Journal of Managerial Modelling in Sustainable Development, Volume 1(1), Summer 2025, 235-254

Journal of Managerial Modelling in Sustainable Development

Journal Homepage: http://MMSD.Semnan.ac.ir

ISSN:

Semnan University

A Hybrid Framework of Clustering and Machine Learning for Evaluating Financial Sustainability and Transparency in the Capital Market

Omid Valizadeh•

PhD Candidate in Operations Research, Department of Management, Faculty of Administrative and Economic Sciences, Ferdowsi University of Mashhad, Mashhad, Iran.

omid.valizadeh@mail.um.ac.ir

Mahsa Akhavan Rad

PhD Candidate in Operations Research, Department of Management, Faculty of Administrative and Economic Sciences, Ferdowsi University of Mashhad, Mashhad, Iran.

mahsa.akhavanrad@mail.um.ac.ir

Atieh Jafarian

PhD Student, Université de Rennes, CNRS, CREM - UMR 6211, F-35000 Rennes, France.

atieh.jafarian@univ-rennes1.fr

ARTICLE INFO

Article type:
Research Full Paper

Article history:
Received: 2025-08-16
Revised: 2025-09-28
Accepted: 2025-10-03

Keywords:
Capital Market;
Financial Sustainability;
Artificial Intelligence;
Machine Learning.

EXTENDED ABSTRACT

Background and Objectives: The capital market, as a fundamental pillar of financing and economic transparency, plays a vital role in the sustainability of firms. Despite notable advances in financial analysis, most prior studies have predominantly focused on one-dimensional evaluations of financial indicators, paying limited attention to structural heterogeneity among firms and the identification of latent patterns. Moreover, research conducted in the Iranian stock market has largely relied on traditional statistical models, which exhibit limited capacity in modelling nonlinear and complex relationships. The objective of this study is to propose an integrated framework for identifying distinct financial clusters and predicting firms' sustainability by employing a hybrid approach that combines clustering techniques with machine learning.

Materials and Methods: Initially, a dataset comprising eight key financial sustainability indicators was collected. After preprocessing which involved outlier removal, imputing missing values using the KNN method, and standardization the data were fed into the K-Means algorithm. To determine the optimal number of clusters, the silhouette index was calculated, identifying three as the optimal solution. The resulting cluster labels were then employed as the target variable in three machine learning models (XGBoost, Random Forest, and Decision Tree), whose performance was evaluated using accuracy, precision, recall, and F1-Score metrics. Finally,

• Corresponding author.
E-mail: omid.valizadeh@mail.um.ac.ir
https://orcid.org/0009-0005-6246-8244

through feature importance analysis, the variables contributing most significantly to cluster differentiation were identified.

Results: The clustering results revealed three distinct structural patterns among firms: a cluster characterized by high profitability, strong economic value added, and low risk; a cluster with high growth and returns but accompanied by elevated risk and cost of capital; and a cluster marked by weak profitability and limited value creation. A comparison of the machine learning models indicated that the XGBoost algorithm achieved the best predictive performance, with an accuracy of 0.97 and an F1-Score of 0.96. Furthermore, the feature importance analysis demonstrated that ROA was the most influential indicator in differentiating clusters across all three models, followed by WACC and EVA as key determinants of financial sustainability.

Conclusion: The present study demonstrates that the hybrid approach of clustering and machine learning can serve as an effective tool for identifying structural heterogeneity in the capital market. The findings suggest that managerial focus on improving ROA and reducing WACC through optimal capital structure management, along with enhancing EVA, constitutes a key strategy for strengthening firms' financial sustainability.

Cite this article:
Valizadeh, O., Akhavan Rad, M. & Jafarian, A. (2025). A Hybrid Framework of Clustering and Machine Learning for Evaluating Financial Sustainability and Transparency in the Capital Market. *Managerial Modelling in Sustainable Development*, *1*(1), 235 -254.
DOI: https://doi.org/10.22075/mmsd.2025.38693.1015

مدل‌سازی مدیریتی در توسعه پایدار

Journal Homepage: http://MMSD.Semnan.ac.ir

ISSN:

دانشگاه سمنان

چارچوب ترکیبی خوشه‌بندی و یادگیری ماشین برای ارزیابی پایداری و شفافیت مالی در بازار سرمایه

امید ولیزاده*

کاندیدای دکتری تحقیق در عملیات، گروه مدیریت، دانشکده علوم اداری و اقتصادی، دانشگاه فردوسی مشهد، مشهد ایران.

omid.valizadeh@mail.um.ac.ir

مهسا اخوان راد

کاندیدای دکتری تحقیق در عملیات، گروه مدیریت، دانشکده علوم اداری و اقتصادی، دانشگاه فردوسی مشهد، مشهد ایران.

mahsa.akhavanrad@mail.um.ac.ir

عطیه جعفریان

دانشجوی دکتری، دانشگاه رِن، مرکز ملی پژوهش‌های علمی فرانسه (*CNRS*)، مرکز پژوهشی اقتصاد و مدیریت (*CREM - UMR 6211*)، رِن، فرانسه

atieh.jafarian@univ-rennes1.fr

اطلاعات مقاله

نوع مقاله:

مقاله کامل علمی- پژوهشی

تاریخ دریافت: ۱۴۰۴/۰۵/۲۵

تاریخ بازنگری: ۱۴۰۴/۰۷/۰۶

تاریخ پذیرش: ۱۴۰۴/۰۷/۱۱

واژه‌های کلیدی:

بازار سرمایه؛

پایداری مالی؛

هوش مصنوعی؛

یادگیری ماشین.

چکیده

سابقه و هدف: بازار سرمایه به‌عنوان رکن اصلی تأمین مالی و ارتقای شفافیت اقتصادی، نقشی اساسی در پایداری بنگاه‌ها ایفا می‌کند. باوجود پیشرفت‌های قابل‌توجه در تحلیل‌های مالی، اغلب مطالعات پیشین بر ارزیابی تک‌بعدی شاخص‌های مالی متمرکز بوده و کمتر به تفاوت ساختاری میان شرکت‌ها و شناسایی الگوهای پنهان توجه داشته‌اند. از سوی دیگر، پژوهش‌های انجام‌شده در بورس ایران عمدتاً مبتنی بر مدل‌های سنتی و آماری بوده‌اند که توان محدودی در مدل‌سازی روابط غیرخطی و پیچیده دارند. هدف این پژوهش ارائه چارچوبی یکپارچه برای شناسایی خوشه‌های متمایز مالی و پیش‌بینی وضعیت پایداری شرکت‌ها با استفاده از رویکرد ترکیبی خوشه‌بندی و یادگیری ماشین است.

روش: ابتدا مجموعه داده‌ای شامل هشت شاخص کلیدی پایداری مالی گردآوری شد. داده‌ها پس از پردازش شامل حذف مقادیر پرت، جایگزینی داده‌های گمشده با روش KNN و استانداردسازی، وارد الگوریتم K-Means شدند. برای تعیین تعداد خوشه‌ها، شاخص سیلوئت محاسبه و مقدار بهینه سه خوشه شناسایی شد. سپس برچسب‌های حاصل به‌عنوان متغیر هدف در سه مدل یادگیری ماشین (XGBoost، جنگل تصادفی و درخت تصمیم) استفاده شد و عملکرد

آن‌ها با معیارهای دقت، صحت، یادآوری و*F1-Score* ارزیابی گردید. درنهایت، با تحلیل اهمیت ویژگی‌ها، متغیرهای اثرگذار در تمایز خوشه‌ها شناسایی شدند.

یافته‌ها: نتایج خوشه‌بندی نشان داد که سه الگوی ساختاری متمایز از شرکت‌ها قابل‌شناسایی است: خوشه‌ای با سودآوری و ارزش‌افزوده اقتصادی بالا و ریسک پایین؛ خوشه‌ای با رشد و بازده بالا همراه با ریسک و هزینه سرمایه زیاد؛ و خوشه‌ای با سودآوری ضعیف و ارزش‌افزوده محدود. مقایسه مدل‌های یادگیری ماشین نشان داد که الگوریتم XGBoost با صحت ۰/۹۷ و F1-Scoreبرابر ۰/۹۶ بهترین عملکرد را در پیش‌بینی خوشه‌ها داشته است. تحلیل اهمیت ویژگی‌ها نیز حاکی از آن بود که ROA در هر سه مدل مهم‌ترین شاخص در تفکیک خوشه‌ها است و پس‌ازآن WACC و EVA به‌عنوان عوامل کلیدی در پایداری مالی شناخته شدند.

نتیجه‌گیری: پژوهش حاضر بیانگر آن است که رویکرد ترکیبی خوشه‌بندی و یادگیری ماشین می‌تواند ابزاری مؤثر برای شناسایی ناهمگنی ساختاری در بازار سرمایه باشد. نتایج نشان می‌دهد که تمرکز مدیران بر بهبود ROA و کاهش WACC از طریق مدیریت بهینه ساختار سرمایه، همراه با افزایش EVA، راهبردی کلیدی در ارتقای پایداری مالی شرکت‌هاست.

استناد: ولیزاده، امید، اخوان‌راد، مهسا و جعفریان، عطیه (۱٤۰٤). چارچوب ترکیبی خوشه‌بندی و یادگیری ماشین برای ارزیابی پایداری و شفافیت مالی در بازار سرمایه. *مدل‌سازی مدیریتی در توسعه پایدار*، ۱(۱)، ۲۳۵-۲۵٤.

DOI: https://doi.org/10.22075/mmsd.2025.38693.1015

ناشر: دانشگاه سمنان

۱. مقدمه

بازار بورس اوراق بهادار یکی از مهم‌ترین مسیرهای تأمین سرمایه برای بنگاه‌های اقتصادی به شمار می‌آید و نقش بسزایی در ارتقای شفافیت مالی، جذب سرمایه‌گذاران و افزایش کارایی نظام اقتصادی دارد. این بازار با ایجاد ارتباط مستقیم میان سرمایه‌گذاران و واحدهای تولیدی، امکان هدایت بهینه منابع مالی به سمت فعالیت‌های مولد را فراهم می‌کند و زمینه حمایت از کسب‌وکارها و توسعه بخش‌های مختلف اقتصادی را مهیا می‌سازد. در ایران، شرکت‌های پذیرفته‌شده در بورس به‌عنوان بازیگران اصلی بازار سرمایه، سهم مهمی در تأمین منابع مالی پروژه‌های بزرگ و تقویت زیرساخت‌های اقتصادی کشور ایفا می‌کنند. شکل‌گیری و گسترش این شرکت‌ها در طول سال‌های گذشته، همواره تحت تأثیر تحولات ساختاری در نظام اقتصادی و مالی کشور قرار داشته است. بورس اوراق بهادار تهران که فعالیت خود را از دهه ۱۳۴۰ آغاز کرده، در طی مسیر تکامل خود فراز و نشیب‌های متعددی را پشت سر گذاشته و اکنون به یکی از ارکان اصلی بازار سرمایه کشور تبدیل‌شده است (محمدی[1] و همکاران، ۲۰۲۱).

باوجود پیشرفت‌های قابل‌توجه در حوزه تحلیل‌های مالی و به‌کارگیری مدل‌های کمی، بخش عمده‌ای از پژوهش‌های پیشین همچنان بر ارزیابی تک‌بعدی شاخص‌های عملکرد مالی متمرکز بوده‌اند؛ به این معنا که هر شاخص به‌صورت منفرد و بدون توجه به تعاملات و روابط میان آن‌ها تحلیل‌شده است. چنین رویکردی، اگرچه می‌تواند تصویری کلی از وضعیت یک شرکت ارائه دهد، اما قادر به شناسایی الگوهای پیچیده و روابط پنهان میان متغیرها نیست. علاوه بر این، کمتر مطالعه‌ای به تفاوت ساختاری میان شرکت‌ها پرداخته است؛ یعنی اختلافات بنیادین در الگوهای مالی که می‌تواند ناشی از تفاوت در ساختار و فرآیندهای کسب‌وکار، ساختار سرمایه، یا شرایط بازار باشد. این مسئله به‌ویژه در بورس ایران که شرکت‌ها ازنظر اندازه، صنعت، چرخه عمر و سایر شرایط اقتصادی با تنوع زیادی مواجه‌اند، اهمیت بیشتری پیدا می‌کند (اوساره[2] و همکاران، ۲۰۲۱). در چنین شرایطی، اغلب تحقیقات انجام‌شده در بازار سرمایه ایران به استفاده از مدل‌های سنتی یا روش‌های آماری کلاسیک بسنده کرده‌اند؛ روش‌هایی که عموماً بر فرضیات خطی و توزیع‌های خاص داده‌ها متکی هستند و توانایی محدودی در مدل‌سازی روابط غیرخطی و پیچیده میان شاخص‌های مالی دارند. این در حالی است که مدل‌های پیشرفته یادگیری ماشین، باقابلیت پردازش حجم زیاد داده و استخراج خودکار ویژگی‌های مؤثر، امکان تحلیل هم‌زمان چندین شاخص و شناسایی الگوهای پنهان را فراهم می‌آورند (غلامزاده[3] و همکاران، ۲۰۲۱)

پایداری مالی شرکت‌ها مفهومی چندبعدی است که فراتر از سودآوری کوتاه‌مدت قرار می‌گیرد و بر توانایی سازمان در ایجاد ارزش اقتصادی پایدار، مدیریت ریسک و حفظ تعادل میان بازده و هزینه تأکید دارد. در این راستا، مجموعه‌ای از شاخص‌های مالی به‌عنوان ابعاد کلیدی پایداری مطرح می‌شوند (وانگ[4] و همکاران، ۲۰۲۰). شاخص ROA[5] و ROE[6] توانایی شرکت در استفاده بهینه از منابع برای ایجاد سود را منعکس می‌کنند. EVA[7] با محاسبه سود مازاد بر هزینه سرمایه، تصویری دقیق‌تر از خلق ارزش واقعی برای سهام‌داران ارائه می‌دهد. GROWTH و RETURN نیز نشان‌دهنده ظرفیت

[1] Mohammadi
[2] Ossareh
[3] Gholamzadeh
[4] Wang
[5] Return on Assets
[6] Return on Equity
[7] Economic Value Added

شرکت برای توسعه آینده و جلب اعتماد سرمایه‌گذاران هستند. در کنار این‌ها، LEV[1] و WACC[2] بیانگر ساختار تأمین مالی و پایداری آن در مواجهه با ریسک‌های اعتباری و هزینه‌های تأمین منابع‌اند. Volatility[3]به‌عنوان شاخص ریسک بازار، پایداری و ثبات جریان سودآوری را تحت تأثیر قرار می‌دهد (ملینا[4] و همکاران، ۲۰۲۳).

بررسی هم‌زمان این مجموعه متغیرها، امکان تحلیل جامع‌تری از پایداری مالی شرکت‌ها را ایجاد می‌کند؛ تحلیلی که نه‌تنها به کارایی عملیاتی و سودآوری توجه دارد، بلکه ابعاد ریسک، ساختار سرمایه و توانایی خلق ارزش پایدار را نیز در نظر می‌گیرد. بااین‌حال، پژوهش‌های گذشته در بازار سرمایه ایران عمدتاً هر یک از این شاخص‌ها را به‌صورت منفرد تحلیل کرده و یکپارچه‌سازی آن‌ها در قالب مدلی جامع صورت نگرفته است. همچنین، شواهد اندکی از به‌کارگیری ترکیبی روش‌های خوشه‌بندی و یادگیری ماشین برای شناسایی الگوهای پایداری و پیش‌بینی وضعیت آتی شرکت‌ها وجود دارد.
بررسی هم‌زمان این مجموعه متغیرها، امکان تحلیل جامع‌تری از پایداری مالی شرکت‌ها را ایجاد می‌کند؛ تحلیلی که نه‌تنها به کارایی عملیاتی و سودآوری توجه دارد، بلکه ابعاد ریسک، ساختار سرمایه و توانایی خلق ارزش پایدار را نیز در نظر می‌گیرد. بااین‌حال، پژوهش‌های گذشته در بازار سرمایه ایران عمدتاً هر یک از این شاخص‌ها را به‌صورت منفرد تحلیل کرده و کمتر تلاشی برای یکپارچه‌سازی آن‌ها در قالب مدلی جامع صورت گرفته است. همچنین، شواهد اندکی از به‌کارگیری ترکیبی روش‌های خوشه‌بندی و یادگیری ماشین برای شناسایی الگوهای پایداری و پیش‌بینی وضعیت آتی شرکت‌ها وجود دارد. با مرور ادبیات موضوع می‌توان دریافت که همچنان چندین خلأ اساسی در این حوزه وجود دارد. نخست، فقدان مدلی جامع که قادر باشد مجموعه متنوعی از شاخص‌های مالی را به‌صورت هم‌زمان و یکپارچه در تحلیل پایداری شرکت‌ها لحاظ کند. دوم، کمبود مطالعاتی که به تفاوت ساختاری میان شرکت‌های بورسی ایران پرداخته و آن را در قالب خوشه‌های متمایز تبیین کرده باشند. سوم، اتکای بسیاری از پژوهش‌ها به روش‌های آماری سنتی و بهره‌گیری محدود از ظرفیت‌های نوین یادگیری ماشین در تحلیل و پیش‌بینی پایداری مالی.

در ادامه، بخش دوم به‌مرور ادبیات پژوهش اختصاص‌یافته و مبانی نظری و مطالعات پیشین مرتبط با پایداری مالی شرکت‌ها بررسی می‌شود. در بخش سوم، روش تحقیق و رویکردهای به‌کاررفته شامل الگوریتم خوشه‌بندی و مدل‌های یادگیری ماشین تشریح می‌گردد. بخش چهارم یافته‌های پژوهش را ارائه می‌دهد که شامل نتایج خوشه‌بندی، تحلیل ویژگی‌های متمایزکننده و مقایسه عملکرد مدل‌هاست. درنهایت، بخش پنجم به بحث و نتیجه‌گیری پرداخته و ضمن تبیین پیامدهای مدیریتی و سیاستی نتایج، پیشنهادهایی برای تحقیقات آینده ارائه می‌کند.

۲. پیشینه پژوهش

۱.۲. روش‌های سنتی تحلیل‌های بازار مالی

روش‌های سنتی تحلیل بازارهای مالی عمدتاً بر مبنای مدل‌های آماری کلاسیک و رویکردهای کمی متداول شکل‌گرفته‌اند. در این رویکردها، داده‌های تاریخی به‌عنوان اصلی‌ترین منبع تحلیل در نظر گرفته می‌شود و روابط میان متغیرها اغلب در قالب مدل‌های خطی و با فروض ساده‌سازی‌شده موردبررسی قرار می‌گیرد (جینگ[5] و همکاران، ۲۰۲۳) چنین روش‌هایی به

[1] Leverage
[2] Weighted Average Cost of Capital
[3] Volatility
[4] Melina
[5] Jing

دلیل شفافیت محاسباتی، سهولت تفسیر و امکان استفاده گسترده، در ادبیات مالی جایگاه ویژه‌ای داشته و طی دهه‌های گذشته به‌طور گسترده به کار گرفته‌شده‌اند. بااین‌حال، محدودیت‌های قابل‌توجهی نیز دارند. وابستگی شدید به فروض سخت‌گیرانه آماری، ناتوانی در شناسایی روابط پیچیده، غیرخطی و پویا و بی‌توجهی به ناهمگنی ساختاری و رفتاری میان شرکت‌ها و صنایع، موجب می‌شود که نتایج حاصل از این روش‌ها در بسیاری از شرایط واقع‌بینانه و قابل‌اتکا نباشد. علاوه بر این، تغییرات سریع محیط‌های اقتصادی و مالی، پویایی بازارها و افزایش سطح ریسک و عدم قطعیت، کارایی روش‌های سنتی را به‌طور محسوسی کاهش داده است. ازاین‌رو، اتکای صرف به چنین ابزارهایی نمی‌تواند پاسخگوی نیازهای تصمیم‌گیری و تحلیل در بازارهای مالی معاصر باشد (چوپرا[1] و همکاران، ۲۰۲۱)

۲.۲. نقش هوش مصنوعی و یادگیری ماشین در تحلیل بازارهای مالی

در سال‌های اخیر، پیشرفت‌های چشمگیر در حوزه هوش مصنوعی و یادگیری ماشین باعث شده است تا این فناوری‌ها به ابزاری بسیار مهم و کارآمد برای تحلیل بازارهای مالی تبدیل شوند. بازارهای مالی، به دلیل حجم بسیار بالای داده‌های تولیدشده روزانه، سرعت تغییرات زیاد و پیچیدگی‌های ذاتی ناشی از تأثیر متقابل عوامل گوناگون اقتصادی، سیاسی و روان‌شناختی، چالش بزرگی برای تحلیل‌گران و سرمایه‌گذاران محسوب می‌شوند. روش‌های سنتی تحلیل مانند مدل‌های آماری کلاسیک و تحلیل‌های تکنیکال ساده اغلب در مواجهه با این پیچیدگی‌ها دچار محدودیت شده و نمی‌توانند به‌طور دقیق رفتار آینده بازار را پیش‌بینی کنند (قاسمیان[2] و همکاران، ۲۰۲۲).

هوش مصنوعی و یادگیری ماشین با توانایی پردازش داده‌های حجیم، پیچیده و چندبعدی، این محدودیت‌ها را تا حد قابل‌توجهی کاهش داده‌اند. این فناوری‌ها قادرند از میان انبوه داده‌های تاریخی و جاری، الگوها و روابط غیرخطی و پیچیده‌ای را کشف کنند که از چشم تحلیل‌های انسانی و مدل‌های سنتی پنهان می‌ماند. به کمک الگوریتم‌های یادگیری ماشین، کامپیوترها می‌توانند قواعدی برای پیش‌بینی روند بازار استخراج کنند، بدون اینکه نیاز باشد قوانین صریحی از پیش تعریف شوند. این یعنی سیستم‌های هوشمند قادرند بر اساس داده‌های گذشته، روندها و رفتارهای آینده بازار را به‌صورت خودکار و با دقت بالاتری پیش‌بینی کنند (شهیدی زندی[3] و رجبی[4]، ۲۰۲۲)

از طرفی، بازارهای مالی به دلیل ماهیت پویا و ناپایدار خود، همواره تحت تأثیر عوامل متعددی مانند تغییرات اقتصادی کلان، سیاست‌های پولی، رویدادهای جهانی و همچنین رفتار سرمایه‌گذاران قرار دارند. الگوریتم‌های یادگیری ماشین می‌توانند با تحلیل سریع و مداوم این داده‌ها و به‌روزرسانی مدل‌های خود، تغییرات بازار را به‌سرعت شناسایی کرده و واکنش مناسب را در پیش‌بینی‌ها لحاظ کنند. این موضوع در شرایطی که بازار دچار نوسانات شدید می‌شود، اهمیت بیشتری پیدا می‌کند (ولیخوان انارکی[5]، ۲۰۲۱).

در ادامه به برخی از مطالعات اخیر در حوزه یادگیری ماشین و بازارهای مالی پرداخته می‌شود:

[1] Chopra
[2] Ghasemian
[3] Shahidi Zandi
[4] Rajabi
[5] Valikhan Anaraki

مالتی بانسال[1] و همکاران (۲۰۲۲) در پژوهش خود به بررسی توانایی روش‌های یادگیری ماشین در پیش‌بینی قیمت سهام پرداختند. آن‌ها از داده‌های قیمتی مربوط به ۱۲ شرکت پیشرو در بازار سهام هند طی هفت سال استفاده کرده و الگوریتم‌های K-NN، رگرسیون خطی و رگرسیون بردار پشتیبان را برای تحلیل داده‌ها به کار گرفتند. نتایج حاصل پس از اجرای گام‌به‌گام روش تحقیق، به‌صورت جدولی و نموداری ارائه شد. یافته‌های پژوهش نشان داد که هرچند این الگوریتم‌ها در مقایسه با روش‌های پیشرفته‌تر دقت کمتری دارند، اما همچنان می‌توانند به‌عنوان رویکردهای پایه برای پیش‌بینی روندهای قیمتی مورداستفاده قرار گیرند.

جاناتان بروگارد[2] و ابوالفضل زارعی[3] (۲۰۲۲) در پژوهشی به بررسی تناقض میان کاربرد گسترده تحلیل تکنیکال توسط فعالان بازار و نظریه‌های کارایی بازار پرداختند. آن‌ها با استفاده از مجموعه‌ای از الگوریتم‌های یادگیری ماشین، اثرگذاری قیمت‌های گذشته بر استخراج قواعد معاملاتی را موردبررسی قراردادند. یافته‌های پژوهش نشان داد که اگرچه این قواعد می‌توانند در دوره‌های اولیه سودآوری خارج از نمونه ایجاد کنند، اما میزان این سودآوری درگذر زمان کاهش‌یافته و درنتیجه، کارایی بازار به‌تدریج افزایش‌یافته است.

غیاثی و همکاران (۱۴۰۴) در مطالعه‌ای کارایی ۱۳۰ شرکت بورسی ایران را طی دوره ۱۳۸۶ تا ۱۴۰۲ موردبررسی قراردادند. آن‌ها ابتدا با استفاده از تحلیل پوششی داده‌ها امتیاز کارایی شرکت‌ها را محاسبه کردند؛ در این مدل، کل بدهی و RETRSIK به‌عنوان متغیرهای ورودی و بازده سالانه سهام و نقد شوندگی به‌عنوان متغیرهای خروجی در نظر گرفته شدند. سپس الگوریتم‌های یادگیری ماشین برای شناسایی و ارزیابی متغیرهای کلیدی مؤثر بر پیش‌بینی عملکرد شرکت‌ها به کار گرفته شدند. نتایج تحقیق نشان داد که الگوریتم XGBoost نسبت به سایر مدل‌ها عملکرد پیش‌بینی بهتری دارد. این پژوهش بیان می‌کند که ترکیب DEA و یادگیری ماشین‌ابزاری مؤثر برای تحلیل و پیش‌بینی کارایی شرکت‌ها است و می‌تواند در بهبود تصمیم‌گیری‌های اقتصادی و مالی سرمایه‌گذاران و مدیران نقش مهمی ایفا کند.

کیوانی[4] و همکاران (۲۰۲۵) در پژوهشی با بهره‌گیری از الگوریتم‌های یادگیری ماشین به پیش‌بینی قیمت سهام پرداختند. داده‌های مورداستفاده شامل اطلاعات پنج شرکت بورسی ایران در بازه زمانی ۲۰۱۱ تا ۲۰۲۱ و ۲۵ عامل مرتبط با تحلیل تکنیکال، بنیادی و اقتصادی بود. در این تحقیق از شبکه‌های بازگشتی (FastRNN) به‌عنوان مدل اصلی برای پیش‌بینی سری‌های زمانی سهام استفاده شد. نتایج پژوهش نشان داد که الگوریتم‌های یادگیری ماشین به‌کاررفته توانسته‌اند دقت بالاتری در پیش‌بینی قیمت سهام ارائه دهند و نسبت به مدل‌های متداول عملکرد بهتری داشته باشند.

میشل کوستولا[5] و همکاران (۲۰۲۳) در پژوهشی تأثیر اخبار مرتبط با همه‌گیری کووید-۱۹ بر بازارهای مالی را بررسی کردند. داده‌های تحقیق شامل بیش از ۲۰۳ هزار مقاله خبری منتشرشده در سه وب‌سایت MarketWatch، نیویورک‌تایمز و رویترز طی دوره ژانویه تا ژوئن ۲۰۲۰ بود. نویسندگان با بهره‌گیری از روش‌های یادگیری ماشین، جریان اطلاعات خبری را به‌عنوان متغیری کلیدی در شکل‌گیری انتظارات بازار موردمطالعه قراردادند. نتایج نشان داد که حجم و محتوای اخبار منتشرشده می‌تواند بر بازدهی شاخص S&P۵۰۰ اثر معناداری داشته باشد و تفاوت در منابع خبری نیز نقش قابل‌توجهی در واکنش بازار ایفا می‌کند.

[1] Malti Bansal
[2] Jonathan Brogaard
[3] Abalfazl Zareei
[4] Keyvani
[5] Michele Costola

ناکگاوا[1] و یوشیدا[2] (۲۰۲۴) مدلی مبتنی بر درخت گرادیان بوستینگ برای داده‌های مقطعی ارائه دادند که با استفاده از معیارهای شباهت بین نمونه‌ها، توانستند دقت پیش‌بینی و سودآوری در تحلیل شاخص‌های بورس جهانی را نسبت به روش‌های پیشین بهبود دهند و نشان دهند این رویکرد می‌تواند به‌عنوان ابزاری مؤثر در تحلیل داده‌های مالی عمل کند.

وانگ[3] (۲۰۲۵) در مقاله‌ای با عنوان «پیش‌بینی قیمت سهام تسلا با استفاده از SVM[4]، درخت تصمیم و جنگل تصادفی» به بررسی عملکرد سه الگوریتم یادگیری ماشین ساده در پیش‌بینی حرکات قیمتی سهام پرداختند. آن‌ها با استفاده از شاخص تکنیکال میانگین متحرک ساده ۲۰ روزه و داده‌های پایه قیمت و حجم معاملات، دریافتند که مدل SVM دقت و تعادل بهتری در پیش‌بینی نسبت به سایر مدل‌ها ارائه می‌دهد.

بررسی ادبیات نشان می‌دهد که اگرچه مطالعات متعددی به پیش‌بینی شاخص‌های مالی یا تحلیل کارایی شرکت‌ها پرداخته‌اند، اما کمتر پژوهشی به‌طور هم‌زمان به شناسایی ناهمگنی ساختاری شرکت‌ها و تبیین عوامل کلیدی پایداری مالی توجه کرده است. پژوهش حاضر این خلأ را با طراحی یک چارچوب ترکیبی مبتنی بر خوشه‌بندی و یادگیری ماشین پر می‌کند که علاوه بر کشف الگوهای متمایز میان شرکت‌ها، توان پیش‌بینی و شفاف‌سازی شاخص‌های اثرگذار را نیز فراهم می‌آورد.

۳. روش تحقیق

روش تحقیق این پژوهش بر پایه یک چارچوب ترکیبی در دو گام اصلی استوار است: ابتدا خوشه‌بندی برای شناسایی الگوهای مالی شرکت‌ها و سپس مدل‌های یادگیری ماشین برای پیش‌بینی و اعتبارسنجی خوشه‌ها به‌کار گرفته شدند. در ادامه، مراحل و الگوریتم‌ها تشریح می‌شوند.

۱.۳. الگوریتم K-Means[5]

الگوریتم K-Means یک روش خوشه‌بندی بدون ناظر مبتنی بر تقسیم‌بندی فضای ویژگی‌ها است که هدف آن، جداسازی مجموعه داده به k زیرمجموعه‌ی مجزا با حداکثر شباهت درون‌پوش‌های و حداقل شباهت بین‌حوزه‌ای هست. این الگوریتم نخستین بار توسط (مک‌کوئین[6]، ۱۹۶۷) معرفی شد. ایده اصلی این روش آن است که مجموعه داده‌ها $X=\{x_1,x_2,...,x_n\}$ با d ویژگی به k خوشه مجزا تقسیم شوند به‌گونه‌ای که هر خوشه s_j داری مرکز ثقل u_j بوده و مجموع مربعات فاصله نقاط از مرکز خوشه خود حداقل شود. تابع هدف طبق روابط زیر تعریف می‌شود:

$$j=\min\sum_{j=1}^{k}\sum_{i=1}^{n}z_{ji}\left|x_j-u_j\right|^2 \quad (۱)$$

$$u_j=1/n_j\sum_{i=1}^{n}z_{ji}x_i \quad (۲)$$

$$z_{ji}=\begin{cases}1 & if \quad x_i\in s_j\\ 0 & else\end{cases} \quad (۳)$$

[1] Kei Nakagawa
[2] Kenichi Yoshida
[3] Jierui Wang
[4] Support Vector Machine
[5] K-Means Clustering Algorithm
[6] MacQueen

در این رابطه، منظور از u_j میانگین اعضای خوشه s_j است. درواقع، کمینه‌سازی این مقدار معادل بیشینه‌سازی میانگین مربع فاصله بین نقاط موجود در خوشه‌های مختلف هست. برای محاسبه فاصله هر داده تا مرکز خوشه، می‌توان از معیارهای گوناگونی استفاده کرد. یکی از رایج‌ترین این معیارها، فاصله اقلیدسی است که در رابطه زیر نشان داده‌شده است.

$$d = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \qquad (۴)$$

در این رابطه،n بیانگر تعداد ویژگی‌های داده و پارامترهای x_k و y_k نشان‌دهنده مقدار ویژگی برای دو رکورد x و yهستند. جزئیات اجرای این الگوریتم مطابق با شرح زیر است:

۱. انتخاب تصادفی K داده از مجموعه به‌عنوان مراکز اولیه خوشه‌ها.

۲. محاسبه فاصله هر داده تا مراکز خوشه و انتساب آن به نزدیک‌ترین مرکز بر اساس یک معیار فاصله مناسب.

۳. به‌روزرسانی مراکز خوشه با محاسبه میانگین داده‌های هر خوشه و تعیین مرکز جدید برای آن.

۴. تکرار مراحل ۲ و ۳ تا زمانی که تغییر قابل‌توجهی در مراکز خوشه‌ها مشاهده نشود یا یکی از شرایط توقف برآورده گردد.

برای انتخاب تعداد مناسب خوشه‌ها، از شاخص سیلوئت استفاده شد. این شاخص میزان شباهت هر داده به خوشه خود را نسبت به نزدیک‌ترین خوشه دیگر اندازه‌گیری می‌کند. مقدار این ضریب بین ۱- و ۱ قرار دارد؛ مقادیر نزدیک به ۱ نشان‌دهنده جدایی و انسجام بهتر خوشه‌ها و مقادیر نزدیک به صفر بیانگر هم‌پوشانی خوشه‌ها است.

۲.۳. الگوریتم *XGBoost*

الگوریتم XGBoost نخستین بار توسط (چن[1] و گوسترین[2]، ۲۰۱۶) معرفی شد و بر مبنای چارچوب تقویت گرادیان توسعه‌یافته است. در این رویکرد، مجموعه‌ای از یادگیرنده‌های ضعیف که غالباً درخت‌های تصمیم هستند به‌صورت مرحله‌ای و متوالی آموزش داده می‌شوند و خروجی هر مرحله برای اصلاح خطاهای مرحله پیشین به کار می‌رود. این فرایند تا برآورده شدن معیار توقف ادامه می‌یابد و در پایان، پیش‌بینی نهایی از ترکیب بهینه پیش‌بینی‌های تمام مدل‌ها حاصل می‌شود. انعطاف‌پذیری بالا در پشتیبانی از مدل‌های خطی و غیرخطی، امکان تنظیم دقیق پارامترها و سرعت‌بالای پردازش از مهم‌ترین مزایای XGBoost به شمار می‌رود.

۳.۳. الگوریتم جنگل تصادفی

جنگل تصادفی یک الگوریتم قدرتمند در حوزه یادگیری ماشین است که با ترکیب مجموعه‌ای از درخت‌های تصمیم‌گیری، دقت پیش‌بینی را در مسائل مختلف ازجمله دسته‌بندی و رگرسیون بهبود می‌بخشد. این روش بر پایه ایده یادگیری جمعی عمل می‌کند؛ به این صورت که چندین درخت تصمیم به‌طور مستقل و بر اساس نمونه‌گیری تصادفی از داده‌ها ساخته می‌شوند. در فرآیند ایجاد هر درخت، در هر گره تنها بخشی از ویژگی‌ها به‌طور تصادفی برای انتخاب معیار تقسیم در نظر گرفته می‌شود. این فرآیند موجب افزایش تنوع میان درخت‌ها، کاهش همبستگی ویژگی‌ها و جلوگیری از بیش‌برازش (Overfitting)می‌شود. در مسائل دسته‌بندی، برآورد نهایی با استفاده از رأی‌گیری اکثریت میان خروجی‌های درخت‌ها تعیین

[1] Chen
[2] Guestrin

می‌شود، درحالی‌که در مسائل رگرسیون، میانگین پیش‌بینی‌ها به‌عنوان خروجی نهایی ارائه می‌گردد. استفاده از مجموعه‌ای از مدل‌های ساده و نسبتاً ضعیف (درخت‌های تصمیم‌گیری) در قالب یک ساختار جمعی، سبب می‌شود که جنگل تصادفی از پایداری بالایی در برابر نویز داده‌ها برخوردار باشد و در طیف گسترده‌ای از شرایط، عملکرد قابل‌اعتمادی ارائه دهد (بریمن[1]، ۲۰۰۱)

٤.٣. الگوریتم درخت تصمیم

درخت تصمیم یکی از الگوریتم‌های پرکاربرد یادگیری ماشین در دسته‌بندی روش‌های نظارت‌شده است که به دلیل ترکیب سادگی، شفافیت و توانایی تفسیر، جایگاه ویژه‌ای در تحلیل داده پیداکرده است. این الگوریتم داده‌ها را به‌صورت سلسله‌مراتبی و در قالب یک ساختار گرافیکی به شکل درخت نمایش می‌دهد که در آن گره‌ها بیانگر شرایط یا آزمون‌های تصمیم‌گیری و شاخه‌ها نشان‌دهنده مسیرهای منتهی به نتایج یا پیش‌بینی‌ها هستند. فرآیند ساخت یک درخت تصمیم با شناسایی و انتخاب ویژگی‌هایی آغاز می‌شود که بیشترین توانایی در تفکیک داده‌ها رادارند. در هر گره تصمیم، مجموعه داده‌ها بر اساس معیار انتخاب‌شده تقسیم‌شده و به شاخه‌های فرعی هدایت می‌شوند. این تقسیم‌بندی به‌صورت تکراری ادامه می‌یابد تا زمانی که داده‌ها به گره‌های برگ برسند؛ گره‌هایی که خروجی یا پیش‌بینی نهایی در آن‌ها ارائه می‌شود. یکی از مزایای برجسته این روش، قابلیت آن در کار با داده‌های عددی و طبقه‌ای به‌صورت هم‌زمان است که انعطاف‌پذیری بالایی در تحلیل انواع داده‌ها ایجاد می‌کند. علاوه بر این، سادگی ساختار درخت تصمیم سبب می‌شود که نتایج آن برای افرادی بدون دانش عمیق آماری یا برنامه‌نویسی نیز قابل‌درک باشد (بریمن و همکاران، ۱۹۸۴).

پس از آماده‌سازی داده‌ها، الگوریتم‌های مختلف روی آن‌ها اعمال و مدل‌های مربوطه با استفاده از معیارهای صحت[2]، دقت[3]، یادآوری[4] و *F1-Score* ارزیابی شدند. در ادامه هر یک از معیارها و نقش آن‌ها در ارزیابی مدل پیش‌بینی ریزش مشتریان بانک بررسی‌شده است.

صحت یکی از شاخص‌های اصلی در ارزیابی مدل‌های یادگیری ماشین و طبقه‌بندی است. این معیار نشان‌دهنده نسبت تعداد پیش‌بینی‌های درست مدل به‌کل پیش‌بینی‌ها بوده و معیاری کلی از میزان تطابق خروجی مدل با واقعیت ارائه می‌دهد. هرچه مقدار صحت بالاتر باشد، عملکرد کلی مدل در پیش‌بینی صحیح نمونه‌ها مطلوب‌تر ارزیابی می‌شود. بااین‌حال، صحت به‌تنهایی همیشه تصویر دقیقی از کیفیت مدل ارائه نمی‌کند، به‌ویژه زمانی که داده‌ها نامتوازن هستند و توزیع کلاس‌ها به‌طور قابل‌توجهی متفاوت است. در چنین شرایطی، استفاده از شاخص‌های تکمیلی مانند دقت، یادآوری و *F1-Score* می‌تواند ارزیابی جامع‌تر و واقع‌بینانه‌تری از کارایی مدل فراهم کند.

دقت در مدل‌های یادگیری ماشین، نشان‌دهنده نسبت پیش‌بینی‌های مثبت صحیح به تعداد کل پیش‌بینی‌های مثبت انجام‌شده توسط مدل است. این شاخص بیان می‌کند که چه میزان از نمونه‌هایی که مدل به‌عنوان مثبت شناسایی کرده، واقعاً مثبت بوده‌اند. دقت بالا به معنای کاهش موارد مثبت کاذب و بهبود اطمینان در پیش‌بینی‌های مدل است. بااین‌حال، دقت به‌تنهایی نمی‌تواند تصویر کاملی از عملکرد مدل ارائه دهد؛ چراکه ممکن است مدلی با دقت بالا اما یادآوری پایین، تعداد

[1] Breiman
[2] Accuracy
[3] Precision
[4] Recall

زیادی از نمونه‌های مثبت واقعی را شناسایی نکند؛ بنابراین، برای ارزیابی جامع عملکرد، دقت باید همراه با معیارهای مکملی مانند یادآوری و*F1-Score* در نظر گرفته شود تا تعادل میان شناسایی صحیح و پوشش کامل نمونه‌های مثبت برقرار گردد.

یادآوری نشان می‌دهد که مدل تا چه اندازه قادر است همه نمونه‌های مثبت واقعی را شناسایی کند. این معیار از نسبت تعداد نمونه‌های مثبت درست شناسایی‌شده به‌کل نمونه‌های مثبت واقعی به دست می‌آید. هرچه مقدار یادآوری بیشتر باشد، یعنی مدل توانسته بخش بیشتری از نمونه‌های مثبت را پیدا کند، حتی اگر برخی پیش‌بینی‌ها نادرست بوده باشد. یادآوری به‌ویژه زمانی اهمیت پیدا می‌کند که از دست دادن هر نمونه مثبت می‌تواند پیامدهای مهم یا پرهزینه‌ای داشته باشد؛ بنابراین، برای ارزیابی کامل عملکرد یک مدل، یادآوری باید همراه با معیارهایی مانند دقت و*F1-Score* امتیاز بررسی شود تا تصویر دقیق‌تری از توانایی مدل به دست آید.

امتیاز *F1-Score* معیاری است که با ترکیب دقت و یادآوری در قالب یک عدد، ارزیابی متوازن‌تری از عملکرد مدل ارائه می‌دهد. این شاخص به‌عنوان میانگین هارمونیک دقت و یادآوری محاسبه می‌شود و زمانی بیشترین کاربرد را دارد که نیاز به برقراری تعادل بین کاهش خطاهای مثبت کاذب و شناسایی کامل موارد مثبت واقعی وجود داشته باشد. مقدار بالای امتیاز *F1-Score*نشان‌دهنده آن است که مدل هم در شناسایی درست نمونه‌های مثبت و هم در پوشش حداکثری آن‌ها عملکرد خوبی داشته است. این معیار به‌ویژه در مسائل با داده‌های نامتوازن اهمیت دارد، زیرا تصویری واقع‌بینانه‌تر از توانایی مدل نسبت به استفاده صرف از دقت یا یادآوری ارائه می‌دهد.

٤. یافته‌ها

داده‌های مورد استفاده در این پژوهش شامل اطلاعات مالی ۱۳۰ شرکت پذیرفته‌شده در بورس اوراق بهادار تهران طی سال‌های ۱۳۸۶ تا ۱۴۰۲ است. این داده‌ها بر اساس هشت شاخص کلیدی پایداری مالی گردآوری و به‌عنوان ورودی مراحل بعدی تحلیل به‌کار گرفته شدند.

٤.١. پاک‌سازی داده‌ها

در فرایند پیش‌پردازش داده‌ها، مجموعه‌ای از مراحل به‌منظور ارتقای کیفیت و قابلیت اتکای مجموعه داده اجرا گردید. نخست، داده‌های تکراری با مقایسه کامل مقادیر تمامی ویژگی‌ها شناسایی و حذف شدند تا از ایجاد سوگیری در مدل جلوگیری شود. سپس، به‌منظور شناسایی نمونه‌های دارای مقادیر غیرعادی، از رویکرد آماری مبتنی بر دامنه بین چارکی (IQR)[1] استفاده شد و رکوردهایی که خارج از محدوده مجاز قرار داشتند، از مجموعه داده کنار گذاشته شدند. در گام بعد، جهت جایگزینی مقادیر گمشده، روش برآورد با نزدیک‌ترین همسایگان (KNN Imputation)[2] به کار گرفته شد. در این روش، هر مقدار گمشده با میانگین مقادیر متناظر در پنج نمونه مشابه جایگزین گردید. سنجش میزان شباهت نمونه‌ها بر اساس فاصله اقلیدسی در فضای ویژگی‌ها صورت گرفت تا ضمن حفظ دقت، از ایجاد انحراف ناشی از انتخاب تعداد نامتناسب همسایگان جلوگیری شود.

[1] Interquartile Range
[2] K-Nearest Neighbors Imputation

۲.٤. انتخاب تعداد خوشه‌ها

برای تعیین k، معیار سیلوئت برای $k = 1, 2, ..., 8$ محاسبه شد. بیشینه سیلوئت در $k = 3$ برابر به دست آمد. جدول ۱ مقادیر سیلوئیت را نشان می‌دهد.

جدول ۱. مقادیر سیلوئت

k	silhouette
3	۰/۵۱
4	۰/۴۱
5	۰/۳۲
6	۰/۲۱
7	۰/۱۷
8	۰/۱۶

۳.٤. پروفایل خوشه‌ها

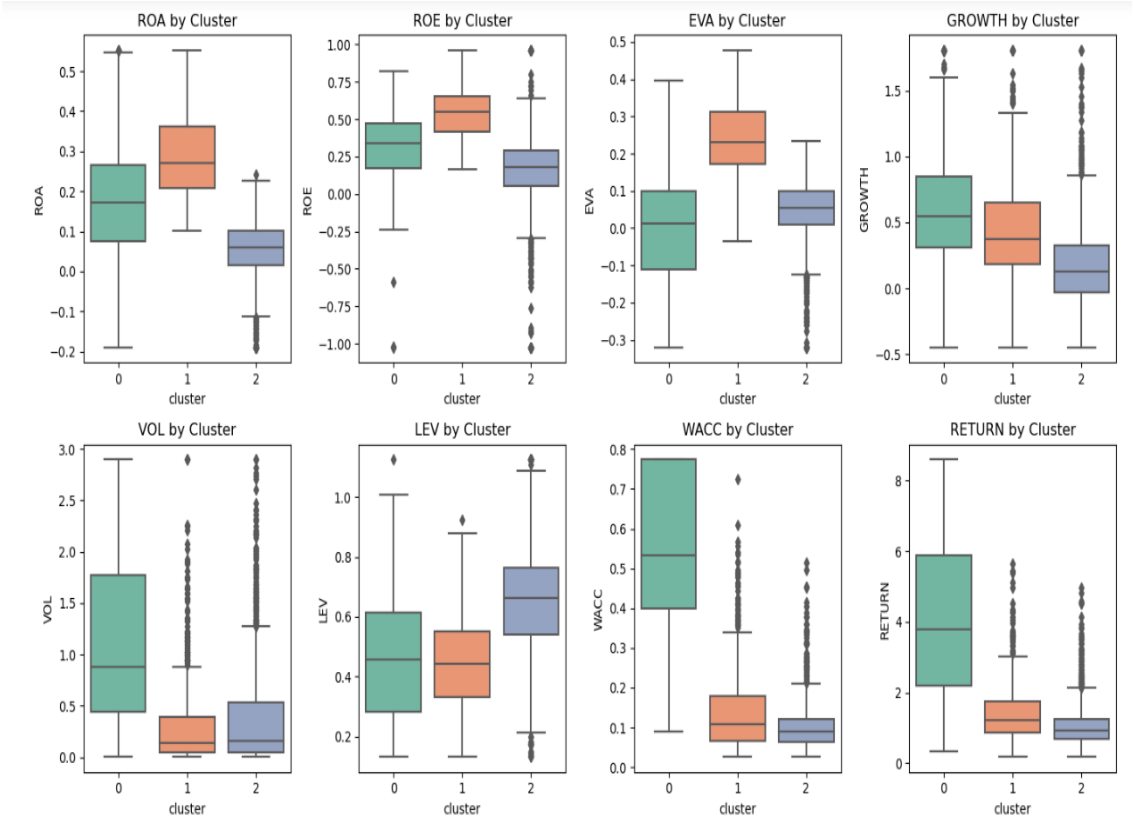
جدول ۲ پروفایل مالی خوشه‌ها را بر اساس میانگین شاخص‌های مالی نشان می‌دهد. خوشه ۱ با سطح بالای سودآوری و ارزش‌افزوده اقتصادی، همراه با اهرم مالی و نوسان پذیری پایین و هزینه سرمایه محدود، بیانگر بنگاه‌های باکیفیت و پایدار است که از ساختار سرمایه بهینه و بازدهی مستمر برخوردارند. خوشه ۰ دارای رشد و بازده بسیار بالا اما همراه با نوسان پذیری و هزینه سرمایه زیاد و ارزش‌افزوده اقتصادی مرزی است؛ الگویی که نشان‌دهنده شرکت‌های پر ریسک و بازده‌محور بوده و برای حفظ پایداری نیازمند مدیریت هزینه سرمایه و کنترل ریسک هستند. خوشه ۲ کمترین سطح سودآوری، رشد و بازده را به همراه اهرم مالی بالا و ارزش‌افزوده اقتصادی محدود دارد؛ وضعیتی که بیشتر ناشی از ضعف عملکرد عملیاتی است تا محدودیت تأمین مالی و بهبود آن مستلزم ارتقای بهره‌وری و بازنگری در ساختار سرمایه است. این تمایزها شواهد روشنی از ناهمگنی ساختاری در پایداری مالی بنگاه‌ها فراهم می‌آورد و مبنایی برای طراحی سیاست‌های متمایز در هر خوشه محسوب می‌شود.

جدول ۲. پروفایل خوشه‌ها

خوشه‌ها	ROA	ROE	EVA	GROWTH	VOL	LEV	WACC	RETURN
0	0.1856	0.3163	-0.0035	0.6260	1.1501	0.4631	0.5469	4.2451
1	0.2919	0.5433	0.2449	0.4591	0.3261	0.4463	0.1430	1.4274
2	0.0532	0.1614	0.0451	0.1733	0.3844	0.6500	0.1015	1.0667

٤.٤. بررسی بصری تمایز ساختاری خوشه‌ها

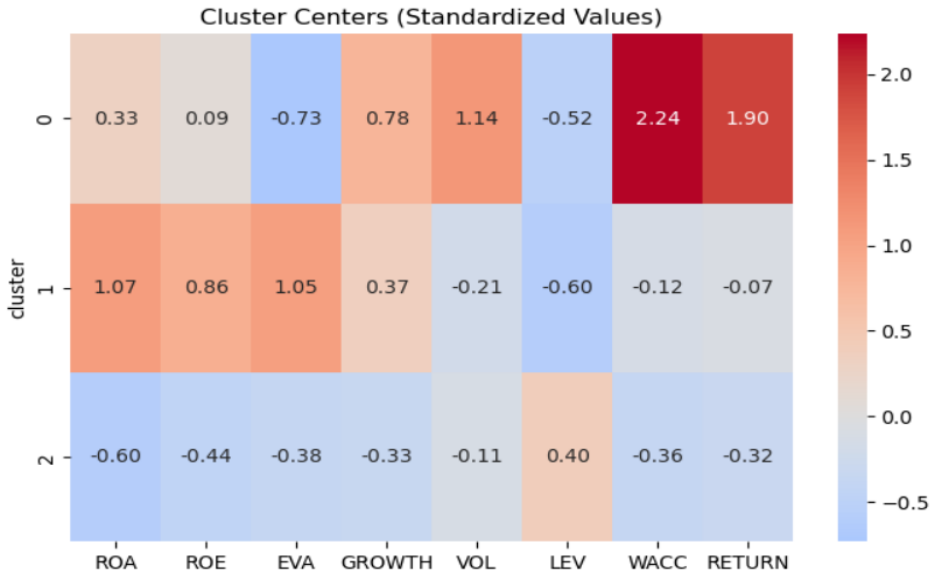
برای سنجش تفاوت‌های ساختاری بین خوشه‌های شناسایی‌شده، از روش‌های مصورسازی مقایسه‌ای استفاده شد. شکل شماره ۱ جعبه نمودار هشت شاخص مالی را برای خوشه‌های مختلف نشان می‌دهد.



شکل ۱. جعبه نمودار شاخص‌های مالی

اختلاف در مقادیر میانه، دامنه بین چارکی و میزان پراکندگی داده‌ها بیانگر این است که هر خوشه ازنظر توزیع شاخص‌ها الگوی خاص خود را دارد.

شکل ۲ نیز نقشه حرارتی مراکز خوشه‌ها را بر اساس مقادیر استانداردشده ارائه می‌دهد که در آن، شدت رنگ میزان فاصله هر شاخص از میانگین کل را مشخص می‌کند؛ رنگ‌های گرم‌تر مقادیر بالاتر و رنگ‌های سردتر مقادیر پایین‌تر را نمایش می‌دهند.



شکل ۲. نقشه حرارتی مراکز خوشه‌ها

شباهت الگوهای مشاهده‌شده در این دو نمودار نشان می‌دهد که خوشه‌ها ازنظر ساختار مالی تفاوت‌های قابل‌توجهی دارند و این امر اعتبار خوشه‌بندی را تأیید می‌کند.

۵.٤. مقایسه الگوریتم‌های یادگیری ماشین برای پیش‌بینی برچسب خوشه‌ها

در این مرحله، داده‌های برچسب خورده حاصل از خوشه‌بندی K-means به‌عنوان ورودی مدل طبقه‌بندی نظارت‌شده به کار گرفته شد. متغیر هدف، برچسب خوشه (۰، ۱، ۲) و متغیرهای ورودی شامل هشت شاخص مالی منتخب به‌علاوه فاصله استانداردشده هر نمونه از مرکز خوشه خود بودند. داده‌ها به نسبت ٪۷۵ برای آموزش و ٪۲۵ برای آزمون، با حفظ توزیع برچسب‌ها در هر مجموعه، تقسیم شدند.

در ادامه به‌منظور ارزیابی قابلیت پیش‌بینی برچسب خوشه‌های شناسایی‌شده، سه الگوریتم یادگیری ماشین شامل XGBoost، جنگل تصادفی و درخت تصمیم بر روی مجموعه داده‌ی برچسب خورده حاصل از مرحله خوشه‌بندی اجرا گردید. ارزیابی مدل‌ها بر اساس داده‌های آزمون و شاخص‌های صحت، یادآوری،F1-Score و دقت انجام شد مطابق جدول ۳ انجام گردید.

جدول ۳. معیارهای ارزیابی الگوریتم‌های یادگیری ماشین

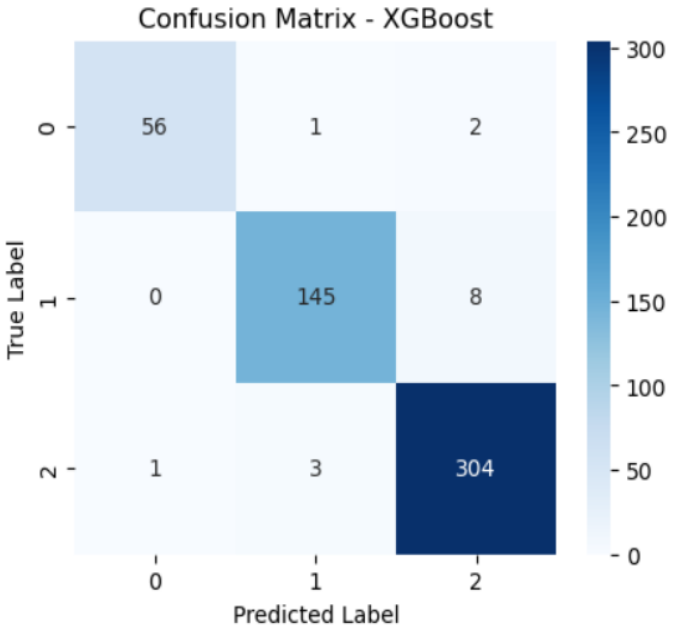
مدل	دقت	یادآوری	F1-Score	صحت
XGBoost	0.975	0.961	0.968	0.971
جنگل تصادفی	0.936	0.942	0.939	0.952
درخت تصمیم	0.897	0.914	0.905	0.923

تحلیل مقایسه‌ای سه مدل یادگیری ماشین تفاوت‌های معناداری را در عملکرد طبقه‌بندی بر اساس چهار شاخص ارزیابی بیان می‌کند. مدل XGBoost در تمامی معیارها برتری محسوسی نسبت به دو الگوریتم دیگر دارد که بیانگر توانایی بالای آن در تعمیم‌پذیری و طبقه‌بندی صحیح شرکت‌ها در تمامی خوشه‌هاست. این برتری عمدتاً ناشی از سازوکار گرادیان بوستینگ است که به‌صورت ترتیبی خطاهای باقی‌مانده مدل‌های ضعیف را اصلاح‌کرده و به مدلی قدرتمند و پایدار منجر می‌شود. مدل جنگل تصادفی، هرچند اندکی ضعیف‌تر از XGBoost عمل کرده، اما عملکردی پایدار و رقابتی ارائه می‌دهد و به‌واسطه ساختار تجمعی خود، از بیش‌برازش جلوگیری کرده و تعاملات پیچیده میان ویژگی‌ها را به‌خوبی مدل می‌کند. در مقابل، درخت تصمیم اگرچه ازنظر تفسیرپذیری و سرعت محاسباتی مزیت دارد، اما پایین‌ترین مقادیر را در تمامی معیارها به ثبت رسانده که نشان‌دهنده محدودیت آن در پیش‌بینی دقیق برچسب خوشه‌ها بدون بهره‌گیری از روش‌های تجمعی است. این نتایج تأکید می‌کند که روش‌های مبتنی بر گرادیان‌بوستینگ و مدل‌های تجمعی در شناسایی ناهمگنی ساختاری پروفایل شرکت‌ها کارآمدتر بوده و می‌توانند مبنای مطمئن‌تری برای تحلیل‌های مدیریتی و تصمیم‌گیری سیاستی فراهم آورند.

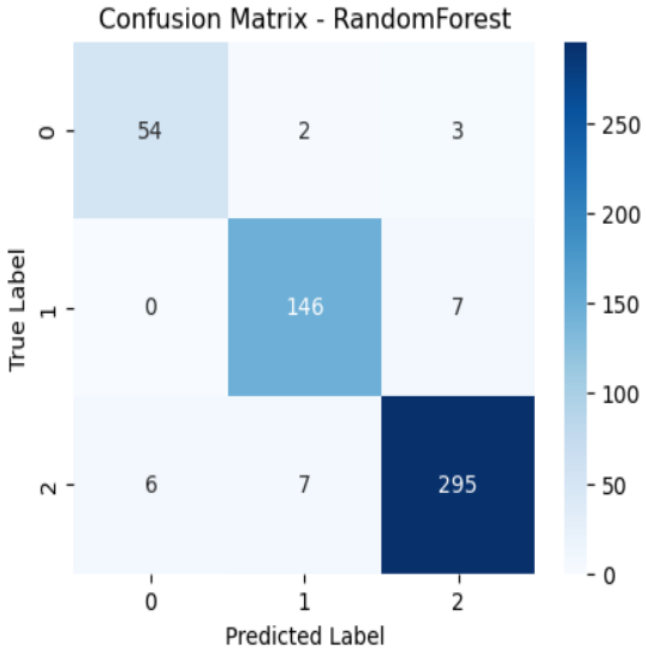
در ارزیابی سه مدل XGBoost، جنگل تصادفی و درخت تصمیم بر اساس ماتریس‌های درهم‌ریختگی، نتایج نشان داد که XGBoost بالاترین دقت را در تفکیک خوشه‌ها ارائه می‌دهد. این مدل در خوشه ۲ تقریباً بدون خطا عمل کرده و تنها تعداد اندکی نمونه به خوشه‌های دیگر تخصیص‌یافته است؛ همچنین در خوشه‌های ۰ و ۱ نیز خطاها حداقلی بوده‌اند. چنین عملکردی بیانگر قدرت بالای این الگوریتم در شناسایی مرزهای تصمیم و مدیریت داده‌های با همپوشانی اندک است. مدل جنگل تصادفی نیز عملکردی قابل‌قبول و نزدیک به XGBoost دارد، اما در مقایسه، خطاهای بیشتری بین خوشه‌های ۲ و سایر خوشه‌ها مشاهده می‌شود. به‌طور خاص، چندین نمونه از خوشه ۲ به‌اشتباه در خوشه‌های ۰ و ۱ قرارگرفته‌اند که نشان‌دهنده

حساسیت این مدل نسبت به داده‌های با ویژگی‌های مرزی است. در مقابل، مدل درخت تصمیم کمترین دقت را در این مقایسه نشان داده است. این مدل خطاهای بیشتری در تمایز خوشه ۲ از خوشه‌های دیگر داشته و در خوشه ۱ نیز نسبت به دو مدل دیگر پیش‌بینی‌های نادرست بیشتری ثبت کرده است. این امر نشان می‌دهد که استفاده از یک درخت منفرد در مقایسه با مدل‌های تجمعی توانایی کمتری در مدل‌سازی روابط پیچیده میان متغیرها دارد.

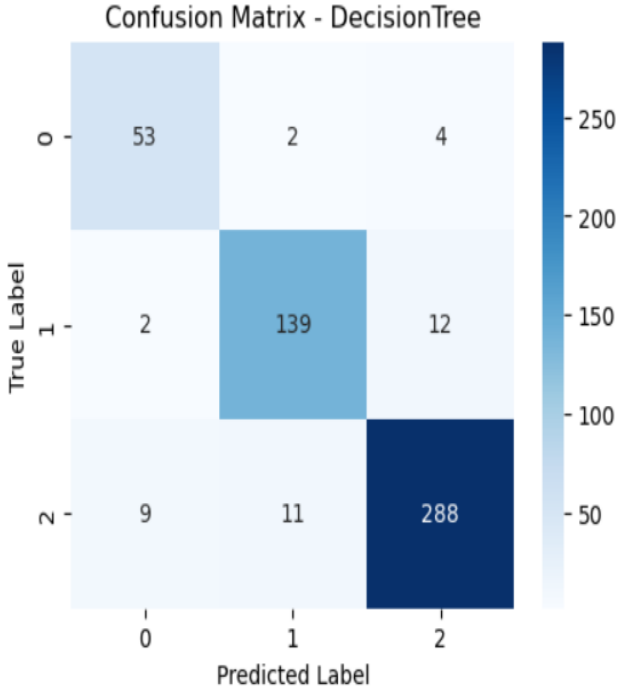
درمجموع، تحلیل ماتریس‌های درهم‌ریختگی حاکی از آن است که مدل‌های تجمعی، به‌ویژه XGBoost، در مسائل چند کلاسه با ساختار داده ناهمگن، نسبت به مدل‌های ساده‌تر برتری محسوسی دارند



شکل ۳. ماتریس درهم ریختگی الگوریتم XGBoost



شکل ٤. ماتریس درهم ریختگی الگوریتم جنگل تصادفی



شکل ۵. ماتریس درهم ریختگی الگوریتم درخت تصمیم

٦.٤. تحلیل ویژگی‌های مهم

به‌منظور شناسایی متغیرهای کلیدی در تمایز خوشه‌های مالی، خروجی اهمیت ویژگی‌ها از سه الگوریتمXGBoost ، جنگل تصادفی و درخت تصمیم مورد ارزیابی تطبیقی قرار گرفت. نتایج حاکی از آن است که ROA در هر سه مدل‌بالاترین سهم را در پیش‌بینی برچسب خوشه‌ها داشته و این همگرایی میان الگوریتم‌ها، نقش بنیادین این شاخص در انعکاس کارایی عملیاتی و سودآوری پایدار بنگاه‌ها را تأیید می‌کند. پس‌ازآن، WACC و EVA به‌ویژه در مدل‌های XGBoost و جنگل تصادفی اهمیت بالایی یافته‌اند که بیانگر حساسیت خوشه‌بندی به ساختار هزینه سرمایه و توانایی ایجاد ارزش مازاد بر هزینه سرمایه به‌عنوان محرک‌های تمایز عملکردی است.

در مدل درخت تصمیم، هرچند ROA همچنان جایگاه نخست را حفظ کرده، اما توزیع اهمیت ویژگی‌ها گرایش بیشتری به WACC و سپس RETURN داشته است. این تفاوت، بازتابی از ماهیت الگوریتم‌ها در شناسایی روابط غیرخطی و وابستگی‌های متقابل میان متغیرهاست؛ جایی که مدل‌هایی مانند XGBoost و جنگل تصادفی تمایل بیشتری به متغیرهای ترکیبی و با اثرات تعاملی نشان می‌دهند، درحالی‌که مدل درخت تصمیم به دلیل ساختار ساده‌تر، بر متغیرهایی با آستانه‌های تفکیک‌پذیر روشن تمرکز می‌کند.

الگوهای مشاهده‌شده در متغیرهای بااهمیت کمتر ازجمله GROWTH،LEV و VOLحاکی از آن است که اگرچه این شاخص‌ها به‌تنهایی قدرت تمایز پایین‌تری دارند، اما در کنار شاخص‌های اصلی می‌توانند به بهبود دقت مدل کمک کنند. به‌طور خاص، سهم پایین‌تر LEV در همه مدل‌ها می‌تواند نشان‌دهنده این باشد که تفاوت‌های ساختاری میان خوشه‌ها بیشتر ناشی از شاخص‌های بازدهی و هزینه سرمایه است تا صرفاً ساختار بدهی. در جدول شماره ۴ اهمیت ویژگی‌های متغیرها ذکرشده است.

جدول ٤. اهمیت ویژگی‌ها

ویژگی‌ها	XGBoost	جنگل تصادفی	درخت تصمیم
ROA	46.64	35.3	61.58
WACC	18.81	15.75	24.44
EVA	9.16	21.5	3.46
RETURN	8.04	8.05	4.4
ROE	3.06	9.85	0.72
GROWTH	4.34	1.65	2.91
LEV	3.25	3.9	0
VOL	3.03	1.11	0.99

طبق جدول شماره ۴ نتایج نشان می‌دهد که شرکت‌ها برای دستیابی به پایداری مالی باید بیشترین تمرکز خود را بر بهبود ROA قرار دهند، زیرا این شاخص مستقیماً بازتاب‌دهنده کارایی عملیاتی و سودآوری پایدار است. همچنین کاهش WACC از طریق مدیریت ساختار سرمایه و انتخاب منابع تأمین مالی بهینه، در کنار افزایش EVA با اجرای پروژه‌هایی که بازدهی آن‌ها فراتر از هزینه سرمایه است، می‌تواند مبنای خلق ارزش بلندمدت قرار گیرد. در کنار این عوامل اصلی، توجه به GROWTH، RETURNو کنترل سطح LEV و VOL نیز هرچند در اولویت‌های بعدی قرار دارند، اما نقش مکمل مهمی در تقویت اعتماد سرمایه‌گذاران و ارتقای انعطاف‌پذیری مالی ایفا می‌کنند.

۵. بحث و نتیجه‌گیری

یافته‌های این پژوهش نشان داد که ترکیب رویکردهای خوشه‌بندی و مدل‌های یادگیری ماشین می‌تواند ابزاری کارآمد برای شناسایی الگوهای پایداری مالی شرکت‌های بورسی ایران باشد. خوشه‌بندی K-Means سه گروه متمایز از شرکت‌ها را بر اساس شاخص‌های مالی شناسایی کرد: بنگاه‌های پایدار با سودآوری بالا و ریسک پایین، شرکت‌های پرریسک با بازده بالا و بنگاه‌های کم‌بازده با ساختار نامطلوب. این تفکیک، شواهد روشنی از ناهمگنی ساختاری در بازار سرمایه ایران ارائه می‌دهد. در مرحله پیش‌بینی، الگوریتم XGBoost بالاترین میزان دقت، یادآوری و F1-Score را ثبت کرد و نسبت به جنگل تصادفی و درخت تصمیم عملکرد بهتری داشت. تحلیل اهمیت ویژگی‌ها نیز نشان داد که شاخص ROA مهم‌ترین متغیر در تمایز خوشه‌ها بوده و پس‌ازآن WACC و EVA نقش برجسته‌ای در تبیین تفاوت‌های ساختاری ایفا می‌کنند. این نتایج بیانگر آن است که بهبود کارایی عملیاتی و مدیریت بهینه هزینه سرمایه می‌تواند بهترین راهبرد برای ارتقای پایداری مالی شرکت‌ها باشد. از منظر کاربردی، نتایج این پژوهش پیامدهای مهمی برای سه گروه از ذی‌نفعان دارد. مدیران شرکت‌ها می‌توانند از نتایج خوشه‌بندی برای مقایسه موقعیت سازمان خود با هم‌گروه‌ها و طراحی راهبردهای بهینه درزمینهٔ سرمایه‌گذاری، تأمین مالی و مدیریت ریسک بهره‌برداری کنند. سرمایه‌گذاران قادر خواهند بود با شناسایی خوشه‌های پرریسک یا پایدار، تصمیمات آگاهانه‌تری در انتخاب سبد سهام اتخاذ نمایند. در سطح کلان، سیاست‌گذاران می‌توانند از نتایج برای طراحی حمایت‌های هدفمند از بنگاه‌های کم‌بازده و ایجاد مشوق‌هایی برای شرکت‌های پایدار استفاده کنند.

بااین‌حال، پژوهش حاضر با محدودیت‌هایی همراه است. نخست آنکه داده‌ها صرفاً از بورس ایران استخراج‌شده و بنابراین، تعمیم نتایج به سایر بازارها باید بااحتیاط صورت گیرد. دوم، داده‌ها ماهیت مقطعی داشته و پویایی زمانی و تغییرات رفتاری بنگاه‌ها در طول دوره‌های بلندمدت را به‌طور کامل پوشش نمی‌دهند. سوم، تمرکز پژوهش بر شاخص‌های مالی مانع از آن

شد که عوامل کیفی همچون حاکمیت شرکتی، نوآوری یا مسئولیت اجتماعی نیز در مدل لحاظ شوند. این محدودیت‌ها زمینه‌ای برای توسعه مطالعات آتی فراهم می‌آورد. در راستای پیشنهادهای آینده، استفاده از مدل‌های یادگیری عمیق مانند LSTMبرای تحلیل سری‌های زمانی مالی و شبکه‌های عصبی گرافی برای بررسی روابط میان شرکت‌ها توصیه می‌شود. همچنین بهره‌گیری از رویکردهای ترکیبی مانند خوشه‌بندی پویا یا Ensemble Clustering می‌تواند به بهبود دقت در شناسایی الگوهای پایداری کمک کند. ورود شاخص‌های غیرمالی نظیر کیفیت مدیریت، ساختار هیئت‌مدیره و میزان مسئولیت اجتماعی شرکت نیز می‌تواند تصویری جامع‌تر از پایداری سازمانی ارائه دهد. درمجموع، این پژوهش نشان داد که چارچوب پیشنهادی مبتنی بر خوشه‌بندی و یادگیری ماشین نه‌تنها قابلیت تبیین تفاوت‌های ساختاری در بازار سرمایه ایران را دارد، بلکه با تطبیق متغیرها و شرایط محیطی، می‌تواند به سایر بازارها و صنایع نیز تعمیم یابد. بدین ترتیب، این رویکرد ظرفیت آن را دارد که به‌عنوان یک ابزار تحلیلی و تصمیم‌یار در ارتقای شفافیت مالی، مدیریت ریسک و پایداری اقتصادی به‌کاررفته گرفته شود.

تعارض منافع

تعارض منافع ندارم.

ORCID

Omid Valizadeh **https://orcid.org/0009-0005-6246-8244**
Mahsa Akhavan Rad **https://orcid.org/0009-0006-2898-3135**
Atieh Jafarian **https://orcid.org/0009-0000-3950-6074**

References

Bansal, M. Goyal, A. & Choudhary, A. (2022). Stock market prediction with high accuracy using machine learning techniques. Procedia Computer Science, 215, 247-265. https://doi.org/10.1016/j.procs.2022.12.028

Brogaard, J., & Zareei, A. (2023). Machine learning and the stock market. *Journal of Financial and Quantitative Analysis*, *58*(4), 1431-1472. https://doi.org/10.1017/S0022109022001120

Ghiyasi, M., Valizadeh O. Joshani B, & Lotfi, M. (1404). Evaluating The Efficiency of Iranian Listed Companies Using a Combined Method of Data Envelopment Analysis and Machine Learning. *The Decision Science and Intelligent Systems, 2*(1), 1-25. [in persian]

Keyvani, F., Nassirzadeh, F., Askarany, D., & Khansalar, E. (2025). A Hybrid Forecasting Model for Stock Price Prediction: The Case of Iranian Listed Companies. *Journal of Risk and Financial Management, 18*(5) 281. https://doi.org/10.3390/jrfm18050281

Costola, M., Hinz, O., Nofer, M., & Pelizzon, L. (2023). Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Research in international business and finance*, *64*, 101881. https://doi.org/10.1016/j.ribaf.2023.101881

Nakagawa, K., & Yoshida, K. (2022). Time-series gradient boosting tree for stock price prediction. International *Journal of Data Mining, Modelling and Management, 14*(2), 110-125. https://doi.org/10.1504/IJDMMM.2022.123357

Wang, J. (2025). Tesla stock prediction with SVM, decision tree and random forest. *Highlights in Business, Economics and Management*, *50*, 342-346. https://doi.org/10.54097/6xrr2902

Mohammadi, S., Saeidi, H., & Naghshbandi, N. (2021). The impact of board and audit committee characteristics on corporate social responsibility: evidence from the Iranian stock exchange. International *Journal of Productivity and Performance Management, 70*(8), 2207-2236. https://doi.org/10.1108/IJPPM-10-2019-0506

Ossareh, A., Pourjafar, M. S., & Kopczewski, T. (2021). Cognitive biases on the iran stock exchange: Unsupervised learning approach to examining feature bundles in investors' portfolios. *Applied Sciences*, *11*(22), 10916. https://doi.org/10.3390/app112210916

Gholamzadeh, M., Faghani, M. and pifeh, A. (2021). Implementing machine learning methods in the prediction of the financial constraints of the companies listed on Tehran's stock exchange. *International Journal of Finance & Managerial Accounting*, *6*(20), 131-144.

Wang, P., Zheng, X., Ai, G., Liu, D., & Zhu, B. (2020). Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. *Chaos, Solitons & Fractals*, *140*, 110214. https://doi.org/10.1016/j.chaos.2020.110214**.**

Melina, Sukono, Napitupulu, H., & Mohamed, N. (2023). A conceptual model of investment-risk prediction in the stock market using extreme value theory with machine learning: a semisystematic literature review. *Risks*, *11*(3), 60. https://doi.org/10.3390/risks11030060

Jing, D., Imeni, M., Edalatpanah, S. A., Alburaikan, A., & Khalifa, H. A. E. W. (2023). Optimal selection of stock portfolios using multi-criteria decision-making methods. *Mathematics*, *11*(2), 415. https://doi.org/10.3390/math11020415.

Chopra, R., & Sharma, G. D. (2021). Application of artificial intelligence in stock market forecasting: a critique, review, and research agenda. *Journal of risk and financial management*, *14*(11), 526. https://doi.org/10.3390/jrfm14110526.

Ghasemian, B., Shahabi, H., Shirzadi, A., Al-Ansari, N., Jaafari, A., Kress, V. R., ... & Ahmad, A. (2022). A robust deep-learning model for landslide susceptibility mapping: A case study of Kurdistan Province, Iran. *Sensors*, *22*(4), 1573. https://doi.org/10.3390/s22041573.

Shahidi Zandi, M., & Rajabi, R. (2022). Deep learning based framework for Iranian license plate detection and recognition. *Multimedia Tools and Applications*, *81*(11), 15841-15858. https://doi.org/10.1007/s11042-022-12023-x

Anaraki, M. V., Farzin, S., Mousavi, S. F., & Karami, H. (2021). Uncertainty analysis of climate change impacts on flood frequency by using hybrid machine learning methods. *Water Resources Management*, *35*(1), 199-223. https://doi.org/10.1007/s11269-020-02719-w.

McQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.* (pp. 281-297).

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). https://doi.org/10.1145/2939672.293978.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. http://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group. https://doi.org/10.1201/9781315139470