BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 30 No. 3. 2025

UDC 004.852:519.23
DOI: 10.62660/bcstu/3.2025.93

# Study of the impact of different categorical feature encoding techniques on cluster structures

**Natalia Kondruk***

PhD in Technical Sciences, Associate Professor
Uzhhorod National University
88000, 14 Universytetska Str., Uzhhorod, Ukraine
https://orcid.org/0000-0002-9277-5131

**Inna Neroda**

Postgraduate Student
Uzhhorod National University
88000, 14 Universytetska Str., Uzhhorod, Ukraine
https://orcid.org/0009-0009-3686-3565

**Abstract.** Categorical features are a common type of data used in data analysis, but their non-metric nature makes it difficult to apply standard clustering algorithms. The relevance of the study is conditioned by the need to assess the impact of different methods of recoding (digitisation) of such features on the effectiveness of cluster analysis. The purpose of the study was to investigate how different techniques of categorical data processing affect the quality and structure of clusters. The methodology included the implementation of three models with different approaches to variable coding: without taking into account domain specifics, considering the content of the features, and with alternating the order of application of clustering and dimensionality reduction approaches. LabelEncoder, OrdinalEncoder, One-Hot Encoding, Mapping, and MultiLabelBinarizer were used for coding. In each of the models, clustering was performed using two algorithms – K-Means and agglomerative clustering, which allowed comparison of their sensitivity to changes in data representation. The t-SNE dimensionality reduction method was used to visualise the cluster structure in two-dimensional space. The quality of clustering was evaluated using the Silhouette Score, Dunn Index, Davies-Bouldin Index, and Calinski-Harabasz Index metrics. The data for the analysis were obtained from an open source and contained information about the psycho-emotional state of students. The study found that the basic recoding of categorical features without considering their semantics and context negatively affected the quality of clustering, reducing the accuracy of the division and complicating the interpretation of the results. Instead, the use of domain-oriented coding approaches ensured the development of clusters with clearer boundaries and a more logical internal structure. In addition, it was found that changing the sequence of clustering and dimensionality reduction affects the preservation of local relationships in the data. It was analysed that different approaches change both the number and quality of clusters, which was reflected in the values of the evaluation metrics. The practical significance of the results lies in the possibility of their application by data analysts and machine learning specialists to improve the accuracy of segmentation of complex categorical data

**Keywords:** data analysis; machine learning; unsupervised learning; automatic object grouping; segmentation

**Article's History:** Received: 11.05.2025; Revised: 11.08.2025; Accepted: 15.09.2025.

**Suggested Citation:**
Kondruk, N., & Neroda, I. (2025). Study of the impact of different categorical feature encoding techniques on cluster structures. *Bulletin of Cherkasy State Technological University*, 30(3), 93-105. doi: 10.62660/bcstu/3.2025.93.

*Corresponding author

## INTRODUCTION

Most of the real data contains categorical features that characterise the qualitative properties of objects, such as colour, product type, geographical location, level of education, or disease category. Unlike numerical variables, which have a natural order and clear metric interpretation, categorical data does not have a numerical equivalent or internal metric structure, which makes it difficult to use them directly in standard clustering algorithms. Most clustering algorithms operate in numerical spaces by applying distance metrics or similarity measures that cannot be directly applied to categorical features. To overcome this problem, a number of coding techniques have been developed that allow transforming categorical variables into numerical representations. Among the most common methods are one-hot encoding, which converts categories to binary vectors, and label encoding, which assigns each category a unique numeric code.

Choosing a coding technique is not a trivial task, as it can significantly affect the results of clustering. It is the way categorical variables are represented that determines how the algorithm interprets distances or similarities between objects, and therefore, how clusters will form. Accordingly, there is a need for a systematic study of various techniques for encoding categorical variables and their impact on cluster development. This includes analysing how different coding approaches change the distribution of objects in clusters, affect their compactness and separability.

Studies confirm that the choice of categorical feature coding technique significantly affects the quality of clustering and the interpretability of the resulting groups. Thus, M. Anitha *et al.* (2025) proposed a new Chi-Square target Encoding (CSTE) approach that uses chi-square statistics to estimate the association between categories and the target variable. This method allows efficient storage of information during the transformation of categorical features into numerical representations, which positively affects the accuracy of clustering. CSTE demonstrated an advantage over conventional methods such as one-hot encoding and fuzzification, achieving an average accuracy gain of 3.94%, and high F1 and AUC values for various data sets.

Similarly, Z. Liang (2025) proposed new encoding methods for highly cardinal categorical variables, in particular, low-level encoding and logistic regression, which allows reducing the dimension of data without losing important characteristics. These methods provide compact and informative feature vectors that help to improve clustering efficiency and reduce the risk of overtraining. The study showed significant improvements in model performance and computational efficiency compared to conventional methods such as one-hot encoding, especially when working with large data sets.

In the field of medical research, H. Hafid & S. Annisa (2025) applied K-Medoids and K-Prototypes techniques to cluster hypertensive patients, demonstrating the importance of successfully selecting categorical variable encoding for accurate grouping of medical data. The researchers noted that the use of the K-Prototypes algorithm, which combines clustering of numerical and categorical features, provides a more adequate development of patient groups compared to conventional methods based only on numerical data. The researchers emphasised that the correct choice of categorical variable coding techniques significantly affects the accuracy and stability of clustering, which is critical for medical applications where segmentation accuracy directly affects the quality of diagnosis and treatment.

Clustering as a method for identifying groups in data largely depends on how categorical features are transformed into numeric spaces. According to A.M. Ikotun *et al.* (2023), the K-Means algorithm remains one of the most common clustering methods, especially in big data environments. However, its efficiency is significantly reduced when working with categorical variables without proper transcoding, which can lead to distortion of the cluster structure and reduce the accuracy of the results. E.K. Tokuda *et al.* (2022) concluded in their paper that agglomerative clustering methods, in particular, those based on different merge criteria (single, average, median, complete, centroid, and Ward), demonstrate flexibility in working with different data types. However, these methods are also prone to detecting false-positive clusters, especially in conditions of homogeneous data distributions. This highlights the importance of carefully choosing categorical feature encoding methods, since incorrect transformation can lead to the development of clusters that do not reflect the actual data structure. However, the literature does not pay enough attention to the direct impact of coding techniques on the cluster structure, which leaves room for further research in this area.

An additional aspect related to preparing data for clustering concerns dimensionality reduction. According to N.E. Kondruk (2023), the use of dimensionality reduction techniques is an important prerequisite for improving the efficiency of machine learning models. The researcher considered several techniques, in particular, Principal Component Analysis (PCA), Isometric Mapping (Isomap), Uniform Manifold Approximation and Projection (UMAP), and t-SNE, and concluded that t-SNE provided the best predictive results for the problem of classifying high-dimensional data and that the use of advanced methods of reducing the dimension of data can contribute to the construction of more efficient models.

Overall, research confirms that conventional methods such as one-hot encoding and label encoding have limitations, especially when working with large data sets. New approaches that consider statistical relationships between categories can improve the quality of clustering and reduce overlap between groups. The purpose of this study was to conduct a comparative analysis of the influence of various techniques for encoding categorical features on the quality and structure of clusters in machine learning problems.

This study examined how different approaches to encoding categorical data can change the distribution of objects in clusters and which methods provide the best performance in different data contexts. In addition, the application of various clustering methods and clustering quality indices was considered to evaluate the effectiveness of different techniques for encoding categorical features using the example of segmenting students according to their psycho-emotional health.

## MATERIALS AND METHODS

Clustering was evaluated using internal, external, and graphical methods. Internal methods were based on the characteristics of the data itself and did not require external information. One of the main indicators was the silhouette coefficient, which evaluated the quality of clustering by analysing the grouping of objects within clusters and their isolation from other clusters. The value varied from -1 to 1, where higher values indicated better cluster separation (Kondruk, 2019). In addition, the Davies-Bouldin index was used, which measured the average similarity between clusters and their resolution (lower values indicate better clustering) (Wegmann *et al.*, 2021). The Akaike Information Criterion (AIC) helped to assess the complexity of the clustering model: lower values indicate a better balance between the quality of clustering and the number of model parameters (Sánchez Vinces *et al.*, 2025).

External methods considered comparisons with reference distributions or correct cluster labels. Graphical methods included a dimensionality reduction procedure for visual evaluation of clusters. The PCA method was previously tested, but due to the nonlinear nature of the data and the lack of linear relationships in the dataset under consideration, it was not applicable. Instead, the nonlinear t-SNE method was used, which better preserves local distances between points and allowed estimating the density and overlap of clusters in two-dimensional space.

The K-Means method is a clustering method designed to split data into k clusters. Its purpose was to split data into a predefined number of clusters (K) so that objects within one cluster have a greater similarity to each other than objects in other clusters (Lloyd, 1982). Agglomerative clustering is a hierarchical analysis method that begins with each object being treated as a separate cluster. The algorithm gradually combines the two nearest clusters at each stage, until all objects form one large cluster. This approach allowed investigating the data structure and visualising relationships between objects (Miyamoto, 2022; Sieranoja & Fränti, 2025; Behzadidoost & Izadkhah, 2025).

For experiments, data on the psychoemotional health of students were taken. Data set based on the survey by V.A. Ashfaq (n.d.), consisted of 87 entries and 21 attributes and contained information about students, their students, and their educational experience, psychological state, and social aspects of life. The study was conducted in accordance with the ethical principles set out in the Declaration of Helsinki (World Medical Association's Declaration of Helsinki, 2013).

In this regard, the study focused on three different models for transcoding categorical variables, each of which offered a unique approach to transforming the original data. Model 1: data was transcoded without considering their nature, which allowed creating a neutral representation of variables for further analysis, followed by data compression and clustering. Model 2: data transcoding was performed considering feature domains, which allowed more accurately reflecting their essence and ensuring correct representation, then data compression and clustering occurs. Model 3: data was transcoded based on its nature, then the data was clustered, and only then the dimensionality reduction method was applied for further visualisation.

In Model 1, label encoder, ordinary encoder, and mapping methods were used to process categorical features, depending on their nature. This approach allowed getting a neutral representation of the source data and minimising the impact of subjective assumptions about their semantics. The label encoder method was used to transcode columns such as "gender", "degree_level", "residential_status", and "campus_discrimination". This method converted attributes to numeric values, while maintaining easy transcoding. It was ideal for signs where there is no natural order, such as gender, educational level, and others. Based on this, the method minimised the complexity of analysis, ensuring ease of working with data. The ordinary encoder method was applied to the "university" and "degree_major" columns. This approach allowed organising categories in a numeric format, which can be important for maintaining the order of values. For example, in the case of universities and educational specialities, the method allowed comparing their unique values with numbers for more convenient further processing.

Category mapping using the dictionary was used for the "academic_year", "cgpa", "sports_engagement", "average_sleep", and "stress_relief_activities" columns. This approach allowed manually mapping text values to numeric values, which was especially convenient for

displaying categories with a natural sequence or simplifying multidimensional data into a single numeric form. For example, in the "academic_year" column, the sequence of academic years is represented in numerical form to maintain a logical order, and the ranges of GPA scores in the "cgpa" column were displayed in numerical averages. Similarly, the degrees of activity in sports, sleep hours, and relaxation activities were transcoded in such a way as to preserve the content of information and increase the convenience for further analysis. Categorical variables were transcoded to numeric format, and the t-SNE method was used to reduce the dimension of the feature space and visualise the data structure.

To analyse clustering using the K-Means method, key metrics were used to assess its quality and determine the optimal number of clusters. Each metric had its own meaning in the decision-making process. Silhouette Score was used to determine the degree of isolation and compactness of clusters. Davies-Bouldin Score allowed estimating the degree of overlap between clusters, where smaller values indicated a clearer separation. The AIC criterion was used to select the model that best balances complexity and clustering quality.

Agglomerative clustering considered the hierarchical structure of data and helped to better identify complex relationships, which was especially noticeable in cluster visualisation. Clustering quality assessment was performed using metrics such as Silhouette Score to assess cluster compactness, and AIC to select the optimal balance between model accuracy and complexity. In addition, a dendrogram was used for agglomerative clustering, a graphical tool that displays the hierarchical process of combining points. The number of clusters that best reflected the data structure can be determined by drawing a horizontal line at a certain distance level.

In Model 2, to ensure the correct representation of data, various transcoding methods were applied for each variable, considering their nature and characteristics. For variables that had a limited set of categories, such as "gender" and "residential_status", the method of simple transcoding using label encoder was used. This approach allowed efficiently representing nominal categories in numeric format, while maintaining their nominal nature without creating additional columns. For variables with an extended set of unique values, such as "university" and "degree_major", the binary encoding method was used – one-hot encoder. This method allowed creating separate binary columns for each category, which avoided logical priority between them and preserved their uniqueness. Additional created columns extended the data set while maintaining its full interpretation.

Variables with ranges of values or ordered categories (academic_year, cgpa, sports_engagement, average_sleep) were transformed using mapping. This allowed converting text ranges to numeric levels or the average value of the range. For example, a scale was created for the CGPA variable that displays the average value of each range, and for "sports_engagement" – numerical levels according to the frequency of sports activities. The "stress_relief_activities" variable contains multiple categories for each respondent, so the "multilabelbinariser" technique was used to process it. This allowed presenting respondents' participation in several activities simultaneously in a binary matrix format. This encoding format allowed considering potential relationships between activities in further analysis. To assess the quality of clustering using the K-Means method, the Silhouette Score and Kalinski-Harabasz Score indices were used, which allowed determining the degree of compactness and resolution of clusters. Similarly, the Calinski-Harabasz index was used for agglomerative clustering, and a visual study of the dendrogram was performed to determine the optimal separation threshold and the number of clusters.

In Model 3, data was transcoded based on the nature of each variable, just as in Model 2. The main difference between these models was the sequence of stages. In Model 2, data was first transcoded, then its dimension was compressed, and only then clustering was performed. This reduced the dimension of data even before clusters were formed, which can simplify the clustering process, but may affect the preservation of the original data structure. In Model 3, by contrast, data transcoding was accompanied by clustering before dimension compression. This change in the sequence allowed focusing on the development of clusters in their original form, and only then applying dimensionality reduction methods to visualise the resulting structures.

To determine the optimal number of clusters in the K-Means method, a sequential analysis of the values of the Silhouette Score, Dunn Index, and Davies-Bouldin index metrics was performed on different cluster numbers. Evaluation of these metrics allowed selecting the number of groups that provided a balance between compactness and cluster resolution in the data under study. In the case of agglomerative clustering, in addition to calculating the Silhouette Score, a visual study of the dendrogram was performed. Changes in cluster structure at different cut-off thresholds were analysed to find the number of groups that best reflects the natural divisions in the data hierarchy. The selected separation level was supposed to provide maximum clarity for clusters without excessive merging or splitting.

## RESULTS AND DISCUSSION

After digitising the categorical data by transcoding them, a correlation matrix was constructed for preliminary analysis of the relationships between features, which is shown in Figure 1.

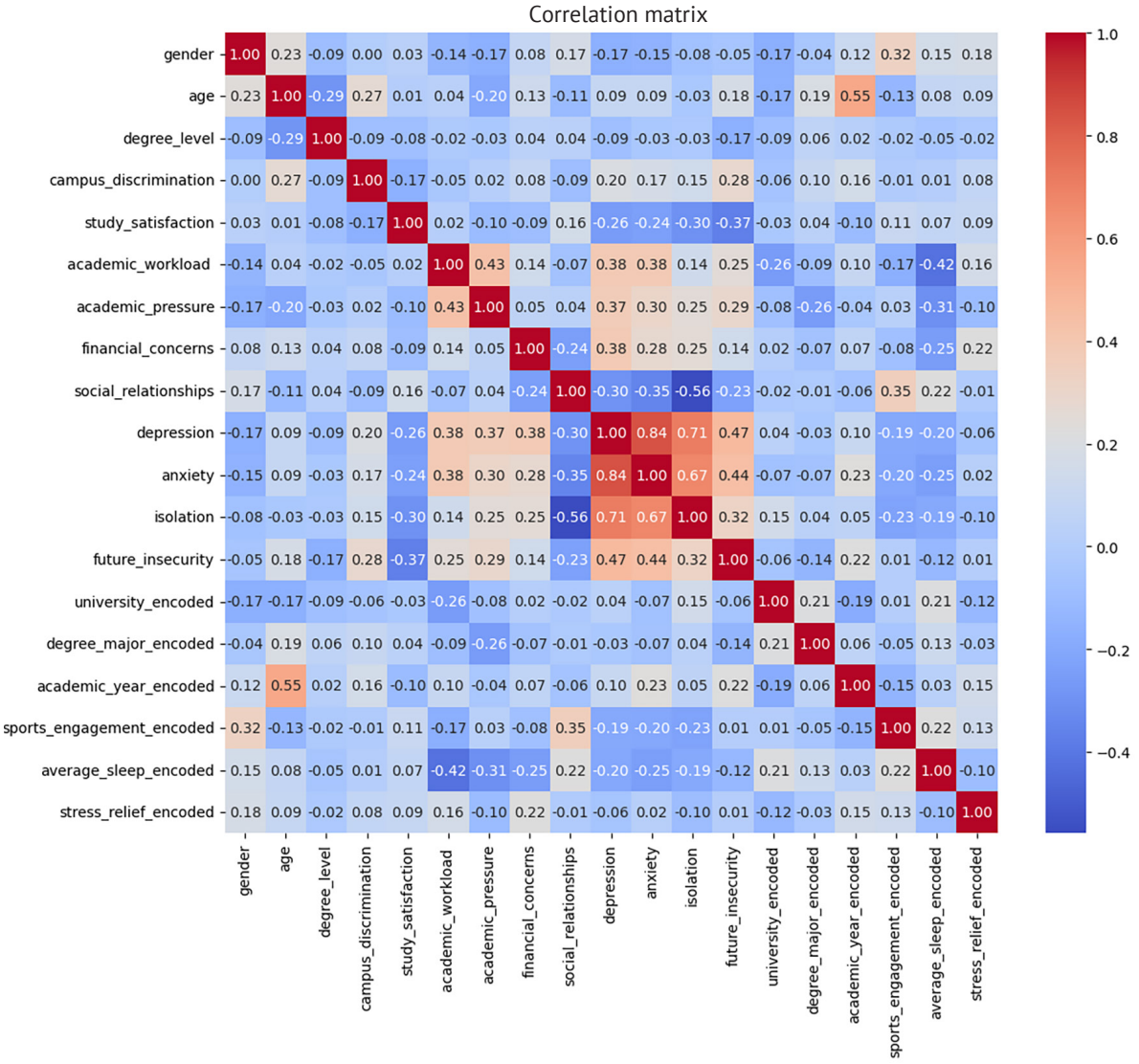


**Figure 1.** Correlation matrix

**Source:** compiled by the authors

The analysis showed that the data has a predominantly nonlinear structure, which influenced the choice of further processing methods. To reduce the data dimension, the PCA method was initially considered, but due to inconsistencies with the assumptions of this method (in particular, the lack of multicollinearity), the t-SNE algorithm was chosen, which better reflects local structures and is able to detect nonlinear dependencies in the data.

**Clustering results based on Model 1.** The K-Means method with 6 clusters demonstrates the division of data into groups characterised by moderate compactness and separation. The Silhouette Score (0.4) reaches a local maximum for six clusters, which indicates the best combination of cluster compactness and distance between them. A low Davies-Bouldin Score (0.8) confirmed that the clusters have a clear structure with minimal overlap. The AIC criterion, which is 804, states that the choice of six clusters provides an optimal balance between the accuracy of the model and its complexity. Despite a certain degree of data noise and overlap, the K-Means method effectively revealed basic patterns and structures that allow data to be divided into groups with relatively clear boundaries.

Analysis of the results of agglomerative clustering for six clusters shows that the algorithm successfully grouped data in a multidimensional space, which is confirmed by the distribution of points in the visualisation. Silhouette Score has a local maximum of six clusters, which indicates the best compactness and clarity of separation. The AIC graph shows the minimum value for six clusters, which confirms the optimal balance between the accuracy of the model and its complexity. The clusters are clearly separated and show moderate compactness. For example, well-defined groups (cluster

0, 1) indicate a clear structure, but some clusters (such as 2 and 5) have more overlap, reflecting more complex relationships in the data. The dendrogram shown in Figure 2, used as an additional tool, clearly reflects the hierarchical process of cluster development and allows confirming the choice of six clusters as the optimal division that preserves the natural structure and relationships between objects. To visualise the spatial distribution of clusters, the t-SNE dimensionality reduction method was used (Fig. 3). This approach allowed for visual assessment of the structure and isolation of clusters in two-dimensional space, which confirms the results of clustering and helps to better understand the internal relationships between groups.

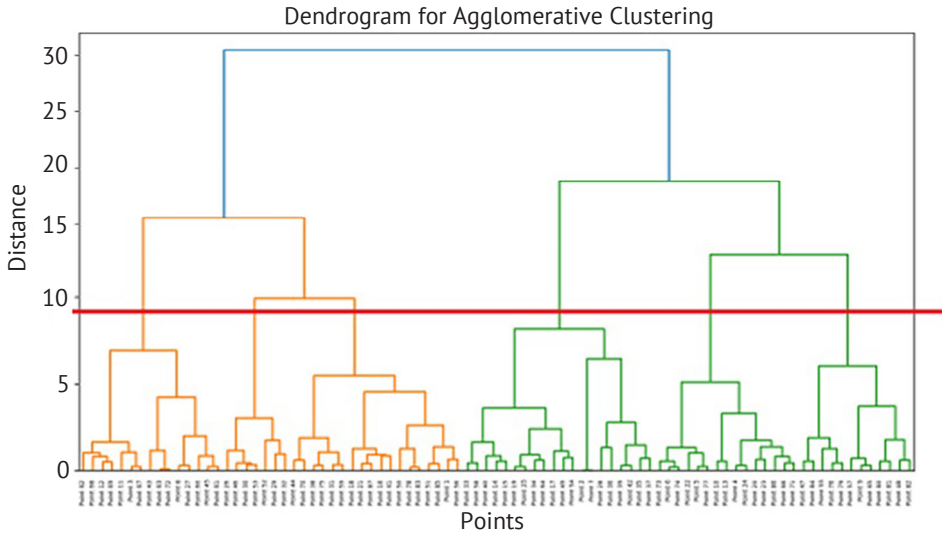


**Figure 2.** Dendrogram for Model 1

**Source:** compiled by the authors

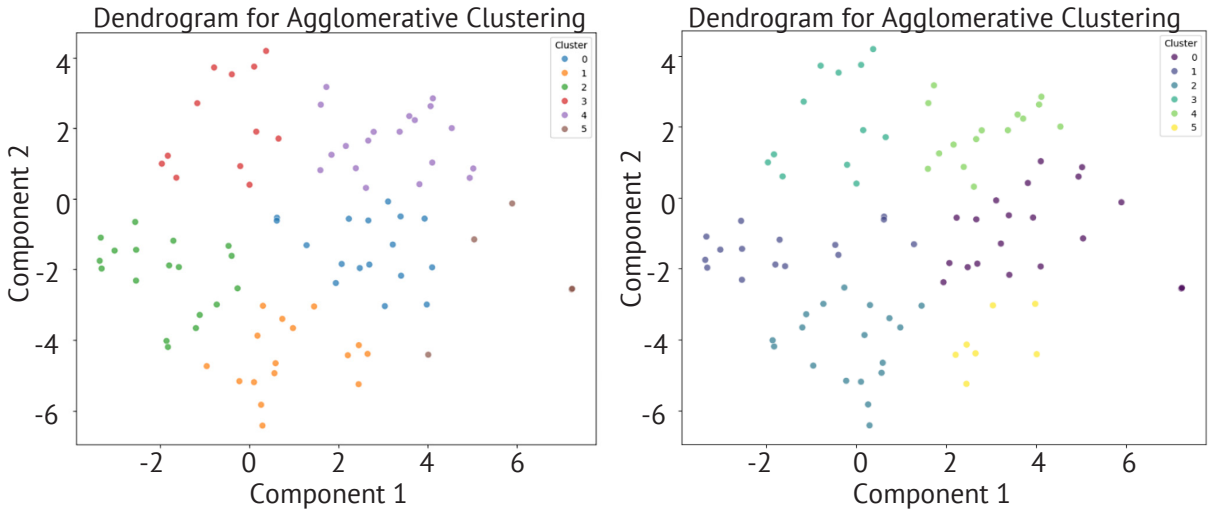


**Figure 3.** Clustering results for Model 1

**Source:** compiled by the authors

**Clustering results based on Model 2.** The results of clustering using the K-Means method with six clusters demonstrated a clear grouping of most data with high compactness and minimal mixing between clusters. Analysis of the Silhouette Score and Kalinski-Harabasz Score shows that the local maximum of the former and the high value of the latter are achieved in six clusters. This indicates optimal compactness and separation of clusters by this number. Thus, six clusters are the optimal choice, since they provide maximum clustering quality.

For agglomerative clustering, analysis of the Calinski-Harabasz Index demonstrated that the maximum ratio between inter- and intra-cluster variance is achieved in five clusters, which indicates an optimal separation of groups with high internal compactness and maximum distance between them. In addition, the dendrogram shown in Figure 4 was used as an additional visual tool that confirmed the selection of five clusters, demonstrating a clear separation of independent groups at the established cut-off threshold.

Thus, the choice of five clusters was consistent with both numerical metrics and visual analysis of the hierarchical data structure. This distribution indicates the presence of homogeneous subgroups with characteristic features, which is important for further analysis. Clear boundaries between clusters ensure the stability of groups and provide grounds for practical interpretation of their characteristics. Thus, the results of agglomerative clustering confirm the mathematical validity of the choice of five clusters and the correspondence of the breakdown to the structure of the original data (Fig. 4).

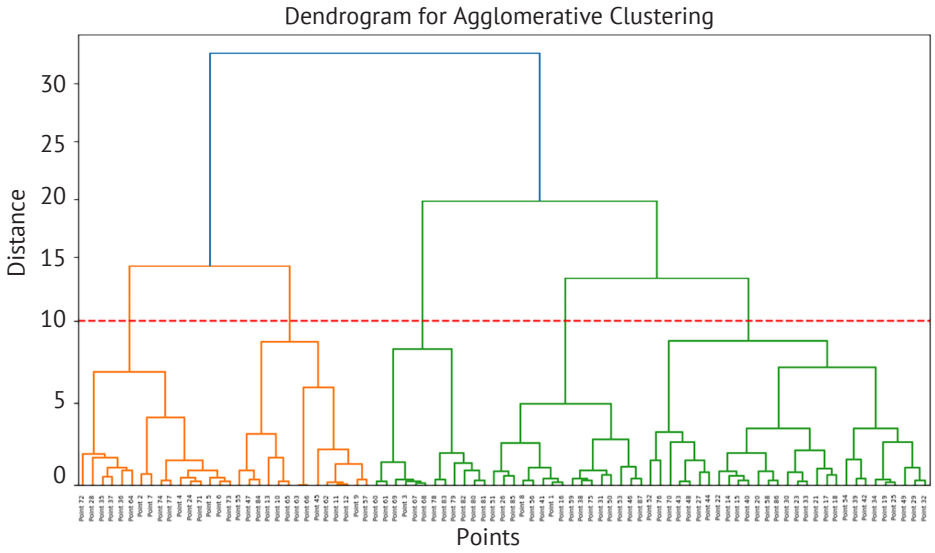


**Figure 4.** Dendrogram for Model 2

**Source:** compiled by the authors

The results of numerical and visual analysis confirm the feasibility of the selected clustering parameters for Model 2. Figure 5 shows visualisations of cluster distributions that illustrate the data structure and segmentation quality obtained by K-Means and agglomerative clustering methods.

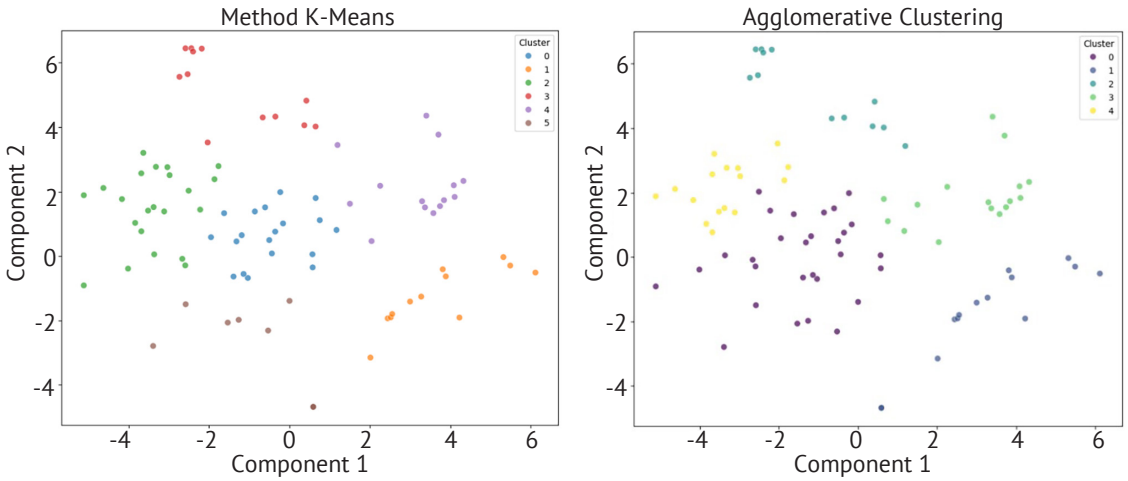


**Figure 5.** Clustering results for Model 2

**Source:** compiled by the authors

For the K-Means method, there is a clear development of six clusters, each of which has a high compactness and is well separated from the others. Minimal overlap between some groups (in particular, clusters 2 and 4) indicates the complexity of the source data structure, but the overall segmentation looks logical and stable. In the case of agglomerative clustering, segmentation into five clusters was optimal, which show better separation between groups compared to Model 1. This suggests that consideration of the content of features in encoding allows more accurately preserving the natural data structure even in a complex multidimensional space.

**Clustering results based on Model 3.** For the K-Means clustering method, it is important to determine the optimal number of clusters that will ensure high-quality data segmentation. The Silhouette Score Index reached a local maximum in six clusters, which indicates high internal cohesion and clear data structure in each cluster. The Dunn Index also showed high values for six clusters, indicating the compactness of the clusters and their good separation. The Davies-Bouldin Index reached a minimum value with the same number of clusters, which indicates low inter-cluster similarity and clustering quality in general. Therefore, the choice of six clusters for K-Means is optimal, since it provides a balance between internal cohesion and group separation. Cluster overlap indicates that some points lie at the boundary between groups, which is natural for complex data with a continuous structure. In general, clustering still allows isolating the main data structure and ensuring their overall distribution into groups.

For agglomerative clustering, the Silhouette Score Index reaches a local maximum for four clusters, which indicates an optimal balance between cohesion within clusters and their separation. The dendrogram analysis shown in Figure 6 confirmed this choice – when divided into four groups, the data form well-defined clusters with a distance between them of about 15 units, which minimises overlap. Fewer clusters lead to less detailed segmentation and loss of local patterns.

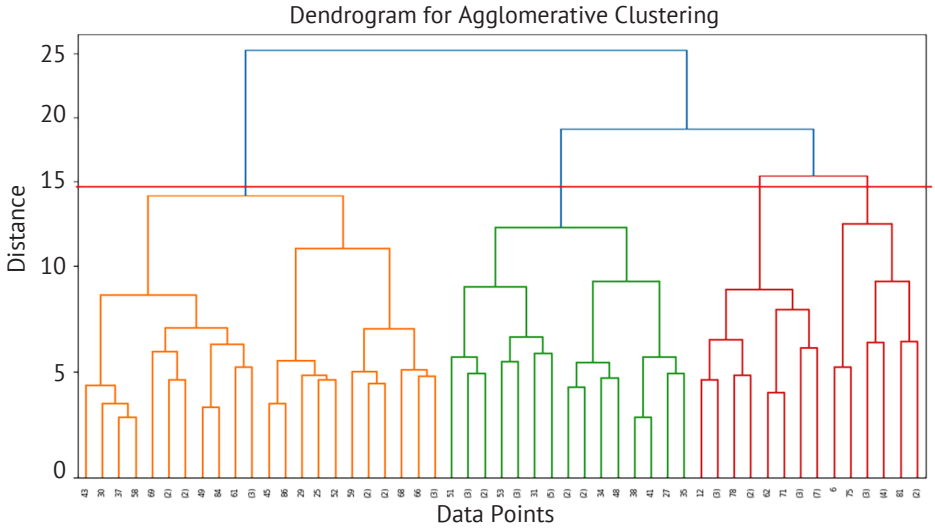


**Figure 6.** Dendrogram for Model 3

**Source:** compiled by the authors

To further confirm the qualitative characteristics of clustering using Model 3, the resulting clusters were visualised. The data distributions shown in Figure 7 using the K-Means method (6 clusters) and agglomerative clustering (4 clusters) allow estimating the spatial structure of groups, their compactness, the degree of separation, and the presence of overlapping zones.

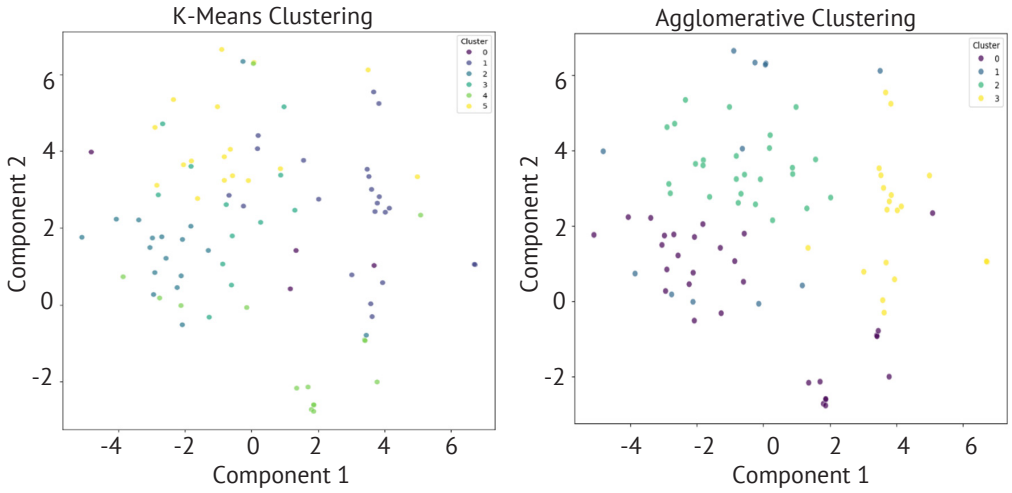


**Figure 7.** Clustering results for Model 3

**Source:** compiled by the authors

Analysis of Model 1 showed that the optimal number of clusters for both clustering methods is six, which was confirmed by quality assessment metrics. For K-Means, metrics indicate acceptable clustering quality, although data noise affected the clarity of boundaries between individual groups. Agglomerative clustering allowed isolating six clusters using a dendrogram. The clarity of the distribution largely depended on the cut-off level, which confirms the influence of the hierarchical structure on clustering results. Within this model, the method provided adequate data segmentation, although partial overlap between certain clusters was observed. Overall, the results of Model 1 showed that clustering methods provided an adequate data structure, but the choice of encoding method affects the accuracy of segmentation and the clarity of cluster boundaries. This highlights the importance of further research on Models 2 and 3, which can eliminate the limitations of Model 1 by more accurately considering the nature of variables or changing the sequence of analysis steps.

Model 2 showed improved clustering accuracy by considering feature domains when transcoding categorical features. This approach ensures correct representation of features, which significantly affects the quality of subsequent stages of analysis. As a result, clustering becomes more structured, with clear boundaries between groups, which confirms the effectiveness of data preparation in this model. The K-Means clustering method allowed creating six compact clusters that show a clear grouping of most data. A slight overlap between some clusters (for example, between 2 and 4) indicates data complexity, but the overall result of clustering is balanced and logical. The model provided efficient segmentation by identifying natural patterns in the data. Agglomerative clustering also showed a noticeable difference: in Model 2, five distinct groups were isolated with less overlap due to the domain nature of the features. This allowed better isolating the data while maintaining its uniformity. In Model 1, partial mixing between groups was more pronounced, which reduced the accuracy of the results. Thus, consideration of feature domains in Model 2 not only improved the quality of clustering, but also made data segmentation more logical and consistent with their natural structure. This highlights the importance of correctly encoding variables to achieve high-quality analysis.

Analysis of Model 3 showed that the sequence of stages – allowed focusing on clusters that best correspond to the original data structure. This approach provides clear grouping while preserving important information about variables. The results of clustering using the K-Means method show the development of six clusters. Although cluster overlap is observed in multiple zones, this effect is due to the natural complexity of the data. Despite this, the method provided high-quality grouping of points into compact groups, which indicates the effectiveness of segmentation and identification of patterns. Agglomerative clustering distinguishes four clusters with a well-structured data distribution. Most clusters have clear boundaries and show a high level of internal cohesion. Small areas of overlap indicate similarity of characteristics between neighbouring groups, but the overall segmentation is clear and logical, which confirms the strength of the method in analysing complex data. Thus, Model 3 provides reliable segmentation of data, helping to preserve its natural structure. This highlights the effectiveness of selecting a sequence of analysis steps and the importance of considering the nature of variables when building a clustering model.

This study demonstrated that a flexible and semantically sound approach to coding – considering the nature of features, their hierarchy, scale, and domain value – can improve data segmentation in clustering tasks. Processing categorical variables is a critical step for classification, clustering, and data analysis tasks. As shown in a large-scale study by F. Matteucci *et al.* (2023), which covered 32 coding methods (including One-Hot, ordinary, Mean-Target, GLMM, Min-Hash, CatBoost, and Sum), the choice of encoder has a significant impact on the quality of results, and its effectiveness depends on a combination of experimental factors: model type, quality metrics, hyperparameter tuning strategy, and how results are aggregated. This highlights the need for a flexible and meaningful approach to coding that considers the nature of features, their hierarchy, and task specifics. However, instead of fully automated solutions, such as the CESAMO method (Valdez-Valenzuela *et al.*, 2024), which automates statistically optimised category coding, retained control over correspondence between categories and their numerical representations. This provided greater transparency and control over the interpretation of results, which is especially important in a limited data set. The importance of adapting coding methods to the characteristics of data was confirmed in research conducted by D. Breskuvienė & G. Dzemyda (2023), who analysed the impact of various coding techniques on classification quality in problems of detecting fraudulent transactions with highly unbalanced categorical data. The researchers pointed out that targeted encoders, such as Weight of Evidence and James-Stein, can significantly improve the performance of models, but the effectiveness of individual methods, such as CatBoost, decreases with a strong imbalance, which emphasises the need to choose an encoding based on the specifics of the task and data structure.

Confirmation of the importance of an adaptive approach to working with categorical variables can be found in the review paper by T. Dinh *et al.* (2024), devoted to the evolution of categorical data clustering methods over the past 25 years. The researchers noted that successful clustering of categorical variables requires not only the correct choice of algorithm, but also careful pre-coding, considering the nature of the variables, otherwise there is a risk of losing important semantic

connections. This approach is fully consistent with the approach in this paper, which combined different transcoding methods depending on the feature content and data logic. Preservation of category semantics allowed for more structured and interpreted clustering.

The problem of clustering categorical and mixed data types was also actively considered in many studies of the last 5 years. In particular, C. Di Nuzzo (2024) focused on improving spectral clustering for such data types by developing new nuclear functions that better consider the specifics of categorical relationships. The paper showed that standard Euclidean approaches may not be sufficient to adequately map internal structures in mixed sets, which is consistent with the observations in this study. For example, in Model 3 in this study, the use of clustering before the dimension reduction step contributed to the preservation of the natural feature structure. Although the researcher focused on spectral clustering, the results highlighted the importance of adapting clustering methods to the nature of variables – an approach that was also followed in model development.

In this context, it is also worth mentioning the study by D. Soemitro & J.F.S.R. Neto (2024), where an alternative approach to spectral clustering of categorical and mixed data was proposed. The researchers avoided conventional preprocessing, particularly one-hot or multi-binary, and instead suggested adding additional nodes to the graph corresponding to feature categories. This approach allowed considering both numerical and categorical information without losing the semantics of features, and also provided interpretability of clustering. This is especially true when using the MultiLabelBinarizer technique, which, although convenient, can lead to excessive dimension and loss of links between categories. The approach proposed by the researchers demonstrated the competitive quality of clustering and, importantly, allowed achieving linear complexity for the case of purely categorical data. Their study reinforced the argument for adaptive method selection based on data type and feature structure.

The choice of encoding method for categorical variables is crucial for the effectiveness of machine learning models, especially in the case of variables with a large number of levels. In a large benchmark study by F. Pargent *et al*. (2022) showed that target encoding, in particular, its regularised versions, provides better predictive results compared to conventional approaches such as ordinal or one-hot encoding. This confirms the importance of choosing an informed coding method based on the structure of variables, which is consistent with the approach to combining multiple coding strategies according to the nature of features.

H.L. Smith *et al*. (2024) drew attention to the potential distortions that can occur when using target encoding in combination with tree models. The researchers have shown that when using target-based coding to construct an ensemble of out-of-bag evaluation (OOB), they can significantly underestimate the classification error and overestimate the importance of variables. These results highlighted the limitations of target coding and the importance of step-by-step control of the feature processing process, an aspect that was considered when designing models and determining the sequence of their processing.

Another important aspect was the consistency between the choice of coding method and the type of machine learning model. A recent study by W. Zhu *et al*. (2024) contains both theoretical and empirical analysis of 14 methods for encoding categorical variables in combination with various types of models. The researchers proved that one-hot encoding is optimal for models that perform affine transformations (for example, multilayer perceptrons), while target encoding and its variants are best suited for models based on decision trees (for example, Random Forest). These results are fully consistent with the observations obtained from this study, that the correspondence between the coding method and the model type significantly affects the final result.

Thus, the approach of this study was based on proven methods of processing categorical variables and at the same time adapts the latest concepts, such as contextual coding, combined coding by variable type and adaptive order of processing stages. It was confirmed that there is no universal method: the effectiveness depends on both the type of problem and the nature of the features. The results obtained coincide with trends in data analysis.

## CONCLUSIONS

In the course of the study, a detailed analysis of three different clustering models was carried out, which differ in approaches to processing and transcoding input categorical data. The results showed that the quality, structure, and clarity of the developed clusters largely depend on the chosen data preparation methods, and on the sequence of their application. This highlights the importance of carefully selecting and combining methods to achieve optimal results in segmentation problems.

Model 1, in which categorical features were transcoded without considering their domain nature, provided basic data separation into six clusters for both K-Means and agglomerative clustering. However, the fuzzy boundaries between clusters and partial overlap indicate that this approach is limited in displaying a complex data structure. This indicates that the method is not sufficiently adapted to the features of categorical variables, which negatively affects the accuracy of segmentation and reduces the quality of clustering in general.

Model 2, in which the transcoding of categorical variables was carried out considering their domain nature, significantly improved clustering results. In this model, six distinct clusters were obtained using the

K-Means method and five clusters using agglomerative clustering. The improvement in clustering quality was confirmed by higher metric values and reduced overlap between groups. This result demonstrated the effectiveness of correct transcoding, which better reflects the internal structure of multidimensional data and contributes to more accurate and logical segmentation.

Model 3, which assumed a change in the sequence of steps – performing clustering before reducing the dimension – showed the best ability to preserve the natural data structure. The K-Means method identified six clusters, while agglomerative clustering identified four clearly structured clusters. A small overlap in some areas does not prevent the preservation of important relationships between features, which is necessary for analysing complex multidimensional sets. This approach highlighted the importance of the correct sequence of analytical steps to preserve key data characteristics.

The results of the study confirmed that the choice of approach to transcoding categorical variables, and the number of clusters, directly affect the cluster structure of data. Different coding strategies can form a different number of clusters, and therefore, create fundamentally different segmentation schemes. Thus, the correct choice of encoding method, considering the domain specifics of features, and the sequence of processing are critical factors for achieving high clustering quality. The results obtained open up prospects for further improvement of algorithms for segmenting complex data and their application in various applied fields – from marketing and healthcare to social research and bioinformatics.

## ACKNOWLEDGEMENTS
None.

## FINANCING
None.

## CONFLICT OF INTEREST
None.

## REFERENCES
[1] Anitha, M., Savarimuthu, N., & Bhanu, S. (2025). Chi-square target encoding for categorical data representation: A real-world sensor data case study. *SN Computer Science*, 6, article number 228. doi: 10.1007/s42979-025-03766-z.
[2] Ashfaq, V.A. (n.d.). *Student mental health survey*. Retrieved from https://www.kaggle.com/datasets/abdullahashfaqvirk/student-mental-health-survey/data.
[3] Behzadidoost, R., & Izadkhah, H. (2025). Identifying effective algorithms and measures for enhanced clustering quality: A comprehensive examination of arbitrary decisions in hierarchical clustering algorithms. *Journal of Classification*, 42, 457-489. doi: 10.1007/s00357-025-09506-5.
[4] Breskuvienė, D., & Dzemyda, G. (2023). Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions. *International Journal of Computers Communications & Control*, 18(3). doi: 10.15837/ijccc.2023.3.5433.
[5] Di Nuzzo, C. (2024). Advancing spectral clustering for categorical and mixed-type data: Insights and applications. *Mathematics*, 12(4), article number 508. doi: 10.3390/math12040508.
[6] Dinh, T., Hauchi, W., Fournier-Viger, P., Lisik, D., Ha, M.-Q., Dam, H.-C., & Huynh, V.-N. (2024). Categorical data clustering: 25 years beyond K-modes. *ArXiv*. doi: 10.48550/arXiv.2408.17244.
[7] Hafid, H., & Annisa, S. (2025). Implementation of K-medoids and K-prototypes clustering for early detection of hypertension disease. *Barekeng: Journal of Mathematics and Its Application*, 19(1), 465-476. doi: 10.30598/barekengvol19iss1pp465-476.
[8] Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178-210. doi: 10.1016/j.ins.2022.11.139.
[9] Kondruk, N.E. (2019). A comparative study of cluster validity indices. *Radio Electronics, Computer Science, Control*, 4, 59-67. doi: 10.15588/1607-3274-2019-4-6.
[10] Kondruk, N.E. (2023). Analysis of dimensionality reduction techniques in machine learning. *Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics*, 42(1), 181-187. doi: 10.24144/2616-7700.2023.42(1).181-187.
[11] Liang, Z. (2025). Efficient representations for high-cardinality categorical variables in machine learning. *ArXiv*. doi: 10.48550/arXiv.2501.05646.
[12] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129-137. doi: 10.1109/TIT.1982.1056489.
[13] Matteucci, F., Arzamasov, V., & Böhm, K. (2023). A benchmark of categorical encoders for binary classification. In *Advances in neural information processing systems 36 (NeurIPS 2023)* (pp. 54855-54875). doi: 10.48550/arXiv.2307.09191.
[14] Miyamoto, S. (2022). *Theory of agglomerative hierarchical clustering*. Singapoure: Springer Nature. doi: 10.1007/978-981-19-0420-2.

[15] Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, 37(5), 2671-2692. doi: 10.1007/s00180-022-01207-6.
[16] Sánchez Vinces, B.V., Schubert, E., Zimek, A., & Cordeiro, R.L. (2025). A comparative evaluation of clustering-based outlier detection. *Data Mining and Knowledge Discovery*, 39(2), article number 13. doi: 10.1007/s10618-024-01086-z.
[17] Sieranoja, S., & Fränti, P. (2025). Fast agglomerative clustering using approximate traveling salesman solutions. *Journal of Big Data*, 12(1), article number 21. doi: 10.1186/s40537-024-01053-x.
[18] Smith, H.L., Biggs, P.J., French, N.P., Smith, A.N., & Marshall, J.C. (2024). Out of (the) bag – encoding categorical predictors impacts out-of-bag samples. *PeerJ Computer Science*, 10, srticle number e2445. doi: 10.7717/peerj-cs.2445.
[19] Soemitro, D., & Neto, J.F.S.R. (2024). Spectral clustering of categorical and mixed-type data via extra graph nodes. *ArXiv*. doi: 10.48550/arXiv.2403.05669.
[20] Tokuda, E.K., Comin, C.H., & Costa, L.D.F. (2022). Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and Its Applications*, 585, article number 126433. doi: 10.1016/j.physa.2021.126433.
[21] Valdez-Valenzuela, E., Kuri-Morales, A., & Gomez-Adorno, H. (2024). Statistical evaluation of categorical encoders for pattern preservation in machine learning tasks. *International Journal of Combinatorial Problems and Informatics*, 15(2), 160-172. doi: 10.61467/2007.1558.2024.v15i2.456.
[22] Wegmann, M., Zipperling, D., Hillenbrand, J., & Fleischer, J. (2021). A review of systematic selection of clustering algorithms and their evaluation. *ArXiv*. doi: 10.48550/arXiv.2106.12792.
[23] World Medical Association's Declaration of Helsinki. (2013). Retrieved from https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki.
[24] Zhu, W., Qiu, R., & Fu, Y. (2024). Comparative study on the performance of categorical variable encoders in classification and regression tasks. *ArXiv*. doi: 10.48550/arXiv.2401.09682.

# Дослідження впливу різних технік кодування категоріальних ознак на структури кластерів

**Наталія Кондрук**

Кандидат технічних наук, доцент
Ужгородський національний університет
88000, вул. Університетська, 14, м. Ужгород, Україна
https://orcid.org/0000-0002-9277-5131

**Інна Нерода**

Аспірант
Ужгородський національний університет
88000, вул. Університетська, 14, м. Ужгород, Україна
https://orcid.org/0000-0002-9277-5131

**Анотація.** Категоріальні ознаки є поширеним типом даних, що використовуються у практиці аналізу даних, проте їх неметричний характер створює труднощі для застосування стандартних алгоритмів кластеризації. Актуальність дослідження зумовлена необхідністю оцінки впливу різних методів перекодування (оцифровування) таких ознак на результативність кластерного аналізу. Метою роботи було дослідити, як різні техніки обробки категоріальних даних впливають на якість та структуру кластерів. Методологія включала реалізацію трьох моделей з різними підходами до кодування змінних: без урахування доменної специфіки, з урахуванням змісту ознак та з чергуванням порядку застосування підходів кластеризації і зменшення розмірності. Для кодування використовувалися LabelEncoder, OrdinalEncoder, One-Hot Encoding, Mapping і MultiLabelBinarizer. У кожній із моделей кластеризація здійснювалася з використанням двох алгоритмів – K-Means та агломеративної кластеризації, що дозволяло порівняти їхню чутливість до змін у представленні даних. Метод зниження розмірності t-distributed Stochastic Neighbor Embedding (t-SNE) застосовувався для візуалізації кластерної структури у двовимірному просторі. Якість кластеризації оцінювалася за допомогою метрик Silhouette Score, Dunn Index, Davies-Bouldin Index та Calinski-Harabasz Index. Дані для аналізу було отримано з відкритого джерела й вони містили інформацію про психоемоційний стан студентів. У ході дослідження було встановлено, що базове перекодування категоріальних ознак без урахування їхньої семантики та контексту негативно впливало на якість кластеризації, знижуючи точність поділу та ускладнюючи інтерпретацію результатів. Натомість використання доменно-орієнтованих підходів до кодування забезпечувало формування кластерів із чіткішими межами та логічнішою внутрішньою структурою. Додатково було виявлено, що зміна послідовності застосування кластеризації та редукції розмірності позначається на збереженні локальних взаємозв'язків у даних. Проаналізовано, що різні підходи змінюють як кількість, так і якість кластерів, що відображається у значеннях оцінкових метрик. Практична цінність результатів полягає у можливості їх застосування фахівцями з аналізу даних та машинного навчання для підвищення точності сегментації складних категоріальних даних

**Ключові слова:** аналіз даних; машинне навчання; навчання без учителя; автоматичне групування об'єктів; сегментація