فصلنامه علمی پژوهشی
دانش حسابداری و حسابرسی مدیریت
دوره ۱۳/ شماره ٤ (پیاپی ۵۲)/ زمستان ۱٤۰۳
صفحه ۱۵ تا ۲۸

# تکنیک‌های داده کاوی و پیش بینی تقلب صورت‌های مالی

**سید جلال احمدی**
دانشجوی دکترای گروه حسابداری، واحد سمنان، دانشگاه آزاد اسلامی، سمنان، ایران
ahmadijalal.acc@gmail.com

**خسرو فغانی ماکرانی**
دانشیار گروه حسابداری، واحد سمنان، دانشگاه آزاد اسلامی، سمنان، ایران (نویسنده مسئول)
kh.makrani@chmail.ir

**نقی فاضلی**
استادیار گروه حسابداری، واحد سمنان، دانشگاه آزاد اسلامی، سمنان، ایران
fazeli.nphd@gmail.com

تاریخ دریافت: ۱۴۰۰/۰۲/۲۰ تاریخ پذیرش: ۱۴۰۰/۰۵/۳۱

**چکیده**

هدف پژوهش حاضر مقایسه تکتیک‌های داده کاوی شبکه عصبی، درخت تصمیم، نزدیک ترین همسایگی و ماشین بردار پشتیبان در پیش بینی صورت‌های مالی متقلبانه و غیر متقلبانه است. روش پژوهش توصیفی - کاربردی و قلمرو زمانی نیز از سال ۱۳۸۷ تا ۱۳۹۶می باشد. در این پژوهش، نسبت‌های مالی برای دو نمونه متقلب و غیر متقلب و روش‌های داده کاوی مورد تجزیه و تحلیل قرار گرفت. فرضیه های آماری نرمال بودن، همگنی و آزمون هم خطی برای نسبت‌های مالی نمونه های متقلب و غیر متقلب، مورد آزمون قرار گرفت. فرضیه نرمال بودن با استفاده از آزمون کولموگروف اسمیرنوف و آزمون شاپیرو ویلک، انجام پذیرفت. سپس ضریب همبستگی پیرسون در خصوص وجود هم خطی مدل برای نسبت‌های مالی و حذف متغیرهای مستقل همبسته مورد بررسی و آزمون قرار گرفت. در مرحله بعد روشهای داده کاوی برای آزمون آنها در پیش بینی تقلب صورتهای مالی و تمایز صورتهای مالی متقلبانه از غیر منتقلبانه به کار برده شده است. به طور کلی، نتایج حاصل نشان می‌دهد که روش‌های داده کاوی در تمایز صورت‌های مالی متقلبانه از غیر متقلبانه موثر هستند. بدین ترتیب که روش شبکه عصبی ۶۹/۴ درصد، درخت تصمیم ۶۵.۴ درصد، نزدیکترین همسایگی ۶۴/۴ درصد و ماشین بردار پشتیبان ۷۸ درصد پیش بینی صحیح داشته اند.

**واژه‌های کلیدی:** تقلب، داده کاوی، نسبت‌های مالی.

## ۱- مقدمه

صورت‌های مالی برای استفاده کنندگان، اطلاعات مالی مفیدی را برای تخصیص منابع به صورت کارآمد و تصمیم گیری های اقتصادی آگاهانه فراهم میکند (کانیگزگرابر[1]،۲۰۱۲). تقلب در حسابداری برای دنیای کسب و کار شایع و پر هزینه است (هایز[2]،۲۰۱۴). زمانی که تقلب حسابداری آشکار شود، قیمت سهام شرکت کاهش می یابد، و در نتیجه سهامداران متحمل زیان می‌شوند(ورگرس و بوربا[3]،۲۰۱۴). تقلب صورت‌های مالی همچنان در حال افزایش است (کراویتز[4]،۲۰۱۲). موارد تقلب مالی شرکت های ایالات متحده آمریکا از ۳۷۶ مورد در سال ۲۰۰۵ به ۷۲۵ مورد در سال ۲۰۱۱ افزایش یافته است(دفتر تحقیقات فدرال[5]،۲۰۱۲). موارد تقلبات مدیریت (مانند اِنرون[6] و دیگران) به شدت تحت تأثیر اقتصاد جهانی و بازارهای سهام قرار دارند (عباسی و همکاران[7]،۲۰۱۲). تشخیص تقلب برای جلوگیری از پیامدهای شدید تقلب صورتهای مالی و ضرر و زیان به ذینفعان حیاتی است. طرح های تقلب، با استفاده از تکنولوژی پیشرفته پیچیده تر شده و در نتیجه شناسایی آن دشوار شده است (رانکیو[8]،۲۰۱۶). علی رغم اهمیت و ضرورت توجه به پدیده صورت های مالی متقلبانه لیست شرکت های متقلب و مصادیق تقلب در صورت های مالی توسط هیچ ارگان یا نهادی در کشور بررسی و ارایه نمی شود. نهادهایی از قبیل سازمان بورس و اوراق بهادار اطلاعات احتمالی مربوط به هر گونه تحریف و به طور خاص تقلب در صورت های مالی را در اختیار عموم و تحلیل گران قرار نمی دهند. موارد تخلف بررسی شده در سازمان بورس اوراق بهادار نیز در صورت صدور رای از طریق محاکم قضایی به طور خصوصی اطلاع رسانی می شود اما برای استفاده عمومی منتشر نمی شود. (کاظمی، سجادی و رحمانی ۱۳۹۵). هدف این پژوهش غیر تجربی کمی، مقایسه تکنیکهای داده کاوی شبکه عصبی، درخت تصمیم، نزدیک ترین همسایگی و ماشین بردار پشتیبان در تمایز صورت‌های مالی متقلبانه از غیر متقلبانه موثر هستند.

## ۲- مبانی نظری و پیشینه پژوهش

### ۲-۱- مبانی نظری

ادبیات حسابداری و حسابرسی هیچ توافقی در خصوص تعریف تقلب صورت های مالی نشان نداده است. وارگر و بوربا[9] (۲۰۱۴) تقلب را به عنوان پنهان سازی عمدی حقایق بااهمیت یا تحریف حقایق برای تحریک دیگران به عمل کردن به زیان خود[10] تعریف کردند. گوئل و گانگولی[11] (۲۰۱۲)، تقلب صورت‌های مالی را به عنوان یک تحریف عمدی یا حذف مبالغ و اطلاعات افشا شده در صورت‌های مالی با هدف فریب دادن استفاده کنندگان صورتهای مالی تعریف کردند. تقلب یک فعالیت بزهکارانه و فریبکارانه با قصد منفعت مالی و سایر منافع می باشد، همچنین تقلب بعنوان یک عمل بزهکارانه در برگیرنده حقه بازی، حیله گری و رفتار غیر منصفانه بوسیله یک شخص فریبکار و متقلب است ( جاویر و همکاران ،۲۰۱۸). تقلب مالی هم بر عملکرد صنایع مختلف و هم بر زندگی روزمره بشر تاثیرگذار است. تقلب می تواند اعتماد به صنعت و کسب و کار را کاهش داده باعث بی ثباتی در سپرده گذاری مردم و در نتیجه تحمیل هزینه اضافی بر زندگی گردد (سادگالی و همکاران ،۲۰۱۹). اغلب تعاریف تقلب اذعان دارند که تقلب به طور کلی و به طور خاص تقلب صورتهای مالی، ارائه عمدی نادرست اطلاعات (مثلا ارائه نادرست یا حذف مبالغ، افشا، طبقه بندی) از ترازنامه، صورت سود و زیان، صورت جریان وجه نقد و یا یادداشتهای همراه یک شرکت ، برای فریب استفاده کنندگان می باشد.

در این پژوهش معیارهای تقلب بر اساس گزارش انجمن بازرسان رسمی تقلب[12] شامل تفاوت زمانی در شناسایی درآمدها و هزینه ها، ثبت درآمدهای واهی یا کم نمایی درآمدها، پنهان کردن بدهیها و هزینه ها یا بیش نمایی بدهیها و هزینه ها، ارزیابی نادرست دارایی‌ها و افشای ناکافی در خصوص رویه های حسابداری و اطلاعات با اهمیت می باشد. بنابراین، گزارشات حسابرسی شرکتها مبنای کار قرار میگیرد. به این صورت که فرض می شود شرکتهایی که گزارش غیر مقبول دارند نسبت به شرکتهایی با گزارش مقبول به احتمال بیشتری مرتکب تقلب می شوند. این روش توسط فرقاندوست حقیقی و بروازی (۱۳۸۸)، مهام، کردستانی و ترابی (۱۳۹۰)، اعتمادی و زلقی (۱۳۹۲) و برزگری و همکاران (۱۳۹۵)، پیشنهاد گردیده است. بنابراین شرکتهایی که گزارش غیرمقبول ارائه داده اند مشخص شده و سپس در صورتی که گزارش حسابرسی آنها دربرگیرنده معیارهای تقلب مشخص شده توسط انجمن بازرسان رسمی تقلب باشد به عنوان شرکت متقلب شناسایی می‌گردند.

[1] Königsgruber, 2012
[2] Hays,2014
[3] Wuerges & Borba, 2014
[4] Kravitz, 2012
[5] Federal Bureau of Investigation [FBI], 2012
[6] Enron
[7] Abbasi, Albrecht, Vance, & Hansen, 2012
[8] Ruankaew, 2016
[9] Wuerges and Borba (2014)
[10] to act to their own detriments
[11] Goel and Gangolly (2012)
[12] Association of Certified Fraud Examiners

**۲-۲- نسبتهای مالی هشداردهنده احتمال تقلب**

آگاهی از علائم تقلب عامل مهمی در جلوگیری و تشخیص تقلب است (ورگرس و بوربا[1]،۲۰۱۴). هرچند جرم تقلب به ندرت دیده می شود، علائم یا شاخص های تقلب مشاهده می شود (آلبرت[2]،۲۰۱۲). در پژوهش حاضر، نسبت های مالی (متغیرهای مستقل) برای پیش بینی خطر بالقوه تقلب صورتهای مالی به دلیل سادگی، انعطاف پذیری و محبوبیت آن در میان جامعه مالی مورد استفاده قرار می‌گیرد. در این پژوهش، ۳۵ نسبت مالی با استفاده از تکنیکهای داده کاوی و اثربخشی آنها در پیش بینی احتمال تقلب صورتهای مالی و تمایز گزارش مالی متقلبانه از غیر متقلبانه آزمون می شود. آگاهی از علائم تقلب، عامل مهمی در جلوگیری و تشخیص تقلب است(ورگرس و بوربا[3]،۲۰۱۴).

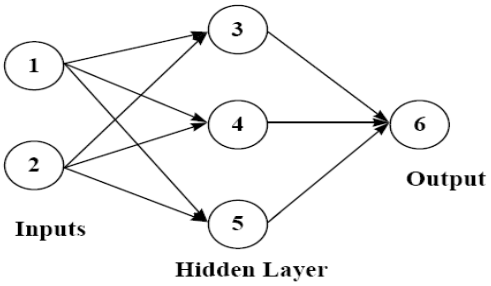
**۲-۳- داده کاوی و کاربردهای آن**

در دو دهه قبل توانایی های فنی بشر در تولید و جمع آوری داده‌ها به سرعت افزایش یافته است. عواملی نظیر استفاده گسترده از بارکد برای تولیدات تجاری، به خدمت گرفتن کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع آوری داده، از اسکن کردن متون و تصاویر تا سیستمهای سنجش از دور ماهواره ای، در این تغییرات نقش مهمی دارند(هربرت.ای. ادلشتاین[4]،۱۹۹۹). بطور کلی استفاده همگانی از وب و اینترنت به عنوان یک سیستم اطلاع رسانی جهانی ما را مواجه با حجم زیادی از داده و اطلاعات می‌کند. این رشد انفجاری در داده‌های ذخیره شده، نیاز مبرم به وجود تکنولوژی های جدید و ابزارهای خودکاری را ایجاد کرده که به صورت هوشمند به انسان یاری رسانند تا این حجم زیاد داده را به اطلاعات و دانش تبدیل کند، داده کاوی به عنوان یک راه حل برای این مسائل مطرح می باشد. در یک تعریف غیر رسمی داده کاوی فرآیندی است، خودکار برای استخراج الگوهایی که دانش را بازنمایی می کنند، که این دانش به صورت ضمنی در پایگاه داده های عظیم، انباره داده[5] و دیگر مخازن بزرگ اطلاعات، ذخیره شده است. داده کاوی بطور همزمان از چندین رشته علمی بهره می برد نظیر؛ تکنولوژی پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه های عصبی، آمار، شناسایی الگو، سیستم های مبتنی بر دانش[6]، حصول دانش[7]، بازیابی اطلاعات[8]، محاسبات سرعت بالا[9] و بازنمایی بصری داده[10] . داده کاوی در اواخر دهه ۱۹۸۰ پدیدار گشته، در دهه ۱۹۹۰ گامهای بلندی در این شاخه از علم برداشته شده و انتظار می رود در این قرن به رشد و پیشرفت خود ادامه دهد(دیوید و همکاران[11]،۲۰۰۱). واژه های «داده کاوی» و «کشف دانش در پایگاه داده»[12] اغلب به صورت مترادف یکدیگر مورد استفاده قرار می گیرند. کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، مفید، و نهایتا الگوها و مدلهای قابل فهم در داده ها می باشد. داده کاوی، مرحله ای از فرایند کشف دانش می باشد و شامل الگوریتمهای مخصوص داده کاوی است، بطوریکه، تحت محدودیتهای مؤثر محاسباتی قابل قبول، الگوها و یا مدلها را در داده کشف می کند(هربرت.ای. ادلشتاین[13]،۱۹۹۹).

**۲-۳-۱- شبکه های عصبی**

شبکه های عصبی از پرکاربردترین و عملی ترین روش های مدل سازی مسائل پیچیده و بزرگ که شامل صدها متغیر هستند می باشد. شبکه های عصبی می توانند برای مسائل کلاس بندی (که خروجی یک کلاس است) یا مسائل رگرسیون (که خروجی یک مقدار عددی است) استفاده شوند. هر شبکه عصبی شامل یک لایه ورودی[14] می باشد که هر گره در این لایه معادل یکی از متغیرهای پیش بینی می باشد. گره های موجود در لایه میانی وصل می شوند به تعدادی گره در لایه نهان[15]. هر گره ورودی به همه گره های لایه نهان وصل می شود. گره های موجود در لایه نهان می توانند به گره های یک لایه نهان دیگر وصل شوند یا می‌توانند به لایه خروجی[16] وصل شوند. لایه خروجی شامل یک یا چند متغیر خروجی می باشد (دیوید و همکاران[17]،۲۰۰۱).

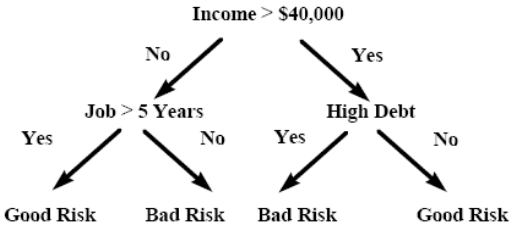
[1] Wuerges & Borba, 2014
[2] Albrech, 2012
[3] Wuerges & Borba, 2014
[4] Herbert A. Edelstein
[5] Data warehouses
[6] Knowledge-based system
[7] Knowledge-acquisition
[8] Information retrieval
[9] High-performance computing
[10] Data visualization
[11] David Hand, Heikki Mannila , Padhraic Smyth.
[12] Knowledge Discovery in Database
[13] Herbert A. Edelstein
[14] Input Layer
[15] Hidden Layer
[16] Output Layer
[17] David Hand, Heikki Mannila , Padhraic Smyth.




**شکل (۱): شبکه عصبی با یک لایه نهان**

### ۲-۳-۲- درختهای تصمیم

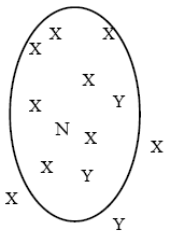
درخت‌های تصمیم روشی برای نمایش یک سری از قوانین هستند که منتهی به یک رده یا مقدار می‌شوند. برای مثال، می-خواهیم متقاضیان وام را به دارندگان ریسک اعتبار خوب و بد تقسیم کنیم. شکل زیر درخت تصمیم را که این مسئله را حل می‌کند نشان می‌دهد و همه مؤلفه‌های اساسی یک درخت تصمیم در آن نشان داده شده است: نود تصمیم، شاخه‌ها و برگ-ها(هربرت.ای. ادلشتاین[1]،۱۹۹۹).




**شکل (۲): درخت تصمیم گیری**

### ۲-۳-۳- نزدیکترین همسایگی

هنگام تلاش برای حل مسائل جدید، افراد معمولا به راه‌حل های مسائل مشابه که قبلا حل شده‌اند مراجعه می‌کنند. نزدیکترین همسایگی K ۲ یک تکنیک دسته‌بندی است که از نسخه‌ای از این متد استفاده می‌کند. در این روش تصمیم‌گیری اینکه یک مورد جدید در کدام دسته قرار گیرد با بررسی تعدادی(k) از شبیه‌ترین موارد یا همسایه‌ها انجام می‌شود. تعداد موارد برای هر کلاس شمرده می‌شوند، و مورد جدید به دسته‌ای که تعداد بیشتری از همسایه‌ها به آن تعلق دارند نسبت داده می‌شود.



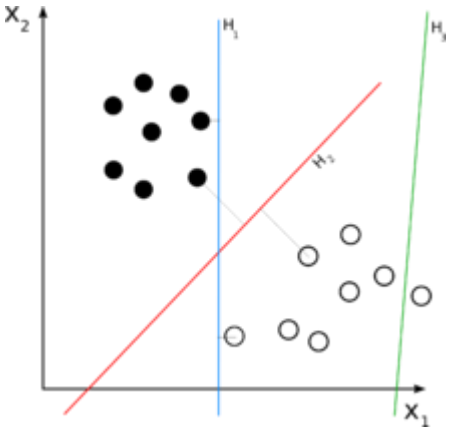

**شکل (۳): محدوده همسایگی**

(بیشتر همسایه ها در دسته X قرار گرفته اند)

### ۲-۳-۴- ماشین بردار پشتیبان

در روش ماشین بردار پشتیبان، هر نمونه داده را به عنوان یک نقطه در فضایn -بعدی روی نمودار پراکندگی داده‌ها ترسیم می کنیم. به طوری‌که مقدار هر ویژگی مربوط به داده‌ها، یکی از مؤلفه‌های مختصات نقطه روی نمودار را مشخص کند. سپس، با ترسیم یک خط راست، داده‌های مختلف و متمایز از یکدیگر را دسته‌بندی می‌نماییم.

در این روش نیز مانند سایر روش های یادگیری ماشین، از داده‌های گروه‌های آموزش و آزمایش استفاده می شود.




**شکل (۴): ماشین بردار پشتیبان**

### ۲-۴- پیشینه‌ی پژوهش

اشرف آکل السایدا[3](۲۰۱۷) در پژوهشی به بررسی اثربخشی ۱۸ نسبت مالی پرداخت. همچنین اثربخشی قانون بنفورد و تکنیک های داده کاوی در تشخیص گزارش مالی متقلبانه از غیرمتقلبانه و پیش بینی احتمال تقلب صورتهای مالی (متغیر وابسته) شرکت های عمومی ایالات متحده که در سازمان بورس و اوراق بهادار ایالات متحده آمریکا ثبت شده اند را مورد بررسی قرارداد. نتایج نشان داد که درخت تصمیم، رگرسیون لجستیک، قانون بنفورد

[1] Herbert A. Edelstein
[2] K-nearest neighbor(k-NN)
[3] Ashraf akl elsayed

و مدل های شبکه عصبی به درستی ۸۹٪، ۹۱٪، ۹۲٪ و ۹۹.۲٪ از موارد تقلب و عدم تقلب را پیش بینی کرده اند و بنابراین می تواند به عنوان ابزار پیش بینی تقلب مورد استفاده قرار گیرند.

وسکی (۲۰۱۳)[1] پژوهشی با عنوان تغییرات قیمت سهم و نسبت قیمت به سود به عنوان شاخص تقلب گزارشات مالی انجام دادند، هدف آزمون فرضیه بازار کارا توسط تعیین محدوده تغییرات در نسبت های قیمت سهم و قیمت به سود قبل از اعلام عمومی تقلب پیش بینی شده یک شرکت بود. نتایج این پژوهش استفاده از معیارهای کمی را برای تشخیص تقلب پیشنهاد می کند. به علاوه محققین می توانند این یافته ها را برای ایجاد یک مدل قوی برای دقت بیشتر تشخیص تقلب به کار ببرند.

ایگو[2] (۲۰۱۷) در پژوهشی با عنوان نسبتهای نقدینگی و سهم بازار به عنوان شاخص تقلب، بررسی کرد که تغییرات در میزان نقدینگی و سهم بازار میتواند شاخصی از فعالیتهای متقلبانه در یک شرکت باشد و اینکه آیا گزارشهای مالی متقلبانه منتشر شده میتواند موجب تغییر در قیمت سهام شرکت شود، تجزیه و تحلیل داده ها نشان داد که بین مولفه های نسبت نقدینگی، نسبت سهم بازار و تغییر در قیمت سهام رابطه معنی داری وجود دار دارد.

خواجوی و ابراهیمی ( ۱۳۹۶) در پژوهشی با عنوان مدلسازی متغیرهای اثرگذار برای کشف تقلب در صورتهای مالی با استفاده از تکنیکهای داده کاوی به بررسی این مساله پرداختند که آیا میتوان از طریق شناسایی و انتخاب متغیرهای اثرگذار در کشف تقلب در صورتهای مالی و با به کارگیری تکنیکهای داده کاوی مدلی برای کشف تقلب در صورتهای مالی شرکتهای پذیرفته شده در بورس اوراق بهادار تهران ارائه کرد. یافته های پژوهش بیانگر وجود شواهدی دال بر عملکرد مناسب مدل‌های پیشنهادی و برتری الگوریتم جنگل تصادفی و شبکه بیزین برای پیشبینی تقلب در صورتهای مالی است. نتایج حاصل از انتخاب ویژگی به روش مبتنی بر همبستگی حاکی از سودمندی متغیرهای نسبت پوشش بهره، نسبت حسابهای دریافتنی به کل داراییها، نسبت موجودی کالا به فروش خالص، نسبت نقدی، لگاریتم طبیعی فروش، نسبت سود خالص به فروش و نسبت جمع داراییهای جاری به کل داراییها برای کشف تقلب بود.

اعتمادی و زلقی (۱۳۹۲) در پژوهشی با عنوان کاربرد رگرسیون لجستیک درشناسایی گزارشگری مالی متقلبانه به بررسی این موضوع پرداختند که داده های صورتهای مالی حسابرسی شده این توانایی را دارند که هرگونه تقلب صورتهای مالی را کشف نمایند. دراین تحقیق، تقلب صورتهای مالی در شرکتهای پذیرفته شده در بورس اوراق بهادار تهران با استفاده از برخی نسبتهای مالی مرتبط تشخیص داده شده است. نمونه آماری تحقیق شامل ۶۸ شرکت درقالب ۳۴ شرکت دارای نشانه های تقلب و ۳۴ شرکت فاقد نشانه های تقلب است. همچنین ۹ نسبت مالی به عنوان پیش بینی کننده های بالقوه برای آزمون انتخاب شده اند. ازروش رگرسیون لجستیک جهت تدوین مدل برای شناسایی عوامل مرتبط با تقلب صورتهای مالی استفاده شده است. این مدل در طبقه بندی صحیح نمونه موردنظر در این تحقیق از نرخ دقت ۸۳.۸ درصد برخوردار بود.

رهنمای رودپشتی (۱۳۹۱) در پژوهشی با عنوان داده کاوی و کشف تقلب های مالی نشان داد که اولا تکنیک های داده کاوی، در شناسایی صورت های مالی متقلبانه سودمند هستند. ثانیا داده کاوی، به عنوان کانون هدایت فکر در مدیریت کسب و کارها جهت کشف تقلب می تواند مورد توجه قرار گیرد.

صفر زاده ( ۱۳۸۹) در پژوهشی با عنوان توانایی نسبت های مالی در کشف تقلب در گزارشگری مالی با استفاد از تحلیل لاجیت در داد های مقطعی به بررسی نقش داده های حسابداری در ایجاد یک الگو برای کشف عوامل مرتبط با تقلب در گزارشگری مالی می پردازد. نتایج تحقیق حکایت از عملکرد مناسب الگو در طبقه بندی شرکت های نمونه داشت به گونه ای که درصد صحت طبقه بندی الگو از ۸۲.۹۸ درصد تجاوز نمود.

## ۳- فرضیه پژوهش

تکنیکهای داده کاوی شبکه عصبی، درخت تصمیم، نزدیک ترین همسایگی و ماشین بردار پشتیبان در تمایز صورت‌های مالی متقلبانه از غیر متقلبانه موثر هستند.

## ۴- روش شناسی پژوهش

این تحقیق از لحاظ هدف کاربردی و ازنوع پژوهشهای تجربی است و با توجه به اینکه از اطلاعات تاریخی استفاده شده است در دسته پژوهشهای شبه آزمایشی قرار می گیرد. جامعه آماری پژوهش حاضر شرکتهای پذیرفته شده در بورس اوراق بهادار تهران و قلمرو زمانی آن هم بین سال‌های ۱۳۸۷ الی ۱۳۹۶ است و داده‌های آن نیز طی این دوره مورداشاره برای تجزیه و تحلیل آماری (توصیفی و استنباطی) با استفاده از بانک های اطلاعاتی و نرم افزارهای موجود گردآوری شده است. متغیر وابسته در این پژوهش، تقلب در صورتهای مالی است که از ماهیت کیفی برخوردار بوده ودارای مقیاس سنجش اسمی است. دراندازه گیری این متغیر، به شرکتهای متقلب عدد یک و به شرکتهای

[1] WESKE

[2] Ego

غیرمتقلب، عددصفر تخصیص داده می شود. متغیرهای مستقل، نسبتهای مالی هستند؛ در ابتدا ۳۵ نسبت مالی انتخاب شدند. که با تحلیل همبستگی و آزمون T منجر به انتخاب نهایی تعداد ۱۹ متغیر مستقل شد؛ که اطلاعات معنادار و غیر همپوشی را ارائه می کنند. سپس آمار توصیفی شامل میانگین، میانه، انحراف او انحراف از میانگین برای هر یک از متغیرهای کمی و به تفکیک نوع شرکت متقلب یا غیرمتقلب بررسی شد. در مرحله بعد برای متغیرهای مستقل، همبستگی بین احتمال تقلب و متغیرهای کمی سنجیده شد. همچنین، با توجه به ضرورت بررسی برخی فرضیات زیربنایی پیش از انجام تحلیل، نرمال بودن(آزمون های کلموگروف اسمیرنف و شاپیرو ویلک) وجود همخطی و همسانی واریانس ها در مورد متغیرهای مستقل مورد بحث، بررسی شد و تحلیل داده ها با استفاده از نرم افزارEXCEL و SPSS۲۰ صورت گرفته است. در نهایت نیز از آماره آزمون کای دو برای تصمیم گیری در مورد معناداری تفاوت توزیع های فراوانی در هر بخش استفاده شد. این تحقیق شامل شرکتهای پذیرفته شده در بورس اوراق بهادار تهران است که شرایط زیر را داشته باشند.

- از تاریخ ۱۳۸۶/۱/۱ به بعد عضو سازمان بورس و اوراق بهادار باشند.
- اطلاعات آنها قابل تهیه و در دسترس باشد.
- جزء شرکت های واسطه مالی، سرمایه گذاری، بانکها، بیمه و بازنشستگی، موسسات اعتباری و شرکتهای چند رشته ای صنعتی نباشند.

## ۵- مدل و متغیرهای پژوهش

نسبتهای مالی شامل وجه نقد عملیاتی به سود خالص، کل دارایی ها به فروش، سود خالص به حقوق صاحبان سهام، فروش به سود خالص، سود قبل از بهره و مالیات به کل دارایی، سود هر سهم، سود ناخالص به فروش، سود عملیاتی به فروش، سود انباشته به کل داراییها، سود خالص به دارایی های ثابت، کل بدهی به حقوق صاحبان سهام، کل بدهیها به کل داراییها، دارایی های جاری به بدهی های جاری، دوره گردش موجودی، دوره پرداخت بدهی، دوره وصول مطالبات، حسابهای دریافتنی به متوسط فروش در هر روز، چرخه تبدیل وجه نقد، Z آلتمن، بدهی های بلندمدت به کل داراییها، سرمایه در گردش به کل دارایی، بدهی بلندمدت به حقوق صاحبان سهام، بهای تمام شده کالای فروش رفته به فروش، فروش به حسابهای دریافتنی، فروش به موجودی، فروش به کل داراییها، فروش به کل داراییهای ثابت و بهای تمام شده کالای فروش رفته به موجودی کالا، مالیات بر درآمد پرداخت شده به سود عملیاتی، هزینه استهلاک به اموال، ماشین آلات و تجهیزات، استقلال هیات مدیره، تعداد جلسات هیات مدیره، درصد مالکیت سهام هیات مدیره، درصد مالکیت سهامداران عمده (بالای ۵ درصد) و درصد مالکیت سهامداران نهادی می باشند.

**۵-۱- برازش مدل شبکه عصبی:** داده ها در دو گروه آموزش و آزمایش تقسیم بندی می شوند.

در گروه آموزش حدود ۷۰ درصد مشاهدات که معادل ۵۷۵ مورد می باشد قرار میگیرد، ۳۰ درصد مابقی داده ها یعنی تعداد ۲۴۵ مورد نیز در گروه آزمایش قرار می گیرند. تابع فعالسازی در لایه پنهان Hyperbolic tangent و در لایه خروجی Softmax می باشد. در این بخش پس از آموزش و برازش مدل شبکه عصبی میزان اعتبار نتایج با توجه به فراوانی پیش بینی های صحیح در هر یک از گروه داده های آزمایش بررسی می شود.

### ۵-۲- برازش مدل درخت تصمیم

روش دیگری که در حوزه ی داده کاوی که مورد استفاده قرار گرفته است روش درخت تصمیم می باشد. هدف بررسی پیش بینی های صحیح شرکت های متقلب و غیرمتقلب می باشد. روش مورد استفاده برای تصمیم گیری ها انتخاب براساس مقدار آماره کای دو (CHAID) بوده است. متغیرهای ورودی شامل متغیرهای مستقل معنادار براساس مدل رگرسیون لجستیک پژوهش احمدی و همکاران (۱۳۹۸) بوده و عبارتند از کل دارایی ها به فروش، سود عملیاتی به فروش، دوره گردش موجودی، چرخه تبدیل وجه نقد می باشند.

### ۵-۳- برازش مدل نزدیک ترین همسایگی

نزدیک‌ترین همسایگی[1](K) یک روش ناپارامتری است که در داده‌کاوی، یادگیری ماشین و تشخیص الگو مورد استفاده قرار می‌گیرد. در این روش نیز مانند سایر روش های یادگیری ماشین، از داده های گروه های آموزش و آزمایش، استفاده می شود. در مرحله پیش پردازش با هدف کسب نتایج بهتر داده ها نرمال سازی می شوند. با نرمال سازی داده ها مقادیر بین «۰» و «۱» قرار می گیرند. به منظور تعیین بهترین مقدار K همسایگی در مدل روی داده های آموزشی با دادن مقادیر مختلف برای K برازش داده شده و ضریب دقت مدل برای هر مقدار K محاسبه می شود.

[1] k-Nearest Neighbors

### ۴-۵- ماشین بردار پشتیبان

در روش ماشین بردار پشتیبان یا همان الگوریتم SVM، هر نمونه داده را به عنوان یک نقطه در فضایn -بعدی روی نمودار پراکندگی داده‌ها ترسیم می کنیم به طوریکه مقدار هر ویژگی مربوط به داده‌ها، یکی از مؤلفه‌های مختصات نقطه روی نمودار را مشخص کند. سپس، با ترسیم یک خط راست، داده‌های مختلف و متمایز از یکدیگر را دسته‌بندی می‌نماییم. در این روش نیز مانند سایر روش های یادگیری ماشین، از داده های گروه های آموزش و آزمایش، استفاده می شود.

در مرحله پیش پردازش با هدف کسب نتایج بهتر داده ها نرمال سازی شده اند. با نرمال سازی داده ها مقادیر بین «۰» و «۱» قرار می گیرند.

پس از آموزش و برازش مدل، میزان اعتبار نتایج با توجه به فراوانی پیش بینی های صحیح در گروه داده های آزمایش بررسی و گزارش می شود.

## ۶- یافته های پژوهش

### ۶-۱- آمار توصیفی

آمار توصیفی که شامل میانگین، میانه، انحراف استاندارد و انحراف از میانگین برای هر یک از متغیرهای کمی و به تفکیک نوع شرکت متقلب یا غیرمتقلب بوده در جدول شمارهٔ (۱)، ارائه شده است. با بررسی این اطلاعات به طور تقریبی تفاوت مقادیر هر یک از این نسبت های مالی مربوط به شرکت های متقلب را در مقایسه با شرکت های غیرمتقلب می توان مشاهده نمود.

**جدول شماره (۱): توصیفی مربوط به متغیرهای مدل**

| شرح متغییرهای مدل پژوهش | متقلب | | | | غیرمتقلب | | | |
|---|---|---|---|---|---|---|---|---|
| | میانگین | میانه | انحراف معیار | انحراف معیار میانگین | میانگین | میانه | انحراف معیار | انحراف معیار میانگین |
| وجه نقد عملیاتی به سود خالص | 1.890 | 1.041 | 14.697 | 0.726 | -4.722 | 0.786 | 168.031 | 8.298 |
| کل دارایی ها به فروش | 1.261 | 1.138 | 0.633 | 0.031 | 1.601 | 1.347 | 1.334 | 0.066 |
| سود خالص به حقوق صاحبان سهام | 0.338 | 0.243 | 1.504 | 0.074 | -0.009 | 0.195 | 3.823 | 0.189 |
| فروش به سود خالص | 37.289 | 7.396 | 140.962 | 6.962 | 199.890 | 6.793 | 2733.651 | 135.005 |
| سود قبل از بهره و مالیات به کل دارایی | 0.168 | 0.147 | 0.127 | 0.006 | 0.115 | 0.108 | 0.157 | 0.008 |
| سود هر سهم | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 |
| سود ناخالص به فروش | 0.245 | 0.205 | 0.156 | 0.008 | 0.196 | 0.173 | 0.173 | 0.009 |
| سود عملیاتی به فروش | 0.177 | 0.142 | 0.153 | 0.008 | 0.104 | 0.105 | 0.292 | 0.014 |
| سود انباشته به کل داراییها | 0.119 | 0.115 | 0.185 | 0.009 | 0.044 | 0.077 | 0.296 | 0.015 |
| سود خالص به دارایی های ثابت | 2.036 | 0.402 | 25.079 | 1.239 | 0.483 | 0.215 | 1.188 | 0.059 |
| کل بدهی به حقوق صاحبان سهام | -0.496 | 1.572 | 31.264 | 1.544 | 3.934 | 1.736 | 23.740 | 1.172 |
| کل بدهیها به کل داراییها | 0.654 | 0.641 | 0.354 | 0.017 | 0.690 | 0.666 | 0.306 | 0.015 |
| داراییهای جاری به بدهی های جاری | 1.439 | 1.291 | 0.982 | 0.048 | 1.223 | 1.109 | 0.576 | 0.028 |
| دوره گردش موجودی | 154.672 | 133.443 | 111.181 | 5.491 | 149.530 | 131.025 | 118.225 | 5.839 |
| دوره پرداخت بدهی | 58.449 | 39.828 | 70.472 | 3.480 | 104.140 | 65.258 | 181.897 | 8.983 |
| دوره وصول مطالبات | 105.983 | 87.470 | 87.614 | 4.327 | 144.872 | 105.021 | 141.660 | 6.996 |
| حسابهای دریافتنی به متوسط فروش در هر روز | 114.186 | 89.525 | 96.339 | 4.758 | 154.972 | 115.548 | 157.942 | 7.800 |
| چرخه تبدیل وجه نقد | 202.206 | 186.036 | 138.822 | 6.856 | 190.262 | 186.333 | 158.838 | 7.844 |
| Z آلتمن | 2.045 | 1.948 | 0.955 | 0.047 | 1.509 | 1.534 | 2.306 | 0.114 |
| بدهی های بلندمدت به کل داراییها | 0.111 | 0.049 | 0.301 | 0.015 | 0.084 | 0.052 | 0.149 | 0.007 |
| سرمایه در گردش به کل دارایی | 0.137 | 0.155 | 0.215 | 0.011 | 0.049 | 0.066 | 0.288 | 0.014 |
| بدهی بلندمدت به حقوق صاحبان سهام | -0.496 | 1.572 | 31.264 | 1.544 | 3.934 | 1.736 | 23.740 | 1.172 |
| بهای تمام شده کالای فروش رفته به فروش | 0.755 | 0.795 | 0.156 | 0.008 | 0.804 | 0.827 | 0.173 | 0.009 |
| فروش به حسابهای دریافتنی | 13.323 | 4.077 | 40.109 | 1.981 | 15.551 | 3.159 | 81.391 | 4.020 |
| فروش به موجودی | 4.332 | 3.618 | 2.881 | 0.142 | 4.878 | 3.677 | 4.128 | 0.204 |
| فروش به کل داراییها | 0.992 | 0.879 | 0.511 | 0.025 | 0.879 | 0.742 | 0.598 | 0.030 |
| فروش به کل داراییهای ثابت | 9.199 | 4.364 | 46.503 | 2.297 | 5.889 | 4.338 | 5.691 | 0.281 |

| شرح متغییرهای مدل پژوهش | متقلب | | | | غیرمتقلب | | | |
|---|---|---|---|---|---|---|---|---|
| | میانگین | میانه | انحراف معیار | انحراف معیار میانگین | میانگین | میانه | انحراف معیار | انحراف معیار میانگین |
| بهای تمام شده کالای فروش رفته به موجودی کالا | 3.348 | 2.583 | 2.561 | 0.126 | 3.967 | 2.700 | 3.705 | 0.183 |
| مالیات بر درآمد پرداخت شده به سود عملیاتی | 0.275 | 0.107 | 2.727 | 0.135 | 0.110 | 0.065 | 0.597 | 0.030 |
| هزینه استهلاک به اموال، ماشین آلات و تجهیزات | 0.355 | 0.108 | 3.328 | 0.164 | 0.107 | 0.100 | 0.069 | 0.003 |
| استقلال هیات مدیره | 0.645 | 0.600 | 0.151 | 0.007 | 0.619 | 0.600 | 0.173 | 0.009 |
| تعداد جلسات هیئت مدیره | 15.688 | 12.000 | 6.117 | 0.302 | 14.815 | 12.000 | 6.067 | 0.300 |
| درصد مالکیت سهام هیات مدیره | 0.664 | 0.727 | 0.227 | 0.011 | 0.739 | 0.700 | 2.454 | 0.121 |
| سهامداران عمده (بالای ۵ درصد) | 0.772 | 0.820 | 0.145 | 0.007 | 0.774 | 0.820 | 0.146 | 0.007 |
| درصد مالکیت سهامداران نهادی | 0.412 | 0.355 | 0.331 | 0.016 | 0.401 | 0.305 | 0.340 | 0.017 |

منبع: یافته های پژوهش

### ۶-۲-آمار استنباطی

با توجه به ضرورت بررسی برخی فرضیات زیربنایی پیش از انجام تحلیل، نرمال بودن، وجود همخطی و همسانی واریانس ها در مورد متغیرهای مستقل مورد بحث، بررسی شده است.

### ۶-۲-۱-آزمون بررسی نرمال بودن متغیرهای کمی

به منظور تعیین نرمال بودن متغیرهای کمی مربوط به نسبت های مالی مورد نظر علاوه بر بررسی شاخص های چولگی و کشیدگی، آماره آزمون های کلموگروف اسمیرنف و شاپیرو ویلک نیز محاسبه شده است. همچنین نمودارهای Q-Qplot (نمودار چندک های توزیع نرمال) و نمودار جعبه ای نیز ترسیم و بررسی شده است. در ارتباط با شاخص های چولگی و کشیدگی انتظار می رود در صورت نرمال بودن متغیرها مقدار حاصل بین ۲- و ۲ باشد. که البته با توجه به این شاخص ها نرمال بودن قابل تایید نیست، بنابراین آماره ی آزمون کلموگروف اسمیرنف و شاپیرو ویلکز بررسی می شود. در این دو آزمون فرضیه صفر بیانگر نرمال بودن متغیر مورد بحث است. چنانچه در هر یک از متغیرهای گروه شرکتهای متقلب یا غیرمقلب مقدار(p-value) حاصل شده، برای این آزمون ها بیش از سطح خطای آزمون یعنی ۰.۰۵ باشد فرضیه ی نرمال بودن پذیرفته می شود. با توجه به توضیحات ارائه شده نتایج حاصل از جدول شماره (۲) نشان دهنده‌ی عدم نرمال بودن متغیرهای مستقل کمی است. البته توجه به این نکته ضروری است که با توجه به حجم بالای اطلاعات موجود استفاده از ضریب همبستگی پیرسون جهت محاسبه ی همبستگی میان متغیرها و آزمون های پارامتری تی جهت مقایسه ی میانگین گروه ها، مشکلی ایجاد نمی کند.

**جدول شماره (۲): بررسی نرمال بودن و همسانی واریانس های متغیرهای کمی مدل پژوهش**

| شرح متغییرهای مدل پژوهش | توصیف متغیرهای پژوهش | | | آماره[1] کلموگروف اسمیرنف[2] | آماره شاپیرو ویلک[3] | سطح معنی داری | بررسی همسانی واریانس | |
|---|---|---|---|---|---|---|---|---|
| | تعدادمشاهدات (N) | چولگی[4] | کشیدگی[5] | | | | آزمون لون | سطح معنی داری |
| وجه نقد عملیاتی به سود خالص | 820 | 774.284 | -27.414 | .446 | .045 | .000 | 3.146 | .077 |
| کل دارایی ها به فروش | 820 | 83.158 | 6.849 | .153 | .598 | .000 | 17.391 | .000 |
| سود خالص به حقوق صاحبان سهام | 820 | 493.864 | -18.279 | .387 | .121 | .000 | 2.290 | .131 |
| فروش به سود خالص | 820 | 789.761 | 27.865 | .444 | .033 | .000 | 5.015 | .025 |
| سود قبل از بهره و مالیات به کل دارایی | 820 | 9.723 | -.719 | .091 | .903 | .000 | .468 | .494 |
| سود هر سهم | 820 | 10.741 | 2.260 | .177 | .785 | .000 | 2.036 | .154 |
| سود ناخالص به فروش | 820 | 5.855 | -.167 | .090 | .940 | .000 | .085 | .771 |
| سود عملیاتی به فروش | 820 | 54.781 | -4.716 | .166 | .712 | .000 | 5.494 | .019 |

[1] Statistic
[2] Kolmogorov-Smirnova
[3] Shapiro-Wilk
[4] Kurtosis
[5] Skewness

| شرح متغییرهای مدل پژوهش | توصیف متغیرهای پژوهش | | | آماره[1] کلموگروف اسمیرنف[2] | آماره شاپیرو ویلک[3] | سطح معنی داری | بررسی همسانی واریانس | |
|---|---|---|---|---|---|---|---|---|
| | تعدادمشاهدات (N) | چولگی[4] | کشیدگی[5] | | | | آزمون لون | سطح معنی داری |
| سود انباشته به کل داراییها | 820 | 54.130 | -4.676 | .160 | .745 | .000 | 6.690 | .010 |
| سود خالص به دارایی های ثابت | 820 | 772.390 | 27.488 | .432 | .035 | .000 | 3.336 | .068 |
| کل بدهی به حقوق صاحبان سهام | 820 | 260.427 | -7.888 | .416 | .156 | .000 | .003 | .957 |
| کل بدهیها به کل داراییها | 820 | 49.758 | 5.283 | .163 | .654 | .000 | .091 | .762 |
| داراییهای جاری به بدهی های جاری | 820 | 80.029 | 6.719 | .176 | .601 | .000 | 2.905 | .089 |
| دوره گردش موجودی | 820 | 23.719 | 3.659 | .126 | .733 | .000 | .847 | .358 |
| دوره پرداخت بدهی | 820 | 153.111 | 10.594 | .281 | .370 | .000 | 19.390 | .000 |
| دوره وصول مطالبات | 820 | 29.203 | 3.787 | .147 | .737 | .000 | 19.754 | .000 |
| حسابهای دریافتنی به متوسط فروش در هر روز | 820 | 50.194 | 4.647 | .155 | .719 | .000 | 14.653 | .000 |
| چرخه تبدیل وجه نقد | 820 | 2.969 | .935 | .056 | .946 | .000 | 3.729 | .054 |
| Z آلتمن | 820 | 306.508 | -13.795 | .184 | .476 | .000 | 2.642 | .104 |
| بدهی های بلندمدت به کل داراییها | 820 | 139.164 | 9.832 | .327 | .346 | .000 | 3.679 | .055 |
| سرمایه در گردش به کل دارایی | 820 | 31.198 | -3.094 | .070 | .849 | .000 | 6.438 | .011 |
| بدهی بلندمدت به حقوق صاحبان سهام | 820 | 260.427 | -7.888 | .416 | .156 | .000 | .003 | .957 |
| بهای تمام شده کالای فروش رفته به فروش | 820 | 5.855 | .167 | .090 | .940 | .000 | .085 | .771 |
| فروش به حسابهای دریافتنی | 820 | 128.735 | 10.651 | .412 | .170 | .000 | 2.118 | .146 |
| فروش به موجودی | 820 | 24.823 | 3.615 | .154 | .724 | .000 | 17.882 | .000 |
| فروش به کل داراییها | 820 | 11.997 | 2.614 | .133 | .803 | .000 | .378 | .539 |
| فروش به کل داراییهای ثابت | 820 | 387.111 | 19.425 | .412 | .088 | .000 | 4.096 | .043 |
| بهای تمام شده کالای فروش رفته به موجودی کالا | 820 | 24.197 | 3.587 | .183 | .701 | .000 | 17.310 | .000 |
| مالیات بر درآمد پرداخت شده به سود عملیاتی | 820 | 724.566 | 26.063 | .411 | .068 | .000 | 1.184 | .277 |
| هزینه استهلاک به اموال، ماشین آلات و تجهیزات | 820 | 409.671 | 20.230 | .471 | .035 | .000 | 6.796 | .009 |
| استقلال هیات مدیره | 820 | .219 | -.477 | .245 | .858 | .000 | .653 | .419 |
| تعداد جلسات هیئت مدیره | 820 | 9.305 | 2.805 | .292 | .606 | .000 | 1.596 | .207 |
| درصد مالکیت سهام هیات مدیره | 820 | 785.610 | 27.728 | .434 | .064 | .000 | 1.750 | .186 |
| سهامداران عمده (بالای ۵ درصد) | 820 | 2.260 | -1.290 | .142 | .909 | .000 | .169 | .682 |
| درصد مالکیت سهامداران نهادی | 820 | -1.412 | .286 | .125 | .894 | .000 | 1.251 | .264 |

منبع: یافته های پژوهش

### ۶-۲-۲- هم‌خطی

در این قسمت ضریب همبستگی میان متغیرهای مستقل نسبت های مالی با یکدیگر محاسبه شده ومتغیرهایی که دارای همبستگی معنادار هستند شناسایی شده اند. در ادامه در صورت لزوم به حذف برخی متغیرها پرداخته شده است.

### ۶-۲-۳- همسانی واریانس‌ها

نتایج حاصل در جدول شمارهٔ (۲)، نشان می دهد که در مورد نسبت های مالی « وجه نقد عملیاتی به سود خالص »،« کل دارایی ها به فروش »،« سود خالص به حقوق صاحبان سهام »،« فروش به سود خالص »،« سود هر سهم »،« سود عملیاتی به فروش »،« سود انباشته به کل داراییها »،« دوره پرداخت بدهی »،« دوره وصول مطالبات »،« حسابهای دریافتنی به متوسط فروش در هر روز »،«Z آلتمن »،«سرمایه در گردش به کل دارایی »،« فروش به موجودی » و « بهای تمام شده کالای فروش رفته به موجودی کالا»، همسانی واریانس های دو گروه برقرار نیست. در مورد سایر متغیرها مقدار (p-value) حاصل شده، بیش از ۰/۰۵ درصد است، از اینرو، وجود همسان بودن واریانس پذیرفته می شود.

### ۶-۲-۴- محاسبه ضریب همبستگی

نتایج حاصل از محاسبه ضریب همبستگی متغیرهای نسبت های مالی و احتمال تقلب نشان می دهد که متغیرهای « وجه نقد عملیاتی به سود خالص »،« کل دارایی ها به فروش »،« کل بدهیها به کل داراییها »،»،« دوره گردش موجودی »،«دوره پرداخت بدهی»،« دوره وصول مطالبات »،«چرخه تبدیل وجه

نقد»،«بهای تمام شده کالای فروش رفته به فروش» و«فروش به حسابهای دریافتنی» همبستگی مثبت با احتمال تقلب دارند و متغیرهای« سود خالص به حقوق صاحبان سهام »،« سود قبل از بهره و مالیات به کل دارایی »،« سود هر سهم »،« سود ناخالص به فروش »،« سود عملیاتی به فروش »،« سود انباشته به کل داراییها »،« دارایی‌های جاری به بدهی های جاری »،« Z آلتمن »، «سرمایه در گردش به کل دارایی »، «فروش به موجودی »،« سهامداران عمده (بالای ۵ درصد)» ،«درصد مالکیت سهامداران نهادی»، رابطه‌ی معکوس با احتمال وقوع تقلب را دارا می باشند.

### ۶-۲-۵- نتایج حاصل از آزمون فرضیه

#### ۶-۲-۵-۱- برازش مدل شبکه عصبی

در گروه آموزش حدود ۷۰ درصد مشاهدات که معادل ۵۷۵ مورد می باشد قرار گرفته است، ۳۰ درصد مابقی داده ها یعنی تعداد ۲۴۵ مورد نیز در گروه آزمایش قرار دارند. تابع فعالسازی در لایه پنهان Hyperbolic tangent و در لایه خروجی Softmax می باشد. پس از آموزش و برازش مدل شبکه عصبی میزان اعتبار نتایج با توجه به فراوانی پیش بینی های صحیح در هر یک از گروه داده های آزمایش بررسی شده است. نتایج حاصل در جدول زیر نشان داده شده است. در داده های بخش آزمایش شبکه، نرخ دقت آزمون برابر ۶۹.۴ درصد بوده است. این مقدار برای داده های گروه آموزش ۶۳ درصد حاصل شده است.

**جدول شماره (۳): برازش مدل شبکه عصبی**

| طبقه بندی | | | | |
|---|---|---|---|---|
| نمونه | مشاهده شده | پیش بینی | | |
| | | ۰ | ۱ | نرخ دقت |
| آموزش | ۰ | ۱۷۹ | ۱۱۰ | %۹.۶۱ |
| | ۱ | ۱۰۳ | ۱۸۳ | %۰.۶۴ |
| | درصد کلی | %۰.۴۹ | %۰.۵۱ | %۰.۶۳ |
| آزمایش | ۰ | ۸۶ | ۳۵ | %۱.۷۱ |
| | ۱ | ۴۰ | ۸۴ | %۷.۶۷ |
| | درصد کلی | %۴.۵۱ | %۶.۴۸ | %۴.۶۹ |

میزان اهمیت هر یک از متغیرهای وارد شده در مدل شبکه عصبی نیز در مقایسه با یکدیگر در جدول شماره (۳) آمده است. بیشترین میزان اهمیت در مدل مربوط به متغیرهای کل دارایی ها به فروش، سود عملیاتی به فروش، دارایی های جاری به بدهی های جاری، دوره وصول مطالبات و Z آلتمن می می باشد.

**جدول شماره (۴):خلاصه نتایج مقایسه‌ای میزان اهمیت هر یک از متغیرهای وارد شده در مدل شبکه عصبی**

| متغیرهای مستقل مهم | | |
|---|---|---|
| شرح متغییرهای مدل پژوهش | ضریب اهمیت | ضریب اهمیت نرمال شده |
| وجه نقد عملیاتی به سود خالص | .055 | 41.5% |
| کل دارایی ها به فروش | .076 | 57.6% |
| سود خالص به حقوق صاحبان سهام | .050 | 38.3% |
| سود هر سهم | .050 | 38.1% |
| سود ناخالص به فروش | .043 | 32.5% |
| سود عملیاتی به فروش | .075 | 57.2% |
| سود انباشته به کل داراییها | .032 | 24.3% |
| کل بدهیها به کل داراییها | .034 | 25.9% |
| داراییهای جاری به بدهی های جاری | .081 | 61.3% |
| دوره گردش موجودی | .028 | 21.0% |
| دوره پرداخت بدهی | .041 | 31.4% |
| دوره وصول مطالبات | .113 | 85.6% |
| چرخه تبدیل وجه نقد | .039 | 29.7% |
| Zآلتمن | .132 | 100.0% |
| سرمایه در گردش به کل دارایی | .058 | 43.8% |
| فروش به حسابهای دریافتنی | .028 | 21.6% |
| فروش به موجودی | .038 | 28.7% |
| سهامداران عمده (بالای ۵ درصد) | .018 | 13.6% |
| درصد مالکیت سهامداران نهادی | .011 | 8.2% |

منبع: یافته های تحقیق

#### ۶-۲-۵-۲- برازش مدل درخت تصمیم

متغیرهای ورودی شامل متغیرهای مستقل معنادار براساس رگرسیون لجستیک بر اساس پژوهش احمدی و همکاران (۱۳۹۸) بوده و عبارتند از کل دارایی ها به فروش، سود عملیاتی به فروش، دوره گردش موجودی، چرخه تبدیل وجه نقد.

براساس درخت حاصل در Node صفر مطابق انتظار ۵۰ درصد شرکتها متقلب و ۵۰ درصد غیرمتقلب هستند. اولین عامل شناسایی شده در این درخت نسبت مالی دوره پرداخت بدهی است. همچنین در رده ی دوم نسبت دارایی های جاری به بدهی های جاری و سرمایه در گردش به کل دارایی و همچنین در رده سوم نسبت کل دارایی ها به فروش، در درخت قرار گرفته است. به طور کلی نتایج حاصل از آزمون مربوطه در جدول شماره (۵) نشان می دهد که پیش بینی های صحیح انجام شده توسط این درخت تصمیم حدود ۶۵ درصد است.

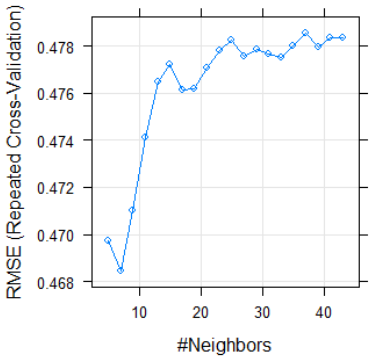
**جدول شماره (۵): خلاصه نتایج آزمون درخت تصمیم**

| دسته بندی | | | |
|---|---|---|---|
| مشاهده شده | پیش بینی | | |
| | ۰ | ۱ | درصد صحیح |
| ۰ | ۲۷۲ | ۱۳۸ | %۶۶.۳ |
| ۱ | ۱۴۶ | ۲۶۴ | %۶۴.۴ |
| درصد کلی | %۵۱.۰ | %۴۹.۰ | %۶۵.۴ |
| Growing Method: CHAID | | | |

بنابراین می توان گفت درخت تصمیم با نرخ دقت ۶۵.۴ درصد توانایی خوبی را در پیش بینی تقلب صورتهای مالی دارد.

### ۶-۲-۵-۳- برازش مدل نزدیک ترین همسایگی

داده های آموزش با دادن مقادیر مختلف برای K برازش داده شده و ضریب دقت مدل برای هر مقدار K محاسبه می شود. نمودار شماره (۲) تغییرات ***RMSE*** را براساس مقادیر مختلف K نشان می دهد. همانطور که ملاحظه می شود K=۷ بهترین حالت ممکن را حاصل می کند.




**نمودار شماره (۱): نمایش برازش مدل نزدیکترین همسایگی**

**جدول شماره (۶): خلاصه نتایج آزمون برازش مدل نزدیک ترین همسایگی**

| دسته بندی | | | | |
|---|---|---|---|---|
| نمونه | مشاهده | پیش بینی | | |
| | | 0 | 1 | درصد صحیح |
| آموزش | 0 | ۲۳۱ | ۷۰ | 23.2% |
| | 1 | ۶۳ | ۲۵۱ | 79.9% |
| | درصد کلی | 78.6% | 78.2% | 78.4% |
| | 0 | 68 | 41 | 62.4% |
| | 1 | 32 | 64 | 66.7% |
| | درصد کلی | 68% | 60.9% | 64.4% |
| متغیر وابسته: تقلب | | | | |

در این روش اعتبار نتایج با استفاده از گروه داده های آزمایش و آموزش قابل محاسبه می باشد. نتایج حاصل برای گروه آموزش نشان می دهد که نرخ دقت آزمون برابر ۷۸.۴ درصد بوده است. همچنین در گروه آزمایش نیز دقت پیش بینی های صورت گرفته برابر ۶۴.۴ درصد می باشد. بنابراین میتوان گفت این روش با نرخ دقت ۶۴.۴ درصد، توانایی خوبی را در پیش بینی تقلب صورتهای مالی دارد.

### ۶-۲-۵-۴- ماشین بردار پشتیبان

پس از آموزش و برازش مدل میزان اعتبار نتایج با توجه به فراوانی پیش بینی های صحیح در گروه داده های آزمایش بررسی و در جدول ... گزارش شده است. نتایج حاصل در جدول زیر نشان داده می دهد که به طور کلی در داده های مورد استفاده برای آزمایش ۷۸ درصد مقادیر درست پیش بینی شده است. همچنین در گروه آموزش نیز ۷۳.۲ درصد پیش بینی صحیح انجام شده است. بنابراین میتوان گفت این روش با نرخ دقت ۷۸ درصد، توانایی خوبی را در پیش بینی تقلب صورتهای مالی دارد.

**جدول شماره (۷): خلاصه نتایج آزمون ماشین بردار پشتیبان**

| نمونه | مشاهده | پیش بینی | | |
|---|---|---|---|---|
| | | ۰ | ۱ | درصد پیش بینی |
| آموزش | 0 | ۲۳۸ | ۹۸ | %۷۰.۸ |
| | 1 | ۶۴ | ۲۱۵ | %۷۷ |
| | درصد کلی | %۷۸.۸ | %۶۸.۶ | %۷۳.۲ |
| آزمون | 0 | ۹۰ | ۲۷ | %۷۶.۹ |
| | 1 | ۱۸ | ۷۰ | %۸۱.۴ |
| | درصد کلی | %۸۳.۳ | %۷۲.۲ | %۷۸ |

## ۷- بحث و نتیجه گیری

داده کاوی بطور همزمان از چندین رشته علمی بهره می برد نظیر؛ تکنولوژی پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه های عصبی، آمار، شناسایی الگو، سیستم های مبتنی بر دانش[1]، حصول دانش[2]، بازیابی اطلاعات[3]، محاسبات سرعت بالا[4]

[1] Knowledge-based system
[2] Knowledge-acquisition
[3] Information retrieval
[4] High-performance computing

و بازنمایی بصری داده[1] . واژه های «داده کاوی» و «کشف دانش در پایگاه داده»[2] اغلب به صورت مترادف یکدیگر مورد استفاده قرار می گیرند. کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، مفید، و نهایتا الگوها و مدلهای قابل فهم در داده ها می باشد. داده کاوی، مرحله ای از فرایند کشف دانش می باشد و شامل الگوریتمهای مخصوص داده کاوی است، بطوریکه، تحت محدودیتهای مؤثر محاسباتی قابل قبول، الگوها و یا مدلها را در داده کشف می کند

هدف اصلی پژوهش حاضر بررسی اثربخشی تکنیکهای داده کاوی شبکه عصبی، درخت تصمیم، نزدیک ترین همسایگی و ماشین بردار پشتیبان درپیش بینی تقلب صورت‌های مالی است. به منظور تحقق این هدف از یک نمونه شرکتهای با احتمال تقلب و شرکتهای غیرمتقلب طبق معیارهای تبیین شده قبلی برای دسته بندی شرکتها؛ استفاده شد. یک مجموعه از ۳۵ نسبت مالی به عنوان پیش بینی کنندگان بالقوه تقلب صورتهای مالی انتخاب شدند.این متغیرها در تحقیقات قبلی با اهمیت ظاهرشدندو از صورتهای مالی انتشاریافته استخراج می شوند. در نهایت ۱۹ متغیر انتخاب شدند. نتایج حاصل از آزمونها که تایید کننده نتایج آزمونهای قبلی است نشان میدهد که صورتهای مالی می تواند جهت پیش بینی تقلب استفاده شود و اعتبار دهندگان، حسابرسان و سایر ذینفعان را یاری رساند. بدین ترتیب که تکنیک شبکه عصبی ۶۹/۴ درصد، درخت تصمیم ۶۵.۴ درصد، نزدیکترین همسایگی۶۴.۴ درصد و ماشین بردار پشتیبان ۷۸ درصد پیش بینی صحیح داشته اند. بنابراین از نظر کارایی، تکنیک ماشین بردار پشتیبان با ۷۸ درصد صحت طبقه بندی بهترین عملکرد را داشته است.

### ۷-۱- مقایسه نتایج حاصل از فرضیه ها

۱) دقت روش نزدیکترین همسایگی (۶۴.۴ درصد) با دقت روش رگرسیون لجستیک (۶۴.۶) تفاوت معناداری ندارد. به بیان دیگر، روش نزدیکترین همسایگی در طبقه بندی صحیح شرکتها مشابه روش رگرسیون لجستیک است. اما دقت روش ماشین بردار پشتیبان (۷۸ درصد) بالاترین دقت و بالاتر از روش شبکه عصبی ( ۶۹.۴ درصد)، درخت تصمیم (۶۵.۴ درصد) و دو روش ذکر شده قبلی است. به بیان دیگر عملکرد روش ماشین بردار پشتیبان در طبقه بندی صحیح شرکتها به شکل معناداری بهتر از عملکرد روشهای نزدیکترین همسایگی، رگرسیون لجستیک، شبکه عصبی و درخت تصمیم است و روش رگرسیون لجستیک، در شناسایی شرکتهای متقلب به شکل معناداری ناموفقتر از روش ماشین بردار پشتیبان است.

۲) نرخ خطای نزدیکترین همسایگی (۳۵.۶) با نرخ خطای رگرسیون لجستیک (۳۵.۲) تفاوت معناداری ندارد. به بیان دیگر روش نزدیکترین همسایگی در طبقه بندی اشتباه شرکتها ، مشابه عملکرد الگوریتم شبکه عصبی است. نرخ خطای ماشین بردار پشتیبان (۲۲ درصد) به شکل معناداری کمتر از نرخ خطای روشهای شبکه عصبی (۳۰.۶) و درخت تصمیم (۳۴.۶) و دو روش ذکر شده قبلی است. به بیان دیگر اشتباه در طبقه بندی شرکتها در ماشین بردار پشتیبان به شکل معناداری کمتر از اشتباه در طبقه بندی شرکتها با استفاده از روشهای نزدیکترین همسایگی، رگرسیون لجستیک، شبکه عصبی و درخت تصمیم است.

۳) حساسیت روش نزدیکترین همسایگی (۶۶.۷ درصد) با حساسیت روش شبکه عصبی (۶۷.۶ درصد) تفاوت معناداری ندارد. به بیان دیگر عملکرد روش نزدیکترین همسایگی در شناسایی شرکتهای متقلب، مشابه روش شبکه عصبی است. معیار حساسیت ماشین بردار پشتیبان (۷۹.۵ درصد) به شکل معناداری بیشتر از معیار حساسیت روشهای رگرسیون لجستیک (۶۱.۴ درصد) و درخت تصمیم (۶۴.۴ درصد) و دو روش ذکر شده قبلی است. به بیان دیگر عملکرد ماشین بردار پشتیبان در شناسایی شرکتهای متقلب به شکل معناداری بهتر از عملکرد روشهای نزدیکترین همسایگی، رگرسیون لجستیک، شبکه عصبی و درخت تصمیم است و روش رگرسیون لجستیک، در شناسایی شرکتهای متقلب به شکل معناداری ناموفقتر از روش ماشین بردار پشتیبان است.

۴) معیار ویژگی ماشین بردار پشتیبان (۷۶.۹) به شکل معناداری بیشتر از ویژگی روشهای رگرسیون لجستیک (۶۸ درصد)، درخت تصمیم (۶۶.۳ درصد)، نزدیکترین همسایگی (۶۲.۴ درصد) و شبکه عصبی (۷۱.۱ درصد) است. به بیان دیگر عملکرد ماشین بردار پشتیبان در شناسایی شرکتهای غیرمتقلب به شکل معناداری بهتر از عملکرد روشهای نزدیکترین همسایگی، رگرسیون لجستیک، شبکه عصبی و درخت تصمیم است. افزون

[1] Data visualization

[2] Knowledge Discovery in Database

بر این ویژگی نزدیکترین همسایگی به شکل معناداری کمتر از ویژگی روش ماشین بردار پشتیبان است. به بیان دیگر نزدیکترین همسایگی در شناسایی شرکتهای غیرمتقلب به شکل معناداری ناموفقتر از روش ماشین بردار پشتیبان است.

۵) این پژوهش میتواند به عنوان زیربنایی برای پژوهشهای آتی استفاده شود و سایر پژوهشگران با استفاده از سایر تکنیک های داده کاوی و یا درنظر قراردادن شرکتهایی مانند شرکتهای سرمایه گذاری، بانکها وبیمه که در نمونه تحقیق قرار نگرفته اند، به پژوهش در این موضوع ادامه دهند. همچنین مطالعه یک صنعت خاص می تواند نتایج ویژه خود را داشته باشد. با ارتقای روشهای مورد استفاده در داده کاوی میتوان داده کاوی را به ابزاری مفیدتر در فرآیند تصمیم گیری مالی تبدیل کرد. کاربرد روشهای داده‌کاوی بر روی نسبتهای مالی و نیز دیگر اطلاعات ، می‌تواند به حسابرسان در کشف تقلب کمک کند، به‌طوری که آنان می‌توانند از نتایج این تحلیلها به‌عنوان یک علامت اولیه هشداردهنده نسبت به احتمال تقلب صورتهای مالی استفاده کنند. کشف نشانگرهای تقلب در صورتهای مالی، اثری بااهمیت بر تعیین تقلب صورتهای مالی دارد. به طور کل نتایج این پژوهش با نتایج پژوهشهای اشرف آکل الساید (۲۰۱۷)، خواجوی (۱۳۹۵) و رهنمای رودپشتی (۱۳۹۱) مبنی بر سودمندی شیوه های داده کاوی برای پیش بینی تقلب در صورتهای مالی مطابقت دارد.

## فهرست منابع


* رهنمای رود پشتی، فریدون (۱۳۹۱)، داده کاوی و کشف تقلب های مالی، فصلنامه علمی و پژوهشی دانش حسابداری و حسابرسی مدیریت، شماره سوم
* جمالی، زهرا؛ برزگری خانقاه، جمال؛ عارف منش، زهره؛ انصاری سامانی، حبیب (۱۳۹۵) بررسی رابطه مکانیزمهای حاکمیت شرکتی و کیفیت حسابرسی بر وقوع تقلب در صورتهای مالی شرکت های پذیرفته شده در بورس اوراق بهادار تهران ، یزد، دانشگاه یزد
* اعتمادی، حسین؛ زلقی، حسن؛ (۱۳۹۲) کاربرد رگرسیون لجستیک در شناسایی گزارشگری مالی متقلبانه، فصلنامه دانش حسابرسی، شماره پنجاه و یک
* مهام, کیهان؛ غلامرضا کردستانی و ابوالفضل ترابی، ۱۳۹۱، ارائه مدلی برای پیش بینی خطر بروز تقلب در گزارشگری مالی، اولین همایش بین المللی اقتصاد سنجی، روشها و کاربردها، سنندج، دانشگاه آزاد اسلامی واحد سنندج.
* فرقاندوست حقیقی، کامبیز؛ برواری، فرید، ۱۳۸۸، بررسی کاربرد روش های تحلیلی در ارزیابی ریسک تحریف صورت های مالی (تقلب مدیریت)، دانش و پژوهش حسابداری، شماره شانزده
* خواجوی، شکراله؛ ابراهیمی، مهرداد (۱۳۹۶) مدلسازی متغیرهای اثرگذار برای کشف تقلب در صورتهای مالی با استفاده از تکنیکهای داده کاوی، فصلنامه حسابدای مالی، شماره سی و سه
* سجادی، سید حسین؛ کاظمی، توحید (۱۳۹۵) الگوی جامع گزارشگری مالی متقلبانه در ایران به روش نظریه پردازی زمینه بنیان، پژوهشهای تجربی حسابداری، سال ششم، شماره ۲۱، پاییز ۱۳۹۵، صص ۱۸۵-۲۰۴
* Abbasi A., Albrecht C., Vance A., & Hansen J. (2012). Metafraud: A meta-learning framework for detecting financial fraud. MIS Quart Manage Inf Syst MIS Quarterly: Management Information Systems, 36(4), 1293-1327.
* Albrecht, W. S. (2012). Fraud examination. Mason, Ohio: South-Western Cengage Learning
* ASHRAF AKL ELSAYED . (2017). Predictability of Financial Statements Fraud-Risk . Northcentral University.
* Association of Certified Fraud Examiners (ACFE). (2016). Report to the nations on occupational fraud and abuse. Austin, TX: The association of certified fraud examiners, Inc.
* Hays, J. B. (2014). An investigation of the motivation of management accountants to report fraudulent accounting activity: Applying the theory of planned behavior. Dissertation Abstracts International Section A, 75.
* Kravitz, R. H. (2012). Auditors' responsibility for detecting fraud. CPA Journal, 82(6), 24-30
* Königsgruber, R. (2012). Capital Allocation Effects of Financial Reporting Regulation and Enforcement. European Accounting Review, 21(2), 283-296. doi:10.1080/09638180.2011.558294
* Goel, S., & Gangolly, J. (2012). Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. Intelligent Systems in Accounting, Finance & Management, 19(2), 75-89. doi:10.1002/isaf.1326
* Mangala, D., & Kumari, P. (2015). Corporate fraud prevention and detection: Revisiting the literature. Journal of Commerce and Accounting Research, 4(1).
* Ruankaew, T. (2016). Beyond the fraud diamond. International Journal Of Busines Management & Economic Research, 7(1), 474-476.
* Wuerges, A. E., & Borba, J. A. (2014). Accounting Fraud: an estimation of detection probability. Revista Brasileira De Gestão De Negócios, 16(52), 466-483. doi:10.7819/rbgn.v16i52.1555

Accounting Knowledge & Management Auditing
Vol. 13/ No. 52/ Winter 2024

# Data Mining Techniques and Forecasting Financial Statement Fraud

**Seyed Jalal Ahmadi**
PhD student in Accounting, Semnan Branch, Islamic Azad University, Semnan, Iran

**Khosro faghani Makrani**
Associate Professor, Department of Accounting, semnan Branch, Islamic Azad University, semnan, Iran.

**Naghi Fazeli**
Assistant Professor , Department of Accounting, Semnan Branch, Islamic Azad University, Semnan, Iran

## Abstract
The purpose of this study is to compare neural network, decision tree, nearest neighbor and support vector machine data mining techniques in predicting fraudulent and non-fraudulent financial statements. The research method is descriptive-applied and time domain from 2008 to 2018. In this study, financial ratios for two fraudulent and non-fraudulent samples and data mining methods were analyzed. Statistical hypotheses of normality, homogeneity and linearity test for financial ratios of fraudulent and non-fraudulent samples were tested. The normality hypothesis was tested using Kolmogorov-Smirnov test and Shapiro Wilk test. Then Pearson correlation coefficient for the existence of the model for financial ratios and elimination of correlated independent variables was tested. Next, data mining methods are used to test them in predicting financial statement fraud and distinguishing fraudulent and non-fraudulent financial statements. In general, the results show that data mining methods are effective in differentiating fraudulent and non-fraudulent financial statements. The neural network method had a correct prediction of 69.4%, decision tree 65.4%, nearest neighbor 64.4% and support vector machine 78%.
**Keywords:** Fraud, Data mining, Financial ratios