

در حوزه امتیازدهی اعتباری، همواره انتخاب الگوریتم‌های مناسب برای بهبود دقت پیش‌بینی و کاهش ریسک‌های مالی، یک چالش اساسی بوده است. پیش از این، مدل‌های طبقه‌بندی پایه به‌تنهایی قادر به ارائه دقت مطلوب در پیش‌بینی امتیاز اعتباری نبودند و نیاز به روشی مؤثرتر احساس می‌شد. در این راستا، پژوهشی با استفاده از مجموعه داده‌های اعتباری استرالیا، پس از پیش‌پردازش داده‌ها و جایگزینی مقادیر گمشده، اقدام به کاهش ابعاد داده‌ها با تحلیل مؤلفه‌های اصلی نمود. سپس، الگوریتم‌های طبقه‌بندی پایه مانند رگرسیون لجستیک، درخت تصمیم، ماشین بردار پشتیبان، نزدیک‌ترین همسایه، شبکه عصبی چندلایه و بیز ساده به‌کار گرفته شدند. علاوه بر آن، الگوریتم‌های گروهی نظیر جنگل تصادفی، درخت تصمیم بسته‌بندی‌شده، طبقه‌بند درخت‌های افزوده، روش تقویت و رأی‌گیری گروهی مورد ارزیابی قرار گرفتند. نتایج نشان داد که الگوریتم‌های گروهی، به‌ویژه جنگل تصادفی و طبقه‌بند درخت‌های افزوده، با دقت بالاتر و خطای کمتر، عملکرد بهتری در پیش‌بینی امتیاز اعتباری دارند. نوآوری این پژوهش در ارائه مقایسه جامع بین الگوریتم‌های پایه و گروهی و شناسایی مدل‌های بهینه برای بهبود دقت پیش‌بینی است که می‌تواند به‌طور مؤثری در کاهش ریسک‌های مالی مؤسسات اعتباری نقش داشته باشد (Parvin and Saleena 2020).

در این مرور ادبیات، مطالعات پیشین بر روی ارزیابی اعتبار مشتریان نشان می‌دهد که مدل‌های رایج همچون رگرسیون لجستیک و تحلیل تمایز خطی به دلیل فرضیات غیرواقعی مانند تفکیک خطی و توزیع نرمال داده‌ها عملکرد ضعیفی داشتند. به کارگیری روش‌هایی مانند شبکه عصبی و ماشین بردار پشتیبان در تلاش بود تا دقت مدل‌های ارزیابی اعتبار را بهبود بخشد، اما این مدل‌های تک‌طبقه‌بند در شناسایی ویژگی‌های خاص مشتریان و کاهش تأثیر نمونه‌های نادر به مشکل برخوردند. برای رفع این محدودیت‌ها، برخی تحقیقات از روش‌های ترکیبی شامل طبقه‌بندهای مجموعه‌ای با رأی‌گیری وزن‌دار استفاده کردند، اما این روش‌ها همچنان در مدیریت داده‌های نامتوازن با مشکل مواجه بودند. در این پژوهش، از نقشه‌های خودسازمان‌ده برای خوشه‌بندی چندلایه داده‌ها بهره گرفته شده و سپس از یک ترکیب طبقه‌بندهای k -نزدیک‌ترین همسایگی و ماشین بردار پشتیبان با رأی‌گیری نرم استفاده شده است. به منظور افزایش دقت و کاهش تأثیر داده‌های نامتوازن، بهینه‌سازی وزن‌ها با الگوریتم sailfish انجام شده که الگوریتمی متاهوریستیک با سرعت همگرایی بالا و قابلیت جستجوی فراگیر است. این ترکیب از خوشه‌بندی خودسازمان‌ده و بهینه‌سازی sailfish نه تنها به طبقه‌بندی دقیق‌تر و شناسایی مؤثر نمونه‌های نادر کمک می‌کند، بلکه پردازش کارآمدتری نیز در مقایسه با روش‌های پیشین دارد (Singh, Kothari et al. 2024).

در تحقیقات گذشته، مسئله مهمی که در مدل‌سازی ریسک اعتباری با روش خوشه‌بندی وجود داشت، انتخاب متغیرهای مناسب برای ایجاد خوشه‌های دقیق بود؛ زیرا انتخاب نادرست متغیرها می‌توانست منجر به خوشه‌های غیرقابل تشخیص و نامؤثر در پیش‌بینی شود. برخی از پژوهش‌ها با استفاده از تحلیل مؤلفه‌های اصلی به کاهش بعد متغیرها پرداختند اما این روش پیچیدگی خوشه‌بندی را افزایش می‌داد و قابلیت تفسیر مدل را کاهش می‌داد. روش‌های دیگر نظیر حذف تدریجی متغیرها با استفاده از رگرسیون خطی نیز به کار گرفته شده‌اند، اما این روش‌ها به دلیل انتخاب‌های تصادفی، دقت خوشه‌بندی و کیفیت پیش‌بینی را تحت تأثیر قرار می‌دادند. مقاله حاضر با معرفی یک روش مبتنی بر یادگیری ماشین، از مقادیر شاپلی برای سنجش اهمیت متغیرها استفاده می‌کند و با انتخاب متغیرهای اصلی در مرحله خوشه‌بندی، سعی دارد دقت پیش‌بینی مدل را افزایش دهد. این رویکرد، از الگوریتم‌های جنگل تصادفی و درخت‌های تقویت‌شده بهره برده و با استفاده از الگوریتم‌های خوشه‌بندی بهینه نظیر k -prototypes و تکنیک‌های اعتبارسنجی، خوشه‌بندی را طوری انجام می‌دهد که نتایج هم از نظر دقت پیش‌بینی و هم از نظر تفسیرپذیری مورد تایید قرار گیرد (Zöllner and Gürtler 2023).

مطالعات پیشین در زمینه ارزیابی ریسک اعتباری عمدتاً بر استفاده از روش‌های خوشه‌بندی و طبقه‌بندی داده‌ها تمرکز داشته‌اند. روش‌های رایج مانند K-Means و سلسله‌مراتبی، علی‌رغم سادگی و تفسیرپذیری، در داده‌های چندبعدی و غیرخطی به محدودیت‌هایی از جمله کاهش دقت در خوشه‌بندی و ناکارآمدی در جداسازی داده‌های نویزی مواجه شده. در این مقاله، یک رویکرد ترکیبی ارائه شده است که خوشه‌بندی K-Means و سلسله‌مراتبی را با استفاده از مختصات فاکتوریال استخراج‌شده از تحلیل مکاتبات چندگانه تلفیق می‌کند. این روش، ابتدا با پیش‌پردازش داده‌ها شامل نرمال‌سازی و کاهش نویز داده‌ها، ساختار داده‌ها را برای تحلیل آماده می‌سازد. سپس، تحلیل مکاتبات چندگانه به منظور کاهش ابعاد و استخراج ویژگی‌های اساسی داده‌ها به کار گرفته می‌شود. در ادامه، خوشه‌بندی K-Means و سلسله‌مراتبی به صورت ترکیبی اعمال شده و کیفیت خوشه‌بندی بر اساس تجزیه اینرسی ارزیابی می‌شود. این روش امکان مدیریت دقیق‌تر داده‌ها، کاهش نویز و بهبود دقت مدل‌های یادگیری ماشین در تحلیل ریسک اعتباری را فراهم کرده است (Jadwal, Pathak et al. 2022).

در زمینه ارزیابی ریسک اعتباری، انتخاب ویژگی به عنوان یکی از چالش‌های اصلی در بهبود دقت پیش‌بینی و کاهش پیچیدگی مدل‌ها شناخته شده است. پیش از این، روش‌های غربالگری مانند رتبه‌بندی همبستگی یا اطلاعات متقابل به دلیل سوگیری در انتخاب ویژگی‌ها و کاهش دقت پیش‌بینی، محدودیت‌هایی داشتند. همچنین، روش‌های ترکیبی دومارحله‌ای مانند استفاده از فیلترهای اولیه و سپس الگوریتم‌های جستجوی تکاملی، اگرچه در برخی موارد به بهبود دقت کمک می‌کردند، اما باعث افزایش پیچیدگی محاسبات و کاهش پایداری مدل‌ها می‌شدند. در این مقاله، رویکردهای یک‌مرحله‌ای مانند حذف بازگشتی ویژگی (RFE) با مدل‌های جنگل تصادفی (RFE-RF) و ماشین بردار پشتیبان (RFE-SVM) بررسی شده‌اند. این روش‌ها ویژگی‌ها را به طور پویا در فرآیند یادگیری انتخاب کرده و امکان پایداری بیشتر انتخاب ویژگی را فراهم می‌کنند. همچنین، استفاده از رگرسیون لجستیک جریمه‌ای (Penalized Logistic Regression) به منظور کاهش ابعاد داده‌ها و کنترل همبستگی متغیرها بررسی شده است. روش RFE-RF در معیارهایی مانند دقت پیش‌بینی و پایداری انتخاب ویژگی عملکرد بهتری نسبت به روش‌های دیگر داشته و الگوریتم RFE-SVM با پایداری بالاتر، گزینه‌ای مناسب برای مجموعه داده‌های با ابعاد بزرگ است (Atif and Salmi 2022).

در زمینه ارزیابی ریسک اعتباری رویکردهای اولیه معمولاً از مدل‌های انفرادی استفاده می‌کردند که به دلیل نبود ترکیب مناسب الگوریتم‌ها و عدم استفاده از تکنیک‌های پیش‌پردازش، دقت پایین‌تری داشتند. برخی تحقیقات نیز از روش‌های ترکیبی مانند ادغام یادگیری نظارت‌شده و بدون نظارت یا استفاده از شبکه‌های عصبی مصنوعی بهره بردند، اما این روش‌ها به دلیل هزینه محاسباتی بالا یا عدم مدیریت مناسب ویژگی‌های غیرمؤثر، در محیط‌های واقعی کارایی مطلوبی نداشتند. استفاده از روش‌های انتخاب ویژگی بر اساس همبستگی یا مدل‌های خطی نیز در برخی مطالعات مطرح شده، اما این روش‌ها قادر به شناسایی روابط غیرخطی بین ویژگی‌ها نبودند. در این میان، مطالعاتی که از روش‌های ترکیبی چندلایه استفاده کردند، توانستند با ادغام خروجی مدل‌های مختلف، عملکرد بهتری ارائه دهند. این روش‌ها به‌ویژه در مدیریت عدم توازن داده‌ها و بهبود معیارهایی نظیر دقت و امتیاز F1 نقش مؤثری ایفا کردند. با این حال، چالش انتخاب ویژگی‌های بهینه و کاهش پیچیدگی مدل همچنان به قوت خود باقی بود. پژوهش حاضر با ارائه رویکردی چندلایه و ترکیب آن با انتخاب ویژگی مبتنی بر اطلاعات متقابل، توانست با بهینه‌سازی ویژگی‌های ورودی و کاهش پیچیدگی محاسباتی، نتایج قابل توجهی در پیش‌بینی ریسک اعتباری ارائه دهد (Emmanuel, Sun et al. 2024).

روش‌های سنتی ارزیابی ریسک اعتباری مانند رگرسیون لجستیک و ماشین بردار پشتیبان با وجود دقت قابل قبول، در مدیریت نمونه‌های پیچیده و اطلاعات ناقص ضعف دارند و معمولاً با تصمیم‌گیری دوکلاسه، ریسک خطا را افزایش می‌دهند. رویکردهای ترکیبی مانند XGBoost و استکینگ نیز با وجود بهبود نسبی، به دلیل استفاده از آستانه‌های ثابت و ناتوانی در مدیریت داده‌های با هزینه بالا، محدودیت‌هایی داشته‌اند. تصمیم‌گیری سه وجهی با افزودن تأخیر در تصمیم‌گیری، به کاهش خطا کمک کرده است، اما تنظیم آستانه‌ها در روش‌های پیشین مبتنی بر فرضیات کلی بوده و داده‌های مالی واقعی را نادیده گرفته‌اند. مقاله حاضر با ترکیب تصمیم‌گیری سه وجهی و استفاده از داده‌های جریان نقدی برای تنظیم آستانه‌ها و بهره‌گیری از مدل بهینه‌سازی چندهدفه، این محدودیت‌ها را برطرف کرده و با ترکیب الگوریتم LightGBM و استکینگ، امکان تصمیم‌گیری دقیق‌تر برای نمونه‌های مرزی و پرهزینه را فراهم ساخته و کارایی مدیریت ریسک اعتباری را افزایش داده است (Li, Zhao et al. 2024).

پیش‌بینی ریسک اعتباری عمده‌تأ بر روش‌های یادگیری ماشین متکی بوده‌اند که شامل تکنیک‌هایی مانند شبکه‌های عصبی، ماشین بردار پشتیبان و رگرسیون لجستیک می‌شوند. این مدل‌ها با وجود توانایی بالا در تحلیل داده‌ها، به دلیل اتخاذ تصمیم‌گیری دوگانه، دچار خطاهایی در دسته‌بندی داده‌های مرزی شده‌اند. برای حل این مشکل، رویکردهای سه‌گانه معرفی شده‌اند که با تقسیم داده‌ها به سه منطقه مثبت، منفی و مرزی، تصمیم‌گیری برای داده‌های نامطمئن را به تأخیر می‌اندازند. روش‌های مختلفی برای یادگیری آستانه در تصمیم‌گیری سه‌گانه استفاده شده است، از جمله استفاده از بهینه‌سازی چندهدفه و مدل‌های مبتنی بر کاهش هزینه‌های تصمیم‌گیری. با این حال، اکثر این رویکردها نیازمند تنظیم دستی پارامترها یا تعریف پیچیده ماتریس‌های هزینه بوده‌اند که قابلیت کاربرد آن‌ها را محدود کرده است. مقاله حاضر با بهره‌گیری از یک چارچوب دو مرحله‌ای و استفاده از الگوریتم بهینه‌سازی ازدحام ذرات برای یادگیری خودکار آستانه‌ها، رویکردی جامع‌تر ارائه کرده است. این روش با ترکیب بهینه اطلاعات و کاهش هزینه تصمیم‌گیری، ضمن افزایش دقت در دسته‌بندی داده‌های مرزی، امکان استفاده از داده‌های تکمیلی در مراحل بعدی را فراهم می‌کند و محدودیت‌های رویکردهای پیشین را برطرف می‌سازد. (Li and Sha 2024)

در ارزیابی ریسک اعتباری، روش‌های سنتی مانند رگرسیون لجستیک و تحلیل تمایز خطی به دلیل سادگی، محبوبیت بالایی داشتند، اما عدم توانایی در مدیریت داده‌های پیچیده و نامطمئن از معایب اصلی آنها بود. مدل‌های پیشرفته‌تر مانند شبکه‌های عصبی مصنوعی و درخت‌های تصمیم‌گیری تقویتی با قدرت تشخیص بالاتر، توانستند این محدودیت را کاهش دهند، اما بیشتر این مدل‌ها تنها از طبقه‌بندی دوتایی استفاده کرده و در مواجهه با اطلاعات ناکافی یا مبهم دچار ضعف می‌شوند. مدل سهرای تصمیم‌گیری با ارائه دسته مرزی برای موارد نامطمئن گامی در مدیریت بهتر این چالش‌ها برداشت، اما این روش نیز به دلیل وابستگی زیاد به تنظیمات ماتریس هزینه‌ها، انعطاف‌پذیری کمی داشت. مقاله حاضر با ترکیب مدل سهرای تصمیم‌گیری و روش‌های یادگیری ماشین مانند تقویت‌کننده گرادیان، و بهینه‌سازی با الگوریتم ژنتیک، یک چارچوب بهینه‌سازی ترتیبی ارائه داده است که بدون نیاز به تنظیمات دستی ماتریس هزینه، آستانه‌های تصمیم‌گیری را بهینه می‌کند و با افزودن تدریجی اطلاعات به موارد مرزی، عدم قطعیت را کاهش داده و عملکرد کلی سیستم را بهبود می‌بخشد (Shen, Zhang et al. 2022).

تصمیم‌گیری سه‌طرفه، بر اساس نظریه مجموعه‌های زیر احتمالی، به عنوان راهکاری مؤثر برای مدیریت عدم قطعیت معرفی شده است. این رویکرد، به جای تصمیم‌گیری قطعی در مدل‌های دوطرفه، با افزودن یک منطقه مرزی، امکان تصمیم‌گیری منعطف‌تر را فراهم می‌کند. پیش از این، مدل‌های دوطرفه مانند رگرسیون لجستیک یا درخت تصمیم برای ارزیابی اعتباری استفاده می‌شدند که مستلزم جمع‌آوری اطلاعات جامع از تمامی متقاضیان و افزایش هزینه‌های پردازش

بود. مقاله حاضر با استفاده از مجموعه‌های زیر احتمالی، یک فرآیند دو مرحله‌ای را پیشنهاد می‌کند. در مرحله اول، با اطلاعات اولیه مانند فرم‌های درخواست، مشتریان به سه دسته تأیید، رد و مرزی تقسیم می‌شوند. مرحله دوم، که مختص موارد مرزی است، شامل جمع‌آوری اطلاعات تکمیلی از طریق مصاحبه و تحلیل مالی است. تحلیل دقیق در این مرحله با استفاده از رگرسیون لجستیک انجام می‌شود که احتمال بازپرداخت را بر اساس داده‌های تکمیلی پیش‌بینی می‌کند. این فرآیند با بهینه‌سازی هزینه و احتمال، از جمع‌آوری اطلاعات غیرضروری جلوگیری کرده و دقت و کارایی تصمیم‌گیری را بهبود می‌بخشد. نوآوری مقاله در استفاده از مجموعه‌های زیر احتمالی برای کاهش هزینه‌های ارزیابی اعتباری و انطباق با الزامات شفافیت در مقررات مالی است (Maldonado, Peters et al. 2020).

- Atif, D. and M. Salmi (2022). "The Most Effective Strategy for Incorporating Feature Selection into Credit Risk Assessment." *SN Computer Science* **4**(2): 96
- Emmanuel, I., et al. (2024). "A machine learning-based credit risk prediction engine system a stacked classifier and a filter-based feature selection method." *Journal of Big Data* using **11**(1): 23
- Jadwal, P. K., et al. (2022). "Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis." *Microsystem Technologies* **28**(12): 2715-2721
- Li, Y. and M. Sha (2024). "Two-stage credit risk prediction framework based on three-way decisions with automatic threshold learning." *Journal of Forecasting*
- Three-Way Decisions Li, Y., et al. (2024). "A Hybrid Credit Risk Evaluation Model Based on and Stacking Ensemble Approach." *Computational Economics*: 1-24
- Maldonado, S., et al. (2020). "Credit scoring using three-way decisions with probabilistic rough sets." *Information Sciences* **507**: 700-714
- Parvin, A. S. and B. Saleena (2020). *An ensemble classifier model to predict credit scoring-comparative analysis*. 2020 IEEE international symposium on smart electronic systems (iSES)(Formerly iNiS), IEEE
- with information Shen, F., et al. (2022). "Sequential optimization three-way decision model gain for credit default risk evaluation." *International Journal of Forecasting* **38**(3): 1116-1128

Singh, I., et al. (2024). "A hybrid metaheuristic optimised ensemble classifier with self-organizing map clustering for credit scoring." Operational Research **24**(4): 1-42

Zöllner, M. and M. Gürtler (2023). "Machine Learning-Based Variable Selection for Clustered Credit Risk Modeling." Available at SSRN 4506537