International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Review on Data Mining Issues, Solution & Techniques

Nandani Sharma[1], Dr. Rajesh Bogey[2], Prof. Ritu Prasad[3]

[1]Research Scholar, Computer Science & Engineering, TIT(Excellence), Bhopal (M.P.)
[2]HOD of CSE Department, TIT(Excellence), Bhopal (M.P.)
[3]Assistant Prof. of CSE, TIT(Excellence), Bhopal (M.P.)

Abstract
Data mining, the process of discovering patterns and knowledge from large datasets, has become a cornerstone of decision-making across various domains including business, healthcare, and finance. This paper reviews the current landscape of data mining applications, exploring the diverse techniques employed and the challenges faced. Key issues include data quality, privacy concerns, and the scalability of mining algorithms in the face of increasingly large and complex datasets. We discuss solutions to these challenges, such as advancements in data cleaning methodologies, privacy-preserving techniques, and the development of scalable algorithms. Additionally, we provide a comprehensive overview of current data mining techniques, including clustering, classification, and association rule mining, and evaluate their effectiveness in addressing real-world problems. The paper aims to provide understanding of both the potential and limitations of data mining, offering insights into future research directions and technological advancements needed to overcome existing hurdles.

Keywords: Data Mining, Challenges, Issues, Solution, Classification, Clustering.

1. Introduction

Data mining is the development of pull-out valuable information from bulky collections of data. It is moreover recognized as information disclosure in databases (KDD)The strategy of information revelation holds activity such as information cleaning, information integration, information choice, information change, information mining, design assessment, and the introduction of information. Data mining is definite as technique of mining information form gigantic sets of data also well-defined as mining knowledge from data. Or in further word Data mining is the progression of removing and discovering pattern in enormous data sets. Investigating data patterns in bulky data writing one or more software. Data mining associations statistics and artificial intelligence to examine big data sets to determine valuable information. Every day, the Internet and many data storage strategies have a substantial impression on business, society, scientific, engineering, medicine and virtually every piece of everyday life. The endless development within the amount of data open could be a result of the mechanization of our society and the fast advancement of persuasive information collection and capacity skill. Data mining turns infinite amounts of data into valued information. For example, Google grips hundreds of millions of searches every day, and each request signifies a transaction as a user status their information needs. Data mining is fetching more prevalent due to the concept of "big data". Data mining is the procedure of mining valuable information from the huge amount of database at any time and at any place. But laterally with the data

mining, it is essential to save and protect the data from annoying hands and from outbreak i.e., to save the data from initiate stolen and from the intrusions. Data mining, frequently familiar as Knowledge Discovery in Databases (KDD). It is a procedure of discovery patterns and other valued material from huge data sets. Various individuals treat information mining as an included for another generally utilized term, information disclosure from information, or KDD, whereas more understanding information mining as just a fundamental step within the handle of information disclosure. The knowledge discovery method as an iterative categorization of the following steps:

- Data cleaning**:** To eliminate Noise and Inconsistent data.
- Data integration**:** Where numerous data sources may be collective.
- Data selection: Where applicable to the exploration task are recovered from the database.
- Data transformation**:** Where records are transformed or collective into suitable for mining by execution immediate or aggregations operations.
- Data mining**:** A vital manner where intellectual methods are practical in order to abstract data patterns.
- Pattern evaluation**:** To classify the truly motivating patterns signifying knowledge based on some interestingness measure.
- Knowledge representation**:** Where beginning and knowledge demonstration techniques are used to existing the mined knowledge to the employer.
-

Figure 1: Data Mining

2. DM Architecture

Data mining is dubbed as a technique of determine or departure existing facts from enormous aggregates of data dropped in several data foundation for instance file systems, databases, data warehouses etc. This fact supports lots of advantages to business approaches, technical, homoeopathic research, managements and distinct. The architecture comprises modules for protected safe-thread communication, database connectivity, systematized data management and competent data analysis for assembly global mining model.

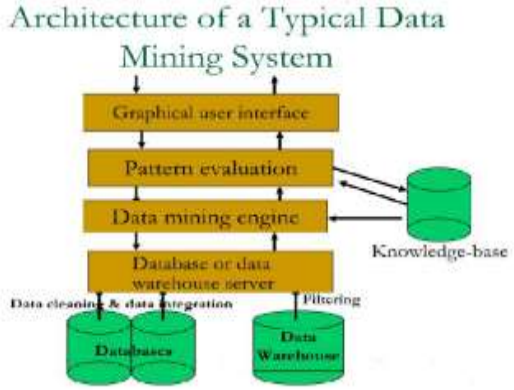


Figure 2: Data Mining Architecture

- **Database, data warehouse, or other information repository:** This component is single or a recognized of databases, data warehouses, spread sheets, or further varieties of material fountains. Data cleaning and data integration procedures may be proficient on the data.[10]

- **Database or data warehouse server**: The essential is in authority for appealing the related data, established on the data mining claim of the user.
- **Knowledge base:** This is the area facts that is second hand to attendant the exploration or approximation the interestingness of subsequent patterns. It embraces concept hierarchies that are used to consolidate attributes or attribute values into diverse levels of abstraction.
- **Data mining engine:** This is an imperative component of the data mining arrangement and superlatively embraces of a set of well-designed sectors for responsibilities for instance characterization, association analysis, classification, evolution and deviation analysis.[10]
- **Pattern evaluation module**: This fragment classically services interestingness production and cooperates through the data mining components thus as to exertion the exploration on stimulating patterns. It can admittance interestingness thresholds deposited in the knowledge base. On the other hand, the pattern assessment component may be involved by the pulling out section, interim on the execution of the data mining technique recycled. Well-organized data mining is likely by assertive the appraisal of pattern interestingness honestly hooked on the removal evolution to confer the exploration to individual the exciting patterns.
- **Graphical user interface:** This component interrelates among handler and the data mining system and authorities the handler to collaborate with the system by demanding a data mining request or assignment, provided that facts to support concentration the exploration, and triumph examining data mining constructed on the transitional data mining grades. This fundamental likewise approves to the handler to cruise database and data warehouse graphics or data structures, estimate extracted patterns, and envision the patterns in dissimilar procedures.

3. Life cycle of Data Mining/Process of Data Mining

Data mining activity is a step-by-step technique that cannot be completed in a single step. In other words, you cannot get the compulsory information from the bulky volumes of data as simple as that. It is not explicit to any industry. Basically, the progression has developed from the knowledge discover processes used extensively in industry. The major intention of data Mining process is to make hefty data projects to run more professionally. The progression including data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge demonstration are to be terminated in the given order.

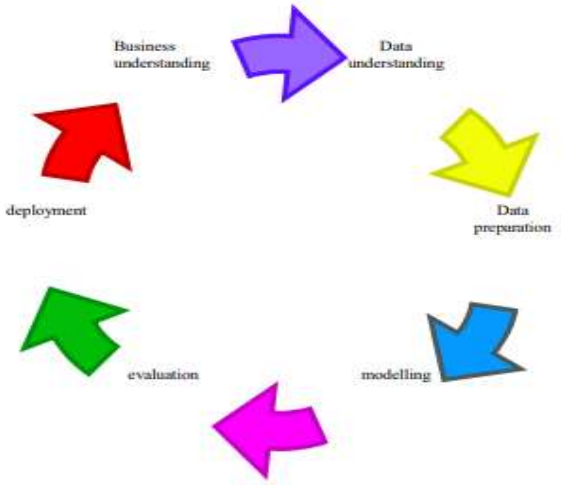


Figure 3: Process of Data mining

3. Working of DM

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.[9]

- **Classes:** Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters:** Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns:** Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.[9]

4. Data Mining Issues and its solution

Data Mining is shows essential role in business and many other areas. It is modern trustworthy technology, still some challenges must be determined. Some of the challenges are highlighted as following:

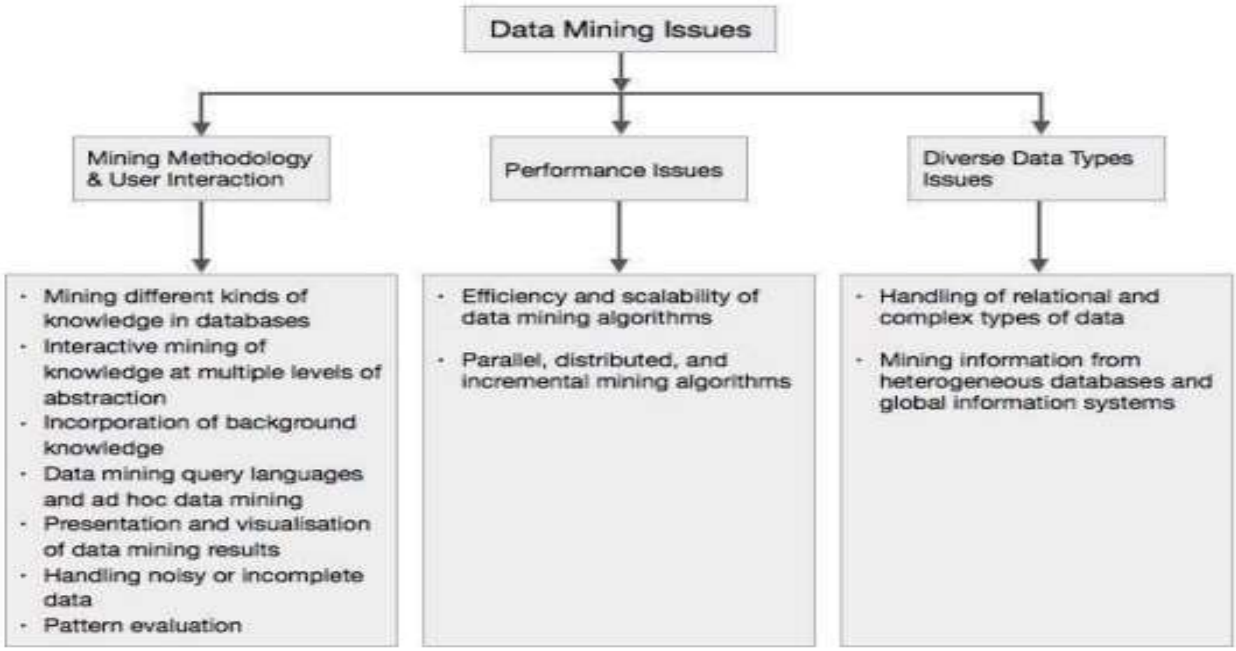


Figure 4: Data Mining Issues

1. **User-friendliness*:*** Eventually a system must hide technological difficulty from the user. To facilitate this, new software, tools, and substructure development is wanted in the extents of grid-maintained workflow organization, resource identification, distribution, and arranging, and user interfaces.
2. **User interface issues:** The knowledge determined by data mining tools is useful providing it is interesting, and above all understandable by the user. Good data visualization facilities the clarification of data mining results, in addition to assistances users better understand their requirements. Many data examining analysis tasks are expressively simplified by the capability to see data in a suitable visual performance. There are many visualization opinions and offers for effective data graphical appearance.

3. **Performance issues**: Many artificial intelligence and statistical methods occur for data exploration and explanation. Be that as it may, these approaches were frequently not arranged for the exceptionally colossal information sets information mining is apportioning with nowadays. Terabyte sizes are mutual. This increases the issues of scalability and productivity of the data mining approaches when treating significantly huge data. Calculations with exponential and indeed medium-order polynomial trouble cannot be of real-world utilize for information mining. Straight calculations are regularly the standard. In same point, examining can be utilized for mining as a substitute of the complete dataset. Though, disquiets for instance completeness and choice of samples may rise. Other topics in the issue of presentation are incremental updating, and similarity programming.
4. **Security, Privacy, and Data Integrity** - Numerous researchers reflected privacy defense in data mining as a significant topic. That's, how to ensure the users' protection whereas their information is being mined. Associated to this point is information mining for assurance of security and protection. One charged state that if we don't resolve the protection issue, information mining will get to be a hostile term to the common open. A few charged reflect the trouble of information judgment appraisal to be critical.

5. Challenges & Solution

1) Heterogeneous Data – Information can be of low quality, tainted, and inadequate. That's why, separated from the complexity of gathering information from diverse information data warehouse, heterogeneous information sorts (HDT) are one of the major information mining challenges. Typically, generally since enormous information comes from distinctive sources, may be consequently collected or manual, and can be subject to different handlers.

Solution: There are two perspectives to an arrangement for this issue. One, we take the conventional approach and prepare each HDT independently as per the classical homogeneous information mining prepare and after that fasten the comes about together. On the other hand, we combine the HDT during the pre-processing organize and after that conduct the data mining handle, treating them as a single substance. This way, after we analyze sentiment through emotion mining, it will chief to more accurate results.

2) Scattered Data - One of the most conspicuous data mining challenges is gathering data from stages through several computing atmospheres. Putting away plenteous amounts of information on a single server isn't adequate, which is why information is kept on local servers. Scattered information may moreover cruel that information is kept in disparate sources for occasion a CRM device or a neighborhood record on an individual computer. This condition frequently offerings itself when a foundation may ought to investigate information from various sources such as HubSpot, a .csv record, and a Prophet database. Businesses are moreover looking at more non-traditional ways to bridge the crevices that their interior from external sources.

Solution: We got to create dispersed assortments of data mining set of rules so that we do not get to transport all of the information to a single centralized store as we are doing presently. We too need the correct conventions and dialects to outline this conveyed information. For presently, this may be finished to very a degree with the help of metadata. One can utilization XML records to store metadata in an exhibit so that heterogeneous databases can be extricated. Analytical mark-up language (PMML) can support with the interchange of models among the dissimilar data storage sites and thus help interoperability, which in turn can support circulated data mining.

3) Data Ethics - Data mining challenges include the question of morals in data gathering to quite a degree. This is dissimilar from data privacy. For occasion, there may not be exact authorization from the special source of the information from where it is accumulated, indeed in case it is on an open stage like a social media channel or an open comment on an internet client audit gathering.

Solution*:* This can be a mastery issue, more than anything else, and one of the conspicuous information mining challenges in a moral AI environment. Much like web site illuminates the client to acknowledge or dismiss treats, or requires authorization to run pop-ups, a trade too must inform the client of what they may utilize their information for. Typically, a duty that businesses have to be address for more straightforwardness with their client.

4) Data Privacy - Information protection may be a genuine issue that emerges in information gathering, only when it comes to social media tuning in and investigation. Social media organizations are beneath the consideration indeed more which in the long run driven to the previous recording for insolvency, and the last mentioned paying a $5 billion fine to the U.S. organizations for information security disturbances. Since of this steady examination, numerous social media stages containing Facebook, Snapchat, and Instagram have choked their information protection frameworks. And this has recognized to posture information mining challenges for social opinion assessment.

Solution*:* This again reductions in the purview of the principles of ideas in data mining. Social media platforms as mentioned above, and even others like Twitter or Amazon Reviews, need to be transparent about their data privacy policies. Another significant way to discourse this issue is to legalize third-party apps that can entrance data through with moreover direct entrance to a user's digital device or incidentally through one of the user's social relations. And thirdly, data scientists need to follow correct protocol when requesting admittance to social media apps and platforms, such as Douyin, which have very inflexible data protection rules and are dissimilar to entrance for the determinations of data mining. At no point should administrations use back channels to access such constrained information.

6. Data Mining Techniques

There are numerous foremost data mining techniques established. The data mining tasks can be classified into two types. The two categories are descriptive tasks and predictive tasks.[8]

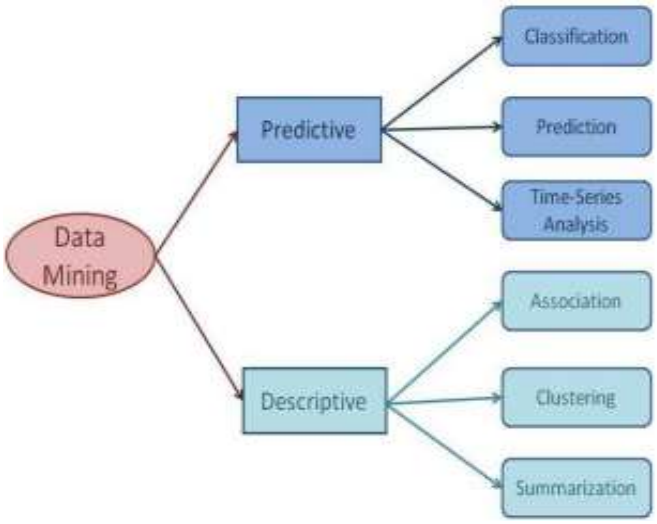


Figure 5: Data Mining Techniques

6.1 Descriptive DM - The descriptive evaluation is recycled to mined data and deliver the newest information on previous or current events. In addition, its experts to observe the data in a meticulous way

so that it would be able to answer simply about "what has occurred?" and "what is happening?". The descriptive data mining use unsupervised learning.

1) Clustering - This is a technique in DM that includes grouping similar data points composed into clusters or groups, the goal to identify patterns and matches in the data without previous knowledge of the construction of the data or the classification of the data points. Clustering can be used in an extensive range of applications, as well as **marketing segmentation, image processing and anomaly detection**, there are several clustering algorithms available but the most common once include- [5]

- Grid-based
- Hierarchy Clustering
- Density-based Clustering

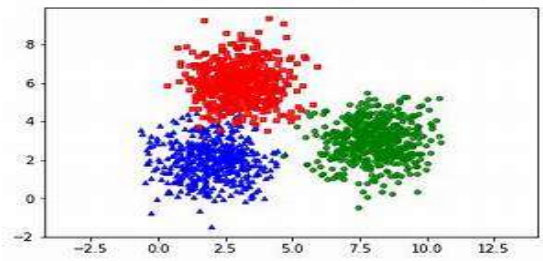
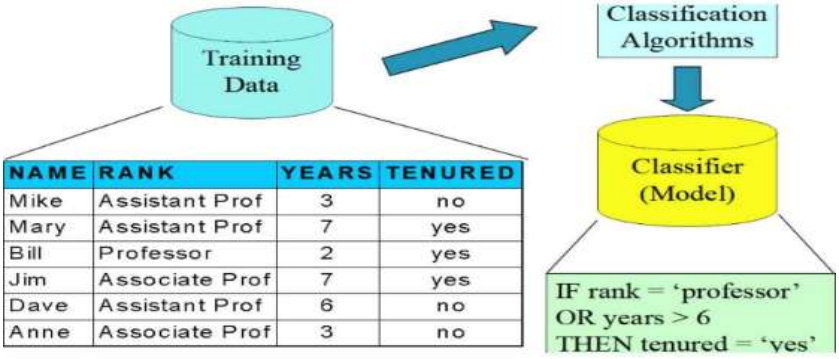


Figure 6: Representation of Clustering

For ex. A vender can use clustering to group consumers based on their buying behavior and demographic info to create targeted marketing operations.

6.2 Predictive DM - On the other hand, the predictive analysis delivers response of the upcoming questions that move across using ancient data as the chief principle for conclusions. The predictive data mining is used for provided that information about "what might happen?" and "why it might happen?". The predictive data mining use supervised learning.

1) Classification - This is often one of the foremost broadly utilized procedures in DM & ML, which includes the recognizable proof of designs in information and the labeling of information into predefined classes or categories. Classification is the strategy of allotting a given information point to a gather or course based on a set of highlights or traits. Classification calculations are utilized to build prescient models that can be utilized to recognize new data based on their highlights, these calculations utilize preparing data to memorize patterns and affiliation between the highlights and classes, and after that apply the learned designs to classify modern information. This method is as a rule utilized in extortion location, client division, spam sifting, hazard appraisals and estimation examination. For occasion, a bank can utilize classification to recognize beguiling exchange developed on a set of predefined traits, for occurrence exchange sum, area and time.



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Figure 7: Representation of Classification

Types of classification models- [5]

- Neural-Networks
- Support-Vector Machine (SVM)
- Bayesian Classification

7. Conclusion

Data mining remains a critical tool for harnessing the power of large and complex datasets to derive actionable insights. Despite its significance, the field faces several key challenges that can impact the effectiveness and reliability of mining operations. These challenges include issues with data quality, scalability, privacy, and interpretability. To address data quality issues, effective data preprocessing and cleaning techniques are essential to ensure that the data used is accurate and consistent. Scalability concerns are mitigated through the development of advanced algorithms and distributed computing methods that can handle large volumes of data efficiently. Privacy and security remain paramount, necessitating the implementation of privacy-preserving techniques such as anonymization and differential privacy to protect sensitive information while still enabling meaningful analysis. Meanwhile, the complexity of modern data mining models often leads to difficulties in interpretability, but advances in explainable AI offer promising solutions for making models more transparent and understandable. Furthermore, balancing model complexity to prevent overfitting and underfitting is crucial for developing robust predictive models. Techniques such as cross-validation, regularization, and ensemble methods contribute to creating models that generalize well to new data. In summary, while data mining presents significant opportunities for innovation and insight, overcoming its challenges requires a multifaceted approach involving improved techniques, technologies, and methodologies. By addressing these issues, the field can continue to advance and provide valuable contributions to various domains, from business and science to technology and beyond. As research progresses, ongoing improvements and innovations will be essential for maximizing the potential of data mining in an increasingly data-driven world.

References

1. Anshu, "Review Paper on Data Mining Techniques and Applications", International Journal of Innovative Research in Computer Science & Technology (IJIRCST), ISSN: 2347-5552, Volume-7, Issue-2, March 2019, DOI: 10.21276/ijircst.2019.7.2.4
2. Han, J. & Kamber, M. (2012). "Data Mining: Concepts and Techniques". 3rd.ed. Boston: Morgan Kaufmann Publishers.
3. Anup Arvind Lahoti, P. L. Ramteke, "Data Mining Technique its Needs and Using Applications", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pg. 572-579, ISSN 2320–088X
4. Dileep Kumar Singh, Vishnu Swaroop, "Data Security and Privacy in Data Mining: Research Issues & Preparation", International Journal of Computer Trends and Technology- volume4Issue2- 2013, ISSN: 2231-2803
5. C. Christy, M. A. Maria Parimala, M. Prema, "The Review on Data Mining Techniques and its Applications", International Journal of Data Mining Techniques and Applications, Volume: 07, Issue: 01, June 2018, Page No.50-54, ISSN: 2278-2419

6. Agu, Edward Onyebueke, Omankwu, Obinnaya Chinecherem and Ngene, Chigozie Chidimma, "Data Mining, Issues and Application", International Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 12, Issue 1, (Series-III) January 2022, pp. 14-17
7. Dr. Malla Reddy Jogannagari, Mrs. Maheshwari Manchala, "Data Mining: Techniques, Tools and its Challenges", International Journal of Creative Research Thoughts (IJCRT), Volume 8, Issue 7 July 2020, ISSN: 2320-2882
8. S.A.R. NIHA, "STUDY OF DATA MINING METHODS AND ITS APPLICATIONS", International Research Journal of Engineering and Technology (IRJET), ISSN: 2395-0056, Volume: 04 Issue: 11, Nov -2017
9. Anu Verma, Jyoti Arora, "A Review on Data Mining, Its Applications and Approaches", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, Volume 5 Issue 6, June 2016
10. B.N. Lakshmi, G.H. Raghunandhan, "A Conceptual Overview of Data Mining", Proceedings of the National Conference on Innovations in Emerging Technology-2011, pp.27-32.