УДК 004.8:004.4:005.94 **doi: 10.26906/SUNZ.2025.3.159-164**

О. В. Шматко, Д. О. Малишенко, О. Б. Волощук

Харківський національний університет радіоелектроніки, Харків, Україна

ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ ІНТЕЛЕКТУАЛЬНОЇ КЛАСИФІКАЦІЇ КЛІЄНТІВ: АРХІТЕКТУРА, РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

Анотація. Актуальність. У сучасних умовах цифрової трансформації бізнес-процесів зростає потреба у розробці інтелектуальних інформаційних систем для аналізу та обробки великих обсягів клієнтських даних. Одним із важливих напрямів є автоматизована класифікація клієнтів з використанням алгоритмів машинного навчання, що дозволяє підвищити ефективність маркетингових стратегій та прийняття управлінських рішень. **Об'єкт дослідження:** процеси класифікації клієнтів в інформаційних системах із використанням методів машинного навчання. **Мета статті:** проєктування, реалізація та дослідження архітектури програмних компонентів інформаційної системи для інтелектуальної класифікації клієнтів з урахуванням вимог до масштабованості, продуктивності та точності алгоритмів класифікації. **Результати дослідження.** У статті запропоновано архітектурну модель інформаційної системи, яка включає модуль збору, обробки та класифікації клієнтських даних. Реалізовано низку програмних компонентів, що інтегрують алгоритми машинного навчання (логістична регресія, дерева рішень, метод опорних векторів). Проведено експериментальні дослідження на основі реального набору даних, що продемонстрували високу точність класифікації та ефективність системи в умовах обмежених обчислювальних ресурсів. **Висновки.** Розроблена інформаційна система забезпечує точну класифікацію клієнтів і може бути застосована в комерційних аналітичних платформах. Результати дослідження можуть бути використані для подальшого вдосконалення інтелектуальних програмних систем аналізу даних.

Ключові слова: інформаційна система; класифікація клієнтів; машинне навчання; програмна архітектура; логістична регресія; дерева рішень; експериментальне дослідження.

Вступ

Постановка проблеми. В умовах сучасного конкурентного ринку ефективна взаємодія з клієнтами є одним із ключових факторів успішного функціонування компаній. Зростання обсягів даних, що генеруються в процесі комунікації з клієнтами, створює нові можливості для аналізу поведінки споживачів і формування персоналізованих бізнес-стратегій [1]. Одним із найважливіших інструментів у цьому контексті є сегментація клієнтів — процес поділу клієнтської бази на окремі групи за спільними характеристиками, такими як демографічні ознаки, купівельна поведінка, місцезнаходження, онлайн-активність або психографічні особливості [2].

Сегментація дозволяє компаніям краще розуміти потреби різних категорій споживачів, оптимізувати маркетингові кампанії, покращити клієнтський досвід і підвищити рівень лояльності до бренду. Сучасні підходи до сегментації дедалі частіше ґрунтуються на використанні методів машинного навчання, що забезпечують більш точний, гнучкий і масштабований аналіз великих обсягів даних. Алгоритми класифікації дозволяють автоматизовано виявляти закономірності в поведінці клієнтів, ідентифікувати ключові відмінності між групами та формувати обґрунтовані рекомендації для прийняття управлінських рішень.

Актуальність теми дослідження зумовлена необхідністю пошуку ефективних програмних рішень для побудови систем класифікації клієнтів із застосуванням сучасних алгоритмів інтелектуального аналізу даних. Такий підхід дає змогу компаніям не лише краще розуміти структуру своєї клієнтської бази, а й значно підвищити ефективність бізнес-процесів, що базуються на персоналізованій взаємодії.

У цьому контексті дана робота спрямована на розробку та реалізацію програмних компонентів для системи класифікації клієнтів компанії з використанням методів машинного навчання. Основна увага приділяється аналізу ключових факторів, які впливають на якість сегментації, вибору відповідних алгоритмів, а також впровадженню архітектурних рішень, здатних забезпечити масштабованість, точність і зручність застосування в реальних умовах бізнесу.

Аналіз останніх досліджень і публікацій. Класичні підходи до сегментації базуються на демографічних і географічних характеристиках [1], [2]. Однак сучасні методи передбачають використання поведінкових і транзакційних даних, а також алгоритмів машинного навчання. Зокрема, у роботах Kotler і Keller [1] та Wedel і Kamakura [2] закладено концептуальні основи сегментації, які стали фундаментом для подальших прикладних досліджень.

Найбільш поширеним методом кластеризації є K-means, який дозволяє ефективно групувати споживачів на основі подібності поведінкових ознак. Його використання описано в дослідженнях Jain [3], Kumar S. [4], Tabianan, K., та інші [5], де продемонстровано ефективність цього алгоритму в електронній комерції, банківській справі та сфері обслуговування. Однак K-means має низку обмежень, зокрема чутливість до вибору початкових центрів кластерів і необхідність заздалегідь визначати кількість груп. У зв'язку з цим у науковій літературі запропоновано кілька модифікацій. Так, в роботі [6] поєднано K-means із методом головних компонент (PCA), а Huang, S. та інші [7] застосували варіант K-medoids для підвищення стійкості кластеризації до викидів.

Ієрархічна кластеризація є ще одним популярним методом, що дозволяє побудову дерева кластерів із можливістю їх динамічного виділення. Дослі-

дження [8] та [9] ілюструють застосування цього методу у телекомунікаціях та роздрібній торгівлі. В роботі [10] представили модифікований варіант ієрархічної кластеризації на основі бінарного розщеплення.

Нечітка кластеризація (fuzzy clustering) дозволяє клієнтам належати до кількох сегментів одночасно з різним ступенем приналежності. Такий підхід продемонстровано в роботі [11]. Використання нейронних мереж для кластеризації розглянуто в дослідженні [12], де алгоритм успішно сегментує клієнтів телекомунікаційної компанії.

Метод опорних векторів (SVM) продемонстрував хороші результати у задачах класифікації клієнтів у сфері e-commerce, як показано у дослідженні [13]. Автори роботи [14] запропонували гібридний підхід, який об'єднує переваги K-means та ієрархічної кластеризації. В роботі [15] застосовано fuzzy clustering у поєднанні з глибокими нейронними мережами, що дозволило досягти високої точності.

Автори статті [16] використали autoencoder-мережі для зменшення розмірності даних перед застосуванням DBSCAN кластеризації. У дослідженні [17] розглянуто вплив попередньої обробки даних на точність сегментації. В роботі [18] виконано порівняння результатів традиційних і гібридних підходів до кластеризації. В роботі [19] розглянуто сегментацію в реальному часі з використанням потокових даних.

В роботі [20] застосовані методи глибинного навчання до аналізу поведінки клієнтів у банківській сфері. Водночас в роботі [21] аналізується застосування нейронних мереж у прогнозуванні відтоку клієнтів. В роботі [22] проводиться аналіз гібридних систем для сегментації в страхових компаніях. Автори роботи [23] пропонують адаптивну модель кластеризації, що самостійно визначає кількість сегментів. В роботі [24] досліджується інтерпретованість моделей класифікації, що важливо для прийняття рішень у бізнесі. В останніх роботах, зокрема [25], розглядаються перспективи інтеграції класифікації клієнтів із CRM-системами.

Таким чином, наукова література демонструє широкий спектр підходів до класифікації клієнтів із використанням машинного навчання. Вибір конкретного методу залежить від структури даних, розміру вибірки та цілей бізнесу. Перспективним є подальше дослідження гібридних моделей і адаптивних алгоритмів, що дозволяють гнучко реагувати на зміни в поведінці клієнтів.

Метою роботи є розробка методичного підходу і програмого засобу для класифікації клієнтів компанії з використанням алгоритмів машинного навчання, що дозволить покращити якість сегментації та забезпечити підтримку управлінських рішень.

Основний матеріал

Архітектура системи кластеризації клієнтів у бізнес-середовищі являє собою багаторівневу структурну модель, яка забезпечує послідовну обробку, трансформацію, аналіз та інтерпретацію клієнтських даних з метою отримання релевантної інформації для прийняття управлінських рішень. Побудова такої архітектури ґрунтується на принципах виявлення знань у базах даних (knowledge discovery in databases, KDD), де кожен етап слугує логічним продовженням попереднього та забезпечує підготовку даних для подальшої аналітичної обробки. Узагальнене представлення цієї архітектури наведено на рисунку 1, що ілюструє основні функціональні блоки системи та напрямки їх взаємодії.

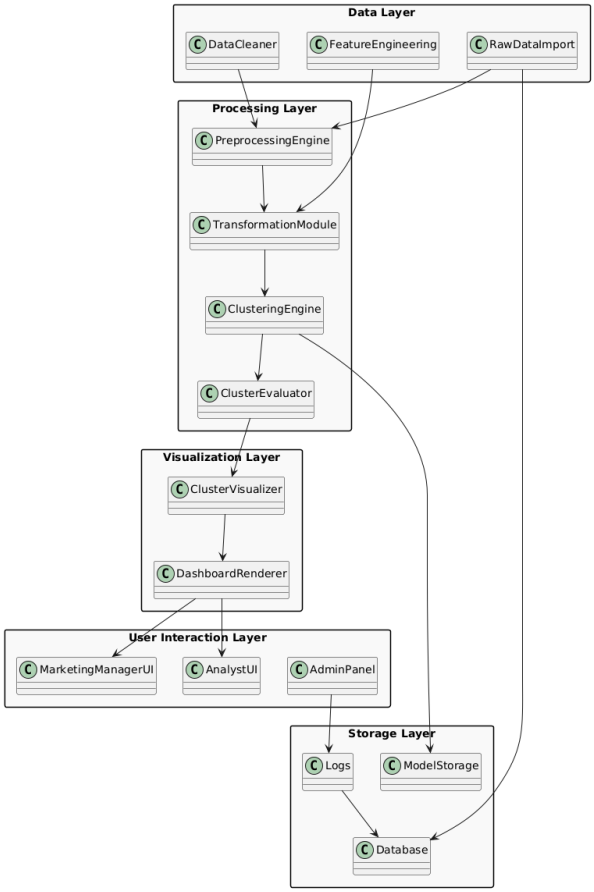

Рис. 1. Архітектура інтелектуальної системи кластеризації клієнтів

Початковим етапом є обробка вхідних даних, які надходять із різноманітних джерел, включаючи CRM-системи, транзакційні бази, маркетингові платформи, файли користувацької активності та інші інформаційні масиви. Вхідні дані можуть мати неоднорідну структуру, містити пропущені або дубльовані значення, а також потребувати перетворення форматів. У зв'язку з цим, першочерговою задачею архітектури є попередня обробка даних (preprocessing), що передбачає очищення, нормалізацію, агрегування та стандартизацію інформації. Результатом цього етапу є підготовлений набір даних, придатний для подальшої аналітики.

Після очищення дані проходять через фазу трансформації, де здійснюється приведення множинних джерел до уніфікованого формату ознак. На цьому етапі можуть виконуватись операції злиття датасетів, розрахунку нових змінних, а також відбір релевантних ознак на основі статистичних або евристичних критеріїв. Сформований у такий спосіб вектор

ознак подається на вхід сегментаційного модуля системи, де реалізується кластеризація клієнтів.

Сегментація являє собою ключовий етап архітектури, в межах якого здійснюється побудова кластерів за допомогою одного або кількох алгоритмів машинного навчання. У залежності від характеру даних та бізнес-задач можуть використовуватись методи K-means, DBSCAN, Fuzzy C-means, ієрархічна кластеризація або самоорганізовувані карти (SOM). Метою даного етапу є виявлення внутрішньої структури даних і формування гомогенних груп клієнтів, які характеризуються схожими поведінковими або демографічними параметрами. Як правило, у результаті первинної кластеризації утворюється велика кількість груп, значна частина яких може бути нерепрезентативною або дублюючою.

Для оптимізації результатів кластеризації наступним етапом є оцінювання сформованих кластерів з метою їх селекції та фільтрації. Оцінка якості кластерів здійснюється на основі внутрішніх метрик, таких як коефіцієнт силуету, індекс Девіса-Боулдіна, індекс Калінського-Харабаза тощо. Цей процес дозволяє відібрати найбільш значущі кластери, що мають внутрішню узгодженість і добре відділені один від одного, забезпечуючи максимальну цінність з точки зору подальшої інтерпретації.

Після того як обрано найбільш релевантні кластери, система переходить до фази візуалізації. На цьому етапі дані подаються у графічному вигляді, що полегшує аналіз результатів класифікації навіть для користувачів, які не мають глибокої технічної підготовки. Візуалізація може включати дво- або тривимірне зображення кластерів (з використанням методів зниження розмірності, таких як PCA або t-SNE), гістограми, діаграми розподілу, а також дашборди з ключовими характеристиками кожної групи. Завдяки цьому користувачі системи отримують зручні інструменти для інтерпретації та подальшого використання кластеризаційних результатів.

Фінальним етапом архітектури є інтерпретація результатів та здобуття знань (manual investigation). На цьому рівні бізнес-аналітики, маркетологи або інші фахівці аналізують класифіковані сегменти, формують висновки, генерують гіпотези та приймають управлінські рішення. Отримані знання можуть бути використані для розробки персоналізованих маркетингових кампаній, прогнозування поведінки клієнтів, покращення клієнтського досвіду або оптимізації роботи з окремими сегментами.

Описана архітектура системи кластеризації клієнтів реалізує принцип наскрізного циклу обробки даних – від початкового збору до практичного застосування знань. Її модульна структура дозволяє масштабування, гнучке налаштування під конкретні задачі та інтеграцію з іншими корпоративними інформаційними системами. Таким чином, вона є ефективним інструментом аналітики в умовах сучасного бізнес-середовища, що характеризується динамічністю, високою конкуренцією та необхідністю прийняття швидких і точних рішень на основі даних.

В роботі розглянуто ключові методи машинного навчання, які активно застосовуються для задач класифікації клієнтів. Увагу зосереджено на чотирьох популярних підходах: дерева рішень (Decision Tree, DT), метод опорних векторів (Support Vector Machines, SVM), метод k-ближчих сусідів (K-Nearest Neighbors, KNN) та випадковий ліс (Random Forest, RF). Дерева рішень (Decision Tree, DT). Метод дерев рішень є одним із найдавніших підходів у машинному навчанні, що активно використовується з 1970-х років. Ключовим елементом алгоритму є визначення нечистоти вузлів. Для цього використовуються показники Gini-індексу або ентропії:

$$Gini = 1 - \sum p_i^2 = 1 - \left(n_g / n\right)^2 - \left(n_b / n\right)^2,$$

де p_i — ймовірність належності до класу, n_g та n_b – кількість об’єктів у позитивному та негативному класах відповідно.

Ентропія:

$$Entropy = -\sum p_i \log p_i = -\left(p_g \log p_g + p_b \log p_b\right).$$

Метод опорних векторів (Support Vector Machines, SVM). Метод опорних векторів є потужним інструментом для класифікації та регресії, який набув широкого використання з 1990-х років. Основна мета SVM – знайти оптимальну гіперплощину, яка максимально розділяє класи у багатовимірному просторі. Рівняння гіперплощини:

$$\vec{w}^T \vec{u} + b = 0,$$

де $\vec{w}$ – вектор ваг, b – зміщення, u – вектор ознак.

Метод k-ближчих сусідів (K-Nearest Neighbors, KNN). Алгоритм KNN базується на концепції просторової близькості даних. Для нового об’єкта визначається відстань до інших точок у навчальній вибірці, після чого він класифікується за більшістю голосів його k найближчих сусідів.

Формула обчислення відстані у n-вимірному просторі:

$$D\left(x_1, x_2\right) = \sqrt{\sum_{i=1}^{n} \left(x_{1i} - x_{2i}\right)^2}.$$

Випадковий ліс (Random Forest, RF). Метод випадкового лісу є ансамблевим методом, що поєднує кілька дерев рішень. Алгоритм формує випадкову вибірку ознак, створює для кожної вибірки окреме дерево, а підсумкове рішення приймається за принципом більшості.

Алгоритм також використовує матрицю близькості між об’єктами, яка конвертується у матрицю відстаней для подальшої обробки (наприклад, за допомогою багатовимірного шкалювання).

З метою реалізації системи кластеризації клієнтів у рамках цієї дипломної роботи було розгорнуто експериментальний стенд, що ґрунтується на мові програмування Python, яка є однією з провідних технологій у сфері машинного навчання, аналізу даних та побудови аналітичних рішень. Завдяки своїй гнучкості, зручному синтаксису та наявності великої кількості спеціалізованих бібліотек Python дозволяє ефективно виконувати всі етапи циклу обробки даних — від завантаження та очищення до побудови моделей класифікації та візуалізації результатів.

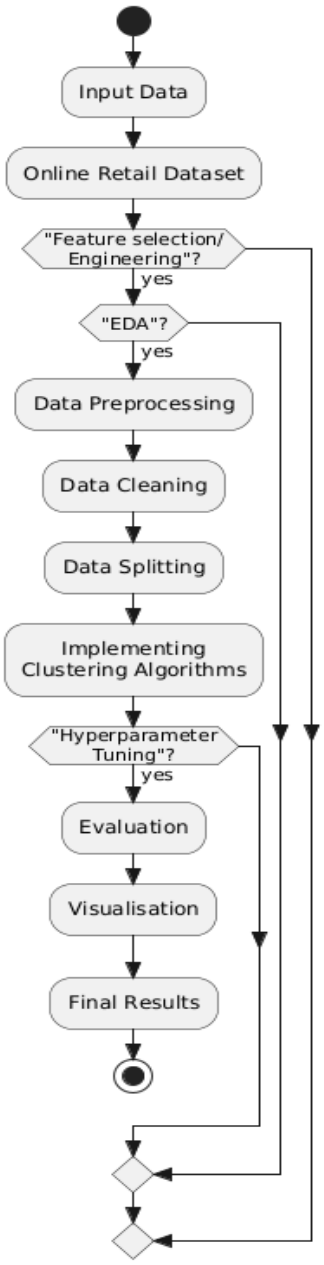

Рис. 2. Потік даних
в інтелектуальній системі кластеризації клієнтів

Потік даних для експериментального середовища представлена на рис. 2. Він охоплює послідовність етапів, необхідних для обробки вхідних даних, навчання моделей кластеризації, їхньої валідації та подальшої інтерпретації. Початковим етапом є завантаження вхідного набору даних — у цьому випадку використано набір онлайн-рітейл-даних, який містить інформацію про клієнтські транзакції. Вхідні дані подаються у систему за допомогою модуля pandas, що використовується для статистичного аналізу та обробки структурованих таблиць.

Після завантаження даних виконується первинне дослідження даних (EDA — Exploratory Data Analysis), у процесі якого аналізуються розподіли, виявляються пропущені значення, аномалії та кореляційні зв'язки. Для цього використовуються бібліотеки pandas, numpy, matplotlib та seaborn, які забезпечують статистичну обробку та графічне представлення розподілів. На основі отриманих результатів виконується відбір ознак і побудова нових змінних (feature engineering), що сприяє покращенню якості класифікації клієнтів. Далі дані проходять етап попередньої обробки, який включає масштабування числових ознак (MinMaxScaler, StandardScaler), кодування категоріальних змінних (LabelEncoder), а також очищення та нормалізацію значень. Після очищення відбувається розділення даних на навчальну та тестову вибірки, що дозволяє здійснити незалежну перевірку якості кластеризації. Модулі train_test_split із sklearn дозволяють провести це розділення із заданою часткою валідаційної вибірки.

У подальшому здійснюється реалізація алгоритмів кластеризації, серед яких застосовано: KMeans, DBSCAN, GaussianMixture, а також PCA — для попереднього зменшення розмірності. Для кожного з алгоритмів виконується налаштування гіперпараметрів із використанням RandomizedSearchCV, що дозволяє підібрати оптимальні конфігурації моделей. У процесі моделювання також застосовуються спеціалізовані функції для оцінювання якості кластеризації — silhouette_score, calinski_harabasz_score, davies_bouldin_score, які дають змогу обрати найкращу модель на основі формальних метрик.

Оцінювання точності та стабільності моделей здійснюється на основі порівняльного аналізу результатів кластеризації. Результати класифікації відображаються у вигляді дво- та тривимірних графіків, побудованих за допомогою matplotlib та seaborn, що дозволяє візуалізувати просторове розташування кластерів і зрозуміти логіку розподілу клієнтів за сегментами. На завершальному етапі формуються підсумкові результати, які можуть бути інтегровані у бізнес-аналітику для створення персоналізованих пропозицій або адаптації маркетингових стратегій. Отримані знання можуть бути збережені у базу даних або передані до зовнішніх інформаційних систем. Створений експериментальний стенд відображає повний цикл класифікації клієнтів — від імпорту даних і підготовки вибірки до кластеризації, оцінювання результатів і побудови висновків, що є базою для прийняття рішень у бізнес-середовищі. У рамках даного дослідження було використано відкритий набір даних зі сфери роздрібної торгівлі, який отримано з UCI Machine Learning Repository — одного з найавторитетніших джерел у сфері машинного навчання. Конкретно було використано Online Retail Dataset, який знаходиться у вільному доступі за посиланням: https://archive.ics.uci.edu/dataset/352/online+retail.

Вибір цього набору даних зумовлений його обсягом, структурованістю, а також широким застосуванням у наукових публікаціях, присвячених задачам кластеризації клієнтів та поведінкової аналітики у сфері електронної комерції. Набір містить 541 909 записів, що охоплюють інформацію про транзакції, здійснені в період з грудня 2010 по грудень 2011 року однією з онлайн-компаній, що базується у Великобританії та займається продажем подарункових товарів. Кожен запис у наборі відповідає окремій товарній позиції у певному замовленні. Опис набору даних представлений в табл. 1.

Таблиця 1 – **Характеристики датасету Online Retail,**

Назва змінної	Формат / Тип значень	Опис
InvoiceNo	Ціле число (6 цифр), текст	Унікальний номер рахунку-фактури. Початок з "C" означає скасування.
StockCode	Ціле число (5 цифр), текст	Ідентифікатор товарної позиції.
Description	Текстовий рядок	Назва або опис товару.
Quantity	Ціле число	Кількість придбаних одиниць товару.
InvoiceDate	Дата і час	Час створення рахунку-фактури.
UnitPrice	Дійсне число (float)	Ціна за одиницю товару (у фунтах стерлінгів).
CustomerID	Ціле число (5 цифр), текст	Унікальний ідентифікатор клієнта.
Country	Текстовий рядок	Назва країни проживання клієнта.

Оцінювання якості кластеризації є критично важливим етапом у процесі сегментації клієнтів, оскільки дозволяє кількісно визначити, наскільки добре сформовані кластери відображають реальні закономірності у даних. У межах цього дослідження для валідації кластеризаційних моделей було використано три основні метрики: Silhouette Score, індекс Калінські-Харабаза (Calinski-Harabasz Index) та індекс Девіса-Боулдіна (Davies-Bouldin Index). Кожна з них забезпечує специфічну оцінку когезійності (внутрішньої узгодженості) кластерів та ступеня їх відокремлення один від одного в багатовимірному просторі ознак.

На першому етапі для кожного клієнта було обчислено значення RFM-показників. Зокрема, Recency визначалась як кількість днів з моменту останньої покупки до дати завершення спостереження, Frequency – як кількість унікальних транзакцій, здійснених клієнтом, а Monetary – як загальна сума витрат за всі замовлення.

На основі отриманих RFM показників було реалізовано поділ клієнтів на п’ять основних категорій (рис. 3): Top customers – клієнти з RFM-скором вище 4.5; High-value customers – більше 4.0; Medium-value customers – більше 3.0; Low-value customers – більше 1.6; Lost customers – нижче або дорівнює 1.6.

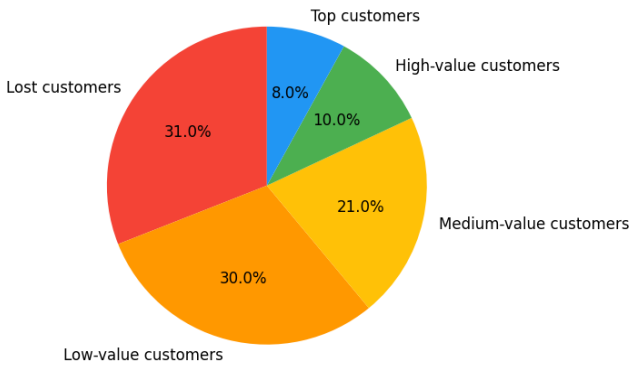

Рис. 3. Розподіл клієнтів за сегментами RFM аналізу

Результати RFM-аналізу показали, що лише близько п’ятої частини клієнтів здійснюють регулярні та високовартісні покупки. Це свідчить про потенціал до оптимізації взаємодії з менш активними клієнтами шляхом впровадження персоналізованих кампаній, систем лояльності та рекомендаційних сервісів. У якості одного з основних алгоритмів кластеризації було застосовано K-Means, який широко використовується завдяки своїй простоті реалізації та ефективності для великих обсягів даних. Для визначення оптимальної кількості кластерів (k) було використано метод "лікоть" (elbow method).

На основі візуального аналізу графіка (рис. 4) було встановлено, що оптимальна кількість кластерів дорівнює 3, оскільки саме при цьому значенні спостерігається найбільш виражений злам інерції. Після визначення параметра k було проведено масштабування ознак за допомогою Min–Max Normalization, щоб привести всі числові значення до одного діапазону. Це дозволяє уникнути зміщення в результатах кластеризації, яке може виникнути через дисбаланс у масштабах окремих ознак, особливо з огляду на те, що K-Means використовує евклідову відстань як метрику подібності.

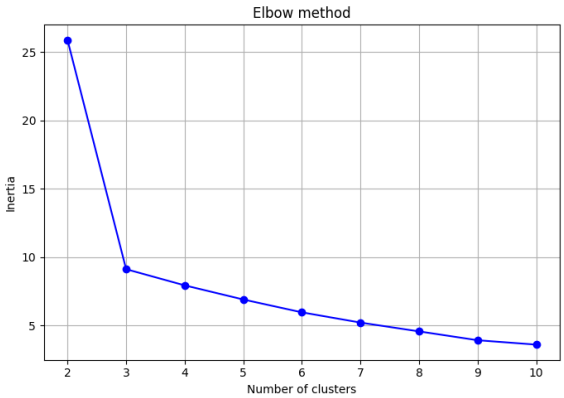

Рис. 4. Визначення кількості кластерів за «elbow method»

Після масштабування було побудовано кластеризаційну модель методом K-Means з трьома кластерами. Результати моделі візуалізовано у вигляді діаграми розсіювання (scatter plot), де кожна точка представляє клієнта, а колір позначає кластер, до якого його віднесено (рис. 5).

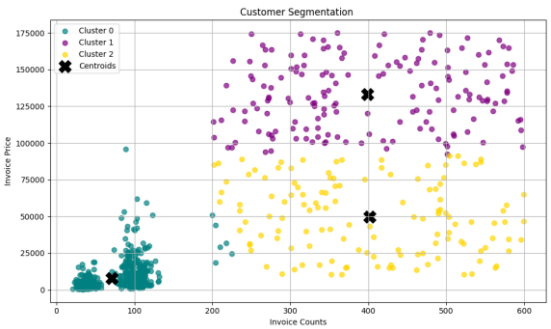

Рис. 5. Діаграма розсіювання (scatter plot) для k=3

Для виконання кластеризації за допомогою моделі Gaussian Mixture Model (GMM) попередньо бу-

ло застосовано метод головних компонент (Principal Component Analysis, PCA) з метою зниження розмірності вхідних даних. На рис. 6 представлено візуалізацію результатів кластеризації, отриманих за допомогою Gaussian Mixture Model.

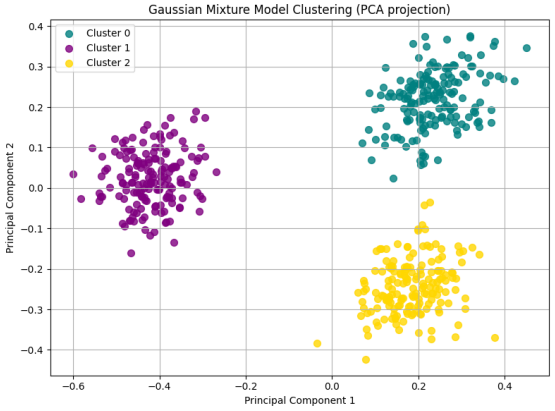

Рис. 6. Візуалізація результатів кластеризації методом Gaussian Mixture Model у просторі P1–P2

Перед застосуванням DBSCAN дані було нормалізовано за допомогою стандартного масштабування, що дозволило вирівняти внесок усіх ознак у обчислення відстаней. Основними параметрами алгоритму є:

- eps — радіус околу точки, який визначає максимальну відстань між точками в межах одного кластера;
- min_samples — мінімальна кількість точок у межах околу, необхідна для того, щоб точку вважати основною (core point).

Для підбору оптимального значення параметра eps було використано K-distance графік, який демонструє відстані до найближчих сусідів для кожної точки. Попри труднощі у точному визначенні зламу (knee point), експериментальним шляхом було встановлено, що значення eps = 0.3 забезпечує найкращу кластеризацію з точки зору розділення даних та стабільності результатів.

Після виконання DBSCAN-кластеризації кожній точці було присвоєно відповідну мітку кластера, а точки, які не мали достатньої кількості сусідів у межах eps, були класифіковані як шумові та отримали мітку -1. Для обчислення загальної кількості кластерів у вибірці враховувалися лише валідні (ненульові) кластери, тобто кількість унікальних міток без урахування шумових точок.

На рис. 7 представлено результати кластеризації методом DBSCAN у двовимірному просторі.

На рис. 8 наведено графічну візуалізацію результатів кластеризації за допомогою BIRCH.

Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) є ефективним засобом ієрархічної кластеризації, спеціально розробленим для обробки великих масивів даних за обмежених ресурсів пам’яті. Однією з ключових переваг BIRCH є його здатність до інкрементального побудування кластерної структури, що забезпечує масштабованість та високу швидкодію під час обробки реальних бізнес-сценаріїв.

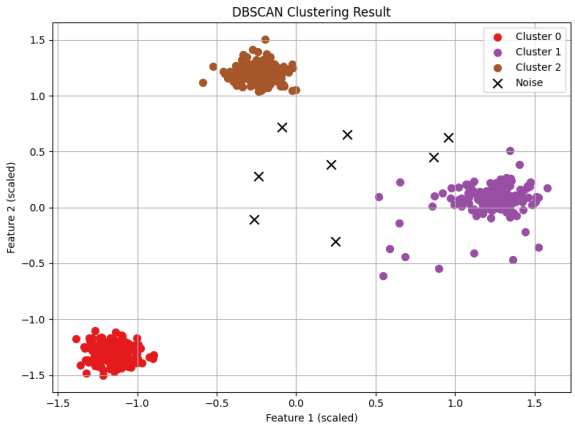

Рис. 7. Візуалізація результатів кластеризації методом DBSCAN

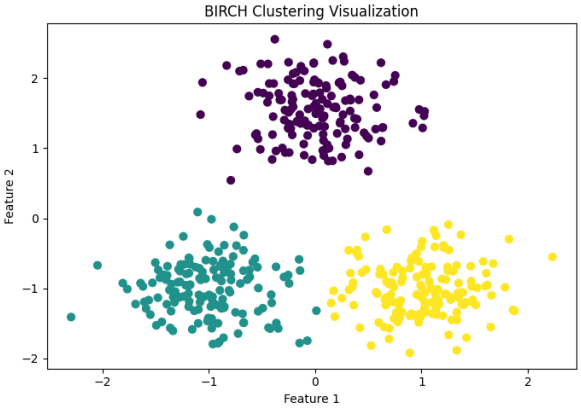

Рис. 8. Візуалізація кластеризації методом BIRCH

Ключовим параметром алгоритму є поріг (threshold), який визначає максимальний допустимий діаметр підкластерів, що формуються на кожному етапі побудови дерева кластерів (CF Tree)

У межах експерименту було протестовано низку порогових значень у діапазоні від 0.01 до 1.0. На основі аналізу метрики Silhouette Score найкращі результати були досягнуті при значенні threshold = 0.01, що дозволило сформувати чітко відокремлені та структурно збалансовані групи. Кількість виявлених кластерів за такого налаштування склала три, що відповідає очікуваній кластерній структурі на основі вхідних ознак.

У межах даного дослідження було проведено порівняльну оцінку ефективності кластеризаційних алгоритмів на основі показника Silhouette Score. До аналізу було включено такі методи: Gaussian Mixture Model (GMM), K-Means, BIRCH, DBSCAN. Обрані алгоритми забезпечили побудову кластерів із мінімальними обчислювальними витратами, що робить їх придатними для обробки масивів даних із невеликою кількістю вимірів. За результатами моделювання було встановлено, що модель GMM у поєднанні з попереднім зниженням розмірності методом PCA забезпечила найвищий показник Silhouette Score, який склав 0.80 (рис. 9). Це свідчить про чітке відокремлення кластерів та мінімальне перекриття між ними. Візуальний аналіз підтвердив високу згуртованість елементів у межах кожної групи, що дозволяє припустити відсутність хибно класифікованих

об’єктів. Такий результат пояснюється здатністю GMM моделювати варіацію даних у кожному кластері, що забезпечує гнучкість та точність розподілу.

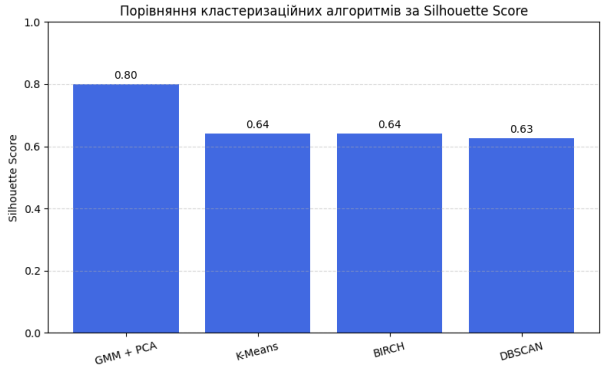

Рис. 9. Порівняння алгоритмів кластеризації

Методи BIRCH та DBSCAN показали помірні результати, зважаючи на те, що обидва орієнтовані на щільність та відстані між точками, а не на параметричне моделювання розподілів. Їхня перевага проявляється в обробці великих масивів даних і роботи з нестандартними формами кластерів, особливо у випадках високої розмірності. Однак у даному дослідженні, завдяки зниженню розмірності за допомогою PCA, GMM зміг продемонструвати кращі результати.

Метод K-Means показав результат зі значенням Silhouette Score на рівні 0.64. При цьому в ході декількох запусків моделі з різними ініціалізаціями центроїдів було зафіксовано варіативність результатів у межах 0.06, що є типовою особливістю цього методу. Вона обумовлена стохастичним характером вибору початкових кластерів, що може впливати на кінцеву якість розбиття, особливо в задачах, пов’язаних із великими наборами даних.

Загалом результати свідчать про ефективність поєднання PCA + GMM для кластеризації клієнтів за комплексними ознаками. Цей підхід дозволяє виявляти глибокі закономірності в поведінці об’єктів без втрати критичної інформації, що робить його перспективним для бізнес-аналітики, персоналізованого маркетингу та систем рекомендацій.

Висновки

Результати роботи підтвердили доцільність та ефективність використання методів машинного навчання для задач класифікації клієнтів. Розроблена система може бути адаптована для використання у сфері електронної комерції, банківського обслуговування, телекомунікацій, а також в інших галузях, де важливим є аналіз споживчої поведінки. Практична реалізація створеного рішення дозволяє інтегрувати його у більш широкі інформаційно-аналітичні платформи з метою автоматизації процесів прийняття управлінських рішень, що відкриває перспективи для подальших досліджень у напрямку інтеграції кластеризації з предиктивними моделями та рекомендаційними системами.

Список літератури

1. Kotler, P., Keller, K. L., Brady, M., Goodman, M., & Hansen, T. (2016). Marketing Management 3rd edn PDF eBook. 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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: https://doi.org/10.1016/j.patcog.2021.107996
8. Jothi, R., Muthukumaran, K. (2022). Telecom Customer Segmentation Using Deep Embedded Clustering Algorithm. In: Alyoubi, B., Ben Ncir, CE., Alharbi, I., Jarboui, A. (eds) Machine Learning and Data Analytics for Solving Business Problems. Unsupervised and Semi-Supervised Learning. Springer, Cham. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: 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 (2076-3417), 15(6).DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: https://doi.org/10.3390/machines11010054
24. Banegas-Luna, A. J., Pe?a-Garc?a, J., Iftene, A., Guadagni, F., Ferroni, P., Scarpato, N., ... & P?rez-S?nchez, 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. DOI: 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. DOI: https://doi.org/10.3390/s24041050

Received (Надійшла) 12.06.2025
Accepted for publication (Прийнята до друку) 20.08.2025

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Шматко Олександр Віталійович – кандидат технічних наук, доцент, доцент кафедри ЕОМ, Харківського національного університету радіоелектроніки, Харків, Україна;
Oleksandr Shmatko – PhD, Ass.prof, Ass.prof of the Department of electronic computers, Kharkiv National University of radio electronics, Kharkiv, Ukraine;
e-mail: oleksandr.shmatko2@nure.ua; ORCID Author ID: https://orcid.org/0000-0002-2426-900X
Scopus Author ID: http://www.scopus.com/inward/authorDetails.url?authorID=6602623478&partnerID=MN8TOARS

Малишенко Дар'я Олександрівна – магістр кафедри ЕОМ, Харківського національного університету радіоелектроніки, Харків, Україна;
Daria Malyshenko – Master degree student, of the Department of electronic computers, Kharkiv National University of radio electronics, Kharkiv, Ukraine;
e-mail: daria.malyshenko@nure.ua; ORCID Author ID: https://orcid.org/0009-0001-4560-2019.

Волощук Олена Борисівна – кандидат технічних наук, доцент, доцент кафедри ЕОМ Харківського національного університету радіоелектроніки, Харків, Україна;
Olena Voloshchuk – PhD, Ass.prof, Ass.prof of the Department of electronic computers, Kharkiv National University of radio electronics, Kharkiv, Ukraine;
e-mail: olena.voloshchuk@nure.ua; ORCID Author ID: https://orcid.org/0000-0002-5912-4126;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57207762084.

Information system for intelligent customer classification: architecture, implementation, and experimental research

Oleksandr Shmatko, Daria Malyshenko, Olena Voloshchuk

Abstract. Relevance. In the context of ongoing digital transformation of business processes, there is a growing demand for intelligent information systems capable of analyzing and processing large volumes of customer data. One of the key directions in this field is automated customer classification using machine learning algorithms, which significantly enhances the effectiveness of marketing strategies and decision-making. **Object of the research:** customer classification processes in information systems using machine learning methods. **Purpose of the article:** to design, implement, and investigate the architecture of software components of an information system for intelligent customer classification, taking into account requirements for scalability, performance, and classification accuracy. **Research results.** The article proposes an architectural model of an information system that includes modules for data collection, preprocessing, and customer classification. Several software components were implemented, integrating machine learning algorithms such as logistic regression, decision trees, and support vector machines. Experimental studies based on a real-world dataset demonstrated high classification accuracy and the system's efficiency under limited computational resources. **Conclusions.** The developed information system ensures accurate customer classification and can be applied in commercial data analytics platforms. The research findings can be used to further improve intelligent software systems for data analysis.

Keywords: information system; customer classification; machine learning; software architecture; logistic regression; decision trees; experimental research.