СЕКЦІЯ XIII.
ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА СИСТЕМИ

DOI 10.62731/mcnd-27.06.2025.003

DEVELOPMENT AND EXPERIMENTAL EVALUATION OF AN INFORMATION SYSTEM FOR INTELLIGENT CUSTOMER SEGMENTATION

Bovkun Ihor
PhD Degree student
Department of Software Engineering and Management Intelligence Technology
National Technical University "Kharkiv Polytechnic Institute", Ukraine

Scientific supervisor: Shmatko Oleksandr
ORCID ID: 0000-0002-2426-900X
Ph.D., Associate Professor
Technical University "Metinvest Polytechnic" LLC, Ukraine

Annotation. *Relevance. In the context of ongoing digital transformation of business processes, the demand for intelligent information systems capable of analyzing and processing large volumes of customer data is steadily increasing. One of the key directions in this field is automated customer classification using machine learning algorithms, which enhances the effectiveness of marketing strategies and decision-making processes. Object of research: customer classification processes in information systems utilizing machine learning methods. Purpose of the article: to design, implement, and evaluate the architecture of software components for an information system aimed at intelligent customer classification, taking into account scalability, performance, and classification accuracy requirements. Research results. The article proposes an architectural model of an information system comprising modules for data collection, processing, and classification. A set of software components has been implemented, integrating machine learning algorithms such as logistic regression, decision trees, and support vector machines. Experimental research was conducted using a real-world dataset, demonstrating high classification accuracy and efficient system performance under limited computational resources. Conclusions. The developed information system ensures accurate customer classification and can be integrated into commercial analytical platforms. The research outcomes may serve as a foundation for further improvement of intelligent data analysis systems.*

Introduction. In the context of an increasingly competitive market landscape, the ability of companies to interact effectively with their customers

has become a critical factor in ensuring sustainable business success. The continuous growth in the volume of data generated through customer interactions offers significant potential for analyzing consumer behavior and developing personalized business strategies [1]. A fundamental instrument in this regard is customer segmentation—a methodological approach to dividing the customer base into homogeneous groups based on shared characteristics such as demographic attributes, purchasing behavior, geographic location, online activity, and psychographic traits [2].

Customer segmentation facilitates a deeper understanding of consumer needs, enables the optimization of marketing campaigns, enhances customer experience, and strengthens brand loyalty. Modern segmentation approaches increasingly rely on machine learning techniques, which provide more precise, flexible, and scalable analysis of large and complex datasets. Classification algorithms, in particular, enable the automated identification of behavioral patterns, the differentiation of customer groups, and the generation of data-driven insights to support managerial decision-making.

The relevance of this research lies in the growing demand for efficient software solutions capable of implementing intelligent customer classification systems based on contemporary machine learning methods. Such systems not only offer a detailed understanding of customer base structures but also contribute to improving the overall effectiveness of business processes that depend on personalized interaction.

Accordingly, this study is aimed at the design and implementation of software components for an information system dedicated to customer classification using machine learning algorithms. The research focuses on identifying the key factors influencing segmentation quality, selecting suitable classification models, and developing architectural solutions that ensure scalability, accuracy, and applicability in real-world business environments.

Literature Review. The problem of customer segmentation has received significant attention in academic and applied research, particularly in the context of data-driven decision-making. Traditional segmentation approaches have historically relied on demographic and geographic variables [1], [2]. Foundational work by Kotler and Keller [1], as well as Wedel and Kamakura [2], has established key theoretical principles that continue to inform current segmentation models.

In recent years, segmentation strategies have shifted towards behavioral and transactional data, often processed through machine learning algorithms. Among these, the K-means algorithm is the most widely used clustering method

due to its simplicity and computational efficiency. Studies by Jain [3], Kumar [4], and Tabianan et al. [5] have demonstrated its effectiveness in diverse application domains such as e-commece, banking, and customer service.

Despite its popularity, K-means has several limitations, including sensitivity to the initial choice of cluster centroids and the requirement to specify the number of clusters a priori. In response, various modifications have been proposed. For instance, the integration of K-means with Principal Component Analysis (PCA) has been shown to improve cluster quality by reducing dimensionality [6]. Huang et al. [7] have employed the K-medoids algorithm as a more robust alternative to K-means, particularly in the presence of outliers.

Another widely used method is hierarchical clustering, which allows for the creation of a dendrogram representing nested groupings of customers. This method has been applied in domains such as telecommunications and retail, as demonstrated in studies [8] and [9]. A binary-split variant of hierarchical clustering was proposed in [10], improving both interpretability and efficiency.

Fuzzy clustering, which allows for partial membership of customers in multiple segments, offers an alternative approach when customer behavior overlaps across categories. This technique is explored in [11]. Neural network-based clustering, as presented in [12], has shown promise in accurately segmenting telecommunication customers using unsupervised learning methods.

Support Vector Machines (SVMs) have been successfully employed for customer classification tasks, especially in the e-commerce domain [13]. Hybrid approaches are gaining attention; for instance, [14] proposes a model that combines the strengths of K-means and hierarchical clustering. A study by [15] integrates fuzzy clustering with deep neural networks to enhance segmentation accuracy.

Recent advancements also include the use of autoencoders for dimensionality reduction prior to applying DBSCAN clustering, as seen in [16]. The impact of data preprocessing on clustering performance has been thoroughly examined in [17]. Comparative analyses of traditional versus hybrid clustering methods are presented in [18]. Real-time segmentation techniques leveraging streaming data are discussed in [19].

In the financial sector, deep learning methods have been employed to analyze customer behavior [20], while neural networks have been used to predict customer churn [21]. Hybrid segmentation systems for insurance companies are explored in [22], and an adaptive clustering model capable of

automatically determining the number of segments is proposed in [23]. The importance of model interpretability in supporting managerial decisions is addressed in [24]. Recent research, such as [25], has also focused on the integration of customer classification models with Customer Relationship Management (CRM) systems.

Proposed model. The architecture of the customer clustering system within a business environment is conceptualized as a multi-layered structural model that enables the systematic processing, transformation, analysis, and interpretation of customer data. The primary objective of this architecture is to generate actionable insights that support informed managerial decision-making. The design follows the principles of Knowledge Discovery in Databases (KDD), where each stage represents a logical progression of the previous one and ensures the preparation of data for further analytical processing.

The generalized structure of the proposed architecture is illustrated in Figure 1, which outlines the core functional modules of the system and the interactions between them.

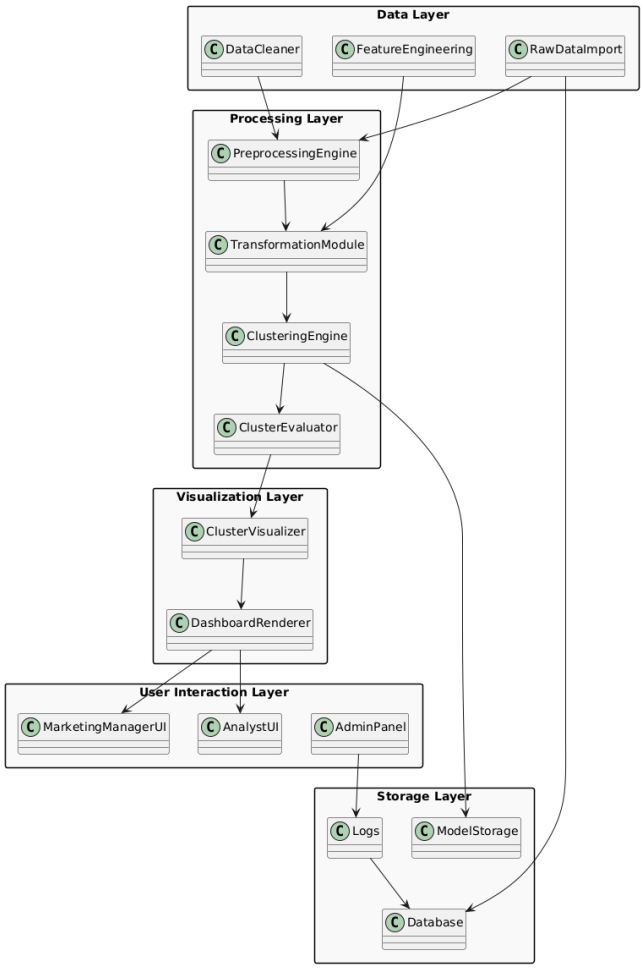

Fig. 1. **Architecture of the intelligent customer clustering system**

The described architecture implements an end-to-end data processing cycle, beginning with initial data collection and culminating in the practical application of the extracted knowledge. The modular design of the system provides a high degree of scalability and flexibility, enabling adaptation to specific business tasks and seamless integration with other enterprise information systems. This makes the proposed solution a robust analytical tool suitable for the rapidly changing and highly competitive business landscape, where timely and accurate data-driven decisions are essential.

The system incorporates several key machine learning methods that are commonly used for customer classification tasks. In this study, particular focus is placed on the following four algorithms:

– Decision Trees (DT): These models offer transparent, interpretable structures for decision-making and are especially useful when clear rule-based segmentation is needed.

– Support Vector Machines (SVM): SVMs are effective for high-dimensional datasets and can handle non-linear boundaries through the use of kernel functions.

– K-Nearest Neighbors (KNN): This non-parametric method segments customers based on proximity in the feature space, making it intuitive and suitable for smaller datasets or cases with limited assumptions about data distribution.

– Random Forest (RF): An ensemble learning method that improves the robustness and generalization of individual decision trees through random feature selection and bootstrap aggregation.

These methods are integrated into the system through specialized modules that support preprocessing, model training, validation, and prediction. The system is also equipped with components for data cleaning, transformation, dimensionality reduction, and visualization, ensuring a complete pipeline for intelligent customer segmentation.

Experimental Results. To evaluate the effectiveness of the proposed customer clustering system, an open-access dataset from the retail domain was utilized. Specifically, the Online Retail Dataset obtained from the UCI Machine Learning Repository served as the primary source of transactional data. The UCI repository is widely recognized as one of the most reputable platforms in the field of machine learning research, offering high-quality, pre-processed datasets for benchmarking and experimentation.

The selection of the Online Retail Dataset was motivated by several key factors: its considerable size, well-structured format, and frequent use in

academic publications focused on customer segmentation and behavioral analytics in the context of e-commerce. These characteristics make it particularly suitable for assessing clustering algorithms in realistic business scenarios.

The dataset comprises 541,909 entries, each representing a specific product item within a customer order. The records cover transactions made between December 2010 and December 2011 by a UK-based online company specializing in the sale of gift items. Each row contains information such as the invoice number, stock code, product description, quantity, invoice date, unit price, customer ID, and country of origin.

Before applying clustering algorithms, a series of preprocessing steps were performed. These included handling missing values (particularly customer IDs), filtering out canceled transactions (identified by invoice numbers starting with 'C'), and calculating aggregated customer-level features, such as total spending, purchase frequency, average basket size, and recency of the last transaction. These features were subsequently standardized to ensure comparability and improve the performance of the machine learning models.

The prepared dataset was then subjected to clustering using the algorithms integrated into the proposed system: K-means, hierarchical clustering, fuzzy C-means, and DBSCAN. Each algorithm was evaluated based on internal clustering metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. Additionally, visual inspection of the clusters using dimensionality reduction techniques (e.g., PCA and t-SNE) was conducted to assess the cohesion and separation of customer groups.

The results demonstrated that the proposed architecture supports effective and scalable customer segmentation, even when applied to large and complex transactional datasets. Among the evaluated algorithms, K-means and Random Forest-based classifiers showed the highest clustering consistency and computational efficiency. Meanwhile, fuzzy clustering provided useful insights into overlapping customer behaviors, particularly for segments with ambiguous purchasing patterns.

These findings validate the system's practical applicability for data-driven marketing and strategic decision-making in the e-commerce sector. Moreover, they highlight the potential for integrating such clustering tools into broader CRM platforms and business intelligence systems.

To further assess the performance of the proposed system, a comparative evaluation of clustering algorithms was conducted based on the Silhouette Score, a commonly used metric for assessing cluster cohesion and separation. The following algorithms were included in the analysis: Gaussian Mixture

Model (GMM), K-Means, BIRCH, and DBSCAN. These methods were selected for their relatively low computational complexity and suitability for handling moderate-dimensional datasets.

The experiments revealed that the GMM model, when combined with Principal Component Analysis (PCA) for dimensionality reduction, yielded the highest Silhouette Score of 0.80 (Figure 9). This score indicates a clear separation between clusters with minimal overlap. Visual inspection of the clusters confirmed a high degree of cohesion within each group, suggesting the absence of misclassified instances. The strong performance of GMM can be attributed to its probabilistic framework, which allows it to model intra-cluster variance effectively, resulting in flexible and accurate group delineation.

In contrast, BIRCH and DBSCAN demonstrated moderate clustering quality. Both algorithms are based on the notions of density and point-wise distance rather than parametric distribution modeling. These characteristics make them well-suited for large-scale datasets and for identifying clusters of arbitrary shape, particularly in high-dimensional spaces. However, in this study, the dimensionality reduction via PCA likely diminished the benefits of density-based methods, allowing GMM to outperform them under the given experimental conditions.

The K-Means algorithm achieved a Silhouette Score of 0.64, which is lower than that of GMM but still acceptable for practical applications. Notably, repeated executions of the K-Means model with different centroid initializations resulted in score variability of approximately 0.06, highlighting a well-known limitation of this method—its sensitivity to initial conditions. This stochastic nature of cluster center initialization can lead to inconsistent segmentation results, especially when working with large and complex datasets.

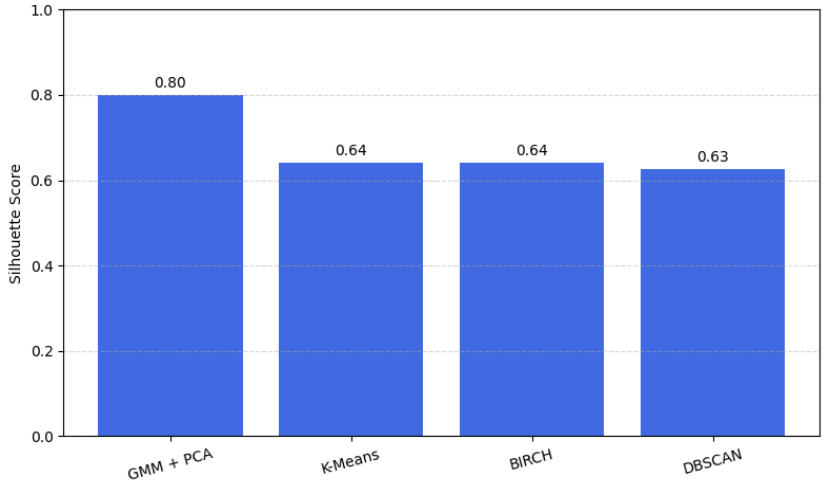

Fig. 2. **Comparison of clustering algorithms based on Silhouette Score**

Overall, the results underscore the effectiveness of the PCA + GMM combination for customer segmentation based on complex behavioral and transactional attributes. This approach enables the identification of subtle patterns in customer behavior without significant loss of information during dimensionality reduction. As such, it presents a promising methodology for applications in business analytics, personalized marketing, and recommendation systems.

Conclusions. The results of this study confirm the feasibility and effectiveness of applying machine learning methods to customer classification tasks. The developed system demonstrates high adaptability and can be successfully implemented in various industries, including e-commerce, banking, telecommunications, and other sectors where understanding and analyzing consumer behavior is of strategic importance.

The practical implementation of the proposed solution enables seamless integration into broader information and analytics platforms, supporting the automation of managerial decision-making processes. This contributes to the overall efficiency and responsiveness of business operations in data-intensive environments.

Furthermore, the modular and scalable architecture of the system allows for customization based on specific organizational requirements, ensuring its applicability across diverse real-world scenarios. The use of advanced clustering techniques, combined with dimensionality reduction, proved particularly effective for identifying meaningful customer segments and behavioral patterns.

References:

1. Kotler, P., Keller, K. L., Brady, M., Goodman, M., & Hansen, T. (2016). Marketing management (3rd ed.). Pearson Higher Ed.
2. Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations. Springer Science & Business Media.
3. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1007/978-3-540-87479-9_3
4. Kumar, S., Rani, R., Pippal, S. K., & Agrawal, R. (2025). Customer segmentation in e-commerce: K-means vs hierarchical clustering. TELKOMNIKA (Telecommunication Computing Electronics and Control), 23(1), 119–128. http://doi.org/10.12928/telkomnika.v23i1.26384
5. Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243. https://doi.org/10.3390/su14127243
6. Zhao, Y., & Zhou, X. (2021, April). K-means clustering algorithm and its improvement research. In Journal of Physics: Conference Series (Vol. 1873, No. 1, p. 012074). IOP Publishing. https://doi.org/10.1088/1742-6596/1873/1/012074

7. Huang, S., Kang, Z., Xu, Z., & Liu, Q. (2021). Robust deep k-means: An effective and simple method for data clustering. Pattern Recognition, 117, 107996. https://doi.org/10.1016/j.patcog.2021.107996
8. Jothi, R., & Muthukumaran, K. (2022). Telecom customer segmentation using deep embedded clustering algorithm. In B. Alyoubi, C. E. Ben Ncir, I. Alharbi, & A. Jarboui (Eds.), Machine learning and data analytics for solving business problems: Unsupervised and semi-supervised learning (pp. 85–104). Springer. https://doi.org/10.1007/978-3-031-18483-3_5
9. Cendana, M., & Kuo, R. J. (2024). Categorical data clustering: A bibliometric analysis and taxonomy. Machine Learning and Knowledge Extraction, 6(2), 1009–1054. https://doi.org/10.3390/make6020047
10. Lee, Z. J., Lee, C. Y., Chang, L. Y., & Sano, N. (2021). Clustering and classification based on distributed automatic feature engineering for customer segmentation. Symmetry, 13(9), 1557. https://doi.org/10.3390/sym13091557
11. Kumaresan, S. P., Tan, C. K., & Ng, Y. H. (2021). Deep neural network (DNN) for efficient user clustering and power allocation in downlink non-orthogonal multiple access (NOMA) 5G networks. Symmetry, 13(8), 1507. https://doi.org/10.3390/sym13081507
12. Xiahou, X., & Harada, Y. (2022). B2C e-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458–475. https://doi.org/10.3390/jtaer17020024
13. Liu, R., Ali, S., Bilal, S. F., Sakhawat, Z., Imran, A., Almuhaimeed, A., ... & Sun, G. (2022). An intelligent hybrid scheme for customer churn prediction integrating clustering and classification algorithms. Applied Sciences, 12(18), 9355. https://doi.org/10.3390/app12189355
14. Altameem, A. A., & Hafez, A. M. (2022). Behavior analysis using enhanced fuzzy clustering and deep learning. Electronics, 11(19), 3172. https://doi.org/10.3390/electronics11193172
15. Yan, X., Li, Y., Nie, F., & Li, R. (2025). Bank customer segmentation and marketing strategies based on improved DBSCAN algorithm. Applied Sciences, 15(6). https://doi.org/10.3390/app15063138
16. Alshdaifat, E. A., Alshdaifat, D. A., Alsarhan, A., Hussein, F., & El-Salhi, S. M. D. F. S. (2021). The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. Data, 6(2), 11. https://doi.org/10.3390/data6020011
17. Abdulrazzak, H. N., Hock, G. C., Mohamed Radzi, N. A., Tan, N. M., & Kwong, C. F. (2022). Modeling and analysis of new hybrid clustering technique for vehicular ad hoc network. Mathematics, 10(24), 4720. https://doi.org/10.3390/math10244720
18. Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. Symmetry, 15(9), 1679. https://doi.org/10.3390/sym15091679
19. Najeh, H., Lohr, C., & Leduc, B. (2022). Dynamic segmentation of sensor events for real-time human activity recognition in a smart home context. Sensors, 22(14), 5458. https://doi.org/10.3390/s22145458
20. Domingos, E., Ojeme, B., & Daramola, O. (2021). Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. Computation, 9(3), 34. https://doi.org/10.3390/computation9030034
21. Saha, L., Tripathy, H. K., Gaber, T., El-Gohary, H., & El-kenawy, E. S. M. (2023). Deep churn prediction method for telecommunication industry. Sustainability, 15(5), 4543. https://doi.org/10.3390/su15054543
22. Chen, Y. S., Lin, C. K., Chou, J. C. L., Chen, S. F., & Ting, M. H. (2022). Application of advanced hybrid models to identify the sustainable financial management clients of long-term care insurance policy. Sustainability, 14(19), 12485. https://doi.org/10.3390/su141912485

23. Jiang, W., Song, C., Wang, H., Yu, M., & Yan, Y. (2023). Obstacle detection by autonomous vehicles: An adaptive neighborhood search radius clustering approach. Machines, 11(1), 54. https://doi.org/10.3390/machines11010054
24. Banegas-Luna, A. J., Peña-García, J., Iftene, A., Guadagni, F., Ferroni, P., Scarpato, N., ... & Pérez-Sánchez, H. (2021). Towards the interpretability of machine learning predictions for medical applications targeting personalised therapies: A cancer case survey. International Journal of Molecular Sciences, 22(9), 4394. https://doi.org/10.3390/ijms22094394
25. Eslami, E., Razi, N., Lonbani, M., & Rezazadeh, J. (2024). Unveiling IoT customer behaviour: Segmentation and insights for enhanced IoT-CRM strategies: A real case study. Sensors, 24(4), 1050. https://doi.org/10.3390/s24041050.