الگوریتم نزدیکترین همسایگی به طور گسترده برای تحقیقات مرتبط با یادگیری ماشین استفاده شده است. طبقه بندی یک موضوع مهم در پردازش کلان داده، علم داده و یادگیری ماشین است. الگوریتم نزدیکترین همسایگی یکی از قدیمی‌ترین، ساده‌ترین و همچنین دقیق‌ترین الگوریتم‌ها برای طبقه‌بندی الگوها و مدل‌های رگرسیون است. بنابر این در چند دهه اخیر مورد مطالعه قرار گرفته است و به طور گسترده در زمینه های مختلف استفاده می شود. در بسیاری از مسائل طبقه‌بندی الگو مانند تشخیص الگو، طبقه‌بندی متن، مدل‌های رتبه‌بندی، تشخیص شی و برنامه‌های کاربردی تشخیص رویداد است [1]

یک الگوریتم ناپارامتریک است. ناپارامتریک به معنای عدم وجود پارامتر بدون توجه به اندازه داده است. در عوض، پارامترها با اندازه مجموعه داده های آموزشی تعیین می شوند. بنابراین این الگوریتم می‌تواند بهترین انتخاب برای هر مطالعه طبقه‌بندی‌ای باشد که دانش قبلی از توزیع داده‌ها کم میباشد.همچنین این الگوریتم یکی از تنبل ترین روش های یادگیری است [2].

عملکرد الگوریتم نزدیکترین همسایگی را با استفاده از ۱۱ معیار فاصله بررسی شده . این تحقیق به بررسی انواع مختلف فاصله‌ها از جمله فاصله اقلیدسی[[1]](#footnote-1)، فاصله ماهالانوبیس[[2]](#footnote-2)، فاصله منهتن [[3]](#footnote-3)، فاصله مینکوفسکی[[4]](#footnote-4)، فاصله چبیشف[[5]](#footnote-5)، فاصله کسینوسی [[6]](#footnote-6)، فاصله همبستگی [[7]](#footnote-7)، فاصله همینگ[[8]](#footnote-8) ، فاصله جاکارد [[9]](#footnote-9)، فاصله اقلیدسی استاندارد [[10]](#footnote-10)و فاصله اسپیرمن [[11]](#footnote-11) پرداخته است. این تحقیق بر روی هشت مجموعه داده مصنوعی باینری با توزیع‌های مختلف تولید شده‌اند، اعمال شده است. داده‌ها قبل از آزمایش نرمال‌سازی شدند.هر مجموعه داده به ۷۰٪ برای آموزش و ۳۰٪ برای تست تقسیم شد. نتایج نشان می‌دهد که فاصله‌های MD، مینکوفسکی، چبیشف، ED، فاصله ماهالانوبیس و فاصله اقلیدسی استاندارد از سایر فاصله‌های آزمایش شده بهتر عمل کرده‌اند.. برای ارزیابی عملکرد KNN، معیارهای دقت، حساسیت و ویژگی برای هر فاصله محاسبه شد. نتایج گزارش شده نشان می‌دهد که استفاده از MD با دقت ۹۷.۸٪، حساسیت ۹۶.۷۶٪ و ویژگی ۹۸.۳۵٪ از سایر فاصله‌های آزمایش شده بهتر عمل کرده است [3].

الگوریتم تصمیم گیری نزدیکترین همسایه به یک نقطه نمونه طبقه بندی نشده طبقه بندی نزدیکترین مجموعه از نقاط طبقه بندی شده قبلی را اختصاص می دهد، بنابراین می توان گفت که نیمی از اطلاعات طبقه بندی در یک مجموعه نمونه بی نهایت در نزدیکترین همسایه موجود است [4]

یافتن کا نزدیک‌ترین همسایه برای یک نمونه آزمون در میان تعداد مشخصی از نمونه طراحی فرآیندی زمان‌بر است، به ویژه وقتی تعداد نمونه بزرگ باشد. یکی از روش ها پیشنهاد شده آن است که مجموعه طراحی با حفظ بیشتر نمونه‌هایی که اطلاعات تمایزدهنده بین کلاس‌ها را ارائه می‌دهند، فشرده‌سازی شود و الگوریتم نزدیک‌ترین همسایه‌ی فشرده [[12]](#footnote-12)را معرفی کرد [5]. در این راستا، مقاله حاضر روش شاخه و حد[[13]](#footnote-13) را برای تسریع محاسبه کا نزدیک‌ترین همسایه‌ها معرفی می‌کند. این روش با تجزیه سلسله مراتبی نمونه‌های طراحی به زیرمجموعه‌های جداگانه و اعمال الگوریتم شاخه و حد بر روی این گروه‌ها، محاسبات فاصله را به طور قابل توجهی کاهش می‌دهد [6].

الگوریتم همسایگی نزدیک نیاز به محاسبه فاصله نمونه طبقه بندی نشده با تمامی نمونه های طبقه بنده شده در مجموعه دیتای های آموزش دیده شده دارد.برای اولین بار یک الگوریتم نزدیکترین همسایه وزن‌دار با فاصله را پیشنهاد شد که با توجه به فواصل بین آن های وزن‌های سنگین‌تر را به همسایگان نزدیک‌تر با اختصاص می‌یافت [7].

 با این حال، اجرای سنتی این روش از نظر محاسباتی پر هزینه است. دو تکنیک موثر، یعنی متراکم سازی و پیش پردازش را مطرح شد تا به طور قابل‌توجهی سرعت الگوریتم همسایگی نزدیک را افزایش دهیم و در عین حال سطح دقت را حفظ کنیم. به خصوص در مواردی که ابعاد فضای ویژگی بالا است ادغام این دو تکنیک با الگوریتم همسایگی نزدیک کمک به سرعت هفت برابری بدون کاهش دقت میکند [8].

الگوریتم جدیدی مبتنی بر الگوریتم همسایگی نزدیک برای طبقه‌بندی پیشنهاد می‌کنند که هدف آن غلبه بر مشکلاتی از جمله کارایی پایین که به علت ماهیت تنبل یادگیری الگوریتم همسایگی نزدیک میباشد و از جمله دیگر محدودیت ها میتوان به محدودیت‌هایی کار با حجم دیتای زیاد و همچنین وابستگی به انتخاب مقدار مناسب برای کا اشاره کرد .

این مشکلات است. روش پیشنهادی، یک مدل kNN برای داده‌ها می‌سازد که جایگزین داده‌ها شده و به عنوان مبنای طبقه‌بندی عمل می‌کند. مقدار k به صورت خودکار تعیین می‌شود، برای داده‌های مختلف متغیر است و از نظر دقت طبقه‌بندی بهینه است. ساخت مدل، وابستگی به k را کاهش داده و طبقه‌بندی را سریع‌تر می‌کند

یکی از الگوریتم‌های پرکاربرد در حوزه طبقه‌بندی، الگوریتم نزدیک‌ترین همسایه است که به دلیل سادگی و کاربرد وسیع در حوزه‌های مختلف شناخته شده است. با این حال، الگوریتم نزدیک‌ترین همسایه با چالشی مهم در زمینه هزینه‌های محاسباتی در هر مورد طبقه‌بندی مواجه است. تحقیقات متعددی به منظور بهبود عملکرد این الگوریتم صورت گرفته‌اند. به عنوان مثال، در مقاله‌ای توسط لگوریتمی معرفی شده است که با بهره‌گیری از خوشه‌بندی مبتنی بر آنتروپی به بهینه‌سازی الگوریتم نزدیک‌ترین همسایه می‌پردازد. برخلاف روش‌های خوشه‌بندی مبتنی بر فاصله سنتی، این الگوریتم داده‌ها را به خوشه‌های جداگانه تقسیم کرده و سعی در کمینه‌سازی آنتروپی در میان همسایگان دارد. این روش با استخراج نقاط نماینده از خوشه‌ها به منظور محاسبه نزدیک‌ترین همسایگان عمل می‌کند. نتایج تجربی بر روی مجموعه داده‌های مختلف UCI[[14]](#footnote-14) نشان می‌دهد که این رویکرد می‌تواند به کاهش قابل توجه هزینه‌های محاسباتی منجر شود.

علاوه بر این، تحلیل آماری آزمون تی[[15]](#footnote-15) از نتایج دقت نشان می‌دهد که عملکرد الگوریتم نزدیک‌ترین همسایه با استفاده از نمایندگان مبتنی بر آنتروپی به طور مشابهی با الگوریتم نزدیک‌ترین همسایه اصلی می‌باشد. این تحقیقات نشان می‌دهد که روش‌های نوین مبتنی بر آنتروپی می‌توانند بهبودهای قابل توجهی در عملکرد الگوریتم‌های طبقه‌بندی سنتی ایجاد کنند [9].

یک روش نوین برای خودکارسازی فرآیند حضور و غیاب استفاده از الگوریتم نزدیک‌ترین همسایه میباشد در این روش با بردارهای ویژگی‌های چهره‌های شناسایی‌شده با آن‌هایی که در یک پایگاه داده از پیش‌پُر شده ذخیره شده‌اند، مقایسه می‌شوند. انعطاف‌پذیری و سادگی الگوریتم نزدیکترین همسایه خود را برای این کار ایده‌آل می‌کند، زیرا چهره ورودی را با یافتن نزدیک‌ترین همسایه خود طبقه‌بندی می‌شوند. دقت سیستم با انتخاب مقدار مناسب کا از طریق ارزیابی تجربی بهبود می‌یابد. این امر آن را به یک راه‌حل قابل اعتماد و قوی برای مدیریت حضور و غیاب در مؤسسات آموزشی، محل‌های کار و سایر محیط‌ها تبدیل می‌کند. پیاده‌سازی این سیستم می‌تواند به بهینه‌سازی فرآیندهای حضور و غیاب، افزایش کارایی و بهبود دقت رکوردها منجر شود و در نتیجه به بهبود بهره‌وری کلی سازمان کمک کند [10].

دسترسی پویا به طیف های فرکانس رادیویی از طریق فناوری رادیو شناسایی می‌تواند به حل مشکل کمبود طیف های فرکانسی کمک کند . در این راستا عملکرد الگوریتم نزدیک‌ترین همسایه از نظر دقت، حساسیت، ویژگی، ضریب F1، ماتریس سردرگمی و عامل کا مورد تحلیل قرار گرفته است. انتخاب مقدار بهینه کا از نتایج به دست آمده، باعث می‌شود الگوریتم نزدیک‌ترین همسایه بتواند به صورت کارآمد و مؤثر وظیفه خود را انجام دهد [11].

نادیده گرفتن داده‌های مفقود می‌تواند کارایی مطالعه را کاهش داده و گاهی اوقات داده ها را تحت تاثیر قرار میدهد. داده‌های نامتعادل نیز تمایل به تحت تاثیر قرار گرفتن توسط کلاس‌های اکثریت دارند و کلاس‌های اقلیت را نادیده می‌گیرند . داده‌های مفقود را با محاسبه مقادیر آماری با استفاده از الگوریتم نزدیک‌ترین همسایه‌ها تکمیل کرده [12].

معاملات سهام یکی از فعالیت‌های مهم در دنیای مالی است. پیش‌بینی بازار سهام به معنای تلاش برای پیش‌بینی ارزش آینده یک سهام یا ابزار مالی دیگر است که در یک بورس مالی معامله می‌شود. با استفاده از داده‌های موجود بازار سهام مدل همسایگی نزدیک آموزش می‌بیند، هوش کسب می‌کند و سپس از اطلاعات یادگرفته شده برای انجام پیش‌بینی دقیق استفاده می‌کند. این مطالعه از قیمت‌ها با فرکانس‌های روزانه و دقیقه‌ای و یک روش یادگیری ماشین به نام کا نزدیکترین همسایه برای پیش‌بینی قیمت سهام برای سرمایه‌گذاری‌های بزرگ و کوچک در سه بازار مختلف بهره می‌برد [13].

فناوری بلاکچین اولین بار در قالب ارزهای دیجیتال به کار گرفته شد و این فناوری مجموعه‌ای از نوآوری‌ها و مسیرهای جدید را در پژوهش‌های علمی فراهم کرده است، از جمله استفاده از داده‌ها برای شناسایی ناهنجاری‌ها یا پیش‌بینی قیمت در بیت‌کوین و اتریوم.

فناوری بلاکچین تکنیک‌های مختلفی را برای خودکارسازی فرآیندهای تجاری ارائه می‌دهد. یکی از پژوهش‌های مهم در این زمینه، مدلی مبتنی بر الگوریتم K-Nearest Neighbor (KNN) است که برای شناسایی تراکنش‌های غیرقانونی در ارزهای دیجیتال پیشنهاد شده است. این مدل از مجموعه داده‌ی Elliptic و الگوریتم KNN برای شناسایی تراکنش‌های غیرقانونی استفاده می‌کند.داده‌ها را به سه کلاس: غیرقانونی، قانونی و ناشناخته دسته‌بندی می‌کند. هر نود دارای 166 ویژگی است که 94 ویژگی اول اطلاعات محلی درباره تراکنش را نشان می‌دهد و 72 ویژگی باقی‌مانده به عنوان ویژگی‌های تجمیعی شناخته می‌شوند. دقت مدل با مقادیر k=2 و k=4 بیش از 90%، یادآوری با k=3 به 56% و دقت پیش‌بینی با k=4 به 78% می‌رسد [14].

یکی دیگر از روش های نظارت شده از ترکیب روش خوشه بندی پویا و الگوریتم همسایگی نزدیکک به وجود آمده (DC-knn) . در ابتدا از فاز خوشه‌بندی بدون نظارت برای کشف اطلاعات جدید از داده‌های اولیه استفاده می‌شود که می‌تواند به بهبود دقت طبقه‌بندی نظارت‌شده کمک کند. این الگوریتم با استفاده از یک تابع هدف جدید، شباهت‌های درون خوشه‌ای و بین خوشه‌ای را در نظر می‌گیرد و به هر ویژگی وزن‌های تطبیقی اختصاص می‌دهد. این رویکرد باعث می‌شود که متغیرهای مهم برای خوشه‌بندی شناسایی شده و متغیرهای نویزآور حذف شوند، در نتیجه دقت طبقه‌بندی به طور قابل‌توجهی افزایش یابد [15].

1. Mehta, A., et al., *Machine Learning‐Based Fault Diagnosis of Self‐Aligning Bearings for Rotating Machinery Using Infrared Thermography.* Mathematical Problems in Engineering, 2021. **2021**(1): p. 9947300.

2. Tamrakar, P. and S.S. Ibrahim. *Lazy learning associative classification with WkNN and DWkNN algorithm*. in *ITM Web of Conferences*. 2021. EDP Sciences.

3. Abu Alfeilat, H.A., et al., *Effects of distance measure choice on k-nearest neighbor classifier performance: a review.* Big data, 2019. **7**(4): p. 221-248.

4. Cover, T. and P. Hart, *Nearest neighbor pattern classification.* IEEE transactions on information theory, 1967. **13**(1): p. 21-27.

5. Hart, P., *The condensed nearest neighbor rule (corresp.).* IEEE transactions on information theory, 1968. **14**(3): p. 515-516.

6. Fukunaga, K. and P.M. Narendra, *A branch and bound algorithm for computing k-nearest neighbors.* IEEE transactions on computers, 1975. **100**(7): p. 750-753.

7. Dudani, S.A., *The distance-weighted k-nearest-neighbor rule.* IEEE Transactions on Systems, Man, and Cybernetics, 1976(4): p. 325-327.

8. Wu, Y., K. Ianakiev, and V. Govindaraju, *Improved k-nearest neighbor classification.* Pattern recognition, 2002. **35**(10): p. 2311-2318.

9. AbdelAzim, H., M. Tharwat, and A. Mohammed. *Efficient Computational Cost Reduction in KNN through Maximum Entropy Clustering*. in *2024 6th International Conference on Computing and Informatics (ICCI)*. 2024. IEEE.

10. Dalwadi, D.C. and U. Jha. *Smart Attendance System Using KNN Algorithm*. in *2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC)*. 2023. IEEE.

11. Somula, L.R. and M. Meena. *K-nearest neighbour (KNN) algorithm based cooperative spectrum sensing in cognitive radio networks*. in *2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*. 2022. IEEE.

12. Cheng, H. *KNN-SVM Classifiers in Complex Diagnosis*. in *Journal of Physics: Conference Series*. 2024. IOP Publishing.

13. Patil, A., G. Padole, and A. Sontakke, *Stock Trend Prediction Using KNN Algorithm.* International Journal of Scientific Research in Science and Technology, 2023. **5**(10): p. 693-699.

14. Elbaghdadi, A., S. Mezroui, and A. El Oualkadi, *K-nearest neighbors algorithm (knn): an approach to detect illicit transaction in the bitcoin network*, in *Integration Challenges for Analytics, Business Intelligence, and Data Mining*. 2021, IGI Global. p. 161-178.

15. Sabri, M., et al., *A Novel Classification Algorithm Based on the Synergy Between Dynamic Clustering with Adaptive Distances and K-Nearest Neighbors.* Journal of Classification, 2024: p. 1-25.

1. **Euclidean Distance** [↑](#footnote-ref-1)
2. Mahalanobis Distance [↑](#footnote-ref-2)
3. **Manhattan Distance** [↑](#footnote-ref-3)
4. Minkowski Distance [↑](#footnote-ref-4)
5. Chebychev Distance [↑](#footnote-ref-5)
6. Cosine Distance [↑](#footnote-ref-6)
7. **Correlation Distance** [↑](#footnote-ref-7)
8. Hamming Distance [↑](#footnote-ref-8)
9. Jaccard Distance [↑](#footnote-ref-9)
10. Standardized Euclidean Distance [↑](#footnote-ref-10)
11. Spearman Distance [↑](#footnote-ref-11)
12. condensed nearest neighbor (CNN) [↑](#footnote-ref-12)
13. branch and bound [↑](#footnote-ref-13)
14. https://archive.ics.uci.edu/ [↑](#footnote-ref-14)
15. T-Test [↑](#footnote-ref-15)