MULTIDISCIPLINARY SCIENCE JOURNAL

RESEARCH ARTICLE
Published Online: September 12, 2024
https://doi.org/10.31893/multiscience.2025123

Uncovering patterns of disadvantage in student socioeconomic data using k-means clustering and association rule mining

Pita Jarupunphol[a] | Chanudda Sumranpit[b] | Mongkhon Thakong[b]

[a]Digital Technology Program, Phuket Rajabhat University, Thailand.
[b]Data Science and Information Technology Program, Udon Thani Rajabhat University, Thailand.

Abstract This study employs a combination of k-means clustering and association rule mining to analyze socioeconomic data from 507 students at a community school in Udon Thani Province, Thailand, highlighting the pronounced economic inequalities impacting educational opportunities. By integrating these data mining techniques, the research categorizes students into four income-based clusters and identifies critical socioeconomic attributes influencing educational access and quality. The application of k-means clustering revealed four distinct economic categories within the student population, with the lowest income cluster significantly lacking in basic amenities such as air conditioning and agricultural vehicles, which are essential for their living and educational environments. Association rule mining was then applied to identify household attributes linked to the 'very low' income classification. Fourteen association rules were discovered, with support values ranging from 0.43 to 0.44, confidence levels between 0.81 and 0.82, and lift values consistently at 1.1. Key findings reveal that the absence of basic amenities such as air conditioning, agricultural vehicles, computers, and televisions are strongly associated with the lowest socioeconomic status. For instance, the rule {AirCondition=N, TV=N} => {Cluster=Very Low} had a support of 0.44, confidence of 0.81, and lift of 1.1, indicating households lacking these items have an 81% probability of being classified in the 'very low' income group. Notably, the presence of electricity did not correlate directly with higher income clusters, indicating that basic infrastructure access alone does not mitigate the broader socioeconomic challenges. These findings underscore the complex interplay between income levels and access to educational and basic living resources. The study's implications are profound, advocating for targeted educational policies and interventions that address the nuanced needs of economically disadvantaged students, thereby enhancing their educational outcomes and socioeconomic mobility. Future research directions include broadening the dataset to multiple schools across different regions and conducting longitudinal studies to reveal long-term effects of socioeconomic factors on education.

Keywords: k-means, association rule mining, socioeconomic data, educational inequality, poverty clustering

1. Introduction

Educational inequality remains a critical and persistent challenge worldwide, especially in areas where economic disparities significantly influence educational opportunities and outcomes (Thapa, 2022). These disparities often determine the quality and extent of educational resources available to students, thus creating a gap that can profoundly affect lifelong learning and earning potential (Ruswa & Gore, 2022). This disparity constrains individual growth and perpetuates a cycle of poverty that impedes socioeconomic development on a broader scale. In this case, economic status is a critical determinant of educational opportunities and outcomes, affecting lifelong learning and earning potential. Students in affluent areas are often better equipped with supportive facilities that are critical for comprehensive education. Conversely, students in poorer areas might struggle with basic infrastructure needs, which can profoundly impact student attendance and engagement (Nisle & Anyon, 2022). The divergence in educational outcomes across various socioeconomic strata is markedly apparent in regions characterized by pronounced economic disparities. Students from economically disadvantaged backgrounds typically encounter limited access to essential supportive resources. The deficiency in supportive living resources among lower socioeconomic groups is not merely a matter of physical materials but extends to encompass less exposure to enriching educational experiences and fewer opportunities for advanced learning that could enhance learning outcomes (Nisle & Anyon, 2022; Thapa, 2022).

The role of data mining in analyzing and predicting socioeconomic outcomes has been increasingly recognized in academic and practical fields (Alsharkawi et al., 2021; Chakravarty & Majumder, 2005; Huang & Xia, 2023). The literature

 Received: July 16, 2024 | Accepted: August 17, 2024

indicates that machine learning can provide significant insights into complex social phenomena such as poverty, which are often influenced by a multitude of interdependent factors. In educational contexts, these algorithms have been used to predict student performance, dropout rates, and resource allocation needs. However, the application to poverty clustering specifically within student populations remains less explored, highlighting a novel aspect of this research. Employing a combination of data mining techniques in this context can leverage the ability to uncover complex patterns that are not readily apparent through conventional analytical methods. Schools and educational policymakers can more efficiently and accurately target interventions and support the students who need them most, bridging the gap between the capabilities of current educational support systems and the real, often nuanced, economic needs of the student population.

1.1. Data Mining in Socioeconomic

This section critically examines prior research that has implemented data mining techniques to address socioeconomic challenges within educational environments. For example, a study by Mwangi et al. (2020) in Kenya used logistic regression to identify poverty indicators. Specifically, Mohamed Nafuri et al. (2022) employed a clustering approach to categorize B40 students in higher education institutions (HEIs), thereby aiming to support governmental initiatives to decrease dropout rates, increase graduation rates, and improve students' socioeconomic status. Additionally, Rajagukguk et al. (2022) utilized a clustering-based analysis of student data, effectively deriving insights to increase student learning outcomes and determining that two clusters were optimal for grouping students on the basis of the analyzed data. Furthermore, Li (2022a) investigated the relationship between students' consumption behaviors and their family economic situations via cluster analysis. This study revealed how these behaviors reflect underlying economic conditions. Concurrently, Li (2022b) demonstrated the feasibility of employing clustering attributes such as consumption intensity and frequency to assist school funding departments in targeted poverty alleviation efforts, effectively distinguishing between "hidden poor" and "fake poor" students.

With respect to applying association rule mining in an academic setting, Sriurai & Nuanmeesri (2024) focused on developing association rules for analyzing student performance via the FP-growth algorithm. This research notably contributes to a prototype for a student performance analysis system, enhancing academic excellence through a mobile application while considering ethical aspects. Moreover, Wang et al. (2022) addressed the challenges in analyzing large, diverse student behavioral data with an improved association rule mining algorithm. This enhancement increased mining efficiency and facilitated a deeper understanding of the link between student behaviors and academic performance. Additionally, Shatnawi et al. (2021) implemented an association rule mining system that successfully generated rules aiding course selection for students, with noteworthy outcomes regarding precision and recall in course predictions.

While clustering is commonly employed to categorize students on the basis of various attributes, applying association rule mining to specifically identify factors contributing to low income among students is novel. In this case, this study aims to apply a hybrid methodology that combines clustering and association rule mining. The objective is to pinpoint critical attributes associated with the lowest socioeconomic status among 507 students at a community school in economically challenging regions of Udon Thani Province, Thailand. This region is characterized by pronounced economic inequalities. This research is expected to effectively combine the clustering of low-income households with association rule mining to uncover and address the factors influencing students' economic challenges.

2. Materials and methods

The methodology encompasses a series of steps beginning with research design, then data collection and preparation, then data clustering and association rule mining, and finally, concluding with essential ethical considerations for the research.

2.1. Research Design

The research design utilized school-provided educational records enriched with socioeconomic information gathered through additional surveys, which are not typically part of school databases. This study adopts a quantitative approach, employing relevant data mining techniques for clustering student data. The aim is to segregate students into distinct groups on the basis of their income levels and average income. After these poverty categories are established, the analysis explores nonincome-related features within these groups. The method of association rule mining is applied to pinpoint features that are significant to particular clusters. This study encompasses 507 students from a community school in Udon Thani Province, known for its socioeconomic diversity, during the 2022--2023 academic year. The selection of this school was influenced by the extensive availability of student data and the notable level of economic challenges faced by its student body.

2.2. Data collection and preparation

Data were sourced predominantly from educational records supplied by the school, encompassing details on student familial backgrounds, socioeconomic factors recognized by the educational institution, and historical evaluations of student requirements. A supplementary survey was also constructed to capture additional socioeconomic information not

customarily recorded in school documentation. This included variables such as the sources and stability of household income, parental employment statuses and types, household composition and responsibilities, and the availability of educational resources at home. After data collection, the data underwent preparatory processing prior to the initiation of the data mining procedure. This research utilized several data preparation techniques to ensure the integrity and usability of the data for analysis (Witten et al., 2016). Specifically, data cleaning was undertaken to address or interpolate missing values, rectify inaccuracies, and standardize data formats suitable for subsequent analysis. In this case, we employed multiple imputation techniques to address missing values, which constituted approximately 5% of our dataset. For continuous variables such as income, we use the multiple imputation by chained equations (MICE) method to preserve the relationships between variables. For categorical variables, we applied mode imputation, replacing missing values with the most frequent category. In cases where more than 20% of a variable's data were missing, we excluded that variable from the analysis to maintain data integrity. Continuous variables were normalized to ensure that all features contributed equally to the clustering algorithm and to mitigate the impact of outliers. Categorical variables were encoded to convert them into a format suitable for our chosen algorithms. This increased the dimensionality of our dataset but preserved the categorical nature of these variables without imposing an ordinal relationship.

2.3. Data Clustering

Clustering techniques are essential tools in data analysis, allowing researchers to group a set of objects so that objects in the same group (or cluster) are more similar to each other than those in other groups. Among the numerous clustering methods available, some of the most commonly used include k-means clustering, hierarchical clustering, DBSCAN (density-based spatial clustering of applications with noise), and spectral clustering (Dol & Jawandhiya, 2023; Li et al., 2022; Witten et al., 2016). Each of these methods has specific applications and strengths depending on the nature of the data and the specific requirements of the analysis. For this research, k-means clustering is selected because of its effectiveness in dealing with numeric datasets (Heidari et al., 2024). K-means clustering is particularly suitable for scenarios where the attributes are numeric and continuous, as it relies on distance metrics (such as Euclidean or Manhattan distance) to determine the 'closeness' of instances (Zhang & Wu, 2024). For example, in the context of our study, variables such as 'income' and 'average income' are numeric and significantly impact socioeconomic analysis. By employing k-means clustering, the data can be segmented into meaningful clusters on the basis of income levels, facilitating a nuanced understanding of each group's economic attributes. Furthermore, the ability of k-means clustering to adapt to the intrinsic structure of data without presuming a cluster count makes it particularly valuable for exploratory data analysis. The objective of k-means clustering is to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a cluster prototype. The objective function to be minimized is as follows:

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \qquad (1)$$

where:
J is the total within-cluster sum of squares (WCSS).
k is the number of clusters.
S_i is the set of all data points assigned to the i-th cluster.
x represents a data point in cluster S_i.
where μ_i is the centroid or mean of the points in S_i.

Following the data preparation, the k-means clustering algorithm was implemented, utilizing the elbow method to ascertain the optimal number of clusters. The elbow method involves plotting the sum of squared distances from samples to their nearest cluster center and identifying a point on the curve that resembles an 'elbow' (Rylko et al., 2024). This point generally represents the stage at which the addition of further clusters does not markedly enhance the model's performance in data representation. The elbow method determines the optimal number of clusters by fitting the model with a range of values for k. The sum of the squared distances of samples to their nearest cluster center is plotted against the number of clusters. This sum of squared distances is the J value computed in the k-means objective function. The goal is to choose a small value of k that still has a low sum of squared distances, and the point where the rate of decrease sharply shifts (the elbow point) can be considered the optimal number of clusters. The elbow method looks for a k such that adding another cluster does not provide much better modeling of the data. After the application of k-means clustering to segment student household incomes, the data effectively categorized the economic diversity present within the student population. The next step involves the visualization of these data, which will facilitate an illustrative representation of the clustering insights.

2.4. Association rule mining

Following the clustering of student household income into designated categories, the subsequent phase of this research entails a detailed examination of the features that influence specific target categories via association rule mining. Association rule mining is used to find relationships between variables in large databases. It is frequently used in market basket analysis to discover combinations of items that frequently cooccur in transactions (Dam et al., 2022; Silva et al., 2019). The primary measures used to evaluate the importance of an association rule are support, confidence, and lift. These measures provide a way to quantify the strength and significance of the relationships discovered through association rule mining, helping to identify the most relevant and useful rules in a dataset.

1- Support measures the frequency or occurrence of an itemset in the dataset. For rule *A→B*, the support is computed as the proportion of transactions in the dataset that contain both *A* and *B*, expressed as *Support(A→B) = the number of transactions containing both A and B/the total number of transactions.*

2- Confidence measures the reliability of the rule's inference. For rule *A→B*, it is the proportion of transactions with item *A* that also contain item *B*, defined as *Confidence(A→B) = Support(A→B)/Support(A).*

3- Lift measures how much more often the antecedent and consequent of rule A→B occur together than we would expect if they were statistically independent. A lift value greater than 1 implies that the presence of *A* has a positive effect on the presence of B in transactions, calculated as *Lift(A→B) = Confidence(A→B)/Support(B)*.

This method is particularly valuable for uncovering hidden patterns that contribute to certain outcomes, in this case, the 'very low' income classification. The approach involves specifying parameters such as minimum support and confidence to generate meaningful and manageable rules. For this study, these thresholds are calibrated to produce approximately 10–15 rules. This range is strategically chosen to ensure that the rules are neither too sparse to miss critical insights nor too numerous to obfuscate meaningful interpretations. Additionally, the lift value, which compares the observed frequency of a rule to that which would be expected if the items were independent, is a critical criterion. A lift value greater than 1 is essential, as it indicates that the association between the features and the 'very low' income category is statistically significant, surpassing what would typically be anticipated under random conditions. The methodological decision to use association rule mining with these specific parameters allows for a focused analysis that identifies and validates the most salient features influencing the 'very low' income classification. This enables researchers not only to identify these key attributes but also to understand their interrelationships and the strength of their impact.

There are alternative methods to association rule mining, such as classification techniques (Hosseinzadeh & Edalatpanah, 2016; Witten et al., 2016). However, we did not consider them for the following reasons: 1) decision trees, although effective for classification, do not explicitly reveal associations between multiple attributes; 2) logistic regression, which is considered for predicting cluster membership, does not offer the same level of interpretable rules as association mining; and 3) neural networks, despite their potential to uncover complex patterns, lack the interpretability essential for our study's goals. Therefore, we chose association rule mining because of its ability to identify multiattribute relationships in a straightforward and interpretable manner, which is crucial for pinpointing key socioeconomic factors linked to low-income classifications. This combination of k-means clustering and association rule mining provides a balance of robust clustering, interpretable results, and the ability to uncover complex relationships in our socioeconomic data.

2.5. Ethical considerations

The research instrument was subjected to exhaustive ethical review by the Human and Animal Research Ethics Committee at Udon Thani Rajabhat University. This full board review, finalized on November 27, 2023, highlighted the essential ethical considerations necessary for research involving human subjects. The authorization, which remains valid until November 27, 2024, guarantees adherence to stringent ethical standards in research, specifically concerning participant consent, confidentiality, and data protection. Additionally, informed consent was secured from the parents or guardians of the participants, and strict confidentiality measures were implemented, with the data being anonymized before analysis. Only data directly relevant to the research questions were collected, minimizing the amount of sensitive information handled. The results are presented in aggregate form to prevent the identification of individual students. No individual cases or examples that could lead to identification are included in the research outputs. The participants were also informed that the data would be retained for a period of 180 days following the completion of the study and were briefed on the secure methods that would be employed for its destruction thereafter.

3. Results

This section presents the outcomes of the clustering analysis and the derived association rules, along with their implications. Additionally, the limitations of the current study are discussed, and directions for future research are proposed.

3.1. Clustering Results

To determine the optimal number of clusters for our data, we applied the k-means clustering algorithm via the elbow method. The elbow method involves plotting the sum of the squared distances of samples to their closest cluster center and looking for an 'elbow' in the curve. This elbow typically indicates a point where adding more clusters does not significantly improve the modeling of the data. In our analysis, we plotted such a curve and determined that four clusters were the best choice because adding more clusters beyond this number provided diminishing returns in terms of reducing the within-cluster sum of squares (WCSS). This selection suggests that four distinct groups adequately capture the variability in our dataset without unnecessary complexity. In this case, we choose K=4 for our k-means clustering method, which is based on the elbow method. Once the optimal cluster count was established at four, we proceeded to use these clusters to classify student income into four descriptive categories: very low, low, medium, and high. Each cluster represents a different income level, with the first cluster corresponding to 'very low' income, followed by 'low', 'medium', and 'high' for the subsequent clusters. Figure 1 shows the elbow plot representing a significant decrease in the cluster sum of squares up to K=4, after which the decrease became marginal. The silhouette analysis further supported this, which presented the highest average silhouette score at K=4, indicating optimal cluster separation and cohesion.

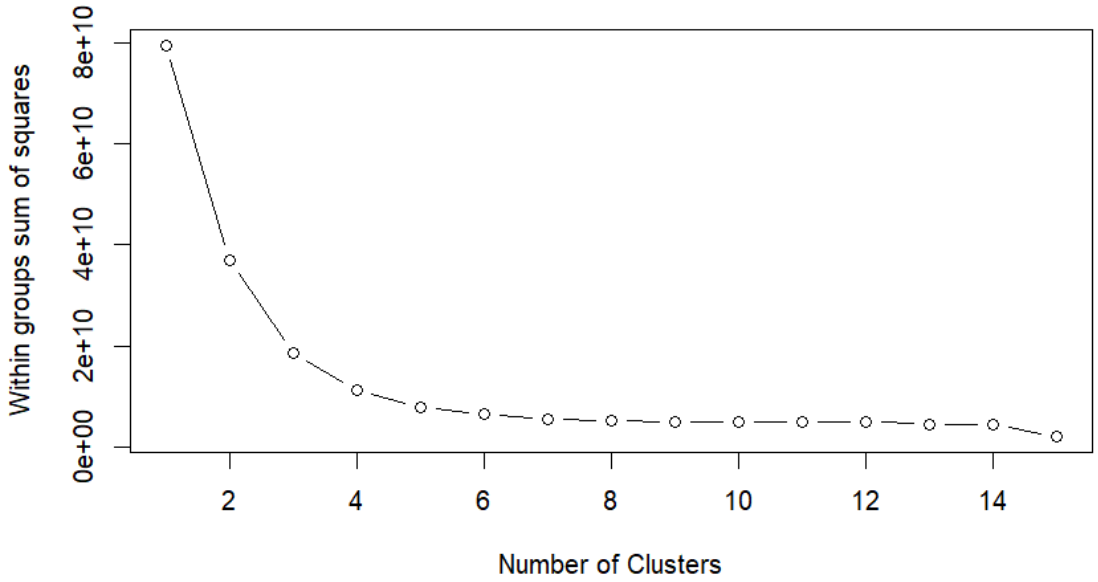

Figure 1 Elbow plot showing a notable reduction in the sum of squares within clusters.

In this analysis, the incomes of student households were segmented via a 4-cluster model developed through k-means clustering, effectively categorizing the economic diversity within the student population. The categorization is as follows:

Cluster 1 (Very Low): This cluster encapsulates households with an annual income of approximately 4,213 baht. It comprises the segment with the most economically disadvantaged students, highlighting a critical group for targeted financial aid and support programs.

Cluster 2 (Low): Households in this cluster have an average annual income of approximately 15,078 baht. This group represents students from low-income families who, while not as economically challenged as those in the first cluster, still require considerable financial support to ensure equal educational opportunities.

Cluster 3 (Medium): This cluster includes households with an average income of approximately 38,888 baht. These students fall into the middle-income category and are likely to experience fewer financial barriers than the first two clusters but still benefit from specific supportive measures to ensure comprehensive educational access.

Cluster 4 (High): The final cluster groups households with a significantly higher average annual income of approximately 113,273 baht. The students in this cluster represent a relatively small, high-income segment of the student body. Their economic advantages suggest a lesser need for financial assistance but provide a baseline for comparative studies on the impact of economic status on educational outcomes.

Figure 2(a) presents a comprehensive visualization of the distribution of student household incomes, segmented into four distinct groups through the application of k-means clustering. This figure effectively illustrates the classification of the entire dataset into economically differentiated clusters, highlighting the diversity of financial backgrounds within the student population. The visualization clearly depicts how student incomes are grouped on the basis of similarity in financial scales, offering a macroscopic view of economic stratification. In addition, Figure 2(b) provides a detailed graphical representation of these four income clusters, complete with delineated boundaries that distinctly separate each cluster. This figure is crucial for visualizing the spatial demarcation between groups, allowing observers to appreciate the relative distances and overlaps between clusters.

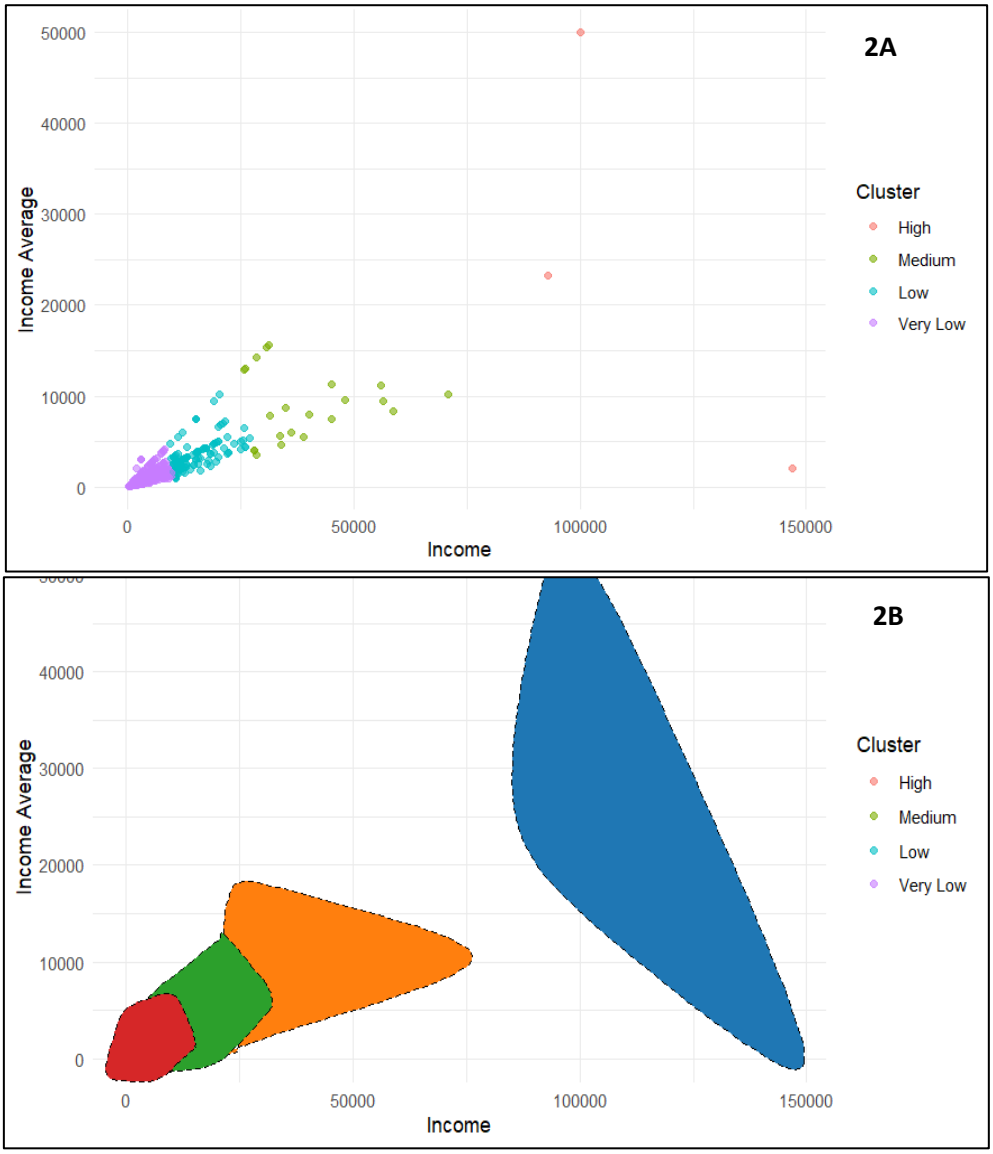

Figure 2 A graphical representation of these four income clusters. 2A: The distribution of student household incomes, segmented into four distinct groups. 2B: Four income clusters, complete with delineated boundaries separating each cluster.

3.2. Association Rules and Implications

We set the minimum support threshold to 0.43 and the minimum confidence to 0.81 on the basis of the following rationale: 1) support of 0.43 ensures that the rules apply to a substantial portion of our dataset (at least 43% of the cases), making them more generalizable; 2) confidence of 0.81 guarantees high reliability of the rules, indicating that when the antecedent occurs, the consequent follows in at least 81% of the cases. In this case, a total of fourteen association rules were identified that link specific household attributes to a 'very low' cluster designation, potentially indicative of lower socioeconomic status. This analysis reveals that conditions such as the absence of air conditioning, agricultural vehicles, computers, television, and a combination of these types of equipment robustly predict this socioeconomic classification. Notably, these rules are not only frequent within the dataset (as evidenced by their support values of approximately 0.44, indicating that 44% of the dataset meets these conditions) but also highly reliable, with confidence levels at or exceeding 0.81, suggesting an 81% probability that these conditions will indeed result in very low cluster classification.

Further dissection of these rules shows that each possesses a lift value slightly over 1.0 (e.g., 1.1), indicating that the observed association is more significant than what could be expected under random conditions. Such insights are crucial, as they highlight the specific attributes—such as the lack of basic amenities—that are strongly linked to the lower segments of the socioeconomic scale, as defined by the cluster analysis. Rules such as {AirCondition=N, TV=N} => {Cluster=Very Low} provide clear, actionable data points that can be leveraged in socioeconomic planning and targeted aid strategies. By understanding these patterns, policymakers and researchers can better address the disparities revealed by the data, focusing efforts and resources on communities characterized by these attributes (Table 1).

Table 1 Association rules linking household attributes to 'very low' income cluster.

Rule	lhs	rhs	support	conf	lift
1	{AirCondition=N, TV=N}	=> {Cluster=Very Low}	0.44	0.81	1.1
2	{AgriculturalVehicles=N, TV=N}	=> {Cluster=Very Low}	0.44	0.81	1.1
3	{Electricity=Y, TV=N}	=> {Cluster=Very Low}	0.44	0.81	1.1
4	{AgriculturalVehicles=N, AirCondition=N, TV=N}	=> {Cluster=Very Low}	0.43	0.82	1.1
5	{Computer=N, AirCondition=N,TV=N}	=> {Cluster=Very Low}	0.44	0.81	1.1
6	{Electricity=Y, AirCondition=N, TV=N}	=> {Cluster=Very Low}	0.44	0.82	1.1
7	{AgriculturalVehicles=N, Computer=N, TV=N}	=> {Cluster=Very Low}	0.44	0.81	1.1
8	{Electricity=Y, AgriculturalVehicles=N, TV=N}	=> {Cluster=Very Low}	0.44	0.81	1.1
9	{Electricity=Y, Computer=N, TV=N}	=> {Cluster=Very Low}	0.44	0.81	1.1
10	{AgriculturalVehicles=N, Computer=N, AirCondition=N, TV=N}	=> {Cluster=Very Low}	0.43	0.82	1.1
11	{Electricity=Y, AgriculturalVehicles=N, AirCondition=N, TV=N}	=> {Cluster=Very Low}	0.43	0.82	1.1
12	{Electricity=Y, Computer=N, AirCondition=N, TV=N}	=> {Cluster=Very Low}	0.44	0.82	1.1
13	{Electricity=Y, AgriculturalVehicles=N, Computer=N, TV=N}	=> {Cluster=Very Low}	0.43	0.81	1.1
14	{Electricity=Y, AgriculturalVehicles=N, Computer=N, AirCondition=N, TV=N}	=> {Cluster=Very Low}	0.43	0.82	1.1

Figure 3 provides a detailed visualization of the results from an association rule mining analysis, demonstrating the discovery of fourteen distinct association rules that correlate specific household attributes with a designation into the 'very low' income cluster. The identified rules exhibit substantial prevalence within the dataset and a high likelihood of accuracy in predicting the 'Very Low' income category. This figure is instrumental in illustrating how specific characteristics of households are strongly associated with the lowest socioeconomic status among the student population. By setting stringent criteria for support and confidence, the analysis ensures that the rules are statistically significant but also relevant and robust in their predictive power. Each rule represents a potential indicator of socioeconomic challenges, providing insights that can be crucial for developing targeted support strategies.

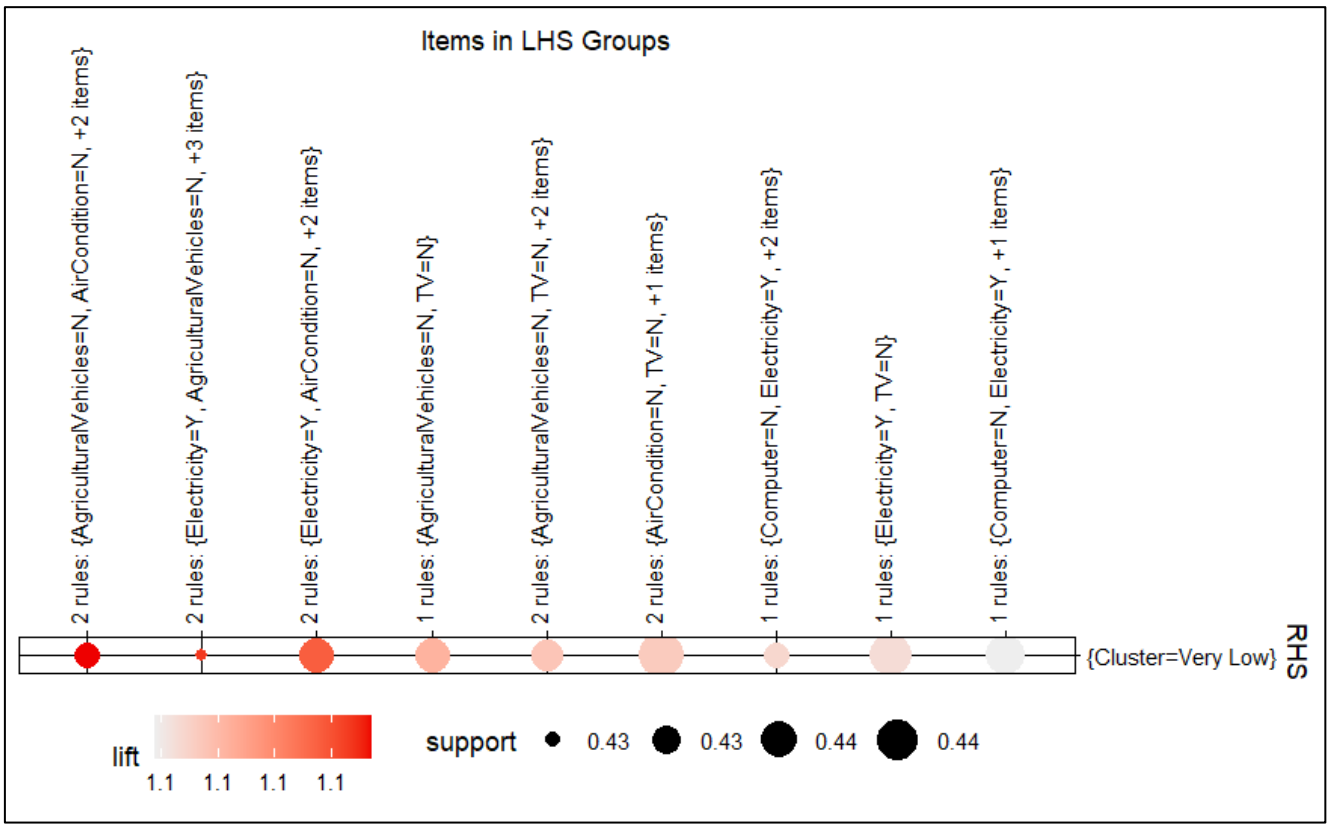

Figure 3 Association rule mining analysis, illustrating fourteen association rules contributing to the 'very low' income.

4. Discussion

The association rules discovered in this study provide valuable insights into the relationship between household attributes and socioeconomic status, particularly for students classified in the 'very low' income cluster. Each rule offers a unique perspective on the factors strongly associated with lower socioeconomic status, which can significantly impact educational outcomes. Table 2 presents selected examples of association rules and their educational implications for students.

Table 2 Examples of association rules and their implications for students.

Association Rules	Description	Implications for Students
{AirCondition=N, TV=N} => {Cluster=Very Low}	Households without air conditioning and television.	Limited disposable income; restricts access to educational programming and news, affecting study conditions and exposure to diverse information.
{AgriculturalVehicles=N, TV=N} => {Cluster=Very Low}	Households lacking agricultural vehicles and television.	Indicates limited means for agricultural productivity and information access; impacts educational resources and broader societal information exposure.
{Electricity=Y, TV=N} => {Cluster=Very Low}	Households with electricity but no television.	While basic infrastructure like electricity is available, the absence of a TV suggests financial constraints; affects access to multimedia educational resources.
{AgriculturalVehicles=N, AirCondition=N, TV=N} => {Cluster=Very Low} {Computer=N, AirCondition=N, TV=N} => {Cluster=Very Low}	Households missing multiple amenities: agricultural vehicles, air conditioning, TV, and possibly computers.	Significant resource constraints impact comfort, access to educational tools (especially digital), and potential family income from agriculture.
{Electricity=Y, AgriculturalVehicles=N, Computer=N, AirCondition=N, TV=N} => {Cluster=Very Low}	Presence of electricity but absence of multiple other amenities.	Basic electrical infrastructure does not compensate for the lack of tools for education and agriculture; poses multiple barriers to student learning and comfort.

These association rules collectively paint a picture of multidimensional poverty, where the absence of various amenities is strongly linked to the lowest income classification. From an educational perspective, these conditions can significantly impact a student's learning environment, access information, and overall educational opportunities. The lack of computers and televisions may limit exposure to diverse information sources and hinder the development of digital literacy skills, which are increasingly crucial in modern education (Saripudin et al., 2020). The absence of air conditioning, particularly in hot climates, could affect study comfort and concentration. The lack of agricultural vehicles in rural areas might indicate reduced family income potential, which could indirectly affect a student's educational resources and opportunities. These findings align with previous research highlighting the significant impact of socioeconomic status on educational outcomes and opportunities (Ruswa & Gore, 2022; Thapa, 2022). Understanding these associations is crucial for educators and policymakers, as emphasized by Nisle & Anyon (2022), who stress the importance of addressing socioeconomic barriers to education as part of comprehensive policy strategies. This highlights the need for targeted interventions that address immediate educational needs and consider the broader socioeconomic factors affecting students' home environments. Such interventions might include providing access to digital learning resources at school, developing programs to improve home study environments, or implementing community-based initiatives to increase overall household resources.

4.1. Limitations and generalizability

The association rules discovered in this study indicate strong correlations between specific household attributes and classification within the 'very low' income cluster. Although these correlations hold statistical significance, they necessitate cautious interpretation to avoid conflating correlations with causation, a common challenge in socioeconomic research (Chakravarty & Majumder, 2005). The identified patterns suggest possible bidirectional relationships and intricate interactions between household attributes and income levels, which merit further investigation. For example, the relationship between the absence of agricultural vehicles and low income is particularly noteworthy in rural or agricultural settings. This condition could stem from either limited income, which hampers investment in agricultural vehicles, thereby reducing agricultural productivity, or, conversely, the lack of these vehicles, which could restrict farming efficiency and income generation. Additionally, the prevalence of electricity in low-income households suggests effective government or community efforts in basic infrastructure provision, but the coexistence of limited amenities indicates persistent financial challenges impacting living standards. These findings lay the groundwork for more focused research and inform policy deliberations, although causal assertions should be cautiously made and backed by additional evidence. Moreover, the study's focus on a single community school in Udon Thani Province could limit the representativeness of the findings across diverse socioeconomic landscapes in Thailand, a limitation noted in similar studies (Blimpo et al., 2018). The specific characteristics of the students at this school might differ from those of a broader population, potentially skewing the results. Furthermore, the reliance on self-report survey data introduces the risk of biases, such as underreporting or overreporting due to social desirability or misunderstandings, a common challenge in socioeconomic research (Mwangi et al., 2020). The sample size, while significant for the context of one school, may not be sufficient for broader, more generalizable conclusions that are applicable nationally or internationally.

4.2. Future Research

Future research should aim to broaden the dataset to include multiple schools across different regions and socioeconomic contexts, not only within Thailand but also globally. This would enhance the generalizability of the findings and enable an exploration of how regional and cultural differences impact educational outcomes related to socioeconomic status (Blimpo et al., 2018; Nisle & Anyon, 2022). The implementation of longitudinal studies could further deepen this understanding by revealing the long-term effects of socioeconomic factors on education and the efficacy of various interventions over time. These studies could track changes in socioeconomic status and educational achievements, providing a dynamic view of the influences impacting educational outcomes.

Additionally, incorporating qualitative research methods, such as interviews and focus groups, could provide deeper insights into the personal experiences of students and families facing socioeconomic challenges. This approach would uncover nuanced factors that quantitative methods might miss, offering a more comprehensive view of the socioeconomic dynamics at play. Research could also explore the specific impacts of technological access on education, assessing how varying levels of access to digital tools affect learning, especially in remote or underserved communities. Comparative studies across different countries or socioeconomic models would further enrich our understanding by highlighting effective educational strategies that could be adapted across different settings, helping to mitigate the educational disparities driven by socioeconomic factors (Mwangi et al., 2020).

Future research on the socioeconomic impacts on education could also greatly benefit from the integration and development of advanced data mining techniques. Ensemble learning methods such as random forests and gradient boosting can enhance the robustness and accuracy of predictions by combining multiple models, offering a richer understanding of the complex factors influencing educational outcomes (Li et al., 2022; Niu et al., 2020). Deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can analyze spatial and temporal data to capture the evolving dynamics of socioeconomic conditions and identify regional disparities and changes over time (Baruah & Organero, 2024; Tian et al., 2023). Network analysis could also play a crucial role in examining the interconnections among various socioeconomic factors, highlighting key influencers and clusters that significantly impact educational outcomes. Moreover, developing hybrid models that merge different data mining approaches could yield comprehensive analytical tools. For example, combining the interpretative ease of association rule mining with the predictive power of neural networks could provide both deep insights and understandable rules that detail complex socioeconomic interactions affecting education.

5. Conclusions

This study's application of k-means clustering and association rule mining has revealed critical socioeconomic attributes that correlate with lower income statuses among students in Udon Thani Province. The findings highlight significant challenges such as the lack of basic amenities, which are strongly associated with the lowest income brackets. These challenges pose considerable barriers to achieving equitable educational outcomes. For example, schools should prioritize resource allocation to ensure that students from very low-income clusters have access to essential educational tools, including technology. In addition, the public and private sectors should collaborate to enhance basic infrastructure in underprivileged areas, ensuring that all students have access to necessary amenities such as electricity and air conditioning. Policymakers should use the insights from this study to formulate targeted educational policies that address the specific needs identified within low-income clusters. This could include subsidies for educational materials or transportation for students from economically disadvantaged backgrounds.

This research contributes to the broader discourse on socioeconomic disparities in education by providing a methodological framework that other researchers can adapt and apply in similar contexts. The findings underscore the importance of addressing socioeconomic barriers to education as part of comprehensive policy and community engagement strategies aimed at fostering educational equity. By continuing to explore and address these critical issues, stakeholders can better support the educational advancement of all students, irrespective of their economic background.

Acknowledgment

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

Ethical considerations

I confirm that I have obtained all the 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 that they have no conflicts of interest.

Funding

This research did not receive any financial support.

References

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

Baruah, R. D., & Organero, M. M. (2024). Explicit context integrated recurrent neural network for applications in smart environments. *Expert Systems with Applications, 124752*. https://doi.org/10.1016/j.eswa.2024.124752

Blimpo, M. P., Evans, D. K., & Lahire, N. (2018). Parental human capital and effective school management: Evidence from The Gambia. *World Bank Policy Research Working Paper* (8628). https://doi.org/10.1596/1813-9450-8628

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

Dam, K. H. T., Given-Wilson, T., Legay, A., & Veroneze, R. (2022). Packer classification based on association rule mining. *Applied Soft Computing, 127*, 109373. https://doi.org/10.1016/j.asoc.2022.109373

Dol, S. M., & Jawandhiya, P. M. (2023). Classification technique and its combination with clustering and association rule mining in educational data mining—A survey. *Engineering Applications of Artificial Intelligence, 122*, 106071. https://doi.org/10.1016/j.engappai.2023.106071

Heidari, J., Daneshpour, N., & Zangeneh, A. (2024). A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers. *Pattern Recognition, 155*, 110639. https://doi.org/10.1016/j.patcog.2024.110639

Hosseinzadeh, A., & Edalatpanah, S. A. (2017). Classification techniques in data mining: Classical and fuzzy classifiers. In A. Adak, D. Manna, & M. Bhowmik (Eds.), *Emerging research on applied fuzzy sets and intuitionistic fuzzy matrices* (pp. 153-188). IGI Global. https://doi.org/10.4018/978-1-5225-0914-1.ch007

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

Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., & Geng, Y. (2022). Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research, 276*, 106238. https://doi.org/10.1016/j.atmosres.2022.106238

Li, S. (2022a). Cluster analysis based on HDWA-Kmeans algorithm assisted identification of poor students. In *Proceedings of IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China. 593-597. https://doi.org/10.1109/ICAICA54878.2022.9844515

Li, S. (2022b). Cluster analysis of students' consumption behavior based on K-means++ algorithm. In *Proceedings of International Conference on Information System, Computing and Educational Technology (ICISCET)*, Montreal, QC, Canada. 154-158. https://doi.org/10.1109/ICISCET56785.2022.00047

Li, Z., Yoon, J., Zhang, R., Rajabipour, F., Iii, W. V. S., Dabo, I., & Radlińska, A. (2022). Machine learning in concrete science: Applications, challenges, and best practices. *Npj Computational Materials, 8*(1), 1–17. https://doi.org/10.1038/s41524-022-00810-x

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences, 12*(19), 9467. https://doi.org/10.3390/app12199467

Mwangi, A. W., Otieno, J., & Ndiritu, J. (2020). A comparison of the socio-economic status of female-headed and male-headed households in Kenya: Use of ordinal logistic regression. *African Journal of Physical Sciences, 4*, 1-22. https://core.ac.uk/download/pdf/328007819.pdf

Nisle, S., & Anyon, Y. T. (2022). An exploration of the relationship between school poverty rates and students' perceptions of empowerment: Student-staff relationships, equitable roles, & classroom sense of community. *Applied Developmental Science, 27*(3), 269-284. https://doi.org/10.1080/10888691.2022.2061490

Niu, T., Chen, Y., & Yuan, Y. (2020). Measuring urban poverty using multi-source data and a random forest algorithm: A case study in Guangzhou. *Sustainable Cities and Society, 54*, 102014. https://doi.org/10.1016/j.scs.2020.102014

Rajagukguk, S. A., & Fudholi, D. H. (2022). Rich kid, poor kid: Students clustering using K-prototype algorithm. In *Proceedings of IEEE International Conference on Information Technology and Digital Applications (ICITDA)*, Yogyakarta, Indonesia. 1-4. https://doi.org/10.1109/ICITDA55840.2022.9971445

Ruswa, A. S., & Gore, O. (2022). Student poverty in South African universities: Promoting the wellbeing and success of students. *Perspectives in Education*, 4-18. https://doi.org/10.38140/pie.v40i4.6379

Rylko, N., Stawiarz, M., Kurtyka, P., & Mityushev, V. (2024). Study of anisotropy in polydispersed 2D micro and nano-composites by elbow and K-means clustering methods. *Acta Materialia, 276*, 120116. https://doi.org/10.1016/j.actamat.2024.120116

Saripudin, A., Wahyudin, D., Hasanah, U., & Supendi, P. (2020). Vocational school teachers' perceptions of e-learning during COVID-19 pandemic. *Jurnal Pendidikan Vokasi, 10*(2), 196-208. https://doi.org/10.21831/jpv.v10i2.32771

Shatnawi, R., Althebyan, Q., Ghaleb, B., & Al-Maolegi, M. (2021). A student advising system using association rule mining. *International Journal of Web-Based Learning and Teaching Technologies, 16*(3), 65-78. https://doi.org/10.4018/IJWLTT.20210501.OA5

Silva, J., Varela, N., Borrero Lopez, L. A., & Rojas Millan, R. H. (2019). Association rules extraction for customer segmentation in the SMEs sector using the Apriori algorithm. *Procedia Computer Science, 151*, 1207–1212. https://doi.org/10.1016/j.procs.2019.04.173

Sriurai, W., & Nuanmeesri, S. (2024). The development of association rules for student performance analysis using FP-Growth algorithm as a guideline for multidisciplinary learning. *Journal of Applied Research on Science and Technology (JARST), 23*(1), Article 1. https://doi.org/10.60101/jarst.2023.253807

Thapa, Y. (2022). Effect of poverty on students' participation in the class at community schools in Nepal. *Orchid Academia Siraha, 1*(1), 111-124. https://doi.org/10.3126/oas.v1i1.52154

Tian, F., Hu, G., Yu, S., Wang, R., Song, Z., Yan, Y., Huang, H., Wang, Q., Wang, Z., & Yu, Z. (2023). An efficient multi-task convolutional neural network for

dairy farm object detection and segmentation. *Computers and Electronics in Agriculture, 211*, 108000. https://doi.org/10.1016/j.compag.2023.108000

Wang, T., Xiao, B., & Ma, W. (2022). Student behavior data analysis based on association rule mining. *International Journal of Computational Intelligence Systems, 15*(1), 32. https://doi.org/10.1007/s44196-022-00087-4

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. https://doi.org/10.1016/c2009-0-19715-5

Zhang, W., & Wu, Z. (2024). E-commerce recommender system based on improved K-means commodity information management model. *Heliyon, 10*(9), e29045. https://doi.org/10.1016/j.heliyon.2024.e29045