Jurnal Matematika UNAND
Vol. **15** No. **1** Hal. 108 – 122
Edisi Januari 2026
ISSN : 2303–291X
e-ISSN : 2721–9410

https://doi.org/10.25077/jmua.15.1.108-122.2026

CLASSIFICATION OF GENDER INEQUALITY IN INDONESIA: UNSUPERVISED AND SUPERVISED LEARNING APPROACHES

MAULIDA NURHIDAYATI*, YUNAITA RAHMAWATI, AJENG WAHYUNI
Universitas Islam Negeri Kiai Ageng Muhammad Besari Ponorogo
email : nurhidayati@uinponorogo.ac.id, rahmawati@uinponorogo.ac.id, ajeng@uinponorogo.ac.id

Received July 31, 2025, Received in revised form January 16, 2026
Accepted January 20, 2026 Available online January 31, 2026

Abstract. *Gender inequality in Indonesia is a multidimensional problem that has a large impact on human development. This study aims to model and classify the level of gender inequality between provinces in Indonesia with a combined approach of unsupervised and supervised learning. Secondary data from 38 provinces in 2024 were analyzed using five methods: K-Means, Self-Organizing Map (SOM), hybrid SOM K-Means, Support Vector Machine (SVM), and Logistic Regression. In the unsupervised approach, the SOM and SOM K-Means methods show better cluster coherence than K-Means. In the supervised approach, the SVM method provides better classification performance compared to logistic regression. In general, SVM was obtained with the highest accuracy, which was 89.47%, surpassing other methods. This research makes a methodological contribution to the use of machine learning for spatial-based gender inequality risk mapping, as well as implications for more precise and adaptive data-based policymaking.*

Keywords: Gender Inequality, Machine Learning, Regional Classification.

1. Introduction

Gender inequality is a complex and multidimensional global issue that reflects differences in access, opportunities, and outcomes between men and women in various areas of life. These problems not only include gaps in education or employment, but also reflect structural inequalities in decision-making, access to resources, and civil and social rights. In Indonesia, this inequality remains a serious challenge in inclusive and equitable human development. Although affirmative policies and gender mainstreaming programs have been implemented since the Reformation era, such as Presidential Instruction No. 9 of 2000 regarding gender mainstreaming in national development, quantitative indicators show that gender disparities remain high in some provinces, especially in the dimensions of economic participation and political representation ([1], [2]).

*Corresponding author

Measurement of gender inequality in Indonesia using the Gender Inequality Index (GII) developed by the Central Statistics Agency (BPS) based on the adaptation of the Gender Inequality Index (GII) belonging to UNDP. The CPI measures inequality on three main dimensions: reproductive health, empowerment, and the labor market. The index value ranges from 0 (perfect equality) to 1 (total inequality). Although these figures seem simple, the social realities underlying gender inequality are complex and differ between regions [3]. In 2024, the national GII will be recorded at 0.421, down 0.026 points from the previous year (0.447). This decline is the fastest improvement over the past six years, more than doubling the achievement in 2023. The most significant improvement occurs in the dimensions of the labor market, with the Labor Force Participation Rate for women increasing by 1.90 points, from 54. 52% in 2023 to 56. 42% in 2024, while men only increased by 0.40 points [2]. The downward trend in Indonesia's GI from 2018 to 2024 in Figure 1 shows consistent improvement.

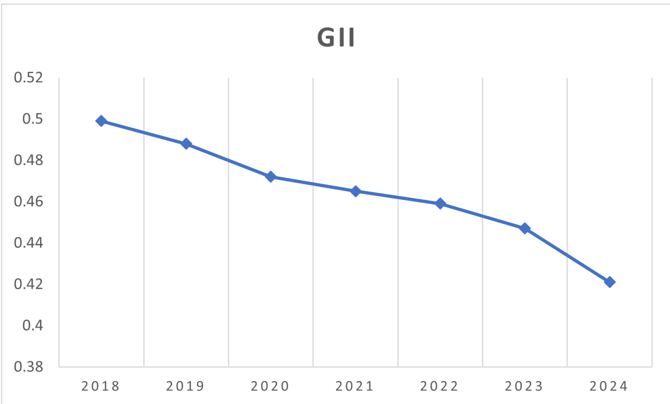

Figure 1. Development of GII in Indonesia

From Figure 1, it is evident that not all regions show an improvement trend, as five provinces experienced increases in GII scores in 2024 compared to 2023, including Maluku, East Kalimantan, DI Yogyakarta, Riau, and Central Kalimantan. However, DKI Jakarta, West Nusa Tenggara, and the Bangka Belitung Islands experienced the most significant decrease in GII [2]. A low GII does not always reflect a good level of equality, but can show equally low participation of both genders [4]. For example, provinces in Eastern Indonesia, such as Papua and East Nusa Tenggara, have relatively low GII scores, but their economic and educational participation is also low. In contrast, provinces such as Bali, Yogyakarta, and DKI Jakarta have high GII scores, reflecting actual disparities in the distribution of development benefits between genders. Although in general, both groups have better access to education and employment.

In previous studies, gender inequality has been significantly associated with development variables such as economic growth, Human Development Index (HDI), poverty, and social welfare. In [5], it is found that an increase in the Gender Inequality Index (GII) was strongly associated with a decrease in per capita income and an increase in women's household poverty. In [6], Duflo was underlining that women's empowerment and economic development have a reciprocal relationship where development can reduce gender inequality, while women's empowerment can accelerate development. However, a review of the literature suggests that these reciprocal relationships may not be strong enough to be sustainable automatically.

Given this complexity, the identification of areas at high risk of gender inequality is important as a basis for policy-making that is data-driven and responsive to local contexts. Machine learning methods such as unsupervised learning (K-Means and Self-Organizing Map/SOM) and supervised learning (Support Vector Machine/SVM and Logistic Regression) have been shown to be effective in various studies to classify regions spatially and predict socio-economic status based on historical data patterns.

Research [7] is a review and meta-analysis of 215 articles (2014–2023) on the use of Machine Learning (ML) for poverty mapping. ML has proven to be effective in uncovering micro-geographic poverty patterns through satellite data analysis. Nighttime light index is a strong indicator, especially when combined with daytime data such as land cover. Random Forest is the most accurate and easy-to-interpret model. The study provides new methodological insights and aims to support data-driven policymaking for more equitable development. Research [8] modeled the Gender Inequality Index (GII) in Indonesia in 2020 using the Naive Bayes algorithm to predict the level of gender inequality based on gender-related indicators, such as access to childbirth, education, labor participation, and women's representation in parliament. Models produce high accuracy: 82.86% on training data and 83.72% on the test data. These results show the potential of the model in predicting GI and as a policy basis to address gender inequality in the regions. Research [9] using the clustering method AGNES and K-Means to group provinces based on reproductive health indicators. K-Means Method More effectively, resulting in three clusters: 15 provinces (medium), 10 (high), and 9 (low). These results are expected to support more targeted policies in improving reproductive health and gender equality in Indonesia.

The variables used in the analysis of gender inequality in this study are derivative indicators that are directly sourced from the components that make up the Gender Inequality Index developed by the Central Statistics Agency (BPS) based on the UNDP Gender Inequality Index (GII) methodology. Conceptually, GII is constructed from three main dimensions, namely reproductive health, empowerment, and the labor market, each of which is represented by specific gender-based indicators ([2], [3]). In this study, these dimensions are operationalized into five indices, namely the Women's Reproductive Health Index, the Women's Empowerment Index, the Women's Labor Market Index, the Male Empowerment Index, and the Male Labor Market Index [2]. These variables not only represent statistical aspects but also illustrate the power structure, distribution of access, and systemic barriers experienced by women in society. With the right analytical approach, this data can be used to construct regional classifications based on the level of inequality: high, medium, low, and very low.

Modeling gender inequality with a combination of clustering and classification techniques offers two main benefits. First, clustering through methods such as SOM and K-Means allows the exploration of regional segmentation based on the similarity of inequality characteristics. It is useful for spatial analysis and region-based intervention planning. Second, supervised learning provides predictive ability to estimate the status of regional inequality with high accuracy, especially through techniques such as SVM that are able to handle non-linear data efficiently. This research fills a gap in the literature by combining unsupervised and supervised learning approaches simultaneously in the analysis of gender inequality. Where this study presents the results of the two approaches side by side. By integrating predictive and exploratory modeling based on GI indicators and spatial data, the study makes a significant methodological contribution.

2. Theoretical Foundations

2.1. *Gender Inequality*

Gender inequality is a form of structural injustice that is reflected in disparities in access and outcomes between men and women in various aspects of life, such as education, employment, health, and political representation[2]. In Indonesia, the measurement of inequality is carried out through the Gender Inequality Index (GII) developed by BPS, which adapts the Gender Inequality Index (GII) from UNDP. The GII combines three main dimensions: reproductive health, empowerment, and participation in the labor market. In the spatial context, the regional classification approach based on gender indicators is important to understand the patterns of inequality spread between provinces.

2.2. *Unsupervised Learning for Clustering*

Unsupervised learning is used to identify hidden patterns or structures in unlabeled data. This study uses three main methods:

2.2.1. *K-Means Clustering*

K-Means is a partition-based clustering method that divides data into k groups by minimizing the distance between data points and the cluster centroid. Although popular for its simplicity, K-Means is sensitive to initial center selection and is not suitable for non-spherical or complex structured data [10].

2.2.2. *Self-Organizing Map (SOM)*

SOM is an unsupervised neural network algorithm used to map high-dimensional data into two-dimensional representations while maintaining topological proximity. SOM excels at handling multivariable social data due to its ability to visualize spatially complex non-linear and complex structures [11]. SOM is also effective for identifying cohesive patterns and data segmentation that simple partitioning methods can not capture.

2.2.3. *Hybrid SOM-KMeans*

This method combines the topological visualization advantages of SOM and the partitioning efficiency of K-Means. The process begins with SOM training to map data to two-dimensional space, then the projection results are used as inputs for K-Means to improve the formation of clusters.

2.3. *Supervised Learning for Classification*

Supervised learning is an approach in machine learning where models are built based on data that has a target label (dependent variable). The two supervised methods used in this study are:

2.3.1. *Support Vector Machine (SVM)*

SVM is a classification algorithm that works by searching for the optimal hyperplane to separate data in high-dimensional space. The main advantage of SVM is its ability to handle non-linear data through the use of kernel functions, as well as its resistance to overfitting [12]. In this study, SVM was implemented using a radial basis function (RBF) kernel, because this kernel is effective in capturing non-linear relationships and complex interactions between social variables that cannot be modeled linearly. SVM is well-suited for complex social data and has non-linear interactions between variables.

2.3.2. *Logistic Regression*

Logistic regression is a linear classification method used to model the probability of a binary or multicategory event. This model is often used in the social sciences due to its intuitive interpretation of coefficients. However, logistic regression has limitations in addressing non-linear relationships and multicollinearity between variables, which are commonly found in multivariate social data [13].

3. Research Methods

This study uses an applied quantitative approach that focuses on the numerical and systematic exploration of the phenomenon of gender inequality. The purpose of this study is to classify provinces in Indonesia based on the characteristics of gender inequality using unsupervised learning (clustering) techniques, as well as to predict regional inequality categories using supervised learning (classification). The developed model is not only exploratory, but also has predictive value that can be used as a data-driven policy planning tool.

The data used in this study is secondary data obtained from the Central Statistics Agency (BPS). The unit of analysis in this study is 38 provinces in Indonesia, with the scope of observation time for the last available year, which is in 2024, considering that at the time this study was conducted, the available data was for 2024. This study uses Gender Inequality Index (GII) data as a dependent variable. Independent variables are taken from derivative indicators from the components that make up the GII, namely the female and male human development index (HDI), the Labour Force Participation Rate (LFPR) of women and men, women's involvement in parliament (WPP), women as professionals (WPro), the average length of schooling of women (MYS) and men, and the contribution of women's income(WIC).

This research consists of three stages, namely:

(1) A categorization of dependent variables was carried out constructed from the value of the Gender Inequality Index (GII) with four classes, namely very low inequality ($GII \leq Q1$), low inequality ($Q1 < GII \leq Q2$), moderate inequality ($Q2 < GII \leq Q3$), and high inequality ($GII > Q3$). This study acknowledges that sensitivity analysis to the quartile limit has not been performed. Therefore, the categorization of gender inequality levels is entirely based on the quantile approach of the Gender Inequality Index (GII) distribution, which was chosen to maintain class balance given the limitations of the sample size ($N = 38$).
(2) Clustering analysis was carried out to identify provincial grouping patterns based on gender indicators. The three methods used in this stage are K-Means Clustering, Self-Organizing Maps (SOM), and Hybrid SOM-K-Means, given that both have advantages in capturing spatial structure and segmenting multivariable data [12].
(3) Classification was carried out to predict gender inequality categories in provinces based on input characteristics. Two classification methods were used in this study, namely Support Vector Machine (SVM) and Logistic Regression.

The entire data processing and analysis process is carried out using R software using cluster libraries, factoextra, kohonen, nnet, and e1071. The steps to implement the analysis include:

(1) Data collection and integration from the www.bps.go.id.
(2) Pre-processing of data by standardizing data because the data have different units.
(3) Clustering and classification analysis.
(4) calculation of accuracy, recall, precision, F1 using the following calculations [14]:

$$\text{Accuracy } (A) = \frac{tp + tn}{tp + tn + fp + fn}, \tag{3.1}$$

$$\text{Precision } (P) = \frac{tp}{tp + fp}, \tag{3.2}$$

$$\text{Recall } (R) = \frac{tp}{tp + fn}, \tag{3.3}$$

$$F_1 = \frac{2 \cdot (P \cdot R)}{P + R}, \tag{3.4}$$

where:

- tp (true positive) is the number of correctly predicted positive instances,
- tn (true negative) is the number of correctly predicted negative instances,
- fp (false positive) is the number of incorrectly predicted positive instances,

- fn (false negative) is the number of incorrectly predicted negative instances.

(5) Determination of the best method. The selection of this quantitative method is based on the ability of statistical approaches to produce Objective Patterns of numerical data and the capabilities Machine Learning in capturing the complexity of non-linear relationships that are often not simply explained by classical models. Thus, this study seeks to contribute a new methodological approach to the study of gender inequality in Indonesia by combining the exploratory and predictive strengths of modern statistical science [12].

4. Result and Discussion

4.1. *Data description*

Before conducting the analysis, the data is first described to obtain an initial understanding of the structure and characteristics of the analyzed data. This stage includes presenting the mean and standard deviation, as well as looking at the minimum and maximum values of the data.

The GII is classified into four classes so that the basis of classification is based on the values of the 1-3 quartile according to Table 1. Based on Table 1, the results of the classification are obtained as in the following Table 2.

Table 1. Gender Inequality Classification

Yes	**Category**	**Code**	**Border**
1	Very low inequality	1	$GII \leq 0.3805$
2	Low inequality	2	$0.3805 < GII \leq 0.458$
3	Moderate inequality	3	$0.458 < GII \leq 0.5122$
4	High inequality	4	$GII > 0.5122$

Table 2. Results of Gender Inequality Classification

Valid	**Frequency**	**Percent**	**Valid Percent**	**Cumulative Percent**
1.00	9	23.7	23.7	23.7
2.00	9	23.7	23.7	47.4
3.00	11	28.9	28.9	76.3
4.00	9	23.7	23.7	100.0
Total	**38**	**100.0**	**100.0**	

Based on the data presented in Table 2, it is known that there are 9 provinces that are included in the category of very low gender inequality, there are 9 provinces that are included in the category of low gender inequality, there are 11 provinces that are included in the category of moderate gender inequality, and there are 9 provinces that are included in the category of high gender inequality.

Based on the results of the descriptive analysis of 38 observations, an overview of the characteristics of variables related to human development and gender-based socio-economic indicators was obtained. The average value of women's Human Development Index (HDI) (HDI) of 70.57 shows that women's quality of life, viewed from the aspects of education, health, and economy, is still below men, who have an average HDIM of 77.36. This indicates that there is a gender gap in human development achievement. In the WPP variable or the proportion of women in parliament, the average of only 18.44% indicates that women's representation in political decision-making is still low. This is also supported by a relatively high standard deviation value (7.26), showing inequality between regions. Meanwhile, the average number of Women as Professionals (WPro) is at 49.57%, which means that almost half of women are still giving birth outside health facilities, indicating a potential high risk to maternal and infant health.

In terms of education, the average means of years schooling for women (MYSF) is 8.61 years, slightly lower than that of men (MYSM = 9.22 years). This reflects that despite relatively equal access to education, women still have shorter formal learning times. This is also in line with the low economic contribution of women, as seen from the average WIC (Women with Income Contribution) of 33.84%, meaning that only one-third of women contribute directly to household income or the economy. In terms of labor force participation, women have an average LFPRF of 56.82%, far below men's 84.08% (LFPRM). These differences reinforce the findings of inequality in women's access to and participation in the formal and productive sectors. The high standard deviation value in women's labour participation (7.58) also suggests that this inequality is more varied between regions than for men, who have a more uniform distribution (SD = 2.60).

Table 3. Statistical Description Results

	N	Minimum	Maximum	Mean	Std. Deviation
HDIF	38	51.75	83.19	70.5658	6.17458
HDIM	38	59.24	86.21	77.3595	4.61748
WPP	38	8.57	45.24	18.4358	7.25789
WPro	38	24.01	59.82	49.5721	6.73105
MYSF	38	3.40	11.19	8.6079	1.34849
MYSM	38	5.35	11.83	9.2168	1.13948
WIC	38	24.85	44.58	33.8350	4.53193
LFPRF	38	47.05	85.71	56.8203	7.57900
LFPRM	38	76.62	90.37	84.0789	2.60288

4.2. *Clustering Based on K-Means*

Data clustering using the K-means algorithm was carried out using R software. The results of clustering using the K-Means method in Table 4, show that the level of conformity between the formed cluster and the original class label has the same value of 33% with an accuracy rate of 23.68%. This low level of accuracy indicates that the K-Means method has not been able to capture the natural structure of the

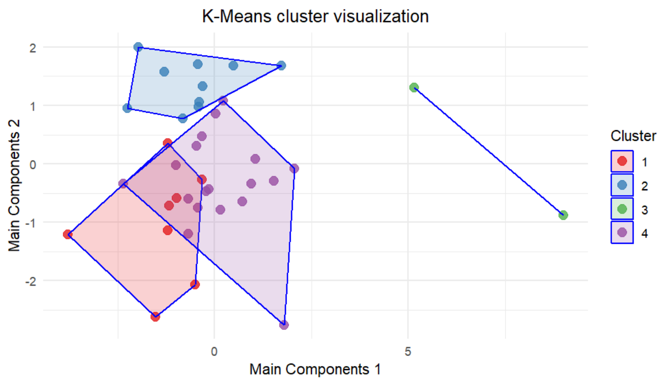

Figure 2. Visualization of GII Clustering based on the K-Means Algorithm Development of GII in Indonesia

data, due to overlap between classes, data distribution that is not clearly separated, or the number of clusters that do not reflect the complexity of the data. Considering that K-Means is an unsupervised algorithm, it is not uncommon for a mismatch with the original label. The result of this low grouping is also supported by the visual results in Figure 2, which shows the results of overlapping.

Table 4. Clustering of GII based on K-Means Algorithm

Predicted	Actual				Percentage Correct	Precision	Recall	F1
	1	2	3	4				
1	3	2	2	1	33%	37.5%	33%	35%
2	2	3	2	3	33%	30.0%	33%	32%
3	0	0	0	2	0%	0.0%	0%	0%
4	4	4	7	3	33%	16.7%	33%	22%
Total	9	9	11	9	**23.68%**	**21.04%**	**25.00%**	**22%**

4.3. *Clustering based on Self-Organizing Maps (SOM)*

The clustering analysis was then carried out using the Self-Organizing Map (SOM) method. This method is used to group complex and multidimensional data into two-dimensional visual forms, making patterns and relationships between data easier to recognize. SOM was chosen because it is able to capture non-linear structures in the data and produce clusters that are more meaningful than conventional clustering methods. The results of clustering with SOM are shown in Table 5.

The results of clustering using the SOM method shown in Table 5 show that the level of conformity between the formed clusters and the original class labels is relatively moderate. Of the total four classes observed, about 39.47% of the data were correctly classified into the appropriate clusters. The 3rd class had the highest

Table 5. Clustering of GII based on SOM

Predicted	Actual				Percentage Correct	Precision	Recall	F1
	1	2	3	4				
1	3	1	2	1	33%	43%	33%	38%
2	0	0	0	0	0%	0%	0%	0%
3	6	7	8	4	73%	32%	73%	44%
4	0	1	1	4	44%	67%	44%	53%
Total	9	9	11	9	**39.47%**	**35.38%**	**37.63%**	**34%**

match rate of 73%, which indicates that most of its members were well grouped into one cluster. However, for the rest of the class, the classification accuracy is much lower at 33%, 44% and even 0%.

4.4. *Clustering based on Hybrid SOM-Kmeans*

The next clustering was carried out using the Hybrid SOM-KMeans approach, which combines the advantages of Self-Organizing Map (SOM) in reducing dimensions and revealing the topological structure of data, with the K-Means algorithm which is effective in forming clusters based on data centers (centroids). The results of clustering with Hybrid SOM-KMeans are shown in Table 6.

Table 6. Clustering of GII based on hybrid SOM and K-Means Algorithm

Predicted	Actual				Percentage Correct	Precision	Recall	F1
	1	2	3	4				
1	4	3	4	1	44%	33.3%	44%	38%
2	2	4	2	4	44%	33.3%	44%	38%
3	3	2	5	2	45%	41.7%	45%	43%
4	0	0	0	2	22%	100.0%	22%	36%
Total	9	9	11	9	**39.47%**	**52.08%**	**39.14%**	**39.01%**

The results of clustering using the hybrid SOM K-means method, shown in Table 6, show that the level of conformity between the formed cluster and the original class label is relatively moderate. Of the total four classes observed, about 39.47% of the data were correctly classified into the appropriate clusters. The 3rd class has the highest match rate, which is 45%. However, for other classes, the accuracy of the classification is much lower. Figure 3 shows a data visualization that shows that there are still intersecting groups.

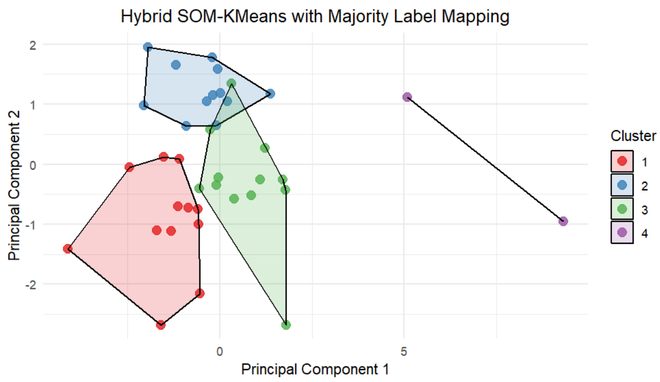

Figure 3. Visualization of GII Clustering based on Hybrid SOM K-Means

4.5. *Classification by Support Vector Machine (SVM)*

The classification process at this stage is carried out using the Support Vector Machine (SVM) method, which is a highly effective supervised learning algorithm to separate data into classes based on optimal decision limits (hyperplane).

Table 7. Clustering of GII based on SVM

Predicted	Actual				Percentage Correct	Precision	Recall	F1
	1	2	3	4				
1	9	0	0	0	100%	100%	100%	100%
2	0	5	2	2	56%	55.6%	56%	56%
3	0	0	11	0	100%	100%	100%	100%
4	0	0	0	9	100%	100%	100%	100%
Total	9	9	11	9	**89.47%**	**88.89%**	**88.89%**	**88.89%**

The classification results using the SVM method shown in Table 7 show that the level of conformity between the formed cluster and the original class label is relatively high. Of the total four classes observed, about 89.47% of the data were correctly classified into the appropriate clusters. The 1st, 3rd and 4th classes had a 100% match rate and only the 2nd class was clustered at 56%.

4.6. *Classification by Logistic Regression*

Classification at this stage is also carried out using logistic regression, which is a classical statistical method that is widely used to model the relationship between one category-dependent variable and one or more independent variables.

The results of the classification using the logistic regression method, shown in Table 8, show that the level of conformity between the formed clusters and the

Table 8. Clustering of GII based on Logistic Regression

Predicted	Actual				Percentage Correct	Precision	Recall	F1
	1	2	3	4				
1	9	0	0	0	100%	100%	100%	100%
2	0	8	2	0	89%	80%	89%	84%
3	0	1	9	2	82%	75%	82%	78%
4	0	0	0	7	78%	100%	78%	88%
Total	9	9	11	9	**86.84%**	**88.75%**	**87.12%**	**87.49%**

original class labels is relatively high. Of the total four classes observed, about 86.84% of the data were correctly classified into the appropriate clusters. The 1st class has a 100% match rate, and the other class is less than 100%.

4.7. *Clustering and Classification Accuracy*

Table 9. Clustering and Classification Accuracy

	Method	**Accuracy**	**Precision**	**Recall**	**F1**
Unsupervised	K-Means	23.68%	21.04%	25.00%	22%
	SOM	36.84%	35.38%	37.63%	34%
	SOM-Kmeans	36.84%	52.08%	39.14%	39.01%
Supervised	SVM	89.47%	88.89%	88.89%	88.89%
	Logistic Regression	86.84%	88.75%	87.12%	87.49%

Unsupervised Learning Outcomes:
In the clustering stage, both SOM and SOM-KMeans produce similar and generally better regional groupings compared to K-Means. Although similar, SOM-Kmeans has a higher average of precision, recall, and F1 values compared to SOM, which indicates in this study, the SOM K-Means hybrid method is a better model compared to SOM. Self-Organizing Maps, as an artificial neural network-based model, are able to map multidimensional data into two-dimensional spaces by maintaining topological proximity between the data. This advantage makes SOM more adaptive to variations in social and spatial indicators than centroid methods such as K-Means, which often fail to capture non-spherical data structures.

Furthermore, the results of the hybrid SOM-KMeans combining the advantages of SOM non-linear mapping and K-Means partitioning efficiency shows that this approach is effective in improving cluster quality and reducing overlap between groups. This is in line with the findings in the study [15] shows that the SOM-KMeans excels in mapping drought projections in China because it is able to handle complex and high-dimensional climate data more effectively than conventional clustering methods. Self-Organizing Map (SOM) efficiently reduce dimensions and uncover non-linear patterns in climate data, while K-Means Fine-tune the formation of clusters based on the structure that the SOM has mapped. The combination of these two stages results in clusters that are more stable, representative, and accurate for modeling spatial and temporal variations of droughts, compared to K-Means or traditional statistical techniques that are less sensitive to spatial complexity and climate change. Instead K-Means, although often used due to its convenience, tends to result in clusters that are less stable and less representative in this context. This can be explained by the nature of K-Means which only considers Euclidean distances and is sensitive to the selection of the initial center point [10]. In the case of spatial and social data such as gender inequality, with very diverse variables, this model is not able to capture the relationships between complex and nonlinear variables.

Supervised Learning Outcomes:
Method Support Vector Machine (SVM) occupy the top position in the classification of gender inequality levels between provinces. SVM has the advantage of handling high-dimensional and complex data, as well as excelling in finding optimal margins between classes, making it ideal for social classification cases with non-linear interacting variables. These findings are reinforced by studies [16] Support Vector Machine (SVM) shows the highest accuracy, i.e. 0.9604 compared to LDA, CART, k-NN, SVM (RBF kernel), and Random Forest. In the context of social data and development, SVM often shows an advantage due to its flexibility in using kernel functions and its resistance to outliers[12]. Instead Logistic Regression, although it is a classic method widely used for category classification, gives lower results. Research [17] use Binary Logistic Regression to predict the opportunity to change land into a built-up area in Kediri City. This method excels at identifying the relationship between spatial variables—such as distance to roads, activity centers, and population density—and land conversion possibilities. The advantage of logistic regression lies in its ability to handle complex independent variables and generate robust probabilistic models to support data-driven spatial planning.

Comparison Results of Unsupervised and Supervised Learning
This study compares the performance of five methods in modeling gender inequality in Indonesia, namely two supervised learning methods (SVM and Logistic Regression), two pure unsupervised learning methods (K-Means and SOM), and one hybrid method (SOM-KMeans). The results show that SVM produces the best classification performance based on evaluation metrics such as accuracy, precision, and F1-score. Meanwhile, in the unsupervised learning approach, SOM and SOM-KMeans provide identical and better grouping results than the classical K-Means method, especially in terms of cluster coherence and topological structure. The results of this study reinforce the importance of selecting algorithms that correspond to the characteristics of complex, multivariate and often non-linear distribution patterns. SVM proved to be the most effective method in classifying gender inequality

status, given its ability to deal with data structures that are difficult for classical models to handle. On the other hand, unsupervised methods such as SOM and SOM-KMeans show high exploratory power in identifying hidden patterns between provinces, providing significant spatial insights in regional segmentation based on gender characteristics. Thus, the combination of predictive (SVM) and exploratory (SOM/SOM-KMeans) approaches results in a comprehensive and adaptive modeling strategy, which can support the evidence-based policymaking process on gender inequality issues. These findings show the great potential of the use of hybrid-based machine learning in social mapping in Indonesia.

5. Conclusion

This study successfully identified and classified the level of gender inequality between provinces in Indonesia by utilizing five machine learning methods consisting of two main approaches: *unsupervised learning* (K-Means, SOM, and hybrid SOM-KMeans) and *supervised learning* (SVM and logistic regression). From the results of the analysis, it can be concluded that:

(1) In the unsupervised learning approach, the SOM and SOM-KMeans methods show advantages in grouping regions based on gender indicators with high cluster coherence, surpassing the performance of the K-Means method which tends to be less stable for multivariate social data. Although similar, SOM-Kmeans has a higher average of precision, recall, and F1 values compared to SOM, which indicates in this study, the SOM K-Means hybrid method is a better model compared to SOM.
(2) In the supervised learning approach, SVM is the best method in terms of gender inequality risk classification, with an accuracy of 89.47%, due to its ability to handle non-linear and high-dimensional data optimally.
(3) In the unsupervised learning and supervised learning approaches, SVM is the best model compared to other methods.
(4) The combination of these two approaches results in a more comprehensive modeling framework, as it can identify latent structures and at the same time provide predictive capabilities based on spatial and social data.
(5) These findings support the use of machine learning methods in public policy, especially in developing gender mainstreaming strategies that are more adaptive to the spatial and social context in Indonesia.

This study fills a gap in the literature on gender inequality by using 2 approaches to unsupervised learning and supervised learning. In addition, the results can be used as a basis for local and central governments to formulate more targeted and precise region-based policies in dealing with the issue of gender inequality.

Bibliography

[1] Nations, U. 2025, *Human Development Report 2025*, In Human Development Reports. United Nations. https://hdr.undp.org/content/human-development-report-2025

[2] Badan Pusat Statistik Indonesia, 2025, *Indeks Ketimpangan Gender (IKG) 2024*, Badan Pusat Statistik Indonesia

[3] Tusianti, E., 2020, *Kajian Penghitungan Indeks Ketimpangan Gender*, BPS-Statistics Indonesia

[4] Kertati, I. 2021. Analisis Indeks Pembangunan Gender (IPG) dan Indeks Pemberdayaan Gender (IDG) Kota Surakarta, *Public Service and Governance Journal*, *2*(1): 01 – 11, https://doi.org/10.56444/psgj.v2i1.597

[5] Sitorus, A. V. Y., 2016, Dampak Ketimpangan Gender terhadap IPM dan Kemiskinan di Indonesia, *Jurnal Ekonomi dan Pembangunan Sosio Informa*, *2*(1): 89 – 101

[6] Duflo, E., 2012., Women Empowerment and Economic Development, *Journal of Economic Literature*, *50*(4): 1051 – 1079, https://doi.org/10.1257/jel.50.4.1051

[7] Lamichhane, B. R., Isnan, M., Horanont, T., 2025, Exploring machine learning trends in poverty mapping: A review and meta-analysis, *Science of Remote Sensing*, *11*: 100200, https://doi.org/10.1016/j.srs.2025.100200

[8] Yolanda, A. M., Adnan, A., Dwiputri, A. A., 2022, Pemodelan Klasifikasi Pada Indeks Ketimpangan Gender (IKG) Tahun 2020 Dengan Metode Naïve Bayes, *Jurnal Keluarga Berencana*, *7*(1): 22 – 31, https://doi.org/10.37306/kkb.v7i1.118

[9] Raihannabil, S. D., Ilyas, H. M. A., Shafira, H. N., Riani, M. A., Hastin, N. N., Siregar, T. K. H., 2024, Perbandingan Agglomerative Nesting dan K-Means untuk Klasterisasi Ketimpangan Gender berdasarkan Dimensi Kesehatan Reproduksi, *Seminar Nasional Official Statistics*, *2024*(1), Article 1, https://doi.org/10.34123/semnasoffstat.v2024i1.1977

[10] Han, J., Kamber, M., Pei, J., 2012, *Data mining: Concepts and techniques*, Morgan Kaufmann

[11] Kohonen, T. 2012. *Self-organizing maps.* Springer Science and Business Media

[12] James, G., Witten, D., Hastie, T., Tibshirani, R., 2013, *An Introduction to Statistical Learning* (Vol. 103), Springer New York, https://doi.org/10.1007/978-1-4614-7138-7

[13] Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X., 2013, *Applied logistic regression.* John Wiley & Sons

[14] Dalianis, H., 2018, Evaluation Metrics and Evaluation, in *Clinical Text Mining: Secondary Use of Electronic Patient Records*, Springer International Publishing: 45 – 53, doi: 10.1007/978-3-319-78503-5_6

[15] Zeng, P., Sun, F., Liu, Y., Wang, Y., Li, G., Che, Y., 2021, Mapping future droughts under global warming across China: A combined multi-timescale meteorological drought index and SOM-Kmeans approach, *Weather and Climate Extremes*, *31*: 100304, https://doi.org/10.1016/j.wace.2021.100304

[16] Yolanda, A. M., Adnan, A., Goldameir, N. E., Rizalde, F. A., 2023, The comparison of accuracy on classification data with machine learning algorithms (Case study: Human development index by regency/city in Indonesia 2020), *AIP Conference Proceedings*, *2536*(1): 060002, https://doi.org/10.1063/5.0118720

[17] Sabila, M. I., Sutikno, F. R., 2024, Pengembangan Infrastruktur Hijau Dalam Pembangunan Green City di Kota Kediri, *Planning for Urban Region and Environment Journal (PURE)*, *13*(4): 195 – 206