پژوهش‌های کاربردی در گزارشگری مالی
سال هفتم، شماره ۱۳، پاییز و زمستان ۱۳۹۷
سازمان حسابرسی
صص ۴۴-۷

بررسی سودمندی روش انتخاب متغیر ریلیف در بهبود نتایج پیش‌بینی فرار مالیاتی با استفاده از داده‌کاوی

محمد نمازی *، محمد صادق‌زاده مهارلویی **

چکیده

پژوهش حاضر به بررسی سودمندی روش‌های ریلیف و داده‌کاوی در پیش‌بینی فرار مالیاتی شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران، با استفاده از داده‌های حسابداری و الگوهای درخت تصمیم، در دو حالت بدون انتخاب متغیرها و با انتخاب متغیرها، می‌پردازد. جامعه آماری پژوهش حاضر کلیه شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران در بازه زمانی ۱۳۸۴ تا ۱۳۹۴ است و نمونه پژوهش برابر با ۱۰۸۱ سال شرکت می‌باشد. از روش‌های آماری تحلیل واریانس یک طرفه، آزمون t-test نمونه‌های مستقل، الگوریتم‌های داده‌کاوی درخت تصمیم و روش انتخاب متغیر ریلیف برای تحلیل داده‌ها استفاده شد. داده‌های پژوهش با استفاده از نرم‌افزارهای SPSS و Weka مورد تجزیه و تحلیل آماری قرار گرفتند. نتایج حاصل از الگوریتم ریلیف نشان داد که متغیرهای نسبت سود عملیاتی به جمع دارایی‌ها، نسبت بازده دارایی‌ها و ارزش بازار شرکت برای پیش‌بینی فرار مالیاتی مناسب‌تر از سایر متغیرها هستند. همچنین، نتایج آزمون تحلیل واریانس نشان داد که تفاوت در دقت پیش‌بینی روش‌های مختلف درخت تصمیم از لحاظ آماری نیز معنادار است. افزون بر این، نتایج نشان داد در هنگام مقایسه هر یک از الگوریتم‌ها به تنهایی در دوحالت با و بدون مرحله انتخاب متغیر، تفاوت تنها در الگوریتم

* استاد حسابداری، دانشگاه شیراز
** دکترای حسابداری، دانشگاه شیراز
نویسنده مسئول: محمد نمازی mnamazi@ rose.sharizu.ac

تاریخ دریافت: ۱۳۹۷/۰۴/۲۰
تاریخ پذیرش: ۱۳۹۷/۱۲/۲۲

LMT معنادار بود و در سایر الگوریتم‌ها، اگرچه دقت نتایج بهتر شده بود، اما این دقت از لحاظ آماری معنادار نبود. به عبارت دیگر، استفاده از روش انتخاب متغیر ریلیف، در هر حالتی موجب به بهبود عملکرد الگوریتم‌ها نمی‌شود.

واژه‌های کلیدی: پیش‌بینی فرار مالیاتی، نسبت‌های مالی، الگوریتم درخت تصادفی، الگوریتم جنگل تصادفی، ریلیف، داده‌کاوی.

مقدمه

به گونه کلی، فرار مالیاتی یک مسئله اقتصادی – اجتماعی مهم بدون توجه به نوع سیستم مالیاتی یا سطح توسعه اقتصادی کشورها در همه جهان است (استانکویشس[۱] ولئوناس[۲]، ۲۰۱۵). شواهد رو به رشدی مبنی بر اینکه فرار مالیاتی و فساد در سال‌های اخیر افزایش یافته است نیز وجود دارد. به عنوان نمونه، گزارش اخیر به وسیله وزارت مالیه اسپانیا[۳] نشان داد که اقتصاد سایه در اسپانیا در طی سال‌های ۲۰۰۸ تا ۲۰۱۲ به میزان ۶/۸ درصد افزایش یافته و به ۲۴/۶ درصد رسیده است (پاپا[۴] و همکاران، ۲۰۱۵). در ایران نیز، آل‌بوسویلم (۱۳۹۰) نشان داد که اندازه نسبی فعالیت‌های زیرزمینی در طی دوره ۱۳۵۷–۱۳۸۷ روندی افزایشی با میانگین ۱۴/۱ درصد داشته است. وی بیان می‌کند که فعالیت‌های زیرزمینی در ایران دارای حجم تقریبی ۳۱ هزار میلیارد ریال است. این امر باعث شده است که مقامات مالیاتی، پیوسته تحت فشار فزاینده برای تشخیص فرارکنندگان مالیاتی، به دنبال مالیات بیشتر و پیش‌بینی رفتارهای غیرمعمول فرارکنندگان مالیاتی باشند.

با توجه به اهمیت فرار مالیاتی در تصمیم‌گیری سرمایه‌گذاران و دولت، به ویژه سازمان امور مالیاتی، استفاده از فن‌ها و روش‌هایی که بتوانند موارد فرار مالیاتی را مشخص نمایند، برای سازمان امور مالیاتی اهمیت زیادی دارد. به این منظور، برای شناسایی فرار مالیاتی در کشورهای خارج، پژوهش‌گران و مقامات مالیاتی با کمک ابزارهای فنّاوری اطلاعات، اقدام به جمع‌آوری و مطابقت اطلاعات در این پایگاه‌های مجزا نموده تا عدم انطباق بین اطلاعات برای آن‌ها مشخص شده و بتوانند فرار مالیاتی را پیش‌بینی کنند (وو[۵] وهمکاران، ۲۰۱۲).

در ایران، اغلب پژوهشگران جهت پیش‌بینی فرار مالیاتی، از متغیرهای اقتصادی استفاده کرده‌اند (به عنوان نمونه، فلاحتی و همکاران، ۱۳۹۱؛ هادیان و تحویلی، ۱۳۹۲)؛ به بیان دقیق‌تر، پژوهش‌های اقتصادی مربوط به فرار مالیاتی با رویکرد کلان (از قبیل متغیرهایی مانند نرخ مالیات، پیچیدگی قوانین و مقررات، نبود سرمایه اجتماعی و تورم، بار مالیاتی، اندازه دولت، درآمد مصرف‌کننده، نرخ تورم و نرخ بیکاری) به بررسی فرار مالیاتی پرداخته‌اند، اما از دیدگاه خرد به این موضوع توجه نشده است. این در حالی است که استفاده از متغیرهای کلان اقتصادی نمی‌تواند در زمینه پیش‌بینی فرار مالیاتی در سطح خرد مورد استفاده قرار گیرد. افزون بر این، اغلب پژوهش های اقتصادی ، برای پیش بینی فرار مالیاتی از فن‌های اماری مانند رگرسیون استفاده کرده‌اند که مبتنی بر فرض وجود رابطه خطی بین متغیر ها است (آل‌بوسویلم، ۱۳۹۰؛ هادیان و تحویلی، ۱۳۹۲). روش‌های خطی به دلیل مفروضاتی که به عنوان پیش‌فرض‌های مورد نیاز برای انجام این آزمون‌ها وجود دارد و با توجه به اینکه رفتار متغیرها را خطی در نظر می‌گیرد، و با رفتار آن‌ها در دنیای واقعی متفاوت است، منجر به بهترین الگو برای پیش‌بینی فرارمالیاتی نمی‌گردد (به‌عنوان نمونه، مین[۶] و لی۷، ۲۰۰۵؛ موکامالا[۸] و همکاران، ۲۰۰۶؛ آلفارو[۹] و همکاران ۲۰۰۸؛ لی[۱۰] و تاو[۱۱]، ۲۰۱۰ را ببینید). به منظور رفع معایب بالا، پژوهش حاضر پرسش‌های زیر را مطرح می‌نماید:

۱. آیا استفاده از متغیرهای خرد و داده‌های حسابداری (به جای متغیرهای اقتصادی) منجر به پیش‌بینی بهتر فرار مالیاتی می‌شود؟

۲. آیا استفاده از روش «ریلیف» منجر به انتخاب بهتر متغیرهای اولیه مربوط به پیش‌بینی فرار مالیاتی می‌شود و پیش‌بینی فرار مالیاتی را بهبود می‌بخشد؟

۳. آیا اگر به جای استفاده از روش‌های آماری خطی، از روش‌های داده‌کاوی همراه با روش ریلیف استفاده شود، پیش‌بینی فرار مالیاتی بهتر انجام می‌شود؟

هدف این پژوهش، پاسخ به پرسش‌های بالاست. اهمیت این مطالعه آن است که با رویکرد جدیدی به پیش بینی فرار مالیاتی می‌پردازد و از دیدگاه خرد و داده‌های حسابداری (به جای داده‌های اقتصاد کلان) استفاده می کند. بنابراین، در پژوهش حاضر از روش‌های داده کاوی و ریلیف جهت پیش‌بینی فرار مالیاتی استفاده می‌شود و عملکرد

الگوریتم‌های مختلف درخت تصمیم با استفاده از داده‌های حسابداری با یکدیگر مقایسه شده و روش بهینه جهت پیش‌بینی فرار مالیاتی با استفاده از داده‌های حسابداری تعیین می‌گردد.

افزون‌بر اهمیت بحث بالا، یکی دیگر از موضوعاتی که در بحث فرار مالیاتی اهمیت اساسی پیدا می‌کند، هزینه جمع آوری اطلاعات لازم برای پیش‌بینی فرار مالیاتی است. بنابراین، اگر بتوان از بین متغیرهای اصلی انتخاب شده برای پیش‌بینی فرار مالیاتی، متغیرهایی که تاثیر بیشتری داشته یا حتی تاثیر معکوس در پیش‌بینی فرار مالیاتی دارند را مشخص نمود و از مدل حذف کرد، می‌توان فرار مالیاتی را کاهش داد و به کارایی اطلاعاتی بیشتری نیز دست یافت. در این پژوهش، این کار با استفاده از روش ریلیف انجام می‌شود. بنابراین، انتظار می‌رود این پژوهش به بهتر و کارا شدن فرآیندهای مربوط به کاهش فرار مالیاتی کمک کند و به گسترش مرزهای دانش و کاربردی در این زمینه نیز بینجامد.

ساختار مقاله حاضر بدین صورت است که در بخش اول به مبانی نظری فرار مالیاتی و داده‌کاوی پرداخته می‌شود. در بخش دوم، پیشینه پژوهش مورد بررسی قرار می‌گیرد. در ادامه، به توصیف روش پژوهش و متغیرهای پژوهش پرداخته می‌شود. سپس، روش‌های تجزیه و تحلیل داده‌ها و یافته‌های پژوهش مطرح می‌شود. سرانجام، نتیجه‌گیری، پیشنهادها و محدودیت‌های پژوهش بیان می‌گردد.

مبانی نظری

فرار مالیاتی

در ارتباط با فرار مالیاتی دو نظریه عمده مدل‌های نئوکلاسیکی[۱۲] و رویکرد نهادگرا[۱۳] وجود دارد. مکتب نئوکلاسیک در اوایل دهه ۱۹۷۰ میلادی با هدف ارائه الگوی کلان پول‌گرایان همراه با اصول موضوعه قوی‌تر به وجود آمد. این پژوهش بر مبنای تئوری مدل‌های نئوکلاسیکی استوار است. می‌توان نقطه شروع مدل نئوکلاسیکی را در مقاله مشهور «فرار از مالیات بر درآمد: یک تحلیل نظری» آلینگهام[۱۴] و ساندمو[۱۵] دانست. این مدل مبتنی بر روش‌شناسی اقتصاد جرم بکر[۱۶] (۱۹۶۸) بوده و سعی می‌کند با استفاده از عوامل اقتصادی به پیش‌بینی فرار مالیاتی بپردازد (امیدی‌پور و همکاران، ۱۳۹۴).

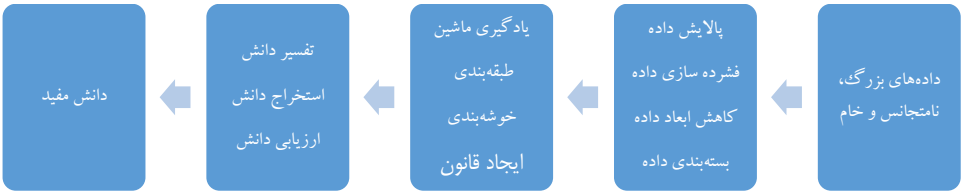
دو فرض کلیدی این مکتب عبارتند از: فرضیه انتظارات عقلایی و فرض تسویه دائم بازارها (شاکری، ۱۳۸۴). در این مدل، عوامل اجتماعی و روانی بر رفتار تمکین یا عدم تمکین مودی تاثیرگذار نیست. به بیانی دیگر، در این روش مودی بر اساس اصل به حداکثررسانی سود مورد انتظار اقدام به فرار مالیاتی می‌کند (امیدی‌پور و همکاران، ۱۳۹۴). در پرتو این نظریه، فعالان اقتصادی در بازار به صورت عقلایی رفتار می‌کنند. هدف آن‌هاحداکثر کردن مطلوبیت است و هر کسی به نسبت قدرت خریدش در بازار قدرت دارد. عقلانیت رفتاری، چنین معنی می‌شود که افراد بر طبق قوانین شناخته شده رفتار می‌کنند. بدون چنین فرضی، رفتار عقلایی قابل تعریف و تفسیر نخواهد بود (یوسفی، ۱۳۹۲).در طرف مقابل و در انتقاد به بی‌توجهی نسبت به عوامل غیراقتصادی تاثیرگذار بر فرار مالیاتی پژوهشگرانی چون آلم[۱۷] و مارتینزوازکوئیز[۱۸](۲۰۰۱)، گرخهانی[۱۹](۲۰۰۲)، تورگلر[۲۰] (۲۰۰۳) و نره[۲۱] (۲۰۰۴) کارکرد عواملی چون فرهنگ، هنجارهای عرفی، اخلاق و ... را مدنظر قرار دادند (امیدی پور و همکاران، ۱۳۹۴).

نهادها، قوانین بازی در جامعه هستند یا به صورت رسمی‌تر، قیودی هستند که از طرف انسان‌ها وضع شده‌اند تا رفتار متقابل افراد را شکل دهند. نهادها همچنین، اقدام دسته‌جمعی جهت کنترل اقدام رفتار فردی تعریف می‌شود (یوسفی، ۱۳۹۲).

داده‌کاوی و درخت تصمیم

داده‌کاوی عبارت است از فرآیند اکتشاف الگو و روندهای منظم و پنهان در داده‌های بزرگ و توزیع شده، با استفاده از مجموعه وسیعی از الگوریتم‌های مبتنی بر علوم ریاضی و آمار. به بیان دیگر، داده‌کاوی یک سیستم واسط بین یک مجموعه از داده و یک برنامه است که داده را به دانش قابل استفاده تبدیل می‌کند. لینف[۲۲] و بری[۲۳] (۲۰۱۱) داده‌کاوی را به دو طبقه داده‌کاوی مستقیم (نظارت شده) و داده‌کاوی غیرمستقیم (غیرنظارتی) تقسیم می‌کند. داده‌کاوی نظارت شده، به توصیف و طبقه‌بندی داده‌ها با توجه به داده خروجی می‌پردازد درحالیکه در داده‌کاوی غیرنظارتی بدون توجه به داده خروجی به دنبال الگویی برای داده‌ها می‌پردازد (خان[۲۴] و همکاران، ۲۰۱۳). دانش موجود در داده‌ها می‌تواند توسط هر کدام از این دو طبقه و یا ترکیبی از آن‌ها استخراج گردد. مطابق با نظر یائو[۲۵] (۲۰۰۱) داده‌کاوی یک رشته جدید بوده و تمرکز انبوه پژوهش‌های صورت‌گرفته دراین زمینه بر

روی توسعه و بهبود الگوریتم‌های موجود و ارزیابی دانش کشف شده در یک فرآیند تک مرحله‌ای است. شکل ۱ فرآیند داده‌کاوی را به صورت شماتیک نشان می‌دهد.



شکل ۱. فرآیند داده‌کاوی در عمل
اقتباس از پال (۲۰۰۴، ۷)

داده‌کاوی شامل الگوریتم‌های متعددی است. برخی از مهمترین آن‌ها که در این مقاله مورد استفاده قرار می‌گیرند، در زیر شرح داده شده می‌شوند. درخت تصمیم یکی از مشهورترین و قدیمی‌ترین روش‌های ساخت الگوی دسته‌بندی است. در الگوریتم‌های دسته‌بندی مبتنی بر درخت تصمیم، دانش خروجی به صورت یک درخت از حالات مختلف مقادیر ویژگی‌ها ارائه می‌شود (صنیعی آباده و محمودی، ۱۳۹۴). نمایش دانش به شکل درخت باعث شده است که دسته‌بندی‌های مبتنی بر درخت تصمیم کاملاً قابل تفسیر باشند (صنیعی آباده و محمودی، ۱۳۹۴). به‌گونه کلی، نقاط قوت درخت تصمیم را می‌توان در فهم آسان مدل ایجاد شده، توانایی ایجاد یک مدل بر مبنای قوانین، عدم نیاز به محاسبات پیچیده، توانایی کار با داده‌های پیوسته و گسسته، اولویت بندی متغیرها، حذف مقایسه‌های غیرضروری و عدم نیاز به تخمین تابع توزیع دانست (جعفری و آذر، ۱۳۹۲ و تسای[۲۶] و چیو[۲۷]، ۲۰۰۹؛ صنیعی آباده و محمودی، ۱۳۹۴). از سوی دیگر، در مواردی که هدف از یادگیری، تخمین تابعی با مقادیر پیوسته است، با تعداد دسته‌های زیاد و نمونه‌های آموزشی کم روبرو هستیم و در صورتی که درخت بزرگ باشد، به دلیل امکان انباشته شدن خطاها از سطحی به سطحی دیگر، درخت تصمیم مناسب نیست (صنیعی‌آباده و محمودی، ۱۳۹۴).

الگوریتم جنگل تصادفی[۲۸]، یک الگوریتم گروهی با مجموعه‌ای از درختان تصمیم است. دقت طبقه بندی روش جنگل تصادفی، با ساخت مجموعه ای از درختان و رأی‌گیری بین آن‌ها برای به دست آوردن رده‌ای با بیشترین تعداد رأی است. الگوریتم جنگل تصادفی می تواند دقت پیش بینی را نسبت به درخت طبقه بند فردی افزایش دهد. در درخت فردی، با تغییرات کوچک در مجموعه آموزش، بی‌ثباتی به وجود می‌آید که باعث اختلال در دقت پیش بینی در نمونه آزمایشی می‌شود. اما گروهی بودن الگوریتم جنگل تصادفی، باعث سازگاری با تغییرات می شود و بی‌ثباتی را از بین می برد (امینی‌خویی و عبداله‌پوری، ۱۳۹۶).

الگوریتم ریشه تصمیم[۲۹]، یک درخت تصمیم است که در آن یک گره ریشه‌ای به دو ساقه تقسیم می‌شود. برای هر ویژگی در داده‌های ورودی یک ریشه تصمیم ساخته می‌شود. الگوریتم ریشه تصمیمی یک الگوریتم ساده است و تنها یک مقایسه در هر ریشه تصمیم برای هر نمونه انجام می‌شود. بنابراین، زمان آزمون برای ریشه تصمیم کوتاه است (هو[۳۰] و همکاران، ۲۰۰۸).

الگوریتم کاهش خطای هرس[۳۱] بوسیله کوئینلن[۳۲] در سال ۱۹۸۷ معرفی شد که پس از آن به عنوان مجموعه قوانین یادگیری بوسیله پژوهشگران مد نظر قرار گرفت. این الگوریتم، یکی از ساده‌ترین استراتژی‌های هرس است که نیازمند یک مجموعه مجزا به عنوان مجموعه هرس دارد که این موضوع اغلب به عنوان ضعف این الگوریتم در نظر گرفته می‌شود. با وجود این، این الگوریتم یک روش پایه است که عملکرد سایر الگوریتم‌ها با آن مورد مقایسه قرار می‌گیرد (الوما[۳۳] و کاریاینن[۳۴]، ۲۰۰۱).

الگوریتم درخت تصمیم تصادفی[۳۵]، به طور تصادفی ویژگی‌های هر گره را انتخاب می‌کند. یک ویژگی گسسته در هر مسیر نمی‌تواند بیش از یک بار مورد استفاده قرار گیرد، اما ویژگی‌های پیوسته می‌توانند چندین بار تکرار شوند ولی با مقادیر متفاوت در هربار تکرار. سرانجام، در طی هر طبقه‌بندی، احتمال درخت میانگین گرفته می‌شود، تا پیش‌بینی نهایی بدست آید (بیرانت[۳۶]، ۲۰۱۶).

الگوریتم LMT[۳۷] دو تکنیک محبوب درخت تصمیم و رگرسیون خطی لجستیک را ترکیب می‌کند. در هر تکرار، رگرسیون‌های لجستیک گره والد به گره فرزند انتقال پیدا می‌کند و گره ساقه همه مدل‌ها را برای تخمین احتمال هر کلاس جمع می‌کند. یک فرآیند هرس، نیز به مدل اعمال می‌شود تا قابلیت تعمیم مدل را بالا ببرد (بیرانت، ۲۰۱۶).

الگوریتم J48 ، گسترش یافته الگوریتم ID3[۳۸] است و احتمالاً منجر به ایجاد درخت تصمیم کوچک‌تری می‌شود. این الگوریتم از رویکرد غلبه و تقسیم که بوسیله هانت[۳۹] و همکاران (۱۹۹۶) ابداع شده است، برای گسترش درخت تصمیم استفاده می‌کند (بهارگاوا[۴۰] و همکاران، ۲۰۱۳).

روش ریلیف

روش انتخاب متغیر ریلیف ازجمله روش‌های انتخاب متغیرهای پیش‌بین مبتنی بر معیار فاصله است. ریلیف از بین داده‌های آموزشی یک نمونه را به‌صورت تصادفی انتخاب می‌کند و سپس فاصله اقلیدسی آن نمونه تا نزدیک‌ترین نمونه از طبقه مشابه و نزدیک‌ترین نمونه از طبقه متفاوت را به دست می‌آورد و سپس این فاصله‌ها را برای به‌روز کردن وزن هر متغیر به کار می‌برد. درنهایت، الگوریتم آن دسته از متغیرهایی را انتخاب می‌کند که وزن آن‌ها از یک حد آستانه‌ی از پیش تعریف‌شده به‌وسیله کاربر، بیشتر است (آتیا[۴۱]، ۲۰۰۱). افزون‌بر این، ریلیف ازجمله روش‌های انتخاب ویژگی فیلتر[۴۲]است که اساساً با رتبه‌بندی نزولی متغیرها عمل می‌کند و به علت سادگی و مؤثر بودن در افزایش دقت طبقه‌بندی، در بسیاری از موارد استفاده می‌شود.

پیشینه

آل‌بوسویلم (۱۳۹۰) به بررسی فرار مالیاتی در ایران طی سال‌های ۱۳۵۷-۱۳۸۷ پرداخته است. وی با استفاده از دو روش علل چندگانه- شاخص چندگانه و تقاضای نقد، حجم اقتصاد زیرزمینی و فرار مالیاتی را محاسبه نموده است. نتایج پژوهش وی نشان داد که

اندازه نسبی فعالیت‌های زیرزمینی در طی دوره مورد پژوهش روندی افزایشی با میانگین ۱۴/۱ درصد داشته است. همچنین، نتایج حاصل از روش تقاضای نقد نشان داد که فعالیت‌های زیرزمینی در ایران، دارای حجم تقریبی ۳۱ هزار میلیارد ریال و فرار مالیاتی دارای میانگین ۲ هزار میلیارد ریال است.

عبداله‌میلانی و اکبرپور روشن (۱۳۹۱) نیز به تخمین حجم اقتصاد غیررسمی و فرار مالیاتی ناشی از آن و تحلیل روند آن‌ها در ایران، به اتکای روش تخمین تابع تقاضای پول با استفاده از الگوی ARDL پرداختند. نتایج این پژوهش نشان داد که حجم اقتصاد غیررسمی و فرار مالیاتی ناشی از آن با وجود نوساناتی در طی دوره ۱۳۷۰-۱۳۸۹، روندی افزایشی داشته است.

هادیان و تحویلی (۱۳۹۲) نیز به بررسی و شناسایی عوامل مؤثر بر فرار مالیاتی در اقتصاد ایران پرداختند. آن‌ها با استفاده از یک الگوی خودتوضیح با وقفه‌های گسترده برای دوره زمانی ۱۳۵۰ تا ۱۳۸۶ به بررسی تأثیر متغیرهای نرخ مالیات، پیچیدگی قوانین و مقررات، نبود سرمایه اجتماعی و تورم بر فرار مالیاتی در اقتصاد ایران پرداختند. نتایج حاصل از این پژوهش نشان داد که در بلندمدت هر چهار متغیر بیان‌شده دارای رابطه مثبت و معنادار با فرار مالیاتی بودند، اما در کوتاه‌مدت تورم نقش با اهمیتی در فرار مالیاتی نداشت.

دستگیر و غریبی (۱۳۹۴) به بررسی و استفاده از روش‌های داده‌کاوی برای ارتقای عملکرد تشخیص فرار مالیاتی در بورس اوراق بهادار تهران پرداختند. نمونه مورد بررسی این پژوهش شامل ۱۲۵ شرکت در دوره زمانی ۱۳۸۳ تا ۱۳۹۰ و روش مورد استفاده آن‌ها قواعد وابستگی و به‌کارگیری الگوریتم پیشینار بود. نتایج پژوهش آن‌ها نشان داد که روش داده‌کاوی مبتنی بر قواعد وابستگی با ایجاد دو مدل با درصد صحت ۹۱٪ بر روی داده‌های آموزش، با صحت ۸۸٪ بر روی داده‌های اعتبارسنجی و با درصد صحت ۸۶٪ بر روی داده‌های آزمون توانسته است موفق به تشخیص فرار مالیاتی شود.

سهرابی و همکاران (۱۳۹۴) نیز به بررسی ارزیابی مالیات عملکرد شرکت‌ها و تحلیل

روندهای مالیاتی با استفاده از الگوریتم‌های داده‌کاوی پرداختند. آن‌ها با استفاده از درخت تصمیم، شرکت‌ها را به سه گروه پرریسک، باریسک مالیاتی متوسط و کم ریسک طبقه‌بندی کردند. نتایج آن‌ها نشان‌دهنده دقت ۸۰٪ الگوریتم داده‌کاوی مورد نظر بود.

حمیدی و همکاران (۱۳۹۴) به بررسی جایگاه جرایم مالیاتی در جلوگیری از فرار در نظام مالیات بر ارزش افزوده به صورت موردی در استان قزوین پرداختند. بدین منظور ابتدا جایگاه جرایم مالیاتی و عوامل موثر بر کاهش فرار مالیاتی با استفاده از ادبیات پژوهش، نظر خبرگان و مودیان مالیات بر ارزش افزوده شناسایی شد. سپس با استفاده از تکنیک دیمتیل فازی و ANP، رابطه‌های ممکن و شدت تاثیر روابط و اهمیت هر یک از عوامل مشخص گردید. نتایج پژوهش نشان داد که جرایم مالیاتی از دیگر عوامل موثر بر کاهش فرار مالیاتی اثر می‌پذیرد، ولی بر هیچ‌یک از شاخص‌ها اثر نمی‌گذارد.

سامعی‌راد و شاه‌بهرامی (۱۳۹۵) با استفاده از الگوهای پردازش موازی به بهبود کارایی الگوریتم‌های تشخیص تقلب مالیاتی پرداختند. نتایج پژوهش آنان نشان داد که با استفاده از الگوهای پردازش موازی می‌توان کارایی برنامه‌های کشف تقلب‌های مالیاتی را به طور قابل ملاحظه‌ای بهبود بخشید.

تقوی‌فرد و همکاران (۱۳۹۶) به بررسی تحلیل آینده‌نگر تشخیص فرار مالیاتی مودیان مالیات بر ارزش افزوده با استفاده از الگوریتم‌های طبقه‌بندی و خوشه‌بندی پرداختند. دوره زمانی پژوهش حاضر سال‌های ۱۳۸۸ تا ۱۳۹۳ بود. الگوریتم‌های مورد بررسی شامل سه روش Naive Bayes، درخت تصمیم و KNN بود. همچنین الگوریتم‌های خوشه‌بندی شامل دو روش میانگینK [۴۳] و کی‌مدوئیدس [۴۴] بود. نتایج پژوهش نشان داد که از میان الگوریتم‌های طبقه‌بندی، روش درخت تصمیم با خطای ۰/۰۵ و خطای آموزش (اعتبارسنجی) ۰/۱۷ نتیجه بهتری کسب کرد.

اخیراً دهقان و همکاران (۱۳۹۷) به بررسی توضیح فرار مالیاتی با استفاده از تئوری چشم‌انداز پرداختند. آن‌ها با درنظر گرفتن تئوری‌های چشم‌انداز و تئوری مطلوبیت انتظاری در سناریوهای مختلف، بدین نتیجه دست یافتند که تفاوت بسیار زیادی در زمینه جریمه

فرار مالیاتی بین این دو تئوری وجود دارد. همچنین، جرایم به دست آمده با استفاده از تئوری مطلوبیت انتظاری بسیار بزرگ‌تر بوده و با افزایش احتمال حسابرسی در هر دو نظریه، نرخ جریمه مالیاتی کاهش می‌یابد. افزون بر این، در سناریوهای ترکیبی افزایش احتمال حسابرسی شدن و افزایش وزن احتمال، محاسبات با استفاده از تئوری چشم انداز، نشان ازکاهش نرخ جرایم مالیاتی و همچنین حساسیت جرایم به وزن احتمال حسابرسی دارند.

وو و همکاران (۲۰۱۲) با استفاده از داده‌کاوی درصدد افزایش عملکرد کشف فرار مالیاتی برآمدند. آن‌ها یک چارچوب نظارتی ارائه کردند تا مغایرت‌های بین گزارش‌های ارزش‌افزوده، را که نیازمند حسابرسی بیشتری بود، مشخص نمایند. نتایج این پژوهش نشان داد که تکنیک داده‌کاوی مورد استفاده به‌درستی توانست میزان تشخیص فرار مالیاتی را افزایش دهد و درنتیجه با به‌کارگیری مؤثر آن می‌توان زیان‌های ناشی از فرار مالیاتی را کاهش داد.

رحیمی‌کیا و همکاران (۲۰۱۷) به کشف فرار مالیاتی با استفاده از سیستم‌های هوشمند ترکیبی و داده‌کاوی در صنایع مواد غذایی و نساجی بورس اوراق بهادار تهران پرداختند. نتایج آن با رگرسیون لجستیک مورد مقایسه قرار گرفت. نتایج حاصل از مقایسه شبکه عصبی و رگرسیون لجستیک نشان داد که استفاده از شبکه عصبی دارای دقت بالاتری بوده و این تفاوت از لحاظ آماری معنادار است.

اخیراً گوماگیاس[۴۵] و همکاران (۲۰۱۸) به بررسی درک رفتار فرار مالیاتی شرکت‌های ریسک‌گریز با استفاده از الگوریتم دیپ‌کیولرنینگ [۴۶] پرداختند. نتایج پژوهش آن‌ها نشان داد که مدل آن‌ها قادر به مشخص کردن رفتار مورد انتظار فرار مالیاتی است. همچنین، مدل فوق قادر به محاسبه درجه ریسک‌گریزی یک واحد تجاری با در نظر گرفتن تخمین‌های تجربی فرار مالیاتی و ارزیابی سیاستهای مالیاتی با درآمدهای موردانتظار بود.

دیدیمو[۴۷] و همکاران (۲۰۱۸) به توصیف سیستم پشتیبانی تصمیم برای کشف فرار مالیاتی بر مبنای زبان دیداری و تکنیک‌های پیشرفته مصورسازی شبکه پرداختند. این

سیستم به استفاده‌کنندگان از آن، این امکان را می‌دهد که گراف‌های زیرمجموعه نمونه‌های مشکوک را با توجه به تطابق با الگوهای موجود رسم کرده و باعث ادغام نتایج و ایجاد شاخص‌هایی برای طبقه‌بندی کردن مالیات‌دهندگان بر مبنای ریسک مالی می‌شود. نتایج نشان داد که سیستم پیشنهادی موثر می‌باشد.

اخیراً زنگنه و همکاران (۲۰۱۸) نیز به بررسی عوامل موثر بر فرار مالیاتی یا استفاده از روش فازی دیماتل پرداختند. نتایج این پژوهش نشان داد که در میان عوامل موثر، کمبود قانون‌گذاران، نفوذ موسسات غالب که مالیات پرداخت نمی‌کنند و میزان زیاد معافیت‌ها بالاترین اثر را در فرار مالیاتی دارند.

نقد پژوهش‌های پیشین

بررسی پژوهش‌های انجام‌شده در داخل و خارج از کشور نشان می‌دهد که در اکثر این مطالعات فرار مالیاتی با استفاده از چند متغیر محدود اقتصادی انجام شده و اهمیتی به نحوه انتخاب متغیرها داده نشده است. در صورتی‌که یکی از مباحث مهم در این خصوص، انتخاب متغیرهای مرتبط با موضوع است که بتوان با استفاده از کمترین تعداد متغیر، بهترین پیش‌بینی را مشخص کرد. افزون براین، در هیچ‌کدام از پژوهش‌های انجام‌شده، اهمیت متغیرهای پیش‌بین مورد مطالعه قرار نگرفته است. به‌عبارت دیگر اینکه کدام متغیرها بیشترین تأثیر را در پیش‌بینی فرار مالیاتی داشته‌اند، مشخص نشده است. همچنین بیشتر پژوهش‌های صورت گرفته در زمینه فرار مالیاتی در سطح کلان بوده و از متغیرهای اقتصادی استفاده کرده‌اند و کمتر به بررسی اطلاعات حسابداری شامل متغیرهای مالی، و استفاده از گزارش حسابرس در این باره پرداخته‌اند.

ازاین‌رو، می‌توان نقطه قوت و متمایزکننده پژوهش حاضر نسبت به پژوهش‌های داخلی و خارجی را در سه مورد خلاصه کرد: اول، این پژوهش با بررسی ادبیات تحقیق، اقدام به شناسایی متغیرهایی که در پژوهش‌های پیشین به‌عنوان متغیرهای مؤثر در کشف فرار مالیاتی شناخته شده‌اند، می‌نماید و به ویژه اطلاعات حسابداری را مورد استفاده قرار می‌دهد. دوم، پس از شناسایی متغیرهای مؤثر در پژوهش‌های پیشین، با استفاده از روش

ریلیف اقدام به انتخاب متغیرهای پیش‌بین بهینه در خصوص فرار مالیاتی شرکت‌های بورس اوراق بهادار تهران می‌کند. سوم، الگوهای مختلف درخت تصمیم را با یکدیگر مقایسه می‌کند و دقت و عملکرد پیش‌بینی آن‌ها با یکدیگر نیز مورد مقایسه قرار می‌گیرد تا الگوی بهینه فرار مالیاتی مشخص گردد.

فرضیه‌ها

بر اساس مبانی نظری و پیشینه تحقیق ارایه‌شده در این مطالعه، فرضیه‌های زیر مطرح می‌شوند:

فرضیه اول: بین دقت پیش‌بینی فرار مالیاتی، با استفاده از متغیرهای پیش‌بین بهینه انتخاب شده از میان اطلاعات حسابداری در روش ریلیف و استفاده از کلیه متغیرهای پیش‌بین در درخت تصمیم ریشه تصمیم، تفاوت معناداری وجود دارد.

فرضیه دوم: بین دقت پیش‌بینی فرار مالیاتی، با استفاده از متغیرهای پیش‌بین بهینه انتخاب شده از میان اطلاعات حسابداری در روش ریلیف و استفاده از کلیه متغیرهای پیش‌بین در درخت تصمیم کاهش خطای هرس، تفاوت معناداری وجود دارد.

فرضیه سوم: بین دقت پیش‌بینی فرار مالیاتی، با استفاده از متغیرهای پیش‌بین بهینه انتخاب شده از میان اطلاعات حسابداری در روش ریلیف و استفاده از کلیه متغیرهای پیش‌بین در درخت تصمیم تصادفی، تفاوت معناداری وجود دارد.

فرضیه چهارم: بین دقت پیش‌بینی فرار مالیاتی، با استفاده از متغیرهای پیش‌بین بهینه انتخاب شده از میان اطلاعات حسابداری در روش ریلیف و استفاده از کلیه متغیرهای پیش‌بین در درخت تصمیم جنگل تصادفی، تفاوت معناداری وجود دارد.

فرضیه پنجم: بین دقت پیش‌بینی فرار مالیاتی، با استفاده از متغیرهای پیش‌بین بهینه انتخاب شده از میان اطلاعات حسابداری در روش ریلیف و استفاده از کلیه متغیرهای پیش‌بین در درخت تصمیم LMT، تفاوت معناداری وجود دارد.

فرضیه ششم: بین دقت پیش‌بینی فرار مالیاتی، با استفاده از متغیرهای پیش‌بین بهینه انتخاب شده از میان اطلاعات حسابداری در روش ریلیف و استفاده از کلیه متغیرهای پیش‌بین در درخت تصمیم J48، تفاوت معناداری جود دارد.

روش‌شناسی

این پژوهش از نوع پژوهش‌های کمی است و جنبه کاربردی دارد. طرح پژوهش آن از نوع شبه‌تجربی و با استفاده از رویکرد پس‌رویدادی (از طریق اطلاعات گذشته) است. از روش پس‌رویدادی زمانی استفاده می‌شود که پژوهشگر پس از وقوع رویدادها به بررسی موضوع می‌پردازد. افزون بر این، امکان دستکاری متغیرهای مستقل وجود ندارد (نمازی، ۱۳۹۵). جامعه آماری این پژوهش، کلیه شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران است. نمونه آماری پژوهش، آن دسته از شرکت‌های مربوط به جامعه آماری خواهند بود که در بازه ۱۳۸۴-۱۳۹۴ در بورس فعالیت داشته و اطلاعات مربوط به صورت‌های مالی و یادداشت‌های توضیحی آن‌ها موجود باشد. فعالیت‌های از نوع سرمایه‌گذاری نداشته باشند و در زمینه استخراج نفت و نیز کشاورزی (به دلیل ماده ۱۴۱ قانون مالیاتهای مستقیم) فعالیت نکنند. پایان سال مالی آن‌ها منتهی به پایان اسفند باشد و طی سال‌های ۱۳۸۴ تا ۱۳۹۴، تغییر سال مالی نداده باشد و اطلاعات مالی مورد نیاز برای انجام این پژوهش را در دوره زمانی ۱۳۸۴ الی ۱۳۹۴ به‌طور کامل ارائه کرده باشند. با توجه به بررسی‌های انجام شده، تعداد ۱۰۸۱ سال- شرکت در دوره زمانی ۱۳۸۴ الی ۱۳۹۴ حائز شرایط فوق بوده و مورد بررسی قرار گرفته است.

در این پژوهش برای جمع‌آوری داده‌ها و اطلاعات از روش‌های کتابخانه‌ای و میدانی استفاده می‌شود. مبانی نظری پژوهش از کتب، مجلات و سایت‌های تخصصی فارسی و لاتین مانند پایگاه مجلات تخصصی نور ، پایگاه اطلاعات علمی جهاد دانشگاهی، پایگاه استنادی علوم جهان اسلام، science direct، springer، J store و proquest و داده‌های مالی مورد نیاز با مراجعه به سایت سازمان بورس اوراق بهادار تهران، صورت‌های مالی شرکت‌ها و هم‌چنین با استفاده از نرم‌افزار ره‌آورد نوین گردآوری شد.

متغیرها

متغیرهای استفاده شده در این پژوهش برای پیش‌بینی فرار مالیاتی را می‌توان به سه دسته متغیرهای مربوط به اعداد صورتهای مالی، متغیرهای نسبتهای مالی و متغیر مربوط به بند مالیاتی در گزارش حسابرسی تقسیم کرد. در این خصوص از بین تمام متغیرهای شناسایی‌شده، متغیرهایی که بیشتر در ادبیات مربوط به پیش‌بینی فرار مالیاتی استفاده شده و داده‌های مورد نیاز برای سنجش آن‌ها از طریق پایگاه‌های اطلاعاتی سازمان بورس و اوراق بهادار و همچنین نرم‌افزار ره‌آورد نوین در دسترس بود، انتخاب شدند. این متغیرها به شرح زیر هستند:

اعداد صورت‌های مالی شامل سود عملیاتی، سود خالص، فروش خالص، رشد فروش خالص، مجموع دارایی‌ها، رشد دارایی‌ها، حقوق صاحبان سهام، رشد حقوق صاحبان سهام، رشد سود عملیاتی، رشد سود خالص و مالیات ابرازی می‌باشد. این متغیرها در پژوهش‌های گالمور و لابرو (۲۰۱۵)، لیزوسکی[۴۸] (۲۰۱۰)، مهرانی و سیدی (۱۳۹۳)، دستگیر و غریبی (۱۳۹۴) نیز به‌منظور پیش‌بینی فرار مالیاتی استفاده شده است.

نسبت‌های مالی استفاده شده در این پژوهش به پنج دسته نسبت‌های نقدینگی، نسبت‌های سودآوری، نسبت‌های فعالیت، نسبت‌های اهرمی و نسبت‌های ارزش بازار تقسیم می‌شوند. نسبت‌های نقدینگی شامل نسبت جاری، نسبت آنی و نسبت خالص سرمایه در گردش به دارایی‌ها می‌باشد. از این نسبت‌ها در پژوهش‌های مو (۲۰۰۳)، باقرپور ولاشانی و همکاران (۱۳۹۱)، دستگیر و غریبی (۱۳۹۴)، به‌منظور پیش‌بینی فرار مالیاتی استفاده شده است. نسبت‌های سودآوری مورد استفاده در این پژوهش شامل بازده دارایی‌ها، نسبت حاشیه سود ناخالص، نسبت حاشیه سود عملیاتی، نسبت حاشیه سود خالص، سود عملیاتی به مجموع دارایی‌ها، رشد نسبت سود عملیاتی به مجموع دارایی‌ها است. از این نسبت‌ها در پژوهش‌های گالمور و لابرو (۲۰۱۵)، مهرانی و سیدی (۱۳۹۳)، دستگیر و غریبی (۱۳۹۴)، رحیمی‌کیا و همکاران (۱۳۹۴) به‌منظور پیش‌بینی فرار مالیاتی و اجتناب مالیاتی استفاده شده است. نسبت‌های فعالیت مورد استفاده در این پژوهش شامل گردش موجودی کالا، نسبت گردش دارایی‌های ثابت، گردش مجموع دارایی‌ها است. از این نسبت‌ها در

پژوهش‌های مو (۲۰۰۳) و رحیمی‌کیا و همکاران (۱۳۹۴) به‌منظور پیش‌بینی فرار مالیاتی استفاده شده است. نسبت‌های اهرمی مورد استفاده در این پژوهش عبارتند از: نسبت بدهی، نسبت بدهی به حقوق صاحبان سهام و نسبت توانایی پرداخت بهره (نسبت پوشش هزینه-های مالی). از این نسبت‌ها در پژوهش‌های گالمور و لابرو (۲۰۱۵)، باقرپور ولاشانی و همکاران (۱۳۹۱)، پورحیدری و سروستانی (۱۳۹۱)، مهرانی و سیدی (۱۳۹۳)، دستگیر و غریبی (۱۳۹۴) و رحیمی‌کیا و همکاران (۱۳۹۴)، تیلور و ریچاردسون (۲۰۱۳)، ایزگیارتا[۴۹] (۲۰۱۵) به‌منظور پیش‌بینی فرار مالیاتی استفاده شده است. این نسبت‌ها شامل ارزش بازار، نسبت ارزش بازار به ارزش دفتری سهام و سود هر سهم می‌باشد. از این نسبت‌ها در پژوهش‌های گراهام و همکاران (۲۰۱۴)، گالمور[۵۰] و لابرو[۵۱] (۲۰۱۵)، باقرپور ولاشانی و همکاران (۱۳۹۱)، پورحیدری و سروستانی (۱۳۹۱)، مهرانی و سیدی (۱۳۹۳)، دستگیر و غریبی (۱۳۹۴)، رضایی و جعفری نیارکی (۱۳۹۴) به‌منظور پیش‌بینی فرار مالیاتی نیز استفاده شده است.

گزارش حسابرسی نیز به عنوان یک متغیر مورد مطالعه قرار می‌گیرد. گزارش حسابرسی به عنوان یک متغیر دو وجهی در نظر گرفته می‌شود. که در صورت وجود بند مالیاتی در گزارش حسابرسی عدد یک، و نبود بند مالیاتی در گزارش حسابرسی عدد صفر منظور می‌گردد. از این متغیر در پژوهش باقرپور ولاشانی و همکاران (۱۳۹۱) نیز استفاده شده است.

متغیر وابسته در این پژوهش، فرار مالیاتی است که با توجه به درصد اختلاف سود مشمول مالیات ابرازی و سود مشمول مالیات قطعی شرکت شناسایی می‌گردد. طبق ماده ۱۹۴ قانون مالیات‌های مستقیم مصوب ۱۳۹۴/۴/۳۱ اگر سود مشمول مالیات قطعی با سود مشمول مالیات ابرازی واحد تجاری بیش از ۱۵٪ اختلاف داشته باشد، واحد تجاری دارای فرار مالیاتی است. از این متغیر در پژوهش دستگیر و غریبی (۱۳۹۴) نیز استفاده شده است.

روش تجزیه و تحلیل داده‌ها

انتخاب و استخراج متغیرهای مناسب جهت رسیدن به بهترین نتیجه در پیش‌بینی، از مباحث چالش‌برانگیز در دو دهه اخیر بوده است. از دیدگاه تئوری، یادگیری بر اساس

تعداد متغیرهای پیش‌بین بیشتر باعث می‌شود تا دقت پیش‌بینی بالا رود. با وجود این، شواهد تجربی بیانگر آن است که این امر همواره صادق نیست؛ زیرا تمام متغیرها، برای تشخیص و پیش‌بینی مهم نیستند و یا برخی از آن‌ها به‌طورکلی در پیش‌بینی نامربوط هستند (لیندنبام[۵۲] و همکاران، ۲۰۰۴). با توجه به این که عامل‌های بسیاری ازجمله کیفیت داده‌ها در موفقیت یک الگوریتم یادگیری مؤثر است، اگر داده‌ها حاوی متغیرها و یا اطلاعات تکراری و نامربوط باشند و یا حاوی اطلاعات دارای پارازیت و نامطمئن باشند، کسب دانش از آن داده‌ها مشکل می‌شود (هال[۵۳]، ۲۰۰۰). افزون بر این، کاهش تعداد متغیرهای پیش‌بین نامربوط یا اضافی، افزون بر کاهش زمان اجرای الگوریتم‌های آموزشی، به مفهومی عمومی‌تر منجر می‌شود. سایر مزایای بالقوه انتخاب و استخراج متغیرهای پیش‌بین شامل تسهیل درک و تجسم داده‌ها، کاهش الزامات اندازه‌گیری و ذخیره اطلاعات، کاهش اضافه‌بار ابعاد[۵۴] و بهبود عملکرد پیش‌بینی و فراهم کردن بینش بهتر در مورد مفهوم زیربنایی از طبقه‌بندی دنیای واقعی است (تسای، ۲۰۰۹). در این پژوهش از روش ریلیف جهت انتخاب متغیرهای پیش‌بین بهینه استفاده شده است. علت انتخاب متغیر ریلیف، مطابق با پژوهش نمازی و همکاران (۱۳۹۵) سادگی و مؤثر بودن در افزایش دقت طبقه‌بندی داده‌ها می‌باشد.

روش هلد آوت[۵۵] یکی از مهم‌ترین روش‌های تعیین روایی پژوهش در پژوهش‌های داده‌کاوی است که در آن داده‌ها به دو دسته مجموعه آموزشی و آزمایشی تقسیم می‌شوند. اما این روش یک تخمین‌گر بدبینانه[۵۶] است زیرا تنها بخشی از داده‌ها برای آموزش به روش پیش‌بینی ارائه شده است و هر چه تعداد نمونه بیشتری برای مجموعه آزمایشی خارج شود، سوگیری برآورد بیشتر می‌شود. از طرفی، نمونه‌های آزمایشی کوچک‌تر (با تعداد کمتر) به معنای بیشتر بودن فاصله اطمینان دقت است. بنابراین، روش مزبور روش مناسبی نیست (کوهاوی[۵۷]، ۱۹۹۵). به منظور برطرف کردن بخشی از این مشکلات می‌توان از روش روایی متقابل[۵۸] استفاده نمود که نتایج بسیاری از پژوهش‌های

انجام شده حاکی از عملکرد بهتر این روش است (آرلوت[۵۹] و سلیسه ۶۰، ۲۰۱۰: ۴۰). در این روش در نهایت از تمام نمونه‌ها هم به‌عنوان داده‌های آموزشی و هم به‌عنوان داده‌های آزمایشی استفاده می‌شود. افزون براین، استفاده از این روش، از بروز مشکل بیش‌برازش[۶۱] و مشکلات مربوط به نتایج برون‌نمونه‌ای[۶۲] جلوگیری می‌کند (آرلوت و سلیسه، ۲۰۱۰). از طرف دیگر، این روش برای برآورد نرخ خطای واقعی کاملاً قابل اتکا و کافی است (هو، ۲۰۱۰). از این‌رو، در پژوهش حاضر به‌منظور بررسی تعمیم‌پذیری پیش‌بینی‌های انجام شده، از روایی متقابل ۱۰ بخشی[۶۳] استفاده شده است.

یافته‌ها

آمار توصیفی

آمار توصیفی متغیرهای پژوهش در جدول۱ ارائه شده است. نتایج جدول۱ نشان می‌دهد که در شرکت‌های مورد مطالعه، به طور متوسط نیمی از شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران دارای فرار مالیاتی هستند. میانگین نسبت بدهی به حقوق صاحبان سهام حدود ۲ است. بدین معنی که تامین مالی شرکت بیش از آن‌که از طریق صاحبان سهام باشد، از طریق بدهی انجام شده است. همچنین، به طور متوسط کمتر از ۴۰٪ حسابرسان، به بیان مشکل مالیاتی شرکت‌ها در گزارش حسابرسی پرداخته‌اند. میانگین نسبت بدهی نشان دهنده آن است که بیش از نیمی از دارایی‌های شرکت‌ها از محل بدهی تامین شده است. نسبت خالص سرمایه در گردش به جمع دارایی‌ها برابر با ۰/۱۱۶ می‌باشد که بدین معناست که بیشتر دارایی‌های شرکت از نوع دارایی‌های بلندمدت است. از سوی دیگر نسبت حاشیه سود ناخالص برابر با ۰/۲۶۶ است که نشان‌دهنده این است که بهای تمام شده کالای فروش رفته در شرکت‌ها تقریباً برابر با ۷۴٪ فروش شرکت‌ها می‌باشد.

جدول ۱- آمار توصیفی متغیرها

متغیر	کمینه	میانگین	بیشینه	متغیر	کمینه	میانگین	بیشینه
ارزش بازار	۱۰/۰۲	۱۱/۶۵	۱۴	سود عملیاتی	-۲۵۴۰۶۷۳	۳۹۵۶۰۶/۴۱	۱۴۳۳۴۲۶۴
نسبت ارزش بازار به ارزش دفتری	۰/۲۱	۵/۸۱	۷۰/۸۷	سود خالص	-۴۶۹۲۲۲۲	۲۶۶۶۹۵/۷۳	۱۵۷۶۰۵۱۲
نسبت بدهی	۰/۱	۰/۶۳	۱/۵۷	نسبت خالص سرمایه در گردش به مجموع دارایی‌ها	-۰/۵۶۸	۰/۱۱۶	۰/۸۲
نسبت بدهی به حقوق صاحبان سهام	-۳۹۷/۶	۲/۰۵	۳۰۳/۸	فروش	۰	۲۱۸۵۴۸۹/۴	۱۰۷۴۲۰۹۶۱
بازده دارایی‌ها	-۰/۳۱۲	۰/۱۲۶	۰/۶۳۹۴	رشد فروش	-۰/۷۷	۰/۱۹۶	۴/۶۵
سود هر سهم	-۴۰۱۹/۹۳	۹۱۷/۹۱	۹۳۲۹/۲	کل دارایی‌ها	۲۲۷۲۵	۲۶۱۳۴۸۲/۹	۱۰۳۴۴۵۵۲۶۲
نسبت جاری	۰/۲۲۲	۱/۳۶۸	۱۰/۸۴۹	رشد کل دارایی‌ها	-۰/۴۹۰	۰/۱۷۶	۲/۱۶۸
نسبت آنی	۰/۰۷۱	۰/۸۳۷	۸/۳۵۸	حقوق صاحبان سهام	-۳۵۳۲۵۱۵	۷۵۹۰۲۲/۴۱	۳۱۱۳۲۵۱۸
نسبت گردش موجودی کالا	۰/۲۵۱	۳/۲۶۲	۲۲/۲۳۹	رشد حقوق صاحبان سهام	-۱۱/۸۳۱	۰/۴۸۴۶	۵۶/۸۳
نسبت سود عملیاتی به مجموع دارایی‌ها	-۰/۳۲۱۱	۰/۱۶۴	۰/۶۷۱	رشد نسبت سود عملیاتی به مجموع دارایی ها	-۲۵/۴۸	۰/۲۳۷	۲۹/۵۵۱
نسبت گردش کل دارایی‌ها	۰	۰/۹۰۵	۴/۸۹۵	رشد سود عملیاتی	-۳۶/۷۸	۰/۴۳۳	۳۷/۸۹۰
نسبت حاشیه سود ناخالص	-۰/۸۵۸	۰/۲۶۶	۰/۸۳۰	رشد سود خالص	-۱۹۱/۷۳۳	۰/۲۲۵	۱۲۵/۲۶۵
نسبت حاشیه سود عملیاتی	-۱/۱۳۷	۰/۱۹۵	۲///۰۶۹	مالیات ابرازی	۰	۲۸۷۶۷/۴	۱۰۵۸۹۶۳
نسبت حاشیه سود خالص	-۱/۱۳	۰/۱۵۳	۲/۰۳۴	بند مالیاتی	۰	۰/۳۹	۱
نسبت توانایی پرداخت بهره	-۱۰/۳۰۹	۲۱۰/۹۹	۱۴۳۱۵۰/۶	فرار مالیاتی	۰	۰/۵۴	۱
نسبت گردش دارایی ثابت	۰	۶/۱۳	۶۷/۳۴۹				

منبع: یافته‌های پژوهشگر

آمار استنباطی

نتایج حاصل از الگوریتم‌های مختلف به کار برده شده در دو حالت بدون انتخاب متغیرهای پیش‌بین و با انتخاب متغیرهای پیش‌بین در این پژوهش به منظور پیش‌بینی فرار مالیاتی در جدول ۲ نشان داده شده است. لازم به ذکر است که نتایج حاصل از آزمون ریلیف نشان داد که متغیرهای بهینه به ترتیب عبارتند از: نسبت سود عملیاتی به مجموع

دارایی‌ها، نسبت بازده دارایی‌ها، ارزش بازار شرکت، گردش موجودی کالا، نسبت بدهی، نسبت حاشیه سود ناخالص، نسبت گردش مجموع دارایی‌ها، نسبت خالص سرمایه در گردش به مجموع دارایی‌ها، نسبت حاشیه سود عملیاتی، مالیات ابرازی، نسبت گردش دارایی ثابت، سود هر سهم، نسبت حاشیه سودخالص، جمع دارایی‌ها، نسبت ارزش بازار به ارزش دفتری، سود عملیاتی، جمع حقوق صاحبان سهام، سودخالص، نسبت آنی، فروش خالص، نسبت جاری و رشد دارایی‌ها. بنابراین، الگوریتم‌های پژوهش، یک بار با تمام متغیرها و بار دیگر با متغیرهای انتخاب شده توسط ریلیف اجرا گردید و سپس نتایج این دو حالت، به منظور مشخص کردن تفاوت در عملکرد الگوریتم‌ها، با یکدیگر مقایسه شد. بدین منظور نتایج حاصل از ۱۰۰ بار تکرار هر الگوریتم در هر یک از حالات (در مجموع ۱۲۰۰ نتیجه) که با استفاده از نرم‌افزار Weka به دست آمده است با استفاده از نرم‌افزار SPSS با یکدیگر مقایسه شد و با توجه به آزمون تحلیل واریانس، آزمون توکی و آزمون ف t-test نمونه‌های مستقل به بررسی تفاوت بین عملکرد الگوریتم‌ها با یکدیگر پرداخته‌شد.

آماره درصد طبقه‌بندی صحیح میزان صحت الگوها را اندازه‌گیری می‌کند که با توجه به نتایج مندرج در جدول ۲، بالاترین میزان صحت الگوهای مورد استفاده در زمینه پیش‌بینی فرار مالیاتی در هر دو حالت را روش جنگل تصادفی و کمترین آن را درخت تصادفی دارا است. لازم به ذکر است که در حالت وجود مرحله انتخاب متغیرهای پیش‌بین، درصد طبقه‌بندی صحیح در تمام الگوریتم‌ها به جز درخت تصادفی و ریشه تصمیم[۶۳] افزایش یافته است. از دیگر آماره‌های مورد استفاده برای بررسی عملکرد یک الگوریتم آماره کاپا[۶۴] است که یک معیار تصحیح شده بر مبنای تصادف (شانس) برای تطابق بین طبقه‌بندی و کلاس‌های صحیح می‌باشد که مقدار بالاتر از صفر این آماره نشان‌دهنده این موضوع است که طبقه‌بندی‌کننده مورد نظر بهتر از حالت تصادفی کار می‌کند. هرچه این میزان بیشتر باشد، نشان‌دهنده عملکرد بهتر الگوریتم استفاده شده است. با توجه به آماره‌های مندرج در جدول ۲، الگوریتم جنگل تصادفی دارای بهترین عملکرد

در پیش‌بینی فرار مالیاتی و الگوریتم درخت تصادفی دارای ضعیف‌ترین عملکرد است. نتایج موجود در جدول ۲ نیز نشان می‌دهد که با افزودن مرحله انتخاب متغیرهای پیش‌بین آماره کاپا، به جز در مورد الگوریتم ریشه تصمیم، افزایش یافته که به طور کلی بیانگر بهتر شدن مدل پیش‌بینی می‌باشد.

جدول ۲- نتایج حاصل از الگوریتم‌ها

حالت		کاهش خطای هرس	درخت تصادفی	جنگل تصادفی	LMT	J48	ریشه تصمیم
۱	درصد طبقه بندی صحیح	۶۷/۵۳	۶۰/۵۹	۷۰/۲۱	۶۶/۹۷	۶۸/۳۶	۶۵/۵۸
	درصد طبقه بندی نادرست	۳۲/۴۶	۳۹/۴۰	۲۹/۷۸	۳۳/۰۲	۳۱/۶۳	۳۴/۴۱
	آماره کاپا	۰/۳۴۵	۰/۲۰۶۵	۰/۴۰۰۶	۰/۳۳۶	۰/۳۶۳۸	۰/۳۰۸۶
۲	درصد طبقه بندی صحیح	۶۸/۴۵	۶۲/۸۱	۷۱/۲۳	۶۹/۰۱	۶۸/۵۴	۶۵/۵۸
	درصد طبقه بندی نادرست	۳۱/۵۴	۳۷/۱۸	۲۸/۷۶	۳۰/۹۸	۳۱/۴۵	۳۴/۴۱
	آماره کاپا	۰/۳۶۱	۰/۲۵۳۱	۰/۴۲	۰/۳۷۱۲	۰/۳۷۱۵	۰/۳۰۸۶
منبع: یافته‌های پژوهشگر							

جدول ۳ نشان‌دهنده ماتریس درهم ریختگی حاصل از بررسی داده‌های پژوهش است. این ماتریس، چگونگی عملکرد الگوریتم دسته‌بندی را با توجه به مجموعه داده ورودی به تفکیک انواع دسته‌های مساله طبقه‌بندی نشان می‌دهد. اگر شرکت دارای فرار مالیاتی بوده باشد، با علامت مثبت و در غیر این‌صورت با علامت منفی نشان داده می‌شود. بنابر این عددی که در هر دو سمت آن دارای علامت مثبت است، بدین معناست که شرکت دارای فرار مالیاتی بوده و روش پیش‌بینی نیز به طور صحیح اقدام به پیش‌بینی آن نموده است که اصطلاحاً به این مورد «مثبت‌های واقعی[۶۵]» گفته می‌شود. همین مورد نیز در رابطه با علامت منفی صادق است. بدین معنا که شرکت دارای فرار مالیاتی نبوده و الگوریتم داده‌کاوی نیز به طور صحیح اقدام به پیش‌بینی آن نموده است. در مواردی که اعداد دارای علامتهای متفاوت در سطرها و ستون‌های ماتریس باشند، بدین معناست که الگوریتم نتوانسته به طور صحیح اقدام به پیش‌بینی فرار مالیاتی کند.

جدول ۳- نتایج ماتریس درهم‌ریختگی الگوریتم‌ها

		کاهش خطای هرس		درخت تصادفی		جنگل تصادفی		LMT		J48		ریشه تصمیم	
		+	−	+	−	+	−	+	−	+	−	+	−
بدون مرحله انتخاب متغیر پیش‌بین	+	۴۱۵	۱۷۲	۳۷۲	۲۱۵	۴۲۲	۱۶۵	۴۰۲	۱۸۵	۴۱۰	۱۷۷	۳۹۲	۱۹۵
	−	۱۷۹	۳۱۵	۲۱۱	۲۸۳	۱۵۷	۳۳۷	۱۷۲	۳۲۲	۱۶۵	۳۲۹	۱۷۷	۳۱۷
با مرحله انتخاب متغیر پیش‌بین	+	۴۳۲	۱۵۵	۳۷۶	۲۱۱	۴۳۳	۱۵۴	۴۴۱	۱۴۶	۳۹۲	۱۹۶	۳۹۲	۱۹۵
	−	۱۸۶	۳۰۸	۱۹۱	۳۰۳	۱۵۷	۳۳۷	۱۸۹	۳۰۵	۱۴۴	۳۵۰	۱۷۷	۳۱۷

منبع: یافته‌های پژوهشگر

با توجه به اینکه معیارهای مختلفی برای ارزیابی عملکرد الگوریتم‌های درخت تصمیم وجود دارد، جدول ۴ معروف‌ترین معیارهای ارزیابی و نتایج آن را برای هر کدام از الگوریتم‌های درخت تصمیم، در دو حالت با و بدون مرحله انتخاب متغیرهای پیش‌بین استفاده شده در این پژوهش، نشان می‌دهد. این معیارها شامل مثبت‌های واقعی، مثبت‌های کاذب۶۶، دقت۶۷، فراخوانی۶۸، F-Measure، ناحیه منحنی مشخصه عملکرد سیستم۶۹ و ناحیه منحنی دقت فراخوانی ۷۰ می‌باشد. از میان معیارهای پیش‌گفته، معیار دقت یا نرخ صحیح طبقه‌بندی نشان‌دهنده این موضوع است که چند درصد از کل مجموعه رکوردهای آزمایشی به درستی طبقه‌بندی شده است. معیار فراخوانی، کارایی الگوریتم را با توجه به تعداد رخداد آن طبقه نشان می‌دهد. حال آنکه معیار دقت اساساً مبتنی بر دقت پیش‌بینی الگوریتم می‌باشد و مبین این موضوع است که به چه میزان می‌توان به خروجی‌های الگوریتم اعتماد کنیم. معیار F-measure ترکیب معیارهای فراخوانی و دقت را نشان می‌دهد و در مواردی به کار می‌رود که نتوان اهمیت ویژه‌ای را برای هر یک از دو معیار فراخوانی و دقت نسبت به یکدیگر قائل شد. این معیار از طریق دو برابر کردن حاصل ضرب معیارهای دقت و فراخوانی تقسیم بر مجموع آن‌ها محاسبه می‌شود.

معیار ناحیه زیر منحنی (AUC) نشان‌دهنده سطح زیر منحنی مشخصه عملکرد سیستم است که روشی برای بررسی کارایی الگوریتم‌ها می‌باشد. هر چه این مقدار بزرگتر باشد، کارایی نهایی الگوریتم مطلوب‌تر ارزیابی می‌شود. منحنی‌های مشخصه عملکرد سیستم، رفتار یک الگوریتم را بدون توجه به توزیع دسته یا هزینه خطا نشان می‌دهند، بنابراین کارایی الگوریتم را از این عوامل جدا می‌کنند. متاسفانه در حالی که منحنی مشخصه عملکرد سیستم، یک تکنیک با ارزش مصورسازی است، اما در انتخاب الگوریتم مناسب کمک کمی می‌کند. معیار ناحیه زیر منحنی برای یک الگوریتم، که به صورت تصادفی طبقه نمونه مورد بررسی را تعیین می‌کند، برابر ۰/۵ است. همچنین بیشترین میزان این معیار برابر با یک بوده و برای وضعیتی رخ می‌دهد که در آن الگوریتم کلیه نمونه‌های را بدون هرگونه اشتباهی تعیین نماید. با توجه به جدول ۴، در تمام معیارهای ارزیابی عملکرد، الگوریتم جنگل تصادفی دارای بهترین عملکرد و درخت تصادفی دارای ضعیف‌ترین عملکرد در هر دو حالت با و بدون مرحله انتخاب متغیرهای پیش‌بین بوده است. در ادامه، به منظور بررسی تفاوت بین عملکرد الگوریتم‌های مختلف درخت تصمیم از آزمون t با نمونه‌های مستقل استفاده شد. در این راستا، از نتایج معیار F-measure موزون حاصل از ۱۰ بار تکرارِ روایی متقابل ۱۰ بخشی (روایی متقابل ۱۰ بخشی با ۱۰ بار تکرار) استفاده به عمل آمد که منجر به ایجاد ۱۰۰ داده در مورد هر الگوریتم در هر یک از حالت‌ها شد (در مجموع ۱۲۰۰ نتیجه). سپس به بررسی نرمال بودن نتایج حاصل از هر دوحالت با و بدون مرحله انتخاب متغیرهای پیش‌بین پرداخته شد.

جدول ۴- نتایج معیارهای مختلف الگوریتم‌ها

	ریشه تصمیم		J48		LMT		جنگل تصادفی		درخت تصادفی		کاهش خطای هرس	
حالت	۱	۲	۱	۲	۱	۲	۱	۲	۱	۲	۱	۲
مثبت‌های واقعی	۰/۶۵۶	۰/۶۵۶	۰/۶۸۴	۰/۶۸۵	۰/۶۷۰	۰/۶۹۰	۰/۷۰۲	۰/۷۱۲	۰/۶۰۶	۰/۶۲۸	۰/۶۷۵	۰/۶۸۵

مثبت‌های کاذب	۰/۳۴۶	۰/۳۴۶	۰/۳۱۹	۰/۳۱۱	۰/۳۳۳	۰/۳۲۱	۰/۳۰۱	۰/۲۹۲	۰/۳۹۹	۰/۳۷۴	۰/۳۳۱	۰/۳۲۵
دقت	۰/۶۵۷	۰/۶۵۷	۰/۶۸۴	۰/۶۹۰	۰/۶۷۱	۰/۶۸۹	۰/۷۰۳	۰/۷۱۲	۰/۶۰۶	۰/۶۲۹	۰/۶۷۵	۰/۶۸۴
فراخوانی	۰/۶۵۶	۰/۶۵۶	۰/۶۸۴	۰/۶۸۵	۰/۶۷۰	۰/۶۹۰	۰/۷۰۲	۰/۷۱۲	۰/۶۰۶	۰/۶۲۸	۰/۶۷۵	۰/۶۸۵
F-Measure	۰/۶۵۶	۰/۶۵۶	۰/۶۸۴	۰/۶۸۶	۰/۶۷۰	۰/۶۸۹	۰/۷۰۲	۰/۷۱۲	۰/۶۰۶	۰/۶۲۹	۰/۶۷۵	۰/۶۸۳
منحنی مشخصه عملکرد سیستم	۰/۶۳۶	۰/۶۳۶	۰/۶۹۴	۰/۶۹۴	۰/۷۱۴	۰/۷۰۰	۰/۷۶۹	۰/۷۷۱	۰/۶۲۴	۰/۶۳۱	۰/۷۱۱	۰/۷۱۱
ناحیه منحنی دقت فراخوانی	۰/۶۰۳	۰/۶۰۳	۰/۶۵۸	۰/۶۶۳	۰/۶۸۳	۰/۶۵۸	۰/۷۵۸	۰/۷۶۱	۰/۵۸۶	۰/۵۸۹	۰/۶۸۲	۰/۶۷۸
منبع: یافته‌های پژوهشگر												

جدول ۵ نتایج مرتبط با بررسی همگنی واریانس‌ها با استفاده از آزمون لون را نشان می‌دهد؛ مطابق با اطلاعات مندرج در این جدول، حاکی از بیشتر بودن سطح معناداری آماره لون از ۰/۰۵ است، می‌توان گفت که مشکلی در خصوص همگنی واریانس‌ها در هر دو حالت وجود ندارد.

جدول ۵-آزمون همگنی واریانس F- measure

حالت	آماره لون	درجه آزادی ۱	درجه آزادی ۲	معناداری
بدون مرحله انتخاب متغیرهای پیش‌بین	۱/۰۹۸	۵	۵۹۴	۰/۳۶۰
با مرحله انتخاب متغیرهای پیش‌بین	۰/۵۳۹	۵	۵۹۴	۰/۷۴۷
منبع: یافته‌های پژوهشگر				

برای بررسی تفاوت بین عملکرد الگوریتم‌های مختلف درخت تصمیم، نتایج آزمون تحلیل واریانس در جدول ۶ نشان داده شده است. مطابق با آماره F مندرج در این جدول، تفاوت میانگین گروه‌ها تایید و در مقابل فرض یکسان بودن آنها رد می‌شود. به عبارت دقیق‌تر، در هر دو حالت، دست کم یکی از گروه‌ها از نظر میانگین نمره مورد نظر با سایر گروه‌ها متفاوت است.

جدول ۶- نتایج تحلیل واریانس F- measure

		مجموع مربعات	درجه آزادی	میانگین مجموع مربعات	آماره F	سطح معناداری
بدون مرحله انتخاب متغیرهای پیش‌بین	بین گروه‌ها	۰/۴۸۶	۵	۰/۰۹۷	۵۳/۳۰۰	۰/۰۰۰
	درون گروه‌ها	۱/۰۸۳	۵۹۴	۰/۰۰۲		
	کل	۱/۵۶۹	۵۹۹	–		
دارای مرحله انتخاب متغیرهای پیش‌بین	بین گروه‌ها	۰/۵۲۵	۵	۰/۱۰۵	۶۲/۸۶۸	۰/۰۰۰
	درون گروه‌ها	۰/۹۹۳	۵۹۴	۰/۰۰۲		
	کل	۱/۵۱۸	۵۹۹	–		

جدول ۷- نتایج آزمون توکی

الگوریتم	فراوانی	زیر مجموعه‌های همگون (در سطح خطای ۵٪)				زیر مجموعه‌های همگون (در سطح خطای ۵٪)			
		۱	۲	۳	۴	۱	۲	۳	۴
درخت تصادفی	۱۰۰	۰/۶۱				۰/۶۲			
ریشه تصمیم	۱۰۰		۰/۶۵				۰/۶۵		
J48	۱۰۰			۰/۶۷				۰/۶۷	
کاهش خطای هرس	۱۰۰			۰/۶۷				۰/۶۸	
LMT	۱۰۰			۰/۶۸				۰/۶۹	
جنگل تصادفی	۱۰۰				۰/۷۰				۰/۷۱
معناداری	–	۱/۰۰	۱/۰۰	۰/۹۴۴	۱/۰۰	۱/۰۰۰	۱/۰۰۰	۰/۵۳۴	۱/۰۰
منبع: یافته‌های پژوهشگر									

مطابق جدول ۷، نتایج آزمون توکی نیز نشان داد که می‌توان شاخص F-measure موزون را بر حسب میانگین‌ها در سطح هر یک از الگوریتم‌های درخت تصمیم در حالت بدون انتخاب مرحله انتخاب متغیر به چهار زیرگروه همگن تقسیم کرد. ب عبارت دیگر، آزمون توکی درخت تصادفی را در گروه اول، ریشه تصمیم را در گروه دوم، J48، کاهش خطای هرس و LMT را در گروه سوم و در نهایت جنگل تصادفی را در گروه چهارم طبقه‌بندی کرده است. افزون بر این، در حالت وجود مرحله انتخاب متغیرهای

پیش‌بین نیز تعداد گروه‌ها به چهار گروه تقسیم شده است. در این حالت، نیز آزمون توکی، درخت تصادفی را در گروه اول، ریشه تصمیم را در گروه دوم، J48 و کاهش خطای هرس و LMT را در گروه سوم و جنگل تصادفی را در گروه چهارم طبقه‌بندی کرده است. به منظور بررسی وجود تفاوت در عملکرد الگوریتم‌های مختلف درخت تصمیم در دو حالت با و بدون مرحله انتخاب متغیرهای پیش‌بین، در ابتدا آماره نمونه‌های مستقل و سپس آزمون t-test برابری میانگین آنها انجام شده است. جدول ۸ میانگین، انحراف معیار و میانگین خطای استاندارد را برای هر دو حالت در هر کدام از الگوریتم‌ها نشان می‌دهد. جدول ۹، آزمون t-test برابری میانگین را نشان می‌دهد. همان‌طور که نشان داده شده است. نتایج حاصل از آزمون نشان می‌دهد که در حالت اجرای الگوریتم‌ها در دو حالت با و بدون مرحله انتخاب متغیرهای پیش‌بین در تمام الگوریتم‌ها جز الگوریتم LMT تفاوت معناداری وجود ندارد. به عبارت دیگر، از میان شش فرضیه مطرح شده در این پژوهش، تنها فرضیه پنجم رد نمی‌شود. به بیانی دیگر، تنها هنگامی که از الگوریتم LMT استفاده می‌شود، مرحله انتخاب متغیر دارای تاثیر معنادار بر نتایج الگوریتم و بهتر شدن این نتایج می‌شود.

جدول ۸- آماره نمونه‌های مستقل

	ریشه تصمیم		J48		LMT		جنگل تصادفی		درخت تصادفی		کاهش خطای هرس	
حالت	۱	۲	۱	۲	۱	۲	۱	۲	۱	۲	۱	۲
میانگین	۰/۶۵۱۲	۰/۶۵۱۲	۰/۶۷۳۷	۰/۶۸۰۹	۰/۶۷۲۰	۰/۶۸۹۳	۰/۷۰۲۹	۰/۷۱۱۶	۰/۶۱۰۷	۰/۶۱۸۹	۰/۶۷۷۵	۰/۶۷۹۵
انحراف معیار	۰/۴۲۳۸	۰/۴۲۳۸	۰/۰۴۱۸۴	۰/۰۴۰۰۸	۰/۰۳۷۷۰	۰/۰۳۶۸۰	۰/۰۴۲۴۸	۰۴۱۱۹۰	۰/۰۵۱۱۶	۰/۰۴۲۸۲	۰/۰۳۹۳۵	۰/۰۴۱۷۳
میانگین خطای استاندارد	۰/۰۰۴۲۴	۰/۰۰۴۲۴	۰/۰۰۴۱۸	۰/۰۰۴۰۱	۰/۰۰۳۷۷	۰/۰۰۳۶۸	۰/۰۰۴۲۵	۰/۰۰۴۱۲	۰/۰۰۵۱۲	۰/۰۰۴۲۸	۰/۰۰۳۹۳	۰/۰۰۴۱۷
منبع: یافته‌های پژوهشگر												

جدول ۹: آزمون t-test برابری میانگین‌ها برای متغیر F-measure موزون

		آزمون لون برای برابری واریانس‌ها		آزمون t-test برای برابری میانگین‌ها						
		آماره f	معناداری	آماره t	درجه آزادی	معناداری (دوطرفه)	تفاوت میانگین‌ها	خطای استاندارد تفاوت	فاصله	
									پایین	بالا
ریشه تصمیم	با فرض برابری واریانس‌ها	۰/۰۰۰	۱/۰۰۰	۰/۰۰۰	۱۹۸	۱/۰۰۰	۰/۰۰۰۰۰	۰/۰۰۵۹۹	-۰/۰۱۱۸۲	۰/۰۱۱۸۲
	بدون فرض برابری واریانس‌ها			۰/۰۰۰	۱۹۸	۱/۰۰۰	۰/۰۰۰۰۰	۰/۰۰۵۹۹	-۰/۰۱۱۸۲	۰/۰۱۱۸۲
J48	با فرض برابری واریانس‌ها	۰/۱۶۱	۰/۶۸۹	-۱/۲۴۴	۱۹۸	۰/۲۱۵	-۰/۰۰۷۲۱	۰/۰۰۵۷۹	-۰/۰۱۸۶۳	۰/۰۰۴۲۲
	بدون فرض برابری واریانس‌ها			-۱/۲۴۴	۱۹۸	۰/۲۱۵	-۰/۰۰۷۲۱	۰/۰۰۵۷۹	-۰/۰۱۸۶۳	۰/۰۰۴۲۲
LMT	با فرض برابری واریانس‌ها	۰/۰۵۵	۰/۸۱۴	-۳/۲۸۹	۱۹۸	۰/۰۰۱	-۰/۰۱۷۳۳	۰/۰۰۵۲۷	-۰/۰۲۷۷۱	-۰/۰۰۶۹۴
	بدون فرض برابری واریانس‌ها			-۳/۲۸۹	۱۹۸	۰/۰۰۱	-۰/۰۱۷۳۳	۰/۰۰۵۲۷	-۰/۰۲۷۷۱	-۰/۰۰۶۹۴
جنگل تصادفی	با فرض برابری واریانس‌ها	۰/۱۷۰	۰/۶۸۱	-۱/۴۶۸	۱۹۸	۰/۱۴۴	-۰/۰۰۸۶۹	۰/۰۰۵۹۲	-۰/۰۲۰۳۶	۰/۰۰۲۹۸
	بدون فرض برابری واریانس‌ها			-۱/۴۶۸	۱۹۷	۰/۱۴۴	-۰/۰۰۸۶۹	۰/۰۰۵۹۲	-۰/۰۲۰۳۶	۰/۰۰۲۹۸
درخت تصادفی	با فرض برابری واریانس‌ها	۰/۹۰۷	۰/۳۴۲	-۱/۲۳۶	۱۹۸	۰/۲۱۸	-۰/۰۰۸۲۵	۰/۰۰۶۶۷	-۰/۰۲۱۴۱	۰/۰۰۴۹۱
	بدون فرض برابری واریانس‌ها			-۱/۲۳۶	۱۹۲	۰/۲۱۸	-۰/۰۰۸۲۵	۰/۰۰۶۶۷	-۰/۰۲۱۴۱	۰/۰۰۴۹۱
کاهش خطای هرس	با فرض برابری واریانس‌ها	۰/۳۵۷	۰/۵۵۱	-۰/۳۵۳	۱۹۸	۰/۷۲۴	-۰/۰۰۲۰۳	۰/۰۰۵۷۴	-۰/۰۱۳۳۴	۰/۰۰۹۲۹
	بدون فرض برابری واریانس‌ها			-۰/۳۵۳	۱۹۷	۰/۷۲۴	-۰/۰۰۲۰۳	۰/۰۰۵۷۴	-۰/۰۱۳۳۴	۰/۰۰۹۲۹

منبع: یافته‌های پژوهشگر

بحث و نتیجه‌گیری

در چند دهه اخیر فرار مالیاتی تبدیل به یکی از دغدغه‌های دولت‌ها شده است. با این حال، پژوهش‌های انجام شده در رابطه با فرار مالیاتی، معمولاً از دیدگاه اقتصادی به بررسی موضوع پرداخته‌اند و کمتر پژوهشی اقدام به بررسی فرار مالیاتی با استفاده از اطلاعات

حسابداری نموده است. از سوی دیگر، عمدتاً پژوهش‌های انجام شده داخلی نیز تنها یک صنعت خاص و یک محدوده زمانی کوتاه در بورس اوراق بهادار تهران را مدنظر قرار داده‌اند. افزون بر این، استفاده از رویکرد داده‌کاوی و روش ریلیف نیز به عنوان یکی از روش‌های جدید کمتر مورد استفاده پژوهشگران در این زمینه قرار گرفته است. از این رو، این پژوهش با معرفی کاربرد رویکردهای داده‌کاوی و روش ریلیف در بحث فرار مالیاتی، به شناسایی روش‌ها و الگوریتم‌های کاراتر برای بررسی فرار مالیاتی در شرکت‌ها پرداخته است. بنابراین، این پژوهش به دنبال پاسخگویی به این پرسش است که آیا می‌توان با استفاده از داده‌های حسابداری و روش‌های داده‌کاوی و ریلیف به پیش‌بینی فرار مالیاتی در بورس اوراق بهادار تهران پرداخت و آیا میان الگوریتم‌های مختلف این روش‌ها، تفاوتی در کارایی آنها وجود دارد یا خیر؟ آیا می‌توان با استفاده از روش‌های انتخاب متغیر با استفاده از تعداد کمتری متغیر به همان نتایج یا نتایج بهتری دست یافت؟

نتایج پژوهش مندرج در جداول ۲ و ۳ بیانگر وجود شواهدی دال بر عملکرد مناسب الگوهای پیشنهادی برای پیش‌بینی فرار مالیاتی است؛ به بیان دقیق‌تر، تحلیل داده‌های پژوهش نشان داد که بهترین روش‌های الگوریتم درخت تصمیم برای پیش‌بینی فرار مالیاتی، که از عملکرد بالاتری برخوردار هستند، به ترتیب عبارتند از: روش جنگل تصادفی (برابر با ۰/۷)؛ روش J48 (برابر با ۰/۶۸)؛ روش کاهش خطای هرس (برابر با ۰/۶۷)؛ روش LMT (برابر با ۰/۶۶)؛ روش ریشه تصمیم (برابر با ۰/۶۵)؛ و روش درخت تصادفی (برابر با ۰/۶۰۴). این موضوع نشان می‌دهد که در وضعیت کشور ایران روش جنگل تصادفی بهترین روش برای پیش‌بینی فرار مالیاتی و روش درخت تصادفی، دارای بدترین عملکرد برای پیش‌بینی فرار مالیاتی است. افزون بر این نتایج حاصل از آزمون تحلیل واریانس ارایه شده در جدول ۷ نشان داد که بین میزان دقت پیش‌بینی الگوهای مختلف درخت تصمیم، تفاوت معناداری وجود دارد. به عبارت دیگر، انتخاب از میان الگوریتم‌های مختلف داده‌کاوی، منجر به پیش‌بینی‌های متفاوتی می‌شود. بنابراین، استفاده از نمونه‌های مختلف الگوریتم‌ها، با توجه به موضوع مورد پژوهش، از اهمیت بسیاری برخوردار است.

با افزودن مرحله انتخاب متغیرهای پیش‌بین به الگوریتم‌ها با استفاده از آزمون ریلیف، نتایج نشان داد که متغیرهای بهینه برای پیش‌بینی فرار مالیاتی به ترتیب عبارتند از: نسبت سود عملیاتی به مجموع دارایی‌ها، نسبت بازده دارایی‌ها، ارزش بازار شرکت، گردش موجودی کالا، نسبت بدهی، نسبت حاشیه سودناخالص، نسبت گردش مجموع دارایی‌ها، نسبت خالص سرمایه در گردش به مجموع دارایی‌ها، نسبت حاشیه سود عملیاتی، مالیات ابرازی، نسبت گردش دارایی ثابت، سود هر سهم، نسبت حاشیه سودخالص، جمع دارایی‌ها، نسبت ارزش بازار به ارزش دفتری، سود عملیاتی، جمع حقوق صاحبان سهام، سودخالص، نسبت آنی، فروش خالص، نسبت جاری و رشد دارایی‌ها. همچنین، نتایج جدول ۷ الی ۹ نشان داد که افزودن روش ریلیف تنها در الگوریتم LMT منجر به بهبود مدل شده است. نتایج این پژوهش نیز همانند پژوهش‌های باقرپور ولاشانی و همکاران (۱۳۹۱)، دستگیر و غریبی (۱۳۹۴)، رحیمی کیا و همکاران (۱۳۹۴)، سامعی راد و شاه‌بهرامی (۱۳۹۵)، تقوی‌فرد و همکاران (۱۳۹۶)، وو و همکاران (۲۰۱۲)، رحیمی کیا و همکاران (۲۰۱۷) و دیدیمو و همکاران (۲۰۱۸) نشان داد که استفاده از رویکردهای داده‌کاوی می‌تواند منجر به بهبود روش‌های کشف فرار مالیاتی گردد. افزون بر این، بدلیل یافت نشدن مطالعه‌ای که به بررسی انتخاب متغیرهای پیش‌بین بهینه در زمینه فرار مالیاتی بپردازد، امکان مقایسه نتایج پژوهش حاضر با سایر پژوهش‌ها در این زمینه وجود نداشت، اما با نتایج پژوهش‌هایی مانند تسای (۲۰۰۹)، ستایش و همکاران (۱۳۹۵) و نمازی و همکاران (۱۳۹۵) که نشان از بهبود عملکرد مدل در صورت انتخاب و استفاده از متغیرهای بهینه دارند، کاملاً مطابقت ندارد.

نتایج پژوهش نشان داد که حسابداران و حسابرسان می‌توانند به منظور کشف فرار مالیاتی از الگوریتم‌های مورد استفاده در این پژوهش استفاده نمایند. از طرف دیگر، نتایج این پژوهش نشان داد که با استفاده از اطلاعات حسابداری در سطح خرد (شرکت‌ها و داده‌های حسابداری) می‌توان اقدام به پیش‌بینی فرار مالیاتی نمود. همچنین نتایج این پژوهش نشان داد که برخی از متغیرهای حسابداری (مانند نسبت سود عملیاتی به مجموع دارایی‌ها، نسبت بازده دارایی‌ها و ارزش بازار شرکت) جهت پیش‌بینی فرار مالیاتی بهتر از سایر متغیرها عمل می‌کنند. این در حالی است که رویکردهای اقتصادی نگاهی کلی به فرار مالیاتی داشته و در سطح خرد اقدام به بررسی فرار مالیاتی نمی‌نمایند، اما با استفاده از اطلاعات حسابداری و فن‌های داده‌کاوی می‌توان به پیش‌بینی فرار مالیاتی در سطح خرد پرداخت.

براساس یافته‌های پـژوهش منـدرج در جـداول ۴ و ۷، بـه مسـئولان اقتصـادی و مالیـاتی کشور پیشنهاد می‌شود که از الگوهای مورد اشاره در پژوهش حاضر جهت پـیش‌بینـی فـرار مالیاتی در شرکت‌ها و جلوگیری از این اقدام ناپسند اجتماعی و اقتصادی بهره برنـد. افـزون بر این، به پژوهشگران آینده پیشنهاد می‌شود که با استفاده از سایر تکنیک‌های داده‌کـاوی (به‌عنوان نمونه، ماشین بردار پشتیبان و شـبکه‌هـای عصـبی مصـنوعی) بـه مطالعـه در زمینـه پیش‌بینی فرار مالیاتی بپردازند. به حسابداران و مجامع حرفه‌ای نیز پیشـنهاد مـی‌شـود کـه بـا توجه به اینکه مدل‌های غیرخطی نتایج بهتـری در پـیش‌بینـی فـرار مالیـاتی دارنـد، اقـدام بـه تـرویج و آمـوزش اسـتفاده از ایـن روش‌هـا در میـان حسـابداران و حسابرسـان نمـوده و پژوهش‌های لازم در این زمینه را مورد حمایت مادی و معنوی خود قرار دهند.

در انجام این پژوهش محدودیت‌هایی نیز وجود داشته است به عنـوان نمونـه، علـی‌رغـم اهمیت نوع صنعت در پیش‌بینی فرار مالیاتی، به‌دلیل حجم محـدود جامعـه، پژوهشـگران در خصوص تطبیق شرکت‌های دارای فرار مالیاتی و بدون فرار مالیـاتی از نظـر نـوع صـنعت بـا محـدودیت مواجـه بودنـد. همچنـین، ایـن پـژوهش دارای محـدودیت مربـوط بـه فـن‌هـای داده‌کاوی است. با وجود این سعی شد روایی و پایـایی پـژوهش در حـد ممکـن خدشـه‌دار نگردد.

یادداشت‌ها:

1.Stankevicius
2. Leonas
3. Spanish Finance Ministry
4. Pappa
5. Wu
6. Min
7. Lee
8. Mukkamala
9. Alfaro
10. Lee and
11 .Tow
12. Neoclassical Models
13. Institutional Approach
14. Allingham
15. Sandmo
16. Becker
17. Alm
18. Martinez-Vazquez
19. Gerxhani
20. Torgler
21. Nerre
22. Linoff
23. Berry
24. Khan
25. Yao
26. Tsai
27. Chiou
28. Random Forest

29. Decision Stump
30. Hu
31. Reduced-error pruning
32. Quinlan
33. lomaa
34. Kaariainen
35. Random Decision Tree
36. Birant
37. Logistic Model Tree
38. InteractiveDichotomizer
39. Hunt
40. Bhargava
41. Atiya
42. Filter
43. k-means
44. k-medoids
45. Goumagias
46. Deep Q-Learning
47. Didimo
48. Lisowsky
49. Isgiyarta
50. Gallemore
51. Labro
52. Lindenbaum
53. Hall
54. Curse of Dimensionality
55. holdout
56. Pessimistic
57. Kohavi
58. Cross Validation
59. Arlot
60. Celisse
61. Overfitting
62. Out-of-Sample
63.10-fold cross validation
64. Kappa
65.True Positive Rate
66. False Positive Rate
67.Precision
68. Recall
69.Receiver Operating Characteristics (ROC)
70. Precision-Recall curves

منابع

آل‌بوسویلم، مسلم (۱۳۹۰)، اندازه‌گیری و تحلیل اقتصاد زیرزمینی و فرار مالیاتی در ایران (پایان‌نامه کارشناسی ارشد، دانشکده علوم اداری و اقتصاد دانشگاه اصفهان).

امیدی‌پور، رضا؛ پژویان، محمدی، جمشید ، تیمور و معمارنژاد، عباس (۱۳۹۴)، برآورد حجم اقتصاد زیرزمینی و فرار مالیاتی: تحلیل تجربی در ایران. *پژوهشنامه مالیات*، ۲۸ (۷۶)، ۹۴-۶۹.

امینی‌خویی، زهره و عبداله‌پوری، علیرضا (۱۳۹۶)، طبقه بندی ترافیک شبکه با استفاده از الگوریتم جنگل تصادفی بهبودیافته، *علوم رایانشی*، ۵، ۲۴-۳۸.

باقرپورولاشانی، محمدعلی؛ باقری، مصطفی؛ خادم، حمید و حسینی پور، رضا (۱۳۹۱)، بررسی عوامل مالی و غیرمالی موثر بر گریز مالیاتی با استفاده از تکنیک های داده‌کاوی:

صنعت خودرو و ساخت قطعات. *مطالعات تجربی حسابداری مالی*، ۳۴، ۱۰۳-۱۲۸.

پورحیدری، امید و سروستانی، امیر (۱۳۹۱)، بررسی تأثیر ویژگی های شرکت، نوع صنعت و مالکیت نهادی بر اختلاف مالیات ابرازی و قطعی شرکتهای پذیرفته‌شده در بورس اوراق بهادار تهران. *پژوهشنامه مالیات* . ۲۰(۱۴)، ۷۷-۶۱.

تقوی‌فرد، سیدمحمدتقی، رئیسی وانانی، ایمان و پناهی، ریحانه (۱۳۹۶)، تحلیل آینده‌نگر تشخیص فرار مالیاتی مؤدیان مالیات بر ارزش افزوده با استفاده از الگوریتم‌های طبقه‌بندی و خوشه‌بندی. *پژوهشنامه مالیات*، ۳۵، ۱۱-۳۵.

جعفری، بهزاد و عادل، آذر (۱۳۹۲)، درخت تصمیم فازی؛ رویکردی نوین در تدوین استراتژی. *پژوهش‌های مدیریت عمومی*، ۱۹، ۲۵-۳۹.

حمیدی، ناصر؛ محمدزاده، امیر و محمدی، فاطمه (۱۳۹۴)، بررسی جایگاه جرایم مالیاتی در جلوگیری از فرار در نظام مالیات بر ارزش افزوده (مطالعه موردی: استان قزوین). *پژوهشنامه مالیات*، ۲۷: ۱۴۷-۱۶۶.

دستگیر، محسن و غریبی، مریم (۱۳۹۴)، کاربست روش‌های داده‌کاوی به منظور ارتقای عملکرد تشخیص فرار مالیاتی. *پژوهشنامه مالیات*، ۲۸، ۹۵-۱۱۶.

دهقان، سحر؛ موسوی جهرمی، یگانه و عبدلی، قهرمان (۱۳۹۷)، تئوری چشم‌انداز؛ رهیافتی نوین در توضیح پدیده فرار مالیاتی. *تحقیقات اقتصادی*، ۵۳(۱)، ۲۳-۱.

رضایی، فرزین و جعفری نیارکی، روح‌اله (۱۳۹۴)، رابطهٔ بین اجتناب مالیاتی و تقلب در حسابداری شرکتها. *پژوهشنامه مالیات*، ۲۳ (۲۶)، ۱۰۹-۱۳۴.

سامعی‌راد، مهدی و اسداله شاه‌بهرامی (۱۳۹۵)، بهبود کارایی الگوریتم‌های تشخیص تقلب مالیاتی با استفاده از الگوهای پردازش موازی، *پژوهشنامه مالیات*، ۲۹، ۱۱-۳۲.

سهرابی، بابک؛ رئیسی وانانی، ایمان و وحیده قانونی شیشوان، روح‌اله (۱۳۹۴)، ارزیابی عملکرد شرکت‌ها و تحلیل روندهای مالیاتی با استفاده از الگوریتم‌های داده‌کاوی. *تحقیقات مالی*، ۱۷(۲): ۲۳۸-۲۱۹.

شاکری، عباس (۱۳۸۴)، مروری تاریخی بر روند شکل‌گیری نظریه‌های اقتصاد کلان. *فصلنامه پژوهش‌های اقتصادی ایران*، ۲۳، ۶۹-۹۳.

صنیعی‌آباده، محمد و محمودی، سینا (۱۳۹۴)، داده‌کاوی کاربردی. چاپ دوم، تهران، ایران: انتشارات نیاز دانش.

عبداله‌میلانی، مهنوش و اکبرپورروشن، نرگس (۱۳۹۱)، فرار مالیاتی ناشی از اقتصاد غیررسمی در ایران. *پژوهشنامه مالیات*، ۱۳(۶۱)، ۱۴۱-۱۶۷.

فلاحتی، علی؛ نظیفی، مینو و عباسپور، سحر (۱۳۹۱)، مدل سازی اقتصاد سایه‌ای و تخمین فرار مالیاتی در ایران با استفاده از شبکه عصبی مصنوعی. *فصلنامه تحقیقات توسعه اقتصادی*، ۲ (۶): ۵۸-۳۳.

مهرانی، ساسان و سیدی، سیدجلال (۱۳۹۳)، بررسی تاثیر مالیات بر درآمد و حسابداری محافظه‌کارانه بر اجتناب مالیاتی شرکت‌ها، *دانش حسابداری و حسابرسی مدیریت*، ۱۰، ۱۳-۳۳.

نمازی، محمد؛ کاظم‌نژاد، مصطفی و نعمت‌الهی، محمدمهدی (۱۳۹۵)، مقایسه روش‌های مختلف انتخاب متغیرهای پیش‌بین برای پیش‌بینی بحران مالی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران. *فصلنامه مهندسی مالی و مدیریت اوراق بهادار*، ۷ (۲۹): ۱۹۳-۲۱۲.

هادیان، ابراهیم و تحویلی، علی (۱۳۹۲)، شناسایی عوامل موثر بر فرار مالیاتی در اقتصاد ایران. *فصلنامه برنامه ریزی و بودجه*. ۱۸(۲) : ۵۸-۳۹.

یوسفی، محمدقلی (۱۳۹۲)، بررسی نظرات مکاتب نئوکلاسیک، اتریشی و نهادگرایی درباره مکانیزم بازار. *فصلنامه پژوهش‌های اقتصادی ایران*، ۵۵، ۴۷-۹۱.

Alfaro, E.; García, N.; Gámez, M.; & D. Elizondo (2008). Bankruptcy Forecasting: An Empirical Comparison of Adaboost and Neural Networks. *Decision Support Systems*, 45, 110–122.

Allingham, M. G., & Sandmo, A. (1972). Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics*, 1(3-4), 323-338.

Arlot, S. & A. Celisse (2010). A Survey of Cross-validation Procedures for Model selection, Statistics Surveys, 4, 40-79.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and New Results, IEEE Transactions on Neural Networks, 12(4), 929-935.

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 Algorithm for Data Mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6).

Birant, D. (2011). Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models. *Journal of Environmental Informatics*, 17(1), 46-53.

Didimo, W., Giamminonni, L., Liotta, G., Montecchiani, F., & Pagliuca, D. (2018). A Visual Analytics System to Support Tax Evasion Discovery. Decision Support Systems, 110, 71-83.

Elomaa, T., & Kaariainen, M. (2001). An Analysis of Reduced Error Pruning. *Journal of Artificial Intelligence Research*, 15, 163-187.

Gallemore, J., & Labro, E. (2015). The Importance of the Internal Information Environment for Tax Avoidance. *Journal of Accounting and Economics*, 60(1), 149-167.

Goumagias, N. D., Hristu-Varsakelis, D., & Assael, Y. M. (2018). Using Deep Q-learning to Understand the Tax Evasion Behavior of Risk-averse Firms. *Expert Systems with Applications*, 101, 258-270.

Goumagias, N. D., Hristu-Varsakelis, D., & Assael, Y. M. (2018). Using Deep Q-learning to Understand the Tax Evasion Behavior of Risk-averse Firms. *Expert Systems with Applications*, 101, 258-270.

Graham, J. R., Hanlon, M., Shevlin, T., & Shroff, N. (2013).

Incentives for Tax Planning and Avoidance: Evidence from the Field. *The Accounting Review*, 89(3), 991-1023.

Hall, M. A. (2000) Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of the Seventeenth International Conference on Machine Learning (June 29 - July 02). P. Langley, Ed. Morgan Kaufmann Publishers, San Francisco, CA, 359-366.

Hu, W., Hu, W., & Maybank, S. (2008). Adaboost-based Algorithm for Network Intrusion Detection. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38(2), 577-583.

Hu, Y. C. (2010). Analytic Network Process for Pattern Classification Problems Using Genetic Algorithms, Information Sciences, 180(13), 2528–2539.

Isgiyarta, J. (2014). Tax Avoidance through Thin Capitalization (Evidence from Indonesian firms). *International Journal of Research in Business and Technology*, 5(3), 692-699.

Khan, D. M., Mohamudally, N., & Babajee, D. K. R. (2013). A Unified Theoretical Framework for Data Mining. *Procedia Computer Science*, 17, 104-113.

Kohavi, R. (1995). A study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection, IJCAI'95

Proceedings of the 14th international joint conference on Artificial intelligence, 1137-1143.

Lee, M-C. & C. To (2010). Comparison of Support Vector Machine and Back propagation Neural Network in Evaluating the Enterprise Financial Distress. *International Journal of Artificial Intelligence & Applications*, 1(3), 31-43.

Lindenbaum, M.; Markovitch, S.; & D. Rusakov (2004). Selective Sampling for Nearest Neighbor Classifiers, *Machine Learning*, 2, 125-152.

Linoff, G. S., & Berry, M. J. (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. John Wiley & Sons.

Lisowsky, P. (2010). Seeking shelter: Empirically modeling tax Shelters Using Financial Statement Information. *The Accounting Review*, 85(5), 1693-1720.

Min, J. H. & Lee, Y (2005). Bankruptcy prediction using Support Vector Machine with Optimal Choice of Kernel Function Parameters. Expert Systems with Applications, 28, 603–614.

Mo, P. L. L. (2003). Tax Avoidance and Anti-avoidance Measures In Major Developing Economies. Greenwood

Publishing Group.

Mukkamala, S.; Sung, A. H.; Ribeiro, B.; & A. Vieira (2006). Computational Intelligent Techniques for Financial Distress Detection. *Journal of Computational Intelligence Research*, 2(1), 60-65.

Pal, S. K. (2004). Soft Datamining, Computational Theory of Perceptions, and Rough-fuzzy Approach. *Information Sciences*, 163(1-3), 5-12.

Pappa, E., Sajedi, R., & Vella, E. (2015). Fiscal Consolidation with Tax Evasion and Corruption. *Journal of International Economics*, 96, S56-S75.

Rahimikia, E., Mohammadi, S., Rahmani, T., & Ghazanfari, M. (2017). Detecting Corporate Tax Evasion Using a Hybrid Intelligent System: A case study of Iran. *International Journal of Accounting Information Systems*, 25, 1-17.

Sandmo, A. (2005). The Theory of Tax Evasion: A Retrospective View. *National Tax Journal*, 1, 643-663.

Stankevicius, E., & Leonas, L. (2015). Hybrid Approach Model for Prevention of Tax Evasion and Fraud. *Procedia-Social and Behavioral Sciences*, 213, 383-389.

Taylor, G., & Richardson, G. (2013). The Determinants of Thinly Capitalized Tax Avoidance Structures: Evidence

from Australian firms. *Journal of International Accounting, Auditing and Taxation*, 22(1), 12-25.

Tsai, C. F. & Y.J. Chiou (2009). Earnings Management Prediction: a Pilot Study of Combining Neural Networks and Decision Trees, *Expert Systems with Applications*, 36(3), 7183–7191.

Tsai, C. F. (2009). Feature Selection in Bankruptcy Prediction, *Knowledge-Based Systems*, 22(2), 120–127.

Wu, R. S., Ou, C. S., Lin, H. Y., Chang, S. I., & Yen, D. C. (2012). Using Datamining Technique to Enhance Tax Evasion Detection Performance. *Expert Systems with Applications*, 39(10), 8769-8777.

Yao, Y. Y. (2001). On Modeling Data Mining with Granular Computing. In Computer Software and Applications Conference, 2001. COMPSAC 2001. 25th Annual International (638-643).

Zanganeh, M., Ashouri sheikh, E., & Abdollahi, A. (2018). Studying and Identifying the Effective Factors on Tax Evasion by Fuzzy DEMATEL-method, *Journal of Optimization in Industrial Engineering*, 11(2), 116-125.