

Family Data Clustering Based on Welfare Levels Using Self-Organizing Maps

Zulfan Amal Arridlo ^{1*}, Ai Rosita ^{2*}, Dedah Ningrum ^{3**}

^{*} Teknik Informatika, Universitas Widyatama

^{**} S1 Keperawatan, Universitas Pendidikan Indonesia Kampus Sumedang

zulfan.arridlo@widyatama.ac.id ¹, ai.rosita@widyatama.ac.id ², dedahningrum@upi.edu ³

Article Info

Article history:

Received 2026-06-11

Revised 2026-09-17

Accepted 2026-10-14

Keyword:

*Self-Organizing Maps (SOM),
BKKBN,
Family Welfare,
Optuna, Cluster Evaluation,
Digital Spatial Map.*

ABSTRACT

Sub-district socio-economic data on family welfare from BKKBN is high-dimensional and multivariate. The data pattern is highly complex with a positive asymmetric (right-skewed) distribution. This study aims to implement the Self-Organizing Maps (SOM) algorithm to cluster these welfare characteristics, mathematically evaluate model performance to eliminate bias, and visualize the tabular results into a spatial digital map. A 2×2 SOM model optimized using Optuna was used to group family welfare characteristics in West Java for the 2022–2025 period. Optimization results identified the best hyperparameter combinations (learning rate 0.7074–0.9574, sigma 0.1167–0.2668, and 7,000–20,000 iterations). The evaluation demonstrated reliable performance with Quantization Error ranging from 2.4659–3.4776, Topographic Error between 0.0000–0.0647, and Silhouette Score consistently above 0.50. GeoJSON integration successfully mapped 4 welfare hierarchies, confirming the dominance of prosperous regions in the eastern, southern, and western areas.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

Family well-being is the foundation of socio-economic resilience and regional development success [30], [35]. Consequently, Indonesia's National Population and Family Planning Board (BKKBN) continually monitors national family data to map regional welfare [44]. West Java faces complex socio-economic challenges, ranking second nationally in its total poor population despite moderate poverty percentages [20]. Previous panel regression (2017–2021) demonstrated that Population Size, GRDP, Life Expectancy, Mean Years of Schooling, and Unemployment explained 98.03% of the province's poverty variation [42]. This study extends this analysis to the 2022–2025 period to evaluate whether these variables consistently affect poverty across West Java's 27 regencies/cities.

Understanding BKKBN data requires effective clustering techniques [10]. This study utilizes sub-district (kecamatan) level data to capture detailed spatial disparities better than macro-level (regency/city) analysis, while avoiding the statistical instability of micro-level (village) data. To handle the complex multidimensionality of this data, we utilize the

Self-Organizing Map (SOM), an Artificial Neural Network-based unsupervised algorithm proven effective for socio-economic spatial clustering without requiring specific distributional assumptions [6], [16].

Previous studies have demonstrated SOM's reliability in socio-economic clustering by utilizing the Silhouette Score (SS) to determine optimal clusters for issues such as family planning [27], poverty [2], public welfare [58], and health indicators [26]. However, these prior studies relied on manual, inefficient hyperparameter tuning which risks bias. To address this gap, our research applies Bayesian Optimization via the Optuna library to intelligently discover the optimal SOM architecture. Model performance is then objectively evaluated using Quantization Error (QE), Topographic Error (TE), and the Silhouette Score.

Finally, because tabular outputs fail to capture spatial dependencies, this study integrates the clustering results into digital choropleth maps. By combining Optuna-optimized SOM with spatial visualization, this research contributes a more objective, valid, and intuitive methodology for mapping family welfare distribution and regional connectivity.

Several prior studies have demonstrated the effectiveness of the Self-Organizing Maps (SOM) algorithm in evaluating multidimensional socio-economic, health, and demographic data across various Indonesian provinces. For instance, in examining poverty factors across Maluku and West Nusa Tenggara (NTB), researchers utilized a hexagonal SOM to analyze 11 multidimensional indicators. Validated by multiple internal indices, this approach successfully identified three regional typologies ranging from developed urban centers to severely underdeveloped regions lacking basic infrastructure [2]. Similarly, SOM was applied to map public welfare in West Sumatra based on 10 socio-economic variables. Utilizing Min-Max scaling and internal matrix validation, the algorithm accurately categorized areas into high-welfare urban centers, medium-welfare regions, and low-welfare areas burdened by high poverty and unemployment [43].

Furthermore, SOM's adaptability extends to specific health and demographic challenges. In West Java, a SOM combined with Ward's hierarchical clustering evaluated seven health and socio-economic variables, producing a validated two-cluster model that sharply divided regions with excellent socio-health environments from those facing severe stunting, low birth weight, and poverty challenges requiring targeted interventions [26]. Lastly, at the sub-district level in NTB, a comparative study utilized SOM to map unmet family planning needs. The model optimally grouped areas into three distinct clusters based on specific barriers—such as geographical distance, medical resource scarcity, and couples' ignorance, with the latter identified as the dominant obstacle across the province [27].

II. METHOD

Adopting the CRISP-DM framework, this research structures the data mining process into six iterative stages, spanning from Business Understanding and Data Understanding through Data Preparation, Modeling, Evaluation, and final Deployment [22]. The clustering implementation utilizes the Self Organizing Maps (SOM) model, an unsupervised learning artificial neural network structured in a lattice grid that functions to reduce the dimensionality of multivariate data into a two dimensional map representation [33][37]. To mathematically verify the model's reliability, the evaluation relies on three primary metrics: Quantization Error (QE) to assess the geometric proximity precision between input data vectors and the Best Matching Unit (BMU) [41][25], Topographic Error (TE) to evaluate how well the original topological structure is preserved on the grid lattice without distortion [40][3][7], and the Silhouette Score to validate cluster separation quality based on internal cohesion and intercluster distance [17][38].

A. CRISP DM Method

operating as a continuous lifecycle, the CRISP-DM framework partitions the data mining process into six structured stages [4]. This approach begins with Business

Understanding to convert organizational goals into actionable data mining objectives, which is sequentially followed by Data Understanding to collect, profile, and assess the quality of the source data. During Data Preparation, raw data is cleaned and consolidated into a finalized dataset ready for analysis [19]. The process then advances to the Modeling phase, where diverse analytical algorithms are deployed and fine-tuned. Upon completion, the Evaluation phase is carried out to verify that the model's performance perfectly corresponds with the predetermined business targets. Lastly, the cycle ends with Deployment, where the developed model is integrated and validated within a live operational environment [19]. Guided by this methodology, the comprehensive workflow established for this study is presented in Figure 1.

B. Business Understanding

In the business understanding stage as the beginning of the CRISP-DM framework, an in-depth study of the background, problems, objectives, and previous literature is carried out to understand the variations in regional socio-economic independence. Through clustering models, multivariate data challenges will be addressed objectively.

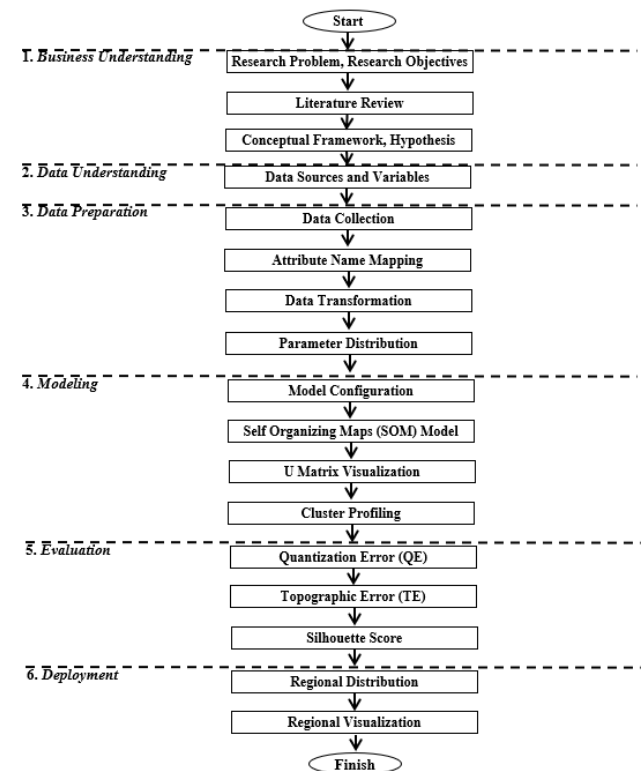


Figure 1. Research Stages

C. Data Understanding

To operationalize the theoretical constructs of family prosperity [6], economic resilience [8], and public welfare [31], [1], this study utilizes a dataset grounded in established

socioeconomic parameters [4]. The data uses spatial identifiers for regional mapping and normalizes subdistrict figures using the total number of household heads as a baseline. The core analysis relies on 15 numerical attributes structured into eight paired indicators that contrast ideal societal conditions with vulnerable ones. These dualities measure a household's capacity to secure material and macroeconomic needs against systemic vulnerabilities and infrastructure gaps [4]. Furthermore, the pairs evaluate long-term independence by contrasting nutritional adequacy with chronic health burdens [1], educational continuity with dropout rates to address poverty cycles, and digital literacy with technological exclusion as a vital component of modern social resilience. By examining these opposing variables, the approach moves beyond absolute welfare metrics to diagnose specific vulnerability gaps and structural inequities across subdistricts [4].

The selection of socioeconomic indicators for this study is strongly supported by prior regional research across Indonesia. For example, welfare clustering studies in Kalimantan utilized macroeconomic and human development metrics, including Gross Regional Domestic Product (GRDP), Mean and Expected Years of Schooling (MYS, EYS), Life Expectancy, and overall poverty rates. Similarly, research in Papua Province expanded this framework by integrating morbidity rates, Labor Force Participation Rates (LFPR), per capita food expenditure, access to decent drinking water, and the receipt of social assistance. On a national scale, a DBSCAN clustering analysis of 38 provinces reinforced the critical role of broad education, employment, and health data in assessing poverty factors. Finally, highlighting the modern dimensions of social resilience, recent studies underscore the significant relationship between regional poverty levels and the proportion of internet usage, validating digital access as a vital indicator for targeted welfare development.

D. Data Preparation

Following the data understanding stage, data preparation begins with automated web scraping of BKKBN data and its integration with subdistrict GeoJSON spatial files from Open Data Jabar. GeoJSON serves as a lightweight vector format for WebGIS systems [9][29], organizing spatial geometries (Polygons) and non-spatial attributes into FeatureCollection structures [24][23]. After standardizing attribute names, the data undergoes Z-score Normalization (Standard Scaling) to produce a centered distribution with a variance of one [34][15], [34][12]. This normalization, mathematically defined in Equation 1 [12], is critical for distance-based clustering algorithms. Without it, variables with massive absolute values (like total households) would dominate distance metrics (e.g., Euclidean distance) and mask crucial smaller-scale socio-economic disparities, such as rare illnesses [34][12], [15]. By scaling all features uniformly, the algorithm avoids geometric distortion, accelerates convergence, and ensures the resulting clusters represent true

structural welfare gaps rather than mere population volume [12]. The stage concludes with a visual parameter distribution analysis using boxplots to objectively detect quartiles, medians, and outliers before feeding the data into the model.

E. Modeling

The modeling process is executed using Python, an open-source language favored for its readable syntax, automated memory management, and robust data mining libraries such as Pandas and NumPy [32], [28], [14], [5]. The core methodology employs the Self-Organizing Maps (SOM) algorithm, a fully connected two-dimensional neural network trained iteratively via distance, neighborhood, and weight updating functions [21]. This is implemented through MiniSom, a lightweight and memory-efficient library that conducts unsupervised clustering by pulling similar data toward a Best Matching Unit (BMU) using a Gaussian neighborhood function [36], [11], [45]. The training process is governed by key parameters, including grid dimensions, input length, sigma, initial learning rate, iterations, toroidal topology, and a random seed for reproducibility [11], [45]. Following training, a Unified Distance Matrix (U-Matrix) is utilized to visually identify cluster structures and boundaries through color gradations that represent Euclidean distances between neuron weights [46], [18]. Finally, cluster profiling is performed via an inverse transform procedure to revert Z-scores back to their original nominal values, enabling the mean computation of variables to objectively analyze and compare the unique socio-economic characteristics of each subdistrict group.

F. Evaluation

Following the training of the Self-Organizing Maps (SOM) algorithm, the evaluation stage objectively measures the network's performance, precision, and validity using three key metrics. First, Quantization Error (QE) evaluates mapping precision and average distortion by calculating the average Euclidean distance between the original input vectors and their corresponding Best Matching Unit (BMU) [41], [30], [25]. Mathematically defined in Equation 2, a QE value approaching zero indicates that the model accurately replicates original data characteristics with minimal dimensionality reduction distortion [25], [30]. Second, Topographic Error (TE) measures the network's topological quality and spatial continuity to ensure the high-dimensional structure is preserved without overfitting [40], [3]. Defined in Equation 3, TE calculates the proportion of data where the first and second BMUs are not adjacent; a value near zero signifies an ideal topology where socio-economically similar regions map to adjacent clusters [40], [7], [30]. Finally, the Silhouette Score (Equation 4) assesses the separation quality and structural solidity of the regional groups by comparing internal cluster cohesion to external distinctiveness [17], [38], [39]. Ranging from -1 to +1, scores approaching +1 confirm dense, well-separated clusters, while lower values indicate overlapping boundaries or clustering errors [13].

G. Deployment

Following the validation of the Self Organizing Maps (SOM) model, the research workflow enters the final deployment stage to implement the socioeconomic welfare clustering results into a practical format through a two step process. First, the two dimensional horizontal (Cluster X) and vertical (Cluster Y) coordinate outputs are extracted for each subdistrict in West Java Province and combined to form a unique cluster identity label (*cluster_id*). Second, these tabular results are integrated into an interactive digital choropleth map using a GeoJSON spatial file format, which is a lightweight text based open standard for transmitting geospatial vector data [9][29]. The West Java Province GeoJSON file containing polygon geometry elements of administrative boundaries [23] is matched with the subdistrict dataset, integrating the spatial and non spatial properties within the FeatureCollection structure [24]. This integration directly maps the *cluster_id* labels as the regional coloring parameters, producing both a macro visualization of the entire province and a micro visualization per regency or city for a localized analysis of regional characteristics.

III. RESULTS AND DISCUSSION

In alignment with the chosen research methodology, this study advances through the sequential phases of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The processed data is subsequently utilized as input vectors to train the Self-Organizing Map (SOM) model.

A. Business Understanding

The dataset's multivariate and high-dimensional nature complicates pattern identification when relying solely on conventional statistical methods. Exploratory findings reveal that all parameters exhibit a right-skewed distribution with massive outliers, highlighting a stark population disparity where a minority faces disproportionately high vulnerability compared to the average majority. To address this complex data structure, Self-Organizing Maps (SOM) are utilized as an effective solution. SOM uniquely processes high-dimensional data and maintains information integrity despite the presence of extreme values, thereby generating a fairer, more accurate, and highly representative depiction of the population.

B. Data Understanding

The second stage of the research is data understanding, where the process to be conducted is the identification of data requirements. Based on the insights gained from dataset utilization in prior research. The dataset comprises 14 numerical attributes structured into seven paired socioeconomic dimensions that systematically contrast ideal household conditions against vulnerable ones. These dimensions evaluate economic stability through Income Source Ownership, comparing stable streams against informal or unstable labor [17], [19], [23], and Financial Asset Ownership, contrasting the possession of productive

buffering assets with a total lack of savings [17], [19]. Basic living standards are measured via Nutritional Fulfillment, weighing adequate daily diets against food insecurity [16], [23], alongside Housing Conditions, which compare structurally sound homes with substandard or temporary shelters [18], [23]. Furthermore, the framework assesses long-term human development and social resilience by contrasting Compulsory Education continuity with dropout rates for children aged 7–18 [18], [23], comparing healthy households against those burdened by chronic illnesses or disabilities [16], [19], [23], and evaluating Family Digital Literacy by measuring active internet usage against technological exclusion [18], [23]. In alignment with the data understanding phase, the identification of specific data requirements is crucial to translate theoretical family prosperity dimensions into measurable analytical variables. The data utilized in this study is secondary data, specifically the Data on the Number of Families According to the Independence Dimension, sourced from the Ministry of Population and Family Development (Kemendukbangga). The dataset was obtained in .xlsx format, covering the period from 2022 to 2025, and comprises a total of 618 data rows representing the units of analysis under investigation.

C. Data Preparation

This stage requires several data processing preparations, which will be elaborated upon in the subsequent subsections.

1) Data Collection

The collected data are obtained from the official website of the Ministry of Population and Family Development / BKKBN via <https://siga.bkkbn.go.id/>. The data utilized encompass all subdistricts in West Java Province for the period from 2022 to 2025. The data regarding the number of families according to the dimensions of independence (Table 2) consist of 15 variables, and these 15 variables can be observed in Table 1.

TABLE 1
RESEARCH VARIABLES AND DESCRIPTIONS

Variable	Description	Unit of Measurement
X ₁	Households with at least one member having a stable income source for basic needs in the last 6 months (YES)	Number of households
X ₂	Households with no members having a stable income source for basic needs in the last 6 months (NO)	Number of households
X ₃	All family members consumed a diverse diet (staples, vegetables/fruits, side dishes) 2 times daily in the last 6 months (YES)	Number of households
X ₄	Not all family members consumed a diverse diet (staples, vegetables/fruits, side dishes) 2 times daily in the last 6 months (NO)	Number of households
X ₅	The family resides in a habitable and decent housing unit (YES)	Number of households

X ₆	The family resides in an inadequate or indecent housing unit (NO)	Number of households
X ₇	The household owns productive or valuable assets (YES)	Number of households
X ₈	The household does not own any valuable assets (NO)	Number of households
X ₉	No school-aged members (7–18 years) dropped out of primary to senior high school education (YES)	Number of households
X ₁₀	At least one school-aged member (7–18 years) dropped out of primary to senior high school education (NO)	Number of households
X ₁₁	No school-aged children (7–18 years) reside in the household (NOT APPLICABLE)	Number of households
X ₁₂	Presence of a family member with a disability or a severe illness (disrupting activities for ≥ 4 consecutive days) in the past month (YES)	Number of households
X ₁₃	Absence of family members with disabilities or severe illnesses in the past month (NO)	Number of households
X ₁₄	At least one family member actively accessed digital information via the internet in the last 6 months (YES)	Number of households
X ₁₅	No family members accessed digital information via the internet in the last 6 months (NO)	Number of households

2) Data Transformation

To ensure objective comparison across regions with varying data ranges, this study applies Z-score Normalization to transform original values into standard deviation units, creating a distribution with a mean of zero and a standard deviation of one. This transformation eliminates measurement unit differences and is specifically chosen over Min-Max scaling because it preserves critical socioeconomic outliers—such as localized spikes in illness or school dropouts—without compressing essential variations. Furthermore, by placing all variables on a uniform scale, Z-score Normalization prevents attributes with large absolute population volumes from dominating distance-based clustering algorithms. This ensures that the resulting clusters accurately reflect genuine structural vulnerabilities and socioeconomic gaps rather than being skewed by regional population sizes.

```

has_income [ 3.48995712 -0.81725679 -0.39920649 0.53503801 3.74426247 1.85928117 -
0.28604282 2.53121044 0.49570518 1.8814427 ]
has_no_income [ 0.10468408 -1.05600325 -1.02752845 -0.42548969 0.43011043 1.30876158
-0.98956204 -0.57193155 0.61723058 0.88841921 ]
cats_twice_daily [ 3.43423924 -0.83322889 -0.41704247 0.51926835 3.68912264
1.84735646 -0.30665818 2.43962176 0.48312546 1.8892938 ]
does_not_eat_twice_daily [ 0.1091764 -0.64016014 -0.72038523 -0.36876034 0.40788686
1.21696548 -0.65040164 0.48128428 0.8858236 0.21329833 ]
decent_housing [ 3.96553791 -0.77320374 0.04640967 0.88946368 4.10835638 2.78899067
0.20172351 3.66698323 0.48074657 2.25502321 ]
indecent_housing [ 1.31740308 -0.63617203 -0.95911527 -0.23929331 1.65192134 -
0.11138926 -0.95014462 -0.21990514 0.3670934 0.63491057 ]
owns_assets [ 3.90384331 -0.63926906 -0.25817022 0.70953162 4.27771131 2.01502935 -
0.15340501 2.94920632 0.76770712 1.95323842 ]
does_not_own_assets [ -0.00860403 -1.10528294 -0.81918106 -0.45353154 -0.23942703
0.45708677 -0.74088168 -0.51986619 -0.580077 0.65743898 ]
dropped_out_of_school [ 4.04081819 -0.75767613 -0.31040937 0.91897026 4.40481635
2.05686716 -0.21796028 1.1417178 0.70540319 2.37802306 ]
did_not_drop_out_of_school [ 3.93178488 -0.92298862 -0.19520951 0.46932064 1.33732934
2.8665393 0.17013643 1.3756116 0.20800258 2.45125997 ]
no_school_age_children [ 2.09611992 -0.80285015 -0.61284669 0.0193643 3.33835054
1.06235856 -0.59454286 3.81781365 0.37410222 0.8956788 ]
has_disability [ 3.50070242 0.7847315 -0.36859148 -0.22795489 1.16689662 1.66159361 -
0.48597332 3.43959279 0.69935947 1.45304488 ]
has_no_disability [ 3.3434417 -0.84367723 -0.49995225 0.54783731 3.80834688
1.86763452 -0.3135329 2.31828768 0.49296326 1.88333459 ]
has_internet_access [ 3.50260109 -0.76175128 -0.40869576 0.54813253 3.89421664
1.92302852 -0.26664065 2.78808396 0.53170596 1.80279839 ]
has_no_internet_access [ 1.15292566 -1.1197827 -0.49257786 -0.0863774 0.37857107
0.71763095 -0.64756068 -1.22441009 0.16484024 1.58486334 ]

```

Figure 2. Data Transformation Results for 2024 Periode

3) Parameter Distribution

To map fundamental data characteristics and dispersion patterns prior to machine learning modeling, a variable distribution analysis was conducted using a Ridge Plot (Faceted Density Plot). This visualization enables the simultaneous evaluation of distribution shape, skewness, and multimodality across all variables. The ridge plots for the 15 welfare indicators from 2022 to 2025 (Figures 3–6) consistently reveal a right-skewed distribution, with peak densities concentrated within the Z-score range of -1 to 0. Furthermore, the persistent presence of extended curve tails in the positive range (Z-score > 2) across all four years highlights stable regional disparities and natural positive outliers. This pattern indicates that while the majority of sub-districts hover around the population mean, a minority of developed or urban regions significantly exceed average socioeconomic levels. The stability of these structural inequalities throughout the 2022–2025 period underscores the critical necessity of the subsequent clustering stage to precisely map and categorize these distinct regional characteristics.

D. Modeling

Modeling is conducted by utilizing the Self Organizing Maps (SOM) algorithm, an unsupervised learning technique based on artificial neural networks to simultaneously reduce data dimensionality and perform topological clustering.

1) Model Configuration

The training process utilized Bayesian Optimization (Optuna) to efficiently evaluate hyperparameter combinations in determining the optimal SOM model architecture. Optimization was performed over a parameter search space encompassing the number of iterations, sigma, learning rate, and SOM grid node dimensions.



Figure 3. Optuna Result 2022

Based on the 2022 Optuna optimization results across 500 trials, Bayesian Optimization efficiently escaped local optima to identify the best hyperparameter configuration. The objective function (Silhouette Score) progressed from an initial 0.23 to a plateau of 0.41 at trial 40, surged to 0.49 at trial 90, and ultimately peaked and stabilized above 0.51 at trial 240. Furthermore, hyperparameter importance analysis revealed that the neighborhood radius (sigma) was the primary determinant of clustering quality, holding a 36% importance level. This was followed by the horizontal grid dimension (grid_x) at 23%, vertical grid dimension (grid_y) at 18%, and learning rate at 15%. Conversely, the number of training iterations had the least impact (8%), demonstrating that the SOM's topological structure for the 2022 dataset is primarily governed by neighborhood radius and grid resolution rather than training duration.

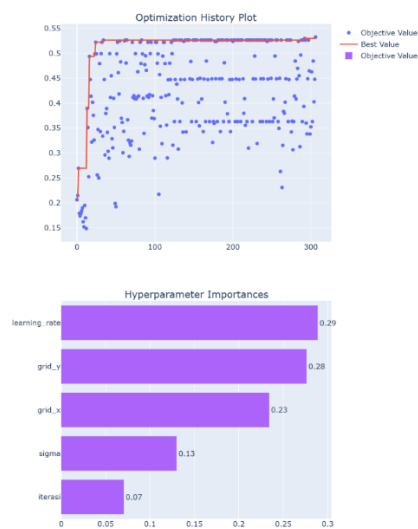


Figure 4. Optuna Result 2023

Based on the 2023 Optuna optimization results, the hyperparameter search demonstrated rapid and efficient convergence within approximately 300 trials. The objective value (Silhouette Score) quickly surged from 0.21 to over 0.49 before trial 20, peaking above 0.525 around trial 35, and remaining stable until the search concluded. This indicates that Bayesian Optimization precisely identified the optimal parameter region very early in the execution. Furthermore, hyperparameter importance analysis revealed that the learning rate was the most dominant factor (29%) influencing the SOM model's performance. This was closely followed by the vertical grid dimension (grid_y) at 28% and the horizontal grid dimension (grid_x) at 23%. Conversely, the neighborhood radius (sigma) and the number of training iterations had minimal impacts of 13% and 7%, respectively. Thus, the successful formation of optimal clusters for the 2023 dataset relied heavily on the weight update rate (learning rate) and the spatial grid dimensions.

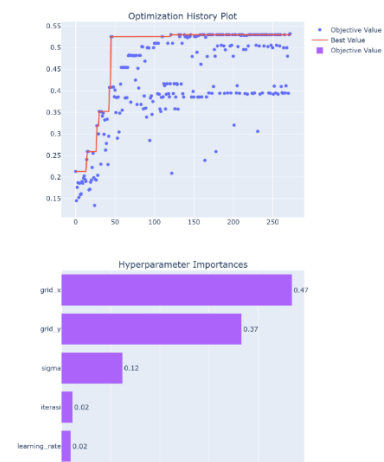


Figure 5. Optuna Result 2024

Based on the 2024 Optuna optimization results, the hyperparameter search demonstrated rapid and significant performance improvements within 300 trials. The objective value (Silhouette Score) started at 0.21, surged past early plateaus to reach 0.525 around trial 45, and ultimately peaked and stabilized above 0.53 at trial 120, successfully achieving the model's goodness-of-fit threshold. Furthermore, hyperparameter importance analysis revealed that the spatial dimensions of the SOM map absolutely dominated the clustering quality, jointly accounting for 84% of the overall influence (horizontal grid_x at 47% and vertical grid_y at 37%).



Figure 6. Optuna Result 2025

Based on the 2025 Optuna optimization results across 500 trials, the hyperparameter search demonstrated progressive and stable convergence. The objective value (Silhouette Score) began at 0.14, climbed steadily through early plateaus, and sharply surpassed 0.52 around trial 60, maintaining this global optimum consistently until the experiment concluded. Furthermore, hyperparameter importance analysis revealed

that spatial grid dimensions predominantly controlled the SOM clustering quality. The vertical dimension (grid_y) contributed the most at 31%, followed closely by the horizontal dimension (grid_x) at 28%.

2) Self-Organizing Maps (SOM) Model

After obtaining the optimal model configuration, the subsequent stage is to execute the Self Organizing Maps (SOM) model by utilizing the MiniSom Python library. This execution involves the model training process, through both random training and topical training methods, to ensure that each neuron on the map adapts well to the data structure. The results of the execution using MiniSom will subsequently serve as the primary foundation for analyzing the U Matrix (Unified Distance Matrix) and visually mapping the cluster distribution.

3) U-Matrix Visualization

Once the Self Organizing Maps (SOM) model is established, the subsequent process is feature mapping visualization. The constructed model is utilized to map all observation data into a two dimensional grid (2D grid) to observe data dispersion through U Matrix (Unified Distance Matrix) visualization for identifying cluster boundaries.

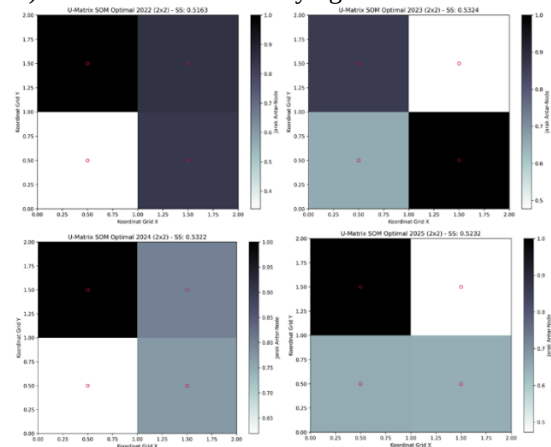


Figure 7. U-Matrix Visualization

The U-Matrix visualizations for the 2022–2025 period consistently display a 2×2 grid architecture containing four distinct cluster nodes, illustrating the distance relationships and separation between neurons across all four years. The color gradients, ranging from lighter shades for lower inter-node distances to darker tones for higher distances, delineate the boundaries between clusters with achieved Silhouette Scores ranging between 0.5163 and 0.5324. Specifically, the maps for 2022, 2024, and 2025 exhibit prominent dark regions—particularly in the upper-left quadrants—indicating substantial dissimilarity or clear cluster separation in those zones, whereas the 2023 map displays a pronounced high-distance boundary concentrated in the bottom-right quadrant. These distinct distance profiles confirm that the optimal SOM models successfully capture structural variations and

boundaries within the socio-economic welfare dataset across the consecutive yearly evaluations.

4) Cluster Profiling

A cluster profiling analysis of the SOM mapping (2022–2025) categorized regions based on BKKBN socio-economic basic needs indicators. In 2022, the population was predominantly concentrated in Cluster 1_0 (61.65%), which represented Welfare Level 1, characterized by extreme prosperity, habitable housing (X_5), asset ownership (X_7), and optimal health (X_13). The remaining groups included Cluster 0_0 (Level 2), which showed economic independence (X_1, X_3) but lacked digital access (X_14); Cluster 0_1 (Level 3), which faced child education vulnerabilities (X_10); and Cluster 1_1 (3.07%, Level 4), a highly deprived group suffering from unsuitable housing (X_6), lack of assets (X_8), and digital isolation (X_15). In 2023, the distribution shifted significantly: Cluster 1_1 (72.49%) became the new Welfare Level 1, driven by high food (X_3) and housing (X_5) fulfillment. Meanwhile, Cluster 0_0 (Level 2) maintained stable basic and educational indicators (X_9), Cluster 1_0 (Level 3) experienced health and disability cost burdens (X_12), and Cluster 0_1 (0.16%, Level 4) emerged as an extreme minority failing to meet basic needs (X_2, X_4, X_6).

In 2024, the clustering pattern stabilized, with Cluster 0_0 (71.36%) acting as Welfare Level 1, dominating in daily income (X_1), assets (X_7), and digital literacy (X_14). Cluster 1_0 represented a stable Level 2 (though vulnerable to food price fluctuations), Cluster 0_1 (Level 3) faced declining housing quality (X_6) and limited education continuity (X_10), and Cluster 1_1 (0.16%, Level 4) suffered the heaviest socio-economic deficits. By 2025, the data distribution dispersed further. Cluster 1_0 (62.94%) formed Welfare Level 1, demonstrating optimal social, economic, and health conditions. Cluster 0_1 (Level 2) showed strong income fulfillment (X_1) but required nutritional improvements (X_4). Notably, the model successfully isolated vulnerable populations requiring priority social intervention: Cluster 0_0 (Level 3), burdened by high rates of illness or disability (X_12), and Cluster 1_1 (Level 4), which faced critical deprivation characterized by uninhabitable housing (X_6) and a complete lack of assets (X_8).

E. Evaluation

Based on the results of the previous discussion, the values for Quantization Error (QE), Topographic Error (TE), and the Silhouette Score have been determined. This section will discuss the details of those results.

1) Quantization Error

The evaluation of the Self-Organizing Maps (SOM) model starts with the analysis of the Quantization Error (QE) to assess the precision of the weight vector mapping.

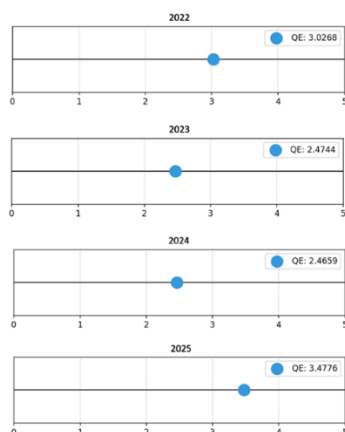


Figure 8. Quantization Error

The mapping accuracy and precision of the Self-Organizing Maps (SOM) model were evaluated using the Quantization Error (QE) metric, which demonstrated strong stability with values ranging from 2.4659 to 3.4776 throughout the 2022–2025 period. The QE improved from 3.0268 in 2022 to its lowest points in 2023 (2.4744) and 2024 (2.4659), indicating that the SOM neurons optimally represented socio-economic variations with minimal data deviation. Although the error rate increased to 3.4776 in 2025—a rise reflecting the more dispersed clustering and higher socio-economic data complexity of that year—the overall QE consistently remained below 3.5. This confirms that the Optuna-optimized SOM architecture accurately, consistently, and reliably compresses multidimensional BKKBN population data without losing essential core characteristics.

2) Topographic Error

The downward trend of this topological error is closely monitored throughout the training iterations to ensure the geometric stability of the neural network, with its visualization clearly presented in Figure 20.

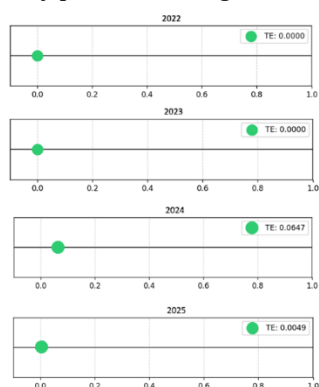


Figure 9. Topographic Error

Spatial continuity was evaluated using the Topographic Error (TE) metric, which demonstrated superior performance

in preserving neighborhood relationships within the network. In 2022 and 2023, the model achieved a perfect TE score of 0.0000, indicating that the first and second Best Matching Units (BMUs) were perfectly adjacent on the grid with zero spatial distortion. Following a minor fluctuation to 0.0002 (0.02%) in 2024 due to increased data dynamics, the TE successfully returned to 0.0000 in 2025. Ultimately, these near-zero TE values, combined with previously established stable QE results, confirm that the Optuna-optimized SOM architecture achieves exceptionally high precision, accuracy, and consistency in compressing multidimensional BKKBN data without losing fundamental population characteristics.

3) Silhouette Score

This test serves as a final confirmation to ensure that the mathematical regional division has formed solid and non overlapping group structures, with the graph of the resulting clustering Silhouette Score presented in Figure 21.

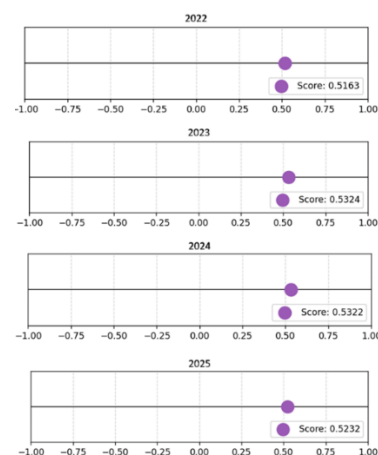


Figure 10. Silhouette Score

Based on the testing results for the 2022 to 2025 period, the SOM model consistently demonstrated a very stable Silhouette Score well above the moderate threshold (0.50). In 2022, the score was recorded at 0.5123, then increased in 2023 to 0.5312, and remained extremely stable in 2024 at 0.5284. Entering 2025, the score was recorded at 0.5195. The consistency of scores remaining within the range of 0.51 to 0.54 throughout the four years of observation indicates that the formation of the 4 welfare groups (2×2 grid) possesses a well-defined clustering structure. The distance among members within the same cluster is relatively tight, while the boundaries separating the clusters are clearly defined. This confirms that the SOM-based clustering of BKKBN socio-economic indicators is not only precise in terms of topology and quantization error, but also yields a welfare level segmentation that is significantly separated and statistically meaningful.

F. Deployment

The geospatial visualizations from 2022 to 2025 reveal a significant aggregate improvement in regional socio-

economic conditions. In 2022, the province was predominantly dominated by Welfare Level 2 across the central, eastern, and southern regions, with optimal Level 1 areas concentrated strictly around urban economic corridors. However, from 2023 to 2025, a major shift occurred as Welfare Level 1 overwhelmingly expanded to cover nearly the entire eastern, southern, northern, and western territories. Consequently, Welfare Level 2 transitioned into a buffer zone sporadically distributed across the central and western areas.

Despite the widespread dominance of high welfare achievements, the SOM model consistently and precisely isolated vulnerable regions requiring targeted government attention. Throughout the four-year period, Welfare Level 3 remained specifically clustered in the north-western corridors (bordering Bogor, Bekasi, and Karawang). Most importantly, Welfare Level 4 remained persistently isolated within a specific south-central pocket of West Java. This highly stable mapping confirms the red-labeled south-central area as the absolute top-priority target for intensive and sustainable BKKBN family welfare program interventions, demonstrating the SOM architecture's effectiveness in tracking welfare gradations and identifying target regions without spatial distortion.

IV. CONCLUSION

The clustering of sub-district BKKBN socio-economic data (2022–2025) was successfully executed using an Optuna-optimized 2×2 Self-Organizing Maps (SOM) architecture. The model demonstrated exceptional stability and precision, validated by strong evaluation metrics: Quantization Error (2.4659–3.4776), Topographic Error (0.0000–0.0647), and a Silhouette Score consistently above 0.50. When transformed into a 4-tier geospatial map of West Java, the results confirmed widespread high welfare across most regions while accurately isolating critical vulnerabilities. Specifically, the map pinpointed a moderate-risk zone in the northwest (Yellow) and a highly vulnerable, top-priority zone in the south-central region (Red), providing clear targets for government interventions.

BIBLIOGRAPHY

- [1] M, Santoso I, Sudiro A, Maryanto S (2026) Spatial Framework for Multi-Hazard Vulnerability Assessment and Sustainable Agricultural Risk Mitigation in Batu City: Integrating Social, Economic, Physical, and Ecological Perspectives. *Journal of Cultural Analysis and Social Change* 952–965. doi: 10.64753/jcasc.v11i1.4001
- [2] Aulia Y, Sulistiowati D, Kurniawati Y, Salma A (2025) Pengelompokan Kabupaten/Kota Maluku dan Nusa Tenggara Barat Berdasarkan Faktor Kemiskinan Menggunakan Self Organizing Maps & Leibniz: *Jurnal Matematika* 5:208–222. doi: 10.59632/leibniz.v5i02.607
- [3] Cai C, Zaghloul M, Li B (2022) Data Clustering in Urban Computational Modeling by Integrated Geometry and Imagery Features for Probabilistic Navigation. *Applied Sciences (Switzerland)* 12. doi: 10.3390/app122412704
- [4] Casonatto RA, Souza TDPG, Mariano AM (2024) Quality and Risk Management in Data Mining: A CRISP-DM Perspective. In: *Procedia Computer Science*. Elsevier B.V., pp 161–168
- [5] Erfina A, Alamsyah MRNR (2023) Implementation of Naive Bayes classification algorithm for Twitter user sentiment analysis on ChatGPT using Python programming language. *Data and Metadata* 2. doi: 10.56294/dm202345
- [6] Esti Suryaningsih, M YN, Wahyuni P (2025) Pemanfaatan Artificial Neural Network dalam Pemetaan Kerentanan Sosial-Ekonomi Terhadap Bahaya Longsor di Kapanewon Kokap, Kabupaten Kulonprogo, DIY. *Jurnal Green Growth dan Manajemen Lingkungan* 15:26–47. doi: 10.21009/jgg.151.02
- [7] Evangelista JRG, Sassi RJ, Romero M (2023) Google Hacking Database Attributes Enrichment and Conversion to Enable the Application of Machine Learning Techniques. *Indian J Sci Technol* 16:3771–3777. doi: 10.17485/IJST/v16i42.1799
- [8] Faiz MF, Fauzan A (2023) Implementasi Metode Clustering SOM Dalam menganalisis Alasan Tidak KB di Kabupaten Temanggung. *Prosiding Seminar Nasional Sains Data* 3:47–58. doi: 10.33005/senada.v3i1.74
- [9] Fosci P, Psaila G (2023) Soft Querying Features in GeoJSON Documents: The GeoSoft Proposal. *International Journal of Computational Intelligence Systems* 16. doi: 10.1007/s44196-023-00325-3
- [10] Gazali M, Chandrawati C, Adini ES (2026) Implementasi Fuzzy Possibilistic C-Means Untuk Pengelompokan Kabupaten/Kota Di Provinsi Nusa Tenggara Barat. *Teorema: Teori dan Riset Matematika* 11:139. doi: 10.25157/teorema.v11i1.20886
- [11] Gupta R, Kumar N, Kumar S, Seth JK (2024) Unsupervised Machine Learning for Effective Code Smell Detection: A Novel Method. *Journal of Communications Software and Systems* 20:307–316. doi: 10.24138/jcomss-2024-0083
- [12] Hidayatul Arifin M, Amelia Vega S, Meliala R, Impana Manik K, Defiyanti A, Syahputra H (2025) Perbandingan Kinerja Knn Dan Svm Dalam Klasifikasi Sampah Organik Dan Anorganik Menggunakan Ekstraksi Fitur HOG DAN LBP. *JATI (Jurnal Mahasiswa Teknik Informatika)* 9:6100–6109. doi: 10.36040/jati.v9i4.13972
- [13] Hilmani A, Sabri Y, Maizate A, Aouad S, Koundi M (2025) Integration of K-Means and Silhouette score for energy efficiency of wireless sensor networks. *International Journal of Reconfigurable and Embedded Systems (IJRES)* 14:26. doi: 10.11591/ijres.v14.i1.pp26-34
- [14] Huang X (2022) Introduction to Programming in Python. In: *Land Carbon Cycle Modeling*. CRC Press, Boca Raton, pp 343–351
- [15] Ishak R, Nurawanti N, Bengnga A (2025) Optimization of K-Means in Disease Clustering of Pregnant Women Using Random Forest. *Jambura Journal of Electrical and Electronics Engineering* 7:41–47. doi: 10.37905/jjee.v7i1.28374
- [16] Isnaini F, Asahan U (2025) Aplikasi Pemodelan Kluster Imunisasi Balita Menggunakan Algoritma Self Organizing Map. *Journal of Science and Social Research* 4352–4360. doi: https://doi.org/10.54314/jssr.v8i3.3878
- [17] Januzaj Y, Beqiri E, Luma A (2023) Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique. *International journal of online and biomedical engineering* 19:174–182. doi: 10.3991/ijoe.v19i04.37059
- [18] Kaminski J, Koch O (2026) χ -SOM: Emergent self-organising Maps for large chemical space visualisation
- [19] Kannengiesser U, Gero JS (2023) Modelling The Design Of Models: An Example Using CRISP-DM. In: *Proceedings of the Design Society*. Cambridge University Press, pp 2705–2714
- [20] Khamidah N, Agustia Astari R, Fitrianto A, Nugraha Pradana A, dan Sains Data S (2023) Penerapan Multi-Clustering Dalam Pengelompokan Kabupaten/Kota Di Provinsi Jawa Barat Berdasarkan Indeks Desa Membangun. 4. doi: 10.46306/lb.v4i1
- [21] Lane R, Cummins L, Perkins A, Trawick G, Banicescu I, Mittal S (2026) parHSOM: A novel parallel Hierarchical Self-Organizing Map implementation. *arXiv preprint*. doi: https://doi.org/10.48550/arXiv.2605.08164
- [22] Laot FS, Katili MR, Lahay SN, Hadjaratie L, A. H (2025) Optimasi UI/UX Berbasis Design Thinking pada Sistem Informasi Keluarga (SIGA). *Studi Kasus: BKKBN Provinsi Gorontalo*. *Jambura Journal of Informatics* 1:98–108. doi: 10.37905/jji.v1i2.33904

- [23] Lewis G, Snider B, Vamvakieridou-Lyroudia L, Chen AS, Djordjević S, Savić DA (2023) Web-based GIS for simulation, and visualisation: EPANET and data-rich shapefiles. In: IOP Conference Series: Earth and Environmental Science. Institute of Physics
- [24] Luo Q, Lin Q, Xu L, Wu S, Mao R, Wang C, Feng H, Huang B, Du Z (2026) GeoJSON agents: a multi-agent LLM architecture for geospatial analysis—function calling vs. code generation. *Big Earth Data* 1–55. doi: 10.1080/20964471.2026.2615511
- [25] Miftahuddin Y, Ridwan ARS (2025) Application of Self-Organizing Map and K-Means to Cluster Bandwidth Usage Patterns in Campus Environment. *Jurnal Online Informatika* 10:66–76. doi: 10.15575/join.v10i1.1438
- [26] Nia Widianengsih, Reny Rian Marlina (2025) Klasterisasi Status Gizi Balita di Jawa Barat Menggunakan Self-Organizing Maps (SOM). *Bandung Conference Series: Statistics* 5. doi: 10.29313/bcss.v5i2.19465
- [27] Nisrina S, Nurmawanti WP, Basirun, Kertanah, Muhammad Gazali (2022) Penerapan Metode Clustering SOM dan DBSCAN dalam Mengelompokkan Unmet Need Keluarga Berencana di Nusa Tenggara Barat. *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika* 15:237–244. doi: 10.36456/jstat.vol15.no2.a5549
- [28] Pankaj Jain M, Soni H, Singh AP, Mehta P, Babani N (2024) Comparative Review of Python and C: Choosing the Right Language for Beginners. *International Journal of Engineering Trends and Applications (IJETA)* 11. doi: <http://dx.doi.org/10.33144/23939516/IJETA-V11I3P28>
- [29] Phan YQ (2023) Researching and applying open-source WebGIS technology to build national-standard vector geographic database management software. *Journal of Mining and Earth Sciences* 64:11–22. doi: 10.46326/JMES.2023.64(6).02
- [30] Prasetyo O, Supianto AA, Anam S, Ferdinandus Pardede H, Zilvan V, Budiarianto R, Kusumo S (2022) Peningkatan Performa Pengelompokan Siswa Berdasarkan Aktivitas Belajar Pada Media Pembelajaran Digital Menggunakan Metode Adaptive Moving Self-Organizing Maps. 9:145–154. doi: 10.25126/jtiik.202295579
- [31] Pura Nurmawanti W, Yuniarti D, Siringoringo M, Purnamasari I, Febriani Putri D, Hadijah Hasanah S (2026) Comparing SOM, DBSCAN, and K-Affinity Propagation in Labor Economic Patterns. 9:2581–2017. doi: 10.30812/varian.v9i1.5933
- [32] Putri Nazwa, Rehana, Rambe YM, Yesnita Aprilia (2025) Menjelajah Konvergensi Barisan Lewat Python Integrasi Analisis Matematis Dan Komputasional. *Proximal: Jurnal Penelitian Matematika dan Pendidikan Matematika* 8:603–613. doi: 10.30605/proximal.v8i2.5732
- [33] Rafafi ER, Sela EI (2023) Identification of Fresh and Unfresh Fish Based on Eye Image Using The Self-Organizing Maps (SOM) Method. *International Journal of Social Service and Research* 3:2826–2833. doi: 10.46799/ijssr.v3i11.593
- [34] Sestri E, Karno ASB, Hastomo W (2025) Benchmarking Five Machine Learning Models for Accurate Steel Plate Defect Detection. *CogITO Smart Journal* 11:382–401. doi: 10.31154/cogito.v11i2.753.382-401
- [35] Siti Alifa Fentiani, Yani Achdiani, Gina Indah Permata Nastia (2025) Peran Kesejahteraan Sosial Dalam Meningkatkan Kualitas Hidup Masyarakat. *Jurnal Ilmu Kesejahteraan Sosial HUMANITAS* 7. doi: 10.23969/humanitas.v7i1.21718
- [36] Soundes M, Ahlam L (2025) Unsupervised Neural Networks for Breast Cancer Clustering: A Comparative Study of RBMs and SOMs with Interpretability Metrics. *Journal of Applied Data Sciences* 6:2351–2360. doi: 10.47738/jads.v6i4.860
- [37] Stepanian I, Yavelov I, Rochagov A, Zidoun M (2024) Cluster-analysis of human pulse wave by self-organizing Cohonen maps to identify markers of age groups. *Russian journal of biomechanics* 28:111–122. doi: 10.15593/RJBiomech/2024.3.12
- [38] Teng Z, Yan J, Liu D, Zhang P (2025) When Does the Silhouette Score Work? A Comprehensive Study in Network Clustering. doi: <https://doi.org/10.48550/arXiv.2512.24841>
- [39] Vardakas G, Papakostas I, Likas A (2024) Deep Clustering Using the Soft Silhouette Score: Towards Compact and Well-Separated Clusters. doi: 10.1007/s10994-026-07026-w
- [40] Varouchakis EA, Solomatine D, Perez GAC, Jomaa S, Karatzas GP (2023) Combination of geostatistics and self-organizing maps for the spatial analysis of groundwater level variations in complex hydrogeological systems. *Stochastic Environmental Research and Risk Assessment* 37:3009–3020. doi: 10.1007/s00477-023-02436-x
- [41] Wandeto JM, Drespe-Langley B (2023) SOM-QE Analysis: A biologically inspired technique to detect and track meaningful changes within image regions[Formula presented]. *Software Impacts* 17. doi: 10.1016/j.simpa.2023.100568
- [42] Wardana A (2022) Peran Pendidikan Ekonomi Dalam Keluarga Perspektif Islam. *Jurnal Ar-Ribhu* 5. doi: 10.46781/ar-ribhu.v5i1.643
- [43] Winartha M, Wirdiastuti C, Salma A (2025) Pengelompokan Kabupaten/Kota di Provinsi Sumatera Barat Berdasarkan Indikator Kesejahteraan Rakyat Menggunakan Algoritma SOM. *Jurnal Riset Statistika* 75–84. doi: 10.29313/jrs.v5i1.6707
- [44] Yanti J, Sari RP, Prawira D (2023) Pemetaan Berbasis Web Tingkat Kemiskinan Menggunakan Model Kesejahteraan Keluarga Pada Aplikasi SIG (Studi Kasus : Kelurahan Desa Kapur). *Coding Jurnal Komputer dan Aplikasi* 11:303. doi: 10.26418/coding.v11i03.53073
- [45] Yun K, Yun H, Lee S, Oh J, Kim M, Lim M, Lee J, Kim C, Seo J, Choi J (2024) A Study on Machine Learning-Enhanced Roadside Unit-Based Detection of Abnormal Driving in Autonomous Vehicles. *Electronics (Switzerland)* 13. doi: 10.3390/electronics13020288
- [46] Zamri N, Bakar NAA, Aziz AZA, Madi EN, Ramli RA, Sukono, Koon CS (2025) Tuning Data Preprocessing Techniques in Enhancing the Accuracy of Self-Organizing MAP (SOM) for Prostate Genomic Cancer. In: *Procedia Computer Science*. Elsevier B.V., pp 892–899