Clustering of Higher Education Institutions Based on Tuition Fee and Living Cost of International Students Using the Gaussian Mixture Models Method

Nadya Aulya Putri[1], Hevlie Winda Nazry[2]
Faculty of Computer Science and Information Technology, Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
Email: [1]nadyaaulyaputri66@gmail.com, [2]hevliewindanazry@umsu.ac.id

ABSTRACT

The wide variation in tuition fees and living costs among higher education institutions across different countries creates significant challenges for prospective international students in making informed enrollment decisions. This study applies the Gaussian Mixture Models (GMM) method to cluster higher education institutions based on the financial burden of international students. The dataset was obtained from the Kaggle platform, comprising 1,407 records covering multiple universities worldwide, with four primary features: Tuition_USD, Living_Cost_Index, Rent_USD, and Visa_Fee_USD. The research methodology encompassed data collection, preprocessing with Z-Score normalization, determination of the optimal number of clusters using the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), and parameter estimation via the Expectation-Maximization (EM) algorithm. The GMM method was selected for its probabilistic approach, which is well-suited to modeling complex and multivariate data distributions. The experimental results demonstrate that the GMM model with K=7 clusters effectively identifies seven distinct university groups with meaningfully differentiated cost profiles, converging at iteration 139 with BIC = 11,196.28 and AIC = 10,650.36. The identified clusters range from affordable public universities to exclusive elite institutions, offering nuanced insights into the financial landscape of international higher education. These findings confirm that GMM-based soft clustering adequately uncovers latent structural patterns in right-skewed, asymmetric international student cost data, outperforming hard clustering methods such as K-Means in handling distributional complexity.

Article Info

Article history:

Received June, 2, 2026
Revised June 21, 2026
Accepted July 12, 2026

Keyword: Gaussian Mixture Models; Soft Clustering; International Student; Tuition Fee & Living Cost; BIC.

Corresponding Author:
Name: Nadya Aulya Putri
Department: Information Technology Study Program
Institution: Universitas Muhammadiyah Sumatera Utara
Address: Medan
Email: nadyaaulyaputri66@gmail.com

1. INTRODUCTION

The internationalization of higher education has expanded considerably over the past two decades, with millions of students each year choosing to study abroad in destinations such as the United States, United Kingdom, Australia, Canada, and Germany [1]-[5]. For international students, navigating the financial side of studying overseas is one of the most critical factors in their enrollment decision, as they must account for multiple cost components simultaneously including tuition, living expenses, rent, and visa fees all of which vary widely across institutions and countries, and can significantly affect accessibility and student welfare [6]-[8].

Previous studies have largely approached these financial challenges through qualitative policy reviews or traditional statistical methods such as regression analysis. Although these methods provide valuable insights, they are not well-suited for uncovering hidden patterns within the complex, multidimensional cost structure of international higher education[9]-[11]. Clustering, as a form of unsupervised machine learning, offers a more

ISSN: -

systematic approach to grouping institutions by shared cost characteristics and enabling more nuanced classification of their financial profiles [12]-[15].

To address this, the present study applies Gaussian Mixture Models (GMM), a probabilistic framework that employs the Expectation-Maximization (EM) algorithm for parameter estimation and soft cluster assignments. Model selection is guided by the Bayesian Information Criterion (BIC). Key contributions include the application of GMM to a dataset spanning more than 20 countries, Z-Score normalization to ensure scale consistency, and a systematic model comparison across K=2 to K=7, resulting in a meaningful classification of institutions by their financial cost profiles [16]-[18].

2. METHOD

This study employs a quantitative research design with a descriptive-exploratory approach. The research aims to identify and characterize latent cost structures in international student expenditure data through probabilistic clustering, without testing directional hypotheses between variables. The analytical pipeline encompasses five sequential stages: (1) data collection, (2) data preprocessing, (3) feature engineering and normalization, (4) GMM modeling with EM parameter estimation, and (5) model evaluation and cluster interpretation.

2.1 Research Framework

The methodological framework integrates principles from data mining, machine learning, and applied statistics. Gaussian Mixture Models are selected as the primary analytical tool on the basis of two key justifications. First, GMM is a model-based clustering approach that accommodates distributional heterogeneity across clusters a critical property given that cost features such as Tuition_USD are known to exhibit severe right skewness. Second, GMM's soft assignment mechanism generates probabilistic cluster membership scores, providing richer information than the hard assignments produced by K-Means[19]-[21].

2.2 Gaussian Mixture Models

A Gaussian Mixture Model assumes that the observed data X are generated from a mixture of K Gaussian components, each characterized by a mean vector μk, a covariance matrix Σk, and a mixing coefficient πk. The model's probability density function is [22]-[24]:

$$p(x) = \Sigma\, \pi k \cdot N(x|\, \mu k,\, \Sigma k) \tag{1}$$

where $\Sigma\ \pi k = 1$ and $\pi k \geq 0$. Parameter estimation is performed via the EM algorithm, which iterates between the Expectation step (E-step), computing the posterior probability (responsibility) of each data point belonging to each component, and the Maximization step (M-step), updating the model parameters to maximize the expected log-likelihood.

2.3 Model Selection Criteria

The Bayesian Information Criterion (BIC) is used as the primary criterion for selecting the optimal number of clusters. BIC is defined as [25]-[27]:

$$BIC = -2 \cdot ln(L) + k \cdot ln(N) \tag{2}$$

where L denotes the maximized likelihood, k the number of free parameters, and N the sample size. Lower BIC values indicate better-fitting models with appropriate complexity. The Akaike Information Criterion (AIC) is computed as an auxiliary criterion: $AIC = -2 \cdot \ln(L) + 2k$ [28], [29].

2.4 Implementation Environment

All analyses are implemented in Python 3.x using the following libraries: NumPy for numerical computation; pandas for data manipulation, StandardScaler, and evaluation metrics, and Matplotlib and Seaborn for visualization. The development environment is Visual Studio Code (VSCode).

ISSN: -

3. DATASET AND PREPROCESSING

3.1 Dataset Description

The dataset used in this study is the Cost of International Education dataset, sourced from the Kaggle open data platform. It comprises 1,407 records across institutions in more than 20 countries, covering tuition fees, living cost indices, rental costs, visa fees, and other financial indicators. Table 1 describes the four features selected for clustering analysis.

Table 1. Selected Features for Clustering Analysis

No	Feature	Description
1	Tuition_USD	Annual tuition fee in US dollars (USD)
2	Living_Cost_Index	City-level cost of living index (composite index)
3	Rent_USD	Monthly accommodation cost in USD
4	Visa_Fee_USD	Student visa processing fee in USD

3.2 Exploratory Data Analysis

Descriptive statistics reveal considerable distributional heterogeneity across features. Tuition_USD exhibits strong right skewness (skewness ≈ 2.1), with a mean of $15,872 but a median of only $8,500, indicating that a minority of elite institutions substantially inflate the sample mean. Visa_Fee_USD displays a bimodal distribution, with two dominant modes near $100–150 and $400–500, corresponding to different national visa fee regimes. Rent_USD and Living_Cost_Index exhibit more symmetric distributions with skewness values near zero, suggesting more homogeneous distributions across the dataset.

Pearson correlation analysis reveals that Rent_USD and Living_Cost_Index share the strongest positive association ($r = 0.72$), consistent with the notion that high cost-of-living cities command higher rental prices. Tuition_USD correlates moderately with Rent_USD ($r = 0.64$) and Visa_Fee_USD ($r = 0.50$). Visa_Fee_USD displays weaker correlations with all other features, particularly with Living_Cost_Index ($r = 0.24$), indicating that visa pricing follows its own national policy logic rather than tracking broader economic conditions.

3.3 Data Preprocessing

Data quality verification confirmed that the dataset contained no missing values, duplicate entries, or structural anomalies, and no data cleaning was required. Z-Score standardization was applied to all four features prior to clustering:

$$z = (x - \mu) / \sigma \tag{3}$$

This transformation ensures that all features contribute equally to the clustering process regardless of their original measurement scales, which span several orders of magnitude. Post-normalization verification confirmed that all features attained a mean of 0 and a standard deviation of 1.

4. RESULTS AND DISCUSSION

4.1 Parameter Initialization

The initial covariance matrix for each component was set to the global sample covariance, and mixing coefficients were initialized to 1/K to reflect uniform a priori cluster proportions. These initialization choices are designed to promote faster convergence and reduce the risk of suboptimal local maxima in the EM optimization landscape.

4.2 Optimal Cluster Selection via BIC and AIC

GMM models were estimated for $K \in \{2, 3, 4, 5, 6, 7\}$. Both BIC and AIC exhibited monotonically decreasing trends with increasing K, with the minimum attained at K=7 for both criteria (BIC = 11,196.28; AIC = 10,650.36). Table 2 summarizes the model selection results.

Table 2. BIC and AIC Values by Number of Clusters

K	BIC	AIC	Converged at Iteration
2	14,301.63	13,875.21	86
3	12,880.09	12,263.45	82
4	11,986.23	11,178.87	61
5	11,839.07	10,841.09	86
6	11,231.31	10,742.51	200
7	**11,196.28**	**10,650.36**	**139**

** Optimal model (lowest BIC and AIC)*

ISSN: -

The log-likelihood convergence curves for all K values display the expected pattern: rapid improvement during early iterations followed by asymptotic stabilization. The K=6 model required 200 iterations to converge the maximum allowed suggesting potential instability in the optimization landscape, which further supports the selection of K=7 as the more stable and parsimonious solution.

4.3 Cluster Characterization

The GMM solution with K=7 yields seven clusters with distinct and interpretable cost profiles. Table 3 presents the cluster characteristics based on raw-scale means of the four cost features.

Table 3. Cluster Characteristics and Interpretation

Cluster	N	Tuition (USD)	LCI	Rent (USD)	Visa (USD)	Profile
1	483	$38,263	66	$1,207	$103	Affordable, most common
2	53	$47,625	44	$1,870	$460	High tuition, very high visa
3	250	$3,972	73	$340	$106	Mid-tier institutions
4	108	$5,285	68	$961	$159	Premium urban universities
5	44	$31,771	77	$1,802	$466	Subsidized, expensive cities
6	437	$23,258	70	$1,672	$268	Renowned universities, developed nations
7	32	$1,629	98	$2,200	$93	Exclusive, highest total cost

Note: LCI = Living Cost Index

Several findings merit specific discussion. Cluster 1 (n=483) is the largest and represents the most accessible segment of the international education market, characterized by low-to-moderate tuition and modest living costs. This cluster is predominantly composed of universities in Germany, South Korea, France, and other European and East Asian countries where tuition subsidies or regulated fee structures maintain affordability.

Cluster 2 (n=53) is a particularly noteworthy segment: despite possessing the highest mean tuition ($47,625), its living cost index (44) is the lowest across all clusters, indicating that these universities are located in economically less expensive cities or regions. This cluster is likely dominated by private institutions in developing countries that charge international-rate tuition while benefiting from low local living costs. The very high visa fees ($460) suggest countries with stringent immigration regimes.

Cluster 7 (n=32) represents the most exclusive segment with the highest living cost index (98) and the lowest tuition ($1,629). This seemingly paradoxical combination likely reflects elite public or semi-public institutions particularly in Singapore, Switzerland, and Scandinavian countries where government-controlled tuition is low but metropolitan living costs are extremely high.

Cluster 6 (n=437) is the second largest and encompasses renowned universities in developed English-speaking nations, displaying high tuition ($23,258), high living costs (70), and high rental costs ($1,672). USA, UK, and Australian institutions largely populate this cluster.

4.4 Geographical Distribution of Clusters

Geographical analysis confirms that cluster membership follows expected regional patterns. USA and Canada exhibit the highest cluster diversity, with significant proportions of universities in Clusters 4 and 6, reflecting the premium cost structure of North American higher education. Germany, South Korea, France, and Italy are predominantly classified in Cluster 1, consistent with their state-subsidized tuition policies. Singapore and Switzerland demonstrate higher proportions of high-living-cost clusters (Clusters 7 and 6), consistent with their status as global financial centers.

4.5 Model Evaluation

The GMM model with K=7 demonstrates strong convergence behavior and interpretable cluster profiles. The BIC penalty for model complexity ensures that the seven-cluster solution is not an artifact of overfitting. The soft cluster assignment mechanism inherent to GMM is particularly appropriate for this dataset, as cost distributions across clusters show partial overlap a property that deterministic methods like K-Means cannot adequately represent. The distributional assumptions of GMM (Gaussian components) are approximately met for Rent_USD (R^2=0.972) and Living_Cost_Index (R^2=0.973) based on Q-Q plot analysis, while Tuition_USD (R^2=0.864) and Visa_Fee_USD (R^2=0.792) deviate more substantially, consistent with their right-skewed and bimodal characteristics respectively.

5. CONCLUSION

This study applied Gaussian Mixture Models to cluster 1,407 international higher education institutions based on four cost features: Tuition_USD, Living_Cost_Index, Rent_USD, and Visa_Fee_USD. The GMM model with K=7, selected via the Bayesian Information Criterion, identified seven meaningful and distinct cost-profile clusters, converging at iteration 139 with BIC = 11,196.28 and AIC = 10,650.36.

The clusters reveal a nuanced spectrum of international educational costs, from affordable state-subsidized institutions in continental Europe and East Asia (Cluster 1) to exclusive metropolitan universities with very high living costs (Cluster 7), and prestigious institutions in English-speaking developed nations (Cluster 6). Notably, the analysis uncovered a counterintuitive cluster (Cluster 2) combining high tuition with low living costs, and another (Cluster 7) with very low tuition but extremely high living costs patterns that simple univariate or regression-based analyses would likely obscure.

The superiority of GMM over hard clustering methods (e.g., K-Means) is demonstrated by its ability to model overlapping, non-spherical, and right-skewed distributions through probabilistic soft assignment characteristics that are inherent to international education cost data. The results provide actionable information for policymakers, university administrators, and prospective students seeking to navigate the global financial landscape of international higher education.

Future research should address several limitations of the current work: (1) the dataset is temporally limited to a single cross-section, and longitudinal tracking could reveal temporal dynamics in cost clustering; (2) additional features such as university ranking, acceptance rates, and quality-of-life indices would enable richer multi-dimensional profiling; and (3) comparative analysis against DBSCAN, Hierarchical Clustering, and Spectral Clustering would provide empirical validation of GMM's relative advantages.

REFERENCES

[1] Ahmad, A., Hasan, M., & Ghorbanpour, M. (2024). Education cost as a new fickle in higher education for students learning via quantitively multinomial logistic regression. *Scientific Reports*, 14(1), 29947.

[2] Agiakloglou, C. and Tsimpanos, A., "Evaluating the performance of AIC and BIC for selecting spatial econometric models," *Journal of Spatial Econometrics*, vol. 4, no. 1, pp. 1–35, 2023, doi: 10.1007/s43071-022-00030.

[3] Alalawi, S. J. S., Shaharanee, I. N. M., and Jamil, J. M., "Clustering student performance data using K-means algorithms," *Journal of Computational Innovation and Analytics*, vol. 2, no. 1, pp. 41–55, 2023, doi: 10.32890/jcia2023.2.1.3.

[4] Cheng, L., & You, C. (2016). Analysis of tuition growth rates based on clustering and regression models. *Educational Data Analysis*, 5(2), 44–58.

[5] Dias Lopes, D. A., "How do tuition fee increases affect international mobility? The case of European Union students in England," *European Journal of Education*, vol. 59, no. 1, pp. 1–15, 2024, doi: 10.1111/ejed.12592.

[6] Gormley, I. C., Murphy, T. B., and Raftery, A. E., "Model-based clustering," *Annual Review of Statistics and Its Application*, vol. 10, pp. 573–595, 2023, doi: 10.1146/annurev-statistics-033121-115326.

[7] Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan algoritma K-Means untuk clustering data obat-obatan pada RSUD Pekanbaru. *Jurnal Nasional Teknologi dan Sistem Informasi*, 5(1), 17–24.

[8] Gutema, D. M., Pant, S., and Nikou, S., "Exploring key themes and trends in international student mobility research — A systematic literature review," *Journal of Applied Research in Higher Education*, vol. 16, no. 3, pp. 843–861, 2024, doi: 10.1108/JARHE-05-2023-0195.

[9] Haslbeck, J. M. B., Vermunt, J. K., and Waldorp, L. J., "The impact of ordinal scales on Gaussian mixture recovery," *Behavior Research Methods*, vol. 55, no. 4, pp. 2143–2160, 2022, doi: 10.3758/s13428-022-01883-8.

[10] Hemati, M., "Factors influencing international students' choice of educational destination: A systematic review through the socioecological model lens," *Journal of International Students*, vol. 14, no. 2, pp. 58–79, 2024, doi: 10.32674/jis.v14i2.8160.

[11] Ikotun, A., Ezugwu, A., Abualigah, L., Abuhaija, B., and Heming, J., "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[12] Kurniawan, H., Defit, S., & Sumijan. (2020). Data mining menggunakan metode K-Means clustering untuk menentukan besaran uang kuliah tunggal. *Journal of Applied Computer Science and Technology*, 1(2), 80–89.

[13] Liu, R., "Data analysis of educational evaluation using K-means clustering method," *Mathematical Problems in Engineering*, vol. 2022, Art. no. 3762431, 2022, doi: 10.1155/2022/3762431.

[14] Liu, T., Hassan, A., Bin, M. A., and Anuar, M., "Systemic review: Financing of higher education internationalization," *Perspectives in Education*, vol. 41, no. 4, pp. 275–292, 2023, doi: 10.38140/pie.v41i4.7340.
[15] Lo, A. D., Chang, D. F., and Chang, A., "Exploring concurrent relationships between economic factors and student mobility in expanding higher education achieving 2030," *Sustainability*, vol. 14, no. 21, Art. no. 14612, 2022, doi: 10.3390/su142114612.
[16] Lundin, H. (2024). Tuition fees for international students: A policy instrument of cost sharing and control or simply income generation. *Higher Education Policy*, 37(1), 88–107.
[17] Luo, Y., "International student mobility and its broad impact on destination countries: A review and agenda for future research," *Journal of Studies in International Education*, vol. 27, no. 1, pp. 45–63, 2023, doi: 10.1177/09504222221150766.
[18] Mahfouz, T., Rahman, M., and Nikou, S., "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, Art. no. 759, 2021, doi: 10.3390/e23060759.
[19] Milson, A., Herwindiati, D. E., & Perdana, N. J. (2024). Penerapan klasifikasi suara sebagai autentikasi keamanan sistem login menggunakan Gaussian Mixture Models. *Computatio: Journal of Computer Science and Information Systems*, 8(1).
[20] Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences*, 12(19), 9467.
[21] OECD. (2025). *Education at a Glance 2025: OECD Indicators*. https://doi.org/10.1787/1C0D9C79-EN
[22] Pangastuti, S. S., Fithriasari, K., Iriawan, N., & Suryaningtyas, W. (2021). Data mining approach for educational decision support. *EKSAKTA: Journal of Sciences and Data Analysis*, 2(1), 33–44.
[23] Riyono, J., Puspa, S. D., & Pujiastuti, C. E. (2022). Simulasi clustering provinsi di Indonesia dalam penyebaran Covid-19 menggunakan algoritma Gaussian Mixture Model. *MAJAMATH: Jurnal Matematika dan Pendidikan Matematika*, 5(1), 43–60.
[24] Saqr, M., López-Pernas, S., and Helske, J., "An introduction and tutorial to model-based clustering in education via Gaussian mixture modelling," in *Learning Analytics Methods and Tutorials*, Springer, 2023, pp. 135–164, doi: 10.48550/arXiv.2306.06219.
[25] Schofield, C. (2024). The impact of the cost-of-living crisis on online student engagement and future study plans. *Higher Education Research & Development*, 43(2), 410–425.
[26] Shahapure, K. R. and Nicholas, C. K., "Cluster quality analysis using silhouette score," in *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Sydney, Australia, 2020, pp. 747–748, doi: 10.1109/DSAA49011.2020.00096.
[27] Ummami, R., & Winarno, B. (2023). Gaussian Mixture Model dengan algoritme Expectation Maximization untuk pengelompokan data distribusi air bersih di Jawa Barat. *PRISMA: Prosiding Seminar Nasional Matematika*, 6, 745–750.
[28] Wang, Z., "Higher education management and student achievement assessment method based on clustering algorithm," *Mathematical Problems in Engineering*, vol. 2022, Art. no. 4703975, 2022, doi: 10.1155/2022/4703975.
[29] Widiarina, W., Mariskhana, K., & Sintawati, I. D. (2024). Application of data mining for clustering human development index based on West Java Province 2017–2022. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 8(1), 44–53.