Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Application of K-Means on Human Rights, Demographic, Economic, and Crypto Investment Data

Akianus Wenda [1]*, Antonius R Kopong Notan[2], Shalwa Azizah Rananda Sudirman [3], T. Ferdiansyah[4], Tegar Surya Pratama[5], Zurnan Alfian[6]

[1]Faculty of Computer Science, Information Technology Study Program, Pamulang University, South Tangerang, Indonesia
akianuswenda@gmail.com[1]*

Abstract

This study combines the K-Means Clustering and Decision Tree methods to analyze multidomain data covering economic and social human rights, demographics, poverty, crypto investment, and sustainable financing in Indonesia's financial services sector. Data was obtained from various credible sources such as the National Commission on Human Rights (Komnas HAM), the Central Statistics Agency (BPS), the Financial Services Authority (OJK), and scientific publications (2019–2023), then processed through missing value handling, outlier detection, and normalization using Min-Max Scaling and Z-score. K-Means was used to group regions based on the similarity of socio-economic and financial indicators, while Decision Tree was used to classify financial entities based on ESG (Environmental, Social, and Governance) scores. Model evaluation was conducted using WCSS, Silhouette Score, Davies-Bouldin Index, and classification accuracy. The results show the formation of clusters representing different levels of inequality and sustainability in Indonesia. This approach contributes to understanding the dynamics of multidimensional development and provides a basis for more adaptive and sustainable policies in the socio-economic and financial sectors.

Keywords: *Data Mining; K-Means; Decision Tree; ESG; Human Rights; Sustainable Finance.*

1. Introduction

In today's digital age and era of data openness, the use of data mining methods has become increasingly important in analyzing complex social and economic phenomena. Indonesia, as a developing country with high social dynamics, faces various challenges related to equitable development, economic inequality, and the fulfillment of human rights (HAM) in economic and social aspects. Meanwhile, the rapid development of financial technology, such as cryptocurrency investments, signifies a shift in societal economic behavior that requires understanding from both policy and social perspectives. The complexity of the interrelationships between these factors demands an analytical approach capable of efficiently grouping multidimensional data to uncover hidden patterns not visible through conventional analysis.

This research is relevant because it integrates various indicators from different domains, such as the implementation of economic and social human rights, demographic structure, poverty levels, and cryptocurrency investment trends. The grouping or clustering of this data is expected to provide a comprehensive picture of Indonesia's socio-economic conditions in recent years, particularly during the 2020–2024 period. Using the K-Means clustering approach, this study aims to identify groups of regions or entities with similar characteristics based on these indicators, thereby supporting decision-making processes by the government, investors, and academics.

The data sources used in this study are real and come from official and credible institutions. Data on the implementation of economic and social rights were obtained from reports by the National Commission on Human Rights (Komnas HAM) and the Ministry of Law and Human Rights [1]. Demographic data were obtained from the Central Statistics Agency (BPS), specifically regarding the population by age group and gender in Bogor Regency in 2020 [2]. Information on the poverty line was collected from BPS reports for the 2008 period [3]. Meanwhile, data related to cryptocurrency investment activities and the list of venture capital approaches were obtained from the Coinstack Patterns Crypto VC Reach Out List for 2023, which is an international blockchain community-based publication and has been used in various digital asset investment analyses [4].

Methodologically, the K-Means algorithm was chosen for its ability to cluster data based on feature similarity without requiring initial labels. The K-Means algorithm is a non-hierarchical algorithm derived from data clustering methods. The K-Means algorithm begins with the initial formation of cluster partitions, which are then iteratively refined until no significant changes occur in the cluster partitions[5]. In the socio-economic context of Indonesia, previous studies have applied K-Means to map poverty[6] and analyze the spread of economic

digitalization[7]. However, few studies have integrated human rights data, demographics, poverty, and cryptocurrency investment simultaneously within a single analytical framework using clustering methods.
Thus, this study offers a new contribution to the utilization of multidomain data mining to identify complex socio-economic patterns in Indonesia. The integration of data from the social and digital sectors is expected to build a more comprehensive understanding of the dynamics of ongoing development.

Based on this background, the research questions in this study are as follows:

a. What are the patterns of relationships between the implementation of economic and social human rights, demographics, poverty, and crypto investment in Indonesia during the 2020–2024 period?
b. How can clustering results using the K-Means algorithm group regions or entities based on these indicators in a meaningful and informative way?

2. Research Method

This study uses an exploratory quantitative approach by applying the K-Means Clustering algorithm to group regions or entities based on multidimensional data from the social, economic, demographic, and digital investment sectors. The main objective of this method is to discover hidden patterns in the data that cannot be identified through ordinary descriptive analysis. This exploratory approach allows researchers to identify inherent structures in the data without any prior assumptions about the distribution or relationships between variables, making it particularly suitable for complex and multidimensional datasets. The quantitative research design ensures that the analysis is based on numerically measurable data that can be statistically tested, thereby enhancing the objectivity and replicability of the findings.

2.1. Types and Sources of Data

The data used in this study is quantitative secondary data collected from various official and credible sources, covering the period 2020–2024. The selection of actual data from leading institutions ensures the validity and reliability of the information used in the analysis. Details of the data and its sources are as follows:

a. Data on the Implementation of Economic and Social Rights: This data includes various indicators that reflect the fulfillment of basic rights in the economic and social fields, such as access to education, health, employment, and a decent standard of living. The data is obtained from reports by the National Commission on Human Rights (Komnas HAM) and the Ministry of Law and Human Rights[1]. Relevant indicators may include the rate of complaints regarding economic/social human rights violations, human rights compliance indices, or data related to government programs in fulfilling these rights.
b. Demographic Data: Demographic information focuses on population characteristics, which are essential for understanding the socio-economic context of the region. Data comes from the Central Statistics Agency (BPS)[2], specifically regarding the population by age group (e.g., productive, non-productive) and gender. The specific data used is for Bogor Regency in 2020, but it can be expanded to other regions if data is available and relevant for the clustering scope. Other demographic variables such as migration rates or population density can also be considered if data permits.
c. Poverty Data: This data is crucial for measuring economic well-being. Information on the poverty line and the percentage of the poor population is collected from BPS reports for the 2008 period [3]. This data will be used to identify regions with varying poverty levels, which will then be integrated with other indicators.
d. Crypto Investment Data: As an indicator of the rapidly developing digital economy, crypto investment data provides insights into the adoption of financial technology. Information related to crypto investment activities, such as the number of investors, transaction volume, or popular types of crypto assets, as well as a list of Venture Capital approaches, is taken from the Coinstack Patterns Crypto VC Reach Out List for 2023 [4]. This data will be filtered to focus on activities in Indonesia or relevant regions if granular data is available.
e. Sustainable Finance Data: As an important indicator in the transition to a green economy, sustainable finance data provides an overview of the financial services sector's commitment to sustainability principles. The information is obtained from official reports by the Financial Services Authority (OJK) covering policies, supervision, and the implementation of sustainable finance in Indonesia. Additionally, the Green Bond Market Summary from the Climate Bonds Initiative (CBI) is used to obtain data on the number of green bond issuances, market volume, and global trends in green investments. Internationally indexed scientific publications from the period 2019–2023 also serve as a source to support the classification of entities based on ESG (Environmental, Social, and Governance) indicators. This data will be selected and adjusted to focus on the financial sector in Indonesia or entities that have direct relevance to national sustainable financing policies.

2.2. Research Variables

Based on the above data sources, the variables to be used in this study include:

a. Economic and Social Human Rights Variables: Index of economic and social human rights fulfillment, number of cases of economic/social human rights violations, government budget for economic/social human rights programs.
b. Demographic Variables: Population by age group (0–14 years, 15–64 years, >65 years), sex ratio, population density.
c. Economic Variables (Poverty): Poverty line (Rp), percentage of poor population, Gini Ratio.
d. Crypto Investment Variables: Number of active crypto investors, monthly/quarterly crypto transaction volume, average investor portfolio value, number of crypto/blockchain startups in the region.
e. Sustainable Finance Variables: Number of green bond issuances, total value of green bonds per year, number of financial institutions implementing ESG policies, average ESG score per entity, and compliance rate with OJK sustainable finance regulations.

These variables will be normalized or standardized before the clustering process to ensure that each variable contributes equally to the calculation of distances between data points, avoiding the dominance of variables with larger value scales.

2.3. Research Stages

These research stages will be carried out systematically and structurally to ensure the accuracy and validity of the clustering results. The sequence of these stages reflects best practices in data analysis using clustering algorithms. These stages are described in detail as follows:

2.3.1. Data Collection

The initial stage involves collecting data from various identified sources. This process includes accessing official databases, public reports, and scientific publications. Data will be collected in a format that allows for efficient integration and pre-processing, such as spreadsheet formats (CSV, Excel). It is important to ensure consistency in the time periods of the data across all variables for accurate comparative analysis. If data is not available for the same period, approaches such as interpolation or extrapolation may be considered, but with an acknowledgment of their limitations.

2.3.2. Data Pre-processing

After the data is collected, the data pre-processing stage is a crucial step in preparing the data before applying the K-Means algorithm. This stage includes:

a. Data Cleaning: Identifying and handling missing values, outliers, and noisy data. Methods for handling missing values may include imputation (e.g., using the mean, median, or regression) or removing rows/columns with too many missing values. Outliers can be detected using statistical methods such as Z-score or IQR (Interquartile Range) and handled by trimming, winsorization, or data transformation.

b. Data Integration: Combining data from various sources into a single dataset. This involves matching keys (e.g., region code, year) to ensure that data from different variables refer to the same entity.

c. Data Transformation:

1. Normalization/Standardization: Apply normalization techniques (e.g., Min-Max Scaling) or standardization (e.g., Z-score Normalization) to change the scale of variables so that they have a uniform range of values. This is important because K-Means is very sensitive to variable scales:

$$x_{normalized} = \left(\frac{x_{old} - x_{min}}{x_{max} - x_{min}}\right) \quad (1)$$

$$x_{new} = \frac{x_{old} - \mu}{\sigma} \quad (2)$$

Description:

(a) $x_{normalized}$: This is the value of the data after normalization. This value will be within the desired target range (for example, between 0 and 1).
(b) x_{old}: This is the original value (old data) of the feature to be normalized.
(c) x_{min}: This is the minimum value of the feature observed in the entire dataset before normalization.
(d) x_{max}: This is the value that should be the new maximum value of the normalized data.
(e) x_{new}: This is the data value after standardization (often referred to as the Z-score).
(f) μ (mu): This is the mean of the feature observed in the entire dataset.
(g) σ (sigma): This is the standard deviation of the feature observed in the entire dataset.

2. Dimensionality Reduction: If there are too many variables, techniques such as Principal Component Analysis (PCA) can be used to reduce the dimensionality of the data while retaining most of the variance. This can improve computational efficiency and reduce noise in the data[8].

2.3.3. Determining the Optimal Number of Clusters (K)

One of the challenges in K-Means is determining the optimal number of clusters (K). This stage is very important for obtaining meaningful clustering results. Several methods that will be used include:

b. Elbow Method: This method involves calculating the Sum of Squared Errors (SSE) for various values of K. SSE is the sum of the squared distances between each data point and its cluster centroid. An SSE vs. K graph is then created, and the "elbow" (the point where the decrease in SSE begins to slow significantly) indicates the optimal value of K.\

c. Silhouette Score Method: This method measures how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where higher values indicate better clustering. The K value that yields the highest silhouette score is considered optimal[9].

2.3.4. Implementation of the K-Means Algorithm

After the optimal K is determined, the K-Means algorithm is implemented. The main steps of this algorithm are as follows:

a. Initialize Centroid: Select K initial data points randomly as the initial cluster centroids. Various initialization strategies can be used, such as K-Means++ for better initial centroid selection and to avoid local optima issues.

b. Cluster Assignment (Assignment Step): Each data point is assigned to the cluster whose centroid has the closest Euclidean distance. The Euclidean distance between two points: $p = p_1, p_2, \ldots, p_n$ and $q = q_1, q_2, \ldots, q_n$ in n-dimensional space is calculated as: $d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ,,, + (q_n - p_n)^2}$

c. Centroid Update (Update Step): The centroid of each cluster is recalculated as the average of all data points assigned to that cluster.

d. Iteration: Steps 2 and 3 are repeated iteratively until convergence is achieved, i.e., when cluster assignments no longer change significantly or centroid changes are below a certain threshold, or the maximum number of iterations is reached[10].

The flowchart for the K-Means implementation process can be illustrated as follows:

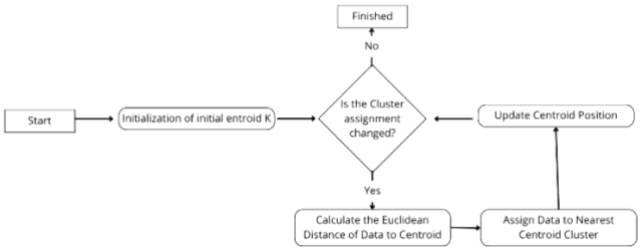

Figure 1: K-Means Implementation Process Flowchart

2.3.5. Analysis and Interpretation of Clustering Results

Once clustering is complete, this stage involves an in-depth analysis of the characteristics of each cluster formed. This includes:

a. Cluster Profiling: Analyzing the average values or distribution of variables within each cluster to understand the unique characteristics of each group[11]. For example, Cluster 1 may be characterized by low poverty levels, high crypto adoption, and good human rights fulfillment, while Cluster 2 may have opposite characteristics.
b. Cluster Visualization: Using data visualization techniques such as scatter plots (e.g., using PCA or t-SNE for dimension reduction to 2D or 3D) to graphically display the clusters[12]. This visualization helps in understanding the separation between clusters and the density of data within each cluster.
c. External Validation (Optional): If external label data is available (e.g., regional classification based on development categories), the clustering results can be validated by comparing them with the external labels[13].

2.3.6. Algorithm Performance Evaluation

The performance of the K-Means algorithm will be evaluated using internal cluster validation metrics. This metric helps measure the quality of clustering without requiring ground truth labels[14]. Some relevant metrics are:

a. Within-Cluster Sum of Squares (WCSS): Similar to SSE used in the elbow method, it measures cluster compactness. A lower WCSS value indicates a more compact cluster[15].
b. Silhouette Score: As previously explained, it measures how well each data point fits into its own cluster and how poorly it fits into neighboring clusters[16].
c. Davies-Bouldin Index: Measures the ratio between dispersion within clusters and distance between clusters. A lower value indicates better clustering[16].

2.3.7. Formulation of Conclusions and Recommendations

The final stage is to formulate conclusions based on the findings of the clustering analysis. The conclusions will address the research question, namely how the patterns of relationships between the variables are and how clustering can group entities in a meaningful way. Based on the conclusions, policy recommendations will be developed for the government, investors, and academics, taking into account the specific characteristics of each identified cluster. These recommendations will be practical and actionable, providing guidance for targeted interventions.

With this structured research process, it is hoped that this study will yield robust findings and make a significant contribution to understanding the socio-economic dynamics in Indonesia, particularly in the context of human rights, demography, poverty, and cryptocurrency investment. All stages will be documented in detail, including the use of statistical software or programming languages (e.g., Python with scikit-learn, pandas, and matplotlib libraries) for data analysis and visualization.

3. Results and Discussion

Results of K-Means clustering applied to four different data domains: Economic and social human rights, crypto venture capital investment, population demographics, and poverty line. Each clustering result will be accompanied by a 2D visualization using Principal Component Analysis (PCA) and a summary of cluster characteristics saved in Excel format.

3.1. Human Rights Data Clustering

The application of the K-Means algorithm to human rights data (including population, unemployment, and illiteracy) identifies groups that show different patterns related to economic and social human rights conditions. With K=3 clusters, each cluster represents unique characteristics, which can be interpreted from their original features. Visualization of the results of human rights data clustering in 2-dimensional space after dimension reduction using PCA.

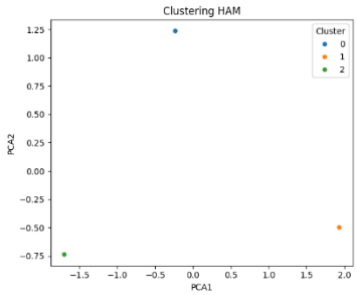

Figure 2: Visualization of Human Rights Clustering Results

Tabel 1: Human Rights Clustering Results

Tahun	Penduduk_Juta	Pengangguran (%)	Buta_Huruf (%)	Cluster
1980	147,49	6,1	27,9	2
1990	179,379	7,4	15,83	0
2000	203,456	4,77	10,08	1

The table and visualization show the results of K-Means clustering on Human Rights (HR) data from 1980, 1990, and 2000. Since the amount of data and clusters are the same (K=3), each year forms its own cluster, showing significant differences in HR characteristics between decades.

a. Cluster 2 (Year 1980 - Green in the plot): Characterized by the lowest population and highest illiteracy rate (27.9%). Its position in the plot (negative PCA1 and PCA2) is highly isolated, reflecting the most challenging initial conditions.
b. Cluster 0 (1990 – Blue in the plot): Shows an increase in population and the highest unemployment rate (7.4%), but a significant decrease in illiteracy (15.83%) compared to 1980. Its position in the plot (PCA1 close to zero, PCA2 positive) reflects this transition.
c. Cluster 1 (Year 2000 - Orange in the plot): Represents the best human rights conditions with the highest population, lowest unemployment (4.77%), and lowest illiteracy (10.08%). Its position in the plot (PCA1 positive, PCA2 negative) indicates significant progress from previous years.

This clustering successfully highlights the evolution of economic and social human rights conditions in Indonesia every decade, with a clear trend of improvement from 1980 to 2000, particularly in literacy and unemployment rates.

3.2. Crypto Venture Capital Data Clustering

Clustering Crypto Venture Capital (VC) data aims to group companies or entities based on their crypto investment characteristics. With K=3 clusters, this grouping can help identify different types of investors or firms in the blockchain and cryptocurrency ecosystem.

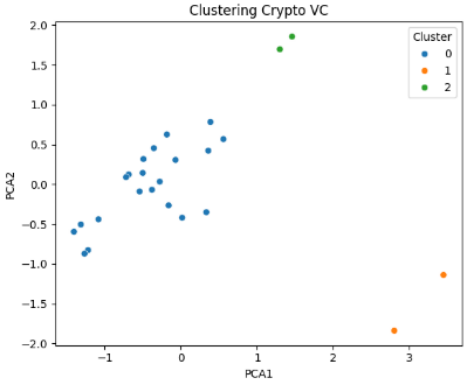

Figure 3: Visualization of Crypto Clustering Results

Tabel 2: Crypto Clustering Results

Rank	Firm	Fund Size (USD)	Crypto Investments	Location	Cluster
1	A16Z Crypto	7,565,000,000	141	SF	1
2	Binance Labs	7,500,000,000	200	Hong Kong	1
3	Multicoin	2,800,000,000	98	Austin	0
4	Paradigm	2,500,000,000	81	SF	0
5	Galaxy Interactive	2,100,000,000	80	NYC	0
6	Polychain	1,965,000,000	148	SF	0
7	Blockchain Capital	1,900,000,000	133	SF	0
8	Arrington Capital	1,600,000,000	88	SF	0
9	Blockchange	1,600,000,000	79	NYC	0
10	Dragonfly	1,522,000,000	109	SF	0
11	Coinbase Ventures	1,500,000,000	244	SF	2
12	Pantera	1,500,000,000	150	SF	0
13	HiveMind Capital	1,500,000,000	9	NYC	0
14	Haun Ventures	1,500,000,000	5	SF	0
15	Animoca Brands	1,500,000,000	230	Hong Kong	2
16	Electric Capital	1,430,000,000	71	SF	0
17	ParaFi Capital	1,200,000,000	83	NYC	0
18	1Confirmation	1,200,000,000	32	SF	0
19	ConsenSys Mesh	1,000,000,000	118	DC	0
20	Borderless Capital	1,000,000,000	74	Atlanta	0
21	Fundamental Labs	1,000,000,000	103	Shanghai	0
22	GSR	1,000,000,000	91	London	0
23	BH Digital	1,000,000,000	11	London	0
24	Digital Finance Group	1,000,000,000	71	Singapore	0
25	JRR Group	1,000,000,000	19	Switzerland	0

The table and visualization show the results of clustering data on the top Crypto Venture Capital (VC) firms. This clustering successfully grouped VCs into three different investment strategy archetypes based on their fund size and number of investment portfolios.

a. Cluster 1 (The Titans - Orange in the plot): Characterized by the highest absolute Fund Size (approximately $7.5 billion). Its position on the plot (extremely positive PCA1) is significantly separated to the right, reflecting an investment strategy dominated by massive capital strength to fund the largest deals.
b. Cluster 2 (Prolific Investors - Green on the plot): Characterized by the highest number of investments (over 230 portfolios), despite a more moderate fund size. Its position on the plot (PCA1 near zero, PCA2 extremely positive) is far above, reflecting a “spray and pray” strategy focused on quantity and broad market coverage.
c. Cluster 0 (Core Group - Blue on the plot): Represents the majority of top-tier VC firms. This cluster demonstrates a balance between large fund size and significant investment volume, without extreme values in either metric. Its clustered position in the middle of the plot (PCA1 and PCA2 are not extreme) indicates this is the standard profile for successful crypto VCs.

This clustering successfully maps the competitive landscape of crypto VCs, highlighting that there is no single way to become a top player. This analysis clearly distinguishes between strategies focused on capital strength (Cluster 1), strategies focused on quantity and diversification (Cluster 2), and a balanced core approach (Cluster 0).

3.3. Clustering of Population Demographic Data

K-Means clustering analysis was performed on the demographic data of the population of Bogor City based on age groups, using male and female population data from 2020 and 2021. The results effectively grouped the population structure into three distinct segments that reflect the stages of the life cycle.

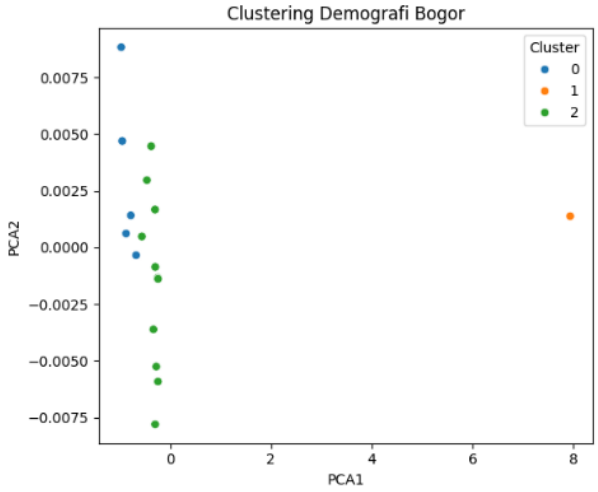

Figure 4: Population Demographics Visualization

Tabel 3: Population Demographics Results

Kelompok_Umur	Laki_Laki_2020	Laki_Laki_2021	Perempuan_2020	Perempuan_2021	Cluster
0 - 4	233323	234861	222793	224332	2
05-Sep	235903	235163	224914	224477	2
Oct-14	238337	236182	222707	221247	2
15 - 19	243751	243617	228881	228667	2
20 - 24	254659	255457	238875	238666	2
25 - 39	251583	253340	238155	239452	2
30 - 34	253269	255303	239628	241060	2
35 - 39	225487	227847	210955	212906	2
40 - 44	208288	210893	200420	202758	2
45 - 49	182798	186882	173311	177567	2
50 - 54	152814	156871	142632	147055	2
55 - 59	117528	121554	108267	112748	0
60 - 64	84244	87909	78102	82280	0
65 - 69	55454	58501	50936	54058	0
70 - 74	29949	32563	30141	32779	0
75+	22351	23981	26613	28560	0
Total	2789738	2820924	2637330	2668612	1

Table 3 and Figure 4 above present the results of cluster analysis of demographic data for the population of Bogor City, grouped by age segment. The purpose of this analysis is to simplify and map the city's population structure by grouping age ranges that show similar demographic characteristics, particularly in terms of population size.

a. Cluster 2 (Child and Productive Age - Green in the plot): Includes the age group from 0 to 54 years. This cluster is characterized by the highest population size in each age group. Its clustered position on the left side of the plot (negative PCA1 value) indicates that this group forms the main population base and serves as the demographic backbone of the city.
b. Cluster 0 (Elderly - Blue on the plot): Covers the age group of 55 years and above. The main characteristic of this cluster is a significantly lower population size compared to the productive age cluster. Its position on the plot, separated from Cluster 2, although still on the left side, reflects the characteristics of the non-productive age group forming the peak of the population pyramid.
c. Cluster 1 (Outlier/Anomaly - Orange on the plot): Represented by a single unidentified data point in the table. Its position far to the right of the plot (extreme positive PCA1 value) makes it a clear outlier. This indicates a group with highly unique demographic characteristics that differ drastically from the general population structure, or may represent aggregate (total) data that was mistakenly included in the analysis.

This clustering successfully highlights the pyramid structure of Bogor City's population, clearly distinguishing between the large productive age population base (Cluster 2) and the smaller elderly population peak (Cluster 0). The presence of an anomaly cluster (Cluster 1) indicates the existence of a segment with very different characteristics that requires further investigation.

3.4. Economic Data Clustering

K-Means clustering analysis was performed on provincial poverty line data across Indonesia, using poverty line values for urban, rural, and combined areas. The results effectively grouped provinces into three distinct economic tiers that reflect disparities in the cost of living and regional welfare levels.

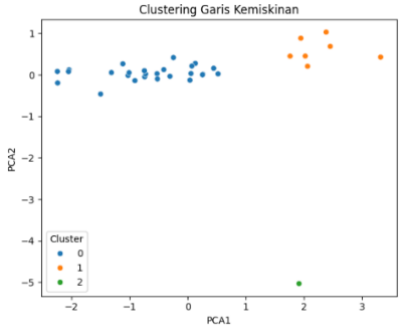

Figure 5: Economic Visualization

Tabel 4: Economic Results

Provinsi	Perkotaan	Perdesaan	Gabungan	Cluster
NAD	266168	229237	239873	1
Sumatera Utara	218333	171922	195331	0
Sumatera Barat	226343	179755	193733	0
Riau	247923	210519	229371	1
Jambi	223527	162434	182229	0
Sumatera Selatan	229552	175556	196452	0
Bengkulu	224081	170878	189607	0
Lampung	203685	160734	172332	0
Bangka Belitung	250240	242441	246169	1
Kepulauan Riau	289541	231580	262232	1
DKI Jakarta	290268	0	290268	2
Jawa Barat	190824	155367	176216	0
Jawa Tengah	184704	152531	168168	0
DI Yogyakarta	208655	169934	194830	0
Jawa Timur	183408	155432	169112	0
Banten	197328	156494	181076	0
Bali	190026	158206	176569	0
NTB	193241	148998	167536	0
NTT	199006	126746	139731	0
Kalimantan Barat	179261	150968	158834	0
Kalimantan Tengah	196354	180671	186003	0
Kalimantan Selatan	199416	166676	180263	0
Kalimantan Timur	257862	205255	237979	1
Sulawesi Utara	175628	162433	168025	0
Sulawesi Tengah	196229	160527	168160	0
Sulawesi Selatan	160220	127938	138334	0
Sulawesi Tenggara	151471	139065	141919	0
Gorontalo	154982	143584	147154	0
Sulawesi Barat	156041	141701	146492	0
Maluku	213969	180087	188931	0
Maluku Utara	213505	176757	187671	0
Papua Barat	244807	230254	233570	1
Papua	264635	213548	225195	1
INDONESIA	204896	161831	182636	0

The table and visualization above present the results of cluster analysis on Poverty Line data in 34 provinces in Indonesia. This analysis aims to group provinces with similar economic characteristics, based on Poverty Line levels in Urban, Rural, and Combined areas, to map the structure of welfare and cost of living across the archipelago.

a. Cluster 0 (Mainstream Provinces - Blue in the plot): This is the majority cluster encompassing most provinces in Indonesia, such as the majority of provinces in Java, Sulawesi, and parts of Sumatra. This cluster is characterized by moderate Poverty Line levels that tend to approach the national average. Its clustered position in the middle-left of the plot reflects the “standard” or mainstream economic profile in Indonesia.
b. Cluster 1 (High Cost of Living Provinces - Orange in the plot): Consists of provinces that are generally rich in natural resources or are new economic centers, such as Riau Islands, East Kalimantan, and Papua. A characteristic of this cluster is a poverty line that is significantly higher than the national average. Its isolated position on the right side of the plot (positive PCA1 value) clearly indicates the high cost of living required in these regions.
c. Cluster 2 (Unique Metropolitan Areas - Green on the plot): This cluster contains only one member, namely DKI Jakarta. This province is an outlier or extreme anomaly for two reasons: (1) It has the highest Poverty Line value in Indonesia, and (2) Its structure is entirely urban (it has no rural components). Its highly isolated position at the bottom of the plot (very negative PCA2 value) reinforces its unique status as the only metropolitan center whose structure cannot be compared to other provinces.

This clustering successfully maps the economic levels of provinces in Indonesia based on minimum living costs. This analysis effectively distinguishes three groups: the majority of provinces with average living costs (Cluster 0), a group of resource-rich provinces with high living costs (Cluster 1), and the special case of DKI Jakarta as a metropolitan center (Cluster 2). These results provide a clear picture of regional economic disparities in Indonesia.

3.5. Clustering of Sustainability Financing Data

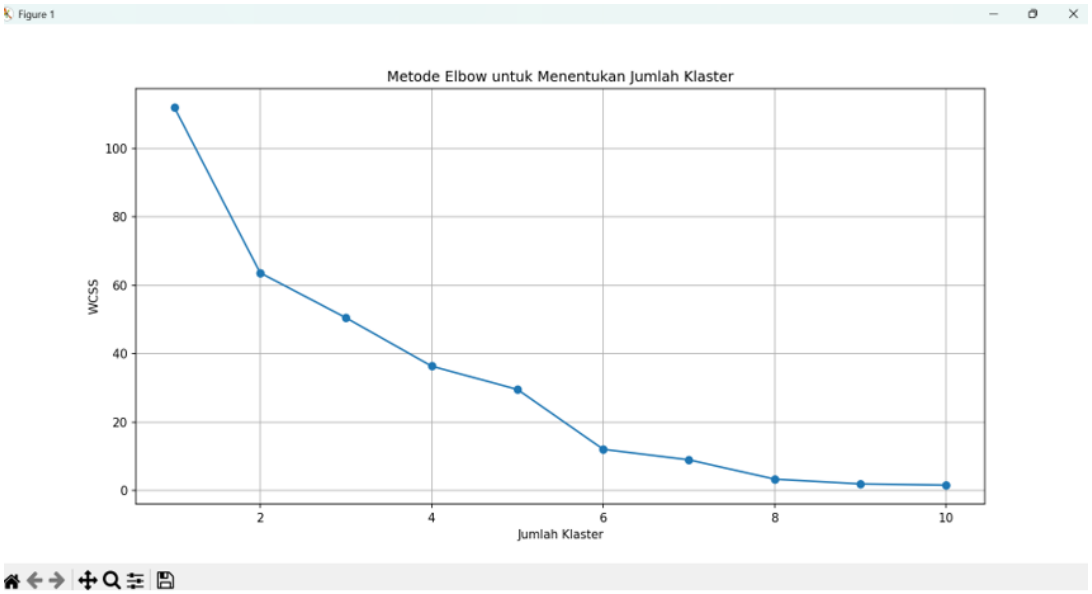

Figure 6: Elbow method results

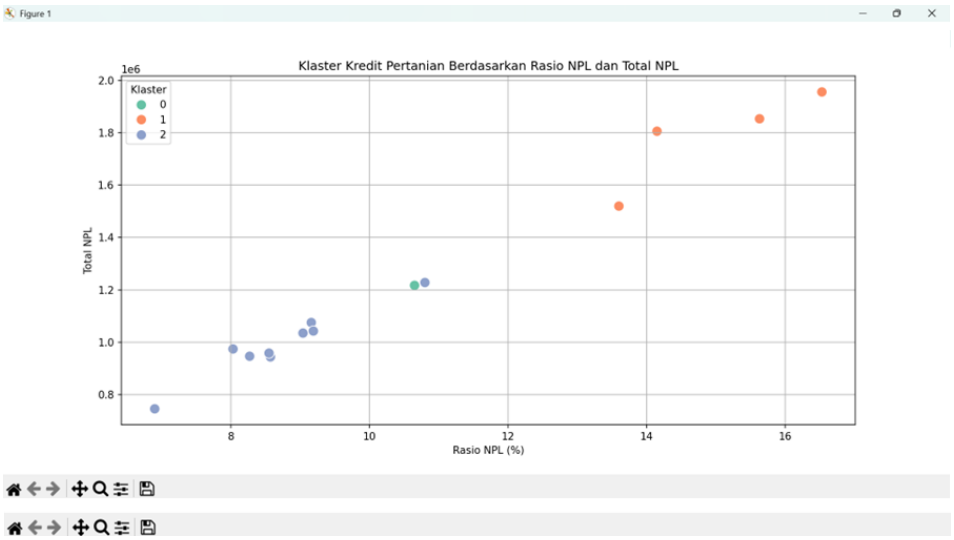

Figure 7: K-Means algorithm clustering results

To determine the optimal number of clusters in the clustering process, the Elbow Method was used by observing the WCSS (Within-Cluster Sum of Squares) value against the number of clusters. The graph shows that there is a sharp decrease in WCSS up to the third cluster, then it levels off. This indicates that the optimal number of clusters is 3, because after that point, the decrease in WCSS is no longer significant. Selecting the number of clusters at this "elbow" point is important to avoid overfitting while maintaining the accuracy of data grouping. Clustering was performed based on two variables, namely the Non-Performing Loan (NPL) Ratio and Total NPL for the agricultural sector. The results produced three clusters that separated regions/provinces based on their credit risk levels. Cluster 0 represents groups with low NPL ratios and small total NPLs, indicating regions with relatively safe credit risk. Cluster 1 consists of regions with high NPL ratios and total NPLs, indicating high credit risk. Cluster 2 lies between the two, indicating regions with moderate risk. These results are useful for credit risk analysis and decision-making in agricultural sector financial policies.

4. Conclusion

This study empirically validates the effectiveness of the K-Means algorithm as a robust analytical framework for investigating multidimensional socio-economic phenomena in Indonesia. The application of this unsupervised learning method to four different data domains—human rights, crypto venture capital investment, demographics, and economics—has successfully extracted significant inherent patterns. Specifically, this analysis has successfully (1) identified the evolutionary trajectory of economic and social human rights conditions across decades; (2) mapped the taxonomy of investment strategies in the crypto asset ecosystem, divided into capital archetypes, quantity, and balance; (3) segment demographic population structures into different life cycle cohorts; and (4) classify provinces into

economic strata reflecting regional living cost disparities. The algorithm's ability to generate coherent and interpretable clusters across each domain underscores its utility in reducing the complexity of large-scale data.

The aggregation of these findings comprehensively addresses the research problem formulation. Although not designed for causal inference, the clustering results implicitly indicate the interconnectivity between social conditions, financial technology adoption, and economic-demographic structures. Thus, this study fulfills its primary objective of providing an empirical mapping that illustrates the heterogeneity of conditions in Indonesia. The main theoretical and methodological contributions of this study lie in its approach, which integrates domains that are traditionally analyzed separately. This offers a conceptual model for a more holistic and data-driven analysis of multidimensional development, transcending conventional disciplinary boundaries.

However, the interpretation of the research results needs to consider several methodological limitations. The main limitation is the cross-sectional study design with temporal heterogeneity in the data sources, which prevents longitudinal analysis and direct cross-domain causal inferences. Additionally, the selection of the number of clusters (K) and the sensitivity of the K-Means algorithm to centroid initialization are factors that can influence the granularity of the clustering results. Therefore, future research agendas can be directed toward three main aspects: first, harmonizing data within a uniform time frame for more valid comparative analysis; second, applying longitudinal analysis to track dynamics and transitions between clusters over time; and third, exploring inferential methods to test hypotheses of causal relationships between variables from different domains. These steps will deepen our understanding of the complex and sustainable dynamics of Indonesia's development.

Acknowledgement

This is a text of acknowledgements. Do not forget people who have assisted you on your work. Do not exaggerate with thanks. If your work has been paid by a Grant, mention the Grant name and number here.

References

[1] Prof. Miriam Budiardjo, Fundamentals of Political Science, 4th ed., vol. 500. Jakarta: Gramedia Pustaka Utama, 2008.
[2] Atmini Dharuri, Mathematics 1, 1st ed., vol. 300. Bogor: Quadra, 2023.
[3] Rudy Badrudin, Regional Autonomy Economics, 1st ed., vol. 207. Yogyakarta: UPP STIM YKPN, 2012.
[4] CRYPTO ACADEMY, MASTERING ALTCOINS, vol. 172. PT. Academy, 2024.
[5] A. Sulistiyawati and E. Supriyanto, "Implementation of the K-means Clustering Algorithm in Determining Outstanding Students," vol. 15, no. 2.
[6] W. Setya and A. Nugraha, "Clustering Poverty Mapping in West Java Province Using the K-Means Algorithm," Scientific Journal of Wahana Pendidikan, January, vol. 2023, no. 2, pp. 234–244, doi: 10.5281/zenodo.7567622.
[7] S. Sugiarti, F. Rahmiyatun, R. Oktayani, R. T. Aliudin, and E. N. Aina, "Analysis of the Influence of Working Capital Turnover on Profitability at PT Gudang Garam Tbk," Equity: Journal of Economics, vol. 10, no. 1, pp. 13–23, Jun 2022, doi: 10.33019/equity.v10i1.84.
[8] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," Pamukkale University Journal of Engineering Sciences, vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/pajes.2021.62687.
[9] P. Vania and B. Nurina Sari, "Comparison of the Elbow and Silhouette Methods for Determining the Optimal Number of Clusters in Rice Production Clustering Using the K-Means Algorithm," Jurnal Ilmiah Wahana Pendidikan, vol. 9, no. 21, pp. 547–558, 2023, doi: 10.5281/zenodo.10081332.
[10] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive Review of K-Means Clustering Algorithms," International Journal of Advances in Scientific Research and Engineering, vol. 07, no. 08, pp. 64–69, 2021, doi: 10.31695/ijasre.2021.34050.
[11] D. R. Yuniartha, N. A. Masruroh, and M. K. Herliansyah, "An evaluation of a simple model for predicting surgery duration using a set of surgical procedure parameters," Inform Med Unlocked, vol. 25, Jan 2021, doi: 10.1016/j.imu.2021.100633.
[12] J. J. Sylvia, "A genealogical analysis of information and technics," Information (Switzerland), vol. 12, no. 3, Mar 2021, doi: 10.3390/info12030123.
[13] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," January 30, 2025, Elsevier Ltd. doi: 10.1016/j.heliyon.2025.e41953.
[14] Y. Sopyan, A. D. Lesmana, and C. Juliane, "Analysis of the K-Means Algorithm and Davies Bouldin Index in Finding the Best Cluster for Divorce Cases in Kuningan Regency," Building of Informatics, Technology and Science (BITS), vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2697.
[15] S. Quiz, "Training the Model Full Code Elbow Method Drawbacks Frequently Asked Questions Quiz Time Are you up for a challenge in the realm of the Elbow Method for Finding the Optimal Number of Clusters in K-Means? Challenge yourself here! What Is the Elbow Method in K-Means Clustering?"
[16] Y. Hasan, "Silhouette Score and Davies-Bouldin Index Measurement on K-Means and Dbscan Cluster Results," 2024.