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

نوید جواهری (کارشناسی ارشد مهندسی مالی، دانشکده فنی مهندسی، دانشگاه میبد) navid.xperia2014@gmail.com
نجمه نشاط[1] (دانشیار گروه مهندسی صنایع، دانشکده فنی مهندسی، دانشگاه میبد) neshat@meybod.ac.ir
عباسعلی جعفری ندوشن (استادیار گروه مهندسی صنایع، دانشکده فنی مهندسی، دانشگاه میبد) a.jafari@meybod.ac.ir

چکیده:

افزایش سودآوری و کاهش میزان ریسک، مستلزم انتخاب مسیر هوشمند سرمایه‌گذاری ضمن بهره‌گیری از تحلیل داده می‌باشد؛ لذا امروزه ارائه تکنیکی در قالب سیستم پشتیبان تصمیم‌گیری مدیریت سبد سهام ضمن درک بهتر جایگاه علم‌داده، ضروری است که در پژوهش حاضر، علاوه بر دستیابی به این مهم، ضمن ترکیب روش‌های علم‌داده با مدل مارکویتز، مدل‌های طبقه‌بندی و پیش‌بینی نیز ایجاد شده‌اند که در تشخیص تقلب مالی شرکت‌ها موثر می‌باشند؛ همچنین به‌منظور خوشه‌بندی شرکت‌های فعال در بورس اوراق بهادار تهران، الگوریتم‌های سلسله‌مراتبی و کامینز استفاده گردیده‌اند. در خصوص طبقه‌بندی داده‌ها، الگوریتم طبقه‌بندی ماشین‌بردار پشتیبان‌خطی با دقت هفتاد درصد در مقایسه با الگوریتم‌های درخت‌تصمیم،نایوبیز،نزدیک‌ترین همسایه، پرسپترون چندلایه شناسایی شده است و در خصوص ساخت مدل‌های پیش‌بینی، درخت تصمیم با حداقل میزان خطا، در مقایسه با الگوریتم‌های نایوبیز،نزدیک‌ترین همسایه،پرسپترون چندلایه شناسایی گردیده است. داده‌های اولیه در پژوهش حاضر شامل بیست عنوان شاخص مالی و داده‌های روزانه قیمت سهام شرکت‌ها در فضای برنامه‌نویسی پایتون می‌باشد.

واژگان کلیدی: مدیریت سبد سهام، مدل‌سازی، علم‌داده، طبقه‌بندی، خوشه‌بندی، بخش‌بندی

۱. مقدمه

در بسیاری از سازمان‌ها مانند بانک‌ها، سازمان‌های مالیاتی و موسسات بزرگ حسابداری و حسابرسی داده‌های مالی جمع‌آوری می‌شوند و «با افزایش حجم داده‌های ذخیره شده در سازمان‌ها، پایگاه‌های داده و سایر مخازن اطلاعاتی، ساخت ابزارهایی برای تجزیه و تحلیل و استخراج دانش معنادار از چنین حجم وسیعی از داده‌ها بسیار حیاتی می‌شود» [1]. حال بازارهای مالی به عنوان موتور محرک اقتصاد هر کشور، به ویژه بازار مبادله سهام شرکت‌ها، بایستی 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 کرده و دستیابی به این مهم با استفاده از نرم افزارهای تحلیل داده ممکن می‌باشد؛ داده‌کاوی به عنوان زیر مجموعه‌ای از علم‌داده، برای مدیریت داده‌های گسترده و افزایش کارایی شرکت و بینش تجاری در تجارت مالی بسیار مهم است[2]؛ در حقیقت «استفاده از تکنیک‌های داده‌کاوی بر روی داده‌های مالی می‌تواند به حل چالش‌های طبقه‌بندی و پیش‌بینی کمک نموده و در عین حال فرآیند تصمیم‌گیری را برای افراد دخیل در امور مالی هر کشوری آسان‌تر نماید» [3]؛ بنابراین بررسی الگوریتم‌های داده‌کاوی و ارائه 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 نمی‌باشند. بنابراین این پرسش ضروری است که چه میزان تحلیل اطلاعات این بازارها همچون تحلیل اطلاعات سهام شرکت‌ها، با استفاده از ابزارهای نوین همچون زبان برنامه‌نویسی پایتون، همانند کشورهای پیشرفته در کشور ما نیز در حال انجام است؟ پس از بررسی پاسخ این پرسش، بایستی به دنبال ایجاد یا بهبود مسیری با هدف تسریع در فرآیند تحلیل بازارهای مالی برآمد و تا حد امکان به گسترش مرزهای دانش در این زمینه پرداخت؛ به همین علت در طی پژوهش جاری این مهم در قالب طراحی و پیاده‌سازی سیستم پشتیبان تصمیم‌گیری مدیریت سبد سهام با استفاده از ابزارهای علم‌داده با یک رویکرد ترکیبی نوین از مطرح‌ترین الگوریتم‌های داده‌کاوی، مورد هدف قرار گرفته است؛ بنابراین، هدف اصلی از پژوهش حاضر، ادغام دانش فراگرفته شده در مهندسی مالی با تجربه متخصصین علم‌داده، بمنظور ایجاد الگویی با حداکثر سرعت پیاده‌سازی و دریافت خروجی با حداقل خطا است؛ این الگو برای تحلیل‌گران و سرمایه‌گذاران در بازارهای مالی که همواره بازده و ریسک را توامان در نظر می‌گیرند و به دنبال کسب بیشترین بازده و

---

[1] نویسنده مسئول

تاریخ دریافت: تاریخ پذیرش:

تحمل کمترین ریسک به هنگام تشکیل سبد سهام هستند، راهگشا خواهد بود [4]. بنابراین در پژوهش حاضر، با توجه به اهمیت مبحث مدیریت سبد سهام به عنوان یکی از زمینه‌های اصلی و تاثیرگذار در مهندسی مالی بر اساس پیشینه پژوهش‌های صورت پذیرفته، یک سیستم پشتیبان تصمیم‌گیری در مدیریت سبد سهام با بهره‌گیری از ابزارهای علم‌داده در قالب ترکیب الگوریتم‌های مطرح در این زمینه شامل بخش‌بندی، طبقه‌بندی، پیش‌بینی و مدل مارکویتز، ارائه شده است؛ در حقیقت تکنیکی در قالب یک سیستم پشتیبان تصمیم‌گیری در حل مسائل مدیریت سبد سهام با ترکیبی از تکنیک‌های داده‌کاوی ارائه شده است، این مهم در شرایطی است که نیاز به تحلیل حجم زیادی از اطلاعات در ابعاد بسیار بالا در قالب کلان داده‌ها و اخذ خروجی در مدت زمان محدود وجود دارد. مهم آنکه رویه در نظر گرفته شده به گونه‌ای است که تحلیل‌گر قادر خواهد بود با در نظر گرفتن جامعه آماری و شاخص‌های مورد نظر خود، اقدام به حل مسائل طبقه‌بندی، خوشه‌بندی یا پیش‌بینی نماید. از ویژگی‌های برجسته تکنیک پیشنهادی، قابلیت تعمیم‌دهی آن می‌باشد، زیرا امکان تنظیم پارامترهای مدل بر حسب فروض مسئله مورد نظر در مدل‌های نهایی، فراهم شده است. در پژوهش حاضر، ابتدا مروری بر پژوهش‌های پیشین صورت می گیرد و خلاصه‌ای از آنها به شکل کاربردی در جهت کمک به درک جایگاه علم‌داده و تعیین جزییات مسئله اصلی، ارائه می شود. در گام بعدی، به بررسی مبانی نظری مورد نیاز می پردازیم، سپس داده‌های مورد نیاز به منظور پیاده‌سازی الگوریتم‌ها با هدف ساخت مدل‌های نهایی جمع‌آوری و گام‌های پیش‌پردازش داده برای داده‌های مذکور اجرا می‌شود. در گام بعدی به پیاده‌سازی و سنجش مدل‌های خوشه‌بندی، طبقه‌بندی و پیش‌بینی پرداخته می‌شود که در این گام مدل‌های مورد نظر هدف اصلی پژوهش ایجاد، ترکیب و سنجیده خواهند شد و در آخرین مرحله، تفسیری در قالب تحلیل خروجی مدل‌های ساخته شده، به منظور روشن‌تر شدن کاربرد سیستم ایجاد شده، ارائه می شود.

۲. مبانی نظری

۱.۲. مدیریت سبد سهام یا مدیریت پرتفوی: لغت پرتفوی به بیان ساده به ترکیبی از دارایی‌ها گفته می‌شود که توسط سرمایه‌گذار برای سرمایه‌گذاری تشکیل می‌شود؛ در این تعریف، سرمایه‌گذار یک فرد یا یک موسسه است. از نظر تکنیکی، پرتفوی در برگیرنده مجموعه‌ای از دارایی‌های واقعی و مالی سرمایه‌گذاری شده یک سرمایه‌گذار است. پرتفویی ایده‌آل است که با فرض یکسان بودن بازده بین تمامی پرتفوهای مورد انتظار، واریانس یا ریسک آن از همه کمتر باشد، یا از بین تمامی پرتفوهایی با واریانس یا ریسک مشابه، بیشترین بازده را داشته باشد. مارکویتز به صورت کمی نشان داد که چرا و چگونه تنوع‌سازی پرتفوی می‌تواند باعث کاهش ریسک پرتفوی شود؛ او در مسئله انتخاب پرتفوی استاندارد خود، معتقد است سرمایه‌گذاران انتخاب‌های خود را بر اساس دو معیار بازدهی و ریسک انجام می‌دهند. در این مدل، بازدهی پرتفو معادل میانگین وزنی بازدهی اجزای پرتفوی و ریسک آن برابر با انحراف معیار بازدهی دارایی‌های موجود می‌باشد. 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. مفروضات مدل مارکویتز عبارت‌اند از: ۱. سرمایه‌گذاران ریسک‌گریز هستند و مطلوبیت مورد انتظار افزایشی است؛ ۲. سرمایه‌گذاران پرتفو را بر اساس میانگین و واریانس مورد انتظار 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 [4]. بنابراین می‌توان گفت بازده یک سبد، بازده وزنی سهام پایه است. بر اساس مدل مارکویتز، اگر σ_p ریسک پرتفوی باشد و n تعداد سهام پایه باشد، سپس: تابع ۱)

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

که در معادله فوق σij برابر است با کوواریانس بین قیمت سهم i و j همچنین Wi و Wj وزن تخصیص داده شده به سهم i و j می‌باشد. وزن تخصیص داده شده در تابع(۱) در حقیقت وزن هر سهم در خوشه می‌باشد که به روش زیر محاسبه می‌شود: تابع ۲)

$$W_j = \frac{M_j}{\sum M_j}$$

که در فرمول فوق Mj ارزش بازار سهم j می‌باشد. به بیان ساده‌تر بازدهی هر خوشه برابر است با میانگین موزون بازدهی سهام موجود در آن خوشه که وزن هر سهم، ارزش بازار آن سهم در آن خوشه است [4].

۲.۲. علم‌داده و داده‌کاوی[۲]: علم‌داده مبتنی بر تحلیل داده‌ها و شامل علوم گوناگونی همچون تحلیل کلان داده‌ها، داده‌کاوی و ... می‌باشد، همچنین به زیردسته‌هایی از جمله هوش مصنوعی[۳]، یادگیری ماشین[۴]، داده‌کاوی[۵]، یادگیری عمیق[۶] و ... تقسیم‌بندی می‌شود؛ همچنین داده‌کاوی به عنوان فرآیند کشف دانش[۷] یا فرآیند کشف داده[۸] نیز شناخته می‌شود؛ فرآیندی است که مستلزم مطالعه و تجزیه و تحلیل داده‌ها از منابع مختلف، ارزیابی و ترکیب آنها به اطلاعات مفیدتر و مهم‌تر می‌باشد. داده‌کاوی به منظور پیش‌بینی، محبوب‌ترین نوع داده‌کاوی با بیشترین خروجی برای فعالیت‌های تجاری است [3]. روش‌های اصلی داده‌کاوی در پژوهشی به ۶ گروه اصلی تقسیم‌بندی شده‌اند که عبارت‌اند از: طبقه‌بندی و پیش‌بینی، خوشه‌بندی، تجزیه و تحلیل‌توالی، تشخیص داده‌های پرت و متن‌کاوی؛ که از میان روش‌های مذکور، خوشه‌بندی از جمله روش‌هایی است که به‌طور گسترده و فشرده مورد مطالعه بسیاری از محققان داده‌کاوی قرار گرفته است [5]. روش‌های داده‌کاوی در پژوهش حاضر نیز به ۳ دسته کلی بخش‌بندی یا خوشه‌بندی، طبقه‌بندی و پیش‌بینی تقسیم‌بندی شده است. داده‌کاوی عمدتا برای اهداف تجاری استفاده می‌شود و در قالب ارائه تکنیک مطرح می‌شود؛ خروجی آن به عنوان یک الگو قابل استخراج است و نوع داده‌ها در آن عموما ساختار

یافته است؛ این در حالی است که علم داده عمدتا 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 ساختار باشد [6]. در پژوهشی در سال ۲۰۰۸، فرآیند کشف دانش از طریق داده‌کاوی را به چهار عنوان زیر تقسیم‌بندی نموده است:

۱. انتخاب، ۲. پیش‌پردازش، ۳. داده‌کاوی و ۴. تفسیر [7]؛ در پژوهشی دیگر در سال ۲۰۲۰ نیز این مراحل را به ۹ مرحله تقسیم‌بندی نموده‌اند: 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. ارایه نتایج [8]. با توجه به مرور ادبیات و با هدف تلخیص و ترکیب ادبیات مروری گوناگون و با بررسی تجربه خبره در این زمینه، در پژوهش حاضر، مراحل داده‌کاوی در قالب پنج مفهوم مستقل با توجه به مباحث کتاب راهنمای علم‌داده پایتون[۹] [9] بررسی شده است.

۳.۲. سنجش خطای مدل‌ها: این مرحله شامل سنجش مدل‌ها و الگوهای ساخته‌شده می‌باشد که به وسیله شاخص‌های اندازه‌گیری خطا همچون مجذور میانگین مربعات خطا[۱۰]، میانگین قدرمطلق خطا[۱۱]، ضریب تعیین[۱۲]، میانیگن مربعات خطا [۱۳]و ماتریس اغتشاش[۱۴] الگوهای ساخته شده مورد ارزیابی قرار خواهندگرفت. در ادامه توابع این شاخص‌ها به جهت یادآوری، قابل مشاهده هستند:

تابع ۳)

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

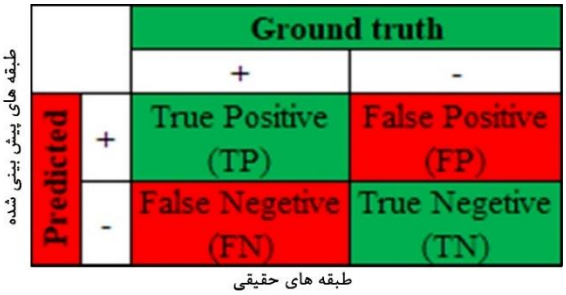
تابع ۴)

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

تابع ۵)

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

در فرمول فوق $\hat{y}$ برابر است با مقدار پیش‌بینی شده از y یا مقادیر ستون هدف و $\overline{y}$ برابر است با میانگین مقادیر ستون هدف یا مقادیری که قرار بر پیش‌بینی آنها می‌باشد. همچنین ماتریس آشفتگی نیز که در شکل(۱) قابل مشاهده است، چهار معیار مورد نظر خود را در قالب چهار نسبت از طریق مقایسه برچسب طبقه‌های پیش‌بینی شده با برچسب طبقه‌های حقیقی محاسبه می‌نماید که در پژوهش حاضر از معیار دقت[۱۵] که ترکیبی است از سه معیار دیگر، استفاده شده است.



شکل ۱- ماتریس آشفتگی

در ماتریس فوق مقدار ضریب خطای معیار دقت برابر است با: تابع۶)

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

۴.۲. الگوریتم درخت تصمیم: ساختار درخت تصمیم شبیه به ساختار یک فلوچارت می‌باشد، درخت‌تصمیم یک تکنیک طبقه‌بندی یا پیش‌بینی است که بالاترین گره در درخت، ریشه می‌باشد و برگ‌ها نشان‌دهنده دسته‌ها و توزیع دسته‌ها هستند. هر آزمون آزمایشی بر یک گره، یک ویژگی را مشخص می‌کند و هر شاخه‌ای که از این گره خارج می شود نتیجه این آزمون را نشان می‌دهد [10]؛ در این روش برخلاف شبکه‌های عصبی، الزامی برای عددی بودن داده‌ها در درخت‌تصمیم وجود ندارد [11].

۵.۲. نایو بیز: نایوبیز تکنیک و روشی از خانواده روش‌های طبقه‌بندی بر مبنای احتمال بر اساس بکارگیری قضیه بیز و فرض اسقلال بین متغیرها است. اگر مشاهدات و داده‌ها از نوع پیوسته باشند، از مدل احتمالی با توزیع گاوسی یا نرمال برای متغیر داده‌های جمع‌آوری شده می‌توان بهره جست [12].

۶.۲. نزدیک‌ترین همسایه: این روش به عنوان یکی از ساده‌ترین روش‌های طبقه‌بندی به عنوان طبقه‌بندی‌کننده تنبل[۱۶] نیز شناخته می‌شود، زیرا مرحله‌ای خاص جهت یادگیری ندارد و به هنگام طبقه‌بندی از تمامی داده برای یادگیری استفاده می‌نماید. این روش متعلق بر کلاس یادگیری مبتنی بر نمونه است؛ در این روش، طبقه‌بندی فقط با نگاه کردن به K نزدیک‌ترین مثال‌ها در مجموعه داده‌های آموزشی (عموما از نظر معیار فاصله اقلیدسی[۱۷] یا بر اساس انوع دیگری از معیارهای فاصله همچون همینگ[۱۸] یا منهتن[۱۹]) در مورد داده‌ای که طبقه‌بندی برای آن در حال پیاده‌سازی است، انجام می‌شود؛ سپس با توجه به نمونه‌های مشابه K، محبوب‌ترین هدف (بر اساس رای اکثریت) به عنوان برچسب طبقه‌بندی انتخاب می‌شود [13].

۷.۲. ماشین‌بردار پشتیبان: یک روش یادگیری با نظارت[۲۰] است که برای طبقه‌بندی استفاده می‌شود؛ این روش مشابه شبکه‌های عصبی قادر است تقریبی با درجه دقت مطلوب برای هر تابع چند متغیره به دست آورد. بنابراین به منظور مدل‌سازی سیستم‌ها و فرآیندهای غیرخطی و پیچیده، از جمله شناسایی فعالیت‌های مرتبط با تقلب مالی، بسیار مفید است [9].

۸.۲. پرسپترون چندلایه: به منظور حل بسیاری از مسائل ریاضی که بر اساس معادله‌های پیچیده غیرخطی حل می‌شوند، یک شبکه عصبی پرسپترون چند لایه می‌تواند به سادگی با تعریف اوزان و توابع مناسب مورد استفاده قرارگیرد. یک پرسپترون چندلایه از سه نوع لایه تشکیل می‌شود که شامل لایه ورودی[۲۱]، لایه‌های میانی یا پنهان[۲۲] و لایه خروجی[۲۳] است. نرون‌ها عناصر تشکیل‌دهنده لایه‌ها در شبکه‌های عصبی می‌باشند؛ عناصر هر لایه با تمام عناصر لایه‌های دیگر در ارتباط است ولی با دیگر عناصر همان لایه ارتباطی ندارد. موضوع اصلی مورد بحث در این شبکه‌ها، تعیین تعداد لایه‌های پنهان و تعداد نورون‌های هر لایه می‌باشد. در یک 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 سیگنال های ورودی متعلق است که هر سیگنال در وزن متناظر w_i ضرب شده و مقادیر حاصل در داخل نورون با هم جمع می‌شوند و مقدار خروجی محاسبه می‌گردد: [14] تابع ۷)

$$NET = X_1W_1 + X_2W_2 + ... + X_nW_n = \sum X_i.W_i$$

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

۱۰.۲. بازده سهام:

اگر اطلاعات مربوط به جریانات نقدینگی(سود نقدی، مزایای حق‌تقدم و سهام‌جایزه) در دسترس نباشد و در تابع فوق این ارقام وارد نشود، خروجی کسر عبارت است از بازده قیمت سهام که در پژوهش حاضر با توجه به کمبود اطلاعات، این مقدار محاسبه گردیده است. به عبارتی ساده‌تر، بازده کل سهام برابر است با مجموع بازده نقدی و بازده قیمت [15].

حال لازم به ذکر است که سوال و هدف ابتدایی این پژوهش آن است که چه میزان می‌توان به بهبود درک جایگاه الگوریتم‌های داده‌کاوی با دیدگاه مالی در بستر علم‌داده ضمن مطالعه پژوهش‌های پیشین پرداخت که به این منظور به مرور دقیق پژوهش‌های پیشین پرداخته شد و در پایان این مرور، شکافی در قالب سوال اصلی پژوهش شناسایی گردید که این شکاف عبارت است از نحوه ارائه 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 داده‌کاوی ضمن ترکیب با مدل مارکویتز؛ که در همین راستا چالشی با هدف نحوه بکارگیری الگوریتم‌های خوشه‌بندی، طبقه‌بندی و پیش‌بینی برای پژوهشگران این پژوهش در مسیر تشکیل سبد سهام، ایجاد گردید که طی مسیر پژوهش، این سوالات مورد بررسی و پیاده‌سازی قرار گرفته‌اند.

۳. پیشینه پژوهش

۱.۳. پیشینه پژوهش‌های داخلی

در زمینه خوشه‌بندی شرکت‌های فعال در بازار سهام، پژوهشی در سال ۲۰۱۱ بررسی شده است که در آن با استفاده از داده‌های موجود در صورت‌های مالی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران، با استفاده از روش خوشه‌بندی C-means به بخش‌بندی این شرکت‌ها پرداخته شده است که نتایج نشان‌داد بخش بزرگی از شرکت‌های مذکور در سبد سهام ترکیبی قرار گرفته‌اند، در حالیکه تمایل رفتاری آن‌ها به سهام رشدی است؛ در این پژوهش از طریق مطالعه کتب و مقالات بین‌المللی و در نهایت مشورت با پژوهشگران، شاخص‌های انتخابی برای بخش‌بندی سهام شرکت‌ها انتخاب شده‌اندکه عبارت‌اند از: نسبت قیمت به درآمد هر سهم[۲۴]، ارزش بازار به ارزش دفتری[۲۵]، شاخص ریسک سیستماتیک[۲۶]، سود هر سهم[۲۷]، سود تقسیمی، مومنتوم[۲۸]، نرخ رشد پنج ساله درآمد هر سهم، میزان تغییر پنج ساله نسبت قیمت به درآمد هر سهم، بازده حقوق صاحبان سهام[۲۹]، نسبت بدهی به حقوق صاحبان سهام[۳۰]، نسبت قیمت سهام به فروش[۳۱] [16]. در پژوهشی دیگر در سال ۲۰۱۵ نیز پژوهشگر به دنبال حذف نویز اطلاعات ماتریس ضرایب همبستگی از طریق روش‌های خوشه‌بندی به منظور بهینه‌سازی سبد سرمایه‌گذاری بوده است؛ به این منظور از دو روش خوشه‌بندی اتصال واحد و اتصال میانگین و بر اساس بازده روزانه ۸۰ شرکت بورس اوراق بهادار تهران در بازه زمانی ۱۳۸۵ تا اسفند ۱۳۹۲ استفاده نموده است؛ نتایج نشان داد که در نهایت ریسک کمتری به سرمایه‌گذار تحمیل می‌شود و همچنین با هدف تحلیل حساسیت نتایج دو روش، یک بررسی بوت استرپینگ در تعداد سهام و افق‌های زمانی مختلف انجام گرفت که در بیشتر حالت‌ها، بهبود عملکرد سبد سرمایه‌گذاری مشاهده شده است[17]. در پژوهشی در سال ۲۰۲۰ نیز، ترکیبی از روش‌های خوشه‌بندی و پیش‌بینی به‌کار گرفته شده است به صورتی که یک مدل بهینه‌سازی پرتفوی مبتنی بر پیش‌بینی، برای بررسی نتایج وضعیت پرتفوی در بورس اوراق بهادار تهران ایجاد گردیده است. بدین منظور ابتدا از روش داده‌کاوی برای پیش‌بینی فرآورده‌های نفتی و صنایع شیمیایی با استفاده از داده‌های خوشه‌بندی بازار سهام استفاده شده است؛ در گامی دیگر به منظور تخمین هر شاخص صنعت با استفاده از تابع پایه شعاعی و شبکه عصبی پرسپترون چندلایه، از عوامل موثری مانند قیمت نفت خام، نرخ ارز، نرخ بهره جهانی، قیمت طلا و شاخص S&P500 بهره گرفته شده است. در نهایت، با مقایسه نسبت‌های

اعتبارسنجی در یک بازار صعودی با استفاده از الگوریتم‌های خوشه‌بندی K-means، شبکه عصبی و فازی، بهترین الگوریتم برای پیش‌بینی شاخص‌ها برای هر صنعت شناسایی گردیده است. نتایج نشان داد که الگوریتم پرسپترون چند لایه بالاترین دقت را داشته و بهترین گزینه برای وضعیت پرتفوی می‌باشد؛ با این حال، الگوریتم فازی C-means بهترین خوشه ها 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[18]. در پژوهشی دیگر در سال ۲۰۱۹ پس از مطالعه مقالات بین‌المللی و بررسی شاخص‌های گوناگون و مشورت با اساتید و صاحب نظران در زمینه مالی و حسابداری، پنج شاخص نهایی برای خوشه‌بندی انتخاب شد که عبارت‌اند از: نسبت قیمت به درآمد هر سهم، شاخص ریسک سیستماتیک، سود هر سهم، بازدهی[۳۲] و نسبت ارزش بازار به ارزش دفتری[۳۳] که به منظور قابل انجام بودن عمل خوشه‌بندی و محاسبه فاصله درست میان سهام، داده‌ها را بر اساس شاخص‌های اصلی نسبت قیمت به درآمد هر سهم و نسبت ارزش بازار به ارزش دفتری فیلتر نموده‌اند و تنها سهام حاوی این ویژگی‌ها برای خوشه‌بندی در نظر گرفته شده است[4]؛ در سایر پژوهش‌ها در این زمینه نیز می‌توان به پژوهشی در سال ۲۰۱۹ اشاره نمود که در آن از پنج نسبت مالی آلتمن که از سایت مرکز بورس اوراق بهادار تهران استخراج شده‌اند به شرح زیر استفاده شده است: نسبت سرمایه در گردش به کل دارایی‌ها، نسبت سود قبل از مالیات به کل دارایی‌ها، نسبت سود انباشته به کل دارایی‌ها، نسبت ارزش بازار سهام به ارزش دفتری کل بدهی‌ها، نسبت فروش کل به کل دارایی‌ها؛ نتیجه نهایی بیان می‌کند که در این مسئله روش‌های فرا ابتکاری در مقایسه با روش‌های معمول، کاراتر عمل نموده‌اند و به بهینه سراسری منجر شده است؛ هچنین نتایج مذکور با نتایج حاصل از تفکیک شرکت‌های عضو بورس بهادار تهران با استفاده از روش تعیین ورشکستگی آلتمن مقایسه شده است و توسط این روش نیزتایید شده است[19]. در پژوهش دستگیر و شفیعی در سال ۲۰۱۹، پیش‌بینی ورشکستگی مشهورترین موضوع در زمینه کاربرد روش‌های داده‌کاوی در حل مسائل مالی می‌باشد و از بین روش‌های موجود در داده‌کاوی نیز شبکه عصبی بیشترین استفاده را به خود اختصاص داده است[8]. در پژوهشی دیگر در سال ۲۰۱۹، عملکرد پنج مدل محبوب آماری و یادگیری ماشین را با هدف تشخیص تقلب صورت‌های مالی در مقایسه با یکدیگر بررسی نموده است؛ اهداف پژوهش شرکت‌هایی هستند که بین سال‌های ۲۰۱۱ و ۲۰۱۶ صورت‌های مالی متقلبانه و غیر متقلبانه را تجربه کرده‌اند. نتایج نشان می‌دهد که شبکه عصبی مصنوعی[۳۴] نسبت به شبکه بیزی[۳۵]، تحلیل تشخیصی[۳۶]، رگرسیون لجستیک[۳۷] و ماشین‌بردار پشتیبان عملکرد بهتری دارد. هدف این مقاله ارائه یک مدل کمی برای شناسایی گزارشگری مالی متقلبانه در شرکت‌های ایرانی با استفاده از الگوریتم‌های طبقه‌بندی است. این مدل، تلاش برای پنهان کردن اطلاعات و یا ارائه اطلاعات نادرست در پرونده‌های سالانه با اوراق بهادار و بورس تهران را شناسایی می‌کند. به بیان این پژوهش از میان نوزده فاکتور پیش‌بینی کننده بررسی شده، تنها ۹ پیش‌بینی‌کننده به‌طور پیوسته توسط الگوریتم‌های طبقه‌بندی مختلف انتخاب و استفاده می‌شوند که عبارت‌اند از: بهره‌وری کارکنان، حساب‌های دریافتنی به فروش، بدهی به حقوق صاحبان سهام، موجودی به فروش، فروش به کل دارایی‌ها، بازده حقوق صاحبان سهام، بازده فروش، بدهی‌ها به هزینه های بهره، و دارایی ها به بدهی‌ها. این یافته‌ها، پژوهش‌های تقلب در صورت‌های مالی را گسترش می‌دهند و می‌توانند توسط متخصصان و تنظیم‌کننده‌ها برای بهبود مدل‌های ریسک تقلب استفاده شوند. نرم‌افزار مورد استفاده در این پژوهش، IBM SPSS Modeler 18 می‌باشد[14]. نتایج حاصل از پژوهشی در سال ۲۰۱۹ نیز نشان داده است که توانگری مالی با دقت قابل قبولی، پیش‌بینی پذیرند و مدل استخراج شده با استفاده از درخت تصمیم دقت و قابلیت بسیار بالایی در تخمین را دارا می‌باشد. از میان ۴ مدلی که در این پژوهش مورد استفاده قرار گرفته‌اند(درخت تصمیم، نایوبیز، شبکه عصبی، نزدیک‌ترین همسایه)، درخت تصمیم از بالاترین دقت و کمترین مقدار خطا برخوردار است و همچنین متغیرهای تصمیم به خوبی متغیر هدف را تعیین می‌کنند. بدترین عملکرد نیز مربوط به مدل نایوبیز شناسایی گردیده است[20].

۲.۳. پیشینه پژوهش‌های خارجی

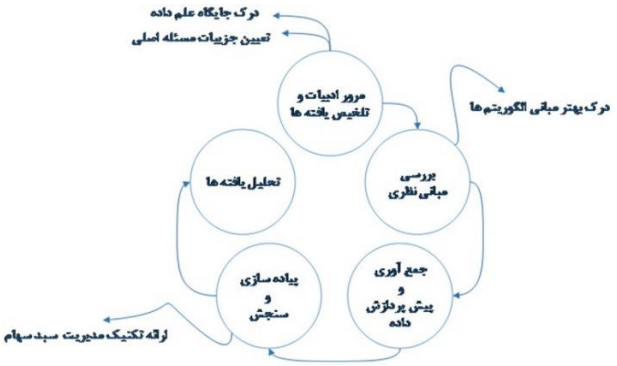
در سال ۲۰۱۰ در مقاله‌ای تحت عنوان «خوشه‌بندی داده‌های بازار سهام هند جهت مدیریت پرتفو» از داده‌کاوی برای طبقه‌بندی سهام در خوشه‌های مختلف استفاده شد، پس از طبقه‌بندی، سهام مورد نظر برای پرتفوها از بین این خوشه‌ها انتخاب شد؛ این روش منجر به کاهش ریسک متنوع‌سازی سبد سهام شده است؛ نتایج نشان داد روش تحلیلی K-means خوشه های متراکم‌تری نسبت به روش‌ها SOM[۳۸] و فازی ایجاد می‌کند. شاخص‌های مورد نظر برای خوشه‌بندی در پژوهش مذکور عبارت‌اند از: نسبت قیمت به درآمد هر سهم، نسبت قیمت به ارزش دفتری، نسبت قیمت به سود نقدی هر سهم[۳۹]، ارزش شرکت به سود شاخص عملکرد با توجه به سرمایه بازار[۴۰] [15]. حال مهم آنکه اگرچه روبه‌رویی با کلان داده‌ها دانش بیشتری را نیازمند است ولی مزایایی نیز دارد؛ به بیان پژوهشی در سال ۲۰۱۵: "کلان داده‌ها نیز از پنج طریق باعث تغییر امور مالی می‌شوند: ایجاد شفافیت، تجزیه و تحلیل ریسک، تجارت الگوریتمی، استفاده از داده‌های مصرف کننده و تغییر فرهنگ [21]. در سال ۲۰۱۸ در پژوهشی بر روی استفاده از داده‌کاوی در پیش‌بینی سهام، مدیریت پرتفوی، تحلیل ریسک سرمایه‌گذاری، همچنین پیش‌بینی ورشکستگی و نرخ ارز خارجی و شناسایی تقلب

مالی تمرکز شده است [3]. در پژوهشی در سال ۲۰۱۹ نیز بیان شده است که اگرچه فضای اقتصادی مدرن امروز، غرق در داده‌های بزرگ است، با این حال تحلیل‌های اقتصادی هنوز در مراحل ابتدایی خود قرار دارند و بسیاری از مسائل مهم به طرز ناقصی بررسی شده‌اند یا اصلا بررسی نشده‌اند؛ به همین منظور پژوهشگر در پژوهش مذکور، پژوهش‌های جدید اقتصادسنجی را در مورد داده‌های بزرگ و مدل‌سازی سری‌های زمانی ارائه داده است. به شکل خلاصه پژوهشگر معتقد است که کلان‌داده در تحلیل‌های اقتصادی و مدل‌سازی اقتصادی تاثیر بسزایی دارد [22]. در پژوهشی دیگر در سال ۲۰۲۰ نیز به خوشه‌بندی و طبقه‌بندی سری‌های زمانی با استفاده از تحلیل داده‌های توپولوژیکی با کاربردهای مالی پرداخته است [23]. در پژوهشی در سال ۲۰۲۱، بازنگری کاملی بر این مقالات در امور مالی ارائه شده است؛ به بیان پژوهش مذکور مباحثی که برای هر دو دسته هوش مصنوعی و یادگیری ماشین مطرح هستند در سه دسته زیر می‌باشند: ۱.ساخت پرتفوی، ارزش‌گذاری و رفتار سرمایه‌گذار، ۲. تقلب و ناتوانی مالی، ۳. پیش‌بینی و برنامه‌ریزی. در پژوهش مذکور مقالاتی معرفی شده‌اند که به بررسی شناسایی تقلب مالی با استفاده از تکنیک‌های داده‌کاوی پرداخته‌اند که با توجه به تعدد این مقالات، از بیان جزییات آنها در مرورپیشینه حاضر خودداری شده است [24]. همچنین با هدف توسعه پژوهش‌های پیشین، پژوهشی در سال ۲۰۲۲ بررسی شده است که در گام اول آن، معماری یک سیستم مدیریت مالی هوشمند به طور کامل مورد بررسی قرار می‌گیرد و معماری جدیدی از یک سیستم پشتیبانی مدیریت مالی هوشمند مبتنی بر داده‌کاوی توسعه می‌یابد؛ در گام دوم، به تعریف و ساختار انبار داده و داده‌کاوی و همچنین نحوه استفاده از استراتژی و فناوری داده‌کاوی در مدیریت مالی پرداخته شده است. به عقیده پژوهشگر، داده‌کاوی در ارتباط با فناوری و توسعه یک الگوریتم داده‌کاوی هوشمند در حال بررسی است؛ نقص‌های الگوریتم داده‌کاوی هوشمند از طریق تجزیه و تحلیل و خلاصه سازی الگوریتم کشف می‌شود و در نهایت یک الگوریتم بهبودیافته برای رفع ایرادات در این پژوهش پیشنهاد شده است. در گام بعدی آزمایش‌های مرتبط بر روی الگوریتم بهبودیافته انجام می‌شود و آزمایش‌ها نشان می‌دهد که الگوریتم مذکور مزایای خاصی دارد. سپس چهارچوب طراحی اولیه برای مدیریت مالی هوشمند ایجاد می‌شود و کاربرد یک مدل داده‌کاوی در سیستم پشتیبانی تصمیم معرفی می‌شود [25]. امسال نیز در پژوهشی در سال ۲۰۲۴ بیان شده است که پیشرفت‌های تکنولوژیکی منجر به دیجیتالی شدن اقتصاد شده است، از همین رو یک مدل یادگیری ماشین با استفاده از شبکه‌های عصبی برای پیش‌بینی قیمت بیت‌کوین ارائه شده است که در این مسیر اطلاعات قیمت پایانی ۶۰ روز این رمز ارز به منظور پیش‌بینی روز ۶۱ام جمع‌آوری شده است، نتیجه آنکه انجام یک مطالعه عمیق در مورد همبستگی بین ارزهای رمزنگاری شده، تجزیه و تحلیل را برای هدایت خواندن به این چارچوب مرتبط با پیش‌بینی برای یک سناریوی پیچیده مالی تقویت می‌کند [30]. در پژوهشی دیگر در سال ۲۰۲۴، ۱۲ روش پیش‌بینی نوسانات ارزهای دیجیتال به شکل جامع مطالعه شده‌اند و نتیجه‌گیری شده است 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 می‌توانند به خوبی مدل‌های پیچیده عمل کنند [31].

با توجه به خلاصه ارائه شده پژوهش‌های مروری توسط نویسندگان پژوهش حاضر، روند کلی مدل‌سازی و شاخص‌های مورد نیاز در این پژوهش شناسایی گردید که در بخش‌های بعدی این موارد مطرح گردیده‌اند.

۴. روش شناسی پژوهش:

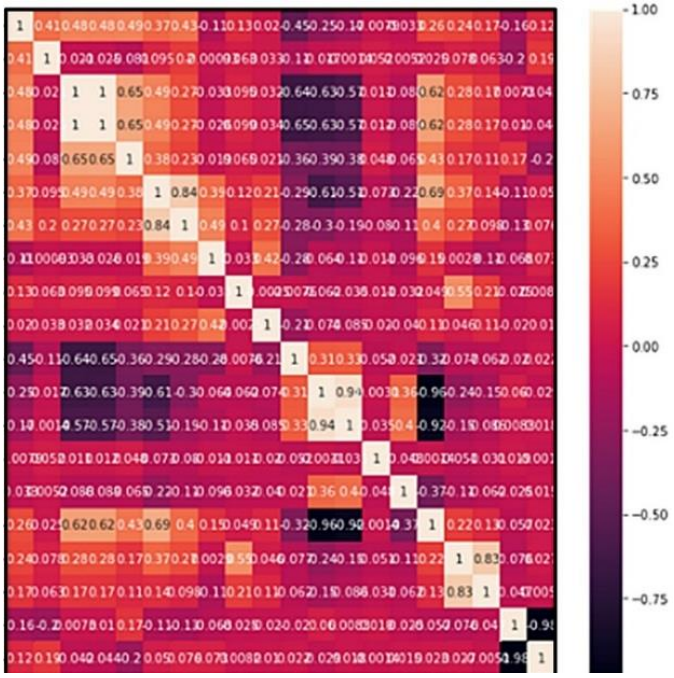
فرآیند کلی اجرای پژوهش در قالب شکل(۲) به جهت درک بهتر مسیر، قابل مشاهده است:



شکل ۲- نمودار فرآیند اجرای پژوهش

در مرحله جمع‌آوری داده‌ها به عنوان اولین گام اجرایی، داده‌های مربوط به ۱۸ نسبت مالی شرکت‌های فعال در بازار بورس اوراق بهادار تهران تهیه می شود؛ از ۱۸ نسبت مذکور، ۹ نسبت از مرور ادبیات پژوهش استخراج شده است. لازم به ذکر است که از میان تمامی نسبت‌های مالی شناسایی‌شده در مرور ادبیات، با توجه به عدم امکان استخراج داده و عدم امکان محاسبه نسبت‌ها برای تعداد بالای شرکت‌های مورد نظر در بازه زمانی انجام پژوهش (بدون بهره‌گیری از نرم‌افزارهای بانک استخراج داده)، ۹ نسبت مالی نهایی مذکور گزینش شده‌اند، ۹ نسبت دیگر نیز به جهت دارا بودن پژوهش حاضر از نوآوری لازم نسبت به پژوهش‌های پیشین، در پژوهش حاضر در نظر گرفته‌شده‌اند؛ این نسبت‌ها در حقیقت نسبت‌هایی هستند که همواره در منابع آموزشی علمی مرور شده توسط پژوهشگران، به چشم می‌خورند. تعداد شرکت‌های مورد نظر پژوهش حاضر نیز با توجه به اینکه ۸۹۲ نماد فعال در قالب ۵۵۸ شرکت در بازار بورس تهران موجود می‌باشد، با استفاده از وبسایت "شرکت مدیریت فناوری بورس تهران۴۱" اسامی ۸۹۲ نماد بازار بورس اوراق بهادار تهران

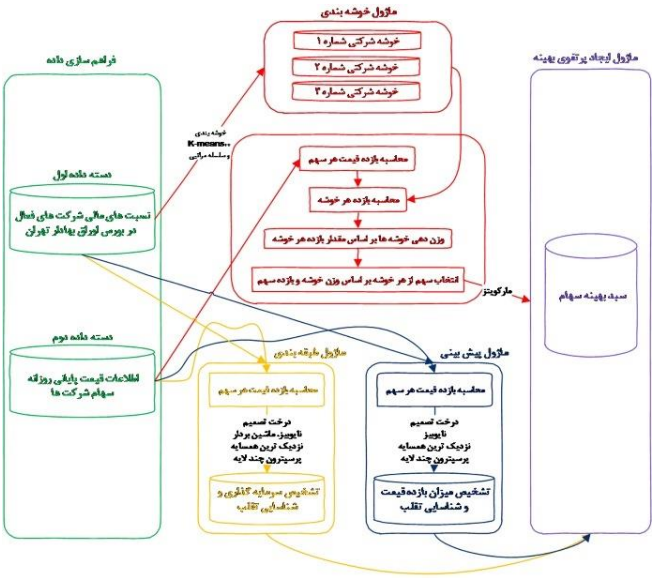
در قالب "لیست همه نمادهای بازار عادی" استخراج گردیده است، ضمن بهره‌گیری از جدول مورگان فریمن، تعداد نمونه مورد نیاز حداقل ۲۲۶.۴۴ شرکت محاسبه گردید که با توجه به حداقل تعداد بدست آمده و پس از بررسی داده‌های خام در مرحله پیش‌پردازش، تعداد ۲۷۱ شرکت در پژوهش حاضر مورد بررسی قرار گرفته است. با توجه به توضیحات مذکور، نسبت‌های مورد نظر پژوهش حاضر به شرح زیر می‌باشند: سود انباشته به دارایی‌ها، سود قبل از بهره و مالیات به دارایی، گردش دارایی‌ها[19] بازده دارایی‌ها[27],[26],[4] بدهی به حقوق صاحبان سهام[16] حاشیه سود خالص و نسبت بدهی[26] گردش دارایی‌های ثابت و نسبت جاری[27] و سایر نسبت‌های مالی مطرح شامل: حاشیه سود عملیاتی، حاشیه سود قبل از مالیات، گردش حساب‌های دریافتنی، دوره وصول مطالبات، جمع بدهی‌ها به جمع دارایی‌ها، نسبت بدهی‌های بلند مدت، نسبت نقدینگی، نسبت جریان نقد سرمایه‌گذاری به درآمد، نسبت جریان نقد تامین مالی به درآمد؛ نسبت‌های مالی مذکور از طریق نرم‌افزار "بورس ویو[۴۲]" در سال مالی ۱۳۹۹ استخراج شده‌اند که در خصوص این دسته داده‌ها، داده استخراجی شامل ۱۸ نسبت مالی در ستون و ۲۷۱ عنوان شرکت در سطر می‌باشد؛ به منظور جمع‌آوری داده‌ها ابتدا اطلاعات ۴۹۹ شرکت با سال مالی منتهی به برج‌های متفاوت استخراج شد، سپس نسبت‌های موجود برای ۳۰۶ شرکت با سال مالی منتهی به برج ۱۲ سال ۱۳۹۹ و شامل ۳۷ عنوان نسبت مالی گردآوری گردید، در نهایت با توجه به فقدان اطلاعات کامل برای شرکت‌ها و یا نسبت‌های مالی، داده‌ای شامل ۲۷۶ شرکت و ۱۸ نسبت مالی فراهم گردید که پس از گردآوری دسته دوم داده‌ها نیز ۵ شرکت حذف و در نهایت اطلاعات ۲۷۱ شرکت باقی مانده است؛ دسته دوم داده‌ها مربوط به داده‌های قیمت پایانی روزانه سهام شرکت‌های فعال در بازار بورس اوراق بهادار تهران که از طریق وبسایت "مرکز پردازش اطلاعات مالی ایران[۴۳]" استخراج شده‌اند، می‌باشد. این داده‌ها در بازه زمانی فروردین ۱۳۹۹ تا فروردین ۱۴۰۰، مورد نظر پژوهش حاضر بوده است. در خصوص این دسته، داده نهایی شامل ۲۷۱ عنوان شرکت به همراه قیمت پایانی در هر روز، قیمت روز قبل و تاریخ می‌باشد. در نهایت با استفاده از قیمت پایانی روزانه و قیمت روز قبل هر سهم، بازده قیمت روزانه سهام شرکت‌ها و در نهایت میانگین بازده قیمت سالانه شرکت‌ها بر اساس مبانی نظری مذکور محاسبه گردیده است. دلیل انتخاب سال مالی ۱۳۹۹ برای هر دو دسته داده، وجود حداکثری اطلاعات لازم نسبت به سال‌های دیگر مالی شرکت‌ها بوده است. حال اولین گام پس از تهیه داده، شناسایی و اصلاح داده‌های پرت و یا از دست‌رفته یا ناموجود به وسیله توابع گوناگون می‌باشد که در خصوص داده‌های این پژوهش، با توجه به حساسیت اعداد(نسبت‌های مالی) و عدم امکان حذف یا درج اعداد توسط تحلیل‌گر، جایگزین با داده‌های خالی، تصمیم بر آن شد تا شرکت‌هایی که برای برخی نسبت مالی خارج از داده هستند و یا نسبت‌هایی که برای عمده شرکت‌ها مقداری ناموجود دارند، از مجموعه داده اولیه حذف شوند؛ گام بعدی بررسی همبستگی بین معیارها، شاخص‌ها و متغیرها می باشد که در ماتریس خروجی شکل (۳)، قطر اصلی نشان‌دهنده 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 صرفا برای دو شاخص (حاشیه سود قبل از مالیات) و (حاشیه سود خالص) دیده شده است که با توجه به تعاریف متفاوت این دو شاخص، می‌توان این هم‌بستگی را نادیده گرفت و صرفا شباهت عددی موجب این همبستگی گردیده است.



شکل ۳- ماتریس همبستگی بین شاخص‌ها

در گام بعدی نسبت به بی مقیاس‌سازی داده‌ها اقدام می‌شود، هدف بی‌مقیاس‌سازی در حقیقت یکسان‌سازی داده‌ها به شکلی است که بتوان تمامی داده را یکجا و با یک دید تفسیری مورد مطالعه قرار داد که این مهم از دو طریق در پژوهش حاضر قابل انجام بوده است(نرمال‌سازی و استانداردسازی) که استانداردسازی روش بهتری نسبت به نرمال‌سازی برای داده‌های مورد نظر پژوهش حاضر شناسایی و صورت پذیرفته است.

شکل (۴) الگوریتم‌ها و مراحل کلی پژوهش را نشان می‌دهد:

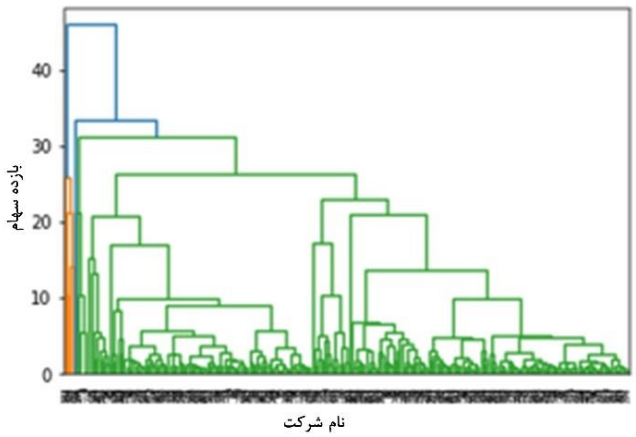


شکل ۴- نمودار روش‌شناسی پژوهش

۵. پیاده‌سازی و اجرا

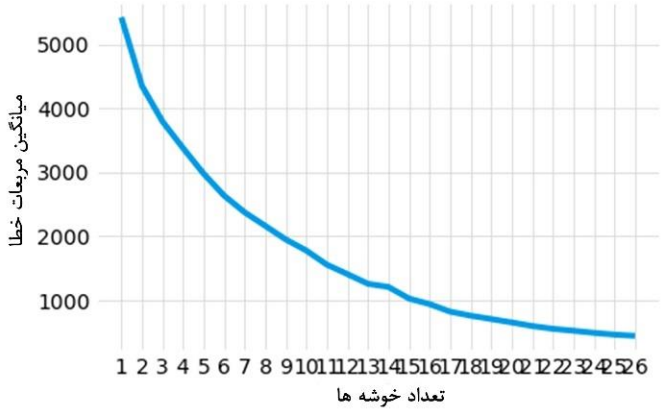
۱.۵. خوشه‌بندی شرکت‌ها

در اولین گام از ارائه مدل‌ها، شرکت‌های فعال در بازار سهام مورد نظر پژوهش حاضر بر اساس نسبت‌های مالی برگزیده مذکور در بخش چهار، خوشه‌بندی شده‌اند تا شرکت‌هایی با رفتار مالی گوناگون تفکیک گردند و پس از آن ضمن محاسبه بازده و وزن کسب شده هر خوشه بر اساس بازده، از هر خوشه تعدادی سهام با بازده بیشتر، استخراج گردد.



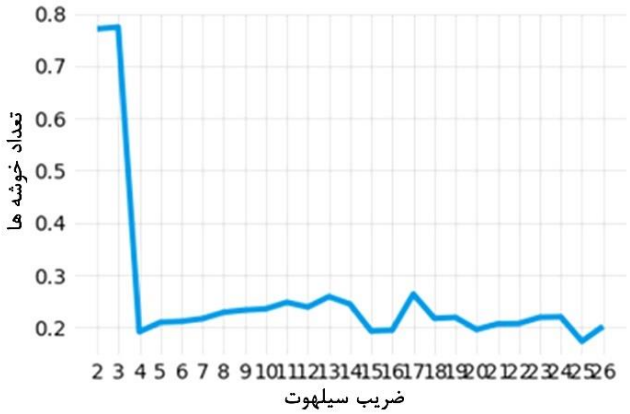
شکل ۵- دندوگرام خروجی الگوریتم خوشه‌بندی سلسله‌مراتبی

خروجی خوشه‌بندی سلسسله‌مراتبی در شکل (۵) نشان می‌دهد که شرکت‌های فعال در بازار سهام را می‌توان بر اساس رفتار مالی آنها، بر اساس شاخص‌های مورد نظر پژوهش حاضر، به ۲ خوشه تقسیم‌بندی نمود. نکته حائز اهمیت آنکه در این روش خوشه‌بندی، تعداد خروجی لزوما تعداد بهینه اصلی نمی‌باشد و صرفا یک تحلیل بر هر نوع داده است. از سوی دیگر به منظور تشخیص تعداد خوشه‌ها در الگوریتم کامینز از طریق روش آرنج، با توجه به آنکه ۲۷۰ شرکت مورد نظر این پژوهش می‌باشد، فرض اولیه بر آن شد که حداکثر ۲۷ خوشه نهایی به منظور خوشه‌بندی شرکت‌ها در نظر گرفته شود؛ بدیهی است که این رقم، خارج از تصور به منظور خوشه‌بندی شرکت‌های فعال در بازار بورس اوراق بهادار تهران می‌باشد. در این روش، الگوریتم از ۱ خوشه تا ۲۷ خوشه، داده‌ها را بر اساس روش خوشه‌بندی کامینز خوشه‌بندی می‌نماید و سپس مقدار خطای «مجموع مربعات خطا» را برای هر بار فرآیند اجرای الگوریتم، محاسبه می‌نماید. خروجی این روش در نمودار شکل (۶)[۴۴] قابل مشاهده است:



شکل ۶- تصویر نمودار SSE

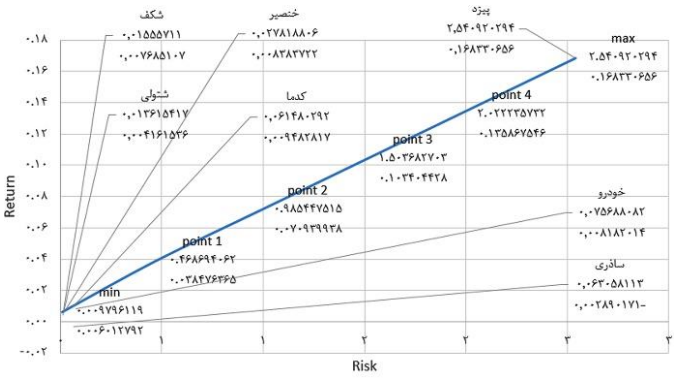
همانطور که در شکل (۶) قابل مشاهده است محور افقی بیانگر تعداد خوشه‌ها و محور عمودی بیانگر مقدار خطا است که با کمی دقت برداشت خواهد شد که پس از تعداد خوشه‌ی ۱۰ یا ۱۱، مقدار خطا تغییر خاصی نخواهد داشت؛ در نتیجه مقدار k در این روش برابر است با ۱۰ یا ۱۱؛ به جهت افزایش اعتبار تحلیل، مقدار تشخیصی داده توسط نرم‌افزار نیز با دو مقدار ۱۰ و ۱۱ تعیین شده است که این مقادیر در تکرارهای گوناگون بدست آمده است. همچنین به منظور تشخیص تعداد خوشه‌ها در این روش از طریق روش ضریب نیم‌رخ، همانند روش آرنج، ۲۷ مرتبه خوشه‌بندی انجام می‌شود و هر بار مقدار ضریب نیم‌رخ ثبت می‌گردد. مقدار این ضریب بین ۱- تا ۱+ متغیر است که این مقدار هرچقدر به یک نزدیکتر باشد، خوشه‌بندی بهتری صورت پذیرفته است.



شکل ۷- نمودار ضریب نیم‌رخ

همانطور که در نمودار فوق قابل مشاهده است، محور افقی بیانگر تعداد خوشه‌ها و محور عمودی بیانگر مقدار ضریب نیم‌رخ می‌باشد. تحلیل

نمودار فوق با توجه به آنچه گفته شد ضریب نیم‌رخ به ازای مقادیر نزدیکتر به یک بهتر است، نشان می‌دهد تعداد خوشه بهینه برای داده مورد نظر برابر با ۳ خوشه می باشد. لازم به ذکر است که مقدار ضریب نیم‌رخ بدست آمده به ازای ۲ و ۳ خوشه در تکرار این روش، بسیار نزدیک به هم بوده و هر دو این اعداد می‌توانند تعداد خوشه مناسب در روش کامینز باشند. بنابراین، چهار مقدار ۲، ۳، ۱۰ و ۱۱ برای تعیین تعداد خوشه مورد نظر در روش کامینز به عنوان K اولیه مورد نیاز، شناسایی شد. مهم‌آنکه عدم حذف داده‌های پرت در داده‌های مورد تحلیل پژوهش حاضر در گام پیش‌پردازش به دلیل مذکور(عدم امکان حذف و یا ایجاد داده‌های جدید تصادفی برای نسبت‌های مالی)، منجر به ایجاد رقمی بزرگ (۱۰ و ۱۱) در روش آرنج گردیده است. با توجه به تعداد خوشه‌های معقول مورد انتظار برای شرکت‌های فعال در بورس و همچنین محدودیت در پیش‌پردازش کامل داده‌ها، به نظر می‌رسد از میان اعداد مذکور، ۲ یا ۳ خوشه عدد مناسبی برای خوشه‌بندی باشد. به منظور سنجش، الگوریتم ضریب نیم‌رخ پیاده‌سازی شده است، خوشه‌بندی با روش کامینز مجددا به صورت دستی صورت پذیرفت و خروجی بر اساس ضریب نیم‌رخ ثبت گردیده است که نتایج نشان می‌دهد خوشه‌بندی به ازای تعداد خوشه ۳ و تکرار ۱۰۰ مرتبه، نتیجه بهتری را در مقایسه با تعداد خوشه‌های مذکور دیگر، ارائه خواهد داد که این نتیجه صحت نمودار SSE را تصدیق می‌نماید. با توجه به خروجی این بخش، سهام مربوط به هر خوشه در نرم‌افزار اکسل مشخص شد؛ در گام بعدی وزن هر سهم در هر خوشه با توجه به مباحث نظری مذکور در این خصوص، همچنین بازده هر خوشه محاسبه شد، سپس وزن هر خوشه با توجه به بازده خوشه‌ها محاسبه گردید. در نهایت با توجه به اوزان کسب شده خوشه‌ها و میانگین بازده قیمت سالانه سهام شرکت‌ها، تعداد ۷ سهم شامل: پیزد، کدما، خودرو، خنصیر، ساذری، شکف و شتولی انتخاب گردید. در آخرین مرحله از این گام، در نرم‌افزار اکسل، ضمن بهره‌گیری از روش مارکویتز، به تشکیل ۶ سبد سرمایه با در نظرگیری مقادیر ریسک و بازده گوناگون بین دو نقطه با حداکثر و حداقل بازده پرداخته شد که از بیان جزییات این گام و فرمول‌های تدوین شده در نرم‌افزار اکسل به جهت جلوگیری از نشر مطالب تکراری، خودداری شده است؛ بدیهی است با تغییر مقدار گام‌ها بین دو نقطه حداقل و حداکثر، می‌توان تعداد نقطه بیشتری را به عنوان سبدهای سرمایه با مقادیر بازده و ریسک گوناگون تشکیل داد؛ در این پژوهش گام‌ها به نحوی تعیین شده‌اند تا ۴ نقطه میانی ایجاد گردد. در نمودار شکل(۸)، مقادیر مربوط به ریسک و بازده هر سبد و هر سهم مشخص شده است، به‌گونه‌ای که در آن خط آبی رنگ نشان‌دهنده مقادیر مربوط به ریسک و بازده ۶ سبد مذکور می‌باشد و نقاط مربوط به ریسک و بازده هر سهم نیز به‌صورت جداگانه نشان داده شده است. از مشاهده این نمودار نیز این‌گونه برداشت می‌شود که از بین سهام شرکت‌های منتخب، هرگاه تصمیم به خرید تک سهمی 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 دارند، بنابراین در سبد نیز ارزش بیشتری خواهند داشت.



شکل ۸- تصویر نمودار سبدهای گوناگون به ازای ریسک و بازده مورد انتظار 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 سالانه یک شرکت را شناسایی می‌نماید و در صورت اجرای مدل به سرمایه‌گذار در جهت تصمیم‌گیری به جهت خرید یا عدم خرید به عنوان یکی از فاکتورهای مدیریت سبد خرید، کمک شایانی می‌نماید؛ همچنین این مدل در تشخیص تقلب شرکت‌ها نیز موثر خواهد بود. به منظور آنکه نرم‌افزار قادر باشد هر ردیف داده را به گروه مناسب یا نامناسب نسبت دهد، نیاز به ستون جدیدی با عنوان ستون هدف می‌باشد، حال نیاز به آن است که سرمایه‌گذار تصمیم‌گیری نماید که میزان بازده قیمت مورد انتظار چقدر باشد تا سهمی مناسب سرمایه‌گذاری شود؛ به همین علت بایستی بررسی شود حد میانی بازده قیمت سهام در داده‌های موجود چقدر است تا ستون هدفی با مقادیر برابر به ازای دو برچسب، موجود باشد؛ این مقدار ممکن است نزدیک به بازده مورد انتظار سرمایه‌گذار باشد و یا به دلیل کمبود اطلاعات در پایگاه‌های داده، همچون پژوهش حاضر، مقداری کمتر از مقدار مورد انتظار عموم سرمایه‌گذاران داشته باشد. به این منظور جدول(۱) با هدف مکان‌یابی نقطه میانه بازده سهام تشکیل شده است:

جدول ۱- مکان‌یابی نقطه میانه بازده سهام به ازای مقادیر مختلف بازده

میزان بازده	تامین شرط	عدم تامین شرط
>۰.۰۵	۵	۲۶۶
>۰.۰۱	۶	۲۶۵
>۰.۰۰۵	۴۹	۲۲۲
>۰.۰۰۴	۸۹	۱۸۲
>۰.۰۰۳	۱۳۴	۱۳۷
>۰.۰۰۲	۱۷۶	۹۲

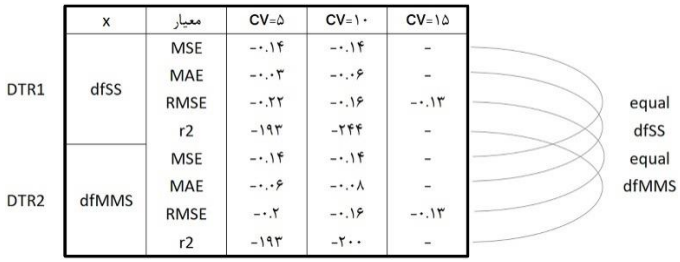
جدول فوق نشان داد که مقدار بازده ۰.۳ درصد، مقداری است که به ازای آن ۱۳۴ سهم بازدهی بیشتر از این مقدار و ۱۳۷ سهم بازدهی کمتر از این مقدار دارند؛ بنابراین شرکت‌هایی که بازده بیشتر دارند برچسب ۱ و شرکت‌هایی که بازده کمتر دارند برچسب ۰ دریافت خواهند نمود. مهم آنکه با در نظر گرفتن بازده ۰.۳ درصد ، مقادیر ستون هدف مقداری متوازن خواهند داشت و بنابراین به علت حساسیت داده‌های نسبت‌ها، نیازی به شبیه سازی داده‌ها به منظور ایجاد داده‌ای متوازن نخواهد بود. در مرحله بعدی داده‌ها با روش استانداردسازی[۴۵]، بی‌مقیاس شده‌اند؛ اگرچه در ادامه در پرسپترون چندلایه، نیاز به بی‌مقیاس‌سازی با روش نرمال‌سازی[۴۶] می‌باشد که جداگانه صورت پذیرفته است. حال بایستی آخرین مرحله قبل از پیاده‌سازی الگوریتم‌های طبقه‌بندی تحت عنوان تقسیم به داده‌های آموزش و تست صورت پذیرد، به این منظور ۳۰ درصد داده‌ها به صورت تصادفی و توسط نرم‌افزار به عنوان داده تست و مابقی به عنوان داده آموزش در نظر گرفته شده است. در مرحله بعدی به مدل‌سازی پرداخته شده است که نتایج در ادامه ارائه شده است؛ مهم آنکه محاسبه میزان امتیاز مدل در تمامی الگوریتم‌ها ضمن بهره‌گیری از ماژول کراس[۴۷] و همچنین مقدار ضریب خطای معیار، ماتریس آشفتگی، که پیش‌تر در مبانی نظری تشریح گردیده است، صورت پذیرفته است و با توجه به حجم مطالب، مطالعه در خصوص نحوه تنظیم معیارهای برنامه‌نویسی به مخاطب واگذار گردیده است و به این منظور منبع شماره ۹ پیشنهاد میگردد.

در مرحله اول طبقه‌بندی با الگوریتم درخت تصمیم با معیار آنتروپی[۴۸] صورت پذیرفته است که در این مرحله پس از بررسی ۳۰ خروجی امتیاز مدل برابر است با ۰.۶؛ در مرحله دوم طبقه‌بندی با الگوریتم درخت تصمیم با معیار جینی[۴۹] انجام شده است، امتیاز مدل ساخته شده توسط درخت تصمیم در این مرحله نیز پس از بررسی ۳۰ خروجی برابر است با ۰.۵۴؛ در مرحله سوم طبقه‌بندی با الگوریتم گوسین نایوبیز صورت پذیرفت، در این الگوریتم هم از روش cross با ۳۰ مرتبه تکرار بهره گرفته شده است که امتیاز خروجی برابر است با ۰.۵؛ در مرحله بعدی طبقه‌بندی با الگوریتم نزدیک‌ترین همسایه، متغیری با عنوان «وزن‌ها[۵۰]» وجود دارد که در پژوهش جاری دو مقدار «یکنواخت[۵۱]» در مرحله چهارم و «فاصله[۵۲]» در مرحله پنجم برای این پارامتر 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[28]، خروجی در مراحل چهارم و پنجم پس از ۳۰ بار تکرار مدل برابر است با ۰.۶۷ و ۰.۶۵ که این مقادیر برتری مقدار فاصله را بر یکنواخت نشان می‌دهد. در مرحله بعد، در الگوریتم ماشین‌بردار پشتیبان متغیری با عنوان «کرنل[۵۳]» به منظور چندمنظوره سازی الگوریتم وجود دارد؛ این تابع امکان ایجاد هسته‌های سفارشی را علاوه بر هسته‌های مشترک فراهم نموده است و به سه شکل(خطی[۵۴]، غیرخطی[۵۵]، چند جمله‌ای[۵۶]) قابل تنظیم می‌باشد[28]. در پژوهش حاضر هر سه نوع هسته‌های مذکور در مراحل ۶، ۷ و ۸ در نظر گرفته شده است تا صرفا بهترین نوع هسته به تشخیص نرم‌افزار بر اساس میزان دقت خروجی محاسبه گردد؛ بدیهی است مخاطب پژوهش جاری قادر خواهد بود از تنظیم متغیرها به شکل دقیق‌تر در هر کدام از الگوریتم‌ها به جهت دریافت خروجی بهتر استفاده نماید. مهم آنکه تعداد تکرار الگوریتم cross، بایستی بررسی شود و بر اساس مقدار خروجی بهترین تعداد تکرار تشخیص داده شود. در نهایت بهترین خروجی حاصل پس از ۱۰ بار تکرار بر اساس هسته خطی برابر است با ۰.۶۷. به منظور مدل‌سازی پرسپترون چندلایه، دو متغیر «فعال‌سازی[۵۷]» و «حل‌کننده[۵۸]» به منظور تغییر از حالت پیش‌فرض قابل بررسی هستند؛ متغیر فعال‌سازی با مقدار identity[۵۹] هم علاوه بر مقدار پیش‌فرض که عبارت است از relu[۶۰] به ترتیب در مراحل دهم و نهم بررسی شده است و متغیر حل کننده با مقدار lbfgs[۶۱] جایگزین با مقدار اصلی بررسی شده است. متغیر «max_iter» هم تعداد تکرار شبکه را مشخص می‌نماید که می‌توان شبکه را تا ۱۵۰۰۰ بار تکرار نماید. همچنین عملکرد این الگوریتم با روش بی‌مقیاس‌سازی به روش استانداردسازی در مقایسه با روش بی‌مقیاس‌سازی به روش نرمال‌سازی در یازدهمین مرحله مدل‌سازی بررسی شده است ولی همانطور که مورد انتظار بوده است، در روش استانداردسازی خروجی بهتری نسبت به روش نرمال‌سازی اتخاذ گردیده است.

در نهایت، مقایسه نتایج امتیاز خروجی مدل‌ها فوق نشان می‌دهد که الگوریتم ماشین‌بردار پشتیبان در مرحله ششم، با اختلاف قابل توجهی نسبت به سایر الگوریتم‌ها توانسته است مقادیر ستون هدف را از منظر طبقه‌بندی، پیش‌بینی نماید. لازم به یادآوری است که هرگاه در هر کدام از الگوریتم‌های بررسی‌شده، تنظیم متغیر دقیق‌تری صورت پذیرد، مدل ساخته شده از میزان دقت بیشتری برخوردار خواهد بود.

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 در نهایت، برترین الگوریتم در این زمینه با توجه به معیارهای سنجش مدل‌های پیش‌بینی انتخاب گردیده است. در مرحله اول پیش‌بینی با الگوریتم درخت تصمیم صورت پذیرفته است که در شکل(۹) مقادیر ارزیابی مدل به ازای دفعات تکرار ۵، ۱۰ و ۱۵ مرتبه برای دو الگوریتم درخت‌تصمیم با داده‌های بی‌مقیاس‌شده به دو روش استانداردسازی[۶۲] و نرمال‌سازی[۶۳] قابل مشاهده می‌باشد. استفاده از داده‌های نرمال، صرفا جهت اطلاع مخاطب در نظر گرفته شده است و با توجه به دلیل مذکور در بخش‌های قبلی، در پژوهش حاضر توجیح علمی ندارد. در نهایت مشاهده شده

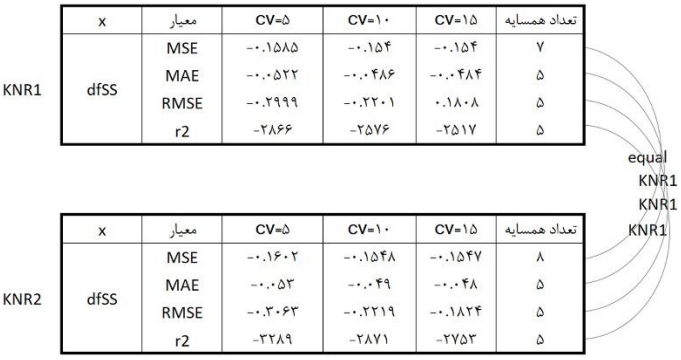
است که با اختلاف بسیار کمی، خروجی اخذ شده با هر دو نوع داده، تقریبا یکسان می‌باشد. در سمت راست تصویر، مدل برتر با توجه به هر معیار ارزیابی نشان داده شده است.



x	معیار	CV=۵	CV=۱۰	CV=۱۵
dfSS	MSE	-۰.۱۴	-۰.۱۴	-
	MAE	-۰.۰۳	-۰.۰۶	-
	RMSE	-۰.۲۲	-۰.۱۶	-۰.۱۳
	r2	-۱۹۳	-۲۴۴	-
dfMMS	MSE	-۰.۱۴	-۰.۱۴	-
	MAE	-۰.۰۶	-۰.۰۸	-
	RMSE	-۰.۲	-۰.۱۶	-۰.۱۳
	r2	-۱۹۳	-۲۰۰	-

شکل ۹- مقایسه مقادیر خطا با دو روش بی‌مقیاس‌سازی با الگوریتم پیش‌بینی درخت تصمیم

در مورد الگوریتم نزدیک‌ترین همسایه، مشابه با الگوریتم قبلی، مقادیر اخذ شده از خروجی دو مدل ساخته شده با داده بی‌مقیاس‌شده به روش استانداردسازی در شکل(۱۰) قابل مشاهده است و در سمت راست تصویر، مدل برتر با توجه به هر معیار ارزیابی نشان داده شده است. مقایسه نشان می‌دهد مدل ساخته شده با عنوان KNR1 مدل برتری نسبت به KNR2 می‌باشد.

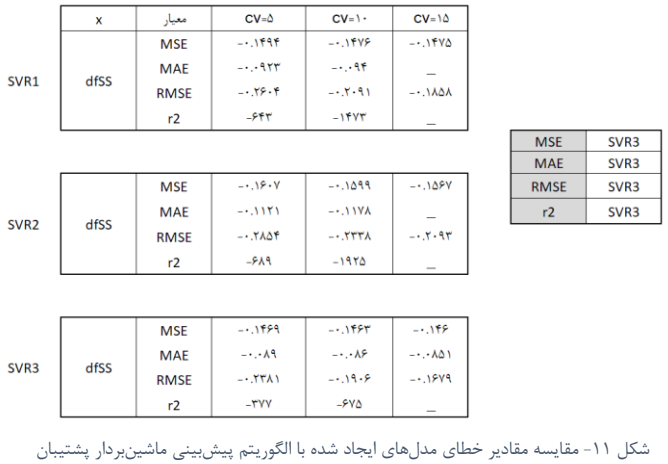


x	معیار	CV=۵	CV=۱۰	CV=۱۵	تعداد همسایه
dfSS	MSE	-۰.۱۵۸۵	-۰.۱۵۴	-۰.۱۵۴	۷
	MAE	-۰.۰۵۲۲	-۰.۰۴۸۶	-۰.۰۴۸۴	۵
	RMSE	-۰.۲۹۹۹	-۰.۲۲۰۱	۰.۱۸۰۸	۵
	r2	-۲۸۶۶	-۲۵۷۶	-۲۵۱۷	۵

x	معیار	CV=۵	CV=۱۰	CV=۱۵	تعداد همسایه
dfSS	MSE	-۰.۱۶۰۲	-۰.۱۵۴۸	-۰.۱۵۴۷	۸
	MAE	-۰.۰۵۳	-۰.۰۴۹	-۰.۰۴۸	۵
	RMSE	-۰.۳۰۶۳	-۰.۲۲۱۹	-۰.۱۸۲۴	۵
	r2	-۳۲۸۹	-۲۸۷۱	-۲۷۵۳	۵

شکل ۱۰- مقایسه مقادیر خطای مدل‌های ایجاد شده با الگوریتم پیش‌بینی نزدیک‌ترین همسایه

در مورد الگوریتم ماشین‌بردار پشتیبان، از آنجاکه الگوریتم حاضر با سه نوع هسته سفارشی قادر به تحلیل داده می‌باشد، سه مدل با هسته خطی، چندجمله‌ای و غیرخطی همچون الگوریتم طبقه‌بندی به ترتیب ایجاد شده است و هر مدل با معیارهای خطای الگوریتم‌های پیش‌بینی مذکور سنجیده شده است. مقایسه خروجی مقادیر مختلف خطا در شکل(۱۱) نشان می‌دهد که سومین مدل یا مدل با هسته سفارشی غیرخطی بهترین نتیجه را از بین سه مدل ساخته شده ارائه داده است.



x	معیار	CV=۵	CV=۱۰	CV=۱۵
dfSS	MSE	-۰.۱۴۹۴	-۰.۱۴۷۶	-۰.۱۴۷۵
	MAE	-۰.۰۹۲۳	-۰.۰۹۴	_
	RMSE	-۰.۲۶۰۴	-۰.۲۰۹۱	-۰.۱۸۵۸
	r2	-۶۴۳	-۱۴۷۳	_

x	معیار	CV=۵	CV=۱۰	CV=۱۵
dfSS	MSE	-۰.۱۶۰۷	-۰.۱۵۹۹	-۰.۱۵۶۷
	MAE	-۰.۱۱۲۱	-۰.۱۱۷۸	_
	RMSE	-۰.۲۸۵۴	-۰.۲۳۳۸	-۰.۲۰۹۳
	r2	-۶۸۹	-۱۹۲۵	_

x	معیار	CV=۵	CV=۱۰	CV=۱۵
dfSS	MSE	-۰.۱۴۶۹	-۰.۱۴۶۳	-۰.۱۴۶
	MAE	-۰.۰۸۹	-۰.۰۸۶	-۰.۰۸۵۱
	RMSE	-۰.۲۳۸۱	-۰.۱۹۰۶	-۰.۱۶۷۹
	r2	-۳۷۷	-۶۷۵	_

شکل ۱۱- مقایسه مقادیر خطای مدل‌های ایجاد شده با الگوریتم پیش‌بینی ماشین‌بردار پشتیبان

در خصوص الگوریتم پرسپترون چندلایه نیز نتایج در قالب شکل(۱۲) قابل مشاهده می‌باشد:

	x	معیار	CV=۵	CV=۱۰	CV=۱۵
MLP	dfSS	MSE	-۵.۹۲	-۷.۸۲	_
		MAE	-۰.۶۳	-۰.۶۷۳۶	_
		RMSE	-۲.۳۵	-۱.۹۲	-۱.۶۱
		r2	-۵۱۵۷۸۴	-۸۸۸۵۸۷	_

شکل ۱۲- مقادیر مختلف خطا به ازای دفعات مختلف تکرار الگوریتم پیش‌بینی پرپسترون چندلایه

در نهایت، در آخرین گام از ایجاد مدل‌های پیش‌بینی، برترین مدل‌های هر الگوریتم با یکدیگر مقایسه گردیده‌اند تا بهترین مدل پیش‌بینی تشخیص داده شود؛ مقایسه نتایج خروجی مدل‌های فوق نشان می‌دهد که الگوریتم درخت‌تصمیم توانسته است با حداقل مقادیر خطا، بهترین مدل را از میان سایر الگوریتم‌ها ارائه دهد؛ مهم آنکه مقدار خطای این الگوریتم جزو حداقل مقادیر ممکن خطا یا مقداری بسیار نزدیک به صفر برای سه معیار اول و مقداری بسیار نزدیک‌تر به یک در مقایسه با سایر الگوریتم‌ها برای معیار آخر می‌باشد.

بهترین نمره	MLO	SVR3	KNR1	DTR1	
DTR۱	-۵.۹۲	-۰.۱۴۶	-۰.۱۵۴	-۰.۱۴	MSE
DTR۱	-۰.۶۳	-۰.۸۵۱	-۰.۰۴۸۴	-۰.۰۳	MAE
DTR۱	-۱.۶۱	-۰.۱۶۷۹	۰.۱۸۰۸	-۰.۱۳	RMSE
DTR۱	-۵۱۵۷۸۴	-۳۷۷	-۲۵۱۷	-۱۹۳	r2

شکل13 - جدول خلاصه مقایسه خروجی مدل‌های پیش‌بینی

از جدول ۱۳ نتیجه‌گیری می‌شود که الگوریتم درخت‌تصمیم توانسته است با حداقل مقادیر خطا، بهترین مدل را از میان سایر الگوریتم‌ها ارائه دهد؛ مهم آنکه مقدار خطای این الگوریتم جزو حداقل مقادیر ممکن خطا یا مقداری بسیار نزدیک به ۰ برای سه معیار اول و مقداری بسیار نزدیک‌تر به ۱ در مقایسه با سایر الگوریتم‌ها برای معیار آخر می‌باشد.

۶. نتیجه‌گیری و پیشنهادها:

در پژوهش حاضر، ابتدا پس از مرور منابعی در زمینه کاربرد علم‌داده در حل مسائل مالی، جایگاه علم‌داده در حل اینگونه مسائل تبیین گردید؛ خروجی این بخش، رهنمودی در جهت تشخیص موضوع اصلی پژوهش حاضر ایجاد نموده است؛ این موضوع به نحوی شناسایی گردیده است که یکی از نیازهای مهم تحلیل‌گران مالی در سطوح مختلف اعم از دانشجویان این رشته و همچنین افراد متخصص با بهره‌گیری از تکنیک‌های علم‌داده در قالب سیستم پشتیبان‌تصمیم‌گیری در مدیریت سبد سهام مرتفع گردد. از این رو، ۶ سبد سهام با ترکیب ۷ سهم منتخب با در نظرگرفتن مقادیر مختلف بازده و ریسک ضمن ترکیب "ماژول خوشه‌بندی" مذکور و مدل مارکویتز ایجاد گردیده است؛ در این مسیر، تعداد زیادی از شرکت‌های فعال در نظر گرفته شده در بازار بورس اوراق بهادار تهران در کمترین

زمان ممکن از نظر رفتار بنیادین مورد تحلیل قرار گرفته و در خوشه‌های مجزا تفکیک گردیدند. از سویی دیگر وزن خوشه‌ها با توجه به بازده قیمت سالانه هر خوشه محاسبه گردید، سپس از هر خوشه متناسب با اوزان کسب شده، سهمی برگزیده شد که از نظر بازده قیمت سالانه، از وزن بیشتری نسبت به سایر سهام شرکت‌ها برخوردار باشد. بنابراین دو مزیت نوین با ارائه تکنیک ساخته‌شده ایجاد گردیده است که عبارت‌اند از: ۱. ایجاد خوشه‌های تفکیکی منجر به ایجاد خوشه‌هایی متشکل از سهام شرکت‌ها با رفتار یکسان از منظر شاخص‌های مالی، در هر خوشه، گردیده است که بدیهی است تحلیل رفتار مذکور با توجه به تعدد شرکت‌ها برای یک تحلیل‌گر با روش‌های مرسوم قبلی بسیار پیچیده و زمان‌بر می‌باشد؛ از سویی دیگر تعداد شرکت‌های در نظر گرفته شده در این پژوهش، تحلیل دقیق‌تری را نسبت به پژوهش‌های پیشین ایجاد نموده است؛ ۲) در مسیر خوشه‌بندی، سهم یک شرکت صرفا بر اساس میزان بازده انتخاب نشده است 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 تحت عنوان "مدل ماشین‌بردار پشتیبان‌خطی" بر اساس بهترین عملکرد معیار سنجش(خروجی ماتریس آشفتگی) با مقدار ۶۷ درصد تطبیق با داده‌های واقعی یا با ۶۷ درصد دقت پیش‌بینی، شناسایی گردید. 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 مورد انتظار بازده قیمت سرمایه‌گذار می‌باشد، پیش‌بینی نماید، این مدل در برخی گام‌های کشف تقلب به‌هنگام حسابرسی شرکت‌ها در قالب کلان موثر می‌باشد؛ به‌صورتی که اگر عدم تطابق بین مقدار بازده قیمت سهم شرکتی با خروجی مدل مذکور مشاهده شود، با ۶۷ درصد اطمینان می‌توان بیان داشت که در صورت‌های مالی شرکت مورد نظر، تقلبی صورت پذیرفته است که این مهم نیازمند حسابرسی جزیی‌تر صورت‌های مالی شرکت خواهد شد. در خصوص ایجاد "ماژول پیش‌بینی" نیز، از میان ۷ مدل نهایی، برترین مدل تحت عنوان «مدل پیش‌بینی درخت-تصمیم» بر اساس کمترین میزان خطا از منظر ۴ معیار خطا، شناسایی گردید؛ مدل ساخته‌شده قادر است با دریافت مقادیر نسبت‌های مالی یک شرکت، مقدار بازده قیمت سهام شرکت را پیش‌بینی نماید که این مهم نیز در شناسایی تقلب شرکت‌ها بسیار موثر خواهد بود و نسبت به مدل طبقه‌بندی ارائه شده، مقدار قابل تامل‌تری را منعکس می‌نماید. 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 نماید. که در این مسیر، مشابه با جیانسی در پژوهش خود در سال ۲۰۰۷، نتیجه‌گیری شد [29] که استفاده از الگوریتم ماشین‌بردار پشتیبان برای داده‌های عظیم، منجر به افزایش کارایی در مسیر طبقه‌بندی خواهد شد؛ همچنین مشابه با ناندا و همکارانش در پژوهشی در سال ۲۰۱۰ که قبلا به آن پرداخته شده است[15]، نتیجه‌گیری شد که انتخاب سهام از بین خوشه‌های تشکیل شده از سهام گوناگون شرکت‌ها، موجب کاهش ریسک متنوع سازی سبد سهام خواهد شد؛ از سویی دیگر مشابه با پژوهشی در سال ۲۰۲۴، نتیجه‌گیری شده است که هیچ گاه بهترین روش واحد برای پیش‌بینی نوسانات وجود ندارد و همانطور که پیش‌تر نیز بیان شد، لازم است تنظیم معیارها متناسب با افق پیش‌بینی صورت پذیرد [31].

در نهایت به عنوان هدف اصلی پژوهش، از ترکیب سه ماژول ارائه شده، تکنیکی در قالب یک سیستم پشتیبان تصمیم‌گیری در مدیریت سبد سهام برای تحلیل‌گران در سطوح مختلف علمی فراهم گردیده است؛ علاوه بر آنکه هر یک از سه گام طی شده، خود به تنهایی یک ماژول در مدیریت سبد سهام به شمار می‌روند. سیستم مذکور قادر خواهد بود به عنوان یک مدل جامع برای داده‌هایی در فواصل زمانی گوناگون و شاخص‌های مورد نظر تحلیل‌گر، قابلیت تعمیم‌دهی داشته باشد و بدین صورت سرمایه‌گذاران را در مسیر تحلیلی سریع و با دقت کافی، یاری نماید. عملکرد تکنیک نهایی این‌گونه است که سرمایه‌گذار ابتدا تعدادی از شرکت‌های مورد نظر خود را با توجه به شنیده‌ها، تحلیل‌ها و اخبار انتخاب می‌نماید؛ با بهره‌گیری از مدل پیش‌بینی مذکور و داده‌های تاریخی نسبت‌های مالی، بازده قیمت شرکت‌ها را پیش‌بینی می‌نماید؛ سپس با بهره‌گیری از مدل طبقه‌بندی مذکور، نسبت به حفظ یا حذف سهم شرکت‌ها با توجه به طبقه پیش‌بینی شده، اقدام می‌نماید و در آخرین مرحله، شرکت‌هایی که موفق به پذیرش شرط شده‌اند را با بهره‌گیری از مدل خوشه‌بندی، تفکیک نموده و سبد سهام متنوعی را با توجه به میزان ریسک و بازده مورد انتظار خود تشکیل می‌دهد. مهم آنکه همانطورکه گفته شد، مدل‌های طبقه‌بندی و پیش‌بینی ارائه شده در کنار کاربرد در تشکیل و مدیریت سبد سهام، قادر خواهند بود در مبحث پیش‌بینی تقلب احتمالی شرکت‌ها نیز موثر واقع شوند. طی انجام پژوهش حاضر، محدودیت‌های نیز وجود داشت که عبارت‌اند از: ۱. به‌منظور بررسی پژوهش‌ها محدودیت تاریخی در نظر گرفته نشده است؛ چراکه در صورت فیلتر نمودن تاریخ مطالعات، احتمال از دست رفتن پژوهش‌هایی با موضوعات خاص وجود دارد؛ ۲. مبانی نظری مالی در این پژوهش با توجه به مخاطب هدف، صرفا به صورت یادآوری مباحث پایه و یا تشریح کامل مباحث تخصصی می‌باشد؛ ۳. با توجه به فراوانی مباحث مبانی نظری در زمینه علم‌داده و داده‌کاوی، گزیده‌ای از مبانی به شکل کاربردی برای درک بهتر روند اجرای الگوریتم‌ها بیان شده است و مطالعه در خصوص مابقی مبانی به خواننده واگذار گردیده است؛ ۴. با توجه به نقص منابع داده و نبود برخی اطلاعات، صرفا شرکت‌هایی در

نظر گرفته شده‌اند که تمامی اطلاعات مورد نیاز پژوهش جاری برای آنها موجود بوده است؛ ۵. با توجه به گستردگی معیارهای موجود در هر الگوریتم، برخی از آنها به منظور جلوگیری از تکرار مطالب در طول تدوین متن پژوهش، همزمان با تشریح مراحل پیاده‌سازی به شکل مختصر تشریح گردیده‌اند؛ همچنین در ادامه به منظور بهبود مدل‌های ساخته شده در پژوهش حاضر، گزینه‌هایی به پژوهشگران آتی پیشنهاد گردیده است تا در آینده صورت پذیرد: ۱. با بهره‌گیری از داده‌های سایر متغیرهای مالی(همچون سایر نسبت‌های مالی و ...) با فرض موجود بودن اطلاعات مورد نیاز، به بهبود کیفیت خوشه‌بندی شرکت‌های فعال در بازار سهام نسبت به پژوهش حاضر، پرداخته شود؛ ۲. ضمن محاسبه بازده کل سهام شرکت‌های فعال در بازار سهام (جمع بازده قیمت و بازده نقدی)، نسبت به بهینه‌سازی کیفیت داده‌های اولیه ستون هدف در مسیر مدل‌سازی و در نتیجه بهبود خروجی مدل‌های طبقه‌بندی و پیش‌بینی ایجاد شده در پژوهش حاضر، اقدام شود؛ ۳. ضمن بهره‌گیری از دانش برنامه‌نویسی پایتون، به وسیله تنظیم دقیق‌تر پارامترها در هر الگوریتم، به بهینه سازی مسیر مدل‌سازی پرداخته شود؛ ۴. مدل‌های ایجاد شده در بستر زبان برنامه‌نویسی، در قالب نرم‌افزارهایی که بهره‌گیری از آن برای کاربر ساده می‌باشد، 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 گردد. ۶. با استناد به پژوهش جاری، مدلی مشابه متشکل از ماژول‌هایی پیشرفته‌تر در بستر بانک‌های اطلاعاتی آنلاین، ایجاد گردد تا پژوهشگران آتی قادر به بهره‌گیری سریع‌تر از ماژول‌ها باشند. ۷. ماژول‌های ساخته شده در این پژوهش را با داده‌های سایر کشورها بکارگیری نمایید تا صحت عملکرد ماژول‌ها در شرایط گوناگون مدیریتی سنجیده شود.

پانوشت‌ها

[2] Data Science & Data Mining [3]Artificial Intelligence (AI)
[4] Machine Learning (ML) [5] Data Mining [6] Deep learning
[7] knowledge discovery [8] Data discovery
[9] Python Data Science Handbook [Book]
[10] Root Mean Square Error (RMSE) [11] Mean Absolute Error (MAE)
[12] R2 [13] Mean Squared Error (MSE) [14] Confusion Matrix
[15] Accuracy [16] Lazy Classifiers [17] Euclidean [18] Hamming
[19] Manhattan [20] Supervised Learning
[21] Input Layer [22] Hidden Layer [23] Output Layer [24] P/E: Price to Earning [25] P/B [26] Beta [27] EPS: Earning per Share
[28] 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 دوازده ماه سال مورد نظر پژوهش می‌باشد.
[29] ROE: Return on Equity [30] D/E [31] P/S [32] Return
[33] M/B [34] Artificial Neural Network [35] Bayesian network
[36] Discriminant Analysis [37] logistic regression
[38] نگاشت خود سازمانده
[39] P/CEPS: Price to Cash EPS
[40] EV/ EBIDTA = Enterprise Value to earnings before interest, taxes, depreciation and amortization [41] tsetmc.ir
[42] bourseview.com [43] fipiran.ir [44] Sum of Squares due to Error = SSE
[45] Standard Scaler (SS) [46] Min Max Scaler (MMS) [47] cross
[48] entropy [49] gini [50] weights [51] uniform [52] distance
[53] kernel [54] Linear [55] rbf [56] poly [57] Activation
[58] Solver
[59] فعال‌سازی بدون عملیات، مفید برای پیاده‌سازی گلوگاه خطی، f(x) = x را بر می‌گرداند
[60] تابع واحد خطی اصلاح شده f(x)=max (0, x) را بر می‌گرداند
[61] یک بهینه‌ساز در خانواده روش های شبه نیوتنی است
[62] dfSS: df Standard Scaler [63] dfMMS: df Minmax Scale

منابع (References)

1. Pyle, D., 2003. Business modeling and data mining. Elsevier. San Francisco: Morgan Kaufmann.

2. Dean, J., 2014. Big data, data mining, and machine learning: value creation for business leaders and practitioners. John Wiley & Sons.

3. Kunnathuvalappil Hariharan, N., 2018. Applications of Data Mining in Finance. International Journal of Innovations in Engineering Research and Technology, 5(2), pp.72-77.

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). International Journal of Nonlinear Analysis and Applications, 15(4), pp. 293-309. [In Persian]. https://doi.org/10.22075/IJNAA.2022.28778.3985.

5. Abdullah, Z., & Hamdan, A. R., (2015). Hierarchical Clustering Algorithms in Data Mining. International Journal of Computer and Information Engineering, 9(10), pp.2194-2199. https:doi.org/10.5281/zenodo.1109341

6. Nokarto,. 2021. from Nokartohttps://nokarto.com/ The difference between data science and data mining. [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), 25-34. [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, March. Research on application of distributed data mining in anti-money laundering monitoring system. In 2010 2nd International Conference on Advanced Computer Control (Vol. 5, pp. 133-135). IEEE.

11. Han, J., Kamber, M. and Pei, J., 2012. Data Mining Third Edition 3.5. 2 Data Transformation by Normalization.

12. Paikari, E., 2023. Bug Report Quality Prediction and the Impact of Including Videos on the Bug Reporting Process. University of California, Irvine. [In Persian]

13. Boschetti, A. and Massaron, L., 2015. Python data science essentials. Packt Publishing Ltd. Third Edition. 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). Journal of Natural Environmental Hazards, 9(24), pp.129-150. [In Persian]. http://doi.org/10.22111/JNEH.2020.31217.1551.

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), p.8793-8798. 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. Journal of Modern Processes in Manufacturing and Production, 7(4), pp.17-30. [In Persian].

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

18. Sayadi, M., 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., 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. [In Persian]

20. Abbasi, I., Dehekhani, H., Khozin, A., Bazadeh, F., (2019). Designing a smart model for predicting financial wealth in insurance companies (data mining approach): Scientific Research Quarterly Journal of Investment Science, 9(34), pp.211-229 (In Persian).

21. Razin., (2015). 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. Journal 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. Journal 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 Ming and Man-Machine Management. Wireless Communications and Mobile Computing.
https://doi.org/https://doi.org/10.1155/2022/2717982

26. Karami, A., 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. scikit-learn., (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 (Doctoral dissertation, SRI CHANDRASEKHARENDRA SARASWATHI VISWA MAHAVIDYALAYA).

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, Volume 151, pp.111132.
https://doi.org/10.1016/j.asoc.2023.111132.

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

Navid Javaheri, navid.xperia2014@gmail.com

M.Sc. degree, Department of Industrial Engineering, Faculty of Engineering, Meybod University, Meybod, Iran

Najmeh Neshat (corresponding author), neshat@meybod.ac.ir

Associate Professor, Department of Industrial Engineering, Faculty of Engineering, Meybod University, Meybod, Iran

Abbass Ali Jafari Nodushan, a.jafari@meybod.ac.ir

Assistant Professor, Department of Industrial Engineering, Faculty of Engineering, Meybod University, Meybod, Iran

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 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 a number of 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.

Keywords: portfolio management, modeling, data science, classification, clustering, segmentation

Received: 2022. . , Accepted: 2022. . .

* MSc Student in Industrial Engineering-Financial Systems, Faculty of Technical Engineering, University of Meybod, Yazd, Iran.