Evaluasi Klaster dan Pola Asosiasi Keterampilan untuk Interpretasi Domain Profesi ICT

Rafif Pramatya Ramadhan[1], Mardiana Andarwati[2]
Sistem Informasi, Fakultas Teknologi Informasi, Universitas Merdeka Malang
e-mail: rafifpramatyar5@gmail.com

Kata Kunci:

*Bundle Skill*s
Frequent Itemset Mining
Hierarchical Clustering
K-Medoids
Lowongan Kerja ICT

ABSTRAK

Kesenjangan kompetensi *Information and Communication Technology* (ICT) dan minimnya analisis kombinasi keterampilan (*bundle skills*) mendorong penelitian mengidentifikasi pola keterampilan lowongan kerja ICT menggunakan *clustering* berbasis *binary hard skill presence matrix* dengan Jaccard distance; *soft skill* sebagai atribut interpretasi. Data diperoleh dari integrasi delapan *dataset* Kaggle: 44.709 data setelah data *selection*, 29.740 setelah *preprocessing* dan ekstraksi *skill*, sampel akhir 15.000 data. Ekstraksi menghasilkan 101 *hard skill* (reduksi ke 45 *via frequency threshold*) dan 17 *soft skill*. K-Medoids dan *Hierarchical Clustering* (AGNES) dievaluasi menggunakan *Silhouette Score* dan *Cluster Balance Ratio* (CBR). Meski AGNES k=2 memiliki *silhouette* lebih tinggi (0,1356 vs 0,0861), distribusinya timpang (CBR=0,0019) sehingga tidak representatif; K-Medoids k=6 (CBR=0,5059) dipilih sebagai model akhir, meski keduanya di bawah ambang 0,25 sebagai batasan penelitian. *Frequent Itemset Mining* (Apriori) menunjukkan asosiasi kuat pada lima dari enam *cluster*, *lift* tertinggi pada kombinasi *hadoop, java, scala*, dan *spark* (659,76) pada *cluster Big Data*. Keenam *cluster* dipetakan ke domain WEF *Future of Jobs* 2025: *Big Data Specialists, Full Stack Engineers, Data Analysts and Scientists, Software and Applications Developers, Database and Network Professionals,* dan *AI and ML Specialists*. Pola *lift* menunjukkan dua orientasi kebutuhan kompetensi ICT: penguasaan mendalam pada ekosistem teknologi tertentu, dan penguasaan data generalis lintas konteks.

Keyword:

Bundle Skills
Frequent Itemset Mining
Hierarchical Clustering
K-Medoids
ICT Job Postings

ABSTRACT

This study addresses the ICT competency gap and limited bundle-skill analysis by identifying skill patterns in ICT job postings using Jaccard-distance clustering on a binary hard skill matrix, with soft skills as an interpretive attribute. Integrating eight Kaggle datasets yielded 44,709 records after selection, 29,740 after preprocessing and extraction, and a final sample of 15,000 records comprising 101 hard skills (reduced to 45 via frequency threshold) and 17 soft skills. K-Medoids and Hierarchical Clustering (AGNES) were compared using Silhouette Score and Cluster Balance Ratio (CBR). Although AGNES (k=2) had a higher silhouette (0.1356 vs 0.0861), its distribution was highly imbalanced (CBR=0.0019); K-Medoids (k=6, CBR=0.5059) was thus selected, though both values fell below the 0.25 threshold, a study limitation. Frequent Itemset Mining (Apriori) revealed strong associations in five of six clusters, with the highest lift (659.76) for hadoop, java, scala, and spark in the Big Data cluster. Clusters were mapped to WEF Future of Jobs 2025 domains: Big Data Specialists, Full Stack Engineers, Data Analysts and Scientists, Software and Applications Developers, Database and Network Professionals, and AI and ML Specialists. The lift patterns indicate two orientations of ICT competency: deep mastery of specific technology ecosystems, and generalist data competency across contexts.

Copyright © SENASIF 2026
ISSN: 2598-0076

I. PENDAHULUAN

Transformasi digital telah membawa perubahan besar pada struktur pasar tenaga kerja, terutama karena organisasi semakin banyak menggunakan teknologi dan otomatisasi dalam berbagai proses kerja. Pada pekerjaan *Information and Communication Technology* (ICT), perubahan tersebut dapat dilihat dari keterampilan yang mulai banyak dicantumkan dalam lowongan kerja. *The Future of Jobs Report* yang diterbitkan oleh *World Economic Forum* memproyeksikan bahwa profesi seperti *software and application developers, data analysts,* serta *AI and machine learning specialists* termasuk dalam kelompok pekerjaan yang mengalami pertumbuhan paling cepat [1]. Hal tersebut sejalan dengan temuan *McKinsey Global Institute* yang menunjukkan bahwa penggunaan kecerdasan buatan dapat meningkatkan produktivitas sekaligus mendorong kebutuhan terhadap tenaga kerja dengan keterampilan tinggi, terutama pada bidang STEM [2]. Namun, pertumbuhan kebutuhan tenaga kerja tersebut tidak terjadi secara merata. Perkembangan teknologi juga mengubah jenis pekerjaan dan keterampilan yang dibutuhkan, sehingga dampaknya bukan hanya berupa berkurangnya pekerjaan, tetapi juga perubahan pada komposisi keterampilan di pasar kerja.

Kondisi tersebut menunjukkan bahwa pertumbuhan pekerjaan berbasis teknologi tidak selalu diikuti oleh penyerapan tenaga kerja yang seragam. Selain jumlah tenaga kerja, relevansi keterampilan terhadap kebutuhan pekerjaan menjadi aspek penting dalam memahami dinamika pasar kerja digital. [3]. *Data Federal Reserve Bank of New York* menunjukkan bahwa tingkat pengangguran lulusan Ilmu Komputer dan Teknik Komputer di Amerika Serikat masing-masing mencapai 6,1% dan 7,5% [4]. Kondisi ini menunjukkan bahwa permasalahan tenaga kerja ICT tidak hanya berkaitan dengan jumlah tenaga kerja, tetapi juga dengan kesesuaian keterampilan yang dimiliki dengan kebutuhan pekerjaan. Perubahan struktur pekerjaan akibat digitalisasi membuat perusahaan semakin membutuhkan tenaga kerja dengan beberapa keterampilan yang dapat digunakan secara bersamaan [5]. Dengan demikian, kebutuhan industri tidak lagi hanya berfokus pada satu keterampilan tertentu, tetapi pada kombinasi keterampilan yang dapat mendukung pekerjaan secara lebih menyeluruh.

Kebutuhan keterampilan ICT tersebut pada dasarnya saling berhubungan dan dapat membentuk pola kombinasi tertentu. Sebagai contoh, pekerjaan *Data Analyst* tidak hanya membutuhkan kemampuan SQL, tetapi juga dapat membutuhkan Python dan Tableau dalam satu posisi pekerjaan. Kombinasi seperti ini dikenal sebagai *bundle skills*. Izzatul Mula dan Ristiani dalam studinya menunjukkan bahwa perkembangan teknologi membuat kebutuhan keterampilan kerja semakin bersifat multidisipliner [6]. Liu et al. juga menemukan bahwa keterampilan dalam pasar kerja digital memiliki pola kemunculan bersama atau *skill co-occurrence* yang dapat menggambarkan hubungan kebutuhan kompetensi antarprofesi [7]. Selain itu, Harahap dan Santoso menunjukkan bahwa data lowongan kerja dapat digunakan untuk menemukan pola tertentu dalam kebutuhan tenaga kerja digital [8]. Temuan tersebut menunjukkan bahwa hubungan antar

Copyright © SENASIF 2026
ISSN: 2598-0076

keterampilan perlu diperhatikan dalam menganalisis kebutuhan tenaga kerja. Salah satu pendekatan yang dapat digunakan adalah *clustering*, karena metode ini dapat mengelompokkan lowongan kerja berdasarkan kemiripan keterampilan sehingga pola *bundle skills* dapat dilihat dengan lebih jelas pada masing-masing kelompok.

Meskipun demikian, penelitian sebelumnya masih memiliki beberapa keterbatasan. Harahap dan Santoso [8] lebih berfokus pada struktur pasar kerja serta hubungan antarposisi melalui *Social Network Analysis* dan belum secara khusus menganalisis kombinasi keterampilan. Wahyudi et al. telah menggunakan K-Medoids untuk mengelompokkan lowongan kerja, tetapi penggunaan semantic embedding dengan Sentence Transformer membuat keterampilan individual tidak terlihat secara langsung sehingga lebih sulit untuk mengetahui keterampilan apa saja yang menjadi ciri dari setiap *cluster* [9]. Sementara itu, Liu et al. [7] telah mengidentifikasi pola *skill co-occurrence* dalam jumlah data yang besar, tetapi belum menggabungkannya dengan proses *clustering* untuk membentuk kelompok pekerjaan berdasarkan kemiripan keterampilan. Selain itu, kekuatan hubungan antar keterampilan pada masing-masing kelompok juga belum dianalisis secara khusus.

Berdasarkan celah tersebut, penelitian ini menggunakan pendekatan *clustering* dengan *binary hard skill presence matrix* dan *Jaccard distance* untuk mengelompokkan lowongan kerja ICT berdasarkan kemiripan *hard skill* yang dimiliki. Dua algoritma *clustering*, yaitu K-Medoids dan *Hierarchical Clustering* (AGNES), dibandingkan untuk melihat algoritma yang mampu menghasilkan segmentasi paling representatif. Setelah *cluster* terbentuk, analisis dilanjutkan menggunakan *Frequent Itemset Mining* (*Apriori*) dengan *support* dan *lift* untuk menemukan kombinasi *hard skill* yang sering muncul sekaligus melihat kekuatan hubungan antar keterampilan dalam setiap *cluster*. Dengan pendekatan tersebut, penelitian ini tidak hanya melihat keterampilan yang muncul secara bersamaan, tetapi juga mencoba melihat keterampilan mana yang memiliki hubungan lebih kuat dalam kelompok lowongan tertentu. Data yang digunakan berasal dari delapan dataset publik ICT pada platform *Kaggle*. Oleh karena itu, hasil penelitian ini hanya menggambarkan pola yang terdapat pada dataset yang dianalisis dan tidak dapat digunakan untuk menyatakan kondisi pasar kerja ICT Indonesia secara keseluruhan.

II. METODE

Penelitian ini menggunakan pendekatan analisis data sekunder berbasis data mining untuk mengidentifikasi pola keterampilan pada lowongan kerja *Information and Communication Technology* (ICT). *Tools* utama dalam penelitian ini menggunakan bahasa pemrograman python, untuk proses pengolahan data dilakukan di lingkungan kerja *Google Colaboratory* (Google Colab) dengan versi python yang tersedia pada *platform* (versi 3.12). Variabel utama berupa *hard skill* yang direpresentasikan dalam bentuk biner melalui *binary hard skill presence matrix*, dengan nilai 1 menunjukkan bahwa suatu keterampilan terdapat dalam lowongan dan nilai 0 menunjukkan bahwa

Copyright © SENASIF 2026

ISSN: 2598-0076

keterampilan tidak ditemukan. *Hard skill* digunakan sebagai variabel pembentuk *cluster*, Selain *hard skill*, penelitian juga mengidentifikasi *soft skill* sebagai atribut interpretatif. *Soft skill* tidak digunakan dalam pembentukan *cluster* karena sebagian keterampilan tersebut muncul pada proporsi tinggi di berbagai lowongan sehingga kemampuan pembeda antar*cluster* menjadi rendah. Dengan demikian, *hard skill* berfungsi sebagai variabel pembentuk *cluster*, sedangkan *soft skill* digunakan untuk membantu interpretasi karakteristik kompetensi yang terbentuk.

Data diperoleh melalui teknik dokumentasi dari delapan *dataset* publik lowongan kerja ICT di Kaggle, kemudian digabungkan menjadi satu *dataset* menggunakan *Power Query Excel*. Dari 44.742 lowongan awal, setelah penghapusan deskripsi yang tidak memadai dan duplikat diperoleh 44.709 lowongan, dengan atribut *title* dan *description* sebagai sumber identifikasi keterampilan.

Tahap pengolahan data diawali dengan *preprocessing* teks yang mencakup penghapusan URL, alamat e-mail, karakter khusus, dan spasi berlebih, kemudian dilanjutkan dengan *case folding*. Ekstraksi keterampilan dilakukan menggunakan *dictionary mapping* dengan kamus *hard skill* dan *soft skill* yang disusun secara manual dengan menggunakan struktur klasifikasi *European Skills, Competences, Qualifications and Occupations* (ESCO) sebagai referensi konseptual. Proses ini digunakan untuk memetakan variasi penulisan keterampilan ke dalam *canonical skill* agar setiap keterampilan yang memiliki makna sama diperlakukan sebagai atribut yang seragam. Setelah proses ekstraksi dan normalisasi, diperoleh 101 *canonical hard skill* dan 17 *canonical soft skill*. Lowongan yang memiliki kurang dari dua *hard skill* terkait ICT dikeluarkan dari tahap pemodelan sehingga diperoleh 29.740 lowongan kerja. Selanjutnya diterapkan teknik *frequency threshold* diterapkan untuk mengurangi dimensi dan sparsity dengan mempertahankan *hard skill* yang memiliki frekuensi kemunculan tertinggi. Pendekatan penyaringan berdasarkan frekuensi ini mengacu pada prinsip *Frequent Itemset Mining* yang diperkenalkan oleh Agrawal et al. [10], sehingga 45 *hard skill* dengan frekuensi kemunculan tertinggi dipertahankan untuk analisis *clustering*.

Populasi analisis terdiri atas 29.740 lowongan kerja yang telah memenuhi kriteria minimal *dua hard skill* ICT. Sebanyak 15.000 lowongan digunakan sebagai sampel pada tahap pengukuran jarak dan *clustering*. Pengambilan sampel tersebut dilakukan untuk mengurangi beban komputasi karena perhitungan *Jaccard distance* dilakukan secara berpasangan atau pairwise dan meningkat secara kuadratik terhadap jumlah data. Sampel akhir menghasilkan *hard binary skill matrix* berukuran 15.000 × 45 sebagai input pengukuran Jaccard distance dan *clustering*, sedangkan *soft binary skill matrix* berukuran 15.000 × 17 disimpan sebagai atribut interpretatif dan tidak digunakan dalam pembentukan *cluster*.

Pengukuran kemiripan antar lowongan kerja menggunakan Jaccard *distance*, dipilih karena data *hard skill* bersifat biner dan sparse sehingga metrik berbasis selisih nilai berisiko menimbulkan *double-zero problem*, yaitu ketidakhadiran bersama suatu atribut yang keliru diperlakukan sebagai indikasi kemiripan. Jaccard *distance* antara dua objek A dan B dinyatakan sebagai:

Copyright © SENASIF 2026
ISSN: 2598-0076

$$d_j(A,B)1 - \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

dengan $|A \cap B|$ adalah *Intersection,* jumlah *hard skill* yang bernilai 1 pada kedua objek secara bersamaan dan $|A \cup B|$ adalah *Union,* jumlah *hard skill* yang bernilai 1 pada salah satu atau kedua objek. Nilai berkisar 0 (identik) hingga 1 (tidak ada kesamaan).Hasil pengukuran digunakan untuk membentuk *distance matrix* sebagai input pada proses *clustering*.

Proses *clustering* dilakukan menggunakan K-Medoids dan *Hierarchical Clustering* dengan pendekatan *Agglomerative Nesting* (AGNES). K-Medoids diimplementasikan menggunakan pendekatan *Partitioning Around Medoids* (PAM). Dalam penentuan jumlah *cluster* digunakan konfigurasi *Cluster* Validity Index (CVI), yaitu ukuran internal yang mengevaluasi kohesi intra-*cluster* dan separasi antar-*cluster* berdasarkan struktur data itu sendiri tanpa memerlukan label eksternal [11]. Pendekatan umum yang digunakan dalam CVI adalah dengan mengevaluasi rentang kandidat nilai k secara bertahap (*Sweep k*). Pengujian dilakukan pada rentang jumlah *cluster* (k=2) hingga (k=8). AGNES menggunakan *average linkage* dan diuji pada rentang (k=2) hingga (k=14), rentang yang lebih luas karena struktur *dendrogram* AGNES memungkinkan evaluasi berbagai titik potong hierarki. Kedua algoritma menggunakan *distance matrix* hasil perhitungan Jaccard. Pengujian beberapa nilai (k) dilakukan untuk memperoleh konfigurasi yang memberikan hasil pengelompokan terbaik berdasarkan evaluasi internal.

Evaluasi hasil *clustering* menggunakan *Silhouette Score* dan *Cluster Balance Ratio* (CBR). *Silhouette Score* mengukur kohesi dan separasi *cluster* dengan interpretasi mengacu skema Kaufman dan Rousseeuw [12] : nilai di atas 0,70 menunjukkan struktur kuat, 0,51 – 0,70 struktur normal, 0,26 – 0,50 struktur lemah, dan ≤0,25 tidak terdapat struktur substansial secara geometris. Konfigurasi *k* dengan *Silhouette Score* tertinggi dipilih sebagai kandidat terbaik per algoritma. Karena nilai *silhouette* dapat meningkat secara artifisial ketika terbentuk *cluster* kecil yang terisolasi dari mayoritas data yaitu kondisi yang umum pada data biner sparse berdimensi tinggi sesuai dengan studi Pavlopaulus et al. [13], maka dari itu solusi dari kondisi tersebut menggunakan evaluasi multi kriteria yaitu menilai keseimbangan distribusi anggota *cluster Cluster Balance Ratio* (CBR) dengan rumus:

$$CBR = n_min / n_max \quad (2)$$

Yaitu rasio jumlah anggota *cluster* terkecil terhadap jumlah anggota *cluster* terbesar. Nilai CBR yang mendekati 1 menunjukkan distribusi yang lebih merata, sedangkan nilai yang mendekati 0 menunjukkan ketimpangan distribusi yang tinggi. Apabila *Silhouette Score* dan CBR menghasilkan keputusan yang berbeda antar-algoritma, representativitas distribusi anggota diprioritaskan sebagai kriteria dominan, karena *cluster* dengan distribusi sangat timpang tidak memberikan dasar yang memadai untuk analisis pola keterampilan.

Pada model *clustering* terpilih, dilakukan identifikasi *hard skill* dominan berdasarkan proporsi kemunculan per *cluster*, dilanjutkan analisis *co-occurrence* pasangan *hard skill* yang sering muncul bersamaan. Analisis ini diperdalam menggunakan *Frequent Itemset Mining* melalui

Copyright © SENASIF 2026

ISSN: 2598-0076

algoritma *Apriori* dengan *minimum support* 0,15 dan panjang *itemset* maksimum 5 *skill* per *bundle*, dijalankan terpisah pada tiap *cluster* menggunakan pustaka *mlxtend*. Nilai ambang tersebut mempertimbangkan karakteristik data biner berdimensi tinggi dan sparse, mengikuti temuan bahwa ambang minimum *support* yang terlalu tinggi berisiko menghasilkan itemset yang tidak representatif, sedangkan ambang terlalu rendah menghasilkan itemset berlebihan dan kurang bermakna [14]. *Support* suatu itemset X dirumuskan sebagai:

$$support(X) = frekuensi\ kemunculan\ X\ /\ total\ transaksi \quad (3)$$

Sedangkan *lift* pasangan item X dan Y dirumuskan sebagai:

$$lift(X,Y) = support(X \cup Y)\ /\ (support(X) \times support(Y) \quad (4)$$

Nilai *lift* = 1 menunjukkan kemunculan X dan Y independen, *lift* > 1 menunjukkan asosiasi positif yang semakin kuat seiring meningkatnya nilai, dan *lift* < 1 menunjukkan asosiasi negatif [15]. *Support* mengukur seberapa dominan suatu kombinasi *hard skill* dalam *cluster*, sedangkan *lift* mengukur seberapa jauh kombinasi tersebut menyimpang dari asumsi independensi. Kedua metrik digunakan bersamaan karena frekuensi kemunculan tinggi tidak selalu mencerminkan asosiasi yang kuat.

Interpretasi kecenderungan bidang pekerjaan ICT dilakukan dengan memetakan *signature skill* setiap *cluster* ke domain profesi yang diadaptasi dari *World Economic Forum Future of Jobs Report* 2025 [1] yaitu ditampilkan pada Tabel 1 Adaptasi Domain Profesi ICT Berdasarkan World Economic Forum.

Tabel 1. Adaptasi Domain Profesi ICT Berdasarkan World Economic Forum

Domain Profesi WEF (World Economic Forum)	Domain Profesi Penelitian
Big Data Specialist	Big Data Specialist
AI and Machine Learning Specialists	AI and Machine Learning Specialist
Software Application Developer	#1 Software Application Developer
Software Application Developer	#2 Mobile Developers
Security Management Specialists	Security Management Specialist
Data Warehousing Specialists	Data Warehousing Specialists
Data Analysts and Scientists	# 1 Data Analysts and Scientists (BI)
Data Analysts and Scientists	#2 Data Analysts and Scientists (Statistical)
Full Stack Engineers	Full Stack Engineers
Database and Network Professionals	Database and Network Professionals

Interpretasi ini dimaksudkan untuk menggambarkan kecenderungan domain profesi pada tingkat *cluster* dan bukan klasifikasi profesi formal terhadap lowongan kerja maupun individu tertentu.

III. HASIL DAN PEMBAHASAN

Karakteristik Data dan Hasil Transformasi

Tabel 2 merangkum tahapan pengolahan data penelitian. Dari 44.709 lowongan kerja hasil seleksi dan pembersihan, proses ekstraksi *skill* berbasis *dictionary mapping* menghasilkan 29.740 lowongan kerja dengan minimal dua *hard skill* ICT teridentifikasi. *Frequency threshold* mereduksi

Copyright © SENASIF 2026
ISSN: 2598-0076

101 *hard skill canonical* menjadi 45 *hard skill* dengan frekuensi tertinggi, sedangkan *soft skill* tetap dipertahankan sebanyak 17 *canonical* sebagai atribut interpretatif. *Stratified sampling* menghasilkan 15.000 data (50,44% dari 29.740) yang digunakan pada tahap *distance measurement* dan *clustering*, menghasilkan *hard binary skill matrix* berukuran 15.000 × 45 dan *soft binary skill matrix* berukuran 15.000 × 17.

Tabel 2. Rincian Dataset Penelitian

Keterangan	Jumlah
Data selection dan proses data cleaning, remove duplicate	44.709
Data Preprocessing + Data Extraction *hard* & *soft*	29.740
Jumlah data setelah sampling (50,44%)	15.000
Jumlah *Hard skill* Canonical awal 101, dilakukan frequency threshold	45
Jumlah *Soft skill* Canonical	17
Dimensi *hard* binary *skill* matrix	15.000 × 45
Dimensi *soft* binary *skill* matrix	15.000 × 17

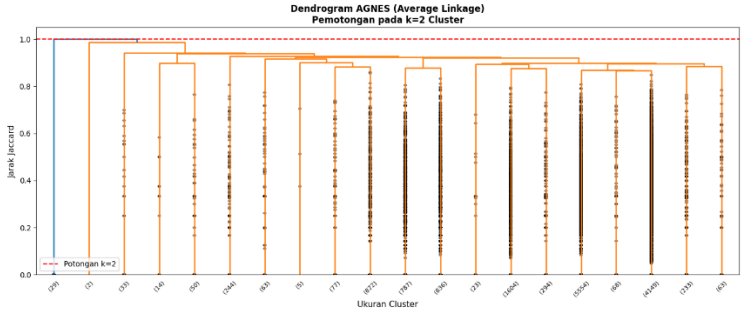
Hasil *Clustering* dan Pemilihan Model Akhir

Sweep k pada K-Medoids (rentang 2–8) menghasilkan *Silhouette Score* tertinggi pada k=6 sebesar 0,0861, sedangkan *sweep k* pada AGNES *average linkage* (rentang 2–14) menghasilkan *Silhouette Score* tertinggi pada k=2 sebesar 0,1356. Tabel 3 menyajikan perbandingan kedua konfigurasi terbaik tersebut beserta distribusi anggota dan *Cluster Balance Ratio*.

Tabel 3. Evaluasi Hasil Silhouette Score dengan Cluster Balance Ratio

Algoritma	k	Silhouette Score	Distribusi Anggota	*Cluster Balance Ratio*
Hierarchical (AGNES)	2	0.1356	C1: 29 (0.2%), C2: 14.971 (99.8%)	0.0019
K-Medoids (PAM)	6	0.0861	C1: 2.524 (16.8%), C2: 2.050 (13.7%), C3: 3.467 (23.1%) C4: 1.754 (11.7%) C5: 2.716 (18.1%) C6: 2.489 (16.6%)	0.5059

Nilai *silhouette* AGNES k=2 yang lebih tinggi tidak dapat langsung diartikan sebagai kualitas pemisahan yang lebih baik. Pemeriksaan *dendrogram* menunjukkan bahwa pemotongan pada k=2 menghasilkan satu *cluster* kecil berisi 29 lowongan kerja dan satu *cluster* besar berisi 14.971 lowongan kerja (CBR=0,0019), pola yang konsisten dengan kecenderungan *silhouette* meningkat secara artifisial ketika *cluster* kecil terisolasi dari mayoritas data pada representasi biner sparse berdimensi tinggi [13], dapat dilihat pada gambar 1 yang menyajikan *Dendogram* AGNES.



Gambar 1. Dendogram Hierarchical Clustering (AGNES)

Copyright © SENASIF 2026

ISSN: 2598-0076

Sebaliknya, K-Medoids k=6 menghasilkan distribusi yang jauh lebih proporsional (CBR=0,5059). Kedua nilai *silhouette* 0,0861 pada K-Medoids dan 0,1356 pada AGNES berada di bawah ambang 0,25 sehingga menurut skema Kaufman dan Rousseeuw [12] tergolong tidak menunjukkan struktur *cluster* yang substansial secara geometris. Kondisi ini dilaporkan secara terbuka sebagai batasan validitas geometris penelitian, *cluster* tetap digunakan untuk tujuan interpretasi pola kompetensi sesuai prosedur evaluasi multi-kriteria yang telah ditetapkan.

Mengikuti kriteria bahwa representativitas distribusi anggota diprioritaskan ketika kedua metrik berbeda arah, K-Medoids k=6 ditetapkan sebagai model *clustering* akhir dengan distribusi anggota yang tersebar secara merata tanpa dominasi satu *cluster* tertentu ditunjukan pada Tabel 4 sebagai model *cluster* yang terpilih.

Tabel 4. Hasil Keputusan Evaluasi Pemilihan Clustering

Cluster	**Jumlah Lowongan Kerja**	***Silhouette* Mean per *Cluster***
1	C1: 2.524 (16.8%),	0.0602
2	C2: 2.050 (13.7%),	0.0582
3	C3: 3.467 (23.1%)	0.1588
4	C4: 1.754 (11.7%)	0.0722
5	C5: 2.716 (18.1%)	0.0988
6	C6: 2.489 (16.6%)	0.0303
K=6	**15.000 / 100%**	***Mean Silhouette*: 0.0861**

Hasil Bundle Skill Frequent Itemsets Mining

K-Medoids k=6 ditetapkan sebagai model *clustering* akhir, analisis dilanjutkan untuk mengidentifikasi kombinasi *hard skill* yang muncul secara bersamaan pada masing-masing *cluster*. Analisis *co-occurrence* digunakan sebagai tahap awal untuk melihat keterampilan yang sering muncul bersama dalam satu lowongan kerja. Namun, frekuensi kemunculan bersama belum menunjukkan kekuatan asosiasi antar-keterampilan karena keterampilan yang memiliki frekuensi individual tinggi dapat menghasilkan frekuensi *co-occurrence* yang tinggi pula. Oleh karena itu, analisis dilanjutkan menggunakan *Frequent Itemset Mining* dengan algoritma *Apriori* untuk mengidentifikasi kombinasi *hard skill* berdasarkan nilai *support* dan mengukur kekuatan asosiasinya menggunakan *lift*.

Support menunjukkan proporsi lowongan kerja dalam suatu *cluster* yang memuat seluruh keterampilan dalam satu *itemset*, sedangkan *lift* menunjukkan tingkat keterkaitan kemunculan bersama dibandingkan dengan kondisi yang diharapkan apabila keterampilan tersebut muncul secara independen. Dengan demikian, *support* digunakan untuk melihat seberapa sering suatu kombinasi ditemukan, sedangkan *lift* digunakan untuk menilai seberapa kuat keterkaitan kombinasi tersebut. *Hasil Frequent Itemset Mining* kemudian dirangkum dengan memilih tiga *bundle skill* dengan nilai *lift* tertinggi pada masing-masing *cluster*. Pemilihan tersebut memungkinkan pola asosiasi keterampilan yang paling kuat pada setiap *cluster* dibandingkan secara langsung tanpa menampilkan seluruh *itemset* yang dihasilkan oleh algoritma *Apriori*. Hasil analisis disajikan pada Tabel 5.

Tabel 5. Hasil Frequent Itemset (Support dan Lift) Hard skill Per Cluster

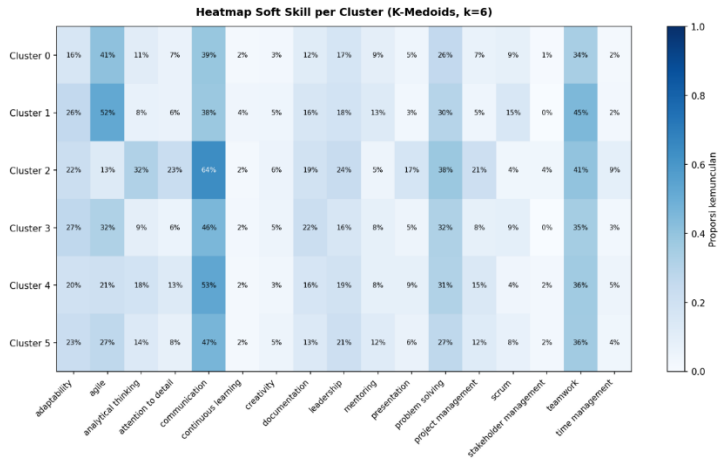
Copyright © SENASIF 2026
ISSN: 2598-0076

Cluster	*Bundle skill*	Jumlah *Skill*	*Support* (%)	*Lift*
C1	hadoop + java + scala + spark	4	15.3%	659.76
C1	hadoop + python + scala + spark	4	18.4%	313.48
C1	hadoop + kafka + python + spark	4	16.2%	298.53
C2	css + html + javascript	3	19.4%	510.30
C2	aws + docker + kubernetes + python	4	16.9%	160.58
C2	aws + docker + kubernetes	3	22.2%	115.58
C3	power bi + python + sql + tableau	4	19.3%	16.64
C3	power bi + sql + tableau	3	29.8%	14.03
C3	excel + sql + tableau	3	22.2%	11.45
C4	c++ + linux + python	3	16.2%	26.51
C4	c++ + linux	2	26.2%	23.31
C4	git + linux	2	17.6%	15.02
C5	python + sql	2	55.7%	1.75
C5	etl + sql	2	16.0%	1.61
C5	machine learning + python	2	16.1%	1.10
C6	artificial intelligence + machine learning + python	3	20.6%	18.18
C6	artificial intelligence + machine learning	2	31.0%	14.92
C6	machine learning + nlp	2	16.9%	11.59

Tabel 5 menunjukan pola asosiasi *hard skill* yang kuat pada lima *cluster*, dengan nilai *lift* tertinggi pada kombinasi $hadoop + java + scala + spark$ di C1 (659,76), $css + html + javascript$ di C2 (510,30), $c++ + linux + python$ di C4 (26,51), $artificial\ intelligence + machine\ learning + python$ di C6 (18,18), dan $power\ bi + python + sql + tableau$ di C3 (16,64). Sebaliknya, C5 menunjukkan pola asosiasi lemah dengan nilai *lift* mendekati 1 pada seluruh kombinasinya dan proporsi *support* nya setengah menguasai *cluster*, menunjukkan bahwa kombinasi atribut tersebut sering muncul dalam *cluster*, namun kekuatan asosiasinya relatif lebih rendah dibandingkan kombinasi pada *cluster* lainnya dan menjadi lintas *skill python* dan *sql*.

Hasil Pemetaan Soft Skill

Pemetaan *soft skill* dilakukan untuk memberikan gambaran mengenai kompetensi nonteknis yang muncul pada masing-masing kelompok kebutuhan *skill* ICT. Hasil pemetaan disajikan dalam bentuk heatmap dan tabel proporsi kemunculan *soft skill* pada setiap *cluster*.



Gambar 2. Visual Heatmap Soft Skill Per Cluster

Gambar 2 menyajikan visualisasi distribusi *soft skill* pada masing-masing *cluster* berdasarkan proporsi kemunculannya. Perbedaan intensitas warna menunjukan variasi tingkat

Copyright © SENASIF 2026
ISSN: 2598-0076

kemunculan *soft skill* antar *cluster*. Hasil pemetaan selanjutnya disajikan pada tabel 6 dalam top *soft skill* dominan per *cluster*.

Tabel 6. Proporsi Top-5 Soft Skill Per Cluster

Cluster	**Top-5 *Soft skill***
C1	Agile (41%), communication (39%), teamwork (34%), problem solving (26%), leadership (16%)
C2	Agile (52%), Teamwork (45%), Communication (38%), Problem solving (29%), Adaptability (26%)
C3	communication (64%), teamwork (40%), problem solving (38%), analytical thinking (31%), leadership (23%)
C4	communication (46%), teamwork (35%), problem solving (32%), agile (32%), adaptability (26%)
C5	communication (53%), teamwork (36%), problem solving (31%), agile (21%), adaptability (19%)
C6	communication (46%), teamwork (36%), problem solving (27%), agile (27%), adaptability (22%)

Soft skill yang memiliki proporsi kemunculan tertinggi pada masing-masing *cluster* digunakan untuk melengkapi karakteristik *cluster* yang telah diidentifikasi berdasarkan *hard skill* per *cluster*.

Karakteristik Enam Cluster

Model K-Medoids k=6 menghasilkan enam *cluster* dengan *signature skill, bundle skill* dominan, dan *soft skill* pendamping, sebagaimana dirangkum pada Tabel 7 dan diinterpretasikan terhadap domain profesi WEF *Future of Jobs* 2025.

Tabel 7. Pemetaan Karakteristik Cluster Berdasarkan Signature skill, Bundle skill, Soft skill, dan Domain Profesi ICT

Cluster	***Signature skill***	***Bundle skill* Dominan**	***Soft skill* Dominan**	**Domain Profesi ICT (WEF 2025)**
C1	Scala, Hadoop, Spark, Redshift, Kafka, Airflow, Snowflake, ETL	hadoop + java + scala + spark (*lift* 659.76)	Agile (41%) Communication (39%)	Big Data Specialists
C2	Docker, Kubernetes, TypeScript, CSS, React, JavaScript, HTML, CI/CD	css + html + javascript (*lift* 510,30)	Agile (52%) Teamwork (45%)	Full Stack Engineers
C3	Excel, Power BI, Tableau, SAS, SQL, SQL Server, Oracle, ETL	power bi + python + sql + tableau (*lift* 16,64)	Communication (64%) Analytical Thinking (31%)	Data Analysts and Scientists (BI)
C4	C++, Linux, Git, Unix, Bash, C#, Java, Deep Learning	c++ + linux + python (*lift* 26,51)	Communication (46%) Problem Solving (32%)	*Soft*ware and Applications Developers
C5	SAS, SQL, Oracle, Python, ETL, SQL Server, .NET, Snowflake	python + sql (*lift* 1,75)	Communication (53%) Teamwork (36%)	Database and Network Professionals
C6	Artificial Intelligence, NLP, Deep Learning, Machine Learning, TensorFlow, PyTorch, Databricks, Azure	ai + ml + python (*lift* 18,18)	Communication (46%) Problem Solving (27%)	AI and ML Specialists

Copyright © SENASIF 2026
ISSN: 2598-0076

Cluster 1 (2.524 lowongan, 16,8%) memiliki *signature skill* Scala, Hadoop, Spark, Redshift, Kafka, Airflow, dan Snowflake yang mencerminkan satu rangkaian kerja pemrosesan data skala besar, data masuk melalui Kafka, diproses Spark dan Hadoop, disimpan pada Redshift atau Snowflake dan didukung kombinasi $hadoop + java + scala + spark$ dengan kekuatan *lift* 659,76. Dominasi *soft skill* Agile (41%) dan *Communication* (39%) selaras dengan koordinasi lintas tim yang dituntut pipeline data berskala besar. Karakteristik ini konsisten dengan profil *Big Data Specialists*, yang pada laporan WEF diproyeksikan sebagai profesi dengan pertumbuhan tercepat periode 2025–2030.

Cluster 2 (2.050 lowongan, 13,7%) memiliki *signature skill* Docker, Kubernetes, TypeScript, CSS, React, JavaScript, HTML, dan CI/CD, mencerminkan integrasi pengembangan *front-end* dengan infrastruktur berbasis *container*. Kombinasi $css + html + javascript$ dengan kekuatan *lift* 510,30 dan $aws + docker + kubernetes$ ber*lift* di atas 100 (Tabel 5), memperkuat keterikatan antara keterampilan *front-end, cloud*, dan *containerization*. *Soft skill* Agile (52%) dan *Teamwork* (45%) selaras dengan siklus rilis iteratif berbasis CI/CD. *Cluster* ini konsisten dengan profil *Full Stack Engineers.*

Cluster 3 (3.467 lowongan, 23,1%) adalah *cluster* terbesar anggotannya dengan *signature skill* Excel, Power BI, Tableau, SAS, SQL, SQL Server, Oracle, dan ETL. Kombinasi $power\ bi + sql + tableau$ dengan kekuatan *lift* 14,03 dan kehadiran SQL yang konsisten pada seluruh *bundle* utama menunjukkan SQL sebagai elemen penghubung analisis dan visualisasi data. *Soft skill Communication* (64%, tertinggi di seluruh *cluster*) dan *Analytical Thinking* (31%) selaras dengan kebutuhan menerjemahkan hasil analisis kepada pemangku kepentingan non-teknis. *Cluster* ini konsisten dengan profil *Data Analysts and Scientists*.

Cluster 4 (1.754 lowongan, 11,7%) memiliki *signature skill* C++, Linux, Git, Unix, Bash, C#, dan Java. Kombinasi $c++ + linux + python$ dengan kekuatan *lift* 26,51 dan linux hadir pada seluruh *bundle* utama dengan *lift* konsisten di atas 15 (Tabel 5), menandakan Linux sebagai penghubung utama antarketerampilan sistem. *Soft skill Communication* (46%) dan *Problem Solving* (32%) selaras dengan tuntutan penyelesaian masalah teknis pada pengembangan perangkat lunak tingkat sistem. *Cluster* ini konsisten dengan profil *Software and Applications Developers.*

Cluster 5 (2.716 lowongan, 18,1%) memiliki *signature skill* SAS, SQL, Oracle, Python, ETL, SQL Server, .NET, dan Snowflake, namun berbeda dari kelima *cluster* lain karena seluruh kombinasi *skill*-nya menghasilkan $python + sql$ mencatatkan *support* tertinggi seluruh *cluster* (55,7%) tetapi *lift* hanya 1,75. Pola ini mengindikasikan Python dan SQL sebagai kompetensi data umum yang tidak eksklusif terhadap satu domain, dan *cluster* ini merepresentasikan segmen operasional generalis (pengelolaan basis data dan integrasi data) alih-alih spesialisasi teknologi tertentu. *Soft skill Communication* (53%) dan *Teamwork* (36%) selaras dengan koordinasi rutin lintas fungsi. *Cluster* ini konsisten dengan profil *Database and Network Professionals*.

Copyright © SENASIF 2026

ISSN: 2598-0076

Cluster 6 (2.489 lowongan, 16,6%) memiliki *signature skill* Artificial Intelligence, NLP, Deep Learning, Machine Learning, TensorFlow, PyTorch, Databricks, dan Azure. Kombinasi $artificial\ intelligence + machine\ learning + python$ dengan kekuatan *lift* 18,18. *Soft skill Communication* (46%) dan *Problem Solving* (27%) selaras dengan kebutuhan menyampaikan hasil eksperimen model kepada pihak non-teknis. *Cluster* ini konsisten dengan profil *AI and ML Specialists*.

Interpretasi Pola Lintas Cluster

Berdasarkan pola *support* dan *lift* pada Tabel 5, temuan ini menunjukkan dua orientasi kebutuhan kompetensi ICT pada *dataset* yang dianalisis. Orientasi pertama adalah spesialisasi mendalam pada ekosistem teknologi tertentu, tercermin pada *Cluster* 1, 2, 4, dan 6 yang ditandai *lift* tinggi dan *signature skill* yang relatif eksklusif terhadap *platform* tertentu seperti Hadoop dan Spark, Docker dan Kubernetes, C++ dan Linux, TensorFlow dan PyTorch. Orientasi kedua adalah penguasaan data generalis lintas konteks, tercermin pada *Cluster* 5 yang ditandai *lift* mendekati satu dan kombinasi Python-SQL yang bersifat umum. *Cluster* 3 berada pada posisi peralihan dengan *lift* kombinasi utamanya ada pada (14,03–16,64) tergolong moderat, namun kehadiran SQL yang konsisten pada seluruh *bundle* utamanya mencerminkan sifat fondasional yang sejalan dengan *Cluster* 5. Pembagian ini menunjukkan bahwa kebutuhan kompetensi ICT pada *dataset* yang dianalisis bukan satu spektrum spesialisasi tunggal, melainkan dua orientasi yang sama-sama muncul.

Posisi Temuan terhadap Penelitian Terdahulu

Penelitian Harahap dan Santoso menghasilkan tiga *cluster* menggunakan K-Means dan *Social Network Analysis* dengan *silhouette score* 0,617 [8], jauh lebih tinggi dari 0,0861 pada penelitian ini. Perbedaan ini perlu dibaca berimbang: penggunaan K-Medoids pada penelitian ini lebih sesuai secara konseptual dengan karakteristik data biner bersparsitas tinggi karena medoid merupakan objek aktual dan tidak bergantung pada nilai rata-rata sebagaimana K-Means, namun kesesuaian konseptual tersebut tidak serta-merta menghasilkan pemisahan *cluster* yang lebih kuat secara geometris. Dengan demikian, temuan ini tidak dapat digunakan untuk menyatakan keunggulan K-Medoids atas K-Means secara umum, dukungan terhadap kesesuaian representasi data bersifat konseptual, bukan empiris pada aspek validitas geometris.

Penelitian Wahyudi et al. [9] menghasilkan 13 *cluster* menggunakan *Sentence Transformer* dengan tingkat kesesuaian yang divalidasi praktisi HR sebesar 88%. Penelitian ini tidak melakukan validasi eksternal setara, sehingga tingkat kesesuaian kedua pendekatan tidak dapat dibandingkan langsung. Namun, penggunaan *binary skill presence matrix* memungkinkan identitas domain tiap *cluster* ditelusuri langsung melalui *skill* yang secara eksplisit hadir atau tidak hadir, berbeda dari representasi vektor kontinu yang mengaburkan keberadaan *skill* individual. Tanpa validasi eksternal,

Copyright © SENASIF 2026
ISSN: 2598-0076

interpretasi domain pada penelitian ini tetap merupakan interpretasi berbasis data dan belum setara dengan validasi praktisi.

Penelitian Liu et al. [7] mengidentifikasi pola *skill co-occurrence* pada skala besar tanpa mengintegrasikan segmentasi *cluster* maupun pengukuran kekuatan asosiasi ke dalamnya. Penelitian ini memperluas pendekatan tersebut dengan menambahkan *Frequent Itemset Mining* dan nilai *lift* pada tingkat *cluster*, menunjukkan bahwa *support* tinggi tidak selalu berarti asosiasi kuat sebagaimana ditunjukkan perbandingan *Cluster* 1 (*support* 15,3%, *lift* 659,76) dengan *Cluster* 5 (*support* 55,7%, *lift* 1,75).

Di antara tiga celah penelitian terdahulu yang menjadi dasar rancangan penelitian ini, penambahan ukuran kekuatan asosiasi di luar frekuensi kemunculan bersama merupakan celah yang paling kuat didukung temuan empiris, sehingga kontribusi utama penelitian ini terletak pada penggabungan segmentasi berbasis *cluster* dengan pengukuran kekuatan asosiasi keterampilan menggunakan *lift*.

IV. SIMPULAN

Bagi industri, pemetaan *bundle skill* per *cluster* memberikan gambaran kombinasi keterampilan yang secara konsisten muncul bersama pada dataset yang dianalisis, sebagai salah satu referensi analitik dalam mempertimbangkan pengembangan kompetensi tenaga kerja. Bagi pencari kerja, pola ini memberikan informasi mengenai kombinasi keterampilan yang relevan pada masing-masing kecenderungan domain, alih-alih hanya keterampilan tunggal.

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam membaca temuannya. Pertama, kedua nilai *silhouette* (K-Medoids 0,0861 dan AGNES 0,1356) berada di bawah ambang 0,25 sehingga tergolong tidak menunjukkan struktur *cluster* yang substansial secara geometris menurut skema Kaufman dan Rousseeuw [12]; keterbatasan ini diakui secara terbuka sebagai batasan validitas geometris, terpisah dari validitas interpretasi pola kompetensi yang dihasilkan. Kedua, data penelitian bersumber dari delapan dataset publik ICT pada platform Kaggle sehingga pola yang teridentifikasi menggambarkan karakteristik dataset yang dianalisis, bukan representasi pasar kerja ICT Indonesia maupun kebutuhan industri ICT secara menyeluruh. Ketiga, interpretasi domain profesi bersifat kecenderungan pada tingkat *cluster* berdasarkan pemetaan WEF *Future of Jobs* 2025, bukan klasifikasi profesi formal terhadap lowongan kerja maupun individu tertentu, dan belum melalui validasi eksternal oleh praktisi sebagaimana dilakukan pada penelitian sebelumnya dari Wahyudi et al [9].

Berdasarkan keterbatasan tersebut, penelitian selanjutnya disarankan melengkapi evaluasi *cluster* dengan metrik internal tambahan seperti Davies-Bouldin Index atau Calinski-Harabasz Index, melakukan analisis sensitivitas terhadap ambang minimum *support* dan *lift*, memperluas cakupan data dengan dataset lokal yang mencerminkan konteks pasar kerja ICT nasional, serta

Copyright © SENASIF 2026

ISSN: 2598-0076

mengembangkan analisis temporal untuk memproyeksikan arah perubahan kebutuhan keterampilan ICT ke depan.

DAFTAR RUJUKAN

[1] World Economic Forum, "The Future of Jobs Report." Accessed: Sep. 24, 2025. [Online]. Available: https://www.weforum.org/publications/the-future-of-jobs-report-2025/

[2] K. Ellingrud *et al.*, "Generative AI and the future of work in America | McKinsey." Accessed: Apr. 29, 2026. [Online]. Available: https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america#/

[3] F. Pusterla and U. Renold, "Does ICT affect the demand for vocationally educated workers?," *Swiss J. Econ. Stat.*, vol. 158, no. 1, p. 22, Dec. 2022, doi: 10.1186/s41937-022-00101-8.

[4] "Federal Reserve Bank of New York." Accessed: Sep. 28, 2025. [Online]. Available: https://www.newyorkfed.org/research/college-labor-market#--:explore:unemployment

[5] M. K. Omar, I. H. Mohd, O. Z. W. Zahari, and N. Azmy, "Hybrid *Skill*s Framework in The IR4.0 and Recruitment Decisions in Hiring Graduates: A Malaysian Employer's Perspective," *Int. J. Res. Innov. Soc. Sci.*, vol. VIII, no. VII, pp. 252–273, 2024, doi: 10.47772/IJRISS.2024.807021.

[6] Izzatul Mula and A. Ristiani, "Transformasi Struktur Pekerjaan dan Kebutuhan Keterampilan di Era Teknologi AI dan Otomatisasi di Pasar Global," *Nian Tana Sikka J. Ilm. Mhs.*, vol. 3, no. 1, pp. 155–167, Jan. 2025, doi: 10.59603/niantanasikka.v3i1.665.

[7] Z. Liu, J. M. Clarke, B. Rohenkohl, and M. Barahona, "Patterns of co-occurrent *skill*s in UK job adverts," *PLOS Complex Syst.*, vol. 2, no. 2, p. e0000028, Feb. 2025, doi: 10.1371/journal.pcsy.0000028.

[8] T. A. J. Harahap and C. B. Santoso, "ANALISIS TREN LOWONGAN PEKERJAAN *SOFT*WARE ENGINEERING DI INDONESIA DENGAN *CLUSTERING* DAN SOCIAL NETWORK ANALYSIS," *IDEALIS Indones. J. Inf. Syst.*, vol. 8, no. 2, pp. 160–171, Jul. 2025, doi: 10.36080/idealis.v8i2.3518.

[9] T. A. Wahyudi, A. Arifia, and A. Adi Suryanto, "Application of Text Mining and K-Medoids for Job *Clustering* Based on Job Description Analysis," *J. Teknol. Dan Manaj. Inform.*, vol. 11, no. 1, pp. 100–109, Jul. 2025, doi: 10.26905/jtmi.v11i1.15863.

[10] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, in SIGMOD '93. New York, NY, USA: Association for Computing Machinery, Jun. 1993, pp. 207–216. doi: 10.1145/170035.170072.

Copyright © SENASIF 2026
ISSN: 2598-0076

[11] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "*Cluster* validity indices for automatic *clustering*: A comprehensive review," *Heliyon*, vol. 11, no. 2, p. e41953, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41953.

[12] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to *Cluster* Analysis," in Wiley Series in Probability and Statistics. Wiley, Mar. 1990. doi: 10.1002/9780470316801.

[13] J. Pavlopoulos, G. Vardakas, and A. Likas, "Revisiting *Silhouette* Aggregation," in *Discovery Science*, D. Pedreschi, A. Monreale, R. Guidotti, R. Pellungrini, and F. Naretto, Eds., Cham: Springer Nature Switzerland, 2025, pp. 354–368. doi: 10.1007/978-3-031-78977-9_23.

[14] E. Hikmawati, N. U. Maulidevi, and K. Surendro, "Minimum threshold determination method based on dataset characteristics in association rule mining," *J. Big Data*, vol. 8, no. 1, p. 146, Nov. 2021, doi: 10.1186/s40537-021-00538-3.

[15] H. Chen, M. Yang, and X. Tang, "Association rule mining of aircraft event causes based on the Apriori algorithm," *Sci. Rep.*, vol. 14, no. 1, p. 13440, Jun. 2024, doi: 10.1038/s41598-024-64360-6.

Copyright © SENASIF 2026

ISSN: 2598-0076