربردند.

نسویژواسکایا و همکاران[1](۲۰۲۳) مبادله دقت در مقابل تفسیرپذیری در مدل تشخیص تقلب را ارائه کردند. مقاله ابتدا بر ترجمه سؤالات عملی مطرح‌شده در صنعت بانکداری در هر مرحله از فرآیند مدیریت تقلب به ارزیابی عملکرد مورد نیاز برای طراحی یک مدل تشخیص تقلب متمرکز است. ثانیاً، طیف وسیعی از رویکردهای یادگیری ماشین را در نظر می‌گیرد که به این ویژگی‌ها می‌پردازد: عدم تعادل بین عملیات متقلبانه و غیر متقلبانه، فقدان برچسب‌های کاملاً قابل‌اعتماد، پدیده رانش مفهومی، و مبادله اجتناب‌ناپذیر بین دقت و تفسیرپذیری تشخیص. این بررسی پیشرفته تا حدودی بر رقابت فناوری بین مدل‌های یادگیری ماشین جعبه سیاه بهبودیافته با تفسیر پس خطی و مدل‌های قابل تفسیر ذاتی که برای به دست آوردن دقت تقویت‌شده‌اند، روشن می‌کند. در نهایت، بر این موضوع بحث می‌کند که چگونه رویکردهای ترکیبی ملموس و امیدوارکننده می‌تواند پاسخ‌های عمل‌گرایانه و کوتاه‌مدت به بانک‌ها و سیاست‌گذاران ارائه دهد، بدون اینکه ذینفعانی را که دارای ریسک‌های اقتصادی و اخلاقی در این مسابقه فناوری هستند، ببلعد.

چن و همکاران[2](۲۰۲۳) تشخیص تقلب مالی شرکت‌های بورسی در چین: رویکرد یادگیری ماشینی را مورد بررسی قراردادند. به عنوان محور نظارت بر بازار سرمایه، تقلب در گزارش‌های مالی روند توسعه اعداد بسیار زیاد، معاملات پیچیده و ابزارهای پنهان را در سال‌های اخیر نشان داده است. برای بهبود کارایی حسابرسی و کاهش وابستگی به داده‌های غیرمالی، مطالعه تنها از داده‌های اصلی ساختاریافته در گزارش مالی برای ایجاد یک مدل شناسایی تقلب جدید استفاده می‌کند که می‌تواند به‌سرعت تقلب را در چین شناسایی کند. مطالعه اثربخشی تشخیص دو الگوریتم یادگیری ماشینی طبقه‌بندی تک و پنج الگوریتم یادگیری گروهی را در تشخیص تقلب مقایسه می‌کند. در مقایسه با الگوریتم‌های یادگیری ماشینی طبقه‌بندی تک، نتایج نشان می‌دهد که الگوریتم‌های یادگیری گروهی عموماً در تشخیص تقلب برای شرکت‌های بورسی چینی بهتر هستند و الگوریتم انباشتگی بهترین عملکرد را دارد. نتایج مطالعه شواهد مستقیمی برای کشف سریع تقلب با استفاده از داده‌های خام گزارش مالی و الگوریتم‌های یادگیری مجموعه ارائه می‌دهد. مطالعه ابتدا یک مدل شناسایی تقلب گزارشگری مالی مبتنی بر الگوریتم انباشته را برای شرکت‌های بورسی در چین پیشنهاد می‌کند که رویکردی ساده و مؤثر برای سرمایه‌گذاران، تنظیم‌کننده‌ها و مدیریت ارائه می‌دهد. همچنین می‌تواند مرجعی برای تشخیص سایر سناریوهای تقلب ارائه کند.

روش‌شناسی پژوهش

در مرحله کیفی این پژوهش شاخص‌های مؤثر بر تقلب در صورت‌های مالی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران با رویکرد فراترکیب به دست آمد و در پی آن ابعاد تقلب در صورت‌های مالی در چهار بُعد عملکرد مالی، کارایی مدیریت، راهبری شرکت و سلامت مالی شناسایی شدند. در این پژوهش، شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران طی سال‌های ۱۳۹۲ تا ۱۴۰۲ که صورت‌های مالی خود را به بورس اوراق بهادار تهران ارائه نموده‌اند، جامعه آماری را تشکیل می‌دهند. نمونه تحقیق نیز به روش حذفی از میان شرکت‌های یادشده انتخاب می‌شود. بر همین اساس شرکت‌های واسطه‌گری مالی،سرمایه‌گذاری، هلدینگ، لیزینگ و بانک‌ها به دلیل ماهیت عملیات آن‌ها، شرکت‌هایی که سال مالی آن‌ها منتهی به ۲۹ اسفندماه نیست به دلیل همگون کردن اطلاعات مورداستفاده و شرکت‌هایی که شرایط مورد نیاز برای محاسبه متغیرهای پژوهش را نداشته و دسترسی به اطلاعات و صورت‌های مالی آن‌ها میسر نباشد، حذف می‌شوند. در نهایت ۲۰۰ شرکت به‌عنوان نمونه نهایی انتخاب شدند و نمونه آماری تحقیق به دو گروه عمده تقسیم شدند:

گروه اول: شرکت‌های مشمول تقلب

در ایران سازمان و یا نهادی جهت شناسایی شرکت‌های مشمول تقلب وجود ندارد. هم‌چنین معیار واحد و استانداردی که مورد توافق همه باشد جهت سنجش تقلب در صورت‌های مالی وجود ندارد. ازاین‌رو به منظور تحلیل حساسیت و قابلیت اتکا بودن نتایج از معیار زیر جهت انتخاب شرکت‌های مشمول تقلب در صورت‌های مالی استفاده خواهد شد .

معیار اصلی

با توجه به استاندارد ۲۴۰ و ۴۵۰ استانداردهای حسابرسی ، شرکتی که دارای گزارش حسابرسی تعدیل‌شده باشد احتمال تقلب در آن نسبت به شرکت‌های دارای گزارش مقبول بیشتر است . سپس از بین شرکت‌های با گزارش حسابرسی تعدیل‌شده شرکتی که یکی از مصادیق زیر را داشته باشد به عنوان شرکت‌های مشمول تقلب شناسایی می‌گردد.(هالبونی، ۲۰۱۵؛ چن و همکاران،۲۰۱۳؛ گاکانیز،۲۰۰۹؛ آتا و سیریک،۲۰۰۹؛

[1] Nesvijevskaia

[2] Chen

پینگ سان و همکاران،۲۰۱۴؛ تشدیدی و همکاران، ۱۳۹۸؛ اعتمادی و زلقی، ۱۳۹۲؛ مرادی و همکاران، ۱۳۹۳؛ فرقاندوست حقیقی و همکاران، ۱۳۹۳)

۱) شناسایی نادرست درآمدها و اندازه‌گیری درآمدهای تحقق‌یافته،
۲) بیش نمایی در دارایی‌ها و موجودی‌های پایان دوره،
۳) شناسایی نادرست هزینه‌ها و عدم اندازه‌گیری هزینه‌های محقق شده،
۴) کم نمایی در بدهی‌ها و استفاده نادرست و متقلبانه از حساب‌های ذخیره،
۵) عدم تهیه صورت‌های مالی با فرض توقف فعالیت شرکتت‌هایی که فرض تداوم فعالیت آنها به شکل اساسی زیر سؤال رفته و در گزارش‌های حسابرسی بازتاب یافته است .
۶) وجود تعدیلات سنواتی بااهمیت(مطابق دستورالعمل نشریه ۱۵۰ سازمان حسابرسی) و صورت‌های مالی تجدید ارائه‌شده
۷) وجود اختلاف مالیاتی با حوزه مالیاتی مطابق یادداشت ذخیره مالیات بر درآمد، پرونده مالیاتی و بند شرط گزارش حسابرسی.

گروه دوم: شرکت‌های غیر متقلب

تمام شرکت‌هایی که مشمول تقلب نمی‌باشند به عنوان شرکت‌های غیر متقلب در نظر گرفته خواهد شد. لیکن به منظور رعایت تطابق ، در هرسال به تعداد شرکت‌ها مشمول تقلب در هر صنعت شرکت‌های غیر متقلب نیز انتخاب خواهد شد. به عبارتی شرکت‌های غیر متقلب در بعد مکان و زمان با شرکت‌های مشمول متقلب برابر می‌باشند . درواقع Fraud یک متغیر مجازی است با مقدار یک برای شرکت‌های مشمول تقلب و صفر برای سایر شرکت‌ها.

جدول ۱: خلاصه متغیرهای پیش بینی کننده

شاخص عملکرد مالی financial performance indices	یائو و همکاران[1](۲۰۱۹)	شاخص های سودآوری
	.تارجو و همکاران[2](۲۰۲۳)	کیفیت دارایی
	.کارشناسان و همکارن (۱۳۹۶)	هموارسازی سود
	.دهدار و مرادی(۱۴۰۰)	سودمندی اطلاعات
	ساوانگاریراک و تانتاماتی[3] (۲۰۲۲)	نسبت سود ناخالص به سود کسب و کار اولیه
	.نیکبخت و پناهی(۱۴۰۱)	شاخص های رشد
	یوسریانتی و همکاران[4](۲۰۲۰)	اندازه شرکت
	سیفی، میرانی(۱۴۰۲)	کیفیت حسابرس
شاخص سلامت مالی financial health indices	زیولیادی و مکاران۵(۲۰۲۳)	تاثیر اهرم
	ساوانگاریراک و تانتاماتی[6] (۲۰۲۲)	نسبت سود ناخالص به سود کسب و کار اولیه، نسبت بدهی بلند مدت به کل دارایی ها و کل دارایی ها.
	تارجو و همکاران[7](۲۰۲۳)	جریان نقدی، یعنی عملیات جریان نقدی به بدهی جاری، نسبت جریان نقد عملیاتی، جریان نقد آزاد، عملیات جریان نقدی به‌کل بدهی، روزهای پرداختنی معوق و تغییر موجودی کالا به عملیات جریان نقدی
شاخص کارایی مدیریت management efficiency indices	ژان[8] (۲۰۲۱)	نسبت بدهی به دارایی گردش کل دارایی گردش دارایی جاری نسبت سهامداری

[1] Yao
[2] Tarjo
[3] Sawangarreerak, S.; Thanathamathee
4 Yusrianti1
[5] Zuliadi
[6] Sawangarreerak, S.; Thanathamathee
[7] Tarjo
[8] Jan

شاخص راهبری شرکتی corporate governance indices	آریوم و همکاران[1](۲۰۲۳)	حاکمیت شرکتی
	ماندال[2] (۲۰۲۳)	سیستم‌های کنترل داخلی و حاکمیت شرکتی
	سوبیر و همکاران[3](۲۰۲۰)	استقلال هیئت‌مدیره، تخصص هیئت‌مدیره ، تمرکز مالکیت، دوگانگی مسئولیت مدیرعامل
	گانگ و همکاران[4](۲۰۲۲)	حاکمیت سهامداران عمده، نسبت سرمایه‌گذاری نهادی خارجی و مدیران مستقل
متغیرهای کنترلی $control\ variables$	سن شرکت	
شاخص صنعت industry dummy	متغیر مجازی صنعت که یک متغیر مجازی است با مقدار یک برای صنعت مربوطه و صفر برای سایر صنایع	
year dummy	متغیر دامی سال که یک متغیر مجازی است با مقدار یک برای سال مربوطه و صفر برای سایر سال‌ها،	

مدل های داده کاوی

داده کاوی رویکردی است که برای استخراج داده های معنی دار از یک مجموعه داده معین با استفاده از یک یا چند رویکرد مانند تکنیک های آماری، یادگیری ماشینی، ریاضی یا هوش مصنوعی استفاده می شود. در میان این رویکردها، انواع مختلفی از تکنیک‌ها را می‌توان برای کلاهبرداری مالی به کار برد، مانند Naïve Bayes (NB)، ماشین بردار پشتیبانی (SVM)، رگرسیون لجستیک (LR)، و بسیاری دیگر(الحشدی و ماگالینگام[5]، ۲۰۲۱). به طور کلی، داده‌کاوی معمولاً برای کشف تقلب‌های مالی استفاده می‌شود که می‌توان آن‌ها را به شش دسته طبقه‌بندی، تجسم، تشخیص بیرونی، خوشه‌بندی، رگرسیون و پیش‌بینی طبقه‌بندی کرد (آشتیانی و راحمی[6]، ۲۰۲۱). علاوه بر این، اعتقاد بر این است که در دو سال گذشته شاهد عملیات گسترده ای از فعالیت های کلاهبرداری با هدف قرار دادن یک سازمان از هر ۳ سازمان بوده است. اما غیر منتظره ترین چیز این است که تنها ۱۰٪ از این کلاهبرداری های مالی به طور تصادفی کشف می شوند(مالیک و همکاران[7]، ۲۰۲۲).

انتخاب ویژگی

انتخاب ویژگی یک کار ضروری برای انجام در آماده‌سازی داده‌ها است زیرا دردسر کاهش ابعاد یک مسئله جدی است که ممکن است به بیش از حد برازش منجر شود. انتخاب ویژگی به حذف ویژگی های اضافی و نامربوط می پردازد(مالیک و همکاران[8]، ۲۰۲۲). مجموعه ویژگی های بزرگ منجر به هزینه محاسباتی بالا، برازش بیش از حد و عملکرد ضعیف می شود. تکنیک‌های طرح‌ریزی زیرفضا مانند تجزیه و تحلیل مؤلفه اصلی (PCA) می‌تواند ابعاد مجموعه ویژگی را کاهش دهد(سند[9]، ۲۰۲۲). هدف از این کار حفظ ویژگی ها در شکل اصلی است. انتخاب ویژگی (FS) تلاش می‌کند تا بزرگی ابعادی را حل کند، بدون اینکه قابلیت تفسیر ویژگی‌ها را از دست بدهد. الگوریتم های FS را می توان به عنوان روش های بسته بندی، فیلتر و تعبیه شده طبقه بندی کرد(رازک و همکاران[10]، ۲۰۲۲) .

طبقه بندی

تکنیک‌های مختلف طبقه‌بندی یادگیری ماشین برای شناسایی تراکنش‌های متقلبانه همانطور که قبلاً بحث شد، استفاده شده است. با این حال، هیچ الگوریتم بهینه ای برای یک مسئله خاص وجود ندارد(مالیک و همکاران[11]، ۲۰۲۲). بنابراین، چهار مدل داده کاوی مختلف از ادبیات انتخاب شدند زیرا عملکرد امیدوارکننده‌ای را در زمینه تشخیص تقلب نشان می‌دهند(الحشدی و ماگالینگام[12]، ۲۰۲۱).

[1] Arum
[2] Mandal
[3] Subair
[4] Gong et al.
[5] Al-Hashedi and Magalingam,
[6] Ashtiani, M.N.; Raahemi
[7] Malik
[8] Malik
[9] Sanad
[10] Razak
[11] Malik
[12] Al-Hashedi and Magalingam,

درخت تصمیم

درخت تصمیم[1] در داده‌کاوی، مدلی است که جهت نمایش طبقه کننده‌ها و رگرسیون‌ها استفاده می‌شود. همان‌طور که از نام آن مشخص است، این درخت از تعدادی گره و شاخه تشکیل‌شده است. در درخت تصمیمی که عمل طبقه‌بندی را انجام می‌دهد، برگ‌ها بیانگر کلاس‌ها هستند. در هر یک از گره‌های دیگر (گره‌های غیر برگ) با توجه به یک یا چند صفت خاصه تصمیم‌گیری صورت می‌گیرد (اسماعیلی، ۱۳۹۳). الگوریتم‌های متعددی برای ساخت درخت تصمیم وجود دارند که در این پژوهش با توجه به عمومیت استفاده از الگوریتم C5.0 در مباحث مربوط به درخت تصمیم، از الگوریتم C5.0 که آخرین نسخه C4.5 است استفاده می‌شود.C5.0، یکی از مهم‌ترین و عمومی‌ترین و درعین‌حال آخرین نسخه‌ی الگوریتم درخت تصمیم و روش تولید قوانین است که به تولید قوانین آنگاه / اگر پرداخته و به فرآیند تصمیم‌گیری کمک می‌نماید.

قاعده بیز

یکی از فرمول‌های مهم احتمال، فرمول احتمال بیز[2] است که می‌توانیم به کمک آن احتمال برچسب کلاس یک نمونه از داده‌ها را تخمین بزنیم. استفاده از این قانون برای طبقه‌بندی، دقت و سرعت خوبی را در پایگاه داده‌های بزرگ به همراه دارد. در این روش فرض بر این است که تأثیر مقدار یک صفت خاصه بر روی برچسب کلاس مستقل از مقادیر دیگر صفات خاصه است و این موضوع استقلال شرطی[3] کلاس نامیده می‌شود. در نظریه بیز احتمالات شرطی زیر راداریم:

$$P(C|A) = \frac{P(A,C)}{P(A)} \quad (۱)$$

$$P(A|C) = \frac{P(A,C)}{P(C)} \quad (۲)$$

بر اساس نظریه بیز رابطه زیر برقرار است:

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} \quad (۳)$$

با استفاده از نظریه بیز، احتمال $P(C|A1, A2, A3, \dots An)$ برای تمام دسته‌های موجود محاسبه می‌شود که به‌صورت زیر است:

$$P(A1, A2, A3, \dots An|C) = \frac{P(A1,C)P(A2,C)\dots P(An,C)}{P(C)} \quad (۴)$$

دسته‌ای که به ازای آن اندازه احتمال $P(C|A1, A2, A3, \dots An)$ بیشینه می‌شود، انتخاب خواهد شد. در حقیقت ازآنجایی‌که مخرج رابطه ۴ برای تمام دسته‌ها یکسان است، درنتیجه عملاً هدف پیدا کردن دسته‌ای است که صورت رابطه را بیشینه می‌کند(صنیعی و همکاران،۱۳۹۳).

شبکه‌های عصبی

شبکه‌های عصبی[4] نوعی مدل‌سازی ساده از سیستم‌های عصبی واقعی هستند. حوزه‌ی کاربرد این شبکه‌ها آن‌چنان گسترده است که از کاربردهای طبقه‌بندی گرفته تا کاربردهایی نظیر تخمین و آشکارسازی و ... را شامل می‌شود. در حقیقت یک شبکه مجموعه‌ای از ساختارهای ساده است که در کنار یکدیگر سیستم پیچیده‌ی نهایی را توصیف می‌کنند. مجموعه‌ای از گره‌ها و اتصالات بین آن‌ها دو مؤلفه‌ی اصلی تشکیل‌دهنده‌ی شبکه‌ها هستند. هرگاه واحد محاسباتی شبکه محسوب می‌شود که می‌تواند پردازشی ساده و یا پیچیده را بر روی ورودی‌ها انجام دهد. در شبکه‌های عصبی مصنوعی، هر گره به‌عنوان یک نرون مصنوعی در نظر گرفته می‌شود. می‌توان گفت نحوه‌ی مدل کردن نرون از اساسی‌ترین نکات کلیدی در کارایی شبکه‌ی عصبی است اما نحوه‌ی برقراری اتصالات شبکه نیز فاکتور مهمی است. یکی از ساده‌ترین و درعین‌حال کارآمدترین اتصال پیشنهادی، مدل پرسپترون چندلایه[5]یا به‌اختصار*MLP* است که از یک‌لایه‌ی ورودی، یک یا چندلایه‌ی پنهان [6] و یک‌لایه‌ی خروجی تشکیل یافته است. تعداد نورون‌های هر لایه مستقل از تعداد نورون‌های لایه‌های دیگر است (اسماعیلی،۱۳۹۳).

نزدیک‌ترین همسایه [7](*KNN*)

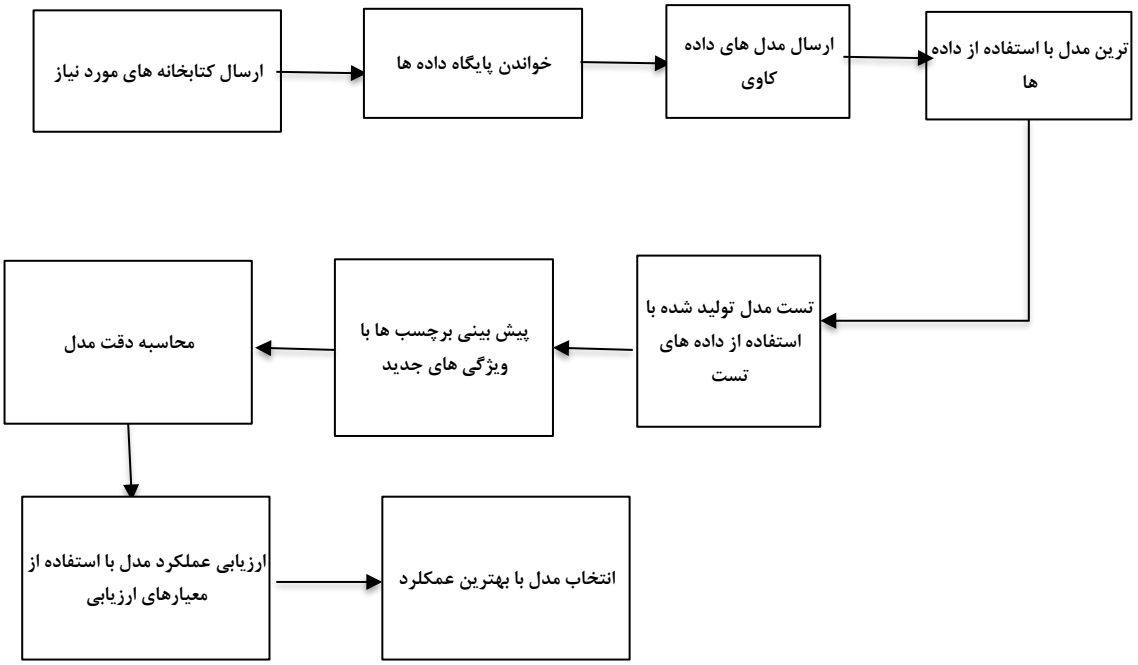
روش K نزدیک‌ترین همسایه یک گروه شامل K رکورد از مجموعه رکوردهای آموزشی که نزدیک‌ترین رکوردها به رکورد آزمایشی باشند را انتخاب کرده و بر اساس برتری دسته یا برچسب مربوط به آن‌ها در مورد دسته رکورد آزمایشی مزبور تصمیم‌گیری می‌نماید. به عبارت ساده‌تر این روش دسته‌ای را انتخاب می‌کند که در همسایگی انتخاب‌شده بیشترین تعداد رکورد منتسب به آن دسته باشند. بنابراین دسته‌ای که از همه دسته‌ها بیشتر در بین K نزدیک‌ترین همسایه مشاهده شود، به عنوان دسته رکورد جدید در نظر گرفته می‌شود. ایده اصلی روش KNN این است که اگر موجودی مثل اردک راه برود و مثل اردک

[1] Decision Tree
[2] Bayesian Theorem
[3] Conditional Independence
[4] Neural Networks
[5] Perceptron Multilayer
[6] Hidden Layer
[7] K-Nearest Neighbor

quack quack کند پس حتماً یک اردک است. یکی از روش‌های محاسبه فاصله بین رکورد جدید و رکوردهای آموزشی، استفاده از فاصله اقلیدسی بین آنها است. فاصله اقلیدسی بین دو رکورد p و q با استفاده از رابطه زیر محاسبه می‌شود:

$$d(p,q) = \sqrt{\sum i}\,(p_i - q_i)^2 \qquad (۵)$$

انتخاب مقدار k درروش دسته‌بندی KNN بسیار مهم و کلیدی است. اگر مقدار K خیلی کوچک انتخاب شود، الگوریتم به نویز حساس می‌شود. درواقع نویزها نزدیک آن رکورد ممکن است ایجاد اشتباه کنند. اگر مقدار K خیلی بزرگ انتخاب شود، ممکن است در میان نزدیک‌ترین همسایه‌ها رکوردهایی از دسته‌های دیگر نیز قرار گیرند. یکی از ایده‌هایی که برای حل این مشکل ارائه‌شده، تعریف فاکتور وزن است. این فاکتور وزنی برابر $1/d^2$ است که مقدار d بیانگر فاصله هر رکورد تا رکورد ورودی است. به‌این‌ترتیب فاصله‌ها برای الگوریتم اهمیت پیدا می‌کنند و این وزن دهی سبب می‌شود که به رکوردهایی که نزدیک‌تر به رکورد ورودی(رکورد جدید)هستند، اهمیت بیشتری داده شود(صنیعی و همکاران،۱۳۹۳). مراحل ساخت مدل‌ها در پایتون در نمودار زیر نشان داده‌شده است:



نمودار ۱: مراحل ساخت مدل‌های داده‌کاوی

معیارهای ارزیابی

از معیارهای ارزیابی زیر استفاده شد: صحت، دقت. در توصیف این معیارهای ارزیابی، از تعاریف زیر استفاده خواهیم کرد: موارد مثبت کاذب: شرکت‌های متقلبی که به‌اشتباه در شرکت‌های غیرمتقلب طبقه‌بندی‌شده‌اند. منفی‌های کاذب (FN): شرکت‌های غیرمتقلبی که به اشتباه در شرکت‌های متقلب طبقه‌بندی‌شده‌اند. مثبت درست: شرکت‌های متقلبی که به‌درستی به‌عنوان متقلب طبقه‌بندی‌شده‌اند. منفی درست (TN): شرکت‌های غیرمتقلبی که به‌درستی به‌عنوان غیر متقلب طبقه‌بندی‌شده‌اند. صحت در معادله ۶ نشان داده‌شده است:

$$accuracy = \frac{(\frac{TP}{TP}+FN)+(\frac{TN}{TN}+FP)}{2} \qquad (۶)$$

دقت اغلب به مقادیر پیش‌گوی مثبت مربوط می‌شود و نسبت تعداد نمونه‌های مثبت درست طبقه‌بندی‌شده به تمام نمونه‌های طبقه‌بندی‌شده با برچسب مثبت می‌باشد و در معادله۷ نشان داده‌شده است.

$$Precision = \frac{TP}{TP+FP} \qquad (۷)$$

بازخوانی: نرخ گزارش مثبت درست، نسبت نمونه‌های مثبت درست طبقه‌بندی‌شده به تمام نمونه‌های درست می‌باشد و در معادله ۸ نشان داده‌شده است.

$$Recall = \frac{TP}{TP+FN} \qquad (۸)$$

و معیار f که نسبتی از دقت و بازخوانی است و در معادله ۹ نشان داده‌شده است.

$$F-Measure = 2.\frac{Precision.Recall}{Precision+Recall} \quad (۹)$$

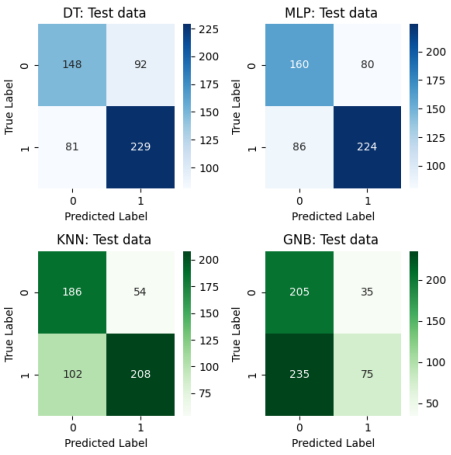
معیار AVP دقت متوسط با ناحیه زیر منحنی دقت- بازخوانی تعریف می‌شود، و میانگین دقت میانگین صرفاً میانگین این مقیاس در تمام کلاس‌های شی است.

و در نهایت امتیاز AUC: که عددی واحد است و عملکرد طبقه‌بندی کننده را در تمام آستانه‌های طبقه‌بندی ممکن خلاصه می‌کند. این امتیاز سطح زیر منحنی ROC را اندازه‌گیری کنید. امتیاز AUCمنحنی ROC نشان می دهد که طبقه‌بندی کننده تا چه اندازه کلاس‌های مثبت و منفی را تشخیص می دهد. می‌تواند مقادیری از ۰ تا ۱ بگیرد.

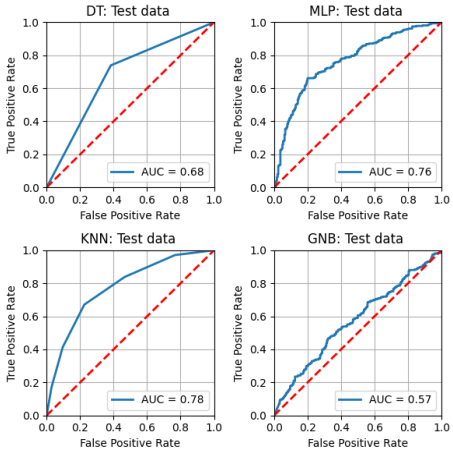
یافته‌های پژوهش

مدل‌های داده‌کاوی این پژوهش برای ۲۲۰۰ داده سال – شرکت اجراشده که تعداد ۵۶ درصد شرکت‌ها متقلب و ۴۴ درصد غیر متقلب بوده‌اند، که تقریباً نسبت متعادلی را نشان می دهد و نیازی به هموارسازی داده وجود ندارد. بنابراین مدل برای ۶۰ درصد داده‌های ترین و برای ۴۰ درصد باقیمانده تست شده است. در شکل زیر ماتریس درهم‌ریختگی(confusion matrix) و نمودار ROC برای داده‌های ترین و تست چهار مدل داده‌کاوی مورد نیاز پژوهش ارائه‌شده است.

استفاده از داده‌های آموزشی برای ارزیابی عملکرد مدل ممکن است با خطر بروز سویه همراه باشد. در بسیاری از موارد، مدل‌ها به جای یادگیری الگوهای واقعی، تمایل به حفظ نمونه‌ها دارند که این امر منجر به بیش برازش [1] می‌شود. به منظور جلوگیری از این مشکل، در این پژوهش از اعتبارسنجی متقابل با ۱۰ تکه [2] استفاده شده است. در این روش، برای هر تکه (fold)، مدل با استفاده از ۹ تکه باقی‌مانده آموزش داده شده و سپس بر روی تکه جدا شده آزمون می‌شود. در نهایت، میانگین عملکرد مدل محاسبه خواهد شد.



شکل ۱: ماتریس درهم‌ریختگی برای داده‌های تست



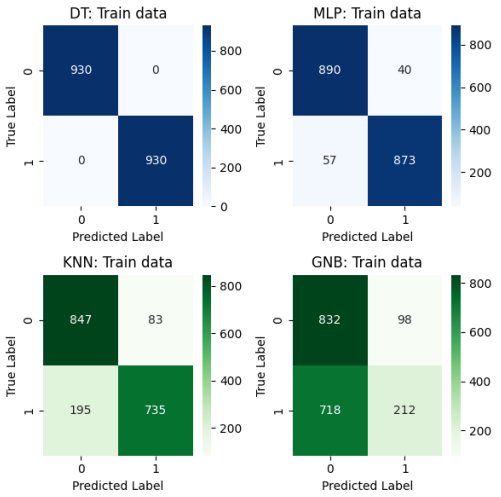
شکل ۲: نمودار ROC برای داده‌های تست

جدول ۲: معیارهای ارزیابی داده‌های تست

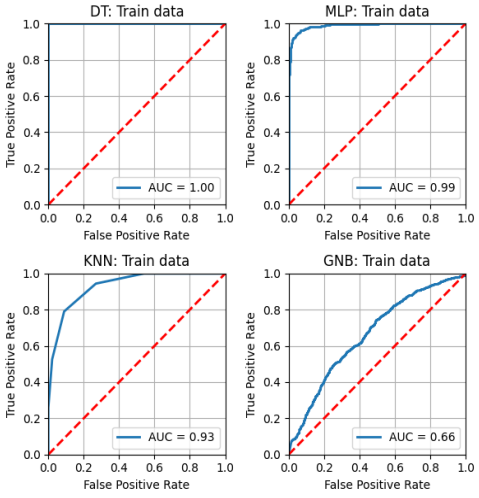
مدل‌ها	معیارهای ارزیابی					
	Acc	Pre	Rec	AUC	AVP	FSCORE
درخت تصمیم(DT)	۰/۶۹	۰/۷۴	۰/۷۱	۰/۶۸	۰/۷۳	۰/۷۳
نزدیک‌ترین همسایه (KNN)	۰/۷۲	۰/۶۷	۰/۷۹	۰/۷۸	۰/۷۸	۰/۷۳
شبکه عصبی (MLP)	۰/۷۰	۰/۷۲	۰/۷۴	۰/۷۶	۰/۷۸	۰/۷۳
قاعده بیز(GNB)	۰/۵۱	۰/۲۴	۰/۶۸	۰/۵۷	۰/۶۴	۰/۳۶

[1] overfitting

[2] fold Cv-10



شکل ۳: ماتریس درهم‌ریختگی برای داده‌های ترین



شکل ۴: نمودار ROC برای داده‌های ترین

جدول ۳: معیارهای ارزیابی داده‌های ترین

مدل‌ها	معیارهای ارزیابی					
	Acc	Pre	Rec	AUC	AVP	FSCORE
درخت تصمیم(DT)	۱	۱	۱	۱	۱	۱
نزدیک‌ترین همسایه (KNN)	۰/۸۵	۰/۷۹	۰/۹۰	۰/۹۳	۰/۹۱	۰/۸۴
شبکه عصبی (MLP)	۰/۹۵	۰/۹۴	۰/۹۶	۰/۹۹	۰/۹۹	۰/۹۵
قاعده بیز(GNB)	۰/۵۶	۰/۲۳	۰/۶۸	۰/۶۶	۰/۶۴	۰/۳۴

در جدول ۳ مقادیر امتیاز F ،قاعده بیز بسیار پایین است، به این معنی که عملکرد این مدل مورد تائید نیست، در حالی که، تمام معیارهای ارزیابی سه مدل دیگر مؤثر هستند، که به این معنی است که درخت تصمیم(DT) ، نزدیک‌ترین همسایه(KNN) و شبکه عصبی (MLP) پیش‌بینی خوبی در تشخیص تقلب شرکت‌های بورسی دارد. بر اساس معیارهای ارزیابی درخت تصمیم بهترین عملکرد و پس‌ازآن شبکه عصبی و در نهایت نزدیک‌ترین همسایه عملکردهای خوبی را در تشخیص تقلب ارائه می‌کنند.

همانطور که پیش ازاین اشاره شد در این پژوهش برای برآورد معیارهای ارزیابی عملکرد از رویکرد اعتبار سنجی متقابل با ۱۰ تکه و در ۱۰ اپک استفاده شده است؛ بنابراین برای بررسی فرض نرمال بودن توزیع معیارهای عملکرد محاسبه شده از آزمون کولموگوروف- اسمیرنوف (KS)استفاده گردید. نتایج حاصل نرمال بودن توزیع برازش شده را رد می کند. بنابراین برای مقایسه معیارهای ارزیابی عملکرد از آزمون رتبه علامتدار ویلکاکسون استفاده شده است. این آزمون مبتنی بر آماره کای دو است که در جدول ۴ نشان داده شده است:

جدول ۴: نتایج آزمون ویلکاکسون

کای دو	درجه آزادی	سطح معنی داری
۲۲/۴۶	۳	۰.۰۰۰

این آزمون نشان می دهد نتایج ارزیابی بدست آمده از روش های مختلف برای پیش بینی تقلب تفاوت معنی دار با هم دارند و در جدول ۵ رتبه های هر یک از روش های پیش بینی تقلب نشان داده شده است:

جدول ۵: رتبه های روش های پیش بینی

مدل	حجم نمونه	میانگین رتبه	رتبه نهایی
knn	۱۰	۲۶/۴۰	۲
DT	۱۰	۲۰/۰۵	۳
MLP	۱۰	۲۹/۰۱	۱
GNB	۱۰	۶/۴۵	۴

بر اساس رتبه های نشان داده شده بهترین برآورد متعلق به شبکه عصبی است.

نتیجه‌گیری

صورت‌های مالی آینه وضعیت مالی شرکت‌ها است که اطلاعات مفیدی را در مورد سلامت مالی شرکت منعکس می‌کند. گزارشگری مالی متقلبانه تحریف عمدی اعداد باهدف فریب استفاده‌کنندگان از صورت‌های مالی رخ دهد. اگر نیرویی از هیئت‌مدیره برای دستیابی به مقدار مشخصی از درآمد وجود داشته باشد و درآمد کسب‌شده با پاداش مدیریت مرتبط باشد، محرک‌های انگیزشی کافی برای گزارشگری مالی متقلبانه ایجادشده است. تقلب در صورت‌های مالی به اعتبار شرکت آسیب می‌زند و بر بازار مالی، کشور و اقتصاد جهانی تأثیر منفی می‌گذارد. بنابراین، نیاز به ابزارهای هوشمند و جدید برای توانمندسازی حسابرسان در کشف داده‌های تقلبی بسیار ضروری است. تکنیک داده‌کاوی موجود می‌تواند تا حد زیادی یک سیستم پشتیبانی تصمیم برای مدیران در تشخیص تقلب در صورت‌های مالی فراهم کند. نتایج این تحقیق به گروه‌های مختلف به‌ویژه حسابرسان و سرمایه‌گذاران این امکان را می‌دهد تا با دقت در مورد شرکت‌هایی که در حال حسابرسی هستند یا در آن سرمایه‌گذاری می‌کنند، تصمیم بگیرند. استفاده از این تکنیک‌ها می‌تواند به حسابرسان کمک کند تا با استفاده از ابزارهای خودکار در زمان و هزینه صرفه‌جویی کنند و این مراحل بارکاری آنها را به میزان قابل‌توجهی کاهش می دهد. تحقیق اخیر سناریوهای واقعی را شبیه‌سازی کرده است تا نشان دهد چگونه می‌توان از این تکنیک‌ها به‌درستی در حسابرسی و کشف تقلب استفاده کرد. نتایج این پژوهش بهتر از نتایج ارائه شده در مقاله خدر و همکاران[1](۲۰۲۳) است، مدل پیش‌بینی قدرتمند برای تقلب در صورت‌های مالی او با الگوریتم‌های یادگیری به دقت نهایی ۹۶.۰۵ درصد در تشخیص FSF رسیدند، همچنین کرکاس و همکاران[2](۲۰۲۴) مرزهای تشخیص تقلب در صورت‌های مالی با مدل‌های زبان بزرگ را بادقت ۶۷ درصد برآورد کردند. در حالیکه یافته های پژوهش ما بر اساس درخت تصمیم قدرت ۱۰۰ درصدی در پیش بینی تقلب در صورت‌های مالی دارد شاید علت این امر همان طور که چن و همکاران[3](۲۰۲۳) بیان می کنند این باشد که الگوریتم‌های یادگیری گروهی عموماً در تشخیص تقلب برای شرکت‌های بورسی چینی بهتر هستند.

مطالعات بیشتر را می‌توان با افزودن متغیرهای بیشتر، مانند گنجاندن سیستم‌های فناوری اطلاعات در سازمان، که به نظر می‌رسد یک عامل تعیین‌کننده باشد، انجام داد. مطالعات تجربی تطبیقی نیز می‌تواند زمانی انجام شود که یک مجموعه داده از سایر بازارهای سهام به‌دست‌آمده باشد. علاوه بر این، روش و روش مورداستفاده در این مقاله می‌تواند به سایر حوزه‌ها برای اهداف مختلف و نه‌تنها برای کشف تقلب گسترش یابد.

فهرست منابع

دهدار، فرهاد و مرادی، محمدحسین. (۱۴۰۰). بهره گیری از رویکرد داده کاوی و شناسایی تقلب در صورت‌های مالی .*چشم انداز حسابداری و مدیریت* ۴ (۴۵) : ۱۱۳-۱۲۳.

کارشناسان، علی، بهرامی نسب، علی، و ممشلی، رضا. (۱۳۹۸). نقش کیفیت سود در شناسایی گزارشگری مالی متقلبانه. پژوهش‌های تجربی حسابداری، ۸(۳ (پیاپی ۳۱)): ۳۳۶-۳۱۳.

کریمی خلیفه محله، پریسا.(۱۴۰۱). بررسی تأثیر کیفیت حسابرسی مستقل بر احتمال وقوع تقلب مالی در شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران.

نیکبخت, محمدرضا، پناهی, سونیا. (۱۴۰۱). بررسی جامع عوامل درون سازمانی اثرگذار بر تقلب‌های مالی: شواهدی از بورس اوراق بهادار تهران .پژوهشهای کاربردی در گزارشگری مالی ,۱۱(۲).

شوشتری، زهرا، یاورزاده محمدرضا. (۱۳۹۹). آمار کاربردی ویژه دانشجویان و محققان علوم انسانی. موسسه مطالعات و پژوهشهای بازرگانی

تشدیدی الهه، سپاسی سحر، اعتمادی، حسین، . (۱۳۹۸). ارائه رویکردی نوین در پیش‌بینی و کشف تقلب صورت‌های مالی با استفاده از الگوریتم زنبور عسل، مجله دانش حسابداری، ۳(۱۰): ۱۶۷-۱۳۹

رحمانیان کوشککی, عبدالرسول ,غفاری شوشتری, غلامعباس. (۱۴۰۲). تاثیر کاهش مکانیزم نظارتی بر رابطه میان حق الزحمه غیرعادی حسابرسان و تقلب در صورت‌های مالی .پژوهش‌های حسابرسی حرفه‌ای

Ashtiani, M.N.; Raahemi, B. **2021**. Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. IEEE Access, 10, 72504–72525.

Ali, A.; Abd Razak, S.; Othman, S.H.; Eisa, T.A.E.; Al-Dhaqm, A.; Nasser, M.; Elhassan, T.; Elshafie, H.; Saif, A. 2022. Financial Fraud Detection Based on Machine Learning: A Systematic

[1] Khedr
[2] Kirkos, et al.
[3] Chen

differences between data mining techniques and experts' judgments," Knowledge- Based Systems, vol. 89, pp. 459–470.

Mark Eshwar Lokanan, Kush Sharma.2022. Fraud prediction using machine learning: The case of investment advisors in Canada, Machine Learning with Applications, Volume 8,

Malik, E.F.; Khaw, K.W.; Belaton, B.; Wong, W.P.; Chew, X. 2022. Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture. Mathematics, 10, 1480. https://doi.org/10.3390/math10091480

Mohammadi, M.; Yazdani, S.; Khanmohammadi, M.H.; Maham, K. 2020. Financial reporting fraud detection: An analysis of data mining algorithms. Int. J. Financ. Manag. Account., 4, 1–12.

Nesvijevskaia, A., Ouillade, S., Guilmin, P., & Zucker, J. 2021. The accuracy versus interpretability trade-off in fraud detection model. Data & Policy, 3, E12. doi:10.1017/dap.2021.3

Paulo Sérgio Gomes Macedo, H.C.I. 2022. Vieira, E.S. A model to detect financial statement fraud in Portuguese companies by the auditor. Contaduría Adm., 67, 185–209.

Sanad, Z., Al-Sartawi, A. (2021). Financial Statements Fraud and Data Mining: A Review. In: Musleh Al-Sartawi, A.M., Razzaque, A., Kamal, M.M. (eds) Artificial Intelligence Systems and the Internet of Things in the Digital Era. EAMMIS 2021. Lecture Notes in Networks and Systems, vol 239. Springer, Cham.

Schreiber-Gregory, D.; Bader, K. 2018. Logistic and Linear Regression Assumptions: Violation Recognition and Control. In Proceedings of the SESUG Conference, St. Pete Beach, FL, USA, 14–17 October; pp. 1–6.

Strelcenia, E.; Prakoonwit, S. 2023. Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation. AI, 4, 172–198

Wadhwa, A.V.K.; Kumar, S. 2020. Financial Fraud Prediction Models: A Review of Research Evidence. Int. J. Sci. Technol. Res., 9, 677–680

Literature Review. Appl. Sci., 12, 9637. https://doi.org/10.3390/app12199637

Alfaiz, N.S.; Fati, S.M. 2022. Enhanced Credit Card Fraud Detection Model Using Machine Learning. Electronics, 11, 662.

Alsinglawi, M.M.A.S.M.A.O.; Almari, M.O.S. 2021. Predicting Fraudulent Financial Statements Using Fraud Detection Models. Acad. Strateg. Manag., 20, 1–17.

Andrew, C.; Robin. 2022. Detecting Fraudulent of Financial Statements Using Fraud S.C.O.R.E Model and Financial Distress. Int. J. Econ. Bus. Account. Res. (IJEBAR), 6, 211–222

Chen, Y.; Wu, Z. 2023. Financial Fraud Detection of Listed Companies in China: A Machine Learning Approach. Sustainability, 15, 105. https://doi.org/10.3390/su15010105

Deebak, B.; Memon, F.H.; Dev, K.; Khowaja, S.A.;Wang,W.; Qureshi, N.M.F. 2022. TAB-SAPP: A trust-aware blockchain-based seamless authentication for massive IoT-enabled industrial applications. IEEE Trans. Ind. Inform., 19, 243–250.

du Toit, E. 2023. The red flags of financial statement fraud: a case study. Journal of Financial Crime, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JFC-02-2023-0028

Khaled Gubran Al-Hashedi, Pritheega Magalingam, 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019,Computer Science Review,Volume 40.

Khedr, A.M.; Arif, I.; P V, P.R.; El-Bannany, M.; Alhashmi, S.M.; Sreedharan, M. 2021. Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. Intell. Syst. Account. Financ. Manag., 28, 3–34

Kirkos, Efstathios and Boskou, Georgia and Chatzipetrou, Evrikleia and Tiakas, Eleftherios and Spathis, Charalampos, 2024. Exploring the Boundaries of Financial Statement Fraud Detection with Large Language Models (May 27,). Available at SSRN: https://ssrn.com/abstract=4842962 or http://dx.doi.org/10.2139/ssrn.4842962

Kumar, S.; Ahmed, R.; Bharany, S.; Shuaib, M.; Ahmad, T.; Tag Eldin, E.; Rehman, A.U.; Shafiq, M. 2022. Exploitation of Machine Learning Algorithms for Detecting Financial Crimes Based on Customers' Behavior. Sustainability, 14, 13875.

Kulikova, L.; Satdarova, D. 2016. Internal control and compliance-control as effective methods of management, detection and prevention of financial statement fraud. Acad. Strateg. Manag. J., 15, 92.

Li, H.; Wong, M.L. 2015. Financial fraud detection by using Grammar-based multi-objective genetic programming with ensemble learning. In Proceedings of the 2015 IEEE Congress on Evolutionary Computation (CEC), Sendai, Japan, 25–28 May; IEEE: New York, NY, USA, pp. 1113–1120.

Lin, A. Chiu, S.Y.Huang, andD. C.Yen, 2015.Detecting the financial statement fraud: The analysis of the

Accounting Knowledge & Management Auditing
Vol. 17/ No. 65/ Spring 2027

Detecting financial fraud statements of companies listed on the Tehran Stock Exchange using data mining

Ahmad Khaksaran
PhD student, department of accounting, Sanandaj branch, Islamic Azad University, Sanandaj , Iran.
ahmadkhaksaran33@gmail.com

Omid.Mahmoodi Khoshroo
**Assistant Professor, Department of Accounting, Sanandaj branch, Islamic Azad University, Sanandaj, Iran.
(corresponding author: omid.khoshroo@gmail.com)

Karim Ghaderzadeh
Assistant Professor, Department of Accounting Payam noor ,Tehran , Iran.
K.ghaderzadeh@pnu.ac.ir

Peyman Amini
Assistant Professor, Department of Accounting, Sanandaj branch, Islamic Azad University
p.amini@uok.ac.ir

Abstract
Fraudulent financial statements not only result in significant losses for investors but also undermine trust in accounting firms, thereby creating a crisis of confidence. Furthermore, such fraudulent activities can entrap a company's financial health in a detrimental cycle, hindering the long-term sustainable development of the broader economy. Consequently, the identification of fraudulent financial statements is crucial for enhancing the sustainability of our socio-economic landscape. This study aims to detect fraud in the financial statements of companies listed on the Tehran Stock Exchange through the application of data mining techniques. The research sample comprises 200 listed companies over an 11-year period, spanning from 2013 to 2023. Evaluation criteria were established for four distinct models: Decision Tree, Neural Network, Nearest Neighbor, and Bayesian Rule. The results indicate that the F-score values for the Bayesian Rule model are considerably low, rendering its performance inadequate. In contrast, the other three models—Decision Tree (DT), Nearest Neighbor (KNN), and Neural Network (MLP)—demonstrated effective predictive capabilities in detecting fraud among listed companies. The findings suggest that existing data mining techniques can significantly enhance decision support systems for managers in identifying fraudulent activities within financial statements.
Key Words: Financial fraud statements, Tehran Stock Exchange, data mining