تحلیل آینده‌نگر تشخیص فرار مالیاتی مؤدیان مالیات بر ارزش افزوده با استفاده از الگوریتم‌های طبقه‌بندی و خوشه‌بندی

سیدمحمدتقی تقوی‌فرد[۱]
ایمان رئیسی وانانی[۲]
ریحانه پناهی[۳]

تاریخ دریافت: ۱۳۹۵/۱۰/۲۷، تاریخ پذیرش: ۱۳۹۶/۷/۱۲

چکیده

فرار مالیاتی یکی از دغدغه‌های مستمر نظام‌های مالیاتی به‌خصوص در کشورهای در حال توسعه می‌باشد. هدف از دریافت مالیات بر ارزش افزوده، شفاف‌سازی تدریجی مبادلات اقتصادی و هم‌چنین ایجاد منبع درآمدی جدید، ثابت و قابل اتکاء برای تأمین هزینه‌های دولت است و ضرورت دارد این شفاف‌سازی از مرحله خرید مواد اولیه تا تولید و فروش کالا صورت گیرد تا بتوان مالیات را به‌درستی وصول کرد. هوش تجاری به‌طورکلی و داده‌کاوی به‌طور خاص، ابزارهای مؤثری برای افزایش کارایی و اثربخشی تشخیص فرار از پرداخت مالیات هستند. در این پژوهش بر اساس اطلاعات موجود در اظهارنامه‌های مالیاتی مؤدیان مالیات بر ارزش افزوده در سال‌های مورد مطالعه (۱۳۹۳–۱۳۸۸) که از سوی سازمان امور مالیاتی کشور حسابرسی شده‌اند و روش‌های داده‌کاوی شامل الگوریتم‌های طبقه‌بندی Naive Bayes، Decision Tree و K-Nearest Neighbor و الگوریتم‌های خوشه‌بندی K-means و K-medoids پیش‌بینی فرار مالیاتی مؤدیان انجام شد، سپس با استفاده از شاخص سیلوئت (Silhouette)، نتایج به‌دست آمده اعتبارسنجی شد. این نتایج می‌تواند به سازمان امور مالیاتی کشور جهت برنامه‌ریزی برای تشخیص فرار مالیاتی کمک کند.

واژه‌های کلیدی: فرار مالیاتی، مالیات بر ارزش افزوده، داده‌کاوی، طبقه‌بندی، خوشه‌بندی

۱. دانشیار مدیریت صنعتی، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبائی Dr. taghavifard@gmail.com
۲. استادیار مدیریت صنعتی، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبائی imanraeesi@atu.ac.ir
۳. کارشناس ارشد مدیریت فناوری اطلاعات، دانشگاه علامه طباطبائی (نویسنده مسئول) panahi.reihaneh@gmail.com

۱- مقدمه

کارکردهای اصلی ابزار مالیات را می‌توان به‌طور اصلی در تأمین درآمدهای عمومی کشور، گسترش عدالت (بازتوزیع ثروت)، تخصیص مجدد منابع، ثبات اقتصادی و تجهیز منابع مالی و تشکیل سرمایه برشمرد و این در حالی است که در حال حاضر، یکی از مشکلات مدیریت مالیاتی در جهان، بحث فرار و عدم تمکین مالیاتی است. تحقیقات جهانی نشان از فرار مالیاتی و سایر موارد عدم تمکین مالیاتی بین ۱۰ تا ۱۵ درصد دارد. با این وجود، هرچند رقم دقیقی در مورد میزان فرار مالیات در ایران و گروه‌هایی که تمکین نمی‌کنند، وجود ندارد ولی کارشناسان، فرار مالیاتی را در ایران بیش از ۵۰ درصد حجم واقعی مالیات وصولی می‌دانند (برزگری خانقاه و فیض‌پور، ۱۳۹۲).

بسیاری از مسائل مربوط به شناسایی تقلب شامل مقادیر بسیار زیاد اطلاعات است (لاندین، کوارنستومر و جانسون، ۲۰۰۳). پردازش این داده‌ها در روش‌های تراکنش‌های تقلبی به تحلیل آماری و الگوریتم‌های سریع و مؤثر نیاز دارد، که در میان آن‌ها داده‌کاوی روش‌های مناسبی را ارائه نموده و تفسیر داده را تسهیل می‌بخشد و به بهبود درک فرآیندهای مرتبط با داده کمک می‌کند (گلن، ۲۰۰۷).

در سال‌های پیشین به‌دلیل عدم وجود بسترهای نرم‌افزاری و سخت‌افزاری برای دریافت اظهارنامه و واریز الکترونیکی مالیات، امکان دسترسی به داده‌های طبقه‌بندی شده الکترونیکی شرکت‌ها برقرار نبود، به‌این دلیل در دوره فوق حرکت به سمت ایجاد یک سیستم هوشمند نرم‌افزاری برای کشف فرار مالیاتی و طراحی معیار آن امکان‌پذیر نبوده است. با پیاده‌سازی زیرساخت‌های نرم‌افزاری و سخت‌افزاری مختلف در سازمان امور مالیاتی کشور امکان طراحی و ایجاد ساختارهای هوشمند مختلفی در کنار سیستم‌های فوق برقرار شده است (رحیمی کیا و همکاران، ۱۳۹۴).

داده‌کاوی می‌تواند عوامل مؤثر بر فرار مالیاتی را شناسایی و مدل‌هایی را جهت کشف میزان احتمالی فرار مالیاتی مؤدیان مالیاتی ارائه دهد. بنابراین دانش ایجاد شده از فرآیند داده‌کاوی راهکارهایی را به سیاست‌گذاران حوزه مالیاتی در خصوص قانون‌گذاری و چهارچوبی را به ممیزان مالیاتی جهت رسیدگی کارا و اثربخش ارائه می‌دهد. در حوزه عملیاتی (رسیدگی و ممیزی)، داده‌کاوی پا را فراتر گذاشته و می‌تواند مبنایی را جهت پیاده‌سازی سیستم حسابرسی مبتنی بر ریسک در سازمان امور مالیاتی کشور فراهم آورد. بر این اساس، ممیزان مالیاتی می‌توانند قبل از نمونه‌گیری و انجام رسیدگی، کلیه مؤدیان مالیاتی را در بازه‌ای از ریسک پایین تا بالا (به لحاظ احتمال فرار مالیاتی) دسته‌بندی و برنامه‌های رسیدگی خود را بر اساس این دسته‌بندی تدوین و اجرا نمایند. به این ترتیب، ممیزان مالیاتی می‌توانند در خصوص این سؤال

که کدامیک از مؤدیان نیاز به بررسی بیشتری دارند، تصمیمات مناسبی را اتخاذ نمایند (باقرپور ولاشانی، باقری و همکاران، ۱۳۹۱).

بنابراین ضروری به نظر می‌رسد تا به این موضوع نگاهی دوباره داشته و سعی در شناسایی خلأهای موجود با بهره‌گیری از روش‌های نوین نمود.

۲- بیان مسأله

فرار مالیاتی پدیده‌ای است که با درجه‌های متفاوت در هر نظام مالیاتی و با هر نظام اقتصادی رواج دارد. شدت و ضعف شیوع این پدیده به عوامل متعددی مانند نوع مالیات، ماهیت، اثرات و پیامدهای اقتصادی و اجتماعی آن، فرهنگ، مذهب، نظام ارزشی و هنجارهای اجتماعی غالب بر جامعه، عوامل اقتصادی، هزینه‌های تمکین و توان سازمان مالیاتی در کشف این پدیده و برخورد با آن بستگی دارد. به‌طورکلی، پدیده فرار مالیاتی از این جهت قابل تأمل است که ضمن کاهش درآمدهای دولت و افزایش سطح شکاف مالیاتی، نسبت درآمدهای مالیاتی به تولید ناخالص داخلی(GDP)[1] را تنزل می‌دهد (آلم و وازکوئز، ۲۰۰۱). فرار از مالیات[2] و تقلب مالیاتی[3] یکی از دغدغه‌های مستمر برای اداره امور مالیاتی به‌خصوص در کشورهای در حال توسعه است. حقیقت این است که مالیات‌ها تنها منبع سرمایه دولت نیستند، با این حال نشانه بسیار مهمی درباره تعهد و کارایی دولت است که آیا دولت از پس فعالیت‌هایش برمی‌آید و دستیابی به سایر منابع درآمد را محدود می‌کند یا خیر (گنزالس و ولاسکوئز، ۲۰۱۳).

هدف از دریافت مالیات بر ارزش افزوده، شفاف‌سازی فرآیند و فعالیت‌های اقتصادی موجود در کشور است و ضرورت دارد این شفاف‌سازی از مراحل اولیه ورود مواد اولیه تا تولید نهایی کالا باشد تا بتوان مالیات را دریافت کرد. نبود زیرساخت‌های لازم برای دریافت اطلاعات فعالیت‌ها و عدم رعایت قوانین توسط مؤدیان، دریافت مالیات را با مشکل مواجه می‌کند.

از دیدگاه نظری، مهمترین مسئله‌ای که زمینه‌های فرار مالیاتی را در سیستم مالیات بر ارزش افزوده فراهم می‌کند آن است که اخذ این مالیات اساساً مبتنی بر مفهوم ارزش خالص است. لذا از آنجا که با اعمال روش تفریقی، امکان بازگشت وجوه مالیاتی پرداخت شده در مراحل مختلف فرآیند تولید و توزیع به وجود می‌آید، احتمال تقلب مالیاتی نیز مطرح می‌شود (موسوی جهرمی و همکاران، ۱۳۸۸).

مؤدیان در زمانی که کالا می‌خرند، مالیات بر ارزش افزوده آن را باید پرداخت کنند و زمانی که کالا را

1. Gross Domestic Product (GDP)
2. Tax Evasion
3. Tax Fraud

می‌فروشند از خریداران، مالیات بر ارزش افزوده دریافت می‌کنند و مابه‌التفاوت مالیات فروش و خرید را به دولت پرداخت می‌کنند. اما مؤدیان با فاکتورهای غیر واقعی خرید (که یکی از مصادیق تقلب و فرار مالیاتی است)، مالیات پرداخت نشده را به عنوان مالیات پرداخت شده محسوب کرده تا به عنوان اعتبار از مالیات فروش کسر و مالیات کمتری پرداخت کنند.

هوش تجاری[1] و به‌طورکلی داده‌کاوی[2] به‌طور خاص، ابزارهای مؤثری برای افزایش کارایی و اثربخشی تشخیص فرار از پرداخت مالیات هستند (فدایرو و همکاران، ۲۰۰۸). هدف کلی این پژوهش، ارائه چهارچوبی جهت پیش‌بینی فرار مالیاتی مؤدیان مالیات بر ارزش افزوده بر اساس روش‌های داده‌کاوی شامل الگوریتم‌های طبقه‌بندی[3] و خوشه‌بندی[4] است.

۳- پیشینه پژوهش

از آنجا که فناوری داده‌کاوی از قابلیت‌های پیش‌بینی[5] و طبقه‌بندی فراوانی برخوردار است می‌تواند فرآیند تصمیم‌گیری در مسائل مالی را تسهیل نماید. کاربرد روش‌های داده‌کاوی با توجه به مطالعات مرتبط و ماهیت آن‌ها، می‌تواند طیف گسترده‌ای شامل پیش‌بینی ورشکستگی، تخمین ریسک اعتباری، وضعیت تداوم فعالیت، درماندگی مالی، پیش‌بینی عملکرد واحد تجاری و انواع تقلب را در برگیرد، بنابراین در این پژوهش جهت پیش‌بینی فرار مالیاتی به کار گرفته شده است. در ادامه برخی از مهم‌ترین پژوهش‌های داخلی و خارجی صورت گرفته در خصوص موضوع پژوهش در قالب جدول‌های ۱ و ۲ بیان می‌شود.

1. Business Intelligence
2. Data Mining
3. Classification
4. Clustering
5. Prediction

جدول (۱) – مروری بر مهم‌ترین پژوهش‌های داخلی در خصوص داده‌کاوی در فرار مالیاتی

عنوان پژوهش	محقق	سال	تکنیک	نتیجه
بررسی عوامل مالی و غیرمالی مؤثر بر گریز مالیاتی با استفاده از تکنیک‌های داده‌کاوی: صنعت خودرو و ساخت قطعات	باقرپور ولاشانی، باقری، خادم و حسینی‌پور	۱۳۹۱	درخت تصمیم[۲] C۵,۰ و شبکه‌های عصبی مصنوعی[۳]	احتمال فرار مالیاتی در مؤدیانی که عملکرد مناسبی ندارند، زیاد است
پیش‌بینی گزارش حسابرس مستقل در ایران: رویکرد داده‌کاوی	باقرپور ولاشانی، ساعدی، مشکانی و باقری	۱۳۹۱	درخت تصمیم، شبکه‌های عصبی مصنوعی و رگرسیون لجستیک[۴]	میانگین دقت مدل حاصل از تکنیک درخت تصمیم از دو روش دیگر بیشتر است
ارزیابی مالیات عملکرد شرکت‌ها و تحلیل روندهای مالیاتی با استفاده از الگوریتم‌های داده‌کاوی	سهرابی، رئیسی وانانی و قانونی شیشوان	۱۳۹۴	خوشه‌بندی و طبقه‌بندی	برتری روش خوشه‌بندی مبتنی بر چگالی (DBSCAN)[۵]

2. Decision Tree
3. Artificial Neural Networks
4. Logistic Regression
5. Density Based Spatial Clustering of Applications with Noise (DBSCAN)

منبع: یافته‌های تحقیق

جدول (۲)- مروری بر مهم‌ترین پژوهش‌های خارجی در خصوص داده‌کاوی در فرار مالیاتی

عنوان پژوهش	محقق	سال	تکنیک	نتیجه
روش‌های داده‌کاوی برای تشخیص صورت‌های مالی جعلی[1]	کرکوس، اسپاتیس و مانولوپوس	۲۰۰۷	درخت تصمیم‌گیری، شبکه عصبی مصنوعی و شبکه باور بیزینی[3]	برتری شبکه باور بیزینی
تشخیص تکاملی تقلب در صورت‌های مالی	ژو و کاپور	۲۰۱۱	مجموعه مدل‌های رگرسیونی، درخت تصمیم‌گیری، شبکه عصبی و شبکه بیزینی و مدل سطح پاسخ[4]	ترکیب مدل‌ها نتایج بهتری را در پی خواهد داشت
استفاده از روش‌های داده‌کاوی به منظور افزایش عملکرد تشخیص فرار از پرداخت مالیات	ووا، اوو، لین، چنگ و ین	۲۰۱۲	یک سیستم مبتنی بر داده‌کاوی	استفاده از سیستم‌های مبتنی بر داده‌کاوی موجب بهبود چشمگیری در تشخیص کسب و کارهای فراری در حوزه مالیات بر ارزش افزوده می‌شود
تعیین و شناسایی مالیات دهندگان با فاکتورهای جعلی با استفاده از روش‌های داده‌کاوی	گنزالس و ولاسکوئز	۲۰۱۳	خوشه‌بندی وطبقه‌بندی	بهره‌گیری از تکنیک‌های خوشه‌بندی و طبقه‌بندی در شناسایی متغیرهای کلیدی مؤثر در شرکت‌های کوچک و شرکت‌های بزرگ و متوسط و ارائه مدلی برای پیش‌بینی تقلب مالیاتی

1. Fraudulent Financial Statements
2. Bayesian Belief Network (BBN)
3. Response Surface Method

منبع: یافته‌های تحقیق

۴ - روش‌شناسی پژوهش

پژوهش حاضر از نظر هدف، کاربردی است، زیرا به پیش‌بینی و تبیین عوامل مؤثر بر شناسایی و تشخیص فرار مالیاتی در اظهارنامه‌های مالیاتی مالیات بر ارزش افزوده بر اساس روش‌های داده‌کاوی می‌پردازد. مبانی نظری پـژوهش از طریـق روش مطالعه کتابخانه‌ای یعنی مطالعه مقاله‌ها، کتاب‌ها، پژوهش‌های انجام شده در گذشته و استفاده از نظرات خبرگان حوزه مالیاتی تـدوین شده است.

هم‌چنین، فرآیند استاندارد داده‌کاوی[1] CRISP-DM در این پژوهش مورد استفاده قرار گرفته است. این مدل از شش مرحله زیرکه به‌صورت یک فرآیند حلقه‌ای است تشکیل می‌شود (شهرابی، ۱۳۹۲: ۱۱۹).

۱- تعریف مسأله[2]

۲- تحلیل داده‌ها[3]

۳- آماده‌سازی داده‌ها[4]

۴- مدل‌سازی[5]

۵- ارزیابی[6]

۶- توسعه[7]

۵- روش تجزیه و تحلیل داده‌ها

جامعه آماری پژوهش حاضر شامل اشخاص حقیقی و حقوقی مشمول نظام مالیات بر ارزش افزوده در شهر و استان تهران می‌باشند که تعداد ۳۱۰۱ مؤدی مالیاتی (دوره مالیاتی) به عنوان نمونه از میان ۲۰،۰۰۰ مؤدی حسابرسی شده انتخاب شده‌اند. از آنجا که اظهارنامه مالیات بر ارزش افزوده به صورت دوره‌ای (فصلی) ارائه و هر مؤدی در هر سال مالی مؤظف به ارائه چهار دوره اظهارنامه است، بنابراین باتوجه به حجم عظیم داده‌ها و اینکه در این پژوهش بخشی از تمام داده‌های سال‌های مورد مطالعه مورد استفاده قرار می‌گیرد، ممکن است تعداد مؤدیان کمتر از تعداد دوره‌های ذکر شده باشد، چرا که این امکان وجود دارد که چندین دورهٔ مالیاتی متعلق به یک مؤدی مالیاتی باشد، بنابراین نمی‌توان با قطعیت عنوان کرد جامعه آماری شامل ۳۱۰۱ مؤدی است. داده‌های مورد نیاز از طریق اطلاعات فرم اظهارنامه‌های مالیاتی مؤدیان مالیات

1. Cross-Industry Standard Process for Data Mining (CRISP-DM)
2. Business Understanding
3. Data Understanding
4. Data Preparation
5. Modeling
6. Evaluation
7. Development

بر ارزش افزوده در سال‌های مورد مطالعه (۱۳۹۳-۱۳۸۸) که از سوی سازمان امور مالیاتی کشور مورد حسابرسی قرار گرفته اند؛ بهره‌گیری از نظر خبرگان مالیاتی بر روی متغیرهای موجود در فرم اظهارنامه؛ گزارش‌های حسابرسی آن اظهارنامه‌ها و دانشی که در ذهن خبرگان برای شناسایی تقلب به‌کار گرفته می‌شود، تأمین شده و ۵ نفر از میان کارمندان سازمان امور مالیاتی کشور (شهر تهران) که ویژگی‌های زیر را دارا می‌باشند، برای این کار انتخاب می‌شوند.

- مدرک تحصیلی: لیسانس به بالا
- تجربه کاری: ۱۵ سال به بالا
- زمینه کاری: بخش ارزش افزوده
- رشته شغلی: کارشناس ارشد و بالاتر

جدول (۳)- متغیرهای منتخب اظهارنامه مالیاتی و نسبت‌های مالی

متغیر	نام متغیر	شرح متغیر
X1	فروش مشمول ابرازی	فروش (کالا و خدمت) ی که طبق قانون مشمول مالیات است
X2	نرخ مالیات	نرخ مالیاتی که به‌موجب قانون تعیین می‌گردد و برای هر سال میزان مشخصی می‌باشد
X3	کل مالیات ابرازی فروش اظهارنامه	مالیات و عوارضی که به فروش کالا و خدمات بر اساس نرخ قانونی تعلق می‌گیرد و مؤدی به‌موجب اظهارنامه ابراز می کند
X4	خرید مشمول ابرازی	خرید (کالا و خدمت) ی که طبق قانون مشمول مالیات است
X5	کل مالیات ابرازی خرید اظهارنامه (اعتبار مالیاتی)	مالیاتی که به خرید کالا و خدمات بر اساس نرخ قانونی تعلق می‌گیرد و توسط مؤدی پرداخت و به‌موجب اظهارنامه ابراز می‌شود
X6	بدهی/ فروش	(کل مالیات ابرازی فروش اظهارنامه – کل مالیات ابرازی خرید اظهارنامه (اعتبار مالیاتی)) / فروش مشمول ابرازی
X7	مالیات فروش/ فروش	کل مالیات ابرازی فروش اظهارنامه / فروش مشمول ابرازی
X8	خرید/ فروش	(خرید ابرازی مشمول+ خرید ابرازی معاف) / (فروش ابرازی مشمول+ فروش ابرازی معاف)

متغیر	نام متغیر	شرح متغیر
X9	کل مالیات فروش برگ مطالبه	مالیات و عوارضی که به فروش کالا و خدمات بر اساس نرخ قانونی تعلق می‌گیرد و فروش مؤدی توسط مأمور مالیاتی به‌موجب بررسی اسناد و مدارک تعیین شده است
X10	کل مالیات خرید برگ مطالبه	مالیات و عوارضی که به خرید کالا و خدمات بر اساس نرخ قانونی تعلق می‌گیرد و خرید مؤدی توسط مأمور مالیاتی به‌موجب بررسی اسناد و مدارک مورد تائید قرار گرفته است
X11	مالیات پرداخت شده/ بدهی	(پرداختی مالیات قبل از قطعی+ پرداختی عوارض قبل از قطعی) / (کل مالیات ابرازی فروش اظهارنامه – کل مالیات ابرازی خرید اظهارنامه(اعتبار مالیاتی))

منبع: یافته‌های تحقیق

همان‌گونه که در جدول ۳ نشان داده شده، متغیرهای مؤثر در پیش‌بینی تقلب مالیاتی با استفاده از اطلاعات موجود در اظهارنامه‌ها، گزارش‌های حسابرسی و نسبت‌های مالی که با استفاده از اظهارنامه‌ها و گزارش‌های حسابرسی تعریف و بر اساس نظر خبرگان مالیاتی تائید شده است، انتخاب شده‌اند.

در ابتدا نیاز است داده‌ها پالایش شوند که این کار با اجرای مراحل پیش‌پردازش داده‌ها و طی مراحل زیر انجام شد:

۵-۱- آماده‌سازی داده‌ها

در ابتدا سعی در شناسایی داده‌های خارج از محدوده[1] با تکنیک شش سیگما[2] شد و چون داده‌های خارج از محدوده به مؤدیان بزرگ (مؤدیانی که حجم فعالیت اقتصادی آنان بیشتر از صد میلیارد تومان است) تعلق دارد، بنابراین با جداکردن آن‌ها از مؤدیان کوچک و متوسط، تعداد کل مؤدیان به ۲۲۶۰ مؤدی کاهش یافت. در این مرحله سطرهایی از داده‌ها که بعضی از مشخصه‌های آن‌ها خالی است و اطلاعات مربوط به آن مشخصه‌ها به‌طور کامل توسط مؤدیان پر نشده‌اند، حذف شده و نهایتاً تعداد کل سطرها در این مرحله به ۱۴۰۹ سطر رسید. هم چنین لازم به ذکر است، چون هر مؤدی یک شماره اقتصادی مخصوص به‌خود دارد و با این شماره وارد سیستم ثبت‌نام و تسلیم اظهارنامه می‌شود و اظهارنامه خود را پر می‌کند، بنابراین امکان وجود داده‌های تکراری از بین می‌رود.

1. Outlier Data
2. Six Sigma

۵-۲- نرمال‌سازی داده‌ها

نرمال‌سازی، تغییر مقیاس داده‌ها به گونه‌ای است که آن‌ها را به یک فاصلهٔ کوچک و معین نگاشت می‌کند و باعث می‌شود که داده‌ها با مقیاس بزرگ، نتایج را به سمت خود منحرف نکنند. در این پژوهش با استفاده از روش مینیمم-ماکزیمم[1]، نرمال‌سازی انجام شد و داده‌ها در بازه عددی بین صفر و یک قرار گرفتند. چون متغیر نرخ مالیات برای سال‌های مختلف مورد مطالعه، اعداد ثابت مخصوص به آن سال‌ها و در بازه بین صفر و یک قرار دارد، بنابراین نیازی به نرمال‌سازی ندارد.

جدول (۴)- سه نمونه از مقادیر اصلی داده‌ها

مالیات پرداخت شده/بدهی	کل مالیات خرید برگ مطالبه	کل مالیات فروش برگ مطالبه	خرید/فروش	مالیات فروش/فروش	مالیات قابل پرداخت	کل مالیات ابرازی خرید اظهارنامه	خرید مشمول ابرازی	کل مالیات ابرازی فروش اظهارنامه	نرخ مالیات	فروش مشمول ابرازی
۱/۳۹۷	۱۰/۵۲۳/۸۶۸	۱/۴۹۶/۶۲۸/۶۸۹	۰/۰۸۹	۰/۰۷۷	۰/۰۶۹	۹۸/۸۲۲/۵۳۴	۱/۲۳۵/۲۸۱/۶۷۵	۱/۰۶۹/۱۰۸/۵۸۷	۰/۰۸	۱۳/۸۹۶/۲۹۵/۷۶۲
۱/۰۰۰	۶۰۹/۲۶۴/۲۹۸	۱/۶۱۴/۰۷۴/۵۲۴	۰/۶۹۷	۰/۰۲۴	۰/۰۱۲	۶۳۰/۰۱۸/۰۲۶	۳۹/۱۲۰/۱۷۱/۹۸۳	۱/۳۴۶/۴۵۸/۴۷۸	۰/۰۳	۵۶/۱۱۹/۳۱۶/۳۹۵
۱/۰۰۰	۶۶۴/۴۳۲/۵۱۲	۲/۴۱۳/۰۲۶/۳۱۲	۰/۲۷۳	۰/۰۶۰	۰/۰۴۳	۶۶۴/۷۷۸/۱۱۲	۱۱/۰۷۹/۶۳۵/۲۰۰	۲/۴۳۱/۶۱۶/۱۷۰	۰/۰۸	۴۰/۵۲۶/۹۳۶/۱۶۷

منبع: یافته‌های تحقیق

1. Min-Max

جدول (۵)- سه نمونه از مقادیر نرمال‌شده داده‌ها

فروش مشمول ابرازی	نرخ مالیات	کل مالیات ابرازی فروش اظهارنامه	خرید مشمول ابرازی	کل مالیات ابرازی خرید اظهارنامه	مالیات قابل پرداخت	مالیات فروش/ فروش	خرید/ فروش	کل مالیات فروش برگ مطالبه	کل مالیات خرید برگ مطالبه	مالیات پرداخت شده /بدهی
۰/۰۲۱۱	۰/۰۸	۰/۰۲۹۰	۰/۰۰۲۰	۰/۰۰۴۴	۰/۸۶۵۱	۰/۰۷۶۸	۰/۰۰۰۹	۰/۰۱۶۰	۰/۰۰۰۵	۰/۹۹۱۹
۰/۰۸۵۱	۰/۰۳	۰/۰۳۶۵	۰/۰۶۳۷	۰/۰۲۸۱	۰/۸۵۵۴	۰/۰۲۳۹	۰/۰۰۶۹	۰/۰۱۷۲	۰/۰۲۶۹	۰/۹۹۱۸
۰/۰۶۱۵	۰/۰۸	۰/۰۶۵۹	۰/۰۱۸۰	۰/۰۲۹۶	۰/۸۶۰۶	۰/۰۵۹۹	۰/۰۰۲۷	۰/۰۲۵۷	۰/۰۲۹۳	۰/۹۹۱۸

منبع: یافته‌های تحقیق

جدول ۴ سه رکورد از مقادیر اصلی داده‌ها و جدول ۵ مقادیر نرمال شده همان داده‌ها را به‌طور نمونه نشان می‌دهد.

۵-۳- مدل‌سازی

داده‌کاوی به دو نوع هدایت شده[1] و غیر هدایت شده[2] تقسیم می‌شود. داده‌کاوی هدایت شده، دارای متغیر هدف خاص و از پیش تعیین شده است که به‌دنبال الگویی خاص می‌گردد. در حالی که هدف داده‌کاوی غیر هدایت شده، یافتن الگوها یا تشابهات بین گروه‌هایی از اطلاعات، بدون داشتن متغیر هدف خاص و یا مجموعه‌ای از دسته‌ها و الگوهای از پیش تعیین شده می‌باشد (شهرابی، ۱۳۹۲: ۱۲).

روش‌های داده‌کاوی مورد استفاده در این پژوهش شامل الگوریتم‌های طبقه‌بندی شامل سه روش K-Nearest Neighbor (KNN) و Naive Bayes، Decision Tree است که از جمله الگوریتم‌های داده‌کاوی هدایت شده‌اند و الگوریتم‌های خوشه‌بندی شامل دو روش k-means و k-medoids است که از جمله الگوریتم‌های داده‌کاوی هدایت نشده‌اند که در ادامه به‌صورت مختصرتوضیح داده شده‌اند.

طبقه‌بندی

طبقه‌بندی شامل بررسی ویژگی‌های یک شیء جدید و تخصیص آن به یکی از مجموعه‌های از قبل تعیین شده می‌باشد.

۱- درخت تصمیم: درخت تصمیم در حقیقت یک گراف با ساختار درخت است که هر رأس آن نشان‌دهنده

1. Directed
2. Undirected

یک آزمون یا مقایسه مقدار یک متغیر می‌باشد و یال‌هایی که از آن رأس خارج می‌شود، نشان‌دهنده تصمیمی است که در مقابل هر نتیجه به‌دست آمده از آزمون گرفته می‌شود. در این روش تلاش می‌شود تا مشاهدات بـه زیرگروه‌هایی تقسیم شوند. از جمله مزیت‌های این روش مستقل بودن آن از چگونگی توزیع داده‌ها و وابستگی متغیرهای ورودی می‌باشد. در حقیقت، درخت تصمیم یک مدل مفهومی ساده را با استفاده از تعدادی تصمیم ساده ایجاد می‌کند. الگوریتم یادگیری در این روش بسیار سریع است (بریمنر و همکاران، ۲۰۰۵).

۲- Naive Bayes: یکی از فرمول‌های مهم احتمال، فرمول احتمال بیز[1] است که می‌توان به کمک آن، برچسب کلاس یک نمونه از داده‌ها را تخمین زد. استفاده از این قانون برای طبقه‌بندی، دقت و سرعت خوبی را در پایگاه داده‌های بزرگ به‌همراه دارد. در این روش فرض بر این است که تأثیر مقدار یک صفت خاصه بر روی برچسب کلاس، مستقل از مقادیر دیگر صفات خاصه است و این موضوع استقلال شرطی[2] کلاس نامیده می‌شود (نوروزی، ۱۳۹۴).

۳- KNN: یکی از کاربردهای رایج الگوریتم KNN، تشخیص الگو است. برای یک داده آزمایشی، الگوریتم به دنبال k نمونه از نزدیک‌ترین نمونه‌ها می‌گردد. نزدیکی دو نمونه با به‌دست آوردن تشابه و یا فاصله میان این دو نمونه محاسبه می‌شود. هر نمونه می‌تواند از انواع داده‌ها تشکیل شده باشد که باید تشابه میان آن‌ها بررسی شود. پس از یافتن این k داده مشابه با نمونه آزمایشی، با رأی اکثریت برچسب کلاس داده آزمایشی انتخاب می‌شود (نوروزی، ۱۳۹۴).

خوشه‌بندی

خوشه‌بندی به عمل تقسیم جمعیت ناهمگن به تعدادی از زیر مجموعه‌ها یا خوشه‌های همگن گفته می‌شود و یک تکنیک داده‌کاوی هدایت نشده است که با شکستن پایگاه داده‌های پیچیده به خوشه‌های ساده‌تر با استفاده از تکنیک خوشه‌بندی می‌توان برای بهبود عملکرد تکنیک‌های هدایت شده، استفاده کرد. با انتخاب اندازه‌گیری‌های متفاوت فواصل، خوشه‌بندی را می‌توان برای هر نوع داده‌ای به کار برد. این الگوریتم برای نشان‌دادن نزدیکی یا دوری داده‌ها بر معیارهای شباهت تکیه می‌کند. در خوشه‌بندی هیچ دسته از پیش تعیین شده‌ای وجود ندارد و داده‌ها صرفاً بر اساس تشابه گروه‌بندی می‌شوند (شهرابی، ۱۳۹۲: ۱۸).

۱- k-means: پارامتر k (تعداد خوشه) را به عنوان ورودی گرفته و مجموعهٔ n شیء را به k خوشه افراز می‌کند. به‌طوری که سطح شباهت داخلی خوشه‌ها بالا بوده و سطح شباهت اشیاء بین خوشه‌ها پایین

1. Bayesian Theorem
2. Conditional Independence

باشد. شباهت هرخوشه نسبت به متوسط اشیاء آن خوشه سنجیده شده که این متوسط مرکز خوشه نیز نامیده می‌شود.

۲- k-medoids: در الگوریتمk-medoids قبل از اینکه فاصله داده‌های دیگر از هر مرکز خوشه محاسبه شود، k نقطه به صورت تصادفی از n داده به‌عنوان مرکز خوشه مشخص می‌شود که مرکز مشخص شده میانه است. سپس، هر نقطه به نزدیک‌ترین خوشه نسبت داده می‌شود. این روش تکرارشونده برای تغییر مراکز خوشه ادامه می‌یابد تا بهترین خوشه‌بندی حاصل شود (سهرابی، رئیسی وانانی و قانونی شیشوان، ۱۳۹۴؛ ونتاین، زونگ‌شنگ و ان، ۲۰۱۳).

۶- یافته‌های پژوهش

الگوریتم‌های مورد استفاده در پژوهش حاضر در محیط نرم افزاری متلب[1] اجرا شده‌اند که نتایج آن‌ها به‌شرح زیر می‌باشد:

۶-۱- اعتبارسنجی طبقه‌بندی

با استفاده از متغیر هدفی با نام فرمول شناسایی تقلب که توسط حسابرسان (در ذهن) برای شناسایی مؤدیانی که بر اساس گزارش‌های حسابرسی، اطلاعات درستی از فعالیت‌های خود ارائه نکرده‌اند (و یا به عبارت دیگر مؤدیان متقلب) به‌کار می‌رود، مؤدیان در سه خوشه مؤدیان کم‌ریسک، ریسک متوسط و پرریسک قرار می‌گیرند.

شناسایی تقلب=کل مالیات فروش برگ مطالبه تقسیم بر کل مالیات ابرازی فروش اظهارنامه

مؤدی کم‌ریسک	۰ ≤ شناسایی تقلب ≤ ۱/۲
مؤدی با ریسک متوسط	۱/۲ ≤ شناسایی تقلب ≤ ۱/۵
مؤدی پرریسک	۱/۲ ≤ شناسایی تقلب

۹۰ درصد داده‌ها (۱۲۶۸ سطر از ۱۴۰۹ سطر داده) به صورت تصادفی به منظور انجام روش‌های مختلف طبقه‌بندی و ۱۰درصد داده‌ها (۱۴۱ سطر از ۱۴۰۹ سطر داده) به منظور اعتبارسنجی و با توجه به نتایج حسابرسی داده‌های اصلی با نسبت‌های ۷۵، ۶ و ۱۸درصد از خوشه‌های داده‌های اصلی، انتخاب می‌شوند. با توجه به این موضوع که ۷۵ درصد داده‌های اصلی را مؤدیان کم‌ریسک، ۶ درصد آن‌ها را مؤدیانی با ریسک متوسط و ۱۸ درصد باقیمانده را مؤدیان پرریسک شامل می‌شوند، بنابراین به صورت تصادفی و با توجه به نسبت‌ها، داده‌های اعتبارسنجی از میان داده‌های اصلی انتخاب شدند تا به خوبی بیانگر داده‌های اصلی باشند و نتایج دقیق‌تری به‌دست آید.

1. MATLAB R2014a

جدول (۶)- مقایسه خطاهای طبقه‌بندی

الگوریتم طبقه‌بندی	خطا	خطای آموزش
Decision Tree	۰/۰۵۴۴	۰/۱۷۰۲
Naive Bayes	۰/۸۳۴۴	۰/۷۹۴۳
K-Nearest Neighbor (KNN)	۰	۰/۲۴۱۱

مطابق جدول ۶، نتایج به‌دست آمده از مقایسه خطاهای سه روش Naive Bayes، Decision Tree و KNN که از روش‌های طبقه‌بندی به شمار می‌روند، حاکی از آن است که روش Naive Bayes در این مدل به دلیل خطای بالا کارایی ندارد، هم چنین خطای روش KNN به دلیل برازش بیش از حد[1] صفر شده است. به دلیل اینکه خطای روش درخت تصمیم و خطای آموزش (اعتبارسنجی) حاصل از آن نسبت به سه روش دیگر کم‌تر است، حاکی از برتری این روش نسبت به سه روش دیگر می‌باشد.

تحلیل درخت تصمیم

شکل ۱ بیانگر درخت تصمیمی است که با استفاده از ۹۰ درصد داده‌ها ساخته شده است.

شکل (۱)- درخت تصمیم طبقه‌بندی مؤدیان مالیات بر ارزش افزوده

منبع: یافته‌های تحقیق

جدول۷، نتایج حاصل از درخت تصمیم ساخته شده را نشان می‌دهد. خطای درخت تصمیم به میزان ۰/۰۵ بدین مفهوم است که درخت تصمیم ایجاد شده، در ۹۵ درصد موارد طبقه‌بندی صحیحی را انجام می‌دهد.

1. Over Fit

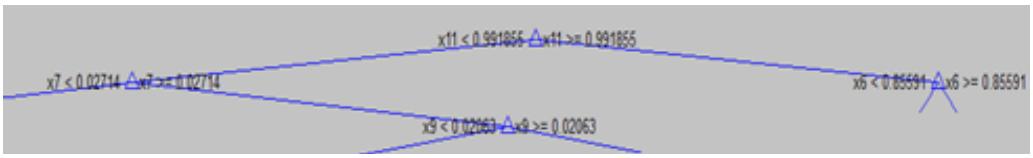
جدول (۷)- نتیجه طبقه‌بندی درخت تصمیم

سطح هرس (Pruning Level)	خطا (ResubLoss)
۱۹	۰/۰۵۴۴

منبع: یافته‌های تحقیق

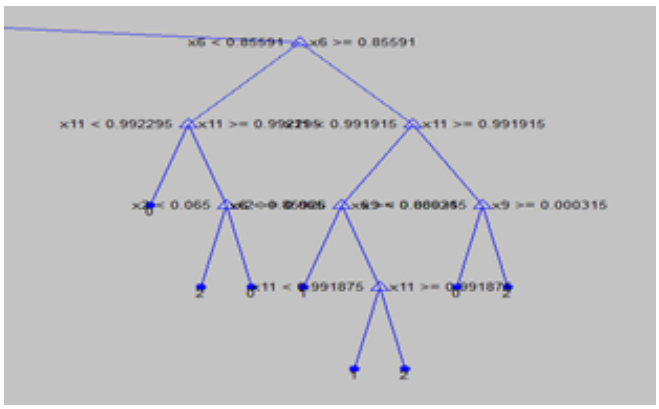
همان‌گونه که در شکل ۲ مشاهده می‌شود، مؤلفه‌ای که توانسته بیشترین تفکیک را در درخت تصمیم ایجاد کند، مالیات پرداخت شده/بدهی (x11) است. در سطح بعدی (سطح ۲)، متغیرهای مالیات قابل پرداخت/فروش مشمول مالیات (x6) و مالیات فروش/فروش (x7) منجر به تفکیک درخت تصمیم شده‌اند.

شکل (۲)- شاخه مالیات پرداخت شده/ بدهی در درخت تصمیم



شکل ۳ بیانگر این است که تمامی ۵۸ داده رسیده به مؤلفه مالیات قابل پرداخت/فروش مشمول مالیات (x6) با سه سطح تفکیک به نتیجه رسیده و درخت در شاخه (x6) به پایان می‌رسد. مالیات فروش/فروش (x7) با استفاده از کل مالیات فروش برگ مطالبه (x9) و فروش مشمول ابرازی (x1) تفکیک را ادامه می‌دهد.

شکل (۳)- شاخه مالیات قابل پرداخت/فروش مشمول مالیات در درخت تصمیم.

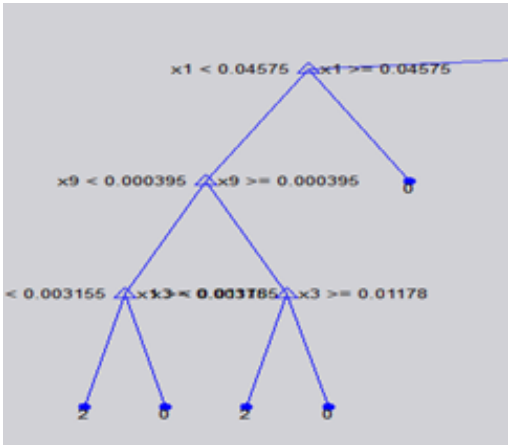


منبع: یافته‌های تحقیق

همان‌گونه که در شکل ۴ مشاهده می‌شود، در صورتی که (x1<30158457097.2345) x1<0.04575

باشد، درخت تصمیم به این نتیجه می‌رسد که این مؤدی کم‌ریسک است. در صورتی‌که x1≥0.04576 (x1≥30158457097.2345) باشد، درخت تصمیم ابتدا با استفاده از کل مالیات فروش برگ مطالبه (x9) و سپس کل مالیات ابرازی فروش اظهارنامه (x3) و فروش مشمول ابرازی (x1) به نتایج نهایی می‌رسد و در شاخه (x1) نیز درخت به پایان می‌رسد.

شکل (۴)-شاخه فروش مشمول ابرازی در درخت تصمیم



منبع: یافته‌های تحقیق

این درخت تصمیم شامل ۱۰۱ گره است که سهم هر یک از متغیرها در تعداد گره‌ها به شرح جدول ۸ می‌باشد.

جدول (۸)- سهم هریک از متغیرهای مستقل در تعداد گره‌های درخت تصمیم

شرح متغیر	تعداد گره‌ها
کل مالیات فروش برگ مطالبه	۲۷
کل مالیات ابرازی فروش اظهارنامه	۱۹
فروش مشمول ابرازی	۱۶
مالیات پرداخت شده/ بدهی	۸
خرید مشمول ابرازی	۷
خرید/ فروش	۷
کل مالیات ابرازی خرید اظهارنامه (اعتبار مالیاتی)	۶
بدهی/ فروش	۶

شرح متغیر	تعداد گره‌ها
مالیات فروش/ فروش	۲
نرخ مالیات	۲
کل مالیات خرید برگ مطالبه	۱

منبع: یافته‌های تحقیق

همان‌گونه که در جدول ۸ مشاهده می‌شود، چون اصل تقلب در فروش و مالیات فروش صورت می‌گیرد، بنابراین بیشترین سهم تقسیم درخت تصمیم به ترتیب متعلق به کل مالیات فروش برگ مطالبه با ۲۷ گره، کل مالیات ابرازی فروش اظهارنامه با ۱۹ گره و فروش مشمول ابرازی با ۱۶ گره می‌باشد. به عبارت دیگر با این سه ویژگی با درصد بالایی می‌توان تقلب را تشخیص داد.

پس از فروش و مالیات فروش، خرید و مالیات خرید، متغیرهای بعدی هستند که تقلب روی آن‌ها صورت می‌گیرد، بنابراین مؤلفه‌های مالیات پرداخت شده/بدهی با ۸ گره، خرید مشمول ابرازی و خرید/فروش هر کدام با ۷ گره، کل مالیات ابرازی خرید اظهارنامه و مالیات قابل پرداخت/فروش مشمول مالیات هر کدام با ۶ گره، در سطح بعدی شناسایی تقلب قرار می‌گیرند. هم‌چنین چون نرخ مالیات برای هر سال مشخص عدد معینی است (بدین مفهوم که در هر سال عدد تعیین شده متعلق به آن سال می‌باشد)، امکان تقلب در آن وجود ندارد ولی در این پژوهش برای بررسی تاثیرگذاری آن بر سایر متغیرهای منتخب و رفتار متقلبانه در مالیات بر ارزش افزوده به‌کار می‌رود.

۶-۲- اعتبارسنجی خوشه‌بندی

شاخص‌های اعتبارسنجی برای سنجش میزان صحت نتایج خوشه‌بندی به منظور مقایسه بین روش‌های مختلف خوشه‌بندی مورد استفاده قرار می‌گیرند. در این پژوهش از شاخص اعتبارسنجی سیلوئت استفاده شده است که در ادامه بطور مختصر شرح داده می‌شود.

شاخص سیلوئت

شاخص سیلوئت، یکی از معیارهای متداول اعتبارسنجی خوشه‌بندی است که دو معیار فواصل درون‌خوشه‌ای و برون‌خوشه‌ای را همزمان در نظر می‌گیرد.

جدول (۹)- تفسیر مقادیر میانگین معیار سیلوئت

میانگین معیار سیلوئت	تفسیر
۱-۰/۷	ساختار قوی

میانگین معیار سیلوئت	تفسیر
۰/۵۱–۰/۷	ساختار منطقی (مناسب)
۰/۲۵–۰/۵	ساختار ضعیف
<۰/۲۵	هیچ ساختار قابل توجهی وجود ندارد.

منبع: پازوکی، سپهری، صابری فیروزی، ۱۳۹۳

جدول ۹ تفسیر مقادیر مختلف معیار سیلوئت را نشان می‌دهد. با توجه به داده‌ها و پس از پیاده‌سازی ارزیابی سیلوئت که یکی از تکنیک‌های ارزیابی به منظور تعیین تعداد خوشه های بهینه است، تعداد بهینه دو خوشه برای الگوریتم‌های k-means و k-medoids انتخاب شد.

جدول (۱۰)- نتایج اعتبارسنجی خوشه‌بندی شاخص سیلوئت

الگوریتم خوشه‌بندی	تعداد خوشه	اعتبار سنجی شاخص سیلوئت
K-means	۲	Silhouette K-means =۰/۹۰۹۳
K-medoids	۲	Silhouette K-medoids =۰/۹۰۹۳

منبع: یافته‌های تحقیق

با توجه به نتایج مقادیر شاخص سیلوئت حاصل از الگوریتم‌های خوشه‌بندی که در جدول ۱۰ بیان شده و مقایسه آن‌ها با تفسیر محدوده این مقادیر که در جدول ۹ به آن اشاره شد، نتایج اعتبارسنجی خوشه‌بندی حاکی از عملکرد خوب هر دو الگوریتم خوشه‌بندی k-means و k-medoids و بیانگر برتری الگوریتم خوشه‌بندی k-means بر الگوریتم خوشه‌بندیk-medoids است.

پس از پیاده‌سازی ارزیابی سیلوئت به منظور تعیین تعداد خوشه بهینه، تعداد بهینه دو خوشه برای الگوریتم k-means به‌دست آمد که جدول ۱۱، بیانگر مراکز این خوشه‌ها است.

جدول (۱۱)- مراکز خوشه‌های به‌دست آمده از الگوریتم k-means

فروش مشمول ابرازی	نرخ مالیات	کل مالیات ابرازی فروش اظهارنامه	خرید مشمول ابرازی	کل مالیات ابرازی خرید اظهارنامه	مالیات قابل پرداخت	مالیات فروش/ فروش	خرید/ فروش	کل مالیات فروش برگ مطالبه	کل مالیات خرید برگ مطالبه	مالیات پرداخت شده/ بدهی
۰/۰۲۴۳۵	۰/۰۵۰۹۴	۰/۰۱۹۳۲	۰/۰۱۳۷۶	۰/۰۱۵۶۳	۰/۸۵۲۵۲	۰/۰۴۵۸۹	۰/۰۱۲۹۰	۰/۰۰۸۹۸	۰/۰۱۲۶۲	۰/۹۹۰۸۸
۰/۳۱۹۶۱	۰/۰۶۹۸۷	۰/۳۲۴۳۷	۰/۲۱۷۲۱	۰/۳۱۸۴۴	۰/۸۵۵۷۴	۰/۰۵۷۰۷	۰/۰۰۷۶۲	۰/۲۰۵۴۳	۰/۲۴۳۱۵	۰/۹۹۱۷۰

منبع: یافته‌های تحقیق

بنابراین با استفاده از شاخص‌های مالیاتی، رفتار خوشه‌های به دست آمده که در جدول ۱۱ بیان شده، تحلیل می‌شوند.

تحلیل رفتار خوشه‌ها

خروجی الگوریتم‌های طبقه‌بندی در نرم‌افزار متلب، مراکز خوشه‌ای است که الگوی خاصی از رفتار داده‌ها درون آن‌ها پنهان است و این امر بر عهده تحلیل‌گر است که به این الگوهای پنهان دست یابد. بنابراین تحلیل مراکز خوشه با نظر خبرگان مالیاتی انجام می‌گیرد و در خوشه اول چهار شاخص و در خوشه دوم سه شاخص که بیش از همه نشان دهنده الگوی رفتاری داده‌های متعلق به آن خوشه است معیار تحلیل قرار می‌گیرد.

با توجه به اینکه در خوشه اول نسبت اکثر شاخص‌ها از جمله چهار شاخص اصلی به شرح زیر، کمتر از حد نرمال (استاندارد تعیین شده بر اساس نظر حسابرسان) تعریف شده برای هر شاخص می‌باشد، بنابراین می‌توان نتیجه گرفت که خوشه اول به عنوان خوشه کم‌ریسک محسوب می‌شود.

- نسبت خرید مشمول مالیات ابرازی به فروش مشمول مالیات ابرازی ۵۲ درصد و کمتر از یک است.
- اختلاف بین مالیات فروش مطالبه شده (رسیدگی شده در حسابرسی) و مالیات فروش ابرازی مؤدی ۱۸ درصد و کمتر از ۲۰ درصد حد نرمال است.
- اختلاف بین مالیات خرید تأیید شده (رسیدگی شده در حسابرسی) و مالیات خرید ابرازی مؤدی ۱۹ درصد و کمتر از ۲۰ درصد حد نرمال است.
- مالیات فروش مطالبه شده (رسیدگی شده در حسابرسی) نسبت به فروش ابرازی در مقایسه با مالیات فروش ابرازی به فروش ابرازی ۱۸ درصد و کمتر از ۲۰ درصد حد نرمال است.

با توجه به اینکه شاخص‌های مطرح شده دارای کمترین اختلاف با حد نرمال تعریف شده می‌باشد، در نتیجه این خوشه، کم‌ریسک محسوب شده و به عبارت دیگر مؤدیان قرار گرفته در این خوشه از جمله مؤدیان با ریسک کم تلقی می‌گردند و چون اظهارات این دسته از مؤدیان نزدیک به واقعیت بوده و در صورت رسیدگی به پرونده آن‌ها اختلاف کمی بین مالیات ابرازی و پرداخت شده با مالیات مطالبه شده در راستای رسیدگی مشاهده می‌شود و این اختلاف کم از دیدگاه تحلیل هزینه فایده به صرفه نخواهد بود (به عبارت دیگر به میزانی که هزینه و زمان برای رسیدگی به پرونده این دسته از مؤدیان صرف می‌شود، مالیات به دست آمده کفایت هزینه‌های پیشین را نخواهد داشت)، پس در نتیجه باید به این مؤدیان اعتماد نمود، و اظهارنامه ارائه شده آنان را بدون رسیدگی پذیرفت.

بر اساس استدلال مطرح شده در بررسی رفتار خوشه اول، خوشه دوم نیز چنین تحلیل می‌شود:

با توجه به اینکه در خوشه دوم نسبت اکثر شاخص‌ها از جمله سه شاخص اصلی به شرح زیر، بیشتر از حد نرمال (استاندارد تعیین شده بر اساس نظر حسابرسان) تعریف شده برای هر شاخص می‌باشد، بنابراین می‌توان نتیجه گرفت که خوشه دوم به عنوان خوشه پرریسک محسوب می‌شود.

- اختلاف بین مالیات فروش مطالبه شده (رسیدگی شده در حسابرسی) و مالیات فروش ابرازی مؤدی ۶۱ درصد به دست آمده و بسیار بیشتر از ۲۰ درصد حد نرمال است.
- اختلاف بین مالیات خرید تأیید شده (رسیدگی شده در حسابرسی) و مالیات خرید ابرازی مؤدی ۳۰ درصد به دست آمده و بسیار بیشتر از ۲۰ درصد حد نرمال است.
- مالیات فروش مطالبه شده (رسیدگی شده در حسابرسی) نسبت به فروش ابرازی در مقایسه با مالیات فروش ابرازی به فروش ابرازی ۶۰ درصد به دست آمده و بسیار بیشتر از ۲۰ درصد حد نرمال است.

با توجه به اینکه شاخص‌های مطرح شده دارای اختلاف زیادی با حد نرمال تعریف شده می‌باشد، در نتیجه این خوشه پرریسک محسوب شده و به عبارت دیگر مؤدیان قرار گرفته در این خوشه از جمله مؤدیان دارای ریسک بالا تلقی می‌گردند و چون اظهارات این دسته از مؤدیان غیر واقعی است، باید رسیدگی به پرونده مالیاتی آنان در اولویت قرار گیرد. از طرف دیگر رسیدگی مؤدیان پرریسک موجب عدم تکرار تخلف آن‌ها در دوره‌های مالیاتی بعدی خواهد شد و از این منظر موجب کاهش هزینه‌های حسابرسی دوره‌های مالیاتی آتی می‌شود.

۶-۳- نتایج طبقه‌بندی بعد از خوشه‌بندی

بعد از انجام خوشه‌بندی، با استفاده از دو خوشه به‌دست آمده از الگوریتم k-means، سه الگوریتم طبقه‌بندی Naive Bayes، Decision Tree و KNN روی ۹۰ درصد داده‌ها (۱۲۶۸ سطر از ۱۴۰۹ سطر داده) به منظور انجام روش‌های مختلف طبقه‌بندی و۱۰درصد داده‌ها (۱۴۱ سطر از ۱۴۰۹ سطر داده) به منظور اعتبارسنجی اجرا می‌شود و داده‌های اعتبارسنجی با توجه به نتایج خوشه‌ها با نسبت‌های ۹۴/۵ و ۵/۵ درصد از خوشه‌های کم‌ریسک و پرریسک، انتخاب می‌شوند.

جدول (۱۲)- مقایسه خطاهای طبقه‌بندی-ترکیب خوشه‌بندی و طبقه‌بندی

الگوریتم طبقه‌بندی	خطا	خطای آموزش
Decision Tree	۰/۰۰۳۹	۰/۰۳۵۵
Naive Bayes	۰/۰۵۶۰	۰/۰۰۷۱
K-Nearest Neighbor (KNN)	۰	۰/۰۹۲۲

منبع: یافته‌های تحقیق

جدول ۱۲، نتایج به‌دست آمده از مقایسه خطاهای سه روش Naive Bayes، Decision Tree و KNN را نشان می‌دهد. شکل ۵ بیانگر درخت تصمیمی است که با استفاده از ۹۰ درصد داده‌ها ساخته شده است.

شکل (۵)- درخت تصمیم مؤدیان مالیات بر ارزش افزوده با استفاده از نتایج خوشه‌بندی

منبع: یافته‌های تحقیق

نکته قابل توجه این است که در تشکیل درخت تصمیم با استفاده از نتایج خوشه‌بندی، تعداد زیادی از متغیرها دخالتی ندارند، متغیرهای ایجاد کننده این درخت تصمیم به شرح جدول ۱۳ می‌باشند.

جدول (۱۳)- سهم متغیرها در تعداد گره‌های درخت تصمیم- ترکیب طبقه‌بندی و خوشه‌بندی

شرح متغیر	تعداد گره‌ها
فروش مشمول ابرازی	۱
کل مالیات ابرازی خرید اظهارنامه (اعتبار مالیاتی)	۱
کل مالیات فروش برگ مطالبه	۱
کل مالیات خرید برگ مطالبه	۱

منبع: یافته‌های تحقیق

جدول ۱۴، نشان‌دهنده نتایج حاصل از درخت تصمیم با استفاده از ترکیب خوشه‌بندی و طبقه‌بندی است.

جدول (۱۴)- نتیجه طبقه‌بندی درخت تصمیم- ترکیب خوشه‌بندی و طبقه‌بندی

سطح هرس (Pruning Level)	خطا (ResubLoss)
۴	۰/۰۰۳۹

منبع: یافته‌های تحقیق

میزان خطا به میزان ۰/۰۰۳ بدین مفهوم است که درخت تصمیم ایجاد شده با استفاده از نتایج خوشه‌بندی، در ۹۹/۷ درصد موارد، طبقه‌بندی صحیحی را انجام می‌دهد (به عبارت دیگر بر اساس این درخت تصمیم تنها برای ۰/۳ درصد از مؤدیان طبقه‌بندی صحیحی از نظر کم‌ریسک و یا پرریسک بودن انجام نمی‌شود).

۷- نتیجه‌گیری و پیشنهادات

در پژوهش حاضر، متغیرهای منتخب برای داده‌کاوی، از طریق داده‌های اظهارنامه‌های مالیاتی مؤدیان مالیات بر ارزش افزوده شهر تهران در سال‌های مورد مطالعه (۱۳۹۳-۱۳۸۸) از سوی سازمان امور مالیاتی کشور مورد حسابرسی قرار گرفته‌اند.

پس از آماده‌سازی و پاک‌سازی داده‌ها، شامل شناسایی داده‌های خارج از محدوده با تکنیک شش سیگما، حذف سطرهای دارای مقادیر مفقوده و نرمال سازی داده‌ها، بررسی روی ۱۴۰۹ سطر از داده‌های مؤدیان انجام شد.

به منظور داده‌کاوی، الگوریتم‌های طبقه‌بندی شامل سه روش Naive Bayes، Decision Tree و KNN و الگوریتم‌های خوشه‌بندی شامل دو روش k-means و k-medoids مورد استفاده قرار گرفتند.

از میان الگوریتم‌های طبقه‌بندی، روش درخت تصمیم با خطای ۰/۰۵ و خطای آموزش (اعتبارسنجی) ۰/۱۷ نتیجه بهتری کسب کرد. از میان متغیرها، کل مالیات فروش برگ مطالبه، کل مالیات ابرازی فروش اظهارنامه و فروش مشمول ابرازی، بیشترین سهم تقسیم درخت تصمیم را به خود اختصاص دادند. به‌عبارت دیگر چون اصل تقلب در فروش و مالیات فروش صورت می‌گیرد، با این سه ویژگی با درصد بالایی می‌توان تقلب را تشخیص داد.

با توجه به داده‌ها و پس از پیاده‌سازی ارزیابی سیلوئت که یکی از تکنیک‌های ارزیابی به منظور تعیین تعداد خوشه بهینه است، تعداد بهینه دو خوشه برای الگوریتم‌های k-means و k-medoids انتخاب شد.

از میان الگوریتم‌های خوشه‌بندی، روش k-means با شاخص اعتبارسنجی سیلوئت ۰/۹۱ به عنوان روش برگزیده در خوشه‌بندی انتخاب شد. بر اساس تحلیل‌های صورت گرفته بر روی شاخص‌های مالیاتی مراکز دو خوشه به‌دست آمده از روش k-means، این دو خوشه با نام‌های کم‌ریسک و پرریسک نام‌گذاری شدند.

با توجه به نتایج حاصل از طبقه‌بندی درخت تصمیم با استفاده از نتایج خوشه‌بندی با الگوریتم k-means، درخت تصمیم حاصل دارای خطای ۰/۰۰۳ و خطای آموزش (اعتبارسنجی) ۰/۰۳۵ است. خطای ۰/۰۰۳ بدین مفهوم است که به وسیله درخت تصمیم ایجاد شده با استفاده از نتایج خوشه‌بندی، در ۹۹/۷ درصد

موارد طبقه‌بندی صحیحی انجام می‌شود (به‌عبارت دیگر بر اساس این درخت تصمیم تنها برای ۰/۳ درصد از مؤدیان طبقه‌بندی صحیحی از نظر کم‌ریسک و یا پرریسک بودن انجام نمی‌شود.)

نتایج حاکی از آن است که استفاده از روش‌های داده‌کاوی به سازمان امور مالیاتی کشور و حسابرسان این امکان را می‌دهد که با استفاده از حداقل زمان و هزینه، گزارشی را ارائه کنند که مستند به روش‌های علمی بوده و از اتکا پذیری و اطمینان بالایی برخوردار است.

پیشنهادها و توصیه‌های سیاستی حاصل از این پژوهش به شرح زیر است:

- به‌کار بردن الگوریتم‌های طبقه‌بندی و خوشه‌بندی برای تشخیص و پیش‌بینی تقلب مالیاتی در سایر منابع مالیاتی (مالیات بر حقوق، مالیات بر اجاره املاک و ...).
- افزایش جامعه آماری مورد مطالعه برای قوی‌تر شدن نتایج پیش‌بینی (زیرا بدون شک در پروژه‌های داده‌کاوی هرچه اندازه نمونه بزرگ‌تر باشد، قوانین به‌دست آمده دقیق‌تر و نتایج به‌دست آمده قابلیت بیشتری برای تعمیم به کل جامعه را دارد).
- علاوه بر اطلاعات موجود در فرم اظهارنامه، اطلاعات مربوط به ریز فاکتورهای خرید و فروش نیز به‌صورت الکترونیکی از مؤدیان اخذ شود، که این امر به قوی‌تر شدن جامعه اطلاعاتی و تشخیص دقیق‌تر کمک می‌کند.
- از پذیرش اظهارنامه‌هایی که تمامی فیلدهای مربوط به آن پر نشده‌اند، جلوگیری شود.

فهرست منابع

۱. باقرپور ولاشانی، محمدعلی، باقری، مصطفی، خادم، حمید و رضا حسینی‌پور (۱۳۹۱). بررسی عوامل مالی و غیرمالی مؤثر بر گریز مالیاتی با استفاده از تکنیک‌های داده کاوی: صنعت خودرو و ساخت قطعات. مطالعات تجربی حسابداری مالی، ۱(۳۴): ۱۲۸-۱۰۳ .

۲. باقرپور ولاشانی، محمدعلی، ساعدی، محمدجواد، مشکانی، علی و مصطفی باقری (۱۳۹۱). پیش‌بینی گزارش حسابرس مستقل در ایران: رویکرد داده‌کاوی. دهمین همایش حسابداری ایران، دانشگاه الزهرا.

۳. برزگری خانقاه، جمال و محمدعلی فیض‌پور (۱۳۹۲). حسابرسی مالیاتی مبتنی بر ریسک: تجاربی از کشورهای توسعه یافته و در حال توسعه. پژوهش حسابداری.

۴. پازوکی، مینا، سپهری، محمدمهدی و مهدی صابری فیروزی (۱۳۹۳). کشف ساختارهای خوشه‌های پنهان در بیماران مبتلا به سیروز کبدی بر پایه نشانه‌های آزمایشگاهی. فصلنامه گوارش. ۱۹(۳) : ۱۹۱- ۱۹۷.

۵. رحیمی کیا، اقبال، محمدی، شاپور و مهدی غضنفری (۱۳۹۴). تشخیص فرار مالیاتی با استفاده از سیستم هوشمند ترکیبی. پژوهشنامه مالیات. ۲۳ (۲۶) :۱۳۶-۱۶۴.

۶. سهرابی، بابک، رئیسی وانانی، ایمان و وحیده قانونی شیشوان (۱۳۹۴). ارزیابی مالیات عملکرد شرکت‌ها و تحلیل روندهای مالیاتی با استفاده از الگوریتم‌های داده‌کاوی. تحقیقات مالی. ۱۷(۴۰): ۲۱۹-۲۳۸ .

۷. شهرابی، جمال (۱۳۹۲). داده‌کاوی. تهران: جهاد دانشگاهی، واحد صنعتی امیرکبیر.

۸. موسوی جهرمی، یگانه، طهماسبی بلداجی، فرهاد و نرگس خاکی (۱۳۸۸). فرار مالیاتی در نظام مالیات بر ارزش افزوده: یک مدل نظری. پژوهشنامه مالیات. ۱۷(۵) :۲۷-۳۸.

۹. نوروزی، مستوره و محمدرضا تقوا (۱۳۹۴). سیستم ارزیابی هزینه‌های درمانی با استفاده از روش‌های داده‌کاوی. دانشگاه علامه طباطبائی، مدیریت صنعتی گرایش تولید.

10. Alm, J. & J. M. Vazquez (2001). Institutions, Paradigms, and Tax Evasion in Developing and Transition Countries. Paper Presented for the Conference Public Finance in Developing and Transition Countries, Georgia State University, U.S.A.

11. Bremner, D., Demaine, E., Erickson, J., Iacono, J., Langerman, S., Morin, P., & G. Toussaint (2005). Output-sensitive Algorithms for Computing Nearest-neighbor Decision Boundaries. Discrete and Computational Geometry 33 (4): 593-604.

12. Fadairo, S. A. Williams, R. Trotman, R., & A. Onyekelu-Eze (2008). Using Data Mining to Ensure Payment Integrity. Journal of Government Financial Management, 57: 22-24.
13. Glenn, J. M. (2007). Making Sense of Data, a Practical Guide to Exploratory Data Analysis and Data Mining. Wiley Interscience.
14. Gonzalez, P. C., & J. D. Velasquez (2013). Characterization and Detection of Taxpayers with False Invoices Using Data Mining Techniques. Expert Systems with Applications, 40: 1427-1436.
15. Kirkos, E., Spathis, C. & Y. Manolopoulos (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. Expert Systems with Applications, 32(4): 995-1003.
16. Lundin, E., Kvarnstrom, H., & E. Jonsson (2003). Synthesizing Test Data for Fraud Detection Systems. In Proceedings of the 19th Annual Computer Security Applications Conference. 384-394.
17. Wentian, J., Sheng, Z. G. & Z. En (2013). Improved k-medoids Clustering Algorithm under Semantic web. Proceedings of the 2nd International Conference on Computer Science and Electronic Engineering (ICCSEE 2013).
18. Wu, R. S., Ou, C. S., Chang, S. I., & D. C. Yen (2012). Using Data Mining Technique to Enhance Tax Evasion Detection Performance. Expert Systems with Applications, 39: 8769-8777.
19. Zhou, W., & G. Kapoor (2011). Detecting Evolutionary Financial Statement Fraud. Decision Support Systems, 50(3): 570- 575 .