Sharif Industrial Engineering and Management Journal (2025), 41(1), 12-27

Sharif University of Technology

https://sjie.journals.sharif.edu/

Research Article

Designing a Decision Support System for Portfolio Management using Data Science Methods (Tehran Stock Exchange)

Navid Javaheri, Najmeh Neshat * and Abbass Ali Jafari Nodushan

Department of Industrial Engineering, Faculty of Engineering, Meybod University, Meybod, Iran.
* corresponding author: (neshat@meybod.ac.ir)

Article Info

Article history:

Received:22 November 2024

Revised:02 April 2024

Accepted:24 Jun 2024

Keywords:

Portfolio management,
modeling,
data science,
classification,
clustering,
segmentation.

Abstract

Increasing profitability and reducing risk always requires choosing a smart investment path while taking advantage of data analysis; Therefore, it is necessary to provide a technique in the form of a decision support system for stock portfolio management while better understanding the position of data science. In this research, while combining data science methods with the Markowitz model, classification and forecasting models have also been created. Detecting financial fraud of companies is also effective; Hierarchical and Cummins algorithms have also been used in order to cluster active companies in the Tehran Stock Exchange. In terms of data classification, the linear support vector machine classification algorithm has been identified with 70% accuracy in comparison with the decision tree, Nyobies, nearest neighbour, and multilayer perceptron algorithms, and in terms of building prediction models, the decision tree with the minimum amount of error, in comparison with Nyobies algorithms, is the closest Neighbor, multilayer perceptron has been identified. The primary data in the current research includes twenty titles of financial indicators and daily data of stock prices of companies in the Python programming space. It was concluded that choosing stocks from among the clusters formed by different stocks of companies will reduce the risk of diversifying the stock portfolio. The mentioned system will be able to generalize as a comprehensive model for data in different time intervals and the indicators desired by the analyst, thus helping investors in a fast and accurate analytical way. The performance of the final technique will be such that the investor first selects several desired companies according to the heard, analysis, and news; It predicts the price performance of companies; Then, it separates the companies that have succeeded in accepting the condition by using the clustering model and separates the portfolio of various stocks. It forms according to the amount of risk and expected return. Importantly, as mentioned, the presented classification and forecasting models, in addition to being used in the formation and management of the stock portfolio, will be able to be effective in predicting the possible fraud of companies.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

To Cite this article:

Javaheri, N., Neshat, N. and Jafari Nadoushan, A.A. 2025. Designing a decision support system for portfolio management using data science methods (Tehran stock exchange), Sharif Industrial Engineering and Management Journal, 41(1), 12-27. https://doi.org/10.24200/j65.2024.63158.2375

E-ISSN: 2676-475X

مجله علمی مهندسی صنایع و مدیریت شریف

https://sjie.journals.sharif.edu

مقاله‌ی پژوهشی

طراحی سیستم پشتیبان تصمیم‌گیری مدیریت سبد سهام با بهره‌گیری از روش‌های علم داده (بازار بورس اوراق بهادار تهران)

نوید جواهری، نجمه نشاط * و عباسعلی جعفری ندوشن

گروه مهندسی صنایع، دانشکده‌ی فنی مهندسی، دانشگاه میبد، یزد.

*نویسنده مسئول (neshat@meybod.ac.ir)

اطلاعات مقاله

تاریخ دریافت : ۱۴۰۲/۰۹/۰۱

تاریخ اصلاحیه : ۱۴۰۳/۰۱/۱۴

تاریخ پذیرش : ۱۴۰۳/۰۴/۰۴

واژگان کلیدی:

مدیریت سبد سهام،
مدل‌سازی،
علم داده،
طبقه‌بندی،
خوشه‌بندی،
بخش‌بندی.

چکیده

افزایش سودآوری و کاهش میزان ریسک، مستلزم انتخاب مسیر هوشمند سرمایه‌گذاری ضمن بهره‌گیری از تحلیل داده است؛ لذا امروزه ارائه‌ی روشی در قالب سیستم پشتیبان تصمیم‌گیری مدیریت سبد سهام، ضمن درک بهتر جایگاه علم داده، ضروری است؛ لذا در پژوهش حاضر، علاوه‌بر دستیابی به این مهم، ضمن ترکیب روش‌های علم داده با مدل مارکویتز، مدل‌های طبقه‌بندی و پیش‌بینی نیز ایجاد شده‌اند، که در تشخیص تقلب مالی شرکت‌ها مؤثر هستند. همچنین، به‌منظور خوشه‌بندی شرکت‌های فعال در بورس اوراق بهادار تهران، الگوریتم‌های سلسله‌مراتبی و کامینز استفاده شده‌اند. درخصوص طبقه‌بندی داده‌ها، الگوریتم طبقه‌بندی ماشین بردار پشتیبان خطی با دقت ۷۰٪ در مقایسه با الگوریتم‌های درخت تصمیم، نایوبیز، نزدیک‌ترین همسایه، و پرسپترون چندلایه و درخصوص ساخت مدل‌های پیش‌بینی، درخت تصمیم با کمینه‌ی میزان خطا، در مقایسه با الگوریتم‌های نایوبیز، نزدیک‌ترین همسایه، پرسپترون چندلایه شناسایی شده‌اند. داده‌های اولیه در پژوهش حاضر، شامل ۲۰ عنوان شاخص مالی و داده‌های روزانه‌ی قیمت سهام شرکت‌ها در فضای برنامه‌نویسی پایتون بوده‌اند.

۱. مقدمه

در بسیاری از سازمان‌ها، مانند بانک‌ها، سازمان‌های مالیاتی و مؤسسات بزرگ حسابداری و حسابرسی داده‌های مالی جمع‌آوری می‌شوند و «با افزایش حجم داده‌های ذخیره‌شده در سازمان‌ها، پایگاه‌های داده و سایر مخازن اطلاعاتی، ساخت ابزارهایی برای تجزیه‌وتحلیل و استخراج دانش معنادار از چنین حجم وسیعی از داده‌ها بسیار حیاتی می‌شود».[۱] حال بازارهای مالی به‌عنوان موتور محرک اقتصاد هر کشور، به ویژه بازار مبادله‌ی سهام شرکت‌ها، باید به سرعت با دانش و ابزار جدید تطبیق یابند؛ چرا که در مواجه با حجم انبوه داده در بازارهای مالی در قالب کلان‌داده‌ها، سرعت و دقت، هر چه بیشتر نقش کلیدی ایفاء می‌کنند و دستیابی به این مهم با استفاده از نرم‌افزارهای تحلیل داده ممکن است. داده‌کاوی به‌عنوان زیرمجموعه‌ای از علم داده[1]، برای مدیریت داده‌های گسترده و افزایش کارایی شرکت و بینش تجاری در تجارت مالی بسیار مهم است؛[۲] در حقیقت «استفاده از روش‌های داده کاوی[2] بر روی داده‌های مالی می‌تواند به حل چالش‌های طبقه‌بندی و پیش‌بینی کمک کند و در عین حال فرآیند تصمیم‌گیری را برای افراد دخیل در امور مالی هر کشوری آسان‌تر سازد».[۳] بنابراین، بررسی الگوریتم‌های داده کاوی و ارائه‌ی یک الگو برای تحلیلگران بازارهای مالی، با توجه به اهمیت تحلیل اطلاعات و سرعت تولید داده در بازارهای مالی، بسیار حائز اهمیت است. بازارهای مالی ایران نیز از این قاعده مثتثنی نیستند. بنابراین، این پرسش ضروری است که چه میزان تحلیل اطلاعات بازارهای مالی، همچون تحلیل اطلاعات سهام شرکت‌ها، با استفاده از ابزارهای نوین مانند زبان برنامه‌نویسی پایتون، همانند کشورهای پیشرفته در کشور ما نیز در حال انجام است؟ پس از بررسی پاسخ پرسش مذکور، باید به دنبال ایجاد یا بهبود مسیری با هدف تسریع در فرآیند تحلیل بازارهای مالی بر آمد و تا حد امکان به گسترش مرزهای دانش در این زمینه پرداخت. به همین علت، در طی پژوهش حاضر، این مهم در قالب طراحی و پیاده‌سازی سیستم پشتیبان تصمیم‌گیری مدیریت سبد سهام با استفاده از ابزارهای علم داده با یک رویکرد ترکیبی نوین از مطرح‌ترین الگوریتم‌های داده کاوی، هدف پژوهش بوده است؛ بنابراین، هدف اصلی از پژوهش حاضر، ادغام دانش فراگرفته‌شده در مهندسی مالی با تجربه‌ی متخصصان علم داده، به‌منظور ایجاد الگویی با بیشینه‌ی سرعت پیاده‌سازی و دریافت خروجی با کمینه‌ی خطاست؛ که برای

[1] Data science

[2] Data mining

استناد به این مقاله:

جواهری، نوید، نشاط، نجمه و جعفری ندوشن، عباسعلی. ۱۴۰۴. طراحی سیستم پشتیبان تصمیم‌گیری مدیریت سبد سهام با بهره‌گیری از روش‌های علم داده (بازار بورس اوراق بهادار تهران). مهندسی صنایع و مدیریت شریف، ۴۱(۱)، صص.۱۲-۲۷. https://doi.org/10.24200/j65.2024.63158.2375

شاپا الکترونیکی: ۲۶۷۶-۴۷۵X

تحلیلگران و سرمایه‌گذاران در بازارهای مالی که همواره بازده و ریسک را توأمان در نظر می‌گیرند و به دنبال کسب بیشترین بازده و تحمل کمترین ریسک به هنگام تشکیل سبد سهام هستند، راهگشا خواهد بود.[۴] بنابراین در پژوهش حاضر، با توجه به اهمیت مبحث مدیریت سبد سهام به‌عنوان یکی از زمینه‌های اصلی و تأثیرگذار در مهندسی مالی براساس پیشینه‌ی پژوهش‌های صورت‌پذیرفته، یک سیستم پشتیبان تصمیم‌گیری در مدیریت سبد سهام با بهره‌گیری از ابزارهای علم داده در قالب ترکیب الگوریتم‌های مطرح در این زمینه، شامل: بخش‌بندی، طبقه‌بندی، پیش‌بینی، و مدل مارکویتز ارائه شده است. در حقیقت، روشی در قالب یک سیستم پشتیبان تصمیم‌گیری[۳] در حل مسائل مدیریت سبد سهام با ترکیبی از روش‌های داده کاوی ارائه شده و این مهم در شرایطی بوده است که نیاز به تحلیل حجم زیادی از اطلاعات در ابعاد بسیار بالا در قالب کلان‌داده‌ها و اخذ خروجی در مدت زمان محدود وجود دارد. مهم آنکه رویه‌ی در نظر گرفته‌شده به گونه‌ای است که تحلیلگر قادر خواهد بود با درنظرگرفتن جامعه‌ی آماری و شاخص‌های موردنظر خود، اقدام به حل مسائل طبقه‌بندی، خوشه‌بندی، یا پیش‌بینی کند. از ویژگی‌های برجسته‌ی روش پیشنهادی، قابلیت تعمیم‌دهی آن است، زیرا امکان تنظیم پارامترهای مدل بر حسب فرض‌های مسئله‌ی موردنظر در مدل‌های نهایی فراهم شده است. در پژوهش حاضر، ابتدا مروری بر پژوهش‌های پیشین صورت گرفته و خلاصه‌ای از آن‌ها به شکل کاربردی در جهت کمک به درک جایگاه علم داده و تعیین جزییات مسئله‌ی اصلی ارائه شده است. در گام بعدی، به بررسی مبانی نظری موردنیاز پرداخته شده است، سپس داده‌های موردنیاز به‌منظور پیاده‌سازی الگوریتم‌ها با هدف ساخت مدل‌های نهایی جمع‌آوری و گام‌های پیش‌پردازش داده برای داده‌های مذکور اجرا شده است. در گام بعدی، به پیاده‌سازی و سنجش مدل‌های خوشه‌بندی، طبقه‌بندی، و پیش‌بینی پرداخته شده و مدل‌های موردنظر هدف اصلی پژوهش، ایجاد، ترکیب، و سنجیده شده‌اند. در آخرین مرحله، نیز تفسیری در قالب تحلیل خروجی مدل‌های ساخته‌شده، به‌منظور روشن‌ترشدن کاربرد سیستم ایجادشده ارائه شده است.

۲. مبانی نظری

۱.۲. مدیریت سبد سهام یا مدیریت پرتفوی

لغت پرتفوی[۴]، به بیان ساده به مجموعه‌ای از دارایی‌ها گفته می‌شود، که توسط سرمایه‌گذار برای سرمایه‌گذاری تشکیل می‌شود؛ که در آن، سرمایه‌گذار یک فرد یا یک مؤسسه است. از نظر تکنیکی، پرتفوی در برگیرنده‌ی مجموعه‌ای از دارایی‌های واقعی و مالی سرمایه‌گذاری‌شده‌ی یک سرمایه‌گذار است. پرتفویی ایده‌آل است که با فرض یکسان‌بودن بازده بین تمامی پرتفوهای موردانتظار، واریانس یا ریسک آن از همه کمتر باشد، یا از بین تمامی پرتفوهایی با واریانس یا ریسک مشابه، بیشترین بازده را داشته باشد. مارکویتز[۵] به‌صورت کمّی نشان داد که چرا و چگونه تنوع‌سازی پرتفوی می‌تواند باعث کاهش ریسک پرتفوی شود. ایشان در مسئله‌ی انتخاب پرتفوی استاندارد خود، معتقد است که سرمایه‌گذاران انتخاب‌های خود را براساس دو معیار بازدهی و ریسک انجام می‌دهند. در مدل مذکور، بازدهی پرتفو معادل میانگین وزنی بازدهی اجزاء پرتفوی و ریسک آن برابر با انحراف معیار بازدهی دارایی‌های موجود است. وزن کارای پرتفوها، مجموعه‌ای از پرتفوهاست که در هر سطح از ریسک، بیشترین بازدهی را ارائه می‌کنند. به بیان مدل مارکویتز، تشکیل یک پرتفوی متنوع، میزان ریسک را تا حد زیادی کاهش می‌دهد. فرضیه‌های مدل مارکویتز عبارت‌اند از:

۱) سرمایه‌گذاران ریسک‌گریز هستند و مطلوبیت موردانتظار، افزایشی است؛ ۲) سرمایه‌گذاران پرتفو را براساس میانگین و واریانس موردانتظار عایدی انتخاب می‌کنند؛ ۳) هر سرمایه‌گذاری تا بی‌نهایت قابل تقسیم است؛ و ۴) سرمایه‌گذاران افق زمانی یک دوره‌ای دارند و این دوره برای همه مشابه است.[۴] بنابراین، می‌توان گفت بازده یک سبد، بازده وزنی سهام پایه است. براساس مدل مارکویتز، اگر σ_p ریسک پرتفوی و n تعداد سهام پایه باشد، آنگاه رابطه‌ی ۱ برقرار است:

$$\sigma_p^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij} \qquad (۱)$$

که در آن، σij برابر با کوواریانس بین قیمت سهم i و j و همچنین Wi و Wj وزن تخصیص داده‌شده به سهم i و j هستند. وزن تخصیص داده‌شده در تابع ۱، در حقیقت وزن هر سهم در خوشه است، که به صورت رابطه‌ی ۲ محاسبه می‌شود:

$$W_j = \frac{M_j}{\sum M_j} \qquad (۲)$$

که در آن، Mj ارزش بازار سهم j است. به بیان ساده‌تر، بازدهی هر خوشه برابر با میانگین موزون بازدهی سهام موجود در آن خوشه است، که وزن هر سهم، ارزش بازار آن سهم در آن خوشه است.[۴]

۲.۲. علم داده و داده کاوی

علم داده مبتنی بر تحلیل داده‌ها و شامل علوم گوناگونی همچون تحلیل کلان داده‌ها، داده کاوی و ... است. همچنین به زیردسته‌هایی از جمله: هوش مصنوعی (AI)[۶]، یادگیری ماشین (ML)[۷]، داده کاوی، یادگیری عمیق[۸]، و ... تقسیم‌بندی می‌شود؛ همچنین داده کاوی به عنوان فرآیند کشف دانش[۹] یا فرآیند کشف داده[۱۰] نیز شناخته می‌شود؛ فرآیندی است که مستلزم مطالعه و تجزیه‌وتحلیل داده‌ها از منابع مختلف، ارزیابی و ترکیب آن‌ها به اطلاعات مفیدتر و مهم‌تر است. داده کاوی به‌منظور پیش‌بینی، محبوب‌ترین نوع داده کاوی با بیشترین خروجی برای فعالیت‌های تجاری است.[۲] روش‌های اصلی داده کاوی در پژوهشی به ۶ گروه اصلی تقسیم‌بندی شده‌اند، که عبارت‌اند از: ۱) طبقه‌بندی، ۲) پیش‌بینی، ۳) خوشه‌بندی، ۴) تجزیه‌وتحلیل توالی، ۵) تشخیص داده‌های پرت، و ۶) متن‌کاوی؛ که از میان آن‌ها، خوشه‌بندی از جمله روش‌هایی است که به‌طورگسترده و فشرده توسط بسیاری از پژوهشگران داده کاوی مطالعه شده است.[۵] روش‌های داده کاوی در پژوهش حاضر نیز به ۳ دسته‌ی کلی: ۱) بخش‌بندی یا خوشه‌بندی، ۲) طبقه‌بندی، و ۳) پیش‌بینی تقسیم‌بندی شده‌اند. داده کاوی عمدتاً برای اهداف تجاری استفاده و در قالب ارائه‌ی روش مطرح می‌شود. خروجی آن به‌عنوان یک الگو قابل استخراج است و نوع داده‌ها در آن عموماً ساختاریافته است؛ این در حالی است که علم داده، عمدتاً برای اهداف علمی استفاده می‌شود و همواره به عنوان یک رشته مطرح است. نوع داده‌ها در علم داده می‌تواند به شکل ساختاریافته، نیمه‌ساختاریافته، و یا بدون

[3] Decision support system
[4] Portfolio
[5] Markowitz
[6] Artificial Intelligence
[7] Machine Learning
[8] Deep learning
[9] knowledge discovery
[10] Data discovery

ساختار باشد.[۶] زرندی و کاظمی (۲۰۰۸)، [۷] فرآیند کشف دانش از طریق داده کاوی را به ۴ عنوان تقسیم‌بندی کرده‌اند:

۱) انتخاب، ۲) پیش‌پردازش، ۳) داده کاوی، و ۴.) تفسیر. دستگیر (۲۰۱۹)،[۹] نیز مراحل اخیر را به ۹ مرحله تقسیم‌بندی کرده است: ۱) شناسایی هدف، ۲) انتخاب داده‌ها، ۳) آماده‌سازی داده‌ها، ۴) ارزیابی داده‌ها، ۵) قالب‌بندی پاسخ، ۶) انتخاب ابزار، ۷) الگوسازی، ۸) اعتبارسازی یافته‌ها، و ۹) ارائه‌ی نتایج.[۸] با توجه به مرور ادبیات و با هدف تلخیص و ترکیب ادبیات مروری گوناگون و با بررسی تجربه‌ی خبرگان در این زمینه، در پژوهش حاضر، مراحل داده کاوی در قالب ۵ مفهوم مستقل با توجه به مباحث کتاب راهنمای علم داده‌ی پایتون[۱۱] بررسی شده است.

۳.۲. سنجش خطای مدل‌ها

مرحله‌ی کنونی، شامل سنجش مدل‌ها و الگوهای ساخته‌شده است، که به وسیله‌ی شاخص‌های اندازه‌گیری خطا، همچون مجذور میانگین مربعات خطا (RMSE)[۱۲]، میانگین قدرمطلق خطا (MAE)[۱۳]، ضریب تعیین[۱۴]، میانیگن مربعات خطا (MSE)،[۱۵] و ماتریس اغتشاش[۱۶] الگوهای ساخته‌شده ارزیابی شده‌اند. در ادامه، توابع شاخص‌های مذکور به جهت یادآوری، در روابط ۳ الی ۵ ارائه شده‌اند:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}[y_t - \hat{y}] \quad (۳)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_t - \hat{y})^2, RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_t - \hat{y})^2} \quad (۴)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_t - \hat{y})^2}{\sum_{i=1}^{N}(y_t - \bar{y})^2} \quad (۵)$$

که در آن‌ها، $\hat{y}$ برابر با مقدار پیش‌بینی‌شده از y یا مقادیر ستون هدف و $\overline{y}$ برابر با میانگین مقادیر ستون هدف یا مقادیری است، که قرار بر پیش‌بینی آن‌هاست. همچنین ماتریس آشفتگی نیز که در شکل ۱ مشاهده می‌شود، چهار معیار موردنظر خود را در قالب چهار نسبت از طریق مقایسه‌ی برچسب طبقه‌های پیش‌بینی‌شده با برچسب طبقه‌های حقیقی محاسبه می‌کند، که در پژوهش حاضر از معیار دقت[۱۷]، که ترکیبی از سه معیار دیگر است، استفاده شده است.

شکل ۱. ماتریس آشفتگی.

در ماتریس آشفتگی، مقدار ضریب خطای معیار دقت مطابق رابطه‌ی ۶ محاسبه می‌شود:

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

۴.۲. الگوریتم درخت تصمیم[۱۸]

ساختار درخت تصمیم شبیه به ساختار یک فلوچارت است. درخت تصمیم، یک شیوه‌ی طبقه‌بندی یا پیش‌بینی است، که بالاترین گره در درخت، ریشه است و برگ‌ها، نشان‌دهنده‌ی دسته‌ها و توزیع دسته‌ها هستند. هر آزمون آزمایشی بر یک گره، یک ویژگی را مشخص می‌کند و هر شاخه‌ای که از آن گره خارج شود، نتیجه‌ی آزمون اخیر را نشان می‌دهد.[۱۰] در روش مذکور برخلاف شبکه‌های عصبی، الزامی برای عددی‌بودن داده‌ها در درخت تصمیم وجود ندارد.[۱۱]

۵.۲. نایو بیز[۱۹]

نایو بیز، شیوه و روشی از خانواده‌ی روش‌های طبقه‌بندی برمبنای احتمال به‌کارگیری قضیه‌ی بیز و فرض استقلال بین متغیرهاست. اگر مشاهده‌ها و داده‌ها از نوع پیوسته باشند، از مدل احتمالی با توزیع گاوسی یا نرمال برای متغیر داده‌های جمع‌آوری‌شده می‌توان استفاده کرد.[۱۲]

۶.۲. نزدیک‌ترین همسایه[۲۰]

روش نزدیک‌ترین همسایه، به‌عنوان یکی از ساده‌ترین روش‌های طبقه‌بندی به‌عنوان طبقه‌بندی‌کننده‌ی تنبل[۲۱] نیز شناخته می‌شود، زیرا مرحله‌ای خاص جهت یادگیری ندارد و به‌هنگام طبقه‌بندی از تمامی داده‌ها برای یادگیری استفاده می‌کند. روش مذکور متعلق به کلاس یادگیری مبتنی بر نمونه است و در آن، طبقه‌بندی فقط با نگاه‌کردن به K نزدیک‌ترین مثال‌ها در مجموعه داده‌های آموزشی (عموماً از نظر معیار فاصله‌ی اقلیدسی[۲۲] یا براساس انوع دیگری از معیارهای فاصله همچون همینگ[۲۳] یا منهتن[۲۴]) در مورد داده‌ای که طبقه‌بندی برای آن در حال پیاده‌سازی است، انجام می‌شود؛ سپس با توجه به نمونه‌های مشابه K، محبوب‌ترین هدف (براساس رأی بیشینه) به‌عنوان برچسب طبقه‌بندی انتخاب می‌شود.[۱۳]

۷.۲. ماشین بردار پشتیبان[۲۵]

یک روش یادگیری با نظارت[۲۶] است، که برای طبقه‌بندی استفاده می‌شود؛ و مشابه شبکه‌های عصبی قادر است تقریبی با درجه‌ی دقت مطلوب برای هر تابع چندمتغیره به‌دست آورد. بنابراین به‌منظور مدل‌سازی سیستم‌ها و فرآیندهای غیرخطی و پیچیده، از جمله شناسایی فعالیت‌های مرتبط با تقلب مالی، بسیار مفید است.[۹]

۸.۲. پرسپترون چندلایه[۲۷]

به‌منظور حل بسیاری از مسائل ریاضی، که براساس معادله‌های پیچیده‌ی غیرخطی حل می‌شوند، یک شبکه‌ی عصبی پرسپترون چندلایه می‌تواند به

[11] Python Data Science Handbook [Book]
[12] Root Mean Square Error
[13] Mean Absolute Error
[14] R^2
[15] Mean Squared Error
[16] Confusion Matrix
[17] Accuracy
[18] Decision tree
[19] Naive Bayes
[20] Nearest Neighbor
[21] Lazy Classifiers
[22] Euclidean
[23] Hamming
[24] Manhattan
[25] Support vector machines
[26] Supervised Learning
[27] Multiluyer perceptron

سادگی با تعریف اوزان و توابع مناسب استفاده شود. یک پرسپترون چندلایه، از سه نوع لایه تشکیل می‌شود: ۱) لایه‌ی ورودی[۲۸]، ۲) لایه‌های میانی یا پنهان[۲۹]، و ۳) لایه‌ی خروجی[۳۰]. نرون‌ها عناصر تشکیل‌دهنده‌ی لایه‌ها در شبکه‌های عصبی هستند. عناصر هر لایه با تمام عناصر لایه‌های دیگر در ارتباط است، ولی با دیگر عناصر همان لایه ارتباطی ندارد. موضوع اصلی موردبحث در شبکه‌ها، تعیین تعداد لایه‌های پنهان و تعداد نورون‌های هر لایه است. در مدل مذکور، ورودی‌هایی وارد نورون می‌شوند که با x_i و به شکل خلاصه با بردار x نمایش داده می‌شوند. هر یک از ورودی‌های نورون به یکی از سیگنال‌های ورودی متعلق است، که هر سیگنال در وزن متناظر w_i ضرب و مقادیر حاصل در داخل نورون با هم جمع و مقدار خروجی (مطابق رابطه‌ی ۷) محاسبه می‌شود:[۱۴]

$$NET = X_1W_1 + X_2W_2 + \ldots + X_nW_n = \sum X_i.W_i \qquad (۷)$$

۹.۲. نسبت‌های مالی

نسبت‌های مالی، مقادیر عددی هستند که به‌منظور بررسی وضعیت شرکت‌ها از صورت‌های مالی آن‌ها، مانند: صورت سود و زیان، ترازنامه، صورت جریان نقدینگی، و ... استخراج می‌شوند. نسبت‌های مالی به بیان روابط بین اقلام موجود در صورت‌های مالی شرکت‌ها می‌پردازند و همواره به‌عنوان یکی از متداول‌ترین گزینه‌ها برای تحلیل‌های مالی در نظر گرفته می‌شوند. در پژوهش حاضر، از بیان مبانی نظری مربوط به نسبت‌های مالی، به‌منظور جلوگیری از تکرار بیان مباحث خودداری شده است.

۱۰.۲. بازده سهام

اگر اطلاعات مربوط به جریان‌های نقدینگی (سود نقدی، مزایای حق‌تقدم، و سهام‌جایزه) در دسترس نباشند و در تابع اخیر، ارقام مذکور وارد نشود، خروجی کسر عبارت از بازده قیمت سهام است، که در پژوهش حاضر با توجه به کمبود اطلاعات، این مقدار محاسبه شده است. به عبارتی ساده‌تر، بازده کل سهام برابر با مجموع بازده نقدی و بازده قیمت است.[۱۵]

حال این تذکر لازم است که سؤال و هدف ابتدایی پژوهش حاضر آن است که چه میزان می‌توان به بهبود درک جایگاه الگوریتم‌های داده کاوی با دیدگاه مالی در بستر علم داده، ضمن مطالعه‌ی پژوهش‌های پیشین پرداخت؛ که به این منظور، به مرور دقیق پژوهش‌های پیشین پرداخته شده است. در پایان نیز شکافی در قالب سؤال اصلی پژوهش شناسایی شده است، که عبارت از نحوه‌ی ارائه‌ی سیستم پشتیبان تصمیم‌گیری در تشکیل سبد سرمایه‌گذاری، شامل سهام شرکت‌های فعال در بازار بورس اوراق بهادار تهران با توجه به بازده قیمت هر سهم و هر خوشه با بهره‌گیری از الگوریتم‌های داده کاوی ضمن ترکیب با مدل مارکویتز است؛ که در همین راستا، چالشی با هدف نحوه‌ی به‌کارگیری الگوریتم‌های خوشه‌بندی، طبقه‌بندی، و پیش‌بینی برای پژوهشگران پژوهش حاضر در مسیر تشکیل سبد سهام ایجاد شده است، که طی مسیر پژوهش، سؤال‌های ذکرشده، بررسی و پیاده‌سازی شده‌اند.

۳. پیشینه‌ی پژوهش

۱.۳. پیشینه‌ی پژوهش‌های داخلی

در زمینه‌ی خوشه‌بندی شرکت‌های فعال در بازار سهام، عباسی و نیکبخت (۲۰۱۸)،[۱۶] با استفاده از داده‌های موجود در صورت‌های مالی شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران، با استفاده از روش خوشه‌بندی C-means به بخش‌بندی شرکت‌های مذکور پرداخته و نشان داده‌اند که بخش بزرگی از آن‌ها در سبد سهام ترکیبی قرار گرفته‌اند، در حالی که تمایل رفتاری آن‌ها به سهام رشدی بوده است. در پژوهش اخیر، از طریق مطالعه‌ی کتب و نوشتارهای بین‌المللی و درنهایت مشورت با پژوهشگران، شاخص‌های انتخابی برای بخش‌بندی سهام شرکت‌ها انتخاب شده‌اند، که عبارت‌اند از: نسبت قیمت به درآمد هر سهم (P/E)[۳۱]، نسبت قیمت به ارزش دفتری (P/B)[۳۲]، شاخص ریسک سیستماتیک (β)[۳۳]، سود هر سهم (EPS)[۳۴]، سود تقسیمی، مومنتوم[۳۵]، نرخ رشد ۵ ساله‌ی درآمد هر سهم، میزان تغییر ۵ ساله نسبت قیمت به درآمد هر سهم، بازده حقوق صاحبان سهام (ROE)[۳۶]، نسبت بدهی به حقوق صاحبان سهام (D/E)[۳۷]، نسبت قیمت سهام به فروش (P/S)[۳۸]. سلطانی‌نژاد (۲۰۱۵)،[۱۷] نیز به دنبال حذف نویز اطلاعات ماتریس ضرایب همبستگی از طریق روش‌های خوشه‌بندی به‌منظور بهینه‌سازی سبد سرمایه‌گذاری بوده است؛ به این منظور از دو روش خوشه‌بندی اتصال واحد و اتصال میانگین و براساس بازده روزانه‌ی ۸۰ شرکت بورس اوراق بهادار تهران در بازه‌ی زمانی ۱۳۸۵ تا اسفند ۱۳۹۲ استفاده کرده و دریافته است که درنهایت، ریسک کمتری به سرمایه‌گذار تحمیل می‌شود و همچنین با هدف تحلیل حساسیت نتایج دو روش، یک بررسی بوت استرپینگ[۳۹] در تعداد سهام و افق‌های زمانی مختلف انجام شده است، که در بیشتر حالت‌ها، بهبود عملکرد سبد سرمایه‌گذاری مشاهده شده است. همچنین صیادی و امیدی (۲۰۲۰)،[۱۸] ترکیبی از روش‌های خوشه‌بندی و پیش‌بینی را استفاده کرده‌اند، به صورتی که یک مدل بهینه‌سازی پرتفوی مبتنی بر پیش‌بینی، برای بررسی نتایج وضعیت پرتفوی در بورس اوراق بهادار تهران ایجاد شده است. بدین منظور، ابتدا از روش داده کاوی برای پیش‌بینی فرآورده‌های نفتی و صنایع شیمیایی با استفاده از داده‌های خوشه‌بندی بازار سهام استفاده شده است. در گامی دیگر، نیز به منظور تخمین هر شاخص صنعت با استفاده از تابع پایه‌ی شعاعی و شبکه‌ی عصبی پرسپترون چندلایه، از عوامل مؤثری مانند: قیمت نفت خام، نرخ ارز، نرخ بهره جهانی، قیمت طلا، و شاخص S&P ۵۰۰ استفاده شده است. درنهایت، با مقایسه‌ی نسبت‌های اعتبارسنجی در یک بازار صعودی با استفاده از الگوریتم‌های خوشه‌بندی K-means، شبکه‌ی عصبی و فازی، بهترین الگوریتم برای پیش‌بینی شاخص‌ها برای هر صنعت شناسایی شده و نتایج نشان داده است که الگوریتم پرسپترون چندلایه، بالاترین دقت را داشته و بهترین

---

[۲۸] Input Layer
[۲۹] Hidden Layer
[۳۰] Output Layer
[۳۱] Price to Earning
[۳۲] Price to Book value
[۳۳] Systematic risk
[۳۴] Earning Per Share

[۳۵] در بازار بدین معناست که یک روند قیمتی تمایل دارد باقی بماند تا زمانی که یک نیروی خارجی مانع آن شود. در پژوهش مذکور شاخص مومنتوم درصد تغییر قیمت در ۱۲ ماه سال موردنظر پژوهش است.

[۳۶] Return On Equity
[۳۷] Debt to Equitu
[۳۸] Price to Sales ratio
[۳۹] Spring Boot

گزینه برای وضعیت پرتفوی بوده است؛ با این حال، الگوریتم فازی C-means بهترین خوشه‌ها را تولید کرده است. نتایج عملی نشان داد که سهام نفت سپاهان و پتروشیمی خارگ، مهم‌ترین سهام در کوتاه‌مدت بوده و سهام پتروشیمی خارگ، پتروشیمی فن‌آوران، و پالایش نفت تهران سهم بیشتری در سبد سهام در میان‌مدت یا بلندمدت داشته‌اند. زمانی و همکارش (۲۰۲۳)، پس از مطالعه‌ی نوشتارهای بین‌المللی و بررسی شاخص‌های گوناگون و مشورت با اساتید و صاحب نظران در زمینه‌های مالی و حسابداری، ۵ شاخص نهایی برای خوشه‌بندی انتخاب کرده‌اند، که عبارت‌اند از: ۱) نسبت قیمت به درآمد هر سهم، ۲) شاخص ریسک سیستماتیک، ۳) سود هر سهم، ۴) بازده[40]، و ۵) نسبت ارزش بازار به ارزش دفتری (M/B)[41]، که به منظور قابل انجام‌بودن عمل خوشه‌بندی و محاسبه‌ی فاصله‌ی درست میان سهام، داده‌ها را براساس شاخص‌های اصلی نسبت قیمت به درآمد هر سهم و نسبت ارزش بازار به ارزش دفتری فیلتر کرده و فقط سهام حاوی ویژگی‌های اخیر برای خوشه‌بندی را در نظر گرفته‌اند.[۴] در سایر پژوهش‌ها در این زمینه نیز می‌توان به پژوهش صادقی آرانی و همکارش (۲۰۱۹) اشاره کرد، که در آن از ۵ نسبت مالی آلتمن، که از سایت مرکز بورس اوراق بهادار تهران استخراج شده‌اند، به این شرح استفاده شده است: ۱) نسبت سرمایه در گردش به کل دارایی‌ها، ۲) نسبت سود قبل از مالیات به کل دارایی‌ها، ۳) نسبت سود انباشته به کل دارایی‌ها، ۴) نسبت ارزش بازار سهام به ارزش دفتری کل بدهی‌ها، و ۵) نسبت فروش کل به کل دارایی‌ها. نتیجه‌ی نهایی بیان کرده است که روش‌های فرا ابتکاری در مقایسه با روش‌های معمول، کاراتر عمل کرده و به بهینه‌ی سراسری منجر شده‌اند؛ همچنین نتایج به‌دست‌آمده با نتایج حاصل از تفکیک شرکت‌های عضو بورس بهادار تهران با استفاده از روش تعیین ورشکستگی آلتمن مقایسه و توسط روش مذکور نیز تأیید شده است.[۱۹] دستگیر و شفیعی (۲۰۱۹)، پیش‌بینی ورشکستگی را مشهورترین موضوع در زمینه‌ی کاربرد روش‌های داده کاوی در حل مسائل مالی دانسته و از بین روش‌های موجود در داده کاوی نیز شبکه‌ی عصبی بیشترین استفاده را به خود اختصاص داده است.[۸] محمودی و همکاران (۲۰۲۰)، عملکرد پنج مدل محبوب آماری و یادگیری ماشین را با هدف تشخیص تقلب صورت‌های مالی در مقایسه با یکدیگر بررسی کرده‌اند. اهداف پژوهش اخیر شرکت‌هایی بوده‌اند که بین سال‌های ۲۰۱۱ و ۲۰۱۶ صورت‌های مالی متقلبانه و غیرمتقلبانه را تجربه کرده‌اند. نتایج ایشان نشان داده است که شبکه‌ی عصبی مصنوعی[42] نسبت به شبکه‌ی بیزی[43]، تحلیل تشخیصی[44]، رگرسیون لجستیک[45]، و ماشین‌بردار پشتیبان عملکرد بهتری داشته‌اند. هدف ایشان، ارائه‌ی یک مدل کمّی برای شناسایی گزارشگری مالی متقلبانه در شرکت‌های ایرانی با استفاده از الگوریتم‌های طبقه‌بندی و شناسایی شرکت‌هایی بوده است، که برای پنهان‌کردن اطلاعات و یا ارائه‌ی اطلاعات نادرست در پرونده‌های سالانه با اوراق بهادار و بورس تهران تلاش کرده‌اند. به بیان پژوهش اخیر، از میان ۱۹ عامل پیش‌بینی‌کننده‌ی بررسی‌شده، فقط ۹ عامل پیش‌بینی‌کننده به‌طورپیوسته توسط الگوریتم‌های طبقه‌بندی مختلف انتخاب و استفاده شده‌اند، که عبارت‌اند از: ۱) بهره‌وری کارکنان، ۲) حساب‌های دریافتی به فروش، ۳) بدهی به حقوق صاحبان سهام، ۴) موجودی به فروش، ۵) فروش به کل دارایی‌ها، ۶) بازده حقوق صاحبان سهام، ۷) بازده فروش، ۸) بدهی‌ها به هزینه‌های بهره، و ۹) دارایی‌ها به بدهی‌ها. یافته‌های پژوهش اخیر، تقلب در صورت‌های مالی را گسترش داده و توانسته است توسط متخصصان و تنظیم‌کننده‌ها برای بهبود مدل‌های ریسک تقلب استفاده شود. نرم‌افزار استفاده‌شده‌ی ایشان، IBM SPSS Modeler ۱۸ بوده است.[۱۴]

نتایج حاصل از پژوهش عباسی و همکاران (۲۰۱۹) نیز نشان داده است که توانگری مالی با دقت قابل‌قبولی، پیش‌بینی‌پذیر بوده و مدل استخراج‌شده با استفاده از درخت تصمیم، دقت و قابلیت بسیار بالایی در تخمین داشته است. از میان ۴ مدلی که در پژوهش اخیر استفاده شده‌اند (درخت تصمیم، نایو بیز، شبکه‌ی عصبی، و نزدیک‌ترین همسایه)، درخت تصمیم بالاترین دقت و کمترین مقدار خطا را داشته است. همچنین متغیرهای تصمیم به‌خوبی متغیر هدف را تعیین کرده‌اند. بدترین عملکرد نیز مربوط به مدل نایو بیز شناسایی شده است.[۲۰]

۲.۳. پیشینه‌ی پژوهش‌های خارجی

ناندا[46] و همکاران (۲۰۱۰)، در نوشتاری با عنوان «خوشه‌بندی داده‌های بازار سهام هند جهت مدیریت پرتفو» از داده کاوی برای طبقه‌بندی سهام در خوشه‌های مختلف استفاده کرده و پس از طبقه‌بندی، سهام موردنظر برای پرتفوها از بین خوشه‌های مذکور انتخاب شده است. روش مذکور منجر به کاهش ریسک متنوع‌سازی سبد سهام شده و نتایج نشان داده است که روش تحلیلی K-means، خوشه‌های متراکم‌تری نسبت به روش‌های SOM[47] و فازی ایجاد می‌کند. شاخص‌های موردنظر برای خوشه‌بندی در پژوهش مذکور عبارت بوده‌اند از: نسبت قیمت به درآمد هر سهم، نسبت قیمت به ارزش دفتری، نسبت قیمت به سود نقدی هر سهم (P/C EPS)[48]، ارزش شرکت به سود شاخص عملکرد با توجه به سرمایه بازار (EV/ EBIDTA)[49].[۲۰] حال مهم آن است که اگرچه روبرویی با کلان‌داده‌ها دانش بیشتری را نیازمند است، ولی مزایایی نیز دارد. رازین[50] (۲۰۱۵)، در نوشتار خود به این نتیجه رسیده است که "کلان‌داده‌ها از پنج طریق باعث تغییر امور مالی می‌شوند: ۱) ایجاد شفافیت، ۲) تجزیه‌وتحلیل ریسک، ۳) تجارت الگوریتمی، ۴) استفاده از داده‌های مصرف‌کننده، و ۵) تغییر فرهنگ.[۲۱] هاری‌هاران[51] (۲۰۱۸)، در پژوهش خود بر روی استفاده از داده کاوی در پیش‌بینی سهام، مدیریت پرتفوی، تحلیل ریسک سرمایه‌گذاری، همچنین پیش‌بینی ورشکستگی و نرخ ارز خارجی و شناسایی تقلب مالی متمرکز شده است.[۳] دایبولد[52] و همکاران (۲۰۱۹) بیان کرده‌اند که اگرچه فضای اقتصادی مدرن امروز، غرق در داده‌های بزرگ است، با این حال تحلیل یا اصلاً بررسی نشده‌اند؛ به همین منظور ایشان در پژوهش خود، پژوهش‌های جدید اقتصادسنجی را در مورد داده‌های بزرگ و مدل‌سازی سری‌های زمانی ارائه کرده‌اند. به شکل خلاصه آن‌ها معتقد بوده‌اند که کلان‌داده در تحلیل‌های اقتصادی و مدل‌سازی اقتصادی تأثیر بسزایی دارد.[۲۲] مجومدار

---

[40] Return
[41] Market value/Book value
[42] Artificial Neural Network
[43] Bayesian network
[44] Discriminant Analysis
[45] logistic
[46] Nanda
[47] نگاشت خود سازمانده
[48] Price to Cash EPS
[49] Enterprise Value to Earnings Before Interest, Taxes, Depreciation and Amortization
[50] Razin
[51] Hariharan
[52] Diebold

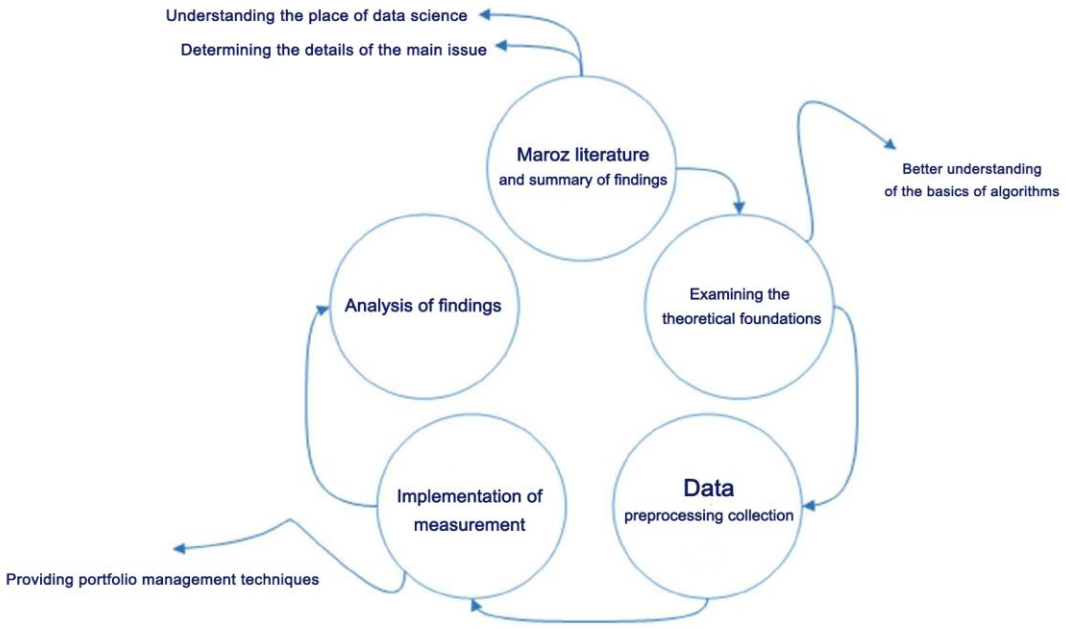
و لاها [53] (۲۰۲۰)، نیز به خوشه‌بندی و طبقه‌بندی سری‌های زمانی با استفاده از تحلیل داده‌های توپولوژیکی با کاربردهای مالی پرداخته‌اند.[۲۳] گودل[54] و همکاران (۲۰۲۱)، در بازنگری کاملی بر نوشتارهای ارائه‌شده در امور مالی دریافته‌اند که مباحثی در هر دو گروه هوش مصنوعی و یادگیری ماشین در ۳ دسته مطرح هستند: ۱) ساخت پرتفوی، ارزش‌گذاری و رفتار سرمایه‌گذار، ۲) تقلب و ناتوانی مالی، ۳) پیش‌بینی و برنامه‌ریزی. در پژوهش مذکور نوشتارهایی معرفی شده‌اند که به بررسی شناسایی تقلب مالی با استفاده از روش‌های داده کاوی پرداخته‌اند، که با توجه به تعدد آن‌ها، از بیان جزئیات‌شان در مرور پیشینه‌ی حاضر خودداری شده است: در گام اول، معماری یک سیستم مدیریت مالی هوشمند کاملاً بررسی شده و معماری جدیدی از یک سیستم پشتیبانی مدیریت مالی هوشمند مبتنی بر داده کاوی توسعه یافته است؛ در گام دوم، به تعریف و ساختار انبار داده و داده کاوی و همچنین نحوه‌ی استفاده از راهبرد و فناوری داده کاوی در مدیریت مالی پرداخته شده است. به عقیده‌ی ایشان، داده‌کاوی در ارتباط با فناوری و توسعه‌ی یک الگوریتم داده کاوی هوشمند در حال بررسی است؛ نقص‌های الگوریتم داده کاوی هوشمند از طریق تجزیه‌وتحلیل و خلاصه‌سازی الگوریتم کشف و درنهایت یک الگوریتم بهبودیافته برای رفع ایرادها پیشنهاد شده است. در گام بعدی، آزمایش‌های مرتبط بر روی الگوریتم بهبودیافته انجام شده و آزمایش‌ها نشان داده‌اند که الگوریتم مذکور مزایای خاصی داشته‌اند. سپس چهارچوب طراحی اولیه برای مدیریت مالی هوشمند ایجاد و کاربرد یک مدل داده کاوی در سیستم پشتیبانی تصمیم معرفی شده است.[۲۵] همچنین، کیانو[55] (۲۰۲۴) در پژوهش خود بیان کرده است که پیشرفت‌های تکنولوژیکی منجر به دیجیتالی‌شدن اقتصاد شده است، از همین رو یک مدل یادگیری ماشین با استفاده از شبکه‌های عصبی برای پیش‌بینی قیمت بیت‌کوین ارائه کرده است، که در مسیر آن، اطلاعات قیمت پایانی ۶۰ روز رمز ارز بیت‌کوین به‌منظور پیش‌بینی روز ۶۱ اُم جمع‌آوری شده است، نتیجه آنکه انجام یک مطالعه‌ی عمیق در مورد همبستگی بین ارزهای رمزنگاری‌شده، تجزیه‌وتحلیل را برای هدایت خواندن به این چارچوب مرتبط با پیش‌بینی برای یک سناریوی پیچیده‌ی مالی تقویت می‌کند.[۳۰] دودک[56] و همکاران (۲۰۲۴)، در نوشتار خود، ۱۲ روش پیش‌بینی نوسان‌های ارزهای دیجیتال را به شکل جامع مطالعه کرده و دریافته‌اند که بهترین روش واحد برای پیش‌بینی یک ارز دیجیتال وجود ندارد و مدل‌های مختلف بسته به ارز دیجیتال خاص، انتخاب متریک خطا، و افق پیش‌بینی، عملکرد بهتری دارند؛ ولی مدل‌های خطی ساده نیز می‌توانند به‌خوبی مدل‌های پیچیده عمل کنند.[۳۱]

با توجه به خلاصه‌ی ارائه‌شده‌ی پژوهش‌های مروری توسط نویسندگان پژوهش حاضر، روند کلی مدل‌سازی و شاخص‌های موردنیاز در پژوهش حاضر شناسایی شده است، که موارد مذکور در بخش‌های بعدی مطرح شده‌اند.

۴. روش‌شناسی پژوهش

فرآیند کلی اجرای پژوهش در شکل ۲، جهت درک بهتر مسیر، مشاهده می‌شود.

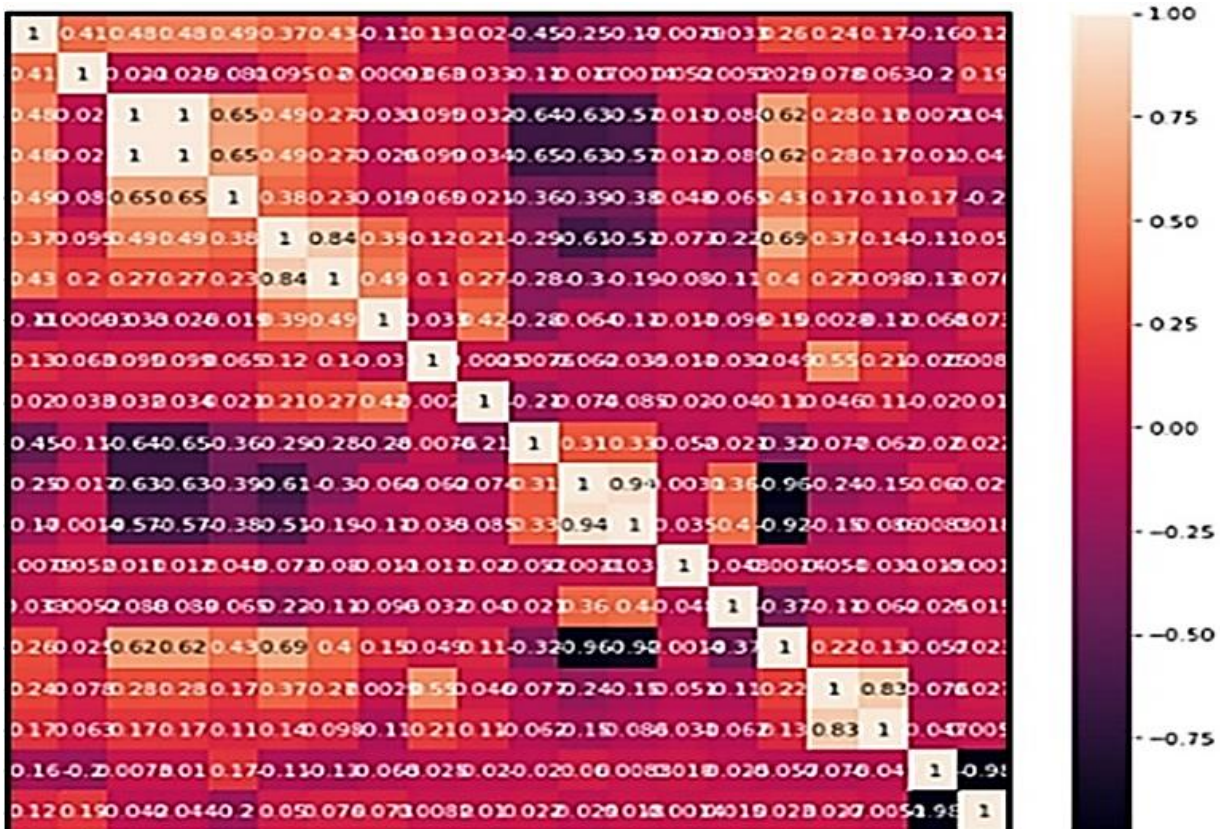
در مرحله‌ی جمع‌آوری داده‌ها به‌عنوان اولین گام اجرایی، داده‌های مربوط به ۱۸ نسبت مالی شرکت‌های فعال در بازار بورس اوراق بهادار تهران تهیه شده است؛ که از میان آن‌ها، ۹ نسبت از مرور ادبیات پژوهش استخراج شده است. این تذکر لازم است که از میان تمامی نسبت‌های مالی شناسایی‌شده در مرور ادبیات، با توجه به عدم امکان استخراج داده و عدم امکان محاسبه‌ی نسبت‌ها برای تعداد بالای شرکت‌های موردنظر در بازه‌ی زمانی انجام پژوهش (بدون بهره‌گیری از نرم‌افزارهای بانک استخراج داده)، ۹ نسبت مالی نهایی مذکور گزینش شده‌اند، و ۹ نسبت دیگر نیز به جهت برخورداری پژوهش حاضر از نوآوری لازم نسبت به پژوهش‌های پیشین، در پژوهش حاضر در نظر گرفته شده‌اند. نسبت‌های مذکور در حقیقت نسبت‌هایی هستند که همواره در منابع آموزشی علمی مرورشده توسط پژوهشگران، به چشم می‌خورند. تعداد



شکل ۲. نمودار فرآیند اجرای پژوهش.

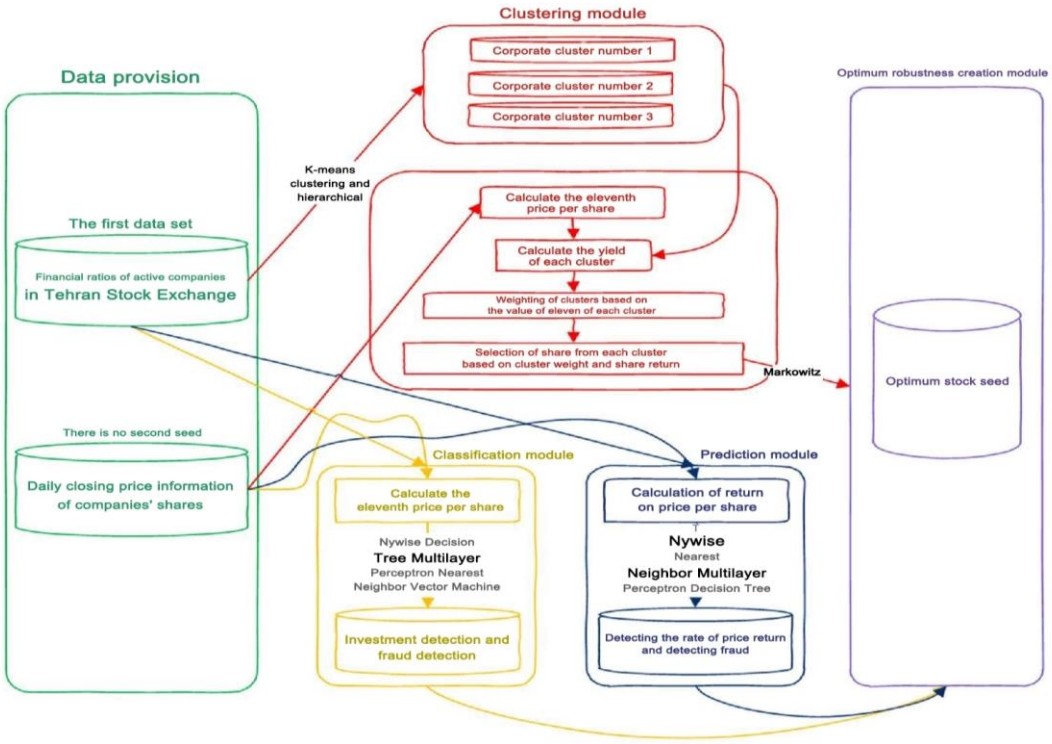
[53] Majumdar, & Laha
[54] Goodell
[55] Ciano
[56] Dudek

شرکت‌های موردنظر پژوهش حاضر نیز با توجه به اینکه ۸۹۲ نماد فعال در قالب ۵۵۸ شرکت در بازار بورس تهران موجود است، با استفاده از وبسایت "شرکت مدیریت فناوری بورس تهران ۴۱" اسامی ۸۹۲ نماد بازار بورس اوراق بهادار تهران در قالب "فهرست همه‌ی نمادهای بازار عادی" استخراج شده است، ضمن بهره‌گیری از جدول مورگان فریمن، تعداد نمونه‌ی موردنیاز، ۲۲۶/۴۴ شرکت محاسبه شده است، که با توجه به کمینه‌ی تعداد به‌دست‌آمده و پس از بررسی داده‌های خام در مرحله‌ی پیش‌پردازش، ۲۷۱ شرکت در پژوهش حاضر بررسی شده‌اند. با توجه به توضیحات اخیر، نسبت‌های موردنظر پژوهش حاضر به این شرح هستند: سود انباشته به دارایی‌ها، سود قبل از بهره و مالیات به دارایی، گردش دارایی‌ها،[۱۹] بازده دارایی‌ها،[۴ و ۲۶-۲۷] بدهی به حقوق صاحبان سهام،[۱۶] حاشیه‌ی سود خالص و نسبت بدهی،[۲۶] گردش دارایی‌های ثابت و نسبت جاری،[۲۷] و سایر نسبت‌های مالی مطرح، شامل: حاشیه‌ی سود عملیاتی، حاشیه‌ی سود قبل از مالیات، گردش حساب‌های دریافتی، دوره‌ی وصول مطالبات، جمع بدهی‌ها به جمع دارایی‌ها، نسبت بدهی‌های بلندمدت، نسبت نقدینگی، نسبت جریان نقد سرمایه‌گذاری به درآمد، نسبت جریان نقد تأمین مالی به درآمد؛ نسبت‌های مالی مذکور از طریق نرم‌افزار "بورس ویو[۵۷]" در سال مالی ۱۳۹۹ استخراج شده‌اند، که درخصوص این دسته داده‌ها، داده‌ی استخراجی شامل ۱۸ نسبت مالی در ستون و ۲۷۱ عنوان شرکت در سطر بوده است. به‌منظور جمع‌آوری داده‌ها، ابتدا اطلاعات ۴۹۹ شرکت با سال مالی منتهی به برج‌های متفاوت استخراج شده‌اند. سپس نسبت‌های موجود برای ۳۰۶ شرکت با سال مالی منتهی به برج ۱۲ سال ۱۳۹۹ و شامل ۳۷ عنوان نسبت مالی گردآوری شده‌اند. درنهایت، با توجه به فقدان اطلاعات کامل برای شرکت‌ها و یا نسبت‌های مالی، داده‌ایی شامل ۲۷۶ شرکت و ۱۸ نسبت مالی فراهم شده‌اند، که پس از گردآوری دسته‌ی دوم داده‌ها نیز ۵ شرکت حذف و درنهایت اطلاعات ۲۷۱ شرکت باقی مانده است. دسته‌ی دوم داده‌ها مربوط به داده‌های قیمت پایانی روزانه‌ی سهام شرکت‌های فعال در بازار بورس اوراق بهادار تهران بوده است، که از طریق وبسایت "مرکز پردازش اطلاعات مالی ایران[۴۳]"، در بازه‌ی زمانی فروردین ۱۳۹۹ تا فروردین ۱۴۰۰ استخراج شده‌اند. درخصوص دسته‌ی اخیر، داده‌ی نهایی شامل ۲۷۱ عنوان شرکت به همراه قیمت پایانی در هر روز، قیمت روز قبل و تاریخ بوده است. درنهایت، با استفاده از قیمت پایانی روزانه و قیمت روز قبل هر سهم، بازده قیمت روزانه‌ی سهام شرکت‌ها و درنهایت میانگین بازده قیمت سالانه‌ی شرکت‌ها براساس مبانی نظری مذکور محاسبه شده است. دلیل انتخاب سال مالی ۱۳۹۹ برای هر دو دسته داده، وجود بیشینه‌ی اطلاعات لازم نسبت به سال‌های دیگر مالی شرکت‌ها بوده است. حال اولین گام پس از تهیه‌ی داده، شناسایی و اصلاح داده‌های پرت و یا از دست‌رفته یا ناموجود به وسیله‌ی توابع گوناگون است، که درخصوص داده‌های پژوهش حاضر، با توجه به حساسیت اعداد (نسبت‌های مالی) و عدم امکان حذف یا درج اعداد جایگزین با داده‌های خالی توسط تحلیلگر، تصمیم بر آن شد تا شرکت‌هایی که برای برخی نسبت‌های مالی خارج از داده هستند و یا نسبت‌هایی که برای عمده‌ی شرکت‌ها مقداری ناموجود دارند، از مجموعه‌ی داده‌های اولیه حذف شوند. گام بعدی، بررسی همبستگی بین معیارها، شاخص‌ها، و متغیرها بوده است، که در ماتریس خروجی شکل ۳، قطر اصلی نشانگر همبستگی یک شاخص با خود و برابر با ۱، و سایر سلول‌ها نیز هر چقدر رنگ روشن‌تری به خود اختصاص داده باشند، همبستگی بیشتری با شاخص متناظر در سلول خود دارند. همبستگی نامتعارف فقط برای دو شاخص (حاشیه‌ی سود قبل از مالیات) و (حاشیه‌ی سود خالص)



شکل ۳. ماتریس همبستگی بین شاخص‌ها.

[57] bourseview.com



شکل ۴. نمودار روش‌شناسی پژوهش.

مشاهده شده است، که با توجه به تعاریف متفاوت دو شاخص مذکور، می‌توان این همبستگی را نادیده گرفت و فقط شباهت عددی موجب این همبستگی شده است.

در گام بعد، نسبت به بی‌مقیاس‌سازی داده‌ها اقدام شده است. هدف بی‌مقیاس‌سازی در حقیقت یکسان‌سازی داده‌ها به شکلی است که بتوان تمامی داده‌ها را یکجا و با یک دید تفسیری مطالعه کرد؛ که این مهم از دو طریق در پژوهش حاضر انجام شده است (نرمال‌سازی و استانداردسازی)، که استانداردسازی روش بهتری نسبت به نرمال‌سازی برای داده‌های موردنظر پژوهش حاضر، شناسایی و صورت پذیرفته است.

در شکل ۴، الگوریتم‌ها و مراحل کلی پژوهش مشاهده می‌شوند.

۵. پیاده‌سازی و اجرا

۱.۵. خوشه‌بندی شرکت‌ها

در اولین گام از ارائه‌ی مدل‌ها، شرکت‌های فعال در بازار سهام موردنظر پژوهش حاضر براساس نسبت‌های مالی برگزیده‌ی مذکور در بخش چهار، خوشه‌بندی شده‌اند، تا شرکت‌هایی با رفتار مالی گوناگون تفکیک شوند. سپس ضمن محاسبه‌ی بازده و وزن کسب‌شده‌ی هر خوشه براساس بازده، از هر خوشه، تعدادی سهام با بازده بیشتر استخراج شده است.

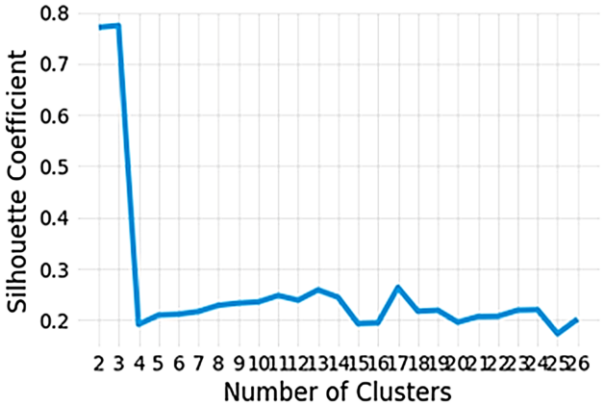
خروجی خوشه‌بندی سلسله‌مراتبی نشان می‌دهد که شرکت‌های فعال در بازار سهام را می‌توان براساس رفتار مالی آن‌ها، براساس شاخص‌های موردنظر پژوهش حاضر، به ۲ خوشه تقسیم‌بندی کرد. نکته‌ی حائز اهمیت آن است که در روش خوشه‌بندی مذکور، تعداد خروجی لزوماً تعداد بهینه‌ی اصلی نیست و فقط یک تحلیل بر هر نوع داده صورت گرفته است. از سوی دیگر، به‌منظور تشخیص تعداد خوشه‌ها در الگوریتم کامینز از طریق روش آرنج، با توجه به آنکه ۲۷۰ شرکت، موردنظر پژوهش حاضر بوده‌اند، فرض اولیه بر آن شد که بیشینه‌ی ۲۷ خوشه‌ی نهایی به‌منظور خوشه‌بندی شرکت‌ها در نظر گرفته شود؛ بدیهی است که این رقم، خارج از تصور به‌منظور خوشه‌بندی شرکت‌های فعال در بازار بورس اوراق بهادار تهران بوده است. در روش پیشنهادی الگوریتم از ۱ خوشه تا ۲۷ خوشه، داده‌ها براساس روش خوشه‌بندی کامینز خوشه‌بندی و سپس مقدار خطای «مجموع مربعات خطا» برای هر بار فرآیند اجرای الگوریتم محاسبه شده است؛ که نمودار خروجی آن در شکل ۵ (SSE)[58] مشاهده می‌شود.

مطابق شکل ۶، محور افقی بیانگر تعداد خوشه‌ها و محور عمودی بیانگر مقدار خطاست، که با کمی دقت برداشت خواهد شد که پس از ۱۰ یا ۱۱ خوشه، مقدار خطا تغییر خاصی نکرده است؛ در نتیجه مقدار k در روش ذکرشده برابر با ۱۰ یا ۱۱ است. لذا، به جهت افزایش اعتبار تحلیل، مقدار تشخیصی داده توسط نرم‌افزار نیز با دو مقدار ۱۰ و ۱۱ تعیین شده است، که این مقادیر در تکرارهای گوناگون به‌دست آمده است. همچنین به‌منظور تشخیص تعداد خوشه‌ها در روش پیشنهادی از طریق روش ضریب نیم‌رخ، همانند روش آرنج، ۲۷ مرتبه خوشه‌بندی انجام و هر بار مقدار ضریب نیم‌رخ ثبت شده است. مقدار این ضریب بین ۱- تا ۱+ متغیر بوده است، که این مقدار هر چقدر به ۱ نزدیک‌تر باشد، خوشه‌بندی بهتری صورت پذیرفته است.

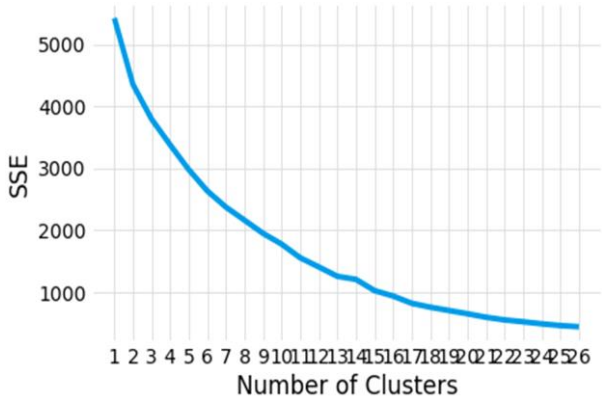
مطابق نمودار اخیر، محور افقی بیانگر تعداد خوشه‌ها و محور عمودی بیانگر مقدار ضریب نیم‌رخ است و با توجه به آنچه گفته شد، ضریب نیم‌رخ به ازاء مقادیر نزدیک‌تر به ۱ نشان می‌دهد که تعداد خوشه‌ی بهینه برای داده‌ی موردنظر مساوی ۳ است. این تذکر لازم است که مقدار ضریب نیم‌رخ به‌دست‌آمده به ازاء ۲ و ۳ خوشه در تکرار این روش، بسیار نزدیک به هم بوده و هر دو عدد مذکور می‌توانند تعداد خوشه‌ی مناسب در روش کامینز باشند. بنابراین، چهار مقدار ۲، ۳، ۱۰، و ۱۱ برای تعیین تعداد کامینز به‌عنوان K

[58] Sum of Squares due to Error

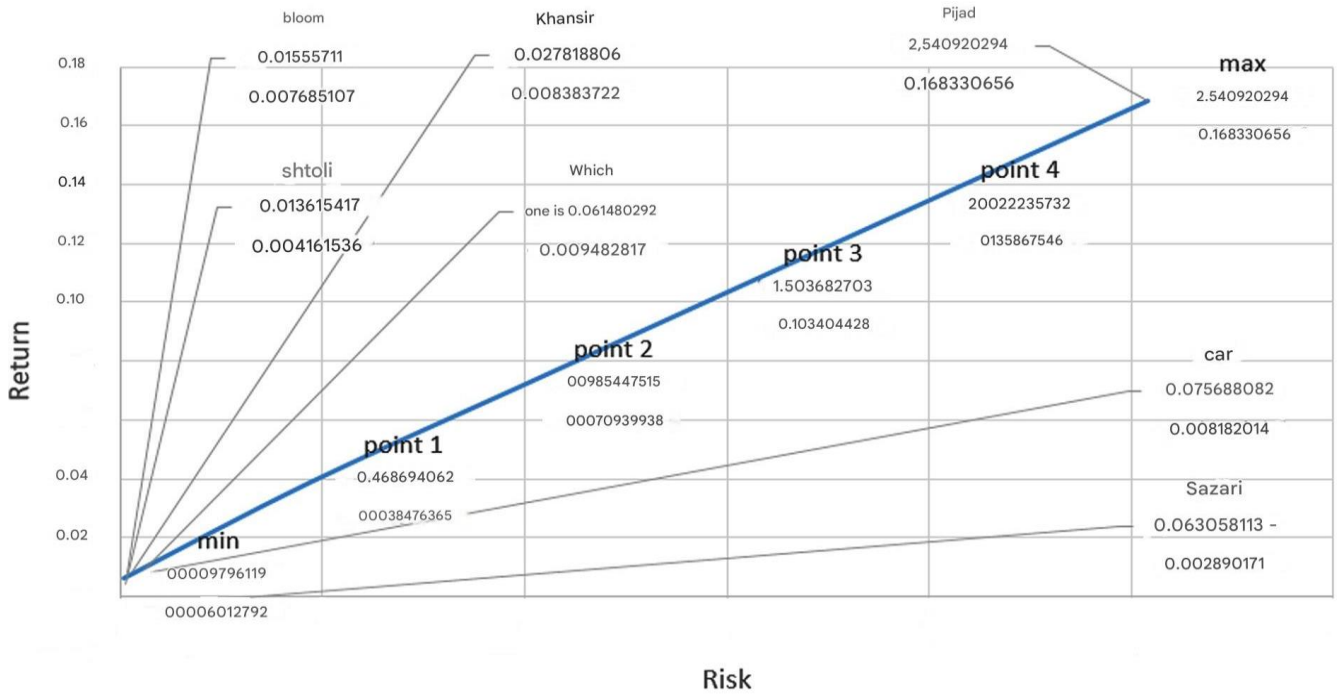
اولیه‌ی موردنیاز، شناسایی شد. مهم آنکه عدم حذف داده‌های پرت در داده‌های موردتحلیل پژوهش حاضر در گام پیش‌پردازش به‌دلیل مذکور (عدم امکان حذف و یا ایجاد داده‌های جدید تصادفی برای نسبت‌های مالی)، منجر به ایجاد رقمی بزرگ (۱۰ و ۱۱) در روش آرنج شده است. با توجه به تعداد خوشه‌های معقول موردانتظار برای شرکت‌های فعال در بورس و همچنین محدودیت در پیش‌پردازش کامل داده‌ها، به‌نظر می‌رسد از میان اعداد مذکور، ۲ یا ۳ خوشه، عدد مناسبی برای خوشه‌بندی باشد. به‌منظور سنجش الگوریتم ضریب نیم‌رخ، پیاده‌سازی و خوشه‌بندی با روش کامینز مجدداً به صورت دستی انجام شده است. همچنین خروجی براساس ضریب نیم‌رخ ثبت شده است، که نتایج نشان می‌دهند خوشه‌بندی به ازاء ۳ خوشه و ۱۰۰ مرتبه تکرار، نتیجه‌ی بهتری را در مقایسه با تعداد خوشه‌های ذکرشده‌ی دیگر ارائه داده است؛در نتیجه راستی‌آزمایی نمودار SSE تأیید می‌شود. با توجه به خروجی بخش حاضر، سهام مربوط به هر خوشه در نرم‌افزار اکسل مشخص شده است. در گام بعد، وزن هر سهم در هر خوشه با توجه به مباحث نظری اشاره‌شده در این خصوص، با توجه به وزن‌های کسب‌شده‌ی خوشه‌ها و میانگین بازده قیمت سالانه‌ی سهام شرکت‌ها، ۷ سهم شامل: پیزد، کدما، خودرو، خنصیر، ساذری، شکف، و شتولی انتخاب شده‌اند. در آخرین مرحله از این گام، در نرم‌افزار اکسل، ضمن بهره‌گیری از روش مارکویتز، ۶ سبد سرمایه با درنظرگرفتن مقادیر ریسک و بازده گوناگون بین دو نقطه با بیشینه و کمینه‌ی بازده تشکیل شد، که از بیان جزئیات گام کنونی و فرمول‌های تدوین‌شده در نرم‌افزار اکسل به جهت جلوگیری از نشر مطالب تکراری خودداری شده است. بدیهی است با تغییر مقدار گام‌ها بین دو نقطه‌ی کمینه و بیشینه، می‌توان تعداد نقطه‌ی بیشتری را به‌عنوان سبدهای سرمایه با مقادیر بازده و ریسک گوناگون تشکیل داد. در پژوهش حاضر، گام‌ها به نحوی تعیین شده‌اند تا ۴ نقطه‌ی میانی ایجاد شود. در شکل ۷، نمودار مقادیر مربوط به ریسک و بازده هر سبد و هر سهم مشخص شده است؛ به‌گونه‌ای که در آن خط آبی‌رنگ، نشان‌دهنده‌ی مقادیر مربوط به ریسک و بازده ۶ سبد مذکور است و نقاط مربوط به ریسک و بازده هر سهم نیز به‌صورت جداگانه مشاهده می‌شود؛ که مطابق آن، از بین سهام شرکت‌های منتخب، هرگاه تصمیم به خرید تک‌سهمی در مقایسه با ریسک و بازده سبد تشکیل شده باشد، سهام شرکت‌های پیزد و با اختلاف زیاد کدما، نسبت به سهام سایر شرکت‌ها ارجحیت بیشتری داشته‌اند، بنابراین در سبد نیز ارزش بیشتری داشته‌اند.



شکل ۶. نمودار ضریب نیم‌رخ.



شکل ۵. نمودار SSE.



شکل ۷. نمودار سبدهای گوناگون به ازاء ریسک و بازده موردانتظار متفاوت + ریسک و بازده هر سهم.

۲.۵. ماژول طبقه‌بندی

در این گام از پژوهش، ماژولی ارائه شده است که ارتباط میان نسبت‌های مالی و مقدار میانگین بازده قیمت سالانه‌ی یک شرکت را شناسایی و در صورت اجرای مدل، به سرمایه‌گذار در جهت تصمیم‌گیری به جهت خرید یا عدم خرید به‌عنوان یکی از عوامل مدیریت سبد خرید، کمک شایانی می‌کند. همچنین مدل ذکرشده در تشخیص تقلب شرکت‌ها نیز مؤثر خواهد بود. به‌منظور آنکه نرم‌افزار قادر باشد هر ردیف داده را به گروه مناسب یا نامناسب نسبت دهد، نیاز به ستون جدیدی با عنوان ستون هدف است، لذا لازم است که سرمایه‌گذار تصمیم‌گیری کند که میزان بازده قیمت موردانتظار چقدر باشد تا سهمی مناسب سرمایه‌گذاری شود؛ به همین علت باید بررسی شود حد میانی بازده قیمت سهام در داده‌های موجود چقدر است تا ستون هدفی با مقادیر برابر به ازاء دو برچسب، موجود باشد. این مقدار ممکن است نزدیک به بازده موردانتظار سرمایه‌گذار باشد و یا به‌دلیل کمبود اطلاعات در پایگاه‌های داده، همچون پژوهش حاضر، مقداری کمتر از مقدار موردانتظار عموم سرمایه‌گذاران داشته باشد. به این منظور، جدول ۱، با هدف مکان‌یابی نقطه‌ی میانه‌ی بازده سهام تشکیل شده است.

جدول اخیر نشان می‌دهد که مقدار بازده ۰/۳٪، مقداری است که به ازاء آن، ۱۳۴ سهم بازدهی بیشتر از مقدار نقطه‌ی میانه و ۱۳۷ سهم بازدهی کمتر از آن داشته‌اند؛ بنابراین شرکت‌هایی که بازده بیشتر دارند، برچسب ۱ و شرکت‌هایی که بازده کمتر دارند، برچسب ۰ دریافت کرده‌اند. مهم آنکه با درنظرگرفتن بازده ۰/۳٪، مقادیر ستون هدف مقداری متوازن خواهند داشت و بنابراین به‌علت حساسیت داده‌های نسبت‌ها، نیازی به شبیه‌سازی داده‌ها به‌منظور ایجاد داده‌ای متوازن نخواهد بود. در مرحله‌ی بعد، داده‌ها با روش استانداردسازی (SS)[۵۹]، بی‌مقیاس شده‌اند؛ اگرچه در ادامه در پرسپترون چندلایه، نیاز به بی‌مقیاس‌سازی با روش نرمال‌سازی (MMS)[۶۰] است، که جداگانه صورت پذیرفته است. حال باید آخرین مرحله‌ی قبل از پیاده‌سازی الگوریتم‌های طبقه‌بندی با عنوان تقسیم به داده‌های آموزش و آزمون صورت پذیرد؛ به این منظور ۳۰٪ داده‌ها به‌صورت تصادفی و توسط نرم‌افزار به‌عنوان داده‌ی آزمون و مابقی به عنوان داده‌ی آموزش در نظر گرفته شده‌اند. در مرحله‌ی بعد نیز به مدل‌سازی پرداخته شده است، که نتایج آن در ادامه ارائه شده است. مهم آنکه محاسبه‌ی میزان امتیاز مدل در تمامی الگوریتم‌ها ضمن بهره‌گیری از ماژول کراس[۶۱] و همچنین مقدار ضریب خطای معیار ماتریس آشفتگی، که پیشتر در مبانی نظری تشریح شده است، صورت پذیرفته و با توجه به حجم مطالب، مطالعه درخصوص نحوه‌ی تنظیم معیارهای برنامه‌نویسی به مخاطب واگذار و نوشتار واندرپلاس[۶۲] (۲۰۱۶)،[۹] برای مطالعه پیشنهاد شده است.

جدول ۱. مکان‌یابی نقطه‌ی میانه‌ی بازده سهام به ازاء مقادیر مختلف بازده.

Not ok	ok	Rate of Return
266	5	>0.05
265	6	>0.01
222	49	>0.005
182	89	>0.004
137	134	>0.003
92	176	>0.002

در مرحله‌ی اول، طبقه‌بندی با الگوریتم درخت تصمیم با معیار آنتروپی[۶۳] صورت پذیرفته است، که در این مرحله پس از بررسی ۳۰ خروجی، امتیاز مدل برابر با ۰/۶ بوده است. در مرحله‌ی دوم، طبقه‌بندی با الگوریتم درخت تصمیم با معیار جینی[۶۴] انجام شده است، که امتیاز مدل ساخته‌شده توسط درخت تصمیم در این مرحله نیز پس از بررسی ۳۰ خروجی برابر با ۰/۵۴ بوده است. در مرحله‌ی سوم، طبقه‌بندی با الگوریتم گوسین نایو بیز صورت پذیرفته است، که در آن هم از روش cross با ۳۰ مرتبه تکرار با امتیاز خروجی برابر با ۰/۵ و در مرحله‌ی بعدی، طبقه‌بندی با الگوریتم نزدیک‌ترین همسایه، متغیری با عنوان «وزن‌ها[۶۵]» وجود دارد، که در پژوهش حاضر دو مقدار «یکنواخت[۶۶]» در مرحله‌ی چهارم و «فاصله[۶۷]» در مرحله‌ی پنجم برای پارامتر ذکرشده در نظر گرفته شده است. الگوریتم با دریافت مقدار یکنواخت، برای تمامی نقاط در یک محل وزن برابری را در نظر می‌گیرد و با دریافت مقدار فاصله، برای همسایگان نزدیک‌تر، مقدار وزن بیشتری در مقایسه با همسایگان دورتر در نظر خواهد گرفت.[۲۸] خروجی در مراحل چهارم و پنجم پس از ۳۰ بار تکرار مدل برابر با ۰/۶۷ و ۰/۶۵ بوده است، که مقادیر اخیر برتری مقدار فاصله را بر مقدار یکنواخت نشان می‌دهند. در مرحله‌ی بعد، در الگوریتم ماشین بردار پشتیبان، متغیری با عنوان «کرنل[۶۸]» به منظور چندمنظوره‌سازی الگوریتم وجود دارد؛ که امکان ایجاد هسته‌های سفارشی را علاوه‌بر هسته‌های مشترک فراهم می‌کند و به سه شکل (خطی[۶۹]، غیرخطی[۷۰]، چندجمله‌ای[۷۱]) قابل تنظیم است.[۲۸] در پژوهش حاضر، هر سه نوع هسته‌های مشترک در مرحله‌های ۶ الی ۸ در نظر گرفته شده است، تا فقط بهترین نوع هسته به تشخیص نرم‌افزار براساس میزان دقت خروجی محاسبه شود. بدیهی است مخاطب پژوهش حاضر قادر خواهد بود از تنظیم متغیرها به شکل دقیق‌تر در هر کدام از الگوریتم‌ها به جهت دریافت خروجی بهتر استفاده کند. مهم آنکه تعداد تکرار الگوریتم cross باید بررسی شود و براساس مقدار خروجی، بهترین تعداد تکرار تشخیص داده شود. درنهایت، بهترین خروجی حاصل پس از ۱۰ بار تکرار براساس هسته‌ی خطی برابر با ۰/۶۷ بوده است. به‌منظور مدل‌سازی پرسپترون چندلایه، دو متغیر «فعال‌سازی[۷۲]» و «حل‌کننده[۷۳]» به‌منظور تغییر از حالت پیش‌فرض قابل بررسی هستند. متغیر فعال‌سازی با مقدار identity[۷۴] هم علاوه‌بر مقدار

[59] Standard Scaler
[60] Min Max Scaler
[61] cross
[62] VanderPlas
[63] entropy
[64] gini
[65] weights
[66] uniform
[67] distance
[68] kernel
[69] Linear
[70] rbf
[71] poly
[72] Activation
[73] Solver
[۷۴] فعال‌سازی بدون عملیات، مفید برای پیاده‌سازی گلوگاه خطی، $f(x) = x$ را بر می‌گرداند.

پیش‌فرض که عبارت است از relu[۷۵] به ترتیب در مراحل دهم و نهم و متغیر حل‌کننده با مقدار lbfgs[۷۶] جایگزین با مقدار اصلی بررسی شده‌اند. متغیر «max_iter» هم تعداد تکرار شبکه را مشخص می‌کند، که می‌توان شبکه را تا ۱۵۰۰۰ بار تکرار کند. همچنین عملکرد الگوریتم اخیر با روش بی‌مقیاس‌سازی به روش استانداردسازی در مقایسه با روش بی‌مقیاس‌سازی به روش نرمال‌سازی در یازدهمین مرحله‌ی مدل‌سازی بررسی شده است؛ ولی همان‌طور که موردانتظار بود، در روش استانداردسازی، خروجی بهتری نسبت به روش نرمال‌سازی اتخاذ شده است.

درنهایت، مقایسه‌ی نتایج امتیاز خروجی مدل‌های اخیر نشان می‌دهد که الگوریتم ماشین بردار پشتیبان در مرحله‌ی ششم، با اختلاف قابل‌توجهی نسبت به سایر الگوریتم‌ها توانسته است مقادیر ستون هدف را از منظر طبقه‌بندی پیش‌بینی کند. لازم به یادآوری است که هرگاه در هر کدام از الگوریتم‌های بررسی‌شده، تنظیم متغیر دقیق‌تری صورت پذیرد، مدل ساخته‌شده، میزان دقت بیشتری خواهد داشت.

۳.۵. ماژول پیش‌بینی

در پژوهش حاضر، به فراخور مسئله و هدف اصلی پژوهش (ماژول تشخیص میزان بازده قیمت و شناسایی تقلب)، به بررسی الگوریتم‌های پیش‌بینی مذکور پرداخته شده است. همچنین درنهایت، برترین الگوریتم در این زمینه با توجه به معیارهای سنجش مدل‌های پیش‌بینی انتخاب شده است. در مرحله‌ی اول، پیش‌بینی با الگوریتم درخت تصمیم صورت پذیرفته است، که در جدول ۲، مقادیر ارزیابی مدل به ازاء دفعات تکرار ۵، ۱۰، و ۱۵ مرتبه برای دو الگوریتم درخت تصمیم با داده‌های بی‌مقیاس‌شده به دو روش استانداردسازی (dfSS)[۷۷] و نرمال‌سازی (dfMMS)[۷۸] مشاهده می‌شود. استفاده از داده‌های نرمال، فقط جهت اطلاع مخاطب در نظر گرفته شده است و با توجه به دلیل مذکور در بخش‌های قبل، در پژوهش حاضر توجیح علمی ندارد. درنهایت مشاهده شده است که با اختلاف بسیار کمی، خروجی اخذشده با هر دو نوع داده، تقریباً یکسان بوده است.

در مورد الگوریتم نزدیک‌ترین همسایه، مشابه با الگوریتم قبلی، مقادیر اخذشده از خروجی دو مدل ساخته‌شده با داده‌ی بی‌مقیاس‌شده به روش استانداردسازی در جدول ۳ ارائه شده است. مقایسه نشان می‌دهد که مدل ساخته‌شده با عنوان KNR ۱ نسبت به مدل با عنوان KNR ۲ برتری دارد.

در مورد الگوریتم ماشین بردار پشتیبان، از آنجا که الگوریتم حاضر با سه نوع هسته‌ی سفارشی قادر به تحلیل داده است، سه مدل با هسته‌ی خطی، چندجمله‌ای، و غیرخطی همچون الگوریتم طبقه‌بندی به ترتیب ایجاد و هر مدل با معیارهای خطای الگوریتم‌های پیش‌بینی اخیر سنجیده شده است. مقایسه‌ی خروجی مقادیر مختلف خطا در جدول ۴ ارائه شده است، که سومین مدل یا مدل با هسته‌ی سفارشی غیرخطی، بهترین نتیجه را از بین سه مدل ساخته‌شده ارائه داده است. درخصوص الگوریتم پرسپترون چندلایه نیز نتایج در جدول ۵ ارائه شده است.

درنهایت، در آخرین گام از ایجاد مدل‌های پیش‌بینی، برترین مدل‌های هر الگوریتم با یکدیگر مقایسه شده‌اند، تا بهترین مدل پیش‌بینی تشخیص داده

جدول ۲. مقایسه‌ی مقادیر خطا با دو روش بی‌مقیاس‌سازی با الگوریتم پیش‌بینی درخت تصمیم.

	x	index	cv=5	cv=10	cv=15
DTR1	dfSS	MSE	-0.14	-0.14	-
		MAE	-0.03	-0.06	-
		RMSE	-0.22	-0.16	-0.13
		r2	-193	-244	-
DTR2	dfMMS	MSE	-0.14	-0.14	-
		MAE	-0.06	-0.08	-
		RMSE	-0.2	-0.16	-0.13
		r2	-193	-200	-

جدول ۳. مدل برتر با توجه به هر معیار ارزیابی.

	x	index	cv=5	cv=10	cv=15	result
KNR1	dfSS	MSE	-0.1585	-0.154	-0.154	7
		MAE	-0.0522	-0.0486	-0.0484	5
		RMSE	-0.2999	-0.2201	0.1808	5
		r2	-2866	-2576	-2517	5
KNR2	dfSS	MSE	-0.1602	-0.1548	-0.1547	8
		MAE	-0.053	-0.049	-0.048	5
		RMSE	-0.3063	-0.2219	-0.1824	5
		r2	-3289	-2871	-2753	5

جدول ۴. مقایسه‌ی مقادیر خطای مدل‌های ایجادشده با الگوریتم پیش‌بینی ماشین بردار پشتیبان.

	x	index	cv=5	cv=10	cv=15
SVR1	dfSS	MSE	-0.1494	-0.1476	-0.1475
		MAE	-0.0923	-0.094	_
		RMSE	-0.2604	-0.2091	-0.1858
		r2	-643	-1473	_
SVR2	dfSS	MSE	-0.1607	-0.1599	-0.1567
		MAE	-0.1121	-0.1178	_
		RMSE	-0.2854	-0.2338	-0.2093
		r2	-689	-1925	_
SVR3	dfSS	MSE	-0.1469	-0.1463	-0.146
		MAE	-0.089	-0.086	-0.0851
		RMSE	-0.2381	-0.1906	-0.1679
		r2	-377	-675	_

[۷۵] تابع واحد خطی اصلاح شده $f(x)=\max(0, x)$ را بر می‌گرداند.

[۷۶] یک بهینه‌ساز در خانواده روش های شبه نیوتنی است.

[77] df Standard Scaler

[78] df Minmax Scale

جدول ۵. مقادیر مختلف خطا به ازاء دفعات مختلف تکرار الگوریتم پیش‌بینی پرسپترون چندلایه.

	x	index	cv=5	cv=10	cv=15
MLP	dfSS	MSE	-5.92	-7.82	_
		MAE	-0.63	-0.6736	_
		RMSE	-2.35	-1.92	-1.61
		r2	-515784	-888587	_

جدول ۶. جدول خلاصه‌ی مقایسه‌ی خروجی مدل‌های پیش‌بینی.

	DTR1	KNR1	SVR3	MLP	The best model
MSE	-0.14	-0.154	-0.146	-5.92	DTR1
MAE	-0.03	-0.0484	-0.851	-0.63	DTR1
RMSE	-0.13	0.1808	-0.1679	-1.61	DTR1
r2	-193	-2517	-377	-515784	DTR1

شود. مقایسه‌ی نتایج خروجی مدل‌های اخیر مطابق جدول ۶ نشان می‌دهد که الگوریتم درخت‌تصمیم توانسته است با کمینه‌ی مقادیر خطا، بهترین مدل را از میان سایر الگوریتم‌ها ارائه دهد. مهم آنکه مقدار خطای الگوریتم درخت تصمیم جزء کمینه‌ی مقادیر ممکن خطا یا مقداری بسیار نزدیک به صفر برای سه معیار اول و مقداری بسیار نزدیک‌تر به ۱ در مقایسه با سایر الگوریتم‌ها برای معیار آخر بوده است.

۶. نتیجه‌گیری و پیشنهادها

در پژوهش حاضر، ابتدا پس از مرور منابعی در زمینه‌ی کاربرد علم داده در حل مسائل مالی، جایگاه علم داده در حل این‌گونه مسائل تبیین شده است. خروجی بخش کنونی، رهنمودی در جهت تشخیص موضوع اصلی پژوهش حاضر ایجاد کرده است. این موضوع به نحوی شناسایی شده است که یکی از نیازهای مهم تحلیلگران مالی در سطوح مختلف اعم از دانشجویان این رشته و همچنین افراد متخصص با بهره‌گیری از شیوه‌های علم داده در قالب سیستم پشتیبان تصمیم‌گیری در مدیریت سبد سهام مرتفع شود. از این رو، ۶ سبد سهام با ترکیب ۷ سهم منتخب با درنظرگرفتن مقادیر مختلف بازده و ریسک ضمن ترکیب "ماژول خوشه‌بندی" مذکور و مدل مارکویتز ایجاد شده است؛ در این مسیر، تعداد زیادی از شرکت‌های فعال در نظر گرفته‌شده در بازار بورس اوراق بهادار تهران در کمترین زمان ممکن از نظر رفتار بنیادین تحلیل و در خوشه‌های مجزا تفکیک شده‌اند. از سویی دیگر، وزن خوشه‌ها با توجه به بازده قیمت سالانه‌ی هر خوشه محاسبه شده است. سپس از هر خوشه متناسب با اوزان کسب‌شده، سهمی برگزیده شده است که از نظر بازده قیمت سالانه، وزن بیشتری نسبت به سایر سهام شرکت‌ها داشته است. بنابراین دو مزیت نوین با ارائه‌ی روش ساخته‌شده ایجاد شده است، که عبارت‌اند از: ۱) ایجاد خوشه‌های تفکیکی که منجر به ایجاد خوشه‌هایی متشکل از سهام شرکت‌ها با رفتار یکسان از منظر شاخص‌های مالی در هر خوشه شده است؛ که بدیهی است تحلیل رفتار مذکور با توجه به تعدد شرکت‌ها برای یک تحلیلگر با روش‌های مرسوم قبلی بسیار پیچیده و زمان‌بر است. از سویی دیگر، تعداد شرکت‌های در نظر گرفته‌شده در پژوهش حاضر، تحلیل دقیق‌تری را نسبت به پژوهش‌های پیشین ایجاد کرده است. ۲) در مسیر خوشه‌بندی، سهم یک شرکت فقط براساس میزان بازده انتخاب نشده و تنوع رفتاری شرکت‌ها نیز در مسیر انتخاب سهام به منظور تشکیل سبد در نظر گرفته شده است؛ که این مهم، تنوع لازم سهام در مدل مارکویتز را بهبود بخشیده است. در بخش ایجاد "ماژول طبقه‌بندی" نیز از میان ۱۱ مدل نهایی، برترین مدل با عنوان "مدل ماشین بردار پشتیبان خطی" براساس بهترین عملکرد معیار سنجش (خروجی ماتریس آشفتگی) با مقدار ۶۷٪ تطبیق با داده‌های واقعی یا با ۶۷٪ دقت پیش‌بینی شناسایی شده است. مدل ساخته‌شده قادر است با دریافت مقادیر شاخص‌های مالی یک شرکت در قالب نسبت‌های مالی، مقدار بازده قیمت سهام شرکت را به صورت «پذیرش شرط» یا «عدم پذیرش شرط»، که شرط همان مقدار موردانتظار بازده قیمت سرمایه‌گذار است، پیش‌بینی کند. مدل پیشنهادی در برخی گام‌های کشف تقلب به‌هنگام حسابرسی شرکت‌ها در قالب کلان مؤثر است؛ به‌صورتی که اگر عدم تطابق بین مقدار بازده قیمت سهم شرکتی با خروجی مدل مذکور مشاهده شود، با ۶۷٪ اطمینان می‌توان بیان کرد که در صورت‌های مالی شرکت موردنظر، تقلبی صورت پذیرفته است، که این مهم نیازمند حسابرسی جزئی‌تر صورت‌های مالی شرکت خواهد شد. درخصوص ایجاد "ماژول پیش‌بینی" نیز، از میان ۷ مدل نهایی، برترین مدل با عنوان «مدل پیش‌بینی درخت تصمیم» براساس کمترین میزان خطا از منظر ۴ معیار خطا شناسایی شده است. مدل ساخته‌شده قادر است با دریافت مقادیر نسبت‌های مالی یک شرکت، مقدار بازده قیمت سهام شرکت را پیش‌بینی کند، که این مهم نیز در شناسایی تقلب شرکت‌ها بسیار مؤثر خواهد بود و نسبت به مدل طبقه‌بندی ارائه‌شده، مقدار قابل تأمل‌تری را منعکس می‌کند. همچنین به هنگام تشکیل سبد سرمایه، سرمایه گذار قادر خواهد بود در کنار بررسی داده‌های نسبت‌های مالی یک شرکت، مقدار بازده قیمت را به سرعت و بدون نیاز به داده‌های تاریخی شاخص قیمت سهام پیش‌بینی کند؛ که در این مسیر، مشابه با نوشتار جیانسی (۲۰۰۷)،[۲۹] نتیجه‌گیری شد که استفاده از الگوریتم ماشین بردار پشتیبان برای داده‌های عظیم، منجر به افزایش کارایی در مسیر طبقه‌بندی خواهد شد. همچنین مشابه با پژوهش ناندا و همکارانش (۲۰۱۰)،[۱۵] که قبلاً به آن پرداخته شده است، نتیجه‌گیری شد که انتخاب سهام از بین خوشه‌های تشکیل‌شده از سهام گوناگون شرکت‌ها، موجب کاهش ریسک متنوع‌سازی سبد سهام خواهد شد. از سویی دیگر، مشابه با پژوهش دودک و همکاران (۲۰۲۴)،[۳۱] نتیجه‌گیری شده است که هیچ‌گاه بهترین روش واحد برای پیش‌بینی نوسان‌ها وجود ندارد و همان‌طور که پیشتر نیز بیان شد، لازم است تنظیم معیارها متناسب با افق پیش‌بینی صورت پذیرد.

درنهایت، به‌عنوان هدف اصلی پژوهش، از ترکیب سه ماژول ارائه‌شده، روشی در قالب یک سیستم پشتیبان تصمیم‌گیری در مدیریت سبد سهام برای تحلیلگران در سطوح مختلف علمی فراهم شده است. علاوه‌بر آنکه هر یک از سه گام طی‌شده، خود به تنهایی یک ماژول در مدیریت سبد سهام به شمار می‌روند، سیستم مذکور قادر خواهد بود به‌عنوان یک مدل جامع برای داده‌هایی در فواصل زمانی گوناگون و شاخص‌های موردنظر تحلیلگر، قابلیت تعمیم‌دهی داشته باشد و بدین صورت سرمایه‌گذاران را در مسیر تحلیلی سریع و با دقت کافی، یاری کند. عملکرد روش نهایی این‌گونه است که سرمایه‌گذار ابتدا تعدادی از شرکت‌های موردنظر خود را با توجه به شنیده‌ها، تحلیل‌ها، و اخبار انتخاب می‌کند؛ بعد با بهره‌گیری از مدل پیش‌بینی مذکور و داده‌های تاریخی نسبت‌های مالی، بازده قیمت شرکت‌ها را پیش‌بینی می‌کند. سپس با استفاده از مدل طبقه‌بندی مذکور، نسبت به حفظ یا حذف سهم شرکت‌ها با توجه به

طبقه‌ی پیش‌بینی‌شده، اقدام می‌کند و در آخرین مرحله، شرکت‌هایی که موفق به پذیرش شرط شده‌اند را با استفاده از مدل خوشه‌بندی، تفکیک می‌کند و سبد سهام متنوعی را با توجه به میزان ریسک و بازده موردانتظار خود تشکیل می‌دهد. مهم آنکه، همان‌طورکه گفته شد، مدل‌های طبقه‌بندی و پیش‌بینی ارائه‌شده در کنار کاربرد در تشکیل و مدیریت سبد سهام قادر خواهند بود در مبحث پیش‌بینی تقلب احتمالی شرکت‌ها نیز مؤثر واقع شوند. طی انجام پژوهش حاضر، محدودیت‌های نیز وجود داشت که عبارت‌اند از:

۱) به‌منظور بررسی پژوهش‌ها، محدودیت تاریخی در نظر گرفته نشده است؛ چرا که در صورت فیلترکردن تاریخ مطالعات، احتمال از دست‌رفتن پژوهش‌هایی با موضوعات خاص وجود دارد؛ ۲) مبانی نظری مالی در پژوهش حاضر با توجه به مخاطب هدف، فقط به‌صورت یادآوری مباحث پایه و یا تشریح کامل مباحث تخصصی بوده است؛ ۳) با توجه به فراوانی مباحث مبانی نظری در زمینه‌ی علم داده و داده کاوی، گزیده‌ای از مبانی به شکل کاربردی برای درک بهتر روند اجرای الگوریتم‌ها ارائه و مطالعه درخصوص مابقی مبانی به خواننده واگذار شده است؛ ۴) با توجه به نقص منابع داده و نبود برخی اطلاعات، فقط شرکت‌هایی در نظر گرفته شده‌اند که تمامی اطلاعات موردنیاز پژوهش جاری برای آن‌ها موجود بوده است؛ ۵) با توجه به گستردگی معیارهای موجود در هر الگوریتم، برخی از آن‌ها به‌منظور جلوگیری از تکرار مطالب در طول تدوین متن پژوهش، هم‌زمان با تشریح مراحل پیاده‌سازی به شکل مختصر تشریح شده‌اند. همچنین در ادامه، به‌منظور بهبود مدل‌های ساخته‌شده در پژوهش حاضر، گزینه‌هایی به پژوهشگران آتی پیشنهاد شده است تا در آینده به آن‌ها پرداخته شود: ۱) با بهره‌گیری از داده‌های سایر متغیرهای مالی (همچون سایر نسبت‌های مالی، و ...) با فرض موجودبودن اطلاعات موردنیاز، به بهبود کیفیت خوشه‌بندی شرکت‌های فعال در بازار سهام نسبت به پژوهش حاضر پرداخته شود. ۲) ضمن محاسبه‌ی بازده کل سهام شرکت‌های فعال در بازار سهام (جمع بازده قیمت و بازده نقدی)، نسبت به بهینه‌سازی کیفیت داده‌های اولیه‌ی ستون هدف در مسیر مدل‌سازی و در نتیجه بهبود خروجی مدل‌های طبقه‌بندی و پیش‌بینی ایجادشده در پژوهش حاضر اقدام شود. ۳) ضمن بهره‌گیری از دانش برنامه‌نویسی پایتون، به وسیله‌ی تنظیم دقیق‌تر پارامترها در هر الگوریتم، به بهینه‌سازی مسیر مدل‌سازی پرداخته شود. ۴) مدل‌های ایجادشده در بستر زبان برنامه‌نویسی، در قالب نرم‌افزارهایی که استفاده از آن برای کاربر ساده است، پیاده‌سازی و تصویرسازی شوند. ۵) در مسیر پیش‌پردازش داده‌های خام، داده‌های از دست‌رفته و یا پرت، شناسایی و نسبت به جایگزینی آن‌ها اقدام شود و نتایج با مسیر اصلی در پژوهش حاضر مقایسه شود. ۶) با استناد به پژوهش حاضر، مدلی مشابه متشکل از ماژول‌هایی پیشرفته‌تر در بستر بانک‌های اطلاعاتی آنلاین ایجاد شود تا پژوهشگران آتی قادر به استفاده‌ی سریع‌تر از ماژول‌ها باشند. ۷) ماژول‌های ساخته‌شده در پژوهش حاضر با داده‌های سایر کشورها استفاده شود، تا راستی‌آزمایی عملکرد ماژول‌ها در شرایط گوناگون مدیریتی سنجیده شود.

References - منابع

1. Pyle, D., 2003. *Business Modeling and Data Mining*. Elsevier Press, Netherlands. https://doi.org/10.1016/B978-1-55860-653-1.X5000-3.
2. Dean, J., 2014. *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. John Wiley & Sons Press, United States. ISBN: 978-1-118-69178-6. https://www.nrigroupindia.com/e-book
3. Kunnathuvalappil Hariharan, N., 2018. Applications of data mining in finance. *J. of Innovations in Engineering Research and Technology*, *5*(2), pp.72-77. https://www.researchgate.net/publication/354224201_APP LICATIONS_OF_DATA_MINING_IN_FINANCE.
4. Zamani, M., Pakmaram, A., Rezaei, N. and Abdi, R., 2023. The role of information behavior in financial reporting (With an emphasis on information entropy theory). *J. of Nonlinear Analysis and Applications*, *15*(4), pp. 293-309. https://doi.org/10.22075/IJNAA.2022.28778.3985. [In Persian].
5. Abdullah, Z. and Hamdan, A., 2015. Hierarchical clustering algorithms in data mining. *J. of Computer and Information Engineering*, *9*(10), pp.2194-2199. https:doi.org/10.5281/zenodo.1109341
6. Nokart., 2021. *The Difference Between Data Science And Data Mining*. from Nokartohttps://nokarto.com/. [In Persian].
7. Fazel Zarandi, M. and Kazemi, A., 2008. Application of rough set theory in data mining for decision support systems (DSSs). *QIAU*, (1), pp. 25-34. https://civilica.com/doc/790867/ [In Persian].
8. Dastgir, Sh., 2019. Data mining technology, a new approach in the financial field. *Auditing Knowledge*, *11*(4), 0-0.
9. VanderPlas, J., 2016. *Python data science handbook: Essential tools for working with data*. "O'Reilly Media, Inc." .https://www.oreilly.com/library/view/python-data-science/9781491912126/

10. Zhang, C.W. and Wang, Y.B., 2010. Research on the application of distributed data mining in anti-money laundering monitoring system. *In 2010 2nd International Conference on Advanced Computer Control*, pp. 133-135). IEEE.

11. Han, J., Kamber, M. and Pei, J., 2012. Data Mining: concepts and techniques, 3rd Edn, Data Transformation by Normalization, *Elsevier*, ISBN 978-0-12-381479-1.

12. Paikari, E., 2023. Bug report quality prediction and the impact of including videos on the bug reporting process, *University of California, Irvine*. https://escholarship.org/uc/item/9tn4t5gs [In Persian]

13. Boschetti, A. and Massaron, L., 2015. *Python Data Science Essentials*. Packt Publishing Ltd. 3rd Edn, Retrieved May 17, 2022, from https://www.oreilly.com/library/view/python-data-science/9781789537864/

14. Mahmoudi, L., Razvian, M.T., Ghorchi, M. and Ostadtaghizadeh, A., 2020.Application of multilayer perceptron (MLP) neural network model in urban vulnerability zoning with emphasis on earthquake (a case study on municipal district 20 in Tehran), *J. of Natural Environmental Hazards*, *9*(24), pp.129-150. http://doi.org/10.22111/JNEH.2020.31217.1551. [In Persian].

15. Nanda, S.R., Mahanty, B. and Tiwari, M.K., 2010. Clustering Indian stock market data for portfolio management. *Expert Systems with Applications*, *37*(12), pp.8793-8798., *37*(12), https://doi.org/10.1016/j.eswa.2010.06.026

16. Abbasi, A. and Nikbakht, M., 2018. Identification and clustering outsourcing risks of aviation part-manufacturing projects in aviation industries organization using k-means method. *J. of Modern Processes in Manufacturing and Production*, *7*(4), pp.1730. https://civilica.com/doc/1146207/ [In Persian].

17. Soltani Nejad. D., 2015. Optimizing the investment portfolio using clustering methods: comprehensive humanities portal. *J. of Asset Management and Financing*. *4*(4), pp.1-16. [In Persian].

18. Sayadi, M. and Omidi, M., 2020. Prediction-based portfolio optimization model for Iran's oil-dependent stocks using data mining methods. *Iranian Journal of Economic Studies*, *8*(1), pp.207-234. https://doi.org/10.1016/j.jeconom.2019.04.017

19. Sadeghi Arani, Z. and Mohgar, A., 2019.Presenting a hybrid clustering model of Tehran Stock Exchange member companies: meta-heuristic algorithms approach. *Farda Management Scientific Research Journal*, *18*(59), pp.18-3. https://www.sid.ir/paper/386061/en [In Persian]

20. Abbasi, I., Dehekhani, H., Khozin, A. and Bazadeh, F., 2019. Designing a smart model for predicting financial wealth in insurance companies (data mining approach). *J. of Investment Science*, *9*(34), pp.211-229.[In Persian].

21. Razin, A., 2015. *Big buzz about big data: Five Ways Big Data Is Changing Finance*, Forbes. https://www.forbes.com/sites/elyrazin/2015/12/03/big-buzz-about-big-data-5-ways-big-data-is-changing-finance/

22. Diebold, F. X., Ghysels, E., Mykland, P. and Zhang, L., 2019. Big data in dynamic predictive econometric modelling. *J. of Econometrics, 212*(1), pp.1-3. https://doi.org/10.1016/j.jeconom.2019.04.017

23. Majumdar, S. and Laha, A. K., 2020. Clustering and classification of time series using topological data analysis with applications to finance. *Expert Systems with Applications*, 162, pp.113868. https://doi.org/10.1016/j.eswa.2020.113868

24. Goodell, J. W., Kumar, S., Lim, W. M. and Pattnaik, D., 2021. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *J. of Behavioral and Experimental Finance*, 32, pp.100577. https://doi.org/10.1016/j.jbef.2021.100577

25. Lai, M., 2022. *Smart financial management system based on data mining and man-machine management*. Wireless communications and mobile computing, 1, p.2717982. https://doi.org/https://doi.org/10.1155/2022/2717982

26. Karami, A. and Moradi. M., 2006. Investigation of linear and non-linear relationships between financial ratios and stock returns in Tehran Stock Exchange. *Accounting and Auditing Reviews*. *3*(2), pp.84-102. https://doi.org/10.1108/eb043404

27. Namazi, M. and Rostami, N., 2006. The relationship between financial ratios and stock returns in Tehran Stock Exchange Market. *The Accounting and Auditing Review*, *13*(44), pp.105-127.

28. Crammer, K. and Singer, Y., 2022. *Support vector machines.* Retrieved September 27, 2022, from https://scikitlearn/stable/modules/svm.html

29. Jayanthi, M. K., 2009. Object oriented analysis and design of learning objects and applications of agent based reusable learning objects in e-Learning system design. *Computer Science and Engineering*: *King*

Khalid University. https://www.researchgate.net/publication/344672573_Object_Oriented_Analysis_and_Design_of_Learning_Objects_Application_of_Agent_Based_Reusable_Learning_objects_in_e-Learning_System_Design

30. Ciano, T., 2024. Bitcoin price prediction and machine learning features: New financial scenarios, Reference Module in Social Sciences, *Elsevier*, ISBN 9780443157851, From https://doi.org/10.1016/B978-0-44-313776-1.00194-X.

31. Dudek, G., Fiszeder, P., Kobus, P. and Orzeszko, W., 2024. Forecasting cryptocurrencies volatility using statistical and machine learning methods: A comparative study, *Applied Soft Computing*, 151, pp.111132. https://doi.org/10.1016/j.asoc.2023.111132.