پژوهش‌های حسابداری مالی
سال هشتم، شماره دوم، پیاپی (۲۸)، تابستان ۱۳۹۵
تاریخ وصول: ۱۳۹۴/۱۰/۱۹
تاریخ پذیرش: ۱۳۹۵/۸/۱۱
صص: ۲۴-۱

بررسی سودمندی طبقه‌بندی‌کنندۀ جنگل‌های تصادفی و روش انتخاب متغیر ریلیف در پیش‌بینی بحران مالی: مطالعۀ شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران

محمد حسین ستایش[1]*، مصطفی کاظم‌نژاد، محمد حلاج*****

* دانشیار حسابداری، دانشگاه شیراز، شیراز، ایران
setayesh@shirazu.ac.ir
** دانشجوی دکتری حسابداری، دانشگاه شیراز، شیراز، ایران
mkazemi5166@gmail.com
*** دکتری حسابداری، دانشگاه شیراز، شیراز، ایران
smhallaj62@gmail.com

چکیده

پژوهش حاضر به پیش‌بینی بحران مالی شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران با استفاده از طبقه‌بندی‌کنندۀ غیرخطی جنگل‌های تصادفی می‌پردازد. در این راستا، پس از بررسی متون پژوهش و شناسایی ۶۹ متغیر پیش‌بین اولیه، از روش انتخاب متغیر ریلیف برای شناسایی متغیرهای پیش‌بین بهینه استفاده شد. یافته‌های تجربی مربوط به بررسی ۹۵ شرکت - سال سالم (بدونِ درماندگی مالی) و ۹۵ شرکت - سال (درماندۀ مالی) پذیرفته‌شده در بورس اوراق بهادار تهران در سال‌های ۱۳۸۰ تا ۱۳۹۲ بیانگر عملکرد بهتر جنگل‌های تصادفی نسبت به رگرسیون لجستیک است. به بیان دیگر، در صورت استفاده از این طبقه‌بندی‌کننده، به‌طور معناداری، میانگین دقت افزایش و خطای نوع اول و دوم کاهش می‌یابد. افزون بر این، یافته‌های پژوهش بیانگر سودمندی روش انتخاب متغیر ریلیف در پیش‌بینی بحران مالی است. به عبارت دیگر، در صورت استفاده از متغیرهای منتخب روش ریلیف (نسبت به استفاده از ۶۹ متغیر اولیه)، به‌طور معناداری، میانگین دقت افزایش وخطای نوع اول و دوم کاهش می‌یابد.

واژه‌های کلیدی: طبقه‌بندی‌کنندۀ جنگل‌های تصادفی، روش انتخاب متغیر ریلیف، پیش‌بینی بحران مالی.

۱ـ نشانی مکاتباتی نویسنده مسؤول: شیراز، دانشگاه شیراز، دانشکده اقتصاد، مدیریت و علوم اجتماعی، گروه حسابداری.

مقدمه

سرمایه‌گذاران و اعتباردهندگان، تمایل زیادی برای پیش‌بینی بحران مالی[1] و ورشکستگی[2] شرکت‌ها دارند، زیرا درصورت ورشکستگی، هزینه‌های زیادی به آن‌ها تحمیل می‌شود. افزون بر این، به‌دلیل اینکه بحران مالی اغلب مقدم و در بسیاری از موارد تسریع‌کنندهٔ ورشکستگی است، ایجاد یک مدل قوی و موفق برای شناسایی و پیش‌بینی شرکت‌های دارای بحران مالی که ممکن است ورشکسته شوند، در پیشگیری یا حداقل در کاهش پیشرفت فرآیند ورشکستگی نقش بسزایی ایفا می‌کند [۴۴]. تاکنون در اغلب پژوهش‌های انجام‌شده در بورس اوراق بهادار تهران، از روش‌های خطی برای پیش‌بینی بحران مالی استفاده شده است. در پژوهش‌هایی نیز از روش‌های غیرخطی از قبیل شبکه‌های عصبی مصنوعی[3]، ماشین‌بردار پشتیبان[4]، تحلیل پوششی داده‌ها و الگوریتم ژنتیک استفاده شده که یافته‌های این پژوهش‌ها حاکی از عملکرد بهتر روش‌های غیرخطی نسبت به روش خطی است. با وجود این، علی‌رغم عملکرد مناسب روش غیرخطی جنگل‌های تصادفی[5] در پیش‌بینی، تاکنون پژوهشی که به پیش‌بینی بحران مالی شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران پرداخته باشد، مشاهده نشد. افزون بر این، به‌رغم اهمیت انتخاب متغیرهای پیش‌بین[6] در عملکرد پیش‌بینی بحران مالی و ورشکستگی، تاکنون پژوهش‌های اندکی در زمینهٔ انتخاب متغیرهای پیش‌بین و روش‌های آن انجام شده است. در مقابل، در بسیاری از پژوهش‌های داخلی و خارجی انجام‌شده در زمینهٔ ورشکستگی و بحران مالی، مرحلهٔ انتخاب متغیرهای پیش‌بین، نادیده گرفته شده و متغیرهای پیش‌بین بدون ضابطه و صرفاً با توجه به پژوهش‌های گذشته انتخاب شده است که این امر به انتخاب متغیرهای پیش‌بین غیربهینه و در برخی موارد، متغیرهای پیش‌بین نامناسب منجر می‌شود. یافته‌های پژوهش لو [۴۸] نیز گویای آن است که انتخاب متغیرهای پیش‌بین و روش‌های آن، نسبت به انتخاب مدل پیش‌بینی، تأثیر بیشتری بر میانگین دقت پیش‌بینی دارد.

با توجه به اهمیت پیش‌بینی بحران مالی و ورشکستگی و نقاط ضعف روش خطی، این پژوهش درصدد پیش‌بینی بحران مالی شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران با استفاده از طبقه‌بندی‌کنندهٔ غیرخطی جنگل‌های تصادفی است. افزون بر این، با توجه به اهمیت انتخاب متغیرهای پیش‌بین در پیش‌بینی از روش انتخاب متغیر ریلیف[7] برای انتخاب متغیرهای بهینه استفاده شده است.

مبانی نظری پژوهش

شیوه‌های آماری تک‌متغیری[8]، از اولین شیوه‌هایی بودند که برای پیش‌بینی ورشکستگی و بحران مالی استفاده شدند. آلتمن [۱۹] تحلیل به تک‌متغیره انتقاد کرد و تحلیل ممیزی چندگانه[9] را که در آن چندین نسبت مالی، همزمان در پیش‌بینی ورشکستگی بررسی می‌شد، پیشنهاد نمود. در اغلب پژوهش‌های بعدی از روش تجزیه و تحلیل خطی لجستیک[10] استفاده شده

[1] Financial Distress
[2] Bankruptcy
[3] Neural Networks
[4] Support Vector Machine
[5] Random Forests
[6] Feature (Variable) Selection
[7] Relief
[8] Univariate Analysis
[9] Multivariate Discriminant Analysis
[10] Logistic Regression

است. به‌طور کلی، یکی از مشکلات آماری در تجزیه و تحلیل داده‌های حسابداری از جمله نسبت‌های مالی، مسألهٔ انتخاب فنون آماری با توجه به توزیع اطلاعات و نسبت‌هاست. اغلب مطالعات گذشته بر روی نسبت‌های مالی، از رگرسیون یک یا چندمتغیره استفاده و در اکثر آن‌ها توزیع نسبت‌ها نرمال فرض شده است. داشتن اطلاعاتی در زمینهٔ توزیع فراوانی نسبت‌های مالی برای تصمیم‌گیری در رابطه با تعیین ابزارهای آماری مناسب برای متغیر مورد مطالعه ضروری است [۶۳]. هنگام به‌کارگیری نسبت‌های مالی غالبا توزیع داده‌ها نرمال فرض شده و از روش‌های پارامتریک استفاده شده است. در متون اقتصاد و آمار، ابزارهای آماری متعدد با مفروضات متفاوت در زمینهٔ توزیع داده‌های مورد بررسی وجود دارد. به‌عنوان نمونه، در آزمون t که برای ارزیابی معناداری متغیرها در رگرسیون حداقل مربعات به کار می‌رود، از فرض نرمال بودن استفاده می‌شود. با وجود این، ابزارهایی از قبیل لاجیت که برای پیش‌بینی بحران مالی استفاده می‌شود، مفروضات دیگری برای توزیع داده‌ها در نظر می‌گیرد. البته اغلب ابزارهای آماری موجود برای تجزیه و تحلیل اطلاعات صورت‌های مالی بر این فرض استوار هستند که داده‌های مورد بررسی از توزیع نرمال تبعیت می‌کنند. بنابراین، در صورت نرمال نبودن توزیع داده‌ها، پژوهشگر در بررسی و تفسیر نتایج با مشکل روبه‌رو می‌شود [۳۳]. دروری [۳۰] نیز معتقد است که استفاده از بسیاری از ابزارهای آماری به نرمال بودن داده‌های مورد بررسی بستگی دارد، پس اگر توزیع نسبت‌ها نرمال نباشد، تفسیر و بررسی نتایج با مشکلاتی روبه‌رو خواهد شد. یافته‌های بسیاری از پژوهش‌های تجربی، از قبیل بوگن و دروری [۲۴]، فرکا و هاپ‌وود [۳۴] و محمودآبادی و برزگر [۱۳] حاکی از نرمال نبودن نسبت‌های مالی مورد بررسی است.

با توجه به مشکلات روش‌های خطی آماری، پژوهش‌های زیادی بر استفاده از شیوه‌های هوش مصنوعی (شبکه‌های عصبی، ماشین‌بردار پشتیبان، الگوریتم ژنتیک، تحلیل پوششی داده‌ها، شبکه‌های بیز، جنگل‌های تصادفی و ...) و مقایسهٔ عملکرد آن‌ها با روش‌های آماری تأکید داشت و به طور کلی، یافته‌های اغلب این پژوهش‌ها، حاکی از برتری روش‌های هوش مصنوعی نسبت به مدل‌های خطی آماری است.

به‌طور کلی، روش‌های غیرخطی از قبیل شبکه‌های عصبی و جنگل‌های تصادفی چندین مزیت مهم در مقایسه با مدل‌های آماری از قبیل رگرسیون خطی دارند. معایب رگرسیون خطی نسبت به روش‌های غیرخطی به شرح زیر است [۲۸]:

- *ماهیت خطی رگرسیون:* یک عیب مهم رگرسیون‌های خطی این است که رگرسیون هیچ شاخص مستقیمی را مبنی بر اینکه آیا داده‌ها در حالت خطی به بهترین صورت نشان داده می‌شود، ارائه نمی‌کند. با توجه به ماهیت علوم اجتماعی، در بسیاری از حالت‌ها، تحلیل آماری خطی نامناسب است.
- *از پیش مشخص کردن مدل:* استفاده از مدل‌های رگرسیون، مستلزم از پیش مشخص کردن مدل پایه است. این کار باعث حل آسان‌تر مسأله می‌شود، اما نیازمند حدس‌های زیاد است.
- *مفروضات رگرسیون:* عملکرد مدل‌های رگرسیون خطی وابسته به مفروضات گوناگونی از

قبیل نبود روابط خطی چندگانه و توزیع نرمال باقیمانده‌هاست.

- *عدم انطباق‌پذیری:* رگرسیون چندمتغیره در حالتی که اجزای مدل را نتوان به‌وسیلهٔ حدس مشخص کرد، دارای خاصیت انطباق‌پذیری با داده‌ها نیست. این مورد بدان معنست که در رگرسیون باید مدل کلی پژوهش و اجزای آن از قبل مشخص باشد؛ در غیر این‌صورت با توجه به داده‌های موجود، ممکن است مدل به دست آمده بهترین مدل برای پیش‌بینی متغیر وابسته نباشد و مدل ارائه‌شده و نتایج حاصل دارای سوگیری باشد.

با توجه به ضعف روش‌های خطی و مزایای روش‌های غیرخطی در پیش‌بینی، به‌عنوان نمونه قابلیت انطباق بیشتر با مسائل جهان واقعی، عملکرد پیش‌بینی بهتر و عدم وابستگی به مفروضات خاص [۳۶]، پژوهش حاضر به پیش‌بینی بحران مالی با استفاده از طبقه‌بندی‌کنندهٔ غیرخطی جنگل‌های تصادفی می‌پردازد.

به رغم انجام پژوهش‌های زیادی با استفاده از روش‌های غیرخطی برای پیش‌بینی بحران مالی، تأکید بیشتر پژوهش‌های انجام شده بر انتخاب مدل‌های بهینه برای پیش‌بینی بوده و کمتر بر انتخاب متغیرهای بهینه برای پیش‌بینی تأکید شده است. مرحلهٔ انتخاب متغیرهای پیش‌بین، عموماً قبل از آموزش مدل‌های پیش‌بینی انجام می‌شود. با وجود این، در اغلب پژوهش‌های داخلی و خارجی انجام‌شده در حسابداری، این مرحله نادیده گرفته شده و متغیرهای پیش‌بین به‌صورت نظام‌مند انتخاب نشده است. این امر به انتخاب متغیرهای پیش‌بین غیربهینه و در برخی موارد، متغیرهای پیش‌بین نامناسب منجر می‌شود [۶۱].

انتخاب و استخراج متغیرهای مناسب به‌منظور رسیدن به بهترین نتیجه در پیش‌بینی، از مباحث چالش برانگیز در دو دههٔ اخیر بوده است. از دیدگاه نظری، یادگیری براساس تعداد متغیرهای پیش‌بین بیشتر باعث می‌شود تا دقت پیش‌بینی بالا رود. با وجود این، شواهد تجربی بیانگر آن است که این امر همواره صادق نیست؛ زیرا تمام متغیرها، برای تشخیص و پیش‌بینی مهم نیستند و یا برخی از آن‌ها به‌طور کلی در پیش‌بینی نامربوط هستند [۴۷]. با توجه به اینکه عامل‌های بسیاری از جمله کیفیت داده‌ها در موفقیت یک الگوریتم یادگیری مؤثر است، اگر داده‌ها حاوی متغیرها و یا اطلاعات تکراری و نامربوط[1] باشند و یا حاوی اطلاعات دارای پارازیت و و نامطمئن باشند، محتوای اطلاعاتی آن داده‌ها مورد شک و تردید قرار می‌گیرد [۳۵]. افزون بر این، کاهش تعداد متغیرهای پیش‌بین نامربوط یا اضافی، علاوه بر کاهش زمان اجرای الگوریتم‌های آموزشی، به مفهومی عمومی‌تر منجر می‌شود. سایر مزایای بالقوه انتخاب و استخراج متغیرهای پیش‌بین شامل تسهیل درک و تجسم داده‌ها، کاهش الزامات اندازه‌گیری و ذخیرهٔ اطلاعات، کاهش اضافه‌بار ابعاد[2] ابعاد[2] و بهبود عملکرد پیش‌بینی و فراهم کردن بینش بهتر در مورد مفهوم زیربنایی از طبقه‌بندی دنیای واقعی است. اضافه‌بار ابعاد به مجموعه‌ای از مشکلات اشاره دارد که در تحلیل داده‌ها در ابعاد بالا رخ می‌دهد و در ابعاد کوچک (همانند سه بعد فیزیکی)، متفاوت است. این مشکلات از جهات مختلف مانند نمونه‌گیری، تحلیل عددی، یادگیری ماشین و ... قابل بررسی است. معمول‌ترین مشکل

[1] Redundant
[2] Curse of Dimensionality

این است که با افزایش ابعاد (تعداد متغیرها)، حجم فضا به‌سرعت افزایش می‌یابد و تعداد داده‌ها به نسبت فضا اندک[1] می‌گردد. برای رفع این مشکل نیاز است که همزمان با افزایش ابعاد، تعداد مشاهدات نیز به‌صورت نمایی افزایش یابد [۶۱]. به‌طور کلی، اندک بودن منطقی متغیرهای پیش‌بین و دقت بالای پیش‌بینی (و اندک بودن خطای نوع اول و دوم) از مهم‌ترین معیارهای کیفیت یک مدل پیش‌بینی محسوب می‌شود [۱].

با توجه به اینکه یکی از مشکلات اصلی در پیش‌بینی بحران مالی، وجود تعداد زیاد متغیرهای پیش‌بین بالقوهٔ قابل استفاده و نبود توافق جامع در خصوص متغیرهای بهینه است، این امکان وجود دارد که برخی از متغیرهای مورد استفاده حاوی اطلاعات دارای پارازیت بوده و در نتیجه عملکرد پیش‌بینی تحت تأثیر قرار گیرد [۳۹ و ۶۱]. در اغلب پژوهش‌های انجام شده، متغیرهای پیش‌بین، بدون ضابطه و صرفاً براساس مطالعات گذشته انتخاب شده است که این روش دارای معایب زیادی است؛ به‌عنوان نمونه، فرض می‌شود متغیری که در پژوهش خاصی، مناسب بوده، همواره مناسب خواهد بود. با وجود این، در بیشتر مواقع، متغیرها زمانی که در نمونه دیگر یا با روش متفاوتی استفاده شود، ویژگی‌های اولیه خود را از دست می‌دهد. در این زمینه، استفاده از روش‌های نظام‌مند انتخاب متغیرهای پیش‌بین بهینه، مزیت دارد [۳۹]. انتخاب متغیرهای بهینه به عنوان مرحلهٔ پیش‌پردازش (قبل از انجام پیش‌بینی)، افزون بر فیلتر کردن متغیرهای نامربوط از داده‌های اولیه، منجر به بهبود عملکرد پیش‌بینی می‌شود [۶۱]. بنابراین، می‌توان مشکل اصلی را در روش نادرست انتخاب متغیر برای پیش‌بینی بحران مالی شرکت‌ها دانست. به منظور حل این مشکل در این پژوهش از روش نظام‌مند انتخاب متغیر ریلیف به منظور انتخاب متغیرهای بهینه استفاده شده و به بررسی این موضوع پرداخته شده است که آیا متغیرهای منتخب روش مزبور از بین متغیرهای اولیه، تأثیر مثبت و معناداری بر عملکرد پیش‌بینی بحران مالی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران دارد یا خیر. روش انتخاب متغیر ریلیفیک روش رتبه‌بندی و انتخاب متغیر است که به‌صورت فیلتر[2] استفاده می‌شود و برای انتخاب متغیر در داده‌های با دو طبقه (مثلا درماندهٔ مالی در مقابل سالم) مورد استفاده قرار می‌گیرد. روش‌های انتخاب متغیری که به‌صورت فیلتر عمل می‌کنند، بر اساس رابطهٔ ریاضی بین داده‌های آموزشی و فارغ از طبقه‌بندی‌کننده، به انتخاب متغیر می‌پردازند. سرعت این روش‌ها اغلب نسبت به روش‌های مبتنی بر رپر[3] که با استفاده از یک طبقه‌بندی‌کننده به انتخاب متغیر می‌پردازند، بالاتر است [۲۰].

پیشینهٔ پژوهش

پیشینهٔ داخلی

در اغلب پژوهش‌های انجام شده در بورس اوراق بهادار تهران از روش خطی برای پیش‌بینی بحران مالی و ورشکستگی استفاده شده است. سلیمانی‌امیری [۱۱] با استفاده از مدل رگرسیون چندگانه و پنج نسبت مالی، به پیش‌بینی بحران مالی پرداخت. یافته‌های پژوهش حاکی از آن بود که مدل ارائه‌شده قادر است تا سه سال قبل از بحران مالی، پیش‌بینی

[1] Sparse

[2] Filter
[3] Wrapper

صحیحی از بحران مالی ارائه دهد. مهرانی و همکاران [۱۷] کاربرد الگوهای تشخیص درماندگی مالی شیراتا و زیمسکی را در بورس اوراق بهادار تهران در دو صنعت داروسازی و نساجی بررسی کردند. نتیجهٔ پژوهش برای الگوی شیراتا پیش‌بینی صحیح ۹۴/۷٪ و برای الگوی زیمسکی ۹۷/۴٪ به‌دست آمد. رهنمای رودپشتی و همکاران [۹] به بررسی کاربرد مدل‌های پیش‌بینی ورشکستگی آلتمن و فالمر پرداختند. نتایج به‌دست آمده حاکی از آن بود که در پیش‌بینی یک شرکت، تفاوت معناداری بین نتایج دو مدل وجود دارد. مدل آلتمن در پیش‌بینی ورشکستگی، محافظه‌کارانه‌تر از مدل فالمر عمل می‌کند. پورحیدری و کوپایی [۳] به پیش‌بینی بحران مالی شرکت‌ها با استفاده از مدل مبتنی بر تابع تفکیکی خطی پرداختند. نتایج بررسی نشان داد که تا پنج سال قبل از بحران مالی می‌توان با استفاده از مدل، با دقت نسبتاً بالا آن را پیش‌بینی کرد. یافته‌های پژوهش دستگیر و همکاران [۶] بیانگر آن بود که شرکت‌های درماندهٔ مالی، سودهای خود را در سه سال قبل از ورشکستگی به شکل افزایشی مدیریت می‌کنند و شرکت‌های درماندهٔ مالی بیشتر از شرکت‌های سالم از طریق فعالیت‌های واقعی سودهای خود را مدیریت می‌کنند، در حالی که شرکت‌های سالم این کار را بیشتر از طریق اقلام تعهدی انجام می‌دهند. همچنین، میزان محافظه‌کاری مشروط در شرکت‌های درمانده بیشتر از شرکت‌های غیردرمانده بود. پناهی و همکاران [۲] مدلی برای پیش‌بینی ورشکستگی ارائه کردند. در این مدل از نسبت‌های مالی الگوی آلتمن به همراه نسبت جاری استفاده شده است. برآورد مدل به سه روش مدل احتمال خطی، مدل لاجیت و مدل پروبیت صورت گرفته است. براساس اطلاعات سال ۱۳۸۲ مدل برآوردشده توانسته است با دقت ۷۸ درصد وضعیت شرکت‌های یادشده در سال ۱۳۸۷ را درست پیش‌بینی کند.

با مشخص شدن معایب روش خطی و مزایای روش‌های غیرخطی، پژوهش‌هایی نیز با استفاده از شیوه‌های هوش مصنوعی در بورس اوراق بهادار تهران انجام شده است. راعی و فلاح‌پور [۸]، مکیان و همکاران [۱۴] و نیکبخت و شریفی [۱۸] به پیش‌بینی درماندگی مالی شرکت‌ها با استفاده از شبکه‌های عصبی مصنوعی پرداختند. یافته‌های پژوهش حاکی از عملکرد بهتر شبکه‌های عصبی مصنوعی نسبت به روش خطی بود. یافته‌های پژوهش سعیدی و آقایی [۱۰] حاکی از برتری شبکه بیز نسبت به مدل رگرسیون لجستیک بود. یافته‌های پژوهش موسوی شیری و طبرستانی [۱۵] حاکی از آن بود که مدل طراحی‌شده با استفاده از تحلیل پوششی داده‌ها قابلیت پیش‌بینی وقوع درماندگی مالی در شرکت‌های تولیدی پذیرفته شده در بورس اوراق بهادار تهران را تا دو سال قبل از وقوع آن دارد. یافته‌های پژوهش فدایی‌نژاد و اسکندری [۱۲] حاکی از آن بود که استفاده از الگوریتم ژنتیک در افزایش دقت پیش‌بینی ورشکستگی مؤثر است، ولی مقایسهٔ مدل‌های الگوریتم ژنتیک و بهینه‌سازی تجمعی ذرات نشان داد که از نظر آماری نمی‌توان اثبات کرد که یکی از این روش‌ها بر دیگری برتری دارد. یافته‌های پژوهش حسینی و رشیدی [۵] حاکی از عملکرد بهتر رگرسیون لجستیک نسبت به درخت تصمیم بود.

با بررسی پیشینهٔ پژوهش، تأکید بیشتر پژوهش‌های انجام شده بر انتخاب مدل‌های مناسب پیش‌بینی بوده و کمتر به انتخاب متغیرهای بهینه برای پیش‌بینی تأکید شده است. به طور کلی، یافته‌های

اغلب این پژوهش‌ها، حاکی از برتری روش‌های هوش مصنوعی نسبت به مدل‌های خطی آماری است. با این وجود، پژوهشی که با استفاده از جنگل‌های تصادفی به پیش‌بینی بحران مالی در شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران پرداخته باشد، مشاهده نشد. همچنین پژوهشی مشاهده نگردید که با استفاده از روش ریلیف به انتخاب متغیرهای پیش‌بین پرداخته باشد. با توجه به کاستی‌های پژوهشی موجود در بورس اوراق بهادار تهران و تأکید پژوهش‌ها بر مدل‌های پیش‌بینی و انتخاب متغیرها صرفاً براساس پژوهش‌های گذشته، پژوهش حاضر به بررسی عملکرد جنگل‌های تصادفی در پیش‌بینی بحران مالی و همچنین سودمندی روش انتخاب متغیر ریلیف در این زمینه می‌پردازد.

پیشینهٔ خارجی

شیوه‌های آماری تک‌متغیری، از اولین شیوه‌هایی بودند که برای پیش‌بینی ورشکستگی و بحران مالی استفاده شد. در سال ۱۹۶۶ بیور به منظور ارزیابی توان نسبت‌های مالی در پیش‌بینی بحران مالی از تحلیل تک‌متغیری استفاده کرد. یافته‌های پژوهش بیور حاکی از آن بود که تفاوت معناداری بین نسبت‌های مالی شرکت‌های درماندهٔ مالی و سالم وجود دارد. آلتمن [۱۹] تحلیل تک‌متغیره را مورد انتقاد قرار داد و تحلیل ممیزی چندگانه را که در آن چندین نسبت مالی، همزمان در پیش‌بینی ورشکستگی بررسی می‌شد، پیشنهاد کرد. اولسون [۵۴] برای ایجاد الگوی خود از روش تجزیه و تحلیل لجستیک استفاده کرد. وی در مدل خود از ۹ متغیر مستقل استفاده کرد.

با توجه به مشکلات روش‌های خطی آماری، پژوهش‌های زیادی بر استفاده از شیوه‌های هوش مصنوعی (شبکه‌های عصبی، ماشین‌بردار پشتیبان، الگوریتم ژنتیک و ...) و مقایسهٔ عملکرد آن‌ها با روش‌های آماری تأکید داشت. ادوم و شارادا [۵۳] برای اولین بار از شبکه‌های عصبی مصنوعی برای پیش‌بینی ورشکستگی استفاده کردند. یافته‌های پژوهش آنان حاکی از دقت و توان پیش‌بینی بهتر شبکه‌های عصبی مصنوعی نسبت به تحلیل ممیزی چندگانه بود. مینولی [۵۱] با استفاده از ماشین بردار پشتیبان، به پیش‌بینی ورشکستگی شرکت‌ها پرداختند. یافته‌های پژوهش آن‌ها نشان داد که ماشین بردار پشتیبان نسبت به مدل‌های آماری سنتی از عملکرد بهتری برخوردار است. یافته‌های پژوهش شین و همکاران [۵۹] نیز حاکی از عملکرد بهتر ماشین بردار پشتیبان نسبت به شبکه‌های عصبی مصنوعی بود. پژوهش‌های زیادی با استفاده از سایر روش‌های هوش مصنوعی از قبیل جنگل‌های تصادفی [۳۸، ۵۲]؛ الگوریتم ژنتیک [۵۸]؛ تحلیل پوششی داده‌ها [۳۲]، شبکه‌های بیز [۵۷] پرداختند. یافته‌های اغلب این پژوهش‌ها حاکی از برتری روش‌های هوش مصنوعی نسبت به مدل خطی آماری بود.

تأکید بیشتر پژوهش‌های انجام‌شده بر انتخاب مدل‌های بهینه برای پیش‌بینی بوده و کمتر به انتخاب متغیرهای بهینه برای پیش‌بینی تأکید شده است. به طور کلی، یافته‌های اغلب این پژوهش‌ها، حاکی از برتری روش‌های هوش مصنوعی نسبت به مدل‌های خطی آماری است. تسای [۶۱] به مقایسهٔ پنج روش متداول انتخاب متغیر مورد استفاده در پیش‌بینی ورشکستگی (شامل آزمون t، ماتریس همبستگی[1]، رگرسیون گام به گام[2]، تحلیل مؤلفه‌های اصلی[3] و

[1] Correlation Matrix
[2] Stepwise Regression
[3] Principle Component Analysis (PCA)

تحلیل عاملی[1]) پرداخت. یافته‌های پژوهش حاکی از سودمندی روش‌های انتخاب متغیر و تفاوت معنادار بین آن‌ها بود. ونگ و همکاران [۶۲] به بررسی سودمندی انتخاب متغیرهای پیش‌بین در پیش‌بینی ورشکستگی پرداختند. یافته‌های پژوهش حاکی از سودمندی متغیرهای انتخاب شده در پیش‌بینی بود. لیانگ و همکاران [۴۶] به بررسی سودمندی انتخاب متغیرهای پیش‌بین بر پیش‌بینی بحران مالی پرداختند. یافته‌های این پژوهش حاکی از آن بود که در اغلب موارد، انتخاب متغیرهای پیش‌بین باعث بهبود عملکرد پیش‌بینی می‌شود.

فرضیه‌های پژوهش

با توجه به مبانی نظری و پیشینه پژوهش، دو فرضیهٔ زیر تدوین و آزمون شده است؛ فرضیهٔ اول به بررسی سودمندی روش غیرخطی جنگل‌های تصادفی (نسبت به روش خطی) در پیش‌بینی بحران مالی می‌پردازد. فرضیهٔ دوم نیز سودمندی روش انتخاب متغیر ریلیف و قدرت پیش‌بینی متغیرهای منتخب آن (نسبت به استفاده از ۶۹ متغیر اولیه) در پیش‌بینی بحران مالی را ارزیابی می‌کند.

۱. تفاوت معناداری بین عملکرد جنگل‌های تصادفی و رگرسیون لجستیک برای پیش‌بینی بحران مالی در زمان استفاده از ۶۹ متغیر اولیه وجود دارد.

۲. تفاوت معناداری بین عملکرد جنگل‌های تصادفی و رگرسیون لجستیک برای پیش‌بینی بحران مالی در زمان استفاده از متغیرهای منتخب ریلیف وجود دارد.

۳. تفاوت معناداری بین عملکرد پیش‌بینی جنگل‌های تصادفی در زمان استفاده از متغیرهای منتخب روش ریلیف و استفاده از کلیهٔ متغیرهای اولیه وجود دارد.

۴. تفاوت معناداری بین عملکرد پیش‌بینی رگرسیون لجستیک در زمان استفاده از متغیرهای منتخب روش ریلیف و استفاده از کلیهٔ متغیرهای اولیه وجود دارد.

روش انجام پژوهش

جامعه و نمونهٔ آماری پژوهش

جامعهٔ آماری این پژوهش، کلیهٔ شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران طی دورهٔ زمانی ۱۳۸۰ تا ۱۳۹۲ است. با توجه به اینکه ملاک درماندگی مالی در این پژوهش، ماده۱۴۱ قانون تجارت شرکت‌هاست، ابتدا فهرستی از شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران که بین سال‌های ۱۳۸۰ تا ۱۳۹۲ دچار درماندگی مالی شده بودند (مشمول مادهٔ ۱۴۱ قانون تجارت بودند) تهیه شد. از بین این فهرست، ۹۵ شرکت – سال تولیدی که اطلاعات مورد نیاز آن‌ها در دسترس بود، به‌عنوان شرکت‌های درماندهٔ مالی انتخاب شد. در ادامه، ۹۵ شرکت – سال سالم (بدونِ درماندگی مالی) از بین کلیهٔ شرکت‌های تولیدی و سالم پذیرفته شده در بورس اوراق بهادار تهران که کلیهٔ اطلاعات مورد نیاز برای انجام پژوهش را به بورس اوراق بهادار ارائه کرده بودند، به صورت تصادفی انتخاب شد.

روش گردآوری داده‌ها و اطلاعات

در این پژوهش برای جمع‌آوری داده‌ها و اطلاعات از روش‌های کتابخانه‌ای و میدانی استفاده - شده است. مبانی نظری پژوهش از کتاب‌ها، مجلات و سایت‌های تخصصی فارسی و لاتین گردآوری شده و داده‌های مالی مورد نیاز با مراجعه به سایت سازمان بورس اوراق بهادار تهران (www.codal.ir)، صورت‌های

[1] Factor Analysis (FA)

مالی شرکت‌ها و همچنین با استفاده از نرم‌افزارهای تدبیرپرداز و ره‌آورد نوین گردآوری شده است. در مرحلهٔ اول با بررسی متون و پیشینهٔ پژوهش، حدود ۱۵۰ متغیر پیش‌بین (مستقل) شناسایی شد. از بین متغیرهای شناسایی شده، ۶۹ متغیری که بیشتر در ادبیات ورشکستگی استفاده شده و داده‌های مورد نیاز برای سنجش آن‌ها از طریق پایگاه‌های اطلاعاتی سازمان بورس و اوراق بهادار و همچنین نرم‌افزارهای تدبیرپرداز و ره‌آورد نوین در دسترس بود، انتخاب شد. در این راستا، سعی شد که ابعاد سودآوری، کارایی، اهرم مالی، نقدینگی، نسبت‌های مبتنی بر هر سهم، نسبت‌های مبتنی بر جریان وجوه نقد و نسبت‌های بازار در نظر گرفته شود. نگارهٔ (۱)، میانگین این متغیرها را در شرکت‌های درماندهٔ مالی و غیردرماندهٔ (سالم) نشان می‌دهد. در ادامه، با استفاده از روش انتخاب متغیر ریلیفاز بین ۶۹ متغیر ذکر شده، متغیرهای بهینه، انتخاب شده است. این روش که یکی از رایج‌ترین رویه‌های کاهش ابعاد یک مسئله است، منجر به انتخاب متغیرهای بهینه از بین متغیرهای اولیه می‌شود.

نگارهٔ ۱. متغیرهای استفاده‌شده و مقایسهٔ میانگین آن‌ها در شرکت‌های درماندهٔ مالی و سالم

#	متغیر	منبع	میانگین درمانده	میانگین غیردرمانده	#	متغیر	منبع	میانگین درمانده	میانگین غیردرمانده
۱	(Ca+STI)/CL	[۳۱]	۰/۰۵۲	۰/۱۵۲	۲	NI/SE	[۶۴]	−۰/۰۸۲	۰/۳۶۸
۳	(R+Inv)/TA	[۳۱]	۰/۶۱۴	۰/۵۰۹	۴	NI/TA	[۱۹]	−۰/۰۱۱	۰/۱۷۱
۵	P/S	[۲۱]	۰/۴۱۸	۰/۲۰۴	۶	OCF	[۴۰]	۴۵۱۵/۴۶۵	۶۰۶۲۷۵/۲
۷	R/S	[۶۴]	۰/۵۳۲	۰/۳۱۸	۸	OCF/SE	[۴۱]	۰/۱۹۱	۰/۴۵۲
۹	Ca/CL	[۶۴]	۰/۰۳۳	۰/۰۷۶	۱۰	OCF/CL	[۶۰]	۰/۰۴۵	۰/۴۷۴
۱۱	Ca/TA	[۲۷]	۰/۰۲۴	۰/۰۳۶	۱۲	OCF/IE	[۴۲]	۱/۰۳۱	/۴۸۶۶۸/۷۲
۱۶	CGS/Inv	[۴۵]	۲/۴۲۶	۳۴۶/۹۱۱	۱۴	OCF/S	[۳]	۰/۰۸۴	۰/۲۴۴
۱۵	CA/CL	[۵۴]	۱/۰۰۸	۱/۳۹۵	۱۶	OCF/TA	[۴۰]	۰/۰۳۵	۰/۱۸۶
۱۷	CA/S	[۲۷]	۱/۲۰۲	۰/۷۶۲	۱۸	OCF/TL	[۴۰]	۰/۰۳۸	۰/۳۹۲
۱۹	CA/TA	[۲۷]	۰/۶۸۵	۰/۶۲۹	۲۰	OCF/NI	[۴۹]	۷/۶۵۶	۱/۰۲۵
۲۱	CL/SE	[۶۴]	۷/۷۲۲	۱/۶۲۴	۲۲	OCF/OI	[۴۹]	−۰/۷۵۸	۱/۲۳۱
۲۳	CL/TA	[۵۸]	۰/۶۵۶	۰/۴۹۴	۲۴	ORPS	[۶۰]	۵۲۴۸/۲۱۶	۷۱۹۶/۵۲۴
۲۵	CL/TL	[۶۰]	۰/۸۴۵	۰/۸۵۹	۲۶	OI/S	[۳]	۰/۰۲۴	۰/۲۶۵
۲۷	D/CS	[۴۵]	۰/۱۲۹	۰/۷۴۸	۲۸	OI/TA	[۴۵]	۰/۰۲۴	۰/۲۱۷
۲۹	D/NI	[۴۹]	۰/۵۳۲	۰/۱۸۸	۳۰	PIC/SE	[۳۱]	۰/۹۶۸	۰/۴۷۲
۳۱	EPS	[۴۱]	−۲۱۵/۶۸۵	۸۵۵/۲۶۴	۳۲	P/OCF	[۲۰]	−۰/۶۶۲	۶/۶۳۴
۳۳	EBIT/IE	[۵۱]	−۲۸/۸۴۷	۴۸۱۶/۴۵۲	۳۴	QA/CL	[۲۷]	۰/۴۶۲	۰/۷۷۸
۳۵	EBIT/S	[۵۱]	−۰/۱۴۲	۰/۲۴۱	۳۶	QA/Inv	[۲۱]	۲/۳۶۸	۳۰۹۵/۱۲
۳۷	EBIT/TA	[۶۴]	−۰/۰۵۶	۰/۱۶۲	۳۸	QA/TA	[۲۷]	۰/۳۱۷	۰/۳۵۱
۳۹	FA/(SE+LTD)	[۵۱]	۰/۸۷۵	۰/۵۵۷	۴۰	R/Inv	[۳۱]	۵/۳۶۲	۱۶۳/۳۷۵

#	متغیر	منبع	میانگین درمانده	میانگین غیردرمانده	#	متغیر	منبع	میانگین درمانده	میانگین غیردرمانده
۴۱	FA/TA	[۶۰]	۰/۲۳۵	۰/۲۷۲	۴۲	RE/Inv	[۲۶]	۵/۲۰۳	۳۱۳۲/۲۳
۴۳	GP/S	[۶۴]	۰/۱۳۶	۰/۳۴۵	۴۴	RE/SC	[۶۰]	۰/۰۸۷	۱/۳۲۷
۴۵	IE/GP	[۳۱]	۱/۵۴۶	۰/۱۴۸	۴۶	RE/TA	[۱۹]	−۰/۰۲۵	۰/۲۱۷
۴۷	IE/S	[۵۱]	۰/۱۱۲	۰/۰۴۲	۴۸	S/Ca	[۴۱]	۱۱۸/۴۲۲	۶۷/۵۲۴
۴۹	IE/TE	[۴۱]	۰/۰۹۸	۰/۰۳۷	۵۰	S/FA	[۵۱]	۴/۴۶۲	۵/۴۱۷
۵۱	Inv/WC	[۲۹]	−۱/۲۷۴	۰/۱۴۹	۵۲	S/SE	[۴۵]	۴/۱۷۱	۲/۴۲۷
۵۳	Inv/S	[۴۲]	۰/۵۷۴	۰/۲۸۲	۵۴	S/TA	[۱۹]	۰/۶۴۲	۰/۸۶۷
۵۵	LTD/SE	[۶۴]	۰/۶۷۵	۰/۱۸۵	۵۶	SE/TA	[۵۱]	۰/۲۰۵	۰/۴۴۱
۵۷	LTD/TA	[۴۹]	۰/۱۲۲	۰/۰۶۸	۵۸	SE/TL	[۴۰]	۰/۲۶۱	۰/۹۷۳
۵۹	MVE/TA	[۵۶]	۰/۴۰۱	۰/۷۶۳	۶۰	Size(log TA)	[۵۴]	۱۲/۲۱۲	۱۴/۳۱۲
۶۱	MVE/TL	[۵۶]	۰/۵۴۲	۱/۷۲۱	۶۲	TIBL/TL	[۲۲]	۰/۱۵۲	۰/۱۲۷
۶۳	MVE/SE	[۵۶]	۲/۱۹۴	۱/۷۴۲	۶۴	TL/TA	[۴۵]	۰/۸۱۲	۰/۵۴۱
۶۵	NAPS	[۶۰]	۰/۰۰۱	۰/۰۰۲	۶۶	WC/S	[۲۷]	−۰/۲۱۴	۰/۱۴۷
۶۷	NI/GP	[۲۹]	−۱/۸۱۸	۰/۶۲۷	۶۸	WC/TA	[۵۴]	−۰/۰۰۵	۰/۱۴۱
۶۹	NI/S	[۶۴]	−۰/۰۳۴	۰/۲۲۱					

CA: دارایی‌های جاری، NI: سود خالص، Ca: موجودی نقد، OI: سود عملیاتی، CL: بدهی‌های جاری، QA: دارایی‌های آنی، PIC: سرمایهٔ پرداخت شده، R: حساب‌ها و اسناد دریافتنی، EBIT: سود قبل بهره و مالیات، RE: سود انباشته، FA: دارایی‌های ثابت، S: درآمدها، GP: سودناخالص، SC: سرمایه، IE: هزینه‌های مالی، SE: حقوق صاحبان سهام، Inv: موجودی‌ها، STI: سرمایه‌گذاری‌های کوتاه‌مدت، TA: مجموع دارایی‌ها، LTD: بدهی‌های بلندمدت، TL: مجموع بدهی‌ها، MVE: ارزش بازار سهام، WC: سرمایه در گردش، OCF: جریان نقد عملیاتی، D: سود تقسیمی، TIBL: مجموع بدهی‌های بهره‌دار، NAPS: خالص دارایی‌های هر سهم، ORPS: درآمد عملیاتی هر سهم

منبع: یافته‌های پژوهشگر

متغیر وابسته این پژوهش نیز درماندگی مالی است که به‌صورت یک متغیر مجازی (یک در صورت درماندهٔ مالی بودن و صفر در صورت سالم بودن) مورد سنجش قرار گرفت. مبنای درماندهٔ مالی یا سالم بودن شرکت‌ها، شرط موجود در ماده ۱۴۱ قانون تجارت است. بر اساس این ماده، «اگر بر اثر زیان‌های وارده، حداقل نصف سرمایهٔ شرکت از میان برود، هیئت مدیره مکلف است بلافاصله مجمع عمومی فوق‌العادهٔ صاحبان سهام را دعوت کند تا موضوع انحلال یا بقای شرکت مورد شور و رأی واقع شود. هرگاه مجمع مزبور رأی به انحلال شرکت ندهد، باید در همان جلسه و با رعایت مقررات ماده ۶ این قانون، سرمایهٔ شرکت را به مبلغ سرمایهٔ موجود کاهش دهد.» بنابراین، اگر شرکتی، مشمول شرط این ماده باشد، به‌عنوان درماندهٔ مالی و در غیر این صورت به‌عنوان بدونِ درماندگی مالی (سالم) قلمداد خواهد

شد. این معیار در اغلب پژوهش‌های داخلی، مانند راعی و فلاح‌پور [۷ و ۸]، سعیدی و آقایی [۱۰]، موسوی‌شیری و طبرستانی [۱۵]، مکیان و همکاران [۱۴]، نیکبخت و شریفی [۱۸]، پورحیدری و کوپایی [۳] و اعتمادی و همکاران [۳۱] نیز استفاده شده است.

طبقه‌بندی‌کننده‌های استفاده‌شده

طبقه‌بندی‌کنندهٔ جنگل‌های تصادفی

در دهه‌های اخیر پژوهش‌های زیادی در خصوص یادگیری تجمیعی[1] انجام شده است. یادگیری تجمیعی به روش‌هایی اشاره دارد که چندین مدل به منظور پیش‌بینی، ترکیب می‌شود. این رویکرد، بخش قابل توجهی از پژوهش‌های اخیر را به خود اختصاص داده و نتایج خوبی از آن گزارش شده است. مزیت رویکرد تجمیعی نسبت به مدل‌های انفرادی[2] شامل افزایش دقت و پایداری[3] است. از مزایای دیگر رگرسیون تجمیعی سادگی در پیاده‌سازی و ترکیب چند روش پیش‌بینی است که ترکیب آن‌ها باعث ایجاد یک روش غیرخطی می‌شود. همچنین تجمیع هر روش پیش‌بینی دلخواه است. افزون بر این، روش مزبور، کاهش میزان واریانس خطا و میزان سوگیری را در پی خواهد داشت. در روش‌های تجمیعی، چند مدل انفرادی ترکیب می‌شود تا یک مدل جدید واحد تشکیل شود. عموماً این مورد با گرفتن میانگین یا میانگین موزون الگوهای انفرادی انجام می‌شود، اما رویه‌های ترکیب دیگر نیز امکان پذیر است [۵۰].

طبقه‌بندی در جنگل‌های تصادفی، از طریق ترکیب (تجمیع) تصادفی درخت‌های CART انجام می‌شود [۶۵]. درخت تصمیم یکی از روش‌های ناپارامتری طبقه‌بندی است. این روش با به کارگیری شیوه‌های ساده، یک الگوی طبقه‌بندی را برای مشاهدات موجود ارائه می‌کند. الگوی معرفی شده به‌وسیلهٔ این روش، از ساختاری ساده و قابل درک برای تصمیم‌گیری برخوردار است. درخت تصمیم یک روش ساده و توانمند برای طبقه‌بندی است که یک گراف غیرچرخشی شبیه درخت دارد که این درخت با مجموعه‌ای از سؤال‌ها نشان داده می‌شود. معمولاً هر سؤال با توجه به یک متغیر مطرح می‌شود. یک گراف درخت تصمیم از سه جزء اصلی ریشه[4]، گره داخلی[5] و گره خارجی[6] (برگ) تشکیل شده است و روند بدین گونه است که ابتدا یک متغیر کمکی به عنوان ریشه انتخاب می‌گردد و با توجه به یک سری از سؤال‌ها و ویژگی‌ها به چندین گره داخلی تقسیم می‌شود. روش CART یکی از انواع درخت تصمیم است که یک گراف غیرچرخشی شبیه درخت با تقسیم‌های دوتایی بر اساس متغیرهای کمکی را برای معرفی یک الگوی رده‌بندی و تشخیصی معرفی می‌کند. در روش CART فرآیند بدین‌گونه است که ابتدا یک متغیر کمکی به عنوان ریشه انتخاب و با توجه به اهداف مطالعه به چندین گره داخلی تقسیم می‌شود. هر گره داخلی نیز مانند ریشه به گره دیگری تقسیم می‌شود تا در نهایت به هر گره یک رده از متغیر پاسخ منتسب گردد. این گره‌ها برگ نامیده می‌شود. به منظور انتخاب متغیرهای مهم در الگوی رده‌بندی درختی، از شاخص جینی استفاده می‌شود. به‌طور کلی، درخت

[1] Ensemble
[2] Single models
[3] Robustness
[4] Root
[5] Internal Node
[6] External Node (Leaf)

تصمیم انفرادی مستعد بیش‌برازش[1] بوده و قدرت تعمیم‌پذیری اندکی دارد. از معایب دیگر درخت تصمیم انفرادی می‌توان به ناپایداری نتایج حاصل از آن نسبت به وجود نویز در داده‌های ورودی اشاره کرد. در هنگام تشکیل یک درخت تصمیم، تغییر کوچکی در الگوهای یادگیری می‌تواند باعث تغییرات اساسی در ساختار آن درخت گردد. برای حل این مشکلات، الگوریتم جنگل‌های تصادفی که یک روش یادگیری تجمیعی مبتنی بردسته‌ای از درخت‌های تصمیم است، پیشنهاد شده است [۲۵].

الگوریتم جنگل‌های تصادفی از ترکیبی از درخت‌های تصمیم مستقل برای مدل‌سازی داده‌ها و ارزیابی اهمیت متغیرها استفاده می‌کند. هر درخت تصمیم در یک جنگل با استفاده از نمونه‌ای خودسازمانده[2] از داده‌ها تشکیل می‌شود. در درخت استاندارد هر گره از طریق بهترین انشعاب[3] (تجزیه) از بین تمام متغیرها تفکیک می‌شود. در جنگل‌های تصادفی، هر گره از طریق بهترین زیرمجموعه از متغیرهایی که به صورت تصادفی در هر گره انتخاب شده است، تفکیک می‌شود. این راهبرد در مقایسه با راهبرد بسیاری از طبقه‌بندی‌کننده‌ها از قبیل تحلیل ممیزی، ماشین‌بردار پشتیبان و شبکه‌های عصبی بهتر و در مقابل بیش‌برازش مقاوم‌تر است. برای تشکیل هر درخت، دستهٔ متفاوتی از الگوهای موجود، با در نظر گرفتن جایگزینی دوباره هر الگوی انتخاب شده، انتخاب می‌شود. اندازهٔ این دستهٔ نمونه‌برداری شده برابر تعداد کل الگوهای موجود خواهد بود. این طریقهٔ نمونه‌برداری معمولاً در حدود یک‌سوم از الگوهای موجود را بیرون از دسته قرار می‌دهد. هر درخت بر اساس دسته الگوی انتخاب شده، تا ماکزیمم عمق از پیش تعیین شده رشد داده می‌شود. این عمق بر اساس حداقل تعداد الگوها در هر گره انتهایی، تعیین می‌شود. بر اساس الگوریتم جنگل‌های تصادفی، در مرحلهٔ رشد هر درخت، درهر گره، دسته‌ای از ویژگی‌ها (متغیرها) به صورت تصادفی انتخاب و بهترین انشعاب در میان دسته ویژگی انتخاب شده برای تشکیل گره‌های جدید بعدی در نظر گرفته می‌شود [۶۵].

جنگل‌های تصادفی دارای چندین مزیت نسبت به سایر روش‌های مدلسازی است. متغیرهای مورد استفاده می‌تواند پیوسته یا طبقه‌ای باشد. به‌دلیل ایجاد تعداد زیاد درخت و میانگین‌گیری در اجرای جنگل‌های تصادفی، این طبقه‌بندی‌کننده به نتایج با تعصب اندک و تغییرپذیری کم، ولی پیش‌بینی‌های دقیق منجر می‌شود. این راهبرد به‌طور معناداری بهتر از طبقه‌بندی‌کننده‌های تحلیل ممیزی، ماشین‌بردار پشتیبان و شبکه‌های عصبی است و در برابر بیش‌برازش[4]، مقاوم است. افزون بر این، جنگل‌های تصادفی، فقط دو شاخص اصلی (تعداد متغیرها در هر گره و تعداد درختان در جنگل) دارد و معمولاً به ارزش‌های آن‌ها چندان حساس نیست. با توجه به این مزایا کاربرد الگوریتم جنگل‌های تصادفی در حال

[1] یکی از مهم‌ترین وظایف در یادگیری ماشینی، انطباق مدل با مجموعهٔ داده‌هاست، به نحوی که پس از آن، مدل بتواند پیش‌بینی قابل اعتمادی از داده‌های دیده نشده، انجام دهد (روایی خارجی). در بیش‌برازش، مدل به‌جای انطباق بر رابطهٔ عمومی داده‌ها، بر تک‌تک داده‌ها (که برخی از آن‌ها نیز پارازیت هستند)، برازش می‌شود. در این حالت کیفیت پیش‌بینی مدل برای داده‌های آموزش داده نشده، به شدت کاهش می‌یابد. بیش‌برازش معمولاً زمانی اتفاق می‌افتد که تعداد شاخص‌های مسئله به نسبت داده‌های موجود، زیاد باشد و روش یادگیری نتواند به‌خوبی رابطهٔ داده‌ها را آموزش ببیند.

[2] Bootstrap

[3] Split

[4] Over fitting

افزایش است [۶۵]. برای اجرای این روش از نرم‌افزار weka استفاده شده است.

رگرسیون لجستیک

به منظور ارزیابی طبقه‌بندی‌کنندهٔ جنگل‌های تصادفی، معیارهای ارزیابی عملکرد حاصل از پیش‌بینی با این روش، با دقت پیش‌بینی رگرسیون لجستیک، مقایسه شده است. رگرسیون لجستیک شبیه به رگرسیون معمولی است، با این تفاوت که روش تخمین ضرایب در آن‌ها یکسان نیست. رگرسیون لجستیک، به جای حداقل کردن مجذور خطاها (که در رگرسیون معمولی انجام می‌شود)، احتمالی که یک واقعه رخ می‌دهد را حداکثر می‌کند. در رگرسیون لجستیک، مدل زیر برآورد می‌شود که در آن P_i احتمال وقوع یک حالت خاص است [۱۶]:

$$\text{Log}\,[P_i/(1-P_i)]=a+B_1X_1+B_2X_2+\ldots+B_kX_k$$

در بیشتر پژوهش‌های انجام‌شده در زمینهٔ پیش‌بینی بحران مالی و ورشکستگی برای ارزیابی مدل‌های مبتنی بر هوش مصنوعی، از یک مدل خطی استفاده می‌شود. با توجه به دلایل زیر، در این پژوهش از مدل خطی رگرسیون لجستیک به عنوان مدل مقایسه‌ای استفاده می‌شود:

۱. دربین مدل‌های آماری، در پژوهش‌های انجام‌شده در زمینهٔ بحران مالی، از همه بیشتر استفاده شده و مشهورتر است [۱۰، ۸].

۲. رگرسیون لجستیک نسبت به تحلیل ممیزی (مدل خطی دیگری که در پیش‌بینی ورشکستگی استفاده زیادی از آن می‌شود) ابزار قوی‌تری است [۱۶].

روش انتخاب متغیر ریلیف

روش ریلیف که اساساً با رتبه‌بندی نزولی متغیرها عمل می‌کند، به علت سادگی و مؤثر بودن در افزایش دقت پیش‌بینی، در بسیاری از موارد استفاده می‌شود. ایدهٔ اصلی این روش بدین صورت است که برای هر متغیر یک امتیاز رتبه‌بندی را که نشان‌دهندهٔ میزان جداسازی داده‌های مثبت و منفی (طبقه‌ها) است، محاسبه می‌کند. الگوریتم به ازای هر دادهٔ آموزشی، نزدیک‌ترین دادهٔ هم‌کلاس با آن (نزدیکترین برخورد[1]) و نزدیک‌ترین داده با کلاس مخالف آن (نزدیکترین خطا[2]) را جست‌وجو می‌کند. پس از آن، امتیاز هر متغیر با محاسبه تفاوت یا نسبت مجموع فاصله نزدیک‌ترین برخورد داده‌ها با مجموع فاصلهٔ نزدیک‌ترین خطای داده‌ها که بر روی آن متغیر نگاشت شده است، به‌دست می‌آید. فاصلهٔ یاد شده در این الگوریتم، فاصلهٔ اقلیدسی است [۲۰].

روش انتخاب متغیر ریلیف از جمله روش‌های انتخاب متغیر مبتنی بر معیار فاصله است. در این روش، اگر یک متغیر به ازای نمونه‌های درون یک طبقه، مقدار یکسان و به ازای نمونه‌های دیگر طبقه‌ها مقادیر مختلفی داشته باشد، وزن بالاتری می‌گیرد. ریلیف از بین داده‌های آموزشی، یک نمونه را به‌صورت تصادفی انتخاب می‌کند و سپس فاصلهٔ اقلیدسی آن نمونه تا نزدیک‌ترین نمونه از طبقهٔ مشابه و نزدیک‌ترین نمونه از طبقهٔ متفاوت را به‌دست می‌آورد و سپس این فاصله‌ها را برای به‌روز کردن وزن هر متغیر به‌کار می‌برد. در نهایت، الگوریتم آن‌دسته از متغیرهایی را انتخاب می‌کند که وزن آن‌ها از یک حد آستانهٔ از پیش تعریف‌شده به‌وسیلهٔ کاربر،

[1] Nearest Hit
[2] Nearest Miss

بیشتر است [۲۰]. در واقع، رتبه‌ای که ریلیف به هر متغیر می‌دهد بر اساس میزان نقش آن متغیر در جداسازی نمونه‌های متفاوت همسایه است. این الگوریتم برای هر نمونهٔ آموزشی به دنبال نزدیک‌ترین همسایه که با آن، هم‌کلاس (طبقه) است، می‌شود که به این نزدیک‌ترین همسایه، نزدیک‌ترین برخورد گفته می‌شود. سپس نزدیک‌ترین همسایه که طبقهٔ آن با طبقهٔ نمونهٔ آموزش مخالف است را پیدا می‌کند که به این همسایه، نزدیک‌ترین خطا گفته می‌شود. رتبه‌ای که به هر متغیر داده می‌شود بر اساس نسبت مجموع فاصلهٔ نمونه‌های آموزشی تصویرشده روی هر متغیر از نزدیک‌ترین برخورد همسایه برای هر نمونهٔ آموزشی به مجموع فاصلهٔ نزدیک‌ترین همسایهٔ خطا برای هر نمونهٔ آموزشی است [۲۰].

ایدهٔ کلیدی الگوریتم ریلیف (الگوریتم ۱)، برآورد کیفیت متغیرها بر این اساس است که چگونه (به چه خوبی) ارزش متغیرها در نمونه‌های نزدیک به یکدیگر تشخیص داده می‌شود. در این راستا، با انتخاب یک نمونهٔ تصادفی Ri (خط ۳)، ریلیف به جست‌وجوی دو تا از نزدیک‌ترین همسایهٔ آن نمونهٔ تصادفی می‌پردازد. یکی در گروه مشابه که نزدیک‌ترین برخورد (H) نامیده می‌شود و دیگری از گروه متفاوت که نزدیک‌ترین خطا (M) نامیده می‌شود (خط ۴). ریلیف، برآورد کیفیت W[A] را برای تمام متغیرهای پیش‌بین، بسته به ارزش آن‌ها برای Ri، M و H به‌روز می‌کند (خطوط ۵ و ۶). اگر نمونه‌ها (Ri) وH دارای مقادیر متفاوتی از متغیر پیش‌بین A باشند، آن‌گاه متغیر پیش‌بین A، دو نمونه متعلق به یک گروه را تفکیک (به‌صورت متمایز، طبقه‌بندی) می‌کند که این امر مطلوب نیست. بنابراین، کیفیت برآورد W[A] کاهش داده می‌شود. در مقابل، اگر نمونه‌ها (Ri) وM دارای مقادیر متفاوتی از متغیر پیش‌بین A باشد، آن‌گاه متغیر پیش‌بین A، دو نمونه از گروه‌های متفاوت را تفکیک (به‌صورت متمایز، طبقه‌بندی) می‌کند که این امر مطلوب است و بنابراین، کیفیت برآورد W[A] افزایش داده می‌شود. کل فرآیند، m مرتبه تکرار می‌شود که m یک شاخص مشخص‌شده به‌وسیلهٔ کاربر است [۵۵].

ورودی: برای هر نمونهٔ آموزشی، برداری از ارزش‌های متغیرهای پیش‌بین و ارزش هر طبقه (گروه وابسته).

خروجی: بردارw از برآوردهای کیفیت متغیرهای پیش‌بین.

۱. تمام وزن‌ها W [A] را برابر صفر قرار بده.
۲. برای i=1 تا m شروع کن.
۳. یک نمونهٔ تصادفی R_i انتخاب کن.
۴. نزدیک‌ترین برخورد و نزدیک‌ترین خطا را پیدا کن.
۵. برای A از یک تا a انجام بده.
۶. W [A]:= W [A] - diff(A, R_i H) /m +diff (A, R_i , M) /m;
۷. پایان

الگوریتم ۳-۱: الگوریتم اولیهٔ ریلیف

شایان ذکر است که این روش، کلیهٔ متغیرهای اولیه را رتبه‌بندی می‌کند. پس از رتبه‌بندی نزولی تمامی متغیرها، K متغیر برتر این رتبه‌بندی انتخاب می‌شود. انتخاب K براساس یک حد آستانه یا تفاوت معنادار امتیاز دو متغیر متوالی تعیین می‌گردد. در این پژوهش از ۱۰ متغیر اول رتبه‌بندی برای پیش‌بینی استفاده شده است، زیرا امتیاز متغیر یازدهم در رتبه‌بندی ریلیف، تفاوت زیادی با امتیاز متغیر دهم داشت؛ به‌بیان دیگر، اضافه کردن متغیر یازدهم، کمک

چندانی به پیش‌بینی نمی‌کرد. ۱۰ متغیر منتخب روش ریلیف عبارت‌اند از: متغیرهای ردیف‌های ۵۶، ۶۴، ۴۶، ۲۳، ۲۸، ۴، ۳۷، ۶۰، ۲۱ و ۲۶ در نگارهٔ ۱.

روایی متقابل

در این پژوهش، به منظور بررسی تعمیم‌پذیری پیش‌بینی‌های انجام شده به‌وسیلهٔ طبقه‌ندی‌کننده‌ها از روایی متقابل ۱۰ بخشی[1] استفاده می‌شود. در روش holdout که در اغلب پژوهش‌های حسابداری و مالی (به‌ویژه در ایران) استفاده شده است، داده‌ها به دو دسته به نام مجموعهٔ آموزشی و مجموعهٔ آزمایشی تقسیم می‌شود. این روش‌ها دارای محدودیت‌های بارزی هستند. روش holdout یک تخمین‌گر بدبینانه است، زیرا فقط بخشی از داده‌ها برای آموزش به روش پیش‌بینی ارائه شده است. هر چه تعداد نمونهٔ بیشتری برای مجموعهٔ آزمایشی خارج شود، تعصب برآورد بیشتر می‌شود. از طرفی، نمونه‌های آزمایشی کوچک‌تر (با تعداد کمتر) به معنای این است که فاصله اطمینان دقت، بیشتر خواهد بود. بنابراین، روش مزبور، روش مناسبی نخواهد بود [۴۳]. در مقابل، روش روایی متقابل، به دلیل سادگی، شفافیت و جامعیت، یک راهبرد مناسب است و نتایج بسیاری از پژوهش‌های انجام شده حاکی از عملکرد بهتر این روش است. در این راستا، در پژوهش حاضر به منظور بررسی تعمیم‌پذیری پیش‌بینی‌های انجام شده از روایی متقابل ۱۰ بخشی استفاده می‌شود. روایی متقابل ۱۰ بخشی برای برآورد نرخ خطای واقعی کاملاً قابل اتکا و کافی است [۳۷]. در این روش، نمونهٔ اصلی به ۱۰ دسته نمونهٔ فرعی مختلف تقسیم می‌شود. ۹ نمونهٔ فرعی به‌عنوان نمونه‌های آموزشی استفاده می‌شود و نمونهٔ فرعی باقی‌مانده به‌عنوان نمونهٔ آزمایشی، مورد آزمون قرار می‌گیرد. این شیوه تا حدی تکرار می‌شود که هر یک از ۱۰ نمونهٔ فرعی به‌عنوان نمونهٔ آزمایشی مورد آزمون قرار گیرد. در این پژوهش، روایی متقابل ۱۰ بخشی، با استفاده از اجزای مختلف مجموعهٔ داده‌ها، به‌طور مستقل، ۱۰ بار انجام شده است (روایی متقابل ۱۰ بخشی با ۱۰ بار تکرار که منجر به صد نتیجه در هر بار اجرای الگوریتم می‌شود). یافته‌های اغلب پژوهش‌ها [۴۳] حاکی از آن است که در مسائل دنیای واقعی، روایی متقابل ۱۰ بخشی، بهترین روش انتخاب مدل است. مزیت روش مزبور، این است که تمام نمونه‌ها در نهایت هم به‌عنوان داده‌های آموزشی و هم به‌عنوان داده‌های آزمایشی استفاده خواهد شد. افزون بر این، استفاده از روایی متقابل، از بروز مشکل بیش‌برازش و مشکلات مربوط به نتایج برون‌نمونه‌ای[2] جلوگیری می‌کند.

روش آزمون فرضیه‌ها

به‌منظور ارزیابی عملکرد پیش‌بینی، از معیارهای ارزیابی (شامل میانگین دقت و خطاهای نوع اول و دوم) استفاده می‌شود. به منظور آزمون فرضیهٔ اول (ارزیابی عملکرد طبقه‌بندی‌کنندهٔ جنگل‌های تصادفی)، معیارهای ارزیابی مربوط به پیش‌بینی بحران مالی با استفاده از این روش با معیارهای ارزیابی حاصل از پیش‌بینی با روش رگرسیون لجستیک مقایسه می‌شود. به منظور آزمون فرضیهٔ دوم (ارزیابی سودمندی روش انتخاب متغیر ریلیف)، معیارهای ارزیابی حاصل از پیش‌بینی با استفاده از متغیرهای منتخب این روش با معیارهای ارزیابی حاصل از عدم انجام مرحلهٔ انتخاب متغیرهای پیش‌بین (پیش‌بینی با ۶۹ متغیر اولیه) مقایسه می‌شود.

[1] 10-fold cross validation

[2] Out-of-Sample

شایان ذکر است که برتری معیارهای عملکرد پیش‌بینی در زمان استفاده از متغیرهای منتخب روش ریلیف نسبت به ۶۹ متغیر اولیه و معنادار بودن این تفاوت از نظر آماری، بیانگر سودمندی روش انتخاب متغیر است؛ زیرا، در این صورت، افزون بر کاهش تعداد متغیرهای پیش‌بین، عملکرد پیش‌بینی بهبود یافته است.

انتخاب متغیرهای پیش‌بین و پیش‌بینی بحران مالی به‌وسیلهٔ نرم‌افزار Weka نسخهٔ ۷-۳ انجام شده است. به‌منظور آزمون فرضیه‌های پژوهش نیز از آزمون t زوجی (براساس صد دقت حاصل از اجرای روایی متقابل ۱۰ بخشی با ۱۰ بار تکرار) در نرم‌افزار SPSS نسخهٔ ۲۱ استفاده شده است. شایان ذکر است که در این پژوهش، از داده‌های یک سال قبل شرکت‌ها برای پیش‌بینی بحران مالی استفاده شده است.

یافته‌های تجربی پژوهش

به منظور آزمون فرضیهٔ اول، میانگین معیارهای ارزیابی (میانگین دقت، خطای نوع اول و دوم) مربوط به پیش‌بینی بحران مالی با استفاده از جنگل‌های تصادفی و رگرسیون لجستیک در صورت استفاده از ۶۹ متغیر اولیه مقایسه می‌شود. از آزمون t زوجی برای آزمون این فرضیه و بررسی وجود تفاوت معنادار بین عملکرد پیش‌بینی این دو طبقه‌بندی‌کننده، استفاده شده است. در این راستا، از دقت‌های حاصل از روایی متقابل ۱۰ بخشی با ۱۰ بار تکرار استفاده شد که منجر به ایجاد ۱۰۰ دقت در هر پیش‌بینی می‌شود. نگارهٔ (۲)، معیارهای ارزیابی مربوط به پیش‌بینی بحران مالی را با استفاده از جنگل‌های تصادفی و رگرسیون لجستیک در حالت استفاده از ۶۹ متغیر اولیه نشان می‌دهد. فرض صفر در این فرضیه، عبارت است از این که در زمان استفاده از ۶۹ متغیر اولیه، تفاوت معناداری بین عملکرد (معیارهای ارزیابی) پیش‌بینی جنگل‌های تصادفی و رگرسیون لجستیک وجود ندارد یا به بیان دیگر، میانگین دقت، خطای نوع اول و دوم پیش‌بینی این دو طبقه‌بندی کننده، یکسان است. با توجه به آمارهٔ t و مقدار احتمال مربوطه، در سطح معناداری ۰/۰۵ تفاوت معناداری بین عملکرد پیش‌بینی جنگل‌های تصادفی و رگرسیون لجستیک در زمان استفاده از ۶۹ متغیر اولیه وجود دارد. با توجه به بهتر بودن معیارهای عملکرد جنگل‌های تصادفی نسبت به رگرسیون لجستیک و معنادار بودن آن از نظر آماری می‌توان استنباط کرد که در زمان استفاده از ۶۹ متغیر اولیه، جنگل‌های تصادفی، عملکرد بهتری نسبت به رگرسیون لجستیک دارد.

نگارهٔ ۲. عملکرد جنگل‌های تصادفی و رگرسیون لجستیک در حالت استفاده از کل متغیرها

عملکرد / طبقه‌بندی‌کننده	جنگل‌های تصادفی	رگرسیون لجستیک	آماره t	مقدار احتمال
میانگین دقت	۰/۹۰	۰/۸۳	۶/۶۵۲	۰/۰۰۰
میانگین خطای نوع اول	۰/۰۹	۰/۱۸	۷/۴۵۴	۰/۰۰۰
میانگین خطای نوع دوم	۰/۱۰	۰/۱۶	۵/۲۴۷	۰/۰۰۰

منبع: یافته‌های پژوهش

نگارهٔ (۳)، نتایج آزمون فرضیهٔ دوم را نشان می‌دهد. در این فرضیه میانگین معیارهای ارزیابی مربوط به پیش‌بینی بحران مالی با جنگل‌های تصادفی و رگرسیون لجستیک در حالت استفاده از متغیرهای منتخب ریلیف مقایسه شده است. فرض صفر در این فرضیه، عبارت است از اینکه در زمان استفاده از ۱۰ متغیر منتخب ریلیف تفاوت معناداری بین عملکرد (معیارهای ارزیابی) پیش‌بینی جنگل‌های تصادفی و رگرسیون لجستیک وجود ندارد؛ یا به بیان دیگر، میانگین دقت، خطای نوع اول و دوم پیش‌بینی این دو طبقه‌بندی‌کننده، یکسان است. با توجه به آمارهٔ t و مقدار احتمال مربوطه، در سطح معناداری ۰/۰۵ تفاوت معناداری بین عملکرد پیش‌بینی جنگل‌های تصادفی و رگرسیون لجستیک در زمان استفاده از متغیرهای منتخب ریلیف وجود دارد. با توجه به بهتر بودن معیارهای عملکرد جنگل‌های تصادفی نسبت به رگرسیون لجستیک و معنادار بودن آن از نظر آماری می‌توان استنباط کرد که در زمان استفاده از متغیرهای منتخب ریلیف، جنگل‌های تصادفی، عملکرد بهتری نسبت به رگرسیون لجستیک دارد.

نگارهٔ ۳. عملکرد جنگل‌های تصادفی و رگرسیون لجستیک در حالت استفاده از متغیرهای منتخب ریلیف

عملکرد / طبقه‌بندی‌کننده	جنگل‌های تصادفی	رگرسیون لجستیک	آماره t	مقدار احتمال
میانگین دقت	۰/۹۳	۰/۸۸	۴/۳۲۸	۰/۰۰۰
میانگین خطای نوع اول	۰/۰۸	۰/۱۴	۴/۶۵۷	۰/۰۰۰
میانگین خطای نوع دوم	۰/۰۶	۰/۱۰	۳/۵۶۸	۰/۰۰۰

منبع: یافته‌های پژوهش

در فرضیه سوم، معیارهای ارزیابی مربوط به پیش‌بینی بحران مالی با استفاده از طبقه‌بندی‌کننده جنگل‌های تصادفی در دو حالت (براساس ۱۰ متغیر منتخب روش ریلیف و ۶۹ متغیر اولیه) مقایسه می‌شود. از آزمون t زوجی برای آزمون فرضیه‌ها و بررسی وجود تفاوت معنادار بین عملکرد پیش‌بینی جنگل‌های تصادفی در این دو حالت، استفاده شده است. در این راستا، از دقت‌های حاصل از روایی متقابل ۱۰ بخشی با ۱۰ بار تکرار استفاده شد که منجر به ایجاد ۱۰۰ دقت در هر پیش‌بینی می‌شود. نگارهٔ (۴)، معیارهای ارزیابی مربوط به پیش‌بینی بحران مالی را با استفاده از جنگل‌های تصادفی در دو حالت (براساس ۱۰ متغیر منتخب روش ریلیف و ۶۹ متغیر اولیه) نشان می‌دهد. فرض صفر در این فرضیه، عبارت است از اینکه در زمان پیش‌بینی با جنگل‌های تصادفی تفاوت معناداری بین عملکرد (معیارهای ارزیابی) پیش‌بینی با استفاده از ۶۹ متغیر اولیه و متغیرهای منتخب ریلیف وجود ندارد؛ یا به بیان دیگر، میانگین دقت، خطای نوع اول و دوم پیش‌بینی جنگل‌های تصادفی با این دو دسته متغیر، یکسان است. با توجه به آمارهٔ t و مقدار احتمال مربوطه، در سطح معناداری ۰/۰۵تفاوت معناداری بین معیارهای عملکرد (به استثنای خطای نوع اول) پیش‌بینی جنگل‌های تصادفی در زمان استفاده از متغیرهای منتخب روش ریلیف و استفاده از کلیهٔ متغیرهای اولیه وجود دارد. بنابراین، با توجه به بهتر بودن معیارهای عملکرد در حالت استفاده از متغیرهای منتخب روش ریلیف نسبت به استفاده از کل متغیرها و معنادار بودن

آن از نظر آماری (به استثنای خطای نوع اول) می‌توان استنباط کرد که روش انتخاب متغیر ریلیف، تأثیر مثبت و معناداری بر عملکرد پیش‌بینی طبقه‌بندی‌کنندهٔ جنگل تصادفی دارد؛ زیرا به رغم کاهش تعداد متغیرهای پیش‌بین (از ۶۹ به ۱۰ متغیر)، عملکرد پیش‌بینی بهتر شده است.

نگارهٔ ۴. عملکرد جنگل‌های تصادفی بر اساس متغیرهای منتخب و متغیرهای اولیه

عملکرد / متغیرها	براساس متغیرهای منتخب	براساس متغیرهای اولیه	آمارهٔ t	مقدار احتمال
میانگین دقت	۰/۹۳	۰/۹۰	۳/۷۵۷	۰/۰۰۰
میانگین خطای نوع اول	۰/۰۸	۰/۰۹	۰/۷۸۲	۰/۴۳۶
میانگین خطای نوع دوم	۰/۰۶	۰/۱۰	۳/۸۶۵	۰/۰۰۰

منبع: یافته‌های پژوهش

در فرضیهٔ چهارم، معیارهای ارزیابی مربوط به پیش‌بینی بحران مالی با استفاده از طبقه‌بندی کننده رگرسیون لجستیک در دو حالت (براساس ۱۰ متغیر منتخب روش ریلیف و ۶۹ متغیر اولیه) مقایسه می‌شود. نگارهٔ (۵)، معیارهای ارزیابی مربوط به پیش‌بینی بحران مالی را با استفاده از رگرسیون لجستیک در دو حالت (براساس ۱۰ متغیر منتخب روش ریلیف و ۶۹ متغیر اولیه) نشان می‌دهد. فرض صفر در این فرضیه، عبارت است از اینکه در زمان پیش‌بینی با رگرسیون لجستیک، تفاوت معناداری بین عملکرد (معیارهای ارزیابی) پیش‌بینی با استفاده از ۶۹ متغیر اولیه و متغیرهای منتخب ریلیف وجود ندارد؛ یا به بیان دیگر، میانگین دقت، خطای نوع اول و دوم پیش‌بینی رگرسیون لجستیک با این دو دسته متغیر، یکسان است. با توجه به آمارهٔ t و مقدار احتمال مربوطه، در سطح معناداری ۰/۰۵ تفاوت معناداری بین عملکرد پیش‌بینی رگرسیون لجستیک در زمان استفاده از متغیرهای منتخب روش ریلیف و استفاده از کلیهٔ متغیرهای اولیه وجود دارد. بنابراین، با توجه به بهتر بودن معیارهای عملکرد در حالت استفاده از متغیرهای منتخب روش ریلیف نسبت به استفاده از کل متغیرها و معنادار بودن آن از نظر آماری می‌توان استنباط کرد که روش انتخاب متغیر ریلیف، تأثیر مثبت و معناداری بر عملکرد پیش‌بینی رگرسیون لجستیک دارد؛ زیرا به رغم کاهش تعداد متغیرهای پیش‌بین (از ۶۹ به ۱۰ متغیر)، عملکرد پیش‌بینی بهتر شده است.

نگارهٔ ۵. عملکرد رگرسیون لجستیک بر اساس متغیرهای منتخب و متغیرهای اولیه

عملکرد / متغیرها	براساس متغیرهای منتخب	براساس متغیرهای اولیه	آماره t	مقدار احتمال
میانگین دقت	۰/۸۸	۰/۸۳	۴/۳۴۸	۰/۰۰۰
میانگین خطای نوع اول	۰/۱۴	۰/۱۸	۴/۲۴۶	۰/۰۰۰
میانگین خطای نوع دوم	۰/۱۰	۰/۱۶	۴/۵۶۸	۰/۰۰۰

منبع: یافته‌های پژوهش

خلاصه و نتیجه‌گیری

پیش‌بینی بحران مالی و ورشکستگی در تصمیم‌گیری‌های مالی از اهمیت بسزایی برخوردار است. پیش‌بینی بحران مالی و ورشکستگی همواره به عنوان موضوعی حیاتی مدنظر بوده و به طور وسیعی در ادبیات حسابداری و مالی مورد مطالعه قرار گرفته است. به‌رغم عملکرد بهتر طبقه‌بندی‌کنندهٔ جنگل‌های تصادفی نسبت به شیوه‌های آماری و بسیاری از شیوه‌های هوش مصنوعی در پیش‌بینی بحران مالی و ورشکستگی، تاکنون پژوهشی با استفاده از این طبقه‌بندی‌کننده در بورس اوراق بهادار تهران انجام نشده است. بنابراین، در پژوهش حاضر به ارزیابی این طبقه‌بندی‌کننده برای پیش‌بینی بحران مالی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران پرداخته شد. افزون براین، وجود تعداد متغیرهای زیاد، نه‌تنها بر عملکرد پیش‌بینی‌کننده اثر نامساعد می‌گذارد، بلکه زمان اجرای الگوریتم یادگیری را نیز تحت تأثیر قرار می‌دهد. هرچه تعداد متغیرها بیشتر شود، زمان اجرای الگوریتم یادگیری نیز بیشتر می‌شود. همچنین بعد بالای داده‌ها می‌تواند به مسألهٔ اضافه‌بار ابعاد منجر شود. بنابراین، در این پژوهش به‌منظور انتخاب متغیرهای پیش‌بین بهینه از روش ریلیف استفاده شد.

یافته‌های تجربی مربوط به بررسی ۹۵ شرکت – سال سالم (بدونِ درماندگی مالی) و ۹۵ شرکت – سال (درماندهٔ مالی) پذیرفته شده در بورس اوراق بهادار تهران در سال‌های ۱۳۸۰ الی ۱۳۹۲ حاکی از عملکرد بهتر جنگل‌های تصادفی نسبت به رگرسیون لجستیک است. به عبارت دیگر، در صورت استفاده از این طبقه‌بندی‌کننده، به‌طور معناداری، میانگین دقت افزایش و خطای نوع اول و دوم کاهش می‌یابد. به‌نظر می‌رسد یکی از دلایل برتری عملکرد این روش غیرخطی نسبت به رگرسیون لجستیک، احراز نشدن مفروضات رگرسیون خطی در داده‌های مورد بررسی و همچنین ماهیت غیرخطی روابط پیچیده بین متغیرهای بررسی شده است. افزون براین، یافته‌های پژوهش حاکی از سودمندی روش انتخاب متغیر ریلیف در پیش‌بینی بحران مالی است. دلیل برتری معیارهای ارزیابی عملکرد در حالت انجام مرحلهٔ انتخاب متغیرها نسبت به عدم انجام این مرحله، مسألهٔ اضافه‌بار ابعاد است. به نظر می‌رسد اضافه کردن متغیرهای بیشتر، پارازیت (نویز) و در نتیجه خطا را افزایش می‌دهد و اضافه کردن متغیرها تا یک حد معین می‌تواند به بهبود پیش‌بینی کمک کند و اضافه کردن بیشتر متغیرها منجر به مسألهٔ اضافه‌بار ابعاد می‌شود.

پیشنهادهای پژوهش

با توجه به یافته‌های این پژوهش، مبنی بر برتری عملکرد طبقه‌بندی‌کنندهٔ جنگل‌های تصادفی در پیش‌بینی بحران مالی، به افراد و مؤسسه‌های زیر پیشنهاد می‌شود که در پیش‌بینی بحران مالی، از این روش‌ها استفاده کنند:

۱. حسابرسان به‌عنوان اعتباردهندگان به اطلاعات صورت‌های مالی شرکت‌ها در ارزیابی فرض تداوم فعالیت.

۲. بانک‌ها و سایر مؤسسه‌های اعتباری در تصمیم‌گیری در مورد اعطا یا عدم اعطای اعتبار و شرایط آن و همچنین تعیین سیاست‌هایی برای نظارت بر وام‌های موجود.

۳. سرمایه‌گذاران و سهامداران در تصمیم‌گیری‌های سرمایه‌گذاری مبنی بر خرید و فروش سهام، کاهش خطر سبد سرمایه‌گذاری و ارزیابی ریسک شرکت.

۴. مدیریت در انجام اقدام‌هایی پیشگیرانه به‌منظور جلوگیری از ورشکستگی و مدیریت بهتر.

منابع

۱- ابریشمی، حمید. (۱۳۸۷). *مبانی اقتصادسنجی*. جلددوم، چاپ پنجم، تهران: انتشارات دانشگاه تهران.

۲- پناهی، حسین، اسدزاده، احمد و علیرضاجلیلی مرند. (۱۳۹۳). پیش‌بینی پنج‌سالهٔ ورشکستگی مالی برای شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران. *تحقیقات مالی*، دورهٔ ۱۶، شمارهٔ ۱، صص. ۵۷-۷۶.

۳- پورحیدری، امید و مهدی کوپایی حاجی. (۱۳۸۹). پیش‌بینی بحران مالی شرکت‌ها با استفاده از مدل مبتنی بر تابع تفکیکی خطی. *پژوهش‌های حسابداری مالی*، سال دوم، شمارهٔ اول، صص. ۳۳-۴۶.

۴- چالاکی، پری و مرتضی یوسفی. (۱۳۹۱). پیش‌بینی مدیریت سود با استفاده از درخت تصمیم‌گیری. *مطالعات حسابداری و حسابرسی*، شمارهٔ ۱، صص. ۱۱۰-۱۲۳.

۵- حسینی، سیدمحسن و زینبرشیدی. (۱۳۹۲). پیش‌بینی ورشکستگی شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران با استفاده از درخت تصمیم و رگرسیون لجستیک. *پژوهش‌های حسابداری مالی*، سال پنجم، شمارهٔ ۱۷، صص. ۱۰۵-۱۲۸.

۶- دستگیر، محسن، حسینزاده، علی‌حسین، خدادادی، ولی و سیدعلی واعظ. (۱۳۹۱). کیفیت سود در شرکت‌های درماندهٔ مالی. *پژوهش‌های حسابداری مالی*، شمارهٔ ۴، صص. ۱-۱۶.

۷- راعی، رضا و سعید فلاح‌پور. (۱۳۸۳). پیش‌بینی درماندگی مالی شرکت‌ها با استفاده از شبکه‌های عصبی مصنوعی. *تحقیقات مالی*، شمارهٔ ۱۷. صص. ۳۹-۶۹.

۸- راعی، رضا و سعید فلاح‌پور. (۱۳۸۷). کاربرد ماشین‌بردار پشتیبان در پیش‌بینی درماندگی مالی شرکت‌ها با استفاده از نسبت‌های مالی. *بررسی‌های حسابداری و حسابرسی*، دورهٔ ۱۵، شمارهٔ ۵۳. صص. ۱۷-۳۴.

۹- رهنمای رودپشتی، فریدون، علی‌خانی، راضیه و مهدی مران‌جوری. (۱۳۸۸).بررسی کاربرد مدل‌های پیش‌بینی ورشکستگی آلتمنوفالمر در شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران. *بررسی‌های حسابداری و حسابرسی*، دورهٔ۱۶، شمارهٔ ۵۵. صص. ۱۹-۳۴.

۱۰- سعیدی، علی و آرزو آقایی. (۱۳۸۸). پیش‌بینی درماندگی مالی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران با استفاده از شبکه‌های بیز. *بررسی‌های حسابداری و حسابرسی*، دورهٔ ۱۶، شمارهٔ ۵۶. صص. ۵۹- ۷۸.

۱۱- سلیمانی امیری، غلامرضا. (۱۳۸۲). نسبت‌های مالی و پیش‌بینی بحران مالی شرکت‌ها در بورس اوراق بهادار تهران. *تحقیقات مالی*، شمارهٔ ۱۵. صص. ۱۲۱- ۱۳۶.

۱۲- فدایی‌نژاد، محمداسماعیل و رسول اسکندری. (۱۳۹۰). طراحی و تبیین مدل پیش‌بینی

ورشکستگی شرکت‌ها در بورس اوراق بهادار تهران. *تحقیقات حسابداری و حسابرسی*، شمارهٔ ۹، صص. ۳۸-۵۵.

۱۳- محمودآبادی، حمید والهه برزگر. (۱۳۸۸). بررسی نحوه توزیع آماری نسبت‌های مالی در شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران. *پیشرفت‌های حسابداری دانشگاه شیراز*، دورهٔ اول، شمارهٔ اول، پیاپی ۵۷/۳، صص. ۱۷۱-۱۸۹.

۱۴-. مکیان، سید نظام‌الدین، المدرسی، سیدمحمدتقی و سلیم کریمی‌تکلو. (۱۳۸۹). مقایسهٔ مدل شبکه‌های عصبی مصنوعی با روش‌های رگرسیون لوجستیک و تحلیل ممیزی در پیش‌بینی ورشکستگی شرکت‌ها. *فصلنامهٔ پژوهش‌های اقتصادی*، سال دهم، شمارهٔ دوم. ص‌ص. ۱۴۱-۱۶۱.

۱۵- موسوی‌شیری، محمود و محمدرضا طبرستانی. (۱۳۸۸). پیش‌بینی درماندگی مالی با استفاده از تحلیل پوششی داده‌ها. *تحقیقات حسابداری*، شمارهٔ دوم. صص. ۱۵۸- ۱۸۷.

۱۶- مؤمنی، منصور و علی فعال‌قیومی. (۱۳۸۶). *تحلیل‌های آماری با استفاده از SPSS*. چاپ اول، تهران: انتشارات کتاب نو.

۱۷- مهرانی، ساسان، مهرانی، کاوه، منصفی، یاشار و غلامرضا کرمی. (۱۳۸۴). بررسی کاربردی الگوهای پیش‌بینی ورشکستگی زیمسکی و شیراتا در شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران. *بررسی‌های حسابداری و حسابرسی*، سال دوازدهم، شمارهٔ۴۱. صص.۱۳۱-۱۰۵.

۱۸- نیکبخت، محمدرضا و مریم شریفی. (۱۳۸۹). پیش‌بینی ورشکستگی مالی شرکت‌های بورس اوراق بهادار تهران با استفاده از شبکه‌های عصبی مصنوعی. *مدیریت صنعتی*، دورهٔ ۲، شمارهٔ ۴. صص. ۱۶۳- ۱۸۰.

19- Altman, E. I. (1968). Financial Ratios, Discriminant Analysis, and the Prediction ofCorporate Bankruptcy.*Journal of Finance*, Vol. 23, No. 4, Pp. 589-609.

20- Atiya, A. F. (2001). Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results. *IEEE Transactions on Neural Networks*, Vol. 12, No. 4, Pp. 929-935.

21- Back, B., Laitinen, T., Sere, K (1996). Neural network and genetic algorithm for bankruptcy prediction, *Expert Systems with Applications*, Vol. 11, No. 4, Pp. 407–413.

22- Barniv, R., Anurag, A., and R. Leach (1997). Predicting the outcome following bankruptcy filing: A three state classification using NN, *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 6, Pp.177–194.

23- Beaver, W. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, Vol. 4, Pp. 71-111.

24- Bougen, P. D., and J. C. Drury. (1980). UK Statistical Distributions of Financial Ratios. *Journal of Business, Finance and Accounting*, Vol. 7, No. 1, Pp. 39- 47.

25- Breiman, L. (2001). Random Forests. *Machine Learning*. Vol. 45, No. 1, Pp. 5-32.

26- Bryant, S.M. (1997). A case-based reasoning approach to bankruptcy prediction modeling, *Intelligent Systems in Accounting, Finance and Management*, Vol. 6, Pp. 195–214.

27- Deakin, E. (1972). A Discriminant Analysis of Predictors of Business Failure. *Journal of Accounting Research*, Vol. 10, No. 1, Pp. 167-179.

28- DeTienne, K. B., DeTienne, D. H., and S. A. Joshi. (2003). Neural Networks as Statistical Tools for Business Researchers. *Organizational Research Methods*, Vol. 6, No. 2, Pp. 236-265.

29- Dimitras, A. I., Slowinski, R., Susmaga, R., and C. Zopounidis (1999). Business failure prediction using rough sets, *European Journal of Operational Research*, Vol. 114, Pp. 263-280.

30- Drury, J. C. (1978). Financial Ratio Distribution for 1976: A Note. *Journal of Management Studies*, Vol. 15, No. 2, Pp. 241–254.

31- Etemadi, H., Anvary Rostamy, A. A., and H. Farajzadeh Dehkordi. (2009). A Genetic Programming Model for Bankruptcy Prediction: Empirical Evidence from Iran. *Expert Systems with Applications*, Vol. 36, No. 2, Pp. 3199–3207.

32- Fernandez- Castro, A., and P. Smith. (1994). Toward a general nonparametric model of corporate performance. *Omega: The International Journal of Management Science*, Vol. 22, No. 3, Pp. 237-249.

33- Foster, G. (1986).Financial Statement Analysis.*Prentice-Hall*.Inc, New Jersey.

34- Frecka, T. J. and W. S. Hopwood. (1983). The Effects of Outliers on the Cross-Sectional Distributional Properties of Financial Ratios. *The Accounting Review*, Vol. 58, No. 1, Pp. 115-128.

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

36- Hoglund, H. (2012). Detecting Earnings Management with Neural Networks. *Expert Systems with Applications*, Vol. 39, Pp. 9564-9570.

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

38- Jabeur, S. B. and Y. Fahmi. (2014). Default Prediction for Small-Medium Enterprises in France: A comparative approach. *South African Journal of Business Management*, 40 (1), Pp. 21-32

39- Jardin, P. (2010). Predicting Bankruptcy Using Neural Networks and Other Classification Methods: The Influence of Variable Selection Techniques on Model Accuracy. *Neurocomputing*, Vol. 73, Pp. 2047–2060.

40- Jones, S., and D. A. Hensher (2004). Predicting firm financial distress: A mixed logit model, *Accounting Review*, Vol. 79, No. 4, Pp. 1011-1038.

41- Karels, G. V., and A. J. Prakash. (1987). Multivariate normality and forecasting of business bankruptcy, Journal of Business Finance & Accounting,Vol. 14, No. 4, Pp. 573-593.

42- Kim, M., and D. Kang. (2010). Ensemble with Neural Networks for Bankruptcy Prediction. *Expert Systems with Applications*, Vol. 37, Pp. 3373–3379.

43- Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, Pp. 1137-1143.

44- Leano, H. J. (2004). *Discriminant Analysis, Factor Analysis and Linear Regression Analysis to Classify Financially Distressed Firms and Predict Bankruptcy Using Financial Ratios and Macroeconomic Predictors*. M. A Thesis, Lamar University.

45- Lee, K. C., Han, I., and Y. Kwon (1996). Hybrid neural network models for bankruptcy predictions, *Decision Support Systems*, Vol. 18, Pp. 63–72

46- Liang, D., Tsai, C. H., and H. T. Wu. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems*, Vol. 73, Pp. 289–297.

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

48- Lo, S. C. (2010). The Effects of Feature Selection and Model Selection on the Correctness of Classification, *Proceedings of the 2010 IEEE IEEM*, pp. 989-993.

49- McKee, T. E. (2003). Rough sets bankruptcy prediction models versus auditor signaling rates, *Journal of Forecasting*, Vol. 22, Pp. 569–589.

50- Mendes-Moreia, J., Soares, C., Jorge, A. M.; and J. F. D. Sousa. (2012). Ensemble Approaches for Regression: A Survey. *ACM Computing Surveys*, Vol. 45, No. 1, pp. 1- 40.

51- Min, J. H. and Y. Lee. (2005). Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters. *Expert Systems with Applications*, Vol. 28, Pp. 603–614.

52- Mukkamala, S., Tilve, G. D., Sung, A. H., Ribeiro, B., and A. S. Vieira. (2006). Computational Intelligent Techniques for Financial Distress Detection. *International Journal of Computational Intelligence Research*, Vol. 2, No. 1, Pp. 60-65.

53- Odom, M. D. and R. Sharda. (1990). A Neural Network Model for Bankruptcy Prediction. *IJCNN International Joint Conference on Neural Networks*, Vol. 2, Pp. 163-168.

54- Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, Vol. 18, No. 1, Pp. 109- 131.

55- Robnik-Sikonja, M., and I. Kononenko. (1997). An Adaptation of Relief for Attribute Estimation in Regression. *Machine Learning, Proceedings of 14th International Conference on Machine Learning (ICML'97)*, Pp. 296–304.

56- Ryu, Y. U., and W.T. Yue (2005). Firm bankruptcy prediction: Experimental comparison of isotonic separation and other classification approaches, *IEEE Transactions On Systems, Management and Cybernetics-Part A: Systems and Humans*, Vol. 35, No. 5, Pp. 727–737.

57- Sarkar, S. and R. S. Sriram. (2001). Bayesian Models for Early Warning of Bank Failures. *Management Science*, Vol. 47, No. 11, Pp. 1457-1475.

58- Shin, K. S. and Y. J. Lee. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, Vol. 23, No. 3, Pp. 321–328.

59- Shin, K., and Lee, T. S., and H. Kim. (2005). An Application of Support Vector Machines in Bankruptcy Prediction Model. *Expert Systems with Applications*, Vol. 28, Pp. 127–135.

60- Sun, J., Jia, M., and H. Li. (2011). AdaBoost Ensemble for Financial Distress Prediction: An Empirical Comparison with Data from Chinese Listed Companies. *Expert Systems with Applications*, Vol. 38, No. 8, Pp. 1- 8.

61- Tsai, C. (2009). Feature Selection in Bankruptcy Prediction. *Knowledge-Based Systems*, Vol. 22, No. 2, Pp. 120–127.

62- Wang, G., Ma, J., and S. Yang. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, Vol. 41, No. 5, Pp. 2353-2361.

63- Williamson, R. W. (1984). Evidence on the Selective Reporting of Financial Ratios. *The Accounting Review*, Vol. 59, No. 2, Pp. 296-298.

64- Yang, Z., You, W., and Ji. G. (2011). Using Partial Least Squares and Support Vector Machines for Bankruptcy Prediction. *Expert Systems with Applications*, Vol. 38, No. 7, Pp. 8336–8342

65- Yeh, C. C., Chi, D. j., and Y. R. Lin. (2014). Going-Concern prediction Using Hybrid Random Forests and Rough Set Approach. *Information Sciences*, Vol. 254, Pp. 98–110.

The Usefulness of Random Forest Classifier and Relief Features Selection in Financial Distress Prediction: Empirical Evidence of Companies Listed on Tehran Stock Exchange

*** M. H. Setayesh**
Associate professor of Accounting, Shiraz University, Shiraz, Iran
M. Kazemnezhad
PhD student of Accounting, Shiraz University, Shiraz, Iran
M. Hallaj
PhD Accounting, Shiraz University, Shiraz, Iran

Abstract
The Purpose of this research is investigating the usefulness of random forest classifier and relief features selection in financial distress prediction of companies listed on Tehran Stock Exchange. In this regard, through reviewing literature, 69 predictive features (variables) were specified as the initial features based on the popularity in the literature and the availability of the necessary data. By using relief method, optimal variables were selected from initial variables. In overall, the experimental results of investigating 95 financially distressed and 95 non-financial distressed in 2002 to 2014, indicated that random forest outperforms the logistic regression. In other words, the application of this classifier, increases the mean of accuracy, and reduces the occurrence of type I and type II errors. Furthermore, the results confirmed the usefulness of relief method in predicting financial distress. In other words, using selected variables of this feature selection method (relative to using 69 initial variables) increases the mean of accuracy, and reduces the occurrence of type I and type II errors

Keywords: Random Forest Classifier, Relief Features Selection method, Financial Distress Prediction.

* setayesh@shirazu.ac.ir