فصلنامه راهبرد مدیریت مالی
دانشگاه الزهرا
سال دوازدهم، شماره چهل و ششم، پاییز ۱۴۰۳
صفحات ۲۳۰-۲۱۳

مقاله پژوهشی

مقایسه دقت هوشمندی الگوریتم‌های مبتنی بر داده‌کاوی جهت برآورد قیمت سهام[۱]

حسین کیانی‌زاده[۲]، علی باغانی[۳]، محسن حمیدیان[۴]

تاریخ دریافت: ۱۴۰۲/۰۲/۱۹ تاریخ پذیرش: ۱۴۰۳/۰۵/۱۹

چکیده

حجم اطلاعات بازار سرمایه به طرز چشمگیری در حال گسترش می‌باشد و بدون استفاده از الگوریتم‌های داده‌کاوی و مدل‌های کلان داده، بهره‌برداری از این داده‌ها امکان‌پذیر نخواهد بود. مطالعات گذشته بیانگر امکان پیش‌بینی قیمت سهام توسط مدل‌های یادگیری ماشین می‌باشد؛ اما دقت پیش‌بینی این مدل‌ها مورد ارزیابی قرار نگرفته است. هدف از این پژوهش مقایسه دقت هوشمندی پنج الگوریتم پرکاربرد داده‌کاوی شامل شبکه عصبی، رگرسیون لجستیک، نزدیکترین همسایه k، ماشین بردار پشتیبان و اعتبارسنجی ضربدری می‌باشد. از بین ۳۸۵ شرکت فعال در بورس اوراق بهادار تهران، ۷۲ شرکت به روش حذف سیستماتیک انتخاب و دقت مدل‌های فوق برای پیش‌بینی قیمت سهام بر روی داده‌های روزانه سهام منتخب برای سال‌های ۱۳۸۸ تا ۱۳۹۹ پیاده‌سازی شده است. متغیر قیمت سهام به عنوان متغیر وابسته و متغیرهای قیمت باز شدن، قیمت بسته شدن، بالاترین قیمت، پایین‌ترین قیمت و حجم معاملات، قیمت روزانه ارز آزاد، قیمت طلا و قیمت نفت به عنوان متغیر مستقل استفاده شده است. برای ارزیابی دقت برآورد قیمت سهام از سه شاخص R^2، MSE و $RMSE$ استفاده شده و از تحلیل واریانس با استفاده از آماره F برای برازش دقت مدل‌ها و از آماره t برای مقایسه دو به دو مدل‌ها با یکدیگر استفاده شده است. نتایج پژوهش نشان داد از بین الگوریتم‌های هوشمند استفاده شده، الگوریتم ماشین بردار پشتیبان بیشترین قدرت برآورد قیمت سهام را به خود اختصاص داده است.

واژگان کلیدی: بورس اوراق بهادار، برآورد، قیمت سهام، الگوریتم‌های هوشمند، یادگیری ماشین، داده‌کاوی.
طبقه‌بندی موضوعی: *C02,C13,C53,C55,C58,G17*

---

۱. کد DOI مقاله: 10.22051/jfm.2024.40333.2685
۲. دانشجوی دکتری، گروه مدیریت مالی، واحد بین الملل کیش، دانشگاه آزاد اسلامی، جزیره کیش، ایران. Email: kianizaceh@gmail.com
۳. استادیار، گروه مدیریت مالی، واحد بین الملل کیش، دانشگاه آزاد اسلامی، جزیره کیش، ایران. Email:ali.baghani.85@gmail.com
۴. دانشیار، گروه مدیریت مالی، واحد بین الملل کیش، دانشگاه آزاد اسلامی، جزیره کیش، ایران. Email: hamidian_2002@yahoo.com

مقدمه

برآورد صحیح قیمت سهام در بازار سرمایه می‌تواند نخستین گامِ یک سرمایه‌گذاری درست باشد. پیش‌بینی قیمت سهام شرکت‌ها در شرایط متلاطم اقتصادی بسیار دشوار است و شناخت متغیرهای تأثیرگذار و استفاده درست از مدل‌های پیش‌بینی قیمت سهام می‌تواند متخصصین را در این حوزه یاری نماید (افشاری راد، ۱۳۹۷). در سال‌های اخیر با پیشرفت سخت افزاری و نرم افزاری در حوزه فناوری اطلاعات، استفاده از مدل‌های یادگیری ماشین در بازار سهام توسعه یافته است و از الگوهای یادگیری ماشین برای برآورد قیمت سهام استفاده شده است (لانگ، ۲۰۱۹).

مطالعات پیشین در حوزه مدل‌های یادگیری ماشین برای پیش‌بینی قیمت سهام را می‌توان به دو دسته کلی تقسیم نمود: الف- استفاده از یک الگوریتم یادگیری ماشین برای پیش‌بینی قیمت سهام و ب- بهینه سازی مدل یادگیری ماشین برای برآورد قیمت سهام (نتایج پژوهش حاضر) و در همه این مطالعات امکان استفاده از مدل استفاده شده برای پیش‌بینی قیمت سهام تأیید گردیده است، ولی مقایسه دقت پیش‌بینی این مدل‌ها علی‌رغم ضرورت و اهمیت آن، در مطالعات پیشین کمتر مورد توجه قرار گرفته است.

هدف از این پژوهش مقایسه دقت هوشمندی الگوریتم‌های مبتنی بر داده‌کاوی برای پیش‌بینی قیمت سهام می‌باشد؛ از این رو ابتدا با بررسی مطالعات پیشین، پرکاربردترین مدل‌های یادگیری ماشین[1] مشتمل بر ماشین‌های بردار پشتیبان[2]، شبکه‌های عصبی، نزدیک‌ترین همسایه‌های k[3]، رگرسیون لجستیک[4] و اعتبارسنجی متقابل[5] احصاء گردیده و سپس، با استفاده از روش‌های آماری، دقت مدل‌های موصوف در برآورد قیمت سهام در بورس اوراق بهادار تهران مورد سنجش قرار گرفته است.

لذا در پژوهش حاضر، علاوه بر ارزیابی امکان پیش‌بینی قیمت سهام توسط الگوریتم‌های داده کاوی که در پژوهش‌های پیشین به صورت مستقل انجام گرفته است، ارزیابی میزان دقت مدل‌های موصوف، نوآوری اصلی این پژوهش این است و راهنمایی است برای فعالان بازار سرمایه که با استفاده از کدام مدل می‌توانند به پیش‌بینی بهتری دست پیدا کنند.

با بررسی ادبیات موضوع و مطالعه پژوهش‌های پیشین، برای مقایسه دقت مدل‌های استفاده شده در برآورد قیمت سهام از سه شاخص R^2، MSE و RMSE استفاده شده است (سرمد، ۱۳۹۹) و سپس، با استفاده از تحلیل واریانس مبتنی بر آماره F برازش دقت مدل‌ها و از آماره t برای مقایسه دو به دوی مدل‌ها با یکدیگر استفاده شده است. با طبقه‌بندی شاخص‌های موصوف این نتیجه حاصل گردید که ماشین بردار پشتیبان دارای بیشترین دقت در پیش‌بینی قیمت سهام می‌باشد و پس از آن به ترتیب شبکه‌های عصبی، رگرسیون لجستیک، اعتبارسنجی متقابل و نزدیک‌ترین همسایه‌های k قرار گرفته‌اند.

1. Machin learning
2. Support Vector Machines (SVM)
3. K Nearest Neighbors (KNN)
4. Logistic Regression (LR)
5. Cross Validation (CV)

در ادامه، پس از بررسی مبانی نظری و پیشینه پژوهش، روش‌شناسی پژوهش تشریح گردیده و سپس، با بررسی آزمون فرضیه ها، به بحث و نتیجه گیری پرداخته و راهکارهایی برای مطالعات آتی پیشنهاد گردیده است.

مبانی نظری و مروری بر پیشینه پژوهش

طبق نظریه بازار کارای سرمایه که در اوایل قرن بیستم برای برآورد قیمت سهام مطرح شد، چنین جمع بندی شده است که با مطالعه رفتار پیشین قیمت سهام، امکان دست‌یابی به بازده غیرعادی وجود ندارد و تغییرات قیمت سهام از یک الگوی قابل پیش‌بینی تبعیت نمی‌کند (راعی، ۱۴۰۰). سرمایه‌گذاران مدت‌های طولانی فرضیه بازار کارا را قبول کرده بودند و این نظریه تا دهه ۱۹۸۰ میلادی به خوبی رفتار قیمت سهام در بازار را توضیح می‌داد؛ ولی پس از آن تحولات بازار سهام نیویورک در سال ۱۹۸۷، اعتبار این نظریه را با ابهام مواجه نمود. (آمپوماه، ۲۰۲۰). از این رو پس از دهه ۱۹۹۰ میلادی، توجّه متخصصان بیشتر به یک رفتار آشوب‌گرایانه همراه با نظم متمایل گردید و طراحی مدل‌های غیرخطی برای پیش‌بینی قیمت سهام مورد توجه بیشتری قرار گرفت و مطالعات اخیر نشان دادند که با بررسی رفتار سهام برآورد قیمت امکان‌پذیر می‌باشد (احمدخان بیگی و عبدالوند ۱۳۹۶).

یکی از مشکلات اصلی در برآورد قیمت سهام، حجم بالای داده‌های موجود در بازار سرمایه است. مجموعه داده‌ها به قدری عظیم است که نمی‌توان آن را با استفاده از روش‌های غیر داده‌کاوی اداره کرد (لانگ [1]، ۲۰۱۹). از آنجا که بازار سرمایه به سرعت به اطلاعات واکنش نشان داده و این اطلاعات به صورت لحظه‌ای قیمت سهام را متأثر می‌نماید، امکان تجزیه و تحلیل حجم عظیم داده‌ها توسط کاربر وجود ندارد (کانتاردزیک، ۱۳۹۲). در همین خصوص لازم است جهت تحلیل اطلاعات از ماشین استفاده نمود؛ از این رو، استفاده از مدل‌های هوشمند داده‌کاوی و یادگیری ماشین روز به روز توسعه یافته و امکان استفاده از تکنیک‌ها و الگوریتم‌های هوشمند و مدل‌های یادگیری ماشین برای برآورد هوشمند قیمت سهام فراهم گردید و محبوبیت داده‌کاوی در دنیای مالی افزایش یافت (احمدخان بیگی و عبدالوند، ۱۳۹۶). در همین راستا، روش‌های تجزیه و تحلیل فنی، پیش‌بینی سری‌های زمانی و داده‌کاوی برای برآورد قیمت سهام گسترش روزافزون یافت (ون لونگ [2]، ۲۰۱۹).

داده‌کاوی و یادگیری ماشین با بررسی حجم عظیمی از داده‌ها و با بهره گیری از الگوریتم‌های خاص برای ساخت مدل‌های خودیادگیرنده مورد استفاده قرار می‌گیرد. برخلاف یک سیستم معمولی که با پیروی از قوانین مشخص یک وظیفه را انجام می‌دهد، یک سیستم هوشمند مبتنی بر داده‌کاوی وظایف خود را از طریق تجربه کردن یاد می‌گیرد و عملکرد آن از طریق یادگیری و تمرین بهبود پیدا می‌کند. الگوریتم‌های یادگیری ماشین با فراگیری از داده‌هایی که دریافت کرده می‌توانند مسائلی را حل کنند که بسیار پیچیده‌تر از آن است که بتوان آن را با برنامه‌نویسی معمولی حل نمود (سیگو [3]، ۲۰۲۰). از الگوریتم‌های یادگیری ماشین می‌توان به الگوریتم‌های طبقه‌بندی، الگوریتم‌های رگرسیون،

1. Long
2. Wen Long
3. Sigo

الگوریتم‌های خوشه‌بندی و الگوریتم‌های کاهش ابعاد اشاره نمود (حوسین [1]، ۲۰۲۱). روش معمول برای داده‌کاوی تقسیم مجموعه داده‌ها به زیرمجموعه‌هایی برای آموزش، اعتبارسنجی و آزمایش است (بالبینگز [2]، ۲۰۱۵). آموزش فرآیندی است که در آن ماشین، الگوی موجود در ورودی را که به صورت مجموعه داده‌های آموزشی است با استفاده از مجموعه قوانینی که نحوه یادگیری را تعریف نموده می‌آموزد. اعتبارسنجی، توانایی الگوریتم را برای ارائه جواب قابل قبول در قبال ورودی‌هایی که جزء مجموعه آموزشی نبوده‌اند و میزان دقت الگوریتم را برای برآورد این داده‌ها مورد سنجش قرار می‌دهد. استفاده از الگوریتم برای انجام وظیفه‌ای که به آن منظور طراحی شده است را آزمایش گویند (باغبانی، ۱۳۹۶).

در طول چند سال اخیر با معرفی ابزارها، سازوکار و پدیده‌های نو اهمیت بازار سرمایه در اقتصاد ایران ارتقاء یافته و برآورد قیمت سهام به عنوان یکی از مسائل جذاب و مهم توجه بسیاری از پژوهشگران دانشگاهی و کارشناسان این حوزه را به خود جلب نموده است و فعالان بازار در صدد بکارگیری روش‌های برآورد قیمت سهام، برای افزایش سود سرمایه خود بوده‌اند؛ لیکن عمده مطالعات به بررسی تحلیلهای تکنیکی و نموداری محدود شده است (افشاری راد، ۱۳۹۷) و استفاده از الگوریتمهای داده‌کاوی کمتر مورد استفاده قرار گرفته و پژوهش‌های انجام گرفته برای پیش‌بینی‌های مالی در این حوزه عمدتاً با محوریت شبکه‌های عصبی مصنوعی و در برخی موارد ماشین‌های بردار پشتیبان صورت گرفته و مقایسه دقت چند مدل پرکاربرد به صورت یکجا صورت نگرفته است که این موضوع یکی از نوآوری این پژوهش می‌باشد و هدف از انجام این پژوهش مقایسه دقت چند مدل پرکاربرد با یکدیگر در برآورد قیمت سهام می‌باشد.

مدل‌های انتخاب شده برای پیش‌بینی قسمت سهام در این پژوهش جزء مدل‌های پرکاربرد در حوزه یادگیری ماشین می‌باشد؛ به طوری که با جستجوی هریک از واژگان در پایگاه جستجوی علمی گوگل https://scholar.google.com این موضوع به خوبی مشخص می‌شود. جدول زیر نتیجه جستجوی هریک از واژگان می‌باشد. برای جستجوی کلی صرفاً نام واژه جستجو شده است و برای جستجوی تخصصی در حوزه بازار سرمایه نام مدل به علاوه stock price جستجو شده است.

جدول ۱. فراوانی کاربرد مدل‌های استفاده شده در پژوهش در مطالعات علمی

عنوان	تعداد مطالعات کلی	تعداد مطالعات در بازار سرمایه
Machine learning	۴٫۹۷۰٫۰۰۰	۷۲۴٫۰۰۰
Logistic Regression (LR)	۳٫۶۵۰٫۰۰۰	۹۰٫۰۰۰
Artificial Neural Networks (ANN)	۲٫۴۴۰٫۰۰۰	۷۴٫۳۰۰
Support Vector Machines (SVM)	۳٫۲۴۰٫۰۰۰	۸۲٫۹۰۰
Cross Validation (CV)	۴٫۲۲۰٫۰۰۰	۱۶۰٫۰۰۰
K Nearest Neighbors (KNN)	۲٫۱۳۰٫۰۰۰	۴۹٫۲۰۰

منبع: نتایج پژوهش

1. Houssein
2. Ballings

تعاریف مدل‌های به کار گرفته شده در پژوهش حاضر عبارتند از:

الگوریتم ماشین بردار پشتیبان: یکی از روش‌های یادگیری با نظارت است که برای طبقه‌بندی و دسته‌بندی الگوها از طریق کم کردن خطای‌های طبقه‌بندی و بهینه کردن الگوها استفاده می‌شود (نیوتا[1]، ۲۰۱۷).

الگوریتم شبکه‌های عصبی: سیستم‌ها و روش‌های محاسباتی نوین برای یادگیری ماشینی، نمایش دانش و در انتها، اعمال دانش به دست آمده در جهت بیش‌بینی پاسخ‌های خروجی از سامانه‌های پیچیده هستند (سیگو، ۲۰۲۰)

رگرسیون لجستیک: یک مدل آماری رگرسیون برای متغیرهای وابسته دوسویی است. این مدل را می‌توان به عنوان مدل خطی تعمیم‌یافت‌های که از تابع لوجیت به عنوان تابع پیوند استفاده می‌کند و خطایش از توزیع چندجمله‌ای پیروی می‌کند، به‌حساب‌آورد (ژائو[2]، ۲۰۱۹).

الگوریتم نزدیک‌ترین همسایه: یک روش آماری است که برای طبقه‌بندی آماری و رگرسیون استفاده می‌شود. در حالت طبقه‌بندی با توجّه به مقدار مشخص شده برای K به محاسبه فاصله نقطه‌ای که می‌خواهیم برچسب آن را مشخص کنیم با نزدیک‌ترین نقاط می‌پردازد و با توجّه به تعداد حداکثری این نقاط همسایه، در رابطه با برچسب نقطه موردنظر تصمیم‌گیری می‌شود (پیریونسی[3]، ۲۰۲۰).

اعتبارسنجی متقابل: یک روش ارزیابی مدل است که تعیین می‌نماید نتایج تحلیل آماری بر مجموعه‌ای از داده‌ها تا چه میزان امکان تعمیم دارد و از داده‌های آموزشی مستقل می‌باشد. این روش عموماً برای پیش‌بینی تغییرات استفاده می‌شود تا تعیین شود مدل موردنظر به چه میزان در عمل مفید خواهد بود (ون لونگ[4]، ۲۰۱۹).

پیشینه پژوهش‌های صورت گرفته در خصوص مدل‌های مورد استفاده در این پژوهش به صورت خلاصه در جداول ۲ و ۳ ارائه گردیده است:

جدول۲. پیشینه پژوهش‌های داخلی

نویسنده و سال	عنوان مقاله	نتیجه
زارعی؛ محمدیان؛ حاضری نیری و باشکوه اجیرلو (۱۳۹۷)	مقایسه روش‌های شبکه‌های عصبی فازی با شبکه‌های عصبی موجک فازی در پیش‌بینی قیمت سهام	نتایج حاصل از مدل با استفاده از تابع هزینه بروز رسانی شده انجام‌گرفته است. نتایج پژوهش در مقایسه شبکه‌های عصبی فازی موجک و شبکه‌های عصبی فازی بیانگر آن است که میزان اطمینان پیش‌بینی قیمت سهام بانک‌ها با استفاده از شبکه‌های عصبی موجک فازی بالای ۹۰ درصد و با شبکه‌های عصبی فازی بالای ۸۰ درصد است. در نتیجه برای پیش‌بینی قیمت سهام شبکه‌های عصبی موجک فازی از درصد اطمینان بالاتری نسبت به شبکه‌های عصبی فازی برخوردار می‌باشد.

1. Nivetha
2. Zhou
3. Piryonesi
4. Wen Long

نویسنده و سال	عنوان مقاله	نتیجه
بیات و باقری (۱۳۹۷)	پیش‌بینی قیمت سهام بااستفاده از الگوریتم کرم شب‌تاب (FA)	محاسبات انجام خطای کمتر از ۶٪ را برای پیش‌بینی نشان می‌دهد. بنابراین الگوریتم کرم شب تاب قابلیت پیش‌بینی قیمت سهام را داراست.
احمدخان بیگی و عبدالوند (۱۳۹۸)	پیش‌بینی قیمت سهام با رویکرد ترکیبی شبکه‌های عصبی مصنوعی و الگوریتم رقابت استعماری مبتنی بر تئوری آشوب	نتایج حاکی از آن است که رویکرد پیشنهادی از عملکرد بهتری نسبت به سایر رویکردهای پیشین برخوردار می‌باشد.
رجب پور؛ تقوا و حسین زاده یزدی (۱۳۹۸)	پیش‌بینی قیمت سهام شرکت‌های بورس اوراق بهادار تهران با استفاده از شبکه های عصبی مصنوعی	نتایج این پژوهش نشان می‌دهد که معماری ۳ لایه با ۸ نورون در لایه اول، ۴ نورون در لایه دوم و ۲ خروجی و معماری ۲ لایه با ۱۲ نورون در لایه اول و ۲ خروجی مدل‌های مناسبی می‌باشند.
مکیان و موسوی (۱۳۹۹)	پیش‌بینی قیمت سهام شرکت فرآورده های نفتی پارس با استفاده از شبکه‌های عصبی و روش رگرسیونی؛ مطالعه موردی: قیمت سهام شرکت فرآورده‌های نفتی پارس	نتایج حاصل از مدل شبکه‌های عصبی خطای کمتر، توضیح‌دهندگی بیشتر و بنابراین قدرت پیش‌بینی بالاتری نسبت به مدلهای رگرسیونی دارد.
تاجیک (۱۳۹۹)	بررسی تحلیل تکنیکال هوشمند با استفاده از رگرسیون تعمیم یافته شبکه عصبی	معامله گری با استفاده از شبکه عصبی VAM و EMV بهتر از نتیجه‌های معامله گری است که توسط VAMA و EMV بدون شبکه عصبی و همچنین میانگین متحرک ساده و استراتژی‌های خرید و فروش عمل می‌کند
فلاح شمس، دلنواز اصغری (۱۳۹۹)	پیش‌بینی شاخص کل بورس اوراق بهادار با استفاده از شبکه‌های عصبی	نتایج پیش‌بینی همه مدل‌های استفاده شده در این پژوهش نیز حاکی از آن است که در بورس اوراق بهادار تهران پیش‌بینی کوتاه مدت با فاصله زمانی کمتر، مناسب‌تر از پیش‌بینی بلندمدت با فاصله زمانی طولانی‌تر است
خالو زاده (۱۳۹۹)	مدل‌سازی غیرخطی و پیش‌بینی رفتار قیمت سهام در بورس ایران	در بخش پیش‌بینی ابتدا از یک شبکه عصبی پیش رو ۳ لایه با توپولوژی ۱-۵-۱-۵ برای پیشبینی یک روز بعد برای سری زمانی قیمت و بازده و شاخص قیمت برابر ۰/۹۹۷۹، ۰/۶۰۱۱ و ۰/۹۹۸۹ بدست آمده است. سپس از یک شبکه سه لایه با اتصالات برگشتی برای پیش‌بینی با افق ۳۰ روز برای سری زمانی قیمت استفاده نموده که نتیجه R^2=78/76 از آن حاصل شده اشت.
افشاری راد، علوی و سینایی (۱۳۹۹)	مدلی هوشمند برای پیش بینی روند سهام با استفاده از روش های تحلیل تکنیکال	پنج روش هوشمند ماشین بردار پشتیبان خطی، ماشین بردار پشتیبان کرنل گوسی، درخت تصمیم، نزدیک ترین K همسایه و نئیو بیز استفاده شده و این نتیجه حاصل شد که روش پیشنهادی به طور متوسط نرخ پیش بینی صحیح ۹۷ درصد دارد.

منبع: نتایج پژوهش

جدول۳. پیشینه پژوهش‌های خارجی

نویسنده و سال	عنوان مقاله	نتیجه
لانگ، سان و تیان (۲۰۱۹)	یک روش جدید هسته گرافیکی برای پیش‌بینی گرایش قیمت سهام بر اساس مفهوم مالی و شباهت ساختاری	با مقایسه صحت پیش‌بینی کرنل با هسته‌های دیگر مثل هسته خطی، نشان داده شد که روش پیشنهادی حداقل به ٪۵ روی دقت عمل می‌کند که یک ارتقا کاملاً معنی‌دار است
لین (۲۰۱۸)	تحلیل تکنیکال و پیش‌بینی سود سهام	تجزیه بازده سهام مبتنی بر اتورگرسیون برداری نشان می‌دهد که منشا اقتصادی قدرت پیش‌بینی، عمدتاً از تغییرات زمانی در جریان نقدی آینده (یعنی کانال جریان نقدی) می‌آید.

نویسنده و سال	عنوان مقاله	نتیجه
وانگ و گوپتا (۲۰۲۰)	پیش‌بینی قیمت آتی سهام (اس‌اندپی ۵۰۰) با استفاده از شبکه‌های عصبی و نویززدایی موجک	شبکه‌های عصبی اگر با آموزش و داده‌های ورودی مناسب همراه باشند، امکان دستیابی به سودآوری بالا را در صندوق‌های قابل معامله در بورس (مبتنی بر اس‌اندپی ۵۰۰) فراهم می‌کنند.
حافظی، شهرابی و هداوندی (۲۰۱۵)	مطالعه سیستم‌های شبکه‌های عصبی خفاشی در پیش‌بینی قیمت سهام: مطالعه موردی در قیمت سهام DAX	داده‌های حاصل از مدل بسیار دقیق و قابل اعتماد هستند و می‌توان از آن‌ها به عنوان یک ابزار مناسب برای پیش‌بینی داده‌ها و ارزها و قیمت‌ها در دوره‌های بعدی استفاده کرد.
وانگ و گوپتا (۲۰۱۹)	پیش‌بینی قیمت آتی سهام (اس‌اندپی ۵۰۰) با استفاده از شبکه‌های عصبی و نویززدایی موجک	نتایج حاصل از این پژوهش بیانگر آن است که شبکه‌های عصبی اگر با آموزش و داده‌های ورودی مناسب همراه باشند، می‌توانند سودآوری بالاتر در صندوق‌های قابل معامله در بورس (مبتنی بر اس‌اندپی ۵۰۰) را به همراه داشته باشند.
حسینیان، دیرار، حسین و محمد (۲۰۲۱)	شبکه‌های عصبی مصنوعی برای پیش‌بینی بازار سهام: مروری جامع	با تجزیه و تحلیل ۴۸ مقاله پژوهشی روش‌های پیش‌بینی قیمت سهام مبتنی بر شبکه‌های عصبی مصنوعی را بررسی نموده‌اند
کارا، بیاسقلو و بایکان (۲۰۱۹)	پیش‌بینی جهت حرکت شاخص قیمت سهام بورس استانبول با مدل‌های شبکه‌های عصبی – فازی و ماشین‌های بردارپشتیبان (SVM)	از اطلاعات روزانه ۱۹۹۷ تا ۲۰۰۷ همراه با ده شاخص فنی به عنوان متغیرهای ورودی مدل استفاده شده است. شبکه‌های عصبی – فازی ۷۵/۷۴٪ و مدل ماشین‌های بردار پشتیبان ۷۱/۵۲٪ را به خوبی پیش‌بینی نمودند و حاصل این پژوهش، عملکرد بهتر شبکه‌های عصبی – فازی در قیاس با مدل ماشین‌های بردار پشتیبان بوده و بهترین عملکرد پیش‌بینی مربوط به سال ۲۰۰۱ می‌باشد.
رودیگر گنزالس و همکاران (۲۰۱۹)	استفاده از شبکه‌های عصبی مصنوعی به منظور بهبود سیستم‌های معامله گری بر مبنای تجزیه و تحلیل تکنیکال	در این پژوهش تکنیک CAST معرفی شده است که به معنی پیش‌بینی دقیق تری از مقدار RSI با استفاده از شبکه عصبی مصنوعی سه لایه feed-forward می‌باشد.
لیائو و وانگ (۲۰۲۱)	پیش‌بینی مدل شاخص بورس جهانی بوسیله شبکه عصبی متأثر از زمان تصادفی	در این مدل یک تکنیک داده‌کاوی بر مبنای یادگیری ماشین امیدبخش، جهت کشف روابط پیش‌بینی بین متغیرهای اقتصادی و مالی فراوان ارائه شده است.
کائو، لگیو و شنیدرجانز (۲۰۱۵)	مقایسه بین مدل فاما و مدل فرانسوی و شبکه‌های عصبی مصنوعی در پیش‌بینی بازار سهام چین	شبکه‌های عصبی مصنوعی در مقایسه با مدل‌های خطی مقایسه شده بهتر عمل می‌نماید.
سیگو و همکاران (۲۰۲۰)	استفاده از یادگیری ماشین با استفاده از تجزیه و تحلیل داده‌های پیش‌بینی کننده بازار سهام هند	به تجزیه و تحلیل تصادفی بودن الگوی حرکت بی ثبات ترین پنجاه سهم بورس NSE-Nifty با استفاده از روش یادگیری ماشین گروهی پرداخته اند
آمپوماه و همکاران (۲۰۲۰)	ارزیابی مدل‌های یادگیری ماشین مبتنی بر درخت تصمیم در پیش‌بینی جهت حرکت سهام	مدل‌های یادگیری ماشین (ML) معمولاً بهتر از مدل‌های آماری و اقتصادسنجی عمل می‌کنند. همچنین مدل‌های یادگیری ماشینی گروهی در ادبیات نشان داده شده است که می‌توانند عملکرد بهتری را نسبت به مدل‌های یادگیری ماشینی تکی تولید کنند.

منبع: نتایج پژوهش

فرضیه پژوهش

دقت پیش‌بینی مدل‌های ماشین‌های بردار پشتیبان، شبکه‌های عصبی، روش نزدیک‌ترین همسایه های k، رگرسیون لجستیک و اعتبارسنجی متقابل با یکدیگر تفاوت دارد.

روش شناسی پژوهش

روش گردآوری اطلاعات در این پژوهش روش کتابخانه‌ای است. مباحث تئوریک پژوهش از مسیر مطالعه منابع، نشریات، منابع داخلی و خارجی موجود در کتاب‌ها و استفاده از اینترنت جمع‌آوری شده است. داده‌های پژوهش از نرم‌افزار ره آورد نوین و پایگاه اطلاع رسانی tsetmc.com به صورت روزانه استخراج شده است و از آمارهای پایگاه اطلاع‌رسانی طلا و ارز برای استخراج نرخ ارز آزاد و قیمت هر اونس طلا و قیمت نفت استفاده شده است. جامعه آماری این پژوهش شامل شرکت‌های پذیرفته شده در تابلوی اصلی بورس برای دوره زمانی ۱۳۸۸ تا ۱۳۹۹ است. به علت گستردگی حجم جامعه آماری و وجود برخی ناهماهنگی‌ها میان اعضاء جامعه، برای گزینش نمونه از روش حذف سیستماتیک استفاده شد (سرمد، ۱۳۹۹) و معیارهای اعمال شده برای به‌دست آوردن نمونه به شرح زیر است:

شرکت قبل از سال ۱۳۸۸ در بورس اوراق بهادار تهران پذیرفته شده باشد.

سال مالی شرکت منتهی به پایان اسفندماه هر سال باشد. (به لحاظ افزایش قابلیت مقایسه)

شرکت در طی سال‌های ۱۳۸۸ تا ۱۳۹۹ تغییر سال مالی نداده باشد.

در طول سال مالی دارای وقفه معاملاتی بیشتر از ۶ ماه نباشند.

بیش از ششصد هزار سهم آنها در هر سال معامله شده باشد.

با اعمال معیارهای فوق، تعداد ۷۲ شرکت فعال در تابلوی اصلی بورس به عنوان نمونه آماری انتخاب گردید.

جدول ۴. نحوه محاسبه نمونه در پژوهش

فیلتر	تعداد شرکت ها	مانده شرکتها
تعداد کل سهام موجود در بازار در سال ۱۳۹۹	۳۸۵	۳۸۵
تعداد سهام پذیرفته شده پس از سال ۱۳۸۸	۲۷۳	۱۱۲
شرکتهایی که سال مالی آن اسفند هر سال نیست	۱۵	۹۷
شرکتهایی که تغییر سال مالی داشته اند	۲	۹۵
شرکتهایی که بیش از شش ماه وقفه معاملاتی داشته اند	۴	۹۱
شرکتهایی که کمتر از ۶۰۰ هزار سهم در سال معامه کرده اند	۱۹	۷۲
تعداد شرکتهای باقی مانده	۷۲	

منبع: نتایج پژوهش

با توجه به اینکه ۷۲ شرکت پس از اعمال فیلتر باقی مانده است، کل ۷۲ سهم به عنوان نمونه مورد مطالعه این پژوهش در نظر گرفته شد. به این ترتیب ۸۶۴ سال شرکت به عنوان نمونه این پژوهش در نظر گرفته شده است.

پس از بررسی‌های فراوان برای شناسایی متغیرها، ۸ متغیر تاثیرگذار بر قیمت سهام در دو گروه به شرح زیر دسته‌بندی شدند.

گروه اول: متغیرهای فنی شامل بالاترین قیمت سهم، پایین‌ترین قیمت سهم، قیمت باز شدن سهم، قیمت بسته شدن سهم و حجم معاملات سهم.

گروه دوم: متغیرهای اقتصادی شامل نرخ ارز، قیمت هر اونس طلا و قیمت نفت در سبد اوپک

ورودی‌های هر کدام از شرکت‌ها به طور متوسط شامل ۲،۰۰۰ داده هستند که ابتدا پس از انتخاب متغیرها و جمع آوری داده‌ها، به منظور آماده‌سازی داده‌ها برای آموزش و آزمایش هر کدام از متغیرها با استفاده از رابطه (۱) نرمال گردید تا تاثیر اعداد بزرگ کاهش یابد.

$$\widehat{S_i} = \frac{(S_i - S_{min})}{S_{max} - S_{min}} \quad (۱)$$

پس از نرمال‌سازی داده‌ها، توسط روش k-Fold Cross-Validation، داده‌های آموزشی و داده‌های آزمایشی تفکیک و ۸۰ درصد داده‌ها به عنوان داده‌های آموزشی و ۲۰ درصد مابقی به عنوان داده‌های آزمایشی استفاده و الگوریتم ۱۰ مرتبه تکرار گردید. در این روش از تمامی داده‌ها برای یادگیری مدل استفاده می‌شود و هر بار بخشی از داده‌ها به عنوان مجموعه آزمایش توانایی مدل را ارزیابی می‌کند.

سپس، برای ارزیابی دقت برآورد سه شاخص R^2 ، MSE و RMSE طبق جدول۵، مورد استفاده قرار گرفت و در پایان بر اساس تحلیل واریانس با استفاده از آماره F برای برازش دقت مدل‌های برآورد سهام و از آماره t برای مقایسه دو به دو مدل‌های پیش‌بینی با یکدیگر استفاده شد. تحلیل‌ها با استفاده از نرم‌افزار Matlab انجام گرفته است.

جدول۵. معیارهای ارزیابی دقت مدل

معیار	رابطه
MSE	$MSE = \frac{\sum_{i=1}^{n}(e_i)^2}{n}$
RMSE	$\sqrt{\frac{1}{N} * \sum_{i=1}^{N}(f_i - y_i)^2}$
R^2	$R^2 = 1 - \frac{SSE}{SSy}$

منبع: روش‌های پژوهش در علوم رفتاری

رگرسیون لجستیک

رگرسیون لجستیک نوعی از رگرسیون است که متغیرهای پیش بین (مستقل) می‌تواند در مقیاس کمی و مقیاس مقوله‌ای مورد استفاده قرار گیرد؛ لیکن متغیر وابسته مقوله‌ای دو سطحی است. این دو مقوله بیانگر عضویت یا عدم عضویت در یک گروه می‌باشد. (نیوتا[1]، ۲۰۱۷) این مدل رگرسیون، شبیه رگرسیون معمولی است با این تفاوت که روش تخمین ضرایب یکسان نمی‌باشد و به جای حداقل کردن مجذور خطاها

1. Nivetha

(که در رگرسیون معمولی انجام می‌شود، احتمالی که یک واقعه رخ می‌دهد را حداکثر می‌نماید)، در رگرسیون لجستیک از مفهوم بخت برای میزان متغیر وابسته استفاده می‌شود. در اصطلاح آماری بخت، احتمال رخداد یک پیشامد (P_i) بر احتمال عدم رخداد ($I - P_i$) آن پیشامد را نشان می دهد به نحوی که احتمال بین ۰ و ۱ در نوسان است؛ در حالی که بخت می‌تواند بیش از یک باشد. واژه کلیدی در تحلیل رگرسیون لجستیک سازه‌ای به نام لوجیت است که لگاریتم طبیعی بخت می‌باشد (پیریونسی، ۲۰۲۰).

رگرسیون لجستیک به صورت فرمول (۲) تعریف می‌شود:

$$Z_i = In\left[\frac{P_i}{1-P_i}\right] = \beta o + \sum \beta i Xi \qquad (۲)$$

در معادله ۲، In بیانگر لگاریتم طبیعی است. در مدل رگرسیون لجستیک احتمال رخداد پیشامد بر اساس رابطه (۳) محاسبه می‌شود:

$$P_i = IIi\left(x1 x2, \dots xk\right) = e^{\beta o} + \frac{\sum \beta i Xi}{1+e^{\beta o+\sum \beta i Xi}} \qquad (۳)$$

یکی از منافع رگرسیون لجستیک بی نیازی آن به مفروضات محدود کننده آماری در رابطه با متغیرهاست. مسئله اصلی که در این مدل وجود دارد این است که تغییرات یکسان وضعیت موجود همیشه تغییرات یکسانی را در احتمال به دست آمده به دنبال ندارد. این موضوع زمانی که احتمال نزدیک مقادیر ۱ یا ۰/۵ باشد اهمیت بیشتری پیدا می‌کند (زارعی، ۱۳۹۷).

الگوریتم شبکه‌های عصبی

در سال‌های اخیر پیش‌بینی و مدل‌سازی به وسیله شبکه‌های عصبی رشد یافته است. یک شبکه‌های عصبی مصنوعی، مجموعه‌ای از نرونهای به هم متصل در لایه‌های مختلف (چند لایه،MLP) است که اطلاعاتی را برای یکدیگر ارسال می‌کنند. ساده ترین شکل شبکه فقط دو لایه دارد، لایه ورودی و لایه خروجی. ارتباط بین هر ورودی و خروجی به وسیله وزنی که حاکی از اهمیت نسبی ورودی برای محاسبه ارزش خروجی است، تعیین می گردد. بر این اساس، ارزش نرون خروجی Net از ورودی مشاهده tام از طریق رابطه (۴) به دست می‌آید:

$$Net_t = a_0 x_0 + a_1 x_{1t} + a_2 x_{2t} = \sum_{i=0}^{2} a_i x_{it} \qquad (۴)$$

پس از آن نرون خروجی، ارزش حاصله را با استفاده از یک تابع فعال‌سازی (محرک) یا تبدیل که با f نشان داده می‌شود- پردازش می‌کند (مکیان و همکاران،۱۳۹۰). تابع فعال‌سازی خطی ساده‌ترین شکل شبکه‌های عصبی است. ارزش به دست آمده از رابطه (۴) با یک تابع فعال‌سازی خطی، خروجی نهایی شبکه برای مشاهده t را به صورت رابطه (۵) می‌سازد.

$$y_t = f(Net_t = a_0 x_{0T} + a_1 x_{1t} + a_2 x_{2t}) = a_0 x_{0t} + a_1 x_{1t} + a_2 x_{2t} \qquad (۵)$$

تقریباً تمام شبکه‌های عصبی در بخش‌هایی از شبکه از توابع فعال‌سازی غیرخطی استفاده می‌کنند. این مسئله باعث می‌شود شبکه الگوهای غیرخطی مناسبی از مجموعه داده‌های پیچیده تولید نماید (زارعی، ۱۳۹۷). رایج‌ترین تابع فعال‌سازی مورد استفاده در ادبیات شبکه‌های عصبی تابع سیگموئیدی است:

$$f(x) = \frac{1}{1+e^{-x}} \quad (۶)$$

در این پژوهش، برای پیش‌بینی از طریق شبکه‌های عصبی، تابع فعال‌سازی لایه میانی و لایه خروجی به ترتیب سیگموئید و خطی در نظر گرفته شده است (زارعی، ۱۳۹۷).

از مزیت‌های شبکه‌های عصبی می‌توان به قابلیت آن در حل مسائل غیرخطی اشاره کرد. هم چنین شبکه‌های عصبی ابزار مناسبی برای مواردی است که در آنها جواب‌ها مهم تر از درک روابط علت و معلولی است. اما شبکه‌های عصبی قادر به ارائه معادله‌ای با ضرایب معین نمی‌باشد؛ ولی رگرسیون از جهت برآورد ضرائب متغیرهای تاثیرگذار بر هدف ما برتری بهتری دارد (بوناکرسو۱، ۲۰۱۷).

ماشین‌های بردار پشتیبان

نخستین روش طبقه‌بندی الگوها در سال ۱۹۳۶ توسط فیشر۲ بر اساس کاهش خطای طبقه‌بندی داده‌های آموزشی برای بهینه‌سازی الگوها ارائه گردید. سپس ولادیمیر وپنیک۳ در سال ۱۹۶۵ الگوریتم ماشین‌های بردار پشتیبان را مبتنی بر نظریهٔ آماری یادگیری ارائه نمود (حسینی نسب، ۱۳۹۲). الگوریتم ماشین‌های بردار پشتیبان کاربردهای زیادی در تقریب توابع از جمله خوشه‌بندی، طبقه‌بندی و رگرسیون دارد. الگوریتم ماشین بردار پشتیبان در زمان ایجاد، تنها شامل طبقه‌بندی دو کلاسه بود که در ادامه به طبقه‌بندی چند کلاسه نیز تعمیم یافت.

ماشین‌های بردار پشتیبان داده‌ها را به صورت بردار پردازش می‌کنند و همواره در میان تمام ابرصفحات تفکیک داده‌ها، ابرصفحه‌ای را انتخاب می‌کنند که بیشترین تفکیک‌پذیری با بیشترین حاشیه را میان داده‌های کلاس‌های مختلف حاصل می‌کنند. ابر صفحه مطلوب به نحوی انتخاب می‌شود که فاصله آن از نزدیک ترین داده حداکثر شود. به نزدیک ترین نقاط آموزشی با بیشترین حاشیهٔ ابر صفحه، بردارهای پشتیبان اطلاق می‌شود. چنین ابر صفحه‌ای در صورت وجود، ابر صفحه با بیشترین حاشیه نامیده می‌شود.

در صورتی که نقاط آموزشی را به صورت $\left[x_i، y_i\right]$، بردار ورودی را $x_i \in Rn$ و ارزش طبقه را $y_i \in \{-1،1\}$ ، $i = 1، \ldots، l$ نشان دهیم، آنگاه صفحه بهینه ای که طبقات تصمیم‌گیری باینری را تفکیک می‌کند، به صورت رابطه (۷) خواهد بود.

$$y = sign(\sum_{i=1}^{N} y_i a_i (X.X) + b) \quad (۷)$$

که در آن y خروجی معادله، y_i ارزش طبقه نمونه آموزشی و X_i ضریب داخلی هستند. بردار $x = \left(x_1، x_2، \ldots، x_n\right)$ بیانگر یک داده ورودی و بردارهای $X_i: i = 1، \ldots، N\ Xi$ بردارهای پشتیبان هستند. در رابطه ۱ پارامترهای b، a_iتعیین کننده ابرصفحه هستند.

1. Bonaccorso
2. Fisher
3. Vladimir N. Vapnik

در صورتی که امکان تفکیک داده‌ها به صورت خطی نباشد، رابطه (۷) به صورت رابطه (۸) تغییر می‌کند.

$$Y = sign(\sum_{i=1}^{N} y_i a_i K(X.X_i) + b) \quad (۸)$$

تابع $K(X.X_i)$ تابع کرنلی[1] است که برای ماشین‌هایی با سطوح مختلف تصمیم‌گیری غیرخطی مبتنی بر داده‌های موجود، ضربهای داخلی تولید می‌کند (ابراهیمی هروی، ۱۳۹۶). برای مثال سه نوع تابع کرنل که در الگوریتم ماشین‌های بردار پشتیبان به کار می‌روند عبارتند از: تابع کرنل پایه شعاعی، تابع کرنل پرسپترون چندلایه و تابع کرنل چند جمله‌ای. قابلیت بکارگیری کرنلها به عنوان نقطه قوت کلیدی ماشین‌های بردار پشتیبان باعث می‌شود ماشین‌های بردار پشتیبان، انعطاف بالایی برای حل انواع مختلف مسائل با دشواری‌های مختلف داشته باشند (نیوتا، ۲۰۱۷).

بعضی از ویژگی‌های ماشین‌های بردار پشتیبان عبارتند از: آموزش آسان ماشین‌های بردار پشتیبان، نتایج خوب با نمونه‌های آموزشی اندک، عدم‌نیاز به بهینه‌یابی محلی برخلاف شبکه‌های عصبی و یافتن بهترین مدل. البته برای دستیابی به این مزیت‌ها باید در انتخاب تابع هسته‌ای مناسب دقت لازم صورت گیرد. (لانگ[2]، ۲۰۱۹)

الگوریتم اعتبار سنجی ضربدری (k-f)

در الگوریتم اعتبارسنجی ضربدری، داده‌ها به K زیرمجموعه تقسیم می‌شوند. از این K زیرمجموعه، در هر بار یک زیر مجموعه برای اعتبارسنجی و K-1 زیر مجموعه دیگر برای آموزش مورد استفاده قرار می‌گیرند. این فرایند K بار تکرار می‌شود و همه داده‌ها دقیقاً یک بار برای آموزش و یک بار برای اعتبارسنجی استفاده می‌شوند. در پایان میانگین حاصل از این K بار اعتبارسنجی به عنوان تخمین پایانی برگزیده می‌شود. البته می‌توان از روش‌های دیگر نیز برای ترکیب نتایج استفاده کرد. بطور معمول از ۱۰-Fold برای طبقه‌بندی داده‌ها استفاده می‌شود.

سعی می‌شود در روش K-Fold طبقه‌ای نسبت داده‌های هر کلاس در هر زیرمجموعه و مجموعه اصلی یکسان باشد.

مساحت زیر منحنی مشخصه عملیاتی گیرنده AUC به عنوان یک معیار عملکرد طبقه‌بندی کننده مناسب و دقیق دیده می‌شود. AUC می‌تواند از ۰٫۵ تا ۱ ارزش داشته باشد. یک مقدار ۰٫۵ به این معنی است که پیش‌بینی‌ها دقیق نیستند، در حالی که مقدار ۱ به این معنی است که پیش‌بینی‌ها عالی هستند. مزیت AUC بر دیگر معیارهای عملکرد (مانند درصد به طور کامل طبقه‌بندی نشده این است که AUC همه مقادیر قطع‌شده را شامل می‌شود به صورت رابطه (۹) تعریف می‌شود (امامی، ۱۳۹۳)

$$AUC = \int_0^1 \frac{TP}{(TP+FN)} d\frac{FP}{(FP+TN)} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} \quad (۹)$$

1. Kernel
2. Long

الگوریتم نزدیک ترین همسایه‌های K

الگوریتم نزدیکترین همسایه‌های K الگوریتمی است که در آن بر اساس الگوهای آموزش، داده‌های آزمون طبقه‌بندی شوند. در این روش داده‌های مشابه در نزدیکی یکدیگر قرار می‌گیرند، به این ترتیب فاصله بین داده‌ها بر اساس عدم مشابهت آنها اندازه گیری و تفسیر می‌شود. به این ترتیب داده‌های کنار هم، همسایه نامیده شده و فاصله هر داده جدیدی که به الگوریتم معرفی می‌شود، با دیگر داده‌ها محاسبه و در دسته‌ای که در نزدیکترین فاصله قرار دارد، جای می‌گیرد (لین[1]، ۲۰۱۸). K به تعداد همسایگانی که در فرایند دسته بندی نقش مهمی ایفا می‌کنند اتلاق می‌گردد. تعداد صحیح متغیر K بستگی به اندازه داده‌ها به عنوان پارامتری مهم در این فرایند دارد. یکی از روش های معمول در انتخاب صحیح، استفاده از الگوریتم تپه نوردی است (سیگو، ۲۰۲۰). در این روش میزان k را مرتب افزایش داده تا به بهترین انتخاب برسد. معیار بهبود در این روش فاصله‌ها می‌باشند که از طریق مجموع قدر مطلق فاصله‌ها، فاصله اقلیدسی یا فاصله همینگ برای تعیین این معیار استفاده می‌شود. به طور خلاصه مراحل زیر برای انجام روند برآورد با الگوریتم نزدیک ترین همسایه‌های K مورد استفاده قرار می‌گیرد:

الف: تعیین تعداد k اولیه در نرم افزار

ب: محاسبه فاصله بین نمونه مورد بررسی با تمام داده‌های ورودی

ج: مرتب سازی فاصله‌ها بر اساس کمترین آن‌ها و مشخص نمودن نزدیکترین‌ها بر حسب k

د: استفاده از بیشترین نزدیکترین همسایه‌ها جهت برآورد مقادیر جدید (لانگ، ۲۰۱۹).

تجزیه و تحلیل داده‌ها و آزمون فرضیه ها

در جدول زیر شاخص‌های آمار توصیفی پنج مدل مورد استفاده در پژوهش ارائه گردیده است. بر اساس شاخص‌های موصوف میانگین R^2 در روش ماشین بردار پشتیبان دارای بیشترین میزان و معادل ۰/۹۴ با انحراف معیار ۰،۰۱ می‌باشد. پس از آن به ترتیب مدل شبکه‌های عصبی مصنوعی، رگرسیون لجستیک، اعتبارسنجی متقابل و در نهایت، نزدیکترین همسایه‌های k قرار دارند. سایر آماره‌های توصیفی نیز پایایی و روایی داده‌ها را تأیید می‌نماید.

جدول ۶. آمار توصیفی مدل‌های مورد استفاده در پژوهش

شاخص های آمار توصیفی		میانگین	انحراف معیار	حداقل	حداکثر	میانه	چولگی	کشیدگی
ماشین بردار پشتیبان	R^2	۰،۹۴	۰،۰۱	۰،۹۱	۰،۹۹	۰،۹۶	۰،۲۴	۱،۱۴-
	MSE	۹۶،۲۴	۲۲،۱۷	۵۸،۳۳	۱۳۱،۹۳	۹۷،۲۸	۰،۲۴-	۱،۱۶-
	RMSE	۵۵،۴۹	۱۸،۸۲	۲۳	۸۵	۵۷	۰،۲۶-	۱،۱۵-
رگرسیون	R^2	۰،۸۹	۰،۰۲	۰،۸۳	۰،۹۵	۰،۸۹	۰،۰۱-	۱،۲۱-

1. Lin

شاخص های آمار توصیفی		میانگین	انحراف معیار	حداقل	حداکثر	میانه	چولگی	کشیدگی
لجستیک	MSE	۱۲۱٫۰۷	۱۷٫۹	۸۷٫۰۶	۱۵۰٫۸۵	۱۲۲٫۵۳	-۰٫۱۳	-۰٫۹۴
	RMSE	۸۵٫۴۶	۱۶٫۸۹	۵۳	۱۱۳	۸۶٫۵	-۰٫۱۴	-۰٫۹۳
اعتبارسنجی متقابل	R^2	۰٫۸۷	۰٫۰۲	۰٫۸۳	۰٫۹۲	۰٫۸۷	-۰٫۰۲	-۱٫۱۹
	MSE	۱۵۲٫۱۹	۱۷٫۹	۱۱۹٫۹۷	۱۸۳٫۵۵	۱۵۲٫۴۲	-۰٫۰۲	-۰٫۹۸
	RMSE	۱۱۵٫۱۳	۱۶٫۹۹	۸۴	۱۴۵	۱۱۵٫۵	-۰٫۰۵	-۰٫۹۸
نزدیکترین همسایه های k	R^2	۰٫۸۵	۰٫۰۲	۰٫۸۲	۰٫۸۸	۰٫۸۵	۰٫۰۱	-۱٫۱۵
	MSE	۱۸۱٫۳۳	۲۳٫۸۴	۱۳۷٫۴	۲۱۹٫۱	۱۸۱٫۷۴	-۰٫۱۷	-۱٫۰۵
	RMSE	۱۴۲٫۷۹	۲۲٫۶۹	۱۰۱	۱۷۹	۱۴۳٫۵	-۰٫۱۶	-۱٫۰۵
شبکه های عصبی	R^2	۰٫۹۱	۰٫۰۲	۰٫۸۸	۰٫۹۶	۰٫۹۱	-۰٫۰۱	-۱٫۲۱
	MSE	۹۸٫۱۳	۱۹٫۵۱	۶۶٫۴۹	۱۳۱٫۱۷	۹۸٫۱۵	۰٫۰۱	-۱٫۲۸
	RMSE	۶۳٫۷۸	۱۸٫۵۸	۳۳	۹۵	۶۴	۰٫۰۱	-۱٫۲۷

منبع: نتایج پژوهش

بررسی فرضیه پژوهش

فرضیه پژوهش بیان می‌کند که "دقت پیش‌بینی مدل‌های ماشین‌های بردار پشتیبان، شبکه‌های عصبی، روش نزدیک‌ترین همسایه‌های k، رگرسیون لجستیک و اعتبارسنجی متقابل با یکدیگر تفاوت دارد". برای بررسی اعتبار این فرضیه از آزمون تفاوت میانگین‌ها (تحلیل واریانس) استفاده‌شده است که با توجه به نتایج مندرج در جدول زیر سطح معنی‌داری در شاخص R^2برابر با ۰/۰۴۴۰ و کمتر از ۰/۰۵ درصد است و فرضیه H1 مبنی بر اینکه" بین دقت مدل‌های ماشین بردار پشتیبان، اعتبارسنجی متقابل، رگرسیون لجستیک، شبکه‌های عصبی و نزدیکترین همسایه اختلاف معنی‌داری وجود دارد." پذیرش و فرض صفر پژوهش مبنی بر عدم اختلاف معنی‌دار رد می‌شود. این موضوع برای شاخص‌های MSE و RMSE نیز صادق است و سطح معنی‌داری در این دو شاخص نیز به ترتیب ۰/۰۲۷ و ۰/۰۳۱ است که کمتر از ۰/۰۵ است.

با بررسی میزان دقت R^2 در جدول زیر میزان دقت مدلهای یادگیری ماشین قابل رتبه‌بندی می‌باشد به نحوی که مدل ماشین‌های بردار پشتیبان با ۰/۹۴۴ رتبه اول، مدل شبکه های عصبی با ۰/۹۰۸ رتبه دوم، مدل رگرسیون لجستیک با ۰/۸۸۸ رتبه سوم، مدل اعتبارسنجی متقابل با ۰/۸۶۸ رتبه چهارم و روش نزدیک ترین همسایه‌های K با ۰/۸۴۸ در رتبه پنجم از نظر میزان دقت قرار دارد. این موضوع در دو شاخص MSE و RMSE نیز با بزرگتر شدن اعداد در جدول همین موضوع را تصدیق می‌نماید. یعنی عدد کوچکتر در هر دو شاخص بیانگر بودن دقت بیشتر مدل می‌باشد.

جدول ۷. نتایج آزمون آماری تحلیل واریانس در خصوص دقت مدل‌های یادگیری ماشین

	SVM	GANN	Log	EM	KNN	آزمون F	سطح معنی داری
R^2	۰٫۹۴۴	۰٫۹۰۸	۰٫۸۸۸	۰٫۸۶۸	۰٫۸۴۸	۳٫۵	۰٫۰۴۴۰
MSE	۹۶٫۲۴	۹۸٫۱۳	۱۲۱٫۰۷	۱۵۲٫۱۹	۱۸۱٫۳۳	۶٫۳۸	۰٫۰۲۷
RMSE	۵۵٫۴۹	۶۳٫۷۸	۸۵٫۴۶	۱۱۵٫۱۳	۱۴۲٫۷۹	۵٫۶۴	۰٫۰۳۱

منبع: نتایج پژوهش

بحث و نتیجه گیری

بر اساس بررسی‌های صورت گرفته، در پژوهش حاضر امکان استفاده از مدل‌های یادگیری ماشین برای برآورد قیمت سهام در بازار سرمایه وجود دارد و دقت برآورد قیمت سهام در مدل‌های مختلف با یکدیگر متفاوت می‌باشد. برای بررسی اعتبار این فرضیه از آزمون تفاوت میانگین‌ها (تحلیل واریانس) استفاده‌شده است که با توجه به نتایج مندرج در جدول، سطح معنی‌داری در شاخص R^2 برابر با ۰/۰۴۴۰ و کمتر از ۰/۰۵ درصد است و فرضیه H1 مبنی بر اینکه "بین دقت مدل‌های ماشین بردار پشتیبان، اعتبارسنجی متقابل، رگرسیون لجستیک، شبکه‌های عصبی و نزدیکترین همسایه اختلاف معنی‌داری وجود دارد". پذیرش و فرض صفر پژوهش مبنی بر عدم اختلاف معنی‌دار رد می‌شود.

بر اساس نتایج پژوهش مدل ماشین بردار پشتیبان دارای بیشترین دقت در برآورد قیمت سهام می‌باشد و با R^2 ۰٫۹۴۴ در رتبه اول قرار دارد. سایر مدل‌ها به ترتیب عبارتند از روش شبکه‌های عصبی مصنوعی با ۰٫۹۰۸، رگرسیون لجستیک با ۰٫۸۸۸، اعتبارسنجی متقابل با .۸۶۸ و نزدیکترین همسایه‌های k با ۰٫۸۴۸ . این رتبه‌بندی بر اساس روش MSE و روش RMSE نیز تأیید می‌شود. به طوری که در جدول فوق مشاهده می‌شود در روش ماشین بردار پشتیبان به عنوان بهترین روش میزان این دو شاخص کمترین مقدار را به خود اختصاص داده و با کاهش دقت با همان ترتیبی که در R^2به آن دست یافتیم مقدار شاخص بیشتر می‌شود.

نتایج حاصل از این پژوهش با نتایج تجربی قبلی از این بابت مشابهت دارد که مدل‌های استفاده شده امکان پیش‌بینی قیمت سهام را فراهم می‌آورد که در جداول مربوط به پیشینه پژوهش عنوان گردید؛ لیکن از آنجا که از روش مقایسه‌ای بین مدل‌های مورد استفاده در این پژوهش، در پژوهش‌های پیشین وجود ندارد، نمی‌توان در خصوص قبول یا رد پژوهش‌های پیشین مطلبی بیان نمود؛ چرا که در واقع مقایسه طیفی از مدل‌های یادگیری ماشین نوآوری اصلی این پژوهش بوده است.

پیشنهاد برای پژوهش‌های آتی

با توجه به حجم بسیار زیاد داده‌های موجود در بازار سرمایه، عدم امکان محاسبه دستی این داده‌ها و امکان پیش‌بینی قیمت سهام در بورس اوراق بهادار از طریق مدل‌های یادگیری ماشین به سازمان بورس، شرکت بورس اوراق بهادار و فرابورس، شرکتهای تأمین سرمایه، کارگزاری، سبدگردان و مشاور سرمایه گذاری پیشنهاد می‌شود از مدل‌های یادگیری ماشین جهت بهبود سرمایه گذاری خود و مشتریان و سرمایه گذاران استفاده نمایند.

همچنین، در معاملات الگوریتمی که روش جدیدی در بازار سرمایه ایران به حساب می‌آید، می‌توان از مدل‌های یادگیری ماشین جهت انجام خودکار معاملات استفاده نمود که به طور قطع بر افزایش راندمان معاملات تأثیر مثبت خواهد داشت.

با توجه به گستردگی روش‌های یادگیری ماشین، در پژوهش‌های آینده می‌توان سایر روش‌های یادگیری ماشین نظیر جنگل تصادفی، سایر روش‌های شبکه عصبی را جهت پیش‌بینی قیمت سهام مورد استفاده قرار داد.

در این پژوهش از قیمت باز، بسته، بالاترین و پایین‌ترین قیمت و همچنین نرخ ارز، قیمت نفت و قیمت طلا به عنوان متغیرهای مستقل استفاده شده است، محققین محترم می‌توانند از اندیکاتورهایی نظیر؛ Simple Moving Average (SMA), Exponential Moving Average (EMA), Weighted Close (WTCL), Weighted 10-day Moving Average (WMA), Moving Average Convergence Divergence (MACD), Stochastic Oscillator, Average True Range (ATR), Average Directional Index (ADX), Relative Strength Index (RSI), Momentum, Williams R%, Commodity Channel Index (CCI), Money Flow Index (MFI), On Balance Volume (OBV), Accumulation/Distribution (A/D), Rate Moving Average (VMA), Variable Moving Average (VMA) Weighted, Parabolic SAR, به عنوان متغیر مستقل استفاده نمایند.

در این پژوهش از داده‌های روزانه برای یادگیری استفاده شده است؛ محققین محترم می‌توانند از داده های چهارساعته، ساعتی و لحظه‌ای برای یادگیری استفاده نمایند. این داده‌ها در پایگاه اطلاع‌رسانی TSETMC موجود می‌باشد و قابل استفاده است؛ لیکن نیازمند قدرت پردازش بیشتر سیستم می‌باشد که با توجه به رشد فناورانه در حوزه رایانه این موضوع قابل انجام می‌باشد.

نتایج حاصل شده برای دوره زمانی ۱۲ ساله ۱۳۸۸ تا ۱۳۹۹ صادق است، لذا برای بررسی اینکه آیا مدل مزبور برای همه بازه‌های زمانی کاربرد دارد یا خیر، می‌توان دوره‌های زمانی کوتاه‌تر یا طولانی‌تر و انجام تقسیم‌بندی‌های جزئی‌تر را ملاک عمل قرار داد. برای مثال، می‌توان دوره زمانی دوازده ساله را به شش دوره دوساله یا دو دوره شش‌ساله اقدام به پیش‌بینی نمود و سپس، نتایج را مقایسه کرد.

پژوهشگران محترم می‌توانند در پژوهش‌های آتی از نسبت‌های مالی و سایر متغیرهای بنیادی به عنوان متغیر مستقل استفاده نمایند و تأثیر این متغیرها را در پیش‌بینی قیمت سهام مورد سنجش قرار دهند.

ترکیب کردن شبکه عصبی و ماشین بردار پشتیبان، به این صورت که باقیمانده‌های پیش‌بینی با شبکه عصبی توسط ماشین بردار پشتیبان پیش‌بینی گردد یا استفاده از سایر روش‌های پیش‌بینی به صورت ترکیبی می‌تواند به ارتقای دقت پیش‌بینی کمک نماید.

ملاحظات اخلاقی

حامی مالی: مقاله حامی مالی ندارد.
مشارکت نویسندگان: تمام نویسندگان در آماده‌سازی مقاله مشارکت داشته‌اند.
تعارض منافع: بنا بر اظهار نویسندگان در این مقاله هیچ‌گونه تعارض منافعی وجود ندارد.
تعهد کپی‌رایت: طبق تعهد نویسندگان حق کپی‌رایت رعایت شده‌است.

References

Ahmadkhan Beigi, Sohail, Abdulvand, Neda. (2016). Stock price forecasting with a combined approach of artificial neural networks and colonial competition algorithm based on chaos theory. *Financial Management Strategy*, 5 (3), 27- 73. Doi: 10. 22051/jfm. 2017. 14635. 1319. (In Persian).

Afshari Rad, Elham, Alavi, Seyed Enayat A.and Sinaii, Hassan Ali. (2017). "An intelligent model for predicting stock trends using technical analysis methods" *Financial Research*, Summer 2017, Volume 20, Number 249. (In Persian)

Ahmadkhan Beigi, Sohail and Abdulvand, Neda. (2016). "Stock price forecasting with a hybrid approach of artificial neural networks and colonial competition algorithm based on chaos theory." *Financial management strategy*, number 5. (In Persian)

Ampomah, E. K; Z. Qin, and G. Nyame. (2020). "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement." *Information*, 11(6), 332.

Ballings, M, D Van den Poel, N Hespeels, and R. Gryp. (2015). "Evaluating multiple classifiers for stock price direction prediction. " *Expert systems with Applications*, 42(20), 7046-7056.

Bonaccorso, G. (2017). Machine learning algorithms. Packt Publishing Ltd.

Ebrahimi Heravi, Behrouz, Kazem Rangzan, Mostafa Kabuli, Hassan Daneshian. (2016). "Comparison of Artificial Neural Network and Fuzzy System Methods in Determining Flood Early Warning Time of the Yellow River Watershed Sub-basin - Khuzestan Province" *Geography and Environmental Planning*, Volume 28, Number 1, Pages 1-20. (In Persian)

Emami, Mohammad, Yatharbi, Seyyed Shahabuddin. (2013). "The use of artificial neural network in the interpretation of barometric test results". *Scientific Research Journal of Imran Modares*, 14th period, special issue (In Persian)

Gardening, Shahnaz. (2016). "Techniques and methods of machine learning on big data." National conference of new technologies in electrical and computer engineering. (In Persian)

Hafezi, R; J. Shahrabi, and E. Hadavandi. (2015). "A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price. ." *Applied Soft Computing*, 29 196-210.

Hosseininasab, Hojjat, Karimi Teklo, Salim and Yousefinejad, Marzieh. (2012). "Comparison of the accuracy of support vector machines and artificial neural networks approaches in predicting the profit per share of companies listed on the Tehran Stock Exchange." *Bi-Quarterly Economic Essays*, 109-134-10 (In Persian)

Houssein, E. H; M. Dirar, K. Hussain, and W. M. mohamed. (2021). "Artificial Neural Networks for Stock Market Prediction: A Comprehensive Review ." Metaheuristics in Machine Learning: Theory and Applications 409-444.

Kantardzik, Mehmed. (2012). Data mining, translated by Amir Alikhanzadeh, Computer Sciences Publishing. (In Persian)

Lin, Q. (2018). "Technical analysis and stock return predictability: An aligned approach." *Journal of financial markets*, 38 103-123.

Long, W; L Song, and Y Tian. (2019). "A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity ." *Expert Systems with Applications*, 118 411-424.

Nivetha, R. Y; and C. Dhaya. (2017). "Developing a prediction model for stock analysis." International Conference on Technical Advancements in Computers and Communications (ICTACC). IEEE.

Piryonesi, S. M; and T. E. El-Diraby. (2020). "Data analytics in asset management: Cost-effective prediction of the pavement condition index." *Journal of Infrastructure Systems*, 26(1), 04019036.

Rai, Reza, Pouyanfar, Ahmed. (1400), Advanced Investment Management. Side Publications (In Persian)

Sarmad, Zahra, Bazargan, Abbas and Hijazi, Elaha. (2019). Research methods in behavioral sciences. Post an ad. (In Persian)

Sigo, M. O; M. Selvam, S. Venkateswar, and C. Kathiravan. (2020). "Application of ensemble machine learning in the predictive data analytics of indian stock market." CIFR Paper Forthcoming.

Wen Long Zhichen Lu Lingxiao Cui. (2019). "Deep learning-based feature engineering for stock price movement prediction.." Knowledge-Based Systems Volume 164 ،163-173.

Zarei, Qasim, Mohammadian, Rana, Hashedi Neiri, Hatef and Ajirlu, Mohammad. (2017) "Comparison of fuzzy neural network methods with fuzzy wavelet neural networks in predicting stock prices of banks listed on the Tehran Stock Exchange" *Journal of Financial Management Strategy*, Fall 6th Year 2017 Number 2. (In Persian)

Zhou, X, L Wang, H Liao, S Wang, and B Lev. (2019). "A prospect theory-based group decision approach considering consensus for portfolio selection with hesitant fuzzy information

COPYRIGHTS

This license allows others to download the works and share them with others as long as they credit them, but they can’t change them in any way or use them commercially.