توسعه مدل شناسایی مؤدیان کم‌اظهار در مالیات بر ارزش افزوده با استفاده از رویکردهای داده‌کاوی

وحید برادران[1]
شیما محمدحسنی[2]

تاریخ دریافت: ۹۶/۴/۲۴، تاریخ پذیرش: ۹۵/۱۱/۲۷

چکیده

تعداد زیاد اظهارنامه مالیاتی، محدودیت منابع و مقرون به صرفه نبودن بررسی تمامی آن‌ها، توسعه روشی هوشمند جهت شناسایی مؤدیان کم‌اظهار را ضروری می‌نماید. در این مقاله، بر اساس نظرات ممیزین مالیاتی، داده‌های هجده متغیر بالقوه مؤثر بر شناسایی کم‌اظهاری مالیات بر ارزش افزوده در یکی از مناطق تهران به همراه نتایج ممیزی آن‌ها جمع‌آوری شده است. روش‌های فیلتری و روش الگوریتم ژنتیک تعداد متغیرهای مؤثر را به ترتیب ده و هفت متغیر شناسایی کرده‌اند. دو روش پایه رده‌بندی «درخت تصمیم» و «k نزدیک‌ترین همسایگی» بر اساس دو نوع متغیرهای مؤثر (روش‌های فیلتری و الگوریتم ژنتیک) برای شناسایی کم‌اظهاری توسعه داده شده و برای توازن داده‌ها دو روش جمعی «بگینگ» و «بوستینگ» استفاده شده است. بررسی دقت پیش‌بینی در دوازده مدل پیش‌بینی (درخت تصمیم و K نزدیکترین همسایگی با دو گروه متغیر مستقل و در سه حالت «عادی»، «بگینگ» و «بوستینگ») نشان می‌دهد، روش‌های جمعی «بگینگ» و «بوستینگ» تأثیری بر پیش‌بینی ندارند و درخت تصمیم ساده با ده متغیر منتخب با روش‌های فیلتری، بیشترین دقت پیش‌بینی و معادل ۱۴٫۸۲٪ را برای تشخیص مؤدیان کم‌اظهار دارد. استخراج قوانین مناسب برای تشخیص مؤدیان کم‌اظهار بر اساس ده متغیر مؤثر بر پیش‌بینی آن‌ها از دیگر نتایج این مقاله است.

واژه‌های کلیدی: داده‌کاوی، مؤدی مالیاتی، کم‌اظهاری مالیات، مالیات بر ارزش افزوده

۱. استادیار گروه مهندسی صنایع، دانشگاه آزاد اسلامی واحد تهران شمال (نویسنده مسئول) V_baradaran@iau-tnb.ac.ir
۲. کارشناسی ارشد مهندسی صنایع، گرایش مدیریت سیستم و بهره‌وری دانشگاه آزاد اسلامی واحد تهران شمال sh.inta@chmail.ir

۱- مقدمه

یکی از منابع اصلی و مهم درآمدی دولت‌ها، مالیات می‌باشد که تحت عناوین مختلف وصول می‌گردد. بخش قابل ملاحظه‌ای از بودجه دولت‌ها از درآمد حاصل از مالیات به دست می‌آید، به‌طوری که در کشورهایی که مالیات از نظام قانونی و مردمی برخوردار است، بیش از ۶۰ درصد بودجه عمومی را مالیات تشکیل می‌دهد. در این میان مالیات بر ارزش افزوده[1] (VAT) در بیش از ۱۳۰ کشور که در مراحل مختلف توسعه اقتصادی قرار دارند، پیاده‌سازی شده، به طوری که در برخی از کشورها حدود ۲۵ درصد از درآمدهای مالیاتی را به خود اختصاص داده است (هریسون و کریلوف، ۲۰۰۵). در ایران نیز قانون مالیات بر ارزش افزوده از نیمه دوم سال ۱۳۸۷ اجرا شده است.

فرار از پرداخت مالیات و تقلّب مالیاتی همواره یک نگرانی دائمی برای سازمان‌های مالیاتی، به‌ویژه در کشورهای در حال توسعه می‌باشد (داویا و همکاران، ۲۰۰۰). از این‌رو یکی از دغدغه‌ها و مسائل پیش‌روی سازمان‌های مالیاتی، توسعه روش‌های قابل اعتماد و دقیق مبتنی بر داده‌ها و اطلاعات پیشین مؤدیان مالیاتی برای تشخیص تقلّب مالیاتی است. انتخاب تصادفی و یا تمرکز بر مؤدیانی که تاکنون در دوره‌های اخیر حسابرسی نشده‌اند و یا انتخاب گزینه‌های مشکوک به تقلّب بر اساس تجربه و دانش حسابرسان و ممیزین از جمله روش‌های کشف موارد مشکوک به تقلّب می‌باشد. توسعه روش‌های مبتنی بر تحلیل‌های آماری، ایجاد سیستم‌های مبتنی بر قاعده[2] یا مدل‌های ریسک از دیگر روش‌های کشف تقلّب مالیاتی است که در سال‌های اخیر مورد استفاده قرار گرفته‌اند (سازمان توسعه و همکاری‌های اقتصادی، ۱۹۹۹).

در بخش مالیات بر ارزش افزوده، هر مؤدی موظف است عملکرد سه‌ماهه خود را تحت عنوان اظهارنامه به سازمان مالیاتی تحویل دهد. مواردی وجود دارد که مؤدیان جهت پرهیز از پرداخت مالیات، مبالغ خود اظهاری را کمتر از مقدار واقعی گزارش می‌دهند. جهت کشف این موارد، ممیزین می‌بایست پرونده‌های منتخب را با دقت بیشتری مطالعه نمایند تا موارد کم‌اظهاری مالیات کشف شوند. حجم بالای تعداد اظهارنامه‌ها، محدودیت منابع انسانی متخصص، زمان‌بر بودن و حتی مقرون به‌صرفه نبودن بررسی برخی از آن‌ها، انجام فرایند شناسایی مؤدیان کم‌اظهار را دشوار نموده است و سازمان‌های مالیاتی از جمله سازمان امور مالیاتی ایران، تنها توانایی بررسی تعداد محدودی از آن‌ها را دارند. از طرف دیگر، در طول سال‌های اجرای قانون مالیات بر ارزش افزوده در کشور، اظهارنامه‌های نسبتاً زیادی توسط ممیزین سازمان مالیاتی کشور بررسی شده و نتایج آن در پایگاه‌های اطلاعاتی مربوطه ثبت شده است. ثبت مشخصات مؤدیان،

1. Value Added Tax
2. Rule-based

ویژگی‌های کسب و کار آن‌ها، داده‌های خود اظهاری و عملکردی و نتیجه بررسی آن‌ها توسط ممیزین، منبع مناسبی برای پردازش اطلاعات و کشف دانش در مورد تقلّب در اظهارنامه‌های مالیات بر ارزش افزوده فراهم نموده است و همچنین توسعه استفاده از صندوق‌های مکانیزه فروش نیز در آینده، اطلاعات با ارزشی را در این زمینه در اختیار سازمان امور مالیاتی کشور قرار خواهد داد.

با توجه به اهمیت مسئله کشف تقلّب مالیاتی و داشتن داده‌های مربوط به پرونده‌های بررسی شده، هدف این پژوهش شناسایی مؤلفه‌های مؤثر بر کشف کم‌اظهاری مالیاتی و ارائه مدلی جهت پیش‌بینی مؤدیان بالقوه کم‌اظهار و انتخاب اظهارنامه غیر واقعی جهت رسیدگی است. مدل پیشنهادی که بر اساس داده‌ها و رفتار مؤدیان در گذشته ساخته می‌شود، قابلیت بررسی پرونده‌های جدید را جهت کشف موارد کم‌اظهار فراهم می‌کند. تکنیک‌های داده‌کاوی[1] با قابلیت کاوش در داده‌های عظیم[2]، امکان استخراج و تولید دانش را از حجم بالایی از داده‌ها جهت شناسایی و تشخیص رفتارهای جعلی و قصور در پرداخت مالیات و در نهایت بهبود استفاده از منابع را فراهم می‌آورند (فایاد و همکاران،۱۹۹۶). لذا توسعه روش‌های مبتنی بر داده‌کاوی، تحقق اهداف این پژوهش را فراهم می‌کند.

۲- مروری بر ادبیات موضوع

مالیات بر ارزش افزوده نوعی مالیات غیر مستقیم عام بر عموم کالاها و خدمات (مگر موارد معاف) است که به‌صورت چند مرحله‌ای از اضافه ارزش کالاهای تولید شده (ارزشی که در هر مرحله به ارزش کالا افزوده می‌شود) یا خدمات ارائه شده در مراحل مختلف تولید و توزیع اخذ می‌شود. این نظام جدید که مبنای آن خوداظهاریِ مالیاتی است بر جلب مشارکت و همکاری هرچه بیشتر مؤدیان مالیاتی تکیه دارد. به همراه خوداظهاری در پرداخت این نوع مالیات، امکان تقلّب و کم‌اظهاری وجود دارد. در تعاریف موجود در ادبیات موضوع، از تقلّب به عنوان سوء استفاده از سود یک شرکت یا سازمان بدون این‌که لزوماً به عواقب قانونی و حقوقی آن توجه شود، نام برده شده است. تقلّب همچنین فرآیندی است که در آن یک یا چند نفر، عمداً و به صورت پنهانی دیگران را از هر چیز با ارزشی، به خاطر منافع شخصی خود محروم می‌کنند (فوآ و همکاران، ۲۰۱۰).

تقلّب مالی به‌طور معمول به دو صورت سوء استفاده مالی و تنظیم صورت‌حساب‌های متقلّبانه ظاهر می‌شود. در نوع اول، فرد یا افراد سعی می‌کنند از دارایی‌های فرد یا افراد دیگر به‌طور غیرقانونی استفاده نمایند. مانند اختلاس و سرقت دارایی‌های مشهود یا نامشهود و موارد دیگر. این کار اغلب با فعالیت‌های

1. Data Mining Techniques
2. Big Data

مجرمانه دیگر مانند درست‌کردن مدارک و سوابق ساختگی یا گمراه‌کننده همراه است (وزارت اقتصاد و دارایی، ۱۳۸۴). صورت حساب‌های متقلّبانه بر اساس دست‌کاری و مخدوش‌کردن حساب‌های مالی ایجاد می‌شود. برای نمونه ایجاد صورت حساب‌های جعلی و به تعویق انداختن ارائهٔ گزارشات مالی از این نوع تقلّبات می‌باشد. کم‌اظهاری در مالیات بر ارزش افزوده از نوع دوم تقلّبات مالی محسوب می‌شود.

تعدادی از محققان در زمینهٔ توسعه و به‌کارگیری روش‌هایی برای کشف انواع تقلّب مالی به پژوهش و مطالعه پرداخته‌اند. نجای و همکاران (۲۰۱۱) ضمن مرور ۴۹ مقاله منتشرشده بین سال‌های ۱۹۹۷ تا ۲۰۱۱، روش‌های کشف تقلّب مالی در سال‌های گذشته را جمع‌آوری کردند. در این مقاله، انواع تقلّب مالی در چهار دسته بانکی، بیمه‌ای، امنیتی و سایر دسته‌بندی شده‌اند و تکنیک‌هایی مانند رده‌بندی، رگرسیون[1]، خوشه‌بندی ، پیش‌بینی، کشف نقاط دور افتاده[2] و مصورسازی[3] از جمله روش‌های داده‌کاوی به عنوان روش‌های کشف تقلّب معرفی شده است.

در اکثر تحقیقات صورت گرفته در خصوص شناسایی تقلّب در اظهارنامه‌های مالیاتی، عدم اعلام مالیات توسط مؤدی به عنوان یک رفتار متقلّبانه مدل‌سازی شده است و کشف مالیات‌های اعلام نشده به‌عنوان یک مسئله محسوب می‌گردد و کمتر تحقیقی با موضوع کم‌اظهاری مالیاتی به عنوان تقلّب انجام شده است (هسو و همکاران، ۲۰۱۵). دیلون و هادزیچ (۲۰۰۹) تحقیقات پیرامون حوزه تقلّب در اظهارنامه‌های مالیاتی را به چهار دسته زیر تقسیم‌بندی نمودند:

- هزینه‌های مستقیم و غیر مستقیم تقلّب
- روابط و مسئولیت‌های افراد در بازار که سهمی از تقلّب دارند
- مکانیزم تقلّب مالی
- نقش تکنولوژی در کشف تقلّب در گزارشات مالی.

به عنوان جمع‌بندی مرور ادبیات تحقیق باید اشاره شود، هرچند به‌منظور تشخیص انواع تقلّب از روش‌های مختلف از جمله رده‌بندی، خوشه‌بندی و کشف قواعد وابستگی استفاده شده، اما در تحقیقات کمی، مسئله بررسی کشف تقلّب در مالیات بررسی شده است. برخی از تحقیقات اخیر مانند هسو و همکاران (۲۰۱۵) متذکر شده‌اند که روش‌های داده‌کاوی به منظور کشف مؤدیانی که سود بیشتری پس از بررسی اظهارنامه‌های ایشان کسب می‌شود، روش‌های بسیار کارآمدتری نسبت به روش‌های دستی و

1. Clustering
2. Outlier Detection
3. Visualization

سنتی می‌باشند. همچنین تحقیقات کمی در این حوزه انجام شده که در آن‌ها ابتدا متغیرهای تأثیرگذار به روش انتخاب ویژگی[1] شناسایی شده باشند. در این تحقیقات نویسندگان عموماً به استفاده از یک روش انتخاب ویژگی اکتفا کرده‌اند و یا مانند کرکاس و همکاران (۲۰۰۷) و رویسانکار و همکاران (۲۰۱۱) تنها از روش‌های آماری به این منظور استفاده کرده‌اند.

مسئله انتخاب ویژگی، یکی از مسائلی است که در مبحث یادگیری ماشین و همچنین شناسایی آماری الگو مطرح است. این مسئله در بسیاری از کاربردها (مانند طبقه‌بندی) اهمیت به‌سزائی دارد. زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد که بسیاری از آن‌ها یا بلااستفاده هستند و یا این‌که بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر، بالا می‌برد. علاوه بر این باعث می‌شود که اطلاعات غیرمفید زیادی به همراه داده‌های مفید ذخیره شود. گونزالس و ولاسکوئز (۲۰۱۳) استفاده از مدل‌های داده‌کاوی را در حوزه خدمات مالیاتی در کشورهای مختلف مورد مطالعه قرار دادند. جدول ۱، خلاصه روش‌های داده‌کاوی در این حوزه را نشان می‌دهد.

جدول (۱)-تکنیک‌های داده‌کاوی مورد استفاده توسط ادارات مالیاتی برخی کشورها کشور

نام روش	آمریکا	کانادا	استرالیا	انگلیس	بلغارستان	برزیل	پرو	شیلی
شبکه‌های عصبی	√	√		√	√		√	√
درخت تصمیم	√	√	√				√	√
رگرسیون لجستیک	√		√	√	√			
SOM			√					
K میانگین			√					√
ماشین بردار پشتیبان	√		√					

1. Feature Selection

روش‌های نمایشی	√				√			
شبکه‌های بیضی					√			
شعاع همسایگی K					√			
قواعد انجمنی							√	
قواعد فازی							√	
زنجیره مارکوف						√		
سری‌های زمانی		√						
رگرسیون			√					
شبیه‌سازی	√							

منبع: مطالعه گونزالس و ولاسکوئز (۲۰۱۳)

مطالعه جدول فوق بیانگر توسعه روش‌های داده‌کاوی در حوزه خدمات مالیاتی در کشورهای توسعه‌یافته و در حال توسعه است. بررسی پیشینه تحقیق در داخل و خارج از کشور نشان می‌دهد مطالعه محدودی پیرامون توسعه روش‌هایی برای کشف تقلّب یا فرار مالیاتی به‌خصوص تحلیل داده‌های مالیات بر ارزش افزوده در ایران انجام شده و استفاده از روش‌های دیگر داده‌کاوی مانند رده‌بندی جمعی[1] شامل رویکردهای «بگینگ»[2] و «بوستینگ»[3] به منظور افزایش دقت مدل‌سازی به دلیل عدم توازن در تعداد افراد کم‌اظهار در مقابل گروه دیگر، می‌تواند خلأ تحقیقاتی داخل و خارج از کشور باشد. در جدول ۲ جمع‌بندی بررسی‌های صورت پذیرفته در این بخش ارائه شده است. لذا این تحقیق علاوه بر به‌روز بودن، یکی از نیازها و اولویت‌های سازمان امور مالیاتی کشور در بخش مالیات بر ارزش افزوده در حوزه پردازش داده‌ها و اطلاعات می‌باشد و نتایج آن می‌تواند در عمل مورد استفاده سازمان امور مالیاتی کشور قرار گیرد.

۱. از روش‌های رده‌بندی جمعی در داده‌کاوی و رده‌بندی جهت مجموعه داده‌های نامتوازن که در آن تعداد نمونه‌هایی که نمایانگر یک کلاس هستند از نمونه‌های دیگر در کلاس‌های متفاوت، کمتر است استفاده می‌گردد.

2. Bagging

3. Boosting

جدول (۲)- جمع‌بندی بررسی تحقیقات انجام شده در حوزه تقلّبات مالی و داده‌کاوی

نویسنده	سال	هدف	انتخاب ویژگی	روش مدل‌سازی	منبع داده‌ها	مدل مورد استفاده	روش‌های رده‌بندی جمعی	الگوریتم برتر
Glancy -Yadav	۲۰۱۱	پیدا کردن گزارشات مالی متقلّبانه	–	خوشه‌بندی	گزارشات مالی ارائه شده شرکتهای مختلف	خوشه‌بندی سلسله مراتبی	–	–
صفرزاده	۱۳۸۹	ایجاد الگوها برای کشف عوامل مرتبط با تقلّب در گزارش‌گیری مالی	–	رده‌بندی	۱۷۸ شرکت	رگرسیون لجستیک	–	–
Yan-hong	۲۰۰۶	شناسایی موقعیت مالی سازمانی از نظر متقلّبانه یا نرمال بودن	–	خوشه‌بندی، استخراج قواعد وابستگی				
رده‌بندی	۱۵ شرکت مختلف	درخت‌های تصمیم	–	–				

نویسنده	سال	هدف	انتخاب ویژگی	روش مدل‌سازی	منبع داده‌ها	مدل مورد استفاده	روش‌های رده‌بندی جمعی	الگوریتم برتر
Deng	۲۰۰۹	ارائه مدلی جهت پیش‌بینی اظهارنامه‌های مالی متقلّبانه	حذف متغیرهای بی‌تأثیر با استفاده از نظر خبرگان	رده بندی	۸۸ اظهارنامه	ماشین‌های بردار پشتیبان	–	–
Hsu	۲۰۱۵	ارائه مدلی به منظور انتخاب اظهارنامه‌های با سود زیاد	انتخاب ویژگی‌های تأثیرگذار با استفاده از مدل های رده‌بندی	رده بندی	۱۰,۹۴۳ cases	C۴,۵		
بیز ساده								
شبکه‌های عصبی پرسپترون چند لایه								
ماشین‌های بردار پشتیبان	بوستینگ	–						
Kirkos	۲۰۰۷	ارائه مدلی به منظور پیش‌بینی اظهارنامه‌های متقلّبانه	آنالیز واریانس	رده‌بندی	۷۶ کارخانه و شرکت یونانی	درخت تصمیم، شبکه‌های عصبی مصنوعی		

نویسنده	سال	هدف	انتخاب ویژگی	روش مدل‌سازی	منبع داده‌ها	مدل مورد استفاده	روش‌های رده‌بندی جمعی	الگوریتم برتر
شبکه‌های اعتقادی بیز	–	شبکه‌های اعتقادی بیز						
Ravisan-kar	۲۰۱۱	پیش‌بینی تقلّبات در اظهارنامه‌های مالیاتی	انتخاب ویژگی بر اساس تست t	رده بندی	۲۰۲ شرکت چینی	شبکه‌های عصبی مصنوعی		
ماشین‌بردار پشتیبان								
برنامه‌ریزی ژنتیک								
روش جمعی مدیریت داده								
رگرسیون لجستیک								
شبکه‌های عصبی احتمالی	–	شبکه‌های عصبی مصنوعی احتمالی						
Wu	۲۰۱۲	بررسی کارایی روش‌های داده‌کاوی برای کشف تقلّبات مالیاتی	–	استخراج قواعد وابستگی	تعدادی از سازمان‌های مالیات‌دهنده در سال‌های ۲۰۰۳ و ۲۰۰۴	–	–	–

نویسنده	سال	هدف	انتخاب ویژگی	روش مدل‌سازی	منبع داده‌ها	مدل مورد استفاده	روش‌های رده‌بندی جمعی	الگوریتم برتر
Song	۲۰۱۴	بررسی کاربرد روش‌های یادگیری ماشین به منظور ارزیابی ریسک تقلّب اظهارنامه‌های مالیاتی	آنالیز واریانس	رده‌بندی	شرکت‌های چینی	رگرسیون لجستیک		
شبکه‌های عصبی، درخت تصمیم								
ماشین‌های بردار پشتیبان	تجمیعی	رده بندی جمعی حاصل از هر چهار روش						
Moepia	۲۰۱۴	مدلی به منظور کشف تقلّبات مالی	PCA و تحلیل فاکتور	رده بندی	اظهارنامه‌های مالیاتی برخی از شرکت‌های کشور آفریقای جنوبی	ماشین بردار پشتیبان		

نویسنده	سال	هدف	انتخاب ویژگی	روش مدل‌سازی	منبع داده‌ها	مدل مورد استفاده	روش‌های رده‌بندی جمعی	الگوریتم برتر
بیز ساده حساس به هزینه								
k نزدیک‌ترین همسایگی حساس به هزینه	–	ماشین‌های بردار پشتیبان						
Huang	۲۰۱۴	ارائه مدلی پشنهادی بر اساس الگوریتم بدون نظارت نگاشت‌های خودسازمانده جهت کشف گزارشات مالی متقلّبانه	–	رده‌بندی	۱۴۴ اظهارنامه مالی شرکت‌های تایوانی	نزدیک‌ترین همسایگی		
شبکه‌های عصبی								
ماشین بردار پشتیبان								

نویسنده	سال	هدف	انتخاب ویژگی	روش مدل‌سازی	منبع داده‌ها	مدل مورد استفاده	روش‌های رده‌بندی جمعی	الگوریتم برتر
ترکیب دو روش تحلیل تشخیصی و نگاشت‌های خودسازمانده	–	روش پیشنهادی						
Lesot	۲۰۱۲	ارائه مدلی جهت کشف تقلّبات مالی در کارت‌های اعتباری	–	خوشه‌بندی	کارت‌های اعتباری	فازی c میانه و روش خوشه‌بندی سلسله مراتبی	–	–
Dheepa	۲۰۰۹	ارائه یک روش پیوندی به منظور کشف تقلّبات مالی کارت‌های اعتباری	–	خوشه‌بندی	کارت‌های اعتباری	خوشه‌بندی چند سطحی		
ماشین‌های بردار پشتیبان								

نویسنده	سال	هدف	انتخاب ویژگی	روش مدل‌سازی	منبع داده‌ها	مدل مورد استفاده	روش‌های رده‌بندی جمعی	الگوریتم برتر
Collective Animal Behavior	–		–					

۳- روش‌شناسی پژوهش

متدولوژی کریسپ[1] یکی از رویکردهای متداول برای انجام پروژه‌های داده‌کاوی است (هسو و همکاران، ۲۰۱۵). این روش شامل شش مرحله به هم وابسته است. مراحل این روش به ترتیب عبارتند از: درک فضای کسب و کار، درک داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی و توسعه نتایج. با توجه به این‌که برای پاسخ به مسئله این تحقیق از ابزار داده‌کاوی استفاده شده، روش تحقیق نیز بر متدولوژی کریسپ پایه‌گذاری شده است. در ادامه ضمن معرفی مختصری از هریک از مراحل این روش شناسی، مراحل حل مسئله کشف تقلّب‌های کم‌اظهاری در اظهارنامه‌های مالیاتی نیز در قالب این متدولوژی تشریح می‌شود.

۳-۱- درک فضای کسب و کار

این مرحله شامل تشریح جنبه‌های مختلف مسئله داده‌کاوی می‌باشد. ضمن معرفی فضای کسب و کار، مسئله تحقیق، اهداف و ضرورت‌های آن تعیین می‌شوند. همانطور که در مقدمه اشاره شد، مسئله این پژوهش شناسایی مؤلفه‌های مؤثر بر کشف کم‌اظهاری و توسعه روشی جهت شناسایی مؤلفه‌های مؤثر بر کشف پرونده‌های مشکوک به تقلّب و کم‌اظهاری و همچنین مدلی برای پیش‌بینی مؤدیان متقلّب در حوزه مالیات بر ارزش افزوده می‌باشد.

نتایج این پژوهش به ممیزین پرونده‌های مالیاتی کمک خواهد کرد تا قبل از بررسی پرونده، یک ارزیابی اولیه داشته باشند تا در صورت پیش‌بینی مدل برای تقلّب، با دقت و جدیّت بیشتری پرونده را بررسی نمایند. در صورت دقت مناسب مدل پیش‌بینی، عملکرد سازمان‌های مالیاتی تا حد زیادی بهبود خواهد یافت و این امر بر استفاده بهتر منابع از جمله منابع انسانی تأثیر خواهد گذاشت. محدوده این پژوهش اداره کل امور مالیات بر ارزش افزوده شهر تهران می‌باشد که در آن تعدادی از اظهارنامه‌های مالیات بر ارزش افزوده که

1. CRISP-DM

مراحل ممیزی و تعیین تکلیف و نتیجه را طی کرده‌اند، به‌طور نمونه و با روش نمونه‌گیری تصادفی ساده با توزیع احتمال یکنواخت جهت مطالعه و توسعه مدل انتخاب شده‌اند.

۳-۲- درک داده‌ها

رویکردهای داده‌کاوی بر اساس داده‌های جمع‌آوری شده پیرامون مسئله، اقدام به مدل‌سازی و ارائه مدل می‌کنند. لذا بانک اطلاعاتی و مجموعه داده‌ها، مواد خام یک پروژه داده‌کاوی می‌باشند. فعالیت‌های این مرحله شامل جمع‌آوری داده‌های خام، درک داده‌ها، کیفیت‌سنجی و تشریح داده‌ها می‌باشد. داده‌های این پژوهش با توجه به پرونده‌های مالیات بر ارزش افزوده و عملکرد مؤدیان و اظهارنامه مالیات بر ارزش افزوده استخراج شده است. با توجه به محدودیت‌های موجود (تکمیل نبودن پایگاه اطلاعات مؤدیان جهت دسترسی کامل به اطلاعات هویتی، عملکردی، دارایی مؤدیان شامل مواردی نظیر اطلاعات مالی، پولی و اعتباری، معاملاتی، سرمایه‌ای و ملکی اشخاص حقیقی و حقوقی)، امکان جمع‌آوری تمامی مشخصه‌های مؤدیان وجود نداشته و داده‌هایی جمع‌آوری شده‌اند که طبق نظر کارشناسان حوزه مالیات، تأثیر بیشتری بر کشف کم‌اظهاری در اظهارنامه‌ها دارد. در پایگاه داده تشکیل شده، ۴۹۵ مشاهده (مربوط به اظهارنامه بررسی و به نتیجه رسیده) مشتمل بر ۱۸ متغیر مربوط به هر اظهارنامه گردآوری شده است. جدول ۳، ضمن تعریف متغیرهای مستقل، متغیر وابسته تحقیق که تقلّب (کم‌اظهاری) یا عدم تقلّب در اظهارنامه می‌باشد را به‌طور مختصر تعریف می‌کند.

جدول (۳)- تشریح متغیرهای تحقیق

نام	تعریف متغیر	نوع	مقادیر متغیر	نقش
تقلّب	وضعیت تقلّب	اسمی	۲=تقلّب با بیش از ۵۰ درصد انحراف(۱۶٫۸٪)، ۱=تقلّب بین ۱۵ تا ۵۰ درصد انحراف(۲۵٫۳٪)، ۰= عدم تقلّب (۵۸٫۰٪)	وابسته (خروجی)
نوع شخصیت	نوع کسب و کار مشمول قانون مالیات	اسمی	حقیقی (۵۱٫۱٪)، حقوقی (۴۸٫۹٪)	پیش‌بین (ورودی)
وضعیت محل استقرار	محل قرار گرفتن شخص مؤدی	اسمی	داخل حریم شهر (۹۷٫۴٪)، خارج از حریم شهر (۲٫۶٪)	پیش‌بین (ورودی)
فعالیت های انتخاب شده	نوع فعالیت مؤدی	اسمی	تولیدی (۱۰٫۱٪)، خدماتی-توزیعی (۲٫۰٪)، خدماتی (۶۳٫۸٪)، توزیعی (۲۱٫۰٪)، توزیعی-تولیدی (۰٫۶٪)، خدماتی-توزیعی-تولیدی (۱٫۲٪)، خدماتی-تولیدی (۱٫۲٪)	پیش‌بین (ورودی)
حوزه	حوزه مالیاتی مؤدی	اسمی	جنوب تهران (۶٫۵٪)، شرق تهران (۱۴٫۹٪)، شمال تهران (۳۲٫۳٪)، غرب تهران (۲۴٫۲٪)، مرکزتهران (۲۲٫۰٪)	پیش‌بین (ورودی)

نام	تعریف متغیر	نوع	مقادیر متغیر	نقش
مرحله مشمول	دوره‌ی مشمول پرداخت مالیات	اسمی	مشمول مرحله‌دوم (۱۷٫۸٪)، مشمول مرحله‌چهارم (۲۲٫۲٪)، مشمول مرحله‌پنجم (۰٫۸٪)، مشمول مرحله‌اول (۵۹٫۲٪)	پیش‌بین (ورودی)
سال	سال اظهارنامه	اسمی	۱۳۹۰ (۴۸٫۹٪)، ۱۳۹۱(۵۱٫۱٪)	پیش‌بین (ورودی)
موجودی نقد	مقادیر سرمایه نقدی مؤدی	عدد صحیح	[۴؛ ۴۵۱۰۲۸۵]	پیش‌بین (ورودی)
حسابهای دریافتنی	مقادیر مطالبات بنگاه اقتصادی با سر رسیدهای کوتاه مدت	عدد صحیح	[۰؛ ۱۳۴۷۸۹۹۵]	پیش‌بین (ورودی)
دارائی جاری	دارایی جاری مؤدی	عدد صحیح	[۷۹۲؛ ۷۹۸۳۳۵۵۸]	پیش‌بین (ورودی)
دارایی ثابت	مقادیر مالی دارایی‌های ثابت مؤدی	عدد صحیح	[۰؛ ۱۴۸۹۸۲۶۸۱]	پیش‌بین (ورودی)
بدهی جاری	دیون مؤدی با سررسید کمتر از یکسال	عدد صحیح	[۳؛ ۱۲۴۴۸۷۹۹۶]	پیش‌بین (ورودی)
بدهی بلند مدت	دیون مؤدی با سررسید بیش از یکسال	عدد صحیح	[۰؛ ۴۲۹۷۰۷۲۰]	پیش‌بین (ورودی)

نام	تعریف متغیر	نوع	مقادیر متغیر	نقش
سرمایه	حق مالی صاحبان سرمایه نسبت به خالص دارایی‌ها	عدد صحیح	[۱؛ ۴۰۰۰۰۰۰۰]	پیش‌بین (ورودی)
سود و زیان انباشته	سود خالص تقسیم نشده	عدد صحیح	[-۱۲۳۴۰۹۷؛ ۱۳۳۳۹۸۸۸]	پیش‌بین (ورودی)
بدهی حقوق صاحبان سهام	بدهی صاحبان سهام	عدد صحیح	[-۹۸۶۸۵؛ ۵۳۷۰۴۹۴۹]	پیش‌بین (ورودی)
سابقه فعالیت شرکت	تعداد سال‌های فعالیت مؤدی	اسمی	فعالیت بیش از ۱۰ سال (۷۲٫۷٪)، فعالیت ۵ تا ۱۰ سال (۱۷٫۶٪)، فعالیت ۳ تا ۵ سال (۶٫۳٪)، فعالیت حداکثر یک سال (۰٫۴٪)، فعالیت ۱ تا ۳ سال (۳٫۰٪)	پیش‌بین (ورودی)
مالیات ابرازی	مالیات خوداظهاری مؤدی	پیوسته	–	استفاده در تعیین متغیر خروجی
مالیات معین شده	مقدار مالیات تعیین شده توسط سر ممیز	پیوسته	–	استفاده در تعیین متغیر خروجی

منبع: یافته‌های تحقیق

متغیر «مالیات معین شده» برای تعیین وضعیت پرونده از نظر تقلّب در نظر گرفته شده است. بنابر نظر کارشناسان ارشد و خبرگان حوزه مالیاتی کشور از طریق مصاحبه آزاد، در صورتی‌که اختلاف بین «مالیات ابرازی» در اظهارنامه و «مالیات تشخیصی» در رسیدگی، زیاد (بیش از ۵۰ درصد) باشد، پرونده مورد نظر در وضعیت تقلّب در نظر گرفته شده و به ازای متغیر تقلّب (متغیر پاسخ تحقیق) عدد ۲ به معنای تقلّب بیش از ۵۰ درصد اختلاف ثبت شده است. در ازای اختلاف بین ۱۵ تا ۵۰ درصدی مقدار متغیر تقلّب به ازای

مؤدی عدد ۱ به معنای مشکوک به تقلّب ثبت شده است. عدد صفر به ازای متغیر تقلّب (سالم بودن پرونده)، معرف انحراف دو متغیر مذکور کمتر از ۱۵ درصد می‌باشد. سایر متغیرهای جدول ۳، متغیرها و مؤلفه‌های مؤثر بر کشف تقلّب یا همان متغیرهای مستقل تحقیق می‌باشند که باید میزان تأثیر آن‌ها بر شناسایی پرونده‌های تقلّب (متغیر تقلّب) بررسی شوند و مدلی مبتنی بر روابط بین دو گروه متغیر مستقل و وابسته جهت پیش‌بینی موارد تقلّب توسعه داده شود. این مؤلفه‌ها بر اساس نظرات در مصاحبه آزاد از خبرگان نظام مالیات بر ارزش افزوده شهر تهران شناسایی و در این تحقیق لحاظ شده‌اند. تعداد خبرگان مالیاتی مصاحبه شونده برابر با ۱۰ نفر مشتمل بر ممیز با شرایط مدرک تحصیلی فوق لیسانس و با سابقه حداقل ۱۰ سال، سرممیز با تحصیلات لیسانس و سابقه مدیریتی حداقل ۵ سال و ممیزکل با حداقل مدرک لیسانس و سابقه مدیریتی حداقل ۳ سال بوده است.

اعداد داخل پرانتز برای حالات مختلف متغیرهای اسمی در ستون «مقادیر متغیرها» در جدول ۳ بیانگر درصد مشاهدات درون هر گروه از هر متغیر می‌باشد. به‌طور مثال ۳۶۰ شرکت از ۴۹۵ مشاهده مورد بررسی (۷۳ درصد) سابقه کاری و فعالیت بیش از ۱۰ سال داشته‌اند. همچنین اکثر مؤدیان در مرحله اول اجرا مشمول گردیده‌اند. لذا سابقه و تجربه مناسبی در تهیه اظهارنامه و پرداخت مالیات دارند. با بررسی متغیر تقلّب نیز مشخص می‌گردد که از میان ۴۹۵ مؤدی مورد بررسی، تعداد ۲۸۷ رکورد دارای رده صفر یعنی سالم، ۱۲۵ مورد دارای رده یک یعنی تقلّب با اختلاف ۱۵٪ تا ۵۰٪ می‌باشند و تعداد ۸۳ سازمان با تفاوت بیشتر از ۵۰٪ اقدام به تقلّب در پرداخت مالیات‌های خود کرده‌اند.

۳-۳- آماده‌سازی داده‌ها

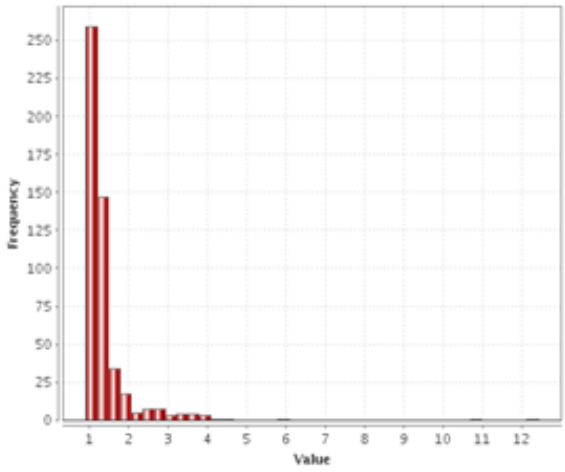
پیش از استخراج دانش، داده‌ها باید برای اجرای تکنیک‌های داده‌کاوی آماده‌سازی شوند. مرحله آماده‌سازی داده‌ها عبارت است از: انتخاب داده‌ها، پاک‌سازی داده‌ها، یکپارچه‌سازی و در نهایت تبدیل داده‌ها. این مرحله یکی از حساس‌ترین مراحل فرآیند کریسپ می‌باشد و به گفته برخی از محققان حتی در شرایطی ممکن است این مرحله، بیش از ۶۰ درصد زمان اجرای پروژه‌های داده‌کاوی را به خود اختصاص دهد (پیله، ۱۹۹۹). در ادامه فعالیت‌های این مرحله تشریح شده است.

ممکن است تعدادی از داده‌های جمع‌آوری شده و مورد استفاده برای داده‌کاوی، داده‌های دور افتاده و یا اشتباه باشند و بر نتایج داده‌کاوی تأثیر داشته باشند. شناسایی و حذف آن‌ها از جمله فعالیت‌های مرحله آماده‌سازی می‌باشد. در این تحقیق داده‌های دورافتاده در مجموعه داده‌ها توسط روش k نزدیک‌ترین همسایگی شناسایی و حذف شده‌اند. به منظور کشف نقاط دور افتاده از شاخص LOF[1] در نرم‌افزار رپیدماینر

1. Local Outlier Factor

نسخه ۵/۱ استفاده گردید. در سال ۲۰۰۰ در تشخیص ناهنجاری، روش LOF به عنوان یک الگوریتم برای پیداکردن نقاط داده غیرعادی با اندازه‌گیری انحراف محلی از یک نقطه داده با توجه به همسایگان خود پیشنهاد شد (برنیگ و همکاران، ۲۰۰۰). این روش بر اساس مفهوم چگالی محلی میباشد به‌طوری که محلی‌بودن با استفاده از k نزدیک‌ترین همسایگی تعیین می‌گردد و از فاصله‌های اقلیدسی به منظور محاسبه چگالی استفاده می‌شود.

شکل (۱)- هیستوگرام شاخص LOF برای شناسایی مشاهدات دورافتاده



منبع: یافته‌های تحقیق

شکل۱، نمودار هیستوگرام مقدار شاخص LOF را برای تمام مشاهدات تحقیق نشان می‌دهد. محور افقی، مقدار شاخص شناسایی داده‌های پرت (LOF) و محور عمودی فراوانی آن را در داده‌ها نشان می‌دهد. همانطور که مشخص است، بیشتر داده‌های تحقیق مقدار LOF کمتر از دو دارند. به‌عنوان یک قانون، مشاهداتی که شاخص LOF بیشتری از ۲/۵ دارند، به‌عنوان داده‌های پرت و دورافتاده در نظر گرفته می‌شوند. به این ترتیب از ۴۹۵ مشاهده موجود تعداد ۲۹ مشاهده به عنوان داده دورافتاده شناسایی و حذف شده‌اند و تعداد رکوردها به ۴۶۶ رکورد کاهش یافته است. بررسی مشاهدات حذف شده نشان می‌دهد، مواردی از مجموعه داده‌های اولیه حذف شده‌اند که حداقل در یکی از متغیرها مقداری بیش از حد متعارف داشته‌اند. در این تحقیق شناخت و حذف داده‌های پرت به منظور پاک‌سازی و با هدف از بین بردن عدم قطعیت در داده‌کاوی مورد استفاده قرار گرفته است.

همانطور که اشاره شد، یکی از اهداف این پژوهش شناسایی متغیرهای (مؤلفه‌ها) مؤثر بر کشف تقلّب می‌باشد. در این بخش با استفاده از روش‌های مختلف آماری، تأثیر متغیرهای بالقوه مؤثر بر متغیر تقلّب

(۱۶ متغیر جدول۳) اندازه‌گیری شده و بر اساس آن مؤثرترین مؤلفه‌ها جهت ورودی به مدل‌های پیش‌بینی وضعیت پرونده‌های مالیاتی تعیین شده است.

در این تحقیق به منظور شناسایی متغیرهای مؤثر بر پیش‌بینی تقلّب و کاهش متغیرهای تحقیق از دو دسته روش انتخاب ویژگی استفاده شده است؛ روشهای مختلف فیلتری شامل روش نسبت سود اطلاعاتی[1]، روش Relief (فیلتری بر پایه نزدیک‌ترین همسایگی) و کای دو[2] و روش‌های پوشاننده مانند الگوریتم ژنتیک[3] هستند.

با توجه به این‌که سه روش فیلتری استفاده شده ممکن است نتایج متفاوتی ارائه کنند، به منظور ارائه یک روش مناسب برای انتخاب ویژگیهای تأثیرگذار، تصمیم گرفته شده که امتیازهای حاصله در سه روش فیلتری نسبت سود اطلاعاتی، روش Relief و روش کای دو با یکدیگر تجمیع شوند و سپس مؤلفه‌های مؤثر بر اساس جمع امتیازات رتبه‌بندی شوند. این روش‌ها بر داده‌های تحقیق اجرا شده و نتایج آن‌ها در جدول (۴) ارائه شده است. میزان تأثیر هرکدام از این متغیرها بر متغیر وابسته در روش‌های فیلتری، مقداری بین صفر و یک اندازه‌گیری شده است. سه متغیر «مرحله مشمول»، «سابقه شرکت» و «نوع شخصیت» به عنوان تأثیرگذارترین متغیرها بر اساس روش Relief شناخته شده‌اند. از طرفی دیگر متغیرهای «سود و زیان انباشته»، «موجودی نقد» و «حوزه»، کمترین تأثیر را بر تقلّب مالیاتی سازمان‌ها داشته‌اند.

تحلیل اوزان روش سود اطلاعاتی در جدول ۴ بیانگر آن است که سه متغیر «دارایی جاری»، «بدهی جاری» و «مرحله مشمول» بیشترین تأثیر را بر متغیر پیش‌بینی دارد. در روش کای دو نیز «مرحله مشمول» به عنوان تأثیرگذارترین متغیر انتخاب شده است. همچنین دو متغیر «بدهی حقوق صاحبان سهام» و «سرمایه» به عنوان دو متغیر تأثیرگذار بعدی شناسایی شده‌اند. از طرف دیگر متغیر «وضعیت محل استقرار» هیچ تأثیری بر شناسایی افراد متقلّب از افراد غیرمتقلّب نداشته است.

جدول (۴)- اوزان تأثیر متغیرهای تحقیق بر متغیر وابسته در روش‌های مختلف

متغیر	کای‌دو	Relief	سود اطلاعاتی	وزن نهایی
مرحله مشمول	۱,۰۰۰	۱,۰۰۰	۰,۸۵۱	۲,۸۵۱
دارائی جاری	۰,۴۱۶	۰,۲۰۴	۱,۰۰۰	۱,۶۲۰

1. Information Gain
2. Chi-square
3. Genetic Algorithm

بدهی جاری	۰,۴۶۴	۰,۱۹۷	۰,۹۲۵	۱,۵۸۶
بدهی حقوق صاحبان سهام	۰,۶۸۰	۰,۱۵۱	۰,۷۲۸	۱,۵۵۹
سرمایه	۰,۴۸۸	۰,۰۸۸	۰,۶۱۰	۱,۱۸۶
دارایی ثابت	۰,۲۹۷	۰,۱۱۷	۰,۶۵۹	۱,۰۷۳
حسابهای دریافتنی	۰,۳۰۰	۰,۰۹۳	۰,۶۱۶	۱,۰۰۹
سابقه فعالیت شرکت	۰,۲۲۹	۰,۵۳۶	۰,۱۴۸	۰,۹۱۲
سود و زیان انباشته	۰,۲۴۹	۰,۰۰۹	۰,۵۲۶	۰,۷۸۴
موجودی نقد	۰,۱۷۷	۰,۰۰۴	۰,۴۲۰	۰,۶۰۱
نوع شخصیت	۰,۱۸۳	۰,۲۶۲	۰,۱۲۶	۰,۵۷۱
سال	۰,۱۵۴	۰,۲۲۶	۰,۱۰۶	۰,۴۸۶
حوزه	۰,۱۹۱	۰,۰۰۰	۰,۱۵۲	۰,۳۴۴
بدهی بلند مدت	۰,۱۳۵	۰,۰۳۷	۰,۱۶۴	۰,۳۳۵
فعالیت های انتخاب شده	۰,۰۵۸	۰,۰۸۳	۰,۰۴۸	۰,۱۸۹
وضعیت محل استقرار	۰,۰۰۰	۰,۰۴۶	۰,۰۰۰	۰,۰۴۶

منبع: یافته‌های تحقیق

هرچند برخی از متغیرها مانند «مرحله مشمول» در سه روش بالاترین رتبهٔ تأثیر را دارد، اما اختلافاتی در خروجی‌های این سه روش وجود دارد. برای تجمیع نتایج سه روش مذکور، وزن‌ها با یکدیگر جمع‌شده و در ستون آخر جدول ۴ ثبت شده‌اند. با مرتب کردن اوزان تجمیعی، ۱۰ متغیری که بیشترین وزن و بالای ۶۰٪ را داشتند، به‌عنوان متغیرهای برتر انتخاب شده‌اند. البته میزان و اولویت هریک از متغیرها بر متغیر پاسخ در روش‌های مختلف پیش‌بینی متفاوت است و اولویت آن‌ها در روش پیش‌بینی به‌طور دقیق‌تر تعیین می‌شود.

علاوه بر روش‌های فیلتری از رویکرد الگوریتم ژنتیک[1] نیز به‌منظور پیدا کردن بهترین مجموعه از متغیرهای مستقل استفاده شده است (جارمولاک و کرو، ۱۹۹۹). در انتخاب زیر مجموعه‌ای از متغیرها توسط الگوریتم ژنتیک، جهش[2] به‌معنی انتخاب یا عدم انتخاب متغیرها در هر بار اجرا می‌باشد. همچنین توارث به معنی تعویض متغیرهای استفاده شده می‌باشد. نحوه کار در این الگوریتم به این ترتیب است که ابتدا تعداد P متغیر انتخاب می‌شوند (تعداد جامعه). سپس برای هر متغیر انتخاب شده جهش اعمال می‌گردد به این ترتیب که متغیرهای استفاده نشده از مجموعه متغیرها خارج می‌شوند و مقدار احتمال آن به Pm (احتمال جهش) تغییر پیدا می‌کند. سپس دو متغیر انتخاب می‌شوند و بین آن‌ها توارث با احتمال Pc صورت می‌پذیرد که نوع این توارث قابل انتخاب است. در این گونه انتخاب ویژگی بر اساس الگوریتم ژنتیک، تابع سازگاری که نمایان‌گر بهینگی می‌باشد، عبارت از یک مدل رده‌بندی کننده است. این به آن معنی است که وقتی مجموعه متغیرهای اولیه بر اساس احتمالات تنظیم شده انتخاب شدند، با استفاده از یک روش رده‌بندی بر روی این متغیرها، مدل‌سازی انجام می‌شود و دقت مدل ارزیابی می‌گردد. سپس انتخاب، جهش و توارث در متغیرهای مجموعه داده تا رسیدن به یک بهینگی در مدل ساخته شده ادامه پیدا می‌کند. در تحقیق حاضر برای بررسی صحت انتخاب متغیرها از روش بیز ساده[3] بهینه شده به منظور رده‌بندی استفاده شده است. به این ترتیب تعداد ۷ متغیری که در جدول ۵ ذکر شده‌اند به عنوان متغیرهای تأثیرگذار در این روش انتخاب شده‌اند.

جدول (۵)- متغیرهای انتخاب شده توسط الگوریتم ژنتیک

نام متغیر	
نوع شخصیت	بدهی بلندمدت
حوزه	سود و زیان انباشته
مرحله مشمول	بدهی حقوق صاحبان سهام
دارایی جاری	

از آنجا که مقیاس داده‌ها ممکن است بر عملکرد روش‌های پیش‌بینی تأثیرگذار باشد، داده‌های هر متغیر قبل از انجام فرآیند مدل‌سازی استاندارد شده‌اند. کسر هر مشاهده از میانگین مربوطه و تقسیم آن بر انحراف

1. Genetic Algorithm
2. Mutation
3. Naive Bayes

معیار متغیر روش بی‌مقیاس‌سازی داده‌ها بوده است.

۳-۴- مدل‌سازی

مرحله مدل‌سازی به عنوان اصلی‌ترین مرحله در پروژه‌های داده‌کاوی محسوب می‌شود که در آن پاسخ مسئله اصلی داده می‌شود. مسئله اصلی این پژوهش توسعه مدلی مناسب برای پیش‌بینی وضعیت تقلّب (کم‌اظهاری) اظهارنامه‌های مالیات بر ارزش افزوده بر اساس متغیرهای مؤثر بر آن (ده متغیر مؤثرتر در جدول ۴ و هفت متغیر منتخب در جدول ۵) است. به‌منظور افزایش دقت پیش‌بینی در این تحقیق جمعاً ۱۲ روش مختلف اجرا و خطای پیش‌بینی آن‌ها محاسبه شده است. دو روش درخت تصمیم و k نزدیک‌ترین همسایگی هرکدام در سه حالت (عادی، جمعی بگینگ و جمعی بوستینگ) بر دو گروه متغیرهای انتخاب شده با روش‌های فیلتری و الگوریتم ژنتیک اجرا شده است. ابتدا هر یک از دو روش درخت تصمیم و k نزدیکترین همسایگی با داده‌های اولیه برای دو دسته متغیر منتخب از روش‌های فیلتری (جدول ۴) و الگوریتم ژنتیک (جدول ۵) اجرا شده است و سپس هریک از این دو روش با دو روش جمعی بگینگ و بوستینگ به تفکیک دو مجموعه متغیر منتخب اجرا شده است.

روش‌های جمعی بگینگ و بوستینگ زمانی به‌کار گرفته می‌شوند که داده‌ها نامتوازن باشند. در هریک از ۱۲ مدل اجرا شده، بهترین مقدار پارامتر مدل‌سازی، به کار گرفته شده است که تنظیمات مربوطه در جدول ۶ ارائه می‌شود.

جدول (۶)- تنظیمات پارامترهای هر مدل

نام مدل	پارامترهای تنظیم شده	مخفف نام
درخت تصمیم	• شاخص انشعاب‌دهی (gain ratio، accuracy، gini index، information gain) • عمق درخت: (۵،۱۰، ۱۵، ۲۰)	DT
بگینگ درخت تصمیم	• شاخص انشعاب‌دهی (gain ratio، accuracy، gini index، information gain) • عمق درخت: (۵،۱۰، ۱۵، ۲۰) • نسبت انتخاب نمونه: (۰,۸، ۰,۹، ۱)	DT-bag

بوستینگ درخت تصمیم	• شاخص انشعاب‌دهی (gain ratio، gini index، information gain، accuracy) • عمق درخت: (۱۰، ۵، ۱۵، ۲۰) • تعداد تکرار (۸، ۹، ۱۰)	DT-boost
K نزدیک‌ترین همسایگی	• K (۵، ۱۰، ۱۵، ۲۰)	KNN
بگینگ K نزدیک‌ترین همسایگی	• K (۵، ۱۰، ۱۵، ۲۰) • نسبت انتخاب نمونه: (۸، ۰، ۹، ۰، ۱)	KNN-bag
بوستینگ K نزدیک‌ترین همسایگی	• K (۵، ۱۰، ۱۵، ۲۰) • تعداد تکرار (۸، ۹، ۱۰)	KNN-boost

منبع: یافته‌های تحقیق

روش درخت تصمیم به‌عنوان یکی از روش‌های پیش‌بینی و دسته‌بندی، بر اساس متغیرهای مستقل و متغیر وابسته اقدام به ساخت درخت سلسله مراتبی می‌کند. هر گره در درخت تصمیم، متناظر با یک متغیر مستقل و هر کمان واسط بین والد به فرزند، نمایانگر یک مقدار ممکن برای آن متغیر است. متغیرها به ترتیب میزان تأثیر آن‌ها در لایه‌های مختلف به سطوح مختلف (کمان‌ها) تقسیم می‌شوند تا درختی با عمق حداکثر برابر تعداد متغیرها ایجاد شود. با استفاده از درخت تصمیم می‌توان مجموعه‌ای قانون[1] (اگر–آنگاه) برای پیش‌بینی متغیر وابسته ایجاد کرد. به‌عنوان مثال اگر متغیرهای مؤثر در دامنه سطوح یک شاخه درخت باشد، احتمال قرار گرفتن پرونده در یکی از وضعیت‌های تقلّب معین می‌شود. روش‌ها و معیارها در روش درخت تصمیم برای انشعاب‌دهی در هر گره توسعه داده شده است که در این تحقیق مطابق جدول ۶ از شاخص‌های «گین»[2]، «اطلاعاتی گین»[3] و «جینی»[4] استفاده شده است.

در روش k نزدیک‌ترین همسایه، یک گروه شامل k نمونه از مجموعه نمونه‌های آموزشی که نزدیک‌ترین نمونه‌ها به نمونه آزمایشی باشند را انتخاب کرده و بر اساس برتری رده یا برچسب مربوط به آن‌ها در مورد

1. Rule
2. Gain Ratio
3. Information Gain
4. Gini Index

دسته نمونه آزمایشی مزبور تصمیم‌گیری می‌نماید. به عبارت ساده‌تر این روش رده‌ای را انتخاب می‌کند که در همسایگی انتخاب شده، بیشترین تعداد نمونه منتسب به آن دسته باشند. بنابراین رده‌ای که از همه رده‌ها بیشتر در بین k نزدیک‌ترین همسایه مشاهده شود، به عنوان رده نمونه جدید در نظر گرفته می‌شود.

به‌منظور ارزیابی کارایی مدل‌های مختلف پیش‌بینی از شاخص‌های ارزیابی دقت رده‌بندی[1] استفاده می‌شود و فرایند اجرا شامل دو مرحله آموزش و ارزیابی دقت مدل است. ابتدا مجموعه داده‌های جمع‌آوری شده به صورت تصادفی به دو مجموعه داده‌های آموزش و آزمایش با نسبت‌های ۷۰ و ۳۰ درصد تقسیم شده است. با استفاده از داده‌های آموزش، پارامترهای درونی مدل تنظیم می‌شود و از داده‌های آزمایش برای اندازه‌گیری دقت مدل‌های پیش‌بینی استفاده می‌شود تا نزدیکی مقدار پیش‌بینی متغیر خروجی توسط مدل با واقعیت (نتیجه‌ای که از قبل معلوم بوده و به مدل نشان داده نشده است) اندازه‌گیری شود. جهت مقایسه مناسب‌تر مدل‌های پیش‌بینی، داده‌های آموزش و آزمایش برای مدل‌های مختلف یکسان بوده و نمونه‌ها برای هر مدل پیش‌بینی تغییر نکرده‌اند.

جدول (۷)- پارامترها و نتایج مدل‌سازی با استفاده از متغیرهای انتخابی توسط سه روش فیلتری

نام روش	بهترین پارامترها	دقت آزمون
درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۵	٪۸۲,۱۴
بگینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۵ و درصد نمونه = ۰,۹ و ۱	٪۸۲,۱۴
بوستینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۵ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۸۲,۱۴
KNN	K = ۲۰	٪۷۱,۴۳
KNN-bag	K۲۰ = و درصد نمونه = ۱	٪۷۱,۴۳
KNN-boost	K۲۰ = و تعداد تکرار = ۸ یا ۹ یا ۱۰	٪۷۱,۴۳
سایر پارامترها		
درخت تصمیم	شاخص انشعاب دهی =ratio gain با عمق ۵	٪ ۰۰,۸۰

1. Accuracy

درخت تصمیم	شاخص انشعاب دهی =gini index با عمق ۵	٪ ۸۰,۰۰
درخت تصمیم	شاخص انشعاب دهی =accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰	٪ ۷۴,۲۹
درخت تصمیم	شاخص انشعاب دهی =ratio gain با عمق ۱۰ و ۱۵ و ۲۰	٪ ۷۲,۱۴
درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۱۰ و ۱۵ و ۲۰	٪ ۷۲,۱۴
درخت تصمیم	شاخص انشعاب دهی =gini index با عمق ۱۰ و ۱۵ و ۲۰	٪ ۷۳,۵۷
بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۵ و درصد نمونه = ۰,۸	٪۷۷,۱۴
بگینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۵ و درصد نمونه = ۰,۸	٪۷۶,۴۳
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۵ و درصد نمونه = ۰,۸	٪۷۳,۵۷
بگینگ درخت تصمیم	شاخص انشعاب دهی = accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۸	٪۷۰,۷۱
بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۸	٪۷۲,۸۶
بگینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق۱۰و۱۵و۲۰ و درصد نمونه = ۰,۸	٪۶۸,۵۷
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۸	٪۶۸,۵۷

بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۵ و درصد نمونه = ۰,۹ و ۱	٪۸۰,۰۰
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۵ و درصد نمونه = ۰,۹	٪۸۰,۰۰
بگینگ درخت تصمیم	شاخص انشعاب دهی = accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۹	٪۵۷/۷۳
بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۹	٪۷۱/۷۵
بگینگ درخت تصمیم	شاخص انشعاب دهی= gain information با عمق۱۰و ۱۵و۲۰و درصد نمونه=۰,۹و۱	٪۷۳,۵۷
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۱۰ و ۱۵ و ۲۰ درصد نمونه = ۰,۹	٪۷۱,۷۵
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۵ و درصد نمونه = ۱	٪۸۰,۷۱
بگینگ درخت تصمیم	شاخص انشعاب دهی = accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۱	٪۷۴,۲۹
بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۱	٪۷۳,۵۷
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۱۰ و ۱۵ و ۲۰ درصد نمونه = ۱	٪۷۵,۰۰
بوستینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق۵ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۸۰,۰۰
بوستینگ درخت تصمیم	شاخص انشعاب دهی=gain information با عمق۱۰و۱۵و۲۰و تعداد تکرار=۸یا۹یا۱۰	٪۷۶,۴۳

بوستینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۵ و تعداد تکرار= ۸یا ۹یا ۱۰	٪۸۰,۰۰
بوستینگ درخت تصمیم	شاخص انشعاب دهی = accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۷۴,۲۹
بوستینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق۱۰ و ۱۵ و ۲۰ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۷۵,۷۱
بوستینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۱۰ و ۱۵ و ۲۰ و تعداد تکرار= ۸یا ۹یا ۱۰	٪۷۷,۱۴
KNN	۵ و ۱۵ = K	٪۶۸,۵۷
KNN	۱۰ = K	٪۶۷,۱۴
KNN-bag	۵ و ۱۵ K = و درصد نمونه =۰,۸ و ۱	٪۶۸,۵۷
KNN-bag	۵ و ۱۰ K = و درصد نمونه =۰,۹	٪۶۷,۸۶
KNN-bag	۱۰ K = و درصد نمونه =۰,۸	٪۶۷,۸۶
KNN-bag	۱۰ K = و درصد نمونه =۱	٪۶۷,۱۴
KNN-bag	۱۵ K = و درصد نمونه =۰,۹	٪۶۷,۲۹
KNN-bag	۲۰ K = و درصد نمونه =۰,۹	٪۷۰,۷۱
KNN-boost	۵ و ۱۵K = و تعداد تکرار = ۸ یا ۹ یا ۱۰	٪۶۸,۵۷
KNN-boost	۱۰K = و تعداد تکرار = ۸ یا ۹ یا ۱۰	٪۶۷,۱۴

منبع: یافته‌های تحقیق

جدول (۸)- پارامترها و نتایج مدل‌سازی با استفاده از متغیرهای انتخابی توسط روش الگوریتم ژنتیک

نام روش	بهترین پارامترها	دقت آزمون
درخت تصمیم	شاخص انشعاب دهی= ratio gain با عمق ۱۰و ۱۵و ۲۰	٪۴۳,۸۱

نام روش	بهترین پارامترها	دقت آزمون
بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۱	٪۷۱,۸۰
بوستینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۵ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۸۰
KNN	= ۱۵K	٪۸۶,۶۷
KNN-bag	K ۱۵= با درصد نمونه = ۰,۹	٪۲۹,۶۹
KNN-boost	K ۱۵= با تعداد تکرار=۸ یا ۹ یا ۱۰	٪۸۶,۶۷
سایر پارامترها		
درخت تصمیم	شاخص انشعاب دهی= ratio gain با عمق ۵	٪۰۰,۸۰
درخت تصمیم	شاخص انشعاب دهی= gain information با عمق ۵	٪۰۰,۷۵
درخت تصمیم	شاخص انشعاب دهی= gini index با عمق ۵ و ۱۰ و ۱۵ و ۲۰	٪۴۳,۷۶
درخت تصمیم	شاخص انشعاب دهی= accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰	٪۱۴,۶۷
درخت تصمیم	شاخص انشعاب دهی= gain information با عمق ۱۰ و ۱۵ و ۲۰	٪۴۳,۷۶
بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۵ و ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۸	٪۵۷,۷۸
بگینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۵ درصد نمونه = ۰,۸	٪۱۴,۷۲
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۵ درصد نمونه = ۰,۸	٪۷۱,۷۵

نام روش	بهترین پارامترها	دقت آزمون
بگینگ درخت تصمیم	شاخص انشعاب دهی = accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۸ و ۰,۹	٪۴۳,۶۶
بگینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۸	٪۰۰,۷۵
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۱۰ و ۱۵ و ۲۰ و درصد نمونه = ۰,۸	٪۱۴,۷۷
بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۵ و درصد نمونه = ۰,۹	٪۵۷,۷۸
بگینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۵ درصد نمونه = ۰,۹	٪۰۰,۷۵
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۵ و۱۰ و ۱۵ و ۲۰ درصد نمونه = ۰,۹	٪۱۴,۷۷
بگینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain با عمق ۵ درصد نمونه = ۱	٪۰۰,۸۰
بگینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۵ درصد نمونه = ۱	٪۲۹,۷۹
بگینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق 5 و ۱۰ و ۱۵ و ۲۰ درصد نمونه = ۱	٪۴۳,۷۶
بگینگ درخت تصمیم	شاخص انشعاب دهی = accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰ درصد نمونه = ۱	٪۱۴,۶۷
بگینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۱۰ و ۱۵ و ۲۰ درصد نمونه = ۱	٪۴۳,۷۶

نام روش	بهترین پارامترها	دقت آزمون
بوستینگ درخت تصمیم	شاخص انشعاب دهی = gain information با عمق ۵ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۰۰,۷۵
بوستینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۵ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۷۱,۷۵
بوستینگ درخت تصمیم	شاخص انشعاب دهی = accuracy با عمق ۵ و ۱۰ و ۱۵ و ۲۰ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۱۴,۶۷
بوستینگ درخت تصمیم	شاخص انشعاب دهی = ratio gain یا gain informa-tion با عمق ۱۰ و ۱۵ و ۲۰ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۵۷,۷۸
بوستینگ درخت تصمیم	شاخص انشعاب دهی = gini index با عمق ۱۰ و ۱۵ و ۲۰ و تعداد تکرار=۸ یا ۹ یا ۱۰	٪۴۳,۷۶
KNN	K ۵ =	٪۱۴,۶۵
KNN	K ۱۰ =	٪۱۴,۶۷
KNN	K ۲۰ =	٪٪۷۱,۶۵
KNN-bag	K ۵ = با درصد نمونه = ۰,۸	٪۵۷,۶۳
KNN-bag	K ۵ = با درصد نمونه = ۰,۹	٪۲۹,۶۴
KNN-bag	K ۵ = با درصد نمونه = ۱	٪۷۱,۶۵
KNN-bag	K ۱۰ = با درصد نمونه = ۰,۸	٪۷۱,۶۵
KNN-bag	K ۱۰ = با درصد نمونه = ۰,۹ و ۱	٪۴۳,۶۵
KNN-bag	K ۱۵ و ۲۰ = با درصد نمونه = ۰,۸	٪۵۷,۶۸
KNN-bag	K ۱۵ = با درصد نمونه = ۱	٪۸۶,۶۷
KNN–bag	K ۲۰ = با درصد نمونه = ۰,۹	٪۸۶,۶۷
KNN–bag	K ۲۰ = با درصد نمونه = ۱	٪۱۴,۶۷

نام روش	بهترین پارامترها	دقت آزمون
KNN-boost	K= ۵ با تعداد تکرار=۸ یا ۹ یا ۱۰	٪۷۱٫۶۵
KNN-boost	K= ۱۰ با تعداد تکرار=۸ یا ۹ یا ۱۰	٪۴۳٫۶۵
KNN-boost	K= ۲۰ با تعداد تکرار=۸ یا ۹ یا ۱۰	٪۱۴٫۶۷

منبع: یافته‌های تحقیق

در این تحقیق برای توسعه مدل پیش‌بینی از نرم‌افزار رپیدماینر، یکی از نرم‌افزارهای قدرتمند داده‌کاوی استفاده شده است. در جدول ۷ پارامترهای مناسب سه روش پیش‌بینی درخت تصمیم و سه روش k نزدیک‌ترین همسایگی به همراه دقت پیش‌بینی آن‌ها زمانی که از متغیرهای انتخابی سه روش فیلتری استفاده شده باشد، نشان داده شده است. به‌طور مشابه در جدول ۸ پارامترهای مدل‌های مختلف پیش‌بینی، زمانی که از متغیرهای انتخابی توسط روش الگوریتم ژنتیک استفاده شود، ارائه شده است. برای افزایش دقت هر مدل، مدل‌ها با پارامترهای متنوع (مانند عمق‌های مختلف در روش درخت تصمیم) اجرا شده و بهترین مقدار دقت در ستون مربوطه ثبت شده است.

۳-۵- ارزیابی نتایج

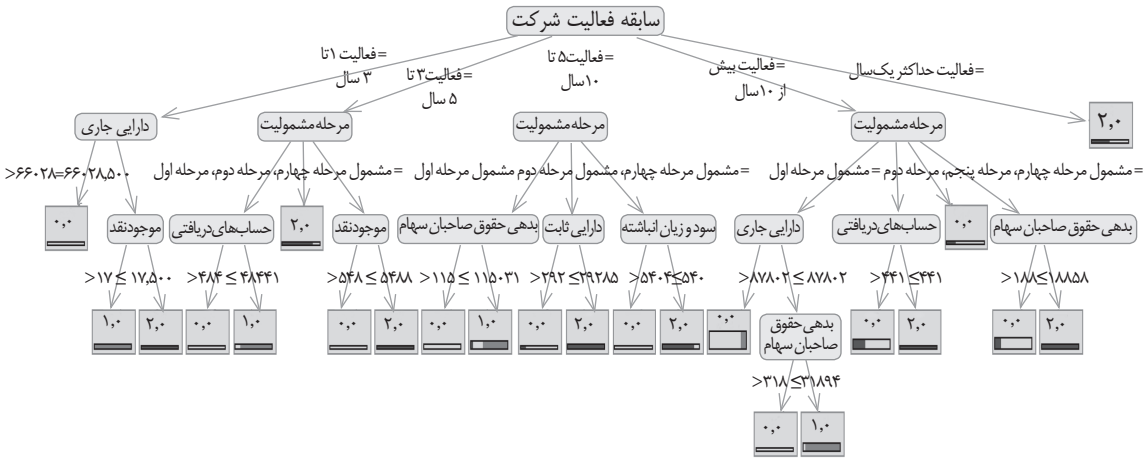
اختلاف بین مقادیر پیش‌بینی و مقدار واقعی، بیانگر دقت مدل پیش‌بینی است. در این پژوهش، درصد تشخیص صحیح متغیر پیش‌بین به عنوان معیار ارزیابی مدل استفاده شده است. یعنی نسبت تعداد تشخیص‌های صحیح مدل پیش‌بین به کل داده‌ها، معیار ارزیابی عملکرد مدل می‌باشد. در ستون آخر جدول‌های ۷ و ۸، دقت پیش‌بینی هر یک از مدل‌های مختلف به تفکیک برای متغیرهای انتخابی روش‌های فیلتری و الگوریتم ژنتیک ارائه شده است.

مطالعه نتایج جدول‌های مذکور نشان می‌دهد، روش پایه درخت تصمیم، بگینگ و بوستینگ، دقتی برابر با یکدیگر (۱۴٫۸۲ درصد) دارند. یعنی روش‌های بگینگ و بوستینگ تأثیری در افزایش دقت روش پایه درخت تصمیم نداشتهاند و همچنین نتایج به دست آمده از روش درخت تصمیم در مقایسه با سه روش مرتبط با k نزدیک‌ترین همسایگی قابل اعتمادتر است. شکل زیر خروجی روش درخت تصمیم را به ازای داده‌های تحقیق نشان می‌دهد. برای ارزیابی بهتر می‌توان در تحقیقات آتی از رویکرد اعتبارسنجی ضربدری[1] نیز استفاده کرد که در آن زیرمجموعه‌های متعددی از داده‌ها تحت عنوان داده‌های آموزش و آزمایش انتخاب می‌شود تا دقت مدل پیش‌بینی با اجراهای متعدد ارزیابی شود. میانگین درصد دقت

1. Cross Validation

پیش‌بینی اجراهای متعدد، مبنای ارزیابی عملکرد مدل می‌باشد.

شکل (۲)- خروجی شبکه درخت تصمیم مسئله کشف تقلّب (یا فرار) مالیاتی



هر کدام از مسیرها در درخت تصمیم فوق منجر به تصمیم‌گیری در مورد یک اظهارنامه مالیات بر ارزش افزوده است. به‌عنوان مثال در صورتی‌که سابقه شرکت مورد بررسی بین یک تا سه سال باشد، لازم است به دارایی جاری شرکت مورد بررسی توجه کرد. در صورتی‌که دارایی جاری آن بیشتر از ۵۰۰٫۶۶۰۲۸ واحد باشد، به احتمال زیاد تقلّبی در آن صورت نگرفته است (متغیر پاسخ مقدار صفر دارد). همان‌طور که در شکل ۲ مشخص است، سابقه فعالیت و مرحله مشمولیت بیشترین تأثیر را در تعیین وضعیت پرونده‌های مالیاتی داشته‌اند. مقبولیت قواعد مستخرج از شکل ۲ بر اساس نظرات ممیزین این حوزه و نتایج حاصل از اجرای آن بر داده‌های آزمون (با دقت بالای ۸۲ درصد) تایید شده است.

۳-۶- اجرا و گسترش نتایج

در مرحله پایانی پروژه داده‌کاوی، دانش و نتایج به دست آمده در فرآیندهای کسب و کار جهت بهبود و نیل به اهداف مرحله اول به کار گرفته می‌شوند. بر اساس نتایج درخت تصمیم می‌توان قوانینی در اختیار ممیزین مالیاتی قرار گیرد تا درجه کم‌اظهاری در اظهارنامه‌های مالیات بر ارزش افزوده را محاسبه کرد. این قوانین از دنبال‌کردن هر یک از مسیرهای شبکه شکل ۲ به‌عنوان یک قانون استخراج می‌شوند. به‌منظور تبیین نحوه استفاده از شکل ۲، سه نمونه از قوانین قابل استخراج از این شکل در ادامه ارائه شده است.

قانون ۱: اگر سابقه فعالیت شرکت ۱ تا ۳ سال باشد و دارایی جاری آن بیشتر از ۵٫۶۶۰۲۸ واحد باشد، آنگاه

اظهارنامه تحویلی تقلّب ندارد (انحراف مقدار اظهاری و تاییدی کمتر از ۱۵ درصد است).

قانون ۲: اگر سابقه فعالیت شرکت ۱ تا ۳ سال باشد و دارایی جاری آن کمتر از ۵٫۶۶۰۲۸ و موجودی نقد آن کمتر از ۵٫۱۷ واحد باشد، آنگاه اظهارنامه تحویلی مشکوک به تقلّب است (انحراف مقدار اظهاری و تاییدی بیشتر از ۱۵ درصد و کمتر از ۵۰ درصد است).

قانون ۳: اگر سابقه فعالیت شرکت ۱ تا ۳ سال باشد و دارایی جاری آن کمتر از ۵٫۶۶۰۲۸ و موجودی نقد آن بیشتر از ۵٫۱۷ واحد باشد، آنگاه اظهارنامه تحویلی تقلّب دارد (انحراف مقدار اظهاری و تاییدی بیشتر از ۵۰ درصد است).

در صورتی‌که سازمان امور مالیاتی کشور از رویکرد حاضر با داده‌های بیشتری اقدام به ساخت مدل پیش‌بینی نماید، قطعاً قوانین معتبرتری برای تعیین وضعیت پرونده‌ها قبل از بررسی حاصل خواهد کرد.

۴- نتیجه‌گیری و پیشنهادات

وجود تقلّب و کم‌اظهاری در اظهارنامه‌های مالیات بر ارزش افزوده یکی از مشکلات و مسائلی است که سازمان مالیاتی کشور با آن مواجه هستند. توسعه روش‌های مکانیزه برای پیش ارزیابی اظهارنامه‌ها و پیش‌بینی میزان تقلّب در آن‌ها به استفاده از منابع انسانی و کاهش هزینه‌های سازمان مالیاتی و افزایش بهره‌وری کمک شایانی خواهد نمود. با توجه به آمار به‌دست آمده از سازمان امور مالیاتی کشور در خصوص فرار از مالیات و تقلّب مالیاتی مؤدیان، می‌توان چنین برداشت نمود که سیستم‌هایی به منظور بررسی بیشتر داده‌های موجود در این سازمان‌ها نیاز است. زیرا یکی از مهم‌ترین مشکلاتی که سازمان‌های مالیاتی با آن مواجه هستند عدم همکاری مؤدیان مالیاتی در خصوص پرداخت صحیح مالیات بر ارزش افزوده می‌باشد که در نهایت از این داده‌ها می‌توان دانش مناسبی را به منظور تسهیل در بازرسی‌های مؤدیان استخراج نمود. در این پژوهش از طریق مصاحبه، مؤلفه‌های بالقوه مؤثر بر شناسایی اظهارنامه‌های کم‌اظهار تعیین شده‌اند. داده‌های ۴۹۵ اظهارنامه مالیات بر ارزش افزوده و عملکردی مربوطه به‌طور تصادفی جمع‌آوری شده است. از اختلاف مقدار خوداظهاری (مالیات ابرازی) و مالیات تعیین‌شده نهایی به عنوان معیار کشف تقلّب (متغیر وابسته تحقیق) استفاده شده و تقلّب (کم‌اظهاری) شامل سه وضعیت سالم (انحراف کمتر از ۱۵ درصد)، مشکوک به تقلّب (انحراف بین ۱۵ تا ۵۰ درصد) و تقلّب (انحراف بیش از ۵۰ درصد) تعیین شده است.

به منظور پیش‌بینی وضعیت پرونده‌های مالیات بر ارزش افزوده از دو روش پایه درخت تصمیم و k نزدیک‌ترین همسایگی به همراه روش‌های جمعی بگینگ و بوستینگ استفاده شده است. تلاش شده بهترین مقدار برای پارامترهای مدل‌های استفاده شده انتخاب شود تا دقت آن‌ها برای پیش‌بینی ارتقاء

یابد. نتایج نشان می‌دهد، روش درخت تصمیم توسعه داده شده در این مقاله، می‌تواند با دقت ۱۴,۸۲ درصد وضعیت پرونده‌های مالیات بر ارزش افزوده را از منظر کم‌اظهاری پیش‌بینی کند. همچنین قوانین تصمیم‌گیری به عنوان خروجی این مدل برای تصمیم‌گیری ممیزین مالیاتی توسعه داده شده است. نتایج نشان می‌دهد، سابقه فعالیت و مرحله مشمول بیشترین تأثیر را در تعیین وضعیت پرونده‌های مالیاتی داشته‌اند. لذا توجه به این چند متغیر از سوی ممیزین مالیات بر ارزش افزوده می‌تواند در کشف پرونده‌های تقلّب مفید باشد. استفاده از سایر روش‌های پیش‌بینی به منظور افزایش دقت پیش‌بینی، پیاده‌سازی مدل پیشنهادی در این تحقیق بر روی مجموعه داده‌های بیشتر و مقایسه نتایج با نتایج به‌دست آمده در این مقاله و همچنین تعریف، جمع‌آوری و به‌کارگیری متغیرهای مستقل بیشتر برای افزایش دقت پیش‌بینی از جمله پیشنهادات تحقیقات آتی می‌باشد.

فهرست منابع

۱. سازمان امور مالیاتی کشور (۱۳۹۲). کتاب برنامه عملیاتی سال ۱۳۹۲.

۲. وزارت امور اقتصادی و دارایی (۱۳۸۴). سازمان حسابرسی، استاندارد حسابرسی.

3. Breunig, M. M.; Kriegel, H. P.; Ng, R. T.; Sander, J. (2000). LOF: Identifying Density-based Local Outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD: 93–104.

4. Davia, H. R., Coggins, P., Wideman, J., and Kastantin, J. (2000). Accountant's Guide to Fraud Detection and Control, 2nd ed., Wiley.

5. Dillon, D., and Hadzic, M. (2009). A Framework for Detecting Financial Statement Fraud through Multiple Data Sources. In Digital Ecosystems and Technologies, 3rd IEEE International Conference on Digital Ecosystems and Technologies. Istanbul, 692-696.

6. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI magazine, 17(3), 37-54.

7. Gonzalez, P. C., Velasquez, J. D. (2013). Characterization and Detection of Taxpayers with False Invoices using Data Mining Techniques. Expert Systems with Applications, 40. 1427–1436.

8. Harrison, G. and Krelove, R. (2005). VAT refunds: A Review of Country Experience. International Monetary Fund (IMF), http://www.imf.org/external/pubs/ft/wp/ 2005/ wp05218.pdf.

9. Hsu, K. W., Pathak, N., Srivastava, J., Tschida, G., and Bjorklund, E. (2015). Data Mining -based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue. In Real World Data Mining Applications. Springer International Publishing. 221-245.

10. Jarmulak, J. & Craw, S. (1999). Genetic Algorithms for Feature Selection and Weighting", Appears in Proceedings of the IJCAI'99 Workshop on Automating

the Construction of Case -based Reasoners.

11. Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. Expert Systems with Applications, 32(4), 995-1003.
12. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature. Decision Support Systems, 50(3). 559-569.
13. OECD (1999). Compliance Measurement, Practice Note. Centre for Tax Policy and Administration, Tax Guidance Series. General Administrative Principles – GAP004 Compliance, http://www.oecd.org/tax/ administration/ 1908448.pdf.
14. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. arXiv Preprint arXiv.1009.6119, http://arxiv.org/ftp/arxiv/papers/1009/1009.6119.pdf.
15. Pyle, D. (1999). Data Preparation for Data Mining (Vol. 15. Morgan Kaufmann,www.temida.si/~bojan/MPS/materials/Data_preparation_ for _data_mining.pdf.
16. Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of Financial Statement Fraud and Feature Selection using Data Mining Techniques. Decision Support Systems, 50(2), 491-500.