کاربرد هوش مصنوعی در حسابرسی مالیاتی

■ محمود نظرپور[۱]
■ سیدحسین نسل موسوی[۲]
■ میرسعید حسینی شیروانی[۳]

تاریخ دریافت: ۹۸/۹/۳۰
تاریخ پذیرش: ۹۹/۲/۱۴

چکیده:

مالیات در ایران به لحاظ تأمین مالی و ابزار سیاستگذاری از جایگاه مناسبی برخوردار نمی‌باشد. بین عملکرد مالیات و سطح بالقوه آن، شکاف وجود دارد که یکی از دلایل اصلی ایجاد آن، وجود پدیده فرار مالیاتی است. در این میان، شناسایی فرار مالیاتی اشخاص حقوقی و ارائه راهکارهایی به منظور شناسایی و مقابله با آن به دلیل درآمدزایی عمده این پایه مالیاتی از اهمیت بالایی برخوردار است. از طرف دیگر وقتی مأمورین مالیاتی جهت حسابرسی به پرونده مالیاتی اشخاص حقوقی مراجعه می‌کنند با حجم زیاد اطلاعات مواجه می‌باشند که در بعضی مواقع ظرف مهلت کمتر از یک روز باید پرونده‌ها را حسابرسی نمایند که نبود فرصت کافی و عدم تفکیک پرونده‌های پرریسک و کم ریسک و ... باعث کاهش کیفیت حسابرسی مالیاتی می‌گردد و مقدمه افزایش فرار مالیاتی مودیان را فراهم می‌کند. در این پژوهش به منظور پیش‌بینی و مبارزه با فرار مالیاتی، تعداد ۲۸ نسبت مالی مربوط به تعداد ۱۱۸۶ شرکت طی سال‌های ۱۳۹۳ تا ۱۳۹۵ که از سایت سازمان بورس اوراق بهادار استخراج گردید به عنوان ورودی مدل، مبنای آموزش و تست روش‌های ترکیبی شبکه بیزین و الگوریتم SVM قرار می‌گیرد. بعد از ارائه مدل، خروجی مشخص می‌کند که شرکت دارای فرار مالیاتی می‌باشد یا خیر. سازمان مالیاتی می‌تواند با استفاده از مدلهای این پژوهش، قبل از حسابرسی مالیاتی مودیان، تشخیص دهد که مؤدی دارای فرار مالیاتی می‌باشد یا خیر و میزان ریسک مؤدی را تعیین نماید. نتایج حاصل از این تحقیق نشان می‌دهد که به ترتیب، روش‌های الگوریتم ترکیبی، الگوریتم SVM و شبکه بیزین از دقت و کارایی بیشتری در کشف و شناسایی فرار مالیاتی برخوردار هستند.

واژه‌های کلیدی: حسابرسی مالیاتی، الگوریتم SVM، شبکه بیزین، فرار مالیاتی، ریسک مالیاتی

۱. دانشجوی دکترای حسابداری، واحد قائم شهر، دانشگاه آزاداسلامی، قائم‌شهر، ایران
۲. عضو هیات علمی گروه حسابداری، واحد قائم شهر، دانشگاه آزاد اسلامی، قائم‌شهر، ایران (نویسنده مسئول)، ایمیل: nseyedhossein2@yahoo.com
۳. گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد ساری، ساری، ایران، ایمیل: mirsaeid_hosseini@iausari.ac.ir

۱- مقدمه

مالیـات به دلیل اهمیـت آن در جبران مخارج عمومـی و آثار اقتصادی که بـه همراه دارد، همواره مورد توجه تمام دولتها بوده اسـت و مطالعات متعددی پیرامون این مقوله و ابعاد مختلف آن، توسـط پژوهش گران انجام گرفته اسـت که فرار مالیاتی یکی از موارد مهم در این پژوهش‌ها می‌باشـد و برای کشـف و شناسایی آن روش‌های بسیاری ارائه شـود. فرار مالیاتی یکی از اجزای اقتصاد غیررسـمی اسـت که در آن افراد قانون گریز به ازای فعالیتهای قانونی انجام شده تمام یا قسمتی از مالیات خود را نمی‌پردازند. در فرار مالیاتی، افراد قانون گریز بر خلاف قوانین و مقررات کشـور فعالیت اقتصادی یا کار غیرقانونی انجام می‌دهند که از این جهت قابل مشـاهده نیست و در آمارهای رسـمی کشور ثبت نمی‌شـود؛ اما به هر حال بررسی و شناخت علل و عوامل مؤثر بر آن اهمیت دارد. فرار مالیاتی در واردات کشور بیانگر آن است که کالاهایی به طور پنهانی و بدون پرداخت مالیات و عوارض گمرکی وارد کشـور می‌شـوند که این مـورد از مصادیق قاچاق کالا به داخل کشور است (مداح و نعمت الهی، ۱۳۹۰).

بـه طـور کلی، روش‌های سـنتی پیش‌بینی فرار مالیاتـی در صورت‌های مالـی دارای برخی مفروضـات محدود کننده مانند خطی بودن، نرمال بودن و مسـتقل بـودن متغیرهای پیش‌بینی کننده یا ورودی‌ها است. نظر به اینکه در ارتباط با داده‌های مالی، تخطی از این مفروضات متصور اسـت، این روش‌های سـنتی در ارتباط با میـزان کارایی و اعتبـار، دارای محدودیت‌های زیادی هسـتند. اما روش‌های مصنوعی و از آن جمله الگوریتم ژنتیک دارای ناسـازگاری و موارد تخطی کمتری در ارتباط با این مفروضات می‌باشـند. یکی از برتری‌های الگوریتم ژنتیک نسـبت به سایر عدم وابسـتگی این الگوریتم بر فرضیه‌های آماری محدودکننده و نرمال بودن توزیع نسـبت‌ها یا برابری واریانس یا کوواریانس ماتریس نسبت‌ها است (پورزمانی، ۱۳۹۲).

در این بین، جدیدترین روش‌هایی که می‌توان به وسیله آنها فرار مالیاتی را پیش‌بینی و کشف نمود، الگوریتم‌های یادگیری ماشـین است. یادگیری ماشین فرآیند استفاده از داده‌ها است که به طور خودکار یک مدل می‌سازد. این ماشین به عنوان ورودی از مجموعه‌ای از ویژگی‌های شناخته شده استفاده می‌کند و به عنوان خروجی چیزی به عنوان مدلی برای پیش‌بینی ارائه می‌دهد.

با توجه به مطالب بالا، هدف از این تحقیق ارائه مدل جدید برای تشـخیص مودیان که دارای فرار مالیاتی هستند، می‌باشد.

یکـی از مهمتریـن چالش‌های نظام مالیاتی ایران این اسـت که مامـوران مالیاتی در هنگام رسـیدگی به پرونـده مودیان با محدودیتهایی مواجه هسـتند که اصلی‌ترین آنهـا، کمبود زمان و عدم تفکیک مودیان دارای فرار یا فاقد فرار مالیاتی (اصطلاحاً کم ریسـک و پر ریسـک) می‌باشد به نحویکه ممکن اسـت برای رسـیدگی یک پرونده که فاقد فرار مالیاتی می‌باشـد زمان بیشتری اختصاص یابد و برعکس برای رسیدگی به پرونده‌ای که دارای فرار مالیاتی می‌باشد زمان کمتری اختصاص یابد و نهایتاً سـازمان مالیاتی را از رسیدن به اهدافش که مهم‌ترین آن‌ها وصول مالیات می‌باشد باز دارد و یا هزینه‌های وصول مالیات را افزایش دهد بنابراین پیش‌بینی و تعیین ریسک

مودیان جهت افزایش راندمان و کارایی ماموران مالیاتی و نیز افزایش وصولی مالیات، برای سازمان امور مالیاتی کشور بسیار پراهمیت و ضروری می‌باشد.

ایده اصلی مقاله از مقایسه نسبت‌های مالی شرکت با نسبت‌های سال گذشته و همچنین برآورد مالیات سال مورد رسیدگی از روی مالیات سال قبل گرفته شده است. الگوریتمی که در این تحقیق بیان شده است مبتنی بر یک الگوریتم ترکیبی می‌باشد که از ترکیب الگوریتم‌های SVM[1] و شبکه بیزین[2] می‌باشد. این الگوریتم‌ها بدین دلیل انتخاب شدند که هر یک به تنهایی دارای مزایا و معایبی می‌باشند لیکن با ادغام آنها، معایب این روشها پوشانده می‌شود و به دقت بالاتری دست می‌یابیم و همینطور برای انتخاب ویژگی نیز از الگوریتم مجموعه راف ترکیب شده با تحلیل سلسله مراتبی استفاده نمودیم که از مجموعه تمام نسبتهای مالی تنها بتوان ویژگی‌های موثرتر و یا همان نسبت‌های مالی موثرتر را شناسایی نمود.

در این مقاله ابتدا مبانی نظری و پیشینه تحقیق ارائه شده است. در بخش بعدی روش تحقیق (شامل فرضیه‌های پژوهش، جامعه آماری، روش گردآوری داده‌ها و روش تجزیه و تحلیل داده‌ها(ارائه خواهد شد؛ در ادامه یافته‌های تحقیق، تجزیه و تحلیل می‌گردد و نهایتاً نتیجه‌گیری و پیشنهادها ارائه خواهد گردید.

۲- ادبیات موضوع

۲-۱- تعریف فرار مالیاتی

ماناسان[3] (۱۹۹۸) هر گونه تلاش غیرقانونی به منظور نپرداختن مالیات را به عنوان فرار مالیاتی می‌خواند. این امر به طرق مختلف مانند ندادن اطلاعات لازم در مورد عواید و منافع مشمول مالیات، اظهار بیشتر از حد میزان هزینه و استفاده از رسیدهای جعلی صورت می‌پذیرد.

اشنایدر[4] و همکارش (۲۰۰۰) فرار مالیاتی را به عنوان کاهش غیر قانونی مالیات پرداختی از طریق عدم گزارش درآمد یا اعلام نمودن نر خهای تفریق بالاتر تعریف می‌نمایند.

هر وضعیتی از عدم تمکین که منجر به عدول از تکالیف مالیاتی، قانونی اشخاص حقیقی و حقوقی گردد، فرار مالیاتی به حساب می‌آید. از جمله، عدم تسلیم اظهارنامه مالیاتی در موعد مقرر قانونی، گزارش کمتر از واقع درآمد، سود یا ثروت، گزارش بیش از واقع کسورات، ارائه اطلاعات مخدوش، تأخیر در پرداخت به موقع مالیات برای سوء استفاده از شرایط تورمی و به طور کلی هر گونه تلاش غیر قانونی برای نپرداختن مالیات، فرار مالیاتی نامیده می‌شود (ضیایی و طهماسبی،۲۰۰۴).

این در حالی است که مقامات مالیاتی در کنار روش‌های پیمایشی از روش حسابرسی مالیاتی نیز برای تعیین بخشی از شکاف مالیاتی ملی استفاده می‌کنند. این کار بر پایهٔ نمونه‌گیری تصادفی

1. Support Vector Machines (SVM)
2. Bayesian Network (BN)
3. Manasan
4. Schneider

و یا کنترل‌شده صورت می‌پذیرد. سپس، مؤدیانی که در نمونه‌گیری انتخاب شده‌اند، توسط مقامات مالیاتی مورد پیمایش قرار می‌گیرند (فوئست و ریدل، ۲۰۰۹). این روش، مبتنی بر این فرض است که درصد معینی از فرار مالیاتی کشف می‌گردد. با این فرض، با استفاده از اطلاعات مربوط به تخلفات کشف‌شده از حسابرسی پرونده‌های مالیاتی-که به روش نمونه‌گیری انتخاب می‌شوند- و با در نظرگرفتن ضریب احتمال کشف، مقدار فرار مالیاتی برآورد می‌شود (عرب‌مازار یزدی، ۱۳۸۰).

ازآنجاکه مقامات مالیاتی الزاماً متوجه تمامی درآمد پنهان نمی‌شوند، با این روش، به جای شکاف مالیاتی واقعی[1]، شکاف مالیاتی قابل‌کنترل[2] به دست می‌آید. اگر فرآیند نمونه‌گیری تصادفی باشد، می‌توان مستقیماً اطلاعات بازیابی‌شده در حسابرسی‌های مالیاتی را برای همهٔ مؤدیان مالیاتی استنتاج کرد.

تکنیک‌های مدیریت ریسک، صرفاً به حسابرسی مالیاتی محدود نمی‌شوند و بخشی از رویکرد جامع و مشارکتی جهت ارتقا تمکین مالیاتی می‌باشد و به طور چشمگیری موجب تغییر دیدگاه سازمان مالیاتی و مودیان خواهد گردید. به دلیل محدود بودن منابع دولت و تعداد بسیار زیاد مودیان (بالاخص مؤسسات کوچک و متوسط[3])، مدیریت ریسک، ابزاری مؤثر در مدیریت تمکین مالیاتی بوده و در یک سازمان مالیاتی مدرن از اولویت خاصی برخوردار می‌باشد. سازمان مالیاتی متناسب با سطح ریسک تمکین هر مؤدی، یک استراتژی تمکین را برنامه‌ریزی و اعمال می‌نماید (نظیر مکاتبه، اخطار، اعتماد به خود اظهاری، حسابرسی محدود و حسابرسی جامع)، شناسایی و انتخاب مؤدی جهت حسابرسی، به دلیل محدودیت منابع و تعداد بسیار زیاد مودیان، نیازمند روشی سیستماتیک و بر مبنای ریسک می‌باشد (سازمان توسعه و همکاری‌های اقتصادی،۲۰۰۴) اینکه سازمان مالیاتی بتواند تک تک مودیان را مورد ممیزی و رسیدگی قرار دهد غیر ممکن است و بررسی مداوم مودیان کم خطر و تمکین دهنده، هدر دادن منابع توسط دولت می‌باشد. به نوعی می‌توان گفت هزینه فرصت از دست رفته برای یک چنین بررسی‌هایی بسیار بالا است بنابراین منابع باید برای مودیانی که پتانسیل ایجاد درآمد و سود بالایی دارند صرف گردد. سازمان مالیاتی باید با هزینه-منفعت نمودن روش‌ها و تکنیک‌های خود، ۱- تمکین مالیاتی توسط مودیان را به منظور شناسایی و پیشگیری از رفتار مجرمانه مدیریت نماید؛ و ۲- بوسیله ارائه خدمات وآموزش مناسب، مودیان را در تمکین و پرداخت تعهدات مالیاتی‌شان کمک نماید. چنین سیستم مدرنی بر پایه خود اظهاری و تمکین داوطلبانه[4] توسط مودیان پایه ریزی شده است و کارکرد ارزیابی ریسک از کارکرد اجرای حسابرسی تفکیک می‌گردد (کاواجا و همکاران،۲۰۱۱)

سازمان مالیاتی باید حسابرسی را برای مودیان با ریسک بالا استفاده نماید و این تمرکز به احتمال زیاد، درآمدهای بالاتری را درپی خواهد داشت و مسلماً عامل بازدارنده قوی برای

1. True Tax Gap
2. Controllable Tax Gap
3. Small and medium-sized enterprises(SMEs)
4. Voluntary compliance

متخلفان خواهد بود. دیگر حسابرس سازمان مالیاتی درگیر پردازش اظهارنامه مالیاتی نمی‌گردد و این موضوع به صورت خودکار انجام می‌شود. سازمان مالیاتی منابع انسانی و مادی خود را نیز همزمان صرف ارقتا خدمات مالیاتی، آموزش و اطلاع رسانی در جهت مدیریت تمکین و ارتقا تمکین داوطلبانه توسط مودیان می‌نماید.

در یک سیستم نمره دهی ریسک از مشخصات و ویژگیهای مؤدی، جهت شناسایی و ارزیابی ریسک تمکین استفاده می‌شود. سیستم، ریسک اطلاعات و نحوه تمکین مؤدی را مورد تجزیه و تحلیل قرار می‌دهد. چنانچه ریسک از آستانه از پیش تعیین شده پایین‌تر باشد، یک ارزیابی خودکار انجام پذیرفته و برگ تشخیص صادر می‌گردد و چنانچه بیش از آستانه باشد، مؤدی برای حسابرسی و بازنگری به صورت دستی انتخاب می‌گردد (انتخاب جهت حسابرسی مبتنی بر ریسک) و پس از حسابرسی و بررسی‌های لازم برگ تشخیص صادر و ابلاغ می‌شود (سازمان مالیاتی ایران، ۱۳۹۰).

۲-۲- مدل‌های فرار مالیاتی

نظریه‌های مطرح‌شده در باب فرار مالیاتی بسیار فراوان‌اند. عده‌ای از محققان با استفاده از عوامل اقتصادی سعی در توضیح این پدیده نموده‌اند که از آن جمله، می‌توان به مطالعهٔ آلینگهام[1] و ساندمو[2] (۱۹۷۲) اشاره کرد. این محققان در مطالعهٔ مذکور موفق به ارائهٔ مدل استاندارد مالیاتی شدند که یکی از مطرح‌ترین مدل‌ها در مورد فرار مالیاتی است. در این مدل، رفتار یک مؤدی مالیاتی ریسک‌گریز در مورد تصمیم به تمکین و یا عدم تمکین مالیاتی، در لحظهٔ تکمیل اظهارنامه و با توجه به عقلانیت اقتصادی، موردبررسی قرار می‌گیرد. این دسته از نظریات که تنها عوامل اقتصادی را موردتوجه قرار می‌دهند، با توجه به چندبعدی بودن فرار مالیاتی، مورد انتقادات شدیدی قرار گرفتند. این انتقادات باعث تعدیل نظریهٔ آلینگهام و ساندمو گشت؛ به‌گونه‌ای که عده‌ای تلاش کردند که عوامل غیراقتصادی را نیز به مدل پیشین وارد کنند. دراین‌بین، عده‌ای با نگاهی مبتکرانه به موضوع، سعی در ارائهٔ نظریه‌هایی جدید با استفاده از رهیافت‌های علم اقتصاد رفتاری نمودند. بدین‌ترتیب، می‌توان دو جریان عمده را در تبیین رفتار فرار مالیاتی تمییز داد: مدل‌های نئوکلاسیکی، و تحلیل‌های علم اقتصاد رفتاری از فرار مالیاتی. همچنین، گروه دوم را می‌توان در دو دستهٔ کلی جای داد: مدل‌های مبتنی بر تئوری مطلوبیت غیر انتظاری، و مدل‌های مبتنی بر تئوری تعاملات اجتماعی. (خان جان،۱۳۹۳)

۲-۳- عوامل مؤثر بر فرار مالیاتی

دانستن اینکه چرا شکاف مالیاتی ایجاد می‌شود، بسیار مهم است؛ چراکه بدون درک عوامل تعیین کنندهٔ شکاف مالیاتی، ارائه و ارزیابی سیاستها در جهت رفع و یا کاهش آن مشکل

1. Allingham
2. Sandmo

می‌شـود. مطالعات نظری، تجربی، و آزمایشـگاهی پیرامون فرار مالیاتی (یا تمکین مالیاتی) نشان می‌دهنـد که افراد عموماً نسـبت به عواملـی نظیر افزایش نرخ مالیات، افزایش نرخ حسابرسـی، حسابرسـی کاراتر، تکرار حسابرسـی، انتخاب اسـتراتژیک یا هدفمند حسابرسی، افشای عمومی نتایج حسابرسـی، افزایش نرخ جریمه، انتشـار اطلاعات بین سازمانهای دولتی حسابرس، افزایش پاداش برای تمکین مالیاتی، اتصال بیشـتر مالیات و خدمات عمومی، افزایش مشـارکت مؤدیان در تصمیمـات گروهی، عفـو مالیاتی، افزایش پیچیدگی و نااطمینانـی در قوانین، و خدمات بهتر برای مؤدیان، واکنشـی قابل پیش‌بینی دارند. در ادامه، برخی از عوامل مؤثر بر فرار مالیاتی که از مطالعات مختلف استخراج شده، موردبحث قرار می‌گیرد. (نوربخش، ۱۳۹۰)

۲-۳-۱- ضعف در شرایط سیاسی و اجتماعی

یکی از دلایل فعالیت بنگاههادراقتصادغیررسـمی ضعف در شـرایط سیاسی و اجتماعی نظیر بوروکراسی، مقررات پیچیده، فساد مالی و نظام قضائی ضعیف است. عده‌ای از این پدیده به‌عنوان سیاسی سازی زندگی اقتصادی یاد می‌کنند. (فوئست[1] و ریدل[2]، ۲۰۱۹).

جانسـون و همکاران[3] (۲۰۱۶) در این باره بیان می‌کنند: «این نوع ابزار نظارتی شـامل اعمال نیروهـای تنظیمی بر بنگاههای خصوصی، توانایی سـامان دادن و محدود کردن ورود، نظارت بر اسـتفاده از زمین و ملکی که مشـاغل خصوصی در آنها مستقرند، بسـتن مالیات بر مشاغل، حق وارسی و تفتیش بنگاههاومتوقف کردن فعالیت آنها درهنگام تخطی از قوانین است طبق معمول، بسـیاری از سیاسـتمداران از این ابزارها برای پیگیری اهداف خود، نظیر حفظ سـطح اشـتغال در بنگاههای ویژه، حمایت دوسـتانهٔ سیاسـی و مجازات خصمانهٔ سیاسی بنگاههاویارانه دادن به متحدان خویش اسـتفاده می‌کنند. همچنین، سیاسـتمداران می‌توانند از این حق برای ثروتمند شدن خود، از طریق پیشنهاد خلاصی بنگاهها از قوانین، در عوض رشوه استفاده نمایند».

۲-۳-۲- بار مالیاتی[4]

بر اسـاس بحث بسـیاری از محققان، بار مالیاتی یکی از مهمتریـن عوامل تعیین کنندهٔ فرار مالیاتی اسـت. به نظر می‌رسـد که افزایش در نرخ مالیات مؤثر، انگیزهٔ بنـگاه را برای فعالیت در اقتصاد غیررسمی افزایش می‌دهد و بدین ترتیب، بر میزان فرار مالیاتی افزوده می‌شود. (جانسون و همکاران ۲۰۱۶)

۲-۳-۳- نرخ حسابرسی

احتمال حسابرسـی و کشف تخلف از دیگر عواملی است که بر اندازهٔ اقتصاد غیررسمی و فرار مالیاتی اثرگذار اسـت. چنانچه احتمال شناسـایی و دسـتگیری در یک کشور بیشتر باشد، تعداد

1. Fuest
2. Riedel
3. Johnson et al.

۴. بار مالیاتی میزان وجوهی است که توسط افراد پرداخت می‌شود. این مقدار به دو قسمت بار پولی و بار واقعی قابل تقسیم است. بار پولی میزان وجوهی است که توسط مشمول مالیات قابل پرداخت است، اما بار واقعی میزان وجوهی است که توسط پرداخت کننده نهایی پرداخت می‌شود. (جعفری صمیمی،۱۳۸۷)

شـاغلین در اقتصاد غیررسـمی کمتر اسـت. مطالعاتی نظیر مطالعهٔ اندرونی و همکاران (۲۰۱۶) چنین ارتباطی را از تحقیق خود نتیجه‌گیری کرده‌اند.

۲-۳-۴- جریمه پس از کشف تخلف

همچنیـن، مجازات پس از کشـف تخلف از دیگر متغیرهای تأثیرگذار بر فرار مالیاتی اسـت. اندرونی و همکاران (۲۰۱۵) ارتباط بین مشـارکت در اقتصاد غیررسـمی و مجازات پس از کشف را بررسـی کرده و به این نتیجه رسـیدند که ارتباطی منفی بین این دو برقرار اسـت. دراین‌باره، ریچوپان[1] هم به نتایج مشـابهی دسـت یافته است. او با در نظر گرفتن الگوی آلینگهام و ساندمو، اثـر عوامـل مختلف را بر تصمیم مؤدی و درآمد مورد انتظار وی به بحث گذاشـته و نتیجه‌گیری می‌کند که اولاً، هر چه نرخ جریمهٔ مالیاتی بیشتر باشد، مقدار درآمد اظهارشده هم بیشتر خواهد بود و ثانیاً، با افزایش احتمال کشـف تخلف، احتمال اظهار درآمد واقعی نیز افزایش می‌یابد (فرار مالیاتی کاهش می‌یابد).

۲-۴- داده‌کاوی

تعاریف متنوعی از داده‌کاوی[2] در مراجع مختلف و توسـط افراد مختلف ارائه شـده اسـت. از جمله:

فرآیند کشف الگوهای مفید از داده‌ها را داده‌کاوی می‌گویند.

فرآینـد انتخـاب، کاوش و مدل بندی داده‌های حجیم، جهت کشـف روابط نهفته با هدف به دست آوردن نتایج واضح و مفید، برای مالک پایگاه داده‌ها را، داده‌کاوی گویند.

اما تعریفی که در اکثر مراجع به اشتراک ذکر شده عبارت است از "استخراج اطلاعات و دانش و کشف الگوهای پنهان از یک پایگاه داده‌های بسیار بزرگ و پیچیده ". داده‌کاوی یک متدولوژی بسیار قوی و با پتانسیل بالا می‌باشد که به سازمان‌ها کمک می‌کند که بر روی مهمترین اطلاعات از مخزن داده‌های خود تمرکز نمایند.

داده‌کاوی فرایندی اسـت کـه از ابزارهای تحلیلی گوناگونی برای کشـف الگوها و روابط بین داده‌ها استفاده می‌کند که ممکن است برای اعتبار بخشیدن به پیش‌بینی استفاده شود.

۲-۴-۱- مراحل داده‌کاوی

داده‌کاوی شامل مراحل مختلفی می‌باشدکه عبارتنداز:

۱. تعیین اطلاعات گذشته.

۲. تمیـز کـردن داده‌ها و پـردازش اولیه. دراین مرحلـه خطاهای داده‌ها تصحیح می‌شـوند و داده‌های اشتباه جایگزین می‌شوند. این مرحله ممکن است تا ۶۰ درصد از زمان را در برگیرد.

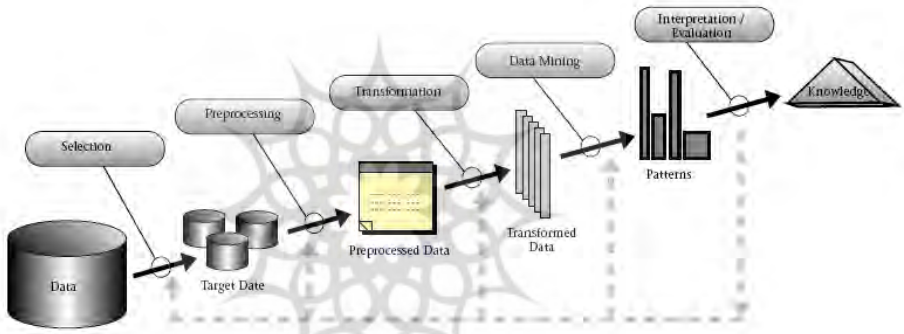
۳. یکپارچه سـازی داده‌ها. معمولاً داده‌ها از منابع متفاوتی جمع‌آوری می‌شوند باید به صورتی درآیند که یک مخزن داده‌های مناسـب ایجاد شود تا بتوان عملیات داده‌کاوی را بهتر انجام

1. Richupan
2. Data Mining

داد.

۴. انتخاب مجموعه داده‌های هدف.
۵. یافتن ویژگیهای مورداستفاده و تعیین ویژگی‌های جدید.
۶. نمایش داده‌ها به صورتیکه بتوان برای داده‌کاوی استفاده نمود.
۷. انتخاب عملیات داده‌کاوی (دسته بندی، خوشه بندی، پیش‌بینی وغیره).
۸. انتخاب روش داده‌کاوی (درخت تصمیم و نظایر آن).
۹. داده‌کاوی و جستجو برای یافتن الگوی مناسب.
۱۰. ارزیابی و تحلیل الگوی به دست آمده و حذف الگوهای نامناسب.
۱۱. تفسیر نتایج داده‌ها و استنتاج از اطلاعات با ارزش.

در شکل زیر می‌توان مراحل داده‌کاوی را به اختصار نشان داد.



۲-۴-۲-یادگیری ماشین

یادگیری ماشین[1] یعنی چگونه می‌توان برنامه‌ای نوشت که از طریق تجربه یادگیری کرده و عملکرد خود را بهتر کند. یادگیری ممکن است باعث تغییر در ساختار برنامه و یا داده‌ها شود. یادگیری ماشین زمینه نسبتاً جدیدی در علوم کامپیوتر است که در حال حاضر دوران رشد و تکامل خود را می‌گذراند. یادگیری ماشین یک زمینه تحقیقاتی بسیار فعال در علوم کامپیوتر است. علوم مختلفی در ارتباط با یادگیری ماشین در ارتباط هستند از جمله: هوش مصنوعی، تئوری اطلاعات، آمار و احتمالات، تئوری کنترل و ... . از دلایل استفاده از یادگیری ماشین برای حل مسائل می‌توان به موارد زیر اشاره کرد:

- بعضی کارها را بدرستی نمی‌توان توصیف نمود؛ در صورتی‌که ممکن است آن‌ها را بتوان بصورت مثال‌های (ورودی/خروجی) معین کرد.
- ممکن است در خیل عظیمی از داده، اطلاعات مهمی نهفته باشد که بشر قادر به تشخیص آن نباشد (داده‌کاوی).

1.Machine Learning

- ممکن است در هنگام طراحی یک سیستم تمامی ویژگی‌های آن شناخته شده نباشد در حالی‌که ماشین می‌تواند حین کار آن‌ها را یاد بگیرد.
- ممکن است محیط در طول زمان تغییر کند. ماشین می‌تواند با یادگیری این تغییرات، خود را با آن‌ها وفق دهد.
- در عمل، نوشتن برنامه‌ای برای تشخیص یک صورت می‌تواند کار مشکلی باشد. زیرا تعریف دقیقی برای آن وجود ندارد و حتی در صورت وجود داشتن نوشتن برنامه‌ای بر اساس آن کار سختی است. در نتیجه بجای نوشتن یک برنامه با دست، تهیهٔ مقدار زیادی نمونه صحیح و اعمال آن به یک الگوریتم یادگیری ماشین برنامه‌ای تولید می‌شود که کار مورد نظر را انجام دهد. این برنامه بسیار با آنچه که با دست نوشته شود متفاوت خواهد بود. اگر این برنامه درست تهیه شده باشد می‌تواند برای نمونه هائی که تاکنون ندیده است نیز خروجی مورد نظر را تولید کند.
- در سال‌های اخیر پیشرفت‌های زیادی در الگوریتم‌ها و تئوری‌های مربوطه بوجود آمده و زمینه‌های تحقیقاتی جدید زیادی پدید آمده‌اند.
- داده‌های آزمایشی زیادی بصورت برخط[1] بوجود آمده‌اند.
- کامپیوترها قدرت محاسباتی زیادی بدست آورده‌اند.
- جنبه‌های عملی با کاربردهای صنعتی بوجود آمده‌اند. (در زمینه پردازش گفتار برنامه‌های مبتنی بر یادگیری از همه روش‌های دیگر پیشی گرفته‌اند).

۲-۴-۳- شبکه بیزین

دسته‌بندی‌کننده بیزین در یادگیری ماشین به گروهی از دسته‌بندی‌کننده‌های ساده بر پایه احتمالات گفته می‌شود که با متغیرهای تصادفی مستقل ساده مفروض میان حالت‌های مختلف و براساس قضیه بیز کاربردی است. به‌طور ساده روش بیز روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده‌است.

براساس ویژگی‌های ذاتی احتمال (به ویژه اشتراک احتمال) دسته‌بندی‌کننده بیز ساده[2] با دریافت تمرین اولیه نتایج خوبی ارائه خواهد کرد. شیوه یادگیری در روش نایو بیز از نوع یادگیری با ناظر[3] است.

شبکه بیزین را می‌توان یک مدل برمبنای احتمال شرطی در نظر گرفت. فرض کنید $X=(x1,\ldots,xn)$ برداری از n ویژگی را بیان کند که به صورت متغیرهای مستقل هستند. به این ترتیب می‌توان احتمال رخداد Ck یعنی $p(Ck|x1,\ldots,xn)$ را به عنوان یکی از حالت‌های کلاس رخدادهای مختلف به ازاء k های متفاوت، به شکل زیر نمایش داد.

1. Online
2. Naive Bayes classifier
3. Supervised learning

$$p(C_k \mid X) = \frac{p(C_k)\, p(X \mid C_k)}{p(X)}$$

همانطور که دیده می‌شود رابطه بالا همان قضیه بیز است.

۲-۵- پیشینه پژوهش

در مطالعاتی که در سطح خرد، به بررسی فرار مالیاتی بین مؤدیان می‌پردازند، عوامل به چشم می‌خورد که سن مؤدیان یکی از آنهااست. مطالعات نشان می‌دهد که مؤدیان پیرتر نسبت به مؤدیان جوان، اطاعت پذیری بیشتری نسبت به قوانین دارند. یکی از این مطالعات، کار تیتل (۲۰۱۸) است. وی ارتباط بین سن و شکاف مالیاتی را به عنوان نتیجهٔ متغیرهای دوران زندگی و تفاوتهای نسلی توضیح می‌دهد. مؤدیان جوانترنسبت به مجازات ریسک‌پذیرتر و کم‌توجه‌ترهستندکه این رفتار، تفاوت‌های اجتماعی و جمعیت شناختی مربوط به دوره‌ای که در آن رشد یافته‌اند را منعکس می‌سازد (تفاوت نسلی).

جنسیت مؤدیان هم از عواملی است که به عنوان یک عامل اثرگذار بر فرار مالیاتی، موردتوجه محققان قرار گرفته است. وُگل (۲۰۱۷)، و ماسُن و کالوین (۲۰۱۶) نشان می‌دهند که سطح اطاعت مؤدیان زن از مؤدیان مرد بیشتر است. البته، در مطالعات جدیدتر نظیر پژوهش جکسُن و میلیرن (۲۰۱۶)، بحث می‌شود که آزادی بیشتر نسلهای جدید زنان، منجر به کاهش این شکاف شده است. بااین‌حال، مطالعات زیاد دیگری (نظیر مطالعهٔ کُلینز و همکاران، ۲۰۱۶) مبین این نکته هستند که شکاف بین سطح اطاعت زنان و مردان هنوز به قوت خود باقی است.

هاشم‌زاده و همکاران (۲۰۱۰) در بررسی تقلب مالیاتی شرکتها و حسابرسی مطلوب، یکی از موارد تقلب را ادعای بیش از حد برای تخفیف مالیات بر ارزش افزوده کالاهای صادراتی می‌دانند و نشان دادند که وجود تقلب نمی‌تواند تصمیمات خروجی واقعی شرکت و سیاست‌های مالیاتی دولت را تحت تأثیر قرار دهد.

ریچاردسون (۲۰۱۰) در پژوهشی تحت عنوان عوامل تعیین کننده فرار مالیاتی، علل فرار مالیاتی را برای ۴۵ کشور منتخب جهان به صورت مقطعی بررسی نموده است. مدل ارائه شده در این تحقیق به این صورت می‌باشد که فرار مالیاتی به عنوان متغیر وابسته و سن، جنسیت، تحصیلات عمومی، سطح درآمد، منبع درآمد کشاورزی (درصد اشتغال در بخش کشاورزی)، منبع درآمد خدماتی (درصد اشتغال در بخش خدمات)، نرخ نهایی مالیات، پیچیدگی سیستم مالیاتی، انصاف مالیاتی و روحیه مالیاتی به عنوان متغیرهای توضیحی لحاظ شده است. نتایج مطالعه حکایت از آن دارد که عوامل غیراقتصادی از جمله پیچیدگی، بیشترین اثر را بر فرار مالیاتی دارند. همچنین نتایج نشان می‌دهد که سطح پایین‌تر پیچیدگی و سطح بالاتر تحصیلات عمومی، انصاف و روحیه مالیاتی موجب سطح پایین‌تری از فرار مالیاتی خواهد شد.

عامل دیگری که بر فرار مالیاتی مؤدیان در سطح خرد تأثیرگذار است، سطح درآمد مؤدیان است. ماسُن و لَوری (۲۰۱۷) و نیز، ویت و وودبری (۲۰۱۷) نشان دادند عموماً مؤدیانی که در سطوح متوسط درآمد قرار دارند، به نسبت مؤدیانی که در سطوح بالا و پائین درآمدی هستند، از اطاعت پذیری بیشتری نسبت به قوانین برخوردارند. مطالعهٔ ریچاردسون و سویر (۲۰۱۱) بر پیچیدگی نتایج تجربی در این زمینه دلالت دارد.

فرانزونی (۲۰۱۴) با بررسی و مرور نظری پژوهش‌های انجام گرفته در زمینه فرار مالیاتی و نیز تمکین مالیاتی در شرکتهای ایتالیا عوامل مختلف تاثیرگذار بر عدم تمکین و فرار مالیاتی را شناسایی نموده و راهکارهای پیشگیری از آن را نیز مورد بررسی و آزمون قرار داده است. بر این اساس و باتوجه به بررسی‌های انجام گرفته محرک‌ها و رویه‌های قانونی دولت سیاست اجرای مالیات را تعیین و بخش عمده از فرار مالیاتی وعدم تمکین مالیاتی برگرفته از خلاهای قانونی بوده است.

تحقیق اشمولر (۲۰۱۵) نشان می‌دهد در کشوری که بخش بزرگی از نیروی کار آن در بخش کشاورزی اشتغال دارند و بخش کوچکی در تجارت مشغول به کارند، مالیات ستانی از درآمد و سود ناموفق خواهد بود. مطالعهٔ پیمایشی والشزکی (۲۰۱۶) از فراریان و پرداخت کنندگان مالیات نشان می‌دهد که کسانی که درآمدشان را از کشاورزی، تجارت‌های مستقل و خویش فرمایی استحصال می‌کنند، فرصت بیشتری برای فرار از مالیات بر درآمد دارند. کمترین فرصت برای فرار مالیات به آن دسته از افراد تعلق دارد که حقوق بگیرند؛ چراکه مالیات به صورت ماهیانه، از حقوقشان کسر می‌شود.

پاپا، سجادی، و ولا[1] (۲۰۱۴)، فرار مالیاتی و فساد را به مدل کینزی جدید با اصطکاک‌های جستجو و تطابق[2] وارد کرده‌اند تا اثر تحکیم مالی مبتنی بر مخارج و مالیات را موردبازنگری قرار دهند. در این مدل، کاهش مخارج عمومی تقاضا برای کالاها در بخش رسمی و غیررسمی را کاهش می‌دهد و لذا، فرار مالیاتی کاهش می‌یابد. افزایش مالیات عوامل را به کار و تولید بیشتر در بخش غیررسمی وامی‌دارد. ازآنجاکه بهره‌وری در بخش غیررسمی کمتر از بخش رسمی است، این سیاست زیان رفاهی و تولید بیشتری به دنبال دارد.

نویدی (۱۳۸۴) در مطالعه‌ای مدل نظری بازی‌های ریاضی در ارتباط با گریز مالیاتی و فساد در سیستم مالیاتی را موردبررسی قرار داده است. در این مدل مرکز کنترل مالیاتی می‌تواند بازرسان مالیاتی را از دو دسته صادق یا غیرصادق به کار گیرد. امتیاز و فرق این دو دسته در اختلاف قیمت بازرسی آن‌هاست. بازرس صادق از قیمت و هزینه بیشتری نسبت به بازرس نوع دیگر برخوردار است. ازآنجاکه اثبات وجود تبانی در کار بازرس ساده نیست، امکان تبانی بازرسان غیرصادق با مالیات‌دهندگان در نظر گرفته شده است. در این مطالعه، عکس‌العمل متقابل مالیات‌دهندگان و بازرسان در مقابل سه استراتژی دولت برای به‌کارگیری بازرسان موردبررسی قرار گرفته است.

1.Pappa, Sajedi & Vella
2.Search and Matching Frictions

نصـر اصفهانی و همـکاران (۱۳۹۱) عوامل مؤثر بر فرار مؤدیـان از پرداخت مالیات با الگوی پستل موردبررسی قرار گرفته است. نویسندگان به دنبال یافتن پاسخ برای این پرسش بوده‌اند که عوامل سیاسـی، اقتصادی، اجتماعی، تکنولوژیکی، قانونی و محیطی -که از عوامل تشکیل‌دهنده محیـط بیرونـی «اجتماعی» و بر گرفته از مدل پسـتل اسـت- تـا چه اندازه بر فـرار مؤدیان از پرداخت مالیات تأثیر دارد. این پژوهش میدانی، با استفاده از پرسش‌نامه نشان می‌دهد که از نقطه نظرمؤدیان مشاغل، عوامل سیاسی و اقتصادی بیشترین تأثیر و عامل محیطی کمترین تأثیر را بر فرار مالیاتی دارند. از نظر کارشناسان کادر تشخیص، عوامل سیاسی، اقتصادی و قانونی بیشترین تأثیر و عامل محیطی کمترین تأثیر را بر فرار مالیاتی داشته‌اند.

دستگیر و غریبی (۱۳۹۴) در کاربست روش‌های داده‌کاوی به منظور ارتقاء عملکرد تشخیص فرار مالیاتی داده‌ها را به طور تصادفی به سـه دسته آموزش، اعتبارسنجی و آزمون تقسیم شدند. نتایج پژوهش نشـان داد که روش‌هـای داده‌کاوی مبتنی بر قواعد وابسـتگی با ایجاد دو مدل یا درصد صحت ۹۱٪ بر روی داده‌های آموزش، با درصد صحت ۸۸٪ بر روی داده‌های اعتبار سـنجی و با درصد صحت ۸۶٪ بر روی داده‌های آزمون توانسته است موفق به تشخیص فرار مالیاتی گردد.

عبدلی و همکاران (۱۳۹۴) بر اهمیت مسـائل اجرایی مالیات سـتانی پرداخته‌اند و با توجه به وجـود اطلاعات نامتقارن در زمینـه مالیات بر درآمد، برای جلوگیری از فـرار مالیاتی (و افزایش درآمد مالیاتی دولت بدون گسترش پایه‌ها و افزایش نرخ‌های مالیاتی) مکانیسم حسابرسی خاصی معرفی و با استفاده از نظریه بازی‌ها و روش میدانی مورد تحلیل نظری و تجربی قرار گرفته است. نتایج بیانگر آن اسـت که اگر گزارش مالیاتی گروهی از مودیان که دارای ویژگیهای شـبیه به هم می‌باشـند، با هم مقایسـه گردیده و آنهایی که از متوسـط گزارش گروه کمتر گزارش داده‌اند با احتمال بیشتر حسابرسی می‌شوند منجر به کاهش فرار مالیاتی، افزایش درآمدهای مالیاتی دولت (در شرایط خاص)، کاهش هزینه‌های حسابرسی و ثبات درآمدهای مالیاتی دولت خواهد شد.

رحیمی کیا و همکاران (۱۳۹۴) فرار مالیاتی را با اسـتفاده از سیسـتم هوشمند ترکیبی مورد بررسی قرار دادند که نتایج نشان داد استفاده از شبکه عصبی دارای دقت بالاتری بوده و این تفاوت از لحاظ آماری معنادار می‌باشد. در شبکه عصبی به ترتیب در صنعت مواد غذایی و نساجی دقت کلی ۷۸,۸۳٪ و ۸۴,۸۵٪ دقت تشـخیص شرکت‌های فراری ۳۱,۸۰٪ و ۳۴,۸۴٪ و دقت تشخیص شرکت‌های سالم ۲۰,۸۷٪ و ۳۶,۸۵٪ می‌باشد.

ناصرآبادی و همکاران (۱۳۹۸) با اسـتفاده از شـبکه‌های مصنوعی، تعداد ۱۶ عامل مربوط به عملکرد سـال ۱۳۹۱ تعداد ۴۰۰ شـرکت را اسـتخراج نمودند. نتایج حاصل از این تحقیق نشان می‌داد شـبکه‌های عصبی مصنوعی با کارایی ۸۲/۵٪ در مقایسـه با رگرسیون خطی چند متغیره، فرار مالیاتی را بهتر کشف می‌کنند.

۲-۶- فرضیه پژوهش:

مدل پیشـنهادی (ترکیبی از الگوریتم SVM و شـبکه بیزین) دقت بیشتری برای پیش‌بینی

فرار مالیاتی نسبت به هر یک از الگوریتم SVM و شبکه بیزین (به تنهایی) دارد.

۳- روش تحقیق

هدف این پژوهش رسیدن به مکانیزمی برای تشخیص مودیان دارای فرار مالیاتی می‌باشد. جامعه مورد مطالعه در این پژوهش عبارت است از شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران و نمونه مورد استفاده متشکل از دو گروه شرکت‌های دارای فرار مالیاتی و شرکت‌های سالم خواهد بود. طبق مفاد ماده ۱۹۴ ق.م.م. در صورتیکه درآمد مشمول مالیات مشخصه قطعی با رقم اظهار شده از طرف مؤدی بیش از ۱۵٪ اختلاف داشته باشدعلاوه بر تعلق جرائم مقرر مربوط که قابل بخشودن نیز نخواهد بود تا سه سال بعد از ابلاغ مالیات مشخصه قطعی از هرگونه تسهیلات و بخشودگی‌های مقرر در قانون مالیاتها نیز محروم خواهند شد. در این پژوهش براساس همین ماده قانون مالیاتهای مستقیم، معیار ۱۵٪ فوق، حدنصاب شناسایی شرکتهای دارای فرار مالیاتی از فاقد فرار مالیاتی قرار می‌گیرد و از رابطه زیر استفاده می‌شود:

$$\%\Delta=100\times(ACCIN - TAXIN)$$

Δ٪: درصد اختلاف درآمد مشمول مالیات ابرازی و درآمد مشمول مالیات مشخصه قطعی مؤدی

ACCIN: درآمد مشمول مالیات ابرازی مؤدی

TAXIN: درآمد مشمول مالیات مشخصه قطعی مؤدی

بدین ترتیب بوسیله معیارهای فوق ابتدا لیستی از شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران که بین سال‌های ۱۳۹۳ تا ۱۳۹۵ مرتکب فرار مالیاتی در صورتهای مالی شده‌اند تهیه می‌شود.

در این تحقیق سعی شده است که یک روش ترکیبی استفاده شود که دارای مزایای هر یک از روش‌ها باشد و بتواند معایب این روش‌ها را با ترکیب کردن آن‌ها پوشش داد. در این تحقیق از روش‌های SVM و بیزین استفاده شده است و یک راهکار برای ترکیب این دو روش ارائه شده است. شبکه بیزین در موارد تحلیل عددی نتایج مناسبی را بدست می‌آورد ولی برای ارتباط دادن نسبت‌های مالی مختلف که در این تحقیق استفاده شده است مناسب نیست و نمی‌تواند به شکل مناسب ارتباط بین ویژگی‌ها را برای پیش‌بینی استفاده نماید از طرفی SVM دقیقاً برعکس می‌تواند ارتباط بین مؤلفه‌ها را به شکل بسیار مناسبی بیابد و با دقت بالایی عملکرد مناسبی را نشان دهد از اینرو در این تحقیق از یک راهکار ترکیبی استفاده شده است که از مزایای هر دو الگوریتم استفاده شود. در این تحقیق برای کاهش ویژگی از مجموعه راف و AHP استفاده شده است.

۳-۱- قلمروپژوهش

قلمرو موضوعی: این پژوهش به دنبال ارائه مدلی ترکیبی بر مبنای الگوریتم‌های یادگیری

ماشـــین (الگوریتم SVM) و شبکه بیزین به منظور پیش‌بینی فرارهای مالیاتی مؤدی با استفاده ازنسبتهای مالی است.

قلمرو مکانی: به منظور تشخیص شرکتهای دارای فرار مالیاتی از شرکتهای فاقد فرار مالیاتی می‌بایســت مالیات شــرکت مورد نظر به مرحله تشخیص رسیده باشــد. از طرف دیگر شرکتهای پذیرفته شده در سازمان بورس اوراق بهادار (به استثناء بانکها، بیمه‌ها، شرکت‌های سرمایه‌گذاری و شــرکتهای میانســالی) به عنوان جامعه تحقیق مورد بررسی قرار گرفت. نظر به اینکه روند رشد (مثبت یا منفی) فرار مالیاتی نیز توسط پژوهشگر محاسبه خواهد شد لذا شرکتهایی که در هر سه ســال ۱۳۹۳ لغایت ۱۳۹۵ در ســایت مربوطه دارای اطلاعات مالی باشند به عنوان جامعه تحقیق مورد بررسی قرار گرفت که تعداد اعضا جامعه ۱۱۸۶ سال/شرکت می‌باشد.

قلمرو زمانی: قلمرو زمانی این پژوهش دوره ســه ســاله حدفاصل ســال‌های ۱۳۹۳ تا ۱۳۹۵ می‌باشد.

۳-۲- روش‌های گرد آوری داده‌ها

در این تحقیق برای جمع‌آوری داده‌ها از روش کتابخانه‌ای و میدانی اســتفاده شــده اســت. اطلاعــات مربوط به مبانی نظری وتئوریــک پژوهش ازکتب ومقالات فارســی ولاتین جمع‌آوری گردیده است. منبع جمع‌آوری دادههای موردنیازاین پژوهش، صورت‌های مالی شرکتهابوده است. داده‌هــای مالی مورد نیــاز این پژوهش ازطریق بانک اطلاعات رایانــه‌ای ره آوردنوین ومراجعه به سامانه جامع اطلاع رسانی ناشران http://www.codal.ir/ و http://www.rdis.ir/ و سیستم اشخاص حقوقی سازمان مالیاتی (Itcms) جمع‌آوری شده است.

داده‌های جمع‌آوری شــده پس ازطبقه‌بندی لازم براســاس متغیرهای موردبررسی، بااستفاده ازنرم‌افزاراکســل وارد رایانه شده اســت. تجزیه وتحلیل نهایی به کمک نرم‌افزارهای مختلف انجام شــده اســت. ازجمله مهمترین مواردی که لازم اســت هنگام جمع‌آوری دادهها به آن توجه کرد، روایــی ابزارهای گردآوری دادهها اســت. منظورازروایی ابزارهای گردآوری دادهها این اســت که ابزارهــا بتواننــد واقعیتهارا به خوبی نشــان دهند ازآنجا که ابزارگــردآوری داده‌ها در این تحقیق، بانک‌های اطلاعاتی تهیه شده توسط سازمان بورس اوراق بهادارتهران و یا کتابخانه سازمان بورس اوراق بهادارتهران است؛ لذا می‌توان به روایی ابزارهای گردآوری دادهها اعتماد کرد.

۳-۳- الگوریتم SVM

ماشــین بردار پشــتیبانی[1] یکی از روش‌های یادگیری بانظارت است که از آن برای طبقه‌بندی و رگرســیون اســتفاده می‌کنند. این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نســبت به روش‌های قدیمی‌تر برای طبقه‌بندی نشــان داده‌است. مبنای کاری دسته‌بندی کننده SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم

1. Support vector machines - SVM

خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. حل معادله پیدا کردن خط بهینه برای داده‌ها به وسیله روش‌های QP که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد. قبل از تقسیمِ خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله تابعِ phi به فضای با ابعاد خیلی بالاترمی‌بریم. برای اینکه بتوانیم مسئله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم از قضیه دوگانی لاگرانژ برای تبدیلِ مسئله مینیمم‌سازی مورد نظر به فرم دوگانی آن که در آن به جای تابع پیچیده phi که ما را به فضایی با ابعاد بالا می‌برد، تابعِ ساده‌تری به نامِ تابع هسته که ضرب برداری تابع phi است ظاهر می‌شود استفاده می‌کنیم. از توابع هسته مختلفی از جمله هسته‌های نمایی، چندجمله‌ای و سیگموید می‌توان استفاده نمود.

الگوریتم SVM، جز الگوریتم‌های تشخیص الگو دسته‌بندی می‌شود. از الگوریتم SVM، در هر جایی که نیاز به تشخیص الگو یا دسته‌بندی اشیا در کلاس‌های خاص باشد می‌توان استفاده کرد.

۳-۴- ساخت شبکه بیزین

شبکه‌های بیزین متعلق به ساختار دیگری از مدل‌های گرافیکی به نام گراف‌های غیرمدور جهت دار هستند که در زمینه‌های آماری، یادگیری ماشینی و هوش مصنوعی بسیار مشهور هستند. شبکه‌های بیزین نمایش و محاسبات موثری از توزیع احتمالاتی مشترک به روی یک سری متغیر تصادفی را فراهم می‌آورند. به علاوه شبکه‌های بیزین شدت ارتباط بین متغیرها را به صورت کمی مدل می‌کنند که اجازه می‌دهند با دسترسی به اطلاعات جدید، اعتقاد شرطی در مورد آنها به صورت خودکار به روز رسانی شود.

۳-۵- انتخاب ویژگی‌های مؤثر با استفاده از مجموعه راف و روش AHP

در این تحقیق از روش مجموعه راف ادغام شده با روش AHP که یک روش تصمیم‌گیری می‌باشد وزن دهی پارامترها صورت گرفته و پارامترهای مؤثر شناسایی می‌شوند.

بسیاری از مفاهیم و تئوری‌های عدم قطعیت نظیر مجموعه‌های فازی، سیستم‌های خاکستری و مجموعه‌های راف، در گذشته معرفی شده و در سال‌های اخیر ابزارهای ریاضی مبتنی بر آن‌ها با سرعت بالایی توسعه یافته‌اند. هریک از این رویکردها، مفاهیم خاص خود را داشته و دارای ویژگی‌های منحصر به خود می‌باشد. تئوری خاکستری به کنترل سیستم‌ها در شرایط کمبود داده‌ها و اطلاعات ناکامل پرداخته و تئوری راف، تقریب و استدلال درباره داده‌ها را بدنبال دارد. داده‌هایی که از دنیای واقعی اخذ می‌گردند معمولاً شامل تمامی انواع نویزها بوده و عدم قطعیت بسیار و اطلاعات غیر کامل فراوانی به همراه دارند. روش‌های سنتی برخورد با این عدم قطعیت نظیر تئوری فازی، تئوری گواه، تئوری احتمالات و نظایر آن، به اطلاعات اضافی مانند توزیع احتمال و تابع عضویت نیازمند هستند. به بیان دیگر، کار با این سیستم‌ها به دلیل حجم بالایی از

داده‌ها مشـــکل است، از این رو به کارگیری سایر تئوری‌ها نظیر تئوری مجموعه‌های راف می‌تواند در این راه کمک کننده باشد.

مجموعه راف ابزاری قابل اســـتفاده از شـــرایط ابهام و عدم قطعیت اســـت که اولین بار توسط پاولاک[1] (۱۹۸۲) ارائه شـــد. این تئوری راف در زمینه‌های مختلفی مورد استفاده قرار می‌گیرد از جمله تجزیه و تحلیل تصمیم‌گیری، سیستم‌های پشتیبان تصمیم. بعد از آقای پاولاک، سه محقق دیگر بنام‌هایژای[2]، خو[3]و ژانگ[4]در سال ۲۰۰۸ اعداد راف را ارائه کردند.

روش AHP (فرایند تحلیل سلسله مراتبی) از روشهای پر کاربرد در تصمیم‌گیری چند معیاره است که هدف آن محاسبه وزن معیارها و گزینه‌های پژوهش تحت یک مدل سلسله مراتبی است. در این مدل ابتدا مقایسات زوجی تشکیل شده و در اختیار خبرگان قرار داده می‌شود تا بر اساس طیف ۱ تا ۹ نظرات خود را نسبت به مقایسه دو به دوی معیارها بیان کنند. برای استفاده از اعداد راف در روش AHP (rough AHP) به طریق زیر عمل گردید:

۱. ابتدا مقایسات زوجی خبره‌ها را از نظر نرخ ناسازگاری بررسی کرده و چنانچه نرخ ناسازگاری کمتر از ۰٫۱ باشد یعنی مقایسه زوجی سازگار است و در صورتیکه بزرگ‌تر از ۰٫۱ باشد باید اعداد مقایسه زوجی اصلاح شود.
۲. ایجاد اعداد راف از اعداد خبره‌ها با استفاده از روابطی که در تئوری گفته شد.
۳. محاسبه وزن فاصله‌ای معیارها با استفاده از روش میانگین هندسی

بعد از پیاده‌سازی گام‌های تئوری راف وزن معیارها به صورت زیر محاسبه می‌شود.

$$w = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$$
$$= \{[2.902,3.556], [0.428,0.506,0.898,1.073], [2.663,3.123], [1.563,1.935], [0.595,0.703], [0.216,0.254]\}$$

در ایـــن تحقیق نیز ما از خبرگانی اســـتفاده نمودیم که بـــرای انتخاب ویژگی‌های مؤثر کمک نمودند و در این حالت بهترین نســـبت‌های مالی انتخاب شـــدند. مقدار وزن‌های محاسبه شده به وسیله روش ترکیبی راف-سلسله مراتبی در روش پیشنهادی به شرح جدول زیر است:

1. Pawlak
2. Zhai
3. Khoo
4. Zhong

Att1: نسبت آنی	Att25: 0.00381343240260058
Att2: سود خالص به فروش	Att24: 0.00427858327479137
Att3: سود ناخالص به کل داراییها	Att20: 0.00618257889138079
Att4: سود خالص به مجموع دارایی‌ها	Att27: 0.00657365841869104
Att5: فروش به دارایی ثابت	Att16: 0.00736528292505495
Att6: سود قبل از بهره و مالیات به فروش	Att14: 0.0073653915490675
Att7: سود قبل از بهره و مالیات به بدهی‌های جاری	Att1: 0.00908759998099143
Att8: نسبت جاری	Att9: 0.00935322974807349
Att9: فروش به مجموع دارایی‌ها	Att17: 0.00956439270753895
Att10: موجودی کالا به بدهی جاری	Att28: 0.0142025287136042
Att11: وجه نقد به جمع بدهی‌ها	Att6: 0.0142627546874609
Att12: بدهی‌های جاری به جمع دارایی‌ها	Att2: 0.0142700407731494
Att13: موجودی کالا به فروش	Att12: 0.0143807664233505
Att14: سرمایه به جمع داراییها	Att23: 0.0185360752681333
Att15: سود قبل از بهره و مالیات به کل داراییها	Att26: 0.0190464464415216
Att16: مجموع بدهیها به مجموع داراییها	Att8: 0.0221482127190511
Att17: حساب‌های دریافتنی به فروش	Att21: 0.0237633163633828
Att18: بهای تمام شده کالای فروش رفته به فروش	Att11: 0.0250140506889554
Att19: سود خالص به دارایی ثابت	Att13: 0.0285061366212412
Att20: هزینه‌های عملیاتی به فروش	Att15: 0.0349809449035666
Att21: موجودی کالا به دارایی جاری	Att4: 0.0378112532648968
Att22: لگاریتم طبیعی دارایی‌ها	Att10: 0.0390482265755149
Att23: وجوه نقد به مجموع دارایی‌ها	Att3: 0.0401621969319288
Att24: موجودی کالا به مجموع دارایی‌ها	Att5: 0.0510222259309194
Att25: دارایی ثابت به جمع دارایی‌ها	Att19: 0.0515867611609809
Att26: بدهی‌ها به حقوق صاحبان سهام	Att7: 0.0523728800527696
Att27: بهای تمام شده کالای فروش رفته به موجودی کالا	Att22: 0.062060175951289
Att28: فروش به حسابهای دریافتنی	Att18: 0.0630793576406058

می‌توان مشاهده نمود که با توجه به این جدول پارامترهای Att18(بهای تمام شده کالای فروش رفته به فروش) و Att22(لگاریتم طبیعی دارایی‌ها) بیشترین تأثیر را در خروجی دارند و این در حالی است که پارامترهایی همچون Att25 (دارایی ثابت به جمع دارایی‌ها) و Att24 (موجودی کالا به مجموع دارایی‌ها) کمترین تأثیر را در خروجی دارند. پارامترهایی که به طور کلی دارای کمترین تأثیر در خروجی می‌باشند باعث می‌شوند که نویز در داده‌ها ایجاد شود و خود علاوه بر اینکه زمان محاسبه را افزایش می‌دهند، دقت را نیز کاهش می‌دهند.

۳-۶- یادگیری شبکه‌های بیزی

یکی از مشکلات استفاده از شبکه‌های بیزی این است که ایجاد کامل شبکه حتی برای

یک خبره هم می‌تواند مشـکل باشـد. بنابراین، تلاش‌های زیادی برای یادگیری شبکه‌های بیزی صورت گرفته اسـت. در هر شبکه بیزی سـاختار آن و جداول احتمالات شرطی تعیین کننده آن هستند. بنابراین، باید بتوان با فرایند یادگیری این دو مورد را تعیین کرد. برای یادگیری خودکار سـاختار شـبکه یکی از روش‌های اصلی بر پایه تعیین وابستگی بین متغیرها بنا نهاده شده است. برای تعیین این وابسـتگی‌ها معیارهای زیادی مانند معیار آنتروپی طراحی شـده‌اند. با استفاده از این معیارها مانند روشـی که قبلاً توضیح داده شد، وابستگی هر متغیر X_i را نسبت به متغیرهای $X_1,...,X_{i-1}$ می‌سـنجیم و آنها را که معیار وابستگی‌شـان از یک آسـتانه‌ای بیشتر بود، به عنوان والدهـای X_i انتخـاب می‌کنیم. پس از تعیین سـاختار، اگـر مقدار همه متغیرهـا به طور کامل قابل مشـاهده باشـند، از تخمین احتمال معمولی (گرفتن تعدادی نمونه و شمردن تعداد اتفاقات یـک رویـداد در این مجموعـه نمونه) اسـتفاده می‌کنیم. اگر بعضـی از متغیرها قابل مشـاهده نباشـند، با استفاده از آموزش یک شـبکه نورونی می‌توان مقادیر جداول احتمالات شرطی را یاد گرفت.

۳-۷- الگوریتم شبکه بیزین در روش پیشنهادی

در این مرحله در ابتدای کار نرمال سازی انجام می‌شود و سپس نتایج استخراج شده از نرمال سـازی، قسـمتی از این داده‌ها به عنوان داده‌های آموزش استفاده شده و مدل شبکه بیزین ایجاد می‌شود و سپس با استفاده از داده‌های تست وزن این الگوریتم محاسبه می‌شود تا اینکه بتوان در ادامه میزان تاثیرگذاری این قسمت از الگوریتم را محاسبه نمود.

۳-۸- تشریح مدل پیشنهادی

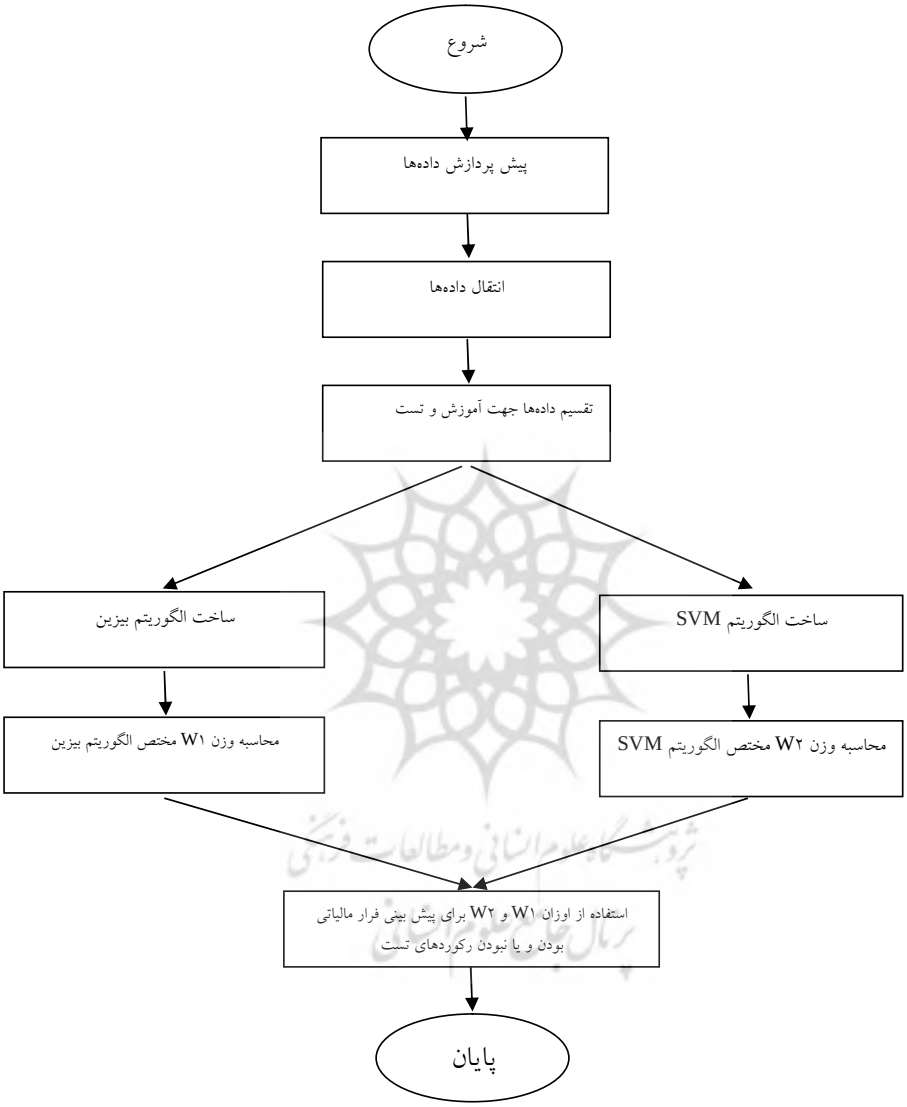
در ایـن تحقیـق قصد داریم تا راهکاری ارائه کنیم که مبتنـی بر الگوریتم SVM و الگوریتم بیزیـن بـوده و به وسـیله آن بتوان تصمیمات مهمی را در زمینه تشـخیص فرارهـای مالیاتی و همچنیـن جلوگیـری از آن‌ها ارائه نمود و برای اینکه بتوان به دقت بالاتری در روش پیشـنهادی دست یافت می‌بایست ابتدا با استفاده از داده‌کاوی سیستم مورد آموزش قرار گیرد.

در مدل پیشنهادی برای هر یک از این الگوریتم‌ها وزنی در نظر گرفته می‌شود و برای محاسبه و پیش‌بینی مقدار محاسـبه شـده توسـط هر یک از این الگوریتم‌ها در وزن آن الگوریتم ضرب می‌شـود و در انتها نتیجه واقعی بدست می‌آید که دارای دقت بالاتری خواهد بود زیرا از فواید هر دو این الگوریتم‌ها استفاده نموده است.

روش پیشنهادی دارای مراحل مختلفی می‌باشد که به ترتیب عبارتند از:

۱- پیش‌پردازش داده‌ها ۲- انتقال داده‌ها ۳- انتخاب ویژگی‌های مؤثر با اسـتفاده از مجموعه راف و AHP ۴- ساخت الگوریتم SVM ۵- ساخت الگوریتم بیزین ۶- آموزش و محاسبه اوزان الگوریتم SVM و الگوریتم بیزین.

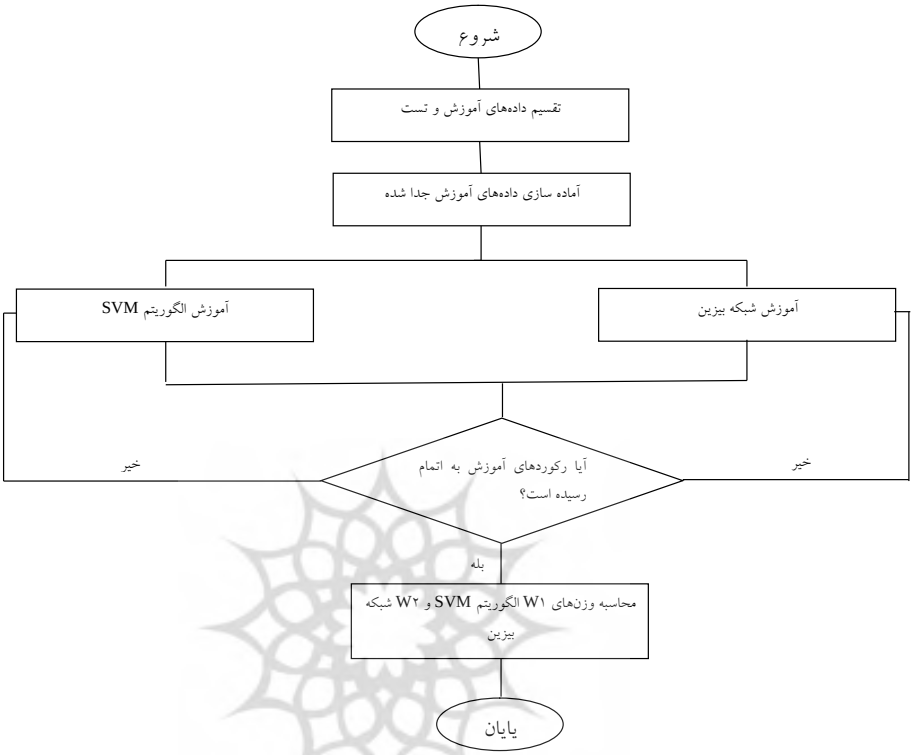
فلوچارت کلی روش پیشنهادی را می‌توان در شکل زیر مشاهده نمود.



۳-۸-۱- آموزش و محاسبه اوزان الگوریتم SVM و الگوریتم شبکه بیزین

در ابتدای کار درصدی از مجموعه دیتاست مورد استفاده برای محاسبه آموزش و محاسبه وزن مورد استفاده قرار می‌گیرد. مدل پیشنهادی یک الگوریتم ترکیبی می‌باشد که از دو الگوریتم SVM و الگوریتم شبکه بیزین استفاده می‌کند و این دو الگوریتم هر کدام سهمی از جواب نهایی را خواهند داشت که بدین شکل دقت سیستم افزایش می‌یابد. می‌توان نمایی از این مرحله را در

.Error! Reference source not found مشاهده نمود.



همانطور که درشکل فوق مشاهده می‌شود الگوریتم پیشنهادی با استفاده از مجموعه داده‌ای در ابتدا به محاسبه وزن می‌پردازد و این محاسبه بدین شکل می‌باشد که هر الگوریتم با استفاده از ۷۰ درصد دیتاست موجود آموزش می‌بیند و در نهایت با توجه به تعداد جواب‌های صحیح امتیاز و یا وزنی به آن تعلق می‌گیرد تا اینکه با استفاده از وزن تعلق گرفته بتوان در مرحله بعدی وزنی را برای خروجی هر الگوریتم در نظر گرفت همانطور که در معماری نیز می‌توان مشاهده نمود بعد از محاسبه وزن که بصورت تقسیم تعداد جواب‌های درست به تعداد کل جواب‌های حدس زده شده می‌باشد، می‌توان میزان تأثیر گذاری هر کدام از این الگوریتم را در خروجی نهایی بهتر تشخیص داد. بعد از محاسبه وزن‌ها، به ازای هر رکورد، پیش‌بینی‌ای توسط الگوریتم شبکه بیزین و الگوریتم SVM صورت می‌گیرد که مقدار پیش‌بینی شده می‌بایست در وزن آن الگوریتم ضرب شود و خروجی نهایی پیش‌بینی الگوریتم برابر است با جمع نتایج هر یک از الگوریتم ضرب در وزن آن الگوریتم که بدین صورت نتیجه نهایی بدست می‌آید و دسته بندی درست صورت می‌گیرد.

۳-۸-۲- روش انتخاب داده‌های آزمون و تست

روشی که برای انتخاب داده های آزمون و تست مورد استفاده قرار گرفته است، روش cross fold K validation و مقدار K برابر مقدار متعارف ۱۰ قرارداده خواهد شد. زیرا این مقدار اثبات شده است که بهترین نسبت برای ارزیابی روش‌های داده‌کاوی می‌باشد که در این تحقیق نیز از همین تعداد استفاده شده است.

داده‌ها در ابتدا توسط برنامه به فرمت مناسب برای تحلیل قرار می‌گیرد یا به عبارتی پیش پردازش ابتدایی صورت می‌گیرد فایلی با فرمت ARFF ایجاد می‌باشد که ساختاری استاندارد برای تحلیل می‌باشد.

پیاده‌سازی در برنامه Visual Studio ۲۰۱۷ و با زبان برنامه نویسی C# صورت گرفته است و در حین کار از کتابخانه Weka نیز کمک گرفته شده است. نرم‌افزار Weka یک واسط همگون برای بسیاری از الگوریتم‌های یادگیری متفاوت، فراهم کرده است که از طریق آن روش‌های پیش پردازش، پس از پردازش و ارزیابی نتایج طرح‌های یادگیری روی همه مجموعه‌های داده موجود، قابل اعمال است. در پیاده‌سازی مدل از این نرم‌افزار به عنوان کتابخانه‌ای در کنار پروژه خود استفاده شد. همچنین با استفاده از کتابخانه ZedGraph، نمودارهای مورد نیاز برنامه ایجاد شده‌اند.

۳-۹- نتایج مدل

خروجی‌های مربوط به اجرای مدل تحقیق در جداول زیر ارائه شده است:

جدول (۱): نسبت درصد صحت پیش‌بینی‌ها و خطای پیش‌بینی‌ها

	درصد صحت پیش‌بینی‌ها	درصد خطای پیش‌بینی‌ها
الگوریتم پیشنهادی	٪۶۰,۵۸	٪۴۳,۷۶
الگوریتم بیزین	٪۵۶,۰۷	٪۴۳,۹۳
الگوریتم SVM	٪۵۶,۲۴	٪۴۳,۷۶

با توجه به نتایج دریافتی می‌توان به وضوح مشاهده نمود که الگوریتم پیشنهادی با ٪۶۰,۵۸ دقت دارای بالاترین دقت صحت و با ٪۴۳,۷۶ اشتباه دارای کمترین میزان اشتباه می‌باشد بنابراین الگوریتم پیشنهادی از باقی الگوریتم‌ها بسیار بهتر عمل می‌کند.

جدول (۲): خطای میانگین مربعات (MSE) برای روش‌های مختلف مورد بررسی

	خطای میانگین مربعات (MSE)
الگوریتم پیشنهادی	۰,۴۹۵
الگوریتم بیزین	۰,۵۲۵
الگوریتم SVM	۰,۵۳۵

می‌توان در جدول فوق مشاهده نمود که روش پیشنهادی دارای کمترین خطای میانگین مربعات می‌باشد و این نشان از عملکرد مناسب‌تر و دقت بالاتر روش پیشنهادی می‌باشد. می‌توان مشاهده کرد که روش پیشنهادی از بیزین بهتر، بیزین از SVM بهتر عمل کرده‌اند.

جدول (۳): خطای میانگین قدر مطلق (MAE) برای روش‌های مختلف مورد بررسی

	خطای میانگین قدر مطلق (MAE)
الگوریتم پیشنهادی	۰,۴۷۷
الگوریتم بیزین	۰,۴۷۸
الگوریتم SVM	۰,۴۷۵

می‌توان مشاهده نمود که از منظر خطای میانگین قدر مطلق، الگوریتم SVM دارای خطای کمتری می‌باشد و روش پیشنهادی تنها از الگوریتم بیزین بهتر عمل کرده است.

جدول (۴): خطای جذر میانگین مربعات (RAE) برای روش‌های مختلف مورد بررسی

	خطای جذر میانگین مربعات (RAE)
الگوریتم پیشنهادی	۱۰۰,۲۹۵
الگوریتم بیزین	۱۰۰,۴۷۹
الگوریتم SVM	۹۹,۶۸۳

در جدول فوق Error! Reference source not found. نیز خطای جذر میانگین مربعات روش پیشنهادی از روش بیزین کمتر و از الگوریتم SVM بیشتر می‌باشد.

برای بررسی دقیق‌تر روش پیشنهادی و دیگر روش‌های مورد بررسی FP، TN، TP و FN نیز مورد محاسبه قرار گرفته است.

جدول (۵): بررسی پارامترهای FP، TN، TP و FN برای روش پیشنهادی و دیگر روش‌های مورد بررسی

	TP	TN	FP	FN
الگوریتم پیشنهادی	۱۹	۷۰۰	۴۴۴	۲۳
الگوریتم بیزین	۶۴	۶۰۱	۳۹۹	۱۲۲
الگوریتم SVM	۱۹۱	۴۷۶	۲۷۲	۲۴۷

می‌توان در جدول فوق Error! Reference source not found.مشاهده نمود که روش پیشنهادی داری مجموع TP+TN بیشتری از دیگر روش‌های مورد بررسی می‌باشد زیرا در واقع تعداد TP و TN هر چه بیشتر باشد یعنی تعداد پیش‌بینی‌های صحیح بیشتری صورت گرفته است. و هر چه تعداد FP و FN کمتر باشد یعنی تعداد پیش‌بینی‌های غلط کمتری صورت گرفته است. تعداد کل داده‌های استفاده شده ۱۱۸۶ رکورد می‌باشد که چون از روش cross k fold استفاده شده است یعنی تمامی قسمت‌های داده‌ها هم برای آموزش و هم برای تست قرار گرفته است که در این حالت اگر هر سطر جمع بسته شود به عدد ۱۱۸۶ منتهی می‌شود زیرا ۱۱۸۶ رکورد مورد تست قرار گرفته است.

از پارامترهای مهم در علم داده‌کاوی، نرخ خطا (FPR[1]) و نرخ صحت (TPR[2]) عملکرد می‌باشد که این پارامترها از طریق روابط زیر نشان داده شده‌اند.

$$FPR = \frac{FP}{FP + TP}$$

$$TPR = \frac{TP}{FP + TP}$$

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

این دو پارامتر برای روش پیشنهادی و دیگر روش‌های مورد بررسی در این تحقیق محاسبه و در جدول (۶) Error! Reference source not found.نشان داده شده‌اند.

جدول (۶): نرخ خطا و نرخ صحت برای روش پیشنهادی و دیگر روش‌های مورد بررسی

	FPR	TPR
الگوریتم پیشنهادی	۰,۳۸۸	۰,۴۵۲
الگوریتم بیزین	۰,۳۹۹	۰,۳۴۴
الگوریتم SVM	۰,۳۶۴	۰,۴۳۶

1. False Positive Rate
2. True Positive Rate

بـا توجـه به جدول فـوق می‌توان دریافت که روش پیشـنهادی دارای نرخ صحت بیشـتری می‌باشـد ولی دارای کمترین نرخ خطا می‌باشد بنابراین روش پیشنهادی خطای بسیار کمی دارد بنابراین می‌توان گفت قطعاً دارای دقت بیشـتری می‌باشـد. در اینجا کاملاً قابل مشاهده است که الگوریتم SVM از الگوریتم بیزین بهتر عمل کرده است و دارای نرخ صحت بیشتری می‌باشد.

یکی دیگر از پارامترهای بسـیار مهم، دقت[1]می‌باشد که در اکثر مقالات برای مقایسه روش‌ها اسـتفاده شده اسـت که در این تحقیق نیز برای روش‌های مختلف محاسـبه شده است. دقت با استفاده از رابطه زیر بدست می‌آید.

$$ACC=(TP+TN)/(TP+TN+FP+FN)$$

در جدول (۷) دقت محاسـبه برای روش پیشـنهادی و دیگر روش‌های مورد بررسـی در این تحقیق نشان داده شده است.

جدول (۷): دقت محاسبه شده برای روش پیشنهادی و دیگر روش‌های مورد بررسی

	دقت (Accuracy)
الگوریتم پیشنهادی	۰,۶۰۶
الگوریتم بیزین	۰,۵۶۱
الگوریتم SVM	۰,۵۶۲

همانطور که نتایج می‌توان مشاهده نمود روش پیشنهادی دارای دقت بالاتری از روش بیزین و دیگر روش‌ها می‌باشـد و همچنین صحت یا TPR روش پیشـنهادی نیز از روش بیزین بیشتر می‌باشـد که این نشـان دهنده عملکرد مناسب روش پیشنهادی نسـبت به روش بیزین و دیگر روش‌های مورد بررسی می‌باشد.

در همین جا می‌توان مشاهده نمود که تمامی فرضیه‌ها به اثبات رسیدند زیرا راهکاری در این تحقیق بیان شـده اسـت که ترکیبی از دو روش SVM و بیزین می‌باشد و مشاهده شد که از هر یک از این روش‌ها دارای دقت بیشـتری می‌باشـد و مشاهده شد که دارای نرخ صحت پیش‌بینی بیشتری از دیگر روش‌های مورد بررسی می‌باشد.

در ادامه ماتریس درهم ریختگی[2]مربوط به روش پیشنهادی و دیگر روش‌های مورد بررسی در این تحقیق نشـان داده شده اسـت. جدول یا ماتریس درهم ریختگی، نتایج حاصل از طبقه‌بندی را بر اسـاس اطلاعات واقعی موجود، نمایش می‌دهد. حال بر اساس این مقادیر می‌توان معیارهای مختلف ارزیابی دسته بند و اندازه‌گیری دقت را تعریف کرد.

1. Accuracy
2. Confusion Matrix

جدول Confusion matrix روش پیشنهادی

TP + FN: 42	FN: 23	TP: 19
FP + TN: 1144	TN: 700	FP: 444
FPR: 0.388 TPR: 0.452 Accuracy: 0.606	FN + TN: 723	TP + FP: 463

جدول Confusion matrix روش الگوریتم بیزین

TP + FN: 186	FN: 122	TP: 64
FP + TN: 1000	TN: 601	FP: 399
FPR: 0.399 TPR: 0.344 Accuracy: 0.561	FN + TN: 723	TP + FP: 463

جدول Confusion matrix روش SVM

TP + FN: 438	FN: 247	TP: 191
FP + TN: 748	TN: 476	FP: 272
FPR: 0.364 TPR: 0.436 Accuracy: 0.562	FN + TN: 723	TP + FP: 463

جداول فوق نیز موید این مطلب است که روش پیشنهادی دارای دقت بالاتری می‌باشد زیرا در این جدول دارای مقدار TP و TN بیشـــتری از دیگر الگوریتم‌ها می‌باشد و در کنار آن نیز دارای FP و FN کمتری از دیگر الگوریتم‌های مورد بررسی در اینجا می‌باشد. زیرا هر چه یک الگوریتم دارای TP و TN بیشـــتری باشـــد یعنی دارای فـــرار مالیاتی بودن و یـــا نبودن‌های در مجموعه داده‌های تست را به مقدار بیشتری درست تشخیص داده است و FP و FN نشان دهنده برعکس این قضیه یعنی پیش‌بینی اشتباه می‌باشد.

۴- نتیجه‌گیری و توصیه‌های سیاستی

به طور کلی، فرار مالیاتی یکی از اجزای اقتصاد غیررسمی است که در آن، افراد قانون‌گریز به ازای فعالیت‌های قانونی انجام‌شده، تمام یا قسمتی از مالیات خود را نمی‌پردازند. این در حالی است که روش‌های پیش‌بینی فرار مالیاتی از اهمیت زیادی برخوردار است. در این میان، روش‌های سنتی پیش‌بینی فرار مالیاتی در صورت‌های مالی، دارای برخی مفروضات محدود کننده مانند خطی بودن، نرمال بودن و مستقل بودن متغیرهای پیش‌بینی کننده یا ورودی‌ها است. نظر به اینکه در ارتباط با داده‌های مالی، تخطی از این مفروضات متصور است، این روش‌های سنتی در ارتباط با میزان کارایی و اعتبار، دارای محدودیت‌های زیادی هستند. اما روش‌های مصنوعی و از آن جمله الگوریتم ژنتیک دارای ناسازگاری و موارد تخطی کمتری در ارتباط با این مفروضات می‌باشند. یکی از برتری‌های الگوریتم ژنتیک نسبت به سایر عدم وابستگی این الگوریتم بر فرضیه‌های آماری محدودکننده و نرمال بودن توزیع نسبت‌ها یا برابری واریانس یا کوواریانس ماتریس نسبت‌ها است. در این بین جدیدترین روش‌هایی که می‌توان به وسیله آنها فرار مالیاتی را پیش‌بینی و کشف نمود، الگوریتم‌های یادگیری ماشین است. یادگیری ماشین فرآیند استفاده از داده‌ها است که به طور خودکار یک مدل می‌سازد. این ماشین به عنوان ورودی از مجموعه‌ای از ویژگی‌های شناخته شده استفاده می‌کند و به عنوان خروجی، مدلی برای پیش‌بینی ارائه می‌دهد.

ایده اصلی مقاله از مقایسه نسبت‌های مالی شرکت با نسبت‌های سال گذشته و همچنین برآورد مالیات سال مورد رسیدگی از روی مالیات سال قبل گرفته شده است. الگوریتمی که در این تحقیق بیان شده است مبتنی بر یک الگوریتم ترکیبی می‌باشد که از ترکیب الگوریتم‌های SVM[1] و شبکه بیزین[2] می‌باشد. این الگوریتم‌ها بدین دلیل انتخاب شدند که هر یک به تنهایی دارای مزایا و معایبی می‌باشند لیکن با ادغام آنها، معایب این روشها پوشانده می‌شود و به دقت بالاتری دست می‌یابیم و همینطور برای انتخاب ویژگی نیز از الگوریتم مجموعه راف ترکیب شده با تحلیل سلسله مراتبی استفاده نمودیم که از مجموعه تمام نسبتهای مالی تنها بتوان ویژگی‌های موثرتر و یا همان نسبت‌های مالی موثرتر را شناسایی نمود.

نتایج حاصل از اجرای مدل‌ها نشان می‌دهد که الگوریتم پیشنهادی با ۶۰,۵۸٪ دقت، دارای بالاترین دقت صحت و با ۴۳,۷۶٪ اشتباه دارای کمترین میزان اشتباه می‌باشد. کاملاً مشخص است که الگوریتم پیشنهادی از باقی بهتر و از خود الگوریتم SVM، الگوریتم بیزین که الگوریتم‌های پایه روش پیشنهادی می‌باشند، نیز بسیار بهتر عمل می‌کند.

به طور کلی و خلاصه نتایج تحقیق نشان می‌دهد که ماموران مالیاتی در هنگام رسیدگی به پرونده مودیان با محدودیت‌هایی مواجه هستند که مهم‌ترین آن‌ها، کمبود زمان و عدم تفکیک مودیان دارای فرار یا فاقد فرار مالیاتی (اصطلاحاً کم ریسک و پر ریسک) می‌باشد به نحویکه ممکن است برای رسیدگی یک پرونده که فاقد فرار مالیاتی می‌باشد زمان بیشتری اختصاص یابد

1. Support Vector Machines (SVM)
2. Bayesian Network (BN)

و برعکس برای رسیدگی به پرونده‌ای که دارای فرار مالیاتی می‌باشد زمان کمتری اختصاص یابد و نهایتاً سازمان مالیاتی را از رسیدن به اهدافش که مهم‌ترین آن‌ها وصول مالیات می‌باشد باز دارد و یا هزینه‌های وصول مالیات را افزایش دهد. در این تحقیق جهت شناسایی فرار مالیاتی مودیان، با استفاده از اطلاعات پرونده مالیاتی و همچنین اطلاعات استخراجی از سایت بورس اوراق بهادار؛ مدل‌هایی مبتنی بر یادگیری ماشین و هوش مصنوعی ارائه شد که چنانچه سازمان مالیاتی از این مدل‌ها استفاده نماید می‌تواند قبل از رسیدگی مالیاتی، مودیان را به دو دسته دارای فرار مالیاتی و فاقد فرار مالیاتی تقسیم بندی کند تا با برنامه‌ریزی بهتر و دقیق‌تر، زمان بیشتر برای رسیدگی به پرونده مالیاتی مودیان دارای فرار مالیاتی اختصاص یابد و درنهایت با هزینه کمتر، مالیات بیشتر وصول نموده و باعث افزایش کارایی سازمان مالیاتی شود.

سرعت و دقت بالای روشهای یادگیری ماشین و همچنین قابلیت تغییر و تکمیل ورودیهای سیستم، ویژگیهایی هستند که سازمان مالیاتی در صورت استفاده از هوش مصنوعی، می‌تواند از آنها به نحو مطلوب استفاده نماید و دقت تشخیص فرار مالیاتی را بیش از پیش نماید.

لازم به توضیح است که از تحقیقات مشابه می‌توان به دو مورد اشاره کرد: در مقاله جمشیدی نوید و همکاران (۱۳۹۸) مدل رگرسیون خطی چندمتغیره با مدل شبکه عصبی مصنوعی برای کشف فرار مالیاتی اشخاص حقوقی مورد مقایسه قرار گرفته که در نتیجه‌گیری تحقیق، مشخص شد که شبکه عصبی مصنوعی کارایی بیشتری دارد.

در مقاله دستگیر و غریبی (۱۳۹۴)، روش‌های داده‌کاوی به منظور ارتقای عملکرد تشخیص فرار مالیاتی را مورد استفاده قرار داده است.

تفاوت این مقاله با دو مقاله قبلی این است که الگوریتمی که در این تحقیق بیان شده است مبتنی بر یک الگوریتم ترکیبی می‌باشد که از ترکیب الگوریتم‌های SVM[1] و شبکه بیزین[2] می‌باشد.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

پیشنهاد برای تحقیقات آتی

پیشنهاد ۱

در این روش پیشنهادی از مجموعه راف و روش تصمیم‌گیری سلسله مراتبی برای انتخاب ویژگی‌های تاثیرگذار استفاده شده است. این روش دارای دقت بالاتری است ولی می‌توان از راهکارهای دیگری نیز برای این کار استفاده نمود که برای نمونه می‌توان به روش ترکیبی ژنتیک و شبکه‌های عصبی اشاره کرد. زیرا ژنتیک و عصبی دارای سرعت بسیار بالاتری از روش استفاده شده در این تحقیق می‌باشد.

پیشنهاد ۲

می‌توان روش پیشنهادی را با بهبود در الگوریتم C۴,۵ نیز استفاده نمود یعنی روش

1. Support Vector Machines (SVM)
2. Bayesian Network (BN)

پیشنهادی را روی C۴,۵ با بهبودی مشابه بهبود که روی SVM در این تحقیق انجام شد، انجام داد تا عملکرد آن افزایش یابد البته نمی‌توان به‌طورقطع گفت که عملکرد آن بهتر می‌شـود بلکه می‌بایست این روش مورد آزمایش قرار گیرد تا صحت عملکرد بررسی شود.

فهرست منابع

الف- منابع فارسی:

۱. مداح، مجید و نعمت الهی، سـمیه (۱۳۹۰). بررسـی و تحلیل اثر نرخ تعرفه بر فـرار مالیاتی در واردات ایران، پژوهشنامه مالیات، شماره ۱۲: ۲۵-۹
۲. ضیائی بیگدلی، محمدتقی؛ فرهاد طهماسـبی بلداجی (۱۳۸۴). مالیات بر ارزش افزوده) مالیات مدرن(تهران: پژوهشکده امور اقتصادی.
۳. خـان جان، علیرضا (۱۳۹۳). پتانسـیل فرار و تقلب در نظام مالیات بر ارزش افـزوده، بازخورد تجربه عملیاتی کشورهای در حال توسعه، مجله اقتصادی.
۴. نصر اصفهانی، مهدی، نصر اصفهانی، مهرناز، و دلوی اصفهان، محمدرضا (۱۳۹۱). بررسـی عوامل مؤثر بر فرار مالیاتی از دیدگاه مؤدیان و کارشناسان مالیاتی. پژوهش‌های مالیه اسلامی، ۱(۱): ۳۶-۲۷.
۵. نویـدی، حمیدرضا. (۱۳۸۴). مدل مالیاتی بازی‌هـای ریاضی در وضعیت گریز مالیاتی و تبانی. مجلهٔ تحقیقات اقتصادی، ۶۹ (۲): ۲۷۶-۲۶۱.
۶. صمدی، علی‌حسـین و تابنده، رضیه (۱۳۹۲). فرار مالیاتی در ایران (بررسـی علل و آثار و برآورد میزان آن). پژوهشنامه مالیات، ۱۹(۳): ۱۰۶-۷۷.
۷. روستا، محمود و حیدریه، سیدعبدالله (۱۳۹۳). رتبه‌بندی علل فرار مالیاتی به روش AHP. پژوهشنامه مالیات، ۲۴(۴): ۱۷۳-۱۵۷.
۸. نوربخش، محسـن (۱۳۷۷). کد اقتصادی، سیسـتم گـردش پول و کالا و نقش آن در وصـول مالیات (تجربهٔ جمهوری اسلامی ایران). مجلهٔ تحقیقات اقتصادی، ۵۲ (بهار و تابستان): ۱-۴۲.
۹. دستگیر، محسن، غریبی، مریم (۱۳۹۴). کاربست روش‌های داده‌کاوی به منظور ارتقای عملکرد تشخیص فرار مالیاتی، فصلنامه پژوهشنامه مالیات، شماره ۲۶: ۱۱۶-۹۵
۱۰. عبدلی، قهرمان، ابریشمی، حمید، حسینی فرد، سیدمحمد (۱۳۹۴). تحلیل نظری و تجربی حسابرسی مالیاتی مطلوب در مالیات بر درآمد جهت کاهش فرار مالیاتی، فصلنامه پژوهشنامه مالیات، شماره ۲۸.
۱۱. رحیمی کیا، اقبال، محمدی، شـاپور، غضنفری، مهدی (۱۳۹۴). تشـخیص فرار مالیاتی با اسـتفاده از سیستم هوشمند ترکیبی، فصلنامه پژوهشنامه مالیات، شماره ۲۶،۱۳۵-۱۶۳.
۱۲. سـازمان مالیاتی ایران (شـهریور ۱۳۹۰). «نگاهی به طرح جامع مالیاتی»، سـتاد طرح جامـع مالیاتی- دفتر مدیریت تغییر
۱۳. قانون مالیاتهای مستقیم، مصوب ۱۳۹۴/۴/۳۱.
۱۴. جعفری صمیمی، احمد (۱۳۸۷)، "اقتصاد بخش عمومی ۲ ". تهران: نشر سمت، ص ۲۰.
۱۵. جمشـیدی نوید، بابک، ناصرآبادی، دلیر، طاهرآبادی، علی اصغر، قنبری، مهرداد (۱۳۹۸). کشـف فرار مالیاتی اشخاص حقوقی: مقایسه کارایی رگرسیون خطی چند متغیره و شبکه‌های عصبی مصنوعی، فصلنامه پژوهشنامه مالیات، شماره ۴۰، ۱۵۷-۱۸۵
۱۶. عرب مازار یزدی، علی (۱۳۸۰). اقتصاد سـیاه در ایران: اندازه، علل و آثار آن در سـه دهه اخیر. نشریه برنامه و بودجه (سازمان مدیریت و برنامه‌ریزی کشور). ۶۲-۶۳: ۶۰-۳.

ب- منابع انگلیسی:

1. Manasan, R. G. (1988). Tax Evasion in the Philippines 1981–1985. Journal of Philippine Development, 27(2), 167- 190.

2. Fuest, C. & Riedel, N. (2019). Tax Evasion, Tax Avoidance and Tax Expenditures in Developing Countries: A Review of the Existing Literature (Report prepared for the UK Department for International Development (DFID), June 19th).
3. Sandmo, A. (2005). The theory of tax evasion: A retrospective view. National Tax Journal, 58(4), 643–666
4. Schneider, F.,& Enste, D. H. (2000). Shadow Economies: Size, causes, and consequences. Journal of Economic Literature, 38(1), 77–114.
5. Khwaja, Munawer sultan,Rajul Awasthi, and Jan Loeprick (2011).»Risk-based tax audits:Approaches and country Experiences«, the international bank for Reconstruction and Development/ The World Bank, 1818 H Street NW, Washington Dc 20433
6. Franzoni, A., (2014). Tax Evasion and social Interactions", CIRANO Working Pa-pers, Journal of Public Economics.18.
7. Hashimzade, Nigar, Huang, Zhanyi, and Myles, Gareth D. (2010). Tax Fraud by Firms and Optimal Auditing, International Review of Law and Economics, Volume 30, Issue 1, 10-17.
8. Richardson, L., (2010), The Underground Economics: Tax Evasion and detections Information, Cambridge University Press.
9. Fuest, C. & Riedel, N. (2009). Tax Evasion, Tax Avoidance and Tax Expenditures in Developing Countries: A Review of the Existing Literature (Report prepared for the UK Department for International Development (DFID), June 19th).

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی