American Journal of Applied Research and AI (AJARAI)

VOLUME 1 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Applied Research and AI (AJARAI))

Volume 1 Issue 2, Year 2026
ISSN: 3142-8932 (Online)
DOI: https://doi.org/10.54536/ajarai.v1i2.8327
https://journals.e-palli.com/home/index.php/ajarai

Revisiting Data Mining for Facebook's Transition to Meta: Challenges and Future Directions

Wael Ghazi Shaaban[1*]

Article Information

Received: June 09, 2026
Accepted: August 24, 2026
Published: September 24, 2026

Keywords
Association-Rule Mining, Big Data Analytics, Clustering, Data Mining, Data Privacy, Meta, Metaverse

ABSTRACT

Facebook's rebranding to Meta represents a strategic transition from a dominant social-media platform toward an integrated, immersive ecosystem built around virtual reality (VR), augmented reality (AR), and interconnected digital spaces. This shift substantially widens the scope, type, and complexity of user-generated data, extending it beyond clicks and text into spatial movement, gesture-based input, device telemetry, and cross-platform behavioural signals. This paper examines which data-mining (DM) techniques are best suited to interpreting this new multi-modal data environment, focusing on clustering and association-rule mining (ARM), and the challenges that accompany their application. Drawing on a structured review of academic and industry literature, the paper discusses how clustering enables behavioural segmentation of VR/AR users while ARM uncovers cross-platform usage patterns that support retention, cross-selling, and seamless user-journey design. The review also identifies four recurring challenges in Meta's data-mining pipeline: privacy and regulatory exposure, data quality and heterogeneity, real-time processing demands, and model interpretability/algorithmic bias. The paper concludes that clustering and ARM offer complementary analytical foundations for Meta but require reinforcement through privacy-preserving data mining (PPDM), anonymization, secure multiparty computation (SMC), incremental and streaming algorithms, and transparent, bias-audited model governance if Meta is to responsibly realise its metaverse ambitions.

INTRODUCTION

Data mining (DM) refers to the process of studying large datasets using statistical analysis and machine-learning (ML) algorithms to discover data structures, patterns, and meaningful insights (Han, Kamber and Pei, 2012). Facebook has heavily invested in DM to analyze users' clicks, location stamps, tag-tracking logs, screen-time records, and textual interactions. As the firm is being rebranded to Meta (Reuters, 2021), it is shifting from a dominant social-media platform toward an integrated ecosystem mounted around VR, AR, and immersive digital spaces. Such a transition significantly expands the scope, type, and complexity of user-generated data to include spatial movements, gesture-based inputs, device telemetry, and cross-platform signals. Against this transformation, re-modelling current Facebook DM techniques becomes essential to making sense of multi-modal data, optimizing user experience, and enabling Meta to build the envisioned "metaverse". Two data-mining techniques particularly relevant to this shift are clustering and association-rule mining (ARM), and their application is inseparable from the wider evolution of big-data analytics and the governance challenges that accompany it.

This paper therefore situates Facebook's transition to Meta within the broader landscape of contemporary data-mining practice: it reviews how clustering and ARM function, traces the evolution of big-data analytics that underpins them, identifies the principal challenges Meta faces in re-engineering its data pipeline, and discusses what these findings imply for the company's analytical and governance strategy.

Research Question

This paper is guided by the following research question: Which data-mining techniques are most relevant to Facebook's transition into Meta's metaverse ecosystem, and what challenges and mitigation strategies emerge from applying these techniques to a multi-modal, sensor-driven data environment? Two supporting questions follow from this: (a) how do clustering and association-rule mining each contribute to interpreting VR/AR-generated user data; and (b) what data-governance measures are required to manage the privacy, quality, real-time processing, and interpretability challenges that accompany this transition?

LITERATURE REVIEW

Evolution of Data Analytics and Big Data

Data analytics has many definitions, most of which refer to the systematic examination of raw data, structured or unstructured, to identify patterns, capture insights, study trends, and visualize them in a way that supports informed decision-making. The practice goes back a hundred years; however, its evolution has accelerated during the Fourth Industrial Revolution (IE4) due to advances in communication, computing power, and cloud storage, leading to increasing demand for Big-Data application in contemporary business analytics (Abuqabita, Al-Omoush

[1] London Metropolitan University, Saudi Arabia
[*] Corresponding author's email: Wael.shaaban@gmail.com

and Alwidian, 2019).

Alongside Big-Data, there is a growing integration of AI and ML into automated analytical tasks including segmentation, forecasting, and anomaly detection, which enables firms to progress from descriptive reporting toward diagnostic, predictive, and prescriptive analytics (Shahnawaz and Kumar, 2025). This integration has produced trends such as batch computing, stream analytics, and interactive analytics across many business domains, with sales and marketing, R&D, and supply chain remaining the dominant areas in a market size expected to exceed USD 100 billion by 2028 (Shahnawaz and Kumar, 2025), illustrated in Figure 1 below.

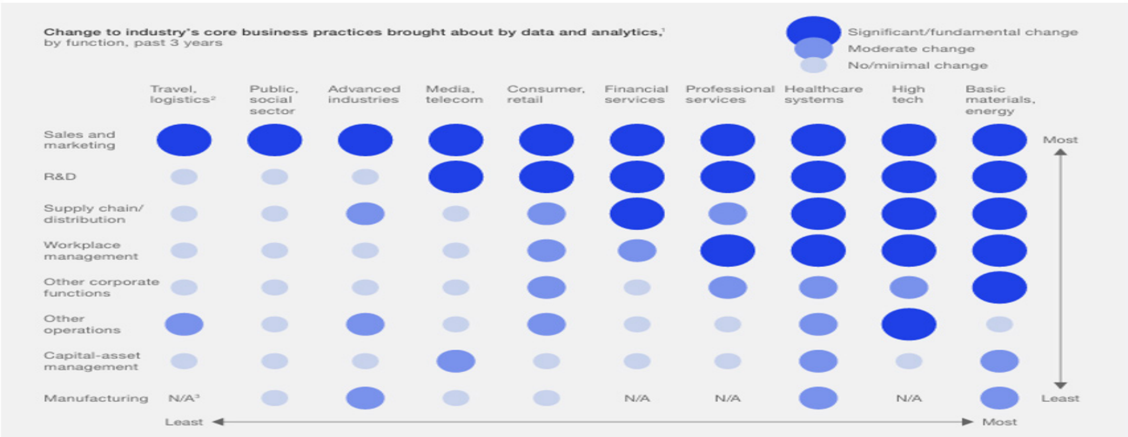

Figure 1: Areas of Data Analytics Application
Source: McKinsey & Company (2018)

Batch computing offers advanced analysis of pre-stored blocks of Big-Data (Tudoran, Nicolae and Brasche, 2017); firms such as Carrefour and Walmart run batch analytics on global sales data to identify demand patterns, inventory turnover rates, and product affinities. Stream analytics enables processing of data blocks sequentially generated at high rates in real time (Zhu, Tang and Li, 2019); Netflix, YouTube, and financial-fraud-detection systems used by banks or payment-gateway providers (e.g., Visa, Mastercard) employ this framework to monitor thousands of transactions occurring every second. Interactive analytics provides the capability to query and visualize Big-Data streams within a very short response time (Garg and Sharma, 2014); firms such as Facebook, LinkedIn, and Google employ this framework to offer instant query processing of user interactions, searches, clicks, and application usage within seconds.

Closely related to this is the emergence of graph analytics, a method designed to analyze relationships and connections among data nodes. This trend is growing in domains where understanding networks and reasoning is crucial, such as cybersecurity, social networking, and e-marketing (Grzegorz *et al.*, 2010; Ren *et al.*, 2014).

These emerging trends are already delivering measurable

Table 1: Big-Data Characteristics and Associated Data-Mining Challenges

Big Data V-Characteristic	**Challenge**	**Impact**	**Primary Data-Mining Stage(s) Impacted**
Veracity	Data Quality (noise, missing data, inconsistent formats)	Low accuracy, missing entries, or corrupted values	Data Collection, Preprocessing
	Privacy, Security & Ethical Issues	Sensitive data exposure, Bias, Regulatory non-compliance	Data Collection, Evaluation, Deployment
Volume	Scalability / Volume / Storage Bottleneck	Large datasets exceed storage and processing capacity	Data Collection, Data Mining
	Computational Complexity and Performance	Algorithms require heavy processing and long run-time	Data Mining, Evaluation, Deployment
Velocity	Velocity / Real-Time Processing	High-speed data streams requiring instant analysis	Data Collection, Data Mining, Deployment
Variety	Data Heterogeneity (Structured / unstructured)	Different data formats (text, images, sensor logs, etc.)	Preprocessing, Integration, Transformation
	Integration & Interoperability	Data spread across systems with inconsistent governance	Data Integration, Transformation, Deployment

Source: Author Findingsw

impact across multiple industries but still carry many challenges (K.U and David, 2014) associated with Big-Data characteristics across the different stages of DM (Shahnawaz and Kumar, 2025), summarised in Table 1 below.

In summary, the latest developments in data analytics, driven by AI and ML and enabled by real-time collection and processing of Big-Data, are reshaping contemporary industries. These innovations support faster decision-making, deeper insight generation, and improved operational resilience, reinforcing the strategic value of data science in addressing real-world business challenges. However, these advantages also introduce difficulties that stretch beyond current technology capabilities, particularly for a company undertaking the scale of transformation that Meta is pursuing.

Clustering as an Unsupervised Technique

Clustering is used to group data points based on the similarity of core characteristics. It allows the discovery of natural structures within large datasets (Singh and Singh, 2024). For Meta, this provides a strategic capability to understand behavioural diversity among VR/AR users, a data scope far bigger than traditional Facebook interactions. The metaverse records include headset-usage patterns, session duration, motion trajectories, avatar interactions, and device-switching logs. Applying clustering to this Big-Data enables grouping users into more meaningful categories such as "social immersion users", "VR gamers", "AR productivity adopters", or "hybrid multi-device explorers". These segments support user-specific UX as well as personalized marketing, including content recommendations and targeted advertising. Clustering, as an unsupervised ML technique, is also suitable because it does not require labelled data (Asghari, 2026); this is ideal where social-data-point categories are still emerging, as is the case with the Meta metaverse (Moujahid and Dornaika, 2025).

Association-Rule Mining (ARM)

ARM focuses on identifying frequent patterns, dependencies, and co-occurrence relationships within datasets (Pinheiro, Guerreiro and Mamede, 2024). In Meta's context, ARM can uncover how users navigate across the company's integrated platforms. For example, it can identify whether certain activities in "Horizon" correlate with increased Instagram engagement, or whether specific VR actions precede purchasing an Oculus device. This type of pattern discovery provides actionable insights for retention strategies, cross-selling, and the design of seamless user journeys across Facebook, WhatsApp, Instagram, and Horizon environments. In addition, ARM can detect behavioural combinations inside VR spaces, such as users who frequently join community events while simultaneously engaging through Messenger invites. These insights are essential to refine social-presence features, community algorithms, and UX recommendations.

Meta's transformation creates a complex, sensor-dependent data environment in which clustering and ARM offer strong, complementary analytical foundations for interpreting immersive patterns and optimizing UX. However, the two techniques carry different strengths and limitations. Clustering supports exploratory analysis in multi-dimensional environments but may struggle as the monitored characteristics become highly heterogeneous in large-scale datasets. ARM excels at identifying co-occurring patterns but requires careful thresholding to avoid bias, excessive correlation, or irrelevant rules, especially in fast-moving VR platforms. Real-time clustering may require incremental algorithms, while ARM may need streaming-based extensions to handle continuous user activity in such spaces. Accordingly, both techniques require substantial computing power and sustained, long-term ML investment.

Challenges of Data Mining in Meta's Transition

Although DM offers significant competitive advantages to firms investing in Knowledge Discovery in Databases (KDD), it introduces complex challenges across its processes: data collection, preparation, mining, and interpretation (Baig *et al.*, 2023). Table 2 below summarises its frequent challenges.

Table 2: Data Mining Challenges

Obstacles in the Data Mining Industry
Security and Social Challenges
Noisy and Incomplete Data
Distributed Data
Complex Data
Performance
Scalability and Efficiency of the Algorithms
Improvement of Mining Algorithms
Incorporation of Background Knowledge
Data Visualization
Mining Dependent on Level of Abstraction

Source: (Baig et al., 2023)

Meta's transition involves dealing with the above-mentioned challenges at different levels of the DM pipeline. Reuters (2021) highlights that Meta is already under substantial scrutiny for its data practices, while the introduction of VR/AR devices increases privacy risks concerning location tracking, movement trajectories, room-mapping, biometrics, and social-presence signals. Questions are raised about consent, intrusive tracking, and algorithmic misuse, especially given the low level of privacy awareness globally (see Figure 1 below). Data-regulatory frameworks — including the European General Data Protection Regulation, the California Consumer Privacy Act (CCPA), the Turkish Personal Data Protection Law (KVKK), and the Saudi Personal Data Protection Law (PDPL) — impose strict constraints on what can be collected and how it is processed. To mitigate these challenges, Meta must adopt measures such as Privacy-Preserving Data Mining (PPDM) (Hewage, Sinha and Naeem, 2023), anonymization (Sweeney, 2002), and Secure Multiparty Computation (SMC) (Tran and Hu, 2019).

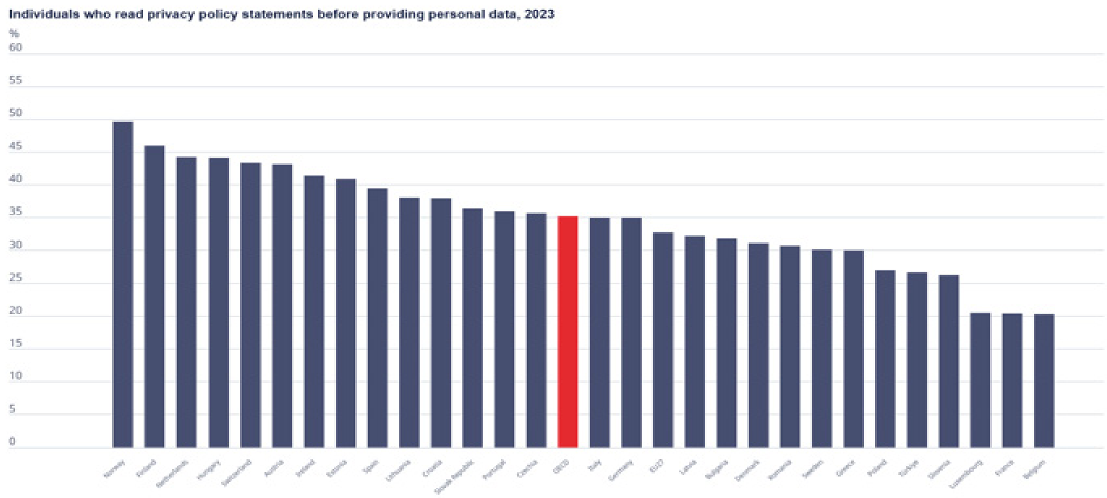

Figure 2: Levels of Privacy Awareness (2023)
Source: OECD

A second major challenge is data quality. Meta's new ecosystem integrates data from many different sources whose types differ in format, structure, frequency, and noise levels. For example, VR telemetry may contain missing values when sensors disconnect, and motion data may require sophisticated filtering. Combining this with Facebook, Instagram, and WhatsApp logs creates a highly fragmented landscape that necessitates correction algorithms to avoid misleading insights. To address this challenge, Koukaras and Tjortjis (2025) suggest a framework for pre-processing and feature engineering.

A third challenge stems from real-time processing of Big-Data, whose volume, velocity, and variety characteristics cause not only data deluge but also affect infrastructure performance when accommodating a growing volume of user interactions (Shahnawaz and Kumar, 2025). As Reuters (2021) notes, Facebook already serves 2.9 billion monthly users, and the new metaverse will address a considerable portion of this number, multiplied by many variables reading live data streams. Traditional batch processing may be insufficient; Meta must rely on distributed computing frameworks run over live stream-processing engines, incremental ML algorithms, and optimized cloud infrastructure to manage this scale (Lalaoui, El Haji and Kounaidi, 2025).

A fourth challenge relates to model interpretability and algorithmic bias. VR/AR data, rich in behavioural and biometric signals, can lead to "black box" models that are difficult to explain, justify, or reconcile with user trust (Atzmueller *et al.*, 2024). In addition, segmentation techniques such as clustering may reinforce social stereotypes if not carefully monitored, potentially producing negative user experiences from overly intrusive algorithms within immersive environments. To address these risks, Meta must incorporate assurance techniques that explain results and audit bias, including human-in-the-loop validation, fairness-ML regulation, and transparent model reporting (Rabonato and Berton, 2025).

In summary, Meta's challenges require not only advanced technical solutions but also strong data governance, regulatory alignment, and ethical oversight. By integrating these measures into its DM processes, Meta can pursue an immersive and responsible data-mining strategy and build greater trust as it shifts from a social-media company into a metaverse-driven ecosystem.

MATERIALS AND METHODS

This paper adopts a qualitative, desk-based literature-review methodology rather than primary data collection, given that its purpose is conceptual: to evaluate which data-mining techniques and challenges are most relevant to Facebook's transition into Meta. Secondary sources were drawn from peer-reviewed journal articles, conference proceedings, and industry reports published mainly between 2019 and 2026, supplemented by authoritative industry commentary (Reuters, McKinsey & Company, OECD, IBM) to ground the analysis in Meta's real-world transition.

Sources were located through academic databases and search engines (e.g., Scopus-indexed journals and Google Scholar) using keyword combinations such as "data mining", "clustering", "association-rule mining", "Big Data challenges", "metaverse data", and "privacy-preserving data mining". Selection criteria prioritised recency, relevance to social-media or immersive-technology contexts, and methodological credibility (peer review or recognised industry authorship). Sources that did not directly inform the paper's two core techniques (clustering and ARM) or its four identified challenge categories (privacy/regulation, data quality, real-time processing, interpretability/bias) were excluded from detailed discussion, though many remain listed in the reference list as background reading.

The analytic approach applied to the selected literature was thematic synthesis: findings were grouped first by data-mining technique (clustering, ARM) and then by challenge category, allowing the two strands to be brought together in the Results and Discussion section to answer the research question. As with any literature-based methodology, the approach is limited by its reliance on secondary sources and publicly available industry commentary rather than primary access to Meta's internal data infrastructure; conclusions about Meta specifically should therefore be read as informed inference rather than empirically verified findings.

RESULTS AND DISCUSSION

The literature reviewed indicates that clustering and ARM are each suited to different, complementary aspects of Meta's data-mining needs. Clustering's strength lies in exploratory, unsupervised segmentation of heterogeneous behavioural data — precisely the kind of headset-usage, motion-trajectory, and device-switching data that VR/AR platforms generate — allowing Meta to build user personas without needing pre-labelled categories (Asghari, 2026; Moujahid and Dornaika, 2025). ARM, by contrast, is better suited to uncovering cross-platform dependencies (for example, links between Horizon activity and Instagram engagement), which supports retention and cross-selling strategies that clustering alone cannot reveal (Pinheiro, Guerreiro and Mamede, 2024). Taken together, the two techniques answer the first supporting research question: clustering interprets "who" the user is behaviourally, while ARM interprets "what sequences" of behaviour matter across Meta's integrated platforms.

The four challenge categories identified in the review — privacy and regulatory exposure, data quality, real-time processing, and interpretability/bias — are not independent of the choice of technique; rather, they compound it. Clustering large, heterogeneous, biometric-rich VR/AR datasets increases both the privacy risk discussed by Reuters (2021) and the interpretability risk discussed by Atzmueller *et al.* (2024), since richer feature sets make both re-identification and "black-box" outcomes more likely. ARM's dependence on frequent-pattern thresholds, meanwhile, is most exposed to the data-quality and real-time-processing challenges: fragmented, noisy, or delayed data from Meta's several platforms can generate spurious or missed association rules unless pre-processing (Koukaras and Tjortjis, 2025) and streaming-based extensions (Lalaoui, El Haji and Kounaidi, 2025) are in place.

This interaction between technique and challenge answers the second supporting research question. Governance measures cannot be treated as an add-on to the analytics; they need to be embedded technique-by-technique. For clustering, this means combining segmentation with anonymization (Sweeney, 2002) and PPDM (Hewage, Sinha and Naeem, 2023) before behavioural personas are used in marketing or UX decisions. For ARM, this means applying SMC (Tran and Hu, 2019) where rules are derived across platforms owned by the same company but drawing on data from users who may not expect cross-platform correlation, alongside fairness auditing and human-in-the-loop review to prevent stereotype-reinforcing rules (Rabonato and Berton, 2025).

Overall, the discussion suggests that Meta's transition is best served not by choosing between clustering and ARM but by deploying them in tandem, wrapped in a governance layer proportionate to the four identified challenges. This reading is consistent with the broader analytics literature, which shows firms progressing from descriptive to predictive and prescriptive analytics only once data-quality and governance foundations are in place (Shahnawaz and Kumar, 2025); Meta's metaverse ambitions make that sequencing more, not less, important.

CONCLUSION

Meta's transformation into a metaverse-driven organization has created a complex, sensor-dependent data environment in which clustering and Association Rule Mining (ARM) provide robust analytical foundations for interpreting immersive user behaviour and optimizing user experience (UX). Clustering facilitates exploratory analysis within high-dimensional virtual environments by identifying hidden behavioural segments, although its effectiveness may decline as monitored characteristics become increasingly heterogeneous and large-scale datasets grow more complex. Similarly, ARM is effective in uncovering co-occurring behavioural patterns and relationships among user activities; however, its performance depends heavily on appropriate threshold selection to avoid generating excessive, biased, or irrelevant association rules, particularly within dynamic virtual reality (VR) ecosystems. Both techniques demand substantial computational resources, continuous machine learning investment, and comprehensive governance mechanisms to ensure reliable and responsible deployment. Based on the literature reviewed, several strategic recommendations emerge for Meta's data-mining approach during its metaverse transition. Privacy-Preserving Data Mining (PPDM), data anonymization techniques, and Secure Multiparty Computation (SMC) should be integrated as default safeguards before behavioural clusters or association rules are incorporated into product development or marketing decisions. Furthermore, Meta should invest in incremental clustering algorithms and streaming-based ARM models capable of processing continuous, real-time VR and augmented reality (AR) user interactions instead of relying exclusively on conventional batch-processing methods. Establishing a comprehensive data pre-processing and feature-engineering pipeline is also essential to address missing, noisy, or inconsistent data resulting from sensor failures and the integration of heterogeneous platforms such as Facebook, Instagram, WhatsApp, and Horizon. In addition, institutionalizing human-in-the-loop validation, fairness-oriented machine learning audits, and transparent model reporting can improve interpretability while reducing the risk of algorithmic bias and stereotype reinforcement associated with clustering outcomes. Finally, Meta should proactively align its data collection, storage, and processing practices with major international data protection regulations, including the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), Turkey's Personal Data Protection Law (KVKK), and the Personal Data Protection Law (PDPL), to address increasing regulatory scrutiny surrounding its data governance practices. Collectively, these measures support the development of an immersive, responsible, and trustworthy analytics strategy. They also answer the research question posed at the outset by demonstrating that clustering and Association Rule Mining are the most suitable techniques for analysing Meta's multi-modal metaverse data; however, their effectiveness can only be fully realised when complemented by robust privacy protection, high data quality, real-time processing capabilities, and transparent, interpretable governance frameworks. Achieving this balance provides Meta with the strongest foundation for building user trust as it evolves from a traditional social media platform into a comprehensive metaverse ecosystem.

REFERENCES

Abbasi, A., Sarker, S., & Chiang, R. H. L. (2016). Big data research in information systems: toward an inclusive research agenda. *Journal of the Association for Information Systems, 17*(2), Article 3.

Abuqabita, F., Al-Omoush, R., & Alwidian, J. (2019). A comparative study on big data analytics frameworks, data resources and challenges. *Modern Applied Science, 13*(7).

Achouch, M., Dimitrova, M., Ziane, K., Karganroudi, S. S., Dhouib, R., Ibrahim, H., & Adda, M. (2022). On predictive maintenance in Industry 4.0: overview, models, and challenges. *Applied Sciences, 12*(16), 8081–8103.

Adejoh, J., Owoh, N., Ashawa, M., Hosseinzadeh, S., Shahrabi, A., & Mohamed, S. (2025). An adaptive unsupervised learning approach for credit card fraud detection. *Big Data and Cognitive Computing, 9*(9), Article 217.

Alam, M. N., Laxmi, V., Kumar, N., Kumari, R., & Sharma, S. (2025). Unsupervised machine learning for anomaly detection: a systematic review. *International Journal of Electrical and Electronics Engineering, 12(*7), 34–72.

Ali, Z., Bhaskar, S., & Kannan, S. (2019). Descriptive statistics: Measures of central tendency, dispersion, correlation and regression. *Airway, 2(*3), 120.

Asghari, A. (2026). TQC: An intelligent clustering approach for large-scale, noisy, and imbalanced data. *Machine Learning with Applications, 23*, 100800.

Ashraf, A., Rayed, C. A., Awad, N. A., & Sabry, H. M. (2025). A framework for customer segmentation to improve marketing strategies using machine learning. *Procedia Computer Science, 260*, 616–625.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*(1), 685–725.

Atzmueller, M., Fürnkranz, J., Kliegr, T., & Schmid, U. (2024). Explainable and interpretable machine learning and data mining. *Data Mining and Knowledge Discovery, 38,* 2571–2595.

Baig, S. F. A., Hussain, Nadeem, M. A., Malik, Mushtaq, Khan, K. H., & Abbas. (2023). Data mining methods and obstacles: a comprehensive analysis. *Journal of Computing & Biomedical Informatics, 6*(1), 13.

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *[Journal title not provided], 128*(1), 301–315.

Canhoto, A. I. (2021). Leveraging machine learning in the global fight against money laundering and terrorism financing: an affordances perspective. *Journal of Business Research, 131,* 441–452.

Chullamonthon, P., & Tangamchit, P. (2023). Ensemble of supervised and unsupervised deep neural networks for stock price manipulation detection. *Expert Systems with Applications, 220*, Article 119698.

Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (2007). Unsupervised learning: association rules. *In Data mining*. Springer.

Coker, B., Rudin, C., & King, G. (2021). A theory of statistical inference for ensuring the robustness of scientific results. *Management Science, 67*(10), 6174–6197.

Cooksey, R. W. (2020). *Descriptive statistics for summarising data*. Springer.

Cui, G., Leung Wong, M., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science, 52*(4), 613–622.

Davenport, T. H. (2013). Analytics 3.0. *Harvard Business*

Review, 64–72.

Douaioui, K., Oucheikh, R., Benmoussa, O., & Mabrouki, C. (2024). Machine learning and deep learning models for demand forecasting in supply chain management: a critical review. *Applied System Innovation, 7*(5), 93–117.

Eggert, M. (2019). Big data research-How to structure the changes of the past decade? In *M. Eggert, The art of structuring.* Springer International Publishing.

Eggert, M., & Alberts, J. (2020). Frontiers of business intelligence and analytics 3.0: a taxonomy-based literature review and research agenda. *Business Research, 13*(2), 685–739.

Farman, H., Khan, A. W., Ahmed, S., Khan, D., Imran, M., & Bajaj, P. (2024). An analysis of supervised machine learning techniques for churn forecasting and component identification in the telecom sector. *Journal of Computing & Biomedical Informatics, 7,* Article 478-0701.

Fulk, G. P. (2023). Descriptive statistics, an important first step. *Journal of Neurologic Physical Therapy, 47*(2), 63.

Gao, H., Kou, G., Liang, H., Zhang, H., Chao, X., Li, C.-C., & Dong, Y. (2024). Machine learning in business and finance: a literature review and research opportunities. *Financial Innovation, 10,* Article 86.

Garg, R., Bhaskar, S. B., Das, S., & Harsoor, S. (2020). Probability and inferential statistics. *Airway, 3(*1), 19–24.

Garg, P., & Sharma, V. (2014). An efficient and secure data storage in mobile cloud computing through RSA and hash function. *In Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad*, India.

Ghous, H., Malik, H. H., & Rehman, I. (2023). Deep learning based market basket analysis using association rules. *KIET Journal of Computing and Information Sciences, 6*(2), 14–34.

Goneppanavar, U., Ali, Z., Bhaskar, S. B., & Divatia, J. V. (2019). Types of data, methods of collection, handling and distribution. *Airway, 2*(1), 36–40.

Grzegorz, M., Austern, M. H., Bik, A. J. C., Dehnert, J. C., Horn, I., Leiser, N., & Czajkowski, G. (2010). Pregel: a system for large-scale graph processing. *In Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data* (pp. 135–146).

Guetterman, T. C. (2019). Basics of statistics for primary care research. *Family Medicine and Community Health, 7*(2), 67–74.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: *Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Henke, N., Libarikian, A., & Wiseman, B. (2016). Straight talk about big data. *McKinsey & Company*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/straight-talk-about-big-data

Hewage, U. H. W. A., Sinha, R., & Naeem, M. A. (2023). Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy:a systematic literature review. *Artificial Intelligence Review, 56,* 10427–10464.

IBM. (2022). Four steps to better business decisions. *IBM Corporation.*

K. U., J., & David, J. M. (2014). Issues, challenges and solutions: big data mining. *In Sixth International Conference on Networks & Communications, Cochin, India* (pp. 131–140).

Kaliyadan, F., & Kulkarni, V. (2019). Types of variables, descriptive statistics, and sample size. *Indian Dermatology Online Journal, 10*(1), 82–86.

Kaur, P., Stoltzfus, J., & Yellapu, V. (2018). Descriptive statistics. *International Journal of Academic Medicine, 4(*1), 60–63.

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics* (2nd ed.). The MIT Press.

Koukaras, P., & Tjortjis, C. (2025). Data preprocessing and feature engineering for data mining: techniques, tools, and best practices. *AI, 6*(10), Article 257.

Lalaoui, I. L., El Haji, E., & Kounaidi, M. (2025). The evolution and challenges of real-time big data: a review. *Computer Sciences & Mathematics Forum, 10*(1), Article 11.

Lee, J. S., & Shin, D. C. (2025). Prediction of waste generation using machine learning: a regional study in Korea. *Urban Science, 9*(8), 297–312.

Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. *IEEE, 12*(1), 70434–70463.

Martinez, Escobar, Garca-Daz, M. E., & Pinto-Roa, D. P. (2021). Market basket analysis with association rules in the retail sector using Orange. *CLEI Electronic Journal, 24*(2), Article 12.

McKinsey & Company. (2018). *Analytics comes of age.* https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/analytics%20comes%20of%20age/analytics-comes-of-age.pdf

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning.* The MIT Press.

Moujahid, A., & Dornaika, F. (2025). Advanced unsupervised learning: a comprehensive overview of multi-view clustering techniques. *Artificial Intelligence Review, 58,* Article 234.

Murphy, K. R. (2021). In praise of Table 1: the importance of making better use of descriptive statistics. *Industrial and Organizational Psychology, 14*(4), 461–477.

OECD. (2023). *Privacy and data protection*. https://www.oecd.org/en/topics/policy-issues/privacy-and-data-protection.html

Pinheiro, C., Guerreiro, S., & Mamede, H. S. (2024). A survey on association rule mining for enterprise architecture model discovery. *Business & Information Systems Engineering, 66*, 777–798.

Rabonato, R., & Berton, L. (2025). A systematic review of fairness in machine learning. *AI and Ethics, 5*(3), 1943–1954.

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed,

M. M. S. (2022). Predicting employee attrition using machine learning approaches. *Applied Sciences, 12(*13), 6424–6441.

Razzaq, K., & Shah, M. (2025). Machine learning and deep learning paradigms: from techniques to practical applications and research frontiers. *Computers, 14*(3), 93–120.

Ren, Y., Domeniconi, C., Zhang, G., & Yu, G. (2014). A weighted adaptive mean shift clustering algorithm. I*n Proceedings of the 2014 SIAM International Conference on Data Mining* (SDM), Philadelphia.

Reuters. (2021, October 29). *Facebook changes name to Meta as it refocuses on virtual reality.* https://www.reuters.com/technology/facebooks-zuckerberg-kicks-off-its-virtual-reality-event-with-metaverse-vision-2021-10-28/

Rolf, B., Beier, A., Jackson, I., Müller, M., Reggelin, T., Stuckenschmidt, H., & Lang, S. (2024). A review on unsupervised learning algorithms and applications in supply chain management. *International Journal of Production Research, 63(*5), 1933–1983.

Sadeghi, P., Karimi, H., Lavafian, A., Rashedi, R., Samieefar, N., Shafiekhani, S., & Rezaei, N. (2024). Machine learning and artificial intelligence within pediatric autoimmune diseases: Applications, challenges, future perspective. *Expert Review of Clinical Immunology, 20*(10), 1219–1236.

Sarker, I. H. (2021). Machine learning: algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), Article 160.

Schmitz, G. M. (2024). An unsupervised machine learning analysis of environmental sustainability indicators using the K-means clustering algorithm among 2,485 global corporations. *The International Journal of Sustainability Policy and Practice, 21*(1), 77–100.

Shafie, M. R., Khosravi, H., Farhadpour, S., Das, S., & Ahmed, I. (2024). A cluster-based human resources analytics for predicting employee turnover using optimized artificial neural networks and data augmentation. *Decision Analytics Journal, 11*, Article 100461.

Shahnawaz, M., & Kumar, M. (2025). A comprehensive survey on big data analytics: characteristics, tools and techniques. *ACM Computing Surveys, 57*(8), 1–33.

Sharafuddin, S., & Belik, I. (2022). The evolution of data analytics through the lens of business cases. *Online Information Review, 46*(6), 1167–1183.

Sheard, J. (2018). Quantitative data analysis. In K. Williamson & G. Johanson (Eds.), *Research methods: Information, systems, and contexts* (2nd ed.). Chandos Publishing.

Singh, J., & Singh, D. (2024). A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: taxonomy, challenges, applications and future prospects. *Advanced Engineering Informatics, 62,* Article 40.

Sitaram, R. (2021). An employee segmentation model. Tilburg University, School of Humanities and Digital Sciences.

Starbuck, C. (2023). *The fundamentals of people analytics: With applications in* R. Springer.

Suh, Y. (2023). Machine learning based customer churn prediction in home appliance rental business. *Journal of Big Data, 10,* Article 41.

Sweeney, L. (2002). Achieving K-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 571–588.

Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability, 14*(12), 7243–7258.

Taweesan, A., Kanabkaew, T., Surinkul, N., & Polprasert, C. (2025). Integrating clustering algorithms and machine learning to optimize regional snapshot municipal solid waste management for achieving sustainable development goals. *Environmental Advances, 19,* Article 100607.

Tran, H.-Y., & Hu, J. (2019). Privacy-preserving big data analytics: A comprehensive survey. *Journal of Parallel and Distributed Computing, 134*, 207–218.

Tudoran, R., Nicolae, B., & Brasche, G. (2017). Data multiverse: the uncertainty challenge of future big data analytics. *In Semantic keyword-based search on structured data sources: IKC 2016*. Lecture notes in computer science. Springer.

Veena, S., & Aravindhar, J. (2025). Challenges of machine learning algorithms used in business for decision-making. *In Proceedings of the 3rd International Conference on Optimization Techniques in the Field of Engineering* (ICOFE-2024), Namakkal, India.

Watson, D. S. (2023). On the philosophy of unsupervised learning. *Philosophy & Technology, 36,* Article 28.

Wissuchek, C., & Zschech, P. (2024). Prescriptive analytics systems revised: a systematic literature review from an information systems perspective. *Information Systems and e-Business Management, 23*(2), 279–353.

Xia, W., Jiang, Y., Chen, X., & Zhao, R. (2021). Application of machine learning algorithms in municipal solid waste management: a mini review. *The Journal for a Sustainable Circular Economy, 40*(6), 609–624.

Yin, J., & Fernandez, V. (2020). A systematic review on business analytics. *Journal of Industrial Engineering and Management, 13*(2), 283–295.

Zhang, Y. (2024). Machine learning-based customer segmentation: a comprehensive investigation of techniques, challenges and applications. *E-commerce and Artificial Intelligence.*

Zhu, J. Y., Tang, B., & Li, V. O. K. (2019). A five-layer architecture for big data processing and analytics. *International Journal of Big Data Intelligence, 6*(1), 38.

Zou, Q., Duan, H., Yang, Z., Wang, X., Tian, Y., Hou, H., & Yang, J. (2025). Machine learning-assisted solid waste life-cycle management: applications constraints, and future opportunities. *ScienceDirect, 219*, Article 108320

Appendix

Appendix A: Glossary of Abbreviations

VR: Virtual Reality.
AR: Augmented Reality.
UX: User Experience.
KDD: Knowledge Discovery in Databases.
Data Regulatory Frameworks: globally recognised examples include the European General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), the Turkish Personal Data Protection Law (KVKK), and the Saudi Personal Data Protection Law (PDPL).
PPDM: Privacy-Preserving Data Mining.
SMC: Secure Multiparty Computation — a technique that allows different parties to jointly compute a certain functionality without revealing personal data.
OECD: The Organisation for Economic Co-operation and Development.