Identification of Latent Dimensions of Digital Readiness and Typology of Districts/Cities in Indonesia Using PCA and K-Means Clustering

Jefita Resti Sari[1], Fani Fahira[2]*, Latifah Zahra[3]*, Anwar Fitrianto[4]*, Erfiani[5]*, Kevin Alifviansyah[6]***
*School of Data Science, Mathematics and Informatics, IPB University
jefitarestisari@apps.ipb.ac.id [1], fanifahira@apps.ipb.ac.id [2], zahralatifah@apps.ipb.ac.id [3],
anwarstat@apps.ipb.ac.id [4], erfiani@apps.ipb.ac.id [5], kevinalifviansyah@apps.ipb.ac.id [6]

Article Info

Article history:

Received 2025-10-13
Revised 2025-11-02
Accepted 2025-11-08

Keyword:

IMDI,
Regional Typology,
Digital Readiness,
Principal Component Analysis,
K-Means Clustering.

ABSTRACT

Digital transformation is a key agenda in Indonesia's national development that requires balanced readiness across regions. However, the level of digital readiness among districts and cities still varies widely, highlighting the need for a typology that can comprehensively describe existing disparities. This study aims to identify the latent dimensions of digital readiness and to develop a regional typology of Indonesian districts/cities using Principal Component Analysis (PCA) and K-Means clustering. The data were obtained from the 2024 Indonesian Digital Society Index (IMDI), which consists of four pillars—Infrastructure and Ecosystem, Digital Skills, Empowerment, and Employment—with ten sub-pillars. PCA reduced these correlated indicators into two main latent components, namely Digital Capacity and Participation and Digital Infrastructure Foundation, which together explain 70.4% of the total variance. Cluster validation using the Silhouette Score and Davies–Bouldin Index (DBI) showed that K = 2 yielded the best internal validity (Silhouette = 0.402; DBI = 0.906), but a three-cluster configuration (K = 3) was adopted to obtain a more interpretable typology of high-, medium-, and low-readiness regions (Silhouette = 0.346; DBI = 1.007). Spatial mapping reveals that high-readiness districts are concentrated in Java, Bali, and parts of Sumatra, whereas low-readiness areas dominate eastern Indonesia. These findings confirm persistent digital inequality across regions and provide a quantitative basis for targeted policy interventions, including infrastructure development, digital literacy programs, and innovation ecosystem strengthening, to support an inclusive digital transformation in Indonesia.

I. INTRODUCTION

Digital transformation is a global phenomenon that has fundamentally changed the economic, social, and governmental landscape. Digital readiness is an important factor for the progress of a country, including Indonesia, in facing challenges and taking advantage of opportunities in the use of information and communication technology (ICT) [1]. Digital readiness encompasses various aspects, ranging from ICT infrastructure and public digital literacy to technology adoption in the public and private sectors [2], [3]. The Indonesian government has demonstrated its commitment to accelerating digital transformation through various initiatives aimed at improving the efficiency of public services, transparency, and accountability, while expanding digital-based service and transaction channels. However, at the implementation level, these efforts still face obstacles in the form of digital literacy gaps, resistance to change, and issues of uneven infrastructure readiness between regions [1], [2], [4].

The digital divide remains a significant challenge in Indonesia. Through mapping the Indonesian Digital Society Index (IMDI) 2022–2024, the study found an increase in the aggregate IMDI value, but with an uneven spatial distribution, with digital transformation hotspots concentrated on the island of Java, while the eastern regions lagged behind. This gap is evident in the four pillars of the IMDI, namely Infrastructure and Ecosystem, Digital Skills, Empowerment, and Employment [5]. Similar findings were obtained, which confirm that digital disparities are closely related to factors such as education, income, and the structure of the formal workforce [6], [7].

However, most of the existing research has been limited to the provincial or sectoral level and has mainly analyzed correlations among IMDI indicators without exploring their latent structure. Therefore, there remains a scientific gap in understanding how the IMDI pillars collectively define Indonesia's digital readiness. This study addresses that gap by applying Principal Component Analysis (PCA) to uncover the underlying latent dimensions of digital readiness and by using K-Means Clustering to classify districts/cities into typologies that can inform policy formulation at the national level.

The phenomenon of digital inequality is not unique to Indonesia. A European study based on the Digital Economy and Society Index (DESI) found that countries can be clustered into several levels of digitization, with limited transition between clusters [8], [9]. Also compared Average Linkage and K-Means on provincial welfare indicators in Indonesia, and showed that the clustering method was able to reveal disparities between regions. This reinforces the urgency of data-based typologies to address digital development gaps [10].

This condition has the potential to deepen development disparities and hamper the effectiveness of digital-based public policy implementation, especially in areas with low technology adoption capabilities [1]. Therefore, the role of local governments (regencies/cities) is important in encouraging a more equitable acceleration of digital transformation, with strategies that are in line with local characteristics.

Measuring digital readiness across districts/cities requires an analytical approach that can handle many variables that are interrelated and potentially redundant. Principal Component Analysis (PCA) is relevant for reducing dimensions and extracting latent data structures, thereby facilitating interpretation without losing important information, while K-Means is capable of grouping regional units based on the proximity of characteristics in the main component space [11], [12]. PCA was chosen for its ability to reduce multicollinearity and to simplify multidimensional indicators into a few interpretable latent components, while K-Means enables unsupervised classification of regions with similar readiness profiles.

Various studies show used PCA and K-Means to profile districts/cities in Central Java based on poverty indicators [13]. Applied PCA for waste management zoning in Tapin District, classifying three types of zones which were then used to determine service priorities [14]. Combined K-Means and PCA to group cities/districts socio-economically and showed how the elbow method helps select an economical number of clusters, as well as in other domains such as plantation productivity [15], crime [16], and public health [17]. Although the domains are different, poverty, waste management, and economics use the same approach, namely correlation reduction, dimension interpretation, clustering, and recommendations [18].

At the clustering stage, selecting the number of clusters (K) and validating cluster quality are important for viewing the modelling results. The Elbow Method is used to detect the optimal point [19], [20], while the Silhouette Score assesses the quality of separation between clusters[21]. The Davies-Bouldin Index (DBI) is also commonly used to measure cohesion and separation, where a low value indicates a good cluster [17], [21]. Compared the Elbow and Silhouette Scores in a multisector dataset, showing that both are generally consistent but can differ in certain data. When this occurs, consideration of the context/domain and visualization (e.g., in PCA space) helps decide on the best result [20]. Added the Davies-Bouldin Index (DBI) as a metric for cluster separation/compactness, with a lower DBI indicating a better cluster structure [21]. Improvements in the initialization aspect are also important, the K-Means++ method improves the initial centroid determination so that the results are more stable, and can be combined with Silhouette, Elbow, and even Gap Statistics to determine K [17]. In fact, in the case of the agribusiness industry, it shows that selecting K with Elbow and Silhouette produces K=2, which is operational for location productivity improvement strategies. This approach is parallel to the idea of public policy typology with the acquisition of an optimal number of clusters and produces informative results [15].

From a methodological perspective, research confirms the close relationship between PCA and K-Means [11], while developed a hybrid PCA–KMeans model to improve the interpretability of results [12]. International studies highlight the importance of cross-border digitalization mapping [8], [9], while research in Indonesia by emphasize the relevance of clustering methods in analysing regional disparities [4], [22]. The urgency of developing a digital readiness typology system arises from the need to formulate targeted policies. Regional clustering can help the government identify priority areas, design needs-based interventions, and allocate resources more efficiently. In a broader socio-economic context, grouping districts and cities enables policymakers to pinpoint regions that are most in need of strategic intervention [13]. In the context of other public services, classification based on multivariate indicators helps to prioritize service and infrastructure development [14]. In the realm of government digital transformation, a case study at the East Java Regional Revenue Agency shows the potential for increased administrative efficiency, transparency, and local revenue through process digitalization, but also emphasizes the importance of bureaucratic readiness and public digital literacy to ensure that the benefits are evenly distributed [2]. At the same time, discourse studies on digital society emphasize the need for communication, information, and education strategies to bridge the literacy gap that hinders the use of technology for welfare [1]. These findings reinforce the argument that a one-size-fits-all approach is inadequate.

Instead, a data-based typology is needed to link policy interventions to the readiness profile of each region.

Existing Indonesian studies on digital readiness typically examine spatial patterns of IMDI and detect hotspot/low-spot clusters over time, but they rarely uncover the latent structure that jointly organizes IMDI's pillars into a small number of interpretable dimensions for nationwide policy design. Recent spatiotemporal work confirms persistent regional disparities and clustering around Java versus eastern Indonesia [5], while socio-economic clustering studies at the district level mostly rely on direct indicators or correlations without explicit dimension reduction [18]. To address this gap, we apply PCA to reveal latent components of digital readiness and then derive a policy-oriented typology using K-Means clustering across all districts/cities.

Based on the above description, this study aims to identify the latent dimensions of digital readiness in districts/cities in Indonesia using PCA and to develop a regional typology based on the PCA results using the K-Means clustering method. This study is expected to provide an empirical basis for the government and policymakers in designing focused and effective interventions, so that the acceleration of digital transformation can reach all regions of Indonesia in an inclusive and equitable manner.

II. Method

A. Data

The data used in this study is primary data from the *Indeks Masyarakat Digital Indonesia* (IMDI), or the Indonesian Digital Society Index, to assess digital disparities in various regencies in Indonesia [23]. The data covers 514 districts/cities in Indonesia with 10 sub-pillars from the four pillars of IMDI in 2024. The IMDI is described as having four main pillars: Infrastructure and Ecosystem, Digital Skills, Empowerment, and Employment. The four main pillars are further divided into several sub-pillars that represent various important factors in the development of a digital society. The analysis was performed using Python 3.12 with the pandas, numpy, scikit-learn, and matplotlib libraries.

TABLE I
PILLARS, SUB-PILLARS, AND INDICATORS OF THE INDONESIAN DIGITAL SOCIETY

Pillar	Sub-Pillars	Indicator
Infrastructure and Ecosystem (P1)	(P1.1) Access and Adoption of Digital Technology	Access and use of ICT
		Technology Adoption
	(P1.2) Learning Ecosystem	Numbers of schools from elementary to secondary level with internet access
		Number of ICT higher education institutions
	(P1.3) Government Digitalization	Data of the Electronic-Based Government System Index (SPBE)
Digital Skills (P2)	(P2.1) Complementarity	Respondent's perceptions of communication and Collaboration
		Ability to think critically
	(P2.2) Introduction to ICT	Ability to identify and using ICT tools
		Data literacy
	(P2.3) Security	Digital equipment and content security
		Personal device security
Empowerment (P3)	(P3.1) Users / Consumers	E-commerce consumers
		E-learning users
	(P3.2) Providers / Sellers	Digital financial providers
		E-commerce sellers
		Social Media content creator
		E-learning providers
Employment (P4)	(P4.1) Demand	Digital skills training
		Occupations and digital skill levels
		Level of automation
	(P4.2) Supply	Internet in the workplace
		Availability of digital skills
		Proportion of individuals
		Skills Training

B. *Kaiser-Meyer-Olkin* (KMO) and *Bartlett's Test of Sphericity*

To test the suitability of data before conducting factor analysis, two main tests are used, namely *Kaiser-Meyer-Olkin* (KMO) dan *Bartlett's Test of Sphericity*. The KMO test is used to determine whether the data sample is adequate and suitable for factor analysis. The KMO values ranges from 0 to 1, with an interpretation that a KMO value ≥ 0.6 indicates that the data is suitable for further analysis, while a value below 0.5 indicates that the data is inadequate [24]. Mathematically, the KMO formula is as follows.

$$KMO = \frac{\sum_{i\neq j} r_{ij}^2}{\sum_{i\neq j} r_{ij}^2 + \sum_{i\neq j} p_{ij}^2} \tag{1}$$

Meanwhile, *Bartlett's Test of Sphericity* is used to test the null hypothesis that the correlation matrix is an identity matrix, which means that there is no significant relationship between variables $(p < 0{,}05)$, then the null hypothesis is rejected, so that there is sufficient correlation between variables to conduct factor analysis [24].

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right)\ln|R| \tag{2}$$

The significant value of χ^2 $(p < 0{,}05)$ indicates that the correlation matrix is not an identity matrix, so the data is suitable for further analysis using factor analysis.

C. Principal Component Analysis

Principal Component Analysis (PCA) is widely method used for dimension reduction to simplify correlated variables into a small number of independent principal components without significant loss of information [25]. The goal of PCA is to identify latent dimensions that can explain the variance structure of the data, thus facilitating interpretation and further analysis, such as clustering.

PCA efficiently handles multicollinearity among IMDI indicators and summarizes the multidimensional space into a few latent components that capture the main axes of variation. This improves interpretability and provides noise-reduced inputs for clustering. K-Means is then employed for unsupervised classification due to its scalability and clear centroid-based groups, a combination widely used in public-sector and regional studies [14], [15], [18]. To ensure robustness, we evaluate the optimal K using Elbow, Silhouette, and Davies-Bouldin Index (DBI) and discuss cases where metrics may disagree and domain considerations guide the final K [19], [20].

Mathematically, PCA is obtained through the eigenvalue decomposition of the covariance or correlation matrix. Suppose there is a standardized data matrix X with the size of $n \times p$, then the covariance matrix is:

$$S = \frac{1}{n-1} X^T X \tag{3}$$

The eigen value (λ_i) of the covariance matrix S states the amount of variance that can be explained by i-th component. The eigenvector e_i associated with λ_i is called the principal component.

The i-th component can be written as follows.

$$Z_i = Xe_i \tag{4}$$

Where Z_i is the factor score of each observation.

Research by [26], PCA is compared to autoencoders and other nonlinear techniques, and although autoencoders sometimes excel in representation flexibility, PCA still offers much higher computational efficiency with comparable classification results. Compared to other reduction techniques, PCA appears to be more stable and consistent in producing meaningful representations of the data and supports classification models well [27].

D. K-Means Clustering

K-Means is one of the Unsupervised Learning algorithms that is widely used to solve clustering problems. The procedure of this algorithm is to classify a data set into a certain number of clusters that have been determined beforehand. The main idea is to define k centroids, one for each cluster. These centroids are placed carefully because different locations will produce different results. Therefore, the better option is to place them as far apart from each other as possible [28].

The next step is to take each data point from the data set and connect it to the nearest centroid. When there are no more unprocessed data points, the first step is complete and an initial group has been formed. At this point, it is necessary to recalculate the k centroids as the centres of the clusters generated in the previous step. Once this new k centroid is obtained, regrouping between the data points and the new centroid is performed. This process will continue to repeat in a loop. As a result of this repetitive process, it will be seen that the centroid will change position gradually until it no longer changes. In other words, the centroid will stop moving.

This algorithm aims to minimize the objective function, in this case the Squared Error Function. The objective function can be written as follows.

$$W(S,C) = \sum_{K=1}^{K} \sum_{i \epsilon S_k} ||y_i - c_k||^2 \tag{5}$$

Where S is the partition into k clusters of the set of entities represented by the vectors y_i in the M-dimensional feature space, consisting of S_k cluster that non-empty and non-overlapping clusters, each cluster having a centroid c_k for k = 1, 2,....K.

E. Final Evaluation

The evaluation to determine the optimal number of clusters in the typology of digital readiness of districts/municipalities in Indonesia uses two validation metrics, namely Silhouette Score and Davies-Bouldin Index (DBI). The Silhouette Score assesses the degree of closeness of objects to the cluster they belong to, while the DBI measures the balance between clusters and the separateness between clusters. Both metrics are calculated using the following equations [29]:

$$DBI = \frac{1}{n}\sum_{i=1}^{n} \underset{i \neq j}{max} \left(\frac{S_i + S_j}{d_{ij}}\right) \tag{6}$$

With

n : numbers of clusters

S_i : Average distance between each point within clusters i to its centroid (*intra-cluster dispersion*)

d_{ij} : Distance between centers of cluster i and j (*inter-cluster distance*)

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{7}$$

with

$S(i)$: silhouette score for data i

$a(i)$: the average distance between point i and all points in the same cluster (*cohesion*)

$b(i)$: the average distance between point i and all points in the nearest other cluster (*separation*)

The smaller the DBI value, the better the cluster quality (compact and separated). Meanwhile, the Silhouette Score is between -1 and +1. These two evaluation metrics produce the best number of clusters that can be determined so that the resulting regional typology is representative and in line with the research objectives. [29].

F. Analysis Procedure

This research is exploratory because it aims to find latent patterns and typologies of digital readiness without any initial hypothesis regarding the number of factors or the number of clusters. The entire analysis was carried out in stages, from data cleaning to interpretation of results, with the following stages.

1) Data Pre Processing and Eksplorating

The first step was conducted to ensure the quality of the data to make it suitable for analysis. Missing values were checked for each variable and then imputed using median imputation to keep the distribution stable. Outlier detection is done through boxplot visualization, and their influence is minimized by scale transformation using Robust Scaler, so that extreme values do not distort the multivariate analysis.

2) Feasibility Test for Factor Analysis (PCA)

Before PCA was applied, the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity were conducted. The KMO value is used to assess whether the data has sufficient sample adequacy, while Bartlett's test is used to ensure that there is a correlation between variables.

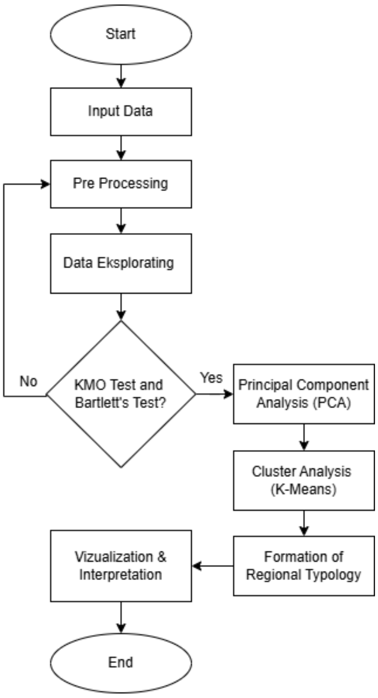
Start
Input Data
Pre Processing
Data Eksplorating
KMO Test and Bartlett's Test?
No
Yes
Principal Component Analysis (PCA)
Cluster Analysis (K-Means)
Formation of Regional Typology
Vizualization & Interpretation
End

Figure 1. Research Flowchart

3) Principal Component Analysis (PCA)

PCA was applied to reduce the correlated IMDI pillar and sub-pillar variables into a number of more concise latent factors. The number of components was determined using a scree plot and cumulative variance proportion. Next, varimax rotation was applied to clarify factor interpretation, so that each pillar or pillar could be identified as having a dominant contribution to a particular factor. The PCA results per pillar were used to form pillar scores, while PCA between pillars produced latent dimensions of digital readiness that formed the basis for typology analysis.

4) Cluster Analysis with K-Means

The factor scores from PCA were used as input for cluster analysis using the K-Means algorithm. The determination of the number of clusters was done exploratively by testing several alternative K values (2-6). Evaluation of cluster quality was conducted through three measures, namely Elbow Method (WCSS), Silhouette Score, and Davies-Bouldin Index (DBI).

5) Establishment of Regional Typology

Regencies and cities were grouped into three main typologies: High-readiness Cluster (regions with positive pillar scores across all dimensions), Medium-readiness Cluster (regions with moderate pillar scores), and Low-readiness Cluster (regions with negative pillar scores in most pillars). The differences in the characteristics of each typology were explored further by calculating the average pillar and sub-pillar scores for each cluster. In this way, it was possible to identify which pillars are the strengths and weaknesses of each typology.

6) *Visualization and Interpretation*

To clarify the results, visualisation is carried out in two forms:

1. A heatmap of average pillar scores per cluster, which shows the comparative differences in strength between pillars.
2. An interactive spatial map (Folium), which shows the geographical distribution of district/city typologies in Indonesia. This visualization is important to reveal spatial patterns and inequalities between regions.

III. Result and Discussion

The research results were obtained through the analysis stages described in the methodology section. The analysis was carried out in stages as follows.

A. *Data Exploration*

Data exploration was conducted to review the relationships between variables in the digital society index. This step aimed to identify correlation patterns that could indicate a close relationship between the pillars of digital readiness.

The Figure 2 illustrates the correlation matrix among the pillars of digital society development. It shows that all pillars are positively correlated, indicating that improvements in one aspect tend to be accompanied by advancements in others. The strongest correlations are observed between the Empowerment Pillar and the Digital Society Index (0.80), as well as between the Digital Skills Pillar and the Digital Society Index (0.78). This suggests that community empowerment and digital skills play a crucial role in shaping the overall progress of a digital society.

In addition, the Infrastructure and Ecosystem Pillar also exhibits a relatively strong relationship with the Digital Society Index (0.76), highlighting that well-established infrastructure and supportive digital ecosystems serve as fundamental enablers of digital development. Meanwhile, the Employment Pillar shows comparatively lower correlations with the other pillars, particularly with the Infrastructure and Ecosystem Pillar (0.01), indicating that digital-based employment opportunities are not yet fully dependent on the existing digital infrastructure or ecosystem.

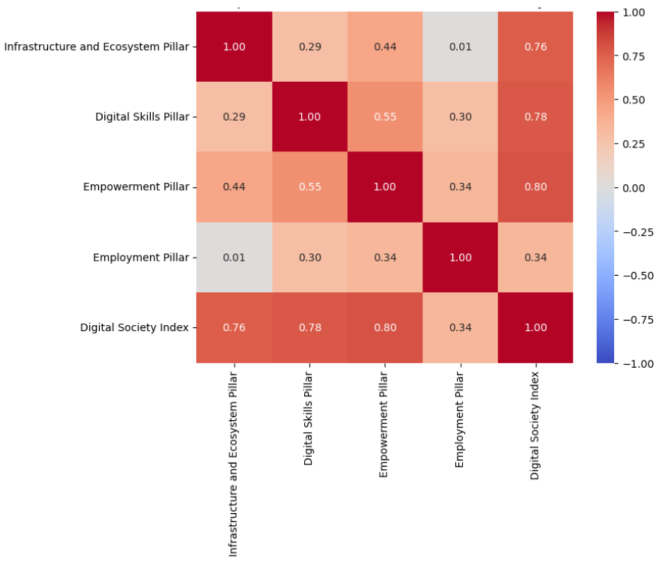

Figure 2. Heatmap of Correlation between Main Pillars and the Digital Society Index

Overall, the correlation pattern reinforces that digital society development is multidimensional and interdependent. Enhancing the digital society index requires not only strengthening infrastructure but also improving digital skills, empowering individuals, and fostering technology-driven employment opportunities.

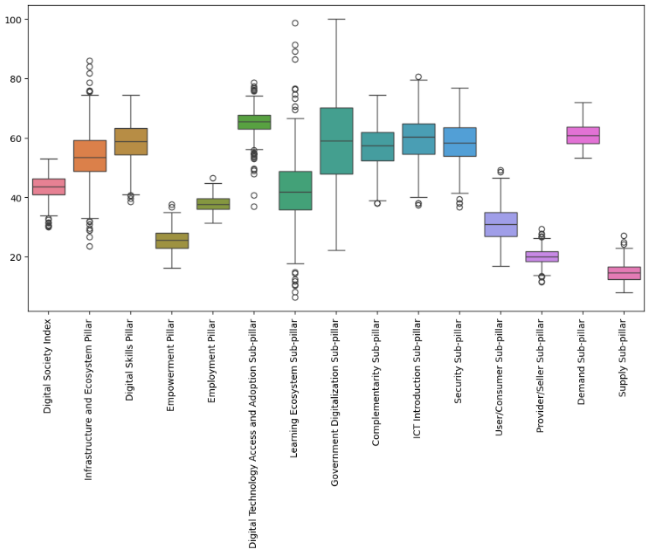

Figure 3. Subpillar Outlier Boxplot

The digital readiness data showed a heterogeneous distribution among variables, with sharply different medians between pillars, many significant outliers, and skewness. This justifies the need for PCA to reduce the dimension and K-Means with metric validation to find a robust cluster structure.

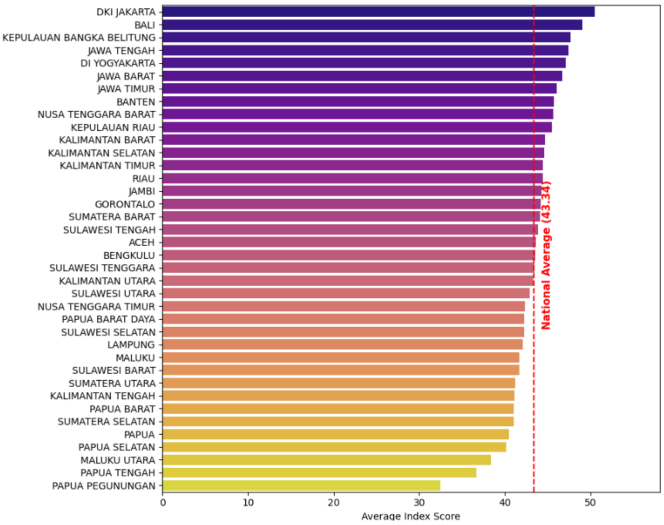

Figure 4. Average Digital Society Index Rankings by Province

By looking at the average ranking of the digital society index across Indonesian provinces, it can be seen that DKI Jakarta, Bali, and Bangka Belitung Islands occupy the highest positions, reflecting relatively better digital readiness, while provinces in the eastern region such as South Papua, North Maluku, Central Papua, and Papua Mountains are at the lowest position. The red dashed line marks the national average of 43.34, which shows that most provinces in Java, Bali, and some provinces in Sumatra are above average, indicating a digital divide between regions, especially between western and eastern Indonesia.

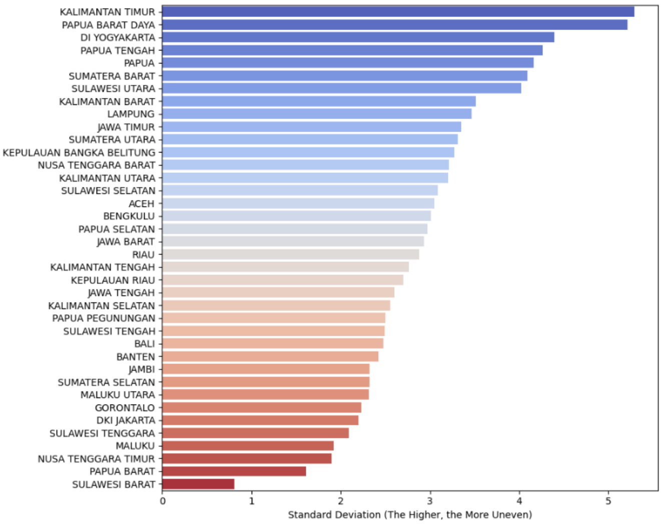

Figure 5. Level of Disparity (Std Dev) in Provincial Internal Digital Index

The level of internal digital disparity in each province, measured by the standard deviation of the digital index between districts/cities. Provinces such as East Kalimantan, Southwest Papua, DI Yogyakarta, Central Papua, and Papua have high standard deviations, indicating a large unevenness of digital achievement within the province. In contrast, provinces such as West Sulawesi, West Papua, East Nusa Tenggara, Maluku and Southeast Sulawesi show low standard deviations, meaning that digital achievement between regions is relatively more evenly distributed although the average score may still be low. In general, this graph emphasizes that the challenge of digital development lies not only in increasing the average index per province, but also in equitable distribution of achievements within the province itself.

B. Kaiser-Meyer-Olkin (KMO) and Bartlett's Test of Sphericity

Before PCA is applied, Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity are performed. The KMO value is used to assess whether the data has sufficient sample adequacy, while the Bartlett's test is used to confirm the correlation between variables. Based on the calculation results, the KMO value obtained is 0.831. This value is above the minimum threshold of 0.6, so it can be categorized as adequate according to Kaiser's criteria. This means that the data used in this study have good sample adequacy and support the application of factor analysis. On the other hand, the Bartlett test results produced a significance value of $p < 0.05$, which means that the null hypothesis is rejected. Thus, the correlation matrix formed between subpillars are not identity and are proven to have a relationship that is worthy of analysis.

The combination of these two test results indicates that the data is eligible to proceed to the Principal Component Analysis (PCA) stage. A high KMO value indicates consistent interrelationships between variables, while the significance of Bartlett's test reinforces the evidence of significant correlations. Therefore, the use of PCA in this study can be methodologically justified, both in terms of sample adequacy and the feasibility of correlation between variables.

C. Principal Component Analysis

Principal Component Analysis (PCA) with the aim of reducing the Indonesian Digital Society Index (IMDI) indicators consisting of pillars and sub-pillars. Because several indicators exhibited multicollinearity, all numerical variables were first standardized using the RobustScaler (scikit-learn 1.5) to minimize the influence of extreme values while preserving the central tendency of the data. This preprocessing ensured that each indicator contributed proportionally to the resulting components. PCA was then performed on the correlation matrix to transform the original variables into a smaller number of independent principal components (latent dimensions) that retained most of the original variance. In this study, the component scores obtained from PCA were used as inputs for the subsequent K-Means clustering to identify regional typologies of digital readiness.

TABLE II
PCA ON IMDI PILLARS AND SUB-PILLARS

Pillar	Variance Explained	Sub-pillar	Loading
Infrastructure & Ecosystem	0.644	Access and Adoption of Digital Technology	0.456
		Learning Ecosystem	0.640
		Government Digitalization	0.618
Digital Skills	0.829	Complementarity	0.583
		Introduction to ICT	0.562
		Security	0.587
Empowerment	0.778	Users / Consumers	0.707
		Providers / Sellers	0.707
Employment	0.508	Demand	-0.707
		Supply	0.707

- The Infrastructure and Ecosystem pillar explains 64.4% of the data diversity, with the dominant sub-pillars being the Learning Ecosystem (loading 0.640) and Government Digitalization (loading 0.618). This shows that the availability of digital-based learning ecosystems and the level of digitalization adoption in the government sector have a very large role in determining the strength of a region's digital infrastructure.
- The Digital Skills pillar explains 82.9% of the variance, with relatively even contributions from all sub-pillars (0.562-0.587). People's ability to complete the digital ecosystem, basic knowledge of information and communication technology, and awareness of digital security all play an equally strong role in determining a region's digital readiness.
- The Empowerment pillar explains 77.8% of the data variance, equally dominated by the User/Consumer and Provider/Seller subpillars (both 0.707). This illustrates that the level of community empowerment in the digital realm is largely determined by the interaction between the demand side (users/consumers) and the supply side (providers/sellers). This balance of contributions also confirms that digital empowerment cannot run optimally from only one side but must involve both parties simultaneously.
- The Employment pillar is only able to explain 50.8% of the data variability, indicating a relatively low contribution compared to the other pillars. The Supply sub-pillar is more prominent (0.707) than Demand (-0.707). This condition can be interpreted that the availability of digital jobs may have developed sufficiently but has not been matched by a corresponding level of demand for digital workers.

Overall, the PCA results at the pillar level indicate that Digital Skills is the dimension that best explains variations in digital readiness of communities in Indonesia, followed by Empowerment and Infrastructure and Ecosystem. Meanwhile, the Employment pillar is the dimension with the lowest explanatory contribution. Thus, these findings confirm that strategies to improve digital readiness at the district/city level should prioritize strengthening people's digital skills and empowering the ecosystem of digital users and service providers, while still paying attention to basic infrastructure development. The employment pillar remains important, but a more comprehensive policy approach is needed to bridge the gap between the demand and supply of digital workers.

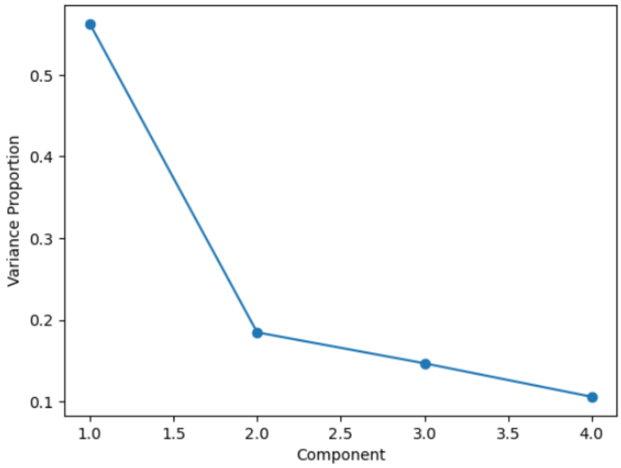

Figure 6. Scree plot of PCA results between IMDI pillars

PCA analysis on the four IMDI pillars resulted produced eigenvalues that were visualized in the form of a scree plot (Figure 6). Based on the graph, it can be seen that the decrease in variance begins to slope after the 2nd component. This indicates that the two main factors are sufficient to represent the diversity of IMDI pillar data. The two factors are cumulatively able to explain more than 70% of the data variation, so it can be used as an adequate representation for further analysis.

TABLE III
RESULTS OF VARIMAX-ROTATION LOADINGS OF THE IMDI PILLARS

Pilar	Factor 1	Factor 2
Infrastructure & Ecosystem	0.026	0.945
Digital Skills	0.542	0.040
Empowerment	0.519	0.216
Employment	0.660	-0.240

The loading values in Table III reflect the correlation between each IMDI pillar and the extracted factors. High positive loadings indicate pillars that move together within the same latent dimension. The strong correlation between Digital Skills, Empowerment, and Employment suggests that improvements in community digital competence and participation are closely linked with labor-market adoption of

ICT. Conversely, the high loading of Infrastructure and Ecosystem on the second factor shows that regional differences in physical and institutional infrastructure contribute independently to variations in digital readiness. The moderate cross-loadings (around 0.2–0.3) imply that while the pillars are interconnected, each still retains partial uniqueness, confirming that PCA successfully disentangles the main sources of variation among the IMDI pillars.

In Table III, Factor 1 is dominated by Digital Skills (0.542), Empowerment (0.519), and Employment (0.660). This factor represents the digital capacity and participation of the community as well as their connection to the world of work. Meanwhile, Factor 2 is dominated by Infrastructure and Ecosystem (0.945), which confirms that variations between regions are mostly explained by differences in digital literacy and basic infrastructure support. Thus, it can be seen that infrastructure is more dominant than existing digital skills.

The two main components are retained because they explain 70.4% of the cumulative variance of the four pillars of IMDI, which form the basis for regional clustering. PCA between pillars successfully reduced the four main pillars of IMDI into two main latent dimensions, namely:

1. Factor 1 (Digital Capacity and Participation) → related to digital skills, community empowerment, and the use of ICT in employment by explaining 45.1% of the variance.
2. Factor 2 (Digital Infrastructure Foundation) → related to infrastructure availability, learning ecosystems, and government support for digitalization in employment by explaining 25.3% of the variance.

The factor scores of these two latent dimensions were then used as input in the clustering analysis to form a typology of districts/cities in Indonesia based on digital readiness.

D. K-means Clustering

To determine the optimal number of clusters, tests were conducted with various values of K (number of clusters). The evaluation used two measures, namely the Silhouette Score and the Davies-Bouldin Index (DBI). Silhouette Score measures the compactness and separateness of the clusters. Values range from -1 to 1, the closer to 1 the better. Meanwhile, DBI measures the similarity between clusters. A smaller value indicates better cluster quality.

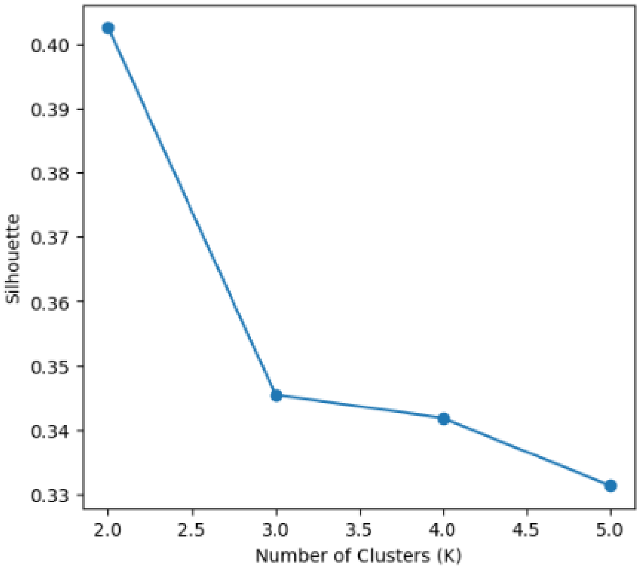

Figure 7. Silhouette Score

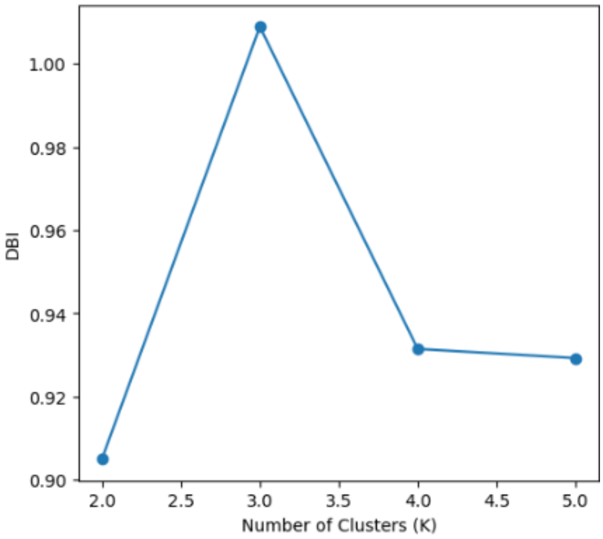

Figure 8. Davies-Bouldin Index (DBI) Value

Or it can be shown in Table IV below.

TABLE IV
SILHOUETTE SCORE AND DBI VALUES FOR VARIOUS K

K	Silhouette Score	DBI
2	0.402	0.906
3	0.346	1.007
4	0.342	0.931
5	0.331	0.930

Based on Table IV, the highest silhouette score was obtained at K=2 (0.402), indicating that two clusters provided the most clear separation compared to other cluster numbers. Meanwhile, the lowest DBI value was obtained at K=2 (0.906), meaning that the two-cluster configuration also produced better separation compared to other clusters. This shows that statistically, two clusters provide the clearest separation between regency/city groups. However, to obtain a better level of policy interpretation, this study chose a three-cluster configuration (K=3). This consideration was based on the clear differences that emerged between regions with High, Medium, and Low digital readiness, which were more representative of conditions in the field.

The Silhouette Score of 0.402 indicates a *moderate* level of cohesion and separation among clusters, meaning that most districts/cities are well grouped with their assigned cluster, although a small number lie near the decision boundary. Meanwhile, the Davies-Bouldin Index (DBI) of 0.906 shows good compactness, as values below 1 generally indicate well-separated clusters. These numerical results confirm that the two-cluster configuration provides a statistically valid and stable partition of digital readiness levels across Indonesian districts/cities. Even when three clusters were adopted for policy interpretation, the corresponding values (Silhouette = 0.346; DBI = 1.007) remained within an acceptable range, ensuring that interpretability is achieved without a significant loss of internal validity.

This approach is in line with the practice of typological analysis in public policy, where three-level segmentation (high-medium-low readiness) is more helpful in formulating a phased intervention strategy. Although the Silhouette value (0.346) and DBI (1.007) at K = 3 are slightly lower than at K = 2, this configuration is still considered adequate because it still shows sufficient separation between clusters and improves the interpretability of results for government decision-making.

1) Distribution of Districts/Cities Clusters

Based on the K-Means results with K=3, the following distribution of districts/cities was obtained.

TABLE V
NUMBER OF DISTRICTS/CITIES PER CLUSTER

Cluster	Number of Districts/Cities	Average	General Characteristics
0	154	47.87	High (relatively better PCA scores across all pillars)
2	240	43.44	Medium (PCA scores are relatively weak but not too high on some pillars)
1	120	38.54	Low (relatively weaker PCA scores across all pillars)

In Table V, cluster 0 is categorized as high because it has a higher average IMDI of 47.87, indicating that the districts/cities in this cluster are relatively more digitally ready. Meanwhile, cluster 1 has a lower average IMDI value of 38,54, so it can be categorized as an area with low digital readiness. Cluster 2 has an average value of 43.44, so it can be categorized as an area that has better digital readiness than cluster 1.

2) Average Digital Society Index and Pillars per Cluster

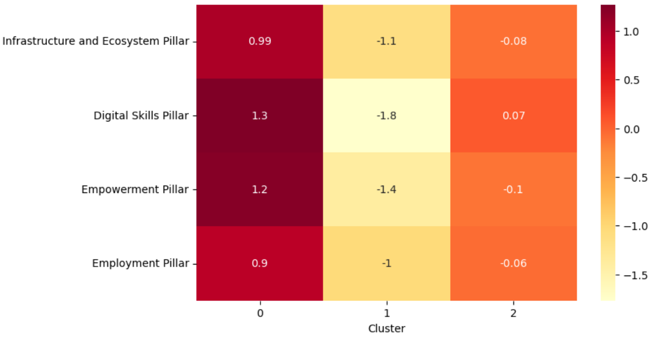

Figure 9. Heatmap of Average Pillars per Cluster

The Figure 9 illustrates the average scores of each digital development pillar across the three identified clusters, revealing distinct characteristics and disparities among them. Cluster 0 (High Category) stands out with the highest values across all four pillars, notably in the Digital Skills Pillar (1.3), Empowerment Pillar (1.2), and Infrastructure and Ecosystem Pillar (0.99). This indicates that regions belonging to this cluster are relatively advanced in terms of digital readiness, supported by strong infrastructure, well-developed ecosystems, and a higher level of digital literacy among the population. Such conditions suggest a mature digital environment that fosters innovation, adaptability, and inclusive participation in the digital economy.

In contrast, Cluster 1 (Low Category) shows the lowest average values across all pillars, including a markedly low score in the Digital Skills Pillar (-1.8). This implies that the regions within this cluster face substantial challenges in developing human capital and digital ecosystems, which may hinder their ability to fully engage in and benefit from digital transformation initiatives. The limited infrastructure and skill base could also contribute to slower adoption of digital technologies and reduced competitiveness in the labor market.

Meanwhile, Cluster 2 (Medium Category) presents scores that are close to the overall average (around zero) across all pillars, reflecting regions with a moderate level of digital development. These areas may possess basic digital foundations but still require targeted interventions to enhance their digital skills, infrastructure, and empowerment initiatives. Overall, the differences among clusters underscore the existence of a digital divide across regions. This highlights the need for differentiated policy strategies strengthening infrastructure and ecosystem support in lagging regions, while promoting innovation and sustainable digital growth in more advanced areas to ensure balanced and inclusive digital transformation.

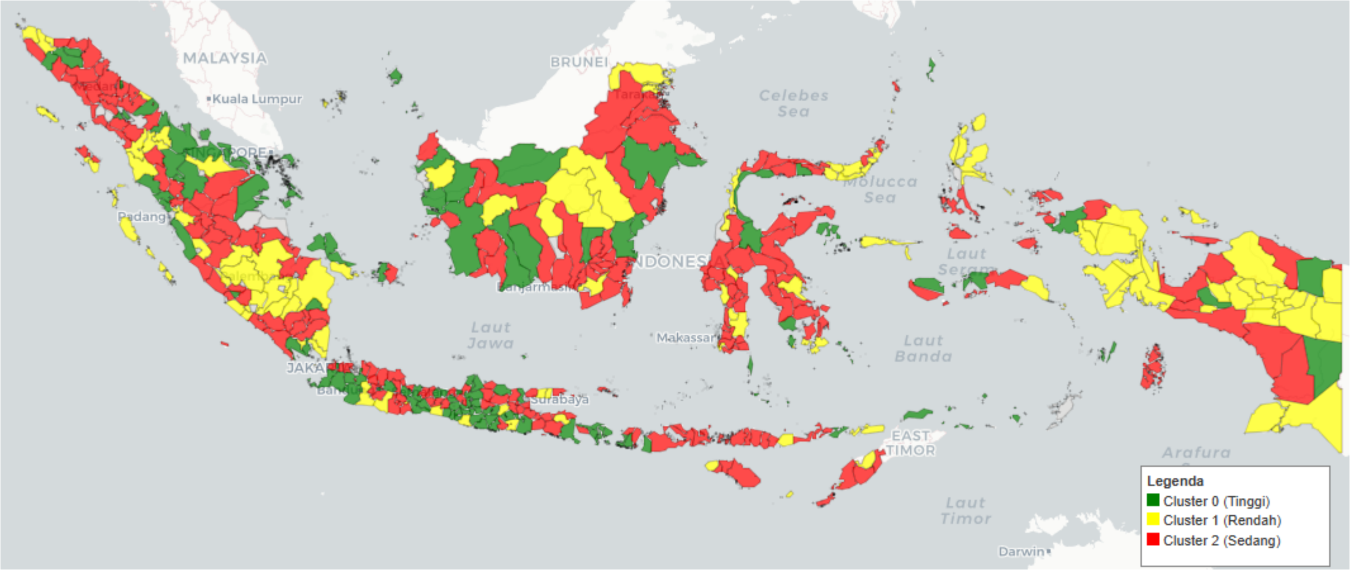

Figure 10. Digital Readiness Typology Map

After identifying the average scores for each pillar, the typology results were then mapped to show the spatial distribution of districts/cities in each cluster. The map in Figure 10 illustrates the clustering results for the average scores of each pillar, yielding three main clusters that describe the level of digital readiness of districts/cities in Indonesia. The map in Figure 10 shows the spatial distribution of each cluster, namely Cluster 0 (High) in green, Cluster 1 (Medium) in yellow, and Cluster 2 (Low) in red.

Cluster 0 (High) marked in green. Areas in this cluster are generally located on the islands of Java and Bali, as well as parts of Sumatra and Kalimantan. These areas have more adequate ICT infrastructure, good access to education, and relatively high levels of digital literacy among the population. The existence of economic and government centers, as well as urban population concentrations, also support stronger digital readiness in these areas.

Cluster 2 (Medium) marked in yellow. This cluster is spread quite widely across various regions of Indonesia, including central Kalimantan, Sulawesi, parts of Nusa Tenggara, and parts of western Papua. Areas in this cluster show a moderate level of digital development, where several pillars such as infrastructure access and technology use are beginning to develop, but still face obstacles in terms of digital literacy equity and innovation ecosystem support.

Cluster 1 (Low) marked in red. Regions with low digital readiness are mostly found in eastern Indonesia, particularly in Maluku, Papua, and parts of East Nusa Tenggara. These regions still face significant limitations in almost all pillars, ranging from ICT network access and human resource quality to the availability of facilities to support digital transformation. This reflects a fairly significant digital divide between eastern and western Indonesia.

Overall, this spatial distribution shows a gradation of digital readiness from west to east, where regions in western Indonesia tend to be more digitally advanced than the central and eastern regions. These findings underscore the importance of inclusive and equitable digital development strategies across regions to reduce the national digital transformation gap.

This mapping provides an important visual illustration, as it makes it clear that digital readiness in Indonesia is uneven. High-readiness districts tend to cluster in areas with better access to infrastructure and resources, while low-readiness districts are generally located in areas with geographical, social and economic barriers.

The typology results provide a strong empirical foundation for targeted government interventions. Districts/cities within the Low-readiness Cluster, which are concentrated in eastern Indonesia, can be prioritized for basic digital infrastructure investment and community digital literacy programs. Areas in the Medium-readiness Cluster should focus on ecosystem strengthening, including MSME digitalization and e-government service integration. Meanwhile, High-readiness Clusters mostly located in Java and Sumatra can focus on innovation, AI adoption, and 5G-based service expansion. Such a tiered intervention design allows policymakers to allocate budgets efficiently, ensuring that transformation efforts correspond with each region's readiness level.

IV. Conclusion

This study successfully identified the latent dimensions of digital readiness of districts/cities in Indonesia through the Principal Component Analysis (PCA) approach. The PCA results show that there are differences in contributions between pillars and sub-pillars in explaining digital readiness, so as to reduce the diversity of indicators into simpler but still representative dimensions. The results of dimension reduction were used as the basis for compiling regional typologies using the K-Means Clustering method. The analysis resulted in three main groups, namely High Cluster, Medium Cluster, and Low Cluster. High Clusters are characterized by positive scores on all pillars, especially Digital Skills and Empowerment, while Medium Clusters and Low Clusters show negative scores indicating relative backwardness in terms of digital literacy, infrastructure, and technology utilization. Spatial distribution reveals that High Clusters are concentrated in Java, Bali, and some major cities in Sumatra and Kalimantan, while Low Clusters are dominant in Eastern Indonesia.

In addition to digital indicators, non-digital socio-economic factors such as regional GDP per capita, education levels, urbanization, and population density—can also influence a region's digital readiness. Incorporating these contextual variables into future multivariate or spatial models (e.g., regression or SEM) would provide a more comprehensive understanding of the determinants of cluster membership and strengthen the policy relevance of the typology.

These findings have important implications for policy formulation. The government and stakeholders can design more targeted interventions according to the characteristics of each typology. Districts/cities in the High Cluster can focus on strengthening digital innovation and increasing global competitiveness, while those in the Low Cluster need to get priority and special attention in ICT infrastructure development, increasing digital literacy, and community empowerment programs. Thus, the digital readiness typology mapping conducted in this study can serve as an empirical basis for efforts to accelerate inclusive and equitable digital transformation throughout Indonesia.

REFERENCES

[1] R. D. Wahyunengseh, T. N. Haryani, P. Susiloadi, and L. Fahmi, "Masyarakat Digital dan Problematika Kesejahteraan: Analisis Isi Wacana Digital," vol. 17, pp. 163–172, 2022.

[2] Polrendyo, I. D. Pramudiana, E. Haryati, and S. Kamariyah, "Digital Transformation In Regional Revenue Management : A Case Study At Bapenda Of East Java Province," no. 95, 2025.

[3] M. Z. Fitry *et al.*, "Analyzing Telecommunication Infrastructure Index as Indicator of Digital Transformation: A Bibliometric Analysis BT - Proceedings of the 4th International Conference on Advances in Communication Technology and Computer Engineering (ICACTCE'24)," C. Iwendi, Z. Boulouard, and N. Kryvinska, Eds., Cham: Springer Nature Switzerland, 2025, pp. 419–432.

[4] Y. Syahidin *et al.*, "The application of unsupervised learning techniques to the clustering method in use of cell phones in Indonesia," in *2022 International Conference on Science and Technology (ICOSTECH)*, 2022, pp. 1–5. doi: 10.1109/ICOSTECH54296.2022.9829089.

[5] I. G. N. M. Jaya *et al.*, "Framework for Monitoring the Spatiotemporal Distribution and Clustering of the Digital Society Index of Indonesia," *Sustain.*, vol. 16, no. 24, pp. 1–22, 2024, doi: 10.3390/su162411258.

[6] D. A. R. Wati, A. Čaplánová, and Ľ. Darmo, "Identification of Digital Divide across Indonesian Provinces: the Analysis of Key Factors," *Statistika*, vol. 104, no. 2, pp. 185–202, 2024, doi: 10.54694/stat.2024.3.

[7] F. Kartiasih, N. Djalal Nachrowi, I. D. G. K. Wisana, and D. Handayani, "Inequalities of Indonesia's regional digital development and its association with socioeconomic characteristics: a spatial and multivariate analysis," *Inf. Technol. Dev.*, vol. 29, no. 2–3, pp. 299–328, Jul. 2023, doi: 10.1080/02681102.2022.2110556.

[8] Ü. Fidan, "Convergence or divergence? Trends in the digitalisation index cluster over the years," *Reg. Stat.*, vol. 14, no. 6, pp. 1050–1068, 2024, doi: 10.15196/RS140602.

[9] B. Zoltán and D. Imre, "Digital development of countries using tiered DEA, tiered Pareto efficiency and Cluster Analysis with data from the 2020 International DESI," *Statisztikai Szle.*, vol. 101, no. 11, pp. 978–998, 2023, doi: 10.20311/stat2023.11.hu0978.

[10] A. L. Yusniyanti, F. Virgantari, and Y. E. Faridhan, "Comparison of Average Linkage and K-Means Methods in Clustering Indonesia's Provinces Based on Welfare Indicators," *J. Phys. Conf. Ser.*, vol. 1863, no. 1, 2021, doi: 10.1088/1742-6596/1863/1/012071.

[11] C. Ding, "K -means Clustering via Principal Component Analysis," 2004.

[12] S. A. Mousavian Anaraki, A. Haeri, and F. Moslehi, "A hybrid reciprocal model of PCA and K-means with an innovative approach of considering sub-datasets for the improvement of K-means initialization and step-by-step labeling to create clusters with high interpretability," *Pattern Anal. Appl.*, vol. 24, no. 3, pp. 1387–1402, 2021, doi: 10.1007/s10044-021-00977-x.

[13] S. N. Mayasari and J. Nugraha, "Implementasi K-Means Cluster Analysis untuk Mengelompokkan Kabupaten/Kota Berdasarkan Data Kemiskinan di Provinsi Jawa Tengah Tahun 2022," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 317–329, 2023, doi: 10.24002/konstelasi.v3i2.7200.

[14] R. Magriaty, K. Murtilaksono, and S. Anwar, "Analisis K-Means Cluster untuk Identifikasi Kawasan Pengelolaan Sampah di Kabupaten Tapin Provinsi Kalimantan Selatan," *J. Reg. Rural Dev. Plan.*, vol. 7, no. 1, pp. 79–90, 2023, doi: 10.29244/jp2wd.2023.7.1.79-90.

[15] A. Fitriani, E. Arfi, and A. Huda, "Penerapan Algoritma K-Means Clustering dalam Memetakan Produktivitas Lokasi Perkebunan Nanas PT Great Giant Pineapple," *J. Math. Comput. Stat.*, vol. 7, no. 2, pp. 215–231, 2024, doi: 10.35580/jmathcos.v7i2.4200.

[16] R. H. Maharrani, P. D. Abda'u, and M. N. Faiz, "Clustering method for criminal crime acts using K-means and principal component analysis," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 34, no. 1, pp. 224–232, 2024, doi: 10.11591/ijeecs.v34.i1.pp224-232.

[17] Muhammad Raqib Syahkur, D. Hartama, and S. Solikhun, "Evaluasi Jumlah Cluster pada Algoritma K-Means++ Menggunakan Silhouette dan Elbow dengan Validasi Nilai DBI dalam Mengelompokkan Gizi Balita," *J. Sains dan Teknol.*, vol. 13, no. 3, pp. 487–496, 2024, doi: 10.23887/jstundiksha.v13i3.86419.

[18] C. I. Amalia, Fitria, and M. Sitorus, "Indonesia Berdasarkan Aspek Sosial Ekonomi Menggunakan Algoritma K-Means," *J. Ilmu Komputer, Sist. Inf. dan Teknol. Inf.*, vol. 2, no. 1, pp. 9–18, 2025.

[19] D. K. Maheswari, "Finding Best Possible Number of Clusters using K-Means Algorithm," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1s4, pp. 533–538, 2019, doi: 10.35940/ijeat.a1119.1291s419.

[20] N. Sutantri, V. Yunitasari, G. R. Sa'adi, S. Ananda, and Z. Alfian, "Analisis Konsistensi Metode Elbow dan Silhouette Score dalam Klasterisasi pada Dataset Multisektor," *J. Innov. Creat.*, vol. 5, no. 2, pp. 10731–10743, 2025, doi: 10.31004/joecy.v5i2.1690.

[21] T. Ikhsan, E. Haerani, F. Wulandari, and F. Syafria, "Clustering Data Penduduk Menggunakan Algoritma K-Means TIN : Terapan Informatika Nusantara," vol. 5, no. 12, pp. 955–963, 2025, doi: 10.47065/tin.v5i12.7328.

[22] A. S. Ahmar, D. Napitupulu, R. Rahim, R. Hidayat, Y. Sonatha, and M. Azmi, "Using K-Means Clustering to Cluster Provinces in Indonesia," *J. Phys. Conf. Ser.*, vol. 1028, no. 1, 2018, doi: 10.1088/1742-6596/1028/1/012006.

[23] P. P. E. S. K. dan Digital, "Data Indeks Masyarakat Digital

Indonesia (IMDI) Tingkat Kabupaten/Kota,” 2024. [Online]. Available: https://imdi.sdmdigital.id/unduh-data

[24] M. S. Bartlett, “Tests Of Significance In Factor Analysis,” *Br. J. Stat. Psychol.*, vol. 3, no. 2, pp. 77–85, Jun. 1950, doi: 10.1111/J.2044-8317.1950.TB00285.X.

[25] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, doi: 10.1098/rsta.2015.0202.

[26] Q. Fournier and D. Aloise, “Empirical comparison between autoencoders and traditional dimensionality reduction methods,” 2019, doi: 10.1109/AIKE.2019.00044.

[27] Z. Shen, “Comparison and Evaluation of Classical Dimensionality Reduction Methods,” *Highlights Sci. Eng. Technol.*, vol. 70, pp. 411–418, 2023, doi: 10.54097/hset.v70i.13890.

[28] T. M. Kodinariya and P. R. Makwana, “Review on determining number of Cluster in K-Means Clustering. International Journal,” *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.

[29] D. Hartama and S. Oktaviani, “Optimization of K-Means and K-Medoids Clustering Using Dbi Silhouette Elbow on Student Data,” *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 11, no. 2, pp. 289–296, 2025, doi: 10.33330/jurteksi.v11i2.3531.