

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА
Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»
Кафедра математичного моделювання та статистики
Освітньо-професійна програма «Економічна кібернетика і Дата Сайнс»
Галузь знань 05 «Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

на тему **«Підвищення ефективності програм лояльності на основі сегментації клієнтів»**
(назва теми)

здобувача Настаулової Ольги Володимирівни
(ПІБ, підпис)

Науковий керівник: кандидат економічних наук, доцент Ольга ОСИПОВА

(підпис)

**Робота допущена до захисту перед екзаменаційною комісією з атестації
здобувачів вищої освіти (ЕК)**

Завідувач кафедри: к. ф.-м. н., професор
Галина ВЕЛИКОІВАНЕНКО

(підпис)

Київ 2024

ЗМІСТ

ВСТУП	3
РОЗДІЛ 1. ОГЛЯД КОНЦЕПЦІЇ ПРОГРАМ ЛОЯЛЬНОСТІ ТА МЕТОДІВ СЕГМЕНТАЦІЇ КЛІЄНТІВ	7
1.1. Поняття та види програм лояльності	7
1.2. Методи сегментації клієнтів	11
1.3. Вплив сегментації клієнтів на ефективність програм лояльності	17
1.4. Програми лояльності в Україні та їх пропозиції	21
РОЗДІЛ 2. МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ СЕГМЕНТАЦІЇ КЛІЄНТІВ	24
2.1. Загальна характеристика методів машинного навчання для сегментації клієнтів ..	24
2.2. Використання методу k-середніх для кластеризації клієнтів	27
2.3. Використання методу DBSCAN для кластеризації клієнтів	30
2.4. Використання методу GMM для кластеризації клієнтів	32
2.5. Порівняння методів кластеризації та критерії оцінки якості кластерів	34
РОЗДІЛ 3. ПРАКТИЧНА РЕАЛІЗАЦІЯ МЕТОДІВ КЛАСТЕРІЗАЦІЇ КЛІЄНТІВ ТА АНАЛІЗ РЕЗУЛЬТАТІВ	40
3.1. Опис набору даних та препроцесинг даних	40
3.2. Реалізація та порівняння трьох методів кластеризації	44
3.3. Аналіз та інтерпретація отриманих сегментів	57
3.4. Рекомендації щодо персоналізованих пропозицій для отриманих сегментів	59
ВИСНОВКИ	65
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	67

ВСТУП

В умовах сучасного конкурентного середовища успішність підприємств значною мірою залежить від здатності формувати довготривалі взаємини з клієнтами. Програми лояльності є одним із ключових інструментів утримання клієнтів і підвищення їхньої задоволеності. Однак, використання уніфікованого підходу до всіх споживачів часто призводить до недостатньої ефективності таких програм. Сегментація клієнтів, як основа персоналізації програм лояльності, є перспективним напрямом для підвищення їх результативності. Актуальність дослідження зумовлена необхідністю розробки інноваційних підходів до сегментації, особливо в умовах цифровізації ринків, що має важливе значення як для світового, так і для українського бізнесу.

Сегментація клієнтів та розробка програм лояльності є важливими напрямками досліджень у сучасному маркетингу. У працях Ф. Котлера основна увага приділяється загальним підходам до сегментації клієнтів, зокрема поведінковим, демографічним, психографічним і географічним характеристикам. Автор наголошує, що успішність сегментації залежить від точності визначення потреб споживачів та їх подальшої інтеграції в маркетингові стратегії. Його робота стала основою для формування уявлень про те, як сегментація впливає на ефективність взаємодії зі споживачами [1].

У статті "Customer Behaviour and B2C Client Segmentation in Data-Driven Society" Жикова П. аналізує вплив цифрової революції на поведінку клієнтів і сегментацію у B2C-середовищі, орієнтованому на дані. Авторка досліджує, як використання великих даних трансформує традиційні моделі управління клієнтами в умовах сучасного фінансового ринку. Основною ідеєю є створення цінності для бізнесу шляхом залучення й утримання прибуткових клієнтів, з

акцентом на оцінку їхньої життєвої цінності (Customer Lifetime Value, CLV) через історичний прибуток і кредитний рейтинг [2].

Робота Марчук І., Орловської С. та Жарської І "Управління лояльністю на ринку B2B" присвячена аналізу концепції лояльності та її значення в контексті бізнес-відносин між компаніями (B2B). У роботі наголошується, що утримання та формування лояльності клієнтів є критично важливими завданнями для компаній, які прагнуть досягти довгострокового успіху. Стаття містить огляд історії розвитку концепції лояльності у B2B-секторі, зокрема важливості програм лояльності як інструменту зміцнення партнерських відносин між компаніями. Наведені приклади й статистичні дані підтверджують ефективність таких програм у підвищенні рівня задоволеності клієнтів, довіри та стабільності ділових відносин. Зокрема, автори звертають увагу на те, що програми лояльності сприяють не лише збільшенню повторних продажів, але й зменшенню ризику втрати клієнтів через високу вартість переходу до конкурентів [3].

Метою роботи є розробка ефективних підходів до сегментації клієнтів на основі сучасних методів кластеризації для персоналізації програм лояльності та підвищення їхньої результативності.

Для досягнення цієї мети поставлені такі завдання:

- Проаналізувати теоретичні основи програм лояльності та методів сегментації клієнтів.
- Вивчити принципи роботи методів кластеризації k-середніх, DBSCAN і GMM.
- Провести сегментацію клієнтів на основі реальних даних, використовуючи зазначені методи кластеризації.
- Оцінити ефективність запропонованих підходів до сегментації для вдосконалення програм лояльності.
- Розробити рекомендації щодо персоналізації програм лояльності на основі отриманих результатів.

Об'єктом дослідження є сегментація клієнтів в умовах взаємодії підприємств з клієнтами.

Предметом дослідження виступають методи кластеризації даних (к-середніх, DBSCAN, GMM) для сегментації клієнтів і їхній вплив на ефективність програм лояльності.

У роботі використані загальнонаукові методи:

— Теоретичний аналіз і синтез – для систематизації знань про програми лояльності, сучасні підходи до сегментації клієнтів і методи кластеризації.

— Методи кластеризації даних:

К-середніх – для групування клієнтів за заданими критеріями;

DBSCAN – для виявлення щільних кластерів та аномалій у даних;

GMM (Gaussian Mixture Models) – для моделювання клієнтських сегментів за ймовірнісним підходом.

— Методи оцінювання якості кластеризації – для аналізу результатів сегментації за допомогою показників, що визначають чіткість і однорідність кластерів.

Теоретична значущість роботи полягає у вдосконаленні підходів до сегментації клієнтів, що сприяє розвитку теорії управління лояльністю. У дослідженні розширюються наукові уявлення про багатовимірний аналіз клієнтської бази, зокрема поведінкові, демографічні та інші характеристики споживачів, які можуть впливати на ефективність маркетингових стратегій.

Методична значущість роботи полягає у розробці інструментарію та методологічних підходів, які можуть бути використані для підготовки спеціалістів у галузі маркетингової аналітики, управління бізнесом та обробки даних, а також для вдосконалення освітніх програм, спрямованих на розвиток практичних навичок у цих сферах.

Практичне значення отриманих результатів полягає у розробці рекомендацій для бізнесу щодо вдосконалення програм лояльності через персоналізацію пропозицій для різних сегментів клієнтів. На основі застосування

сучасних методів кластеризації підприємства отримують інструменти для точного аналізу клієнтської бази, що сприяє оптимізації їхньої маркетингової діяльності. Алгоритм сегментації, розроблений у роботі, може бути використаний компаніями для автоматизації процесів аналізу клієнтів, що забезпечує економію часу та ресурсів. Крім того, результати дослідження дозволяють оцінити ефективність програм лояльності, спрямованих на різні сегменти споживачів, і розробити цільові стратегії, що підвищують рентабельність інвестицій у ці програми. Особливу цінність має застосування таких підходів для українського ринку, де бізнесу необхідно адаптуватися до змінних умов і високої конкуренції.

Інформаційною базою дослідження виступали статистичні дані про клієнтів, які брали участь у маркетингових активностях минулого року, аналітичні звіти, наукові праці вітчизняних та зарубіжних авторів, та технічна документація щодо використання мови Python.

Робота складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків.

РОЗДІЛ 1

ОГЛЯД КОНЦЕПЦІЇ ПРОГРАМ ЛОЯЛЬНОСТІ ТА МЕТОДІВ СЕГМЕНТАЦІЇ КЛІЄНТІВ

1.1. Поняття та види програм лояльності

Програми лояльності - це програми стимулювання клієнтів, які пропонуються роздрібними торговцями та іншими компаніями для заохочення повторних покупок. Ці програми, як правило, винагороджують клієнтів балами, знижками або іншими стимулами за їхні покупки. Чим більше клієнт витрачає, тим більшими будуть його винагороди. Основою цих програм є прагнення створити для клієнта додаткову цінність, яка виходить за межі звичайної покупки товару або послуги, стимулюючи таким чином його прихильність [4].

Програми лояльності пропонують різноманітні переваги, щоб винагородити клієнтів. Ці стимули можуть включати:

- Ранній доступ до нових продуктів та продажів
- Безкоштовні товари або покращені послуги
- Ексклюзивні знижки та спеціальні пропозиції
- Пріоритетна доставка та інші спеціальні послуги

Щоб взяти участь у програмі лояльності, клієнти, як правило, надають свої особисті дані компанії та отримують унікальний ідентифікатор, наприклад, членську картку або номер рахунку. Цей ідентифікатор використовується для відстеження їхніх покупок та відповідного їх винагородження [5].

Ідея лояльності клієнтів не нова і бере свої витоки з концепції утримання клієнтів як дешевшої альтернативи залученню нових. Історично перші програми лояльності виникли в середині XX століття у вигляді систем знижок і бонусів. З часом ці програми еволюціонували, перетворившись із простих механізмів

стимулювання продажів на складні системи управління клієнтським досвідом. Сьогодні програми лояльності стали інтегрованою частиною CRM-систем (Customer Relationship Management) і активно використовують технології великих даних, машинного навчання та персоналізації.

У маркетингу поняття «лояльність» охоплює як емоційну прихильність клієнта до бренду, так і його раціональні рішення залишатися клієнтом компанії. Програми лояльності спрямовані на формування обох аспектів, поєднуючи раціональні стимули (знижки, бонуси, спеціальні пропозиції) з емоційними (створення відчуття унікальності, ексклюзивності та турботи) [6].

З року в рік програми лояльності стають все більш популярними. Згідно з опитуванням Lending Tree, близько 80% американських споживачів є членами принаймні однієї програми лояльності [7].

Програма лояльності також може мати значний вплив на задоволеність клієнтів та загальне збільшення доходів. Наприклад, дослідження McKinsey показало, що клієнти, які є членами програми лояльності, на 59% частіше обирають бренд, а не конкурента, і на 43% частіше купують щотижня. Насправді, дослідження показало, що клієнти, які є членами програми лояльності, на 62% частіше витрачають більше на бренд [8].

Програми лояльності поділяються на кілька видів залежно від механізму їх роботи та цілей, які вони переслідують. Найбільш поширеними є такі:

а. Бонусні програми: Популярний інструмент лояльності, який дозволяє компаніям стимулювати повторні покупки та утримувати клієнтів. За кожну покупку клієнти отримують бонуси або бали, які можуть бути використані для отримання знижок, безкоштовних товарів чи послуг. Такий підхід не лише мотивує клієнтів робити покупки частіше, але й створює у них відчуття винагороди, що сприяє емоційному зв'язку з брендом. Бонуси можуть нараховуватись за різними принципами, наприклад, за витрачену суму, кількість товарів чи виконання спеціальних умов. Зібрані бали клієнти можуть

обміняти на знижки або отримати безкоштовні продукти, що стимулює їх повертатися для нових покупок.

Крім того, бонусні програми можуть бути персоналізованими, надаючи клієнтам індивідуальні пропозиції, знижки до свят чи інші вигоди. Це дозволяє компаніям краще розуміти потреби своїх клієнтів і створювати для них більш привабливі умови для покупок.

Для бізнесів бонусні програми є вигідним інструментом, оскільки вони не тільки підвищують рівень продажів, але й сприяють зростанню клієнтської лояльності, що в свою чергу забезпечує стабільний потік доходів у майбутньому.

- b. Програми накопичення: популярний тип лояльнісних стратегій, який передбачає довгострокове накопичення клієнтом певних ресурсів або балів, які можна обміняти на цінні призи, товари або послуги. Ці програми часто використовуються в таких сферах, як авіаперевезення, ритейл, готельний бізнес, банківські послуги та інші. Основна ідея програми накопичення полягає в тому, що клієнти з часом збирають бали або милі, виконуючи покупки чи використовуючи послуги компанії. Замість того, щоб отримувати негайну винагороду, клієнт накопичує ресурси протягом певного періоду, що створює відчуття досягнення мети і стимулює повторні покупки. Наприклад, клієнти авіакомпаній можуть накопичувати милі за кожен переліт і обмінювати їх на безкоштовні квитки або підвищення класу обслуговування. У ритейлі накопичення балів може призвести до знижок або безкоштовних товарів.

Ці програми працюють на довгостроковій прив'язаності клієнта до бренду, оскільки вони мотивують клієнта залишатися в межах певної екосистеми і постійно повертатися до компанії, щоб накопичити більше балів. Це також допомагає компаніям підтримувати стабільний рівень лояльності серед клієнтів, оскільки бажання "відкупити" накопичені бали або милі часто є сильним стимулом для подальших покупок чи використання послуг.

З часом, програми накопичення можуть пропонувати більш ексклюзивні переваги, наприклад, доступ до спеціальних подій, привілеїв чи персоналізованих пропозицій для найбільш лояльних клієнтів. Це дозволяє створити враження, що клієнт є частиною закритого клубу, що значно підвищує його прив'язаність до бренду.

Завдяки цьому механізму компанії можуть не лише утримувати своїх клієнтів, але й залучати нових, оскільки пропозиція накопичення балів або миль виглядає привабливою для тих, хто шукає додаткові вигоди від своїх витрат. Клубні програми: Клуби лояльності пропонують своїм членам спеціальні переваги, такі як доступ до ексклюзивних товарів, послуг чи подій. Учасництво часто асоціюється з певним статусом, що формує емоційний зв'язок із брендом.

- с. Крос-програми (коаліційні програми): Ці програми розробляються спільно кількома компаніями, що працюють у різних сегментах ринку, але об'єднані спільною метою — залучити клієнтів і підвищити їх лояльність. У таких програмах клієнти можуть накопичувати бали, купуючи товари чи послуги у будь-якого з учасників програми, а потім використовувати їх у будь-якої іншої компанії-партнера. Цей тип програм має кілька переваг як для компаній, так і для клієнтів. Для компаній це можливість об'єднати ресурси і клієнтську базу, а також забезпечити більше цінних пропозицій для споживачів, що стимулює їх до більш активного використання послуг і товарів усіх учасників програми. Для клієнтів це означає більше можливостей для накопичення балів або бонусів, адже вони можуть отримувати їх не тільки від однієї компанії, а від кількох, що підвищує вигідність участі у програмі. Прикладом коаліційних програм є бонусні системи великих банків, коли клієнти можуть накопичувати бали за покупки в магазинах-партнерах або за використання банківських карток. Такі бали можна потім витратити на знижки, подарунки або інші привілеї у будь-якому з партнерів програми. Іншим прикладом є торгові мережі, які об'єднують різні бренди і компанії в одну програму лояльності, що дозволяє клієнтам отримувати вигоди за покупки в різних точках продажу.

Крос-програми дозволяють компаніям розширювати свою клієнтську базу, а також залучати більше споживачів через партнерства з іншими підприємствами. Для клієнтів це зручний спосіб отримати більше вигод, адже вони можуть використовувати свої накопичені бали в різних місцях, що знижує ризик "втратити" бонуси, якщо вони не використовуються в межах однієї компанії. Цей підхід стимулює більш часті покупки та сприяє розвитку довгострокових відносин з брендами [9].

- d. Програми на основі знижок: Традиційні системи знижок для постійних клієнтів залишаються популярними завдяки їх простоті та зрозумілості. Однак вони часто виявляються менш ефективними для формування емоційної лояльності.
- e. Цифрові програми лояльності: З появою нових технологій багато компаній переходять на цифрові платформи для управління програмами лояльності. Це дозволяє інтегрувати дані про клієнтів, використовувати мобільні додатки, створювати персоналізовані пропозиції й інтерактивно взаємодіяти з клієнтами [10].

Ефективна програма лояльності здатна принести компанії низку переваг, зокрема:

- підвищення повторюваності покупок;
- збільшення середнього чека;
- зниження витрат на залучення нових клієнтів;
- зміцнення емоційного зв'язку між клієнтом і брендом;
- зменшення ризику переходу клієнтів до конкурентів.

1.2. Методи сегментації клієнтів

Сегментація клієнтів - це практика поділу клієнтської бази на групи індивідів, які мають схожі характеристики, що стосуються маркетингу, такі як вік, стать, інтереси та звички витрачання.

Компанії, які використовують сегментацію клієнтів, розуміють, що кожен клієнт відрізняється, тому таргетинг усіх своїх клієнтів одним підходом не є ефективним. Натомість маркетингові зусилля організації спрямовані на конкретні, менші групи з повідомленнями, які споживачі вважають релевантними.

Метою сегментації клієнтів є досягнення клієнтів більш персоналізованим повідомленням, яке цікаве для них і призводить до більшої кількості продажів або конверсій клієнтів. Компанії також сподіваються отримати глибше розуміння переваг своїх клієнтів, коли вони вивчають, що кожен сегмент вважає найбільш цінним. Такий підхід дозволяє їм точніше адаптувати маркетингові матеріали до конкретного сегмента.

Сегментація клієнтів ґрунтується на ідентифікації ключових диференціаторів, щоб клієнтів можна було розділити на групи, на які можна орієнтуватися. При визначенні практики сегментації клієнтів враховується така інформація, як демографічні дані, географія, психографічні та поведінкові тенденції клієнта [11].

У маркетингу бізнес-для-бізнесу (B2B) компанії зосереджуються на таких аспектах прийняття рішень: посадах осіб, що приймають рішення, галузі, публічному чи приватному характері компанії, її розмірі та місцезнаходженні, патернах покупок та доступних технологіях.

У маркетингу бізнес-для-споживача (B2C) компанії зосереджуються на профілях, ставленні та стилі життя окремих клієнтів. Вони також можуть враховувати географічне розташування. B2C-компанії, які сегментують клієнтів за географічною ознакою, можуть адаптувати пропозиції до регіональних подій та переваг. B2C-компанії також можуть персоналізувати пропозиції на основі переважаючих мов у кожному регіоні [12].

Підходи до сегментації клієнтів B2B включають вертикальне або горизонтальне вирівнювання. У вертикальній сегментації компанії вибирають певні галузі або посади, які, ймовірно, найбільше зацікавлені їхніми продуктами, а потім зосереджують маркетингові зусилля на цих сегментах. Такий підхід дозволяє компаніям пропонувати послуги, які точно налаштовані на конкретні галузі. Наприклад, потреби фінансового сектору відрізняються від потреб охорони здоров'я. Якщо кожному сегменту пропонуються індивідуальні послуги, це може підвищити рівень прийняття та задоволення.

У горизонтальній сегментації компанії зосереджуються на одній посаді в різних галузях та організаціях. Це дозволяє сильніше зосередитися на потребах певного типу роботи або ролі. Наприклад, фокусування на фінансових директорах може допомогти створити новий маркетинговий матеріал, веб-повідомлення та електронні розсилки, спеціально адаптовані до цієї ролі [13].

Процес сегментації базується на аналізі даних про клієнтів і поділяється на декілька етапів:

- a. Збір даних: аналіз демографічних, поведінкових, психографічних і географічних характеристик клієнтів.
- b. Визначення критеріїв сегментації.
- c. Застосування методів сегментації для розподілу клієнтів на групи.
- d. Аналіз отриманих сегментів і розробка відповідних маркетингових рішень.

Методи сегментації можна поділити на традиційні (класичні) та сучасні (засновані на використанні алгоритмів машинного навчання).

Традиційні методи сегментації:

Демографічна сегментація: Клієнтів поділяють на групи за такими характеристиками, як вік, стать, рівень доходу, освіта, сімейний стан тощо. Цей метод є одним із найпростіших, проте його недоліком є обмеженість у врахуванні індивідуальних потреб і поведінкових особливостей клієнтів.

Географічна сегментація: Цей метод базується на територіальних особливостях клієнтів: місце проживання, кліматичні умови, міський чи сільський

регіон. Географічна сегментація є особливо актуальною для локальних бізнесів і компаній, які працюють на міжнародному рівні.

Поведінкова сегментація: Сегментація проводиться на основі особливостей поведінки клієнтів: частоти покупок, реакції на акції, рівня лояльності, способів використання товару. Поведінковий підхід дозволяє компанії ефективніше налаштовувати програми лояльності.

Психографічна сегментація: Цей метод орієнтується на стиль життя, цінності, мотивацію, інтереси та особистісні характеристики клієнтів. Психографічна сегментація є складною, але дає змогу виявити глибокі причини вибору товарів або послуг.

Соціально-економічна сегментація: Базується на професії, рівні доходу, фінансовій спроможності, що дозволяє адаптувати товарні пропозиції до купівельної спроможності клієнтів [14].

З розвитком великих даних (Big Data) і машинного навчання методи сегментації стали більш точними й адаптивними.

Сучасні методи сегментації:

Кластерний аналіз: Алгоритми кластеризації, такі як метод к-середніх, DBSCAN і GMM, використовуються для автоматичного виявлення груп клієнтів із подібними характеристиками. Ці методи дозволяють сегментувати клієнтську базу без заздалегідь визначених критеріїв, використовуючи лише дані [15].

Регресійний аналіз: метод, який допомагає виявляти залежності між різними факторами, що впливають на поведінку клієнтів. Він дозволяє зрозуміти, як зміни в одних умовах можуть призвести до змін в інших. У маркетинговому контексті цей підхід використовується для прогнозування реакцій клієнтів на різні маркетингові заходи. За допомогою регресійного аналізу можна дослідити, як такі фактори, як ціна, сезонність, рекламні кампанії або знижки, впливають на продажі або поведінку клієнтів. Це дає можливість зрозуміти, які саме аспекти маркетингової стратегії мають найбільший вплив на результати.

Регресійний аналіз також дозволяє прогнозувати, як клієнти можуть відреагувати на майбутні акції чи зміни в ціновій політиці, використовуючи історичні дані. Це дає змогу бізнесу краще планувати свої маркетингові стратегії і оцінювати, які з них принесуть найбільшу вигоду.

Загалом, регресійний аналіз допомагає компаніям розуміти, як різні фактори взаємодіють між собою і як ці взаємозв'язки можна використовувати для підвищення ефективності маркетингових кампаній [16].

Дерева рішень: Цей підхід ґрунтується на побудові дерева з вузлами, які відповідають за певні характеристики клієнтів, що допомагає знайти найважливіші критерії сегментації [17].

Аналіз часових рядів: використовується для вивчення змін у поведінці клієнтів з часом, зокрема для виявлення сезонних коливань у покупках або інших транзакціях. Цей підхід дозволяє зрозуміти, як змінюються показники впродовж певного періоду і виявити закономірності, що можуть допомогти в ухваленні бізнес-рішень. Наприклад, компанії можуть аналізувати сезонність попиту на свої товари чи послуги, щоб краще планувати маркетингові кампанії або управління запасами.

Крім того, за допомогою аналізу часових рядів можна виявити довгострокові тренди, які допомагають прогнозувати зміни в поведінці клієнтів і приймати рішення на основі цих трендів. Такий підхід також дозволяє виявляти аномальні зміни в попиті, що можуть свідчити про нові потреби або зміни в ринку.

Часові ряди також допомагають сегментувати клієнтів залежно від їхньої активності в різні періоди часу. Це дозволяє бізнесу проводити більш таргетовані маркетингові кампанії та пропонувати спеціальні акції або знижки в моменти, коли попит на певні товари є найбільшим.

Загалом, аналіз часових рядів є потужним інструментом для оптимізації стратегії продажів, планування попиту і підвищення ефективності маркетингових заходів [18].

Модельовання на основі ймовірностей (Bayesian Networks): Байєсівські мережі дозволяють враховувати ймовірнісні залежності між різними факторами, що впливають на рішення клієнтів, і використовувати ці дані для створення точних прогнозів.

Завдяки таким моделям можна оцінити, як клієнти можуть відреагувати на певні маркетингові стратегії, наприклад, на знижки, рекламу або акції. Це дозволяє бізнесу планувати ефективніші кампанії та спрогнозувати, хто із клієнтів з найбільшою ймовірністю здійснить покупку або підвищить свою лояльність до бренду.

Крім того, Байєсівські мережі допомагають виявити основні фактори, які впливають на поведінку клієнтів, що дає змогу більш точно налаштовувати маркетингові стратегії під конкретні потреби різних сегментів. Вони також дозволяють створювати персоналізовані пропозиції для кожного клієнта, підвищуючи ймовірність позитивної реакції на маркетингові заходи.

Таким чином, використання Байєсівських мереж дає можливість бізнесу приймати більш обґрунтовані рішення на основі ймовірнісних прогнозів, що дозволяє ефективно управляти маркетинговими кампаніями і підвищувати їх результативність [19].

Вибір методу залежить від кількості доступних даних, цілей аналізу та специфіки бізнесу. Традиційні методи є зручними для невеликих компаній із обмеженими даними, тоді як сучасні підходи краще підходять для великих підприємств із доступом до великих масивів інформації.

Сучасні методи дозволяють отримати більш точні та персоналізовані сегменти, проте їх впровадження може вимагати значних ресурсів, зокрема доступу до даних, спеціалізованого програмного забезпечення та знань у галузі аналітики.

1.3. Вплив сегментації клієнтів на ефективність програм лояльності

Правильна сегментація клієнтів у програмах лояльності є ключовим елементом стратегічного управління взаєминами з клієнтами. Вона базується на комплексному аналізі поведінкових, демографічних та психографічних характеристик споживачів і дозволяє адаптувати програми до потреб кожної окремої групи. Її важливість полягає в тому, що вона забезпечує ефективність маркетингових зусиль, підвищує залученість клієнтів і формує стійку лояльність до бренду. Компанії використовують багато інструментів таких як:

Персоналізація пропозицій та винагород

Сегментація ринку є основою для створення персоналізованих пропозицій, які дозволяють бізнесу ефективно звертатися до різних груп клієнтів, враховуючи їх індивідуальні потреби та вподобання. Завдяки цьому компанії можуть пропонувати своїм споживачам саме ті продукти чи послуги, які найбільше їх цікавлять, що підвищує ймовірність позитивного сприйняття та прийняття пропозиції.

Наприклад, одна група клієнтів може бути зацікавлена в знижках на певні категорії товарів, такі як сезонний одяг чи побутова техніка, в той час як інша група — в програмах накопичення балів або винагород для постійних покупців. Для молодших споживачів можуть бути актуальними акції, що дозволяють економити на покупках, в той час як для старших клієнтів, які орієнтовані на стабільність і вигоду в довгостроковій перспективі, більш привабливими можуть бути програми лояльності та бонуси за накопичені покупки.

Персоналізація пропозицій дозволяє не тільки задовольняти потреби різних клієнтських сегментів, а й підвищує рівень довіри до бренду. Клієнти, які отримують винагороди та вигоди, що відповідають їхнім інтересам, з більшою ймовірністю залишатимуться лояльними до компанії. Вони почуваються цінними,

оскільки бачать, що бренд дбає про їхні індивідуальні переваги. Це, в свою чергу, сприяє зміцненню клієнтських відносин, підвищенню рівня задоволення від обслуговування та збільшенню кількості повторних покупок.

Підвищення залученості клієнтів

Клієнти, яких сегментують на підставі їхнього поведінкового профілю (частота покупок, обсяг витрат, реакція на акції), можуть отримати пропозиції, що відповідають їхньому стилю споживання. Це дозволяє значно підвищити рівень їх залученості, оскільки персоналізовані акції виглядають більш релевантними та привабливими для конкретних груп.

Покращення ефективності маркетингових зусиль

Завдяки сегментації, компанія може визначити, які саме клієнти з більшою ймовірністю відреагують на певні стимули чи пропозиції, що дозволяє уникнути витрат на широкомасштабні рекламні кампанії, які можуть бути менш ефективними.

Наприклад, для активних покупців, які вже регулярно роблять покупки, можна розробити спеціальні пропозиції, такі як ексклюзивні знижки чи бонуси за лояльність, що буде стимулювати їх продовжувати взаємодію з брендом. Такий підхід дає змогу закріпити лояльність тих, хто вже зацікавлений у продуктах компанії, не витрачаючи зайвих ресурсів на безпосереднє залучення нових покупців.

З іншого боку, для менш активних клієнтів, які не роблять покупок так часто, можна використовувати спеціальні нагадування або знижки, які мотивуватимуть їх повернутися до магазину. Це може бути, наприклад, email-розсилка з пропозицією знижки на наступну покупку або нагадування про незавершені покупки в кошику. Такий підхід допомагає не тільки залучити клієнтів до повторних покупок, а й оптимізувати витрати на маркетинг, оскільки маркетингові зусилля спрямовані саме на тих споживачів, які можуть бути зацікавлені в пропозиціях.

Таким чином, сегментація дозволяє маркетологам створювати більш цілеспрямовані і персоналізовані кампанії, що зменшує витрати на рекламу, підвищує ефективність маркетингових активностей та покращує загальну рентабельність бізнесу. Це також дає змогу зберігати баланс між залученням нових клієнтів і підтримкою лояльності наявних, що в довгостроковій перспективі забезпечує стійкий розвиток бренду.

Оптимізація лояльнійної програми

Залежно від сегментів клієнтів, компанія може оптимізувати механізм програми лояльності. Наприклад, для клієнтів, які здійснюють рідкісні покупки, можна створити систему «повернення» балів або пропонувати знижки на наступну покупку для стимулювання повторних покупок. Водночас для найбільш лояльних клієнтів можна розробити ексклюзивні пропозиції, що будуть сприяти утриманню таких споживачів.

Аналіз ефективності програм лояльності

Сегментація дає змогу оцінити ефективність програм лояльності на рівні різних груп клієнтів. Це дозволяє виявляти слабкі місця і коригувати стратегію програми для досягнення кращих результатів. Наприклад, можна з'ясувати, що певний сегмент клієнтів не реагує на поточні акції, і спробувати змінити стратегію для цієї групи.

Збільшення середнього чека та частоти покупок

Таргетування спеціальних пропозицій на основі сегментації може стимулювати клієнтів до здійснення частіших покупок або збільшення середнього чека. Наприклад, сегменти високоприбуткових клієнтів можуть отримати ексклюзивні бонуси за більші покупки або за поповнення рахунку в програмі лояльності.

Підвищення рівня утримання клієнтів

Програми лояльності, орієнтовані на специфічні потреби кожного сегмента, сприяють не лише залученню нових клієнтів, а й утриманню існуючих. Для клієнтів, які регулярно взаємодіють з брендом, може бути розроблена програма статусів чи VIP-послуг, що зміцнить їхню лояльність.

Можливість коригування програм в реальному часі

Однією з головних переваг використання сегментації є можливість швидко коригувати програми лояльності на основі аналізу даних, отриманих про клієнтів. Завдяки постійному моніторингу та аналізу поведінки споживачів, компанії можуть оперативно реагувати на зміни в інтересах і потребах своїх клієнтів, адаптуючи пропозиції та винагороди відповідно до актуальних тенденцій.

Наприклад, якщо сегмент клієнтів активно реагує на певні акції або продукти, компанія може збільшити кількість балів, що нараховуються за покупку товарів із цієї категорії, щоб ще більше заохотити їх до покупок. Якщо ж якась категорія товарів не викликає великого інтересу, можна тимчасово змінити умови програми лояльності, додавши до неї вигідніші пропозиції або знижки, щоб стимулювати більшу активність клієнтів.

Крім того, компанії можуть коригувати програми лояльності в залежності від сезонних змін у поведінці споживачів. Наприклад, під час свят чи акційних періодів можна запроваджувати спеціальні пропозиції, які підвищують активність клієнтів, наприклад, збільшити бонуси за покупки в певних категоріях товарів або пропонувати ексклюзивні знижки для тих, хто зробив покупку в цей період.

Така гнучкість дозволяє компаніям підтримувати інтерес клієнтів на високому рівні, знижуючи ризик втрати лояльності та залучаючи нових покупців. Завдяки здатності швидко коригувати програми лояльності в реальному часі, бізнеси можуть бути більш адаптивними та ефективно реагувати на зміни в ринковій ситуації або поведінці своїх клієнтів [20].

1.4. Програми лояльності в Україні та їх пропозиції

Програми лояльності в Україні стали популярними серед підприємств у різних сферах, від роздрібної торгівлі до банківського сектору. Вони мають на меті заохочення постійних клієнтів та підтримку довгострокових взаємин з брендами.

Ось деякі популярні програми лояльності в Україні та їх пропозиції:

1. Програми лояльності в ритейлі

- *Сільпо*: Одна з найбільших програм лояльності в Україні, яка пропонує бонуси за покупки у магазинах мережі, а також знижки на певні товари. Підключення до програми можливе через пластикові картки або мобільний додаток. Спеціальні пропозиції для учасників програми — персоналізовані знижки та акції.
- *Ашан*: Програма лояльності з можливістю накопичення балів за покупки, які потім можна використовувати для отримання знижок на наступні покупки. Програма також надає доступ до ексклюзивних акцій та пропозицій.
- *Епіцентр*: Картка лояльності дає можливість збирати бали, які обмінюються на знижки. Крім того, учасники програми часто отримують доступ до приватних розпродажів.

2. Програми лояльності в банківському секторі

- *ПриватБанк*: Мобільний додаток "Приват24" пропонує програму лояльності, де клієнти можуть отримувати бонуси за використання карток, платежі через мобільний додаток та здійснення покупок у партнерах банку. Бонуси можна витратити на оплату послуг або отримати знижки у різних магазинах.

- *Монобанк*: Програма монобанку передбачає кешбек на покупки у партнерських мережах, а також на повернення частини коштів за категорії, як-от ресторани, супермаркети, заправки. Для користувачів програми є бонуси у вигляді знижок на фінансові послуги.

- *Ощадбанк*: Програма лояльності надає кешбек за покупки та нарахування бонусів за використання банківських карток, які можна обміняти на подарунки або знижки в партнерських магазинах.

3. Програми лояльності у сфері мобільних операторів

- *Київстар*: Програма лояльності "Київстар Преміум" дозволяє клієнтам отримувати бонуси за використання мобільного зв'язку, а також надає спеціальні пропозиції, такі як знижки на гаджети, квитки на концерти, доступ до ексклюзивних подій та сервісів.

- *МТС Україна (Vodafone)*: У рамках програми лояльності клієнти можуть отримувати бонуси за поповнення рахунку та активне користування послугами. Також є можливість брати участь в акціях і вигравати призи.

4. Програми лояльності у готельному та туристичному секторі

- *Інтерконтиненталь Готелі (IHG)*: Глобальна програма лояльності, яка надає можливість заробляти бали за ночівлі у готелях мережі, які можна обміняти на безкоштовні ночівлі, оновлення номерів або інші переваги.

- *Radisson Rewards*: Схожа програма, яка дозволяє накопичувати бали за перебування в готелях Radisson та обмінювати їх на різні вигоди, включаючи безкоштовні ночівлі, харчування або знижки.

5. Програми лояльності у ресторанах та кафе

- *Макдональдс*: Програма лояльності "Макбонус" дає можливість заробляти бали за покупки, які можна обмінювати на безкоштовні страви або знижки.

- *KFC*: Програма лояльності пропонує бонуси за кожну покупку, а також можливість отримати знижки та акційні пропозиції для учасників.

- *Coffe House*: Програма дає знижки на каву та інші напої за допомогою накопичення бонусів, а також доступ до спеціальних акцій.

6. Онлайн-магазини та платформи

- *Rozetka*: Програма лояльності включає бонуси на покупки, а також знижки на товари у партнерських магазинах. Учасники отримують спеціальні пропозиції в залежності від їхніх покупок.
- *Prom.ua*: Онлайн-платформа також пропонує програму лояльності, яка надає знижки та спеціальні пропозиції для постійних клієнтів.

РОЗДІЛ 2

МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ СЕГМЕНТАЦІЇ КЛІЄНТІВ

2.1. Огляд методів машинного навчання для сегментації клієнтів

Застосування методів машинного навчання суттєво оптимізує процес сегментації даних, надаючи змогу ефективно обробляти величезні обсяги інформації та виявляти навіть найтонші взаємозв'язки між різноманітними змінними, що недоступно для традиційних статистичних методів.

Серед найпоширеніших підходів до сегментації клієнтів можна виділити такі алгоритми машинного навчання, як кластерний аналіз, дерева рішень та лінійні моделі. Вибір конкретного методу залежить від поставлених бізнес-завдань, розміру та структури даних, а також бажаного рівня деталізації сегментів [21].

Дерева рішень:

Класифікація — це двоетапний процес у машинному навчанні, етап навчання та етап прогнозування. На етапі навчання модель розробляється на основі даних навчання. На етапі прогнозування модель використовується для прогнозування відповіді для заданих даних. Дерево рішень є одним із найпростіших і найпопулярніших алгоритмів класифікації для розуміння та інтерпретації.

Алгоритм «Дерево рішень» належить до сімейства алгоритмів керованого навчання. На відміну від інших алгоритмів навчання під наглядом, алгоритм дерева рішень також можна використовувати для вирішення проблем регресії та класифікації.

Цей метод будує ієрархічну структуру дерева шляхом ітеративного розбиття набору даних на основі значень ознак, прагнучи створити однорідні

підмножини. Цей процес, часто званий рекурсивним розбиттям, триває доти, доки не буде досягнуто визначеного критерію зупинки, такого як максимальна глибина або мінімальна кількість зразків на листовому вузлі. Вибір критеріїв розбиття, таких як інформаційний здобуток або критерій Джині, значно впливає на передбачувальну продуктивність дерева.

Менші дерева рішень, як правило, є кращими завдяки їхній інтерпретованості та зниженому ризику перенавчання. Однак складні набори даних можуть вимагати глибших дерев для захоплення складних закономірностей. Щоб зменшити перенавчання, використовуються різні методи обрізки, такі як попередня обрізка (раннє зупинення росту дерева) та наступна обрізка (видалення непотрібних гілок).

Методи ансамблів, такі як випадкові ліси, додатково підвищують передбачувальну точність дерев рішень шляхом комбінування кількох дерев, кожне з яких навчено на різному підмножині даних. Це зменшує дисперсію та покращує узагальнюючу продуктивність.

Метод дерев рішень, хоча й є потужним інструментом для аналізу даних, може не повністю відповідати складності поведінки клієнтів у програмах лояльності. Клієнти часто демонструють динамічні та різноманітні патерни поведінки, які важко укласти в жорсткі правила, властиві деревам рішень. Великі обсяги даних, які збираються в рамках програм лояльності, також можуть ускладнювати побудову ефективних моделей на основі дерев рішень. Крім того, інтерпретація результатів, отриманих за допомогою дерев рішень, може бути складною, особливо для неспеціалістів [22].

Лінійні моделі в сегментації:

Не дивлячись на те що лінійні моделі традиційно асоціюються з задачами класифікації та регресії, їх також можна використовувати для кластеризації, хоча цей підхід має свої особливості та обмеження.

Основні ідеї використання лінійних моделей для кластеризації

1. Лінійні моделі, такі як лінійна регресія або логістична регресія, дозволяють будувати гіперплощини, які розділяють простір даних. Ці гіперплощини можуть бути використані для розділення даних на кластери.

2. Для створення більш складних кластерів можна комбінувати кілька лінійних моделей. Наприклад, можна побудувати декілька гіперплощин, які розділяють дані на різні області, кожна з яких відповідає одному кластеру.

3. Кластеризацію можна розглядати як задачу класифікації з невідомими мітками класів. Тобто, спочатку ми будуємо модель, яка намагається розділити дані на кілька класів, а потім використовуємо цю модель для присвоєння кожному об'єкту кластер [23].

Лінійні методи кластеризації найкраще працюють з даними, які можна розділити на групи за допомогою прямих ліній або гіперплощин. Однак, для більш складних розподілів даних такі методи можуть бути неефективними. Крім того, перед застосуванням лінійних моделей необхідно визначити кількість кластерів, що часто є нетривіальним завданням. Інтерпретація результатів кластеризації, особливо для багатовимірних даних, також може викликати труднощі.

Кластерний аналіз

Кластерний аналіз, як і факторний аналіз, стосується збору даних, у якому змінні не були попередньо розділені на підмножини критеріїв і предикторів. Мета кластерного аналізу полягає в тому, щоб знайти подібні групи суб'єктів, де «схожість» між кожною парою суб'єктів представляє унікальну характеристику групи проти більшої сукупності. Сильна диференціація між групами вказується через окремі кластери; один кластер вказує на надзвичайно однорідні дані.

Кластерний аналіз — це алгоритм неконтрольованого навчання, тобто його використовують не знаючи, скільки кластерів існує в даних до запуску моделі. На відміну від багатьох інших статистичних методів, кластерний аналіз зазвичай використовується, коли немає припущень щодо ймовірних зв'язків у даних. Він надає інформацію про те, де існують асоціації та шаблони в даних [24].

Ієрархічні методи включають агломеративні та дивізійні методи. Агломеративні методи починають з кожного об'єкта в окремому кластері і поступово об'єднують найближчі кластери, поки всі об'єкти не опиняться в одному кластері. Дивізійні методи, навпаки, починають з усіх об'єктів в одному кластері та поступово розділяють його на менші кластери, поки кожен об'єкт не буде належати окремому кластеру.

Розподільні методи припускають, що дані генеруються з певного розподілу ймовірностей і намагаються знайти розподіл, який найкраще відповідає даним. Один з найбільш популярних розподільних методів — це метод k -середніх, який вибирає випадкові центри кластерів і потім повторює дві основні процедури: спочатку призначає кожен об'єкт до найближчого центру, а потім перераховує центри кластерів як середні значення об'єктів у кожному з них.

Щільнісні методи визначають кластери як області з високою щільністю об'єктів, які відокремлені від низькощільних областей. Одним з найпопулярніших методів є DBSCAN, який визначає кластери на основі ϵ -окола та мінімальної кількості точок в цьому колі.

Модельні методи використовують статистичні моделі для опису даних, і одним з таких підходів є модель сумішей Гауссових розподілів, яка припускає, що дані генеруються з суміші кількох гауссових розподілів [25].

2.2. Використання методу k -середніх для кластеризації клієнтів

Метод k -середніх (k -means) є одним з найбільш популярних методів кластеризації, який класифікує об'єкти в k кластерів так, щоб елементи в одному кластері були максимально схожими один до одного, а елементи з різних кластерів максимально відрізнялися. Це досягається шляхом мінімізації суми квадратів

відстаней між точками та центрами кластерів. Алгоритм використовує ітераційний процес, доки не досягне оптимального розподілу [26].

Алгоритм методу k -середніх складається з кількох основних кроків:

Ініціалізація:

- Вибір числа кластерів k , що визначає кількість груп, на які будуть розподілені дані.
- Ініціалізація початкових центрів кластерів (можна вибирати випадкові точки з даних або використовувати більш складні методи, як k -means++).

Призначення кластерів:

- Для кожного об'єкта (клієнта) визначається відстань до кожного з k центрів кластерів.
- Об'єкти призначаються до кластера, центр якого є найближчим.

Оновлення центрів кластерів:

- Після того, як всі об'єкти призначені до своїх кластерів, обчислюється новий центр кожного кластеру, як середнє значення всіх об'єктів, що належать до цього кластера.

Кроки призначення кластерів та оновлення центрів повторюються, поки центри кластерів не перестануть змінюватися (або зміни будуть дуже малі) [27].

1. Визначення центру кластеру:

Нехай $X = \{x_1, x_2, \dots, x_n\}$ — набір об'єктів (клієнтів), де кожен об'єкт x_i має вигляд вектору ознак, наприклад, $x_i = (x_{i1}, x_{i2}, \dots, x_{im})$, де m — кількість ознак.

Кластери позначаються як $C_1, C_2, \dots, C_k$, де k — кількість кластерів.

Центр кластера C_j визначається як середнє значення всіх точок, що належать кластеру C_j :

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2.1)$$

де $|C_j|$ — кількість елементів у кластері C_j ,

μ_j — координати центру цього кластера.

2. Призначення об'єктів до кластерів:

Для кожного об'єкта x_i визначається відстань до всіх центрів кластерів. В якості метрики зазвичай використовують евклідову відстань:

$$d(x_i, \mu_j) = \sqrt{\sum_{l=1}^m (x_{il} - \mu_{jl})^2} \quad (2.2)$$

Об'єкт x_i призначається до кластеру C_j , якщо відстань до його центру μ_j є мінімальною серед усіх відстаней до інших центрів:

$$C_i = \operatorname{argmin}_j \{d(x_i, \mu_j)\} \quad (2.3)$$

Тобто, C_i — це кластер, до якого належить точка x_i .

3. Мінімізація функції варіації (цільова функція):

Алгоритм k -середніх намагається мінімізувати функцію варіації (суму квадратів відстаней від точок до їхніх центрів), що можна записати так:

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} d(x_i, \mu_j)^2 \quad (2.4)$$

Функція J вимірює "якість" кластеризації, де ми прагнемо, щоб точки в кожному кластері були якомога ближче до свого центру.

4. Оновлення центрів:

Після того як об'єкти були призначені до кластерів, центри оновлюються, як було зазначено раніше за допомогою формули (2.1).

Цей процес триває до тих пір, поки зміни в центрах кластерів не стануть дуже малими, або поки не досягнемо максимального числа ітерацій [28].

Метод k -середніх є ефективним інструментом для кластеризації, який дозволяє розділити набір даних на k груп за допомогою мінімізації функції варіації. Однак правильний вибір числа кластерів k і початкових центрів може істотно вплинути на кінцевий результат, тому часто використовують методи ініціалізації та оцінки якості кластеризації для покращення результатів.

2.3. Використання методу DBSCAN для кластеризації клієнтів

Метод DBSCAN (Density-Based Spatial Clustering of Applications with Noise) є популярним методом для кластеризації даних, особливо в тих випадках, коли потрібно працювати з неструктурованими даними або даними з шумом. Він належить до класу алгоритмів, що базуються на концепції щільності: точки, які розташовані в області з достатньою щільністю, об'єднуються в кластер, тоді як точки в рідких областях позначаються як шуми [29].

Алгоритм DBSCAN має два основних параметри:

- Eps (ϵ) — максимальна відстань між двома точками, щоб вони могли належати до одного кластера.
- MinPts — мінімальна кількість точок, які повинні бути в околі точки для того, щоб вона могла бути центром кластера [30].

В алгоритмі DBSCAN точки класифікуються за трьома категоріями:

- Core Points (точки ядра): Точка є точкою ядра, якщо в її Eps-околі міститься хоча б MinPts точок. Вона має достатньо сусідів для утворення кластера.
- Border Points (точки на межі): Точка є точкою на межі, якщо вона не є точкою ядра, але належить до Eps-округу якої-небудь точки ядра.
- Noise Points (шумові точки): Точки, які не є ні точками ядра, ні точками на межі, тобто не мають достатньої щільності для того, щоб бути включеними в кластер [31].

Алгоритм DBSCAN зазвичай використовує евклідову відстань для визначення того, чи є дві точки в Eps-околі одна одної. Евклідова відстань між двома точками $p=(p_1, p_2, \dots, p_n)$ і $q=(q_1, q_2, \dots, q_n)$ у просторі з n вимірів визначається як:

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (2.5)$$

де p_i і q_i — координати точок в n -вимірному просторі.

Алгоритм шукає всі точки, що належать до Eps-округу для кожної точки. Для кожної точки p визначаються всі точки q , для яких виконується умова:

$$d(p, q) \leq \epsilon \quad (2.6)$$

Якщо для точки p кількість сусідів (включаючи її саму) більша або рівна значенню MinPts, то ця точка є точкою ядра [32].

Алгоритм DBSCAN кластеризує дані за допомогою концепції щільності, працюючи наступним чином:

1. Ініціалізація: Спочатку всі точки позначаються як невідвідані.
2. Початкова точка: Вибирається перша невідвідана точка, і для неї знаходяться всі точки, що лежать у її Eps-окрузі (тобто, ті, що знаходяться на відстані не більше ніж Eps від вибраної точки).
3. Визначення типу точки:
 - Якщо в Eps-окрузі точки є більше ніж MinPts точок, то ця точка стає точкою ядра, і на основі неї створюється новий кластер.
 - Якщо точка не є точкою ядра, але знаходиться в Eps-окрузі точки ядра, то вона стає точкою на межі і додається до цього кластера.
 - Якщо точка не має достатньої щільності (менше ніж MinPts точок у її Eps-окрузі), вона вважається шумом і не потрапляє в жоден кластер.
4. Розширення кластера: Якщо точка є точкою ядра, то кластер розширюється, включаючи всі точки, що знаходяться в її Eps-окрузі. Цей процес триває, поки нові точки не виявляться або точками ядра, або точками на межі.
5. Завершення: Алгоритм завершується, коли всі точки або потрапляють до кластерів, або позначаються як шум.

Якщо є набір точок $X = \{x_1, x_2, \dots, x_n\}$, де кожна точка має свої координати в просторі, то DBSCAN класифікує ці точки в два можливих стани:

- Кластер $C \subseteq X$, де кожен елемент $x_i \in C$ є точкою ядра або точкою на межі.
- Шум $N \subseteq X$, де точки не належать до жодного кластера [33].

Однією з основних переваг DBSCAN є те, що він не вимагає вказувати кількість кластерів заздалегідь. Але для оцінки якості кластеризації можна використовувати кілька метрик, таких як:

- Силует (Silhouette score): оцінює, наскільки добре кожна точка вписується в свій кластер порівняно з іншими кластерами.
- Індекс Данна (Dunn index): вимірює мінімальну відстань між точками різних кластерів і максимальний радіус кластера [34].

2.4. Використання методу GMM для кластеризації клієнтів

Метод GMM (Gaussian Mixture Model) — це статистичний підхід до кластеризації, який використовує ймовірнісну модель для представлення даних як суміші кількох нормальних (гауссових) розподілів. Цей метод є розширенням K-means, оскільки він дозволяє кластерам мати різні форми та розміри, а також враховує варіації в даних. GMM кластеризує дані, враховуючи ймовірність того, що точка належить до того чи іншого кластеру [35].

Модель гауссових сумішей (Gaussian Mixture Model):

- Припускається, що дані є результатом змішування кількох нормальних розподілів (гауссів). Кожен кластер можна описати за допомогою нормального розподілу з певними параметрами:
 - Середнє (μ_k) — центри кластерів.
 - Коваріація (Σ_k) — розмір і форма кластерів.
 - Ймовірність (вага π_k) — ймовірність того, що точка належить до конкретного кластеру.

Формула ймовірності для GMM виглядає так:

$$p(x) = \sum_{k=1}^K \pi_k * N(x|\mu_k, \Sigma_k) \quad (2.7)$$

де:

- K — кількість кластерів,
- π_k — ймовірність (вага) k -го кластера,
- $N(x|\mu_k, \Sigma_k)$ — нормальний розподіл для кожного кластеру.

Алгоритм EM (Expectation-Maximization): GMM використовує алгоритм EM (Expectation-Maximization) для оцінки параметрів моделі:

- Е-крок (Expectation): оцінка ймовірностей того, що кожна точка належить до кожного з кластерів (з урахуванням параметрів моделей для кожного кластера).
- М-крок (Maximization): оновлення параметрів кожного кластеру (середнє, коваріація та вага), щоб максимізувати ймовірність даних, що спостерігаються.

Ці кроки повторюються, поки алгоритм не досягне стійкого результату [36].

Основними параметрами для GMM є:

- Кількість кластерів (K): необхідно вибрати кількість кластерів, яку GMM буде використовувати для моделювання даних. Це можна зробити через експериментування або використовуючи критерії оцінки моделей, такі як AIC (Akaike Information Criterion) або BIC (Bayesian Information Criterion).
- Тип коваріації: GMM дозволяє моделювати різні типи коваріацій:
 - Сферична коваріація: всі кластери мають однаковий розмір і форму.
 - Повна коваріація: кожен кластер має власний розмір і форму.
 - Теплова коваріація: усі кластери мають однакову форму, але різні розміри [37].

2.5. Порівняння методів кластеризації та критерії оцінки якості кластерів

Кожен з розглянутих алгоритмів кластеризації — DBSCAN, GMM та K-means — має свої сильні сторони, недоліки та специфічні сценарії застосування.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) є алгоритмом кластеризації, який має як переваги, так і обмеження. Його ключовою особливістю є здатність визначати кластери будь-якої форми, що робить його придатним для аналізу складних даних. Важливою перевагою є робота з "шумовими" точками, які він розпізнає та ігнорує, а також відсутність потреби у заздалегідь визначеній кількості кластерів, що зручно для роботи з невідомими наборами даних.

Однак DBSCAN залежить від правильного вибору параметрів Eps (радіус сусідства) і MinPts (мінімальна кількість точок для формування кластера), що може ускладнювати застосування. Він менш ефективний для даних із кластерами різної щільності та може бути повільним на великих наборах даних [38].

Gaussian Mixture Model (GMM) — це ймовірнісний підхід до кластеризації, який має низку переваг та обмежень. Його головна перевага полягає у гнучкості: GMM дозволяє створювати кластери різних форм і розмірів завдяки використанню різних параметрів коваріації. Крім того, ймовірнісна природа моделі дає змогу враховувати належність точки до кількох кластерів одночасно, що особливо корисно для задач із нечіткими межами між кластерами. GMM добре підходить для даних із нормальним розподілом, що робить його ефективним для кластерів із різними стандартними відхиленнями.

Серед недоліків — необхідність заздалегідь вказувати кількість кластерів, що може бути складним, якщо їх кількість невідома. Також процес налаштування параметрів коваріації вимагає ретельного вибору, оскільки неправильний підхід

може погіршити результати. Окрім цього, через метод ЕМ алгоритм є обчислювально витратним і може бути повільним на великих наборах даних [39].

K-means — популярний алгоритм кластеризації, відомий своєю простотою та швидкістю. Завдяки низькій обчислювальній складності він ефективний для великих наборів даних, а його результати легко інтерпретувати через визначені центри кластерів. Широка підтримка алгоритму в різних бібліотеках робить його доступним і зручним у використанні.

Проте K-means має свої обмеження. Він вимагає попереднього визначення кількості кластерів, що може бути складним для незнайомих даних. Результати алгоритму залежать від початкового вибору центроїдів, що іноді призводить до локального оптимуму. Метод краще працює для сферичних кластерів і може бути неефективним для даних зі складними або несферичними структурами. Крім того, він чутливий до шуму, оскільки викиди можуть впливати на кінцеву кластеризацію [40].

DBSCAN підходить для ситуацій, де важливо виявляти кластери будь-якої форми та працювати з даними, що містять шум. Він ідеальний для реальних даних з нестандартними структурами і варіаціями в щільності. Однак вибір параметрів може бути складним, і він не підходить для випадків, де кластери мають різну щільність.

GMM є потужним методом для моделювання даних, коли кластери мають еліптичні форми або різну величину. Він дозволяє м'яко класифікувати точки та працювати з даними, що мають ймовірнісну природу. Однак для великих наборів даних та складних ситуацій GMM може бути обчислювально важким.

K-means є відмінним вибором для простих задач, де кластери мають конвексну форму, і можна легко визначити кількість кластерів. Він дуже швидкий і ефективний для великих наборів даних, але не підходить для складних або несферичних кластерів і чутливий до початкових умов.

Оцінка якості кластеризації є важливим етапом аналізу результатів алгоритмів кластеризації, оскільки без об'єктивних критеріїв неможливо точно

визначити, наскільки правильно алгоритм розподілив об'єкти на групи. Існує кілька теоретичних підходів і критеріїв для оцінки якості кластеризації. Вони можуть бути поділені на кілька типів залежно від того, чи мають вони доступ до істинних міток (знань) про класи (як у випадку з навчанням з учителем), чи ні (як у випадку з навчанням без учителя). Ось основні критерії:

Внутрішні критерії

Ці критерії не потребують зовнішньої інформації про правильність кластерів (не вимагають міток класів). Вони оцінюють якість кластеризації, спираючись на структуру самих кластерів, їх компактність і відокремленість.

1.1. Індекс силуета (Silhouette Index)

Індекс силуета оцінює, наскільки добре кожен об'єкт відповідає своєму кластеру, порівнюючи відстань між об'єктами в межах одного кластера та відстань до найближчого кластера.

Для кожного об'єкта i розраховуються два параметри:

— Відстань $a(i)$: середня відстань від об'єкта i до всіх інших об'єктів в його кластері:

$$a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, j \neq i} d(i, j) \quad (2.8)$$

де C_i — це кластер, до якого належить об'єкт i ,

$d(i, j)$ — відстань між об'єктами i і j .

— Відстань $b(i)$: середня відстань від об'єкта i до всіх об'єктів найближчого до нього кластера:

$$b(i) = \min_{C_k \neq C_i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \quad (2.9)$$

де C_k — інші кластери, до яких належать об'єкти,

$d(i, j)$ — відстань між об'єктами i та j .

Силует об'єкта i визначається як:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (2.10)$$

Індекс силуета для всіх об'єктів в наборі даних:

$$S = \frac{1}{n} \sum_{i=1}^n s(i) \quad (2.11)$$

де n — кількість об'єктів у даних.

Значення S може варіюватися від -1 до 1, де значення, близьке до 1, вказує на хорошу кластеризацію, близьке до 0 — на погану, а негативне значення — на те, що об'єкти можуть бути неправильно класифіковані [41].

1.2. Внутрішньокластерна дисперсія

Цей критерій оцінює, наскільки сильно відрізняються об'єкти всередині одного кластера. Чим менша дисперсія в кластері, тим він більш компактний.

Внутрішньокластерна дисперсія для кластеру C_k обчислюється як:

$$WCSS(C_k) = \sum_{i \in C_k} ||x_i - \mu_k||^2 \quad (2.12)$$

де μ_k — це центроїд (середнє значення) кластеру C_k ,

x_i — об'єкт в кластері C_k .

Загальна внутрішньокластерна дисперсія для всіх кластерів:

$$WCSS = \sum_k^K WCSS(C_k) \quad (2.13)$$

де K — кількість кластерів. [42]

Зовнішні критерії

Зовнішні критерії вимагають наявності істинних міток класів для порівняння з результатами кластеризації.

2.1. Індекс Ранд (Rand Index)

Індекс Ранд вимірює схожість між двома класифікаціями, порівнюючи пари об'єктів. Індекс Ранд визначається як:

$$RI = \frac{a+b}{\binom{n}{2}} \quad (2.14)$$

де:

- a — кількість пар об'єктів, які в обох класифікаціях належать до одного класу або до різних класів;
- b — кількість пар об'єктів, які належать до різних класів у обох класифікаціях;
- n — загальна кількість об'єктів;
- $\binom{n}{2}$ — кількість всіх можливих пар об'єктів [43].

2.2. Коригований індекс Ранда (Adjusted Rand Index)

Коригований індекс Ранда враховує ймовірність того, що випадково два об'єкти будуть в одному класі, і дає коригування для випадкових результатів. Він визначається як:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad (2.15)$$

де $E[RI]$ — очікуване значення індексу Ранда для випадкових класифікацій,
 $\max(RI)$ — максимальне можливе значення індексу Ранда [44].

2.3. Нормалізована взаємна інформація (NMI)

Нормалізована взаємна інформація визначає, скільки інформації про істинні мітки класів можна отримати за допомогою результатів кластеризації. Вона базується на концепції *взаємної інформації*.

Взаємна інформація між двома змінними X і Y визначається як:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2.16)$$

де $p(x, y)$ — спільна ймовірність для об'єктів x і y ,

$p(x)$ і $p(y)$ — ймовірності для x і y окремо.

Нормалізована взаємна інформація:

$$NMI(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}} \quad (2.17)$$

де $H(X)$ і $H(Y)$ — ентропії для змінних X і Y відповідно [45].

Динаміка кластерів

Існує також потреба оцінювати, як зміна даних або параметрів кластеризації впливає на якість результату. Це дозволяє зрозуміти стабільність кластерів, коли дані змінюються або додаються нові об'єкти.

Крос-валідація

Крос-валідація полягає в розбитті набору даних на кілька підмножин (folds), де одна частина використовується для навчання, а інші — для тестування. Це дозволяє оцінити, наскільки добре кластеризація може узагальнювати на нові дані.

РОЗДІЛ 3

ПРАКТИЧНА РЕАЛІЗАЦІЯ МЕТОДІВ КЛАСТЕРІЗАЦІЇ КЛІЄНТІВ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

3.1. Опис набору даних та препроцесинг даних

У рамках кваліфікаційної роботи здійснено аналіз великого обсягу даних, що стосуються клієнтів роздрібного магазину. Метою дослідження було виявлення споживчих тенденцій, аналіз взаємозв'язків між різними характеристиками клієнтів та їх кластеризація для створення персоналізованих рекомендацій.

Дослідження базувалося на обробці набору даних, який включав такі показники:

- ‘ID’, унікальний ідентифікатор кожного клієнта.
- ‘Year_Birth’, рік народження клієнта (для розрахунку віку).
- ‘Complain’, позначка, чи надходили скарги від клієнта протягом останніх двох років (1,скарга була, 0,не було).
- ‘Dt_Customer’, дата реєстрації клієнта у системі компанії.
- ‘Education’, рівень освіти клієнта.
- ‘Marital’, сімейний статус клієнта.
- ‘Kidhome’, кількість малолітніх дітей у домогосподарстві клієнта.
- ‘Teenhome’, кількість підлітків у домогосподарстві клієнта.
- ‘Income’, річний дохід домогосподарства клієнта.
- ‘MntFishProducts’, сума витрат на рибну продукцію за останні два роки.
- ‘MntMeatProducts’, сума витрат на м'ясну продукцію за останні два роки.
- ‘MntFruits’, сума витрат на фрукти за останні два роки.
- ‘MntSweetProducts’, сума витрат на кондитерські вироби за останні два роки.
- ‘MntWines’, сума витрат на винну продукцію за останні два роки.

- ‘MntGoldProds’, сума витрат на ювелірну продукцію за останні два роки.
- ‘NumDealsPurchases’, кількість покупок зі знижками.
- ‘NumCatalogPurchases’, кількість покупок через каталог (замовлення товарів поштою).
- ‘NumStorePurchases’, кількість покупок, здійснених безпосередньо в магазинах.
- ‘NumWebPurchases’, кількість покупок, здійснених через вебсайт компанії.
- ‘NumWebVisitsMonth’, кількість відвідувань вебсайту компанії протягом останнього місяця.

Збір даних для дослідження здійснювався з використанням відкритої платформи Kaggle, яка є популярним ресурсом для аналітичних та машинних навчальних проєктів. Kaggle забезпечує доступ до якісних, структурованих наборів даних, що дозволяє проводити аналіз у різних галузях [46].

Перед застосуванням методів кластеризації необхідно здійснити детальне ознайомлення з початковим набором даних та виконати їх ретельну підготовку. Цей етап є ключовим у процесі аналізу, оскільки від якості та повноти вхідних даних залежить ефективність та достовірність отриманих результатів. Підготовка даних включає перевірку їх структури, виявлення пропусків, усунення аномалій, нормалізацію значень та трансформацію змінних у зручний для обробки формат. Лише після виконання цих кроків можливе коректне застосування кластеризаційних алгоритмів.

Спочатку потрібно ознайомитись з першими строками датасету за допомогою команди `.head()`.

```
data.head()
```

	Id	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPu
0	1826	1970	Graduation	Divorced	84835.0	0	0	6/16/2014	0	189	...	111	189	218	
1	1	1961	Graduation	Single	57091.0	0	0	6/15/2014	0	464	...	7	0	37	
2	10476	1958	Graduation	Married	67267.0	0	1	5/13/2014	0	134	...	15	2	30	
3	1386	1967	Graduation	Together	32474.0	1	1	11/5/2014	0	10	...	0	0	0	
4	5371	1989	Graduation	Single	21474.0	1	0	8/4/2014	0	6	...	11	0	34	

5 rows × 22 columns

Рисунок 3.1 — Перші рядки датасету

Джерело: розроблено автором на основі [46]

Знання формату даних у змінних є ключовим етапом аналізу, оскільки воно забезпечує правильну обробку та точну інтерпретацію інформації. Різні формати, такі як текстові рядки, цілі числа чи числа з плаваючою комою, мають свої унікальні властивості та підходять для виконання різних обчислень [47]. Розуміння цих особливостей дозволяє уникнути помилок у процесі аналізу, забезпечує коректну обробку даних та полегшує інтеграцію з іншими аналітичними інструментами, бібліотеками чи алгоритмами, які можуть висувати специфічні вимоги до формату. Для перевірки формату змінних у наборі даних скористаємося командою `.info()`:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2240 entries, 1826 to 4070
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Year_Birth             2240 non-null   int64
1   Education              2240 non-null   object
2   Marital_Status         2240 non-null   object
3   Income                 2216 non-null   float64
4   Kidhome                2240 non-null   int64
5   Teenhome               2240 non-null   int64
6   Dt_Customer            2240 non-null   object
7   Recency                2240 non-null   int64
8   MntWines               2240 non-null   int64
9   MntFruits              2240 non-null   int64
10  MntMeatProducts        2240 non-null   int64
11  MntFishProducts        2240 non-null   int64
12  MntSweetProducts       2240 non-null   int64
13  MntGoldProds           2240 non-null   int64
14  NumDealsPurchases      2240 non-null   int64
15  NumWebPurchases        2240 non-null   int64
16  NumCatalogPurchases    2240 non-null   int64
17  NumStorePurchases      2240 non-null   int64
18  NumWebVisitsMonth       2240 non-null   int64
19  Response               2240 non-null   int64
20  Complain               2240 non-null   int64
dtypes: float64(1), int64(17), object(3)
memory usage: 385.0+ KB
```

Рисунок 3.2 — Формат змінних датасету

Джерело: розроблено автором на основі [46]

Набір даних містить 2240 записів і 21 змінну. Типи даних розподілені наступним чином: 17 змінних представлені у вигляді цілих чисел (int64), серед яких рік народження, кількість дітей, витрати на різні категорії продукції та кількість покупок. Одна змінна (Income) містить числові дані з плаваючою комою (float64) і відображає річний дохід клієнтів. Три змінні представлені у текстовому форматі (object), зокрема рівень освіти (Education), сімейний статус (Marital_Status) та дата реєстрації клієнта (Dt_Customer). Усі змінні, окрім Income, мають повні дані, тоді як у цьому стовпці виявлено 24 пропущені значення.

Перед проведенням кластерного аналізу було виконано видалення трьох змінних: Education, Marital_Status та Dt_Customer. Це здійснювалося за допомогою методу drop, який дозволяє видаляти вибрані стовпці з набору даних. Такі дії зазвичай виконуються, коли ці змінні вважаються неінформативними або нерелевантними для конкретного аналізу.

```
data.drop('Education', axis=1, inplace=True)  
data.drop('Marital_Status', axis=1, inplace=True)  
data.drop('Dt_Customer', axis=1, inplace=True)
```

Рисунок 3.3 — Видалення змінних

Джерело: розроблено автором на основі [46]

Education та Marital_Status: Обидві змінні мають текстовий формат і можуть бути складними для прямого використання в алгоритмах кластеризації, які вимагають числових даних. Хоча можливо перетворити ці змінні на числові коди або вектори, вони можуть незначно впливати на результат або навіть ускладнювати аналіз.

Dt_Customer: Дата реєстрації клієнта також може бути не надто значущою для кластеризації, оскільки вона здебільшого відображає часову інформацію, яка не обов'язково корелює зі споживчою поведінкою чи іншими релевантними характеристиками.

Видалення цих змінних дозволяє зменшити розмірність даних, спрощуючи модель і знижуючи ризик внесення шуму до кластерного аналізу. Це забезпечує більш точне формування кластерів на основі ключових змінних, які мають прямий зв'язок із метою дослідження.

Надалі було важливо заповнити пропущені значення у стовпці Income середнім значенням цього стовпця. Такий підхід забезпечує цілісність даних, оскільки пропуски можуть впливати на точність статистичного аналізу або викликати помилки під час застосування алгоритмів машинного навчання. Використання середнього значення як заміни є стандартною практикою, яка дозволяє зберегти загальну структуру даних без значного викривлення результатів.

```
data['Income'].fillna(data['Income'].mean(), inplace=True)
```

Рисунок 3.4 — Заповнення пропущених значень

Джерело: розроблено автором на основі [46]

3.2. Реалізація та порівняння трьох методів кластеризації

Перед початком реалізації методів кластеризації було проведено стандартизацію даних:

```
from sklearn.preprocessing import StandardScaler  
scaler = StandardScaler() # обираємо бажаний метод стандартизації  
scaler.fit(data)  
X_stand = scaler.fit_transform(data)
```

Рисунок 3.5 — Стандартизація даних

Джерело: розроблено автором на основі [46]

Стандартизація даних є важливим етапом попередньої обробки для багатьох алгоритмів машинного навчання, особливо тих, які залежать від відстаней між точками даних. Без стандартизації змінні з різними масштабами та одиницями вимірювання можуть мати нерівний вплив на модель [48].

Метод `scaler.fit(data)` на рис.3.5 обчислює середнє значення та стандартне відхилення для кожної змінної в наборі даних, що необхідно для подальшого процесу стандартизації. Стандартизація передбачає віднімання середнього значення та поділ на стандартне відхилення для кожної змінної. Команда `X_stand = scaler.fit_transform(data)` поєднує два етапи: спочатку обчислюються статистики (середнє та стандартне відхилення), а потім ці значення використовуються для перетворення даних таким чином, щоб кожна змінна мала середнє значення 0 і стандартне відхилення 1. Отриманий результат зберігається у новому масиві `X_stand`.

Після проведення всіх необхідних процедур, спочатку був створений об'єкт класу `KMeans` з параметрами, що визначають кількість кластерів (3) і фіксований генератор випадкових чисел для забезпечення відтворюваності результатів (рис. 3.6). Потім метод `fit()` навчає модель на стандартизованих даних, визначаючи центри кластерів і класифікуючи кожен об'єкт у набір відповідних груп. За допомогою методу `predict()` модель прогнозує, до якого з кластерів належать всі об'єкти, і зберігає ці мітки у змінній `labels`. В кінці виводяться мітки, що вказують на кластер, до якого належить кожен об'єкт. Така кластеризація дозволяє групувати схожі об'єкти в окремі категорії для подальшого аналізу.

```
[ ]
    from sklearn.cluster import KMeans

[ ] kmeans = KMeans(n_clusters=3, random_state=42)

[ ] kmeans.fit(X_stand)

↔ KMeans
   KMeans(n_clusters=3, random_state=42)

[ ] labels = kmeans.predict(X_stand)
    print(labels)

↔ [1 0 0 ... 2 1 1]
```

Рисунок 3.6 — Реалізація алгоритму K-means

Джерело: розроблено автором на основі [46]

Але для визначення оптимальної кількості кластерів у алгоритмах кластеризації, таких як K-means використовується Метод локтя (рис. 3.7). Він ґрунтується на аналізі графіка, який показує залежність інерції (міри якості кластеризації) від числа кластерів. Оптимальна кількість кластерів вибирається на основі точки, де зменшення інерції починає сповільнюватися, що вказує на найбільш ефективний поділ даних. Це дозволяє знайти найкраще число кластерів, яке мінімізує помилки кластеризації.

```
klusters = range(2, 11) # задамо кількість кластерів від 2 до 10 (верхня межа в функції range береться невиключно)
inertias = []
for k in klusters:
    # задаємо метод кластеризації - k-середніх
    model = KMeans(n_clusters=k)

    # оцінюємо модель
    model.fit(X_stand)

    # для кожної k-ті кластерів розраховуємо inertia, розраховані значення додаємо в список inertias
    inertias.append(model.inertia_)

# зобразимо на графіку показник inertia для кожної k-ті кластерів
plt.figure(figsize=(6,3))
plt.plot(klusters, inertias, '-o', color='red')
plt.xlabel('number of clusters')
plt.ylabel('inertia')
plt.xticks(klusters)
plt.show()
```

Рисунок 3.7 — Програмний код до методу локтя

Джерело: розроблено автором на основі [46]

Спочатку був створений діапазон можливих значень кількості кластерів від 2 до 10, після чого для кожного з цих значень оцінюється модель K-means. Для кожної кількості кластерів виконується навчання моделі на стандартизованих даних, після чого обчислюється значення інерції — метрики, яка вимірює суму квадратів відстаней між точками в кластері та центром цього кластера. Це значення додається до списку інерцій для подальшого аналізу. За допомогою візуалізації отриманих даних будується графік, що показує залежність інерції від кількості кластерів. Під час аналізу графіка оптимальна кількість кластерів визначається як точка, де зменшення інерції починає сповільнюватися, що вказує на найбільш ефективний поділ даних на групи. Був отриманий наступний графік:

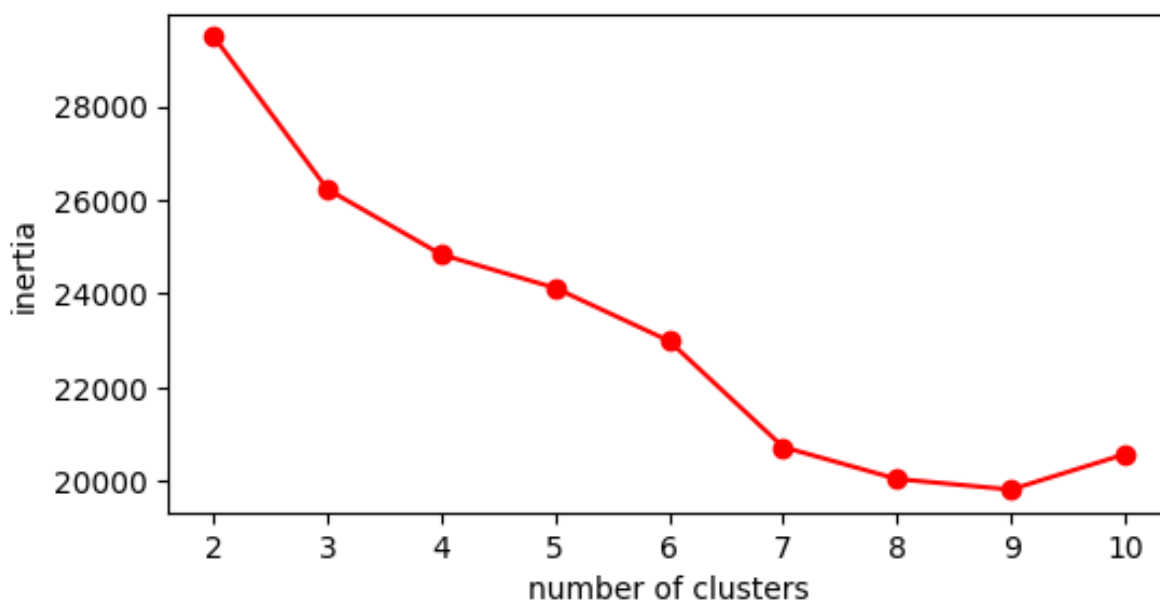


Рисунок 3.8 — Графік для визначення оптимальної кількості кластерів

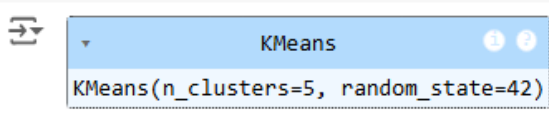
Джерело: розроблено автором на основі [46]

З аналізу графіка видно, що після п'яти кластерів спостерігається значне сповільнення темпу зменшення інерції, що свідчить про оптимальну кількість кластерів для даного набору даних. Це є показником того, що подальше збільшення кількості кластерів не призводить до суттєвого поліпшення якості кластеризації.

Тому варто провести повторну кластеризацію методом K-means з визначеною оптимальною кількістю кластерів:

```
[ ] kmeans = KMeans(n_clusters=5, random_state=42)

[ ] kmeans.fit(X_stand)
```



```
[ ] labels = kmeans.predict(X_stand)
print(labels)
```



Рисунок 3.9 — Реалізація методу K-means для 5 кластерів

Джерело: розроблено автором на основі [46]

Оцінка якості кластеризації є важливим етапом в аналізі даних, оскільки вона дозволяє визначити, наскільки добре алгоритм кластеризації поділяє дані на групи.

В даній роботі методи кластеризації оцінюватимуться за допомогою трьох критеріїв, а саме: інерції, силуетного коефіцієнта(Silhouette Score), індекса Калінскі-Харабаса (Calinski-Harabasz Index).

Для кластеризації k-means, було отримано наступні результати:

```
inertia = kmeans.inertia_
print(f'Inertia: {inertia}')
```

Inertia: 24266.162497826965

Рисунок 3.10 — Інерція для k-means

Джерело: розроблено автором на основі [46]

Результат кластеризації з інерцією 24266.16 свідчить про суму квадратів відстаней між точками даних та центрами їхніх відповідних кластерів. Це значення показує, наскільки добре модель змогла розподілити дані по кластерах. Чим менша інерція, тим краща кластеризація, оскільки це вказує на менші відстані між точками та центрами кластерів, що свідчить про більш чітке та згуртоване поділ даних.

```
from sklearn.metrics import silhouette_score

silhouette_avg = silhouette_score(X_stand, labels)
print(f'Silhouette Score: {silhouette_avg}')

Silhouette Score: 0.18915584208468017
```

Рисунок 3.11 — Силуетний коефіцієнт для *k-means*

Джерело: розроблено автором на основі [46]

Результат силуетного коефіцієнта **0.189** вказує на те, що якість кластеризації є середньою. Силуетний коефіцієнт вимірює, наскільки добре кожен об'єкт належить до свого кластеру, порівнюючи середнє відстані до інших точок у своєму кластері та до точок найближчого іншого кластера.

```
from sklearn.metrics import calinski_harabasz_score

# Обчислення індексу Калінскі-Харабаса
calinski_score = calinski_harabasz_score(X_stand, labels)
print(f'Calinski-Harabasz Index: {calinski_score}')

Calinski-Harabasz Index: 369.65390407908194
```

Рисунок 3.12 — індекс Калінскі-Харабаса для *k-means*

Джерело: розроблено автором на основі [46]

Індекс Калінскі-Харабаса 369.65 вказує на досить хороший результат кластеризації. Індекс Калінскі-Харабаса оцінює баланс між внутрішньокласовою і

міжкласовою дисперсією: чим вищий цей індекс, тим краще розділені кластери. Він обчислюється як відношення між дисперсією між кластерами (яка повинна бути великою) і дисперсією всередині кожного кластера (яка повинна бути малою).

Зазвичай, чим вищий індекс Калінскі-Харабаса, тим кластери є більш чітко відокремленими і внутрішньокласова дисперсія менша. Оскільки значення індексу в цьому випадку досить велике, це свідчить про те, що кластери добре розділені та мають відносно малу дисперсію всередині себе [49].

Наступним методом кластеризації, який буде застосовано, є DBSCAN.

Для початку було виконано пошук оптимальних параметрів для алгоритму кластеризації DBSCAN шляхом перебору різних варіантів параметрів:

```
from sklearn.model_selection import ParameterGrid

params = {'eps': [1.5, 2.5, 2.8], 'min_samples': [7, 10, 15]}
best_score = -1
best_params = None # задамо змінну до циклу

for param in ParameterGrid(params):
    dbscan = DBSCAN(eps=param['eps'], min_samples=param['min_samples'])
    labels = dbscan.fit_predict(X_stand)
    noise_ratio = list(labels).count(-1) / len(labels)
    if noise_ratio < 0.2 and noise_ratio > best_score:
        best_score = noise_ratio
        best_params = param

if best_params is not None:
    print(best_params)
else:
    print("Параметри не знайдено, що відповідають заданим умовам.")

{'eps': 2.8, 'min_samples': 10}
```

Рисунок 3.13 — Підбір параметрів DBSCAN

Джерело: розроблено автором на основі [46]

Спочатку було визначено два параметри для кластеризації: `eps`, який визначає максимальну відстань між точками, що дозволяє їм належати до одного кластера, та `min_samples`, що вказує на мінімальну кількість точок, які повинні бути присутніми для формування кластера. Задається кілька можливих значень

для кожного з цих параметрів. Потім за допомогою `ParameterGrid` перебирались всі можливі комбінації цих параметрів. Для кожної з комбінацій створювалась нова модель DBSCAN, яка кластеризує дані. Після кластеризації обчислюється відсоток точок, що віднесені до шуму (мітка -1), і цей показник порівнюється з певним порогом: шукаються ті параметри, для яких кількість точок, що виявилися шумом, менше 20%.

Було визначено оптимальні параметри для алгоритму DBSCAN: значення `eps` становить 2.8, а `min_samples` — 10. Далі ці параметри будуть використані для виконання кластеризації за допомогою методу DBSCAN.

```
] dbscan = DBSCAN(eps=2.8, min_samples=10)

] labels = dbscan.fit_predict(X_stand)

] data['cluster2'] = labels
```

Рисунок 3.14 — Кластеризація DBSCAN

Джерело: розроблено автором на основі [46]

Після виконання кластеризації було виявлено 5 кластерів, проте один з них, кластер 0, виглядає як шум, оскільки в нього кластеризувалась велика кількість значень.

```
#перевіримо наповненість кластерів
data[['cluster2']].value_counts()
```

count	
cluster2	
0	1668
-1	400
1	142
2	21
3	9

Рисунок 3.15 — Отримані кластери DBSCAN

Джерело: розроблено автором на основі [46]

Більше 50% клієнтів визначається, як шум, що вже може свідчити про неякісну кластеризацію. Тож додатково варто перевірити якість кластеризації вищезазначеними критеріями оцінки:

```
silhouette_avg = silhouette_score(X_stand, labels)
print(f'Silhouette Score: {silhouette_avg}')

Silhouette Score: 0.04710655779144617
```

Рисунок 3.16 — Силуетний коефіцієнт для DBSCAN

Джерело: розроблено автором на основі [46]

У цьому випадку коефіцієнт 0.047 вказує на те, що кластери недостатньо чітко розділені, і є значні труднощі у правильному розподілі точок. Це може свідчити про необхідність коригування параметрів кластеризації або вибору іншого методу кластеризації для покращення результатів.

```
# Обчислення індексу Калінскі-Харабаса
calinski_score = calinski_harabasz_score(X_stand, labels)
print(f'Calinski-Harabasz Index: {calinski_score}')

Calinski-Harabasz Index: 114.42117679857988
```

Рисунок 3.17 — індекс Калінскі-Харабаса для DBSCAN

Джерело: розроблено автором на основі [46]

Значення 114.42 свідчить про наявність певного рівня відокремлення між кластерами, але не такого якісного, як при кластеризації методом к-середніх.

Доцільно застосувати ще один підхід для кластеризації, а саме Gaussian Mixture Model (GMM), оскільки цей метод дозволяє ефективно класифікувати дані, використовуючи ймовірнісні моделі, що можуть враховувати складні структури та варіації в даних, яких не завжди можна досягти за допомогою стандартних алгоритмів, таких як K-Means. Використання GMM дає змогу не лише визначити кластери, але й оцінити ймовірність належності кожного зразка

до того чи іншого кластеру, що є корисним для більш гнучкої та детальної класифікації.

```
from sklearn.decomposition import PCA
from sklearn.datasets import make_blobs
from matplotlib.patches import Ellipse
pca = PCA(n_components=2)
X_2d = pca.fit_transform(X_stand)

n_components = 7 # Очікувана кількість кластерів
gmm = GaussianMixture(n_components=n_components, random_state=42)
gmm.fit(X_2d)
labels = gmm.predict(X_2d)
```

Рисунок 3.18 — GMM

Джерело: розроблено автором на основі [46]

Код на рис 3.17 реалізує двоетапний процес аналізу даних, який включає зниження розмірності та кластеризацію. Спочатку застосовується метод Principal Component Analysis (PCA) для зниження розмірності даних до двох компонент, що дозволяє спростити дані та зберегти основну структуру інформації. Це робиться для полегшення подальшого аналізу та візуалізації даних, зберігаючи найбільш значущі характеристики. Метод `fit_transform` застосовується для трансформації стандартизованих даних в двовимірний простір.

Наступним етапом є кластеризація даних за допомогою Gaussian Mixture Model (GMM). Цей метод використовує ймовірнісну модель для класифікації даних, що дозволяє визначити ймовірність належності кожного зразка до певного кластера. У даному випадку, модель GMM передбачає 7 кластерів, і для кожного зразка передбачає, до якого з кластерів він належить, на основі ймовірнісних характеристик цих кластерів. Результат кластеризації дозволяє присвоїти кожному зразку мітку відповідного кластера, що допомагає у подальшому аналізі структури даних. Таким чином, код поєднує методи зниження розмірності і кластеризації для

виявлення груп в даних та зменшення їх складності, що сприяє більш ефективному аналізу й візуалізації.

Візуалізація отриманих кластерів виглядає наступним чином:

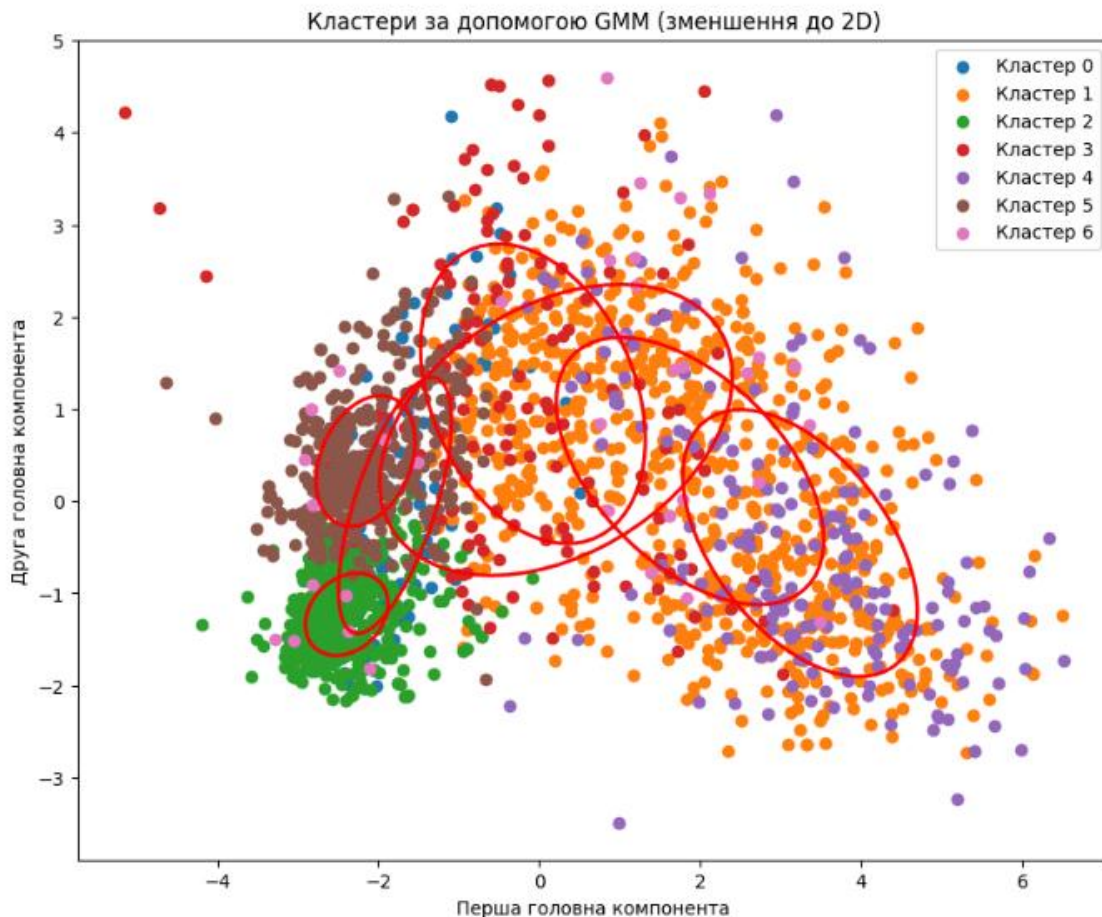


Рисунок 3.19 — Візуалізація отриманих кластерів *GMM*

Джерело: розроблено автором на основі [46]

На графіку, що відображає результати кластеризації методом Gaussian Mixture Model (*GMM*) з двовимірним зниженням вимірності, можна спостерігати різний рівень щільності даних у кожному кластері. Найбільша щільність спостережень фіксується в другому та п'ятому кластерах. Дані в цих кластерах концентруються в певних областях простору, що свідчить про високу схожість спостережень між собою в межах цих груп. Це дозволяє припустити, що ці кластери характеризуються сильними внутрішніми кореляціями і мають меншу варіативність за аналізованими ознаками.

У першому та четвертому кластерах дані розподілені більш розкидано, що вказує на вищу варіативність усередині цих груп. Такі кластери можуть містити спостереження з різними характеристиками, що призводить до менш вираженої щільності в певних частинах простору.

Натомість третій, шостий та нульовий кластери виявляються менш інформативними. Це можна пояснити низькою щільністю точок у цих кластерах, що свідчить про те, що дані в них є більш розрідженими та, ймовірно, мають меншу внутрішню структуру або слабше визначені ознаки. Такий розподіл може вказувати на відсутність чіткої кластерної структури в цих групах, що робить їх менш значущими з точки зору подальшого аналізу. Важливо зазначити, що такі кластери можуть бути також ознакою шумових або аномальних спостережень, які не відповідають типовим патернам у даних.

Додатково варто провести аналіз критеріїв оцінки кластеризації даним методом:

```
silhouette_avg = silhouette_score(X_stand, labels)
print(f'Silhouette Score: {silhouette_avg}')

Silhouette Score: 0.12131705230668827
```

Рисунок 3.20 — Силуетний коефіцієнт для DBSCAN

Джерело: розроблено автором на основі [46]

В даному випадку, значення 0.121 вказує на те, що кластеризація не є надто ефективною, оскільки значення Silhouette Score наближається до 0, що свідчить про наявність певного перекриття між кластерами або недостатньо чітко визначених меж між ними.

```
# Обчислення індексу Калінскі-Харабаса
calinski_score = calinski_harabasz_score(X_stand, labels)
print(f'Calinski-Harabasz Index: {calinski_score}')

Calinski-Harabasz Index: 244.00959616092828
```

Рисунок 3.21 — *індекс Калінські-Харабаса* для DBSCAN

Джерело: розроблено автором на основі [46]

Значення 244.01 є середнім показником для цього індексу. Це може означати, що кластери мають помірний рівень чіткості та відділення, однак для більш точного оцінювання необхідно порівняти його з результатами інших методів кластеризації або з іншими наборами даних.

Порівнюючи три проведені методи кластеризації — К-середніх (K-Means), DBSCAN, та Gaussian Mixture Model (GMM), можна зробити кілька висновків. Метод К-середніх показує найкращі результати за більшістю критеріїв: він має показник Inertia (24266.16) та високий Calinski-Harabasz Index (369.65), що свідчить про добре відокремлені та компактні кластери. Однак, за Silhouette Score (0.189), що вказує на чіткість кластерів, результат не є оптимальним і вказує на те, що кластери не є ідеально відокремленими.

Метод DBSCAN показує найгірший результат за Silhouette Score (0.047), що вказує на погану чіткість кластерів і наявність шуму, оскільки DBSCAN може позначати деякі точки як шум (не належать жодному кластеру). Його Calinski-Harabasz Index (114.42) також є значно нижчим, ніж у К-середніх, що підтверджує гіршу якість кластеризації.

GMM показує кращі результати за Silhouette Score (0.121) та Calinski-Harabasz Index (244.01) порівняно з DBSCAN, але поступається К-середнім, особливо за показником Calinski-Harabasz Index. Це свідчить про те, що кластери, створені за допомогою GMM, є менш чітко відокремленими та компактними, порівняно з тими, що були створені методом К-середніх.

Отже, метод К-середніх видає найкращі результати за більшістю показників, хоча і потребує уваги в контексті чіткості кластерів, оскільки Silhouette Score свідчить про певні проблеми з відокремленням кластерів.

3.3. Аналіз та інтерпретація отриманих сегментів

Враховуючи порівняння результатів кластеризації, виконаної трьома різними методами (К-середніх, DBSCAN та Gaussian Mixture Model), найбільш доцільним для подальшого аналізу є обрання кластерів, отриманих за допомогою методу К-середніх. Це обґрунтовується тим, що метод К-середніх показав найкращі результати за більшістю ключових критеріїв якості кластеризації, зокрема, найнижчий показник Inertia, що свідчить про добре стиснуті і компактні кластери. Крім того, Calinski-Harabasz Index для цього методу є значно вищим порівняно з іншими, що вказує на добре відокремлені групи даних і високу внутрішню однорідність кластерів. Хоча Silhouette Score для К-середніх також не є ідеальним, вказуючи на певні труднощі в чіткому відокремленні деяких кластерів, цей показник є вищим за аналогічні значення для методів DBSCAN та GMM, що вказує на кращу загальну якість кластеризації.

Для візуалізації кластерів було використано програмний код, який спочатку зберігає координати центрів кластерів, які є середніми значеннями всіх характеристик для кожної групи даних. Потім ці координати формуються у вигляді нової таблиці, де кожний рядок представляє центр одного кластера, а колонки відповідають вихідним ознакам даних. В результаті таблиця надає зрозумілу інтерпретацію того, якими характеристиками в середньому володіють об'єкти кожного кластера:

```
centers = kmeans.cluster_centers_

new_data = pd.DataFrame(centers, columns = ['Year_Birth', 'Income', 'Kidhome', 'Teenhome', 'Recency', 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumM',
new_data
```

Рисунок 3.22 — програмний код визначення центрів кластерів

Джерело: розроблено автором на основі [46]

Після застосування цього методу можна візуалізувати отримані дані:

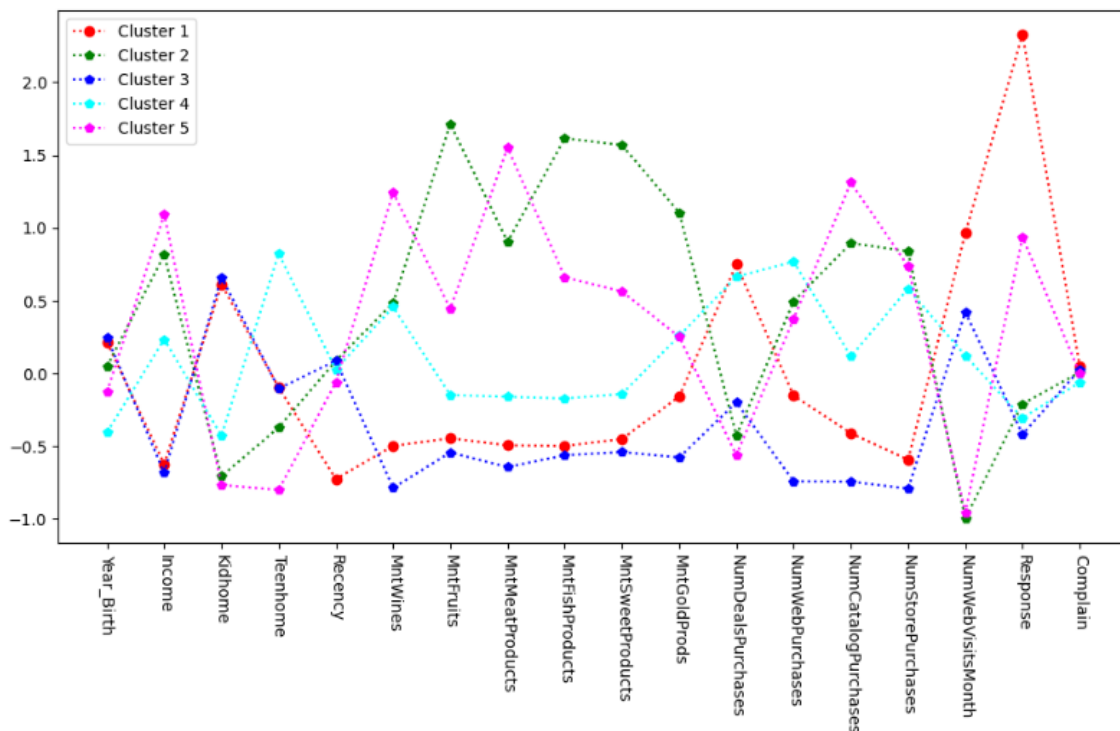


Рисунок 3.23 — візуалізація центрів кластерів

Джерело: розроблено автором на основі [46]

Перший кластер «Лояльні покупці з високим доходом»: Ця група складається з людей, які регулярно роблять значні покупки базових продуктів. Вони не схильні до імпульсивних дій, не шукають акцій та рідко замовляють товари онлайн. Їх характеризує лояльність до компанії, цінність якості та зручності, а також готовність платити більше.

Другий кластер «Імпульсивні покупці»: До цієї групи належать люди різного віку та доходу, які тяжіють до товарів не першої необхідності, таких як алкоголь та солодощі. Вони схильні до імпульсивних покупок, часто шукають акційні пропозиції та замовляють товари онлайн. Їх характеризує схильність до емоційних покупок, пошук вигідних пропозицій та цінність зручності онлайн-замовлень.

Третій кластер «Сім'ї з дітьми»: Цей кластер складається з дорослих, які мають дітей. Їх не характеризує лояльність до певного магазину, адже вони обирають найвигідніші пропозиції. Їхніми основними характеристиками є практичність, орієнтація на бюджет, пошук товарів для дітей та сім'ї, а також цінність зручності та доступності.

Четвертий кластер «Студенти»: Ця група складається з молодих людей з низьким доходом. Їх характеризує прагнення до економії, порівняння цін на веб-сайті, пошук найвигідніших пропозицій та віддавання переваги онлайн-замовленням. Їм важливі вигода та економія, вони шукають інформацію в інтернеті та цінують зручність онлайн-покупок.

П'ятий кластер «Незадоволені покупці»: Ця група складається з людей, які залишали скарги на досвід купівлі. Їх характеризує негативний досвід, нелояльність до компанії та потреба у покращенні обслуговування [50].

3.4. Рекомендації щодо персоналізованих пропозицій для отриманих сегментів

Для кожного з визначених кластерів клієнтів можна запропонувати персоналізовані рішення, які враховують їхні потреби, поведінкові особливості та цінності.

Лояльні покупці з високим доходом, як правило, мають високу вимогливість до якості продукції, зручності покупок і персоналізованого обслуговування. Для цього сегмента клієнтів важливими факторами є не лише функціональні характеристики товарів, але й емоційний досвід взаємодії з

брендом. Тому для збереження та розширення цієї аудиторії бізнесам варто впроваджувати розвинуті програми лояльності, орієнтовані на преміальні рівні.

Програми лояльності з преміальними рівнями дають можливість запропонувати ексклюзивні послуги та пропозиції, які задовольняють високі вимоги таких покупців. Преміальний підхід має на увазі надання персоналізованих знижок, пропозицій, а також доступу до обмежених або ексклюзивних товарів, які не доступні широкій аудиторії. Важливо, щоб ці пропозиції були не лише економічно вигідними, а й формували враження індивідуального підходу до кожного покупця. Наприклад, періодичні персоналізовані знижки, спеціальні подарунки до свят чи ексклюзивний доступ до нових колекцій.

Індивідуалізація взаємодії також може проявлятися через персоналізовані комунікації, що ґрунтуються на даних про покупців. Інформація, зібрана під час попередніх покупок або через анкетування, дозволяє надавати рекомендації, які відповідають інтересам та перевагам кожного покупця, а також запобігають надмірному навантаженню клієнта неприємними чи непотрібними пропозиціями.

Для цього сегмента також важливим є забезпечення зручних сервісів. Сервіси, які забезпечують максимальний комфорт для покупців, можуть включати можливість попереднього замовлення товарів, доставку у зручний для клієнта час чи місце, а також можливість безкоштовної доставки чи повернення товарів. Такі сервіси демонструють не тільки увагу до потреб клієнта, але й підвищують рівень задоволеності та лояльності.

З метою підтримки взаємодії з лояльними покупцями варто використовувати ненав'язливі форми комунікації через email-розсилки або SMS-повідомлення. Ці канали дозволяють інформувати клієнтів про новинки, акції чи спеціальні пропозиції, не створюючи враження надмірного навантаження рекламою. Важливо, щоб ці комунікації були добре таргетовані та мали чітко визначену мету: надання корисної інформації та створення додаткової цінності для клієнта.

У результаті впровадження таких стратегій лояльності підприємства можуть не лише зберегти лояльність заможних клієнтів, але й значно підвищити їхню залученість та тривалість взаємодії з брендом.

Імпульсивні покупці, як правило, характеризуються схильністю до швидких, емоційно мотивованих покупок, часто спонтанних, під впливом вигідних пропозицій, знижок або акцій. Залучення цієї групи клієнтів потребує адаптації маркетингових стратегій до їхнього стилю поведінки, що орієнтований на отримання швидких вигод за мінімальних зусиль. Ось кілька ключових підходів для ефективної роботи з імпульсивними покупцями:

Короткострокові акції та обмежені пропозиції. Імпульсивні покупці швидше приймають рішення про покупку, якщо відчують, що пропозиція обмежена за часом чи кількістю товарів. Стратегії, як-от flash-sale (миттєві акції) або знижки, що діють лише кілька годин, створюють відчуття терміновості, стимулюючи покупця до швидкої покупки. Чітка комунікація про обмеженість пропозиції (наприклад, через таймер на сайті чи в мобільному додатку) може бути ефективним стимулом для імпульсивних покупок.

Крос-продажі та up-selling. Техніки крос-продажу (коли пропонуються супутні товари або аксесуари) і up-selling (коли покупцям пропонується більш дорога чи вдосконалена версія товару) можуть бути дуже ефективними для імпульсивних покупців. Це особливо важливо в онлайн-середовищі, де покупці легко можуть додавати товари до кошика, не замислюючись довго. Пропозиція таких товарів безпосередньо під час оформлення покупки збільшує ймовірність додаткових імпульсних покупок.

Push-сповіщення про вигідні пропозиції. Важливо підтримувати постійний контакт з імпульсивними покупцями через push-сповіщення в мобільних додатках або через браузер. Вони повинні інформувати про актуальні акції, знижки, розпродажі, а також про обмежені за часом або кількістю вигідні пропозиції.

Проте важливо, щоб ці повідомлення були ненав'язливими та релевантними для інтересів покупця, щоб не створювати негативне враження від нав'язливості.

Оптимізація мобільних додатків для швидких покупок. Мобільні додатки для імпульсивних покупців повинні бути максимально зручними, з простими функціями, що дозволяють здійснювати покупку за кілька кліків. Це включає можливість зберігати товари в кошику, швидкий доступ до акційних товарів, а також можливість оплатити покупку одним натисканням через мобільний гаманець чи інші швидкі методи оплати. Функції персоналізованих рекомендацій, що відображають товари на основі попередніх покупок або переглядів, можуть ще більше підвищити ймовірність здійснення імпульсних покупок.

Персоналізація досвіду покупок. Імпульсивні покупці дуже швидко реагують на персоналізовані пропозиції, які відповідають їхнім інтересам і потребам. Використання даних про попередні покупки, переглянуті товари та поведінкові патерни дозволяє створювати таргетовані рекомендації, які мотивують до швидкої покупки. Наприклад, якщо покупець нещодавно переглядав певний товар, то сповіщення про знижку на нього або супутні товари може стати спонукальним фактором для імпульсивної покупки.

Застосування таких стратегій дозволить не лише задовольнити потреби імпульсивних покупців у зручних, швидких та вигідних покупках, але й сприятиме підвищенню їхньої активності та лояльності до бренду. Важливо, щоб усі ці ініціативи були добре синхронізовані, що дозволить створити зручний та привабливий процес покупок, який відповідатиме специфічним потребам цієї групи споживачів.

Сім'ї з дітьми, як правило, є практичними та раціональними споживачами, які надають перевагу доступності та бюджетності. При покупках вони орієнтуються на функціональність, якість і вигідність пропозицій. Для цього сегмента клієнтів важливо забезпечити умови, які дозволяють зекономити час і гроші, а також відповідають їхнім повсякденним потребам. Ось кілька ключових стратегій для ефективної роботи з сім'ями з дітьми:

1. Вигідні набори товарів для сім'ї. Для цієї аудиторії дуже привабливими є пакетні пропозиції, що включають набори товарів для всієї родини — продукти харчування, побутову хімію, засоби гігієни тощо. Наприклад, комплекти для догляду за дітьми або для дому, які включають в себе кілька необхідних товарів за зниженою ціною, допомагають сім'ям заощаджувати на повсякденних покупках і забезпечують зручність у виборі. Пропонуючи спеціальні набори для різних потреб (для дітей різного віку, для родини з малюками або з підлітками), магазини можуть значно полегшити процес покупки для батьків.

2. Програми накопичувальних знижок. Сім'ї з дітьми зазвичай здійснюють покупки регулярно і на великі суми, що робить їх ідеальними кандидатами для програм лояльності з накопичувальними знижками. Такі програми дозволяють отримувати знижки на подальші покупки залежно від суми витрачених коштів або кількості покупок, що стимулює їх здійснювати покупки частіше. Крім того, можна впровадити персоналізовані бонуси для родинних категорій, наприклад, додаткові знижки на товари для дітей або продукти, які часто купуються сім'ями.

3. Передплата на регулярну доставку базових продуктів. Сім'ї, особливо з маленькими дітьми, часто зіштовхуються з необхідністю регулярних покупок базових продуктів, таких як дитячі харчування, підгузки, засоби гігієни та побутова хімія. Для спрощення цього процесу варто впровадити сервіси передплати, що дозволяють автоматично отримувати потрібні товари з регулярною доставкою. Такий сервіс дозволяє зекономити час, уникнути частих походів до магазину та гарантує постійну наявність необхідних товарів вдома.

4. Окремі категорії на сайті для товарів для сімей. Для зручності покупців варто створити спеціальні категорії на вебсайті, які об'єднують всі товари, що можуть бути корисні сім'ям. Це можуть бути підкатегорії для товарів для новонароджених, для дітей дошкільного та шкільного віку, для організації домашнього побуту, а також для забезпечення здоров'я і гігієни родини. Така

структура дозволить батькам швидко знайти потрібні товари без зайвих витрат часу на пошук.

6. Зручність та доступність онлайн-покупок. Оскільки сім'ї з дітьми часто мають обмежений час для походів у магазини, для них важливо забезпечити зручність онлайн-покупок. Це включає можливість легко оформити замовлення через мобільні додатки або сайт, безпечні методи оплати та швидку доставку товарів на зручний час. Крім того, варто впровадити опції зручного повернення товарів, що дозволить знизити ризик і зробить процес покупок більш комфортним.

Студенти прагнуть економити та цінують зручність онлайн-покупок. Для них корисними будуть спеціальні знижки для студентів, підписки на регулярні покупки зі знижками, а також програми гейміфікації, які дозволяють накопичувати бонуси за активність. Просування товарів через соціальні мережі допоможе залучити цю аудиторію.

Незадоволені покупці потребують швидкого та ефективного вирішення своїх проблем, щоб відновити довіру до бренду. Важливо створити систему, яка дозволяє оперативно обробляти скарги та забезпечує відкриту комунікацію з клієнтами. Ключовим аспектом є впровадження ефективної системи роботи зі скаргами, що забезпечує доступність і зручність подачі скарг через різні канали комунікації. Швидка реакція на звернення та чітка інформованість про етапи вирішення ситуації допомагають зняти напругу і продемонструвати клієнтам, що їхнє незадоволення сприймається серйозно.

Для відновлення лояльності важливо також пропонувати компенсації та бонуси. Це може бути часткове повернення коштів, безкоштовний товар чи спеціальні знижки на майбутні покупки. Така практика не тільки дозволяє компенсувати шкоду, але й дає покупцю відчуття, що компанія цінує його довіру.

Реалізація таких персоналізованих підходів сприятиме задоволенню потреб клієнтів, підвищенню їхньої лояльності та зростанню ефективності бізнесу.

ВИСНОВКИ

У межах цього дослідження було проаналізовано сучасні підходи до сегментації клієнтів і оцінено їхній вплив на ефективність програм лояльності в умовах динамічного ринкового середовища. За результатами проведеного аналізу встановлено, що універсальні програми лояльності, які раніше мали значну популярність, вже не можуть забезпечити достатню результативність через зростаючу конкуренцію та різноманітність потреб клієнтів. Сучасні методи сегментації, зокрема алгоритми кластеризації, дозволяють створювати персоналізовані програми лояльності, які враховують унікальні характеристики та специфіку поведінки різних клієнтських сегментів, що, у свою чергу, підвищує їхню ефективність.

В ході дослідження були досягнуті основні цілі, зокрема розроблено нові підходи до сегментації клієнтів, що ґрунтуються на використанні передових методів кластеризації, таких як к-середніх, DBSCAN та GMM. Застосування цих методів дозволило на основі реальних даних виділити п'ять основних сегментів клієнтів, кожен з яких має свої специфічні потреби та вимоги до програм лояльності. Для кожного з цих сегментів були розроблені індивідуальні рекомендації щодо удосконалення програм лояльності:

Лояльні покупці з високим доходом – цей сегмент потребує преміальних пропозицій, індивідуального сервісу та комфортних умов обслуговування, оскільки вони шукають винятковий досвід покупок.

Імпульсивні покупці, які реагують на емоційні стимули, потребують активного використання акційних кампаній, персоналізованих рекомендацій та зручних цифрових інструментів, що сприяють швидкому прийняттю рішень.

Сім'ї з дітьми орієнтовані на вигоду та зручність, тому для них важливо пропонувати спеціальні набори товарів, знижки на сімейні продукти та програми передплати.

Студенти, які цінують економію та активно користуються онлайн-платформами, мають потребу в спеціальних знижках, бонусних програмах та інтерактивних кампаніях у соціальних мережах.

Незадоволені покупці вимагають особливого підходу, що включає не лише ефективну роботу з їхніми скаргами та компенсаціями, але й заходи для відновлення довіри та лояльності до бренду.

Результати дослідження демонструють високу ефективність застосування сучасних методів кластеризації для детальної сегментації клієнтської бази. Завдяки цьому підприємства можуть краще розуміти своїх споживачів, оптимізувати ресурси, що спрямовуються на маркетингові активності, та створювати релевантні пропозиції, що відповідають індивідуальним потребам різних клієнтських сегментів.

Практичні результати дослідження мають значний потенціал для впровадження в бізнес-процеси компаній, оскільки дозволяють підвищити ефективність маркетингових стратегій і, як наслідок, рентабельність інвестицій у розвиток програм лояльності. Ці підходи є особливо важливими для українського ринку, де конкуренція на постійно змінюваному ринку вимагає швидкої адаптації та чіткої орієнтації на потреби клієнтів.

Таким чином, сегментація клієнтів і персоналізація програм лояльності виступають як ключові інструменти сучасного маркетингу, які сприяють зміцненню конкурентних позицій підприємств, формуванню довгострокових взаємин із клієнтами та підвищенню їхньої задоволеності від взаємодії з брендом. Використання розроблених підходів дозволить бізнесам не лише ефективніше взаємодіяти зі своєю клієнтською базою, але й досягати стабільного зростання та розвитку в умовах високої конкуренції.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Armstrong G., Kotler P. Marketing: An Introduction (5th Edition). 5th ed. Prentice Hall, 1999. 205 p.
2. Jílková P. Customer Behaviour and B2C Client Segmentation in Data-Driven Society. *International Advances in Economic Research*. 2020. Vol. 26, no. 3. P. 325–326. URL: <https://doi.org/10.1007/s11294-020-09799-9> (date of access: 24.11.2024).
3. Марчук І., Орловська С., Жарська І. Управління лояльністю на ринку B2B. *Економіка та суспільство*. 2023. № 56. URL: <https://doi.org/10.32782/2524-0072/2023-56-134> (дата звернення: 24.11.2024).
4. Dekay F., Toh R. S., Raven P. Loyalty Programs. *Cornell Hospitality Quarterly*. 2009. Vol. 50, no. 3. P. 371–382. URL: <https://doi.org/10.1177/1938965509338780> (date of access: 24.11.2024).
5. Rudzewicz A. Loyalty programs and consumer loyalty. *Sj-economics scientific journal*. 2018. Vol. 28, no. 1. P. 127–135. URL: <https://doi.org/10.58246/sjeconomics.v28i1.145> (date of access: 24.11.2024).
6. Oliver R. L. Whence Consumer Loyalty?. *Journal of Marketing*. 1999. Vol. 63. P. 33. URL: <https://doi.org/10.2307/1252099> (date of access: 24.11.2024).
7. Papandrea D. With Inflation Soaring, Half of Americans Think Loyalty Programs Are More Important Than Eve. <https://www.lendingtree.com>. URL: [https://www.lendingtree.com/credit-cards/study/loyalty-programs/#:~:text=Six-figure%20earners%20\(57%\),at%20least%20one%20loyalty%20program](https://www.lendingtree.com/credit-cards/study/loyalty-programs/#:~:text=Six-figure%20earners%20(57%),at%20least%20one%20loyalty%20program) (date of access: 24.11.2024).
8. Boudet J., Huang J., Ryter von Difloe. Coping with the big switch: How paid loyalty programs can help bring consumers back to your brand. McKinsey & Company. URL: <https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/coping-with-the-big-switch-how-paid-loyalty-programs-can-help-bring-consumers-back-to-your-brand> (date of access: 24.11.2024).

9. 7 Types of Loyalty Programs: Which is Right for Your Brand? [With Examples]. ebbo. URL: <https://www.ebbo.com/insights/blog/7-types-of-loyalty-programs-which-is-right-for-your-brand/> (date of access: 24.11.2024).
10. Панаско О., Микитенко Н. Програми лояльності покупців на підприємстві торгівлі: теоретичний та практичний ракурси. *Молодий вчений*. 2022. № 1 (101). С. 233–241. URL: <https://doi.org/10.32839/2304-5809/2022-1-101-48> (дата звернення: 24.11.2024).
11. Market Segmentation by Malcolm McDonald | 9780080479194. Perlego | An online subscription for all your academic books. URL: https://www.perlego.com/book/4477695/market-segmentation-how-to-do-it-how-to-profit-from-it-pdf?utm_source=google&utm_medium=cpc&utm_campaign=20933451054&utm_adgroup=162926082892&utm_gad_source=1&utm_gclid=Cj0KCQiAuou6BhDhARIsAIfgrn7UN_mibXgjJsOoYUXvm1PXPpqbhVok8DD3o9USuq0Bt55rRsMJUp0aAkQuEALw_wcB (date of access: 24.11.2024).
12. Shah I. Customer Segmentation. *International Journal for Research in Applied Science and Engineering Technology*. 2024. Vol. 12, no. 1. P. 1586–1591. URL: <https://doi.org/10.22214/ijraset.2024.58144> (date of access: 24.11.2024).
13. Cooil B., Aksoy L., Keiningham T. L. Approaches to Customer Segmentation. *Journal of Relationship Marketing*. 2008. Vol. 6, no. 3-4. P. 9–39. URL: https://doi.org/10.1300/j366v06n03_02 (date of access: 24.11.2024).
14. Customer Segmentation Models: Types, Methods, and Techniques. Contentsquare: all-in-one experience intelligence platform. URL: <https://contentsquare.com/guides/customer-segmentation/models/> (date of access: 24.11.2024).
15. Gore P. A. Cluster Analysis. *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. 2000. P. 297–321. URL: <https://doi.org/10.1016/b978-012691360-6/50012-4> (date of access: 30.11.2024).
16. What Is Regression Analysis in Statistics?. Julius AI | Your AI Data Analyst. URL: <https://julius.ai/articles/regression-analysis-in-statistics> (date of access: 24.11.2024).

17. IBM. What is a Decision Tree? | IBM. IBM - United States. URL: <https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes> (date of access: 24.11.2024).
18. Brillinger D. R. Time Series: General. *International Encyclopedia of the Social & Behavioral Sciences*. 2015. P. 341–346. URL: <https://doi.org/10.1016/b978-0-08-097086-8.42084-2> (date of access: 30.11.2024).
19. ElKalaawy N., Wassal A. Methodologies for the modeling and simulation of biochemical networks, illustrated for signal transduction pathways: A primer. *Biosystems*. 2015. Vol. 129. P. 1–18. URL: <https://doi.org/10.1016/j.biosystems.2015.01.008> (date of access: 24.11.2024).
20. Bwalya K. J., Chris R., Mandla N. Convergence of Wireless Technologies in Consolidating E-Government Applications in Sub-Saharan Africa. *International Journal of ICT Research and Development in Africa*. 2010. Vol. 1, no. 4. P. 15–30. URL: <https://doi.org/10.4018/978-1-60960-042-6.ch004> (date of access: 24.11.2024).
21. 2nd M. S., 1st S. V., Mahalakshmi 3. Machine Learning and Deep Learning. Alexis Press LLC, 2022. 30 p.
22. Data Mining and Knowledge Discovery Handbook / ed. by O. Maimon, L. Rokach. New York : Springer-Verlag, 2005. pp.165-192 URL: <https://doi.org/10.1007/b107408> (date of access: 26.11.2024).
23. Linear Models for Classification, Applied Machine Learning in Python. *Andreas C. Müller - Machine Learning Scientist*. URL: <https://amueller.github.io/aml/02-supervised-learning/06-linear-models-classification.html> (date of access: 26.11.2024).
24. Bratchell N. Cluster analysis. *Chemometrics and Intelligent Laboratory Systems*. 2010. Vol. 6, no. 2. P. 105–125. URL: [https://doi.org/10.1016/0169-7439\(87\)80054-0](https://doi.org/10.1016/0169-7439(87)80054-0) (date of access: 26.11.2024).

25. Tagliaro C., Orel M., Hua Y. Methodological Approaches for Workplace Research and Management. London : Routledge, 2023. URL: <https://doi.org/10.1201/9781003289845> (date of access: 26.11.2024).
26. Li Y., Wu H. A Clustering Method Based on K-Means Algorithm. *Physics Procedia*. 2012. Vol. 25. P. 1104–1109. URL: <https://doi.org/10.1016/j.phpro.2012.03.206> (date of access: 26.11.2024).
27. K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data / A. M. Ikotun et al. Information Sciences. 2022. URL: <https://doi.org/10.1016/j.ins.2022.11.139> (date of access: 26.11.2024).
28. Sinaga K. P., Yang M.-S. Unsupervised K-Means Clustering Algorithm. IEEE Access. 2020. Vol. 8. P. 80716–80727. URL: <https://doi.org/10.1109/access.2020.2988796> (date of access: 26.11.2024).
29. DBSCAN Clustering in ML | Density based clustering - GeeksforGeeks. *GeeksforGeeks*. URL: <https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/> (date of access: 26.11.2024).
30. Satpati S. Clustering by DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Clearly... Medium. URL: <https://medium.com/@satpatishrimanta/clustering-by-dbscan-density-based-spatial-clustering-of-applications-with-noise-clearly-f93c5c72f706> (date of access: 26.11.2024).
31. DBSCAN Algorithm for Document Clustering / R. G. Crețulescu et al. *International Journal of Advanced Statistics and IT&C for Economics and Life Sciences*. 2019. Vol. 9, no. 1. P. 58–66. URL: <https://doi.org/10.2478/ijasitels-2019-0007> (date of access: 26.11.2024).
32. Efendi R., Wahyono T., Widiyarsi I. R. DBSCAN SMOTE LSTM: Effective Strategies for Distributed Denial of Service Detection in Imbalanced Network Environments. *Big Data and Cognitive Computing*. 2024. Vol. 8, no. 9. P. 118. URL: <https://doi.org/10.3390/bdcc8090118> (date of access: 26.11.2024).
33. How to Master the Popular DBSCAN Clustering Algorithm for Machine Learning?. *Analytics Vidhya*.

- URL: <https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/> (date of access: 26.11.2024).
34. Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987. Vol. 20. P. 53–65. URL: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7) (date of access: 26.11.2024).
 35. Maximum Gaussian Mixture Model for Classification / J. Zhang et al. 2016 8th International Conference on Information Technology in Medicine and Education (ITME), Fuzhou, 23–25 December 2016. 2016. URL: <https://doi.org/10.1109/itme.2016.0139> (date of access: 26.11.2024).
 36. Mishra D., Pattipati K. R., Bollas G. M. Gaussian mixture model for tool condition monitoring. *Journal of Manufacturing Processes*. 2024. Vol. 131. P. 1001–1013. URL: <https://doi.org/10.1016/j.jmapro.2024.09.038> (date of access: 26.11.2024)
 37. Gaussian mixture models. scikit-learn. URL: <https://scikit-learn.org/1.5/modules/mixture.html> (date of access: 26.11.2024).
 38. Олійник С. В., Бойко Б. М. Порівняння методів кластеризації для створення цільових груп клієнтів у наборі даних. *Scientific Bulletin of UNFU*. 2023. Т. 33, № 5. С. 77–83. URL: <https://doi.org/10.36930/40330510> (дата звернення: 26.11.2024).
 39. What are some of the pros and cons of GMMs?. AIML.com. URL: <https://aiml.com/what-are-some-of-the-pros-and-cons-of-gmms/> (date of access: 26.11.2024).
 40. K-Means Clustering: 7 Pros and Cons Uncovered. Datarundown. URL: <https://datarundown.com/k-means-clustering-pros-cons/#:~:text=Its%20benefits%20include%20scalability,%20simplicity,boundaries,%20and%20sensitivity%20to%20outliers.> (date of access: 26.11.2024).
 41. Gomez J., León E., Nasraoui O. Minimum Cluster Size Estimation and Cluster Refinement for the Randomized Gravitational Clustering Algorithm. *Lecture Notes in Computer Science*. Berlin, Heidelberg, 2012. P. 51–60. URL: https://doi.org/10.1007/978-3-642-34654-5_6 (date of access: 28.11.2024).

42. Toward automated classification of pathological transcranial Doppler waveform morphology via spectral clustering / S. G. Thorpe et al. PLOS ONE. 2020. Vol. 15, no. 2. P. e0228642. URL: <https://doi.org/10.1371/journal.pone.0228642> (date of access: 28.11.2024).
43. Warrens M. J., van der Hoef H. Understanding the Rand Index. Studies in Classification, Data Analysis, and Knowledge Organization. Singapore, 2020. P. 301–313. URL: https://doi.org/10.1007/978-981-15-3311-2_24 (date of access: 28.11.2024).
44. Yang Y. Temporal Data Clustering. Temporal Data Mining Via Unsupervised Ensemble Learning. 2017. P. 19–34. URL: <https://doi.org/10.1016/b978-0-12-811654-8.00003-8> (date of access: 28.11.2024).
45. Courses Landing. URL:
https://course.ccs.neu.edu/cs6140sp15/7_locality_cluster/Assignment-6/NMI.pdf
 (дата звернення: 28.11.2024).
46. Superstore Marketing Campaign Dataset. Kaggle: Your Machine Learning and Data Science Community. URL: <https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset?resource=download> (date of access: 28.11.2024).
47. Practical Guide on Data Preprocessing in Python. Kaggle: Your Machine Learning and Data Science Community. URL:
<https://www.kaggle.com/code/ajay1216/practical-guide-on-data-preprocessing-in-python> (date of access: 28.11.2024).
48. Gal M., Rubinfeld D. L. Data Standardization. SSRN Electronic Journal. 2018. URL: <https://doi.org/10.2139/ssrn.3326377> (date of access: 28.11.2024).
49. Determining the Quality of a Dataset in Clustering Terms / A. Rachwał et al. Applied Sciences. 2023. Vol. 13, no. 5. P. 2942. URL: <https://doi.org/10.3390/app13052942> (date of access: 28.11.2024).
50. Szczepańska K., Gawron P. Loyalty Programs Effectiveness. Foundations of Management. 2011. Vol. 3, no. 2. P. 89–102. URL: <https://doi.org/10.2478/v10238-012-0044-5> (date of access: 28.11.2024).