ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Data Mining Best Practices and Efficiency in the Large-Scale Data Mining Using Artificial Intelligence and Generative AI

Chandra Sekhar Veluru

Tracy, United States

ABSTRACT

In the awakening era of the technological landscape, the emergence of key technologies such as generative AI and artificial intelligence (AI) has been instrumental in exercising an effective data mining process. The proliferated demand for best data mining practices continues to express how critical it is for growth in many sectors across different landscapes of human engagements and businesses.

Adoption of effective and reliable data mining practices has been an anticipated wave, and features such as machine learning have fostered growth and development. In the literature selection, criteria such as the publication date, studies' relevance to the research topic, and the type of sources, for example, books or peer-reviewed journals, were utilized.

Methodologies such as case studies, statistical analyses, and experiments applied by authors in the literature studies were analyzed as they provided insights on the approaches employed by the authors.

However, the emergence of AI and generative AI has come with a batch of risks that ought to be mitigated immediately. Risks such as breaching data privacy and loss of data protection practices, both on a small and large scale, send a signal of the unimaginable losses and challenges that the economy and other sectors are yet to witness in their pursuit of data mining efficiency. This study aims to conduct a comprehensive analysis of the data field to determine the best data mining practices that can be adopted to hone data efficiency in the diversified and wide data ecosystem. Through an in-depth evaluation of the techniques and methodologies significant to the data mining environment, we seek to establish a framework of insights that will be instrumental to organizations in optimizing data mining practices to attain a reliable operational outcome.

***Corresponding author**

Chandra Sekhar Veluru, Tracy, United States.

Received: April 09, 2024; **Accepted:** April 15, 2024; **Published:** April 23, 2024

Keywords: Data Mining, AI, Generative AI, Data Scalability, Machine Learning, Deep Learning, Data Security, Data Privacy

Introduction

Background

Data mining is the act of using diversified techniques and approaches to exhibit data insights, correlations, and patterns from both structured and unstructured data in a wide pool [1-4]. The use of AI has simplified the insight derivation from a complex data structure by enhancing automation of the most critical data columns from the wide data volume. Previously, cases of data double entry and inaccuracy have surfaced in the data environment, which have immensely contributed to inaccurate insights and wrong decisions [5]. Conversely, the discovery and use of AI technologies such as natural language processing, machine learning, and deep learning have fostered data transformation by guaranteeing data scalability, increased processing speed, and a high degree of accuracy in data mining.

The data mining landscape has recently experienced various phases of significant evolution following constant technological discoveries. The application of machine learning knowledge to the data analysis tools has unleashed an interesting pathway for uncovering data patterns and correlations, influencing quality decision-making, and creating reliable predictive models. The use of methodologies such as predictive models and techniques, including data regression, data set clustering, and classification and association algorithms, has fostered easy mining of data even for complex tasks [6]. The automation feature has made it easy to access data easily, regroup them, and use the comparative rule to embody data sets with similar characteristics for easy analyses and derivation of patterns and correlations.

Objectives

- Understanding of the preexisting data mining practices and how they have been employed to enhance data mining.
- Determine common and extreme hurdles related to handling large volumes of structured and non- structured data sets.
- Explore the effectiveness of the existing methods for leveraging the demands of data mining.
- Highlight how the application of AI and generative AI changes the paradigm of data, influences decision- making, and contributes to more solidified insights.
- Determine future recommendations that can be instrumental

in countering any existing problems and enhancing data mining efficiency.

Rationale

The increasing volume of data in the fast-growing data-driven economy, ranging from sectors such as banking and securities, tourism, healthcare, communication and media, insurance, transport, and education, sets the groundwork for the need for a study of best data mining practices [5]. Drawing from these sector demands, the ability to exercise data mining proficiency has been flagged as one of the competitive advantages among companies in different sectors. However, the development of sophisticated data designs due to emerging volumes of structured and non-structured datasets challenges the efficiency of relying on odd data analysis methods that are not compatible with the new technology algorithms and software. The study seeks to leverage effective practices that would enhance the dissemination of accurate information, foster strategic decisions, and ensure data protection compliance.

Methodology

Search Strategy

In the detailed and diversified environment data, various strategies have been applied strategically to collect enough information. Key words such as "generative AI," "data mining," "data privacy," "data security," "large volumes of data mining," "artificial intelligence," "data domains," and "data mining practices" were key in this research for the great results. Several search engines such as Google, Bing, and databases such as Google Scholar, Scopus, JSTOR, PubMed, Academic Search Premier, and the ACM Digital Library were well utilized, making it easy to compile this research work to the last bit.

Existing materials in these databases provided insights that have continued to trigger the need to create effective and dependable data mining practices that would be flexible enough to incorporate the fastest-growing technological dimension.

Studies' Inclusion and the Exclusion Criteria

On inclusion criteria, the review focused on studies that were published between 2020 and 2024, which is the most updated database. In addition, the focus was on data literature that was written in English, which made it possible to understand and interpret easily.

Studies that focused on data mining analysis areas such as data mining practices as competitive edges, efficiency of best data practices, the impeccable role of AI in the data mining landscape, data privacy, data mining efficiency enhancement, data security, and data compliance were prioritized [7]. The publications used were either peer-reviewed or scholarly publications to accord research credibility and eliminate conspiracies and assumptions about the topic. Conversely, on the exclusion criteria, studies with less information because they were conducted on a small scale and publications not published in English were not considered. In addition, publications that did not directly converge on the research topic or address the concerns were excluded.

Data Extraction Techniques from Literature Materials

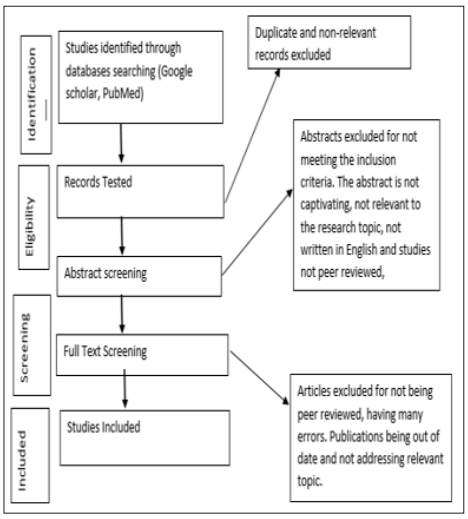
- **Source Identification:** Credible academic research databases such as Google Scholar and JSTROL were considered as they contain large volumes of academic materials such as books, peer-reviewed publications, and journals.
- The search queries-questions that are relevant to the data mining practices and keywords-were used to optimize the search for accurate and reliable results.
- **Literature Search:** The already-written topical key words were fed into the search engines, such as Google, and bravely searched for the relevant content. Features such as publication date and the type of documents needed were used to optimize the relevance of the literature.
- **Analysis of the Introduction and Conclusion Parts:** The introduction and conclusion parts were analyzed and provided solid background insights on the methods used in research, objectives, and rationale.
- **Understanding of Research Scope:** Intensive reading of the research abstract was conducted to determine the scope of the study and the objectives that the researcher discussed.
- **Exploration of Methodology Section:** The study explored the methodology section to understand the approaches and techniques used by the researcher in the collection of information on data mining efficiency. Case studies, interviews, and experimentation methodologies were analyzed to collect enough insights on data mining efficiency.
- **Analysis of Research Key Findings and Methodological Results:** The findings sections provided the insights derived by the author on data mining practices and how scalability and data complexity challenges were addressed using AI and generative AI. Statistical and empirical evidence were also analyzed to validate the literature insights.

Literature Quality Assessment

The quality of the literature studies included in the research is very critical to ensuring the information obtained is reliable and credible for the research topic. The quality of the literature was assessed through various criteria to ensure the accuracy and completeness of the extraneous data [8]. Literature works anchored on clear and objective-oriented methodologies were considered of high value as they presented a clear analysis of the topic with reliable information. In addition, the studies that addressed concerns such as data privacy, security, data mining effective practices, and the intensive role played by artificial intelligence in the data dimension were relevant to this review. In addition, the review prioritized the works published in the last five years to ensure the recent patterns and trends of technological advancements are captured. Materials with a high reputation, such as books and peer- reviewed journals, as evident by the number of citations and their publication category, such as peer reviews, are highly reliable.

Results

Selection Criteria Flowchart



Key Characteristics of the Included Literature Studies

The literature studies included in this review display various characteristics that have made them stand out as reliable. The methodologies employed by the authors, such as theoretical evaluations, empirical frameworks, case studies, and statistical analyses, enable understanding of complex data structures and data mining practices [5]. The included studies analyzed different technological domains such as finance, transport, and healthcare that provide wide coverage, presenting data diversity for easy comparative analysis. In addition, the included studies explored the value gained with the adoption of AI tools such as machine learning and generative AI in fostering data mining effectiveness [9]. With varied publication dates, the studies covered a wide range of data mining and technological trends across different time ranges. Moreover, the studies have addressed various components through their research objectives. The research objectives embraced by these studies made it possible to discuss areas such as AI integration, data mining, data privacy and protection, data scalability, and methodologies and techniques optimization.

Findings Synthesize for Efficient Data Mining Practices

Cleaning the data after collection is essential, and it prevents making wrong decisions. Data collected from various sources is prone to inaccuracy; therefore, the data processing stage should ensure the employment of all technical techniques, such as data normalization and continuous cleaning, to remove possible outliers. The selection of the data mining AI algorithms is critical, as the algorithms influence how effective the data mining process will be [10]. The choice of algorithms from the pool of odd statistical techniques to the most advanced methods such as deep learning and machine learning is determined by the state of the data in terms of complexity, the expected output metrics, and the nature of the data.

The efficiency of the data mining process is also determined by the optimization parameters that directly affect the accuracy of the predictive models.

After data extraction, engineering of the data in the data mining process is significantly essential as it facilitates the data transformation from variable format, which is unfiltered, to desirable data that can facilitate modeling and improve performance. In data engineering, it is recommended to maximize the use of techniques such as feature importance ranking and dimensionality reduction to increase efficiency.

Researchers can select the desired features by optimizing the abilities of machine learning algorithms and applying exploratory data analysis and data domain knowledge. Data cleaning is linked to data engineering to ensure the data is accurate and meets the required standards. Employing assessment measures such as detection of data outliers and anomalies ensures data integrity is preserved. In addition, while handling data on immense scales continues to pose a challenge, the use of GPUs, cloud computing, Spark, and data architectures would be pivotal in fostering parallelization and scalability when working with complex datasets.

Discussion

Key Findings Summary

The review has established that practices such as data engineering, normalization, data cleaning, and validation are key to assembling accurate data that is free of outliers and errors. Employment of AI, generative AI, machine learning tools, and data analysis knowledge plays a vital role in aligning data appropriately to improve efficiency during data mining [11]. Automation of data mining processes has eliminated the complexity of the datasets, even as scalability is considered a challenge to data mining efficiency to date [9]. In addition, the review has found that the combination of various evaluation methods, such as bootstrapping, enhances reliability in data models and data generalization. Balancing both data parallelization and computation facilitates the management of large datasets and eases data processing. To improve the performance of the data models and enhance efficiency in data computing, data miners should prioritize the selection of features that align with the process. Moreover, ethical considerations should be followed to protect organizations from data breaches and streamline the mining process.

Comparison of Review with Other Literature, Its Limitations and Implications

Compared to the previous studies, this review reinforces the importance of artificial intelligence, data optimization, the selection of data features, ethical considerations, and data quality during the data mining process. The comprehensive analysis of AI and generative AI roles in data mining processes supports the findings of the previous research works as it emphasizes the valuable role played by the integration of AI technologies in data mining [5]. This review contributes to quality decision-making by the organizations through the insights discussed herein.

The study also ascertains that adherence to the use of quality and accurate data, ethical data considerations, and the management of AI data privacy risks could facilitate organizations in attaining security, efficiency in their operations, and a competitive edge in the data-driven economy [12]. However, this study is characterized by limitations, such as technological advancement, in which some techniques and methodologies discussed here would not be relevant. The generalization of this study's findings is limited by variations from different sectors, which have unique dataset issues and use different technological infrastructure systems. In addition, the study only takes into consideration technological aspects, leaving aside the possible effects of social and economic parameters.

Future Recommendations

The study recommends the use of comprehensive data mining frameworks to address bias and ethical considerations during the integration of AI technologies. In addition, the study recommends the exploration of more algorithms that can be employed to address the scalability problem in both structured and unstructured complex data sets during the data mining process. Data regulatory measures should be taken into consideration during the data mining process to avoid data breaches and safeguard privacy. Moreover, the technology landscape should be further explored to understand the effects of developing AI technologies in the data mining environment.

Conclusion

Although the concerns of data privacy and protection remain a challenge due to the incorporation of AI, several benefits of AI have been witnessed in the data mining environment. The employment of artificial intelligence and generative AI has brought about automation features that have eased the analysis of complex and large data sets. This study has established that the data mining process is vital in every organization as the technology converts sectors to data-driven economies, creating a demand for reliable and accurate data. The use of machine learning and deep learning

has enhanced the effectiveness of the data mining process, making it easy to derive patterns, trends, and decision- making from data. To foster best data practices, the quality of the data used from different literatures should be assessed. Any data outliers should not be tolerated, as they may result in incorrect decision-making.

References

1. Shukla S (2023) Creative Computing and Harnessing the Power of Generative Artificial Intelligence. Journal Environmental Sciences and Technology 2: 556-579.
2. Alwahedi F, Aldhaheri A, Ferrag MA, Battah A, Tihanyi N (2024) Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. Internet of Things and Cyber Physical Systems 4: 167-185.
3. Golda A, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, et al. (2024) Privacy and Security Concerns in Generative AI: A Comprehensive Survey. IEEE Access 12.
4. Durugkar SR, Raja R, Nagwanshi KK, Kumar S (2022) Introduction to data mining. Data Mining and Machine Learning Applications 1-19.
5. Gupta MK, Chandra P (2020) A comprehensive survey of data mining. International Journal of Information Technology 12: 1243-1257.
6. Shu X, Ye Y (2023) Knowledge Discovery: Methods from data mining and machine learning. Social Science Research 110: 102817.
7. Mengist W, Soromessa T, Legese G (2020) Method for conducting systematic literature review and meta-analysis for environmental science research. Methods X 7: 100777.
8. Namoun A, Alshanqiti A (2020) Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences 11: 237.
9. Bandi A, Adapa PVSR, Kuchi YEVPK (2023) The power of generative AI: A review of requirements, models, input–output formats, evaluation metrics, and challenges. Future Internet 15: 260.
10. Anantrasirichai N, Bull D (2022) Artificial intelligence in the creative industries: a review. Artificial intelligence review 55: 589-656.
11. Frey CB, Osborne M (2023) Generative AI and the future of work: a reappraisal. Brown Journal of World Affairs 30.
12. Dai D, Boroomand S (2022) A review of artificial intelligence to enhance the security of big data systems: state-of-art, methodologies, applications, and challenges. Archives of Computational Methods in Engineering 29: 1291-1309.
13. Curzon J, Kosa TA, Akalu R, El Khatib K (2021) Privacy and artificial intelligence. IEEE Transactions on Artificial Intelligence 2: 96-108.

Copyright: ©2024 Chandra Sekhar Veluru. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.