Author

2026

OPTIMIZING HOTEL GUEST SEGMENTATION USING K-MODES AND KNEEDLE ALGORITHMS



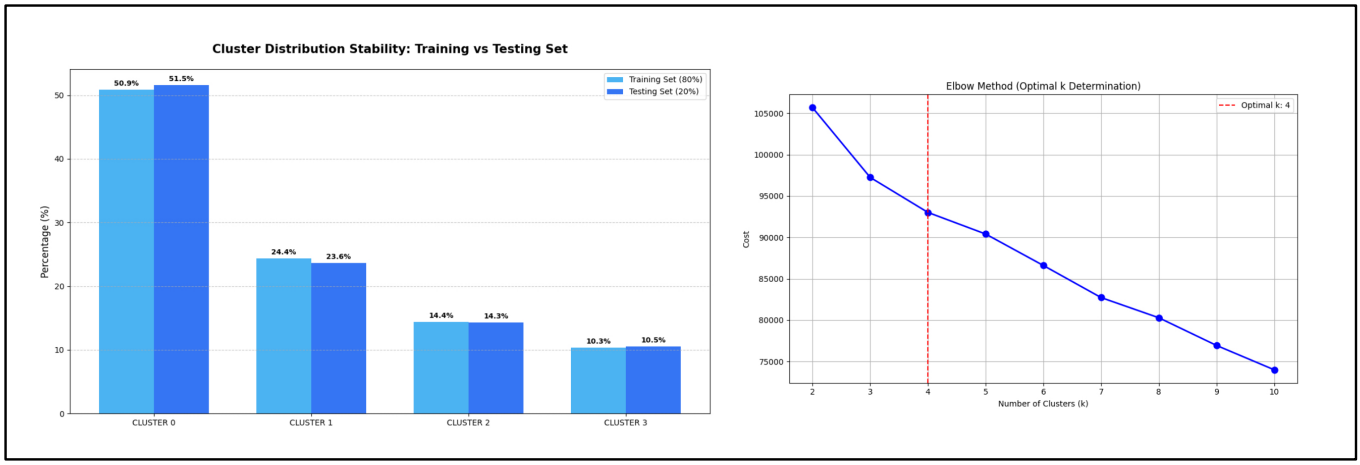
In today's rapidly evolving tourism and hospitality industry, delivering personalized guest experiences has become a fundamental requirement for sustaining customer loyalty and operational excellence. Digital transformation across luxury hotels and resorts enables properties to capture comprehensive guest interaction histories through modern Property Management Systems (PMS).

Every day, these operational databases record tens of thousands of reservation transactions encompassing guest demographics, countries of origin, stay durations, room preferences, and booking channels. Yet, in many hospitality enterprises, these valuable records remain underutilized as static administrative logs rather than actionable business intelligence.

A critical technical barrier in analyzing hotel reservation logs lies in the predominantly nominal and categorical nature of guest attributes. Conventional clustering techniques such as *K-Means* rely on Euclidean distance over continuous numerical variables. Forcing categorical attributes—such as nationality codes, room types, or reservation statuses—into numerical space via *one-hot encoding* inflates feature dimensionality and produces artificial cluster centroids that lack operational meaning.

To overcome this challenge, a study prepared for publication in the **SKANIKA Journal (Sistem Komputer dan Teknik Informatika)** titled ***"Optimasi Segmentasi Tamu Hotel Menggunakan Algoritma K-Modes dan Kneedle pada Data Kategorikal"*** introduces an end-to-end unsupervised machine learning pipeline tailored specifically for high-dimensional categorical hospitality records.

As documented in the SKANIKA journal manuscript, the research investigates **43,492 historical PMS transaction records** collected from four-star hotel and resort properties in Bali. Through a rigorous data preparation pipeline—incorporating anomaly filtering, age-group and length-of-stay discretization, and rare-category consolidation (using a minimum frequency threshold of 10 transactions)—a refined dataset of **43,340 clean guest profiles** is established.



The proposed computational architecture integrates three core innovations:

1. **K-Modes Clustering with Deterministic Cao Initialization:** Instead of computing numerical means, the system evaluates guest similarity using simple matching dissimilarity and selects initial cluster modes based on attribute frequency density, guaranteeing reproducible segmentation.
2. **Objective Cluster Selection via the Kneedle Algorithm:** Eliminating subjective visual guesswork in traditional Elbow plots, the Kneedle algorithm mathematically pinpoints the point of maximum curvature on the normalized cost curve, identifying k = 4 as the optimal number of guest personas.
3. **Cross-Partition Stability & Cramér's V Feature Attribution:** The model is validated across an 80:20 train-test split to verify generalization stability, while bias-corrected Cramér's V quantifies the exact discriminatory power of each guest attribute.

Key Empirical Findings and Real-World Hospitality Impact

Experimental evaluation reported in the SKANIKA journal demonstrates exceptional structural stability across partitions, with cluster proportion deviations between training and testing sets remaining within 0.10% to 0.76%. The framework uncovers four actionable guest personas: (1) Domestic Agency-Reliant Guests (50.9%), (2) East Asia Digital Short-Stayers (24.4%), (3) Oceania High-Value Loyalists (14.4%), and (4) Oceania Digital Explorers (10.3%). Furthermore, Cramér's V analysis reveals that guest gender (V = 0.489) and length of stay (V = 0.425) act as the primary drivers of cluster differentiation. By embedding these empirical personas into hotel Customer Relationship Management (CRM) platforms, hospitality managers can automate targeted promotions, tailor room and amenity packages, and elevate guest satisfaction across Indonesian tourism destinations.