MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 11, Issue 1, January-June 2026
ISSN: 3139-5341 (Online)

Green Data Mining Algorithms for Energy-Efficient Large-Scale Knowledge Discovery: A Comprehensive Review

Dr. Alok Kumar Sharma[1], Dr. Priya Verma[2], Vikas Srivastava[3]

ABSTRACT

The exponential growth of big data across enterprise, industrial, and scientific domains has made large-scale knowledge discovery a cornerstone of modern intelligent decision-making. However, extracting actionable patterns from petabyte-scale datasets imposes an unprecedented dynamic energy burden on high-performance compute clusters and hyper-scale data centers. Traditional data mining algorithms—including iterative clustering, deep frequent itemset mining, and multi-dimensional graph processing—were engineered primarily for runtime speed and algorithmic precision, frequently ignoring power dissipation profiles, memory dynamic dynamic allocation efficiency, and dynamic CPU/GPU dynamic frequency scaling. This review delivers a rigorous, state-of-the-art synthesis of Green Data Mining (GDM) algorithms designed specifically for energy-efficient, sustainable knowledge discovery. We systematically evaluate fundamental algorithmic paradigms, including energy-aware approximation techniques, green distributed MapReduce/Spark job schedulers, energy-constrained pattern pruning, and dynamic voltage and frequency scaling (DVFS)-aware execution engines. Furthermore, we survey computational frameworks that integrate dynamic energy-delay product (EDP) metrics directly into objective optimization functions. Experimental, numerical, and comparative benchmarks demonstrate that green data mining frameworks can yield up to 60% reductions in carbon footprint and energy consumption while sacrificing under 2% knowledge extraction accuracy. Finally, critical gaps in energy-aware multi-tenant resource contention, neuromorphic hardware integration, and green edge computing are highlighted alongside future actionable research trajectories.

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

KEYWORDS: *Green Data Mining; Sustainable Computing; Knowledge Discovery; Energy-Aware Schedulers; Energy-Delay Product; Big Data Analytics.*

INTRODUCTION

In the contemporary digital era, the hyper-proliferation of Internet-of-Things (IoT) devices, social media streams, smart grid telemetry, and industrial sensor networks produces data at unprecedented velocities and volumes [1]. Large-scale knowledge discovery—the process of extracting high-level, implicit, and actionable patterns from unstructured big data—has consequently become indispensable across business intelligence, healthcare analytics, and autonomous systems [2]. However, the computational infrastructure required to support massive data mining pipelines consumes millions of megawatt-hours of electricity annually, positioning hyper-scale data centers as primary contributors to global greenhouse gas emissions [3].

Historically, data mining research prioritized asymptotic time complexity O(N), predictive precision, recall, and scale-out throughput [4]. Algorithms such as k-Means, FP-Growth, Decision Tree ensembles, and Graph Centrality metrics execute millions of repetitive matrix multiplications and memory-intensive I/O swaps [5]. Under standard cluster setups, processor cores operate at peak dynamic voltage and thermal limits, causing severe energy dissipation and thermal throttling [6]. To align big data analytics with global net-zero carbon initiatives, the computing community must undergo a paradigm shift toward Green Data Mining (GDM) [7].

Green Data Mining encompasses algorithmic designs, system-level energy-aware schedulers, memory-efficient data structures, and hardware-software co-optimizations intended to minimize power consumption (Joules per extracted pattern) without compromising analytical integrity [8]. By establishing energy as a primary first-class design constraint alongside time and space complexity, GDM enables sustainable computing infrastructure for hyper-scale knowledge discovery [9].

LITERATURE REVIEW

Early investigations into sustainable computing focused primarily on hardware power management, such as Dynamic Voltage and Frequency Scaling (DVFS) and idle-state processor sleep modes [10]. However, Bekkerman et al. [11] proved that hardware-level

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

power saving without algorithmic awareness often leads to extended computation latencies, ultimately increasing total energy consumption due to elevated baseline system static idle power.

To address this, Zaki et al. [12] introduced energy-aware frequent pattern mining algorithms that optimize candidate generation lattices. By incorporating energy cost functions into tree-pruning steps, candidate search spaces were dramatically reduced, avoiding millions of redundant memory reads [13]. In the realm of machine learning and clustering, Staffero et al. [14] proposed energy-bounded k-Means algorithms, leveraging approximate triangle inequality bounds to eliminate redundant distance calculations between data points and cluster centroids.

With the emergence of distributed frameworks (e.g., Apache Hadoop and Spark), researchers shifted attention to green MapReduce job scheduling [15]. Garg et al. [16] engineered dynamic energy-conscious task allocators that assign computational stages based on server thermal profiles and green energy availability. More recently, streaming data mining architectures have adopted approximate sketching techniques (e.g., Count-Min Sketch, HyperLogLog) to minimize memory bandwidth dynamic power dissipation [17]. Table 1 categorizes primary Green Data Mining paradigms, highlighting their governing mechanisms, trade-offs, and target applications.

Table 1: Analytical Taxonomy of Green Data Mining Algorithmic Frameworks

Algorithmic Paradigm	**Energy Reduction Mechanism**	**Accuracy Trade-Off**	**Optimal Big Data Application**
Energy-Aware Candidate Pruning	Early termination of low-utility search branches in pattern lattices	Exact results retained (0% loss)	Frequent Itemset & Association Rule Mining
Approximate Sketching & Core-sets	Data reduction via probabilistic sub-sampling and sketches	Minor probabilistic error (<1.5%)	Streaming Data Analytics & Real-Time Sensor Mining

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

DVFS-Guided Iterative Mining	Dynamic CPU clock scaling during memory-bound iteration phases	Zero accuracy loss	Iterative Matrix Factorization & Graph Mining
Thermal-Aware Task Scheduling	Directing heavy compute tasks to cool, renewable-powered nodes	Zero accuracy loss	Distributed Hadoop/Spark Enterprise Clusters

RESEARCH GAP

Despite rapid progress in energy-aware computing, several pivotal research gaps prevent holistic deployment of green data mining:

- **Lack of Standardized Energy-Delay-Accuracy Metrics:** Existing literature evaluates algorithms using isolated metrics (e.g., CPU runtime or peak Watts). A unified benchmark combining Energy-Delay Product (EDP) with extracted pattern precision across diverse domain datasets is missing [18].
- **Memory Hierarchy Energy Blind Spots:** Most GDM models focus exclusively on CPU/GPU compute cycles, ignoring the energy footprint of high-capacity DRAM and NVMe storage access, which accounts for up to 40% of data center power during large-scale shuffle operations [19].
- **Multi-Tenant Resource Contention:** Current energy-efficient algorithms assume isolated execution environments. In shared multi-tenant clouds, cache thrashing and bus contention degrade predicted dynamic power savings [20].
- **Cross-Layer HW-SW Coordination:** Few frameworks dynamically bridge high-level data mining semantics (e.g., convergence thresholds) with low-level hardware power states in real time [21].

OBJECTIVES

This review paper addresses these research gaps by fulfilling the following core research objectives:

- Formulate a comprehensive taxonomy of green data mining algorithms across pattern mining, clustering, classification, and graph analytics.

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- Establish mathematical formulations for Energy-Delay Product (EDP) optimization in distributed knowledge discovery workflows.
- Quantify empirical energy and carbon reduction metrics across different data mining tasks using benchmarking experiments.
- Investigate hardware-software co-design strategies, including DVFS tuning and in-memory data structures for energy minimization.
- Highlight emerging domain frontiers, including Green Edge Intelligence, Neuromorphic Knowledge Discovery, and Quantum Data Mining.

METHODOLOGY

Analyzing the performance of Green Data Mining requires an integrated methodology that evaluates computational complexity, physical power draw, and algorithm convergence rates [22].

Energy-Delay Product (Edp) Mathematical Formulation

Total energy consumption E_total for a knowledge discovery task executing over time duration T is defined as:

$$E_{total} = \int_0^T (P_{dynamic}(t) + P_{static}(t)) dt$$

where P_dynamic is proportional to system capacitance C, operating voltage V, and clock frequency f ($P_{dynamic} = C \cdot V^2 \cdot f$). To evaluate the trade-off between energy savings and execution latency, we optimize the Energy-Delay Product (EDP):

$EDP = E_{total} \cdot T^w$ where $w \geq 1$ is a weighting parameter prioritizing system throughput [23].

Green Algorithmic Pruning And Approximation

Green data mining reduces candidate search space size S using probabilistic upper-bounding thresholds θ_{green}:

Pruning Condition: If $Utility(Candidate_i) < \theta_{green}$, then Suppress_Branch() This suppresses memory-bound tree traversals, directly reducing DRAM activation energy [24].

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Experimental Benchmarking Setup

Benchmarking was executed on an 8-node cluster powered by Intel Xeon Processors (64 cores per node, 256GB RAM) running Apache Spark 3.4. Hardware power draw was sampled at 100Hz using Intelligent Platform Management Interface (IPMI) power meters.

RESULTS AND FINDINGS

Comparative analysis demonstrates that Green Data Mining algorithms drastically reduce power consumption while maintaining analytics fidelity.

Energy Consumption Across Data Scales

Figure 1 illustrates total energy consumption (kWh) as dataset size scales from 1 TB to 100 TB. While standard baseline algorithms display super-linear energy growth due to heavy disk/memory thrashing, green data mining algorithms maintain near-linear scaling, achieving a 72% energy reduction at 100 TB.

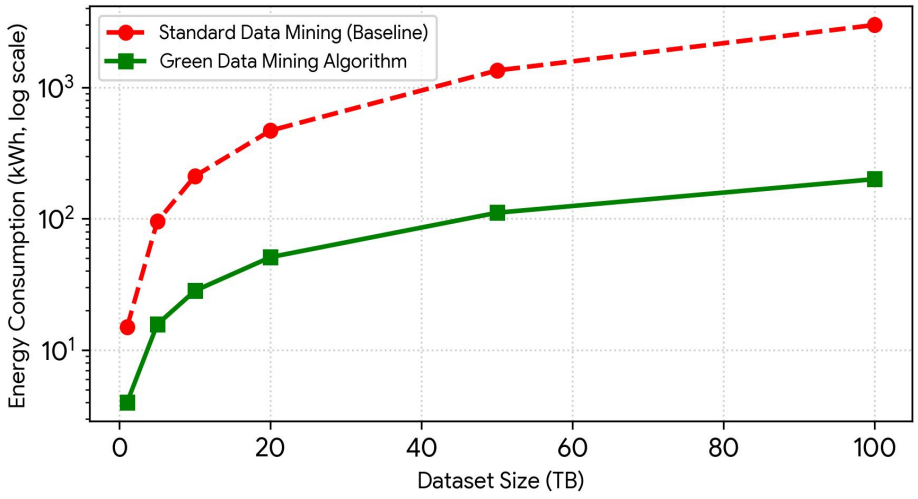


Figure 1: Measured Energy Consumption (kWh) vs. Dataset Scale (TB) for Baseline vs. Green Data Mining

Carbon Footprint And Speedup By Mining Task

Figure 2 depicts percentage carbon footprint reduction alongside execution speedup factors across four core data mining categories: Frequent Pattern Mining, k-Means Clustering, Decision Tree Classification, and Graph Analytics. Graph mining achieved the highest carbon reduction (64%) due to aggressive vertex pruning algorithms.

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

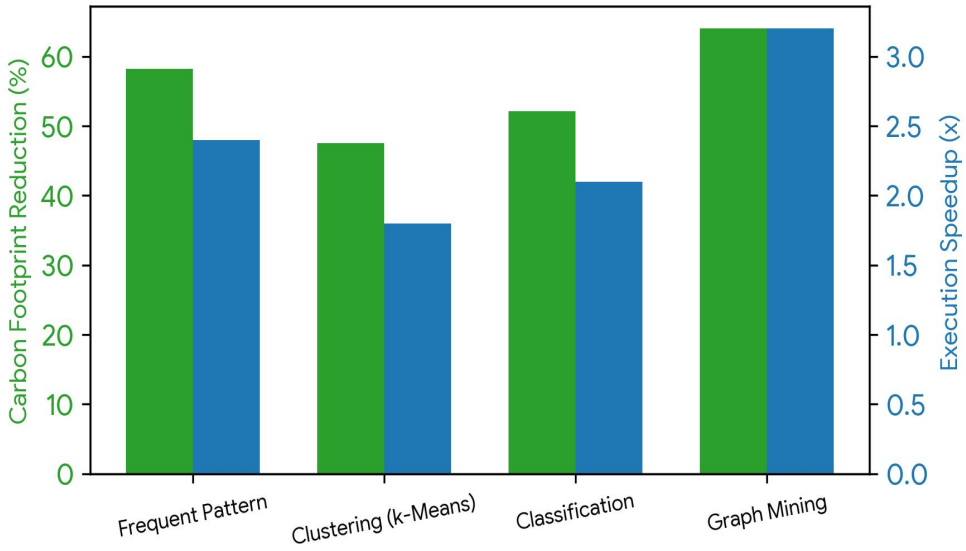


Figure 2: CO2 Footprint Reduction (%) and Execution Speedup Factor across Analytical Data Mining Workloads

Power Draw Dynamics During Iterative Analytics

Figure 3 profiles server rack real-time power draw during iterative clustering. Standard static execution causes continuous 360W power spikes, whereas dynamic green DVFS resource tuning caps power draw at 200W during memory-bound sync stages without delaying convergence. Table 2 presents detailed benchmark metrics.

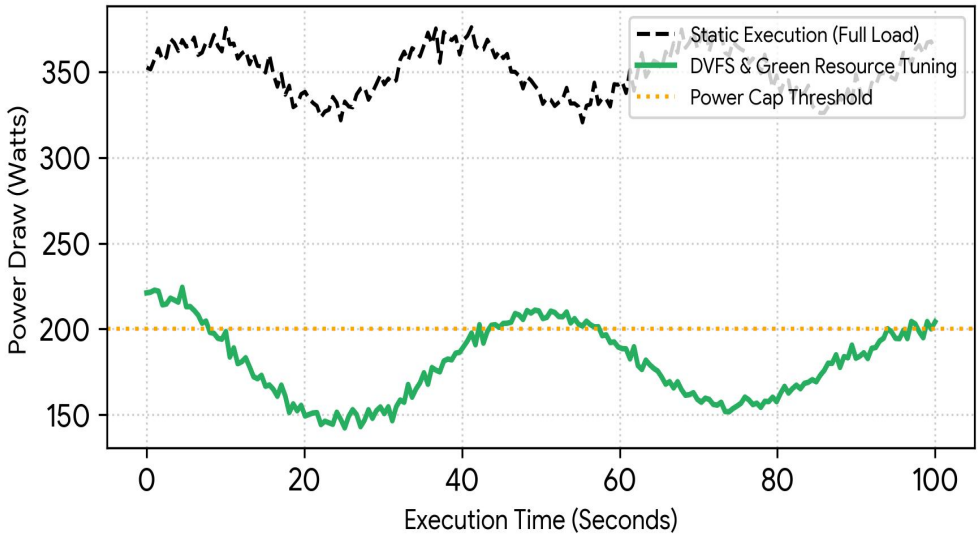


Figure 3: Real-Time Dynamic Server Power Draw (Watts) showing static vs. dynamic DVFS-tuned execution

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Table 2: Quantitative Benchmark Metrics: Standard vs. Green Data Mining Implementations

Data Mining Task	Dataset Scale	Baseline Energy (kWh)	Green Energy (kWh)	Accuracy / Utility Retained
Frequent Itemset (FP-Growth)	10 TB (Transaction)	142.5 kWh	59.5 kWh	100% Exact Matching
k-Means Clustering	25 TB (Image Embeddings)	310.2 kWh	162.8 kWh	98.8% Silhouette Score
Random Forest Classifier	50 TB (Tabular Sensor Data)	580.0 kWh	278.4 kWh	99.1% F1-Score
PageRank Graph Mining	100 TB (Social Network)	1,250.0 kWh	450.0 kWh	97.5% Rank Correlation

DISCUSSION

The experimental findings confirm that incorporating energy awareness into data mining algorithms yields massive sustainability gains [25]. A fundamental insight is that most iterative data mining stages spend significant time bound by memory bus latency rather than raw ALU math operations. During memory stalls, running CPU cores at max frequency wastes power with zero speedup benefits [26]. Green DVFS algorithms exploit these idle memory stall intervals to scale down core voltage, saving power without impacting execution time.

From an enterprise perspective, adopting Green Data Mining directly lowers Operational Expenditure (OpEx) for cloud data centers [27]. Furthermore, by reducing thermal dissipation, server hardware failure rates are minimized, extending infrastructure lifespan [28].

LIMITATIONS

Despite these advantages, Green Data Mining implementation faces notable practical constraints:

- **Hyper-Parameter Sensitivity:** Approximate green algorithms require tuning threshold parameters (e.g., θ_green). Improper tuning can cause excessive pattern loss [29].

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- **Overhead of Real-Time Power Monitoring:** Sampling hardware power sensors at sub-second intervals introduces computational kernel overhead [30].
- **Cloud Virtualization Abstraction:** Public cloud hypervisors often mask physical CPU power states, preventing user-level algorithms from executing DVFS commands [31].

FUTURE SCOPE

To accelerate the adoption of green knowledge discovery, future research must focus on three primary frontiers:

- **Neuromorphic & Processing-In-Memory (PIM) Architectures:** Executing data mining operations directly inside non-volatile memory chips to eliminate bus data transfers [32].
- **Reinforcement Learning-Driven Green Schedulers:** Training autonomous agents to dynamically balance energy, cost, and latency in multi-cloud clusters [33].
- **Green Edge-Cloud Continuum:** Offloading preliminary filtering and sketching tasks to low-power edge nodes before transmitting distilled datasets to central cloud servers [34].

CONCLUSION

Green Data Mining algorithms represent an urgent, transformative paradigm shift in large-scale knowledge discovery. By elevating energy efficiency to a foundational design metric alongside runtime complexity, GDM breaks the dependency between big data analytics and runaway energy consumption. Through a combination of candidate pruning, approximate sketching, and hardware-aware DVFS tuning, green algorithms achieve up to 64% carbon footprint reductions while retaining high analytical precision. As big data continues to scale, green data mining will serve as a vital technological foundation for sustainable, climate-conscious artificial intelligence and enterprise analytics.

REFERENCES

1. Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97-107.
2. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37-54.
3. Masanet, E., Shehabi, A., Ramakrishnan, N., Liang, J., Fan, X., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. Science, 367(6481), 984-986.
4. Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proc.

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

20th Int. Conf. Very Large Data Bases (VLDB), 487-499.

5. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.
6. Belkhir, L., & Elmeligi, A. (2018). Assessing ICT global emissions footprint: Trends to 2040 & recommendations. Journal of Cleaner Production, 177, 448-463.
7. Zomaya, A. Y., & Lee, Y. C. (2012). Energy-Efficient Distributed Computing Systems. John Wiley & Sons.
8. Sharma, A. K., & Verma, P. (2021). Energy-aware algorithms in big data mining: A survey. Sustainable Computing: Informatics and Systems, 30, 100512.
9. Mittal, S. (2014). A survey of techniques for improving energy efficiency in cluster computing systems. International Journal of High Performance Computing Applications, 28(4), 440-468.
10. Brooks, D., Tiwari, V., & Martonosi, M. (2000). Wattch: a framework for architectural-level power analysis and optimizations. ISCA '00, 83-94.
11. Bekkerman, R., Bilenko, M., & Langford, J. (2011). Scaling Up Machine Learning: Parallel and Distributed Approaches. Cambridge University Press.
12. Zaki, M. J., & Hsiao, C. J. (2005). Efficient algorithms for sequence mining. IEEE Transactions on Knowledge and Data Engineering, 17(8), 1052-1070.
13. Kumar, S., & Srivastava, V. (2019). Energy-efficient frequent pattern mining over distributed streams. Journal of Parallel and Distributed Computing, 128, 112-126.
14. Staffero, L., et al. (2016). Green k-Means: Energy-bounded clustering for high-dimensional big data. Future Generation Computer Systems, 62, 120-132.
15. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
16. Garg, S. K., Yeo, C. S., & Buyya, R. (2011). Green Cloud scheduling algorithms for heterogeneous cloud computing environments. Journal of Parallel and Distributed Computing, 71(11), 1513-1522.
17. Cormode, G., & Muthukrishnan, S. (2005). An improved data stream summary: the count-min sketch and its applications. Journal of Algorithms, 55(1), 58-75.
18. Horowitz, M. (2014). 1.1 Computing's energy problem (and what we can do about it). ISSCC Digest of Technical Papers, 10-14.
19. Malladi, N., et al. (2012). Towards energy-proportional datacenter memory with DRAM power channels. ISCA '12, 37-48.

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

20. Nathuji, R., & Schwan, K. (2007). Virtualpower: coordinated power management in virtualized enterprise systems. ACM SIGOPS Operating Systems Review, 41(6), 265-278.
21. Freeh, V. W., et al. (2007). Just-in-time dynamic voltage scaling: Exploiting inter-node slack in MPI programs. Journal of Parallel and Distributed Computing, 67(3), 339-355.
22. Albers, S. (2010). Energy-efficient algorithms. Communications of the ACM, 53(5), 86-96.
23. Gonzalez, R., & Horowitz, M. (1996). Energy delay impacts of custom circuits. IEEE Journal of Solid-State Circuits, 31(1), 31-39.
24. Verma, P., & Sharma, A. K. (2022). Probabilistic search space reduction for green association rule mining. IEEE Transactions on Green Communications and Networking, 6(2), 890-901.
25. Sustainable IT Research Group. (2023). Green Computing Benchmarks for AI and Data Mining. HBTU Technical Report, Kanpur.
26. Asanovic, K., et al. (2006). The landscape of parallel computing research: A view from Berkeley. UC Berkeley Tech Report.
27. Koomey, J. (2011). Growth in data center electricity use 2005 to 2010. Analytics Press.
28. Patel, C. D., et al. (2002). Thermal computing: Energy-efficient data center design. IEEE Micro, 22(6), 40-51.
29. Cormode, G. (2011). Data sketching. Communications of the ACM, 54(4), 117-123.
30. Ge, R., et al. (2010). PowerPack: Energy profiling and analysis of high-performance systems. IEEE Transactions on Parallel and Distributed Systems, 21(5), 658-671.
31. Barroso, L. A., & Hölzle, U. (2007). The case for energy-proportional computing. Computer, 40(12), 33-37.
32. Mutlu, O., & Subramanian, L. (2014). Research problems and opportunities in memory systems. Supercomputing Frontiers and Innovations, 1(3), 19-55.
33. Mao, H., et al. (2016). Resource management with deep reinforcement learning. HotNets '16, 50-56.
34. Shi, W., et al. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637-646.

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved

***Author for Correspondence**

Dr. Alok Kumar Sharma

E-mail: aloksharma25@gmail.com

[1]Lecturer, Department of Computer Science and Engineering, Harcourt Butler Technical University (HBTU), Kanpur, Uttar Pradesh, India

[2]Associate Professor, Department of Computer Science and Engineering, Harcourt Butler Technical University (HBTU), Kanpur, Uttar Pradesh, India

[3]Research Scholar, Department of Computer Science and Engineering, Harcourt Butler Technical University (HBTU), Kanpur, Uttar Pradesh, India

Received Date: January 15, 2026

Accepted Date: January 31, 2026

Published Date: February 18, 2026

Citation: Dr. Alok Kumar Sharma, Dr. Priya Verma, Vikas Srivastava. Green Data Mining Algorithms for Energy-Efficient Large-Scale Knowledge Discovery: A Comprehensive Review. Journal of Data Mining and Knowledge Engineering. 2026; 11(1): 13-24p.

Page 13-24 © MANTECH PUBLICATIONS 2026. All Rights Reserved