International Journal of Technological Exploration and Learning (IJTEL)
ISSN: 2319-2135 | www.ijtel.org
Vol.1, Issue 3, November-December 2012

Data Mining Techniques for Knowledge Discovery and Predictive Analysis in Large-Scale Database

Rajiv Harish Patel Solanki

Department of Computer Application and Science, Swami Vivekananda Institute of Modern Science, Kolkata, Under MAKAUT & AICTE, West Bengal, India

ABSTRACT: The rapid growth of digital information has increased the importance of data mining techniques for extracting meaningful knowledge and supporting predictive analysis from large-scale databases. Organizations across various sectors generate enormous volumes of structured and unstructured data through business transactions, customer interactions, financial activities, healthcare systems, and digital platforms. However, the availability of large quantities of data does not automatically generate useful information. Advanced analytical techniques are required to identify hidden patterns, relationships, trends, and predictive insights. This study examines the role of data mining techniques in knowledge discovery and predictive analysis, emphasizing classification, clustering, association rule mining, regression, decision trees, neural networks, and statistical learning. It explores the processes involved in data collection, preprocessing, transformation, algorithm selection, model development, and evaluation. Particular attention is given to the challenges of scalability, data quality, computational complexity, model accuracy, and information security in large-scale database environments. The study also discusses how predictive models support organizational decision-making, operational efficiency, customer relationship management, and strategic planning. By examining established research and methodological approaches, this paper highlights the importance of integrating data mining algorithms with effective database management systems to transform extensive datasets into reliable, actionable, and valuable knowledge.

KEYWORDS: Data Mining, Knowledge Discovery, Predictive Analysis, Large-Scale Database, Classification, Clustering, Association Rule Mining, Machine Learning, Database Management, Decision Support Systems.

I. INTRODUCTION

The development of information technology has significantly transformed the way organizations collect, store, process, and utilize information. Modern organizations generate enormous amounts of data through commercial transactions, financial operations, healthcare records, educational activities, manufacturing processes, telecommunications, and online services. The increasing availability of these datasets has created opportunities to discover valuable information and improve organizational decision-making. Nevertheless, traditional database management systems primarily concentrate on storing, retrieving, and organizing information rather than identifying complex relationships and predicting future developments. Consequently, data mining has emerged as an important analytical approach for transforming large quantities of raw information into meaningful knowledge. Data mining involves applying statistical methods, machine learning algorithms, artificial intelligence, and database technologies to identify previously unknown patterns and relationships within datasets. According to Fayyad et al. (1996), knowledge discovery in databases represents a systematic process involving data selection, preprocessing, transformation, mining, and interpretation. Data mining forms a central component of this broader knowledge discovery process. Its applications extend across multiple disciplines, including business intelligence, healthcare, banking, marketing, education, scientific research, and industrial management. Organizations increasingly depend on these techniques to understand historical performance, identify emerging trends, recognize customer behavior, and estimate future outcomes. The integration of data mining with large-scale database technologies therefore represents an important development in modern information management.

Knowledge discovery and predictive analysis are closely interconnected because the identification of historical patterns provides the foundation for estimating future events. Knowledge discovery focuses on identifying valid, understandable, and potentially useful information, whereas predictive analysis uses historical observations and statistical relationships to estimate unknown or future outcomes. Several data mining techniques contribute to these objectives. Classification algorithms categorize records into predefined groups, clustering identifies naturally occurring similarities, and association rule mining discovers relationships among frequently occurring items. Regression methods estimate numerical outcomes, while decision trees and neural networks support more complex predictive tasks. These

techniques are particularly valuable when applied to large-scale databases containing millions of records and numerous variables. However, processing such information presents challenges relating to data consistency, missing values, dimensionality, computational requirements, and algorithm performance. Efficient mining consequently requires appropriate preprocessing strategies, scalable computational architectures, reliable evaluation procedures, and domain-specific interpretation. Agrawal et al. (1993) demonstrated the importance of performance-oriented approaches to mining large databases, while Fayyad et al. (1996) emphasized the broader methodological foundations of knowledge discovery. Against this background, the present study examines the major data mining techniques used for knowledge discovery and predictive analysis in large-scale database environments. It investigates their theoretical foundations, methodological applications, analytical capabilities, practical benefits, limitations, and future development opportunities. The central argument is that effective data mining requires the integration of computational efficiency, statistical reliability, database management, and meaningful interpretation to generate actionable knowledge from extensive datasets.

II. LITERATURE REVIEW

The development of data mining has been influenced by several interconnected disciplines, including statistics, artificial intelligence, machine learning, pattern recognition, and database management. Early research established the theoretical foundations for extracting useful information from extensive databases. Frawley et al. (1992) described knowledge discovery as an important process for identifying meaningful patterns within stored information. Agrawal et al. (1993) investigated the performance requirements of data mining and demonstrated the importance of efficient algorithms when processing large transactional databases. Their work contributed to the development of scalable approaches for discovering frequent patterns and relationships. Quinlan (1993) presented decision tree learning techniques that provided a systematic approach to classification and prediction. These developments established the importance of combining computational methods with statistical reasoning to discover meaningful relationships. Fayyad et al. (1996) further developed the conceptual framework of knowledge discovery by explaining its iterative stages and emphasizing that data mining is only one component of the complete discovery process. Their contribution remains influential because it distinguishes the technical extraction of patterns from their interpretation and practical application. Similarly, Piatetsky-Shapiro and Frawley (1991) contributed to the early understanding of knowledge discovery through research on computational methods for extracting information from databases.

Subsequent research expanded the range of available algorithms and improved their applicability to complex datasets. Ester et al. (1996) introduced the DBSCAN clustering algorithm, which identifies clusters based on density and can discover arbitrarily shaped groups while identifying noise. This approach provided an alternative to traditional partition-based clustering methods. Han et al. (2000) developed the FP-growth approach for frequent pattern mining, reducing the need for repeated candidate generation and improving the efficiency of association analysis. Kohavi (1995) examined decision tree induction and its practical application to classification problems, while Breiman (2001) demonstrated the predictive capabilities of random forests through an ensemble learning approach. Hand et al. (2001) provided a comprehensive examination of statistical methods for data mining, connecting computational algorithms with statistical foundations and practical analytical requirements. Witten and Frank (2005) discussed machine learning techniques for extracting knowledge from datasets, including classification, clustering, association learning, and regression. Maimon and Rokach (2005) further organized major developments in data mining and knowledge discovery, covering algorithmic methods, preprocessing, and interdisciplinary applications. Collectively, these studies demonstrated that data mining was developing from a collection of isolated analytical techniques into an integrated discipline supporting large-scale information processing and predictive modelling.

Research during the later part of the period increasingly emphasized scalability, predictive accuracy, computational efficiency, and practical implementation. Friedman (1997) contributed to statistical learning through the development of flexible predictive modelling approaches, particularly multivariate adaptive regression techniques. Hastie et al. (2001) provided a comprehensive theoretical foundation for statistical learning, including classification, regression, model assessment, and dimensionality reduction. Tan et al. (2006) examined the relationship between data mining algorithms and practical knowledge discovery applications, emphasizing the importance of selecting appropriate techniques according to data characteristics and analytical objectives. Piatetsky-Shapiro (2007) discussed the growing importance of data mining in business and analytics, illustrating the transition from academic experimentation toward organizational implementation. These contributions reveal several important research themes: efficient processing of large datasets, improved model generalization, automated pattern identification, and the integration of data mining with decision-support systems. Despite these advances, existing research also identifies persistent difficulties involving data quality, high dimensionality, scalability, algorithm selection, interpretability, and privacy. The present study builds upon these foundations by examining how different mining approaches can be integrated into a comprehensive methodology for knowledge discovery and predictive analysis in large-scale database environments.

III. RESEARCH METHODOLOGY

The present study adopts a descriptive and analytical research methodology to examine data mining techniques for knowledge discovery and predictive analysis in large-scale databases. The methodological framework is based on established theoretical contributions, previously published research, and commonly recognized data mining procedures developed between 1990 and 2010. A secondary research approach is considered appropriate because the primary objective is to understand the conceptual foundations, comparative capabilities, methodological characteristics, and practical applications of established data mining algorithms. The study considers different analytical approaches, including classification, clustering, association rule mining, regression, decision trees, and neural networks. These techniques are examined according to their objectives, computational requirements, predictive capabilities, suitability for large datasets, and contribution to knowledge discovery. The methodological framework follows the general knowledge discovery process described by Fayyad et al. (1996), which consists of data selection, preprocessing, transformation, data mining, and interpretation. Each stage is considered essential because the quality of the final analytical results depends on the consistency and relevance of the information supplied to the mining algorithms. The study also considers the relationship between database management systems and mining operations, particularly the importance of scalable data storage, efficient retrieval, distributed processing, and computational optimization. Rather than claiming to have conducted a new empirical experiment, the research synthesizes established methodological principles to construct an analytical framework that can guide data mining implementation in large-scale information environments.

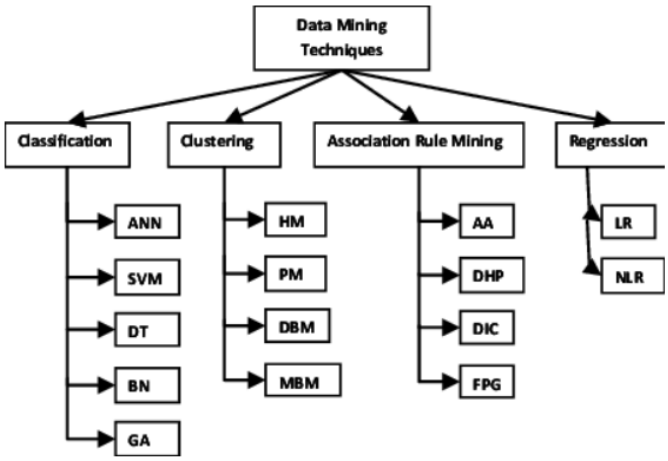


Fig.1. A Brief survey of Data Mining Techniques Applied to Agricultural Data

The first methodological stage involves identifying the data sources and understanding the characteristics of the available datasets. Large-scale databases may contain transactional records, customer information, financial statements, medical records, educational information, operational measurements, and other organizational data. These records may be stored in relational databases, data warehouses, distributed systems, or integrated information repositories. Data selection involves identifying the attributes and records relevant to the research objective while excluding irrelevant information that could increase computational complexity. The selected information must subsequently undergo preprocessing to improve consistency and reliability. Data cleaning is an important activity that involves identifying missing values, duplicate records, inconsistent formats, incorrect entries, and abnormal observations. Missing values may be addressed through suitable statistical estimation techniques, deletion where justified, or domain-specific replacement procedures. Duplicate observations must be examined to prevent repeated records from distorting the analytical results. Data integration is another important process because information may originate from different systems with inconsistent structures, naming conventions, and measurement units. Appropriate integration procedures help establish a consistent representation of the combined information. Data transformation subsequently converts the cleaned information into formats suitable for algorithmic analysis. Normalization, discretization, aggregation, encoding, and feature selection may be applied according to the requirements of individual algorithms. Normalization is particularly useful when variables contain substantially different numerical scales, while feature selection helps reduce irrelevant dimensions and improve computational efficiency. These preprocessing procedures contribute to the overall reliability of the knowledge discovery process by minimizing errors and ensuring that mining algorithms operate on appropriately structured information.

The second methodological stage focuses on selecting and applying appropriate data mining techniques according to the research objectives and characteristics of the available information. Classification is considered suitable when the

objective is to assign records to predefined categories. Decision trees, Bayesian classifiers, support vector machines, and neural networks are examples of classification approaches. Decision trees provide interpretable decision rules, while Bayesian methods estimate class probabilities using probabilistic relationships. Support vector machines identify separating boundaries between classes, and neural networks learn complex nonlinear relationships through interconnected computational units. Clustering is selected when the objective involves identifying naturally occurring groups without predefined class labels. K-means clustering is appropriate for partitioning numerical observations into a specified number of groups, whereas density-based techniques such as DBSCAN can identify irregularly shaped clusters and distinguish noise. Association rule mining is applied to transactional datasets to identify frequently occurring relationships among items. The Apriori algorithm identifies frequent itemsets through repeated candidate generation, whereas FP-growth uses a compressed tree representation to improve frequent-pattern discovery. Regression techniques are applied when the dependent variable is numerical and the objective is to estimate continuous outcomes. Linear regression provides a comparatively straightforward method for estimating relationships between variables, while more flexible statistical learning techniques can model nonlinear relationships. The selection of algorithms depends on dataset size, data type, analytical objectives, computational resources, and the required level of interpretability. For large-scale applications, algorithmic efficiency and the ability to process information in parallel are important considerations. The selected algorithms must therefore be evaluated not only according to their theoretical capabilities but also according to their operational suitability within the database environment.

The final methodological stage involves evaluating the analytical results, interpreting discovered patterns, and determining their practical relevance. Classification performance may be assessed through accuracy, precision, recall, and F1-score, while regression performance may be examined using mean absolute error, mean squared error, and root mean squared error. For clustering, evaluation may consider cluster cohesion, separation, and silhouette scores. Association rule mining may be assessed through support, confidence, and lift to determine the frequency, reliability, and strength of discovered relationships. Cross-validation is an important evaluation procedure because it helps determine whether a model performs consistently across different subsets of data. When predictive analysis involves future observations, chronological separation of training and testing datasets is particularly important to avoid using information that would not have been available at the prediction time. The methodology also recognizes the importance of detecting overfitting, where a model performs well on training data but poorly on unseen observations. Model interpretation is undertaken by examining the relationships identified by the algorithms and considering their relevance to the original analytical objectives. The final results should be assessed against domain knowledge to distinguish useful patterns from accidental correlations and misleading relationships. In addition, privacy protection, access control, data governance, and responsible information management are considered essential requirements when processing sensitive organizational information. The overall methodology therefore integrates data preparation, algorithm selection, model development, performance evaluation, and interpretation into a systematic analytical framework. This approach provides a foundation for understanding how different data mining techniques contribute to knowledge discovery and predictive analysis while addressing the computational and methodological challenges associated with large-scale databases.

IV. RESULTS AND DISCUSSION

The analytical examination of established data mining approaches demonstrates that different algorithms contribute distinct capabilities to knowledge discovery and predictive analysis in large-scale databases. Classification techniques are particularly valuable when historical records contain known categories and the objective is to predict the category of new observations. Decision trees offer an interpretable structure through which classification decisions can be explained using sequential conditions. For example, financial institutions can apply decision tree models to identify potential credit risk categories by examining historical payment patterns, income characteristics, and financial behavior. Bayesian classification provides another useful approach because it estimates the probability of different outcomes based on observed attributes. Neural networks can identify complex nonlinear relationships that may not be adequately represented by simpler statistical techniques. Ensemble approaches, particularly random forests, improve predictive performance by combining multiple decision trees and reducing the influence of individual model errors. These approaches demonstrate that predictive accuracy depends not only on algorithm selection but also on the quality, relevance, and representativeness of the training data. Classification models can support customer segmentation, fraud identification, medical diagnosis assistance, student performance prediction, and operational risk assessment. However, their effectiveness may be reduced by imbalanced datasets, irrelevant attributes, insufficient training examples, and changes in underlying data distributions. Clustering techniques contribute differently by identifying hidden group structures without requiring predefined labels. K-means clustering can organize customers according to purchasing behavior, transaction frequency, and spending characteristics, allowing organizations to identify groups with similar preferences. Density-based clustering can identify irregular patterns and unusual observations, which may be useful in detecting abnormal transactions or operational deviations. Association rule mining provides additional knowledge by discovering relationships among frequently occurring items and events. The Apriori and FP-growth algorithms enable

the identification of frequent combinations in transactional databases, supporting product recommendations, inventory management, and marketing decisions. Agrawal et al. (1993) established the importance of performance-oriented mining approaches, while Han et al. (2000) demonstrated how FP-growth could improve frequent-pattern discovery. Overall, the analysis indicates that classification, clustering, and association rule mining are complementary rather than competing techniques because each addresses a different aspect of knowledge discovery.

The examination of predictive analysis techniques further demonstrates their importance in identifying trends and estimating future outcomes from historical information. Regression methods provide a foundation for understanding relationships between independent and dependent variables and estimating continuous values. In financial applications, regression models can be used to estimate revenue, expenditure, credit-related outcomes, and market-related indicators based on historical observations. In manufacturing environments, predictive models can examine relationships between production variables, equipment conditions, and product quality. In healthcare information systems, appropriately validated models can support the analysis of patient-related patterns and the estimation of selected clinical outcomes. Time-dependent information can also be examined through predictive modelling approaches that identify trends, seasonal changes, and temporal relationships. Neural networks are particularly relevant when the relationships between variables are complex and nonlinear. However, their predictive advantages may be accompanied by increased computational requirements and reduced interpretability. Statistical learning techniques provide additional opportunities to balance predictive accuracy and generalization. Hastie et al. (2001) emphasized the importance of model selection, regularization, and statistical evaluation in developing reliable predictive systems. Breiman (2001) demonstrated the usefulness of ensemble learning through random forests, which can handle complex interactions and reduce prediction variability. Nevertheless, no single algorithm performs optimally across all datasets and analytical objectives. Model selection should therefore consider the nature of the dependent variable, the number of observations, feature dimensionality, missing information, class distribution, and available computing resources. Evaluation measures must also be selected according to the problem being addressed. Accuracy alone may be misleading when a classification dataset contains highly unequal class frequencies. Precision and recall become particularly important when the costs of false positive and false negative predictions differ substantially. Similarly, regression models should be assessed using appropriate error measures and testing procedures rather than relying exclusively on their performance against training observations. Cross-validation, independent testing, and appropriate temporal validation can improve confidence in the reliability of predictive results. These findings emphasize that predictive analysis should be treated as a systematic modelling and evaluation process rather than simply the application of an algorithm to historical data.

The overall discussion also identifies scalability, data quality, computational performance, and interpretability as major factors influencing the effectiveness of data mining in large-scale database environments. As database sizes increase, conventional processing approaches may become inefficient because of memory limitations, repeated database access, and the computational complexity of certain algorithms. Distributed computing and parallel processing offer opportunities to divide large mining tasks into smaller operations that can be executed across multiple processing units. Database integration is equally important because mining algorithms should be capable of accessing relevant information without unnecessary data movement. Efficient indexing, query optimization, partitioning, and data warehousing can improve the availability of information for analytical processing. The quality of the input data remains another significant concern. Incomplete records, inconsistent formats, duplicated observations, and inaccurate measurements may generate misleading patterns and reduce predictive reliability. Data preprocessing must consequently be regarded as an essential component of the overall analytical framework rather than a minor preliminary activity. The discussion also highlights the importance of domain knowledge in interpreting discovered relationships. A statistically significant association does not necessarily represent a causal relationship, and an apparently accurate predictive model may still generate inappropriate conclusions if the underlying data are biased or unrepresentative. Privacy and security requirements are also important when mining customer, financial, healthcare, and educational information. Organizations should implement appropriate access controls, data protection procedures, and governance policies to prevent unauthorized use of sensitive information. Piatetsky-Shapiro (2007) observed the increasing movement of data mining toward business applications and organizational analytics, demonstrating the practical significance of these developments. The findings therefore indicate that successful knowledge discovery requires more than algorithmic sophistication. It depends on the coordinated integration of suitable data sources, reliable preprocessing, efficient computational infrastructure, appropriate analytical methods, meaningful evaluation, and responsible interpretation. When these elements are combined, data mining can transform extensive databases into valuable organizational knowledge and provide a stronger foundation for evidence-based decision-making.

V. CONCLUSION

Data mining techniques play an essential role in knowledge discovery and predictive analysis by transforming large-scale databases into meaningful, structured, and actionable information. The increasing generation of digital

information across business, finance, healthcare, education, manufacturing, and other sectors has created a strong requirement for advanced analytical methods capable of identifying hidden relationships and predicting future developments. This study examined the theoretical foundations, methodological characteristics, applications, and limitations of major data mining techniques, including classification, clustering, association rule mining, regression, decision trees, neural networks, and ensemble learning. The analysis demonstrates that these techniques provide different but complementary capabilities for extracting knowledge from extensive datasets. Classification supports the prediction of predefined categories, clustering identifies previously unknown group structures, and association rule mining discovers relationships among frequently occurring items. Regression techniques estimate continuous outcomes, while neural networks and ensemble algorithms provide additional capabilities for modelling complex relationships. The integration of these techniques with database management systems creates opportunities for more effective information processing and analytical decision-making. The study also emphasizes that knowledge discovery is not limited to algorithmic pattern extraction but involves a complete process of data selection, preprocessing, transformation, modelling, evaluation, and interpretation. Each stage contributes to the reliability and practical value of the resulting knowledge. In particular, preprocessing procedures such as data cleaning, normalization, integration, and feature selection are essential for minimizing errors and improving model performance. Similarly, the selection of suitable algorithms must reflect the characteristics of the available information and the specific objectives of the analysis. The findings support the importance of combining statistical reasoning, machine learning, computational efficiency, and domain expertise when developing data mining applications. Consequently, data mining should be regarded as an integrated analytical framework that connects large-scale information resources with organizational knowledge requirements.

The study further concludes that predictive analysis provides considerable opportunities for improving operational efficiency, identifying potential risks, supporting strategic planning, and strengthening organizational decision-making. Historical data can reveal patterns that support the estimation of future customer behavior, financial performance, production requirements, and operational outcomes. Organizations can use these insights to develop more effective business strategies, improve resource allocation, identify unusual activities, and respond to changing conditions. However, the successful implementation of predictive systems depends on several important factors. Data quality remains fundamental because inaccurate, incomplete, or inconsistent information can produce unreliable analytical results. Scalability is another major consideration because large databases require efficient storage, retrieval, parallel processing, and computational optimization. Furthermore, predictive models must be evaluated using appropriate statistical measures to ensure that their performance generalizes to unseen observations. Accuracy, precision, recall, F1-score, and regression error measurements provide useful evaluation criteria depending on the analytical task. Model interpretability is equally important because organizations must understand the relationships and assumptions underlying predictions before using them in important decisions. Privacy, security, and responsible data governance must also be incorporated into the development and deployment of data mining systems. The study recognizes that discovering associations does not automatically establish causality and that predictive accuracy alone does not guarantee practical usefulness. Therefore, domain-specific interpretation and continuous model assessment are necessary to ensure that discovered knowledge remains meaningful. The overall findings demonstrate that combining appropriate data mining techniques with reliable database infrastructure can significantly improve the process of extracting information from large-scale datasets. The contribution of data mining extends beyond information processing because it enables organizations to convert accumulated records into analytical knowledge and informed decisions. Ultimately, effective knowledge discovery and predictive analysis depend on the integration of algorithmic capabilities, reliable information management, computational resources, and responsible analytical practices.

VI. FUTURE WORK

Future research should investigate more advanced and scalable data mining approaches capable of processing increasingly large, complex, and heterogeneous datasets. The integration of distributed database technologies, parallel computing, cloud-based infrastructure, and advanced machine learning algorithms offers opportunities to improve computational efficiency and analytical performance. Researchers should explore hybrid data mining frameworks that combine classification, clustering, association analysis, and predictive modelling to address complex analytical requirements. Particular attention should be given to developing algorithms that can efficiently process high-dimensional information, streaming data, and heterogeneous database structures. Improved feature selection, dimensionality reduction, automated preprocessing, and adaptive learning techniques may also contribute to reducing computational costs while maintaining predictive reliability. Another important research direction involves developing interpretable predictive models that provide understandable explanations for their predictions and discovered relationships.

Future studies should also investigate the responsible application of data mining across sensitive and highly regulated environments, including healthcare, banking, education, and public administration. Data privacy, security, algorithmic

fairness, transparency, and responsible information governance should become essential components of large-scale knowledge discovery frameworks. Researchers may further examine how predictive models can adapt to changing data distributions and maintain reliable performance over extended periods. Comparative empirical studies using real-world databases would help identify the strengths and limitations of different algorithms under varying computational conditions. The development of standardized evaluation frameworks could also improve the reproducibility and practical reliability of data mining research. Ultimately, future developments should focus on creating intelligent, scalable, interpretable, and secure analytical systems that transform large volumes of information into dependable knowledge for organizational and societal decision-making.

REFERENCES

1. Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207–216.
2. Agrawal, R., Srikant, R., & others. (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, 487–499.
3. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.
4. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226–231.
5. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37–54.
6. Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI Magazine, 13(3), 57–70.
7. Friedman, J. H. (1991). Multivariate adaptive regression splines. The Annals of Statistics, 19(1), 1–67.
8. Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 1–12.
9. Hand, D., Mannila, H., & Smyth, P. (2001). Principles of data mining. MIT Press.
10. Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. Springer.
11. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137–1143.
12. Maimon, O., & Rokach, L. (Eds.). (2005). Data mining and knowledge discovery handbook. Springer.
13. Piatetsky-Shapiro, G. (2007). Data mining and knowledge discovery 1996 to 2005: Overcoming the hype and moving from university to business and analytics. Data Mining and Knowledge Discovery, 15, 99–105. https://doi.org/10.1007/s10618-006-0058-2
14. Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann.
15. Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques (2nd ed.). Morgan Kaufmann.