نوع مقاله: پژوهشی

تاریخ پذیرش: ۱۴۰۰/۰۹/۰۱
تاریخ دریافت: ۱۴۰۰/۱۱/۱۲

استفاده از روش نیمه‌نظارتی انتشار برچسب برای رتبه‌بندی اعتباری مشتریان بانک
(مطالعه موردی بانک پاسارگاد)

مجید ابتیاع
دانشجوی کارشناسی ارشد دانشکدهٔ علوم پایه، گروه ریاضی، دانشگاه آیت‌الله‌العظمی بروجردی
پست الکترونیکی: majid.ebtia@gmail.com

سیدمحمدحسینی*
استادیار دانشکدهٔ علوم پایه، گروه ریاضی، دانشگاه آیت‌الله‌العظمی بروجردی
پست الکترونیکی: sm.hoseini@abru.ac.ir

چکیده

امروزه، بانک‌ها و موسسات مالی مجموعه‌های حجیم از داده‌های مختلف و متنوعی مرتبط با مشتریان خود جمع‌آوری و نگهداری می‌کنند. رتبه‌بندی اعتباری با هدف شناسایی برچسب مشتریان اعتباری اعم از این‌که خوش‌حساب یا بدحساب هستند، سبب کاهش معوقات بانکی و ارائه بهینه و درست تسهیلات به مشتریان می‌شود و به‌عنوان یکی از مهم‌ترین روش‌های تحقیقی و عملیاتی مورد استفاده در بانک‌داری و مهندسی مالی شناخته می‌شود. در این راستا، روش‌های گوناگونی که تحت عنوان روش‌های نظارتی شناخته می‌شوند ابداع و استفاده شده است. در این روش‌ها، مدلی به کمک داده‌های برچسب‌دار موجود، ایجاد و برای پیش‌بینی برچسب داده‌های جدید استفاده می‌شود. بنابراین پیش‌فرض چنین روش‌هایی، وجود پایگاه داده‌ای است که تمام اشیاء داده‌ای آن برچسب داشته باشند. اما از آنجا که تعداد زیادی از داده‌های موجود در بانک‌ها فاقد برچسب هستند، یا برچسب دقیق و کارشناسی شده ندارند، استفاده از روش‌های نیمه‌نظارتی توصیه می‌شود. در روش‌های نیمه‌نظارتی برخلاف روش‌های نظارتی، لزومی به برچسب‌دار بودن تمام داده‌ها نیست و می‌توانند به‌وسیلهٔ مجموعه‌های داده‌ای که بخش زیادی از آن بدون برچسب هستند، مدل‌های قابل اعتمادی ایجاد کنند. روش انتشار برچسب یکی از روش‌های نیمه‌نظارتی موفق و پرکاربرد در این زمینه می‌باشد. در این روش، ویژگی ردهٔ داده‌های برچسب‌دار با یک روش تکراری به داده‌های فاقد برچسب منتشر می‌شود. در این پژوهش، ابتدا با کمک روش درخت تصمیم، ویژگی‌های تاثیرگذار بر وضعیت اعتباری مشتریان شناسایی و با کمک روش انتشار برچسب، مدل طبقه‌بندی مشتریان بانک ساخته می‌شود. به‌منظور نمایش کارایی مدل، از مجموعه داده‌های بانک پاسارگاد استفاده شده است. مدل نهایی بر روی داده‌های بانک پاسارگاد توانست به ۹۹/۷۸ درصد نرخ تشخیص مشتریان بدحساب و دقت کلی ۹۸/۷۷ درصد برسد. همچنین به‌منظور مقایسه، داده‌های یک

* نویسندهٔ مسئول

بانـــک خارجی (آلمانی) نیز مورد بررســـی قرار گرفت. بر اســـاس منابع مورد مطالعه در این تحقیق، بهترین دقت در بین مدل‌های نظارتی توســـط روش درخت تصمیم مبتنی بر خوشـــه‌بندی و الگوریتم ژنتیک بـــا دقت ۸۵/۳۷ درصد گزارش شـــده است. همچنین در بین مدل‌های نیمه‌نظارتی، روش نیمه‌نظارتی گروهی مبتنی بر پرسپترون چندلایه با دقت ۷۵/٤ درصد گزارش شـــده است، در حالی که روش نیمه‌نظارتی پیشنهادی در این مقاله، دقت ۷٦/۸۵ درصدی را کســـب کرد. یافته‌های پژوهش حاکی از عملکرد مطلوب روش نیمه‌نظارتی پیشنهادی است.

واژه‌های کلیدی: روش نیمه‌نظارتی، رتبه‌بندی اعتباری، بانک، انتشار برچسب

۱- مقدمه

رتبه‌بندی اعتباری شـــرکتی یک تحلیل از ریســـک‌های اعتباری در یک شرکت اســـت که نقش حیاتی در مدیریت ریسک مالی ایفا می‌کند. امتیازات اعتباری در همه جا وجود دارنـــد و ابزاری برای تامین‌کننـــدگان وام و تنظیم‌کنندگان وام هســـتند. امتیاز اعتباری به طور گســـترده ابزاری در صنعت مالی است که بارها اثر بخشی خود را در به حداقل رســـاندن ریسک اثبات کرده است. وظیفه اصلی یک امتیاز اعتباری، کمک بـــه ارائه‌دهندگان وام در برآورد ریســـک در حین تطبیق‌پذیری و هزینه برای محصولاتشـــان است. بانک‌ها و نهادهای مالی بخش اصلی نظام مالی هر کشوری را تشـــکیل می‌دهنـــد و نقش مهمـــی در بخش‌های مختلف اقتصادی بـــر عهده دارند. نظام بانکی با فعالیت‌های متنوع و گوناگون از قبیل تجهیز منابع، تامین نقدینگی، ابزارهای پرداخت، اعطای تسهیلات و سایر موارد بر عملکرد اقتصاد کشور تاثیر می‌گذارد. مهم‌ترین فعالیت بانک‌ها، جمع‌آوری منابع مالی و ســـرمایه‌گذاری و تخصیص آن به بخش‌های مختلف اقتصادی اســـت[۱]. در راســـتای ایفای این نقش، بانک‌ها همواره با ریســـک‌های متفاوتی روبه‌رو هستند که یکی از مهم‌ترین آن‌ها ریســـک اعتباری می‌باشـــد. ریسک اعتباری را به طور خلاصه می‌توان احتمال قصور دریافت کنندگان تسهیلات بانکی نسبت به تعهداتشان طبق شرایط توافق شده نسبت به بانک تعریف کرد. بر این اساس لازم است بانک‌ها برای کنترل و کاهش ریسک اعتباری، قبل از پرداخت تســـهیلات به مشتریان وضعیت اعتباری آن‌ها را ارزیابی کنند. به خاطر محدودیت منابع مالی و تسهیلات در اختیار بانک‌ها، ارائهٔ درست و بهینه‌ی تسهیلات و اعتبارات مالی یکی از وظایف بسیار مهم بانک‌ها به شمار می‌رود و بانک‌ها تمایل به اعطای تسهیلات خود به مشتریانی دارند که از ریسک کمتری برخوردار هستند [۲]. ارزیابی درست مشتریان بانک‌ها و موسسات مالی و مدیریت ریسک آن‌ها مباحثی حیاتی هســـتند. بحران‌های به وجودآمده در نظام بانکی کشـــورها ناشـــی از عدم توجه به مدیریت ریســـک اعتباری است و حجم عظیم معوقات بانکی، خود گویای نبود مدل‌های کارا و مناسب جهت اندازه‌گیری آن است. در نتیجه بانک‌ها و موسسات مالی برای ارزیابی ریسک اعتباری و کنترل آن نیازمند یک سیســـتم طبقه‌بندی مشتریان هستند تا با اختصاص رتبهٔ شایســـته به هر مشـــتری، تصمیمات مناسب جهت اعطای تسهیلات گرفته شود [۲]. از گذشته تا کنون پژوهش‌های متفاوتی برای ارائه مدلی مناسب و کارا جهت ارزیابی ریسک اعتباری مشتریان بانک‌ها ارائه شده اســـت که می‌توان به روش‌های آماری، تحلیل ممیزی [۲]، برنامه‌ریزی ریاضی و سیستم‌های خبره اشاره کرد [۲۲].

در عصر حاضر با پیشرفت فناوری‌های نوین اطلاعاتی و کاربردهای وســـیع آن در طبقه‌بندی اطلاعات، مدل‌هایی مبتنی بر هوش مصنوعی و یادگیری ماشـــین مورد توجه بسیاری از پژوهشگران قرار گرفته است. روش‌هایی مانند رگرسیون [۳]، شبکه‌های عصبی [۵]، الگوریتم ژنتیک [۱]، درخت تصمیم [۳]، تحلیل تشـــخیص خطی، ماشین بردار پشتیبان [۲] و بسیاری از روش‌های دیگر که زیرمجموعهٔ داده‌کاوی، یادگیـــری ماشـــین و هـــوش مصنوعـــی قرار می‌گیرند، برای اعتبارســـنجی مشـــتریان و تخمین ریسک اعتباری به‌کار گرفته شده است [۲۲]. روش‌های مذکور را

اصطلاحاً یادگیری نظارتی یا باناظر[1] می‌گویند. روش‌های یاد شـده، عموماً دقت بالایـی دارنـد و روش‌هایی قوی محسـوب می‌شـوند، در عین حال معایبی دارند که از آن جمله می‌توان به پیچیدگی محاسباتی بالا و زمان آموزش طولانی و حساسـیت به متوازن بودن برچسـب‌ها اشاره کرد [۱,۲]. امـا پیش‌فرض اصلـی در روش‌های نظارتی، برچسب‌دار بودن داده‌ها است و بنابراین روش‌های نظارتی را نمی‌توان بر روی داده‌هایی که بخشی یا تمام آن‌ها فاقد برچسب هستند آموزش داد. از طرف دیگر، بخش زیادی از داده‌های موجود در پایگاه‌های داده‌ای و به‌ویژه در بانک‌ها فاقد برچسب هستند. به‌عنوان نمونه، با توجه به فقدان یک سیستم خودکار برچسب‌گذاری در مورد داده‌های مربوط به رتبه‌بندی اعتباری مشـتریان، برچسب خوش‌حساب یا بدحسـاب اکثراً وجود ندارد یا توسط یک کارشناس برای تعداد معدودی از مشـتریان انجام شده است. لازم به ذکر است که برچسب‌دار کردن داده‌ها یک فرآیند هزینه‌بر برای بانک‌ها است. علاوه بر این، برچسبی که توسط کارشناس بانک برای هر مشـتری در نظر گرفته می‌شـود می‌تواند دارای خطای انسـانی باشـد و گاهی ممکن اسـت به‌طور سلیقه‌ای انجام شود. بنابراین استفاده از روش‌های نظارتی بر روی چنین داده‌هایی، یا به‌خاطر فقدان برچسب بخشی از داده‌های آموزشـی ناممکن است، یا در صورت ساخت مدل حتی اگر دقت بالایی داشـته باشـد، بـه دلیل خطای موجود در داده‌ها، نتایج مطلوب و موثری ارایه نمی‌کنند.

بنابـر آنچه در بالا بیان شـد، اسـتفاده از روش‌های یادگیری نیمه‌نظارتی[2] یکی از رویکردها و راه‌حل‌های مهم و مطرح در این گونه موارد است. در روش‌های نیمه‌نظارتی برخلاف روش‌های نظارتی، لزومی به برچسـب‌دار بودن تمـام داده‌ها نیسـت. این روش‌هـا، می‌توانند به‌وسـیلهٔ مجموعه‌های داده‌ای که بخش زیادی از آن بدون برچسب هسـتند، مدل‌های قابـل اعتمادی ایجاد کننـد. دیدگاه‌های نیمه‌نظارتی به چهار گروه شامل مدل‌های مولد، جداسازی بـا چگالی کم، مبتنی بر گراف، تغییر نماینده و فرا ابتکاری تقسیم می‌شوند [۲۰]. الگوریتم‌های نیمه‌نظارتی علاوه بر مزیت مهم مذکور، روش‌هایی سـاده هستند و فهم آسانی دارند و پایدار هستند [۲۰،۱۸،۹].

در ایـن مقاله، مسـئلهٔ رتبه‌بندی اعتباری مشـتریان موسسات مالی مورد مطالعه قرار گرفته است. از آنجا که بخشی از داده‌های مورد بررسی در این مقاله فاقد برچسب بودند، از روش نیمه‌نظارتی انتشار برچسب به‌عنوان یکی از روش‌های نوین یادگیری ماشین بهره گرفته شده است.

پژوهش حاضر به پنج بخش سـازماندهی شده است. پس از مقدمه، در بخش دوم مروری جامع بر مبانی نظری و پیشینه پژوهش خواهیم داشت. سپس در بخش سوم به روش‌شناسی، ویژگی‌های داده‌ها و شناخت آن‌ها پرداخته می‌شـود. در بخش چهارم برآورد مدل و نتایج گردآوری شـده اسـت. نهایتا در بخش پایانی، نتیجه‌گیری و ارایه‌ی پیشنهادات آورده شده است.

۲- مبانی نظری و مروری بر پیشینه پژوهش

قدمـت ارزیابی ریسـک اعتباری به زمـان ایجاد پول برمی‌گردد و زمانی اسـت که افرادی بـه گروه‌ها و افراد مختلف پول قرض می‌دادند و توانایی مالی آن‌ها را در نظر می‌گرفتنـد. اما در آن زمان متغیرها و ویژگی‌های اثرگذار بـر وضعیت مالی و اعتباری افـراد محدود بود و با توجه به شـناخت قبلی، وضعیت اعتباری افراد مشخص می‌شد. اندازه‌گیری و رتبه‌بندی ریسک اعتباری یک موضوع بسیار مهم و واقعی اسـت و نیاز آن امروزه به شـدت احساس می‌شود، کسب و کارهای کلان، بانک‌ها، موسسات مالی و ... به شـدت نیازمند تجزیه و تحلیل صحیح و دقیق ریسک اعتباری خود می‌باشـند. امروزه بخـش عظیمی از داده‌ها را داده‌های بدون برچسـب تشکیل می‌دهند، به همین علت روش‌های نیمه‌نظارتی به‌عنـوان راهکاری کاربردی برای حل این مشـکل توصیه می‌شـود. روش‌های نیمه‌نظارتی عـلاوه بر کاهش هزینه‌هـای محاسـبات و کاهش هزینه

1- Supervised learning
2- Semi-supervised learning

برچسب‌گذاری از تعمیم‌پذیری بهتری برخوردار هستند و پایدار می‌باشند [۲۰،۱۸،۹]. در این زمینه پژوهش‌های داخلی و خارجی مختلفی با روش‌های متفاوت انجام شده است که به برخی از آن‌ها اشاره می‌شود.

طلوعی و همکارانش (۱۳۸۸) با کمک مدل ترکیبی ماشین بردار پشتیبان مبتنی بر دو راهبرد به امتیازدهی اعتباری متقاضیان کارت‌های اعتباری می‌پردازد. آن‌ها با استفاده از تکنیک رتبهٔF، ویژگی‌های مهم و تاثیرگذار را شناسایی و برای پیدا کردن مقادیر پارامترهای بهینهٔ ماشین بردار پشتیبان از جستجوی شبکه استفاده کردند [۱]. الگوریتم ژنتیک به‌عنوان یکی از تکنیک‌های تکاملی مبتنی بر جمعیت شناخته می‌شود و در بهینه‌سازی کاربرد فراوانی دارد. اقبالی و همکارانش (۱۳۹٦) به بررسی ارزیابی عملکرد توابع شایستگی الگوریتم ژنتیک در رتبه‌بندی مشتریان پرداختند و بین عملکرد این الگوریتم با روش‌هایی چون رگرسیون لجستیک و تحلیل پوششی داده‌ها مقایسه‌ای ارایه می‌کند [۲]. کشاورزحداد و آیتی‌گازار (۱۳۸٦) در پژوهشی، مدل‌های رگرسیون لجستیک و درخت تصمیم برای اعتبارسنجی مشتریان بانکی را مقایسه کردند. آن‌ها نشان دادند که خطای نوع اول و دوم برای مدل درخت تصمیم در مقایسه با رگرسیون لجستیک کمتر است [۳]. پویان‌فر و همکارانش (۱۳۹۲) با به‌کارگیری روش حداقل مربعات ماشین بردار پشتیبان مبتنی بر ژنتیک به تخمین رتبهٔ اعتباری مشتریان بانک‌ها می‌پردازد. نتایج این پژوهش با مدل‌های آماری مانند لاجیت و رویکردهای بهینه‌سازی پارامترهای ماشین بردار پشتیبان مقایسه شد و نتایج این پژوهش حاکی از عملکرد مطلوب این مدل می‌باشد [٤]. مهرآرا و همکارانش (۱۳۸۸) با انتخاب نمونهٔ ٤۰۰ تایی از مشتریان بانک پارسیان، به کمک روش‌های رگرسیون و شبکه‌های عصبی به ارزیابی ریسک اعتباری مشتریان پرداختند. آن‌ها نشان دادند که مدل مبتنی بر شبکه‌های عصبی نسبت به مدل رگرسیونی از دقت بالاتری برخوردار است [٥]. حاجی‌کرد و همکارانش در سال ۱۳۹٥، بر روی داده‌های مشتریان بانک تجارت با استفاده از مدلی ترکیبی به بررسی ریسک اعتباری پرداختند. در این پژوهش ابتدا از الگوریتم ژنتیک به‌منظور بهینه‌سازی داده‌های ورودی استفاده شد. سپس مدل، بر اساس روش ماشین بردار پشتیبان به وجود آمد. بهینه‌سازی داده‌های ورودی، منجر به افزایش دقت بر روی داده‌های آزمایشی و تعمیم‌پذیری مدل می‌شود [٦]. شریعت‌پناهی و هاشمی‌برکادهی (۱۳۸۷) با ارائهٔ مدل تحلیل ممیزی، اعتبارسنجی مشتریان بانک صنعت و معدن و همچنین پیش‌بینی نکول شرکت‌های دریافت‌کنندهٔ تسهیلات را انجام داد [۷].

معیار فاصله نقش بسیار مهم و کلیدی در بسیاری از الگوریتم‌های یادگیری ماشین دارد، به طوری که انتخاب تابع فاصله مناسب، تاثیر بسزایی بر عملکرد این الگوریتم‌ها دارد. با گذشت زمان و گسترش الگوریتم‌ها، فاصله‌ها جای خود را به هسته‌ها دادند و انتخاب هستهٔ مناسب و پارامترهای آن به یک مسئله بسیار مهم تبدیل شد. زارع بیدکی و همکارانش (۱۳۹٦) الگوریتم یادگیری نیمه‌نظارتی هسته مرکب که از روش‌های یادگیری مبتنی بر فاصله است را تشکیل داد و از آن برای سنجش فاصله داده‌ها در خوشه‌بندی استفاده کرد [۸]. روش‌های یادگیری نیمه‌نظارتی که مبتنی بر گراف هستند بیشتر اوقات بر روی مسئله‌های تک برچسبی اجرا شده‌اند. کردآبادی و همکارانش (۱۳۹۸) در پژوهشی یک مدل نیمه‌نظارتی ترکیبی که مبتنی بر گراف و یادگیری چند برچسبی است تشکیل و نشان دادند این مدل نسبت به سایر مدل‌ها، به‌ویژه زمانی که تعداد نمونه‌های برچسب‌دار کم است عملکرد بهتری دارد [۹]. کریچن (۲۰۱۷) با الگوریتم بیز ساده به ارزیابی ریسک وام ۹۲٤ پروندهٔ اعتبارات اعطایی که به شرکت‌های صنعتی تونسی توسط یک بانک داده شده بود پرداخت و دقت مناسبی نزدیک به ٦۳ درصد به‌دست آورد. از نتایج حاصل می‌توان نقش وثیقه را در پیش‌بینی ریسک پیش‌فرض نام برد. در حقیقت این شاخص ظرفیت توضیحی برای پیش‌بینی ریسک اعتباری دارد [۱۱].

هانگ و همکاران (۲۰۰۷) برای امتیازدهی اعتباری مشتریان با یک روش استخراج داده بر اساس ماشین بردار پشتیبان، مدل‌هایی بر روی داده‌های دو بانک آلمانی و استرالیایی ارایه دادند. دقت مدل آن‌ها بر روی داده‌های بانک آلمانی نزدیک به ۷۶ درصد و بر روی داده‌های بانک استرالیایی نزدیک به ۸۵ درصد به دست آمد [۱۲]. در یادگیری ماشین باناظر، فرآیند به دست آوردن داده‌های برچسب‌دار پرهزینه و طولانی است. هان (۲۰۲۰) با کمک الگوریتم یادگیری نیمه‌نظارتی که از داده‌های برچسب‌دار و بدون برچسب بهره می‌گیرد، بر اساس رگرسیون لجستیک افزایشی تعمیم یافته به تشخیص ناهنجاری اعتبار شرکتی پرداخت. در این پژوهش نتایج حاصل شده از مطلوبیت خوبی برخوردار بود و همچنین ویژگی‌های مهم و تاثیرگذار شناسایی شدند [۱۳]. لیویریس (۲۰۱۸) در پژوهشی بر روی سه مجموعه داده بانکی استرالیایی، ژاپنی و آلمانی با کمک روش‌های گروهی نیمه‌نظارتی به ارزیابی ریسک اعتباری پرداخت که نتایج بسیار مناسبی حاصل شد [۱۴]. کیم (۲۰۱۹) با یک روش یادگیری گروهی نیمه‌نظارتی مبتنی بر ماشین بردار پشتیبان، عدم پرداخت بدهی در وام‌های بانکی را پیش‌بینی کرد [۱۵]. فنگ (۲۰۲۱) با کمک یک روش دوفازی و ترکیبی شرکت‌های مالی را رتبه‌بندی کرد. نتیجه حاصل شده از این روش ترکیبی، بهبود عملکرد و دقت است [۱۶].

۲-۱ یادگیری ماشین

یادگیری ماشین زیرمجموعه‌ای از هوش مصنوعی است که می‌توان با کمک آن سیستم یا برنامه‌ای طراحی کرد که الگوها را بر اساس تجربه یاد بگیرد [۱۸]. الگوریتم‌های یادگیری ماشین در یک دسته بندی به سه دسته یادگیری بانظارت، بدون نظارت و نیمه‌نظارتی تقسیم می‌شوند [۲۰]. در یادگیری بدون نظارت، تمامی داده‌ها بدون برچسب هستند و هدف از این روش خوشه‌بندی داده‌ها می‌باشد. منظور از خوشه‌بندی یعنی تقسیم مجموعهٔ داده‌ای به چندین زیرمجموعه به‌گونه‌ای که داده‌های داخل هر زیرمجموعه بیشترین شباهت به یکدیگر و کمترین شباهت به داده‌های سایر خوشه‌ها داشته باشد. در یادگیری بانظارت، تمام داده‌ها دارای برچسب هستند و هدف از این روش پیدا کردن مدل یا الگویی است که بتواند ویژگی برچسب هر داده را به طور مطلوب مشخص کند [۱۸]. یکی از روش‌های بسیار مهم در یادگیری نظارتی، درخت تصمیم می‌باشد که برای رگرسیون و دسته‌بندی به‌کار می‌رود. در این روش یک درخت تصمیم با تقسیمات دودویی تشکیل می‌گردد که در آن برای تقسیم‌بندی و انشعاب بر روی داده‌ها از شاخص جینی استفاده می‌شود.

در یادگیری نیمه‌نظارتی، هدف الگوریتم استفاده همزمان از داده‌های برچسب‌دار و بدون برچسب برای تقویت یادگیری است. الگوریتم‌های یادگیری نیمه‌نظارتی به طور کلی به چهار دسته روش‌های مولد، جداسازی کم چگالی، مبتنی بر گراف و فرا ابتکاری تقسیم می‌شوند [۲۰]. در این مقاله، یک الگوریتم تکراری مبتنی بر تکثیر یا انتشار برچسب که در آن همزمان از داده‌های برچسب‌دار و بدون برچسب استفاده می‌شود بررسی می‌گردد. داده‌های برچسب‌دار برای آموزش الگوریتم‌های نظارت‌شده ضروری هستند هرچند اغلب تعداد کمی از آن‌ها موجود است ولی داده‌های بدون برچسب فراوان است.

مجموعه داده کلی به‌صورت $X = \{x^1, \cdots, x^{n+m}\}$ که در آن $x^i \in R^D$، را در نظر بگیرید. در این مجموعه، هر داده دارای $\mathbf{D}$ ویژگی یا متغیر و یک متغیر وابسته یا رده است اما تنها برچسب تعداد n داده از این مجموعه مشخص شده‌اند. بنابراین n دادهٔ برچسب‌دار وجود دارد و بقیه‌ی داده‌ها بدون برچسب هستند. بدون کاستن از کلیت، فرض کنید $(x^1, y_1), \ldots, (x^n, y_n)$ داده‌های برچسب‌دار این مجموعه باشند. برچسب‌های رده تمام داده‌های برچسب‌دار را با بردار $Y_L = (y_1, \cdots, y_n)$ نشان دهید. همچنین فرض کنید C تعداد رده‌های مختلف موجود در داده‌های برچسب‌دار را نشان می‌دهد و از هر رده در مجموعهٔ داده‌ای برچسب‌دار تعدادی موجود است. در نظر

بگیرید $(x^{n+1}, y_{n+1}), \ldots, (x^{n+m}, y_{n+m})$ داده‌های بدون برچسب باشند. به عبارت دیگر y_{n+i} به ازای $i = 1,2, \ldots, m$ نامشخص است. از بردار $Y_U = (y_{n+1}, \cdots, y_{n+m})$ برای برچسب داده‌های با برچسب نامشخص استفاده کنید. در این مجموعه داده‌ای، به دنبال آن هستیم نقاطی از داده‌ها که نزدیک به هم هستند برچسب‌های مشابه داشته باشند.

روش انتشار برچسب به شیوهٔ زیر برچسب داده‌های بدون برچسب را تعیین می‌کند. ابتدا یک گراف کامل وزن‌دار که در آن راس‌ها همهٔ داده‌های مجموعه شامل داده‌های با برچسب و بدون برچسب هستند ایجاد می‌شود. یال‌های این گراف با ω_{ij} که نشان‌دهندهٔ وزن میان دو راس i, j است وزن‌دار می‌شوند. در این مقاله از فاصلهٔ اقلیدسی برای وزن‌دار کردن یال‌ها استفاده شده است. نکتهٔ مهمی که باید به آن اشاره کرد این است که به‌جای فاصله اقلیدسی می‌توان هر معیار فاصله دیگری قرار داد. در اینجا با توجه به راحتی و متداول بودن فاصله اقلیدسی از آن استفاده شده است. وزن یال بین دو راس به‌صورت زیر مقداردهی می‌گردد:

$$\omega_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) = \exp\left(-\frac{\sum_{d=1}^{D}\left(x_d^i - x_d^j\right)^2}{\sigma^2}\right) \qquad (۱)$$

که در آن وزن‌ها به کمک پارامتر σ کنترل می‌شوند اما لازم به‌ذکر است که می‌توان برای هر متغیر یک پارامتر σ_d مخصوص آن به‌کار برد. در این‌صورت داریم

$$\bar{\omega}_{ij} = \exp\left(-\bar{d}_{ij}^2\right) = \exp\left(-\sum_{d=1}^{D}\frac{\left(x_d^i - x_d^j\right)^2}{\sigma_d^2}\right) \qquad (۲)$$

در این مقاله، از یک پارامتر σ برای همهٔ ویژگی‌ها استفاده شده است. این شیوهٔ وزن‌دار کردن یال‌ها به گونه‌ای است که هر چه گره‌ها از لحاظ فاصله اقلیدسی نزدیک‌تر باشند، وزن ω_{ij} بزرگتر است و از منظر هدف الگوریتم انتشار برچسب، دو راسی که وزن بیشتر و در نتیجه فاصله اقلیدسی کمتری دارند شباهت بیشتری وجود دارد و اگر یکی از آن‌ها برچسب‌دار و دیگری بدون برچسب باشد همان برچسب به راس بدون برچسب نسبت داده خواهد شد. در واقع برچسب راس‌های برچسب‌دار از طریق یال‌ها به راس‌های بدون برچسب منتقل و منتشر می‌شود و هرچه وزن یالی بزرگتر باشد انتشار برچسب از طریق آن آسان‌تر خواهد بود. برای انتشار برچسب، ماتریس انتقال T با ابعاد $(n+m) \times (n+m)$ به شکل زیر نیاز خواهد شد

$$T_{ij} = \frac{\omega_{ij}}{\sum_{k=1}^{n+m}\omega_{kj}} \qquad (۳)$$

که در آن T_{ij} درایهٔ سطر i و ستون j است و احتمال پرش از گره j به i را محاسبه می‌کند. همچنین ماتریس Y با ابعاد $(n+m) \times C$ که ماتریس برچسب نامیده می‌شود و سطر i–ام آن نشان‌دهنده توزیع احتمال برچسب راس x^i در رده‌های مختلف می‌باشد. با توجه به این‌که برچسب راس‌های برچسب‌دار معلوم است لذا در ماتریس Y سطرها مربوط به آن‌ها در ستون رده‌شان مقدار ۱ و بقیه ستون‌ها صفر است. الگوریتم انتشار برچسب یک روش تکراری است که با مقداردهی اولیهٔ ماتریس Y شروع و در انتهای الگوریتم، ماتریس Y با احتمال تعلق هر داده به هر یک از رده‌ها به‌عنوان خروجی ارایه می‌شود. ثابت می‌شود که مقدار اولیهٔ سطرهایی از Y که متناظر با راس‌های بدون برچسب است اهمیتی ندارند و با هر مقدار بین صفر و یک، به جواب همگرا می‌شوند (چافیل و همکاران، ۲۰۰۶).

مراحل الگوریتم انتشار برچسب به‌صورت زیر است:

۱- برچسب را انتشار بده: TY را محاسبه و جایگزین Y کن.

۲- سطرهای Y را نرمال کن.

۳- سطرهای مربوط به داده‌های برچسب‌دار را به حالت اول برگردان و به مرحله ۱ برو.

این سه مرحله می‌تواند تا زمان همگرایی Y تکرار شود یا بعد تعداد مشخصی تکرار، تمام شود. همچنین هر معیار دیگری برای توقف الگوریتم می‌تواند استفاده شود. در مرحله ۱، همهٔ راس‌ها برچسب‌های خود را یک واحد منتشر می‌کنند. در مرحله ۲، نرمال شدن سطرهای Y برای تفسیر احتمالاتی برچسب‌ها انجام می‌شود. با توجه

جدول ۱: نتایج پژوهش‌های مشابه در مورد رتبه‌بندی اعتباری مشتریان بانک به کمک روش‌های یادگیری ماشین

منبع اطلاعات	دسته	روش	داده‌ها	مزایا	معایب	دقت
هانگ (۲۰۰۷)	نظارتی	ماشین بردار پشتیبان مبتنی بر جست‌وجوی شبکه و امتیاز F	بانک آلمانی	انتخاب ویژگی‌های مناسب و بهینه‌سازی پارامترها	افزایش هزینه محاسبات و کاهش سرعت الگوریتم	۷۷,۹۲
گتاشیا(۲۰۱۴)	نظارتی	جنگل تصادفی	بانک آلمانی	تفسیر پذیری بالا	پیچیدگی محاسباتی و حساسیت به داده‌ها	۷۷,۲
کشاورزحداد(۱۳۸۶)	نظارتی	درخت طبقه‌بندی و مدل لاجیت	بانک مسکن	بدون پیش‌فرض بر داده‌ها	حساس به تعداد داده‌ها	۸۵,۵
مهرآرا(۱۳۸۸)	نظارتی	شبکه عصبی	مشتریان حقوقی بانک پارسیان	دقت بالا، مقاوم در برابر نوفه و شناسایی روابط غیرخطی	پیچیدگی محاسبات	۸۶
طلوعی(۱۳۸۹)	نظارتی	ماشین بردار پشتیبان مبتنی بر جست‌وجوی شبکه و امتیاز F	مشتریان حقوقی یک بانک داخلی	انتخاب ویژگی‌های مناسب و بهینه‌سازی پارامترها	افزایش هزینه محاسبات و کاهش سرعت الگوریتم	۷۶,۷
پویان‌فر(۱۳۹۲)	نظارتی	ماشین بردار پشتیبان مبتنی بر ژنتیک	بانک آلمانی	افزایش دقت مدل	پیچیدگی محاسبات و افزایش زمان محاسبات	۸۴
محمدیان(۱۳۹۵)	نظارتی	ماشین بردار پشتیبان مبتنی بر ژنتیک	مشتریان حقوقی بانک تجارت	بهینه‌سازی پارامترها	کاهش سرعت الگوریتم	۶۹
اقبالی(۱۳۹۶)	نظارتی	مقایسه توابع شایستگی الگوریتم ژنتیک	مشتریان حقیقی یک بانک داخلی	افزایش دقت مدل	هزینه محاسباتی بالایی دارد	۴۴
لیوبریس(۲۰۱۸)	نیمه‌نظارتی	یادگیری نیمه‌نظارتی ترکیبی	بانک آلمانی	افزایش دقت مدل و ساخت دسته‌بند قوی‌تر	افزایش پیچیدگی مدل	۷۵,۴

به این‌کـه در مرحله ۱ و ۲ ممکن اسـت توزیع احتمالاتی راس‌های برچسـب‌دار به‌خاطر تاثیـر گرفتن از راس‌های مجاور تغییر کرده باشـند در مرحلهٔ ۲، سطرهای مربوطه به داده‌های برچسـب‌دار به حالت ثابـت اولیه برگردانده می‌شـوند. بنابراین به جـای این‌که اجازه داده شـود که راس‌های برچسـب‌دار به مـرور در راس‌های دیگر محو شوند، با دسـت‌زدن به توزیع برچسب‌های آن‌ها، دوباره آن‌ها اصلاح می‌شوند.

در ایـن مقاله به‌منظور تنظیم پارامتر σ به‌صورت زیر عمل می‌شود. ابتدا با کمک الگوریتم کروسکال(کروسکال، ۱۹۵۶)، درخت پوشـای مینیمم بر گـراف کامل پیدا کنید. سـپس کوتاه‌ترین یال از این درخت را به‌گونه‌ای پیدا کنید که بین دو راس با برچسب مختلف است. چنانچه وزن این یال a باشد مقدار $\sigma = a / 3$ لحاظ می‌شود. تقسیم به سه به‌خاطر قانون 3σ در توزیع نرمال است و به‌گونه‌ای است که وزن این یال نزدیک به میانگین توزیع نرمال استاندارد یعنی صفر باشـد به این امید که انتشار محلی برچسب در رده‌هـای مختلف با همین پراکندگی توزیع شـود[۲۰,۲۱].

در جدول ۱، مقایسـه‌ای بین پژوهش‌های مشـابه گزارش می‌شود.

۳- مدل پژوهش، ویژگی‌ها و شناخت آن‌ها

اعتبارسـنجی و ارزیابـی ریسـک اعتبـاری شـامل طبقه‌بندی متقاضیان تسهیلات اعتباری بر اساس ویژگی‌ها و شـرایط متقاضی مثـل وضعیت کاری، میزان حسـاب بانکی، وضعیت مسکن تسهیلات و ... بررسی می‌شود. در ایـن بخش به شـرح مدل به کار برده شـده در این مقاله، برای طبقه‌بندی جامعه آماری متقاضیان تسهیلات اعتباری و تعیین وضعیت اعتباری آن‌ها پرداخته می‌شود. مدلی که بر اساس داده‌ها سـاخته می‌شود را اصطلاحاً یادگیرنده، طبقه‌بند یا دسـته‌بند می‌نامند. جامعه آماری در نظر گرفته شده در این پژوهش مشـتریانی هستند که از بانک مورد نظر تسهیلات دریافت کرده‌اند و آن‌ها را به بانک بازگشت داده یا نداده‌اند.

جدول ۲: مقادیر ویژگی‌های داده‌های آلمانی

نماد	عنوان ویژگی	نوع	مقادیر ویژگی										
g_1	وضعیت حساب	اسمی	بدون حساب	حساب خالی	حساب کمتر از DM 200	حساب بیش از DM 200							
g_2	مدت اعتبار	عددی	عددی (برحسب ماه)										
g_3	سابقه اعتبار	اسمی	بدون سابقه	بازپرداخت به موقع در این بانک	بازپرداخت به موقع تمام تسهیلات	تاخیر در بازپرداخت	حساب بحرانی یا اعتبارات در بانک‌های دیگر						
g_4	هدف تسهیلات	اسمی	کسب و کار	لوازم خانه	تحصیلات	مسافرت	مبلمان و تجهیزات	ماشین جدید	ماشین دست دوم	رادیو تلویزیون	تعمیر	بازآموزی	موارد دیگر
g_5	مقدار اعتبار	عددی	عددی										
g_6	وضعیت پس‌انداز	اسمی	کمتر از 100 DM	بین 100 تا 500 DM	500 DM تا 1000 DM	بیش از DM 1000	نامشخص یا بدون حساب پس‌انداز						
g_7	سابقه کار	اسمی	بیکار	کمتر از ۱ سال	بین ۱ تا ۴ سال	بین ۴ تا ۷ سال	بیش از ۷ سال						
g_8	تعداد اقساط	عددی	عددی										
g_9	وضعیت تاهل و جنسیت	اسمی	مرد مجرد	مرد مطلقه یا جدا شده	مرد متاهل	زن مجرد	زن متاهل یا مطلقه						
g_{10}	سایر بدهکاران یا ضامنین	اسمی	بدون ضامن	متقاضی مشترک	ضامن								
g_{11}	مدت زمان اقامت فعلی	عددی	عددی										
g_{12}	اموال و دارایی‌ها	اسمی	ماشین	بیمه زندگی	املاک و ساختمان	اموال ناشناخته							
g_{13}	سن	عددی	عددی										
g_{14}	سایر تسهیلات	اسمی	بانکی	فروشگاه‌ها	ندارد								
g_{15}	مسکن	اسمی	رایگان	مالک	اجاره								
g_{16}	اعتبارات موجود در این بانک	عددی	عددی										
g_{17}	شغل	اسمی	بیکار/بی‌تجربه یا غیرمقیم	بی‌تجربه و مقیم	کارمند یا متخصص	مدیر/آزاد/کارمند/فوق تخصص/افسر							
g_{18}	تعداد عائله	عددی	عددی										
g_{19}	مالکیت تلفن	اسمی	خیر	بلی									
g_{20}	کارگر خارجی	اسمی	خیر	بلی									
g_{21}	طبقه(رده)	اسمی	بد	خوب									

نمونــه آمــاری ایــن پژوهــش شــامل یــک بانــک خارجی(آلمانــی) که از پایگاه داده یادگیری ماشــین UCI جمع آوری شــده اســت. تعداد نمونه‌های داده‌های بانک آلمانی کــه در این پژوهش اســتفاده شــده ۱۰۰۰ نمونه اســت که هر نمونــهٔ آن دارای ۲۰ ویژگی اســت. به طور کلی ویژگی‌ها به دو بخش کیفی و کمی تقســیم می‌شوند. ویژگی‌های این مجموعه، شــامل ۲۰ ســتون (۱۳ ویژگی اســمی و ۷ ویژگی عددی) اســت که در جدول ۲ گزارش شــده است. در این مجموعه، هر مشتری یک ویژگی هدف یعنی وضعیت اعتباری دارد کــه برای هرکدام از نمونه‌ها با یکی از دو مقدار خوب (مشــتریان خوش‌حســاب) و بد (مشــتریان بدحساب) مشخص شده است. در این پژوهش وضعیت اعتباری به‌صورت دودویی یعنی ۱ برای مشتری خوش‌حساب و ۱- برای مشتری بدحساب در نظر گرفته شده است.

همچنین یک مجموعهٔ داده‌ای داخلی یعنی بانک پاسارگاد نیز بررسی شده است. در این مجموعه، صورت‌های مالی ۲۲۱۸ مشتری حقوقی وجود دارد که هر مشتری دارای ۱۴ متغیر (ویژگی) ۱ اســت و در جدول ۳ ارایه شده است. در هر دو مجموعه بانکی ۸۰ درصد از نمونه‌ها برچسب آن‌ها ناشناخته است.

مدل پژوهش

فرآیند مدل پیشــنهادی در ایــن پژوهش به دو مرحله تقسیم می‌شــود. پس از جمع‌آوری داده‌های مورد نظر و مناسب به پیش‌پردازش آن‌ها پرداخته می‌شود.

● **گام اول:** این مرحله شــامل پاک‌ســازی، نرمال‌سازی و انتخاب ویژگی می‌باشــد. هر دو مجموعه داده بانکی فاقد مقادیر مفقودی و نوفه می‌باشند. با کمک نرم‌افزار یادگیری ماشــین پایتون به پاک‌ســازی داده‌ها که شامل شناسایی داده‌های پرت اســت با روش‌های مختلف پرداخته شــد و همچنین ویژگی‌های اســمی به عددی تبدیل شــدند. برای نرمال‌ســازی و مقیاس‌بندی مجموعــه داده‌ها روش‌های متفاوتی وجود دارد. به طور کلی می‌توان هر ویژگی را در محدوده [۱-،۱] یا [۰،۱] مقیاس بندی کرد.

بعد از نرمال‌ســازی، ویژگی‌های مهــم و تاثیرگذار بر وضعیت اعتباری مشــتریان شناسایی و انتخاب می‌شود. در ایــن مقاله، ایــن کار با کمک روش درخت رگرســیون و دســته‌بندی انجــام می‌شــود. در ایــن روش داده‌ها به شــاخه‌های دودویی تقســیم می‌شود و تقســیم داده‌ها و انشــعاب روی آن‌ها با کمک معیار جینی صورت می‌گیرد ایــن کار تا زمانی ادامه پیدا می‌کند کــه یکی از معیارهای توقف در درخت تصمیم مثل هرس کامل درخت یا رسیدن بــه گره‌های خالــص تحقق پیدا کند. با کمــک معیار جینی ویژگی‌های مهم و تاثیرگذار بر وضعیت اعتباری مشتریان شناسایی و انتخاب می‌شوند.

● **گام دوم:** داده‌هــای حاصل از پیش‌پردازش، به الگوریتم انتشار برچسب ارایه می‌شــود. برای ساخت مدل ابتدا به روش اعتبارسنجی، پارامترهای مدل انتشار برچسب تنظیم و بهترین آن‌ها انتخاب می‌شود. پارامترهای بهینهٔ الگوریتم به‌صورت زیر محاسبه می‌شوند. یکی از زیرمجموعه‌های داده‌های آموزشــی انتخاب می‌شــود. ابتدا یک شــبکه از مقادیــر ابرپارامترها انتخاب می‌شــود و به ازای هر نقطه از این شــبکه، ابرپارامترهای روش مقداردهی می‌شــوند. سپس از روش اعتبارسنجی متقابل برای ارزیابی عملکرد ابرپارامتر در نقطهٔ مذکور بهره برده خواهد شد.

در روش اعتبارســنجی متقابل، داده‌های آموزشی به k قسمت مســاوی تقسیم می‌شــوند. از زیرنمونه‌ها، یکی به‌عنوان داده‌های اعتبارســنجی برای آزمایش مدل حفظ می‌شــود و بقیه زیــر نمونه‌ها مدل را می‌ســازند. بعد از ساخت مدل، زیرنمونهٔ آزمایشــی برای میزان اعتبار مدل ساخته شده به‌کار گرفته می‌شود. به این ترتیب به ازای هر نقطهٔ شبکه، فرآیند اعتبارسنجی متقابل، k بار تکرار می‌شود و در هر بار معیار عملکرد آن نقطه ســنجیده می‌شود. در انتها از میانگین معیارهای به‌دســت آمده برای ســنجش میــزان اعتبار آن نقطه بهره برده می‌شــود. در میان نقاط

جدول ۳: مقادیر ویژگی‌های داده‌های پاسارگاد

نماد	عنوان ویژگی	نوع	مقادیر ویژگی							
p_1	جنسیت	اسمی	مرد						زن	
p_2	مقدار تسهیلات	عددی	عددی							
p_3	مبلغ وجه التزام	عددی	عددی							
p_4	نوع وثیقه	اسمی	اوراق مشارکت	سند ملکی ترهینی- مسکونی/ اداری	سپرده سرمایه‌گذاری	اسناد عادی	چک تضمینی	تعهدنامه	افرازنامه	سفته تضمینی
p_5	سن	عددی	عددی							
p_6	نحوه بازپرداخت	عددی	شش ماه						سه ماهه	
p_7	میزان تحصیلات	اسمی	دیپلم	فوق دیپلم	لیسانس		فوق لیسانس		دکتری	
p_8	میزان سابقه کار	عددی	عددی							
p_9	وضعیت مسکن	اسمی	مالک						اجاره	
p_{10}	سابقه چک برگشتی	اسمی	دارد						ندارد	
p_{11}	حدود درآمد ماهانه	عددی	عددی							
p_{12}	اموال و دارایی فعلی	عددی	عددی							
p_{13}	متوسط سپرده	عددی	عددی							
p_{14}	طبقه(رده)	اسمی	خوب						بد	

مختلف شبکهٔ مذکور، نقطه‌ای که بهترین اعتبار را تولید کرده است به‌عنوان ابرپارامترهای مدل استفاده می‌شود. بعد از انتخاب بهترین پارامترها، الگوریتم انتشار برچسب روی مجموعهٔ داده اعمال و مدل ساخته می‌شود.

معیارهای ارزیابی

برای ارزیابی و مقایسهٔ نتایج حاصل شده و آزمون عملکرد روش‌های ارائه شده از معیارهای زیر بهره گرفته شده است. تعداد نمونه‌های ردهٔ بدحساب که به درستی بدحساب تشخیص داده شده‌اند را با TP، تعداد نمونه‌های ردهٔ خوش‌حساب است که به درستی خوش‌حساب تشخیص داده شده‌اند را با TN، تعداد نمونه‌های ردهٔ بدحساب که به اشتباه خوش‌حساب تشخیص داده شده‌اند را با FP و تعداد نمونه‌های ردهٔ خوش‌حساب که به اشتباه بدحساب تشخیص داده شده‌اند را با FN نشان دهید. در این‌صورت معیارهای زیر به‌منظور عملکرد دسته‌بند به طور متداول استفاده می‌شود.

معیار اول که نرخ تشخیص نمونه‌های بدحساب نامیده می‌شود بسیار مهم و حائز اهمیت می‌باشد چرا که هزینه عدم تشخیص مشتریان بد حساب بسیار بیشتر از هزینه عدم تشخیص مشتریان خوش حساب می‌باشد. هزینه مورد نظر شامل از دست دادن تسهیلات به‌صورت اصل و فرع به همراه هزینه پیگیری مطالبات معوق می‌باشد. اختلاف این معیار از یک را می‌توان به‌عنوان ریسک اعتباری مدل ساخته شده در نظر گرفت.

$$TNR = \frac{TN}{TN + FP} \qquad (۴)$$

معیار دوم که نرخ تشخیص نمونه‌های خوش‌حساب نامیده می‌شود برای شناسایی مشتریان خوش حساب به‌کار می‌رود. پیش‌بینی نادرست و اشتباه مشتریان خوش

حساب سبب کاهش حاشیه سود بانکی می‌شود. اختلاف این معیار از یک را می‌توان به‌عنوان ریسک تجاری مدل مورد استفاده در نظر گرفت.

$$TPR = \frac{TP}{TP + FN} \quad (۵)$$

معیار سوم دقت کلی مدل است، که نشان‌دهنده توانایی مدل در دسته‌بندی کلی مشتریان چه خوش‌حساب و چه بدحساب می‌باشد.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (۶)$$

معیار چهارم نشان‌دهنده پیش‌بینی‌های صحیح انجام شده برای مشتریان بدحساب را نشان می‌دهد.

$$PPV = \frac{TP}{TP + FP} \quad (۷)$$

چون معیارهای دوم و چهارم عکس یکدیگر هستند و افزایش یکی باعث کاهش دیگری و بالعکس می‌شود پس از معیار دیگری استفاده می‌شود تا عملکرد کلی مدل ارزیابی شود این معیار برای داده‌های نامتوازن بسیار مناسب است. این معیار ترکیب دو معیار دوم و چهارم (میانگین هندسی) می‌باشد.

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (۸)$$

معیار ششم نشان‌دهنده خطای نوع اول می‌باشد.

$$EI = \frac{FP}{FP + TN} \quad (۹)$$

معیار هفتم نشان‌دهنده خطای نوع دوم می‌باشد.

$$EII = \frac{FN}{TP + FN} \quad (۱۰)$$

معیار هشتم که بیان‌گر کیفیت کلاس‌بندی می‌باشد.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (۱۱)$$

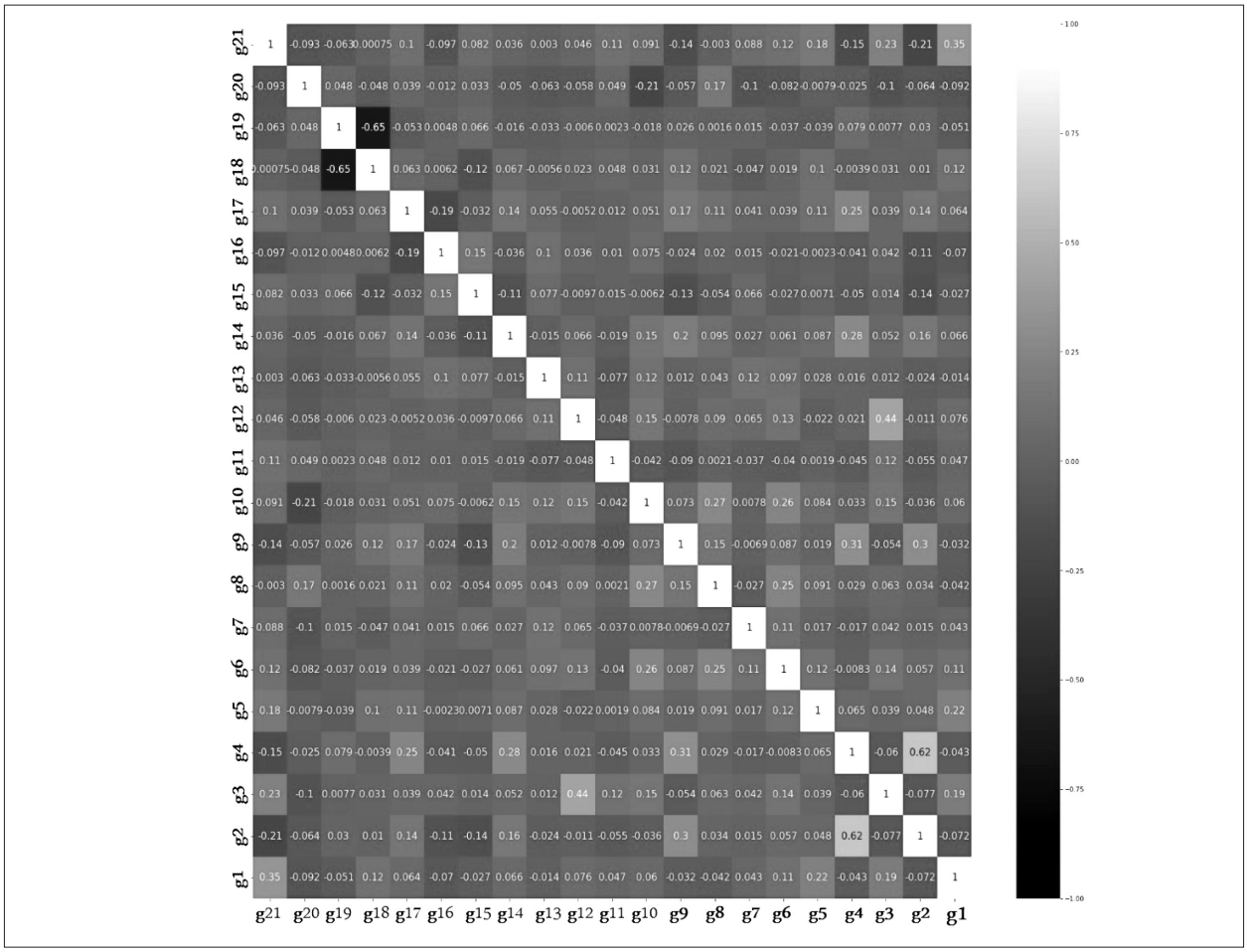
مساحت زیر سطح منحنی کارایی نشان‌دهنده قدرت یک مدل یا آزمون است که این مقدار عددی بین ۰ تا ۱ می‌باشد و هرچه به ۱ نزدیک‌تر باشد مدل کارایی بیشتری دارد. معیار نهم مساحت زیر سطح منحنی کارایی است که با AUC نشان داده می‌شود. در نمودار منحنی کارایی، محور عمودی نشان‌دهنده معیار ارزیابی اول و محور افقی نشان‌دهنده معیار ارزیابی پنجم است.

از آن‌جا که هر یک از معیارهای ارزیابی از اهمیت خاصی برخوردار بوده پس لازم است در ارزیابی عملکرد مدل از هر نه معیار در کنار یکدیگر استفاده شده و مورد بررسی قرار گیرند.

۴- برآورد مدل و نتایج پژوهش

در این قسمت نتایج حاصل از مدل ساخته‌شده بر روی مجموعه‌های داده‌ای بررسی می‌شود. مطابق بخش ۳-۱(مدل پژوهش) در گام اول، داده‌ها پاک‌سازی و نرمال‌سازی می‌شوند. سپس عوامل تاثیرگذار بر وضعیت اعتباری مشتریان یک بار با معیار همبستگی و بار دیگر با روش درخت تصمیم شناسایی می‌شود. با روش همبستگی، ارتباط خطی هر ویژگی با وضعیت اعتباری مشتریان بررسی می‌شود. همچنین ارتباط بین ویژگی‌ها نیز مشخص می‌گردد. ماتریس همبستگی یک ماتریس متقارن است و مقادیر آن بین -1 تا 1 می‌باشد. در شکل ۱، ماتریس همبستگی متغیرهای مختلف داده‌های بانک آلمانی را مشاهده می‌نمایید. هر یک از سطر یا ستون‌های شکل ۱، نشان‌دهنده ویژگی‌های معرفی شده در جدول ۲ است.

در شکل ۱ همبستگی بین ویژگی‌های مختلف بانک آلمانی با یکدیگر نمایش داده شده است. عنوان هر سطر یا ستون بر اساس نماد به‌کار رفته در جدول ۲ می‌باشد. سطر یا ستون‌های g_1 تا g_{20} ویژگی‌های ورودی، و سطر یا ستون g_{21} ویژگی وضعیت اعتباری یا خروجی را نمایش می‌دهد. از شکل ۱، مشخص است که ویژگی g_1 یعنی وضعیت حساب با مقدار ۳۵ درصد، ویژگی g_2 یعنی مدت اعتبار با مقدار ۲۱- درصد و ویژگی g_3 سابقهٔ اعتبار با مقدار ۲۳- درصد بیشترین تاثیر را بر ویژگی رده (برچسب) یعنی وضعیت اعتباری مشتریان دارند. ویژگی‌های سن (g_{13})، مسکن (g_{15})، وضعیت تاهل (g_9)، تعداد اقساط (g_8) و هدف تسهیلات (g_4) به ترتیب کمترین همبستگی با وضعیت اعتباری مشتریان بانک آلمانی دارند. همچنین ویژگی‌های g_2 و g_4 (مدت اعتبار و هدف

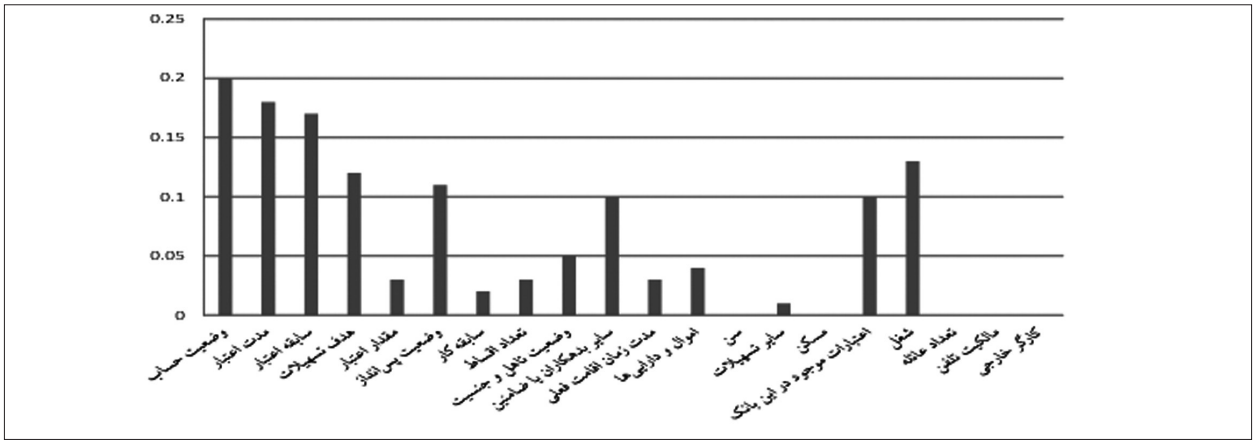


شکل ۱: همبستگی بین ویژگی‌های داده‌های آلمانی

تســـهیلات) با مقدار ٦٤ درصد بیشـــترین تاثیر را بر روی یکدیگر دارند که نشـــان‌دهنده همسبتگی بالایی است و در صورت نیاز می‌توان یکی را به دلخواه حذف کرد.

الگوریتم درخت تصمیم بـــه کمک معیار جینی، ارتباط بین هر ویژگی با ویژگی رده را نشان می‌دهد. معیار جینی یک معیار بر پایه آنتروپی اســـت که برای شـــاخه‌زدن بر روی ویژگی‌ها استفاده می‌شود. هر چقدر مقدار این معیار کمتر باشـــد ویژگی از اهمیت بیشـــتری برخوردار است و بر اساس نســـبت عکس آن، هر ویژگی یک امتیاز در بازه ۰ تا ۱ کســـب می‌کند. شـــکل ۲، اهمیت ویژگی‌ها بر مبنای تاثیرگذاری بر ویژگی رده یعنی وضعیت اعتباری مشتریان را در مجموعهٔ داده‌ای بانک آلمانی نشـــان می‌دهد. محور افقـــی نام ویژگی‌ها و محور عمـــودی اهمیت هر ویژگی را نشان می‌دهد. با توجه به شکل ۲، ویژگی‌های سن، مسکن، تعداد عائله، مالکیـــت تلفن و کارگر خارجی تاثیری نزدیک به صفر بر وضعیت اعتباری مشـــتریان بانک آلمانی دارند و می‌توان آنها را حذف کرد. همچنین تحلیل همبستگی این ویژگی‌هـــای با ویژگی رده نیز ایـــن مطلب را تایید می‌کند. در این پژوهش، ویژگی‌هایی که بر اساس الگوریتم درخت تصمیم اهمیتی کمتر از ۰/۰۰۱دارند حذف شدند.

در شـــکل ۳، همبستگی بین ویژگی‌های بانک پاسارگاد با یکدیگر نمایش داده شده است. عنوان هر سطر یا ستون بر اســـاس نماد به‌کار رفته در جدول ۳ می‌باشـــد. سطر یا ستون‌های p_1 تا p_{13} ویژگی‌های ورودی و سطر یا ستون p_{14} ویژگی وضعیت اعتباری یا خروجی را نمایش می‌دهد. از شکل ۳، مشخص است که ویژگی p_6 (نحوه بازپرداخت) بـــا مقـــدار ۸۵ درصد و ســـتون p_{12} (امـــوال و دارایی) با مقدار ۹۱- درصد بیشترین همبستگی با ویژگی رده یعنی



شکل ۲: اهمیت ویژگی‌های داده‌های آلمانی. منبع: محاسبات محقق

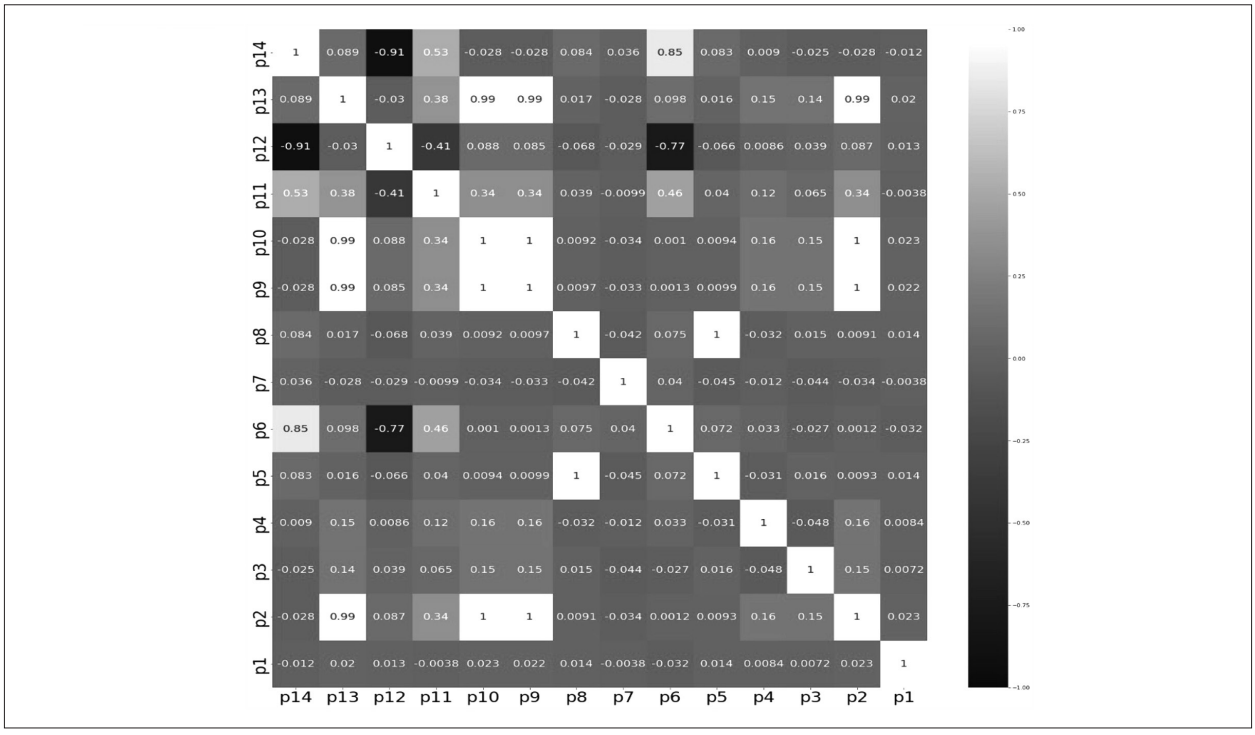
وضعیت اعتباری مشتریان را دارند. همچنین ویژگی p_2 (مقدار تسهیلات) با ویژگی‌های p_9 (سابقه کار)، p_{10} (چک برگشتی) و p_{13} (متوسط سپرده) همبستگی بالا و نزدیک به ۱۰۰ درصد دارد. این مطلب بدین معناست که می‌توان با داشتن ویژگی p_2 به طور دقیق مقادیر ویژگی p_9، p_{10} و p_{13} را تعیین کرد بنابراین می‌توان سه ویژگی اخیر را از مجموعهٔ داده‌ای حذف کرد.

شکل ٤، اهمیت ویژگی‌ها با استفاده از روش درخت تصمیم را در مجموعهٔ داده‌ای بانک پاسارگاد نشان می‌دهد. محور افقی نام ویژگی‌ها و محور عمودی اهمیت هر ویژگی را نشان می‌دهد. با توجه به شکل ٤، غیر از ویژگی‌های نحوهٔ بازپرداخت، سابقه چک برگشتی، حدود درآمد ماهیانه، اموال و دارایی فعلی و متوسط سپرده بقیه ویژگی‌ها اهمیتی نزدیک به صفر دارند. در این پژوهش در مورد داده‌های بانک پاسارگاد، ویژگی‌های کم اهمیت حذف شدند.

پس از انتخاب ویژگی‌های مهم، روش انتشار برچسب بر روی داده‌های جدید اعمال می‌شود. ابتدا برچسب ۲۰ درصد از داده‌های آموزشی حذف شدند. بنابراین در داده‌های آموزشی ۸۰ درصد برچسب‌دار و ۲۰ درصد بدون برچسب هستند. اکنون داده‌های آموزشی در اختیار الگوریتم انتشار برچسب قرار می‌گیرند و مدل ساخته می‌شود. پس از ساخت مدل نهایی، معیارهای ارزیابی مختلفی که در بخش قبل معرفی شدند محاسبه می‌شود. جدول ٤، گزارشی از معیارهای مختلف مذکور بر روی داده‌های بانک آلمانی و پاسارگاد ارایه می‌کند.

اطلاعات گزارش شده در جدول ٤، معیارهای ارزیابی حاصل از اعمال روش نیمه‌نظارتی پیشنهادی در این مقاله بر روی داده‌های بانک آلمانی و بانک پاسارگاد را نمایش می‌دهد. در مورد بانک پاسارگاد ابتدا روش پیشنهادی بر روی داده‌ها با پنج ویژگی مهم یعنی p_6 (نحوهٔ بازپرداخت)، p_{10} (سابقه چک برگشتی)، p_{11} (حدود درآمد ماهیانه)، p_{12} (اموال و دارایی فعلی) و p_{13} (متوسط سپرده) اعمال شد و نتایج بسیار خوبی به‌دست آمد که در جدول ٤ گزارش شده است. علاوه بر این، به‌منظور این‌که نشان داده شود پنج ویژگی مهم و منتخب با روش درخت تصمیم تاثیر بسزایی بر روی نتایج دارد روش پیشنهادی بر روی داده‌های بانک پاسارگاد بدون این پنج ویژگی مهم اعمال شد و نتایج آن در جدول ٤ ارایه شد. در ادامه به بررسی مفصل اطلاعات موجود در جدول ٤ پرداخته می‌شود.

بر اساس نتایج گزارش شده در جدول ٤ با روش انتشار برچسب، نرخ تشخیص نمونه‌های بدحساب یعنی معیار TNR در بین مشتریان بانک آلمانی ۷۱/۸٦ درصد است که نشان‌دهنده طبقه‌بندی قابل قبول مشتریان بدحساب می‌باشد. در مورد بانک پاسارگاد نیز مقدار ۹۹/۷۸ درصد برای این معیار به‌دست آمده است که بسیار مطلوب می‌باشد. همانگونه که قبلا اشاره شد این معیار اهمیت بسیار بالایی برای مدیران بانک خواهد داشت زیرا هزینه



شکل ۳: همبستگی بین ویژگی‌های بانک پاسارگاد

عدم تشخیص مشتریان بدحساب بسیار بیشتر از هزینه عدم تشخیص مشتریان خوش‌حساب می‌باشد. اعطای تسهیلات به مشتری بدحساب منجر به افزایش معوقات بانک خواهد شد. از این رو با توجه به مفهوم این معیار یا بر اساس معیار خطای نوع اول (EI)، ریسک اعتباری مدل حاصل از روش انتشار برچسب بر داده‌های بانک آلمانی ۲۸/۱۴ درصد و بر داده‌های بانک پاسارگاد ۰/۲۲ درصد خواهد بود. بنابراین روش پیشنهادی قادر خواهد بود به نحو مطلوبی مشتریان بدحساب بانک آلمانی و با ریسک بسیار پایینی مشتریان بدحساب بانک پاسارگاد را تشخیص دهد. ریسک اعتباری برای داده‌های بانک پاسارگاد بدون در نظر گرفتن پنج ویژگی مهم افزایش چشمگیری می‌یابد (۶۱/۱۱ درصد) و نشان می‌دهد ویژگی‌های مهم به خوبی تشخیص داده شده‌اند.

از سوی دیگر نرخ تشخیص نمونه‌های خوش‌حساب TPR مشتریان بانک آلمانی ۸۱/۸۶ درصد و برای بانک پاسارگاد ۹۷/۷۸ درصد به‌دست آمد که نشان‌دهنده طبقه‌بندی بسیار مطلوب مشتریان خوش‌حساب در هر دو بانک آلمانی و پاسارگاد می‌باشد. همان‌طور که پیش از این بیان شد ریسک تجاری به کمک این معیار یا بر اساس معیار خطای نوع دوم (EII) قابل محاسبه است. ریسک تجاری روش پیشنهادی بر روی داده‌های بانک آلمانی ۱۸/۱۴ درصد و برای بانک پاسارگاد ۲/۲۲ درصد است که در مورد بانک پاسارگاد بسیار مطلوب است.

معیار ارزیابی MCC که به معیار ارزیابی بهره‌وری مدل نیز شناخته می‌شود هر اندازه که به صد نزدیک‌تر باشد نشان‌دهندهٔ توان پیش‌بینی قابل قبول‌تر است. مقادیر درصدی این معیار با توجه به فرمول (۱۱) بین ۱۰۰- تا ۱۰۰+ است. بر این اساس میزان بهره‌وری مدل پیشنهادی بر روی داده‌های بانک آلمانی ۵۳/۹۸ و برای بانک پاسارگاد ۹۷/۵۶ است. بنابراین مدل پیشنهادی بر روی داده‌های بانک پاسارگاد توانایی بیشتری نسبت به داده‌های بانک آلمانی دارد. همچنین با حذف ویژگی‌های مهم در داده‌های بانک پاسارگاد، افت شدید این معیار به ۲۱/۲۵ نتیجه شده است. در مورد کارایی مدل‌ها از معیار AUC استفاده می‌شود. همان‌گونه که قبلا بیان شد

شکل ۴: اهمیت ویژگی‌های داده‌های آلمانی. منبع: محاسبات محقق

این معیار مساحت زیر منحنی کارایی مدل است و مقدار درصدی آن بین ۰ تا ۱۰۰ درصد خواهد بود هرچه این مقدار به ۱۰۰ درصد نزدیک‌تر باشد مدل کارایی بیشتری خواهد داشت. بر اساس جدول ٤، کارایی مدل نیمه‌نظارتی پیشنهادی بر روی داده‌های آلمانی ۸۱/۵۷درصد و بر روی داده‌های بانک پاسارگاد ۹۹/۹۵ درصد می‌باشد که نشان از کارایی مطلوب روش پیشنهادی است. همچنین با حذف ویژگی‌های مهم از مجموعهٔ داده‌ای بانک پاسارگاد، کارایی مدل به ٦۳/۳۲ کاهش می‌یابد.

توانایی مدل نیمه‌نظارتی پیشنهادی در دقت دسته‌بندی کلی، یعنی Acc، برای مشتریان بانک آلمانی ۷٦/۸۵ درصد، برای بانک پاسارگاد ۹۸/۷۷ درصد و برای داده‌های بانک پاسارگاد بدون ویژگی‌های منتخب ۵۹/٦٦ درصد می‌باشد. این معیار که نرخ تشخیص درست همهٔ برچسب‌ها را می‌سنجد به‌عنوان مبنای مقایسهٔ روش پیشنهادی با سایر روش‌های موجود در پژوهش‌های دیگر مورد توجه قرار گرفت. هانگ و همکارانش در سال ۲۰۰۷ با استفاده از روش نظارتی ماشین بردار پشتیبان بر روی داده‌های بانک آلمانی دقت کلی ۷۷.۹۲ درصد را گزارش کرد]۱۲[. در سال ۱۳۹۲ پویان‌فر و همکارانش با انتخاب ویژگی به کمک الگوریتم ژنتیک و ساخت مدل نظارتی مبتنی بر ماشین بردار پشتیبان، دقت کلی نزدیک به ۸٤ درصد دست یافتند[٤]. لازم به یادآوری است که در روش‌های نظارتی، تمامی داده‌های آموزشی باید برچسب‌دار باشند در حالی‌که در روش‌های نیمه‌نظارتی بخشی از داده‌های آموزشی برچسب‌دار و بقیه بدون برچسب هستند. با مقایسه کلی نتایج جدول ٤ و نتایج پژوهش هانگ، می‌توان گفت که روش‌های نیمه‌نظارتی از عملکرد مناسبی برخوردار هستند. گتاشیا در سال ۲۰۱٤ به کمک روش نظارتی جنگل تصادفی برای رتبه‌بندی مشتریان بانک آلمانی، دقت کلی ۷۷/۲ درصد را ارایه کرد در حالی که روش نیمه‌نظارتی پیشنهادی در این مقاله، دقت کلی ۷۸ درصد را به‌دست آورده است. نتایج حاصل شده در مقایسه با جدول ٤ نشان از عملکرد مناسب روش‌های نیمه‌نظارتی در مقایسه با جنگل تصادفی دارد[۱۷]. لیویریس و همکارانش در سال ۲۰۱۸ با روش‌های نظارت‌شده و نیمه‌نظارت‌شده برای ارزیابی ریسک اعتباری داده‌های بانک آلمانی، دقت کلی ۷۵/٤ درصد را گزارش کرد [۱٤] در حالی که روش نیمه‌نظارتی پیشنهادی در این مقاله، ۲,٦ درصد بهبود در دقت کلی مدل را نشان می‌دهد.

در میان پژوهش‌های انجام شده بر روی مسئلهٔ رتبه‌بندی اعتباری مشتریان بانک، تحقیقاتی متنوعی وجود دارند که بر روی داده‌های بانک آلمانی نتایجی گزارش نکرده‌اند و به‌طور موردی داده‌های یک بانک مشخصی را تحلیل کرده‌اند که در ادامه به برخی از آن‌ها اشاره می‌شود. همچنان از معیار دقت کلی روش‌های ارایه شده برای مقایسه با نتایج حاصل از روش نیمه‌نظارتی پیشنهادی بر روی بانک پاسارگاد استفاده خواهد شد. اطلاعات خلاصه‌ای از آن‌ها در جدول ۱ ارایه شده است. محمدیان و همکارانش در سال ۱۳۹۵ با استفاده از روش

ماشین بردار پشتیبان برای ریسک اعتباری مشتریان بانک تجـارت، دقت کلی ٦٩ درصد را گزارش کرد [٦]. طلوعی و همکارانش در سـال ١٣٨٩، چندین روش از جمله ماشین بردار پشتیبان، رگرسیون لجسـتیک، شبکه بیز و درخت تصمیم بـه رتبه‌بندی مشـتریان بانکی بـه‌کار بردند که به‌عنوان نمونه روش ماشین بردار پشتیبان مبتنی بر شبکه و امتیاز F، دقت ٧٦/٧ را به‌دست آوردند[١٠]. کشاورزحداد و همکارانش در سـال ١٣٨٦ بین مدل‌های لاجیت و درخت تصمیم برای اعتبارسـنجی مشتریان بانکی مسکن درخت تصمیـم با دقت بالاتر ٨٥/٥ درصـد را به‌عنوان مدل بهتر نسـبت به مدل لاجیت معرفی کرد [٣]. اقبالی و همکارانش در سـال ١٣٩٦ با اسـتفاده از توابع شایسـتگی الگوریتم ژنتیک در رتبه‌بندی مشـتریان حقیقی یک بانک داخلی به دقت کلی ٤٤ درصد دسـت یافتند [٢]. مهرآرا و همکارانش در سـال ١٣٨٨ با استفاده از شبکه‌های عصبی مصنوعی، رتبه‌بندی مشتریان حقوقی بانک پارسیان را بررسی کرد و دقت کلی نزدیک به ٨٦ درصد را گزارش کرد [٥]. همانطور که از جدول ٤ نتیجه می‌شود روش نیمه‌نظارتی پیشنهادی در این پژوهش بر روی داده‌های بانک پاسارگاد به مراتب دقت کلی بهتری نسبت به روش‌های مذکور دارد و حتی در صورتی کـه ویژگی‌های مهم از مجموعهٔ داده‌ای این بانک داخلی حذف شـود همچنان نسبت به روش‌های ارایه شده در مراجع [٦ ,٢] عملکرد قابل توجهی دارد.

۵- نتیجه‌گیری و ارائهٔ پیشنهادات

تخصیص امتیاز اعتباری دقیق به مشـتریان بانک‌ها، همواره یک مسئلهٔ بسـیار مهم برای محققان و بانک‌داران اسـت، زیرا تنها یک درصد افزایش دقـت می‌تواند جلوی خسارت‌هایی عظیم به بانک‌ها و موسسات مالی را بگیرد. با توسعهٔ مداوم، گسـترده و پویای صنعت اعتبارسنجی، هر روز این صنعت نقش مهم‌تـری در اقتصاد ایفا می‌کند و اعتباردهندگان به‌منظور توسعه فرآیند مدیریت اعتباری از روش‌هـا، ابزارهـا و ایده‌های جدید اسـتفاده می‌کنند.

سـاخت مدل‌های رتبه‌بنـدی اعتباری از یـک پایگاه داده اعتبـاری را می‌توان به‌عنوان یـک فرآیند داده‌کاوی انجام داد. در ایـن پژوهش بـا اسـتفاده از روش نوینی مبتنی بـر یادگیری نیمه‌نظارتی، مشـتریان بانک آلمانی و بانک پاسـارگاد رتبه‌بندی شـدند و مدلی کارا جهـت ارزیابی وضعیت اعتباری مشـتریان ارائه شـد. نمایش عملکرد و کارایی مناسـب و بسیار نزدیک دسته‌بندهای نیمه‌نظارتی به روش‌های نظارتی از نتایج اصلی این پژوهش می‌باشد. از آنجـا که اعمال روش نیمه‌نظارتی پیشـنهادی بر روی داده‌های بانک آلمانی در مقایسـه بـا تحقیقات موجود بر اسـاس ٩ معیار ارزیابی متنوع، از عملکـرد خوب و قابل دفاعی برخوردار اسـت و همچنین بر روی داده‌های بانک داخلی (پاسـارگاد) نتایج بسـیار مطلوب و موثر به‌دست آورد لـذا توصیه می‌شـود مدیران بانک‌ها و موسسـات مالی یا اعتبارسـنجی به موضـوع روش‌های نیمه‌نظارتی توجهـی خاص بنمایند و از آن برای حل مسـایلی از قبیل رتبه‌بندی اعتباری مشـتریان خود بهره ببرند. همچنین در اینجا از درخت تصمیم برای انتخاب ویژگی استفاده شد که پیشنهاد می‌شود محققان از روش‌های دیگر انتخاب ویژگی مثل فیلترها، بسـته‌بندها، روش‌های تـوکار و فراابتکاری اسـتفاده کنند. این کار سـبب افزایش سـرعت الگوریتم می‌شود و همچنین تاثیر بسزایی در شناسایی و درک بهتر عوامل تاثیرگذار بر وضعیت اعتباری خواهد داشت. از آن جایی که داده‌های بدون برچسب ارزان‌تر هستند و همیشه در دسترس می‌باشند و به راحتی به دست می‌آیند پیشنهاد می‌شـود به کمک الگوریتم‌های دیگر نیمه‌نظارتی و ترکیب آن‌هـا، دسـته‌بندهای قوی‌تری به وجـود آورد. همچنین می‌توان از ترکیب دسته‌بندهای نظارت‌شده و نیمه‌نظارتی مدل‌های جدیدی سـاخت و به نتایج مناسب، تخمین بهتر و تعمیم‌پذیـری بالاتری رسـید. به‌کارگیـری روش‌های گروهی مانند ارتقا و یا رای‌گیری بین دسته‌بندها یا ترکیب دسته‌بندها نیز برای تحقیقات آتی پیشنهاد می‌شود.

مراجع

۱. عباس طلوعی‌اشلقی، فرناز مقدوری‌شـربیانی، فرید دانشگر، «امتیازدهی اعتباری متقاضیان کارت‌های اعتباری بانک‌ها با استفاده از تکنیک ماشین بردار پشتیبان»، کنفرانس شهر الکترونیکی، دوره ۲، ۱۳۸۸.

۲. علی اقبالی، سیدحسـین رضوی‌حاج‌آقا، حنان عموزاد مهدیرجی، «ارزیابی مقایسـه‌ای عملکرد توابـع شایسـتگی الگوریتم ژنتیـک در رتبه‌بندی مشـتریان»، مدیریت صنعتی (دانش مدیریت)، دوره۹، شماره ۲، صفحه ۲۶۴-۲۴۵، ۱۳۹۶.

۳. غلامرضا کشاورزحداد، حسـین آیتی‌گازار، «مقایسه کارکرد مدل لاجیت و روش درخت‌هـای طبقه‌بندی و رگرسـیونی در فرآیند اعتبارسـنجی متقاضیان حقیقی برای اسـتفاده از تسهیلات بانکی»، پژوهش‌های رشد و توسعه اقتصادی، دوره ۷، شماره ۴، صفحه ۹۷-۷۱، ۱۳۸۶.

۴. احمد پویان‌فر، سعید فلاح‌پور، محمدرضا عزیزی، «رویکرد حداقل مربعات ماشین بردار پشتیبان مبتنی بر الگوریتم ژنتیک جهت تخمین رتبه اعتباری مشتریان بانک‌ها»، مجله مهندسـی مالی و مدیریت اوراق بهادار، دوره ۴، شماره ۱۷، صفحه ۱۵۸-۱۳۳، ۱۳۹۲.

۵. محسن مهرآرا، میثم موسایی، مهسا تصوری، آیت حسن‌زاده، «رتبه‌بندی اعتباری مشتریان حقوقی بانک پارسـیان»، فصلنامه مدل‌سازی اقتصادی، دوره ۳، شماره ۹، صفحه ۱۵۰-۱۲۱، ۱۳۸۸.

۶. امین محمدیان‌حاجی‌کرد، ملیحه اصغرزاده‌زعفرانی، مصطفی امام‌دوست، «بررسی ریسک اعتباری مشتریان حقوقی با استفاده از مدل ماشین بردار پشتیبان و مدل هیبریدی الگوریتم ژنتیک- مطالعه موردی بانک تجارت»، مهندسی مالی و اوراق بهادار، دوره ۷، شماره ۲۷، صفحه ۱۷-۳۲، ۱۳۹۵.

۷. سیدمجید شـریعت‌پناهی، سیما هاشمی‌برکادهی، «ارائه مدلی برای اعتبار سنجی مشـتریان در بانک صنعت و معدن»، مطالعات تجربی حسابداری مالی، دوره ۶، شماره ۲۱، صفحه ۸۲-۶۱، ۱۳۸۷.

۸. طاهـره زارع‌بیدکی، محمدتقـی صادقی، حمیدرضا ابوطالبـی، «یادگیری نیمه‌نظارتی کرنل مرکب با اسـتفاده از روش‌های یادگیری معیار فاصله»، پردازش علایم و داده‌ها، دوره ۱۴، شماره ۱، صفحه ۷۱-۵۳، ۱۳۹۶.

۹. مجتبی کردآبادی، محرم منصوری‌زاده، حسـن ختن‌لو، «روش ترکیبی و نیمه‌نظارتی مبتنی بر گراف برای برچسـب زنی خـودکار تصاویر»، مجله ماشـین بینایی و پردازش تصاویـر، دوره ۶، شـماره ۲، صفحه ۸۸-۷۹، ۱۳۹۸.

۱۰. عباس طلوعی‌اشلقی، هاشم نیکومرام، فرناز مقدوری‌شربیانی، «طبقه‌بندی متقاضیان تسـهیلات اعتباری بانک‌ها با اسـتفاده از تکنیک ماشین بردار پشتیبان»، مجله پژوهش‌های مدیریت، شماره ۸۴، ۱۳۸۹.

11. A. Krichene, “Using a naive Bayesian classifier methodology for loan risk assessment evidence from a Tunisian commercial bank”, Journal of Economics, Finance and Administrative Science, Vol. 22, No. 42, pp. 3-24, 2017.
12. C. Huang, M. Chen and C. Wang, “Credit scoring with a data mining approach based on support vector machines”, Expert Systems with Applications, Vol. 33, pp.847-856, 2007.
13. S. Han, “Semi-supervised learning classification based on generalized additive logistic regression for corporate credit anomaly detection”, IEEE Access, Vol 8, pp. 199060-199069, 2020.
14. I.E. Livieris, N. Kiriakidou, A. Kanavos, V. Tampakas and P. Pintelas, “On ensemble SSL algorithms for credit scoring problem”, Informatics Vol. 5, No. 4, 40, pp.1-16, 2018.
15. A. Kim and S. Cho, “An ensemble semi-supervised learning method for predicting defaults in social lending”. Engineering Applications of Artificial Intelligence”, Vol. 81, pp. 193–199, 2019.
16. B. Feng and X. Wenfang, “Adversarial Semi-supervised Learning for Corporate Credit Ratings”, Journal of Software, Vol. 16, No. 6, pp. 259-266, 2021.
17. N. Ghatasheh, “Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study”, International Journal of Advanced Science and Technology, Vol. 72, pp. 19-30, 2014.
18. J. Han, M. Kamber and J. Pei, “Data Mining Concepts and Techniques”, The Morgan Kaufmann Series in Data Management Systems, New York, 2012.
19. V. Papastefanopoulos, S. Karlos and S. Kotsiantis, “Using Semi-Supervised Learning Methods for Credit Score Problem”, Advances in Smart Systems Research, Vol. 6, No. 2, pp. 28-40, 2017.
20. O. Chapelle, B. Schölkopf and A. Zien, “Semi-supervised learning”, The MIT Press, London, 2006.
21. J. B. Kruskal, “On the shortest spanning subtree of a graph and the traveling salesman problem”, In Proceedings of the American Mathematical Society, Vol. 7, pp. 48-50, 1956.
22. Sabzevari H., Soleymani M., Noorbakhsh E.(2006), (n.d.) A Comparison between Statistical and Data Mining Methods for Credit Scoring in Case of Limited Available Data.