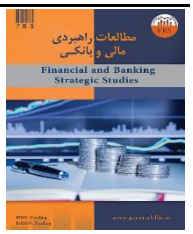
E-ISSN: 2981-1589 | P-ISSN: 2981-1570

Financial and Banking Strategic Studies

www.journal-fbs.com

Fin. Bank. Stra. Stud. Vol. 4, No. 1(2026) 22–41.



Paper Type: Original Article

Anomaly Detection in Commitment of Traders (COT) Report Data Using a Manifold Learning Approach

Parastoo Kabi-Nejad[1,*], Sara Nasher Ahkami[2]

[1] Department of Mathematics, Faculty of Mathematics and Computer Science, Iran University of Science and Technology, Tehran, Iran; parastookabinejad@iust.ac.ir.

[2] Department of Computer Science, Faculty of Mathematics and Computer Science, Iran University of Science and Technology, Tehran, Iran; saranasher8@gmail.com.

Citation:

Kabi-Nejad, P., & Nasher ahkami, S. (2026). Anomaly detection in commitment of traders (COT) report data using manifold learning approach. *Financial and Banking Strategic Studies, 4*(1), 22-41.

Received: 02/12/2025 Reviewed: 19/01/2026 Revised: 09/03/2026 Accepted: 12/05/2026

Abstract

Purpose: The objective of this research is to present a comprehensive framework for identifying anomalies in Commitments of Traders (COT) report data and to investigate their role in detecting economic trends, market disruptions, and sudden changes.

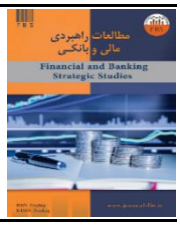
Methodology: In this study, following the preprocessing of the COT data, statistical methods, including the standard score and the interquartile range, were combined with machine learning algorithms, including Isolation Forest and One-Class Support Vector Machine. Then, by creating ensemble anomalies, the points identified as anomalous by all methods were extracted. Subsequently, linear and non-linear dimensionality reduction methods PCA, Isomap, UMAP, and LLE, were applied to the COT dataset, and the One-Class Support Vector Machine, Local Outlier Factor, Isolation Forest, and K-Means algorithms were implemented for anomaly detection.

Findings: The findings demonstrated that the detected anomalies closely coincide with prominent macroeconomic disruptions, notably the 2008 Global Financial Crisis and the economic shock of the 2020 COVID-19 pandemic. Additionally, the consensus framework, grounded in the intersection of model outputs, effectively filtered alarms driven by algorithm-specific sensitivities, thereby yielding a focused subset of 1,638 observations characterized by the highest inter-algorithm consensus.

Originality/Value: This research, by integrating statistical methods, machine learning, and Manifold Learning, provides an accurate and reliable framework for analyzing complex financial market data and creates the foundation for developing interactive dashboards for real-time market monitoring.

Keywords: Anomaly detection, Commitments of traders data, Isolation forest, One-class support vector machine, Manifold learning, Market analysis.

Corresponding Author: parastookabinejad@iust.ac.ir 10.22105/fbs.2026.247779



مطالعات راهبردی مالی و بانکی

دوره ۴، شماره (۱)، (۱۴۰۵)، ۲۲-۴۱

www.journal-fbs.com

نوع مقاله: پژوهشی

تشخیص ناهنجاری در داده‌های گزارش تعهدات معامله‌گران (COT) با استفاده از رویکرد یادگیری منیفلد

پرستو کعبی نژاد[۱،*]، سارا ناشر احکامی[۲]

[۱] گروه ریاضی، دانشکده ریاضی و علوم کامپیوتر، دانشگاه علم و صنعت، تهران، ایران.

[۲] گروه کامپیوتر دانشکده ریاضی و علوم کامپیوتر، دانشگاه علم و صنعت، تهران، ایران.

چکیده

هدف: هدف این پژوهش ارایه یک چارچوب جامع برای شناسایی ناهنجاری‌ها در داده‌های گزارش تعهدات معامله‌گران[۱] و بررسی نقش آن‌ها در شناسایی روندهای اقتصادی، اختلالات بازار و تغییرات ناگهانی است.

روش‌شناسی پژوهش: در این مطالعه، پس از پیش‌پردازش داده‌های *COT*، روش‌های آماری شامل نمره استاندارد و دامنه بین‌چارکی با الگوریتم‌های یادگیری ماشین شامل جنگل ایزوله و ماشین بردار پشتیبان تک‌کلاسه ترکیب شدند. سپس با استفاده از یک راهبرد اجماعی، نقاطی که توسط تمامی روش‌ها به‌عنوان ناهنجار شناسایی شده بودند استخراج شدند. در ادامه، روش‌های کاهش بعد خطی و غیرخطی، *PCA*، ایزومپ، *Umap* و *LLE* را بر مجموعه داده‌های *COT* اعمال نمودیم و سپس الگوریتم‌های ماشین بردار پشتیبان تک کلاسه، عامل پرت محلی، جنگل ایزوله و *k-means* را برای تشخیص ناهنجاری‌ها پیاده‌سازی نموده‌ایم.

یافته‌ها: نتایج نشان داد که ناهنجاری‌های شناسایی‌شده از نظر زمانی با برخی از رویدادهای برجسته اقتصاد کلان، از جمله بحران مالی جهانی ۲۰۰۸ و شوک اقتصادی ناشی از همه‌گیری کووید-۱۹ در سال ۲۰۲۰، هم‌زمانی و تطابق قابل‌توجهی دارند. همچنین، رویکرد اجماعی مبتنی بر اشتراک خروجی الگوریتم‌ها توانست با پالایش هشدارهای ناشی از حساسیت‌های منفردِ هر روش، زیرمجموعه‌ای متمرکز شامل ۱۶۳۸ مشاهده با بالاترین میزان توافق میان‌الگوریتمی استخراج کند.

اصالت/ارزش افزوده علمی: این پژوهش با تلفیق روش‌های آماری، یادگیری ماشین و یادگیری منیفلد، چارچوبی دقیق و قابل‌اعتماد برای تحلیل داده‌های پیچیده بازارهای مالی ارایه می‌دهد و زمینه توسعه داشبوردهای تعاملی برای پایش بازار را فراهم می‌کند.

کلیدواژه‌ها: تشخیص ناهنجاری، داده‌های تعهدات معامله‌گران، جنگل ایزوله، ماشین بردار پشتیبان تک کلاسه، یادگیری منیفلد، تحلیل بازار.

۱- مقدمه

گزارش تعهدات معامله‌گران که به‌صورت هفتگی توسط کمیسیون معاملات آتی کالا (*CFTC*[۲]) منتشر می‌شوند، به‌عنوان منبعی کلیدی برای اطلاعات مربوط به موقعیت‌های نگهداری‌شده توسط دسته‌های مختلف مشارکت‌کنندگان بازار ازجمله پوشش‌دهندگان تجاری[۳]، معامله‌گران غیرتجاری[۴] و معامله‌گران خرد شناخته می‌شوند. این گزارش‌ها شفافیت در بازارهای آتی را افزایش می‌دهند و به‌طور گسترده توسط تحلیلگران، معامله‌گران و پژوهشگران برای بررسی تمایلات و احساسات بازار و رفتار اتخاذ موقعیت مورد استفاده قرار می‌گیرند [1]. ناهنجاری‌ها در

[1] Commitments of Traders (COT)

[2] Commodity Futures Trading Commission (CFTC)

[3] Commercial hedgers

[4] Non-commercial traders

موقعیت‌های معاملاتی که به‌عنوان انحراف‌های ناگهانی یا غیرعادی از الگوهای تاریخی تعریف می‌شوند، می‌توانند نشان‌دهنده رویدادهای بحرانی بازار مانند شوک‌های نقدینگی، حباب‌های سفته‌بازی یا آغاز بحران‌ها باشند [2]. تشخیص چنین ناهنجاری‌هایی برای پژوهش‌های دانشگاهی و کاربردهای عملی در مدیریت ریسک مالی از اهمیت بالایی برخوردار است [3]. مطالعات سنتی داده‌های *COT* تا حد زیادی بر آمار توصیفی، شاخص‌های احساسی یا مدل‌های پیش‌بینی پویایی قیمت متمرکز بوده‌اند. با این حال، تعداد نسبتا محدودی از مطالعات بر تشخیص سیستماتیک ناهنجاری‌ها در بازارهای مختلف و استفاده از روش‌های ترکیبی برای افزایش پایداری نتایج تاکید کرده‌اند [4]. پیشرفت‌های اخیر در روش‌های تشخیص ناهنجاری، هم شامل الگوریتم های آماری ساده مانند نمره استاندارد Z و دامنه بین چارکی و هم الگوریتم‌های پیشرفته یادگیری ماشین مانند جنگل ایزوله و ماشین بردار پشتیبان تک کلاسه می‌شوند. هر یک از این روش‌ها دیدگاه متفاوتی درباره شناسایی و تعریف مشاهدات ناهنجار ارایه می‌کنند: رویکردهای آماری، انحراف‌های عددی شدید را برجسته می‌کنند، درحالی‌که روش‌های یادگیری ماشین، الگوهای پنهان و غیرخطی را شناسایی می‌کنند. با این‌حال، هر روش به‌تنهایی دارای محدودیت‌هایی ازجمله حساسیت به انتخاب پارامترها یا آسیب‌پذیری در برابر نویز است [5]. یادگیری منیفلد (*Manifold Learning*) مجموعه‌ای از روش‌های کاهش بُعد است که با بهره‌گیری از ساختار هندسی داده‌ها، نمایش کم‌بُعدی از داده‌های پُربُعد ایجاد می‌کند. که ساختار داده‌ها را در فضایی با بعد کمتر بازنمایی می‌کند و هم‌زمان ساختار و اطلاعات هندسی نهفته در داده‌ها را حفظ می‌نماید. در واقع، روش‌های یادگیری منیفلد با کشف این ساختار هندسی، داده‌های با ابعاد بالا را به فضایی با بعد پایین‌تر نگاشت می‌کنند تا الگوهای غیرخطی داده‌ها و ذاتا پیچیده داده ها بدون از دست رفتن اطلاعات مهم حفظ شود [12] .در این مقاله، یک رویکرد یکپارچه برای تشخیص ناهنجاری در داده‌های *COT* ارایه می‌دهیم که رویکردهای آماری و یادگیری ماشین را ترکیب می‌کند. به‌طور خاص، آرشیوهای خام اکسل *COT* را به یک مجموعه داده ساختاریافته شامل بیش از ۴۳۳۰۰۰ رکورد هفتگی پیش‌پردازش کرده، چهار الگوریتم تشخیص ناهنجاری را اعمال می‌کنیم و با ترکیب خروجی آن‌ها، یک مجموعه ناهنجاری ترکیبی با اطمینان بالا استخراج می‌کنیم. علاوه بر این، یک داشبورد بصری تعاملی ارایه می‌دهیم که امکان کاوش ناهنجاری‌ها در بازارهای مختلف را از طریق نمودارهای قابل تنظیم و مصورسازی‌های تعاملی برای تحلیلگران فراهم می‌کند. سپس، با اعمال روش های کاهش بعد، الگوریتم‌های ماشین بردار پشتیبان تک کلاسه، عامل پرت محلی، جنگل ایزوله و *k-means* را برای تشخیص ناهنجاری ها به کار گرفتیم.

دستاوردهای این پژوهش در سه بخش قابل بیان است:

۱. یک چارچوب پیش‌پردازش و تشخیص ناهنجاری قابل تکرار برای گزارش‌های *COT* طراحی و پیاده‌سازی کرده‌ایم.
۲. عملکرد مقایسه‌ای روش‌های آماری و روش‌های یادگیری ماشین را ارزیابی می‌کنیم و از تحلیل هم‌پوشانی و تحلیل پایداری به‌عنوان معیارهای ارزیابی استفاده می‌نماییم.
۳. روش های کاهش بعد را به کار می گیریم و سپس با پیاده کردن چهار الگوریتم روی مجموعه داده ها به مقایسه آن ها می پردازیم.

۲- پژوهش‌های مرتبط

پژوهش‌های مرتبط با گزارش‌های تعهدات معامله‌گران *COT* عمدتا بر استفاده از این گزارش‌ها به‌عنوان شاخص‌های احساسات بازار و سیگنال‌های پیش‌بینی‌کننده نوسانات قیمت تمرکز داشته‌اند. مطالعات اولیه عموما از آمار توصیفی یا روش‌های مبتنی بر رگرسیون برای بررسی رفتار معامله‌گران تجاری و غیرتجاری استفاده کرده‌اند. این مطالعات نشان داده‌اند که تغییرات در موقعیت‌های معامله‌گران می‌تواند شاخص‌های پیش‌روی از نقاط تحول بازار و فعالیت‌های سفته‌بازی ارایه دهد.

علاوه بر تحلیل‌های توصیفی، شاخص‌های احساسات استخراج‌شده از داده‌های *COT* به چارچوب‌های پیش‌بینی‌کننده نیز به‌کار رفته‌اند. به‌عنوان‌مثال، برخی مطالعات موقعیت‌های خالص معامله‌گران غیرتجاری را با پویایی‌های آتی قیمت کالاها مرتبط دانسته‌اند و قدرت پیش‌بینی‌کنندگی آن‌ها را در شرایط پرنوسان بازار برجسته کرده‌اند. با این حال، این رویکردها معمولا امکان وجود ناهنجاری‌های ساختاری در رفتار معامله‌گران را که توسط مدل‌های مبتنی بر روند به‌طور کامل قابل شناسایی نیستند، نادیده می‌گیرند [7].

در سال‌های اخیر، پژوهش‌هایی پیرامون تشخیص ناهنجاری در بازارهای مالی به میزان قابل‌توجهی گسترش یافته است. روش‌های آماری مانند آستانه‌های نمره Z و دامنه بین چارکی[1] به دلیل سادگی و قابل تفسیر بودن کماکان محبوبیت دارند. بااین‌حال، این روش‌ها به فرضیات توزیعی حساس هستند و در مواردی که داده‌ها دارای دنباله سنگین یا ساختارهای غیرخطی باشند ممکن است با شکست مواجه شوند. روش‌های یادگیری ماشین چشم‌اندازهای جایگزین ارایه می‌دهند. الگوریتم‌هایی مانند جنگل ایزوله و ماشین‌های بردار پشتیبان تک کلاسه برای تشخیص ناهنجاری در سری‌های زمانی و مجموعه داده‌های مالی چندبعدی به کار گرفته‌شده‌اند. این روش‌ها به‌ویژه برای شناسایی الگوهای پنهان و غیرخطی مفید هستند، اگرچه نیازمند تنظیم دقیق پارامترها و منابع محاسباتی کافی می‌باشند [8]، [9]. علیرغم این پیشرفت‌ها، مطالعاتی به کاربرد راهبردهای ترکیبی که چندین الگوریتم تشخیص ناهنجاری را برای تحلیل نظام‌مند داده‌های *COT* با هم تلفیق می‌کنند پرداخته‌اند[6].

۳- مجموعه داده و پیش‌پردازش

۳-۱- توصیف مجموعه داده

مجموعه داده مورد استفاده در این مطالعه از گزارش‌های تعهدات معامله‌گران که به‌صورت هفتگی توسط کمیسیون معاملات آتی کالا[2] منتشر می‌شوند، استخراج‌شده است. این گزارش‌ها حاوی اطلاعات دقیق در مورد موقعیت‌های سه دسته اصلی معامله‌گران هستند: پوشش‌دهندگان تجاری، سفته‌بازان غیرتجاری و معامله‌گران خرد. با ادغام و تجمیع چندین مجموعه داده در قالب اکسل، یک مجموعه داده ساختاریافته متشکل از بیش از ۴۳۳۰۰۰ رکورد هفتگی به دست آمد که طیف گسترده‌ای از بازارهای آتی و اختیار معامله را پوشش می‌دهد. این مجموعه داده هم ازنظر گستردگی (شامل کلاس‌های دارایی مختلف مانند کالاها، ارزها و شاخص‌های سهام) و هم ازنظر عمق زمانی (شامل داده‌های تاریخی چند دهه‌ای) غنی است [10].

جدول ۱- خلاصه ساختار مجموعه‌داده و ویژگی‌های کلیدی.
Table 1- Summary of dataset structure and key features.

توضیح	ستون
تاریخ هفتگی گزارش	Report_Date
نام بازار و بورس مربوطه	Market_and_Exchange_Names
درصد موقعیت خالص معامله‌گران غیرتجاری	NonComm_Net_pct
برچسب دودویی ناهنجاری شناسایی‌شده با استفاده از Z-Score	Anomaly_zscore
برچسب دودویی ناهنجاری شناسایی‌شده با استفاده از IQR	Anomaly_iqr
برچسب دودویی ناهنجاری شناسایی‌شده با استفاده از جنگل ایزوله	Anomaly_iforest
برچسب دودویی ناهنجاری شناسایی‌شده با استفاده از One-Class SVM	Anomaly_ocsvm
برچسب دودویی ترکیبی که نشان‌دهنده شناسایی ناهنجاری توسط تمام الگوریتم‌ها است	All_algorithms

جدول ۱ نمایی مختصر از مجموعه‌داده ارایه می‌دهد که شامل توصیف ستون‌ها و آمارهای مرتبط بوده و متغیرهای اصلی مورداستفاده در تحلیل را برجسته می‌سازد.

[1] Interquartile Range (IQR)

[2] Commodity Futures Trading Commission (CFTC)

۲-۳- مهندسی ویژگی‌ها

از گزارش‌های خام *COT*، چندین متغیر مشتق‌شده برای تشخیص ناهنجاری ساخته‌شده‌اند که بدین شرح می‌باشند: درصدهای موقعیت خالص معامله‌گران غیرتجاری که موقعیت گیری نسبی را در طول زمان ثبت می‌کند، تغییرات در موقعیت‌های خالص هفتگی که تغییرات ناگهانی در احساسات بازار را برجسته می‌سازد، شناسه‌های بازار، شامل نام کالا و بورس، برای دسته‌بندی ناهنجاری‌ها در سطوح مختلف مورد استفاده قرار گرفتند. این ویژگی‌های مهندسی‌شده می‌توانند تفسیر نتایج تشخیص ناهنجاری را تسهیل کنند و امکان بررسی انحراف‌های معنادار در رفتار معامله‌گران را فراهم آورند.

۳-۳- مراحل پیش‌پردازش

چارچوب پیش‌پردازش شامل مراحل زیر می‌باشد:

۱. پاک‌سازی داده‌ها که همان حذف سطرهای ناقص یا خراب می‌باشد.
۲. استانداردسازی که درواقع یکسان‌سازی نام متغیرها و شناسه‌های بازار در طول سال‌های مختلف می‌باشد.
۳. تبدیل تاریخ که همان تبدیل گزارش هفتگی به قالب زمانی یکپارچه است.
۴. مقیاس‌گذاری ویژگی‌ها که استانداردسازی ستون‌های عددی برای پشتیبانی از الگوریتم‌های آماری و یادگیری ماشین می‌باشد.
۵. تجمیع برچسب‌های ناهنجاری که با ایجاد ستون‌های دودویی، ناهنجاری‌های شناسایی‌شده توسط هر الگوریتم را نشان می‌دهند و بنابراین یک ستون اضافی (*all_algorithms*) برای تحلیل ترکیبی خواهیم داشت.

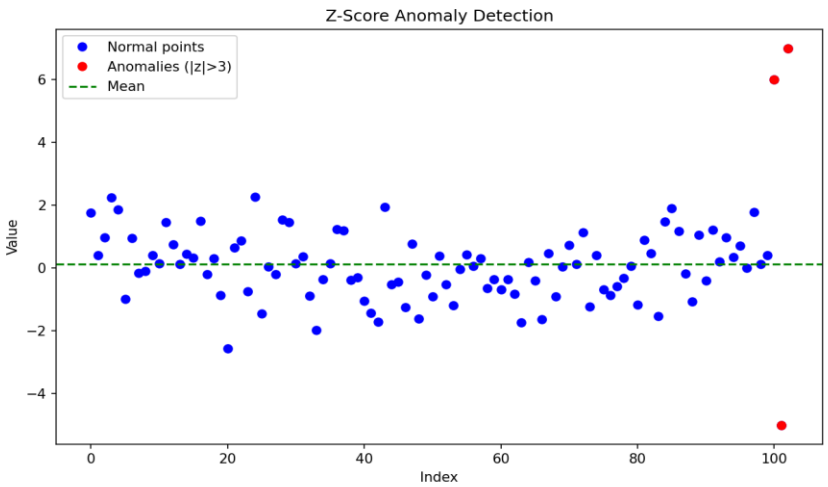
این چارچوب پیش‌پردازش، قابلیت تکرارپذیری را تضمین می‌کند و یک مجموعه داده یکپارچه مناسب برای اعمال روش‌های تشخیص ناهنجاری آماری و یادگیری ماشین فراهم می‌کند.

۴- روش‌ها

۱-۴- روش‌های آماری

از دو روش آماری کلاسیک استفاده نموده‌ایم: نمره Z و دامنه بین چارکی.

نمره Z ناهنجاری‌ها را به‌عنوان نقاطی شناسایی می‌کند که مقدار استانداردشده مطلق آن‌ها از یک آستانه مشخص فراتر می‌رود ($|z| > 3$) . این روش برای شناسایی انحراف‌های شدید کارآمد است، اما به ناهنجاری‌ها بسیار حساس می‌باشد.



شکل ۱- مکانیسم روش امتیاز Z برای تشخیص داده‌های پرت.

Figure 1- Mechanism of the Z-score method for detecting outliers.

تصویری از چگونگی انحراف نقاط داده‌ای که بیش از یک آستانه تعیین‌شده (مثلا $|z| > 3$) از میانگین انحراف دارند و به‌عنوان ناهنجاری علامت‌گذاری می‌شوند.

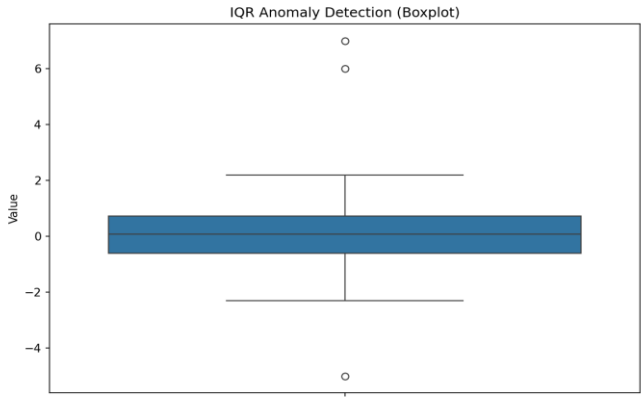
۲-۴- شناسایی ناهنجاری با استفاده از IQR

روش *IQR* ناهنجاری‌ها را به‌عنوان مقادیری شناسایی می‌کند که خارج از بازه‌ی زیر قرار می‌گیرند:

$$[Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR],$$

در این رابطه $Q1$ چارک اول، $Q3$ چارک سوم و $IQR = Q3 - Q1$ دامنه بین‌چارکی داده‌ها است.

در مقایسه با روش *Z-Score*، روش *IQR* از استواری بالاتری در برابر توزیع‌های غیرنرمال و داده‌های دارای نویز برخوردار است. به همین دلیل، این روش به‌ویژه برای داده‌های مالی که اغلب دارای چولگی، دم‌های سنگین و نوسانات شدید هستند، مناسب‌تر است.

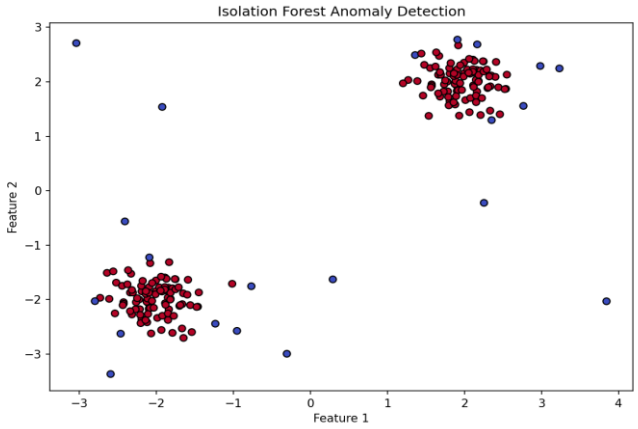


شکل ۲- تشخیص داده‌های پرت در نمودار جعبه‌ای.
Figure 2- Detecting outliers in a box plot.

درواقع **شکل ۲** فرآیند شناسایی داده‌های پرت را با استفاده از نمودار جعبه‌ای نشان می‌دهد. همان‌طور که در این شکل مشاهده می‌شود، مقادیر پرت به‌صورت نقاطی خارج از مرزهای جعبه نمایش داده‌شده‌اند.

۳-۴- روش‌های یادگیری ماشین و منیفلد

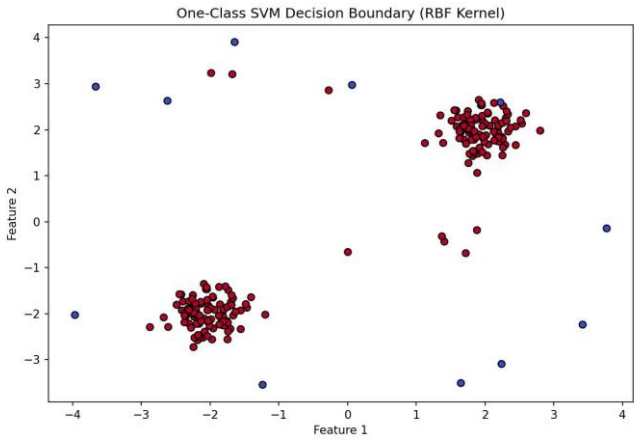
دو مدل پیشرفته‌ی یادگیری ماشین برای شناسایی ناهنجاری‌ها را به‌کارگرفتیم. الگوریتم جنگل ایزوله یک روش مبتنی بر مجموعه‌ای از درخت‌های جداسازی است که ناهنجاری‌ها را از طریق تقسیم‌بندی تصادفی ویژگی‌ها شناسایی می‌کند. این الگوریتم به‌ویژه برای مجموعه‌داده‌های با ابعاد بالا موثر است، زیرا مشاهدات ناهنجار معمولا با تعداد تقسیم‌های کمتری نسبت به داده‌های عادی منزوی می‌شوند. در این پژوهش، این مدل با پارامتر آلودگی برابر با ۰٫۰۱ تنظیم شد.



شکل ۳- ساز وکار الگوریتم جنگل ایزوله (Isolation Forest).
Figure 3- Isolation Forest anomaly detection mechanism.

شکل ۳ نقاط داده ناهنجار در عمق کمتر درخت یا با تعداد تقسیم‌های کمتری به‌سرعت ایزوله می‌شوند که این ویژگی، مبنای الگوریتم جنگل ایزوله برای شناسایی ناهنجاری‌ها را تشکیل می‌دهد.

ماشین بردار پشتیبان تک کلاسه یک روش مبتنی بر کرنل است که یک مرز تصمیم در فضای ویژگی برای تفکیک ناحیه اصلی داده‌ها یاد می‌گیرد. با استفاده از هسته تابع پایه شعاعی[1] یا هسته *RBF* (تابع پایه شعاعی)، این الگوریتم قادر به مدل‌سازی روابط غیرخطی بوده و برای شناسایی انحراف‌های ظریف و پیچیده مناسب است.

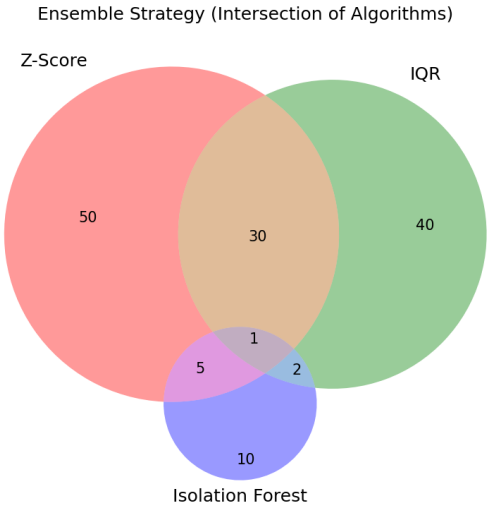


شکل ۴- مرز تصمیم‌گیری الگوریتم ماشین بردار پشتیبان تک‌کلاسه و نقاط ناهنجار خارج از مرز تصمیم‌گیری.
Figure 4-The decision boundary of the single-class support vector machine algorithm and outliers outside the decision boundary.

۴-۴- راهبرد تجمیعی (Ensemble)

ازآن‌جا که هر یک از الگوریتم‌ها ممکن است زیرمجموعه‌های متفاوتی از ناهنجاری‌ها را شناسایی کنند، یک برچسب ناهنجاری تجمیعی با عنوان *all_algorithms* ساخته شد. این برچسب تنها نقاطی را به‌عنوان ناهنجار علامت‌گذاری می‌کند که به‌طور هم‌زمان توسط هر چهار روش شناسایی شده باشند. این راهبرد محافظه‌کارانه موجب افزایش اطمینان به نتایج و کاهش تعداد هشدارهای کاذب می‌شود.

[1] Radial Basis Function (RBF)



شکل ۵- راهبرد ترکیبی و نواحی هم‌پوشانی میان الگوریتم‌های مختلف شناسایی ناهنجاری و نقاط شناسایی‌شده به‌عنوان ناهنجار توسط تمامی روش‌ها.

Figure 5- Combined strategy and overlap regions among different anomaly detection algorithms and points identified as anomalous by all methods.

۵-۴- جزییات پیاده‌سازی

پیاده‌سازی این پژوهش با استفاده از پایتون نسخه ۳٫۱۱ انجام شد و از کتابخانه‌های زیر بهره گرفته‌شده است:

کتابخانه‌های *Pandas* و *numpy* برای پیش‌پردازش داده‌ها و مهندسی ویژگی‌ها، *scikit-learn* برای پیاده‌سازی الگوریتم‌های جنگل ایزوله و ماشین بردار پشتیبان تک‌کلاسه و *matplotlib* و *seaborn* برای مصورسازی داده‌ها استفاده‌شده‌اند.

مجموعه داده نهایی شامل ٤٣٣٥٤٤ رکورد هفتگی است که بازه‌ی زمانی ۱۵ ژانویه ۱۹۸٦ تا ۳۱ دسامبر ۲۰۲٤ را پوشش می‌دهد.

جدول ۲ ویژگی‌های کلیدی و مقایسه روش‌های شناسایی ناهنجاری به‌کار رفته در این پژوهش این جدول نوع روش، نقاط قوت، نقاط ضعف و پارامترهای اصلی چهار الگوریتم شناسایی ناهنجاری را خلاصه می‌کند. روش‌های آماری مانند *Z-Score* و *IQR* ساده و سریع هستند، اما ممکن است ناهنجاری‌های نامحسوس را شناسایی نکنند؛ در مقابل، روش‌های جنگل ایزوله و ماشین بردار پشتیبان تک‌کلاسه قادر به شناسایی الگوهای پیچیده هستند، هرچند به بهای نیاز محاسباتی بالاتر و حساسیت بیشتر نسبت به تنظیم پارامترها باشد.

جدول ۲- مقایسه روش‌های تشخیص ناهنجاری از نظر نوع روش، مزایا، معایب و پارامترهای کلیدی.

Table 2- Comparison of anomaly detection methods in terms of method type, strengths, limitations, and key parameters.

روش	نوع (خطی / غیرخطی)	نقاط قوت	نقاط ضعف	پارامترهای کلیدی
Z-Score	خطی	ساده، سریع و قابل تفسیر	حساس به نویز و داده‌های با توزیع غیرنرمال	آستانه Z = 3
IQR	خطی (دامنه بین‌چارکی)	مقاوم‌تر در برابر داده‌های پرت	ممکن است ناهنجاری‌های ظریف را شناسایی نکند.	Q3، Q1، ضریب = 1.5
Isolation Forest	غیرخطی (مبتنی بر درخت)	توانمند در شناسایی الگوهای پنهان و چندبعدی	وابسته به تنظیم پارامتر آلودگی	contamination = 0.01، n_estimators
One-Class SVM	غیرخطی کرنل (RBF)	قابلیت مدل‌سازی الگوهای غیرخطی پیچیده	هزینه محاسباتی بالا و حساس به پارامترها	γ = scale، ν = 0.01

۵- تظیمات آزمایشی

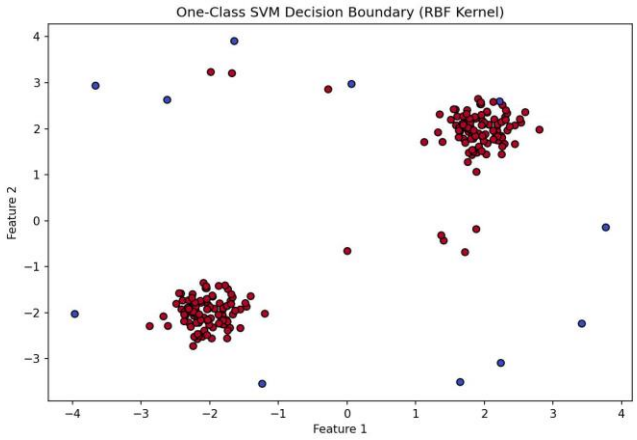
۵-۱- راهبرد ارزیابی

در این پژوهش، به دلیل نبود برچسب‌های واقعی (*Ground Truth*) برای ناهنجاری‌ها، از راهبردهای شبه‌واقعیت مبنا (*Pseudo-Ground Truth*) استفاده شد. به‌طور مشخص، اشتراک خروجی هر چهار الگوریتم به‌عنوان مجموعه‌ای از ناهنجاری‌ها با سطح اطمینان بالا در نظر گرفته شد. افزون بر این، معیارهای هم‌پوشانی مانند اشتراک‌های دوتایی، هم‌پوشانی‌های سه‌تایی و شاخص ژاکارد (*Jaccard Index*) محاسبه شدند تا میزان توافق میان الگوریتم‌ها کمی سازی شود. این رویکرد امکان ارزیابی نسبی دقت و پوشش روش‌های مختلف را فراهم می‌کند.

۵-۲- مطالعه موردی

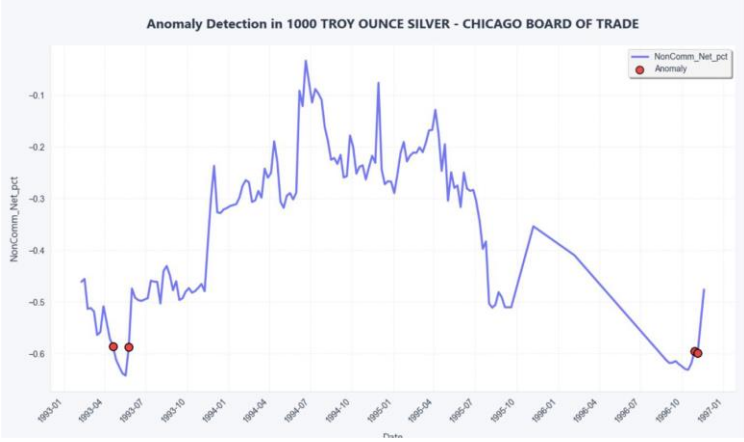
به‌منظور ارایه تحلیلی دقیق‌تر، بازار نقره ۱۰۰۰ اونس تروا در بورس تجارت شیکاگو[1] به‌عنوان مطالعه موردی انتخاب شده است. شکل ۶ سری زمانی درصد موقعیت خالص معامله‌گران غیرتجاری (*NonComm_Net_pct*) را نشان می‌دهد که در آن ناهنجاری‌هایی که به‌طور مشترک توسط تمامی الگوریتم‌ها شناسایی شده‌اند، با رنگ قرمز مشخص شده‌اند.

شکل ٦ نشان می‌دهد که ناهنجاری‌ها اغلب با دوره‌های افزایش نوسان یا شوک‌های بازار، به‌ویژه در بازه زمانی ۱۹۹۳ تا ۱۹۹٦، هم‌زمان بوده‌اند. برای نمونه، در برخی بازه‌های مرتبط با معاملات سفته‌بازانه یا تغییرات قیمتی ناشی از سیاست‌گذاری، رفتار معامله‌گران به‌طور محسوسی از الگوهای معمول فاصله گرفته است. این یافته‌ها نشان می‌دهند که ناهنجاری‌های شناسایی شده می‌توانند به‌عنوان شاخص‌های هشدار اولیه برای تحرکات غیرعادی بازار در قراردادهای آتی نقره مورد استفاده قرار گیرند.



الف-

[1] Chicago Board of Trade (CBOT)



ب-

شکل ۶- تحلیل ناهنجاری‌ها در بازار نقره (CBOT)؛ الف- درصد موقعیت خالص معامله‌گران غیرتجاری، ۱۹۹۳-۱۹۹۶؛ ب- ناهنجاری‌های با اطمینان بالا شناسایی‌شده توسط راهبرد تجمیعی و مشترک میان همه الگوریتم‌های تشخیص.

Figure 6. Anomaly analysis in the CBOT silver market; a. Net position percentage of non-commercial traders, 1993–1996, b. High-confidence anomalies identified by the ensemble strategy and jointly detected by all algorithms.

۳-۵- اثبات پایداری و تحلیل حساسیت

برای ارزیابی پایداری نتایج، یک‌شبه برچسب واقعی[1] تعریف شد که ستون *all_algorithms* به‌عنوان برچسب مرجع مورد استفاده قرار گرفت. در این ساختار، نقاطی که به‌طور هم‌زمان توسط هر چهار الگوریتم به‌عنوان ناهنجاری شناسایی شده بودند، به‌عنوان ناهنجاری‌های واقعی در نظر گرفته شدند.

با استفاده از این مرجع، معیارهای *Precision*،*Recall* و *F1-Score* برای هر الگوریتم محاسبه شد. این معیارها نشان می‌دهند که هر روش تا چه حد دقیق و جامع ناهنجاری‌ها را شناسایی می‌کند. نتایج در **جدول ۳** خلاصه شده‌اند.

روش *Z-Score* بیشترین *Precision* را داشت، به این معنی که بیشتر ناهنجاری‌های گزارش‌شده با مجموعه مرجع هم‌پوشانی داشتند. روش *IQR*، *Recall* بالاتری ارایه داد و بخش بزرگ‌تری از ناهنجاری‌های بالقوه را شناسایی کرد، اما *Precision* کمتری داشت. الگوریتم *Isolation Forest* تعادل خوبی بین *Precision* و *Recall* ایجاد کرد و *F1-Score* بالایی داشت. *One-Class SVM* با این‌که تعداد بیشتری ناهنجاری را شناسایی کرد، *Precision* پایین‌تری داشت که حساسیت آن به نویز و الگوهای پیچیده را نشان می‌دهد. این نتایج بر ارزش راهبرد تجمیعی چند الگوریتمی (*Ensemble*) تاکید دارند، چراکه هم نرخ هشدارهای کاذب را کاهش می‌دهد و هم از غفلت از ناهنجاری‌های مهم جلوگیری می‌کند.

جدول ۳- ارزیابی الگوریتم‌های شناسایی ناهنجاری با استفاده از Precision، Recall و F1-Score نسبت به مرجع مبتنی بر راهبرد تجمیعی.

Table 3- Evaluation of anomaly detection algorithms using Precision, Recall, and F1-Score compared to the reference based on the aggregation strategy.

الگوریتم	درست مثبت (TP)	نادرست مثبت (FP)	نادرست منفی (FN)	درست منفی (TN)	پیش‌بینی مثبت	Precision	Recall	F1-Score
Z-Score	1,638	5,560	0	426,346	7,198	0.228	1.0	0.371
IQR	1,638	31,278	0	400,628	32,916	0.050	1.0	0.095
Isolation Forest	1,638	2,648	0	429,258	4,286	0.382	1.0	0.553
One-Class SVM	1,638	21,926	0	409,980	23,564	0.070	1.0	0.130

[1] Pseudo-ground truth

توضیح ستون‌ها:

TP (مثبت واقعی): تعداد ناهنجاری‌های شناسایی‌شده که در مرجع واقعی نیز ناهنجار بوده اند.

FP (مثبت کاذب): تعداد نقاطی که به اشتباه به‌عنوان ناهنجار شناسایی شده‌اند.

FN (منفی کاذب): تعداد ناهنجاری‌های واقعی که شناسایی نشده‌اند.

TN (منفی واقعی): تعداد نقاط درست شناسایی شده به‌عنوان نرمال.

(*TP+FP*)، (پیش‌بینی مثبت): مجموع نقاطی که الگوریتم به‌عنوان ناهنجار شناسایی کرده است.

Precision: نسبت درست مثبت‌ها به کل پیش‌بینی‌های مثبت.

Recall: نسبت درست مثبت‌ها به کل ناهنجاری‌های واقعی.

F1-Score: میانگین هارمونیک *Precision* و *Recall* معیار کلی عملکرد الگوریتم.

۶- نتایج

۶-۱- تعداد ناهنجاری‌ها در هر بازار

پس از اعمال چهار الگوریتم شناسایی ناهنجاری *Z-Score*، *IQR*، *Isolation Forest* و *One-Class SVM* بر روی مجموعه داده‌های *CFTC*، تعداد ناهنجاری‌های شناسایی‌شده در هر بازار محاسبه شد. برای افزایش پایداری نتایج، ستونی ترکیبی با عنوان *all_algorithms* ایجاد شد که تنها شامل نقاطی است که به‌طور هم‌زمان توسط هر چهار روش به‌عنوان ناهنجاری علامت‌گذاری شده‌اند.

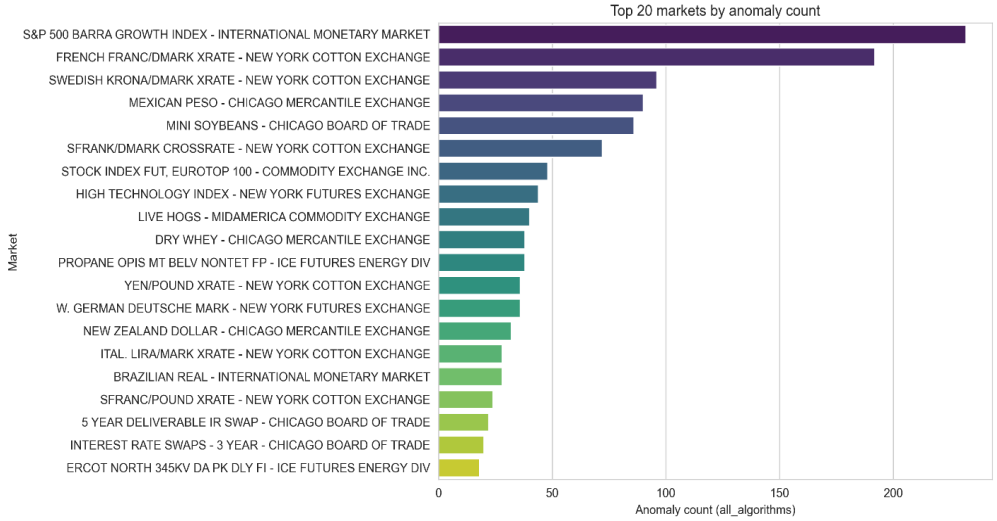
نتایج نشان‌دهنده ناهمگونی قابل‌توجه بین بازارها می‌باشد. برخی بازارها تقریبا بدون ناهنجاری می‌باشند، درحالی‌که بازارهایی مانند شاخص رشد *S&P 500 BARRA* بازار بین‌المللی پول[1]، تعداد زیادی ناهنجاری را نشان دادند. این امر بیانگر آن است که شدت و فرکانس رفتار غیرمعمول معامله‌گران در بازارها به‌طور قابل‌توجهی متفاوت است و اهمیت تمرکز بر بازارهای با نوسان بالا را برجسته می‌کند.

جدول ٤- ده بازار برتر با بیشترین تعداد ناهنجاری‌های شناسایی‌شده.

Table 4- Top ten markets with the highest number of identified anomalies.

نام بازار و بورس	تعداد ناهنجاری‌ها (تمام الگوریتم‌ها)
شاخص رشد S&P 500 BARRA - بازار بین‌المللی پول	323
فرانک فرانسه/نرخ ارز DMARK - بورس پنبه نیویورک	193
کرون سوئد/نرخ ارز DMARK - بورس پنبه نیویورک	96
پزو مکزیک- بورس شیکاگو	90
مینی سویا-بورس شیکاگو (CBOT)	86
فرانک سوئیس / نرخ ارز DMARK - بورس پنبه نیویورک	72
قرارداد آتی شاخص سهام Eurotop 100 - بورس کالای ایالات	48
شاخص فلزات پیشرفته-بورس آتی نیویورک	44
گوسفند زنده-بورس کالا MidAmerica	40
وی‌کشت-بورس شیکاگو	28

[1] International Monetary Market (IMM)



شکل ۷- توزیع ناهنجاری‌ها در ۲۰ بازار برتر.
Figure 7- Distribution of anomalies in the top 20 markets.

نمودار میله‌ای نشان‌دهنده تعداد هفته‌هایی است که تمامی الگوریتم‌ها به‌طور هم‌زمان یک بازار مشخص را ناهنجار تشخیص داده‌اند. این شکل بازارهایی را که بالاترین سطح فعالیت غیرعادی را داشته‌اند برجسته می‌کند.

همان‌طور که در **جدول ۲** نشان داده شده است، بیشترین تعداد ناهنجاری‌ها در شاخص رشد *S&P 500 BARRA* بازار بین‌المللی پول، با ۲۳۲ نقطه پرت شناسایی شده مشاهده می‌شود. این موضوع نشان می‌دهد که بازارهای شاخص سهام، به دلیل حساسیت بالای خود نسبت به تغییرات مالی و اقتصادی جهانی، بیشتر مستعد رفتار غیرعادی معامله‌گران هستند. پس ازآن، بازارهای ارز مانند فرانک فرانسه */DMARK* و کرون سوئد */DMARK* در رتبه‌های بعدی قرار دارند که بازتاب‌دهنده نوسانات تاریخی نرخ‌های ارز است.

به‌طورکلی، نتایج نشان می‌دهند که بازارهای سهام و ارز فرکانس بالاتری از ناهنجاری‌ها را نسبت به بازارهای کالا دارند؛ هرچند این بازارهای کالایی نیز نوسان دارند، اما کمتر تحت‌تاثیر شوک‌های کلان اقتصادی قرار می‌گیرند.

۲-۶- مطالعه موردی نقره CBOT

برای مطالعه موردی دقیق‌تر، بازار نقره (۱۰۰۰ اونس تروی-نقره-بورس تجاری شیکاگو *CBOT*) انتخاب شد. تجزیه‌وتحلیل درصد موقعیت خالص غیرتجاری بین سال‌های ۱۹۹۳ تا ۱۹۹۶ (شکل ۲) چندین نقطه ناهنجاری بحرانی را آشکار می‌کند.

به‌طور خاص، ناهنجاری‌های شناسایی‌شده در سال ۱۹۹۳ و اواخر ۱۹۹۶ هم‌زمان با کاهش شدید یا تغییرات ناگهانی در موقعیت بازار بوده اند. این نقاط دوره‌هایی را برجسته می‌کنند که رفتار معامله‌گران به‌طور قابل‌توجهی از الگوهای عادی منحرف شده است، که اغلب در پاسخ به استرس اقتصادی یا تغییرات ناگهانی در عرضه و تقاضا است. چنین انحرافاتی می‌تواند به‌عنوان شاخص‌های هشداردهنده اولیه برای نوسانات در بازارهای فلزات گران‌بها عمل کند؛ نقره در این بازارها هم به‌عنوان کالای صنعتی و هم در برخی شرایط، به‌عنوان دارایی امن معامله می‌شود.

۳-۶- تحلیل مقایسه‌ای الگوریتم‌ها

تحلیل مقایسه‌ای الگوریتم های پیاده‌سازی‌شده، تفاوت‌های واضحی در عملکرد آن‌ها در شناسایی ناهنجاری ها نشان می دهد.

Z-Score و *IQR*، به‌عنوان روش‌های آماری ساده، در شناسایی ناهنجاری‌های قوی و آشکار، به‌ویژه تغییرات بزرگ و ناگهانی در موقعیت‌های معامله‌گران، موثرتر بودند. این روش‌ها ساده و سریع هستند، اما ممکن است ناهنجاری‌های نامحسوس را از دست بدهند.

Isolation Forest و *One-Class SVM*، به‌عنوان روش‌های یادگیری ماشین، توانایی بالاتری در شناسایی الگوهای غیرعادی پیچیده و کمتر آشکار دارند، به‌ویژه در بازارهایی با نوسانات غیرخطی یا نامتقارن. با این‌حال، این الگوریتم‌ها ممکن است نسبت به روش‌های آماری مجموعه متفاوت و گسترده‌تری از مشاهدات را به‌عنوان ناهنجار علامت‌گذاری کنند. ایجاد ستون تجمیعی *all_algorithms* بر پایه اشتراک خروجی هر چهار الگوریتم، رویکردی محافظه‌کارانه در شناسایی ناهنجاری‌ها اتخاذ می‌کند که با محدود کردن تشخیص به نقاط مورد توافق، از ثبت موارد مشکوک یا ناشی از حساسیت یک مدل منفرد جلوگیری کرده و فراوانی موارد علامت‌گذاری‌شده را به‌طور معناداری کاهش می‌دهد. این مقایسه، مزایای چارچوب چند الگوریتمی را برجسته می‌کند: در بازارهای مالی با نوسان بالا که در آن‌ها الگوهای رفتار غیرعادی همواره آشکار نیستند، ادغام روش‌های آماری و یادگیری ماشین پایداری و قابلیت اطمینان شناسایی ناهنجاری‌ها را به میزان چشمگیری افزایش می‌دهد.

جدول ۵- تعداد ناهنجاری‌های شناسایی‌شده توسط هر الگوریتم و نسبت آن‌ها به کل مجموعه داده

Table 5- Number of anomalies detected by each algorithm and their proportion relative to the entire dataset.

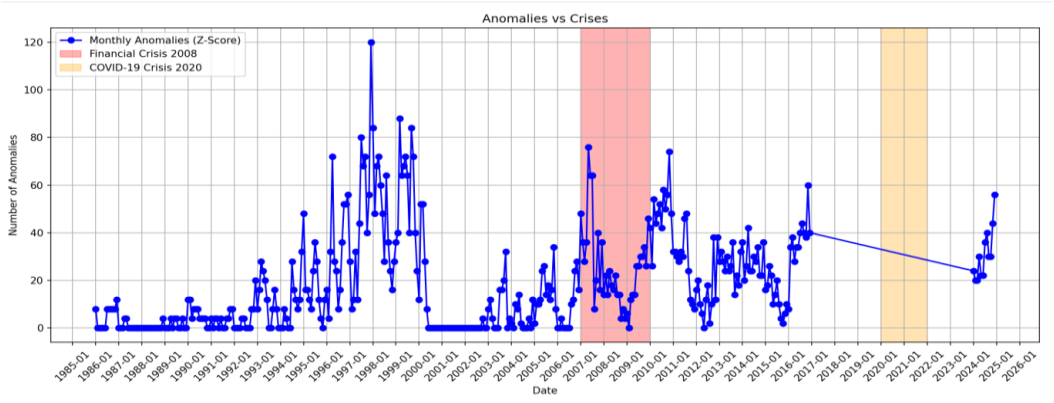
الگوریتم	تعداد ناهنجاری‌های شناسایی‌شده	درصد از کل رکوردها
Z-Score	7,198	1.66%
IQR	32.916	7.59%
Isolation Forest	4.286	099%
One-Class SVM	23.564	5.43%
All Algorithms	1.638	0.38%

همان‌طور که مشاهده می‌شود، روش *IQR* بیشترین تعداد ناهنجاری‌ها را شناسایی کرده است، درحالی‌که *Isolation Forest* موارد کمتری را تشخیص داده که نشان‌دهنده رویکرد محافظه‌کارانه‌تر آن است. ستون *All all_algorithms* نقاطی را نمایش می‌دهد که به‌طور هم‌زمان توسط هر چهار الگوریتم‌ها به‌عنوان ناهنجار شناسایی شده‌اند.

۴-۶- ناهنجاری‌ها و بحران‌های مالی

یکی از یافته‌های کلیدی این پژوهش، هم‌راستایی زمانی ناهنجاری‌های شناسایی‌شده با بحران‌های عمده مالی و اقتصادی جهانی است. تحلیل سری‌های زمانی نشان می‌دهد که در طول بحران مالی جهانی ۲۰۰۸ و همه‌گیری کووید-۱۹ در سال ۲۰۲۰، تعداد ناهنجاری‌ها افزایش محسوسی یافته است. به‌ویژه در بازارهای شاخص مانند *S&P 500* و بازارهای ارز، الگوریتم‌ها تغییرات غیرمعمول در الگوهای موقعیت‌گیری معامله‌گران را شناسایی و علامت‌گذاری کردند..

به همین ترتیب، در بازارهای کالایی مانند نقره و نفت خام، بخش قابل‌توجهی از ناهنجاری‌ها با برخی رویدادهای مهم مرتبط با عرضه و تقاضا و تصمیم‌های سیاست‌گذاری اقتصادی هم‌زمان بود. این هم‌راستایی نشان می‌دهد که ناهنجاری‌های شناسایی‌شده صرفا نوسانات تصادفی نیستند و بازتاب‌دهنده ریسک‌های سیستمیک و بحران‌های کلان اقتصادی هستند.



شکل ۸- هم‌راستایی زمانی ناهنجاری‌ها با بحران‌های مالی جهانی (۲۰۰۸ و ۲۰۲۰).

Figure 8- Temporal alignment of anomalies with the global financial crises (2008 and 2020).

نواحی سایه‌دار دوره‌های بحران را نشان می‌دهند، که در این بازه‌ها تعداد ناهنجاری‌ها به‌طور چشمگیری افزایش‌یافته است.

۶- ۵- نتایج هر الگوریتم

این بخش نتایج عددی و توصیفی هر روش تشخیص ناهنجاری را بر اساس مجموعه‌داده‌ای متشکل از ۴۳۳۵۴۴ رکورد گزارش می‌دهد. هر الگوریتم بر نوع متفاوتی از رفتار غیرعادی تأکید می‌کند و دیدگاه‌های مکملی را در مورد ناهنجاری‌های بازار نمایان می‌سازد.

روش آماری نمره استاندارد Z، ۷۱۹۸ ناهنجاری (حدود ۱٫۶۶ درصد از مجموعه‌داده) را شناسایی کرد. این ناهنجاری‌ها عمدتاً نشان‌دهنده نقاط پرت شدید و تغییرات ناگهانی هستند و این روش را به‌ویژه در تشخیص شوک‌های واضح بازار مؤثر می‌سازد. نتایج نشان داد که در بسیاری از بازارها، تنها تعداد محدودی از نقاط از آستانه $|z| > 3$ فراتر رفته‌اند که نشان می‌دهد اکثر بازارها رفتاری نسبتاً عادی داشته و ناهنجاری‌ها در اطراف بحران‌های بزرگ یا تغییرات سیاستی متمرکز شده‌اند.

روش دامنه بین‌چارکی بیشترین تعداد ناهنجاری را گزارش داده است، ۳۲۹۱۶ ناهنجاری (حدود ۷٫۵۹ درصد از مجموعه‌داده). مقادیر خارج از محدوده $[Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR]$ به‌عنوان ناهنجاری علامت‌گذاری شدند. در مقایسه با نمره Z، روش دامنه بین‌چارکی به دلیل اتکا به آماره‌های مقاوم و ناپارامتری، در بازارهای پرنوسان با توزیع‌های دارای دنباله سنگین، انحرافات بیشتری را علامت‌گذاری کرده است. نتایج نشان داد که در بازارهای آتی ارز و کالاهایی مانند نفت خام و گندم، تعداد بیشتری ناهنجاری شناسایی شده است که با ماهیت نوسانی این بازارها همخوانی دارد.

الگوریتم یادگیری جمعیِ جنگل ایزوله مبتنی بر درخت، با نرخ آلودگی ۰٫۰۱، ۴۲۸۶ ناهنجاری، معادل حدود ۰٫۹۹ درصد از مجموعه‌داده، را شناسایی کرد. برخلاف روش‌های آماری مبتنی بر آستانه، جنگل ایزوله با ایجاد تقسیم‌های تصادفی بر روی ویژگی‌ها و مقادیر آن‌ها، مشاهده‌های قابل‌تفکیک‌تر را شناسایی می‌کند و می‌تواند ساختارهای پیچیده و غیرخطی داده‌ها را نیز در نظر گیرد. تعداد ناهنجاری‌های شناسایی‌شده توسط این روش از روش‌های *IQR* و ماشین بردار پشتیبان تک‌کلاسه کمتر بود که با توجه به تنظیم نرخ آلودگی برابر با ۰٫۰۱ قابل‌انتظار است. بررسی توصیفی خروجی‌ها نشان داد که این روش در بازارهایی با پویایی‌های پیچیده و توزیع‌های نامتقارن، مجموعه‌ای متفاوت از مشاهدات را نسبت به روش‌های آماری علامت‌گذاری کرده است.

روش ماشین بردار پشتیبان تک‌کلاسه (*One-Class SVM*) با هسته *RBF*، ۲۳۵۶۴ ناهنجاری، معادل حدود ۵٫۴۴ درصد از مجموعه‌داده، را شناسایی کرد. این روش مجموعه‌ای متمایز از ناهنجاری‌ها را نسبت به سایر روش‌ها شناسایی نمود و به‌ویژه در بازارهایی با روندهای صعودی یا نزولی بلندمدت، مشاهدات متفاوتی را علامت‌گذاری کرد. همچنین، این روش برخی انحرافات موضعی از الگوی غالب داده‌ها را شناسایی و علامت‌گذاری کرد. این ویژگی می‌تواند قابلیت استفاده از روش مذکور را برای بررسی انحراف‌های موضعی در کنار الگوهای بلندمدت بازار افزایش دهد.

در مجموع، ۱۶۳۸ مشاهده، معادل حدود ٪۰٫۳۸ از مجموعه‌داده، در اشتراک منطقی خروجی‌های چهار روش (*all_algorithms*) قرار گرفتند؛ به این معنا که هر چهار الگوریتم آن‌ها را ناهنجار تشخیص داده‌اند. اگرچه اندازه این مجموعه نسبتا کوچک است، این مشاهده‌ها نمایانگر توافق بالای میان چهار الگوریتم هستند. ازاین‌رو، می‌توانند برای انتخاب مطالعات موردی و انجام بررسی‌های تکمیلی در اولویت قرار گیرند.

۶-۶- تفسیر مقایسه‌ای نتایج

مقایسه کمی و کیفی روش‌ها، تفاوت‌هایی را در الگوی شناسایی ناهنجاری‌ها نشان می‌دهد. روش‌های آماری، شامل نمره استاندارد Z و دامنه بین‌چارکی، از نظر محاسباتی سریع و از حیث تفسیر نسبتاً ساده هستند. روش دامنه بین‌چارکی با علامت‌گذاری ۳۲۹۱۶ مشاهده، بیشترین تعداد ناهنجاری را گزارش کرد. این تفاوت می‌تواند با اتکای این روش به چارک‌ها و ساختار توزیع داده‌ها، به‌ویژه در توزیع‌های چوله یا دارای دم‌های کلفت، مرتبط باشد. نمره استاندارد Z نیز ۷۱۹۸ مشاهده را شناسایی کرد که، با توجه به آستانه $|z| > 3$، عمدتا شامل انحراف‌های بزرگ‌تر از میانگین بوده‌اند.

روش‌های یادگیری ماشین، شامل جنگل ایزوله و ماشین بردار پشتیبان تک‌کلاسه، بر مبنای سازوکارهای متفاوت خود، مجموعه‌هایی متمایز از مشاهده‌ها را به‌عنوان ناهنجار علامت‌گذاری کردند. ماشین بردار پشتیبان تک‌کلاسه ۲۳۵۶۴ ناهنجاری را شناسایی کرد، درحالی‌که جنگل ایزوله ۴۲۸۶ مشاهده را علامت‌گذاری نمود. این تفاوت‌ها می‌تواند ناشی از تفاوت سازوکار الگوریتم‌ها از جمله تفکیک تصادفی مشاهده‌ها در جنگل ایزوله و برآورد مرز تصمیم در ماشین بردار پشتیبان تک‌کلاسه و نیز انتخاب ابرپارامترهایی مانند *contamination*، *nu* و *gamma* باشد. به‌ویژه، تعداد ناهنجاری‌های جنگل ایزوله با تنظیم *contamination=0.01* ارتباط مستقیم دارد.

در اشتراک خروجی هر چهار روش، ۱۶۳۸ مشاهده قرار داشتند؛ به این معنا که هر چهار الگوریتم آن‌ها را ناهنجار علامت‌گذاری کرده‌اند. این مشاهده‌ها بیانگر توافق بالای میان روش‌ها هستند و می‌توانند برای انتخاب موارد تحلیل موردی و انجام بررسی‌های تکمیلی در اولویت قرار گیرند. برای سنجش میزان توافق میان روش‌ها، هم‌پوشانی‌های زوجی و مرتبه‌های بالاتر محاسبه شد و می‌تواند در قالب جدول، نقشه حرارتی یا نمودار ون ارایه شود.

جدول ٦ میزان هم‌پوشانی زوجی میان روش‌های مختلف را گزارش می‌کند. بر اساس شاخص ژاکارد، بیشترین هم‌پوشانی میان نمره استاندارد *Z* و جنگل ایزوله، با مقداری حدود ۰/۵۹، مشاهده شد؛ این نتیجه نشان می‌دهد بخش قابل‌توجهی از مشاهده‌های شناسایی‌شده توسط جنگل ایزوله با خروجی نمره *Z* مشترک بوده‌اند. کمترین هم‌پوشانی نیز میان روش دامنه بین‌چارکی و ماشین بردار پشتیبان تک‌کلاسه مشاهده شد؛ شاخص ژاکارد این دو روش حدود ۰/۰۵ بود.

جدول ۶- همپوشانی‌های زوجی بین الگوریتم‌ها.
Table 6- Pairwise overlaps between algorithms.

الگوریتم ۱	الگوریتم ۲	اشتراک	اجتماع	شاخص ژاکارد	درصد از کل رکوردها
Z-Score	IQR	7,198	32,916	0.219	1.66%
Z-Score	Isolation Forest	4,286	7,198	0.595	0.99%
Z-Score	One-Class SVM	1,894	28,868	0.066	0.44%
IQR	Isolation Forest	4,286	32,916	0.130	0.99%
IQR	One-Class SVM	2,774	53,704	0.052	0.64%
Isolation Forest	One-Class SVM	1,638	26,212	0.062	0.38%

جدول ۶ میزان توافق زوجی میان روش‌ها را بر اساس شاخص ژاکارد نشان می‌دهد. بیشترین هم‌پوشانی میان نمره استاندارد Z و جنگل ایزوله، با شاخص ژاکارد ۰,۵۹۵حدود ۶۰ درصد مشاهده شد. این مقدار نشان می‌دهد بخش قابل‌توجهی از مشاهده‌های علامت‌گذاری‌شده توسط این دو روش مشترک بوده‌اند. کمترین هم‌پوشانی نیز میان دامنه بین‌چارکی (*IQR*) و ماشین بردار پشتیبان تک‌کلاسه، با شاخص ژاکارد ۰/۰۵۲ حدود ۵% به دست آمد. این یافته بیانگر تفاوت قابل‌ملاحظه در مجموعه مشاهده‌های شناسایی‌شده توسط این دو روش است که می‌تواند با سازوکارهای تشخیصی متفاوت و نیز تنظیم ابرپارامترهای آن‌ها مرتبط باشد.

جدول ۷- همپوشانی‌های سه‌گانه بین الگوریتم‌ها.
Table 7- Triple overlaps between algorithms.

ترکیب الگوریتم‌ها	تعداد نقاط مشترک	درصد از کل رکوردها
جنگل ایزوله و IQR و Z نمره	4,286	0.99%
تک کلاسه SVM و IQR و Z نمره	1,894	0.44%
تک کلاسه SVM و جنگل ایزوله و Z نمره	1,638	0.38%
تک کلاسه SVM و جنگل ایزوله و IQR	1,638	0.38%

جدول ۷ هم‌پوشانی مشاهده‌های ناهنجار در ترکیب‌های سه‌گانه الگوریتم‌ها را گزارش می‌کند. در این جدول، هم‌پوشانی هر ترکیب سه‌تایی شامل مشاهده‌هایی است که دست‌کم توسط همان سه الگوریتم علامت‌گذاری شده‌اند؛ بنابراین، ممکن است بخشی از این مشاهده‌ها توسط الگوریتم چهارم نیز شناسایی شده باشند.

بیشترین هم‌پوشانی سه‌گانه میان نمره استاندارد Z، دامنه بین‌چارکی و جنگل ایزوله مشاهده شد. این ترکیب شامل ۴۲۸۶ رکورد، معادل حدود ۰/۹۹% از کل مشاهدات، بود. با توجه به اینکه تعداد کل ناهنجاری‌های شناسایی‌شده توسط جنگل ایزوله نیز ۴۲۸۶ رکورد گزارش شده است، این نتیجه نشان می‌دهد همه مشاهده‌های علامت‌گذاری‌شده توسط جنگل ایزوله در خروجی دو روش نمره استاندارد Z و دامنه بین‌چارکی نیز قرار داشته‌اند.

کمترین هم‌پوشانی سه‌گانه میان دامنه بین‌چارکی، جنگل ایزوله و ماشین بردار پشتیبان تک‌کلاسه، با ۱۶۳۸ رکورد، معادل حدود ۰/۳۸% از کل مشاهدات، مشاهده شد. این مقدار با اشتراک خروجی هر چهار الگوریتم برابر است؛ از این‌رو، نشان می‌دهد تمامی مشاهده‌های مشترک این سه روش، توسط نمره استاندارد Z نیز شناسایی شده‌اند.

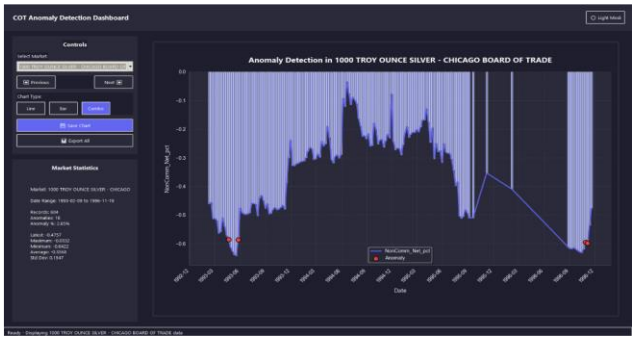
به‌طور کلی، نتایج **جدول ۷** نشان می‌دهد توافق میان روش‌ها به ترکیب الگوریتم‌های مورد مقایسه وابسته است. هم‌پوشانی بالای ترکیب نمره استاندارد Z، دامنه بین‌چارکی و جنگل ایزوله حاکی از آن است که جنگل ایزوله در این مجموعه‌داده عمدتا مشاهده‌هایی را علامت‌گذاری کرده است که دو روش آماری نیز آن‌ها را ناهنجار تشخیص داده‌اند. در مقابل، کاهش هم‌پوشانی در ترکیب‌های شامل ماشین بردار پشتیبان تک‌کلاسه می‌تواند بیانگر تفاوت در مجموعه مشاهده‌های شناسایی‌شده توسط این روش باشد؛ تفاوتی که ممکن است با سازوکار برآورد مرز تصمیم و تنظیم ابرپارامترهای آن مرتبط باشد.

باید توجه داشت که مقادیر درصدی جدول، سهم مشاهده‌های مشترک از کل مجموعه‌داده را نشان می‌دهند و پایین‌بودن آن‌ها به‌خودی‌خود غیرعادی نیست؛ زیرا ناهنجاری‌ها و به‌ویژه ناهنجاری‌های مورد توافق چند الگوریتم، معمولا بخش کوچکی از کل مشاهدات را تشکیل می‌دهند.

در انتهای این بخش، به منظور تکمیل تحلیل و ارایه یافته‌ها به صورت بصری، خروجی گرافیکی تولیدشده توسط برنامه پیاده‌سازی‌شده ارایه شده است. این خروجی شامل نمودارهای خطی، میله‌ای و ترکیبی است که هر یک هدف خاصی در نمایش جنبه‌های مختلف داده‌ها دارند. این شکل‌ها به عنوان مرجع بصری عمل کرده و درک روشن‌تری از الگوها و نقاط ناهنجاری شناسایی‌شده را در اختیار خواننده قرار می‌دهند.

الف-

ب-



ج-

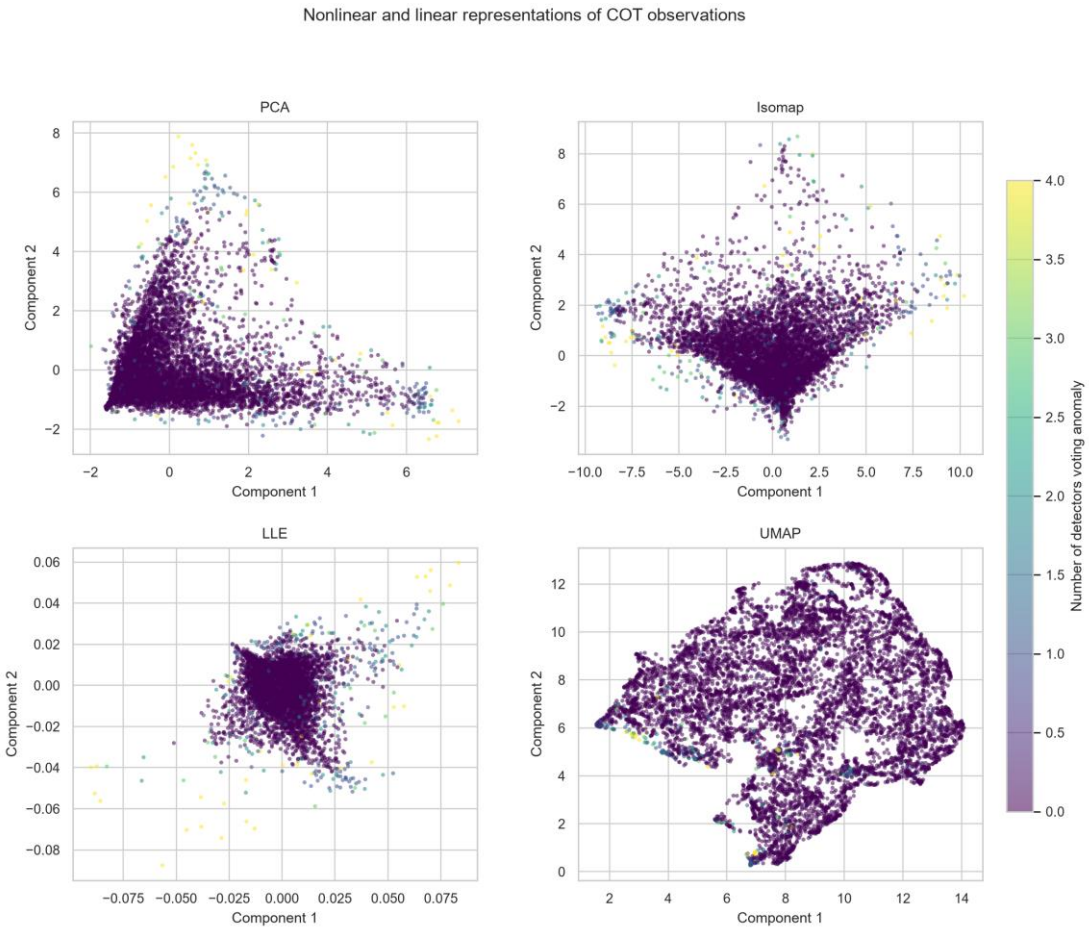
شکل ۹- خروجی‌های گرافیکی چارچوب پیشنهادی؛ الف- نمودار خطی تغییرات داده‌ها در طول زما، ب- نمودار میله‌ای مقایسه تعداد ناهنجاری‌ها، ج-نمودار ترکیبی برای نمایش هم‌زمان روندها و ناهنجاری‌ها.

Figure 9. Graphical outputs of the proposed framework; a. Line chart illustrating temporal data variations, b. Bar chart comparing the number of detected anomalies, c. Combined chart showing both trends and detected anomalies.

شکل ۹ خروجی‌های گرافیکی تولیدشده توسط برنامه، شامل نمودارهای خطی، میله‌ای و ترکیبی که روندها، مقایسه‌ها و نقاط ناهنجاری را نشان می‌دهند.

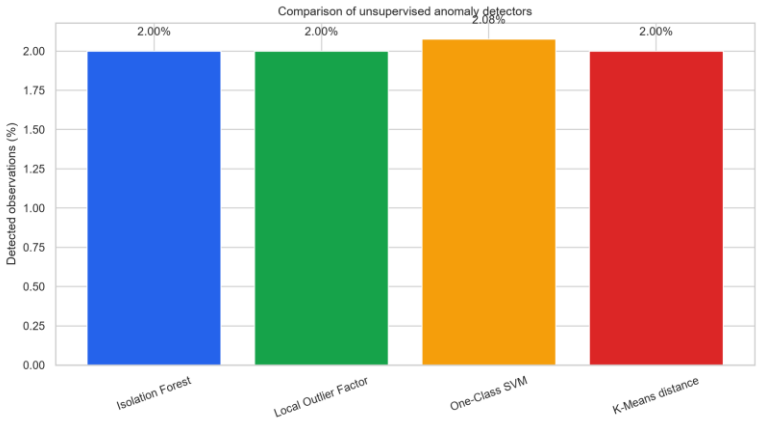
۸- کاهش بعد و روش‌های تشخیص ناهنجاری بدون نظارت

در ابتدا روش‌های کاهش بعد خطی و غیرخطی، *PCA*، ایزومپ، *Umap* و *LLE* را بر مجموعه داده‌های *COT* اعمال نمودیم [12] و سپس الگوریتم‌های ماشین بردار پشتیبان تک کلاسه، عامل پرت محلی، جنگل ایزوله و *k-means* را برای تشخیص ناهنجاری‌ها پیاده‌سازی نموده‌ایم. در انتها میزان هم‌پوشانی جاکارد روش‌ها به صورت جدول بیان شده است.



شکل ۱۰- بازنمایی‌های PCA، Isomap، LLE و UMAP بر اساس تعداد آرای روش ها رنگ‌آمیزی.

Figure 10 – PCA, Isomap, LLE, and UMAP representations colored according to the number of detector votes.



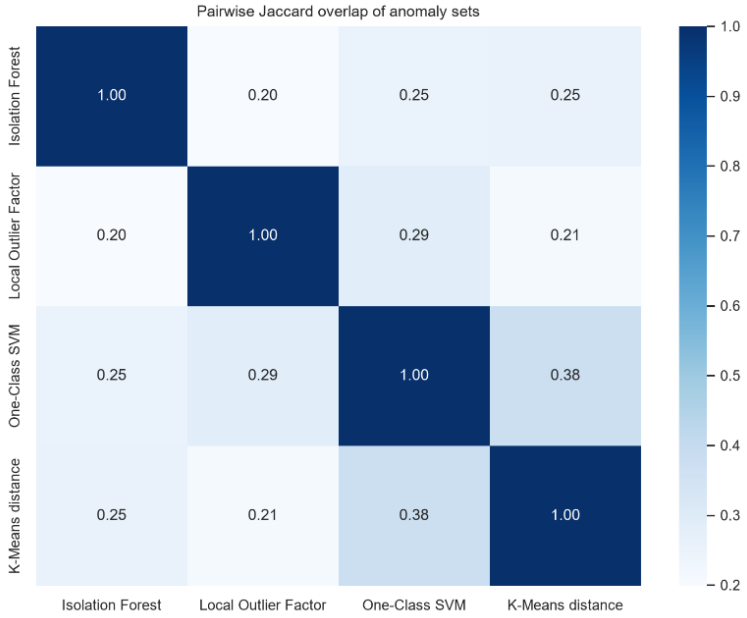
شکل ۱۱- نرخ‌های تشخیص در چهار روش بدون نظارت.

Figure 11- Detection rates of the four unsupervised methods.

جدول ۸- نتایج هر روش

Table 8 – Detector counts and consensus results.

روش	تعداد علامت‌گذاری‌شده	نرخ
جنگل ایزوله	160	2.00%
عامل پرت محلی	160	2.00%
ماشین بردار پشتیبان تک‌کلاسه	166	2.08%
فاصله کی‌میانگین	160	2.00%
اجماع (دست‌کم ۲ روش)	146	1.83%

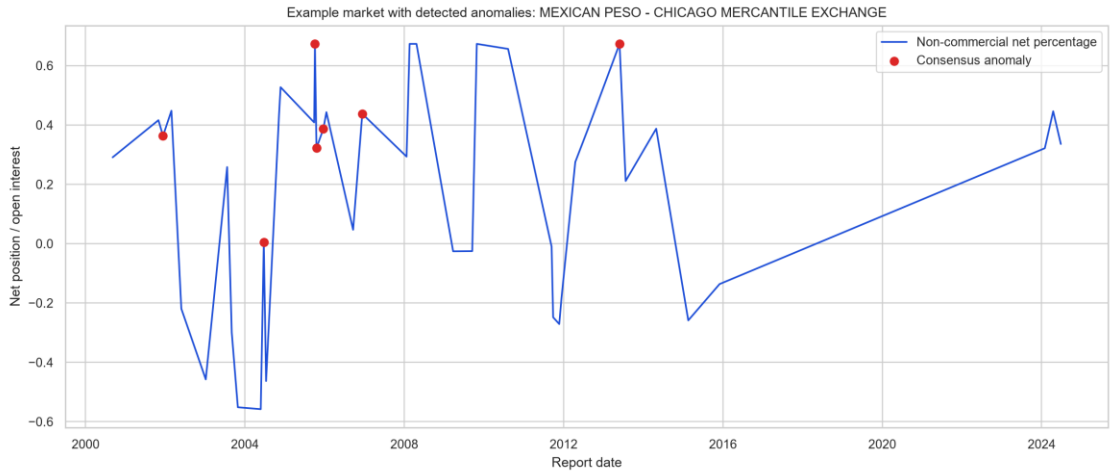


شکل ۱۲- هم‌پوشانی جاکاردِ دوبه‌دو میان مجموعه‌های ناهنجاری.

Figure 12- Pairwise jaccard overlap among the anomaly sets.

شکل ۱۳- بازارهایی با بیشترین تعداد ناهنجاری اجماعی در نمونه تحلیلی.

Figure 13- Markets with the highest number of consensus anomalies in the analytical sample.



شکل ۱۴- نمونه سری بازار همراه با نشانگرهای ناهنجاری اجماعی.

Figure 14- Sample market time series with consensus anomaly markers.

۸- تمرکز در سطح بازار و بررسی زمانی

ناهنجاری‌های اجماعی به‌صورت پراکنده در میان بازارها توزیع شده‌اند و در یک دسته‌بندی واحد و فراگیر از بازار متمرکز نیستند. بیشترین شمار در نمونهٔ محاسباتی مربوط به قراردادهای ارز، شاخص سهام و کالا است؛ از جمله ین ژاپن، دلار استرالیا، شاخص دلار آمریکا و برخی قراردادهای منتخبِ شاخص سهام. شکل ۵ بازارها را بر حسب تعداد ناهنجاری‌های اجماعی رتبه‌بندی می‌کند؛ با این‌حال، نرخ‌ها باید همراه با تعداد مشاهدات تفسیر شوند، زیرا بازارها از نظر دامنهٔ پوشش تاریخی یکسان نیستند.

نمای سری زمانیِ نمونه برای قرارداد پزو مکزیک - بورس کالای شیکاگو (*MEXICAN PESO-CHICAGO MERCANTILE EXCHANGE*) ارایه شده است. نشانگرهای قرمز رنگ، تاریخ‌هایی را مشخص می‌کنند که شرط اجماعِ دست‌کم دو آشکارساز را برآورده ساخته‌اند. این نقاط، گزینه‌هایی برای بررسی و تطبیق با تغییرات هم‌زمان در قیمت، حجم معاملات، متغیرهای اقتصاد کلان یا سازوکارهای گزارش‌دهی به‌شمار می‌روند؛ چرا که صرف داده‌های *COT* به‌تنهایی برای تبیین علت وقوع ناهنجاری کافی نیست.

۹- نتیجه‌گیری

این مطالعه نشان داد که ترکیب روش‌های آماری ساده با الگوریتم‌های پیشرفته یادگیری ماشین بر روی داده‌های *COT*، چارچوب مؤثری برای تشخیص این روش امکان تشخیص ناهنجاری (*Anomaly Detection*) را فراهم می‌کند. فراهم می‌کند. نتایج نشان می‌دهد که نمونه‌های نادر شناسایی‌شده با رویدادهای مالی مهم هم‌راستا هستند، مانند بحران جهانی ۲۰۰۸ و همه‌گیری *COVID-19* در سال ۲۰۲۰، که هم اعتبار علمی و هم اعتبار عملی این رویکرد را تقویت می‌کند. مهم‌ترین دستاورد این کار، توسعه یک ابزار گرافیکی تعاملی است که به تحلیلگران امکان می‌دهد تغییرات غیرمعمول در رفتار معامله‌گران را به‌سرعت شناسایی کنند. علاوه بر این، استفاده از ترکیب چند الگوریتمی، اطمینان را افزایش داده و خطاهای مثبت کاذب در تشخیص نمونه‌های نادر را کاهش می‌دهد. از دیدگاه عملی، چارچوب پیشنهادی می‌تواند به‌عنوان یک سیستم هشدار زودهنگام برای مؤسسات مالی و سیاست‌گذاران عمل کند و همچنین در طراحی سیستم‌های معاملاتی هوشمند و تطبیقی پشتیبانی ایجاد نماید. برای تحقیقات آینده، توصیه می‌شود که در پژوهش‌های آینده می‌توان متغیرهای بیشتری، از جمله قیمت‌ها، حجم معاملات و شاخص‌های کلان اقتصادی، را به چارچوب افزود و روش‌های مدل‌سازی و پیش‌بینی بلادرنگ را نیز توسعه داد تا چارچوب پیشنهادی به یک داشبورد هوشمند برای پایش و تحلیل بازار تبدیل شود.

تشکر و قدردانی

نویسندگان از تمامی افراد موثر در انجام این پژوهش قدردانی می‌نماید.

منابع مالی

این پژوهش بدون حمایت مالی انجام شده است.

تعارض منافع

هیچ‌گونه تعارض منافعی گزارش نشده است.

منابع

[1] Irwin, S. H., & Sanders, D. R. (2011). Index funds, financialization, and commodity futures markets. *Applied economic perspectives and policy*, *33*(1), 1–31. https://doi.org/10.1093/aepp/ppq032

[2] Brooks, C., & Wichmann, R. (2019). *Eviews guide to accompany introductory econometrics for finance*. SSRN. https://ssrn.com/abstract=3466908

[3] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE. https://doi.org/10.1109/ICDM.2008.17

[4] Sato, Y. (2025). Shortage, liquidity, and the effectiveness of monetary policy: Evidence from the us commodity futures markets. *Available at SSRN 5567658*. https://dx.doi.org/10.2139/ssrn.5567658

[5] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1-58. https://doi.org/10.1145/1541880.1541882

[6] Lensen, A., Zhang, M., & Xue, B. (2020). Multi-objective genetic programming for manifold learning: Balancing quality and dimensionality. *Genetic programming and evolvable machines*, *21*(3), 399–431. https://doi.org/10.1007/s10710-020-09375-4

[7] Kritzman, M., & Li, Y. (2010). Skulls, financial turbulence, and risk management. *Financial analysts journal*, *66*(5), 30–41. https://doi.org/10.2469/faj.v66.n5.3

[8] Phatak, A. A., Wieland, F. G., Vempala, K., Volkmar, F., & Memmert, D. (2021). Artificial intelligence based body sensor network framework—narrative review: proposing an end-to-end framework using wearable sensors, real-time location systems and artificial intelligence/machine learning algorithms for data collection, data mining and knowledge discovery in sports and healthcare. *Sports Medicine-Open*, *7*(1), 79. https://doi.org/10.1186/s40798-021-00372-0

[9] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM transactions on knowledge discovery from data (TKDD)*, *6*(1), 1-39. https://doi.org/10.1145/2133360.2133363

[10] Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine learning*, *54*(1), 45-66. https://doi.org/10.1023/B:MACH.0000008084.60811.49

[11] Aulerich, N. M., Irwin, S. H., & Garcia, P. (2014). Bubbles, food prices, and speculation: Evidence from the CFTC's daily large trader data files. In *The economics of food price volatility* (pp. 211-253). University of Chicago Press. https://www.nber.org/system/files/chapters/c12814/c12814.pd

[12] Ma, Y., & Fu, Y. (2012). *Manifold learning theory and applications* (Vol. 434). Boca Raton: CRC press. https://api.pageplace.de/preview/DT0400.9781439871102_A23986384/preview-9781439871102_A23986384.pdf