European Research Studies Journal
Volume XXVII, Special Issue 2, 2024

pp. 72-82

Use of Autoencoder and One-Hot Encoding for Customer Segmentation

Submitted 18/02/24, 1st revision 16/03/24, 2nd revision 20/04/24, accepted 16/05/24

Tomasz Smutek[1], Jan Sikora[2], Sylwester Bogacki[3], Marek Rutkowski[4], Dariusz Woźniak[5]

Abstract:

Purpose: *The research article aims to apply Autoencoder and One-Hot Encoding techniques to the segmentation of retail customers, exploring how these methodologies can contribute to a more refined and actionable segmentation process.*
Design/Methodology/Approach: *The study uses a dataset comprising detailed profiles of 2240 retail customers, applying Autoencoders and One-Hot Encoding to categorize customers into distinct segments. It evaluates Autoencoders' embeddings and compares them with the traditional One-Hot Encoding method. The effectiveness of the segmentation is further analyzed using various clustering algorithms, including K-means, DBSCAN, Louvain Community Detection, Greedy Modularity, and Label Propagation. The research assesses clustering quality using indices such as the Caliński-Harabasz, Davies-Bouldin, and modularity metrics.*
Findings: *Application of the Louvain method with a cut-off parameter of 0.75 using AutoEmbedder revealed three evenly distributed customer groups, albeit with slightly lower Caliński-Harabasz and Davies-Bouldin index values than those obtained by the Greedy method using AutoEmbedder with a cut-off parameter of 0.5. However, the Louvain method exhibited higher modularity, indicating more cohesive segmentation. Comparisons between AutoEmbedder and One-Hot Encoding suggested the superiority of AutoEmbedder in forming customer clusters.*
Practical Implications: *The findings present actionable insights for marketing strategists to develop targeted campaigns based on customer expenditure patterns. By identifying customer segments with similar attributes, businesses can allocate marketing resources more effectively and tailor strategies to meet the specific needs of each segment.*
Originality/Value: *The article introduces a novel comparison between Autoencoder embeddings and traditional One-Hot Encoding in the context of customer segmentation, providing evidence of the former's enhanced capability in creating more meaningful and modular customer groups. It also extends the discussion on clustering quality assessment in the segmentation process, adding value to marketing analytics.*

[1]*Corresponding Author: WSEI University, Lublin, Poland,*
e-mail: Tomasz.Smutek@wsei.lublin.pl;
[2]*WSEI University, Lublin, Poland, e-mail: Jan.Sikora@wsei.lublin.pl;*
[3]*WSEI University, Lublin, Poland, e-mail: Sylwester.Bogacki@wsei.lublin.pl;*
[4]*WSEI University, Lublin, Poland, e-mail: Marek.Rutkowski@wsei.lublin.pl;*
[5]*Wyższa Szkoła Biznesu - National Louis University, e-mail: dwozniak@wsb-nlu.edu.pl;*

Keywords: *Autoencoder, One-Hot Encoding, customer segmentation, machine learning, clustering algorithms.*

JEL codes: *C45, C53, D12, L11, L86, M15.*

Paper type: *Research article.*

1. Introduction

Recent advancements in data science have significantly improved our ability to process and interpret large data sets, a trend that has become increasingly evident in contemporary research. The strategic incorporation of machine learning into customer data analysis has substantially deepened businesses' understanding of diverse consumer demographics.

As outlined in scholarly work, the rise of big data has been instrumental in enhancing consumer group segmentation and identifying key market segments, thereby refining marketing strategies and tailoring them to specific consumer behaviors and needs (Galiano and Coronil, 2022).

The implementation of advanced analytical techniques like GRU networks and machine learning frameworks has significantly enhanced marketing adaptability and innovation, ensuring that business models are closely aligned with consumer preferences (Rymarczyk, Bednarczuk, *et al.*, 2021; Rymarczyk, Golabek, *et al.*, 2021).

The utilization of complex machine learning algorithms such as K-means and DBSCAN has been extensively documented for their effectiveness in deciphering extensive data collections, with literature suggesting a combination of multiple clustering methods for optimal segmentation (Hicham and Karim, 2022).

Innovations such as AutoEmbedder have emerged as critical tools for translating categorical variables into vector representations and streamlining data interpretation. This technique maintains the integrity of relational data while addressing the common challenge of high dimensionality in data analysis (Dahouda and Joe, 2021; Rachwał *et al.*, 2023; Czainska *et al.*, 2021; Kadlubek *et al.*, 2022).

Similarly, Principal Component Analysis (PCA) remains a staple statistical method that simplifies complex datasets by transforming them into new variables or principal components, thereby facilitating pattern recognition with minimal loss of information (Abdulhafedh, 2021; Naveen *et al.*, 2022; Polyakova *et al.*, 2019).

Moreover, sophisticated clustering techniques, such as the Louvain Community Detection and Greedy Modularity, have been examined for their ability to identify

complex patterns within networks, which can be analogous to segmenting customers based on common attributes (Gonzalez-Montesino, Grass-Boada and Armannazas, 2023; Rustamaji *et al.*, 2024). Such methods have been proven to uncover patterns that might not be detectable by traditional analysis methods.

The evaluation of clustering algorithms is also a burgeoning field of study, with metrics such as the Caliński-Harabasz index, the Davies-Bouldin index, and modularity measures providing insights into the quality of clustering results (Wei *et al.*, 2021; Daraghmeh, Agarwal and Jararweh, 2023; Bihari *et al.*, 2024).

Assessing clustering algorithms via diverse indices and metrics constitutes a growing research field, especially relevant for multifaceted and sizable datasets spanning areas from bioinformatics to social network analysis. Prominent among these measures are the Caliński-Harabasz index and the Davies-Bouldin index, alongside modularity metrics, all offering perspectives on the clustering results' caliber. Analyses of various algorithms, focusing on how these metrics convey each method's efficiency across different data compilations, are critical to this discussion.

This paper aimed to apply and evaluate the effectiveness of the Autoencoder and One-Hot Encoding methodologies for retail customer segmentation. The purpose of the study is twofold: to apply these methods to create distinct customer segments based on detailed profile data and to analyze these segments using different clustering algorithms.

This research's innovation lies in comparing Autoencoder and One-Hot Encoding techniques to form a more sophisticated understanding of customer groups. Graph-based clustering algorithms, such as Louvain Community and Greedy Modularity, were applied to the embeddings generated by the Autoencoders. These advanced techniques, examined alongside traditional clustering methods such as K-means and DBSCAN, provide a comprehensive analysis of the patterns within consumer data.

2. Materials and Methods

To evaluate the effectiveness of the clustering methodology, a dataset including retail customer data from (Customer Personality Analysis, no date) was selected for analysis. The retail dataset, which formed the basis for the clustering algorithm tests, included a sample size of 2,240 individuals.

The data features detailed customer profiles, encompassing Customer ID, birthdate, educational background, marital status, annual earnings, household child count, recency of last purchase, presence of any complaints, and expenditure on diverse categories such as wine, fruit, meats, fish, confectionery, and precious metals over a biennial period. It further details consumer behavior, including discounted purchase frequency, promotional engagement levels, e-commerce interactions, catalog utilization, in-store transaction history, and recent web visitation frequency.

The purpose of the research conducted was to apply Autoencoder and One-Hot Encoding for customer segmentation. Autoencoders are neural networks designed to encode input data into what is known as a hidden representation and then decode it in such a way that the reconstructed data is as similar as possible to the original. The Autoencoder consists of two parts: an encoder and a decoder. The encoder function *f* maps the input data x_i to the hidden representation $h(x_i)$:

$$h = f(x_i) = a_f(W x_i + b_h)$$

The above equation calculates the projection of x_i onto the hidden space *h*. The decoder function gg maps the hidden representation *h* back to the reconstruction *r*:

$$r = g(h) = a_g(W'h + b_r)$$

where a_f and a_g are the activation functions of the encoder and decoder, respectively. This equation computes the reconstruction of the input data (Diallo *et al.*, 2021).

Autoencoders are widely used in unsupervised machine-learning tasks. One common task in unsupervised learning is clustering unlabeled data, according to Ohi et al. (Ohi *et al.*, 2020) demonstrate with a dataset of images the structure of AutoEmbedder, which extracts features from high-dimensional data and compresses them into an embedding point of lower dimension where clustering can be performed.

A Deep Convolutional Neural Network (DCNN) architecture accomplishes dimension reduction based on a backpropagation distance loss calculation generated from a Siamese network architecture. This algorithm requires knowledge of whether selected pairs of observations belong to the same group.

It does not compute a loss function for clusters; instead, it minimizes the Euclidean distance between elements belonging to the same group, which is reduced during the backpropagation process. Most embedding generation methods require knowledge of the target variable. An autoencoder eliminates this need, as the loss function is calculated by comparing the reconstruction r with the input data x.

One-hot encoding is a method of representing categorical variables. It applies to qualitative variables that do not allow for an ordinal relationship. An example of such a variable is "color" with potential levels such as "red", "green", and "blue". One-hot encoding transforms each value of the categorical variable into a new column consisting of binary values.

Each observation is assigned a value of 1 in precisely one of the newly created columns. For instance, in the case of the variable "color", after applying One-Hot

Encoding, three new columns would be made: red, green, and blue. An observation that previously had the value "green" will now have a value of 1 in the new "green" column and zeros in the others. This example is depicted in Figure 1.

***Figure 1**. An example of One-Hot Encoding*

ID	Color
N1	red
N2	green
N3	blue
N4	green

ID	Color_red	Color_green	Color_blue
N1	1	0	0
N2	0	1	0
N3	0	0	1
N4	0	1	0

***Source:** Own creation.*

The newly created binary columns are referred to as dummy variables. These columns are linearly dependent (for example, if the first column has a value of 1, it is understood that the remaining columns must have a value of 0); hence, it is generally advisable to remove one of these columns—typically the first—to avoid issues of multicollinearity.

The presentation of the proposed methodology for customer segmentation shall ensue upon clarifying all necessary concepts. The Autoencoder method should be particularly well-suited for dealing with categorical variables. Primarily, the application is most advantageous when the dataset contains categorical variables.

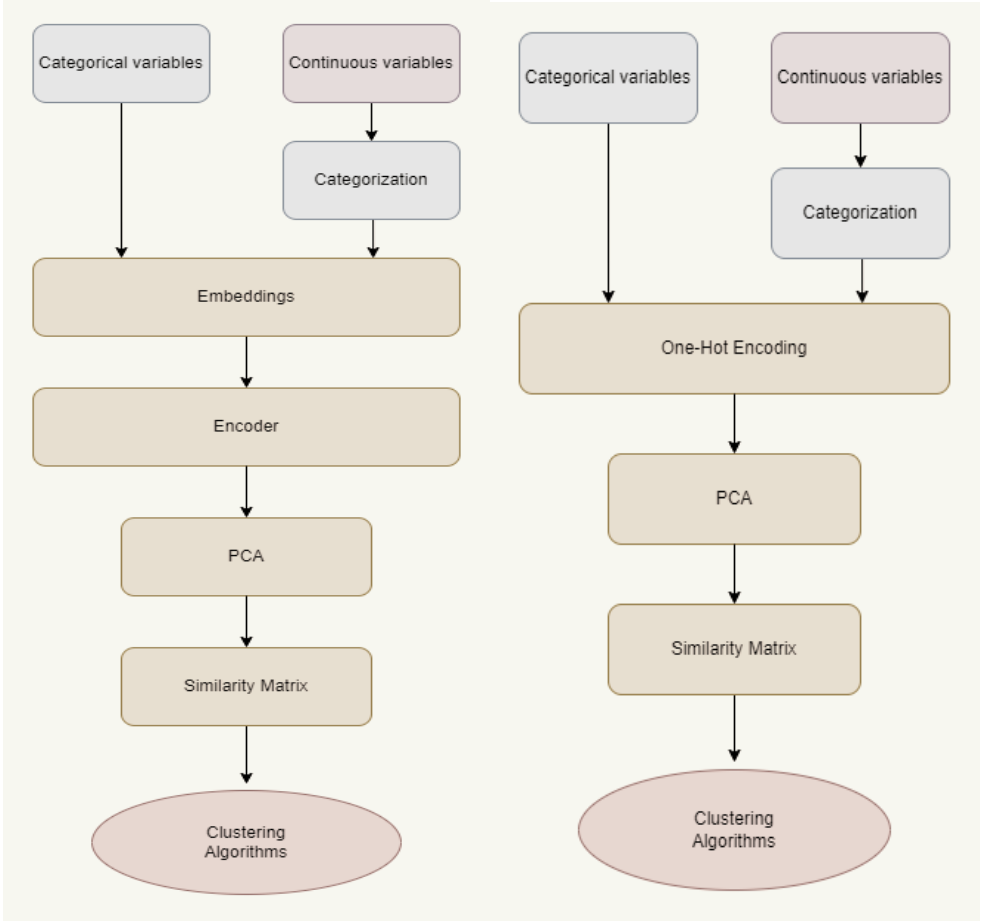
The central issue we address with our solution is the generation of embeddings for categorical variable values. This has been facilitated by a structure known as AutoEmbedder.

The AutoEmbedder was trained on a learning dataset. Upon training, the AutoEmbedder utilizes a layer and an encoder for embeddings. Principal Component Analysis (PCA) is subsequently performed for the encoded data to ensure that 95% of the variance remains explained.

Following this procedure, each observation is represented as a vector of real numbers. We generate a similarity matrix for such encoded observations when selecting an appropriate similarity measure. Subsequently, we can move on to selecting specific clustering algorithms, including graph clustering algorithms, among others.

The methodology is graphically presented with the aid of Figure (2a). This method is compared with a more classical approach, wherein the values of categorical variables are encoded using One-Hot Encoding instead of AutoEmbedder. The remainder of the procedure remains unchanged. Figure (2b) provides a graphical presentation of the methodology utilizing One-Hot Encoding.

Figure 2. *Proposed methodology: a) Autoencoder, b) One-Hot Encoding*



a) *b)*

Source: *Own creation.*

Cosine similarity, scaled from 0 to 1, was chosen as the similarity metric. To identify clusters of customers with comparable characteristics, the algorithms used included K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and three graph-based techniques: Louvain Community, Greedy Modularity, and Label Propagation.

Constructing a graph depicting customer relationships in the dataset was crucial for implementing these graph-based methods. The graph was built from the similarity matrix with the condition that it must be connected. Creating edges between similar customers was managed using a cutoff parameter. An edge was drawn between customers whose similarity met or exceeded the cutoff value. Cutoff values of 0.25, 0.5, and 0.75 were evaluated.

Clusters generated were compared using metrics designed for unsupervised clustering. The Calinski-Harabasz and Davies-Bouldin indices were employed to identify the optimal clustering. Normalized Mutual Information (NMI) and the Fowlkes-Mallows index were utilized to measure the similarity between two clusterings, and modularity was examined for graph-based methods.

The Elbow method, silhouette criterion, and Calinski-Harabasz and Davies-Bouldin indices helped determine the ideal number of clusters for the K-means algorithm. For DBSCAN, parameters were chosen based on the silhouette criterion.

3. Results

In the initial phase of this study, clustering techniques were applied to a customer dataset using AutoEmbedder. The encoding process encountered no errors, and after using AutoEmbedder, the dataset comprised 96 columns. Principal Component Analysis (PCA) was then employed, reducing the number of columns to eight.

With a cut-off parameter of 0.75, the Louvain method identified three groups: group 0 contained 565 customers, group 1 had 929 customers, and group 2 included 746 customers. The Greedy Modularity and Label Propagation methods divided the dataset into two groups. In both methods, group 0 had 1285 customers and group 1 had 955, with similar partitions. Their similarity was assessed using the Fowlkes-Mallows index (~0.98) and Normalized Mutual Information (NMI, ~0.94).

Reducing the Louvain method's cut-off parameter to 0.5 resulted in three groups with less balanced distribution. The Greedy Modularity method yielded two groups: group 0 had 1200 customers, while group 1 had 1040. The Label Propagation method found just one group containing all customers. Lowering the cut-off to 0.25 resulted in two Louvain method groups (group 0: 1048 customers, group 1: 1192).

The Greedy Modularity method similarly formed two groups (group 0: 1131 customers, group 1: 1109). In contrast, the Label Propagation method grouped all customers. The Louvain method with a 0.75 cut-off resulted in the highest modularity (0.566). Using the DBSCAN method, three groups were found (group 0: 950 customers, group 1: 831 customers, group 2: 16 customers), while 443 customers remained unclassified.

AutoEmbedder and One-Hot Encoding were used to compare the clustering for customer grouping. With a cut-off of 0.75, the graph failed to maintain coherence. At a cut-off of 0.5, the Louvain method identified three groups (group 0: 1156 customers, group 1: 1077, group 2: 7). The Greedy Modularity method resulted in two groups (group 0: 1141, group 1: 1099), and the Label Propagation method grouped all customers.

At a cut-off of 0.25, the Louvain method formed three groups, while the Greedy Modularity and Label Propagation methods each resulted in a single group. The highest modularity values (0.336) were achieved with the Louvain and Greedy Modularity methods using a 0.5 cut-off. The DBSCAN method formed three groups, leaving 426 customers unclassified. Graph-based methods and DBSCAN were compared with the K-means method, which identified two groups.

Cluster quality was assessed using the Calinski-Harabasz and Davies-Bouldin indices. The best results for both indices were achieved using the Greedy Modularity method with a 0.5 cut-off, which had values comparable to the K-means method. The NMI and Fowlkes-Mallows parameters measured cluster similarity across different methods. Cluster quality varied, and Table 1 summarizes Calinski-Harabasz and Davies-Bouldin index results across various methods and cut-off parameters.

Table 1. *Values of Calinsky-Harabash and Davies Bouldin indices for different methods and different cutoff parameters*

Clustering	Parameter cut-off	Method	Calinsky Harabash Index	Davies Bouldin Index
AutoEmbedder	0.25	Louvain Community	801.783	1.626
	0.25	Greedy Modularity	796.954	1.641
	0.25	Label Propagation	-	-
	0.5	Louvain Community	415.123	2.358
	0.5	Greedy Modularity	782.355	1.644
	0.5	Label Propagation	-	-
	-	DBSCAN	314.034	2.484
	0.75	Louvain Community	542.745	2.080
	0.75	Greedy Modularity	773.752	1.624
	0.75	Label Propagation	781.243	1.625
One-Hot Encoding	0.25	Louvain Community	148.362	3.287
	0.25	Greedy Modularity	-	-
	0.25	Label Propagation	-	-
	-	DBSCAN	102.272	4.673
	-	K-means	249.258	2.873
	0.5	Louvain Community	133.536	2.736
	0.5	Greedy Modularity	264.791	2.749
	0.5	Label Propagation	-	-

Source: *Own creation.*

When comparing approaches utilizing AutoEmbedder to One-Hot Encoding, the values of the Calinski-Harabasz index are higher with the use of AutoEmbedder, while the Davies-Bouldin index values are lower. This leads to the inference that the clusters formed through the use of AutoEmbedder are superior.

Table 2. *Modularity parameter for different graph algorithms and different cutoff parameters on a set of retail customers*

Clustering	Parameter cut-off	Method	Modularity
One-Hot Encoding	0.5	Louvain Community	0.336
	0.5	Greedy Modularity	0.336
	0.25	Louvain Community	0.063
AutoEmbedder	0.75	Louvain Community	0.566
	0.75	Greedy Modularity	0.463
	0.75	Label Propagation	0.409
	0.5	Louvain Community	0.309
	0.5	Greedy Modularity	0.309
	0.25	Louvain Community	0.205
	0.25	Greedy Modularity	0.202

Source: *Own creation.*

Table 2 shows the Modularity parameter values for different graph methods and cutoff parameters. It becomes clear that clustering with AutoEmbedder leads to higher Modularity values. The Louvain method with a cutoff of 0.75 delivers the highest value.

Using an autoencoder allowed the identification of three evenly distributed groups (via the Louvain method at a 0.75 cutoff). While this clustering has slightly lower Calinski-Harabasz and Davies-Bouldin Index values than the Greedy method with AutoEmbedder and a 0.5 cutoff, it achieves a higher Modularity parameter value.

4. Conclusions

This study explored the application of Autoencoder and One-Hot Encoding in the segmentation of retail customers. By leveraging these methods, it was possible to develop a nuanced understanding of customer groupings, which could be instrumental in tailoring marketing strategies to distinct clusters. Three evenly distributed groups were discerned using the Louvain method with a cut-off parameter of 0.75.

Despite slightly lower Calinski-Harabasz and Davies-Bouldin indices scores than the Greedy method employing AutoEmbedder with a cut-off parameter of 0.5, this approach exhibited a superior modularity value. Such a feature highlights the robustness of the Louvain method in capturing the modular structure within the data, thus potentially revealing more cohesive and meaningful market segments.

The segmentation achieved offers a promising avenue for aligning marketing initiatives with customer clusters. For instance, the Louvain method suggests a customer division that allows marketing efforts to be differentiated based on the customers' spending on products, categorized into low, medium, and high expenditure.

This segmentation method is precious as it aligns marketing strategies with the spending behavior of different customer groups, enabling a more targeted and efficient allocation of marketing resources.

References:

Abdulhafedh, A. 2021. Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. Journal of City and Development, 3(1), 12-30. Available at: https://doi.org/10.12691/JCD-3-1-3.

Bihari, A., Vishwakarma, S., Kumar Bhardwaj, S., Tripathi, S., Agrawal, S., Joshi, P. 2024. Cancer Gene Clustering Using Computational Model. GMSARN International Journal, 18, 252-257.

Customer Personality Analysis (no date). Available at: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis.

Czainska, K., Sus, A., Thalassinos, E.I. 2021. Sustainable survival: Resource management strategy in micro and small enterprises in the rubber products market in Poland during the COVID-19 pandemic. Resources, 10(8), 85.

Dahouda, M.K., Joe, I. 2021. A Deep-Learned Embedding Technique for Categorical Features Encoding. IEEE Access, 9, 114381-114391. Available at: https://doi.org/10.1109/ACCESS.2021.3104357.

Daraghmeh, M., Agarwal, A., Jararweh, Y. 2023. An ensemble clustering approach for modeling hidden categorization perspectives for cloud workloads. Cluster Computing. Available at: https://doi.org/10.1007/S10586-023-04205-5.

Diallo, B., Hu, J., Li, T., Khan, G.A., Liang, X., Zhao, Y. 2021. Deep embedding clustering based on contractive autoencoder. Neurocomputing, 433, 96-107. Available at: https://doi.org/10.1016/J.NEUCOM.2020.12.094.

Galiano Coronil, A. 2022. Behavior as an approach to identifying target groups from a social marketing perspective. International Review on Public and Nonprofit Marketing, 19(2), 265-287. Available at: https://doi.org/10.1007/S12208-021-00298-Z/FIGURES/5.

Gonzalez-Montesino, L., Grass-Boada, D.H., Armannazas, R. 2023. Network Community Detection in Connectomics Data using Graph Theory. Proceedings, 2023 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, 3459-3465. Available at: https://doi.org/10.1109/BIBM58861.2023.10385337.

Hicham, N., Karim, S. 2022. Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering. International Journal of Advanced Computer Science and Applications, 13(10). Available at: www.ijacsa.thesai.org.

Kadłubek, M., Thalassinos, E.I., Domagała, J., Grabowska, S., Saniuk, S. 2022. Intelligent transportation system applications and logistics resources for logistics customer service in road freight transport enterprises. Energies, 15(13), 4668.

Naveen, S., Omkar, A., Goyal, J., Gaikwad, R. 2022. Analysis of Principal Component Analysis Algorithm for Various Datasets. 2022 International Conference on Futuristic Technologies, INCOFT 2022, 1-7. Available at: https://doi.org/10.1109/INCOFT55651.2022.10094448.

Ohi, A.Q., Mridha, M.F., Safir, F.B., Hamid, M.A., Monowar, M.M. 2020. AutoEmbedder: A semi-supervised DNN embedding system for clustering. Knowledge-Based Systems, 204, 106190. Available at: https://doi.org/10.1016/J.KNOSYS.2020.106190.

Polyakova, A.G., Loginov, M.P., Serebrennikova, A.I., Thalassinos, E.I. 2019. Design of a socio-economic processes monitoring system based on network analysis and big data. International Journal of Economics and Business Administration, 7(1), 30-139.

Rachwał, A., Popławska, E., Gorgol, I., Cieplak, T., Pliszczuk, D., Skowron, Ł., Rymarczyk, T. 2023. Determining the Quality of a Dataset in Clustering Terms. Applied Sciences, 13(5), 2942. Available at: https://doi.org/10.3390/APP13052942.

Rustamaji, H.C., Kusuma, W.A., Nurdiati, S., Batubara, I. 2024. Community detection with Greedy Modularity disassembly strategy. Scientific Reports, 14(1). Available at: https://doi.org/10.1038/S41598-024-55190-7.

Rymarczyk, P., Bednarczuk, P., Nowak, R., Cieplak, T. 2021. Methods of Analyzing Consumer Behavior Based on Multi-Source Data. European Research Studies Journal, Vol. 24, (Special Issue 2), 335-345. Available at: https://doi.org/10.35808/ERSJ/2229.

Rymarczyk, P., Golabek, P.S., Rzemieniak, M. 2021. Profiling and Segmenting Clients with the Use of Machine Learning Algorithms. European Research Studies Journal, Vol. 24, (Special Issue 2), 513-522. Available at: https://doi.org/10.35808/ERSJ/2281.

Wei, J., Ma, H., Liu, Y., Li, Z., Li, N. 2021. Hierarchical high-order co-clustering algorithm by maximizing modularity. International Journal of Machine Learning and Cybernetics, 12(10), 2887-2898. Available at: https://doi.org/10.1007/S13042-021-01375-9/TABLES/6.