MULTIDISCIPLINARY SCIENCE JOURNAL

RESEARCH ARTICLE
Published Online: May 27, 2025
https://doi.org/10.31893/multiscience.2025557

Evaluating K-Means and K-Medoids clustering for household poverty analysis using random forests

Pita Jarupunphol[a] | **Suthasinee Kuptabut**[b] | **Wichidtra Sudjarid**[c]

[a]Digital Technology Program, Phuket Rajabhat University, Thailand.
[b]Computer Program, Sakon Nakhon Rajabhat University, Thailand.
[c]Environmental Science Program, Sakon Nakhon Rajabhat University, Thailand.

Abstract This study evaluates the effectiveness of k-means and k-medoids clustering techniques, combined with random forest classification, for household poverty analysis in the Kut Bak District of Thailand. Using data from 301 households encompassing 27 socioeconomic features, a correlation heatmap analysis was performed using Pearson's correlation coefficient to assess the relationships between key socioeconomic factors. The optimal number of clusters was determined using the elbow method, with k-means and k-medoids converging on three clusters representing three distinct economic clusters: low, medium, and high. Clustering performance was assessed using the Silhouette Score (k-means: 0.418, k-medoids: 0.430), Calinski-Harabasz Index (k-means: 46.184, k-medoids: 46.108), and Davies-Bouldin Index (k-means: 2.084, k-medoids: 2.056), demonstrating k-medoids' slight advantage in handling outliers and noise. Random forest classification validated the clustering results with an accuracy of 96% for k-means and 93% for k-medoids. K-means outperformed in recall (93% vs. 86%), while k-medoids excelled in precision (97% vs. 96%), highlighting a trade-off between broader coverage and classification specificity. The model's accuracy was compared to prior studies, demonstrating the robustness of the proposed approach. Key features influencing clustering included total household expenditure (THEX), cost-reduction projects (CRFUP, CRIUP), and total income from project participation (TIPP), emphasizing the interplay of income and agricultural initiatives in economic stratification. Beyond technical results, the study underscores the social implications of poverty classification. The strong correlations among project participation variables suggest that community-driven agricultural programs foster economic resilience. These insights inform tailored interventions: financial assistance for the low-income group, skill development for the medium-income group, and investment opportunities for the high-income group. By integrating machine learning techniques with socioeconomic analysis, this research provides a robust framework for poverty alleviation strategies that are data-driven, equitable, and impactful.

Keywords: household poverty, k-means, k-medoids, random forest, socioeconomic

1. Introduction

Classifying poverty is essential for understanding and mitigating socioeconomic disparities (Huang & Xia, 2023). This classification is pivotal for resource distribution, policymaking, and crafting targeted poverty alleviation strategies. Poverty is complex, involving more than financial hardships, and includes broader socioeconomic elements (Morris et al., 2018). Poverty affects billions globally, often defined by the World Bank (2023a) regarding income. However, such a monetary perspective fails to encompass multi-dimensional poverty, including a lack of access to education, healthcare, and essential services (D'Attoma & Matteucci, 2023). The regional context of poverty varies; it is typically tied to rural underdevelopment in poorer nations, while in wealthier countries, it may relate more to urban inequalities and systemic issues (World Bank, 2023b). Classifying household poverty is vital for effectively tailoring policy and support. It helps pinpoint households urgently needed and determines the most suitable support forms (Vaidyanathan, 2001; Zhang & Dai, 2023). Moreover, multi-dimensional poverty classification provides a fuller understanding of deprivation, considering factors beyond financial constraints like access to education, healthcare, and employment.

In Thailand's Sakon Nakhon Province, the Kut Bak District exemplifies the challenges in poverty classification. It is one of Thailand's poorest areas, predominantly rural, agriculturally reliant, and economically undiversified. The limited market access, dependence on agriculture, and scarce alternative income opportunities in Kut Bak heighten poverty vulnerability, necessitating a nuanced understanding of poverty levels and dimensions. Addressing these challenges requires innovative solutions, such as enhanced data collection and analysis through technology, adaptive models that reflect poverty's dynamic nature, and diverse stakeholder participation to reduce bias. This research provides insights for tailoring interventions that address multidimensional poverty, combining advanced data science techniques with a nuanced understanding of socioeconomic contexts. By integrating clustering methods like k-medoids and machine learning algorithms, this study sets a

 Received: December 4, 2024 | Accepted: March 31, 2025

benchmark for addressing complex poverty challenges in data-limited environments. The methodological framework presented here is not only applicable to poverty analysis but also holds promise for broader socioeconomic research areas such as public health and resource allocation.

1.1. Household Poverty Classification

The classification of household poverty has diversified to include various methods that reflect its complex nature. Historically, poverty was identified through income or consumption metrics, with the World Bank's international poverty line serving as a familiar standard (Mansi et al., 2020). Another approach is the multidimensional poverty index (MPI), created by the Oxford poverty and human development initiative (OPHI), which evaluates poverty across several dimensions such as education, health, and living standards (UNDP, 2023). Furthermore, composite indices classify household poverty by integrating multiple indicators into a single measure, offering a holistic view of poverty (Drago, 2021). For instance, the human poverty index amalgamates different deprivation aspects, while the social vulnerability index considers factors that heighten a household's risk of social and economic difficulties (Chakravarty & Majumder, 2005; Llera-Sastresa et al., 2017; Mah et al., 2023). Asset-based methods gauge poverty by assessing ownership of assets like land, housing, and livestock, which is especially pertinent in rural settings where income data may be less accurate (Carney et al., 2001). The basic needs approach also evaluates if households meet essential needs such as food, shelter, and clothing, typically setting a baseline for these necessities (Allen, 2017).

1.2. Machine Learning for Poverty Analysis

Machine learning, with its ability to analyze vast datasets and uncover intricate patterns, presents a promising avenue for advancing poverty analysis (Mullainathan & Spiess, 2017). This section delves into applying ML techniques in classifying poverty and identifying the multifaceted factors that contribute to it. For instance, Aiken et al. (2022) demonstrated how machine learning algorithms trained on traditional survey data can discern patterns in mobile phone usage that correlate with poverty levels. This approach facilitates more efficient resource allocation to those in dire need, particularly in crisis scenarios where conventional data sources may be insufficient or outdated. Moreover, machine learning techniques are instrumental in unraveling the complex interplay of factors contributing to poverty. Longa et al. (2021) employed machine learning to predict energy poverty in the Netherlands, integrating socioeconomic parameters such as house value, ownership, age, household size, and population density. Their findings underscore that while income remains a critical predictor, additional variables are essential for achieving high prediction accuracy, highlighting the intricate mechanisms underlying energy poverty (Longa et al., 2021). Similarly, Kolak et al. (2020) utilized machine learning to quantify neighborhood-level social determinants of health (SDOH) across the continental United States, developing indices that reflect various dimensions of advantage, isolation, opportunity, and social cohesion. Their research emphasizes the importance of considering multiple dimensions of SDOH rather than relying solely on single deprivation indices for a more nuanced understanding of poverty's complexities (Kolak et al., 2020).

Several studies have also explored using machine learning to examine the relationship between various social determinants and health outcomes (Dunn, 2017; Shin et al., 2018). For example, Shin et al. (2018) applied a machine learning approach to identify pediatric asthma patients at risk of hospital revisits, incorporating biomarkers and sociomarkers (measurable indicators of the patient's social environment). Their findings reveal the substantial impact of social factors on health outcomes, even when accounting for symptom-related features (Shin et al., 2018). Dunn et al. (2017) leveraged Twitter data to map information exposure related to the human papillomavirus (HPV) vaccine, correlating this exposure with state-level vaccine coverage and socioeconomic factors such as poverty and education. Machine learning has shown promise in addressing specific poverty-related issues as well. Delafiori et al. (2021) developed a machine learning-based platform for rapid COVID-19 diagnosis using metabolomics, underscoring the importance of efficient diagnostic methods for resource allocation and risk management, especially in regions where poverty is prevalent (Delafiori et al., 2021). The work of Longa et al. (2021) offers a framework for categorizing energy risk based on income and expenditure, using machine learning to predict energy poverty from socioeconomic parameters. In education, machine learning has been applied to predict dropout risks among vulnerable populations. Sani et al. (2020) used machine learning to forecast dropout risks among B40 students (the bottom 40% income group) in Malaysia, aiming to enhance educational outcomes for these at-risk groups.

1.3. Poverty Clustering

Clustering techniques are pivotal in analyzing multidimensional data, including poverty-related datasets, to effectively identify patterns and categorize data points. Commonly employed clustering methods include k-means, k-medoids, hierarchical clustering, density-based spatial clustering of applications with noise (DBSCAN), and Gaussian mixture models. Each of these methods is tailored to specific data structures and analytical goals, offering unique advantages depending on the nature and distribution of the data (Cady, 2017; Knox, 2018; Witten et al., 2016). For instance, k-means and k-medoids are well-suited for relatively spherical or convex data distributions and require the number of clusters to be defined a priori. Several studies have

conducted comparative analyses of the k-means and k-medoids clustering algorithms across various datasets and evaluation metrics, yielding mixed results regarding their relative performance. While k-medoids is generally recognized for its superior accuracy, robustness to outliers, and ability to handle dissimilarities effectively, the outcomes are not universally consistent and often depend on the specific dataset and problem context. For instance, Soni and Patel (2017) evaluated the two algorithms using the Iris dataset, reporting that k-medoids achieved an accuracy of 92%, surpassing k-means, which scored 88.7%. Similarly, Nurhayati et al. (2018) assessed the algorithms in the context of big data applications, where k-medoids demonstrated an average accuracy of 63.24%, compared to k-means' 52.11%. These findings underscore the advantages of k-medoids in certain scenarios, particularly when dealing with noisy or non-Euclidean data.

However, exceptions to this trend have been documented. Suarna et al. (2021) investigated clustering techniques for categorizing fish cooking menus and found that k-means outperformed k-medoids, as evidenced by a lower Davies–Bouldin Index (DBI) value, which indicates better clustering quality. This observation aligns with the findings of Albayati and Altamimi (2019), who applied both algorithms to identify fake Facebook profiles. In their study, k-means exhibited higher accuracy than k-medoids, highlighting the variability in algorithm performance based on the dataset and application domain. These contrasting results emphasize that the effectiveness of k-means and k-medoids depends on the dataset's characteristics and the specific problem being addressed. As Hennig & Liao (2013) notes, there is no universally superior clustering algorithm. Instead, the choice between k-means, k-medoids, or other clustering methods should be guided by careful consideration of the dataset properties, the nature of the task, and the evaluation metrics relevant to the application at hand. This case-specific approach ensures that the selected algorithm aligns optimally with the requirements of the problem, thereby maximizing clustering performance. In this research, k-means and k-medoids can offer advantages in poverty analysis by identifying natural socioeconomic groupings without requiring predefined labels. Unlike traditional econometric models, clustering methods can capture nonlinear relationships among poverty indicators, including expenditure, income sources, and project participation (Huang & Xia, 2023). Recent studies (Zhang & Dai, 2023) have demonstrated that unsupervised learning enhances the precision of poverty classification by detecting hidden patterns in large-scale economic data. These methods are computationally efficient and widely applicable but may need help with irregular cluster shapes or significant outliers.

Hierarchical clustering excels in scenarios where data exhibits a natural hierarchical structure, such as socioeconomic indicators that region, demographic profiles, or income brackets may group. Unlike partition-based methods, hierarchical clustering does not require a predefined number of clusters, making it particularly useful for exploratory analysis. Similarly, DBSCAN is advantageous for identifying clusters of arbitrary shapes and sizes, making it a robust choice for datasets with noise and outliers, such as those frequently encountered in poverty analysis. Its ability to uncover clusters based on density thresholds adds flexibility in analyzing spatial or temporal poverty data. Gaussian mixture models, which utilize a probabilistic framework, provide a soft clustering alternative by assigning membership probabilities to data points. This approach is instrumental in scenarios where data points may belong to multiple clusters, such as households exhibiting moderate and extreme poverty characteristics. In this study, the choice of k-medoids clustering in a study might be driven by the need for a straightforward, scalable solution that provides clear partitioning of the data into distinct groups, which is often desirable in practical applications like socioeconomic analysis where actionable insights are needed (Arora et al., 2016).

1.4. Poverty Classifications

Poverty classification has significantly evolved with the integration of machine learning techniques, marking a significant shift from traditional statistical analysis methods. These computational approaches add a new layer to poverty analysis by utilizing algorithms that can predict and make decisions from data (Alsharkawi et al., 2021). This approach is instrumental in poverty classification, where the complex interplay of socioeconomic factors presents significant challenges. For example, Huang & Xia (2023) focus on classifying rural relative poverty in China through a questionnaire distributed across 23 impoverished counties using several methods, such as decision tree and logistic model, to assess the impact of various land elements. Several research works also employ a range of machine learning methods to directly classify levels of poverty (Huang et al., 2023; Sihombing & Arsani, 2021), while Muñetón-Santa et al. (2022) propose a novel approach that uses natural language processing (NLP) to analyze the discourse of people impacted by poverty, aiming to pinpoint their specific deprivation levels. Furthermore, Alsharkawi et al. (2021) explore the multidimensional poverty issues in Jordan by deploying various machine-learning models to examine and track the poverty status of Jordanian households.

There are various machine-learning classification techniques, such as logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN), and neural networks (Knox, 2018; Witten et al., 2016). Each method has strengths and applications based on the data's nature and the analysis's requirements. A random forest is also a machine learning technique that constructs many decision trees during training and outputs the class, the mode of the classes (classification), or mean prediction (regression) of the individual trees. It is an ensemble learning algorithm that constructs multiple decision trees during training, combining their outputs through majority voting for classification or averaging for regression to enhance predictive accuracy and mitigate overfitting. Furthermore, it corrects for decision trees' habit of overfitting to their training set, providing a more generalizable model. As such, its versatility and robustness make it effective across diverse fields, such as remote sensing, species distribution modeling, genomics, proteomics, stock market prediction, and biomarker identification

(Hanberry, 2024; Cao et al., 2024; Park et al., 2022). With its ability to handle complex, high-dimensional datasets, Random Forest is particularly well-suited for predicting human capital readiness in low-income areas, where intricate socioeconomic interactions require accurate and reliable analysis.

2. Materials and Methods

This section contains methodological steps for poverty classification, where households might be categorized into multiple poverty levels. The following steps are involved in classifying household poverty in Kut Bak District.

2.1. Data Collection

Field surveys conducted by trained personnel and interviews with household members form the cornerstone of data collection in socioeconomic research, particularly in Kut Bak District. In this study, 301 households were selected using stratified random sampling, ensuring proportional representation across different income levels in the Kut Bak District. This approach reduces selection bias and improves generalizability compared to convenience sampling techniques. The primary objective was to classify these households into four distinct poverty levels. This classification was not just a mere categorization but a step towards understanding the nuanced and multi-dimensional nature of poverty in the area. The study strictly adheres to ethical guidelines, having received formal approval from the ethics committee at Sakon Nakhon Rajabhat University. It was granted clearance through a full board review (approval number HE 65-099), valid from August 31, 2022, to August 31, 2023. This ethical framework safeguards the rights and well-being of all participants involved. In addition to obtaining ethics committee approval, this study adhered to strict confidentiality measures, ensuring that household-level data were anonymized and securely stored. This ethical framework underscores the importance of responsible data collection and analysis when addressing sensitive social issues like poverty.

By categorizing households, we can better understand each group's challenges and effectively tailor interventions. The study employs a nuanced approach to classify households into four distinct poverty levels. This classification is a simple economic assessment and a multidimensional analysis considering various aspects of household well-being. The methodology revolves around three main features:

1- Incomes: This feature covers various sources of household income, including, but not limited to, agricultural earnings, wage labor, and remittances. It provides insight into households' financial stability and earning capacity.

2- Expenses: Household expenses, including routine expenditures, healthcare, educational expenses, and other significant outgoings, offer a window into families' living standards and economic pressures.

3- Project Participation: Whether governmental or non-governmental, engagement in local projects indicates a household's access to external support systems and community resources. It reflects on the social capital and the level of integration within development initiatives.

These main features are further divided into sub-features, providing a layered and detailed perspective on each household's socioeconomic status. The sub-features encompass a range of indicators from basic income and expenditure to involvement in specific community projects, painting a comprehensive picture of the multi-dimensional nature of poverty. Applying the multiclass decision forest method in this context aims to utilize this detailed data, offering a more accurate and insightful classification of poverty levels than traditional methods. This approach is expected to yield significant insights, enabling more effective targeting of poverty alleviation efforts in Kut Bak District and potentially offering a model that can be replicated in other similar contexts. The data features are outlined in Table 1 as follows.

2.2. Data Preparation

Several critical steps are undertaken to prepare the dataset in the preprocessing data analysis stage to ensure the data accuracy and reliability. Data cleaning is conducted initially, but values are present across all columns in the dataset. Next, feature selection is carried out, where relevant features, such as income, expenses, and project participation, are identified based on their significant impact on poverty classification. This step is crucial for focusing the analysis on the most influential variables. Following this, data transformation is applied to normalize or scale the data, ensuring consistency across the dataset and mitigating issues arising from differing scales or distributions.

The preprocessing stage of the data analysis was simplified due to the numerical nature of all variables in the dataset. Given that the data consisted entirely of quantitative measures, there was no requirement for encoding categorical variables or transforming qualitative information into numerical formats. This intrinsic numerical structure of the data eliminated the need for standard preprocessing techniques such as one-hot encoding, label encoding, or ordinal encoding, typically employed to convert categorical data into a format suitable for mathematical operations and machine learning algorithms. The absence of this encoding step in the data preparation phase streamlined the analysis process. It reduced the risk of introducing artifacts or biases sometimes arising from variable encoding procedures. In this case, the dataset contains 27 numerical features, indicating various aspects of household finances and project involvement: 1) income-related (6 columns ending with 'I'): AHI, AGRI, NTI, ILH, SAI, THI; 2) expenditure-related (14 columns ending with 'E' or 'EX'): HCE, UE, ESE, EDUEX, HEX, IPE, SEX, LEX,

GEX, EDEX, TREX, MMFE, CMFE, THEX; and 3) project participation-related (7 columns ending with 'P'): VCP, VSPP, CWPP, CRVCP, CRFUP, CRIUP, TIPP.

Table 1 Features used in this research and their descriptions.

Feature	Description
AHI	Annual Household Income: Total income earned by all household members in a year.
AGRI	Agricultural Income: Income derived from agricultural activities like farming or livestock.
NTI	Non-Tangible Income: Income earned from non-physical or intangible sources such as copyrights, patents, or digital products.
ILH	Income from Land Holdings: Revenue generated from owning land, including leasing or selling.
SAI	Support and Assistance Income: Income received from social support or assistance programs.
THI	Total Household Income: Aggregate of all income sources for the household.
HCE	Household Consumption Expenditure: Total expenditure by the household on consumption goods and services.
UE	Utility Expenditure: Money spent on water, gas, and electricity.
ESE	Essential Services Expenditure (Water, Electricity, Phone): Spending on essential household services, including water, electricity, and telecommunication.
EDUEX	Educational Expenditure: Expenditure on education-related services and materials.
HEX	Healthcare Expenditure: Spending on medical services and health-related products.
IPE	Insurance Premium Expenditure: Payments made for insurance policies.
SEX	Social Event Expenditure: Money spent on weddings, funerals, or religious ceremonies.
LEX	Leisure and Entertainment Expenditure: Expenditure on activities for relaxation and entertainment.
GEX	Gambling Expenditure: Money spent on gambling activities.
EDEX	Energy Drink Expenditure: Spending specifically on energy drinks.
TREX	Travel Expenditure: Expenditure on travel, including tickets, accommodation, and other related costs.
MMFE	Motorcycle Maintenance and Fuel Expenditure: Costs associated with maintaining and fueling motorcycles.
CMFE	Car Maintenance and Fuel Expenditure: Costs related to car maintenance and fuel.
THEX	Total Household Expenditure: Sum of all expenditures incurred by the household.
VCP	Vegetable Cultivation Project Participation: Involvement in projects aimed at cultivating vegetables, typically community driven.
VSPP	Vegetable Seedling Project Participation: Participating in projects focusing on growing vegetable seedlings.
CWPP	Community Welfare Project Participation: Engagement in community welfare projects designed to improve local living conditions.
CRVCP	Cost Reduction in Vegetable Cultivation Project: Benefits derived from initiatives aimed at reducing costs in vegetable cultivation.
CRFUP	Cost Reduction in Fertilizer Usage Project: Cost savings are achieved through projects that reduce fertilizer usage.
CRIUP	Cost Reduction in Insecticide Usage Project: Savings from reduced insecticide use in agricultural projects.
TIPP	Total Income from Project Participation: Total income generated from participation in various community or agricultural projects.

2.3. Correlation Heatmap Analysis

Correlation heatmap analysis is a visual tool representing the pairwise correlations between multiple variables in a dataset (Cady, 2017; Mariani et al., 2021). It displays a matrix where each cell represents the correlation coefficient between two variables, often colored from high to low to indicate the strength and direction of the correlation (e.g., from red for high positive to blue for high negative). This type of analysis is used for your dataset to quickly identify and visualize the strength and direction of relationships between different socioeconomic factors, such as income, expenditure, and project participation. By using a correlation heatmap, closely related variables can be easily spotted, which helps in understanding the underlying structure of the data, reducing dimensionality if redundant variables are identified, and informing the selection of variables for further analysis, like modeling or clustering. This method enhances the efficiency and effectiveness of data analysis by guiding data preprocessing and analytical strategies based on observed relationships.

2.4. Data Clustering

Poverty analysis often involves uncovering hidden patterns or natural groupings in the data, which aligns well with the unsupervised learning paradigm. The choice of clustering methods (k-means and k-medoids) over supervised learning techniques was driven by the exploratory nature of this research and the absence of predefined labels for household income categories (Arora et al., 2016). Unlike supervised learning, which requires labeled data to predict predefined categories, clustering allows for the discovery of meaningful socioeconomic groups based on the inherent structure of the data (Hennig & Liao, 2013). In this study, the goal was to classify households into low, medium, and high income categories in a data-driven manner, without relying on external definitions or subjective thresholds. Clustering methods provided flexibility to identify these categories based on the distribution of household income and related features, making them more suitable for this task. Additionally, clustering methods offer practical advantages when dealing with real-world household income data, which often contains outliers and noise (Sihombing & Arsani, 2021). Furthermore, clustering serves as a valuable preprocessing step for

downstream tasks. Once the clusters were identified, they could be used as labels for training supervised models, such as random forests, to predict poverty levels or other outcomes. This hybrid approach leverages the strengths of both unsupervised and supervised learning, allowing for a comprehensive analysis of household poverty. K-means is a popular clustering algorithm that partitions data into clusters by minimizing the variance within clusters, using the arithmetic mean of points as the center (Knox, 2018). However, k-means is sensitive to outliers and anomalies since extreme values influence the centroid. This distorts the clustering results, especially in datasets with socioeconomic features that often have outlier values.

In contrast, k-medoids clustering is a partitioning technique similar to k-means but uses medoids as the center of each cluster instead of means (Harikumar & Pv, 2015). A medoid is the most centrally located data point within a cluster, making it less sensitive to outliers than centroids. The algorithm, often implemented using the partitioning around medoids (PAM) method, minimizes the sum of dissimilarities between points in a cluster and the medoid (Arora et al., 2016). This robustness also makes k-medoids particularly suitable for datasets with extreme income values or anomalies, as it ensures more reliable clustering outcomes. This method provides representative clustering that maintains the integrity of the analysis even in the presence of outliers. This technique ensures that the results are reliable even when the data contains outliers, a common occurrence in poverty-related datasets.

The elbow method is one of the most straightforward and intuitive ways to determine the optimal number of clusters. It involves plotting the within-cluster sum of squares (WCSS) against the number of clusters (k) and then identifying the 'elbow point' where adding more clusters results in diminishing returns regarding reducing WCSS. Although the elbow method is heuristic and may not always provide the most statistically rigorous result, it has been shown to work well in practice for many real-world applications, especially when the goal is to create interpretable clusters (Knox, 2018). In contrast, methods like the Gap Statistic and Silhouette Score require more complex calculations. They may not always provide a straightforward interpretation, especially when dealing with large or noisy data. Moreover, the elbow method is computationally less intensive compared to other methods. For example, the Gap Statistic requires comparing the observed WCSS with reference datasets generated under a null hypothesis, which can be time-consuming. The Silhouette Score calculates pairwise distances between all points, which can be computationally expensive for large datasets. Since this research aims to use clustering as a preprocessing step for classification with random forests, the elbow method can provide effective clusters.

2.5. Data Clustering Evaluation

The clustering performance evaluation is critical to assess the resulting clusters' quality and applicability to real-world contexts, such as household poverty analysis. In this study, three key evaluation metrics were employed: Silhouette Score (Januzaj et al., 2023), Calinski-Harabasz Index (Hu, 2024), and Davies-Bouldin Index (Wijaya et al., 2021). These metrics were selected due to their complementary nature in assessing cluster cohesion, separation, and overall quality.

1- Silhouette score: This metric measures how well each data point fits within its assigned cluster relative to others. It is computed as the mean silhouette coefficient, which ranges from -1 to 1. A score closer to 1 indicates that clusters are well-separated and internally cohesive, while values near 0 suggest overlap between clusters. Negative values indicate misassigned data points. This metric provides insights into the overall structure and clarity of the clusters.

2- Calinski-Harabasz index: Also known as the Variance Ratio Criterion, this index evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values indicate that clusters are compact and well-separated, ideal for clearly defined groupings. This metric is beneficial for comparing cluster structures across different clustering algorithms.

3- Davies-Bouldin index: This index measures the average similarity between each cluster and the most similar cluster. A lower score indicates better clustering, which signifies more significant dissimilarity between clusters and better-defined boundaries. This metric is sensitive to cluster overlap and is a reliable measure of cluster separation.

In the context of household poverty analysis, these metrics were applied to evaluate clustering results generated by k-means and k-medoids. Each method categorized households into distinct groups based on income characteristics, and the clustering outcomes were assessed using the three metrics to determine which method provided the most meaningful and interpretable segmentation of households. By comparing the metrics across methods, we identified the strengths and weaknesses of each approach in accurately grouping households, offering insights into their economic conditions. This methodology ensures a comprehensive evaluation of clustering techniques, guiding the selection of the most effective method for further analysis.

2.6. Data Classification

Recent studies have increasingly leveraged machine learning techniques to enhance poverty classification accuracy. Huang & Xia (2023) applied a decision tree-based approach to classify rural poverty levels, emphasizing land-use variables. Similarly, Muñetón-Santa et al. (2022) proposed an NLP-driven framework to analyze household discourse and predict deprivation levels. Other studies have explored various classification algorithms, such as Random Forest and SVM, to assess socioeconomic vulnerability (Sihombing & Arsani, 2021; Alsharkawi et al., 2021). However, these approaches rely on predefined

labels, which may introduce biases in dynamic poverty contexts. By contrast, clustering techniques such as K-Means and K-Medoids enable the discovery of natural economic groupings without prior assumptions, providing a more flexible framework for poverty stratification (Zhang & Dai, 2023).

In this study, random forests were employed as a robust classification tool to evaluate and validate the clustering results from both k-means and k-medoids. Socioeconomic datasets, such as this one, often exhibit significant complexity and nonlinearity, incorporating diverse features like income, expenditures, and indirect income from project participation. Random forests, being an ensemble method, are well-suited for handling such multifaceted datasets due to their ability to capture intricate relationships among variables while maintaining robustness against overfitting. Random forests also provide a unique advantage by offering feature importance rankings. These were used to identify the most influential factors driving the clustering outcomes for both k-means and k-medoids. This interpretability is crucial in understanding the socioeconomic dynamics underlying the classifications. Furthermore, random forests excel in handling missing values, allowing them to maintain accuracy even when portions of the data are incomplete. Their ensemble nature, which aggregates predictions from multiple decision trees, ensures stability and reliability, making them an ideal choice for evaluating the complex classification tasks inherent in socioeconomic clustering and predicting outcomes such as economic status classifications.

Random Forest creates multiple bootstrap samples from the original dataset (Knox, 2018). A bootstrap sample is a randomly selected subset of the data, chosen with replacement, typically the same size as the original dataset. This sampling technique is given by $\mathrm{D_i} = \{(\mathrm{x_1}, \mathrm{y_1}), (\mathrm{x_2}, \mathrm{y_2}), \ldots, (\mathrm{x_n}, \mathrm{y_n})\}$ for $\mathrm{i} = 1$ to N, where $\mathrm{D_i}$ is the i-th bootstrap sample, and N is the number of trees. For a classification task, the final prediction is made by majority voting among all the trees in the forest. If Ck(x) denotes the class predicted by the k-th tree for input x, the final class prediction $\hat{y}$ is given by $\hat{y}(x) = mode\{C_1(x), C_2(x), \ldots, C_N(x)\}$.

2.7. Model Performance Evaluation

The effectiveness of the clustering models, k-means and k-medoids, was evaluated using a testing dataset comprising 30% of the data, while the remaining 70% was used for training and optimization. This critical testing phase applied the models to previously unseen data, providing an unbiased measure of their performance in classifying poverty levels into distinct categories such as low, medium, and high economic statuses. The evaluation ensures that the models' precision and dependability are assessed in real-world-like conditions by focusing on testing data.

Four essential metrics were employed to gauge performance comprehensively: accuracy, precision, recall, and F1-score. These metrics offered insights into different aspects of the models' classification capabilities, such as overall correctness (accuracy), ability to avoid false positives (precision), effectiveness in identifying true cases (recall), and the balance between precision and recall (F1-score). Confusion matrices were generated for k-means and k-medoids to validate performance further. These matrices visually represent the models' predictions by summarizing true positives, false positives, true negatives, and false negatives. By leveraging these evaluation techniques, a thorough assessment of the models' ability to classify households into various poverty levels was achieved, ensuring their reliability and practical applicability in socioeconomic analysis.

3. Results

This part presents various analytical outcomes, encompassing the visualization of variable relationships through correlation heatmaps, the reduction of data complexity, the determination of the ideal cluster count, the evaluation of clustering performance, the categorization of socioeconomic information, and the results of the classification process.

3.1. Correlation Heatmap

Figure 1 illustrates a complex network of correlations among various income sources and project participation variables, revealing several noteworthy relationships. The strongest correlations are observed within the cluster of income variables derived from different project participations. Notably, vegetable cultivation project participation (VCP) and total income from project participation (TIPP) exhibit an exceptionally high correlation of 0.96. Similarly, strong correlations are evident between cost reduction in fertilizer usage project (CRFUP) and TIPP (0.94), cost reduction in insecticide usage project (CRIUP) and TIPP (0.92), CRIUP and CRFUP (0.92), CRFUP and VCP (0.89), CRIUP and VCP (0.86), and cost reduction in vegetable cultivation project (CRVCP) and VCP (0.86). While these project-related correlations are particularly robust, other significant relationships warrant attention. Total household income (THI) demonstrates strong correlations with agricultural income (AGRI) at 0.61 and non-tangible income (NTI) at 0.70. Additionally, community welfare project participation (CWPP) shows moderate correlations with both AGRI (0.44) and NTI (0.48).

Total household expenditure (THEX) exhibits substantial correlations with various expenditure categories, including household consumption expenditure (HCE) at 0.53, utility expenditure (UE) at 0.51, essential services expenditure (ESE) at 0.53, educational expenditure (EDUEX) at 0.48, healthcare expenditure (HEX) at 0.44, insurance premium expenditure (IPE) at 0.45, gambling expenditure (GEX) at 0.33, energy drink expenditure (EDEX) at 0.50, and travel expenditure (TREX) at 0.54.

Interestingly, leisure and entertainment expenditure (LEX) displays notable correlations with several project participation variables: VCP (0.32), CRVCP (0.41), CRFUP (0.37), CRIUP (0.33), and TIPP (0.39). In this case, the observed correlations provide critical insights for policymakers. For instance, the strong relationship between project participation variables (e.g., CRFUP, CRIUP) and income measures underscores the potential of targeted agricultural support programs to uplift economic conditions. These findings highlight the interconnectedness of community initiatives and economic resilience, suggesting opportunities for scaling similar interventions in other rural settings.

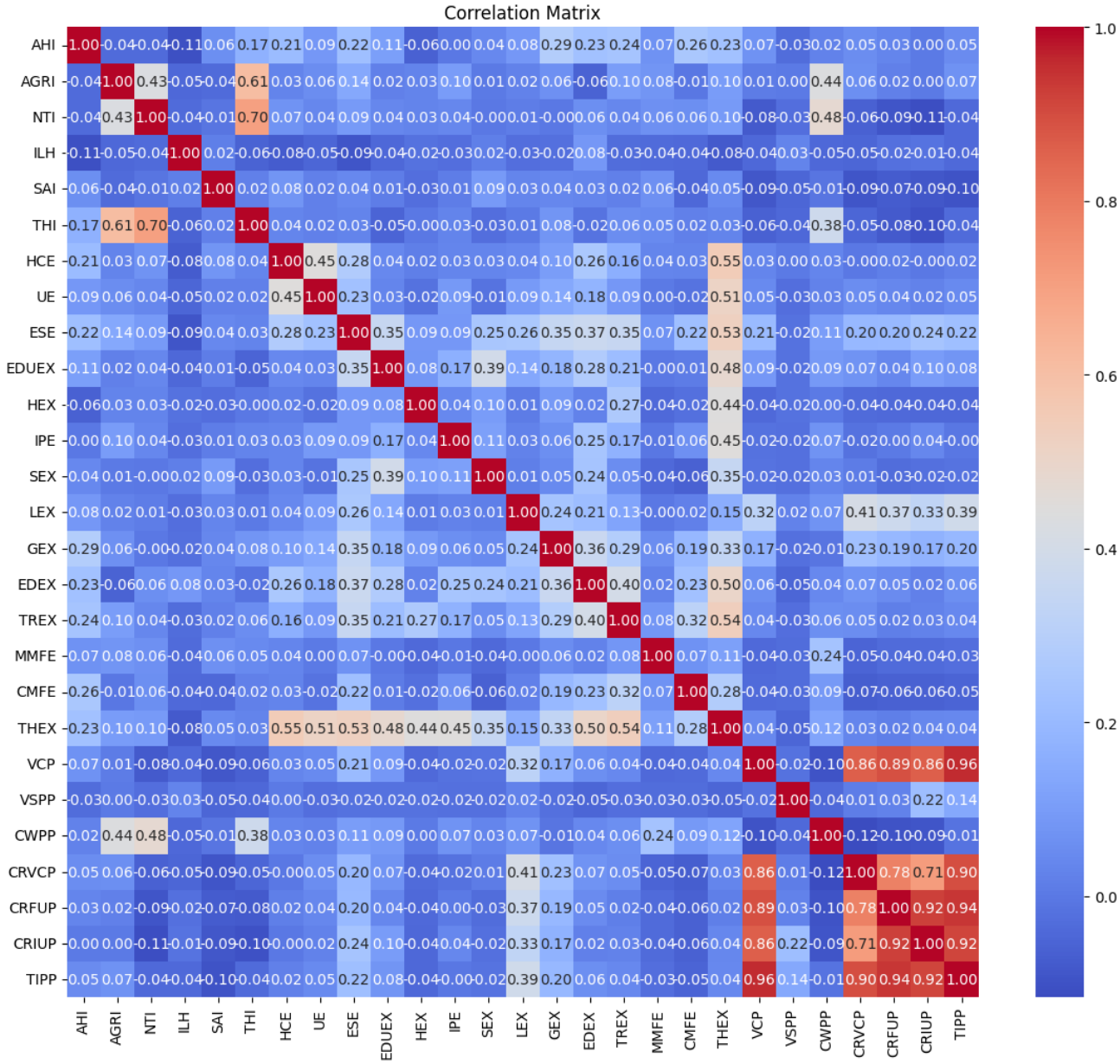


Figure 1 A correlation heatmap representing relationships between the variables.

3.2. Optimal Number of Clusters

The elbow method was employed to determine the optimal number of clusters for both k-means and k-medoids. In both cases, the analysis revealed a clear 'elbow' at *k=3*, indicating the presence of three distinct clusters within the dataset. This optimal clustering aligns with the underlying goal of segmenting the income data into Low, Medium, and High. The inertia, which measures the sum of squared distances between data points and their respective cluster centroids (medoids), demonstrated a significant drop up to *k=3*, followed by a plateau, signifying minimal gains in clustering quality with additional clusters. While both methods reached the same conclusion regarding the optimal cluster count, the paths to achieving this result were distinct. K-means minimizes inertia based on centroids, the average positions of points within clusters, whereas k-medoids focuses on minimizing dissimilarities using actual data points as cluster centers. This fundamental difference affects clustering, particularly in noisy or outlier datasets. Despite this, both methods were effective in uncovering the same cluster

structure, demonstrating their reliability in identifying key patterns within the data. Figure 2 illustrates the application of the elbow technique to determine the ideal number of groupings for (a) k-means and (b) k-medoids clustering.

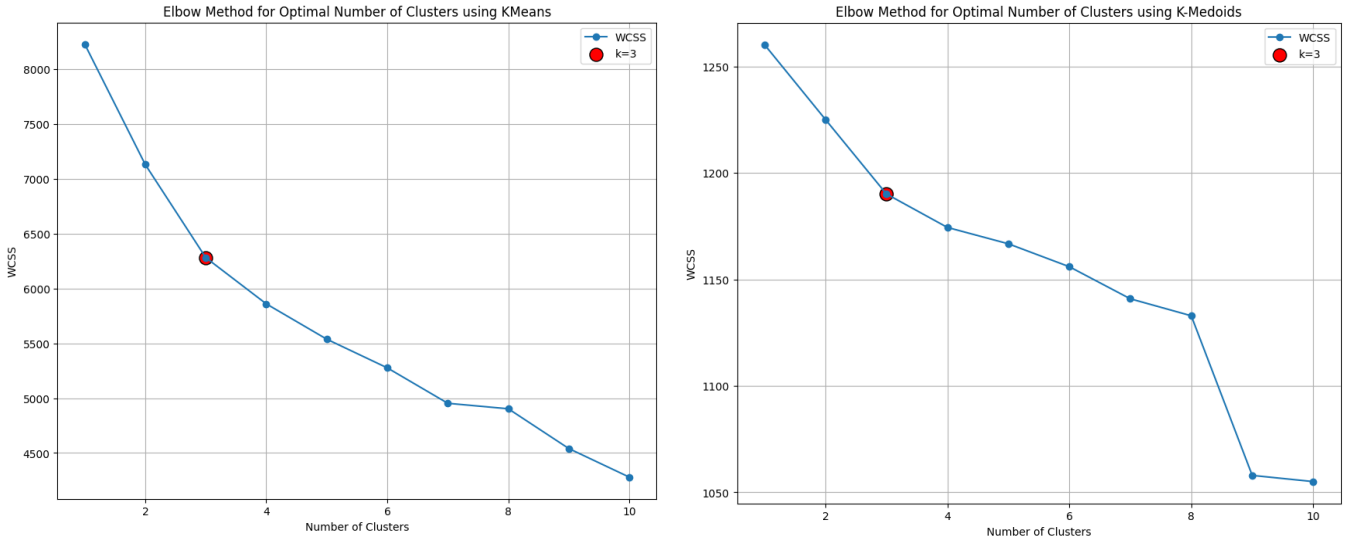


Figure 2 Elbow method for optimal number of clusters using k-means and k-medoids.

3.3. Data Clustering

K-means and k-medoids successfully segmented the dataset into three clusters corresponding to distinct economic statuses: Cluster 0 represents households with a 'Low' economic status, Cluster 1 captures households with a 'Medium' economic status, and Cluster 2 encompasses households with a 'High' economic status. These clusters were derived based on features indicative of income levels, expenditures, and other socioeconomic factors. The clustering results revealed differences in how household incomes were grouped into low, medium, and high-income categories based on Total Household Income (THI). While both methods successfully identified three distinct income groups, the income ranges within each cluster varied due to differences in how the two algorithms determine cluster centers. K-means, which assigns clusters based on centroids (mean values), produced wider income ranges because it is more sensitive to outliers. This sensitivity led to a high-income cluster that contained only a single outlier household at 105,653.33. In contrast, the lower-income clusters spanned -264 to 9,166.67 for the low-income group and 9,572.50 to 41,607.50 for the medium-income group. The inclusion of an extreme outlier in the high-income cluster resulted in a less-even distribution of income groups, making the middle-income category unusually wide.

In contrast, k-medoids formed more compact and structured income ranges because it selects actual data points (medoids) as cluster centers, making it more robust to outliers. Instead of allowing extreme values to distort cluster assignments, K-Medoids grouped households into more realistic and well-separated economic categories. Initially, the income ranges in K-Medoids were not smoothly ordered, requiring an adjustment to ensure continuous progression. After this refinement, the low-income cluster ranged from -264 to 7,115.83, the medium-income cluster spanned 7,405 to 28,309.72, and the high-income cluster covered 36,664.17 to 105,653.33. This more balanced segmentation suggests that k-medoids provides a more reliable classification of household income groups, ensuring that middle-income households are not stretched too broadly and that extreme values do not dominate high-income groups.

The grouping demonstrates the effectiveness of both clustering methods in categorizing households into meaningful and interpretable segments. While both methods produced similar groupings, the underlying clustering mechanisms led to nuanced differences in classification. K-means relies on centroids, which makes it sensitive to outliers and better suited for datasets with relatively uniform distributions. Conversely, k-medoids uses actual data points as cluster centers (centroids), offering greater resilience to noise and outliers. This distinction may result in slight variations in cluster boundaries, with k-medoids potentially being more reliable in datasets with irregular patterns or anomalies. However, both methods provided consistent economic stratifications in this dataset, validating their utility in identifying key household income groups.

The segmentation into three economic strata—low, medium, and high—offers a practical framework for designing targeted interventions. Policymakers and NGOs can use this classification to focus resources effectively. For example, households in the 'low' cluster may require direct financial assistance, while 'medium' and 'high' groups could benefit from capacity-building programs and access to credit facilities. Figure 3 represents (a) k-means and (b) k-medoids clustering for household economy status into low, medium, and high.

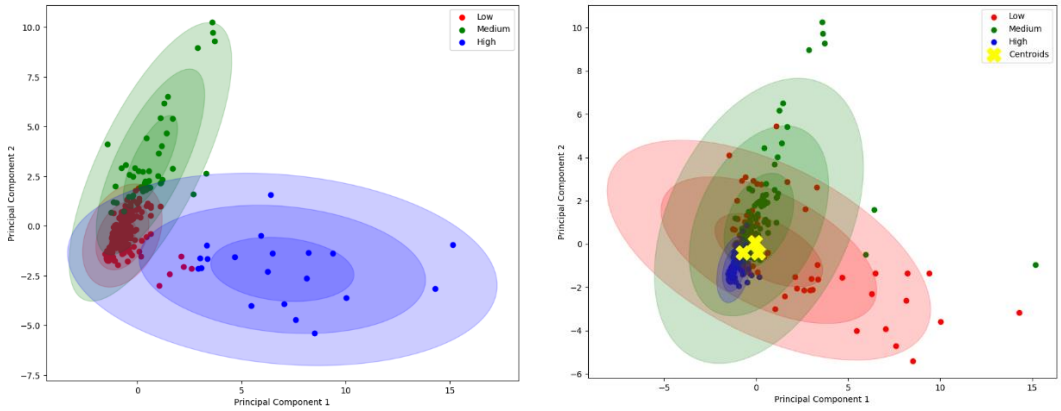


Figure 3 K-means and k-medoids clustering for household economy status.

3.4. Clustering Performance Evaluation

The clustering results for k-means and k-medoids were evaluated using the Silhouette score, Calinski-Harabasz index, and Davies-Bouldin index. Each metric provides unique insights into the clustering quality regarding cohesion and separation. The Silhouette score, which ranges from -1 to 1, evaluates how well-separated clusters are from one another. A higher score indicates stronger cluster cohesion and separation. In this case, k-medoids scores slightly better (0.430) than k-means (0.418), suggesting that k-medoids clusters are marginally more distinct and internally consistent. This may reflect k-medoids' reliance on medoids (actual data points) as cluster centers, which can better handle outliers and provide more robust clustering in some cases. The Calinski-Harabasz index measures the ratio between-cluster dispersion to within-cluster dispersion, with higher values indicating better-defined clusters. Both methods perform nearly identically, with k-means scoring slightly higher (46.184 vs. 46.108). This indicates that both methods create clusters with similar compactness and separation, showing that either method could work well for the dataset. Finally, where lower values are better, the Davies-Bouldin index measures the average similarity between each cluster and the most similar one. K-medoids slightly outperforms k-means (2.056 vs. 2.084), suggesting better-defined clusters with less overlap. Overall, both methods perform well, but k-medoids exhibits a slight advantage in cluster definition and robustness to outliers. At the same time, k-means remains a strong option for its computational efficiency and comparable performance. The choice between the two methods depends on the specific context and importance of these nuances (Table 2).

Table 2 Comparison of clustering performance metrics for k-means and k-medoids.

Metric	K-Means	K-Medoids
Silhouette	0.418	0.430
Calinski-Harabasz	46.184	46.108
Davies-Bouldin	2.084	2.056

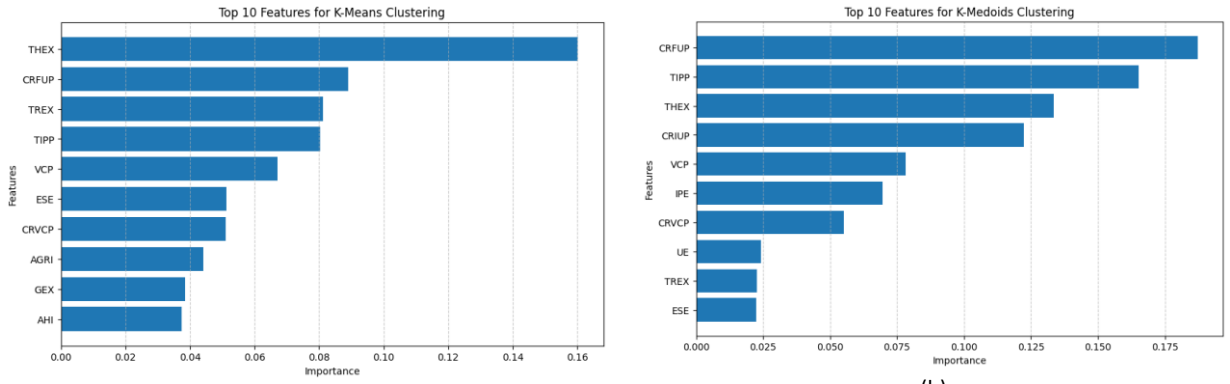
3.5. Data Classification on Socioeconomic Data

The top features identified for k-means and k-medoids clustering reflect how these methods interpret socioeconomic data. In k-means, THEX emerges as the most critical feature, with an importance score of 0.160, emphasizing that overall spending patterns are pivotal in distinguishing economic statuses. This is followed by CRFUP and TREX, highlighting the relevance of cost-saving measures in agricultural projects and travel-related spending as indicators of economic capability. TIPP and VCP rank next, showing the influence of income and engagement in community-driven agricultural activities. Finally, ESE (Essential Services Expenditure) underscores the role of basic utility and telecommunication costs in clustering households.

In contrast, k-medoids assigns the highest importance to CRFUP (0.187) and TIPP (0.165), suggesting a stronger emphasis on financial benefits derived from community projects and participation in agricultural cost-saving initiatives. THEX, while still significant, ranks third (0.133), indicating that total expenditure remains an essential but less dominant factor compared to k-means. CRIUP and IPE rank prominently in k-medoids, reflecting their greater focus on agricultural cost efficiency and financial planning indicators. CRVCP also appears in k-medoids' top features, emphasizing granular agricultural cost-saving activities. Lower-ranked features like TREX and ESE in k-medoids suggest a broader but less detailed analysis of consumption patterns compared to k-means.

The shared emphasis on CRFUP and TIPP across both methods highlights the critical role of agricultural and project-based income in defining economic clusters. However, k-means relies more on aggregate household-level measures like THEX, while k-medoids delves deeper into specific financial and agricultural cost-saving features such as CRIUP and CRVCP. This makes k-medoids better suited for datasets where agricultural and financial efficiency are crucial. At the same time, k-means identifies

broader economic trends driven by total expenditure and project participation. Figure 4 illustrates the top 10 key features influencing the 'low' household economic status, as determined by (a) k-means and (b) k-medoids clustering.



(a) (b)

Figure 4 Top 10 important features contributing to 'low' household economy status based on k-means and k-medoids clustering.

After obtaining the top 10 feature importance rankings for k-means and k-medoids clustering using random forest, it is essential to synthesize and compare these features to validate their relevance in real-world contexts. Such a comparison highlights overlaps and differences in the features prioritized by each method, providing a deeper understanding of how socioeconomic factors influence clustering. By integrating these rankings, we can assess whether the features consistently drive clustering outcomes across both algorithms, strengthening their credibility. This synthesis also bridges the gap between technical findings and practical applications, ensuring the features identified are statistically significant and actionable in addressing real-world challenges.

Table 3 provides a synthesized comparison of the top 10 feature importance rankings for k-means and k-medoids. Alongside their definitions and importance scores, the table offers insights into how each feature reflects household economic status and recommends corresponding policy implications. For instance, Total Household Expenditure (THEX) consistently ranks as the most critical feature, emphasizing the centrality of expenditure patterns in economic stratification. Features like the Cost Reduction in Fertilizer Usage Project (CRFUP) highlight the impact of agricultural initiatives in uplifting rural households. By understanding these features' roles, policymakers can design targeted interventions, such as financial literacy programs or expanded agricultural support schemes, to improve economic resilience across different strata.

3.6. Data Classification Results

When comparing k-means and k-medoids clustering performance metrics, k-means consistently delivers a more balanced performance across crucial evaluation criteria. K-means' accuracy is 96%, slightly higher than the 93% achieved by k-medoids, indicating that k-means is more effective at correctly classifying households into economic groups overall. Similarly, k-means demonstrates superior recall (93% vs. 86%), showcasing its ability to identify a higher proportion of true positive cases within each economic group. This makes k-means a better choice when maximizing the identification of households within each economic stratum is a priority.

Conversely, k-medoids excels in precision, achieving an average score of 97%, compared to 96% for k-means. This indicates that k-medoids is slightly better at minimizing false positives, ensuring that the predicted income groups are more accurate. This strength in precision aligns with k-medoids' robustness to outliers, as it uses medoids (actual data points) as cluster centers, which leads to more accurate and stable classifications in noisy or heterogeneous datasets. While k-means achieved higher recall and overall accuracy, k-medoids demonstrated greater precision, making it suitable for contexts where false positives are critical to avoid. This trade-off highlights the importance of selecting clustering techniques based on specific analytical needs, especially in noisy datasets. For instance, in poverty contexts with potential outliers (e.g., extreme income levels), k-medoids' robustness may provide more reliable classifications.

The F1-Score balances precision and recall and highlights the overall trade-off between these metrics. K-means achieves a marginally higher F1-Score (94%) compared to k-medoids. This underscores k-means' strength as a more balanced model, particularly in datasets requiring a harmonious trade-off between identifying all relevant cases (recall) and minimizing false positives (precision). While k-medoids' higher precision makes it suitable for contexts prioritizing classification accuracy over recall, k-means offers a more holistic approach, ideal for broader socioeconomic analyses. The classification results are not just technically robust but also socially impactful. The ability to correctly categorize households into economic strata makes resource allocation more precise and effective. For instance, households classified as 'Low' could be prioritized for immediate financial assistance, while 'Medium' households might receive skill development or credit access programs to prevent downward mobility. These nuanced classifications ensure that interventions are equitable and tailored to foster long-term

socioeconomic stability. Figure 5 depicts the classification performance of the random forest model for k-means and k-medoids clustering.

Table 3 Synthesized comparison of top 10 feature importance for k-means and k-medoids.

No	Feature	Definition	Importance (k-means)	Importance (k-medoids)	Insights	Policy Implication
1	THEX (Total Household Expenditure)	Aggregate household spending on all categories (essential and non-essential).	0.160076	0.192568	Broad indicator of financial capacity; higher expenditures correlate with higher economic strata.	Focus on financial literacy and expenditure optimization to improve economic stability.
2	CRFUP (Cost Reduction in Fertilizer Usage Project)	Participation in agricultural initiatives aimed at reducing fertilizer costs.	0.08905	0.097298	Key marker of agricultural support impact; strongly influences income and economic status.	Expand cost-reduction programs targeting farming communities to boost economic resilience.
3	TREX (Travel Expenditure)	Household spending on travel-related activities, including transportation and accommodation.	0.081199	0.050828	Reflects travel patterns as indicators of economic capability and lifestyle.	Improve transportation infrastructure and travel-related economic opportunities.
4	TIPP (Total Income from Project Participation)	Total income generated through participation in agricultural and community projects.	0.080356	0.078371	Shows economic benefits from project participation, critical in rural economies.	Encourage project participation through subsidies and incentives to enhance income.
5	VCP (Vegetable Cultivation Project Participation)	Engagement in vegetable cultivation projects, usually community driven.	0.067208	0.084335	Highlights the role of community-driven agricultural initiatives in economic stratification.	Promote agricultural engagement and provide technical support for community projects.
6	ESE (Essential Services Expenditure)	Spending on essential utilities and telecommunication services.	0.051223	0.079113	Utility and telecommunication spending indicate essential financial capability.	Ensure affordable access to utilities and improve telecommunication infrastructure.
7	CRVCP (Cost Reduction in Vegetable Cultivation Project)	Participation in cost-saving initiatives for vegetable cultivation.	0.050974	0.045516	Indicates targeted savings in agriculture, helping financially vulnerable households.	Develop targeted interventions for agricultural efficiency and cost reductions.
8	AGRI (Agricultural Income)	Income derived from agricultural activities like farming or livestock.	0.044039	-	Significant for rural areas where agriculture is a primary income source.	Provide better market access and support for agricultural outputs.
9	GEX (Gambling Expenditure)	Household spending on gambling activities, an indicator of discretionary income.	0.038391	0.058502	Discretionary spending behavior can indicate disposable income availability.	Increase awareness about financial management and its impact on economic status.
10	AHI (Annual Household Income)	The total annual income earned by all household members.	0.03746	-	Reflects overall household earning capacity as a broad economic marker.	Enhance household income generation through employment and skill development programs.
11	IPE (Insurance Premium Expenditure)	Payments made towards insurance policies, reflecting financial planning.	-	0.051236	Shows long-term financial planning behaviors, important for economic stability.	Encourage widespread adoption of insurance schemes to improve financial resilience.
12	CRIUP (Cost Reduction in Insecticide Usage Project)	Savings achieved from reduced insecticide usage in agricultural activities.	-	0.035924	Indicates project-driven cost efficiency in agriculture, aiding in income stability.	Promote eco-friendly and cost-saving agricultural practices for better economic outcomes.

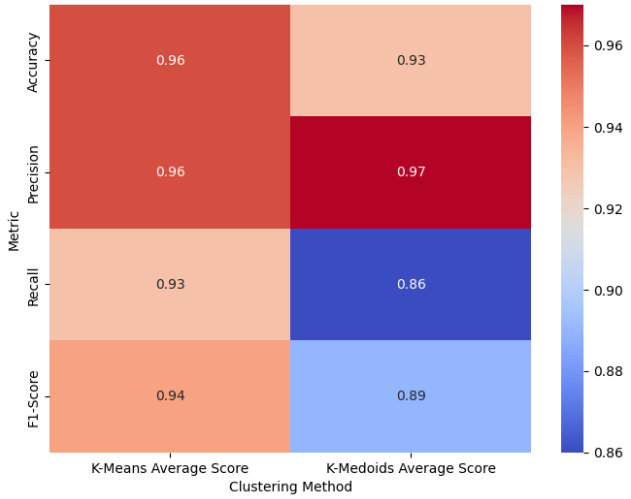


Figure 5 Random forest performance classification for k-means and k-medoids clustering.

3.7. Confusion Matrix Comparison

The confusion matrices in Figure 6 for (a) k-means and (b) k-medoids reveal distinct patterns in their classification outcomes, reflecting the strengths and limitations of each method. For k-means, the diagonal entries in the confusion matrix are consistently higher, indicating that it classifies a more significant proportion of households correctly across all economic groups. This aligns with its superior accuracy (96%) and recall (93%), showcasing its strength in identifying true positives effectively. However, k-means does exhibit occasional misclassifications, particularly between adjacent economic statuses such as 'Low' and 'Medium', which may be attributed to the overlap in socioeconomic characteristics.

On the other hand, the confusion matrix for k-medoids highlights its higher precision (97%) but reveals more pronounced misclassifications in identifying households from the 'Medium' economic group. This is consistent with its lower recall (86%), as k-medoids tends to under-identify some true positive cases, prioritizing the accuracy of predicted groups over comprehensive coverage. While it performs well in classifying 'High' and 'Low' income groups, the 'Medium' group appears to pose challenges, likely due to the overlapping characteristics and reliance on medoids as cluster centers.

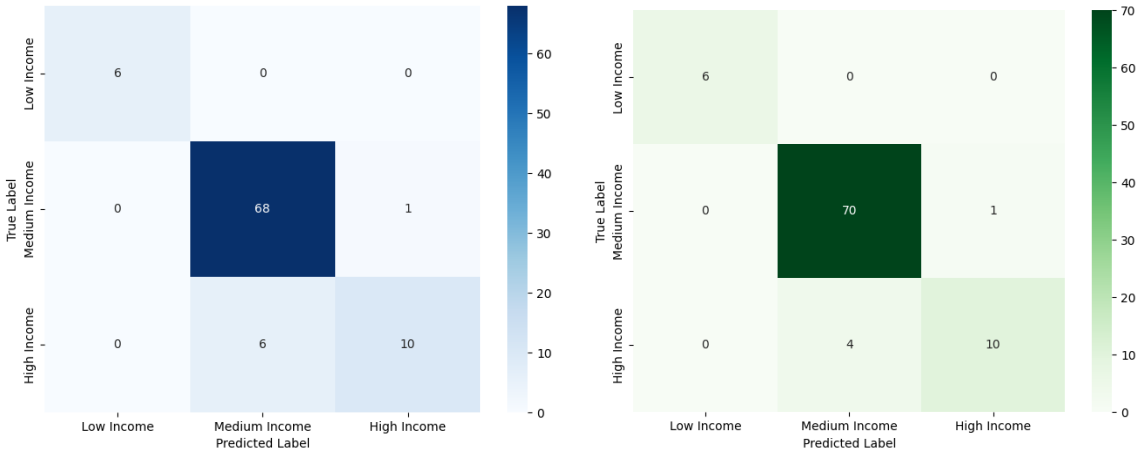


Figure 6 Confusion matrix for low household economy classification.

4. Discussion

This part explores the consequences and potential applications of the study's findings, identifies areas for further investigation, and suggests new avenues of inquiry that could build upon or extend the present work. It also examines how the results might impact current understanding or practices and suggests areas for future research directions.

5. Implications

The implications of these correlation patterns are multifaceted and significant for understanding household economics and the impact of project participation. For example, the strong correlations among various project participation variables (VCP, CRFUP, CRIUP, CRVCP, and TIPP) suggest a synergistic effect, implying that households participating in one agricultural

project are likely to engage in others, potentially leading to cumulative benefits. Policymakers might consider promoting comprehensive agricultural support packages rather than isolated interventions (Zhang & Dai, 2023). Moreover, the strong correlations between total household expenditure (THEX) and various specific expenditure categories reveal a complex spending behavior. This suggests that spending across multiple categories rises proportionally as overall expenditure increases. This insight could be valuable for developing targeted financial management programs or social support initiatives (Vaidyanathan, 2001). These observations align with the multidimensional poverty index principles outlined by UNDP (2023), emphasizing the interplay of economic behaviors and access to essential services in poverty assessments.

Interestingly, the correlations between leisure and entertainment expenditure (LEX) and various project participation variables imply that households engaged in agricultural projects may have more disposable income for leisure activities. This could indicate an improved quality of life associated with project participation. The correlations between THEX and various expenditure categories, including some non-essential ones (e.g., GEX, EDEX), highlight the importance of financial literacy programs. These could help households manage increased income more effectively. Primarily, the strong correlations between project participation variables and income measures suggest that agricultural and cost-reduction projects could be effective targets for policies aimed at improving household economic status.

The comparison underscores that k-means provides a more balanced classification performance across all economic groups, excelling in recall and overall accuracy. In contrast, k-medoids prioritizes precision and stability in classification, making it a better choice for applications where minimizing false positives is critical, even at the expense of missing some true cases. The choice between the two methods depends on the specific requirements of the analysis—whether broader coverage (k-means) or higher specificity (k-medoids) is the priority. This aligns with the findings by Harikumar & Pv (2015), who highlighted k-medoids' robustness in datasets with outliers and anomalies. Similarly, Arora et al. (2016) emphasized the suitability of k-medoids for socioeconomic datasets, where clear group partitioning is critical.

The study underscores the intertwined nature of economic and social factors in poverty analysis. The strong correlation between agricultural project participation and income reflects the potential of community-driven initiatives to uplift rural economies. Social science theories on collective action suggest that these programs provide financial benefits and build community trust, cooperation, and resilience. Future interventions could amplify these effects by fostering partnerships between households, local governments, and NGOs. Moreover, the stratification into economic clusters provides a practical framework for targeted policy design. Aligning technical classifications with social insights ensures that interventions address immediate economic needs and the underlying social determinants of poverty, such as education, healthcare access, and gender equity. This holistic approach is crucial for sustainable poverty alleviation.

Another implication involves common pool resources (CPRs), defined as natural or man-made resources where the use by one person lessens another's ability to utilize the same resource and where excluding users is difficult (Cooperman et al., 2022). CPRs encompass a variety of natural resources and environmental systems, which are integral to this study. These resources include agricultural income (AGRI) and income from land holdings (ILH), which rely on communal land and water use. Moreover, engagement in vegetable cultivation and seedling projects (VCP, VSPP) is linked to CPRs, as these initiatives depend on shared environmental resources. Furthermore, participation in community welfare, organic fertilizer, and cost reduction projects (CWPP, OFPP, CRVCP, CRFUP), along with the Cost reduction in insecticide usage project (CRIUP) and total income from project participation (TIPP), reflects the community's use of and dependence on these shared natural assets.

6. Future Research

This study's application of data mining techniques has revealed complex patterns and relationships within the data, suggesting multiple avenues for further research. Future studies could explore the causal relationships between the identified factors and economic status, employing longitudinal data to track changes over time and assess the long-term impact of specific interventions. This study exemplifies the potential of advanced data analytics and machine learning in enhancing the understanding of complex social issues like poverty. Integrating such technologies into social science research offers opportunities for more precise analysis and developing predictive models to forecast socioeconomic trends and potential crises, thereby enabling preemptive action. Moreover, future research could evaluate the scalability of the proposed methodology for larger and more diverse datasets. This could involve applying the clustering techniques to national or international datasets to assess their robustness and generalizability across varying socioeconomic and geographic contexts.

7. Conclusions

All findings, including the statistical analysis, pattern recognition, and derived rules for poverty classification, are meticulously documented. This comprehensive report is critical for policymakers and other stakeholders in poverty alleviation and common pool resource management. The derived rules and model insights provide a foundation for future research and practical interventions in the field. They offer a structured approach to understanding and addressing household poverty, which can be adapted and applied in similar contexts. This study's findings are not only significant in terms of academic research but also hold substantial implications for policy-making and targeted intervention strategies. High accuracy in poverty classification

enables policymakers and social workers to allocate and manage resources more efficiently and design programs tailored to different poverty groups' needs. For instance, while groups at the extreme ends may require more urgent and direct assistance, intermediate groups might benefit more from sustainable development programs, education, and skill-building initiatives.

Furthermore, this research underscores the importance of leveraging advanced data mining techniques in social sciences, particularly in areas like poverty assessment, where traditional methods may need to address the multi-dimensional nature of the issue. In particular, a combination of k-medoid clustering and random forest classification in poverty classification offers a comprehensive and nuanced understanding of poverty at the household level, enabling more effective and tailored poverty alleviation strategies. By continuing to refine these models and incorporating a broader range of data, future research can further enhance the accuracy and applicability of poverty classification, contributing to more effective poverty alleviation and common pool resource management strategies and socioeconomic development.

Acknowledgment

I appreciate the *Multidisciplinary Science Journal* for providing a well-structured format for authors.

Ethical considerations

I confirm that I have obtained all consent required by the applicable law to publish any personal details or images of patients, research subjects, or other individuals used. I agree to provide the *Multidisciplinary Science Journal* with copies of the consent or evidence that such consent has been obtained if requested.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research was supported by Program Management Unit on Area Based Development (PMU A) under contract no. A11F660115 for the project “Provinclal Poverty Alleviation Operating System Platform”.

References

Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature, 603*(7903), 864-870. https://doi.org/10.1038/s41586-022-04484-9

Albayati, M. B. and Altamimi, A. M. (2019). Identifying fake Facebook profiles using data mining techniques. *Journal of ICT Research and Applications, 13*(2), 107-117. https://doi.org/10.5614/itbj.ict.res.appl.2019.13.2.2

Allen, R. C. (2017). Absolute Poverty: When Necessity Displaces Desire. *American Economic Review, 107*(12), 3690–3721. https://doi.org/10.1257/aer.20161080.

Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., & Alyaman, M. (2021). Poverty Classification Using Machine Learning: The Case of Jordan. *Sustainability, 13*(3), Article 3. https://doi.org/10.3390/su13031412

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Computer Science, 78*, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Cady, F. (2017). *The Data Science Handbook*. Wiley. https://doi.org/10.1002/9781119092919

Carney, S., Gale, W. G., & Olover, M. L. (2001). *Asset Accumulation Among Low-Income Households*. In T. M. Shapiro & E. N. Wolff (Eds.), Assets for the Poor: The Benefits of Spreading Asset Ownership (pp. 165–205). Russell Sage Foundation. http://www.jstor.org/stable/10.7758/9781610444958.10

Centers for Medicare & Medicaid Services (CMS), HHS (2006). Medicare program; revisions to payment policies, five-year review of work relative value units, changes to the practice expense methodology under the physician fee schedule, and other changes to payment under part B; revisions to the payment policies of ambulance services under the fee schedule for ambulance services; and ambulance inflation factor update for CY 2007. Final rule with comment period. *Federal register*, *71*(231), 69623–70251.

Chakravarty, S. R., & Majumder, A. (2005). Measuring Human Poverty: A Generalized Index and an Application Using Basic Dimensions of Life and Some Anthropometric Indicators. *Journal of Human Development, 6*(3), 275–299. https://doi.org/10.1080/14649880500287605

Cooperman, A., McLarty, A. R., & Seim, B. (2022). Drivers of successful common-pool resource management: A conjoint experiment on groundwater management in Brazil. *Global Environmental Change*, 74, 102512. https://doi.org/10.1016/j.gloenvcha.2022.102512

D’Attoma, I., Matteucci, M. (2023). Multidimensional poverty: an analysis of definitions, measurement tools, applications and their evolution over time through a systematic review of the literature up to 2019. *Quality and Quantity*, 2023. https://doi.org/10.1007/s11135-023-01792-8

Delafiori, J., Navarro, L. C., Siciliano, R. F., Melo, G. C. d., Busanello, E. N. B., Nicolau, J. C., … & Catharino, R. R. (2021). Covid-19 automated diagnosis and risk assessment through metabolomics and machine learning. *Analytical Chemistry, 93*(4), 2471-2479. https://doi.org/10.1021/acs.analchem.0c04497

Drago, C. (2021). The Analysis and the Measurement of Poverty: An Interval-Based Composite Indicator Approach. *Economies, 9*(4):145. 10.3390/ECONOMIES9040145

Dunn, A. G., Surian, D., Leask, J., Dey, A., Mandl, K. D., & Coiera, E. (2017). Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. *Vaccine, 35*(23), 3033-3040. https://doi.org/10.1016/j.vaccine.2017.04.060

Harikumar, S., & Pv, S. (2015). K-Medoid Clustering for Heterogeneous DataSets. *Procedia Computer Science*, 70, 226–237. https://doi.org/10.1016/j.procs.2015.10.077

Hennig, C. and Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of*

the Royal Statistical Society Series C: Applied Statistics, 62(3), 309-369. https://doi.org/10.1111/j.1467-9876.2012.01066.x

Hu, Z. (2024). Comparison of k-means, k-medoids, and k-means++ algorithms based on the Calinski-Harabasz index for Covid-19 epidemic in China. *Applied and Computational Engineering, 49*(1), 11-20. https://doi.org/10.54254/2755-2721/49/20241046

Huang, K., & Xia, F. (2023). Classification of Rural Relative Poverty Groups and Measurement of the Influence of Land Elements: A Questionnaire-Based Analysis of 23 Poor Counties in China. *Land, 12*(4), Article 4. https://doi.org/10.3390/land12040918

Huang, W., Liu, Y., Hu, P., Ding, S., Gao, S., Zhang, M. (2023). What influence farmers' relative poverty in China: A global analysis based on statistical and interpretable machine learning methods. *Heliyon, 9*(9), e19525. https://doi.org/10.1016/j.heliyon.2023.e19525.

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 (iJOE), 19*(04), 174-182. https://doi.org/10.3991/ijoe.v19i04.37059

Knox, S. W. (2018). *Machine Learning: a Concise Introduction*. Wiley. https://doi.org/10.1002/9781119439868

Kolak, M., Bhatt, J., Park, Y. H., Padrón, N. A., & Molefe, A. (2020). Quantification of neighborhood-level social determinants of health in the continental United States. *JAMA Network Open, 3*(1), e1919928. https://doi.org/10.1001/jamanetworkopen.2019.19928

Llera-Sastresa, E., Scarpellini, S., Rivera-Torres, P., Aranda, J., Zabalza-Bribian, I., Aranda-Uson, A. (2017). Energy Vulnerability Composite Index in Social Housing, from a Household Energy Poverty Perspective. *Sustainability, 9*(5):691-. https://doi.org/10.3390/SU9050691

Longa, F. D., Sweerts, B., & Zwaan, B. v. d. (2021). Exploring the complex origins of energy poverty in the Netherlands with machine learning. *Energy Policy*, 156, 112373. https://doi.org/10.1016/j.enpol.2021.112373

Mah, J. C., Penwarden, J. L., Pott, H., Theou, O., & Andrew, M. K. (2023). Social vulnerability indices: A scoping review. *BMC Public Health, 23*(1), 1253. https://doi.org/10.1186/s12889-023-16097-6

Mansi, E., Hysa, E., Panait, M., & Voica, M. C. (2020). Poverty—A Challenge for Economic Development? Evidences from Western Balkan Countries and the European Union. *Sustainability, 12*(18), Article 18. https://doi.org/10.3390/su12187754

Mariani, M. C., Tweneboah, O. K., & Beccar-Varela, M. P. (2021). *Data Science in Theory and Practice: Techniques for Big Data Analytics and Complex Data Sets*. Wiley. https://doi.org/10.1002/9781119674757

Morris, M. H., Santos, S. C., & Neumeyer, X. (2018). Understanding poverty. In *Poverty and Entrepreneurship in Developed Economies* (pp. 1–20). Edward Elgar Publishing. https://doi.org/10.4337/9781788111546.00009

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87-106. https://doi.org/10.1257/jep.31.2.87

Muñetón-Santa, G., Escobar-Grisales, D., López-Pabón, F. O., Pérez-Toro, P. A., & Orozco-Arroyave, J. R. (2022). Classification of Poverty Condition Using Natural Language Processing. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 162*(3), 1413–1435.

Nurhayati, N., Sinatrya, N. S., Wardhani, L. K., & Busman, B. (2018). Analysis of K-means and K-medoids' performance using big data technology. In *2018 6th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-6). IEEE. https://doi.org/10.1109/CITSM.2018.8674251

Sani, N. S., Fikri, A., Ali, Z., Zakree, M., & Nadiyah, K. (2020). Drop-out prediction in higher education among B40 students. *International Journal of Advanced Computer Science and Applications, 11*(11). https://doi.org/10.14569/ijacsa.2020.0111169

Shin, E. K., Mahajan, R., Akbilgiç, O., & Shaban-Nejad, A. (2018). Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. *NPJ Digital Medicine, 1*(1). https://doi.org/10.1038/s41746-018-0056-y

Sihombing, P. R., & Arsani, A. M. (2021). Comparison of Machine Learning Methods in Classifying Poverty in Indonesia in 2018. *Jurnal Teknik Informatika, 2*(1), 51–56. https://doi.org/10.20884/1.jutif.2021.2.1.52

Soni, K. G., & Patel, A. (2017). Comparative analysis of k-means and k-medoids algorithm on IRIS data. *International Journal of Computational Intelligence Research, 13*(5), 899-906. https://www.ripublication.com/ijcir17/ijcirv13n5_21.pdf

Suarna, N., Wijaya, Y. A., Mulyawan, M., Hartati, T., & Suprapti, T. (2021). Comparison of k-medoids algorithm and k-means algorithm for clustering fish cooking menu from fish dataset. *IOP Conference Series: Materials Science and Engineering, 1088*(1), 012034. https://doi.org/10.1088/1757-899X/1088/1/012034

UNDP (United Nations Development Programme). (2023). 2023 Global Multidimensional Poverty Index (MPI): Unstacking global poverty: Data for high impact action. New York. https://hdr.undp.org/content/2023-global-multidimensional-poverty-index-mpi#/indicies/MPI

Vaidyanathan, A. (2001). Poverty and Development Policy. *Economic and Political Weekly, 36*(21), 1807–1822. http://www.jstor.org/stable/4410661

Wijaya, Y. A., Kurniady, D. A., Setyanto, E., Tarihoran, W. S., Rusmana, D., & Rahim, R. (2021). Davies Bouldin index algorithm for optimizing clustering case studies mapping school facilities. *TEM Journal*, 1099-1103. https://doi.org/10.18421/tem103-13

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier. https://doi.org/10.1016/c2009-0-19715-5

World Bank. (2023a). World Bank East Asia and the Pacific Economic Update, October 2023: Services for Development. Washington, DC: World Bank. https://doi.org/10.1596/40383

World Bank. (2023b). World Development Report 2023: Migrants, Refugees, and Societies. Washington, DC : World Bank. https://openknowledge.worldbank.org/handle/10986/39696

Zhang, H., & Dai, J. (2023). Poverty improvement policies and household income: Evidence from China. *Heliyon, 9*(11), e21442. https://doi.org/10.1016/j.heliyon.2023.e21442