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.(1)

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 رای‌گیری 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 رای‌گیری نرم استفاده شده است. به منظور افزایش دقت و کاهش تأثیر داده‌های نامتوازن، بهینه‌سازی وزن‌ها با الگوریتم sailfish انجام شده که الگوریتمی متاهیوریستیک 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.(2)

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 غیرقابل‌تشخیص و ناموثر 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.(3)

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 داده ها 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.(4)

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 دو‌مرحله‌ای 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.(5)

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. (6)

1. Parvin AS, Saleena B, editors. An ensemble classifier model to predict credit scoring-comparative analysis. 2020 IEEE international symposium on smart electronic systems (iSES)(Formerly iNiS); 2020: IEEE.

2. Singh I, Kothari D, Aditya S, Rajora M, Agarwal C, Gautam V. A hybrid metaheuristic optimised ensemble classifier with self organizing map clustering for credit scoring. Operational Research. 2024;24(4):1-42.

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

4. Jadwal PK, Pathak S, Jain S. Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis. Microsystem Technologies. 2022;28(12):2715-21.

5. Atif D, Salmi M. The Most Effective Strategy for Incorporating Feature Selection into Credit Risk Assessment. SN Computer Science. 2022;4(2):96.

6. Emmanuel I, Sun Y, Wang Z. A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. Journal of Big Data. 2024;11(1):23.