# PEMETAAN KARAKTERISTIK DOSEN MENGGUNAKAN ALGORITMA K-PROTOTYPES PADA UNIVERSITAS X

**Gulielmus Jason [1)] Dedi Trisnawarman [2)]**

[1,2)] Sistem Informasi Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
email : gulielmus.825210038@stu.untar.ac.id [1)], dedit@fti.untar.ac.id [2)]

## ABSTRAK

*Penelitian ini bertujuan memetakan karakteristik dosen di lingkungan universitas menggunakan algoritma K-Prototypes sebagai pendekatan analitik dalam pengambilan keputusan sumber daya manusia akademik. Dataset yang digunakan terdiri atas 662 entri dosen Universitas Tarumanagara dengan atribut numerik (usia, masa kerja) dan kategorikal (status kepegawaian, jenis kelamin, divisi, dan jenjang pendidikan). Proses pra-pemrosesan mencakup imputasi nilai hilang, normalisasi, serta encoding atribut kategorikal. Jumlah klaster optimal ditentukan menggunakan metode Elbow dan Silhouette Score, yang menghasilkan dua klaster dengan nilai Calinski–Harabasz sebesar 789,26. Klaster pertama merepresentasikan dosen senior berusia lanjut dengan masa kerja panjang, mayoritas berstatus tidak tetap, sedangkan klaster kedua menggambarkan dosen tetap berusia produktif dengan potensi tinggi dalam Tridharma. Hasil ini menunjukkan bahwa algoritma K-Prototypes efektif dalam mengelompokkan data campuran dan memberikan wawasan strategis terkait segmentasi dosen. Temuan ini dapat dimanfaatkan untuk mendukung kebijakan pengembangan SDM akademik, seperti perencanaan karier, pelatihan, serta distribusi beban kerja secara proporsional.*

### Key words

*K-Prototypes, data analitik universitas, karakteristik dosen, clustering, mixed-type data*

## 1. Pendahuluan

Data analitik pada universitas merupakan pendekatan strategis yang bertujuan memahami dinamika kinerja, perilaku, dan karakteristik sivitas akademika melalui eksplorasi data yang kompleks serta heterogen [1]. Salah satu aspek krusial dalam konteks ini adalah pemetaan karakteristik dosen yang memiliki peran sentral terhadap keberhasilan institusi pendidikan tinggi dalam kegiatan pengajaran, penelitian, dan pengabdian kepada masyarakat [2]. Data yang merepresentasikan karakteristik dosen pada umumnya bersifat mixed-type, karena mencakup atribut numerik dan atribut kategorikal. Kondisi tersebut menimbulkan tantangan dalam penerapan metode analisis data konvensional yang umumnya hanya mampu menangani satu tipe data [3].

Dalam ranah data mining, *clustering* merupakan salah satu teknik utama yang digunakan untuk mengidentifikasi pola atau struktur tersembunyi pada data tanpa label [4]. Di antara berbagai algoritma klasterisasi, *K-Prototypes* menjadi salah satu metode yang banyak digunakan karena kemampuannya menggabungkan prinsip *K-Means* untuk data numerik dan *K-Modes* untuk data kategorikal, sehingga mampu menangani data campuran secara efektif [5]. Algoritma ini telah diterapkan dalam berbagai bidang, termasuk pendidikan [6], transportasi [7], kesehatan [8], dan penelitian ilmiah [9], berkat keandalannya dalam mengelompokkan entitas berdasarkan kemiripan atribut numerik maupun kategorikal.

Sejumlah penelitian sebelumnya telah mengevaluasi serta mengembangkan algoritma *K-Prototypes*. Aschenbruck et al. [1] menunjukkan bahwa proses inisialisasi pusat klaster berpengaruh signifikan terhadap kualitas hasil klasterisasi, sehingga diperlukan strategi optimasi untuk menghindari konvergensi pada nilai minimum lokal. Penelitian lain mengusulkan pendekatan multi-view untuk menangani data dengan beragam perspektif atribut [10], serta strategi imputasi untuk mengatasi data yang memiliki nilai hilang (missing values) [11]. Costa et al. [3], melalui studi benchmarking, menemukan bahwa *K-Prototypes* menunjukkan performa yang kompetitif dibandingkan metode berbasis jarak lainnya, seperti KAMILA dan FAMD-KMeans.

Dalam konteks pendidikan tinggi, penerapan algoritma ini telah menghasilkan berbagai temuan yang relevan. Dake et al. [12] mengelompokkan perilaku mahasiswa menggunakan *K-Prototypes* guna mendukung sistem bimbingan akademik adaptif. Palani et al. [13] menerapkannya untuk mendeteksi tingkat keterlibatan rendah mahasiswa dalam lingkungan pembelajaran daring. Sementara itu, Arfianta et al. [14] serta Pişirgen dan Peker [15] memanfaatkan metode serupa untuk mengelompokkan dosen berdasarkan tingkat produktivitas publikasi ilmiah. Hasil-hasil tersebut menunjukkan potensi besar *K-Prototypes* dalam mendukung proses pengambilan keputusan di sektor pendidikan tinggi.

Berdasarkan tinjauan tersebut, penelitian ini bertujuan melakukan pemetaan karakteristik dosen di lingkungan

universitas dengan menggunakan algoritma *K-Prototypes*. Hasil klasterisasi diharapkan dapat memberikan gambaran deskriptif mengenai pola karakteristik dosen yang dapat dimanfaatkan sebagai dasar dalam perumusan kebijakan pengembangan sumber daya manusia akademik, termasuk perencanaan pelatihan, promosi, serta pengalokasian beban kerja secara proporsional.

## 2. Metode Penelitian

### 2.1. Seleksi Data

Pemilihan data dilakukan untuk memastikan hanya atribut yang relevan terhadap karakteristik dosen yang dianalisis. Data yang tidak lengkap atau tidak relevan dieliminasi untuk meningkatkan kualitas analisis.

*Dataset* penelitian terdiri atas 662 entri staf akademik dari Universitas Tarumanagara, Jakarta, Indonesia, yang dikumpulkan pada Oktober 2025. Setiap entri merepresentasikan satu individu staf akademik dengan atribut sebagai berikut:

1. Status – status kepegawaian.
2. L/P – jenis kelamin dosen.
3. Divisi – fakultas atau unit kerja tempat dosen bertugas.
4. Umur/Usia – usia dosen dalam tahun.
5. Masa Kerja – lamanya masa kerja dosen di institusi.
6. Jenjang Pendidikan – tingkat pendidikan tertinggi.

Data telah dianonimkan untuk menjaga kerahasiaan identitas dan melalui proses verifikasi kualitas data agar bebas dari inkonsistensi maupun kesalahan entri.

### 2.2. Pra-Pemrosesan Data

Langkah ini mencakup penanganan nilai hilang (missing values) dan inkonsistensi data. Untuk atribut dengan nilai hilang, dilakukan imputasi menggunakan pendekatan berbasis klaster yang direkomendasikan oleh Aschenbruck et al. [11]. Data duplikat dan atribut yang tidak relevan dieliminasi untuk meningkatkan kualitas analisis.

### 2.3. Transformasi Data

Pada tahap transformasi data, atribut numerik dinormalisasi menggunakan metode *Z-Score Standardization* untuk menyamakan skala antar variabel dan menghindari dominasi atribut dengan rentang nilai besar.

Untuk atribut kategorikal, tidak dilakukan *encoding* secara eksplisit karena algoritma *K-Prototypes* secara internal menangani perhitungan jarak antar kategori menggunakan fungsi ketidaksamaan (*dissimilarity measure*) seperti pada *K-Modes*. Seluruh nilai kategorikal dikonversi ke tipe string guna memastikan kompatibilitas dengan algoritma.

### 2.4. Proses *Clustering* dengan Algoritma *K-Prototypes*

Algoritma *K-Prototypes* digunakan untuk mengelompokkan dosen berdasarkan kemiripan atribut numerik dan kategorikal. Metode ini mengombinasikan pendekatan *K-Means* untuk atribut numerik dan *K-Modes* untuk atribut kategorikal, sehingga mampu menangani data dengan tipe campuran [5]. Fungsi objektif dari algoritma *K-Prototypes* dirumuskan sebagaimana ditunjukkan pada Persamaan (1).

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_{ij} \left[ \sum_{l \in N} (x_{il} - \mu_{jl})^2 + \gamma \sum_{l \in C} d(x_{il}, \mu_{jl}) \right] \quad \ldots\ldots\ldots (1)$$

dengan:

1. $J$= fungsi biaya total yang diminimalkan,
2. $n$= jumlah data,
3. $k$= jumlah klaster,
4. $\delta_{ij} = 1$jika objek ke-$i$ termasuk dalam klaster ke-$j$, dan 0 jika tidak,
5. $N$= himpunan atribut numerik,
6. $C$= himpunan atribut kategorikal,
7. $x_{il}$= nilai atribut ke-$l$ dari data ke-$i$,
8. $\mu_{jl}$= pusat klaster ke-$j$ pada atribut ke-$l$,
9. $d(x_{il}, \mu_{jl})$= fungsi dissimilarity (0 jika sama, 1 jika berbeda),
10. $\gamma$= parameter penyeimbang antara kontribusi atribut numerik dan kategorikal.

Dalam penelitian ini, parameter penyeimbang $\gamma$ ditetapkan mengikuti pendekatan standar algoritma *K-Prototypes*, yang secara otomatis menyesuaikan proporsi antara kontribusi atribut numerik dan kategorikal dalam perhitungan jarak. Strategi inisialisasi pusat klaster menggunakan metode *Cao initialization*, yang menetapkan pusat awal berdasarkan distribusi kepadatan data guna meningkatkan stabilitas dan konsistensi hasil klasterisasi.

Proses iterasi berlanjut hingga nilai fungsi biaya J mencapai konvergensi, yaitu ketika perubahan nilai antariterasi berada di bawah ambang batas tertentu.

### 2.5. Penentuan Jumlah Klaster Optimal

Penentuan jumlah klaster optimal dilakukan menggunakan dua pendekatan, yaitu *Elbow Method* dan *Silhouette Coefficient* [12], [13]. Pada *Elbow Method*, nilai Within-Cluster Sum of Squares (*WCSS*) dihitung untuk berbagai jumlah klaster $k$, sebagaimana ditunjukkan pada Persamaan (2).

$$WCSS(k) = \sum_{i=1}^{k} \sum_{x \in C_i} || x - \mu_i ||^2 \ldots\ldots\ldots\ldots (2)$$

dengan:

1. $WCSS(k)$= total jumlah kuadrat jarak dalam setiap klaster (Within-Cluster Sum of Squares),
2. $C_i$= himpunan anggota pada klaster ke-$i$,
3. $x$= vektor data anggota klaster,
4. $\mu_i$= pusat (*centroid*) klaster ke-$I$,
5. $k$= jumlah klaster yang diuji.

Nilai $k$optimal dipilih pada titik di mana penurunan *WCSS* mulai melandai (membentuk "siku" pada grafik), yang menandakan keseimbangan antara jumlah klaster dan homogenitas data.

*Silhouette Coefficient* digunakan untuk mengevaluasi kualitas klaster dengan membandingkan seberapa baik suatu objek berada dalam klasternya sendiri dibandingkan dengan klaster lainnya. Persamaannya ditunjukkan pada Persamaan (3).

$$s(i) = \frac{b(i) - a(i)}{max\ (a(i), b(i))} \ldots\ldots\ldots\ldots (3)$$

dengan:

1. $a(i)$= rata-rata jarak antara objek $i$dan semua anggota klasternya sendiri,
2. $b(i)$= jarak rata-rata terkecil antara objek $i$dan klaster lain.

Nilai $s(i)$mendekati 1 menunjukkan klaster yang baik, sedangkan nilai mendekati 0 atau negatif menandakan kemungkinan salah klasterisasi.

Hasil perbandingan kedua metode digunakan untuk menentukan jumlah klaster terbaik sebelum dilakukan proses *profiling*.

## 2.6. Interpretasi dan Evaluasi Klaster

Tahap terakhir melibatkan interpretasi hasil klaster untuk mengidentifikasi profil dosen berdasarkan pola karakteristik dominan pada setiap klaster. Validitas hasil klasterisasi dievaluasi menggunakan Indeks *Calinski–Harabasz* (*CH*) [10], yang mengukur keseimbangan antara dispersi antar-klaster dan dalam-klaster. Persamaannya ditunjukkan pada Persamaan (4).

$$CH = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)} \ldots\ldots\ldots\ldots (4)$$

dengan:

1. $Tr(B_k)$= total dispersi antar-klaster (*between-cluster dispersion*),
2. $Tr(W_k)$= total dispersi dalam klaster (*within-cluster dispersion*),
3. $n$= jumlah data,
4. $k$= jumlah klaster.

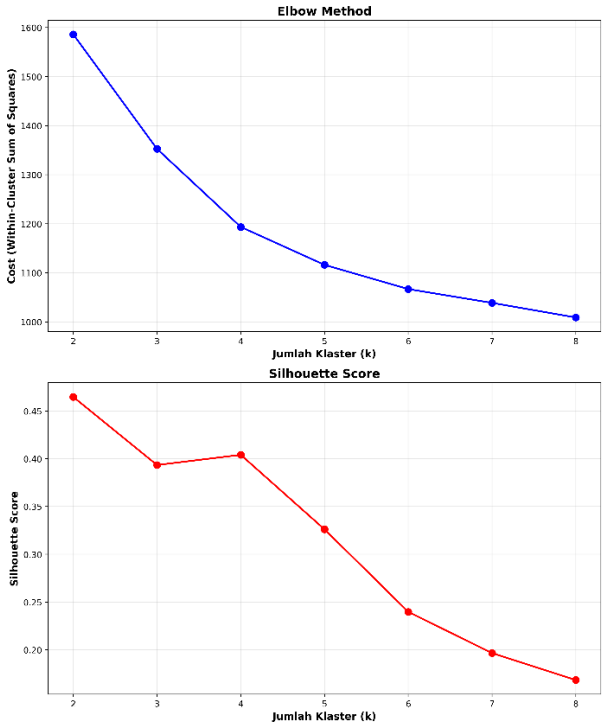
Nilai *CH* yang tinggi menunjukkan klaster yang baik, yaitu ketika jarak antar-klaster besar dan homogenitas dalam klaster tinggi.

Hasil interpretasi ini diharapkan memberikan masukan strategis dalam kebijakan pengembangan dosen dan perencanaan sumber daya manusia di lingkungan universitas.

# 3. Hasil dan Pembahasan

## 3.1. Penentuan Jumlah Klaster Optimal

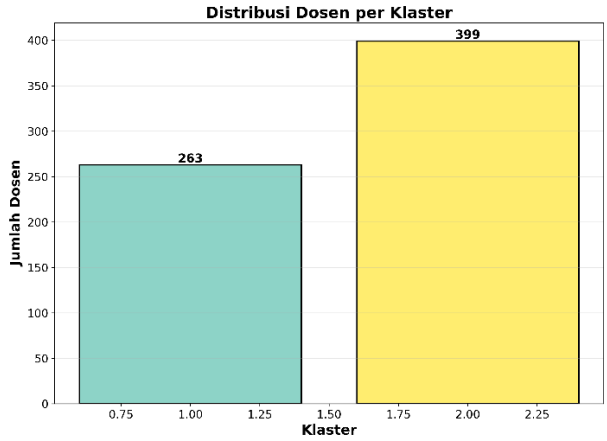
Penentuan jumlah klaster dilakukan menggunakan metode Elbow dan *Silhouette Score* (Gambar 1). Berdasarkan grafik *Elbow Method*, penurunan nilai *cost* mulai melandai pada k = 2, sedangkan nilai *Silhouette Score* tertinggi juga diperoleh pada k = 2 (0.465). Hasil ini menunjukkan bahwa dua klaster merupakan jumlah optimal untuk memisahkan karakteristik dosen secara signifikan tanpa kehilangan kehomogenan dalam masing-masing kelompok.




Gambar 1 Grafik *Elbow Method* dan *Silhouette Score*

## 3.2. Hasil *Clustering*

Proses klasterisasi dilakukan menggunakan algoritma *K-Prototypes* dengan parameter k = 2 dan metode inisialisasi Cao. Proses konvergensi tercapai pada iterasi ke-6 dengan nilai *cost* akhir sebesar 1585.96. Hasil *clustering* menghasilkan dua klaster utama dengan distribusi Klaster 1 sebanyak 263 dosen (39,7%) dan Klaster 2 sebanyak 399 dosen (60,3%) (lihat Gambar 2).

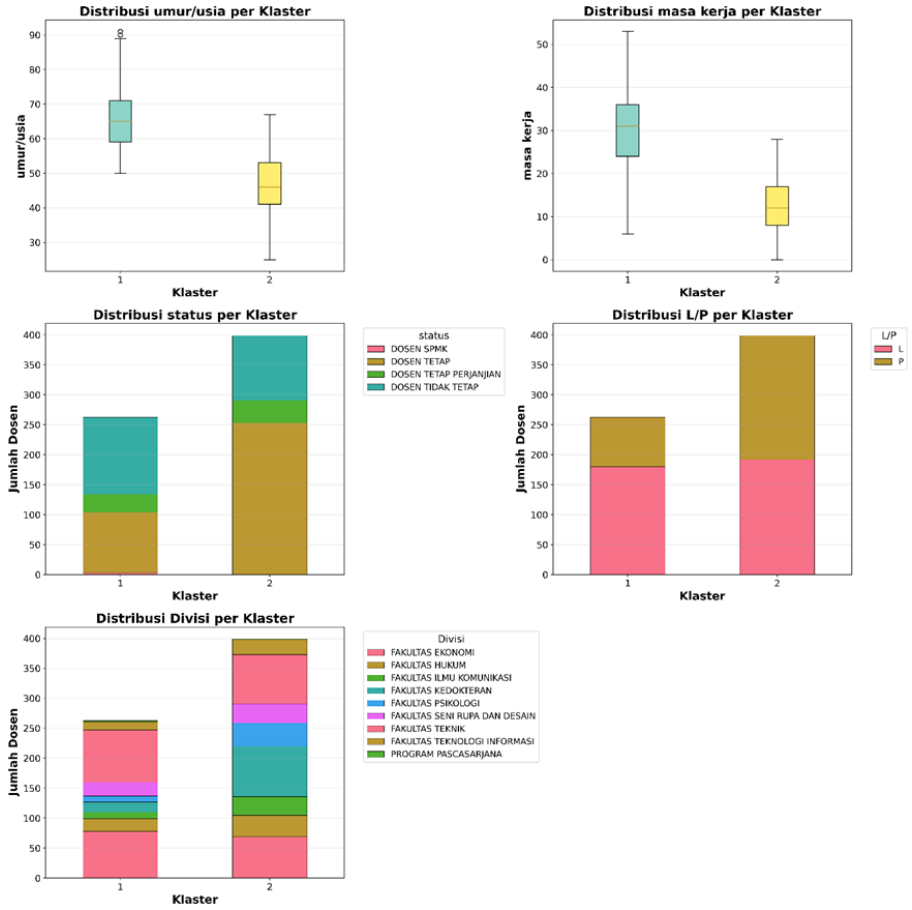



Gambar 2 Distribusi dan Proporsi Dosen per Klaster

Nilai *Calinski–Harabasz* Score sebesar 789.26 menunjukkan bahwa pemisahan antar-klaster cukup baik dan memiliki jarak antargrup yang jelas.

## 3.3. Karakteristik Klaster

Klaster 1 terdiri atas 263 dosen (39,7%), dengan rata-rata usia 65,73 tahun dan masa kerja 29,79 tahun. Berdasarkan visualisasi *boxplot* (Gambar 3), kelompok ini didominasi oleh dosen berusia di atas 60 tahun dengan masa kerja panjang.




Gambar 3 Distribusi Karakteristik Dosen per Klaster

Dari sisi kategorikal, proporsi dosen tetap (51%) dan tidak tetap (49%) relatif seimbang, dengan kecenderungan lebih banyak dosen tetap. Mayoritas berjenis kelamin laki-laki (68,4%), dan banyak berasal dari Fakultas Teknik (33,1%) serta Fakultas Ekonomi (29,7%). Mayoritas memiliki jenjang pendidikan Magister (S2) sebesar 49,8%, diikuti oleh Doktor (S3) sebanyak 42,6%.

Karakteristik tersebut mengindikasikan bahwa Klaster 1 merepresentasikan dosen senior berpengalaman yang telah lama berkarier di institusi, dengan kecenderungan masih aktif mengajar meskipun banyak berstatus tidak tetap.

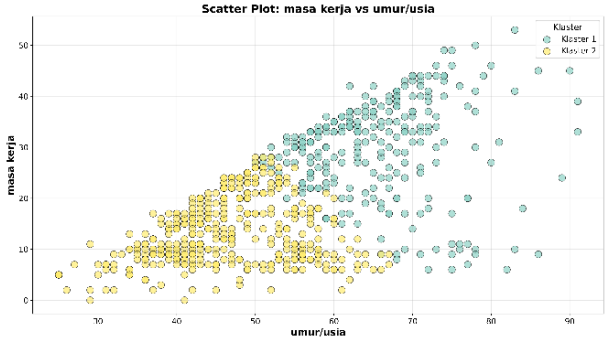
Klaster 2 merupakan kelompok terbesar dengan 399 dosen (60,3%). Rata-rata usia anggota kelompok ini adalah 46,37 tahun dengan masa kerja 13,14 tahun. Berdasarkan distribusi (Gambar 3), dosen dalam klaster ini berada pada rentang usia produktif 40–50 tahun dan masa kerja menengah.

Kategori status menunjukkan dominasi dosen tetap (63,4%), dengan proporsi jenis kelamin perempuan lebih tinggi (51,9%) dibanding laki-laki. Divisi terbanyak berasal dari Fakultas Kedokteran (21,1%), Fakultas

Teknik (20,6%), dan Fakultas Ekonomi (17,3%). Sebagian besar berpendidikan Magister (S2) (67,9%) dan Doktor (S3) (26,6%).

Secara umum, Klaster 2 dapat diidentifikasi sebagai kelompok dosen muda tetap yang aktif dalam kegiatan Tridharma dan berpotensi menjadi tulang punggung institusi dalam jangka panjang.

### 3.4. Analisis Perbandingan Antar Klaster




Gambar 4 *Scatter plot* Masa Kerja terhadap Usia Dosen per Klaster

Gambar 4 menunjukkan perbandingan sebaran usia dan masa kerja antar-klaster. Dosen pada Klaster 1 memiliki korelasi positif kuat antara usia dan masa kerja, mencerminkan pengalaman panjang dalam pengajaran dan penelitian. Sebaliknya, Klaster 2 menunjukkan variasi yang lebih luas, menandakan adanya regenerasi dosen dengan latar belakang pengalaman yang beragam.

Secara umum, hasil ini menggambarkan adanya segmentasi dua kelompok dosen dalam institusi:

1. Kelompok senior (Klaster 1) – dosen berpengalaman dengan masa kerja panjang, cenderung berstatus tidak tetap.
2. Kelompok muda (Klaster 2) – dosen produktif berstatus tetap, mendominasi jumlah dosen aktif.

## 4. Kesimpulan

Penelitian ini menunjukkan bahwa algoritma *K-Prototypes* dapat mengelompokkan data dosen yang memiliki kombinasi atribut numerik dan kategorikal. Berdasarkan hasil analisis, diperoleh dua klaster utama yang merepresentasikan pola karakteristik dosen di lingkungan Universitas Tarumanagara. Klaster pertama menggambarkan kelompok dosen senior berusia lanjut dengan masa kerja panjang dan sebagian besar berstatus tidak tetap, sedangkan klaster kedua merepresentasikan dosen muda berstatus tetap dengan usia produktif dan masa kerja menengah. Nilai *Calinski–Harabasz* sebesar 789,26 menunjukkan bahwa hasil klasterisasi memiliki kualitas pemisahan yang baik dan jarak antargrup yang cukup jelas. Temuan ini menegaskan bahwa *K-Prototypes* dapat digunakan sebagai alat analitik yang bermanfaat untuk memahami segmentasi dosen dan mendukung pengambilan keputusan strategis dalam pengelolaan sumber daya manusia akademik, seperti perencanaan karier, pelatihan, dan distribusi beban kerja. Meskipun demikian, penelitian ini masih terbatas pada satu institusi dan periode data tertentu, sehingga pada penelitian selanjutnya disarankan untuk memperluas cakupan data lintas universitas serta membandingkan performa *K-Prototypes* dengan algoritma lain guna memperoleh hasil klasterisasi yang lebih komprehensif dan generalis.

## REFERENSI

[1] R. Aschenbruck, G. Szepannek, and A. F. X. Wilhelm, “Initialization Strategies for Clustering Mixed-Type Data with the K-Prototypes Algorithm,” Advances in Data Analysis and Classification, 2025. https://doi.org/10.1007/s11634-025-00639-4

[2] S. Bumann, “Research Profile Clusters Among Lecturers in Non-Traditional Higher Education: An Exploratory Analysis in the Swiss Context,” International Journal of Educational Research Open, vol. 3, p. 100182, 2022. https://doi.org/10.1016/j.ijedro.2022.100182

[3] E. Costa, I. Papatsouma, and A. Markos, “Benchmarking Distance-Based Partitioning Methods for Mixed-Type Data,” Advances in Data Analysis and Classification, vol. 17, no. 3, pp. 701–724, 2023. https://doi.org/10.1007/s11634-022-00521-7

[4] A. Diop et al., “Simrec: A Similarity Measure Recommendation System for Mixed Data Clustering Algorithms,” Journal of Big Data, vol. 12, p. 43, 2025. https://doi.org/10.1186/s40537-024-01052-y

[5] G. Szepannek, R. Aschenbruck, and A. Wilhelm, “Clustering Large Mixed-Type Data with Ordinal Variables,” Advances in Data Analysis and Classification, vol. 19, pp. 749–767, 2025. https://doi.org/10.1007/s11634-024-00595-5

[6] M. Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, “Clustering Analysis for Classifying Student Academic Performance in Higher Education,” Applied Sciences, vol. 12, no. 19, p. 9467, 2022. https://doi.org/10.3390/app12199467

[7] A. Mohd, L. E. Teoh, and H. L. Khoo, “Passengers’ Requests Clustering with K-Prototype Algorithm for the First-Mile and Last-Mile Shared-Ride Taxi Service,” Multimodal Transportation, vol. 3, no. 2, p. 100132, 2024. https://doi.org/10.1016/j.multra.2024.100132

[8] P. Liu et al., “A Modified and Weighted Gower Distance-Based Clustering Analysis for Mixed Type Data,” BMC Medical Research Methodology, vol. 24, p. 305, 2024. https://doi.org/10.1186/s12874-024-02427-8

[9] K. Chahuán-Jiménez et al., “Cluster Analysis of Digital Competencies Among Professors in Higher Education,” Frontiers in Education, vol. 10, 2025. https://doi.org/10.3389/feduc.2025.1499856

[10] J. Ji et al., “A Multi-View Clustering Algorithm for Mixed Numeric and Categorical Data,” IEEE Access, vol. 9, pp. 24913–24924, 2021. https://doi.org/10.1109/ACCESS.2021.3057113

[11] R. Aschenbruck, G. Szepannek, and A. F. X. Wilhelm, “Imputation Strategies for Clustering Mixed-Type Data with Missing Values,” Journal of Classification, vol. 40, pp. 2–24, 2023. https://doi.org/10.1007/s00357-022-09422-y

[12] D. K. Dake, E. Gyimah, and C. Buabeng-Andoh, “University Students Behaviour Modelling Using the K-Prototype Clustering Algorithm,” Mathematical Problems in Engineering, 2023. https://doi.org/10.1155/2023/5507814

[13] K. Palani, P. Stynes, and P. Pathak, “Clustering Techniques to Identify Low-Engagement Student Levels,” in Proc. 13th Int. Conf. on Computer Supported Education (CSEDU), 2021, pp. 248–257. https://doi.org/10.5220/0010456802480257
[14] P. Arfianta, K. Fithriasari, and W. T. D. Ary, “Clustering Mixed-Type Data on the Scientific Publication Productivity of ITS Lecturers Using CEBMDC,” in Proc. Int. Conf. on Data Science and Its Applications (ICoDSA), 2025, pp. 1353–1358. https://doi.org/10.1109/ICoDSA67155.2025.11156956
[15] A. Pişirgen and S. Peker, “A Clustering Approach for Classifying Scholars Based on Publication Performance Using Bibliometric Data,” Egyptian Informatics Journal, vol. 28, p. 100537, 2024. https://doi.org/10.1016/j.eij.2024.100537