Journal of
Empirical Research In Accounting

Alzahra University, Faculty of Social Sciences and Economics
Empirical Research in Accounting, Winter 2022, V. 12, No. 46, pp.255-276

Financial Reporting Fraud Scheme Prediction via Machine Learning Approach – Multiclass Classification[1]

Tohid Kazemi[2], Parviz Piri[3]

Received: 2022/06/15
Accepted: 2022/11/20

Research Paper

Abstract

This paper attempts to evaluate the performance of machine learning models in fraudulent financial Reporting schemes prediction via a multi-classification approach and using an unbalanced dataset. Therefore, the financial statements of 134 companies listed on the Tehran Stock Exchange from 2009 to 2021 were investigated by Logistic Regression, Decision Tree, Boosting Algorithms, and Support Vector Machine. Models were programmed with Python and Performance indicators were calculated and compared. Furthermore, the machine learning model's performance was investigated in binary classification with thebalanced dataset to predict each fraud scheme exclusively. According to the results via a multi-classification approach, then thesignificant difference between machine learning models' performance was approved. Support Vector Machin was preferred in multiclass problem space with the unbalanced data set. To predict fraud schemes via binary classification, a significant difference between machine learning models' performance was not approved except to predict the "Overstatement assets and income" scheme. Support Vector Machin was preferred to Logistic Regression and Decision Tree model. The present research attempts to fill the research gap in the research area by developing machine learning models with a multi-classification approach.

Keywords: Fraud Scheme, Fraudulent Financial Reporting, Machine Learning, Multi-Classification.

JEL Classification: M41, M42, G32.

---

1. DOI: 10.22051/JERA.2022.41290
2. Assistant Professor, Department of Accounting, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran. (Corresponding Author), (t.kazemi@alzahra.ac.ir).
3. Associate Professor, Department of Accounting, Faculty of and Economics and Management, Urmia University, Urmia, Iran. (p.piri@urmia.ac.ir).

فصلنامه پژوهش‌های تجربی حسابداری «نشریه علمی»

دانشگاه الزهرا (س)، دانشکده علوم اجتماعی و اقتصادی

پژوهش‌های تجربی حسابداری، سال دوازدهم، شماره ٤٦، زمستان ١٤٠١، صص ٢٧٦-٢٥٥

پیش بینی طرح تقلب در گزارشگری مالی با استفاده از رویکرد یادگیری ماشین در فضای چند کلاسه[1]

توحید کاظمی[2]، پرویز پیری[3]

مقاله پژوهشی

تاریخ دریافت: ۱۴۰۱/۰۳/۲۵
تاریخ پذیرش: ۱۴۰۱/۰۸/۲۹

چکیده

هدف از انجام پژوهش حاضر بررسی عملکرد الگوهای یادگیری ماشین در پیش بینی طرح‌های تقلب مورد استفاده در گزارشگری مالی در فضای چند کلاسه با استفاده از مجموعه داده نامتوازن است. از این رو صورت‌های مالی ۱۳۴ شرکت‌پذیرفته شده در بورس اوراق بهادار تهران در قلمرو زمانی سال ۱۳۸۷ الی ۱۳۹۹ با استفاده از روش‌های رگرسیون لجستیک، درخت تصمیم، الگوریتم گرادیان تقویت شده و ماشین بردار پشتیبان مورد تحلیل و بررسی قرار گرفته‌اند. الگوهای مزبور در محیط پایتون با رویکرد چندکلاسه پیاده سازی و اجرا شدند. معیار ارزیابی عملکرد محاسبه و مقایسه شد. افزون بر این عملکرد الگوهای یادگیری ماشین در تشخیص نوع تقلب در صورت‌های مالی با رویکرد دوکلاسه و بر اساس مجموعه داده متوازن نیز بررسی گردید. نتایج پژوهش نشان می‌دهد تفاوت معنادار در عملکرد الگوهای یادگیری ماشین در فضای چند کلاسه وجود دارد و روش ماشین بردار پشتیبان نسبت به سایر روش‌ها عملکرد بهتری دارد. با تقلیل فضای مسئله به دسته بندی دو کلاسه تفاوت معنادار در عملکرد الگوهای یادگیری ماشین در تشخیص گزارش‌های مالی مشکوک به "بیش نمایی دارایی، کم نمایی بدهی و هزینه"، "بیش نمایی دارایی و کم نمایی هزینه" و "کم نمایی هزینه و بدهی" تایید نشد. با این حال، عملکرد ماشین بردار پشتیبان بر عملکرد روش رگرسیون لجستیک و درخت تصمیم در پیش بینی گزارش‌های مالی مشکوک به "بیش نمایی دارایی و درآمد" ارجح است. پژوهش حاضر با توسعه فضای مسئله با هدف دسته بندی چندکلاسه سعی دارد شکاف تحقیقاتی موجود در قلمرو موضوعی پژوهش را رفع نماید.

واژه های کلیدی: طرح تقلب، گزارشگری مالی متقلبانه، یادگیری ماشین، دسته بندی چند کلاسه.

طبقه بندی موضوعی: G32, M41,M42.

---

۱. DOI: 10.22051/JERA.2022.41290

۲. استادیار، گروه حسابداری، دانشکده علوم اجتماعی و اقتصادی، دانشگاه الزهرا، تهران، ایران. (نویسنده مسئول). (t.kazemi@alzahra.ac.ir)

۳. دانشیار، گروه حسابداری، دانشکده اقتصاد و مدیریت، دانشگاه ارومیه، ارومیه، ایران. (p.piri@urmia.ac.ir)

مقدمه

صورت‌های مالی ارائه شده توسط بنگاه‌های اقتصادی، به عنوان یک منبع اطلاعاتی حاوی اطلاعات سودمند برای تصمیم گیری سرمایه گذاران و اعتباردهندگان در بازارهای مالی است که سبقه فعالیت اقتصادی بنگاه را ارائه می‌دهد. انباشت این اطلاعات در طول زمان منجر به اضافه بار اطلاعاتی شده و تحلیل این اطلاعات را دشوار می‌سازد. از سوی دیگر تقلب در ارائه صورت‌های مالی به عنوان یک پدیده نامطلوب با تاثیر بر قیمت سهام شرکت، باعث تغییرات غیر منطقی قیمت سهام و ارزش سرمایه سهامداران شده و اعتماد سرمایه‌گذاران به نظام گزارشگری مالی را خدشه‌دار می کند. گزارشگری مالی متقلبانه به واسطه گمراه کردن سرمایه-گذاران، اعتباردهندگان و دولت باعث اخلال در نظام توزیع عادلانه ثروت می‌شود. در پی تقلب در گزارشگری مالی، منابع اقتصادی محدود به سمت بنگاه‌های اقتصادی ناموفق هدایت می‌شوند که نتیجه آن اتلاف منابع است (سجادی و کاظمی، ۱۳۹۵).

از این رو کشف و استخراج الگوهای تقلب از بطن صورت‌های مالی به کمک علم داده کمک شایانی به استفاده کنندگان از آن می‌نماید. علم داده از علوم مختلفی از جمله علم آمار، هوش مصنوعی، یادگیری ماشین، شناسایی الگو و پایگاه داده نشات گرفته است که فرآیند آن شامل سه مرحله‌ی آماده سازی داده، یادگیری مدل، ارزیابی و تفسیر مدل است. یادگیری ماشین الگوریتم هایی را ایجاد می کند که رایانه ها به منظور شناسایی و استخراج الگوها از داده‌ها آن‌ها را درک می کنند. یادگیری ماشین فرآیندی دو مرحله‌ای است. در مرحله اول الگوریتم یادگیری ماشین روی مجموعه‌ای از داده‌ها اعمال می‌شود تا الگوهای مفید موجود در داده‌ها را شناسایی کند. در مرحله دوم وقتی مدل ایجاد شد برای تجزیه و تحلیل استفاده می‌شود (کلهر و تیرنی، ۱۴۰۰). این الگوها به روش‌های متفاوتی مانند مدل رگرسیون، شبکه عصبی مصنوعی، درخت تصمیم گیری، ماشین بردار پشتیبان و بوستینگ ارائه می‌شوند.

پژوهشهای پیشین در حوزه پیش بینی تقلب در گزارشگری مالی، وقوع یا عدم وقوع تقلب را مورد بررسی قرار داده‌اند. به دیگر سخن پژوهش‌های پیشین از مزایای روش‌های یادگیری ماشین برای دسته بندی صورت‌های مالی در دو طبقه متقلبانه و سالم بهره جسته‌اند. در این حوزه شناسایی نوع تقلب مغفول واقع شده است. عضویت سازمان بورس و اوراق بهادار در جمع اعضای سازمان بین المللی کمیسیون های اوراق بهادار، توجه ویژه به‌سلامت بازار سرمایه و شناسایی به موقع طرح تقلب، الزام به ارتقاء کیفیت اطلاعات، افزایش تعداد شرکت‌های پذیرفته شده بر لزوم توجه به موضوع کشف و شناسایی طرح تقلب تاکید می‌نماید.

رشد، توسعه و ترویج مفهوم حسابداری قضایی در کشور و لزوم استفاده از روش‌های نوین در فرآیند کشف تقلب و شناسایی طرح تقلب در گزارشگری مالی می‌تواند زمینه آشنایی کارشناسان رسمی دادگستری و حسابرسان مستقل با علم داده و مزایای آن را فراهم آورد. از این رو به منظور رفع شکاف تحقیقاتی موجود، پژوهش حاضر سعی دارد با بهره مندی از مزایای رویکرد یادگیری ماشین چند کلاسه راهکاری جدید در پیش بینی طرح تقلب در صورت‌های مالی در کشور ارائه نماید. استفاده از رویکرد چند کلاسه در محیط مجموعه داده نامتوازن، فضای مسئله را در شرایط واقعی تعریف می‌کند و تمام مشاهدات را در حل مسئله مورد استفاده قرار می‌دهد. افزون بر این به منظور مطالعه بیشتر، عملکرد الگوهای یادگیری ماشین در تشخیص طرح‌های تقلب مورد استفاده در گزارشگری مالی در محیط دو کلاسه با مجموعه داده متوازن بررسی شد. از این رو مسئله پژوهش به شرح زیر قابل طرح است.

«در شرایط واقعی، عملکرد الگوهای یادگیری ماشین در شناسایی طرح‌های تقلب در صورت های مالی چگونه است؟» منظور از شرایط واقعی، تعریف مسئله پژوهش در فضای چند کلاسه و حل آن بر اساس مجموعه داده نامتوازن است. در این مقاله طرح تقلب در صورت‌های مالی با استفاده از روش‌های رگرسیون، درخت تصمیم، الگوریتم گرادیان تقویت شده و ماشین بردار پشتیبان شناسایی شده و عملکرد روش‌های مزبور مقایسه شده است.

مبانی نظری و توسعه فرضیه‌های پژوهش

گزارشگری مالی متقلبانه: طبق گزارش ۲۰۲۲ انجمن بازرسان رسمی تقلب، گزارشگری مالی متقلبانه از لحاظ فراوانی کمتر از دیگر انواع تقلب رخ داده و از لحاظ اثر مالی بیشترین زیان را به شرکت‌ها تحمیل کرده است. طبق تعریف این نهاد، تقلب در صورت‌های مالی ارائه نادرست، حذف اقلام و افشا نکردن کافی اطلاعات به منظور فریب کاربران صورت‌های مالی، به خصوص سرمایه گذاران و اعتبار دهندگان است. علاوه بر این، کمیسیون ملی گزارشگری مالی متقلبانه، (کمیسیون تریدوی) گزارشگری مالی متقلبانه را اقدامات عمدی در حذف یا انجام فعالیت‌هایی که منجر به ارائه نادرست صورت‌های مالی شده و صورت‌های مالی مزبور را گمراه کننده می‌سازد، تعریف کرده است (رامنی و استین بارت، ۱۳۸۷: ۱۷۳).

سجادی و کاظمی (۱۳۹۵) معتقدند مهمترین دلیل ارتکاب تقلب در صورت‌های مالی اعمال فشار به مدیریت برای گزارش سود است. نهادهای حرفه‌ای مانند کوزو و انجمن بازرسان رسمی

تقلب طبقه‌بندی‌های متفاوتی از طرح‌های تقلب در صورت‌های مالی، ارائه کرده‌اند. کوزو (۲۰۱۰)، روش‌های مرسوم تقلب در صورت‌های مالی را در ۷ طبقه اصلی شامل شناسایی نادرست درآمدها، بیش نمایی دارایی‌ها، کم نمایی هزینه‌ها و بدهی‌ها، تخصیص نادرست دارایی‌ها، افشای نامناسب، معاملات با اشخاص وابسته، معاملات درون گروهی و سایر دسته بندی کرده است. انجمن بازرسان رسمی تقلب نیز در گزارش ۲۰۲۲ خود دو طبقه اصلی برای تقلب در صورت‌های مالی در نظر گرفته است. به نظر انجمن مزبور بیش نمایی و کم نمایی خالص دارایی‌ها و درآمدها طبقات اصلی گزارشگری مالی متقلبانه هستند که ممکن است ناشی از تفاوت زمانی، درآمد ساختگی یا کم نمایی درآمد، پنهان سازی بدهی و هزینه یا بیش نمایی بدهی و هزینه، ارزیابی نامناسب دارایی و افشا نامناسب باشد. افزون بر این کرانچر، رایلی، ولز (۲۰۱۱) در یک طبقه بندی کلی، طرح‌های تقلب در صورت‌های مالی را در ۵ طبقه اصلی شامل درآمد ساختگی، تفاوت زمانی، پنهان کردن بدهی و هزینه، افشا نامناسب و ارزیابی نامناسب دارایی‌ها ارائه نموده‌اند.

مسائل چند کلاسه: با عنایت به اینکه طبقه بندی انواع تقلب در حوزه گزارشگری مالی فراتر از دو طبقه می‌باشد لذا مسئله پژوهش حاضر در دسته مسائل چند کلاسه تعریف می‌شود. صنیعی آباده محمودی و طاهرپور (۱۳۹۳)، مسائلی که تعداد دسته‌های آن بیش از دو دسته است را به عنوان مسائل چند کلاسه معرفی می کنند و برای توسعه یک دسته بند دوکلاسه برای دسته بندی مسائل چند کلاسه استفاده از روش "یکی در مقابل یکی" و روش "یکی در مقابل همه" را پیشنهاد می‌کنند. ایشان معتقدند مجموعه داده‌هایی که در آن‌ها ویژگی دسته دارای توزیع نامتوازن باشد در مسائل واقعی مانند تشخیص کلاهبرداری و ناهنجاری بسیار شایع است. در مسائل با دسته‌های نامتعادل، ارزش تشخیص رکوردهای مربوط به دسته نادر بالاتر از تشخیص دسته‌های شایع است. به نظر آنان داده کاوی برای برخورد با مشکل دسته‌های نامتعادل از "راهکار مبتنی بر معیار" و "راهکار مبتنی بر نمونه برداری" استفاده می‌کند. عمادالدین و همکاران (۱۳۹۷) و شریفی راد و نیک نفس (۱۳۹۳) معتقدند یکی از روش های مناسب دسته بندی داده های نامتوازن، استفاده از ماشین بردار پشتیبان است.

پژوهش حاضر مطابق با طبقه بندی انواع تقلب در حوزه گزارشگری مالی سعی دارد در مرحله اول، نوع تقلب واقع شده در صورت‌های مالی را با استفاده از رویکرد یادگیری ماشین در فضای دسته بندی چند کلاسه با استفاده از مجموعه داده نامتوازن بر اساس "راهکار مبتنی بر معیار"

پیش بینی نماید. در مرحله دوم به منظور ساده سازی فرآیند دسته بندی، دسته بندی دو کلاسه با استفاده از مجموعه داده متوازن و بر اساس "راهکار نمونه برداری" اجرا شده است.

رگرسیون لجستیک: رگرسیون لجستیک در دسته‌ای از مدل‌های احتمالی به نام مدل‌های تفکیکی گنجانده می‌شود. در این مدل‌ها فرض بر این است که متغیر وابسته یک مقدار تولید شده از یک توزیع احتمالاتی است که توسط یک تابع از متغیرهای مستقل تعریف می‌شود (آگاروال، ۱۳۹۸: ۲۴۱).

درخت تصمیم: در الگوریتم‌های دسته بندی مبتنی بر درخت تصمیم دانش خروجی به صورت یک درخت از حالات مختلف مقادیر ویژگی‌ها ارائه می‌شود (صنیعی آباده و همکاران، ۱۳۹۳: ۸۸). این الگو مجموعه‌ای از آیا و سپس قوانین دیگر را کد می‌کند. هدف الگوریتم یادگیری درخت تصمیم یافتن مجموعه‌ای از قوانین طبقه بندی است که مجموعه داده‌های آموزشی را به مجموعه‌هایی از نمونه‌ها تقسیم می‌کنند که همان مقدار را برای ویژگی هدف داشته باشند (کلهر و تیرنی، ۱۴۰۰: ۱۱۶). در یک درخت تصمیم برگ‌ها نشان دهنده دسته بندی و شاخه‌ها و گره‌های میانی ویژگی‌های مختلف برای رسیدن به یک کلاس را نشان می‌دهند. درخت تصمیم را می‌توان به کمک مجموعه‌ای از شروط یا قوانین نمایش داد (ویسی، قایدشریف و ابراهیمی، ۱۳۹۸).

الگوریتم گرادیان تقویتی: این الگوریتم از دسته الگوریتم‌های گرادیان تقویتی بوده که از درخت تصمیم باینری به عنوان مبنای پیش بینی کنندگی بهره می‌گیرد و عملکرد بسیار خوبی در دسته بندی، رگرسیون و رتبه بندی دارد. به دلیل پیش بینی دقیق، سرعت زیاد و پشتیبانی از اجرای چندمنظوره و توزیع شده آن، در مسائل دسته بندی بسیار محبوب است (ویسی، قایدشریف و ابراهیمی، ۱۳۹۸).

ماشین بردار پشتیبان: این روش از طریق بی نهایت انطباق غیرخطی بردارهای ورودی به فضای ویژگی با ابعاد بالا یک طبقه بندی کننده دودویی، ابر صفحه های تفکیک کننده بهینه تولید می کند. ماشین بردار پشتیبان از طریق مرزهای طبقه غیرخطی بر مبنای بردارهای پشتیبان، مدلی خطی برای تخمین تابع تصمیم ایجاد می کند. چنانچه داده ها به صورت خطی تفکیک شده باشند، ماشین بردار پشتیبان با آموزش ماشین های خطی، به دنبال یافتن صفحه بهینه ای است که داده ها را بدون خطا و با حداکثر فاصله بین ابر صفحه و نزدیک ترین نقاط آموزشی تفکیک می کند. نزدیک ترین نقاط به ابر صفحه تفکیک کننده بهینه، بردارهای پشتیبان نامیده می شوند

(خواجوی و ابراهیمی، ۱۳۹۶). عمادالدین و همکاران (۱۳۹۷) و شریفی راد و نیک نفس (۱۳۹۳) معتقدند یکی از روش های طبقه بندی داده های نامتوازن، استفاده از ماشین بردار پشتیبان است.

پیشینه پژوهش: مطالعه پیشینه پژوهش نشان می‌دهد، بخشی از تحقیقات درصدد شناسایی صورت‌های مالی متقلبانه با استفاده از نسبت های مالی با تاکید بر روش رگرسیون برآمده‌اند. نتایج حاصل از این پژوهش‌ها حاکی از موفقیت روش رگرسیون در دسته بندی صورت‌های مالی بوده است (کاناپیکین و گروندین ، ۲۰۱۵؛ پرسونس، ۱۹۹۵؛ سپاتیس، ۲۰۰۲؛ فرقاندوست حقیقی، هاشمی و فروغی دهکردی، ۱۳۹۳؛ اعتمادی و زلفی، ۱۳۹۲؛ صفرزاده، ۱۳۸۹). با توسعه کاربرد روش‌های داده کاوی در تحقیقات حسابداری، پژوهشگران به مقایسه عملکرد روش‌های نوین در کشف تقلب روی آوردند.

امیدی، مین، مرادی و پیری (۲۰۱۹)، چن، لیو، وان و چن (۲۰۱۸)، نورماه، جوهری و اسمیث (۲۰۱۷)، لین، چیو، هانگ و ین (۲۰۱۵)، کات سیس و همکاران (۲۰۱۲)، پرلوس (۲۰۱۱)، راویسانکار، راوی، رائو و باس (۲۰۱۱)، ژائو و کاپور (۲۰۱۰)، کی روکس، اسپاتیس و مانولوپولوس(۲۰۰۷) به بررسی عملکرد و کارایی روش‌های مختلف داده کاوی در کشف تقلب در صورت‌های مالی پرداختند. نتایج این دسته از پژوهش‌ها عملکرد موفق روش‌های داده کاوی و یادگیری ماشین در کشف گزارشگری مالی متقلبانه را تایید می‌کند. افزون بر این تحقیق چن و همکاران (۲۰۱۸) و پرلوس (۲۰۱۱) موید عملکرد موفق رویکرد ماشین بردار پشتیبان نسبت به سایر روش‌ها می‌باشد. در تحقیقات داخلی نیز تاراسی، بنی طالبی و زمانی (۱۳۹۸)، خواجوی و ابراهیمی (۱۳۹۶) و مرادی، سلیمانی و باقری (۱۳۹۴) با استفاده از رویکرد درخت تصمیم گیری و شبکه عصبی مصنوعی و بوستینگ موفق به شناسایی صورت‌های مالی متقلبانه شدند.

بخش دیگری از پژوهش‌های خارجی و داخلی در حوزه شناسایی صورت‌های مالی متقلبانه سعی در بهبود عملکرد روش‌های داده کاوی با استفاده از رویکرد ترکیبی دارند (سادگالی و همکاران، ۲۰۱۹، ژان، ۲۰۱۸، ملکی، بحری، جبارزاده و آشتاب، ۱۴۰۰، رضایی، ناظمی و صدرآبادی، ۱۴۰۰). جدیدترین پژوهش‌ها در قلمرو موضوعی پژوهش حاضر در حوزه ارزیابی عملکرد روش‌های یادگیری عمیق در شناسایی تقلب است. نتایج پژوهش‌های مزبور حاکی از موفقیت رویکرد یادگیری عمیق در کشف گزارشگری مالی متقلبانه است (هیوگو و شنگ لینگ، ۲۰۲۲؛ لونگ جان، ۲۰۲۱؛ کراجا، کیم و لس من، ۲۰۲۰).

فرضیه‌های پژوهش

با توجه به مبانی نظری و پیشینه پژوهش و تاکید عمادالدین و همکاران (۱۳۹۷) و شریفی راد و نیک نفس (۱۳۹۳) بر مطلوبیت استفاده از رویکرد ماشین بردار پشتیبان در دسته بندی داده های نامتوازن ، فرضیه‌های پژوهش به شرح زیر ارائه می گردد.

فرضیه اول: عملکرد رویکرد ماشین بردار پشتیبان در پیش بینی طرح‌های تقلب در صورت‌های مالی در فضای چند کلاسه و بر اساس مجموعه داده نامتوازن، به شکل معنادار بهتر از سایر روش‌های یادگیری ماشین است.

فرضیه دوم: عملکرد رویکرد ماشین بردار پشتیبان در پیش بینی طرح‌های تقلب در صورت‌های مالی در فضای دو کلاسه و بر اساس مجموعه داده متوازن، به شکل معنادار بهتر از سایر روش‌های یادگیری ماشین است.

روش‌شناسی پژوهش

پژوهش حاضر از لحاظ هدف در طبقه پژوهش‌های کاربردی قرار گرفته و برای گردآوری داده‌ها از روش کتابخانه ای و اسنادکاوی استفاده شده است. جامعه آماری تحقیق حاضر شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران هستند که در قلمرو زمانی ۸۷/۰۱/۰۱ الی ۱۲/۲۹/ ۹۹ مورد بررسی قرار می گیرند. نمونه آماری شامل شرکت‌هایی است که الف) از تاریخ ۸۷/۰۱/۰۱ به بعد عضو بورس اوراق بهادار تهران باشند. ب) سال مالی آنها منتهی به ۱۲/۲۹ هر سال باشد. ج) در قلمرو زمانی تحقیق، سال مالی خود را تغییر نداده باشند. د) اطلاعات آنها قابل تهیه و در دسترس باشد. هـ) جزو شرکت‌های واسطه مالی، سرمایه گذاری و بانک‌ها نباشند. به این ترتیب مجموعه ۱۳۴ شرکت و ۱۷۴۲ صورت مالی به عنوان نمونه شناسایی شدند.

متغیر وابسته: متغیر وابسته در تحقیق حاضر، طرح تقلب در صورت‌های مالی می‌باشد. فرقاندوست حقیقی و همکاران (۱۳۹۳) و رضایی و همکاران (۱۴۰۰) معتقدند شرکت‌هایی که گزارش حسابرسی آنها مردود، عدم اظهار نظر یا مشروط بوده‌اند با احتمال بیشتری نسبت به شرکت‌هایی که گزارش مقبول دارند دست به تقلب زده‌اند. خواجوی و ابراهیمی (۱۳۹۶) نیز برای تعیین تقلب در صورت های مالی از تطبیق بندهای گزارش حسابرس و علائم خطر تقلب ذکر شده در استاندارد ۲۴۰ حسابرسی با عنوان "مسئولیت حسابرس در ارتباط با تقلب و اشتباه در صورت‌های مالی" بهره جسته اند. این علائم شامل بیش نمایی

موجودی کالا، بیش نمایی حساب ها و اسناد دریافتنی، بیش نمایی دارایی‌های ثابت، بیش نمایی سرمایه گذاری‌ها، کسری ذخیره مطالبات مشکوک الوصول، کسری استهلاک، بیش نمایی درآمدها، بیش نمایی سود، بیش نمایی سود انباشته، کم نمایی حساب‌ها و اسناد پرداختنی، کسری ذخیره مالیات، بدهی‌های احتمالی، کسری ذخیره مزایای پایان خدمت کارکنان، کم نمایی هزینه‌ها، حساب‌ها و اسناد دریافتنی که مدت زیادی از سررسید آن‌ها گذشته است، موجودی راکد، دارایی راکد، اشتباه در به کارگیری استانداردهای حسابداری مرتبط با اندازه گیری، شناسایی، طبقه بندی، ارائه یا افشا می‌باشد.

در پژوهش حاضر طرح تقلب در صورت های مالی به پیروی از فرقاندوست حقیقی و همکاران (۱۳۹۳)، خواجوی و ابراهیمی (۱۳۹۶) و رضایی و همکاران (۱۴۰۰) با مراجعه به بندهای گزارش حسابرس مستقل تعیین شده است. در این شیوه در دسته بندی اولیه چنانچه هر یک از علائم فوق در بندهای گزارش حسابرس وجود داشته باشد به عنوان گزارش مشکوک به تقلب به هر یک از طرح‌های بیش نمایی دارایی، بیش نمایی درآمد، کم نمایی بدهی، کم نمایی هزینه، بیش نمایی بدهی، بیش نمایی هزینه و افشا نامناسب تخصیص داده شد. مواردی که ابهام در قضاوت و تشخیص طرح تقلب وجود داشت به عنوان طبقه سایر شناسایی شدند. در مرحله دوم صحت طرح‌های تخصیص یافته به صورت‌های مالی مشکوک به تقلب توسط یک حسابدار رسمی و شریک موسسه حسابرسی بررسی شد و اصلاحات لازم انجام پذیرفت.

متغیر وابسته در پژوهش حاضر از نوع متغیر اسمی بوده و برای صورت‌های مالی سالم عدد صفر و برای صورت‌های مالی مشکوک به تقلب اعداد یک الی شش، مطابق با جدول یک تخصیص یافته است.

جدول ۱. طرح‌های تقلب در صورت‌های مالی

ردیف	طرح تقلب	کد	فراوانی	درصد فراوانی
۱	گزارش مالی سالم	۰	۸۶۳	٪۵۰
۲	بیش نمایی دارایی، کم نمایی بدهی و هزینه	۱	۴۰۲	٪۲۳
۳	بیش نمایی دارایی و کم نمایی هزینه	۲	۲۱۰	٪۱۲
۴	کم نمایی هزینه و بدهی	۳	۱۴۰	٪۸
۵	بیش نمایی دارایی و درآمد	۴	۶۹	٪۴
۶	افشا نامناسب	۵	۲۵	٪۱
۷	سایر	۶	۳۳	٪۲
جمع (۱۳۴ شرکت و ۱۳ سال)			۱۷۴۲	٪۱۰۰

منبع: یافته‌های پژوهش

نتایج پژوهش، بلیتس و همکاران (۲۰۱۳) نشان می‌دهد کاربرد نمونه‌های کوچک (۵ الی ۲۵ تایی) در مدل‌های دسته بندی منجر به کسب نتیجه مطلوب نخواهد شد و منحنی یادگیری تحت تاثیر عدم اطمینان ناشی از حجم محدود نمونه قرار می‌گیرد. براین اساس مشاهدات طبقه تقلب از نوع افشا نامناسب از نمونه آماری حذف گردید. افزون بر این با عنایت به ابهام موجود در طبقه سایر، مشاهدات این طبقه نیز از نمونه آماری حذف شد. از این رو تعداد نهایی نمونه برابر با ۱۶۸۴ مشاهده می‌باشد.

متغیر مستقل: به اقتباس از پژوهش‌های پیشین (هیوگو و شنگ لینگ، ۲۰۲۲؛ لونگ جان، ۲۰۲۱؛ کات سیس و همکاران، ۲۰۱۲؛ پرلوس، ۲۰۱۱؛ کی روکس، اسپاتیس و مانولوپولوس، ۲۰۱۱؛ رضایی، ناظمی و صدرآبادی، ۱۴۰۰؛ خواجوی و ابراهیمی، ۱۳۹۶؛ اعتمادی و زلفی، ۱۳۹۲؛ صفرزاده، ۱۳۸۹) و تطبیق معیارها با مصادیق گزارشگری مالی متقلبانه و تاثیر آن بر اقلام صورت‌های مالی، متغیرهای مستقل در این پژوهش به شرح جدول دو تعریف شد. داده مربوط به متغیرهای مستقل از بسته بنیادی نرم افزار رهاورد نوین ۳ و سامانه جامع تحلیل بنیادی بورس ویو استخراج گردید. داده استخراج شده مورد پایش قرار گرفت. در صورت وجود داده‌های از دست رفته، ارزش‌های مزبور با مراجعه به سایت کدال و اطلاعات مالی مندرج در صورت‌های مالی، محاسبه و در مجموعه داده پژوهش وارد گردید.

جدول ۲. تعریف عملیاتی متغیرهای پژوهش

متغیر مستقل	نحوهٔ اندازه‌گیری
حاشیه سود ناخالص GPM	فروش / سود ناخالص
حاشیه سود عملیاتی OPM	فروش / سود عملیاتی
حاشیه سود خالص NPM	فروش / سود خالص
بازده دارایی‌ها ROA	دارایی / سود خالص
بازده حقوق صاحبان سهام ROE	دارایی / حقوق صاحبان سهام
سود قبل از بهره و مالیات به دارایی EBIT/A	دارایی / سود قبل از بهره و مالیات
گردش دارایی‌ها ASTRN	دارایی / فروش خالص
گردش دارایی ثابت FXASTRN	دارایی ثابت / فروش خالص
گردش موجودی کالا INVTRN	متوسط موجودی کالا / بهای تمام شده کالای فروش رفته
گردش حساب‌های دریافتنی ACCRECTRN	متوسط حساب های دریافتنی / فروش خالص
جمع بدهی به جمع دارایی‌ها LIB/ASST	جمع دارایی ها / جمع بدهی ها
نسبت بدهی به حقوق صاحبان سهام LIB/EQT	حقوق صاحبان سهام / جمع بدهی ها

متغیر مستقل	نحوهٔ اندازه‌گیری		
نسبت سود انباشته به دارایی‌ها ACUM/ASST	دارایی / سود یا زیان انباشته		
نسبت سود انباشته به سرمایه ACUM/EQT	سرمایه سهام عادی / سود یا زیان انباشته		
نسبت پوشش هزینه بهره INTCOV	سود قبل از بهره و مالیات / هزینه بهره		
نسبت جاری CURRAT	بدهی جاری / دارایی جاری		
نسبت آنی QUICRAT	بدهی جاری/(سرمایه گذاری کوتاه مدت+ حساب دریافتنی+ وجه‌نقد)		
نسبت وجه نقد آزاد به درآمد FCF/REV	فروش خالص / وجه نقد آزاد		
	وجه نقد آزاد = سود خالص + هزینه استهلاک + هزینه مالی (مالیات – ۱) – سرمایه گذاری در دارایی ثابت – سرمایه گذاری در سرمایه در گردش		
نسبت مخارج سرمایه‌ای به درآمد CE/REV	فروش / مخارج سرمایه ای	مخارج سرمایه‌ای = تفاوت خالص ارزش دفتری دارایی‌های ثابت در ابتدا و پایان دوره به علاوه هزینه استهلاک	
نسبت مخارج سرمایه‌ای به سود خالص CE/NP	سود خالص / مخارج سرمایه‌ای		
نسبت وجه نقد عملیاتی به درآمد OCF/REV	فروش خالص / خالص جریان وجوه نقد حاصل از فعالیت های عملیاتی		
نسبت وجه نقد عملیاتی به بدهی OCF/LIB	جمع بدهی‌ها / خالص جریان وجوه نقد حاصل از فعالیت های عملیاتی		
ارزش بنگاه EV	(نقد و معادل نقد– ارزش دفتری بدهی‌ها+ ارزش بازار سهام) ln		
ارزش دفتری BV	(ارزش دفتری حقوق صاحبان سهام) ln		

پس از استخراج داده‌ها، به منظور اطمینان از اهمیت متغیرهای تعریف شده در حل مسئله پژوهش، از مدل XGBoost استفاده گردید. مطابق با جدول سه، نتایج حاصل از اجرای مدل XGBoost حاکی از اهمیت همه متغیرها در پیش بینی طرح تقلب می‌باشد. به عبارت دیگر حضور همه متغیرها برای پیش بینی طرح تقلب لازم و کمک کننده است.

جدول ۳. اهمیت متغیرهای پژوهش در پیش بینی طرح تقلب

متغیر	GPM	OPM	NPM	ROA	ROE	EBIT/A
اهمیت	٪۴/۲	٪۴	٪۴/۱	٪۵/۳	٪۳/۶	٪۷/۲
متغیر	ASTRN	FXASTRN	INVTRN	ACCRECTRN	LIB/ASST	LIB/EQT
اهمیت	٪۴/۱	٪۳/۶	٪۳/۵	٪۵/۱	٪۴/۷	٪۳/۸
متغیر	ACUM/ASST	ACUM/EQT	INTCOV	CURRAT	QUICRAT	FCF/REV
اهمیت	٪۴/۲	٪۴/۶	٪۴/۲	٪۳/۳	٪۳/۳	٪۳/۹
متغیر	CE/REV	CE/NP	OCF/REV	OCF/LIB	EV	BV
اهمیت	٪۳/۳	٪۳/۷	٪۳/۹	٪۴/۲	٪۴/۳	٪۳/۹

منبع: یافته‌های پژوهش

به منظور کاهش ابعاد مساله از روش PCA استفاده شد. در مرحله پیش آزمون کاهش ابعاد مساله منجر به کاهش عملکرد الگوهای یادگیری ماشین در پیش بینی طرح تقلب گردید. از این رو همه متغیرهای مستقل در فضای مسئله به عنوان ویژگی تعریف شد و حفظ گردید.عمل نرمال سازی به روی داده‌ها اعمال شد. داده‌ها برای استفاده در الگوهای یادگیری ماشین به دو دسته آموزش و آزمون تقسیم شدند. الگوهای رگرسیون لجستیک، درخت تصمیم، گرادیان تقویتی و ماشین بردار پشتیبان با رویکرد چندکلاسه در محیط پایتون برنامه نویسی شدند. برای دستیابی به ترکیب بهینه از پارامترها در هر الگو از الگوریتم Grid Search استفاده شد.

رویکرد چند کلاسه: برای حل بسیاری از مسائلی که در عمل با آنها مواجه‌ایم ناچاریم از روش‌های کلاس بندی چند کلاسه استفاده کنیم. به منظور بسط کاربرد الگوهای مزبور در حل مسئله پژوهش، از رویکرد دسته بندی چندکلاسه استفاده شد. برای این منظور از رویکرد توصیه شده توسط کتابخانه متن باز Scikit-Learn طراحی شده توسط پدرگوسا و همکاران (۲۰۱۱) استفاده گردید. این کتابخانه با ساختن n عدد مدل مستقل برای هر کدام از طبقات، مسائل دسته بندی را انجام می‌دهد. به دیگر سخن برای هر دسته یک مدل جداگانه ساخته و سپس آن‌ها را به صورت مستقل پیش‌بینی می کند. پیرو رویکرد پیشنهادی در پژوهش شعبانی و علوی (۱۳۹۲) برای توسعه کاربرد روش ماشین بردار پشتیبان در دسته بندی چند کلاسه از توابع کرنل و تولید n-1 ماشین بردار پشتیبان برای n کلاس استفاده گردید.

معیار ارزیابی عملکرد مدل‌های دسته بندی: در یک فضای دو دسته‌ای، بر اساس اطلاعات جدول پریشانی معیارهای نرخ دقت طبقه بندی، فراخوانی، صحت f1 و AUC محاسبه و در ارزیابی عملکرد مدل‌ها مورد استفاده قرار می گیرد. نرخ دقت طبقه بندی عبارت است از نسبت تعداد صورت های مالی سالم و متقلبی که سلامت یا تقلب آن‌ها به درستی تشخیص داده شده است. فراخوانی به معنای نسبتی از گزارش‌های متقلبانه است که مدل به درستی آن‌ها را متقلبانه شناسایی کرده است. اگر این نرخ پایین باشد، یعنی گزارش متقلبانه به اشتباه، سالم طبقه بندی شده است. صحت، نسبت تعداد گزارش متقلبانه‌ای که به درستی پیش بینی شده است به تعداد کل گزارشات متقلبانه‌ای که به درستی یا به اشتباه توسط مدل شناسایی شده‌اند. اگر نرخ صحت پایین باشد، یعنی گزارشات سالم را به اشتباه، متقلبانه تشخیص داده‌ایم. معیار f1 ترکیبی از معیار

فراخوانی و صحت است و هرچه مقدار آن به یک نزدیک‌تر باشد بهتر است. AUC نشان دهنده سطح زیر نمودار ROC است. هرچه مقدار این عدد به یک نزدیک‌تر باشد بیانگر عملکرد بهتر مدل دسته بندی است.

گراندینی، باگلی و ویسانی (۲۰۲۰) معتقد هستند روش‌های ارزیابی الگوهای دسته بندی چند کلاسه نسبت به الگوهای معمول دو کلاسه متفاوت خواهد بود. ایشان دقت طبقه بندی را به عنوان یکی از شاخص‌های ارزیابی عملکرد دسته بندی چند کلاسه معرفی می‌کنند. برای محاسبه معیار فراخوانی، دقت و f1 در فضای چند کلاسه دو رویکرد ماکرو و میکرو معرفی شده است. در رویکرد ماکرو میانگین معیارهای ارزیابی در سطح دسته‌ها محاسبه می‌شود و در رویکرد میکرو میانگین معیار عملکرد در سطح کل ارائه می‌شود. افزون بر این ضریب کاپای کوهن به عنوان یک معیار، توافق میان دو ارزیاب (شواهد واقعی و پیش بینی) را نشان می‌دهد که برای ارزیابی عملکرد الگوهای چند کلاسه و مجموعه داده نامتوازن مناسب است.

در فرضیه اول، مجموعه داده شامل ۲۴ ویژگی (متغیر مستقل) به طور همزمان در طبقات صورت‌های مالی سالم با کد ۰، مشکوک به تقلب از نوع بیش نمایی دارایی، کم نمایی بدهی و هزینه با کد ۱، بیش نمایی دارایی و کم نمایی هزینه با کد ۲، کم نمایی هزینه و بدهی با کد ۳ و بیش نمایی دارایی و درآمد با کد ۴ تعریف شد. الگوهای رگرسیون لجستیک، درخت تصمیم، گرادیان تقویتی و ماشین بردار پشتیبان بر مجموعه داده اجرا گردید. معیارهای عملکرد محاسبه و با استفاده از آزمون فریدمن مقایسه شد. در فرضیه دوم با استفاده از روش نمونه گیری، چهار مجموعه داده متوازن شامل مشاهدات مربوط به صورت‌های مالی سالم و صورت‌های مالی مشکوک به طرح تقلب مربوطه تعریف گردید. الگوهای رگرسیون لجستیک، درخت تصمیم، گرادیان تقویتی و ماشین بردار پشتیبان در هر چهار مجموعه داده به طور جداگانه اجرا شد. معیارهای عملکرد محاسبه و با استفاده از آزمون فریدمن مقایسه شد.

یافته‌های پژوهش

آمارتوصیفی متغیرهای مستقل مورد استفاده در پژوهش در جدول چهار ارائه شده است.

جدول ۴.آمار توصیفی متغیرهای پژوهش

متغیر	میانگین	میانه	کمینه	بیشینه	انحراف معیار
GPM	۰/۲۶	۰/۲۴	-۰/۳۴	۱	۰/۱۷
OPM	۰/۱۸	۰/۱۵	-۱/۳۹	۱/۵۳	۰/۲
NPM	۰/۱۷	۰/۱۲	-۱/۳۹	۷/۸	۰/۳۱
ROA	۰/۱۳	۰/۱	-۰/۳	۰/۶۸	۰/۱۳
ROE	۰/۱۷	۰/۲۷	-۷۲/۷	۲/۱۵	۲/۲۲
EBIT/A	۰/۱۸	۰/۱۵	-۰/۲۵	۰/۷	۰/۱۴
ASTRN	۰/۹۱	۰/۷۹	۰	۶/۰۵	۰/۵۸
FXASTRN	۶/۰۳	۳/۹۹	۰	۸۷/۹۶	۶/۸۱
INVTRN	۶/۸۶	۲/۵۷	۰	۱۰۰۳/۳۵	۳۷/۹۲
ACCRECTRN	۸۵/۲۸۹	۴/۷۹	۰/۲۱۸	۳۰۸۸۲/۹۹	۱۱۹۲/۷۴
LIB/ASST	۰/۵۸	۰/۶	۰/۰۳	۱	۰/۱۸
LIB/EQT	۲/۷۶	۱/۴۷	۰/۰۳	۳۰۳/۸۲	۱۲/۰۷
ACUM/ASST	۰/۱۵	۰/۱۴	-۱/۰۸	۰/۸۴	۰/۱۹
ACUM/EQT	۰/۰۵	۰/۳۷	-۱۴۰/۶۷	۱/۰۸	۴/۵۱
INTCOV	۵۳/۵۲	۴/۷۸	-۱۹۰۰/۴۳	۶۵۹۹/۹	۳۵۰/۴۴
CURRAT	۱/۵	۱/۲۹	۰/۲۲	۲۷/۱	۱/۲۱
QUICRAT	۰/۸۷	۰/۷۵	۰/۰۴	۲۶/۳۱	۱/۰۱
FCF/REV	۰/۲۵۳	۰/۰۶	-۳/۲۹۳	۲۸۲/۳۷	۶/۸۸۴
CE/REV	۰/۱۶۷	۰/۰۲۵	-۲/۶۸۵	۱۸۲/۲۵۲	۴/۴۴
CE/NP	۰/۶۶۱	۰/۱۷۸	-۷۱/۳۱۷	۱۲۲/۹۷۸	۵/۶۷۱
OCF/REV	۰/۰۲۹	۰/۱۰۳	-۱۴۹/۶۱	۲/۰۹	۳/۶۵۶
OCF/LIB	۰/۱۷۲	۰/۱۴۵	-۴۴/۲۷۴	۲/۵۶۳	۱/۱۴
EV	۷/۵۴	۷/۳	-۰/۴۸۷	۱۴/۷۸۵	۱/۷۷
BV	۱۳/۲۳	۱۳/۰۶	۶/۷۴	۱۹/۹۷	۱/۶۶

منبع: یافته‌های پژوهش

آزمون فرضیه اول: هر یک از الگوهای رگرسیون لجستیک، درخت تصمیم، گرادیان تقویتی و ماشین بردار پشتیبان با هدف پیش‌بینی طرح‌های تقلب در صورت‌های مالی با رویکرد چند کلاسه در محیط پایتون پیاده و در مجموعه داده نامتوازن شامل دسته صورت‌های مالی سالم و چهار دسته مشکوک به طرح تقلب اجرا شد. جدول پنج نتایج حاصل از اجرای روش‌های مزبور در حل مسئله پیش بینی طرح گزارشگری مالی متقلبانه را نشان می‌دهد.

جدول ۵. معیارهای ارزیابی عملکرد الگوهای یادگیری ماشین با رویکرد چندکلاسه

	دقت طبقه بندی	نرخ فراخوانی		صحت		معیار f1		معیار کاپا
		Macro	micro	macro	Micro	macro	micro	
رگرسیون لجستیک	۰/۴۹۵۵	۰/۲۱۰۵	۰/۴۹۵۵	۰/۱۵۹۱	۰/۴۹۵۵	۰/۱۷۱۳	۰/۴۹۵۵	۰/۰۶۴۱
درخت تصمیم	۰/۴۵۷۰	۰/۳۰۰۹	۰/۴۵۷۰	۰/۳۰۹۸	۰/۴۵۷۰	۰/۳۰۲۹	۰/۴۵۷۰	۰/۱۹۹۷
گرادیان تقویتی	۰/۵۲۵۲	۰/۲۷۵۳	۰/۵۲۵۲	۰/۲۷۴۴	۰/۵۲۵۲	۰/۲۶۷۵	۰/۵۲۵۲	۰/۲۰۷۱
ماشین بردار	۰/۵۴۰۱	۰/۲۸۱۲	۰/۵۴۰۱	۰/۳۱۷۰	۰/۵۴۰۱	۰/۲۷۸۸	۰/۵۴۰۱	۰/۲۲۸۴

منبع: یافته‌های پژوهش

به منظور مقایسه عملکرد روش‌های مزبور، با عنایت به عدم تایید نرمال بودن توزیع معیارهای عملکرد با استفاده از آزمون کولموگروف اسمیرنوف در سطح معناداری ۵ درصد، از رویکرد تحلیل واریانس دوطرفه فریدمن استفاده شد. آماره آزمون ۱۱/۵۵ با سطح معناداری ۰/۰۰ موید تفاوت معنادار در عملکرد الگوهای یادگیری ماشین می‌باشد. جدول شش نتایج حاصل از مقایسه عملکرد زوجی روش‌ها با استفاده از آزمون فریدمن را ارائه می‌دهد.

جدول ۶. مقایسه زوجی حاصل از آزمون فریدمن - عملکرد الگوها در فضای چندکلاسه

	آماره آزمون	سطح معناداری		آماره آزمون	سطح معناداری
رگرسیون لجستیک – درخت تصمیم	-۰/۶۲	۰/۳۳	درخت تصمیم – گرادیان تقویتی	-۰/۱۲	۰/۸۵
رگرسیون لجستیک -گرادیان تقویتی	-۰/۷۵	۰/۲۴	درخت تصمیم – ماشین بردار پشتیبان	-۱/۵۰	۰/۰۲
رگرسیون لجستیک – ماشین بردار پشتیبان	-۲/۱۲	۰/۰۰	گرادیان تقویتی- ماشین بردار پشتیبان	-۱/۳۷	۰/۰۳

منبع: یافته‌های پژوهش

جدول شش نشان می‌دهد تفاوت معنادار در عملکرد روش رگرسیون لجستیک، درخت تصمیم و گرادیان تقویتی وجود ندارد. آماره آزمون و سطح معناداری زیر ۵ درصد موید برتری عملکرد رویکرد ماشین بردار پشتیبان نسبت به سایر روش‌هاست.

آزمون فرضیه دوم: مسئله پژوهش در فضای دو کلاسه متوازن با استفاده از راهبرد نمونه گیری تعریف شد. برای انتخاب نمونه، از الگوریتم نمونه گیری Kennard-Stone استفاده شد. به این ترتیب چهار مجموعه داده مجزا (برای هر طرح تقلب یک مجموعه داده) تعریف شد. در هر مجموعه تعداد شرکت‌های سالم برابر با تعداد شرکت‌های مشکوک به طرح تقلب انتخاب

گردید. برای هر طرح تقلب، الگوی رگرسیون لجستیک، درخت تصمیم، گرادیان تقویتی و ماشین بردار اجرا شد. برای مقایسه عملکرد الگوها در پیش بینی صورت‌های مالی مشکوک به هر طرح تقلب با توجه به عدم تایید نرمال بودن توزیع معیارهای عملکرد با استفاده از آزمون کولموگروف اسمیرنوف، از روش تحلیل واریانس دوطرفه فریدمن استفاده شد.

جدول هفت نتایج اجرای روش‌های مزبور برای پیش بینی صورت‌های مالی مشکوک به بیش نمایی دارایی، کم نمایی بدهی و هزینه و آزمون تحلیل واریانس دوطرفه فریدمن را ارائه می‌دهد. نتایج آزمون از تفاوت معنادار در عملکرد الگوهای یادگیری ماشین پشتیبانی نمی‌کند.

جدول ۷. عملکرد الگوی یادگیری ماشین در پیش‌بینی طرح بیش نمایی دارایی، کم نمایی بدهی و هزینه

	الگو	دقت طبقه بندی	نرخ فراخوانی	صحت	F1	AUC
عملکرد الگوهای یادگیری ماشین	رگرسیون لجستیک	۰/۵۵۵	۰/۹۴۳	۰/۵۳۲	۰/۶۸۰	۰/۷۸۱
	درخت تصمیم	۰/۶۱۱	۰/۲۳۱	۰/۹۶۰	۰/۳۷۰	۰/۶۱۱
	گرادیان تقویتی	۰/۶۲۷	۰/۸۱۹	۰/۵۹۵	۰/۶۸۸	۰/۷۰۸
	ماشین بردار پشتیبان	۰/۶۱۰	۰/۷۹۹	۰/۵۸۲	۰/۶۷۱	۰/۷۱۱
نتایج آزمون تحلیل واریانس دوطرفه فریدمن		آماره آزمون	۲/۵۲	سطح معناداری		۰/۴۷

منبع: یافته‌های پژوهش

جدول هشت نتایج اجرای الگوی یادگیری ماشین در پیش بینی صورت‌های مالی مشکوک به بیش نمایی دارایی، کم نمایی هزینه و نتایج آزمون تحلیل واریانس دوطرفه فریدمن را ارائه می‌دهد. نتایج آزمون از تفاوت معنادار در عملکرد الگوهای یادگیری ماشین پشتیبانی نمی‌کند.

جدول ۸ عملکرد الگوی یادگیری ماشین در پیش‌بینی طرح بیش نمایی دارایی وکم نمایی هزینه

	الگو	**دقت طبقه بندی**	**نرخ فراخوانی**	**صحت**	**F1**	**AUC**
عملکرد الگوهای یادگیری ماشین	رگرسیون لجستیک	۰/۵۲۷	۱/۰۰۰	۰/۵۱۵	۰/۶۷۹	۰/۷۸۴
	درخت تصمیم	۰/۵۰۰	۱/۰۰۰	۰/۵۰۰	۰/۶۶۷	۰/۶۸۳
	گرادیان تقویتی	۰/۵۸۹	۰/۹۶۵	۰/۵۵۱	۰/۷۰۱	۰/۷۴۴
	ماشین بردار پشتیبان	۰/۷۲۳	۰/۷۱۵	۰/۷۳۰	۰/۷۲۰	۰/۷۷۵
نتایج آزمون تحلیل واریانس دوطرفه فریدمن		آماره آزمون	۴/۷۱	سطح معناداری		۰/۱۹

منبع: یافته‌های پژوهش

جدول نه نتایج اجرای الگوی یادگیری ماشین در پیش بینی صورت‌های مالی مشکوک به کم نمایی هزینه و بدهی و نتایج آزمون تحلیل واریانس دوطرفه فریدمن را ارائه می‌دهد. نتایج آزمون از تفاوت معنادار در عملکرد الگوهای یادگیری ماشین پشتیبانی نمی‌کند.

جدول ۹. عملکرد الگوی یادگیری ماشین در پیش‌بینی طرح کم نمایی هزینه و بدهی

	الگو	دقت طبقه بندی	نرخ فراخوانی	صحت	F1	AUC
عملکرد الگوهای یادگیری ماشین	رگرسیون لجستیک	۰/۵۰۰	۱/۰۰۰	۰/۵۰۰	۰/۶۶۶	۰/۶۴۵
	درخت تصمیم	۰/۵۹۲	۰/۶۲۹	۰/۵۹۰	۰/۶۰۶	۰/۶۲۰
	گرادیان تقویتی	۰/۵۴۰	۰/۹۲۱	۰/۵۲۴	۰/۶۶۶	۰/۶۸۸
	ماشین بردار پشتیبان	۰/۷۹۰	۰/۷۸۶	۰/۷۸۷	۰/۷۸۵	۰/۸۰۸
نتایج آزمون تحلیل واریانس دوطرفه فریدمن		آماره آزمون	۵/۶۹	سطح معناداری		۰/۱۲

منبع: یافته‌های پژوهش

جدول ده نتایج اجرای الگوی یادگیری ماشین در پیش بینی صورت‌های مالی مشکوک به بیش نمایی دارایی و درآمد و نتایج آزمون تحلیل واریانس دوطرفه فریدمن را ارائه می‌دهد. نتایج آزمون موید تفاوت معنادار در عملکرد الگوهای یادگیری ماشین است. از این رو مقایسه زوجی عملکرد الگوهای یادگیری ماشین مورد بررسی قرار گرفت.

جدول ۱۰. عملکرد الگوی یادگیری ماشین در پیش‌بینی طرح بیش نمایی دارایی و درآمد

	الگو	دقت طبقه بندی	نرخ فراخوانی	صحت	F1	AUC
عملکرد الگوهای یادگیری ماشین	رگرسیون لجستیک	۰/۴۴۳	۰/۴۳۳	۰/۴۳۳	۰/۴۲۲	۰/۳۴۴
	درخت تصمیم	۰/۶۱۱	۰/۹۰۰	۰/۵۸۹	۰/۷۰۸	۰/۵۸۸
	گرادیان تقویتی	۰/۶۳۹	۰/۹۰۰	۰/۶۰۸	۰/۷۲۲	۰/۵۶۹
	ماشین بردار پشتیبان	۰/۸۳۶	۱/۰۰۰	۰/۷۶۳	۰/۸۶۵	۰/۷۹۲
نتایج آزمون تحلیل واریانس دوطرفه فریدمن		آماره آزمون	۱۴/۰۲	سطح معناداری		۰/۰۰

مقایسه زوجی عملکرد الگوهای یادگیری ماشین در پیش‌بینی طرح بیش نمایی دارایی و درآمد

	آماره آزمون	سطح معناداری		آماره آزمون	سطح معناداری
رگرسیون لجستیک – درخت تصمیم	-۱/۳۰	۰/۱۱۱	درخت تصمیم – گرادیان تقویتی	-۰/۴۰	۰/۶۲۴
رگرسیون لجستیک -گرادیان تقویتی	-۱/۷۰	۰/۰۳۷	درخت تصمیم – ماشین بردار پشتیبان	-۱/۷۰	۰/۰۳۷
رگرسیون لجستیک – ماشین بردار پشتیبان	-۳/۰۰	۰/۰۰۰	گرادیان تقویتی- ماشین بردار پشتیبان	-۱/۳۰	۰/۱۱۱

منبع: یافته‌های پژوهش

نتایج ارائه شده در جدول ده نشان می‌دهد به شکل معنادار عملکرد رویکرد ماشین بردار پشتیبان بهتر از روش‌های رگرسیون لجستیک و درخت تصمیم است.

نتیجه‌گیری و بحث

عمده مسائل دسته بندی در دنیای واقعی اعم از مسائل شناسایی بیماری، ناهنجاری‌ها، تقلب و کلاهبرداری در فضای مجموعه داده های چندکلاسه و ناموزن مطرح می‌شوند. نحوه برخورد با این موارد به عنوان یک مسئله چالش برانگیز در داده کاوی شناخته می شود. پژوهش حاضر سعی دارد طرح تقلب مورد استفاده در گزارشگری مالی را در فضای چند کلاسه با استفاده از مجموعه داده نامتوازن پیش بینی نماید.

فرضیه اول عملکرد الگوهای رگرسیون لجستیک، درخت تصمیم، گرادیان تقویتی و ماشین بردار پشتیبان در پیش بینی طرح تقلب در فضای چند کلاسه را مورد بررسی قرار داده است. نتایج حاصل از آزمون فرضیه موید برتری عملکرد الگوی ماشین بردار پشتیبان، نسبت به سایر روش‌ها است. این نتیجه مطابق با مبانی نظری و نتیجه پژوهش عمادالدین و همکاران (۱۳۹۷) و شریفی راد و همکاران (۱۳۹۳) و همسو با نتایج تحقیق چن و همکاران (۲۰۱۸) و پرلوس (۲۰۱۱) می‌باشد.

فرضیه دوم عملکرد الگوهای یادگیری ماشین در پیش بینی نوع تقلب در صورت‌های مالی را با تقلیل فضای مسئله به دسته بندی دوکلاسه و متوازن، بررسی می کند. نتایج نشان می‌دهد تفاوت معنادار در عملکرد الگوهای یادگیری ماشین در تشخیص گزارش‌های مالی مشکوک به "بیش نمایی دارایی، کم نمایی بدهی و هزینه"، "بیش نمایی دارایی و کم نمایی هزینه" و "کم نمایی هزینه و بدهی" وجود ندارد. با این حال، عملکرد ماشین بردار پشتیبان بر عملکرد روش رگرسیون لجستیک و درخت تصمیم در پیش بینی گزارش‌های مالی مشکوک به "بیش نمایی دارایی و درآمد" ارجح است.

بر اساس نتایج پژوهش پیشنهاد می‌گردد کارشناس رسمی دادگستری و نهادهای نظارتی در توسعه ابزارهای کاربردی جهت پیش بینی طرح تقلب در گزارشگری مالی در فضای چند کلاسه و همچنین در شناسایی گزارش‌های مالی مشکوک به "بیش نمایی دارایی و درآمد" در فضای دو کلاسه از الگوی ماشین بردار پشتیبان استفاده نمایند. افزون بر این رویکرد مزبور می-تواند به عنوان ابزار کمکی در تحلیل علایم خطر تقلب و پیش بینی تحریف صورت‌های مالی مورد استفاده حسابرسان قرار گیرد. در تفسیر نتایج پژوهش باید به این نکته توجه داشت که با عنایت به تعریف عملیاتی طرح تقلب در پژوهش حاضر با رجوع به بند گزارش حسابرس و عدم تایید قصد تقلب در مرجع قانونی صالحه، ممکن است با توجه به اختیارات مدیریت در انتخاب

رویه حسابداری موارد مربوط به مدیریت و هموارسازی سود به عنوان صورت‌های مالی مشکوک به تقلب دسته بندی گردد.

پیشنهاد می گردد در پژوهش‌های آتی علاوه بر نسبت‌های مالی از معیارهای غیر مالی در پیش بینی طرح تقلب در صورت‌های مالی استفاده گردد. افزون بر این عملکرد سایر روش‌های داده کاوی مانند الگوریتم‌های فراابتکاری و زیستی در پیش بینی طرح تقلب مورد بررسی قرار گیرد.

منابع

اعتمادی، حسین؛ زلفی، حسن، (۱۳۹۲)، کاربرد رگرسیون لجستیک درشناسایی گزارشگری مالی متقلبانه، *فصلنامه دانش حسابرسی*، ۱۳(۵۱): ۱۴۵-۱۶۳.

آگراوال، چارو، (۱۳۹۸)، *متن کاوی به کمک یادگیری ماشین*، مهدی اسماعیلی، تهران: آتی نگر.

تاراسی، مجتبی؛ بنی طالبی دهکردی، بهاره؛ زمانی، بهزاد، (۱۳۹۸)، پیش بینی گزارشگری مالی متقلبانه از طریق شبکه عصبی مصنوعی، *حسابداری مدیریت*، ۱۲(۴۰): ۶۳-۷۹.

خواجوی، شکرالله؛ ابراهیمی،مهرداد، (۱۳۹۶)،ارائۀ یک رویکرد محاسباتی نوین برای پیش‌بینی تقلب در صورت‌های مالی با استفاده از شیوه‌های خوشه‌بندی و طبقه‌بندی (شواهدی از شرکت‌های پذیرفته‌شده بورس اوراق بهادار تهران)، *پیشرفت‌های حسابداری*، ۹(۲): ۱-۳۴.

رامنی، مارشال؛ استین بارت، پل، (۱۳۸۷)، *سیستم‌های اطلاعاتی حسابداری*، سید حسین سجادی و سید محسن طباطبایی نژاد، اهواز: انتشارات دانشگاه شهید چمران اهواز.

رضائی، مهدی؛ ناظمی اردکانی، مهدی؛ ناصر صدرآبادی، علیرضا، (۱۴۰۰)، پیش بینی تقلب صورت‌های مالی با استفاده از رویکرد کریسپ (CRISP)، *دانش حسابداری و حسابرسی مدیریت*، ۱۰(۴۰): ۱۳۵-۱۵۰.

سجادی، سید حسین؛ کاظمی، توحید، (۱۳۹۵)، الگوی جامع گزارشگری مالی متقلبانه در ایران به روش نظریه پردازی زمینه بنیان، *پژوهش‌های تجربی حسابداری*، ۶(۲۱): ۱۸۵-۲۰۴.

شریفی راد، سمیه؛ نیک نفس، علی اکبر، (۱۳۹۳)، بررسی توابع کرنل الگوریتم SVM در دقت کلاس بندی داده های نامتوازن در بازه‌های مختلف نرخ عدم توازن، *همایش ملی الکترونیکی دستاوردهای نوین در علوم مهندسی و پایه*، اردبیل.

شعبانی، علی؛ علوی، سید محمد،(۱۳۹۲)، ارائه‌روشی برای کلاس‌بندی اهداف دریایی سوناری‌با استفاده‌از الگوریتم‌های چندکلاسه‌ماشین‌بردار پشتیبان، *فصلنامه‌صنایع الکترونیک*، ۴(۱۳): ۱۲-۱۹.

صفرزاده، محمد حسین، (۱۳۸۹)، توانایی نسبتهای مالی در کشف تقلب در گزارشگری مالی تحلیل لاجیت، مجله دانش حسابداری، ۱(۱): ۱۳۷-۱۶۳.

صنیعی آباده،محمد؛ محمودی، سینا؛ طاهرپور، محدثه(۱۳۹۳)، ***داده‌کاوی کاربردی***، تهران: نیاز دانش.

عمادالدین، مریم؛ بدیع، نسرین؛ خفاجه، حمید (۱۳۹۷)، طبقه بندی داده های نامتوازن توسط الگوریتم ماشین بردار پشتیبانی،*کنفرانس بین المللی تحقیقات بین رشته ای در مهندسی برق، کامپیوتر، مکانیک و مکاترونیک در ایران و جهان اسلام*، کرج.

فرقاندوست حقیقی، کامبیز؛ هاشمی، عباس؛ فروغی دهکردی، امین، (۱۳۹۳)، مطالعه رابطه مدیریت سود و امکان تقلب در صورت‌های مالی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران، *دانش حسابرسی*، ۵۶(۱۴): ۴۷-۶۸.

کلهر، جان، تیرنی، برندن، (۱۴۰۰)، *علم داده*، امیر رضا تجلی، امیر محمد رمدانی و امیر علی رمدانی، تهران: شرکت چاپ و نشر بازرگانی.

مرادی، مهدی؛ سلیمانی مارشک، مجتبی؛ باقری، مصطفی،(۱۳۹۴)، بررسی عوامل موثر بر به هنگامی گزارشگری مالی با استفاده از تکنیک‌های شبکه‌های عصبی مصنوعی و درخت تصمیم، *پژوهش های تجربی حسابداری*، ۵ (۱۷): ۱۱۹-۱۳۷.

ملکی کاکلر، حسن؛ بحری ثالث، جمال؛ جبارزاده کنگرلویی، سعید؛ آشتاب، علی، (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.

ویسی، هادی؛ قایدشرف، حمیدرضا؛ ابراهیمی، مرتضی،(۱۴۰۰)، بهبود کارایی‌الگوریتم‌های یادگیری ماشین‌در تشخیص‌بیماری‌های قلبی‌با بهینه‌سازی داده‌ها و ویژگی‌ها، *محاسبات نرم*، ۸(۱۵): ۷۰-۸۵.

هان، ژیاوی، کامبر، پی، میشلین ژان، (۱۳۹۳)، *داده کاوی*، نسترن حاجی حیدری و سیدبهنام خاکباز، تهران: دانشگاه تهران.

References

Aggarwal, ch. (2018**).** *Machine Learning for Text*. Tehran, Ati Negar. (In Persian).

Assciation of Certified Fraud Examiners. (2022). REPORT TO THE NATIONS ON OCCUPATIONAL FRAUD AND ABUSE. https://legacy.acfe.com/report-to-the-nations/2022.

Beasley, M and et al (2010), Fraudulent Financial Reporting 1998 – 2007, COSO.

Chen, J. Liou, W. Chen, W. (2018), Fraud Detection for Financial Statements of Business Groups. *International Journal of Accountion Information Systems*. 7(15): 10-26.

Beleites, Claudia. Ute, Neugebauer. Thomas, Bocklitz. Christoph, Krafft. Jürgen, Popp. (2013). Sample size planning for classification models. *Analytica Chimica Acta*. 760: 25-33.

Craja, p. Kim, A. Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*. 139.

Emaddin, M, Badieh, N, Khafajeh, H. (2017). Classification of unbalanced data by support vector machine algorithm, International conference of interdisciplinary research in electrical, computer, mechanical and mechatronic engineering in Iran and the Islamic world, Karaj. (In Persian).

Etemadi, H. & Zolfi, H. (2014). Application of Logistic Regression in Identifying Fraudulent Financial Reporting, Audit Knowledge, 13 (51): 145-163. (In Persian).

Farqandoost Haghighi, K, Hashemi, A, Foroghi Dehkordi, A. (2013). Study of relationship between profit management and the possibility of fraud in the financial

statements of companies admitted to the Tehran Stock Exchange, Auditing Knowledge. 14(56): 47- 68. (In Persian).

Grandini, M. Bagli, E. Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. White Paper.https://doi.org/10.48550/arXiv.2008.05756.

Han,J. Kamber,M. pei,J.(2015). Data Mining. Tehran. Tehran University. (In Persian).

Jan, Ch. (2018). An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan. *Sustainability* 10(2): 513.

Jan, Ch. (2021). Detection of Financial Statement Fraud Using Deep Learning for Sustainable Development of Capital Markets under Information Asymmetry. *Sustainability* 13(17): 9879.

Kanapickienė, R and Grundienė, Z .(2015). The Model of Fraud Detection in Financial Statements by Means of Financial Ratios , *Social and Behavioral Sciences*, 213:321-327.

Katsis, D. Christos & et al. (2012). Using Ants to Detect Fraudulent Financial Statements. *Journal of Applied Finance & Banking*, 2 (6): 73-81.

Kelleher,J.(2020*), Data Science*. Tehran. Business Publishing Company. (In Persian)

Khajavi, S., Ebrahimi, M. (2017). A Novel Computational Approach to Predict Financial Statements Fraud using Clustering and Classification Techniques: Evidence from Listed Companies in Tehran Stock Exchange. *Journal of Accounting Advances*, 9(2), 1-34. (In Persian)

Kirkos, S., Spathis, Ch., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Journal of Expert Systems with Applications*, 32:995–1003.

Kranacher,M. Riley,R. Wells, J. (2011). *Forensic Accounting and Fraud Examination*. New York: John Willy and Sons.

Lin, C., Chiu, A. & et al . (2015). Detecting the financial statement fraud. *Journal of Knowledge-Based Systems*. 89 (C): 459-470.

Maleki Kakler, H. Bahri Tahal, J, Jabarzadeh Kangarloui, S, Ashtab, A, (2020), The effectiveness of statistical models and machine learning patterns in predicting fraudulent financial reporting, *Financial Economics* (Financial Economics and Development), 15, 54 , 267-292. (In Persian).

Moradi, M., soleymani mareshk, M., Bagheri, M. (2015). Factors Effective on Timeliness of Financial Reporting: Using Synthetic Neural Networks and Decision Trees Techniques. *Empirical Research in Accounting*, 5(3), 119-137. doi: 10.22051/jera.2015.640. (In Persian).

Normah. O. Zulaikha, Amirah., J. Malcolm, S., (2017), Predicting Fraudulent Financial Reporting Using Artificial Neural Network. *Journal of Financial Crime*, 24 (2): 362-387.

Omidi, M, Qingfei, M, Moradinaftchali, V, Piri, M. (2019). The Efficacy of Predictive Methods in Financial Statement Fraud. *Discrete Dynamics in Nature and Society*. https://doi.org/10.1155/2019/4989140

Pedregosa, F. & et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 12: 2825-2830.

Perols, J .(2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *A Journal of Practice & Theory*, 30 (2), 19-50.

Persons, O. (1995). Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of Applied Business Research*, 11:38–46.

Ravisankar, P , Ravi, V., & et al (2011), Detection of financial statement fraud and feature selection using data mining techniques, *Decision Support Systems*, 50(2): 491-500.

Razaie, M., Nazemi Ardakani, M., naser sadrabadi, A. (2021). Predicting financial statement fraud using The CRISP approach. *Journal of Management Accounting and Auditing Knowledge*, 10(40), 135-150. (In Persian)

Romney, M. & Steinbart, P. (2009). *Accounting Information Systems.* Ahvaz, Shahid Chamran University. (In Persian).

Sadgali. I, Sael. N & Benabbou. F.(2019). Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science*, 148:45-54.

Safarzadeh, M. (2012). The Ability of Financial Ratios in Detecting Fradulent Financial Reporting: Logit Analysis. *Journal of Accounting Knowledge*, 1(1), 137-163. (In Persian).

Sajadi, S. & Kazemi, T. (2016). A Comprehensive Pattern of Fraudulent Financial Reporting in Iran, Grounded Theory. *Empirical Research in Accounting*, 6(3), 185-204. doi: 10.22051/jera.2016.2542. (In Persian)

Saniee Abadeh, M, Mahmoudi, M. & Taherpour, M. (2015). *Applied data mining*. Tehran. Niaze Danesh. (In Persian).

Shabani,A. & Alavi, S. M. (2014). Presenting a method for marine sonar classification using support vector machine multi-class algorithms. *Electronics Industries*, 4 (13): 12-19. (In Persian).

Sharifi Rad, S. & Niknafs, A. (2019). Investigating kernel functions of SVM algorithm in the accuracy of imbalanced data classification in different imbalance rate ranges. *National electronic conference of new achievements in engineering and basic sciences*, Ardabil. (In Persian).

Spathis, C. T. (2002). Detecting false financial statements using published data:Some evidence from Greece. *Managerial Auditing Journal*, 17: 179-191.

Tarasi, M., Banitalebi, B., Zamani, B. (2019). Forecasting Fraudulent Financial Reporting Through Artificial Neural Network. *Management Accounting*, 12(40), 63-79. (In Persian).

Veisi, H., Ghaedsharaf, H., Ebrahimi, M. (2021). Improving the Performance of Machine Learning Algorithms for Heart Disease Diagnosis by Optimizing Data and Features. *Soft Computing Journal*, 8(1), 70-85. (In Persian).

Xiuguo, W. Shengyong, D. (2022). Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using DL. *IEEE Access*. 10: 22516-22532.

Zhou, W. , Kapoor, G .(2011), Detecting evolutionary financial statement fraud, *Decision Support Systems*. 50 (3): 570-575.