IJECA (International Journal of Education & Curriculum Application)
Journal homepage: http://journal.ummat.ac.id/index.php/IJECA
ISSN 2614-3380 | Vol. 9, No. 2, August 2026
https://doi.org/10.31764/ijeca.v9i2.38963

Quantitative Evaluation of Data Science Curriculum Structures in Indonesian Universities

Khinsa Fairuz Zahirah*
School of Information System, Universitas Ciputra Surabaya, Indonesia
Author Corresponding: khinsa.zahirah@ciputra.ac.id

ABSTRACT

This study presents a quantitative structural evaluation of undergraduate data science curricula across Indonesian universities using a compositional modeling approach. Existing competency frameworks identify essential domains in data science education, yet limited empirical research evaluates how these competencies are proportionally structured across institutions. Using course-level curriculum data collected from 37 undergraduate programs, courses were mapped into five competency domains: programming, statistical foundations, machine learning, applied practice, and ethics. Each curriculum was operationalized as a normalized compositional competency vector and analyzed using the Curriculum Structural Balance Index (CSBI), a distance-based metric that measures proportional structural imbalance across competency domains. Hierarchical clustering was further applied to identify recurring curriculum typologies. The findings reveal a consistent structural pattern characterized by strong concentration in programming and statistics, moderate variation in machine learning, and limited integration of applied practice and ethics. This recurring configuration, identified as a Foundational-Dominant model, indicates that current programs prioritize technical consolidation over interdisciplinary integration. Unlike conventional curriculum evaluation approaches that primarily assess competency presence or coverage, the proposed CSBI framework quantitatively evaluates proportional competency balance and enables systematic comparison of curriculum structures across institutions. The study contributes a measurable and replicable analytic approach for evaluating interdisciplinary curriculum architecture in higher education.

Keywords: Data Science Curriculum; Curriculum Evaluation; Curriculum Analytics; Compositional Analysis.

Article History:
Received: 01-04-2026
Revised : 07-05-2026
Accepted: 09-05-2026
Online : 01-08-2026

How to Cite (APA style):
Zahirah, K. F. (2026). Quantitative Evaluation of Data Science Curriculum Structures in Indonesian Universities. *IJECA (International Journal of Education and Curriculum Application)*, *9*(2), 518-530. https://doi.org/10.31764/ijeca.v9i2.38963

1. INTRODUCTION

Data science has rapidly institutionalized as an academic field in response to expanding industry demand for professionals capable of extracting insight from complex data (Donoho, 2017; Harden, 2001). As an inherently interdisciplinary field integrating statistics, computing, mathematics, and domain expertise, data science presents distinct challenges in curriculum design (De Veaux et al., 2017; National Academies of Sciences, Engineering, and Medicine, 2018). Recent studies highlight the rapid expansion of data science education alongside advances in artificial intelligence and data technologies, while emphasizing unresolved questions regarding competency definition, proportional emphasis, and curriculum coherence (Avila-Garzon & Bacca-Acosta, 2025; Fadillah et al., 2025). Universities must therefore determine not only which competencies to include, but how to structure and balance technical, applied, and governance components within a coherent program architecture (Li & Yu, 2022; Smaldone et al., 2022).

To guide curriculum development, several international frameworks have articulated core competency domains. The ACM Data Science Task Force (2021) emphasizes programming, data management, and modeling within computing-oriented curricula, while the National Academies of Sciences, Engineering, and Medicine (2018) provide a broader interdisciplinary vision for undergraduate data science education. The EDISON Data Science Framework (2017) further operationalizes competencies into structured bodies of knowledge to support curriculum mapping and design. Despite these advances, comparative analyses show limited international consensus on how competencies should be prioritized or structured within curricula (Schmitt et al., 2025). More importantly, existing frameworks define which competencies should be included but provide limited methodological guidance for quantitatively evaluating how these competencies are proportionally distributed and structurally balanced across programs.

Beyond formal frameworks, discipline-rooted scholarship has shaped the intellectual foundations of data science as an academic field. Donoho (2017) conceptualizes data science as an extension of computational statistics grounded in empirical workflows, while Hardin et al. (2015) emphasize the integration of programming, data acquisition, and modern data practices in undergraduate education. De Veaux et al. (2017) translate these principles into curriculum guidelines that integrate statistics, computing, and domain context within a coherent program structure. Collectively, these contributions underscore that data science education should combine technical foundations, applied context, and professional responsibility. However, while they define what should be taught, they do not address how these components are structurally distributed within curricula.

Existing research on data science education has largely focused on pedagogical strategies, competency descriptions, and individual program case studies (Deza & Deza, 2016; Harden, 2001). Reviews of the field highlight the predominance of descriptive and case-based approaches, with limited application of quantitative methods for analyzing curriculum structure across institutions (Deza & Deza, 2016; Dogucu et al., 2024). While curriculum mapping approaches have been used to identify where competencies are taught (Worden & Bray, 2025), these methods typically assess presence rather than proportional emphasis, limiting their ability to compare structural configurations across programs.

A related body of work highlights persistent challenges in integrating applied practice and ethics/governance within data science curricula (Borenstein & Howard, 2021; Nguyen et al., 2023). Studies show that undergraduate programs often prioritize computing and statistical foundations while allocating less structural emphasis to ethics and domain context (Oliver & McNeil, 2021). Even when these domains are included, their integration is often limited or compartmentalized (Baumer et al., 2022; Colando & Hardin, 2024). This reinforces a critical limitation: the presence of a competency does not guarantee its meaningful structural integration within a curriculum (Borenstein & Howard, 2021; Holmes et al., 2022).

Taken together, the literature demonstrates broad agreement regarding the core competency domains required in data science education, yet a significant methodological gap remains in how curriculum structures are evaluated across institutions. Existing competency frameworks primarily identify which competencies should be included, while curriculum mapping approaches generally assess competency presence or coverage rather than proportional structural distribution. As a result, limited quantitative methods exist for evaluating how competencies are compositionally organized, balanced, or concentrated within curriculum architectures across programs. This limitation is particularly important in interdisciplinary fields such as data science, where curriculum quality depends not only on competency inclusion but also on the proportional relationship among technical, applied, and ethical domains.

Conceptually, this study positions curriculum as a compositional system in which competency domains function as interdependent structural components whose proportional distribution shapes the overall curriculum architecture. From this perspective, compositional analysis provides an appropriate analytical framework for evaluating curriculum balance and structural concentration across institutions. This gap is especially relevant in emerging higher education systems, where institutional capacity, disciplinary heritage, and resource constraints may influence curriculum design in distinctive ways (Li & Yu, 2022). Indonesia provides a useful context for examining these dynamics as undergraduate data science programs continue to expand nationally while adapting global competency expectations to local academic conditions. However, systematic evidence remains limited regarding how Indonesian universities proportionally structure core data science competencies and whether recurring structural patterns emerge across institutions.

This study addresses this gap by conducting a multi-institutional structural evaluation of undergraduate data science–related programs across Indonesian universities. Courses are mapped to five core competency domains derived from internationally recognized frameworks (ACM Data Science Task Force, 2021; De Veaux et al., 2017; EDISON Data Science Framework, 2017; National Academies of Sciences, Engineering, and Medicine, 2018), and each curriculum is operationalized as a compositional competency vector. Using distance-based metrics and clustering analysis, the study enables quantitative comparison of proportional emphasis across programs. To support this evaluation, the study introduces the Curriculum Structural Balance Index (CSBI), which measures the degree of structural balance by quantifying deviation from an equal distribution of competency domains. Accordingly, this study addresses the following research questions: (1) how are core data science competencies proportionally distributed across undergraduate curricula in indonesian universities? (2) what structural patterns of curriculum configuration emerge across institutions? (3) which programs exhibit more balanced competency structures, and which show structural concentration? and (4) how do observed competency distributions align with internationally articulated data science competency frameworks?

By operationalizing curriculum as a compositional competency system and applying distance-based comparison and clustering techniques, this study moves beyond descriptive curriculum mapping toward quantitative structural evaluation. The proposed framework enables systematic detection of concentration patterns and recurring curriculum typologies across institutions, and provides a replicable approach for analyzing curriculum architecture in interdisciplinary program design.

2. METHODS

2.1 Research Design

This study adopts a quantitative structural evaluation design in which curriculum is conceptualized as a compositional system. Rather than treating programs as narrative descriptions of course content, each curriculum is modeled as a measurable distribution of competency domains. This approach enables systematic comparison of curriculum structures across institutions and supports quantitative evaluation of structural balance. Competency domains were derived from internationally recognized data science curriculum frameworks, including ACM Data Science Task Force (2021), the National Academies of Sciences, Engineering, and Medicine (2018) report, and the EDISON Data Science Framework (2017). These frameworks provide domain taxonomies but do not prescribe proportional allocation across domains. In this study, they serve as definitional references for domain construction rather than normative weighting standards. Each curriculum C_i is represented as a compositional competency vector:

$$C_i = (p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5}) \tag{1}$$

where p_{ij} denotes the normalized proportion of required coursework allocated to competency domain j, and:

$$\sum_{j=1}^{5} p_{ij} = 1 \tag{2}$$

The five domains are Programming and Data Handling, Statistical Foundations, Machine Learning and Data Mining, Applied Data Practice, and Ethics and Governance. This representation allows curriculum structures to be analyzed within a metric space of proportional distributions. An equal-weight benchmark vector is defined as:

$$B = (0.2, 0.2, 0.2, 0.2, 0.2) \tag{3}$$

to enable quantitative evaluation of curriculum structure, this study introduces the Curriculum Structural Balance Index (CSBI), which measures the degree of proportional deviation between each program's competency distribution and a neutral equal-weight benchmark. Lower CSBI values indicate more proportionally balanced distributions, while higher values indicate stronger structural concentration in selected domains. The detailed calculation and methodological justification are presented in the quantitative structural comparison subsection.

2.2 Data Collection and Sample Selection

Data were collected from publicly available institutional sources, including official program websites, curriculum structures, course catalogs, and published course descriptions. Data collection was conducted between December 2025 and January 2026 from Pangkalan Data Pendidikan Tinggi (PDDikti) website. Programs were included if they: (1) offered a standalone undergraduate data science degree or a formally designated data science track within a related discipline (e.g., computer science, informatics, statistics); (2) included at least three data science–related courses; and (3) addressed a minimum of three core competency domains.

Elective courses were excluded to ensure comparability based on required curricular structure. An initial screening identified 83 programs explicitly designated as Data Science or formally structured as data science tracks. Of these, 37 programs met the inclusion criteria, comprising 27 private universities and 10 public institutions. Figure 1 presents a donut chart showing the proportion of private and public institutions, while Figure 2 displays a bar chart of program distribution across provinces.

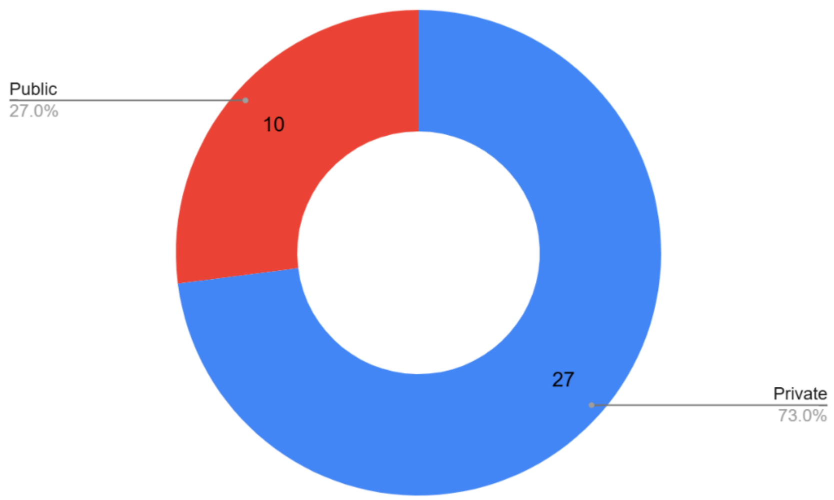

Figure 1. Distribution of Data Science Programs by Institution Type

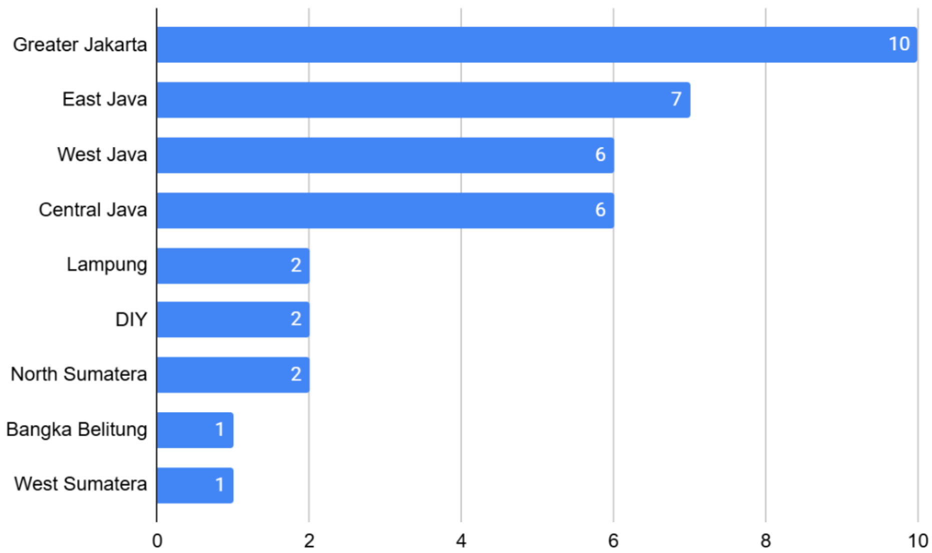

Figure 2. Distribution of Data Science Programs Across Provinces

The unit of analysis was the required course. Each required course was treated as a structural component contributing to the competency distribution of its respective program. This sampling approach ensures that all included programs meet a minimum threshold of disciplinary relevance while preserving comparability across institutional contexts. Courses were mapped to five core competency domains: Programming and Data Handling, Statistical Foundations, Machine Learning and Data Mining, Applied Data Practice, and Ethics and Governance. These domains synthesize competency categories identified across ACM, NASEM, and EDISON frameworks.

Course classification was based on analysis of course titles and official descriptions. When a course addressed multiple competencies, it was assigned to all relevant domains. This multi-label coding approach captures integrative course design while preserving proportional comparability across programs. A predefined coding rubric guided domain classification. The coding rubric was developed to ensure consistent classification across domains, with domain definitions aligned to established competency frameworks. This approach reduces subjectivity in course classification and strengthens the interpretability of the resulting competency distributions.

2.3 Quantitative Structural Comparison

For each program, the proportion of required courses associated with each competency domain was calculated relative to total required coursework. A multi-label coding scheme was employed to account for courses that addressed multiple competencies. Accordingly, proportions were calculated based on total domain assignments per program rather than unique course counts, ensuring that the resulting competency vector summed to one. To evaluate curriculum structure, CSBI was calculated using Manhattan (L1) distance between each program's competency distribution vector and the equal-weight benchmark distribution. Because curriculum was operationalized as proportional distributions summing to one, L1 distance provides a direct measure of aggregate proportional deviation across competency domains (Cha, 2007). Manhattan distance was selected because it is sensitive to overall proportional deviation while being less influenced by extreme values within individual dimensions compared with Euclidean distance.

The equal-weight benchmark distribution was adopted as a neutral reference condition rather than a normative model of ideal curriculum design. This approach enables consistent comparison of structural concentration and balance across institutions without imposing framework-specific weighting assumptions. Alternative approaches such as entropy-based measures or weighted models were not employed because the study focuses on interpretable proportional deviation rather than probabilistic dispersion or predefined normative weighting schemes.

Pairwise Manhattan distances were further computed to assess structural similarity across programs. The resulting distance matrix was analyzed using hierarchical clustering with average linkage to identify recurring curriculum typologies based on shared proportional competency distributions (Everitt et al., 2011). To improve procedural clarity and reproducibility, the overall analytical workflow was organized into four sequential stages. The research procedure consisted of four stages:

a. Identification and selection of undergraduate data science–related programs based on predefined inclusion criteria;
b. Collection of curriculum documents and course descriptions from publicly available institutional sources;
c. Classification of required courses into five competency domains using the predefined coding rubric and construction of normalized compositional competency vectors;
d. Quantitative analysis using CSBI calculation, pairwise Manhattan distance comparison, and hierarchical clustering to identify structural patterns across institutions.

2.4 Reliability and Validity

Course classification was conducted using a predefined coding rubric aligned with established competency frameworks, including ACM Data Science Task Force (2021), the EDISON Data Science Framework (2017), and the National Academies of Sciences, Engineering, and Medicine (2018). Domain definitions and classification criteria were explicitly operationalized prior to coding to ensure consistency across programs. To strengthen classification reliability, the coding process followed a rule-based audit procedure in which ambiguous courses were iteratively reviewed using course descriptions, learning objectives, and domain indicators specified in the rubric. Repeated consistency checks were conducted throughout the coding process to ensure stable classification decisions across institutions. Multi-label assignment was applied when courses addressed multiple competency domains, reducing the risk of forced single-domain categorization in interdisciplinary coursework.

While the study employed a single-rater coding process, inter-rater reliability testing was not applied because the analysis relied on publicly documented curriculum structures and an explicitly rule-based domain classification rubric rather than independent subjective judgment. To mitigate potential classification bias, the coding process incorporated systematic coding audits, repeated consistency checks, and iterative review of ambiguous cases using course descriptions, learning objectives, and predefined domain indicators. These procedures strengthened the transparency, objectivity, and reproducibility of the competency mapping process.

3. RESULT AND DISCUSSION

3.1 Structural Distribution of Competency Domains

Analysis of required curricula across the 37 programs reveals a clear structural hierarchy of competency domains. Programming and Data Handling and Statistical Foundations are consistently emphasized across institutions, indicating their role as core structural components of data science curricula. Machine Learning and Data Mining are also widely included, although their proportional allocation varies across programs. Figure 3 presents the average and median number of courses allocated to each competency domain. Programming and data handling exhibit the highest representation (mean = 6.4; median = 5.0), followed by statistical foundations (mean = 4.1), although the lower median (1.0) suggests uneven distribution across programs. Machine learning occupies a moderate position (mean = 3.7; median = 3.0), indicating relatively consistent inclusion with varying emphasis. In contrast, applied data practice (mean = 2.5; median = 2.0) and ethics and governance (mean = 0.6; median = 0.0) exhibit substantially lower representation.

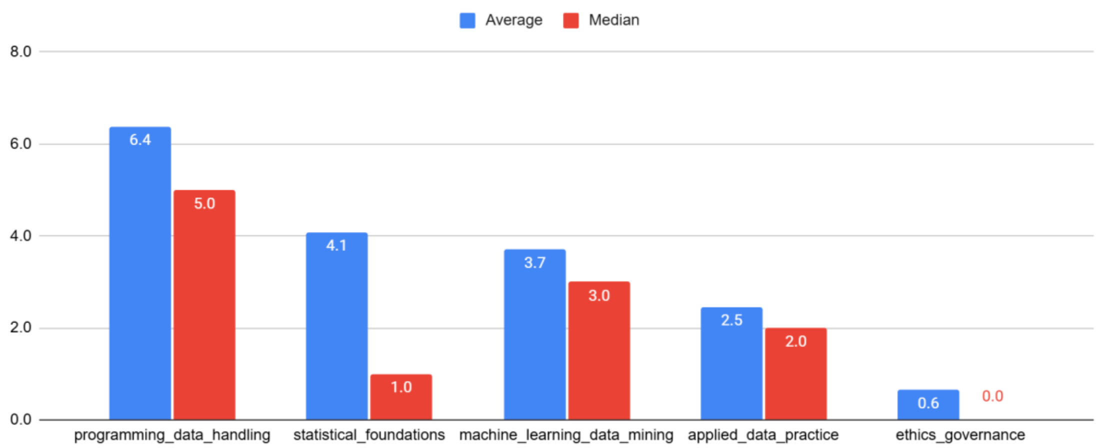

Figure 3. Average and Median Number of Courses by Competency Domain Across Programs

This distribution indicates a systematic structural imbalance in which technical domains dominate curriculum design, while integrative components remain comparatively underrepresented. The substantial gap between technical domains and integrative domains suggests that curriculum emphasis remains strongly oriented toward foundational computational competencies. In particular, the comparatively low representation of applied practice and ethics/governance indicates that interdisciplinary integration is often treated as a secondary structural component rather than a program-wide organizing principle. While such emphasis aligns with the disciplinary foundations of data science (Donoho, 2017; Hardin et al., 2015), it also reflects a prioritization of technical consolidation over cross-domain integration

3.2 Structural Balance and Concentration (CSBI)

To evaluate structural balance, CSBI was computed as the Manhattan (L1) distance between each program's competency distribution and the benchmark distribution. Figure 4 demonstrates a moderately right-skewed distribution of CSBI values, indicating that structurally concentrated curricula are more prevalent than highly balanced configurations across institutions. Most programs are distributed within the mid-range interval (approximately 0.6–0.9), suggesting recurring patterns of moderate concentration rather than extreme specialization. Only a limited number of programs exhibit relatively low CSBI values, indicating more proportionally distributed competency structures. Conversely, the upper tail of the distribution reflects a smaller subset of institutions with strong structural concentration, particularly in programming and statistical foundations.

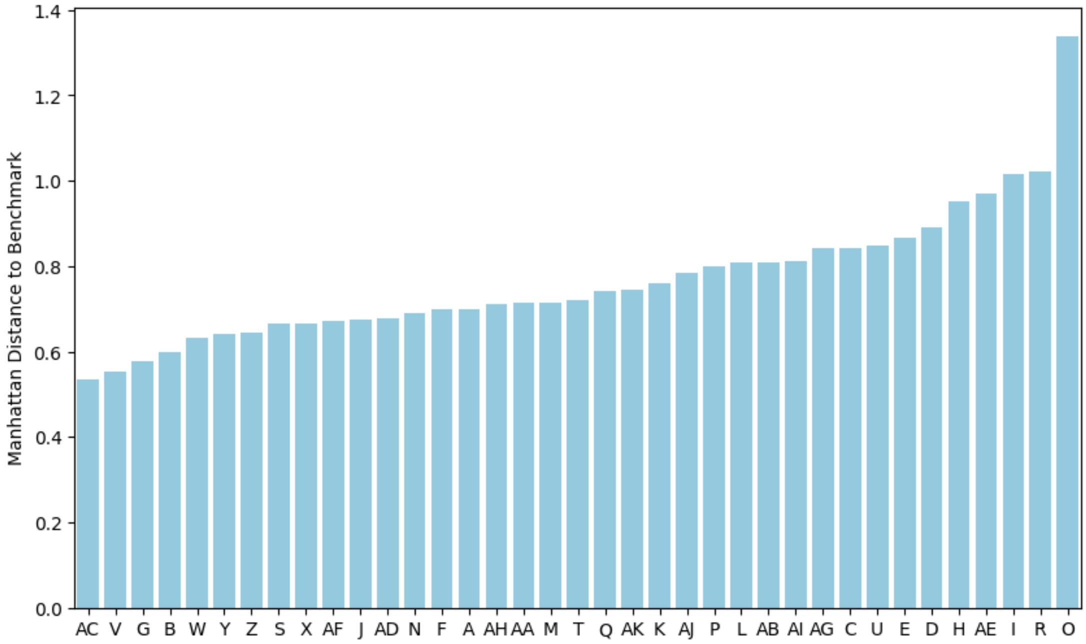

Figure 4. Distribution of CSBI Values Across Programs

Programs with lower CSBI values exhibit more evenly distributed competency structures, while higher values indicate stronger concentration in selected domains. These patterns suggest that while variation exists across programs, curriculum architectures tend to converge toward a shared foundational emphasis rather than highly diverse structural configurations. Importantly, these results quantify structural imbalance rather than merely describing it, enabling direct comparison of curriculum configurations across institutions. This extends prior descriptive analyses of data science education Deza & Deza (2016) by introducing a measurable framework for evaluating curriculum structure.

Vol. 9, No. 2, August 2026, pp. 518-530

3.3 Structural Typologies and Clustering Patterns

Hierarchical clustering based on Manhattan distance matrices reveals the presence of distinct curriculum typologies across institutions (Figure 5). Using average linkage, programs group into clusters that reflect shared structural emphasis patterns rather than random variation.

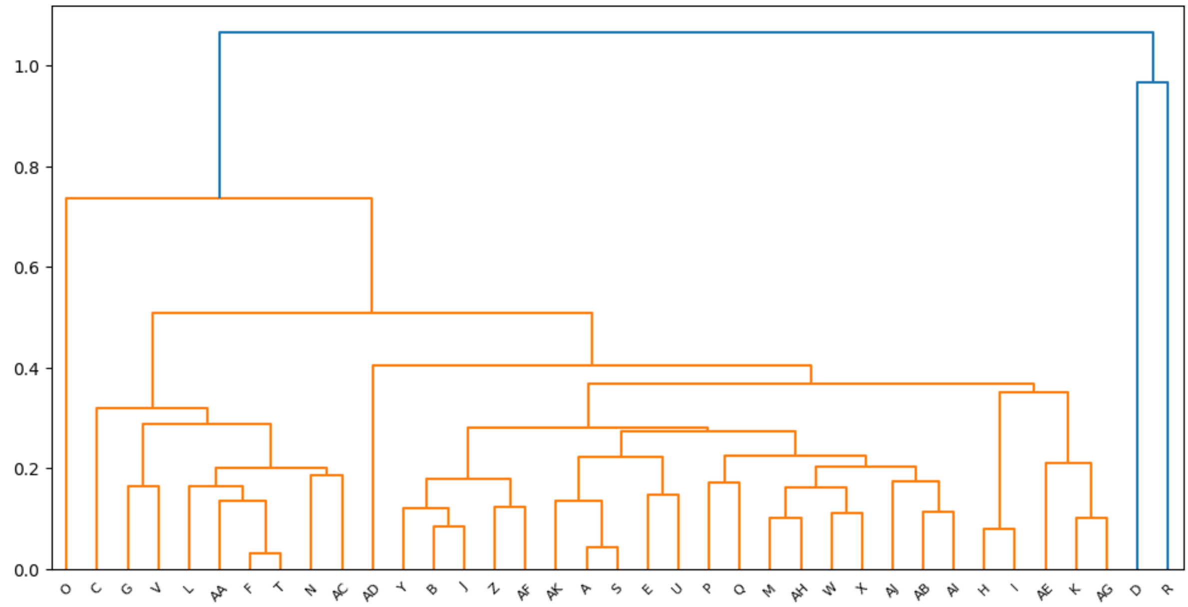

Figure 5. Hierarchical Clustering Dendrogram of Curriculum Structures

Three structural configurations emerge. The first cluster represents relatively balanced curricula with distributed competency emphasis. The second cluster reflects moderately concentrated structures, combining strong foundational emphasis with partial integration of applied domains. The third cluster exhibits highly concentrated configurations, where programming and statistical foundations dominate with minimal representation of applied practice and governance. These patterns demonstrate that curriculum structures are not randomly distributed across institutions but instead converge toward recurring proportional configurations characterized by varying degrees of structural concentration. This supports the identification of a Foundational-Dominant model as the prevailing structural pattern across programs.

The clustering results further clarify the structural meaning of CSBI values across programs. Programs with lower CSBI values tended to cluster within more proportionally distributed curriculum configurations, indicating relatively balanced competency allocation across domains. In contrast, programs with higher CSBI values were more frequently associated with clusters characterized by strong concentration in programming and statistical foundations. This relationship indicates that CSBI not only measures individual structural imbalance but also contributes to distinguishing recurring curriculum typologies across institutions. Accordingly, the integration of CSBI and clustering analysis enables identification of both the degree and the structural pattern of curriculum concentration within undergraduate data science programs.

3.4 Interpretation: The Foundational-Dominant Model

The combined evidence from domain distribution, CSBI analysis, and clustering indicates that undergraduate data science curricula converge toward a Foundational-Dominant structural model. This configuration is characterized by: (1) high proportional emphasis on programming and statistics; (2) consistent but variably weighted inclusion of machine learning; (3) limited structural integration of applied practice; and (4) compartmentalized treatment of ethics and governance. This pattern reflects strong technical consolidation but comparatively weak cross-domain integration. While foundational emphasis supports disciplinary grounding and aligns with established frameworks, excessive concentration may constrain the development of integrative competencies required for real-world data problem-solving.

3.5 Structural Integration Challenges: Applied Practice and Ethics

The underrepresentation of applied practice and governance domains reflects a broader challenge in data science education. International frameworks emphasize the importance of integrating applied and ethical dimensions throughout curricula (De Veaux et al., 2017; National Academies of Sciences, Engineering, and Medicine, 2018), yet empirical evidence shows these domains are often structurally marginalized (Oliver & McNeil, 2021). The present findings reinforce this pattern. Applied components are frequently positioned at later stages of programs, limiting iterative reinforcement, while ethics is typically treated as a standalone component rather than embedded across technical courses. As prior studies suggest, meaningful integration requires program-wide embedding rather than isolated inclusion (Baumer et al., 2022; Colando & Hardin, 2024). The issue is therefore not the absence of these domains, but their structural positioning. This affects how students experience interdisciplinarity, with technical competencies developed early and integrative competencies introduced later or peripherally.

The findings of this study are consistent with prior research indicating that undergraduate data science curricula tend to prioritize computational and statistical foundations over applied and ethical integration. Previous studies have similarly reported that programming, statistics, and machine learning frequently dominate curriculum structures, while ethics and interdisciplinary application receive comparatively limited emphasis (Oliver & McNeil, 2021; Baumer et al., 2022). The present study supports these observations while extending existing research through quantitative structural evaluation across institutions rather than descriptive curriculum mapping alone. Unlike earlier studies that primarily examined competency presence or pedagogical approaches, the current findings demonstrate how proportional concentration patterns systematically recur across curriculum architectures. This suggests that the imbalance identified in prior literature is not only curricular but also structural in nature.

3.6 Institutional Drivers of Curriculum Structure

Clustering patterns further suggest that institutional context influences curriculum structure. Programs embedded within computing-oriented departments tend to exhibit stronger concentration in programming and machine learning, consistent with computing-driven frameworks (ACM Data Science Task Force, 2021). In contrast, programs with interdisciplinary or business orientation display relatively more distributed profiles, although integration of governance remains uneven. These findings indicate that curriculum structure is shaped not only by pedagogical considerations but also by institutional capacity, disciplinary heritage, and resource allocation. In emerging higher education systems, such factors may strongly influence which competencies are structurally emphasized.

3.7 Implications for Curriculum Design

The findings suggest three directions for curriculum design. First, applied learning should be integrated progressively across multiple semesters rather than concentrated at the end of programs. Second, ethical reasoning should be embedded within technical coursework instead of isolated in standalone modules. Third, institutions should adopt periodic structural evaluation using competency-distribution metrics such as CSBI to assess curriculum balance. Overall, while strong foundational emphasis provides technical robustness, proportional distribution determines how interdisciplinarity is experienced throughout the curriculum. Structural evaluation therefore complements existing competency frameworks by revealing patterns of emphasis not captured through coverage-based analysis (Hsu, 2024).

4. CONCLUSION

This study demonstrates that undergraduate data science curricula in Indonesian universities exhibit recurring patterns of structural concentration when evaluated as compositional competency systems. Across institutions, programming and statistical foundations consistently dominate curriculum structures, while applied practice and ethics/governance remain comparatively underrepresented. The clustering results further reveal that these distributions are not institution-specific anomalies but recurring structural configurations shared across programs. The findings support the identification of a Foundational-Dominant curriculum model characterized by strong emphasis on technical foundations combined with limited proportional integration of interdisciplinary and professional competencies. Theoretically, this study contributes by conceptualizing curriculum as a compositional system in which competency domains function as interdependent structural components rather than isolated content categories. Through the introduction of the Curriculum Structural Balance Index (CSBI), the study further provides a quantitative framework for evaluating proportional curriculum balance and structural concentration across institutions.

The results suggest that curriculum quality in interdisciplinary fields depends not only on competency inclusion but also on how competencies are proportionally distributed throughout program structures. Accordingly, curriculum developers should move beyond additive curriculum design approaches toward more integrated proportional models. Applied learning components should be distributed progressively across multiple semesters rather than concentrated at the end of programs, while ethical and governance competencies should be embedded within technical coursework through cross-subject integration strategies. Institutions may also use CSBI-based evaluation periodically to assess curriculum balance and identify structural concentration patterns during curriculum revision processes. Several limitations should be acknowledged. The study evaluates intended curriculum structures derived from publicly available documentation and does not assess instructional implementation, pedagogical quality, or student learning outcomes. Future research should therefore extend this framework through longitudinal analysis, cross-national comparison, and outcome-based evaluation linking curriculum structure with graduate competencies and professional readiness.

ACKNOWLEDGEMENT

The author acknowledges Universitas Ciputra for academic supports and resources provided during this research. The author also appreciates the availability of institutional curriculum data from PDDikti and Indonesian universities, which enabled this study to be conducted.

REFERENCES

ACM Data Science Task Force. (2021). *Computing competencies for undergraduate data science curricula (CCDS 2021)*. Association for Computing Machinery. https://www.acm.org/media-center/2021/march/data-science-curricula-2021

Avila-Garzon, C., & Bacca-Acosta, J. (2025). Curriculum, pedagogy, and teaching/learning strategies in data science education. *Education Sciences, 15*(2), 186. https://doi.org/10.3390/educsci15020186

Baumer, B. S., Garcia, R. L., Kim, A. Y., Kinnaird, K. M., & Ott, M. Q. (2022). Integrating data science ethics into an undergraduate major: A case study. *Journal of Statistics and Data Science Education, 30*(1), 15–28. https://doi.org/10.1080/26939169.2022.2038041

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics, 1*(1), 61–65. https://doi.org/10.1007/s43681-020-00002-7

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences, 1*(4), 300–307. https://www.fisica.edu.uy/~cris/teaching/Cha_pdf_distances_2007.pdf

Colando, S., & Hardin, J. (2024). Philosophy within data science ethics courses. *Journal of Statistics and Data Science Education, 32*(4), 361–373. https://doi.org/10.1080/26939169.2024.2394542

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B., Bray, A., Bressoud, T., Bryant, L., Cheng, L., Francis, A., Gould, R., & Kim, A. Y. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application, 4*, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Deza, M. M., & Deza, E. (2016). *Encyclopedia of distances*. Springer. https://doi.org/10.1007/978-3-662-52844-0

Dogucu, M., et al. (2024). A systematic literature review of undergraduate data science education research. *arXiv preprint*. https://doi.org/10.48550/arXiv.2403.03387

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Donovan, D., Clemmons, A., & Crowe, A. (2022). Data-driven approach to analyze competency teaching in an undergraduate biology program: A case study. *Journal of the Scholarship of Teaching and Learning, 22*(4), 50–71. https://doi.org/10.14434/josotl.v22i4.33192

EDISON Project. (2017). *EDISON data science framework (EDSF)*. https://edison-project.eu/edison/edison-data-science-framework-edsf/

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley. https://doi.org/10.1002/9780470977811

Fadillah, M. A. A., et al. (2025). Bibliometric mapping of data science in education: Trends, benefits, challenges, and future directions. *Social Sciences & Humanities Open, 11*(1), 101600. https://doi.org/10.1016/j.ssaho.2025.101600

Harden, R. M. (2001). Curriculum mapping: A tool for transparent and authentic teaching and learning. *Medical Teacher, 23*(2), 123–137. https://doi.org/10.1080/01421590120036547

Hardin, J., Hoerl, R., Horton, N. J., & Nolan, D. (2015). Data science in statistics curricula: Preparing students to think with data. *The American Statistician, 69*(4), 343–353. https://doi.org/10.1080/00031305.2015.1077729

Holmes, W., Anastopoulou, S., Schaumburg, H., & Mavrikis, M. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Hsu, Y.-C. (2024). Mapping the landscape of data science education in higher education: A comprehensive syllabi analysis. *Education Sciences, 14*(7), 763. https://doi.org/10.3390/educsci14070763

Li, M., & Yu, T. (2022). Methodological issues on evaluating agreement between two detection methods by Cohen's kappa analysis. *Parasites & Vectors, 15*(1), 270. https://doi.org/10.1186/s13071-022-05402-8

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. National Academies Press. https://www.nationalacademies.org/projects/DEPS-CSTB-16-01/publication/25104

Oliver, J. C., & McNeil, T. (2021). Undergraduate data science degrees emphasize computer science and statistics but fall short in ethics training. *PeerJ Computer Science, 7*(1), 441. https://doi.org/10.7717/peerj-cs.441

Ran, X., et al. (2023). Comprehensive survey on hierarchical clustering algorithms. *Artificial Intelligence Review, 56*(8), 8219–8264. https://doi.org/10.1007/s10462-022-10366-3

Schmitt, K. R. B., et al. (2025). Evaluation of EDISON's data science competency framework. *Foundations of Data Science, 5*(2), 177–198. https://doi.org/10.3934/fods.2021031

Smaldone, F., et al. (2022). Employability skills: Profiling data scientists in the digital labour market. *European Management Journal, 40*(5), 671–684. https://doi.org/10.1016/j.emj.2022.05.005

Worden, J. V., & Bray, M. J. (2025). A theory of change approach to curricular revision. *BMC Medical Education, 25*(1), 533. https://doi.org/10.1186/s12909-025-07100-2

Xu, C., & Wu, C.-F. (2025). Interdisciplinary integration competencies in curriculum patterns. *Humanities & Social Sciences Communications, 12*(1), 1683. https://doi.org/10.1057/s41599-025-05950-1