مجله علمی، پژوهشنامه مالیات،
شماره پنجاه و نه (مسلسل ۱۰۷)،
پاییز ۱۴۰۲، صفحه ۲۵-۷

مقاله پژوهشی

انتخاب برای حسابرسی مالیاتی با استفاده از الگوریتم‌های داده‌کاوی

محمد برزگری دهج[۱]
احمد یعقوب‌نژاد[۲]
امیررضا کیقبادی[۳]
آزیتا جهانشاد[۴]

doi 10.61186/taxjournal.33.59.8

چکیده

با تصویب قانون مالیات‌های مستقیم در سال ۱۳۹۴ و اصلاح ماده ۹۷ آن، سازمان امور مالیاتی کشور مکلف است اظهارنامه مالیاتی تسلیمی اشخاصی که شروع سال مالی آنها از ۱۳۹۷/۰۵/۲۷ و به بعد می‌باشد را بپذیرد و صرفاً تعدادی از آنها را براساس شاخص‌های ریسک انتخاب و مورد حسابرسی قرار دهد. یکی از روش‌های تعیین مؤدیان پرریسک مالیاتی استفاده از روش‌های داده‌کاوی می‌باشد که به موجب آن می‌توان براساس اطلاعات هر مؤدی، مؤدیان پرریسک را تعیین نمود. در این تحقیق، اطلاعات اظهارنامه‌های مالیاتی اشخاص حقوقی از سال ۱۳۹۳ تا ۱۳۹۵ برای ارزیابی ریسک مورداستفاده قرار گرفته‌است. الگوریتم‌های مورد استفاده در این پژوهش، روش‌های دسته‌بندی ماشین بردار پشتیبان، شبکه عصبی، درخت تصمیم و نزدیک‌ترین همسایه بوده است. نتایج پژوهش مؤید آن است که الگوریتم شبکه عصبی به عنوان بهترین الگوریتم برای برآورد ریسک اظهارنامه، معرفی می‌شود.

واژه‌های کلیدی: ماده ۹۷ قانون مالیات‌های مستقیم، ریسک اظهارنامه مالیاتی، روش‌های داده‌کاوی، مؤدیان پرریسک مالیاتی

تاریخ دریافت: ۱۴۰۲/۰۷/۲۳، تاریخ پذیرش: ۱۴۰۲/۰۹/۰۸

۱. دانشجوی دکتری، گروه حسابداری، دانشکده اقتصاد و حسابداری، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران، mba9090@iran.ir
۲. دانشیار، گروه حسابداری، دانشکده اقتصاد و حسابداری، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران، (نویسنده مسئول)، yaghoobacc@gmail.com
۳. استادیار، گروه حسابداری، دانشکده اقتصاد و حسابداری، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران، acc1388@gmail.com
۴. دانشیار، گروه حسابداری، دانشکده اقتصاد و حسابداری، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران، az_jahanshad@yahoo.com

مقدمه

سابقه استفاده از رویکردهای نوین حسابرسی مبتنی بر ریسک و بهره‌گیری از تجزیه و تحلیل استراتژیک و ریسک تجاری توسط مؤسسات بزرگ حسابرسی به اواسط دهه نود میلادی برمی‌گردد (Dannell and Schults, 2003). پیشرفت‌های صورت گرفته در مفاهیم نظری حسابرسی، کاربرد گسترده فناوری اطلاعات در تجارت و پدیدار شدن تکنولوژی‌ها، چالش جدیدی را در روش‌های حسابرسی ایجاد کرده است (Efstathios et al, 2007). پیدایش این روش‌های نوین و ایجاد تغییرات بنیادین در قوانین و مقررات مالیاتی در ایران، لزوم به کارگیری روش‌های متفاوت با گذشته را بیش از پیش ضروری نموده است. از جمله این رویکردها استفاده از روش‌های نوین تحلیل داده، است. یکی از تفاوت‌های بین تجزیه و تحلیل سنتی داده‌ها و روش‌های نوین در این است که در تجزیه و تحلیل سنتی، این فرض وجود دارد که فرضیه‌ها شکل گرفته‌اند. اما در روش‌های جدید، الگوها و فرضیه‌ها به صورت اتوماتیک از داده‌های مورد بررسی استخراج می‌شوند و براساس نتایج حاصل شده، از بین حساب‌ها، «حوزه‌های شک» برای تمرکز در رسیدگی مشخص می‌شود. علاوه بر آن، روش‌های سنتی و دستی، به شدت زمان‌بر و پرهزینه هستند؛ بنابراین ارائه روش‌های فناوری اطلاعات که هزینه و زمان را به طرز قابل ملاحظه‌ای کاهش دهد، مورد نیاز است. در بیشتر موارد فناوری نوین اطلاعاتی، بسیار بیشتر از فناوری کنترلی پیشرفت کرده و توسعه یافته است (Arabmazar & Mohamadi, 2008).

در نتیجه این پیشرفت‌ها و توسعه سیستم‌های اطلاعات حسابداری، ارزیابی اظهارنامه مالیاتی اشخاص به عنوان خروجی سیستم‌های حسابداری و تعیین رتبه ریسک آنها، چالش جدیدی پیش‌روی سازمان امور مالیاتی در اجرای مقررات اصلاحیه جدید قانون مالیات‌ها ایجاد کرده است. مسئولان استاندار سازی اعتقاد دارند که فرآیند ارزیابی ریسک به عنوان چهارچوب محوری، کیفیت و اثربخشی حسابرسی را ارتقا خواهد داد و منتج به یک تغییر ضروری در حسابرسی می‌شود (Bell, et al., 2005). ریسک، نوعی تهدید بالقوه قابل پیش‌بینی است. برای جلوگیری از رویدادهای تهدید کننده که احتمال دارد بر امنیت سیستم اثر بگذارد، باید ارزیابی‌های خاصی انجام شود (Moeinedin & Fazelyazdi, 2011). اگر تهدید بالقوه فرآیند اجرای مقررات ماده (۹۷) قانون مالیات‌ها را «پذیرش اظهارنامه‌های نادرست» و یا «انتخاب اظهارنامه‌های درست، برای حسابرسی» تعریف کنیم، لازم است برای کم‌ریسک قلمداد کردن و پذیرش اظهارنامه مالیاتی و یا پرریسک قلمداد نمودن و عدم پذیرش آن، نسبت به ارزیابی آن، با ابزاری مناسب اقدام نمائیم. در این پژوهش، روش‌های داده‌کاوی به عنوان ابزاری برای این منظور معرفی و مورد آزمون قرار گرفته‌اند. به این ترتیب که با استفاده از پرونده‌های رسیدگی شده، دو گروه اظهارنامه‌های مشمول رسیدگی متمم[۱] و اظهارنامه‌های غیر مشمول رسیدگی متمم تشکیل شده است. پس از آن، با شناسایی

۱. موضوع ماده (۱۵۶) قانون مالیات‌های مستقیم

اظهارنامه‌های پرریسک، با استفاده از ابزارهای داده‌کاوی در نمونه پژوهش و مقایسه با نتایج واقعی عملیات حسابرسان امور مالیاتی، نسبت به ارزیابی موفقیت روش‌های معرفی شده در پیش‌بینی ریسک اظهارنامه اقدام شده است.

بیان مسئله

رشد روزافزون دیتا و اطلاعات و رویکرد سازمان امور مالیاتی کشور برای حرکت به سوی فرآیندهای الکترونیکی و هوشمند از یک سو و چالش‌ها و الزامات اصلاحیه جدید قانون مالیات‌های مستقیم از سوی دیگر، به کارگیری ابزارهای متناسب با این تغییرات را ضروری کرده است. استخراج اطلاعات نهان و شناسایی الگوها و روابط داده‌ها در بانک بزرگ اطلاعاتی طرح جامع مالیاتی، که در راستای هوشمندسازی نظام مالیاتی پایه‌ریزی شده است از استراتژی‌های راهبردی سازمان است (INTA, 2020). برای این منظور و تحقق اهدافی چون کشف داده‌های خطا و پرریسک، استفاده از ابزارهای تحلیل داده مدنظر قرار گرفته است. حاجیها (۱۳۸۹) معتقد است نقطه شروع حسابرسی مبتنی بر ریسک، تعیین سطح خطر است که حسابرس هنگام بیان اظهارنظر حسابرسی آماده پذیرش آن است. روش ارزیابی خطر حسابرسی، ممکن است بر برنامه‌ریزی، طرح‌ریزی راهبردهای متعاقب حسابرسی و نتایج نهایی آن اثر بگذارد. استانداردهای حسابرسی، ارزیابی خطر را بر اساس مدل خطر حسابرسی الزامی می‌کند. آنچنان که افستاتیوس و همکاران (۲۰۰۷) در پژوهش خود عنوان می‌کنند، روش‌های داده‌کاوی، دانش یا قانون نهفته در ورای داده‌ها را استخراج می‌کنند تا با آن بتوان مدل‌های مختلف تحلیل‌کننده داده را ساخت. داده‌کاوی یک اصطلاح کلی است که دربرگیرنده روش‌هایی است که به هدف استخراج هوش انسانی از داده‌ها صورت می‌گیرد (Setayesh et al. , 2011).

در این راستا، مسئله اصلی که این تحقیق به بررسی و آزمون آن پرداخته است استفاده از ابزارها و روش‌های داده‌کاوی برای تعیین اظهارنامه‌های مالیاتی مؤدیان با ریسک بالا بوده است.

مبانی نظری و پیشینه تحقیق

براساس اصلاحیه جدید قانون مالیات‌های مستقیم، سازمان امور مالیاتی می‌تواند اظهارنامه مالیاتی دریافتی از مؤدیان را بدون رسیدگی قبول و تعدادی از آنها را براساس معیارها و شاخص‌های تعیین شده به طور نمونه انتخاب و برابر مقررات مورد رسیدگی قرار دهد. مسیحی و همکاران (۱۳۹۸) در تحقیقی تحت عنوان «طراحی مدل انتخاب برای حسابرسی مالیاتی اشخاص حقوقی در نظام مالیات بر ارزش افزوده» تکنیک‌های داده‌کاوی را به منظور دستیابی به مدل بهینه برای انتخاب پرونده‌ها برای حسابرسی ارزش افزوده مورد مطالعه قرار داده‌اند. نتایج

این تحقیق نشان داده است برای رسیدن به نتیجه بهینه در حسابرسی‌های مالیات بر ارزش افزوده، استفاده از ابزارهای داده‌کاوی ضروری است.

در این پژوهش اطلاعات جداول مختلف اظهارنامه‌هایی که در رسیدگی‌های بعدی مشمول تعیین درآمد مشمول مالیات متمم (تعدیل گزارش حسابرسی) شده‌اند، با اطلاعات جداول سایر اظهارنامه‌های مورد آزمایش، مقایسه شده‌اند. این اطلاعات بسیار زیاد و متنوع بوده و انجام مقایسات آنها با الگوهای دستی، کاری ناممکن یا پرهزینه و زمان‌بر است.

در روش‌های داده‌کاوی، انواع مختلفی از توابع هدف تعریف می‌شود و متناسب با هدفی که از داده‌کاوی انتظار می‌رود، یک یا مجموعه‌ای از این توابع هدف انتخاب می‌شود و سعی در حداقل نمودن هزینه‌ای است که از توابع هدف به دست می‌آید. بنابراین مهم‌ترین مزیت مستقیم داده‌کاوی می‌تواند کاهش هزینه و زمان باشد؛ با این وجود مزیت‌های استفاده از آن، تنها به این موارد محدود نمی‌شود. همچنین از طریق داده‌کاوی می‌توان ارتباط بین چندین پارامتر مختلف در یک مجموعه از داده‌ها را که به صورت دستی یا با روش‌هایی غیر از داده‌کاوی غیرقابل تشخیص است، در قالب یک مدل یادگیرنده ارائه داد. این مدل‌های یادگیرنده با وجود داده‌های جدید به صورت هوشمند آموزش داده می‌شود و خود را بهبود می‌دهند (HirshPasek, 2015). عملیات داده‌کاوی معمولاً حداقل به دو نوع مجموعه از داده‌ها نیاز دارد. مجموعه اول که از آن به عنوان داده‌های آموزشی[1] یاد می‌شود، داده‌هایی هستند که از آن‌ها برای کشف الگو و روابط بین پارامترهای مختلف نمونه‌های آموزش و برای برازش مدل استفاده می‌شود. مجموعه دوم، داده‌هایی هستند که روش داده کاوی با آن ارزیابی می‌شود. این داده‌ها بدون در نظر گرفتن برچسب آن، به مدل داده شده و در نهایت از مدل می‌خواهیم که با توجه به فرآیند یادگیری انجام شده، برچسب تخمینی را به عنوان خروجی ارائه دهد. در نهایت با استفاده از برچسب‌های قطعی و برچسب‌هایی که از طریق مدل به دست آمده است فرآیند ارزیابی مدل صورت می‌پذیرد. بدین ترتیب می‌توان میزان دقت روش داده کاوی را اندازه گرفت. به این مجموعه از داده‌ها، داده‌های آزمایش یا آزمون می‌گوییم (Karahoca et al, 2012).

یکی از تکنیک‌های داده‌کاوی، روش‌های باناظر[2] است. در این روش‌ها، ورودی و خروجی داده‌ها مشخص بوده و قرار بر این است که از طریق الگوریتم، روابط مابین ورودی‌ها و خروجی‌ها به صورت یک تابع، آشکار گردیده و در داده‌های جدید، خروجی‌ها را براساس ورودی‌ها پیش‌بینی کنیم. مسائل یادگیری بانظارت به دو دسته‌ی رگرسیون و دسته‌بندی[3] تقسیم می‌شود. مسائل رگرسیون مسائلی هستند که برچسب آن‌ها مقادیر

1. Training Data
2. Supervised
3. Classification

پیوسته است، در صورتی که مسائل دسته‌بندی آن‌هایی هستند که برچسب آن‌ها محدود یا به طور کلی مقادیر گسسته است (Karahoca et al, 2012). در ادامه به برخی از کارهای مرتبط با استفاده از این رویکردها اشاره می‌شود. اسپاتیس (۲۰۰۷) در پژوهشی با استفاده از نمونه‌ای شامل ۱۰۰ شرکت یونانی و با بهره گیری از روش‌های رگرسیون حداقل مربعات معمولی و لجستیک به بررسی تاثیر دعاوی حقوقی و اطلاعات مالی بر گزارش مشروط حسابرسی پرداخته است. تحقیقات ایشان نشان داده است که دعاوی حقوقی و بحران مالی از عوامل موثر بر صدور گزارش مشروط می‌باشد. صحت مدل وی در پیش‌بینی نوع اظهارنظر حسابرس به میزان ۸۷ درصد بوده است.

راکس و همکاران (۲۰۱۸) یکی از اصلی‌ترین اولویت‌های مالیات‌های محلی را کشف تقلب مالیاتی عنوان کرده‌اند که ملزم به تدوین استراتژی‌های مقرون به صرفه مانند روش‌های سیستمی برای مقابله با این مشکل هستند. اکثر کارهای اخیر در زمینه کشف تقلب مالیاتی بر اساس تکنیک‌های یادگیری با نظارت است که از داده‌های برچسب گذاری شده یا با کمک حسابرس استفاده می‌کنند. متأسفانه، حسابرسی اظهارنامه‌های مالیاتی یک فرایند کند و پرهزینه است، بنابراین دسترسی به اطلاعات تاریخی دارای برچسب بسیار محدود است. به همین دلیل، کاربرد روش‌های یادگیری ماشین تحت نظارت برای کشف تقلب مالیاتی به شدت تحت تاثیر قرار می‌گیرد.

پلاکنسیا و همکاران(۲۰۲۰) با توجه به محدودیت‌های موجود در ادارات مالیاتی، مانند: کارکنان، ابزار، زمان و غیره، ادارات مالیاتی به دنبال بازیابی بدهی‌ها در مراحل اولیه هستند. آنها با استفاده از تکنیک‌های یادگیری عمیق مدلی را جهت پیش‌بینی بدهی‌های مؤدیان با احتمال زیاد عدم پرداخت در مدت زمان کوتاه ارائه داده‌اند. برای اندازه گیری عملکرد از معیار شاخص تطابق استفاده کردند و عملکرد به دست آمده ۹۰ درصد بوده است.

با توجه به این که در حوزه این پژوهش، هدف تعیین ریسک اظهارنامه مالیاتی است، این مسئله یک مسئله دسته‌بندی تلقی می‌شود و برچسب آن پرریسک یا کم‌ریسک خواهد بود؛ بنابراین در این حالت، داده‌های ورودی، داده‌هایی خواهند بود که علاوه بر متغیرهای اصلی شامل برچسب پرریسک یا کم ریسک نیز می‌شوند. روش‌های دسته‌بندی به طور کلی تعدد بالایی دارند و هر کدام مزیت‌ها، چالش‌ها و معایب خاص خودشان را دارند. در این پژوهش از چهار روش دسته‌بندی اصلی استفاده شده است. این روش‌ها به ترتیب درخت تصمیم، ماشین بردار پشتیبان، روش‌های مبتنی بر نمونه و شبکه عصبی مصنوعی هستند. همگی روش‌های دسته‌بندی، یادگیری با نظارت هستند و برای پیاده‌سازی و اجرا به داده‌هایی معتبر با برچسب نیاز دارند (Karahoca et al, 2012). برچسب مورد نیاز همان کم‌ریسک بودن یا پرریسک بودن اظهارنامه مالیاتی خواهد بود.

اگر اظهارنامه‌های مالیاتی برچسب نداشته باشند، یعنی اطلاعاتی از این‌که چه اظهارنامه‌هایی پرریسک هستند، نداشته باشیم، نمی‌توانیم از روش‌های یادگیری با نظارت استفاده کنیم و مجبور خواهیم بود از روش‌های

یادگیری بی‌نظارت استفاده کنیم. یادگیری بی‌نظارت در حوزه تعیین ریسک اظهارنامه مالیاتی نمی‌تواند به تشخیص ریسک کمک کند، با این وجود، استفاده از روش‌های خوشه‌بندی باعث می‌شود که بتوان نمونه‌ها را به خوشه‌های متمایز تقسیم کرد. در این حالت هر کدام از این خوشه‌ها با دقت مشخصی نشان‌گر خوشه‌های متمایزی است که نشان‌دهنده فرار مالیاتی، اجتناب از مالیات، صداقت در مالیات، عدم اظهار صادقانه و... هستند.

یکی از روش‌های دسته‌بندی مرسوم، که به طور معمول از آن استفاده می‌شود، دسته‌بندی خطی پرسپترون است. افستاتیوس و همکاران (۲۰۰۷) با استفاده از نمونه‌ای شامل ۴۵۰ شرکت ایرلندی و انگلیسی از سه روش داده‌کاوی شامل پرسپترون چندلایه، درخت تصمیم و شبکه بیزین برای طبقه‌بندی اظهارنظر حسابرسان استفاده نموده‌اند. نتایج این تحقیق، بیانگر عملکرد کلی بالاتر شبکه بیزین نسبت به سایر روش‌ها بوده است.

ساختار درخت تصمیم[1] نوعی یادگیری ماشین نظارت شده است. تکنیک یادگیری ماشین برای استنتاج یک درخت تصمیم از داده‌ها، یادگیری درخت تصمیم نامیده می‌شود که یکی از رایج‌ترین روش‌های داده‌کاوی است. این مدل برای طبقه‌بندی یا پیش‌بینی، براساس سؤالات قبلی به کار می‌رود. گره ریشه، پایه درخت تصمیم است. یک گره به چندین زیرگره تقسیم می‌شود. هر گره متناظر یک متغیر و هر کمان به یک فرزند، نمایانگر یک مقدار ممکن برای آن متغیر است. در صورتی که یک زیرگره به زیرگره‌های بیشتری تقسیم نشود، در واقع نشان دهنده خروجی احتمالی بوده و به آن گره برگ گویند. در واقع یک گره برگ، با داشتن مقادیر متغیرها که با مسیری از ریشه درخت تا آن گره برگ بازخوانی می‌شود، مقدار پیش‌بینی شده متغیر هدف را نشان می‌دهد. در ساختار درخت تصمیم، پیش‌بینی به دست آمده از درخت در قالب یک سری قواعد توضیح داده می‌شود. یادگیری یک درخت می‌تواند با تفکیک کردن یک مجموعه منبع به زیرمجموعه‌هایی براساس یک تست مقدار صفت انجام شود. این فرآیند به شکل بازگشتی در هر زیرمجموعه حاصل از تفکیک تکرار می‌شود. عمل بازگشت زمانی کامل می‌شود که تفکیک بیشتر سودمند نباشد یا بتوان یک دسته‌بندی را به همه نمونه‌های موجود در زیرمجموعه‌ی بدست آمده اعمال کرد (Karahoca et al, 2012). نمازی و صادق‌زاده (۱۳۹۷) قابلیت پیش‌بینی فرار مالیاتی شرکت‌های پذیرفته شده در بورس اوراق‌بهادار تهران را با استفاده از الگوی درخت تصمیم مورد تحقیق قرار داده‌اند. آنها با استفاده از نمونه‌ای متشکل از ۱۰۸۱ شرکت در بازه زمانی ۱۳۸۴ تا ۱۳۹۴ به بررسی الگوریتم‌های درخت تصمیم در پیش‌بینی فرار مالیاتی پرداخته‌اند. براساس نتایج تحقیق آنان روش‌های جنگل تصادفی، کاهش خطای هرس، LMT , J48، ریشه تصمیم و درخت تصادفی به ترتیب از دقت و کارایی بیشتری در پیش‌بینی فرار مالیاتی برخوردار هستند. نتایج این تحقیقات همچنان نشان داده است که کارایی پیش‌بینی روش‌های مختلف درخت تصمیم نسبت به یکدیگر دارای تفاوت معناداری هستند.

1. Decision Tree

بردارهای پشتیبان[1] ابزار دیگری برای استفاده در پیش‌بینی است. ماشین بردار پشتیبان در واقع یک طبقه‌بندی کننده دودوئی است که دو کلاس را با استفاده از یک مرز خطی از هم جدا می‌کند. در این روش با استفاده از تمامی باندها و یک الگوریتم بهینه‌سازی، نمونه‌هایی که مرزهای کلاس‌ها را تشکیل می‌دهند به دست می‌آورند. دینگ و همکاران (۲۰۲۱) با استفاده از روش یادگیری ماشین به ساخت یک مدل ارزیابی خطر هشداردهنده پرداختند که با استفاده از روش ماشین بردار پشتیبان (SVM) به شناسایی شرکت‌هایی نائل شدند که فاکتورهای نادرست صادر می‌کردند (Ding et al, 2021).

از دیگر روش‌های پیش‌بینی روش‌های مبتنی بر نمونه[2] است. در روش‌های مبتنی بر نمونه، از رویه غیرپارامتریک بهره می‌برند. در این روش‌ها هیچ پارامتری کشف نمی‌شود، بلکه در هر مرحله از فرآیند تست، از تمامی داده‌های آموزش به منظور دسته‌بندی استفاده می‌شود و مرحله آموزشی به صورت جداگانه وجود نخواهد داشت. یکی از معروف‌ترین دسته‌بندهای مبتنی بر نمونه، دسته‌بندی k- نزدیک‌ترین همسایه[3] (KNN) است. الگوریتم مذکور روشی بر پایه فاصله است. در این دسته‌بندی، برای کشف برچسب هر داده آزمایش، فاصله آن تا تمامی داده‌های آموزش اندازه‌گیری می‌شود و سپس k- نزدیک‌ترین همسایه انتخاب و برچسب آن داده، برابر با پرتعدادترین برچسب‌ها می‌شود. تشخیص فاصله در اجرای داده‌کاوی در تولید بهترین نتیجه نقش مؤثری دارد.

ابزار دیگر داده‌کاوی برای پیش‌بینی، شبکه عصبی مصنوعی است. شبکه‌های عصبی مصنوعی یا به زبان ساده‌تر شبکه‌های عصبی، سیستم‌ها و روش‌های محاسباتی نوینی برای یادگیری ماشینی، نمایش دانش و در انتها اعمال دانش به دست آمده در جهت پیش‌بینی پاسخ‌های خروجی ازسامانه‌های پیچیده هستند. یکی از پایه‌ای‌ترین مدل‌های عصبی موجود، مدل پرسپترون چند لایه یا به اختصار MLP است که عملکرد انتقالی مغز انسان را شبیه‌سازی می‌کند. از کاربردهای مهم شبکه‌های عصبی مصنوعی می‌توان به پردازش تصاویر، تجزیه و تحلیل در دستگاه‌های پزشکی مثل ضربان‌نگار قلب، شناسایی چهره، اثر انگشت و دستخط، تشخیص نوع صدا، تجزیه و تحلیل انبارها و فروش تجاری، پیش‌بینی و تخمین برآورد و کشف تقلب استفاده کرد؛ بنابراین در زمینه‌ی کشف فرارهای مالیاتی نیز بر روی شبکه‌های عصبی مصنوعی کار شده است. مورونکره و همکاران (۲۰۲۳) به پیش‌بینی فرار مالیاتی با استفاده از مدل‌های یادگیری ماشینی تحت نظارت پرداختند. این تحقیق زمینه‌ای را فراهم می‌کند که در آن، ممیز می‌تواند با استفاده از مدل‌های یادگیری ماشین، بازخورد فوری را دریافت نماید. در این پژوهش، مدل‌های یادگیری ماشین تحت نظارت مانند شبکه عصبی مصنوعی، رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، GaussianNB و XGBoost ارزیابی شده است. با عنایت به نتایج

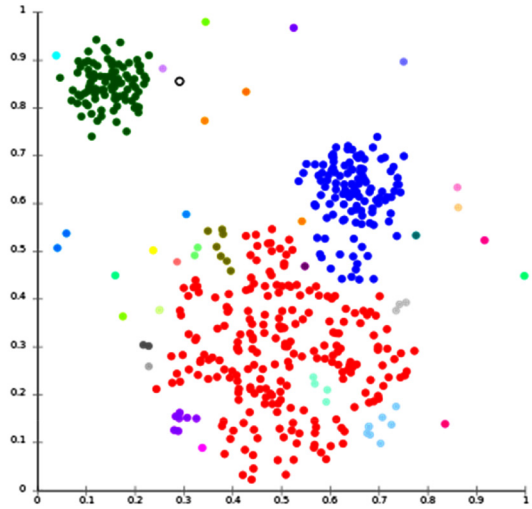
1. Support Vectors
2. Instance-based Approach
3. K-Nearest Neighbor

و بر اساس معیارهای ارزیابی مختلف، شبکه عصبی مصنوعی، قوی‌ترین مدل برای پیش‌بینی فرار مالیاتی بوده است (Murorunkwere et al, 2023).

روزگس و همکاران (۲۰۲۲) با استفاده از روش‌های داده‌کاوی، به بررسی افزایش کارایی تشخیص فرار مالیاتی در کشور لیتوانی پرداختند. این مطالعه مدل‌های مختلفی را برای تقسیم‌بندی، ارزیابی ریسک، الگوهای رفتاری و کشف فرار مالیاتی استفاده کرده است. نتایج نشان می‌دهد که تکنیک داده کاوی می‌تواند به طور موثر فرار مالیاتی را تشخیص دهد و دانش پنهان را استخراج کند. یافته‌های آنها تأیید می‌کنند که رگرسیون لجستیک و روش شبکه عصبی بهترین نتایج را به عنوان مدل‌های تشخیص ارائه می‌دهد. در مورد نمونه‌های آموزشی و اعتبارسنجی، درصد تشخیص صحیح موارد به ترتیب ۸۲. ۳ و ۸۱ درصد بوده است (Ruzgas et al, 2023).

خوشه‌بندی[1] یا آنالیز خوشه، یکی از شاخه‌های یادگیری بی‌نظارت می‌باشد و فرآیندی است که در طی آن، نمونه‌ها به دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند تقسیم می‌شوند که به این دسته ها خوشه گفته می‌شود.

شکل (۱)- خوشه‌بندی مجموعه‌ای از داده‌های بدون برچسب با سه خوشه‌ی کلی



از مهم‌ترین کاربردهای خوشه‌بندی می‌توان به بازیابی اطلاعات، شناسایی بازار تجاری، تقسیم عکس‌ها و خوشه‌بندی داده‌های حجیم اشاره کرد. به عنوان مثال در مسئله فرار مالیاتی، می‌توان عملیات خوشه‌بندی را بر روی داده‌های مالیاتی با تعداد خوشه‌های مختلف انجام داد.

1. Clustering

معیارهای ارزیابی مدل

روش‌های مختلفی برای ارزیابی کیفیت پیش‌بینی‌های یک مدل وجود دارد که سعی می‌کنند خطای پیش‌بینی[1] را برای مقاصد خاصی بر اساس توابع خسارت، امتیاز[2] یا توابع سودمندی[3] به دست آورند. در این قسمت، معیارهای مختلف ارزیابی که بر اساس آن‌ها هر مدل بررسی و با دیگران مقایسه می‌شود توضیح داده شده‌اند.

معیار دقت[4]: این تابع میزان درستی را می‌سنجد که به صورت خطای جهت‌گیری آماری[5] در نظر گرفته می‌شود و به عنوان یک معیار آماری برای آزمایش شناسایی درست یک وضعیت دودویی به کار می‌رود. معادله زیر نحوه محاسبه دقت را نشان می‌دهد.

$$accuracy = \frac{number\ of\ \{true\ positives + \ true\ negatives\}}{number\ of\ \{true\ positives/negatives/\ false\ positives/negatives\}} \quad (۱)$$

در دسته‌بندی‌های چندبرچسبی، تابع دقت زیرمجموعه‌ای را محاسبه می‌کند؛ بنابراین اگر تمامی زیرمجموعه‌ها با برچسب پیش‌بینی شده به درستی تطبیق داده شود، یک و در غیر این صورت صفر برگردانده می‌شود؛ بنابراین اگر y نمایش‌گر برچسب‌های اصلی و $\hat{y}$ برچسب پیش‌بینی شده باشد و n تعداد نمونه‌ها باشد، محاسبه درستی به صورت معادله زیر خواهد بود.

$$accuracy(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} 1(y_i = \hat{y}_i) \quad (۲)$$

معیار صحت[6]–پوشش[7]: صحت قابلیت دسته‌بندی برای برچسب نزدن یک نمونه منفی به عنوان یک نمونه مثبت و پوشش قابلیت دسته‌بندی برای یافتن تمامی نمونه‌های مثبت را نشان می‌دهد و معمولاً با بالا رفتن امتیاز صحت، پوشش کاهش می‌یابد و برعکس؛ بنابراین در طراحی مدل‌های مختلف، به این موضوع توجه می‌شود. در بحث تعیین ریسک اظهارنامه، ترکیب صحت و پوشش هر دو اهمیت پیدا می‌کنند. اگر مدل‌های طراحی شده با صحت بالا باشند ولی پوشش پایینی داشته باشند، تعداد زیادی از اظهارنامه‌های پرریسک شناسایی نشده‌اند. هم‌چنین اگر پوشش بالا ولی صحت پایین باشد، اظهارنامه‌های پرریسک به خوبی شناسایی شده‌اند ولی تعداد اظهارنامه‌های مشکوک به ریسک، بالا بوده و تعداد تشخیص‌های نادرست زیادی وجود خواهد داشت که تشخیص دستی آن‌ها زمان‌بر خواهد بود؛ با این وجود به دلیل این‌که هدف یافتن اظهارنامه‌های پرریسک است،

1. Prediction Error
2. Score
3. Utility Functions
4. Accuracy Score
5. Statistical Bias
6. Precision Score
7. Recall

مدل‌های با بازخوانی بالا به مدل‌های با صحت بالا ترجیح داده خواهد شد.

معیار-F

معیار-F به عنوان میانگین هارمونیک وزن‌دار صحت و پوشش مطرح می‌شود و مطابق معادله زیر محاسبه می‌شود و در بهترین حالت یک و در بدترین حالت صفر خواهد شد و همان‌طور که مشاهده می‌شود، با افزایش β، تأثیر بازخوانی نیز افزایش می‌یابد. معیار-F معمولاً در بازیابی اطلاعات و مسائل جست‌وجو مورد استفاده قرار می‌گیرد و در چارچوب ارزیابی ریسک مطرح شده در این پژوهش اهمیت β با محاسبه مقادیر مختلف وابسته به هدف برای آن برآورده می‌شود.

$$F_\beta = (1+\beta^2)\frac{precision \times recall}{\beta^2 precision + recall} \quad (۳)$$

در مسائل چندبرچسبی و چند کلاسی صحت و پوشش می‌توانند به هر برچسب به طور مستقل اعمال شوند؛ بنابراین نیاز به یک روش متوسط‌گیری روی برچسب‌ها احساس می‌شود. روش‌های مختلفی برای تعیین معیارهای صحت، پوشش و معیار-F وجود خواهد داشت. در این پژوهش از روش‌های معدل وزن‌دار[1]، میکرو[2] و ماکرو[3] استفاده شده است. در روش میکرو، محاسبه معیارها بر اساس جفتِ برچسب و نمونه انجام می‌شود. در روش ماکرو، معیارها بر روی هر برچسب به صورت جداگانه محاسبه و در نهایت بین آن‌ها میانگین گرفته می‌شود و در روش وزن‌دار تعداد هر برچسب نیز در عمل میان‌گیری دخیل می‌شود.

اعتبارسنجی: در یک مسئله دسته‌بندی، داده‌ها به طور کلی به دو مجموعه داده‌های آموزش و داده‌های آزمایش تقسیم می‌شود. داده‌های آزمایش داده‌های بدون برچسبی هستند که می‌خواهیم برچسب آن‌ها را به دست آوریم. داده‌های آموزش داده‌هایی هستند که برچسب آن‌ها وجود دارد؛ با این وجود از همه این داده‌ها به طور مستقیم در فرآیند آموزش استفاده نمی‌شود. در مسائل داده‌کاوی، داده‌های دارای برچسب را به دو دسته داده‌های آموزش و داده‌های اعتبارسنجی[4] تقسیم می‌کنند. داده‌های آموزش برای فرآیند آموزش و یافتن مدل‌ها استفاده می‌شود و داده‌های اعتبارسنجی به منظور ارزیابی هر مدل و انتخاب بهترین مدل به کار می‌رود. این کار هرچند باعث می‌شود که مدل‌های بهتری تولید شوند، باعث می‌شود که داده‌های با ارزش آموزش که می‌توان از آن‌ها برای ایجاد مدل‌های بهتری استفاده کرد، در فرآیند آموزش مورد استفاده قرار نگیرد و تنها برای اعتبارسنجی مورد استفاده قرار گیرند. راه حلی که در استفاده از تکنیک‌های داده‌کاوی مورد استفاده قرار می‌گیرد، اعتبارسنجی

1. Weighted Average
2. Micro Average
3. Macro Average
4. Validation

تقاطعی[1] است. در این حالت، داده‌های آموزش به k دسته تصادفی با اندازه یکسان تقسیم می‌شود و در k مرحله، هر بار یکی از این دسته‌ها به عنوان داده‌های اعتبارسنجی کنار گذاشته می‌شود و از $k-1$ دسته‌ی دیگر به عنوان داده‌های آموزش استفاده می‌شود؛ بنابراین برای هر مدل k مجموعه آموزش و اعتبارسنجی مطرح و جهت ارزیابی نهایی بر روی تمامی این مجموعه‌ها میانگین معیارها درنظر گرفته می‌شود.

سؤال تحقیق

حال با توجه به بررسی روش‌های داده کاوی و ریسک مالیاتی سوالات تحقیق به شرح زیر است:

۱- با عنایت به روش‌های داده کاوی کدام روش‌ها در تعیین ریسک اظهارنامه مالیاتی مؤدیان قابل استفاده هستند؟

۲- توانایی کدام‌یک از مدل‌های داده‌کاوی در پیش‌بینی ریسک اظهارنامه مالیاتی بالاتر است؟

روش‌شناسی و مدل تحقیق

این پژوهش از نوع کاربردی، و روش پژوهش، توصیفی[2] - تجربی[3] است. زیرا برای بررسی موضوع تحقیق، به تحلیل آمار توصیفی و مقایسه نتایج به‌کارگیری ابزارهای داده‌کاوی با داده‌های واقعی حسابرسی ماموران سازمان امور مالیاتی پرداخته است. داده‌های مورد استفاده در این پژوهش، داده‌های سازمان امور مالیاتی کشور که به صورت متمرکز از سال ۱۳۹۳ تا ۱۳۹۵ توسط اظهارنامه‌های الکترونیکی جمع‌آوری شده، بوده است. جدول زیر آمار اظهارنامه‌های تسلیمی مؤدیان در سنوات ۹۳ لغایت ۹۸ را نشان می‌دهد.

جدول (۱) - آمار اظهارنامه‌های تسلیمی مؤدیان در سالهای ۹۳ الی ۹۸

سال	اشخاص حقوقی	اظهارنامه اشخاص حقیقی	تبصره ۱۰۰ اشخاص حقیقی
۱۳۹۸	۳۰۸.۳۰۳	۱.۰۷۵.۳۴۰	۱.۵۲۲.۰۹۹
۱۳۹۷	۳۰۳.۲۹۲	۱.۳۳۷.۰۴۵	۱.۴۲۰.۰۰۶
۱۳۹۶	۲۸۶.۵۳۵	۱.۴۵۸.۳۲۹	۱.۱۶۴.۸۰۱
۱۳۹۵	۲۸۸.۱۸۱	۱.۷۵۶.۹۵۹	۸۰۷.۰۴۸
۱۳۹۴	۲۷۳.۸۲۷	۲.۵۰۸.۳۲۹	۰
۱۳۹۳	۲۵۳.۱۳۶	۲.۳۷۷.۳۵۲	۰

1. Cross-validation
2. Descriptive Method
3. Experimental Method

جامعه آماری این پژوهش، اظهارنامه تسلیمی مؤدیان اشخاص حقوقی در سنوات ۱۳۹۳ تا ۱۳۹۵ بوده است. علت انتخاب این جامعه آماری، در نظر گرفتن گذشت حداقل سه سال و بیشتر از مهلت زمانی پنج ساله انجام رسیدگی‌های ماموران مالیاتی بوده است.[۱] تا فرصت معقولی برای صدور گزارشات متمم و تعدیل گزارش رسیدگی که در این پژوهش به عنوان معیار ارزیابی ریسک اظهارنامه‌های تسلیمی مدنظر قرار گرفته است، فراهم شود. باتوجه به وجود اقلام اطلاعاتی کامل‌تر در اظهارنامه‌های اشخاص حقوقی، این دسته از اظهارنامه‌ها مورد استفاده قرار گرفته‌اند. باتوجه به دسترسی به اطلاعات اظهارنامه‌های تمام مؤدیان، از کل داده‌ها در این پژوهش استفاده شده است.[۲]

اطلاعات اظهارنامه‌های مورد بررسی، در قالب ۱۹ جدول در پژوهش مورد استفاده قرار گرفته‌اند. باتوجه به میزان تاثیر ارقام جداول در مقدار درآمد مشمول مالیات، اولویت‌بندی جداول به صورت زیر بوده است:

۱. اولویت اول اقلام اطلاعاتی مربوط به اطلاعات سود و زیان

۲. اولویت دوم مربوط به مواردی است که از سود و زیان ناشی می‌شود:

جداول فعالیت‌ها و هزینه‌های قابل قبول که شامل جداول مربوط به موجودی کالا و جداول پیمانکاری‌ها است.

جدول جزئیات پذیرش در بورس، جدول درآمدهایی که مالیات آنها به صورت مقطوع پرداخت شده است.

جدول درآمدهای معاف، جدول معافیتها و بخشودگیهای مالیاتی، جدول توسعه، نوسازی و بازسازی واحدهای صنعتی و معدنی.

جدول معافیتها و بخشودگی‌های درآمد حاصل از فعالیتهای خارج از کشور، جدول ثبت کمکهای مالی پرداختی

جدول موجودی مواد و کالا، جدول انواع محصولات اصلی به ترتیب بیشترین فروش، جدول بهای تمام شده کالای فروش رفته

---

۱ . موضوع ماده (۱۵۷) قانون مالیات‌های مستقیم

۲ . درصورت استفاده از فرمول کوکران برای محاسبه تعداد نمونه، محاسبات تعداد نمونه می‌تواند به صورت زیر انجام شود.

$$n = \frac{\frac{z^2 pq}{d^2}}{1 + \frac{1}{N}[\frac{z^2 pq}{d^2} - 1]}$$

که در آن N: حجم جامعه n: حجم نمونه، p: درصد توزیع صفت در جامع یعنی نسبت اظهارنامه‌هایی که دارای ریسک بالا هستند q: درصد اظهارنامه‌هایی که ریسک بالایی ندارند، (اگر میزان p و q مشخص نباشد، از حداکثر مقدار آن یعنی ۰/۵ استفاده می‌کنیم) در سطح خطای ۵ ٪ ، z برار با ۱/۹۶ و برابر با ۳/۸۴ است. d مقدار تفاضل نسبت واقعی صفت در جامعه با میزان تخمین پژوهشگر برای آن صفت در جامعه است. دقت نمونه‌گیری به مقدار این عامل بستگی دارد. در حالتی که مقدار d برابر ۰/۰۵ است نمونه‌گیری دارای بیشترین دقت است.

$$\frac{\frac{(3.84) \times 0.5 \times 0.5}{(0.0025)}}{1 + \frac{1}{815144}[\frac{(3.84) \times 0.5 \times 0.5}{(0.0025)} - 1]} = \frac{384}{1.0004698644} = 384$$

جدول بهای تمام شده کار انجام شده پیمانکاری / خدمات، جدول درآمد ناخالص پیمانکاری / ارائه خدمت
جدول فهرست صادرات و ما به ازاء دریافتی

۳. اولویت بعدی جداول تصمیمات مجمع و جدول مربوط به‌بخش ب (محاسبه مالیات) است
جدول استهلاک زیان سنواتی، جدول گردش حساب سود (زیان) انباشته

۴. اولویت بعدی جدول ترازنامه است

۵. اولویت بعدی اقلام اطلاعاتی است که به شکلی از جداول ۱۸ و ۲۶ اظهارنامه استخراج گردیده

به دلیل مستندات کامل و جامعی که کتاب‌خانه پایتون دارد و هم‌چنین استفاده فراوان آن در مقالات علمی، تحقیقاتی و عملی مختلف، در این پژوهش نیز از این ابزار استفاده شده است. این ابزار، انواع الگوریتم‌های گفته شده در این پژوهش، روش‌های ارزیابی و انواع روش‌های شکست داده و معیارهای آماری را در اختیار قرار می‌دهد. بنابراین به خوبی نیازهای این پژوهش را برطرف می‌کند؛ همچنین این ابزار، مجموعه عظیمی از مثال‌های قابل بازاستفادگی در پژوهش‌های مشابه را دارد و بنابراین برای توسعه‌های بعدی نیز متناسب است. هم‌چنین به دلیل قابلیت‌های زبان پایتون، به خوبی می‌تواند با انواع نرم‌افزارها ترکیب شده و به خروجی مناسب منتج شود. در فاز اول، داده‌های خام جمع‌آوری و طبقه‌بندی شد. سپس داده‌ها وارد نرم‌افزار کردیم. بعد از آن فیلتر نمودن داده‌ها، عمل پاکسازی و پیش پردازش روی داده‌ها انجام گردید. شایان ذکر است که تعداد سطرهای داده‌ای اظهارنامه‌ها ۸۱۵. ۱۴۴ اظهارنامه بود، که با حذف داده‌های تماما صفر، تعداد ۵۲. ۷۶۳ ردیف از داده‌های ورودی حذف شد. بنابراین از تعداد ۷۶۲. ۳۸۱ داده، که صحت برچسب آن‌ها قطعی بود، به منظور اجرای فرآیند داده‌کاوی استفاده شد. بعد از آن، الگوریتم ماشین بردار پشتیبان خطی (LSVM) و... ساخته و در آن و متغیرهای ورودی تعیین شد. سپس الگوریتم اجرا شد تا مدل و خروجی‌های آن ساخته شود. مدل‌ها، الگوریتم‌ها و روش‌های ارزیابی متناسب با هدف‌گذاری فاز اول مشخص شدند. با توجه به این‌که هدف‌گذاری استخراج الگوها به منظور پیش‌زمینه‌ای برای تعیین ریسک اظهارنامه تسلیمی است، مدل مشخص شده به منظور بررسی مدل‌های رده‌بندی[1] است؛ همچنین به منظور ایجاد فهرستی از اظهارنامه‌های با ریسک بالا و بررسی نتایج بر روی آن‌ها، داده‌های ورودی به دو دسته‌ی داده‌های آموزش و آزمایش تقسیم شده‌اند و نتایج توسط الگوریتم‌های مورد نظر استخراج شده‌اند. داده‌های آموزش، شامل دو دسته اظهارنامه تسلیمی که (۱) مشمول مالیات متمم شده و (۲) مشمول متمم نشده‌اند، بوده است. الگوریتم‌های رده‌بندی که در این پژوهش از آن استفاده شده است، روش‌های دسته‌بندی ماشین بردار پشتیبان، MLP، درخت تصمیم و نزدیک‌ترین همسایه بوده است. نتایج این پژوهش، روش MLP را به عنوان بهترین الگوریتم معرفی می‌کند؛ با این وجود پیشنهاد می‌دهد که مستقل از الگوریتم‌های اصلی،

1. Classification

الگوریتم درخت تصمیم به هر شکل بر روی داده‌ها آزمایش شود زیرا ممکن است نتایج بهتری از آن گرفته شود؛ در این صورت مسئله، به صورت یک مسئله قانون محور خواهد بود. در مرحله سوم، عملیات داده‌کاوی با استفاده از طراحی‌های الگوریتمیک دو فاز قبلی انجام شده است. ارزیابی در این فاز از طریق ارزیابی‌های درستی، دقت، فراخوانی، امتیاز f و اطلاعات مشترک انجام می‌شود. همچنین با استفاده از یک بررسی آماری می‌توان الگوهای مختلف را تطبیق داد و از آن‌ها در جهت تشخیص رتبه ریسک اظهارنامه استفاده کرد. ارزیابی به صورت جداگانه بر روی داده‌های آموزش و آزمایش انجام می‌شود؛ بنابراین اختلاف بین آن‌ها می‌تواند خبر از عمومیت بخشی نتایج حاصل از اجرای الگوریتم‌ها بر روی داده‌های آموزش و آزمایش باشد.

تحلیل مدل، خروجی و نتایج نهایی

همان‌طور که پیش از این اشاره شد، در این پژوهش از مدل دسته‌بندی با استفاده از الگوریتم‌های ماشین بردار پشتیبان، یادگیری مبتنی بر نمونه، درخت تصمیم و پرسپترون چندلایه (MLP) استفاده شده و پیاده‌سازی آن در کتابخانه scikit-learn در محیط پایتون بوده است. روش یادگیری مبتنی بر نمونه‌ای که در این پژوهش استفاده شده است، الگوریتم نزدیک‌ترین همسایه است. تمامی این الگوریتم‌ها پیش از این به تفصیل توضیح داده شده است. در صورتی که تعداد نمونه‌ها از صد هزار کم‌تر باشد از روش ماشین بردار پشتیبان استفاده می‌شود و در صورتی که نتایج قابل قبول نباشد، از روش‌های مبتنی بر نمونه استفاده خواهد شد.

معیار دقت[1] در الگوریتم شبکه عصبی (پرسپترون چندلایه) نشان می‌دهد در ۸۵ ٪ موارد، ریسک اظهارنامه (مؤدیانی که حسابرسی به پرونده آنها منجر به تعدیل گزارش حسابرسی یا حسابرسی متمم شده است) به درستی تشخیص داده شده است. به این معنی که چنانچه اظهارنامه پرریسک تشخیص داده شده، در مهلت مقرر در ماده (۱۵۷) قانون مالیات‌های مستقیم مورد رسیدگی متمم قرار گرفته است و چنانچه کم‌ریسک تشخیص داده شده، مورد رسیدگی متمم واقع نشده است. این معیار تمامی خطاها را یکسان درنظر می‌گیرد. یعنی خطای اینکه یک اظهارنامه پرریسک، کم‌ریسک تشخیص داده شود و یا یک اظهارنامه کم‌ریسک، پرریسک تشخیص داده شود، یکسان است. برای غلبه بر این مشکل دو معیار دیگر وجود دارد. معیار صحت[2] که در این الگوریتم نشان می‌دهد ۸۲ ٪ موارد تشخیص پرریسک درست بوده است. به این معنی که در ۸۲ ٪ مواقع، یک اظهارنامه شناسایی شده پرریسک، مورد رسیدگی متمم قرار گرفته است. معیار پوشش[3] که نشان‌دهنده این موضوع است ۸۷ ٪ اظهارنامه‌های رسیدگی شده متمم در آزمون پرریسک تشخیص شده بوده‌اند. این معیار نشان دهنده پوشش روی کل داده‌هاست. معیار F1، میانگین معیار دقت و پوشش است که می‌تواند به عنوان رتبه هر یک از الگوریتم‌ها درنظر گرفته شود. نتایج حاصل از اجرای سایر الگوریتم‌ها در جدول زیر آمده است.

1. Accuracy
2. Precision
3. Recall

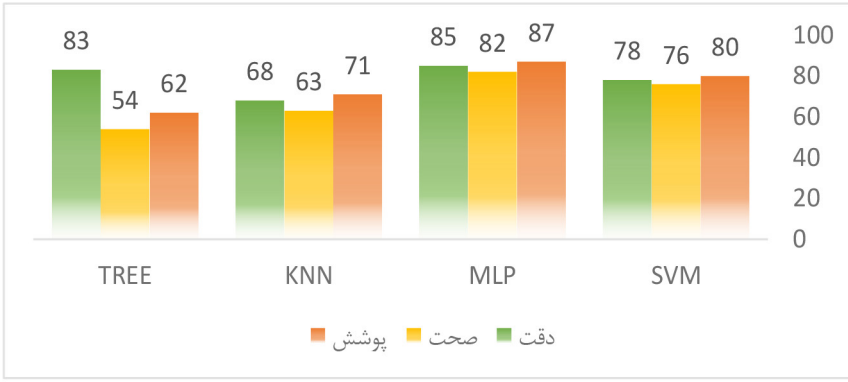
جدول (۲)- نتایج به دست آمده از اجرای الگوریتم‌های پژوهش

الگوریتم	معیار F1	پوشش (Recall)	صحت (Precision)	دقت (Accuracy)
ماشین بردار پشتیبان (ردەبندی Svm)	۰/۷۸	۰/۸۰	۰/۷۶	۰/۷۸
ردەبندی درخت تصمیم decision Tree c4.5	۰/۵۸	۰/۶۲	۰/۵۴	۰/۸۳
ردەبندی نزدیک‌ترین همسایه (knn)	۰/۵۸	۰/۶۲	۰/۵۴	۰/۸۳
شبکه عصبی	۰/۸۵	/۸۷	۰/۸۲	۰/۸۵

منبع: یافته‌های پژوهش

نتایج مربوط به هرکدام از روش‌های داده کاوی در شکل۲ نشان داده شده است.

شکل (۲)- نتایج روش‌های مختلف داده کاوی به صورت نمودار ستونی



منبع: یافته‌های پژوهش

محدودیت‌های پژوهش

اولین محدودیت این پژوهش، عدم اتمام مهلت زمانی پنج ساله موضوع ماده (۱۵۷) قانون مالیات‌های مستقیم برای رسیدگی متمم اظهارنامه‌های سال ۱۳۹۵ در تاریخ انجام پژوهش بوده است. این موضوع باعث می‌شود مواردی وجود داشته باشد که به عنوان اظهارنامه پرریسک تشخیص شده باشد و به علت اتمام مهلت، هنوز مورد رسیدگی متمم قرار نگرفته باشند و به عنوان تشخیص خطای الگوریتم‌ها محسوب شده‌اند. همچنین ممکن است به عنوان اظهارنامه کم‌ریسک تشخیص داده شده و در اجرای الگوریتم‌ها نیز به عنوان تشخیص درست محسوب شده، لیکن در مهلت باقیمانده مورد رسیدگی متمم قرار بگیرند.

محدودیت دیگر این پژوهش، استفاده از رسیدگی‌های متمم به عنوان معیار ارزیابی ریسک اظهارنامه‌های تسلیمی و تهیه برچسب داده‌های آزمون برمبنای آن بوده است. روشن است که درنظر گرفتن این معیار برای تهیه برچسب‌ها، گویای تمام ابعاد ریسک اظهارنامه‌های مالیاتی نیست و جهات بسیار دیگری در این موضوع قابل تصور است. اما باتوجه به اینکه این تحقیق، اولین تحقیق در خصوص بکارگیری ابزارهای نوین تحلیل داده در اجرای مقررات ماده (۹۷) قانون مالیات‌های مستقیم است، می‌تواند به عنوان سرمنشاء انجام تحقیقات دیگر و توسعه جهات دیگر ارزیابی ابعاد ریسک اظهارنامه مالیاتی قرار گیرد.

نتیجه‌گیری و پیشنهادات

ماده (۹۷) قانون مالیات‌های مستقیم را شاید بتوان زیربنایی‌ترین ماده قانونی درخصوص حسابرسی مالیاتی در ایران دانست. با تغییرات این ماده در اصلاحیه قانون[۱]، ساختار جدیدی از حسابرسی مالیاتی پیاده‌سازی شده که اساس آن اظهارنامه مالیاتی است. در ساختار جدید، اظهارنامه یا توسط مؤدی تسلیم و مورد پذیرش واقع می‌شود، یا مردود شده و بر مبنای حسابرسی‌های مالیاتی اصلاح می‌شود و یا اینکه تسلیم نشده و سازمان امور مالیاتی آن را برآورد می‌کند. در هر کدام از حالت‌های فوق، مبنای صدور برگ تشخیص مالیات و تعیین مالیات، اظهارنامه مالیاتی است. آنچه موجب پذیرش، اصلاح و یا برآورد اظهارنامه می‌شود، ارزیابی ریسک آن و داده‌های فعالیت و اطلاعات اقتصادی مؤدیان در طرح جامع مالیاتی است. داده‌کاوی به مفهوم استخراج اطلاعات نهان و شناسایی الگوها و روابط داده‌ها در یک یا چند بانک اطلاعاتی بزرگ، نظیر بانک اطلاعاتی طرح جامع است. بنابراین می‌تواند ابزار مناسبی برای ارزیابی ریسک اظهارنامه باشد.

راه‌کارهای مقابله با فرار مالیاتی از طریق تکنیک‌های هوش مصنوعی و به طور خاص‌تر داده‌کاوی، مورد توجه بسیاری قرار گرفته است. استفاده از حجم عظیم داده‌های مالیاتی که در سالیان طولانی گردآوری می‌شود

---

۱. اصلاحیه قانون مالیات‌های مستقیم مصوب ۱۳۹۴/۰۴/۳۱

و یا استفاده از الگوهای مختلفی که شاید در بررسی تک‌تک افراد، از طریق ممیزی‌های مالیاتی ممکن نباشد و یا با هزینه‌های زیادی همراه شود، از این طریق قابل انجام است. در این حالت حجم عظیم داده‌های مالیاتی که از طریق اظهارنامه‌های مالیاتی فراهم می‌آید در کنار داده‌های حاصل از تشخیص قرار می‌گیرد. در نهایت از طریق تکنیک‌های مختلف داده‌کاوی می‌توان با ضرایب مختلفی که می‌تواند کمک‌رسان حسابرسی مبتنی بر ریسک باشد، به پاسخ رسید.

در این تحقیق، استفاده از ابزارهای داده‌کاوی در ارزیابی ریسک اظهارنامه تسلیمی مؤدیان اشخاص حقوقی در سنوات ۱۳۹۳ تا ۱۳۹۵ مورد پژوهش قرار گرفته است. اولین نتیجه حاصل از پژوهش، این است که ابزارهای داده‌کاوی، مانند روش‌های دسته‌بندی، ماشین بردار پشتیبان، MLP، درخت تصمیم و نزدیک‌ترین همسایه، قابلیت استفاده برای تعیین ریسک اظهارنامه‌های مؤدیان در اجرای مقررات ماده (۹۷) قانون مالیات‌های مستقیم را دارند. بنابراین ضروری است توسعه این ابزارها در ارزیابی ابعاد مختلف ریسک اظهارنامه‌های مالیاتی در پژوهش‌های آتی مورد تحقیق قرار گیرد. همچنین یافته‌های این تحقیق به گونه‌ای که دربخش های مختلف این مقاله توضیح داده شد، نشان می‌دهد روش MLP با ملاک حسابرسی متمم، به عنوان معیار برچسب گذاری داده‌ها، بهترین نتایج را برای تعیین رتبه ریسک اظهارنامه مالیاتی بدست می‌دهد که البته برای استفاده کاربردی، می‌توان نتایج روش MLP را با نتایج روش‌های دیگر بررسی نمود.

فهرست منابع

1. Fazel Yazdi, A. , Moinuddin, M. (2013). Risk Analysis of Accounting Information Systems, Auditor Magazine, No. 60, pp. 106 – 111, (In Persian).
2. Hajiha, Z. (2010). AN Investigation on the Relationship between Inherent and Control Risks in Risk Based Audit Approach, Financial Accouning; 2(6), pp. 95-120, (In Persian).
3. Masihi, M. , Yaghoobnejad, A. , Keyghobadi, A. , Torabi, T. (2019). Using Data Mining Techniques to Measure Tax Risk of Value Added Taxes, Journal of Investment Knowledge, 8(32), pp. 347-363 (In Persian).
4. Mohammadi, T. , Arab Maziar Yazdi, A. , Ghasemi, A, Taklif, A. , jalalpanahi, R. (2020). Balance of Payment Constrained Growth in Tow Developing and Developed Oil-Exporting Economies (Case Study: Iran and Norway) qjerp, 27 (92), pp. 257-296 (In Persian).
5. Namazi, M. , Sadeghzadeh Maharluie, M. (2018). Predicting Tax Evasion by Decision Tree Algorithms FINANCIAL ACCOUNTING, 9(36), pp. 76-100, (In Persian).
6. Setayesh, M. , Ebrahimi, F. , SAIF, S. , Sarikhani, M. (2013). Forecasting The Type of Audit Opinions: A Data Mining Approach, Management Accounting, 5(15), pp. 69-82, (In Persian).
7. Bell, T. B. , Peecher, M. E. , Solomon, I. (2005). The 21st Century Public Company Audit: Conceptual Elements of KPMGs Global Audit Methodology, KPMG, LLP.
8. Ding, N. , Zhang, X. , Zhai, Y. , & Li, C. (2021). Risk Assessment of VAT Invoice Crime Levels of Companies Based on DFPSVM: a case study in China, Risk Management, Palgrave Macmillan, vol. 23(1), pp. 75-96.
9. De Roux, D. , Perez, B. , Moreno, A. , Villamil, M. D. P. , and Figueroa, C. (2018). Tax Fraud Detection for Under-reporting Declarations Using an Unsupervised Machine Learning Approach, In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 215-222.
10. Efstathios. K. , Spathis. Ch. , Nanopoulos. A and Y. Manolopoulos, (2007). Identifying Qualified Auditors Opinions: A Data Mining Approach, Journal of Emerging Technologies Accounting, Vol. 4, pp 183-197.
11. Hirsh-Pasek, Kathy, et al. (2015) Putting Education in Educational Apps: Lessons from the Science of Learning. Psychological Science in the Public Interest 16. 1,

pp. 3-34.

12. The Tax Training, Research and Planninig (2020). Tax organization Development Deed. Tax Library.
13. Karahoca, Adem, Dilek Karahoca, and Mert Sanver. (2012) Survey of Data Mining and Applications (Review from 1996 to Now), Data Mining Applications in Engineering and Medicine: 1.
14. Kiros, Efstathios, Charalambos Spathis, and Yannis Manolopoulos. (2007) Data Mining Techniques for the Detection of Fraudulent Financial Statements, Expert Systems with Applications 32. 4, pp. 995-1003.
15. Murorunkwere, B., F., Haughton, H., Nzabanita, J. Kipkogei, F. & Kabano, I. (2023) Predicting Tax Fraud Using Supervised Machine Learning Approach, African Journal of Science, Technology, Innovation and Development, 15: 6, 731-742.
16. Vuhdj Thwgi (2003) The Influence of Strategic-systems Lens of Auditor Risk Assessments, Working Paper, Arizona State University.
17. Placencia, J. O. , Hallo, M. , & Luján-Mora, S. (2020). Detection of Taxpayers with High Probability of Non-payment: An Implementation of a Data Mining Framework, In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-6. IEEE.
18. Ruzgas T, Kižauskienė L, Lukauskas M, Sinkevičius E, Frolovaitė M, Arnastauskaitė J. (2023). Tax Fraud Reduction Using Analytics in an East European Country, Axioms. 12(3).