

Clustering Language Learning Strategy Patterns in Children with Autism Using K-Prototypes

Akhlis Munazilin¹, Fidy Ayu Pamestri^{*2}

¹Department of Computer Science, Faculty of science and technology, Universitas Ibrahimy, Situbondo, Indonesia

²Department of Informatics, Faculty of Computer Science, Universitas Amikom Purwokerto, Indonesia
e-mail: akhlismunazilin@gmail.com¹, fidyaayu98@gmail.com²

Article History	Abstract
Received : dd Month yyyy Revised : dd Month yyyy Accepted : dd Month yyyy Published : dd Month yyyy Keywords: <i>autism spectrum disorder, language learning strategy, K-Prototypes clustering, exploratory analysis, characteristic patterns.</i>	Children with Autism Spectrum Disorder (ASD) have diverse language learning needs; however, appropriate learning strategies are often selected without considering individual characteristics. This study analyzes language learning strategy patterns for children with ASD using K-Prototypes Clustering. The dataset consisted of 57 records with features including age, gender, speaking ability, severity level, learning media, and learning difficulty. K-Prototypes was selected because it can handle numerical and categorical data simultaneously. The optimal number of clusters was determined using the elbow method based on the cost value. The results produced four clusters with 18, 14, 14, and 11 data points, respectively. Cluster stability assessed using the Adjusted Rand Index (ARI) showed an average value of 0.6838 (SD = 0.1734), indicating adequate stability. Chi-Square and Cramér's V tests showed significant associations between clusters and learning difficulty (V = 0.5496), speaking ability (V = 0.4406), and severity level (V = 0.4182), but not learning strategy (V = 0.1445, p = 0.1845). Learning strategy distributions also showed different dominant patterns across clusters. These findings provide an initial exploratory analysis of language learning strategy patterns among children with ASD.

^{*}Corresponding author : fidyaayu98@gmail.com

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a group of neurodevelopmental conditions characterized by difficulties in social interaction and communication, accompanied by restricted and repetitive patterns of behavior, interests, or activities [1]. Data from the Centers for Disease Control and Prevention (CDC) show that the prevalence of ASD among 8-year-old children in the United States has continued to increase and is estimated at approximately 1 in 31 children in the latest ADDM Network report [2]. In addition to behavioral characteristics, ASD also affects children's learning ability, communication, and social participation; therefore, educational support needs to consider variations in individual needs [3], [4].

One important aspect in children with ASD is language development. Language profiles among children with ASD are highly diverse, ranging from non-verbal and semi-verbal children to verbal children with varying pragmatic abilities [5]. This diversity indicates that language learning strategies cannot be applied uniformly. Several intervention practices have been widely used, such as Applied Behavior Analysis (ABA), the Picture Exchange Communication System (PECS), visual support, and language intervention delivered by speech therapists [6]–[12]. Each approach has a different focus, for example, ABA focuses on shaping behavior and skills through reinforcement, PECS focuses on picture-exchange-based communication, visual support focuses on presenting information visually, and speech therapy focuses on developing communication and articulation abilities.

Previous studies have extensively discussed intervention and learning strategies for children with ASD. Hume et al. [6] and Steinbrenner et al. [7] summarized various evidence-based practices for children and adolescents with ASD, while Wong et al. [8] identified focused intervention practices with empirical support. Nevill et al. [5] showed that the language profiles of children with ASD can vary considerably based on the results of multiple assessment instruments. On the other hand, K-Prototypes was selected because the dataset contains both numerical and categorical variables. Unlike K-Means and K-Modes, which are primarily designed for numerical and categorical data, respectively, K-Prototypes integrates both types of variables within a single clustering framework, making it suitable for identifying clusters based on the mixed characteristics of children with ASD. [14]–[16]. However, most previous ASD studies have focused on intervention effectiveness, language profiles, and communication difficulties, while studies on mixed-data clustering have mainly focused on methodological development and the clustering of datasets containing numerical and categorical variables. These two approaches have generally not been combined to characterize children with ASD based on mixed numerical and categorical characteristics and subsequently examine the distribution of language learning strategies across the resulting clusters. Therefore, this study contributes by applying K-Prototypes to identify characteristic clusters of children with ASD and by describing the distribution of language learning strategies within each cluster.

This study aims to analyze patterns in the characteristics of children with ASD using K-Prototypes Clustering and to explore the language learning strategies that most frequently appear in each cluster. This study is not intended to determine the most effective strategy, but rather to provide an initial data-driven exploratory overview. Therefore, the findings are expected to serve as preliminary considerations in special education discussions, while still requiring validation from teachers, therapists, special education experts, and more comprehensive intervention effectiveness data.

2. METHOD

2.1 Research Flow

The research flow used in this study is shown in Figure 1. The study consisted of eight main stages: data collection, data understanding and preprocessing, exploratory data analysis, clustering data preparation and cluster number selection using the elbow method, K-Prototypes modeling, stability and association testing, analysis of learning strategy patterns per cluster, and conclusion drawing.

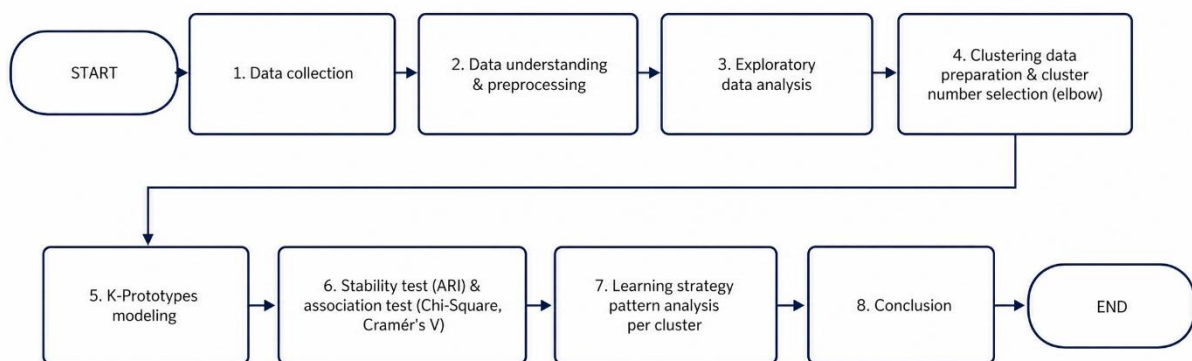


Figure 1. Research Flow

2.2 Dataset

The dataset used in this study is primary data collected through an online survey form completed by parents or caregivers of children with Autism Spectrum Disorder (ASD). Based on the information in the timestamp column, the data collection period lasted from 28 April 2026 to 30 May 2026. The dataset consisted of 57 records of children with ASD aged 5–18 years with an average age of 11.86 years. Overall, the data were dominated by boys, with 47 records (82.46%), while girls accounted for 10 records (17.54%). This condition is consistent with ASD epidemiological data indicating that ASD prevalence is higher among boys than girls [2].

The dataset contains information about children's characteristics and the language learning strategies used. The initial dataset columns consisted of timestamp, age, gender, speaking ability, ASD severity level, learning media, language learning difficulty, and language learning strategy. In this study, the `strategi_belajar` variable was not used as a feature for forming clusters, but was analyzed after the clusters were formed to identify the language learning strategy patterns that appeared most frequently in each group. Thus, the clustering process focused on children's characteristics, while learning strategies were used as the basis for interpreting patterns after the grouping process was completed.

Data ethics were considered by limiting the use of personally identifiable information. The dataset did not contain children's names, addresses, contact numbers, or other personal identifiers that could refer to specific individuals. Therefore, the data were analyzed anonymously, and the results were presented in aggregate form through statistical summaries, distribution tables, and cluster patterns. The data used in this study were limited to academic purposes and were not intended for direct clinical or pedagogical decision-making without validation from relevant experts. Details of the dataset columns are presented in Table 1.

Table 1. Description of Dataset Columns

Variable	Data type	Value	Description
timestamp	Administrative	28 April 2026–30 May 2026	Date and time of submission; removed during preprocessing because it was not used in clustering.
Age	Numerical	5–18 years; average 11.86 years	Age of children with ASD in years.
Gender	Categorical	Male: 47 (82.46%); Female: 10 (17.54%)	Children's demographic characteristics.
Speaking ability	Categorical	Semi-verbal: 22 (38.60%); Non-verbal: 19 (33.33%); Verbal: 16 (28.07%)	Children's speaking ability profile.
Severity level	Categorical	Moderate: 24 (42.11%); Mild: 17 (29.82%); Severe: 16 (28.07%)	Severity level of ASD condition based on survey data.
Learning media	Categorical	Visual: 44 (77.19%); Audio: 9 (15.79%); Audiovisual: 4 (7.02%)	The most frequently used learning media.
Learning difficulty	Categorical	Pronunciation: 25 (43.86%); Comprehension: 17 (29.82%); Vocabulary: 15 (26.32%)	Main difficulty in language learning.
Learning strategy	Categorical	ABA: 25 (43.86%); Visual Method: 15 (26.32%); PECS: 10 (17.54%); Speech Therapy: 7 (12.28%)	Language learning strategies analyzed after the clusters were formed.

2.3 Preprocessing

The preprocessing stage was conducted to ensure that the data were ready for use in K-Prototypes. The processes included: (1) removing the timestamp column because it only served as a time marker for form submission; (2) standardizing column names into *usia*, *jenis_kelamin*, *kemampuan_bicara*, *tingkat_keparahan*, *media_belajar*, *kesulitan_belajar*, and *strategi_belajar*; (3) removing excess spaces in categorical values; (4) simplifying learning media categories into Visual, Audio, and Audiovisual; (5) simplifying strategy names into ABA, Visual Method, PECS, and Speech Therapy; (6) checking missing values and duplicate data; and (7) separating clustering variables from the *strategi_belajar* variable for post-clustering analysis. The inspection results showed that there were no missing values or duplicate data.

The age variable was standardized using *StandardScaler* to place the numerical feature on a comparable scale and prevent its magnitude from disproportionately influencing the dissimilarity calculation relative to categorical attributes. This scaling was applied because age is the only numerical variable used in the clustering process, while the remaining variables are categorical. Categorical variables were not converted into dummy variables because K-Prototypes can process categorical attributes directly. The features used to form clusters were *usia*, *jenis_kelamin*, *kemampuan_bicara*, *tingkat_keparahan*, *media_belajar*, and *kesulitan_belajar*.

2.4 K-Prototypes Clustering

K-Prototypes was selected because the research dataset consisted of a combination of numerical and categorical data. The use of K-Means was less suitable because it is designed only for numerical data, while converting all categorical data into numbers may create artificial ordinal meanings. K-Prototypes combines the principle of K-Means for numerical attributes and K-Modes for categorical attributes, making it more suitable for mixed-type data [14].

The dissimilarity function in K-Prototypes can be formulated as follows:

$$d(x, q) = \sum_{r=1}^{m_r} (x_{ir} - q_{lr})^2 + \gamma \sum_{c=1}^{m_c} \delta(x_{ic}, q_{lc}) \quad (1)$$

In the equation, the first term represents the Euclidean distance for numerical attributes, while the second term represents categorical dissimilarity through a mismatch function. The parameter γ is used to regulate the relative contribution between numerical and categorical attributes. In this study, the numerical attribute was *usia*, while the categorical attributes included *jenis_kelamin*, *kemampuan_bicara*, *tingkat_keparahan*, *media_belajar*, and *kesulitan_belajar*.

2.5 Selection of the Number of Clusters

The number of clusters was selected by testing *k* values from 2 to 6. This range was chosen to evaluate a practical number of clusters while avoiding solutions that were too small to provide meaningful group-level interpretation given the limited sample size. The optimal number of clusters was then determined based on the elbow point of the cost curve. Each *k* value was evaluated using the cost value, which represents the total dissimilarity between the data and the cluster centroids. The number of clusters was selected based on the cost reduction pattern in the elbow method [17], while also considering cluster size to avoid producing groups that were too small for descriptive analysis.

2.6 Cluster Stability Test

The cluster stability test was conducted using the Adjusted Rand Index (ARI), which measures the similarity between two cluster partitions after correcting for the chance of random agreement [18]. The model was rerun using ten *random_state* values, namely 0, 1, 7, 21, 42, 99, 123, 202, 404, and 999. These values were selected to provide a diverse set of random initializations and assess whether the resulting cluster assignments remained consistent across different random seeds. The selected seeds

were fixed and used consistently across repeated runs to ensure reproducibility of the stability analysis. The clustering result for each random_state was compared with the final model result. An ARI value close to 1 indicates increasingly similar cluster partitions, while values near 0 indicate agreement comparable to random partitioning [18]. In this study, an average ARI above 0.60 was used as a study-specific operational criterion for sufficient stability in exploratory analysis, reflecting a substantial level of agreement across repeated clustering runs. This criterion was not treated as a universal threshold, but was adopted to provide a consistent basis for interpreting the stability of the clustering results in this study.

2.7 Chi-Square and Cramér's V Association Test

The association test was conducted between the cluster labels and each categorical variable. The Chi-Square test was used to determine the significance of relationships between variables, while Cramér's V was used as a measure of association strength for categorical data [19], [20]. The interpretation of association strength used in this study was $V < 0.10$ as very weak, $0.10\text{--}0.29$ as weak, $0.30\text{--}0.49$ as moderate, and ≥ 0.50 as strong. In addition, expected frequencies were also examined because the Chi-Square test may be less stable when many cells have expected frequencies below 5 [20].

3. RESULTS AND DISCUSSION

3.1 Selection of the Number of Clusters

The cost testing results for $k=2$ to $k=6$ are presented in Table 2 and visualized in Figure 2. The cost value continued to decrease as the number of clusters increased, because more clusters allowed the data points to be closer to their respective centroids. However, the largest cost reduction occurred from $k=2$ to $k=4$, while after $k=4$ the reduction rate began to slow down. The decrease from $k=3$ to $k=4$ was still 14.74%, while the decrease from $k=4$ to $k=5$ dropped to 8.70% and from $k=5$ to $k=6$ to 6.20%.

Based on this pattern, $k=4$ was selected as the final number of clusters because it was located at the point before the cost reduction began to flatten. In addition, $k=4$ produced cluster sizes of 18, 14, 14, and 11 data points, allowing descriptive interpretation for each group. Larger k values produced smaller groups and could potentially reduce the stability of interpretation in a dataset consisting of only 57 records.

Table 2. Results of Selecting the Number of Clusters

Number of Clusters (k)	Cost Value	Cost Reduction (%)
2	72.1020	-
3	61.4051	14.84
4	52.3511	14.74
5	47.7972	8.70
6	44.8339	6.20

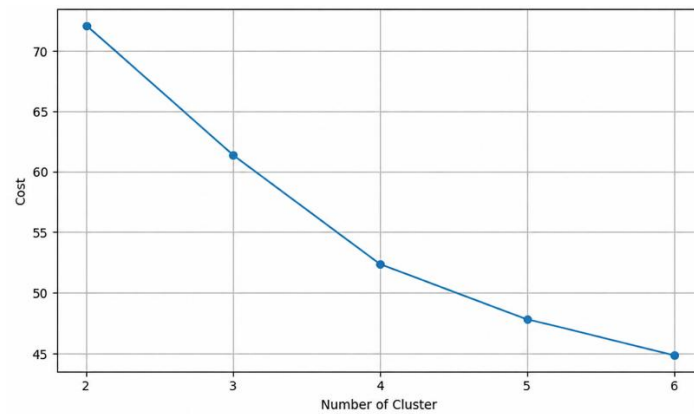


Figure 2. Elbow Method Graph for Selecting the Number of Clusters

3.2 Final Cluster Profile

The final K-Prototypes model with $k=4$ produced a final cost value of 52.3511. The distribution of cluster members showed that no single cluster strongly dominated the entire dataset, although the size of each cluster was not completely equal. The dominant characteristic profile of each cluster is presented in Table 3, while the distribution of cluster members is shown in Figure 3.

Table 3. Final K-Prototypes Cluster Profile ($k=4$)

Characteristic	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Number of records	18 (31.58%)	14 (24.56%)	14 (24.56%)	11 (19.30%)
Average age	8.50 years	14.00 years	15.57 years	9.91 years
Age range	5–10 years	11–18 years	14–18 years	7–11 years
Speaking ability	Semi-verbal	Semi-verbal	Non-verbal	Verbal
Severity level	Moderate	Moderate	Severe	Mild
Learning media	Visual	Visual	Visual	Visual
Learning difficulty	Pronunciation	Vocabulary	Pronunciation	Comprehension

The main differences among clusters were not only related to age, but also to speaking ability, severity level, and learning difficulties. Cluster 0 and Cluster 1 were both dominated by semi-verbal children with moderate severity, but differed in age range and type of learning difficulty: Cluster 0 was younger and predominantly experienced pronunciation difficulties, while Cluster 1 was older and predominantly experienced vocabulary difficulties. Cluster 2 showed a higher support-need profile because it was dominated by non-verbal children with severe severity and pronunciation difficulties. In contrast, Cluster 3 represented children with verbal ability and mild severity, but still showed challenges in language comprehension. Thus, the resulting clusters more clearly reflected variations in language development profiles and the level of support needed by children with ASD.

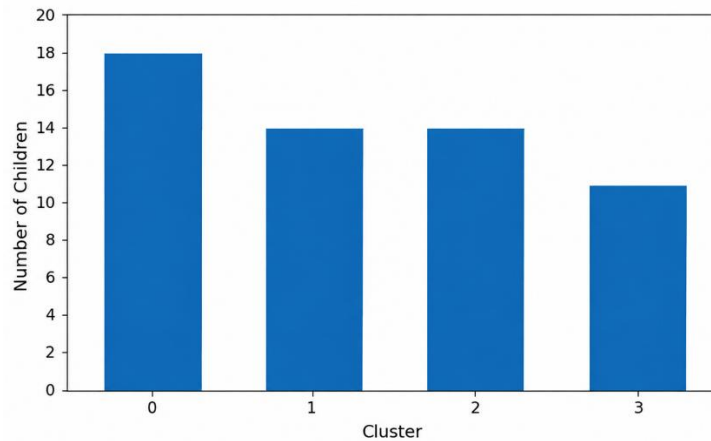


Figure 3. Distribution of Final Cluster Members

Figure 3 shows that Cluster 0 was the largest group with 18 records, while Cluster 3 was the smallest group with 11 records. The differences in cluster sizes were still acceptable for exploratory analysis because none of the clusters contained very few data points. This condition supported the interpretation of cluster patterns, although the results should still be interpreted cautiously because the overall dataset size was limited.

3.3 Cluster Stability Test

The stability test results using ARI are presented in Table 4. Because ARI measures the similarity between two partitions, with values approaching 1 indicating greater agreement [18], the average ARI of 0.6838 indicates that the cluster assignments showed a relatively consistent pattern across different random initializations. The standard deviation of 0.1734 suggests some variation between runs, indicating that the clustering structure was not completely invariant to initialization. Nevertheless, the average agreement remained sufficiently high to support the use of the resulting clusters for exploratory analysis. This interpretation is operational and study-specific rather than a universal threshold for clustering stability. The result therefore suggests that the identified clusters capture reasonably reproducible patterns in the data, while further validation using larger datasets and independent samples would be needed to establish stronger generalizability.

Table 4. Cluster Stability Test Results (Adjusted Rand Index)

Metric	Value
Tested random_state values	0, 1, 7, 21, 42, 99, 123, 202, 404, 999
Average ARI	0.6838
ARI standard deviation	0.1734
Interpretation	Sufficiently stable for exploratory analysis

The ARI standard deviation of 0.1734 indicates that there was still variation in the results when the model initialization changed. This variation is reasonable because the dataset was relatively small and several categories had low frequencies. Therefore, the clustering results can be used to observe pattern tendencies, but are not yet strong enough to serve as the basis for final recommendations without additional validation.

3.4 Chi-Square and Cramér's V Association Test

The association test results between clusters and each categorical variable are presented in Table 5 and visualized in Figure 4. This test was used to identify which variables contributed most to statistically differentiating the cluster profiles.

Table 5. Chi-Square and Cramér's V Association Test Results

Variable	Chi-Square	p-value	Cramér's V	Strength	Significant
Learning difficulty	39.3176	0.0000	0.5496	Strong	Yes
Speaking ability	27.4519	0.0001	0.4406	Moderate	Yes
Severity level	25.3295	0.0003	0.4182	Moderate	Yes
Learning media	8.7906	0.1857	0.1562	Weak	No
Learning strategy	12.5402	0.1845	0.1445	Weak	No
Gender	0.4809	0.9231	0.0000	Very weak	No

The results in Table 5 show that clusters had the strongest association with learning difficulty ($V = 0.5496$), followed by speaking ability ($V = 0.4406$) and severity level ($V = 0.4182$). These three variables were statistically significant, indicating that cluster formation primarily reflected differences in children's learning needs and communication ability profiles. In contrast, learning media, learning strategy, and gender did not show significant associations with the clusters.

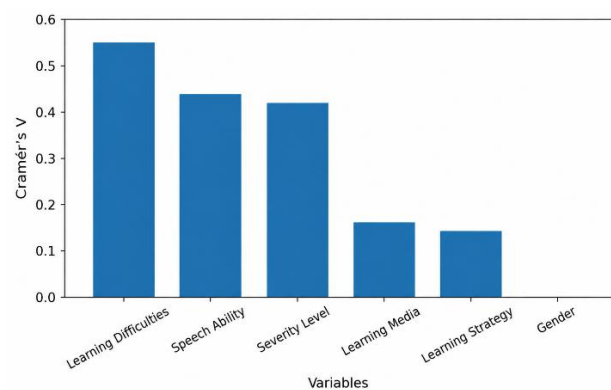


Figure 4. Association Strength (Cramér's V) Between Clusters and Each Variable

Figure 4 illustrates the relative strength of the associations between cluster membership and the variables based on Cramér's V. Learning difficulty showed the strongest association with cluster membership ($V \approx 0.55$), followed by speaking ability ($V \approx 0.44$) and severity level ($V \approx 0.42$). In contrast, learning media and learning strategy showed weaker associations, with Cramér's V values of approximately 0.16 and 0.14, respectively, while gender showed virtually no association. These results indicate that the clusters were primarily differentiated by learning difficulty, speaking ability, and severity level, whereas learning strategy was less strongly associated with cluster membership. In particular, the association between clusters and learning strategy was not statistically significant ($V = 0.1445$; $p = 0.1845$), with 13 expected frequencies below 5, requiring cautious interpretation of the Chi-Square result. Therefore, the figure provides a visual comparison of association strength, while the corresponding significance tests and exact statistical values are reported in Table 5. These findings support the exploratory nature of the study and should not be interpreted as evidence of learning strategy effectiveness.

3.5 Learning Strategy Patterns per Cluster

The distribution of learning strategies in each cluster is presented in Table 6, while the summary of the most frequently appearing strategies is shown in Table 7.

Table 6. Distribution of Learning Strategies per Cluster

Learning Strategy	Cluster 0 (n=18)	Cluster 1 (n=14)	Cluster 2 (n=14)	Cluster 3 (n=11)
ABA	6 (33.33%)	9 (64.29%)	6 (42.86%)	4 (36.36%)
Visual Method	6 (33.33%)	3 (21.43%)	4 (28.57%)	2 (18.18%)
PECS	4 (22.22%)	2 (14.29%)	0 (0.00%)	4 (36.36%)
Speech Therapy	2 (11.11%)	0 (0.00%)	4 (28.57%)	1 (9.09%)

Table 7. Dominant Strategy Patterns per Cluster

Cluster	Dominant Strategy	Percentage	Pattern Strength
Cluster 0	ABA + Visual Method	33.33% each	Combined / parallel pattern
Cluster 1	ABA	64.29%	Strong dominant pattern
Cluster 2	ABA	42.86%	Moderate dominant pattern
Cluster 3	ABA + PECS	36.36% each	Combined / parallel pattern

Tables 6 and 7 show that ABA was the strategy that appeared most frequently in several clusters, particularly in Cluster 1 with a percentage of 64.29%. Cluster 0 showed a combined pattern of ABA and the Visual Method, while Cluster 3 showed a combined pattern of ABA and PECS. Cluster 2 showed ABA as the strategy with the highest percentage, but the Visual Method and Speech Therapy also appeared with relatively close percentages. This pattern indicates variation in the strategies used across groups with different characteristics.

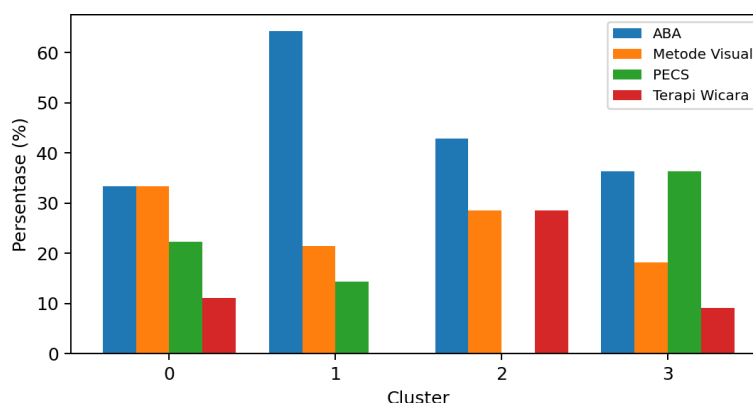


Figure 5. Percentage Distribution of Learning Strategies per Cluster

Figure 5 shows the differences in the percentage of learning strategies across clusters more clearly. The ABA strategy was more prominent in Cluster 1, while Cluster 0 and Cluster 3 showed more dispersed patterns. Because the association test between clusters and learning strategy was not significant, this graph was used as a descriptive visualization.

3.6 Discussion

This study was conducted to identify patterns in the characteristics of children with ASD and the distribution of language learning strategies based on mixed numerical and categorical data. The main results showed that K-Prototypes was able to form four clusters with profiles that could be differentiated based on age, speaking ability, severity level, and learning difficulty. This finding is consistent with the characteristics of ASD as a heterogeneous spectrum, particularly in terms of communication ability and learning support needs [1], [3], [5].

The association test results showed that cluster differences were more strongly explained by learning difficulty, speaking ability, and severity level than by learning strategy. In other words, the resulting clusters represented children's needs profiles more than the intervention choices used. This finding is important because previous studies on evidence-based practices for ASD have shown that various strategies such as ABA, PECS, visual support, and speech-language interventions can be beneficial [6]–[13], but their implementation still needs to consider children's individual characteristics and learning contexts.

The distribution of learning strategies per cluster showed that ABA appeared most frequently in several clusters, while other clusters showed combined patterns with the Visual Method or PECS. However, because *strategi_belajar* was not significantly associated with clusters, these dominant patterns cannot be interpreted as the most effective strategies. The results are more appropriately used as an initial overview of practices that appeared in the data, which can serve as a basis for discussion among teachers, therapists, or researchers in designing further data collection that includes children's language development outcomes. The limitations of this study include the small dataset size, data collection that was limited to a survey, and the absence of information regarding intervention duration, learning intensity, language development evaluation, and external validation on other datasets. Several categories also had low frequencies, which affected the reliability of the Chi-Square test. Future studies should increase the amount of data, include observational data from teachers or therapists, add indicators of strategy effectiveness, and validate clusters on independent datasets so that the results become stronger statistically and practically.

4. CONCLUSION

This study applied K-Prototypes Clustering to analyze characteristic patterns of children with ASD based on mixed numerical and categorical data. The clustering results produced four clusters with a final cost value of 52.3511 and cluster sizes of 18, 14, 14, and 11 records. The stability test showed an average ARI of 0.6838 (SD = 0.1734), indicating that the cluster results can be considered sufficiently stable for exploratory analysis. The association test showed that clusters were significantly associated with learning difficulty, speaking ability, and severity level, but not significantly associated with learning strategy. The distribution of learning strategies showed that ABA frequently appeared, particularly in Cluster 1, as well as combined patterns of ABA with the Visual Method or PECS in other clusters. However, these findings only describe dominant patterns in the data and cannot be interpreted as evidence of the effectiveness of learning strategies. Future studies are recommended to use larger datasets, add indicators of intervention effectiveness, and involve teacher or therapist observations and longitudinal measurements of language development.

REFERENCES

- [1] World Health Organization, "Autism," 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders>
- [2] Centers for Disease Control and Prevention, "Data and Statistics on Autism Spectrum Disorder," 2025. [Online]. Available: <https://www.cdc.gov/autism/data-research/index.html>
- [3] C. Lord et al., "The Lancet Commission on the future of care and clinical research in autism," *Lancet*, vol. 399, no. 10321, pp. 271–334, 2022, doi: 10.1016/S0140-6736(21)01541-5.
- [4] National Institute of Mental Health, "Autism Spectrum Disorder," 2024. [Online]. Available: <https://www.nimh.nih.gov/health/publications/autism-spectrum-disorder>
- [5] R. E. Nevill et al., "Language profiles in young children with autism spectrum disorder: A community sample using multiple assessment instruments," *Autism*, vol. 23, no. 1, pp. 141–153, 2019, doi: 10.1177/1362361317726245.
- [6] K. Hume et al., "Evidence-based practices for children, youth, and young adults with autism: Third generation review," *J. Autism Dev. Disord.*, vol. 51, pp. 4013–4032, 2021, doi: 10.1007/s10803-020-04844-2.
- [7] J. R. Steinbrenner et al., "Evidence-Based Practices for Children, Youth, and Young Adults with Autism," The University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Institute, Chapel Hill, NC, 2020. [Online]. Available: <https://ncaep.fpg.unc.edu/wp-content/uploads/EBP-Report-2020.pdf>
- [8] C. Wong et al., "Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review," *J. Autism Dev. Disord.*, vol. 45, no. 7, pp. 1951–1966, 2015, doi: 10.1007/s10803-014-2351-z.
- [9] M. Gitimoghaddam, B. Chichkine, P. McArthur, and A. G. Sangha, "Applied behavior analysis in children and youth with autism spectrum disorders: A scoping review," *Perspect. Behav. Sci.*, vol. 45, pp. 521–557, 2022, doi: 10.1007/s40614-022-00338-x.
- [10] A. S. Bondy and L. A. Frost, "The Picture Exchange Communication System," *Focus Autistic Behav.*, vol. 9, no. 3, pp. 1–19, 1994, doi: 10.1177/108835769400900301.
- [11] S. M. Rao and B. Gagie, "Learning through seeing and doing: Visual supports for children with autism," *Teach. Except. Child.*, vol. 38, no. 6, pp. 26–33, 2006, doi: 10.1177/004005990603800604.
- [12] A. V. Binns, R. Smyth, A. Andres, J. Lam, and J. O. Cardy, "Looking back and moving forward: A scoping review of research on preschool autism interventions in the field of speech-language pathology," *Autism Dev. Lang. Impair.*, vol. 6, pp. 1–16, 2021, doi: 10.1177/23969415211033171.
- [13] National Institute on Deafness and Other Communication Disorders, "Autism Spectrum Disorder: Communication Problems in Children," 2020. [Online]. Available: <https://www.nidcd.nih.gov/health/autism-spectrum-disorder-communication-problems-children>
- [14] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, 1998, doi: 10.1023/A:1009769707641.
- [15] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, 2013, doi: 10.1016/j.neucom.2013.04.011.
- [16] A. Ahmad and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," *IEEE Access*, vol. 7, pp. 31883–31902, 2019, doi: 10.1109/ACCESS.2019.2903568.
- [17] R. L. Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953, doi: 10.1007/BF02289263.
- [18] L. Hubert and P. Arabie, "Comparing partitions," *J. Classif.*, vol. 2, pp. 193–218, 1985, doi: 10.1007/BF01908075.
- [19] H. Akoglu, "User's guide to correlation coefficients," *Turkish J. Emerg. Med.*, vol. 18, no. 3, pp. 91–93, 2018, doi: 10.1016/j.tjem.2018.08.001.
- [20] A. Agresti, *An Introduction to Categorical Data Analysis*, 3rd ed. Hoboken, NJ: Wiley, 2019.