

Behavioral Profiling of Children with ASD Using Score-Based K-Modes

Dwi Purnomo¹, Shasy Kirana Syaharani^{*2}

¹Department of Computer Science, Faculty of Psychology, Business, and Technology, Universitas YPIB Majalengka, Indonesia

²Department of Informatics, Faculty of Computer Science, Universitas Amikom Purwokerto, Indonesia

e-mail: dwipurnomo@univypib.ac.id¹, shasykiranasyaharani99@gmail.com²

Article History	Abstract
Received: 27 June 2026 Revised: 25 August 2026 Accepted: 4 September 2026 Published: 16 September 2026	Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects communication, social interaction, and behavior in children. This study analyzes children's behavioral profiles using K-Modes Clustering based on indication scores derived from seven ASD-related behavioral indicators. The dataset consists of 50 children's records, processed through data cleaning, missing-value imputation, indication-score formation, KModes clustering, and feature-relationship analysis using Cramer's V. The clustering produced three behavioral profile groups: Low Indication (22%), Medium Indication (44%), and High Indication (34%). Association analysis shows the strongest relationship between speech ability and repetitive movements (Cramer's V = 0.548). These results indicate that K-Modes Clustering can serve as an initial exploratory approach for grouping observed behavioral indicator patterns, though the findings cannot yet be interpreted as diagnostic or clinical conclusions. This exploratory approach supports mapping the level of ASD-related behavioral indications to aid early, nonclinical screening in educational settings.
Keywords: Spectrum Disorder, K-Modes Clustering, Behavioral Profile, Indication Score, Data Mining.	

^{*}Corresponding author: shasykiranasyaharani99@gmail.com

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects an individual's communication skills, social interactions, and behavioral patterns. The characteristics of ASD are generally manifested through difficulties in making eye contact, responding to name calls, interacting with the social environment, and the emergence of repetitive or stereotyped behaviors. The wide variety of symptoms causes each child to exhibit different characteristics, so the identification process requires observing various behavioral indicators [1]. Early identification of behavioral characteristics associated with ASD is crucial because it can facilitate faster and more appropriate intervention. However, the initial identification process often requires lengthy observation and the involvement of professionals, making it difficult to implement in educational settings or for the general public. Various studies have shown that the use of machine learning technology can facilitate behavioral analysis and support more objective decision-making based on available data [2].

The development of data mining and machine learning has been widely utilized in ASD research. Han et al. successfully identified behavioral and neuroanatomical features associated with ASD using a machine learning approach to aid in the identification of characteristics in children with ASD [3]. Another study conducted by Song et al. developed a machine learning-based model to aid behavioral problems in children with ASD [5]. However, most of these studies rely on supervised, classlabeled models built from clinical or genomic data, leaving a gap in exploratory, unsupervised approaches suited to small-scale, categorical, teacher-observed behavioral datasets typically available in educational settings. This study addresses that gap by applying a score-based K-Modes clustering approach to such data.

Most previous studies have used a supervised learning approach that requires class labels as the learning target. However, behavioral data for children with ASD are often categorical and do not always contain complete diagnostic information, making class labels unavailable. This situation makes unsupervised learning approaches a relevant alternative. One method that can be used is clustering, a technique for grouping data based on the level of similarity in characteristics without requiring prior class labels [6].

Clustering approaches have also been used in ASD research to uncover phenotypic and biologically related subtypes [19], [20]. Given that the data used are predominantly categorical, the K-Modes Clustering method was chosen because it is specifically designed to handle categorical data [7]. Compared with alternatives such as K-Prototypes (which requires mixed numeric-categorical data), hierarchical clustering with Gower distance (computationally costly to re-run across many seeds), or latent class analysis (which assumes an underlying probabilistic model and needs larger samples for stable estimation), K-Modes offers a simpler, distribution-free, and computationally efficient option well suited to a small, purely categorical dataset such as the one used here. Through clustering, behavioral patterns with similar characteristics can be identified, providing an initial overview of the child's behavioral profile [8].

The choice of K-Modes Clustering was based on the characteristics of Autism Spectrum Disorder (ASD) datasets, which are dominated by categorical attributes in the form of behavioral responses (e.g., Yes/No). According to Huang (1998), K-Modes is an extension of K-Means specifically designed for categorical data by using the mode as the centroid and simple matching dissimilarity as a distance measure. This approach allows objects with similar behavioral patterns to be grouped into the same cluster without requiring categorical data to be transformed into numerical data, resulting in more accurate and easily interpretable segmentation results [9]. The data used consisted of seven behavioral indicators: eye contact, name response, speech ability, social interaction, repetitive movements, focus, and the ability to follow simple instructions. In addition to the clustering process, this study also implemented an indication score formation to aid interpretation of the clustering results, as well as an analysis of relationships between features using Cramer's V. Based on this background, this study aims to conduct an exploratory analysis of the behavioral profiles of children with Autism Spectrum Disorder indicators using the K-Modes Clustering algorithm based on indication scores. The results are expected to provide an overview of children's behavioral patterns based on observed indicators and serve as an exploratory approach that can support the initial screening process and mapping of children's behavioral profiles.

2. METHOD

2.1 Research Stages

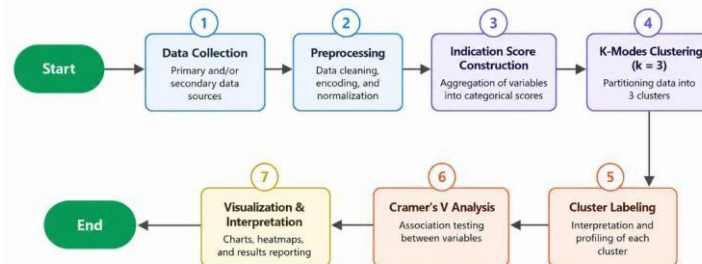


Figure 1. Flowchart of the K-Modes Clustering Analysis Process

1. **Feature Selection**
In this stage, seven behavioral indicators frequently used in observing ASD characteristics were selected: eye contact, name response, speech ability, social interaction, repetitive movements, focus, and simple instructions.
2. **Data Preprocessing** Preprocessing is performed to improve data quality before the clustering process. The preprocessing stages include data cleaning, category standardization, handling missing values using modes, and checking the distribution of categories for each variable. This stage aims to ensure the data is ready for use in the clustering process.
3. **Formation of Indication Scores** Each behavioral category is assigned a score based on its level of indication of ASD characteristics. The scores from all indicators are then summed to produce an Indication Score variable. This indication-score layer is what distinguishes the proposed score-based K-Modes approach from standard K-Modes clustering: rather than clustering solely on raw categorical attributes, cluster interpretation is anchored to a quantitative severity ordering derived from these scores, allowing the resulting clusters to be labeled and ranked (Low/Medium/High Indication) in a more clinically meaningful way. The equation for the indication score is written as:

Where: (SI) = Indication Score, (xi) = score on the first indicator, (n) = number of indications. The indication score is used as the basis for interpreting clustering results.

$$SI = \sum_{i=1}^n \frac{1}{n_i} \quad (1)$$

4. Clustering Using K-Modes

The data clustering process was performed using the K-Modes algorithm. This algorithm is a development of K-Means specifically designed for categorical data by replacing the mean with the mode and using a matching dissimilarity measure [7]. K-Modes was chosen because all research attributes are categorical. This algorithm is capable of grouping objects with similar characteristics based on the mode value in each cluster [10]. The number of clusters used in this study was three ($k = 3$). The determination of $k = 3$ was based on evaluation results using the Elbow method, which showed an elbow point at $k = 3$, with a significant decrease in cost before the curve began to flatten. Therefore, $k = 3$ was considered the optimal point, as adding clusters beyond this value did not provide a significant reduction in cost. Clustering is necessary because it can objectively reveal hidden patterns in multidimensional data without requiring prior class labels. This allows individuals to be grouped based on overall similarities in behavioral characteristics, rather than on a single score [11]. The following is a detailed explanation of each parameter:

- a. Maximum Number of Iterations (max_iter) At each iteration, the algorithm examines all objects and moves them to the most similar cluster, then updates the mode. The process stops when there are no object movements (convergence) or when the maximum number of iterations is reached (max_iter). A value of 100 is sufficient for most datasets.
- b. Number of Initializations (n_init) Because the initial centroid selection is random, results can vary between runs. With (n_init=10), the algorithm is run 10 times and the result with the lowest cost is selected. This reduces the risk of getting stuck in a bad local solution.
- c. Random Seed (random_state) The random state value of 42 is a common convention in data science research and practice. However, this value can be replaced with other values as needed, as long as the value is recorded and used consistently to ensure the reproducibility of experimental results.
- d. Final Cost Value Reported after training is complete. For example, if cost = 320, this means there are a total of 320 attribute mismatches between all objects and their respective cluster modes. The final cost value is used as an indicator to evaluate and compare the quality of clustering results in each run and across different cluster number variations (k).
- e. Dissimilarity Function (simple matching) Unlike K-Means, which uses Euclidean distance, K-Modes calculates the number of mismatched attributes between an object and its cluster mode. For instance, if there are five attributes and three of them differ, the dissimilarity value is 3.

5. Cluster Labeling

The clustering results were then analyzed using the average indication scores for each cluster. The cluster with the lowest average score was labeled "Low Indication," the cluster with a medium score was labeled "Moderate Indication," and the cluster with the highest average score was labeled "High Indication."

6. Analysis of Relationships Between Features

Cramer's V analysis is used to determine the relationship between categorical variables. This method measures the strength of the association between two categorical variables, with values ranging from 0 to 1. The Cramer's V value is calculated using Equation (2):

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \quad (2)$$

Where: (V) = Cramer's V, (χ^2) = Chi-Square value, (n) = sample size, and (k) = minimum number of categories. The closer the value is to 1, the stronger the relationship between the variables.

2.2 Dataset

This study utilizes a dataset of child behavior obtained through observations and teacher interviews within an educational setting. The dataset comprises data on 50 children, covering seven behavioral indicators associated with Autism Spectrum Disorder (ASD) characteristics: eye contact, response to name, speech ability, social interaction, repetitive movements, focus, and the ability to follow simple instructions. As all attributes are categorical, a clustering method capable of handling non-numerical data is required.

Table 1. Data Attributes

No	Attribute	Type	Category/Scale
1	Eye contact when being spoken to	Categorical	Yes, No
2	Response when their name is called	Categorical	Fast, Slow, Unresponsive
3	Speaking ability	Categorical	Smooth, Limited
4	Social interaction	Categorical	Good, Needs Improvement
5	Repetitive Movement	Categorical	Yes, No
6	A child's ability to focus	Categorical	Good, Needs Improvement
7	Understanding simple instructions	Categorical	Yes, No

2.3 Evaluation of Results

The evaluation was conducted using several approaches. First, a descriptive analysis of cluster profiles was performed by identifying the mode of each feature within each cluster based on the KModes centroid results [13]. Second, an indication score was calculated for each object—representing the number of attributes indicating ASD—and then averaged per cluster to measure relative severity levels across groups. Third, cluster stability validation was carried out using the Adjusted Rand Index (ARI) by comparing clustering results across various random seeds, where an ARI value close to 1 indicates high result consistency [14]. Fourth, relationships between features were analyzed using Cramer's V to measure the strength of association between categorical variables [15]. The clustering results were then interpreted to characterize the children's behavioral profiles based on ASD-related indicators, as shown in the following table.

Table 2. Cluster profile

Feature	Cluster 0 (Moderate Indication)	Cluster 1 (High Indication)	Cluster 2 (Low Indication)
Eye contact	Yes	No	Yes
Name response	Fast	Slow	Fast
Speaking ability	Fluent	Limited	Limited
Social interaction	Good	Not enough	Good
Repetitive movement	Yes	Yes	No
Focus	Not enough	Not enough	Good
Simple instructions	Yes	Yes	Yes
ARI Validation	0.7466	0.7466	0.7466

3. RESULTS AND DISCUSSION

3.1 K-Modes Clustering Results

The clustering process was conducted using the K-Modes algorithm with three clusters ($k = 3$). The selection of three clusters aimed to group children's behavioral profiles into categories representing different levels of indication, based on previously established indication scores. The clustering process yielded three

groups, which were subsequently labeled Low Indication, Moderate Indication, and High Indication. These labels were assigned based on the average indication score of each cluster; the cluster with the lowest average score was labeled Low Indication, while the cluster with the highest average score was labeled High Indication.

Table 3. Distribution of Cluster Members

Clusters	Cluster Label	Number of Data Points	Percentage
Cluster 0	Moderate Indication	22	44%
Cluster 1	High Indication	17	34%
Cluster 2	Low Indication	11	22%
Total		50	100%

Based on Table 3, the majority of the data falls into the Moderate Indication cluster, comprising 22 children (44%). Furthermore, the High Indication cluster consists of 17 children (34%), while the Low Indication cluster comprises 11 children (22%). These results indicate that the majority of children in the dataset exhibit certain behavioral characteristics associated with ASD, yet do not show a predominance of indicators across all observed aspects. On the other hand, the presence of a "High Indication" cluster reveals a subgroup of children who display a greater number of ASD-related behavioral indicators compared to the other groups. The distribution of cluster members is visualized using a bar chart to facilitate the interpretation of the clustering results. The visualization shows that the "Moderate Indication" cluster has the largest number of members compared to the other clusters.

3.2 Analysis of Relationships Between Features Using Cramer's V

Following the clustering process, an analysis of the relationships between features was conducted using Cramer's V to determine the strength of association among the observed behavioral indicators. Based on the Cramer's V calculations, several variable pairs were found to exhibit relatively strong relationships. The five highest association values are presented in Table 4.

Table 4. Five Strongest Feature Relationships

Variable 1	Variable 2	Cramer's V
Speaking Ability	Repetitive Movement	0.548
Speaking Ability	Social Interaction	0.480
Eye Contact	Speaking Ability	0.453
Speaking Ability	Focus	0.436
Name Response	Speaking Ability	0.431

Based on Table 4, the strongest association was found between Speech Ability and Repetitive Movements, with a Cramer's V value of 0.548. This value indicates a moderately strong categorical association between the two variables. This study did not conduct accompanying chi-square significance testing; given the modest sample size ($n = 50$), some category combinations may have small expected cell counts, which is a caveat to bear in mind when interpreting association strength (see also Section 3.3). Additionally, the Speech Ability variable showed relatively high associations with Social Interaction (0.480), Eye Contact (0.453), Focus (0.436), and Response to Name (0.431). These findings demonstrate that the aspect of communication is closely linked to various other behavioral indicators.

3.3 Discussion

The results of Cramer's V analysis indicate that the Speech Ability variable has the strongest relationship with several other behavioral indicators, particularly social interaction, name response, and focus. This finding aligns with the ASD literature, which consistently identifies verbal communication impairments as a core characteristic of autism spectrum disorder [16]. Children with speech impairments tend to exhibit difficulties in responding to their names and establishing functional social interactions, making speech ability a central indicator in the development of behavioral profiles in this study. The findings of this study have practical implications for teachers and educational staff in school settings.

The resulting cluster profiles can be used as initial references to identify children requiring further attention, particularly those in the High Indication group who exhibit multiple behavioral indicators related to

ASD simultaneously. This clustering-based approach does not require a clinical diagnostic label, making it a simple and practical initial screening instrument for teachers through daily behavioral observations [17]. The use of K-Modes is considered appropriate because all the attributes used are categorical, allowing the clustering process to be performed without complex numerical transformations [7].

This algorithm preserves the original meaning of the data categories, making the clustering results easier to interpret. However, K-Modes has several weaknesses, including sensitivity to the initial centroid selection and the need to manually determine the number of clusters before the clustering process [18]. Furthermore, this algorithm does not consider weighting between attributes, so each behavioral indicator is treated equally, even though the indicators may differ in clinical significance.

This study has several limitations that require consideration. First, the sample size was relatively small (50 children), so caution is needed when generalizing the clustering results to a broader population. Second, the data were obtained through teacher observations and interviews, which introduces the potential for subjective bias in the behavioral indicators. Third, this study was exploratory in nature and was not intended as a clinical diagnostic tool for ASD, but rather as an initial mapping of children's behavioral profiles based on indicators observed in educational settings. Fourth, the reported cluster percentages and Cramer's V value were not accompanied by formal statistical significance testing or confidence intervals, so these estimates should be interpreted as descriptive rather than as precise population parameters. Fifth, this study did not include a baseline comparison against a simpler approach, such as directly ranking or binning children by raw indication score without clustering; such a comparison would help clarify the specific added value of the K-Modes clustering step and is recommended for future work.

Future research is recommended to use a larger dataset and involve clinical validation by professionals. Future research is recommended to expand the sample size so that the clustering results are more representative of the population of children with diverse ASD characteristics. Furthermore, the use of standardized observation instruments can reduce the potential for subjective bias in data collection. Model development can be improved by incorporating attribute weighting based on the clinical significance of each indicator, as well as involving validation from professionals such as psychologists or therapists to strengthen the interpretation of the grouping results.

4. CONCLUSION

This study aims to conduct an exploratory analysis of children's behavioral profiles based on seven indicators related to Autism Spectrum Disorder (ASD) using the K-Modes Clustering algorithm. The results of the study produced three groups of behavioral profiles, namely Low Indication (22%), Medium Indication (44%), and High Indication (34%), which indicates that most children in the dataset have several behavioral characteristics related to ASD with varying degrees of indication. Cramer's V analysis revealed that speech ability is the variable most closely related to other behavioral indicators, with the highest value found in the relationship between speech ability and repetitive movements of 0.548. This study has limitations in the relatively small number of samples and data sourced from teacher observations that potentially contain subjective bias, so the clustering results cannot be generalized widely and are not intended as a clinical diagnostic tool.

BIBLIOGRAPHY

- [1] H. Yu, D. M. Rizzo, J. P. Hanley, E. L. Coderre and A. P. Prelock, "Identifying neuroanatomical and behavioral features for autism spectrum disorder diagnosis in children using machine learning," *PLoS ONE*, vol. 17, no. 7, pp. 1-19, 2022.
- [2] M. Liao, H. Duan and G. Wang, "Application of Machine Learning Techniques to Detect the Children with Autism Spectrum Disorder," *Hindaw iJournal of Healthcare Engineering*, vol. 2022, no. 1, 2022.
- [3] T. Vargason, R. E. Frye, D. L. McGuinness and J. Hahn, "Clustering of Co-Occurring Conditions in Autism Spectrum Disorder During Early Childhood: A Retrospective Analysis of Medical Claims Data," *International Society for Autism Research*, vol. 12, no. 8, 2019.
- [4] C. Song, Z.-Q. Jiang, L.-F. Hu, W.-H. Li, X.-L. Liu, Y.-Y. Wang, W.-Y. Jin and Z.-W. Zhu, "A machine learning-based diagnostic model for children with autism spectrum disorders complicated with intellectual disability," *Front. Psychiatry*, vol. 13, 2022.
- [5] H. B. Alateyat, S. Cruz, E. Cernadas, M. Tubío-Fungueiriño, A. Sampaio, A. J. González-Villar, A. Carracedo, M. Fernández-Delgado and M. F. Prieto, "A Machine Learning Approach in Autism Spectrum

-
- Disorders: From Sensory Processing to Behavior Problems," *Frontiers in Molecular Neuroscience*, vol. 15, 2022.
- [6] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2012.
- [7] Z. Huang , "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, p. 283–304, 1998.
- [8] A. L. Ahmad and N. Cavus, "A Clinical Validity-Preserving Machine Learning Approach for Behavioral Assessment of Autism Spectrum Disorder," *OBM Neurobiology*, vol. 6, no. 3, pp. 1-40, 2022.
- [9] Z. Huang and M. K. Ng, "A Note on K-modes Clustering," *Journal of Classification*, vol. 20, p. 257–261, 2003.
- [10] L. Ziyang and L. Panchi, "Clustering Algorithm of Quantum Self-Organization Network," *Journal of Applied Sciences*, vol. 5, pp. 270-278, 2015.
- [11] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, p. 645–678., 2005.
- [12] R. L. Thorndike, "Who Belongs in the Family?," *The Psychometrika Society*, vol. 18, p. 267–276, 1953.
- [13] K. S. Dorman and R. Maitra, "An Efficient K-Modes Algorithm for Clustering Categorical Datasets," *The ASA Data Science Journal*, vol. 15, no. 1, pp. 83-97, 2021.
- [14] L. & A. P. ARI: Hubert, L. Hubert and P. Arabie , "Comparing partitions," *Journal of Classification*, vol. 2, p. 193–218, 1985.
- [15] H. Cramer, *Mathematical Methods of Statistics*, Uppsala: Princeton University Press, 1946, p. 282–283.
- [16] H. Tager-Flusberg , "On the nature of linguistic functioning in early infantile autism," *Journal of Autism and Developmental Disorders*, vol. 11, p. 45–56, 1981.
- [17] T. Charman and G. Baird, "Diagnosis of autism spectrum disorder in 2- and 3-year-old children," *Journal of Child Psychology and Psychiatry*, vol. 43, no. 3, p. 289–305, 2008.
- [18] C. K. Syriopoulou- Delli and E. Papaefstathiou, "Review of Cluster Analysis of Phenotypic Data in Autism Spectrum Disorders: Distinct Subtypes or a Severity Gradient Model?," *International Journal of Developmental Disabilities*, vol. 66, no. 1, pp. 13-21, 2019.
- [19] L. E. Gialloreti , R. Enea, . V. D. Micco , D. D. Giovanni and P. Curatolo , "Clustering Analysis Supports the Detection of Biological Processes Related to Autism Spectrum Disorder," *Genes*, vol. 11, no. 12, pp. 118, 2020.
- [20] A. Narita, M. Nagai, S. Mizuno, S. Ogishima, G. Tamiya and M. Ueki, "Clustering by phenotype and genome-wide association study in autism," *Translational Psychiatry*, vol. 10, no. 1, pp. 1-12, 2020.