Київський національний університет імені Тараса Шевченка

Економічний факультет

Кафедра економічної кібернетики

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

«Сегментація та кластеризація користувачів у платній рекламі: застосування методів data science**»**

студента 2 курсу магістратури
спеціальності 051 «Економіка»
ОНП «Економічна кібернетика»
денної форми навчання
Петренко Кароліни Володимирівни

Науковий керівник:
кандидат фізико-математичних наук,
доцент
Кравець Тетяна Вікторівна

Засвідчую, що у цій дипломній роботі
немає запозичень із праць інших авторів
без відповідних посилань

Студентка_____
(підпис)

Роботу допущено до захисту перед ЕК
рішенням кафедри економічної кібернетики
від 07.05.2025 р. протокол №12
Завідувач кафедри:
доктор економічних наук, професор
Ляшенко Олена Ігорівна

_____
(підпис)

КИЇВ – 2025

РЕФЕРАТ

Кваліфікаційна робота магістра містить: 3 розділи, 50 сторінок, 1 таблицю, 18 рисунків, 32 джерела, 1 додаток

Ключові слова: кластеризація, сегментація, реклама, цифрові платформи, персоналізація контенту, аналіз аудиторії.

Об'єкт дослідження: : Патерни взаємодії та споживчої активності користувачів у середовищі цифрової реклами.

Мета дослідження: кластеризація користувачів за допомогою різних методів, виявлення закономірностей у кластерах та використання результатів для персоналізованих рекламних підходів

Методи дослідження: кластерний аналіз, K-means, DBSCAN, ієрархічна кластеризація, індекс Девіса-Болдіна, Метод головних компонент (PCA)

Наукова новизна, теоретична значимість: комплексне застосування та порівняння трьох різних методів кластеризації для сегментації користувачів цифрової платформи з урахуванням поведінкових і демографічних ознак, що дозволило поглибити уявлення про ефективність алгоритмів у прикладних умовах.

Практична цінність: розробка інструментів автоматизованої сегментації користувачів, результати якої можуть бути безпосередньо використані для персоналізації контенту, підвищення залученості та ефективності маркетингових стратегій цифрової платформи.

RESUME

Taras Shevchenko National University of Kyiv,
Faculty of Economics, Faculty of Economics

Key words: clustering, segmentation, advertising, digital platforms, content personalization, audience analysis.

The graduation research of student Karolina Petrenko deals with audience data analysis from digital platforms using clustering and segmentation to improve advertising and content personalization

The work is interesting as it explores effective methods for audience data analysis and user segmentation on digital platforms, contributing to enhanced content personalization and targeted advertising strategies.
Pages 50, tables 1, bibliog. 32, append 1.

ЗМІСТ

Вступ

У сучасних умовах стрімкого розвитку цифрових технологій та зростання кількості онлайн-сервісів, проблема ефективного аналізу даних про користувачів набуває особливої актуальності. Цифрові платформи, незалежно від сфери їхньої діяльності, щодня генерують величезні обсяги інформації, які можуть стати джерелом цінних знань для прийняття стратегічних рішень. Зокрема, зростає потреба у глибокому розумінні структури користувацької аудиторії, її поведінкових моделей, вподобань і взаємодії з платформою. У цьому контексті на перший план виходять методи автоматизованої сегментації — інструменти, що дозволяють систематизувати великі масиви даних, виділяючи в них однорідні групи споживачів без попередньо визначених категорій.

Кластерний аналіз як одна з ключових технологій машинного навчання без учителя демонструє високу ефективність у задачах виявлення прихованої структури в даних. Його використання відкриває нові можливості для аналізу користувацьких баз: від персоналізованих маркетингових стратегій до покращення користувацького досвіду шляхом адаптації сервісів під потреби конкретних сегментів. Застосування кластеризації у сфері аналітики даних є актуальним не лише з точки зору дослідницького інтересу, а й у контексті реальних прикладних потреб цифрових компаній.

Особливістю сучасного етапу розвитку Data Science є те, що ефективність кластеризації залежить не лише від вибору алгоритму, а й від якісної підготовки даних, коректного відбору ознак, обґрунтованого зниження розмірності та налаштування параметрів моделей. У межах цієї дипломної роботи було реалізовано підхід, що поєднує як методологічні основи машинного навчання, так і прикладну аналітику на основі реалістичного набору даних користувачів. Такий підхід дозволяє не лише продемонструвати алгоритмічну спроможність окремих методів, а й оцінити їхню ефективність у реальних сценаріях застосування.

Метою роботи є кластеризація користувачів за допомогою різних методів, виявлення закономірностей у кластерах та використання результатів для персоналізованих рекламних підходів.

Досягнення мети дослідження потребує виконання певного набору **завдань**, а саме:

1. Характеристика сучасного стану та основних трендів цифрового маркетингу;
2. Теоретико-методологічний опис та порівняння методів кластеризації та їх оцінки;
3. Дослідження закономірностей та структури вхідних даних;
4. Поділ даних на кластери та аналіз отриманих результатів;
5. Надання рекомендацій щодо застосування отриманих результатів на практиці.

Об’єкт дослідження — патерни взаємодії та споживчої активності користувачів у середовищі цифрової реклами.

Предмет дослідження — виявлення спільних характеристик та кластеризація користувачів для персоналізації реклами.

Методи дослідження включають кластерний аналіз із використанням алгоритмів K-means, DBSCAN та метод ієрархічної кластеризації, метод головних компонент для зниження розмірності даних, а також індекс Девіса-Болдіна для оцінки якості кластеризації.

Наукова новизна роботи полягає у комплексному застосуванні та порівнянні трьох різних методів кластеризації для сегментації користувачів цифрової платформи з урахуванням поведінкових і демографічних ознак, що дозволило поглибити уявлення про ефективність алгоритмів у прикладних умовах.

Практична цінність дослідження проявляється у розробці інструментів автоматизованої сегментації користувачів, результати якої можуть бути безпосередньо використані для персоналізації контенту, підвищення залученості та ефективності маркетингових стратегій цифрової платформи.

Розділ 1. Теоретичні основи аналізу аудиторії цифрових платформ

1.1. Особливості залучення користувачів через цифрові платформи

У сучасному інформаційному просторі цифрові платформи стали основним інструментом комунікації між брендами та користувачами. Їх популярність зумовлена широким охопленням, зручністю у використанні та можливістю персоналізованої взаємодії. Залучення користувачів через такі платформи базується на аналізі їхньої поведінки, потреб та інтересів, що дозволяє компаніям формувати більш ефективні стратегії комунікації.

Традиційні методи маркетингу, такі як друкована реклама, телебачення та радіо, втрачають свою ефективність, оскільки споживачі дедалі більше часу проводять у цифровому середовищі. [1]. Традиційні канали вже не забезпечують достатнього охоплення аудиторії. Натомість цифрові платформи — зокрема соціальні мережі, блоги, відеоблоги, пошукові системи та мобільні застосунки — відкривають нові можливості для побудови ефективної комунікації з клієнтами в режимі реального часу. Вони дозволяють застосовувати більш індивідуалізований підхід та формувати унікальний досвід для кожного користувача.

Цифровий маркетинг, як напрямок, почав активно розвиватися з середини 1990-х років, коли інтернет став невід'ємною частиною повсякденного життя. Поява перших вебсайтів, таких як Yahoo в 1994 році, дала бізнесам можливість охоплювати величезну аудиторію через нові онлайн-канали. Спочатку маркетинг в інтернеті обмежувався простими банерами та розміщенням реклами на популярних сайтах, але вже на початку 2000-х з'явились нові можливості для взаємодії з користувачами. Розвиток пошукових систем, таких як Google, сприяв виникненню SEO-стратегій, а згодом і контекстної реклами. Цей період можна вважати етапом "цифрової революції", коли бренди почали активно використовувати онлайн-ресурси для досягнення цільової аудиторії. [2]

З часом розвиток мобільних технологій, соціальних мереж та застосунків ще більше змінив ландшафт цифрового маркетингу, відкриваючи нові канали для взаємодії та комунікації з користувачами. Цифровий маркетинг сьогодні

використовує широкий спектр платформ, кожна з яких має свої унікальні можливості для залучення та взаємодії з користувачами. Основні платформи можна умовно поділити на чотири категорії: соціальні мережі, медійна реклама, пошукова реклама та нативна реклама.

Соціальні мережі надають великі можливості для таргетингу та візуального контенту. Facebook та Instagram (платформи Meta) дозволяють запускати персоналізовані кампанії з високим рівнем деталізації — за віком, статтю, інтересами, поведінкою та геолокацією. Їхні переваги включають точний таргетинг, інтерактивні формати та гнучкість у налаштуванні. Недоліком є висока конкуренція, що може підвищувати вартість показів. TikTok став улюбленою платформою молоді — тут важлива віральність, креатив та нативність відеоконтенту. Проте, обмеженість в точному таргетингу та специфічний стиль платформи можуть бути викликом для деяких брендів. Також варто згадати LinkedIn — платформу для B2B-сегменту з можливістю таргетування за професіями, посадами та галузями.

Медійна реклама орієнтується на охоплення через банери, відео та інтерактивні формати. YouTube є провідною платформою для відеореклами, що забезпечує глибоке занурення в контент і сильну аналітику, однак вимагає більше ресурсів на створення відео. Google Display Network (GDN) дозволяє розміщувати банери на тисячах сайтів, охоплюючи широку аудиторію, але її ефективність може знижуватись через так звану "банерну сліпоту". У цьому сегменті також варто виділити Criteo — платформу для динамічного ремаркетингу, яка дозволяє показувати користувачам персоналізовану рекламу на основі їхніх дій на сайті.

Пошукова реклама зосереджена на активних запитах користувача. Google Ads (Search) — один із найефективніших інструментів для залучення клієнтів, які вже мають інтерес до продукту чи послуги. Така реклама має високу конверсію, але конкуренція в пошукових запитах часто підвищує вартість кліку. Альтернативою є Bing Ads, яка забезпечує нижчу вартість за клік і корисна для охоплення аудиторії, що використовує браузери Windows за замовчуванням, зокрема у США та Великій Британії.

Нативна реклама та месенджери зосереджені на природній інтеграції бренду в контент і спілкування. Telegram є популярною платформою для прямого контакту з користувачами через боти, канали та чати. Його перевага — висока довіра, активне залучення та хороша видимість. Недолік — менш розвинена аналітика та складніша автоматизація процесів. Нативна реклама також реалізується через платформи на кшталт Taboola або через публікації у блогах, ЗМІ та спеціалізованих сайтах, де реклама виглядає як частина основного контенту.

Таким чином, кожна з платформ має свої унікальні особливості, переваги й обмеження. Найефективніший підхід — це мультиканальна стратегія, яка поєднує різні формати та канали, орієнтуючись на потреби та поведінкові особливості цільової аудиторії. Це дозволяє створити повноцінну екосистему взаємодії з користувачем на всіх етапах його шляху до покупки.

Цифрові платформи відіграють важливу роль у забезпеченні конкурентоспроможності сучасних компаній, оскільки сприяють не лише залученню нових споживачів, а й утриманню наявної аудиторії шляхом побудови довготривалої комунікації. Можливість оперативного реагування на зворотний зв’язок, персоналізація контенту та формування позитивного користувацького досвіду підвищують рівень задоволеності клієнтів і сприяють зміцненню їхньої лояльності до бренду. Проте низький рівень персоналізації та масовий підхід у цифровому маркетингу, зокрема у вигляді одноманітних рекламних розсилок, може суттєво знижувати ефективність комунікації з цільовою аудиторією. Для окремих сегментів споживачів така реклама не лише втрачає актуальність, а й викликає роздратування, що негативно впливає на сприйняття бренду. У результаті неефективної рекламної кампанії може постраждати як репутація компанії, так і її загальна ринкова позиція. Так, на початку своєї діяльності McDonald’s дотримувався стратегії масового підходу — компанія пропонувала однакові гамбургери для всіх клієнтів, незалежно від їхніх смаків, уподобань чи культурних особливостей. Такий підхід був ефективним у період обмеженої конкуренції та менш вимогливого споживача. Проте зі зростанням глобалізації та цифровізації змінилася й поведінка аудиторії:

користувачі почали очікувати персоналізованого досвіду та адаптації до локальних контекстів. [3] Тому сучасна стратегія цифрового маркетингу повинна враховувати потреби конкретних груп споживачів і базуватися на персоналізованому підході. Вона має охоплювати комплекс заходів із управління конкурентоспроможністю підприємства, включаючи гнучке налаштування каналів комунікації, аналітику споживчої поведінки та адаптацію контенту відповідно до змін зовнішнього середовища. [4]

Персоналізація відіграє ключову роль у формуванні стійкого зв'язку між брендом і споживачем. В умовах високої конкуренції та перенасиченості інформаційного простору, лише ті компанії, які здатні надати релевантний і цінний контент у потрібний момент, можуть утримати увагу користувача. Персоналізовані пропозиції підвищують рівень залучення, сприяють зростанню лояльності клієнтів і, як наслідок, — ефективності маркетингових заходів. Ще у XX столітті Філіп Котлер — один із провідних теоретиків маркетингу — наголошував на необхідності чіткого розподілу ринку на окремі сегменти задля задоволення потреб споживачів. [5] Він виділяв чотири основні приципи сегментації аудиторії: географічна, демографічна, психографічна та поведінкова.

- ***Географічна сегментація*** ґрунтується на місці проживання споживачів — країні, місті, регіоні, кліматі чи густоті населення. Наприклад, мережа кав'ярень може адаптувати своє меню відповідно до кліматичних умов: у південних регіонах робити акцент на холодних напоях, а в північних — на гарячих напоях і випічці. Туристичні агентства можуть рекламувати гірськолижні тури в Карпати саме мешканцям великих міст Західної України, де попит на такий відпочинок традиційно високий.
- ***Демографічна сегментація*** враховує такі характеристики, як вік, стать, рівень доходу, освіта, професія, сімейний стан чи національність. Наприклад, виробники преміальних автомобілів (Tesla, Mercedes) орієнтуються на споживачів з високим доходом та статусною професією. А додатки для

вивчення мов, як-от Duolingo, часто спрямовані на студентів і молодих спеціалістів віком 18–35 років, які прагнуть до саморозвитку.

- ***Психографічна сегментація*** фокусується на внутрішньому світі споживача — стилі життя, інтересах, цінностях, особистих установках. Бренд Apple, зі свого боку, часто асоціюється з креативними, інноваційними та амбіційними людьми, які прагнуть виділятись і цінують естетику. A Lush — виробник натуральної косметики — приваблює тих, хто підтримує етичне споживання, уникає тестування на тваринах і обирає екологічні рішення.
- ***Поведінкова сегментація*** базується на поведінці клієнтів: частоті покупок, ступені лояльності, етапі готовності до покупки, реакції на акції тощо. Наприклад, Netflix аналізує уподобання користувачів (жанри, час перегляду) та пропонує персоналізовані рекомендації. Авіакомпанії створюють програми лояльності, щоб утримувати постійних клієнтів і заохочувати їх повертатись.

У реальному бізнесі компанії часто комбінують кілька типів сегментації для максимально точної роботи з аудиторією. Наприклад, бренд одягу може орієнтуватися одночасно на географічні особливості (холодний клімат), вік аудиторії (молоді люди 20–30 років), стиль життя (урбаністичний, активний) та поведінку (часті покупки онлайн). Такий підхід дозволяє ефективніше розробляти продукти, будувати комунікацію й підвищувати рівень залученості клієнтів.

Ще одним популярним методом сегментації є метод, розроблений М. Шеррингтоном, під назвою «5W». Його суть полягає у поділі аудиторії не по групах, а по відповідях на 5 питань, які в оригіналі мають наступні назви: What? Who? Why? When? Where? [5]

Окрім традиційних типів, існують й інші підходи, які дозволяють більш детально розуміти потреби та переваги споживачів. Один із таких типів — техногенна сегментація, яка орієнтується на використання різних технологій та пристроїв. Цей підхід дозволяє компаніям адаптувати свої стратегії в залежності від того, чи користуються споживачі мобільними пристроями чи працюють через десктопи. Такий аналіз допомагає створювати зручніші пропозиції, враховуючи різні

платформи й формати. Сегментація за життєвим циклом клієнта зосереджена на етапі взаємодії користувача з брендом — чи це новий, постійний або потенційний клієнт. Така сегментація дозволяє брендам розробляти відповідні стратегії: для нових клієнтів пропонуються знижки, для постійних — програми лояльності, а для потенційних — інформативні кампанії. Крім того, існує сегментація за ставленням до бренду, де користувачі можуть бути лояльними, нейтральними чи негативно налаштованими. Розуміння цього дозволяє застосовувати різні стратегії для підтримки та зміцнення взаємодії з кожною групою. Сегментація за емоційним станом або настроєм є досить новим підходом і базується на визначенні емоційного стану споживачів у конкретний момент часу. Це дозволяє брендам створювати кампанії, які відповідають потребам людей в особливі моменти, наприклад, під час свят, стресових ситуацій або важливих подій. Додатково, сегментація за соціальним статусом зосереджується на орієнтації на різні соціальні класи, такі як заможні, середній клас чи студенти. Це дозволяє компаніям адаптувати свої продукти та послуги відповідно до потреб кожної групи, наприклад, преміум-бренди орієнтуються на високий соціальний статус своїх клієнтів, в той час як інші бренди пропонують доступні ціни для середнього і низького класу. Не менш важливою є сегментація за культурними факторами, яка враховує етнічні, релігійні, мовні та інші культурні особливості споживачів. Такий підхід дозволяє брендам адаптувати свої маркетингові кампанії до різних культур і традицій, що є особливо важливим для глобальних компаній. Компанії, що продають продукти харчування або побутові товари, можуть створювати спеціалізовані пропозиції, враховуючи культурні переваги та звичаї своїх клієнтів.

1.2. Концептуальні основи збору даних та аналізу взаємодії користувачів із контентом та рекламою

У сучасному цифровому середовищі аналіз поведінки користувачів набуває все більшого значення для компаній, які прагнуть підвищити ефективність контенту та

рекламних кампаній. З розвитком інтернет-технологій та поширенням соціальних платформ виникла потреба у системному підході до збору, обробки та інтерпретації даних про дії користувачів в онлайні. Це дозволяє не лише фіксувати кількісні показники взаємодії (перегляди, кліки, час на сторінці тощо), а й глибше розуміти мотивації, уподобання та реакції аудиторії. Збір таких даних є основою для прийняття обґрунтованих рішень у сфері цифрового маркетингу. Сучасні аналітичні інструменти, включаючи штучний інтелект і машинне навчання, надають можливість не просто аналізувати минулі взаємодії, а й прогнозувати майбутню поведінку користувачів, формуючи тим самим персоналізовані та ефективні комунікаційні стратегії. Узагальнення підходів до збору й аналізу даних про взаємодію з контентом і рекламою дає змогу створити міцне концептуальне підґрунтя для дослідження цифрової поведінки споживачів. Це, своєю чергою, сприяє підвищенню релевантності контенту, ефективності рекламних кампаній і загальної конкурентоспроможності бренду в онлайн-просторі.

Збір даних про взаємодію користувачів із контентом та рекламою ґрунтується на широкому спектрі цифрових джерел. Кожне з них має свою специфіку, тип інформації, яку надає, і роль у формуванні повної картини поведінки користувача. Розглянемо основні джерела збору даних, які найчастіше використовуються у цифровому маркетингу.

Трекінгові пікселі — це частина коду, що вбудовується в код вебсторінки, email-розсилки або рекламного оголошення для збору даних про дії користувача. Коли користувач відкриває сторінку або лист, що містить піксель, браузер надсилає запит на завантаження цього елемента з сервера компанії, яка його встановила (наприклад, Meta або TikTok) [6]. Разом із цим запитом передається низка параметрів: IP-адреса, тип пристрою та браузера, геолокаційна інформація, URL сторінки, на якій було активовано піксель, час та дата взаємодії, унікальний ідентифікатор користувача (якщо він авторизований або вже ідентифікований раніше через cookie). Ці дані обробляються на сервері і використовуються для подальшого аналізу та оптимізації. Найпоширенішими прикладами таких пікселів є Meta Pixel

(Facebook та Instagram), TikTok Pixel, Google Ads Conversion Pixel, LinkedIn Insight Tag, Twitter Pixel і Pinterest Tag.

Завдяки своїм технічним характеристикам та інтеграції з аналітичними платформами, піксель виконує низку важливих функцій, які дозволяють оптимізувати маркетингові стратегії компанії, а саме:

1. ***Відстеження конверсій.*** Однією з головних функцій пікселя є фіксація цільових дій користувача, таких як реєстрація, оформлення замовлення, підписка або завантаження мобільного додатку. Це дозволяє точно визначити, які рекламні кампанії приносять результат, і забезпечує зворотний зв'язок для коригування стратегії.
2. ***Ретаргетинг.*** Піксель дає змогу визначити користувачів, які вже взаємодіяли з сайтом, але не завершили цільову дію. На основі цієї інформації налаштовується показ персоналізованої реклами, що суттєво підвищує ймовірність повторного залучення та конверсії.
3. ***Формування та сегментація аудиторій.*** За допомогою пікселя можна створювати кастомні аудиторії на основі дій користувачів (відвідування сторінок, додавання товару в кошик тощо), а також схожі аудиторії (lookalike audiences) — групи користувачів із подібними характеристиками до наявних клієнтів. Це підвищує точність таргетингу.
4. ***Оптимізація рекламних кампаній.*** Дані, зібрані через піксель, використовуються для алгоритмічної оптимізації рекламних показів. Системи машинного навчання на основі цих даних автоматично визначають, які користувачі з більшою ймовірністю здійснять цільову дію, і коригують стратегію показу відповідно. [7]
5. ***Вимірювання ефективності реклами.*** Пікселі дозволяють проводити комплексний аналіз ефективності рекламних кампаній: порівнювати оголошення, джерела трафіку, пристрої та географічні регіони. Це створює основу для прийняття обґрунтованих маркетингових рішень.

6. ***Інтеграція з зовнішніми системами.*** Трекінгові пікселі можна інтегрувати з CRM-системами, платформами електронної комерції, email-маркетингом та іншими інструментами бізнес-аналітики. Така інтеграція забезпечує повноцінне бачення взаємодії користувача на всіх етапах воронки продажів.

Таким чином, трекінговий піксель виступає універсальним інструментом у сфері цифрової аналітики, який дозволяє не лише відстежувати поведінку користувачів, а й динамічно впливати на рекламну стратегію в реальному часі.

Попри ефективність, трекінгові пікселі стикаються з низкою обмежень. По-перше, законодавство про захист персональних даних (GDPR, CCPA) зобов'язує компанії отримувати згоду користувача на збір такої інформації [8]. По-друге, користувачі все частіше використовують блокувальники трекерів і приватні режими перегляду. Також, оновлення політик конфіденційності з боку платформ (наприклад, iOS 14+) обмежують можливості відстеження без явного дозволу.

Наступним способом збору даних про користувачів є системи веб-аналітики. Вони є інструментами, що дозволяють збирати, вимірювати, аналізувати та інтерпретувати дані про відвідувачів веб сайтів і мобільних додатків [9]. Вони дають змогу отримати важливу інформацію про поведінку користувачів, джерела трафіку, ефективність маркетингових кампаній. Це дозволяє компаніям приймати обґрунтовані рішення для покращення користувацького досвіду, підвищення конверсій та оптимізації рекламних витрат. Основними функціями систем веб-аналітики є:

- Перша функція будь-якої системи веб-аналітики полягає у зборі даних про користувачів: відвідувані сторінки, час перебування на сайті, кліки, прокручування, введення даних у форми, покупки тощо.
- Системи веб-аналітики можуть визначити, звідки приходять користувачі (джерела трафіку): з пошукових систем, соціальних мереж, реферальних сайтів або безпосередньо. Це дозволяє виміряти ефективність різних каналів маркетингу.

- Система дозволяє налаштувати і відстежувати конверсії (цільові дії, такі як покупки, реєстрації або підписки). Це важливо для визначення ефективності рекламних кампаній і сайту в цілому.
- Використання систем веб-аналітики дозволяє сегментувати користувачів за різними ознаками: географія, демографія, поведінка на сайті, пристрої та інші. Це дає можливість створювати персоналізовані маркетингові стратегії.
- Веб-аналітика часто включає можливість проводити A/B-тестування, шоб тестувати різні версії сторінок, банерів або форм, щоб визначити, який варіант є більш ефективним для досягнення цілей (наприклад, підвищення конверсій).

Розглянемо кілька популярних систем веб-аналітики, які допомагають компаніям ефективно використовувати дані для досягнення своїх цілей.

1. **Google Analytics -** це одна з найпоширеніших і найпотужніших систем веб-аналітики. Вона дозволяє відстежувати відвідування сайтів, джерела трафіку, перегляди сторінок, поведінку користувачів, конверсії та багато іншого. Google Analytics також підтримує створення кастомних звітів і надає можливості інтеграції з іншими інструментами, такими як Google Ads і Google Search Console [10]. Наприклад, компанія, що продає продукцію онлайн, може використовувати Google Analytics для аналізу того, які рекламні канали (наприклад, соціальні мережі або Google Search) приносять найбільше відвідувачів, які продукти користувачі переглядають найчастіше, а також на яких етапах воронки продажів вони залишають сайт.
2. **Hotjar** спеціалізується на дослідженні користувацької поведінки через інструменти теплових карт (heatmaps), записів сесій користувачів і опитувань [11]. Він дає змогу зрозуміти, як користувачі взаємодіють із сайтом на глибшому рівні, що важливо для оптимізації UX/UI. Інтернет-магазин може використовувати Heatmaps в Hotjar, щоб побачити, які частини сторінки привертають найбільшу увагу користувачів і куди вони клікають. Це дозволить зробити сайт більш зручним та ефективним у конверсіях.

3. **Adobe Analytics** є більш потужним інструментом для великих підприємств і брендів. Він пропонує складніші можливості для аналізу даних, зокрема глибоку сегментацію аудиторії, прогнозування та персоналізацію [12]. Велика компанія, що працює в кількох країнах, може використовувати Adobe Analytics для аналізу поведінки користувачів на різних мовних версіях сайту і на різних ринках, а також для оптимізації маркетингових кампаній у різних регіонах.
4. **Mixpanel** фокусується на зборі даних щодо поведінки користувачів на рівні подій [13]. Ця система є корисною для компаній, які хочуть відстежувати не лише відвідування, а й більш конкретні дії користувачів, такі як натискання кнопок або додавання товарів у кошик. Стартап, що пропонує мобільний додаток, може використовувати Mixpanel для аналізу того, які функції додатку користуються найбільшим попитом, скільки користувачів активно використовують додаток, і коли вони припиняють взаємодію з ним.
5. **Matomo** — це відкрите програмне забезпечення для веб-аналітики, яке дозволяє зберігати дані на власних серверах, що робить його ідеальним для тих, хто хоче зберігати повний контроль над своїми даними [14]. Matomo підтримує майже всі функції, які є в Google Analytics, включаючи відстеження подій, конверсій і поведінки користувачів. Компанія, яка прагне до високого рівня конфіденційності та контролю даних, може обрати Matomo для аналізу трафіку та ефективності кампаній без залучення сторонніх сервісів.

Ще одним із поширених інструментів для аналізу та збору даних взаємодії користувачів із контентом і рекламою є **CRM-системи** (такі як Salesforce, HubSpot, Zoho) та **сервіси email-маркетингу** (наприклад, Mailchimp, Klaviyo, eSputnik). Ці платформи забезпечують детальний збір даних про реакції користувачів на контент, що надсилається через email-канали, а також про їхню взаємодію з рекламними повідомленнями, які інтегруються в загальну маркетингову стратегію. Зокрема, фіксуються такі параметри, як відкриття листів, кліки за вбудованими посиланнями, час взаємодії з повідомленням, а також відповіді на персоналізовані пропозиції, зокрема знижки або рекомендації, сформовані на основі попередньої поведінки

користувача [15]. Крім того, інтеграція CRM із email-маркетингом дозволяє відстежувати історію покупок, взаємодії зі службою підтримки, та відповідність користувача до певного сегменту аудиторії (наприклад, за рівнем залученості або життєвим циклом клієнта) [16].

Ці системи не тільки автоматизують процес збору даних, а й забезпечують їхню структуризацію, що дозволяє створювати динамічні сегменти аудиторії для точнішого налаштування рекламного контенту. Наприклад, користувачі, які регулярно відкривають листи з новинами, можуть отримувати більш глибокий контент, тоді як ті, хто реагує лише на знижки, потрапляють до сегменту цінових пропозицій. У такий спосіб CRM та email-сервіси стають не лише каналами комунікації, а й джерелами високоточних аналітичних даних, що дозволяють покращити релевантність контенту та підвищити ефективність рекламних кампаній.

Отже, інструменти збору даних про взаємодію користувачів із контентом і рекламою відіграють вирішальну роль у прийнятті обґрунтованих рішень. Завдяки трекінговим пікселям, системам веб-аналітики, CRM-платформам та email-сервісам компанії можуть не лише збирати поведінкову інформацію, а й формувати сегменти та кластери клієнтів на основі їхніх дій, інтересів та характеристик. Це дає змогу ефективно налаштовувати таргетинг, персоналізувати комунікацію та підвищувати конверсії. У підсумку такі дані стають основою для створення результативних бізнес-стратегій, орієнтованих на зростання прибутку та лояльності клієнтів.

Розділ 2. Методологічні основи кластеризації користувачів

2.1. Основні принципи сегментації та кластеризації

Сегментація та кластеризація є ключовими інструментами в аналізі даних, що дозволяють виділяти окремі групи користувачів із подібними характеристиками або поведінкою. Обидва підходи мають на меті підвищення ефективності прийняття рішень у маркетингу, розробці продуктів, персоналізації сервісів тощо.

Сегментація — це процес поділу користувачів на логічно обґрунтовані групи (сегменти) на основі попередньо визначених характеристик, таких як вік, стать, частота використання продукту, рівень доходу, поведінка у додатку тощо. Вона може бути як вручну налаштованою (rule-based), так і автоматизованою — з використанням методів машинного навчання [17]. Сегментація з аналітичної точки зору — це задача оптимізації, що балансує між збереженням значущої структури даних і забезпеченням інтерпретованості результатів [18]. Вона має велике значення у виявленні прихованих закономірностей, побудові гіпотез і підтримці прийняття управлінських рішень на основі даних.

Сегментація застосовується в різноманітних галузях, зокрема в маркетингу для точного визначення цільової аудиторії, у фінансовому секторі для оцінки кредитних ризиків та виявлення шахрайських операцій, у медицині для класифікації пацієнтів та оптимізації лікувальних заходів, у соціальних науках для аналізу демографічних та поведінкових характеристик, у телекомунікаціях для моделювання користувацької поведінки та запобігання відтоку клієнтів, у промисловості для моніторингу технічного стану обладнання, а також в електронній комерції для аналізу покупецьких патернів і формування рекомендаційних систем.

Кластеризація — це один з методів неконтрольованого машинного навчання, який автоматично знаходить приховані структури в даних без наявності заздалегідь заданих міток. Мета кластеризації — згрупувати користувачів так, щоб об'єкти в межах одного кластеру були максимально подібні один до одного, а між різними кластерами — максимально відмінні.

Кластер — група елементів, що характеризуються загальною спільною властивістю, головна ціль кластерного аналізу — знаходження груп схожих об’єктів у вибірці [19]. Сфера застосування кластерного аналізу є надзвичайно широкою: він використовується в таких галузях, як археологія, медицина, психологія, хімія, біологія, державне управління, філологія, антропологія, маркетинг, соціологія та інших дисциплінах. Водночас, завдяки універсальності цього методу, сформувалась велика кількість різноманітних термінів, методів і підходів, що ускладнює його однозначне застосування та інтерпретацію результатів.

Термін «кластерний аналіз» був вперше запропонований Тріоном (Tryon) у 1939 році і охоплює понад сто різних алгоритмів [20]. Однією з переваг кластерного аналізу є те, що він не вимагає апріорних припущень щодо структури даних, не накладає обмежень на тип представлення об’єктів дослідження та дозволяє працювати з різними типами даних, включно з інтервальними, частотними та бінарними показниками.

Основні принципи ефективної кластеризації включають:

- Гомогенність у межах кластеру - користувачі в одному кластері мають мати схожі характеристики або поведінку.
- Гетерогенність між кластерами - відмінності між різними групами мають бути якомога більшими.
- Інтерпретованість результатів - сформовані кластери повинні мати зрозуміле пояснення з точки зору бізнесу або цілей дослідження.
- Стабільність кластерів - при зміні невеликої кількості даних результат кластеризації не повинен кардинально змінюватися.
- Масштабованість - метод повинен ефективно працювати на великих наборах даних.
- Релевантність ознак - обрані змінні для кластеризації повинні мати логічну та аналітичну цінність.

Попри широке використання кластеризації в задачах аналізу даних, цей підхід має низку суттєвих обмежень, які необхідно враховувати під час його практичного застосування:

1. Значна частина алгоритмів кластеризації, зокрема K-means, потребує попереднього задання кількості кластерів. Оскільки ця інформація зазвичай невідома на початкових етапах дослідження, застосовуються емпіричні методи, такі як метод "лікоть" або аналіз коефіцієнта силуету.
2. Якість кластеризації істотно залежить від обраної метрики подібності між об’єктами, а також від попередньої обробки ознак. Відсутність нормалізації або недоречний вибір метрики може призвести до формування кластерів, що не відображають реальну структуру даних.
3. Алгоритми кластеризації, зокрема ті, що використовують середні значення для визначення центрів кластерів (наприклад, K-means), демонструють низьку стійкість до викидів та аномальних спостережень. Наявність таких об'єктів у вибірці може суттєво змістити центри кластерів, що, своєю чергою, призводить до спотворення загальної структури кластеризації та зниження її якості.
4. Крім того, деякі алгоритми кластеризації, зокрема K-means або GMM, припускають, що кластери мають сферичну або еліптичну форму з приблизно однаковими розмірами. У випадках, коли дані мають складну, нерегулярну або витягнуту структуру (наприклад, у формі ланцюжків або областей різної густини), ці методи не здатні адекватно відобразити природний розподіл даних. У таких ситуаціях ефективнішими можуть бути алгоритми, що не вимагають припущень про форму кластерів, як-от DBSCAN або HDBSCAN, які базуються на щільності розподілу точок і краще справляються зі складними топологіями.

Таким чином, хоча кластеризація є потужним інструментом аналізу даних, її ефективне застосування вимагає критичного підходу до вибору алгоритму,

налаштувань параметрів та інтерпретації отриманих результатів з урахуванням зазначених обмежень.

2.2. Огляд методів Data Science для кластеризації користувачів

Сучасні підходи до кластеризації користуються великою популярністю в маркетингу, заснованому на даних. Серед них особливо виділяються алгоритми кластеризації, зокрема K-means та DBSCAN. Вони ефективно працюють із складними, багатовимірними наборами даних, де традиційні методи виявляються малоефективними.

Алгоритм K-means — один із найпоширеніших у сфері кластеризації. Він поділяє користувачів (або клієнтів) на групи на основі певної поведінки, наприклад, покупок. Для його роботи необхідно заздалегідь визначити кількість кластерів, після чого алгоритм поступово призначає кожен об'єкт до найближчого центроїда, уточнюючи межі кластерів на кожній ітерації. Цей метод активно застосовується у сфері роздрібної торгівлі та e-commerce — наприклад, Amazon використовує K-means для групування клієнтів за частотою покупок, що дозволяє точніше формувати персоналізовані рекомендації та налаштовувати рекламу.

Цей алгоритм спрямований на мінімізацію цільової функції, відомої як функція квадратичної помилки, яка визначається за формулою:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

де:

- C*i* - набір точок у кластері □,
- □□ - центроїд кластера □,
- ||*x* - □□ ||^2 - квадрат евклідової відстані між точкою x та центроїдом. [21]

Проте K-means має свої обмеження, зокрема, чутливість до викидів і припущення, що кластери мають бути схожими за формою (сферичними). У таких випадках ефективнішим виявляється алгоритм **DBSCAN (Density-Based Spatial Clustering of Applications with Noise),** який може виявляти складні, нерівномірні

структури в даних. DBSCAN не вимагає попереднього визначення кількості кластерів і добре справляється з виявленням викидів та "аномальних" користувачів.

DBSCAN розподіляє точки по декільком кластерам та виділяє точки шуму:

- ***Core points (ядерні точки)***: точки, що знаходяться в області щільності, що містить принаймні *MinPts* точок у радіусі *ε.* Параметр *MinPts* (minimum points) визначає мінімальну кількість точок у околі, яку потрібно досягти, щоб вважати цю область «щільною», а саму точку — ядерною. Саме ядерні точки виступають центрами для формування кластерів.
- ***Border points (пограничні точки):*** точки, що знаходяться в межах *ε*-околу ядрової точки, але самі не мають достатньої кількості сусідів для того, щоб бути ядровими.
- ***Noise points (шумові точки):*** точки, що не належать до жодного з кластерів, тобто не є ядровими або пограничними точками.

Алгоритм DBSCAN працює за таким принципом: для кожної точки даних шукається її *ε*-около, тобто всі точки, що знаходяться на відстані не більше ніж *ε* від цієї точки. Якщо кількість точок в *ε*-околі більше або дорівнює □□□□□□, то точка стає ядровою, і всі її сусіди (які також задовольняють умови щільності) додаються до одного з кластерів. Пограничні точки не можуть бути ядровими, але вони приєднуються до кластера, яку вони оточують. Якщо точка не має достатньої кількості сусідів для того, щоб бути ядровою, вона позначається як шум.

У дослідженнях, що ґрунтуються на поведінкових або транзакційних даних користувачів, цей метод дозволяє виявити цінні сегменти, які залишилися поза увагою K-means через наявність шуму у вибірці. Наприклад, були знайдені клієнти з незвичною купівельною поведінкою, до яких бізнес згодом застосував індивідуальні маркетингові стратегії, що сприяло зростанню продажів і утриманню клієнтів [22-25].

Загалом, алгоритми кластеризації дозволяють аналізувати одразу багато параметрів користувача, надаючи глибше розуміння поведінки, ніж традиційні методи. Завдяки машинному навчанню (зокрема K-means і DBSCAN), компанії

можуть знаходити приховані закономірності у великих масивах даних і будувати на їх основі персоналізовані маркетингові кампанії, що значно підвищує ефективність використання ресурсів та рентабельність інвестицій. Перехід до просунутої сегментації, орієнтованої на поведінкові патерни, а не лише демографію, став ключем до точного маркетингу нового покоління.

2.3. Критерії оцінки якості кластеризації

У задачах кластерного аналізу ключовим викликом є не лише формування кластерів, а й перевірка того, наскільки отримане розбиття даних є репрезентативним та релевантним. Оскільки кластеризація є методом навчання без заздалегідь відомих класів, оцінити її точність так, як це робиться в задачах класифікації, неможливо. Тому використовують спеціалізовані показники, а саме: внутрішні метрики, а також візуальні, порівняльні та прикладні підходи [30].

У деяких випадках сам факт поділу на групи не є гарантією успішного результату. Іноді кластери можуть формуватися за випадковими або непринциповими ознаками, що не дають практичної цінності. Тому важливо не лише покладатися на обчислені показники, а й осмислювати зміст кожної групи.

Однією з найпопулярніших метрик оцінки результатів кластеризації є Silhouette score (коефіцієнт силуету). Її часто використовують як основний показник загальної якості кластеризації. Вона оцінює, наскільки близькою є кожна точка до інших у своєму кластері порівняно з найближчим сусіднім кластером. Значення метрики лежить у межах від -1 до 1. Якщо вона близька до 1, це означає, що точка добре "вписалася" у свою групу. Якщо значення близьке до нуля — кластер нечіткий, а якщо від’ємне, то точка, скоріше за все, віднесена до невірного кластера. У середньому вважається значення силуета вище 0,5 хорошим результат, 0,3–0,5 — прийнятним, а рівень нижче 0,2 ознака поганої або невиразної кластерної структури.

Ще один показник, який часто використовують, це індекс Девіса-Болдіна (Davies-Bouldin Index, DBI). Він працює за наступним принципом: розраховується,

наскільки кожна група схожа на найближчу до неї іншу. Вважається, що чим далі кластери розташовані один від одного та чим компактніші вони в середині, тим краще. Менші значення даного показника вказують на більш якісну кластеризацію [24]. Значення в межах від 0 до 1 вважається найкращим, від 1 до 2 — прийнятним, а значення метрики вище 2,5, вказує на те, що кластери або перекриваються, або мають багато внутрішнього шуму. Ця метрика зручна тим, що дозволяє швидко оцінити, чи розділення дійсно має сенс з геометричної точки зору. Особливістю DBI є те, що він дає змогу порівнювати кластери не лише як окремі групи, а як систему, де важлива і форма кластерів, і їхня віддаленість один від одного. Його значення добре поєднується з силуетом, оскільки акцентує увагу на міжкластерній різниці [32].

Інший показник, який використовується для оцінки кластеризації – індекс Калінські-Харабаша (Calinski-Harabasz Index, CHI). Він оцінює, наскільки великою є різниця між відстанями всередині кластерів і між різними кластерами. Іншими словами, даний підхід фіксує "контрастність" розподілу: чи кожен кластер дійсно виділяється як окрема група. Значення метрики не обмежується фіксованим діапазоном і може сягати десятків, сотень або навіть тисяч, залежно від кількості спостережень і розмірності даних. Проте в межах одного набору даних вищі значення CHI вказують на краще розділення. Наприклад, CHI у межах 30–100 зазвичай означає хорошу кластерну структуру, тоді як менше 20 може вказувати на накладання груп або відсутність яскраво виражених меж між ними [31].

У більшості випадків варто розглядати ці три метрики спільно. Наприклад, високе значення CHI при низькому силуеті може вказувати на наявність великих, але слабко пов’язаних груп. Комбіноване тлумачення дозволяє уникнути хибних висновків, особливо коли структура даних складна.

Попри зручність і простоту цих метрик, їх результати не завжди треба сприймати буквально. Наприклад, силует може бути невисоким для кластерів складної форми, навіть якщо вони мають сенс у практичному використанні. А високий CHI ще не гарантує, що групи будуть інтерпретованими або корисними для

бізнесу. Саме тому важливо доповнювати числові оцінки візуальним аналізом. Один з найефективніших способів — зниження розмірності (наприклад, методом головних компонент або t-SNE), щоб зобразити багатовимірні дані у вигляді двовимірного графіка. Це дозволяє побачити форму кластерів, наявність шуму, перекриттів або пустих зон між групами.

Метод головних компонент (PCA) зменшує кількість вимірів, перетворюючи початкові змінні на нові компоненти, які є лінійними комбінаціями ознак. Ці компоненти впорядковані так, що перші з них зберігають найбільшу кількість варіації в даних. На практиці це дозволяє, наприклад, візуалізувати дані з 10 і більше ознаками у вигляді простої двовимірної площини, при цьому зберігаючи найбільш значущі патерни розподілу. Якщо на графіку видно, що точки з одного кластеру згруповані щільно, а від інших відділені, то кластеризація є репрезентативною та якісною. Також PCA дає змогу проаналізувати, які змінні найбільше впливають на кожну компоненту, що може бути корисним для інтерпретації результатів [29].

На відміну від PCA, метод t-SNE (t-distributed Stochastic Neighbor Embedding) не прагне зберегти глобальну структуру даних, а фокусується на збереженні локальних відстаней між точками. Він краще підходить для візуалізації складних, нелінійних структур, наприклад, коли кластери мають незвичну форму у високо-вимірному просторі. T-SNE будує таку карту, де точки, які були близько в оригінальних даних, залишаються близько і на площині. Його основна перевага у здатності демонструвати приховані структури, які не видно в лінійних проєкціях. Проте результат t-SNE не варто використовувати для кластеризації напряму: він більше служить як візуальний інструмент для оцінки форми і відокремленості груп [31].

Після побудови кластерної моделі доречно перейти до профілювання кожної групи, тобто опису типових характеристик, які їх відрізняють. Це може бути середній вік, рівень активності, поширені інтереси або поведінкові шаблони. Профілювання відіграє роль концептуального зв’язку між кількісним аналізом результатів кластеризації та їх предметною інтерпретацією, забезпечуючи перехід

від суто обчислювального підходу до практичного застосування отриманих сегментів [28].

Крім формальних метрик та візуалізації, важливим аспектом є також стійкість кластерної структури. Це означає, що при повторному аналізі з дещо іншими даними, випадковою ініціалізацією або варіативністю параметрів отримані кластери залишаються подібними за складом і характеристиками. Така стабільність підвищує довіру до результатів і свідчить про їхню природну наявність у структурі даних, а не про випадковість, зумовлену алгоритмом.

У дослідженнях часто поєднують декілька методів кластеризації та порівнюють їх результати. Це дозволяє оцінити, чи справді певні сегменти повторюються незалежно від підходу, і чи можна вважати їх стійкими ознаками групування. Такий підхід особливо корисний у задачах з високою кількістю змінних або складними, багатовимірними даними.

Коли доступні справжні мітки (наприклад, у штучно створених наборах або задачах з відомими сегментами), застосовують так звані зовнішні метрики, такі як скоригований індекс Ренда або нормалізована взаємна інформація. Вони порівнюють знайдену кластеризацію з "еталонною", але в реальних прикладних задачах такі мітки трапляються рідко.

Загалом, оцінка кластеризації — це завжди поєднання декількох джерел інформації. Жодна метрика не дає вичерпної відповіді сама по собі. Найкращі результати дає підхід, у якому поєднуються числові показники, візуальна та логічна інтерпретація: наскільки знайдені групи пояснювані, стабільні, і чи відображають вони реальні закономірності в даних.

Розділ 3. Розробка та аналіз моделі кластеризації користувачів цифрової платформи

3.1. Опис вхідних даних цифрової платформи та визначення цілей кластеризації

Для дослідження було використано дані про користувачів платформи для спілкування з ресурсу Kaggle [27] за 2024 рік.

Набір даних складається з даних про 500 користувачів та 10 стовпців, серед яких:

- User ID - унікальний ідентифікатор користувача в базі даних.
- Age - вік користувача.
- Gender - стать користувача, представлена у вигляді категоріальної змінної Male або Female.
- Height - зріст користувача у сантиметрах
- Interests - основні інтереси користувача, представлені у вигляді переліку ключових слів або категорій: “Sports”, “Cooking”, “Reading”, “Music”, “Hiking”, “Movies”.
- Children - кількість дітей у користувача у вигляді числового значення.
- Education Level - рівень освіти: High School, Bachelor’s Degree, Master’s Degree, Ph.D.
- Occupation - професія або сфера діяльності користувача.
- App logins - загальна кількість входів користувача на платформу за весь час.
- Frequency of Usage - частота використання застосунку: Daily, Weekly, Monthly.

Розглянемо детальніше структуру розподілу користувачів платформи за різними показниками.

Основна частина користувачів платформи — це молоді дорослі віком від 24 до 35 років, із середнім віком 26,98 років. Хоча розподіл віку є доволі рівномірним, спостерігається незначне зростання кількості користувачів у групі від 30 до 35 років, що видно на рис.1.

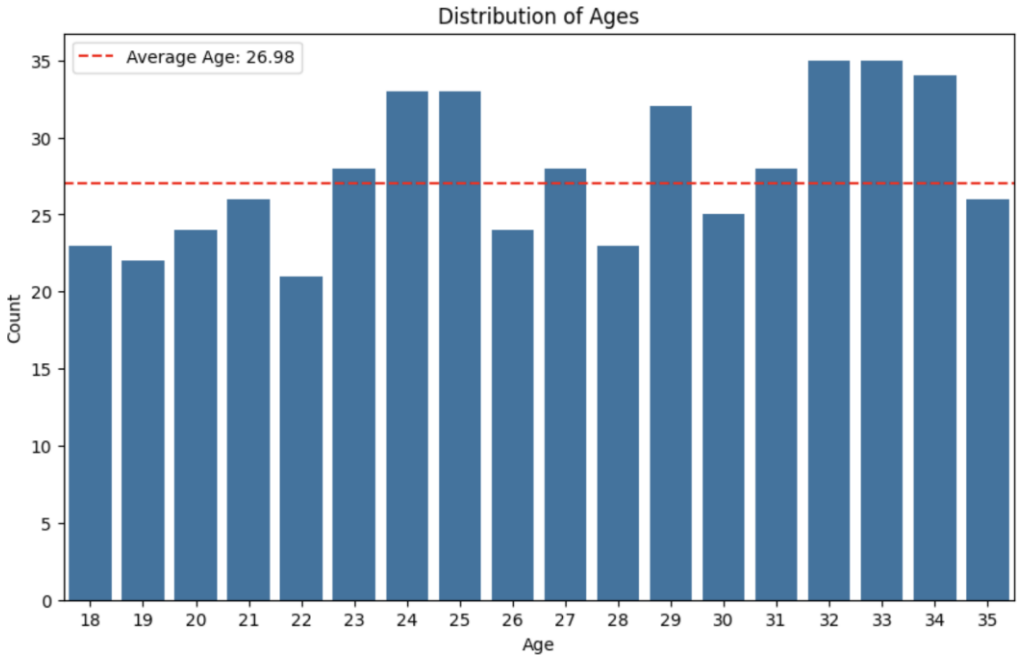

Рис. 1. Розподіл віку користувачів платформи

Джерело: складено автором на основі даних [27]

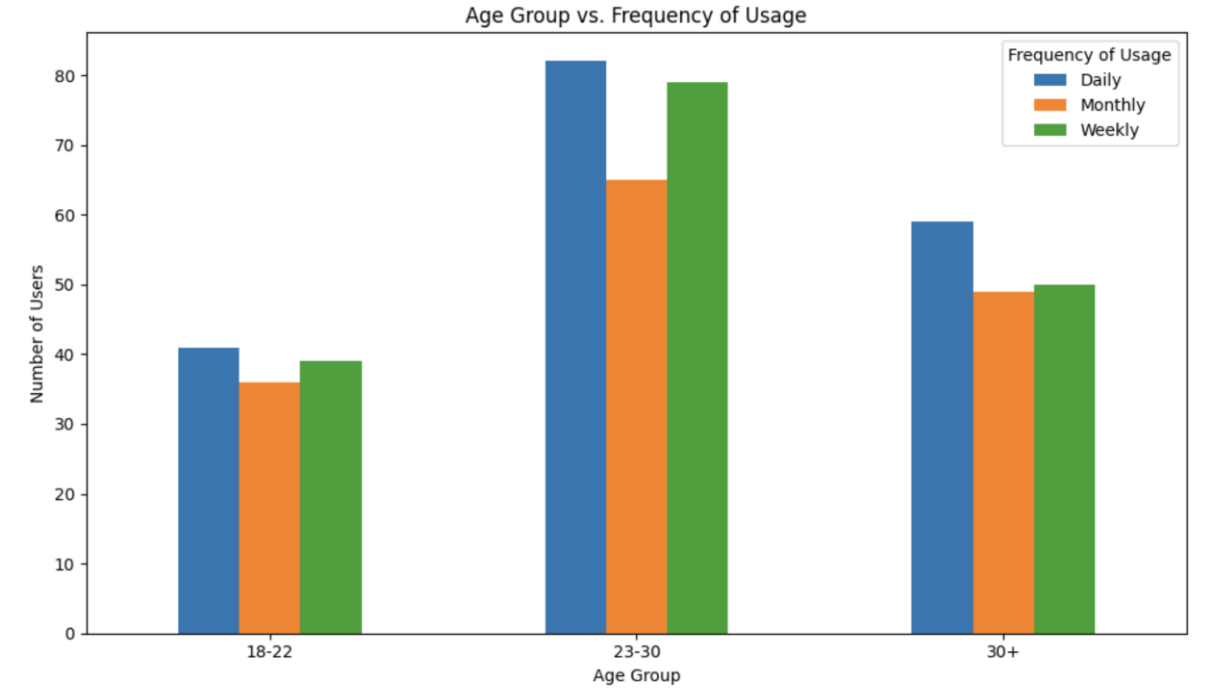

Рис. 2. Частота використання платформи в залежності від вікових груп

Джерело: складено автором на основі даних [27]

З рисунка 2 видно, що найактивніше платформу використовують користувачі віком 23–30 років — як щоденно, так і щотижнево. Вікова група 30+ також демонструє високий рівень залученості, особливо серед щоденних користувачів.

Найнижча активність спостерігається у віковій групі 18–22 роки, незалежно від частоти використання.

Рис. 3 демонструє, що найбільше користувачів обох статей зосереджено у віковій групі 23–30 років, де жінки мають незначну чисельну перевагу. У категорії 18–22 також переважають жінки. У віковій групі 30+ навпаки — більше представників чоловічої статі.

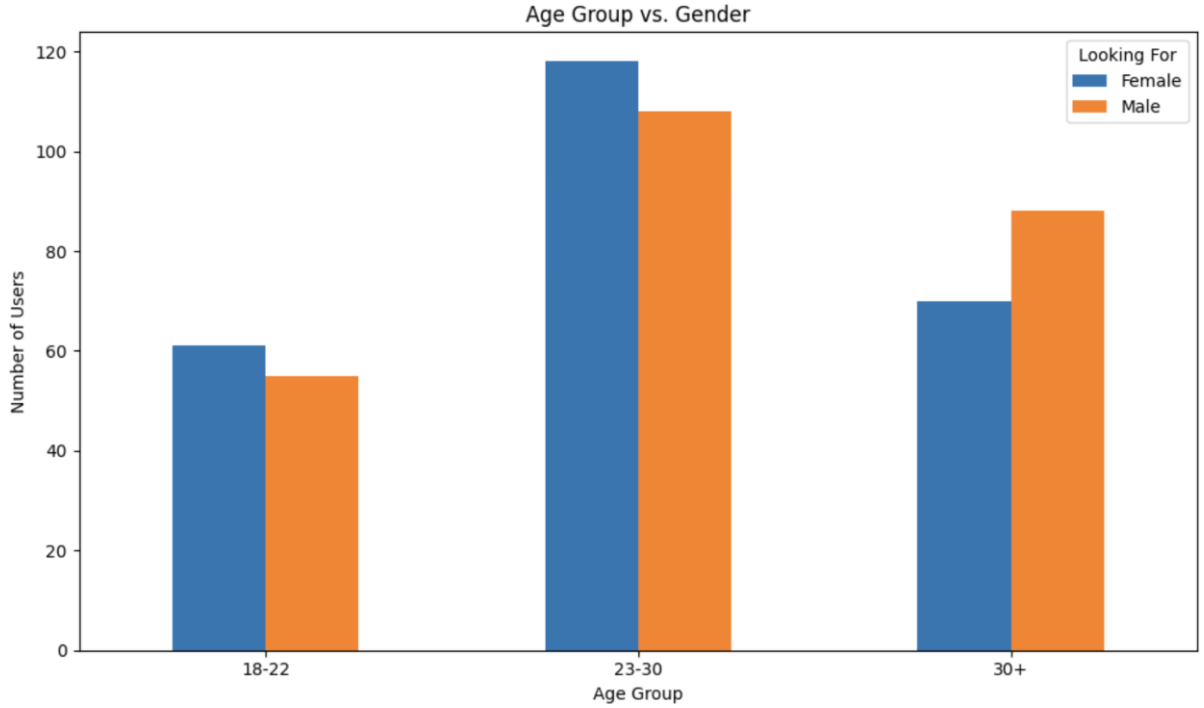

Рис. 3. Розподіл гендеру по віковим групам

Джерело: складено автором на основі даних [27]

Порівняємо різні сегменти користувачів на предмет відмінностей у поведінці та закономірностей використання платформи. На рис. 4 подано візуалізацію даних про частоту використання платформи в залежності від статі користувача.

Найвищий рівень активності в обох груп спостерігається при щоденному використанні, при цьому чоловіки трохи частіше використовують платформу щодня, ніж жінки. Щотижневе користування є досить схожим між обома групами, однак у жінок воно трохи вище. Щомісячна частота використання платформи, навпаки, є трохи вищою серед жінок порівняно з чоловіками. Загалом, поведінка користувачів обох статей має подібну динаміку, але з незначними відмінностями у пріоритетних режимах використання.

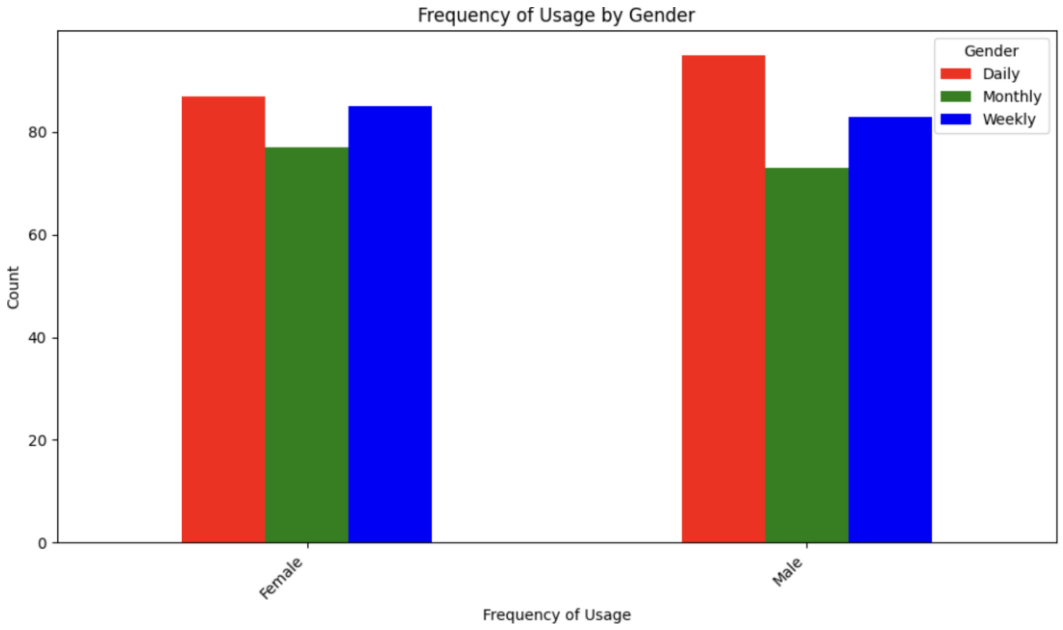

Рис. 4. Розподіл гендеру за частотою використання

Джерело: складено автором на основі даних [27]

За рис. 5. можна побачити, що найвищий рівень щоденного використання спостерігається серед осіб з бакалаврским ступенем і серед випускників середньої школи. Особи з науковим ступенем Ph.D. частіше за інших використовують платформу щотижня, тоді як щоденне і щомісячне використання в цій групі є нижчим. Також варто зазначити, що щомісячне користування є найменш популярним серед усіх освітніх груп, особливо серед бакалаврів та Ph.D. Таким чином, рівень освіти частково впливає на стиль і регулярність використання платформи.

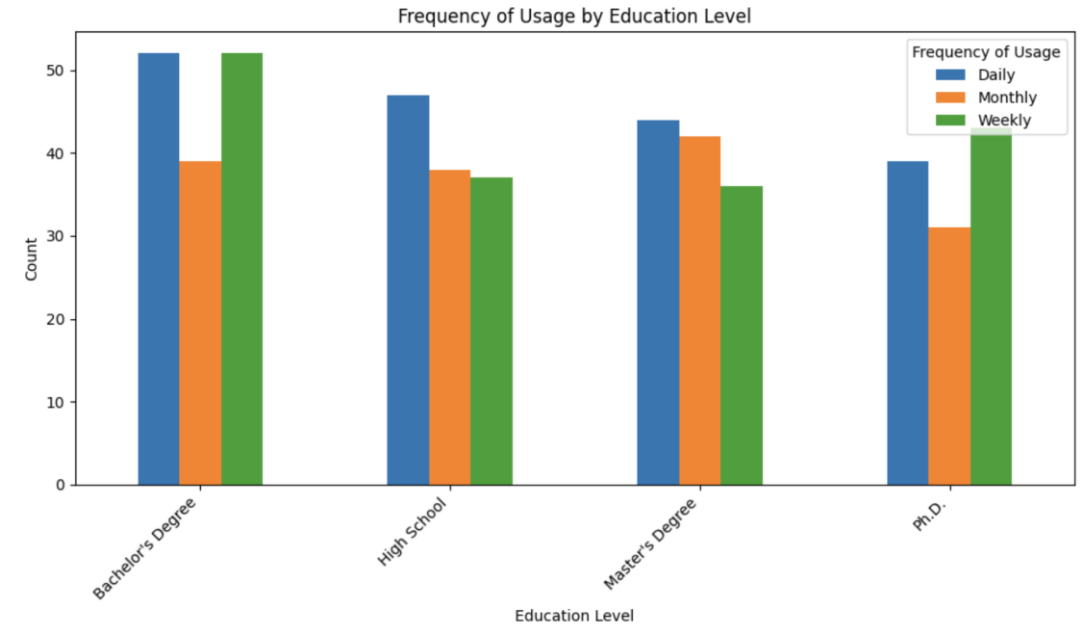

Рис. 5. Розподіл за частотою активності на платформі за рівнем освіти

Джерело: складено автором на основі даних [27]

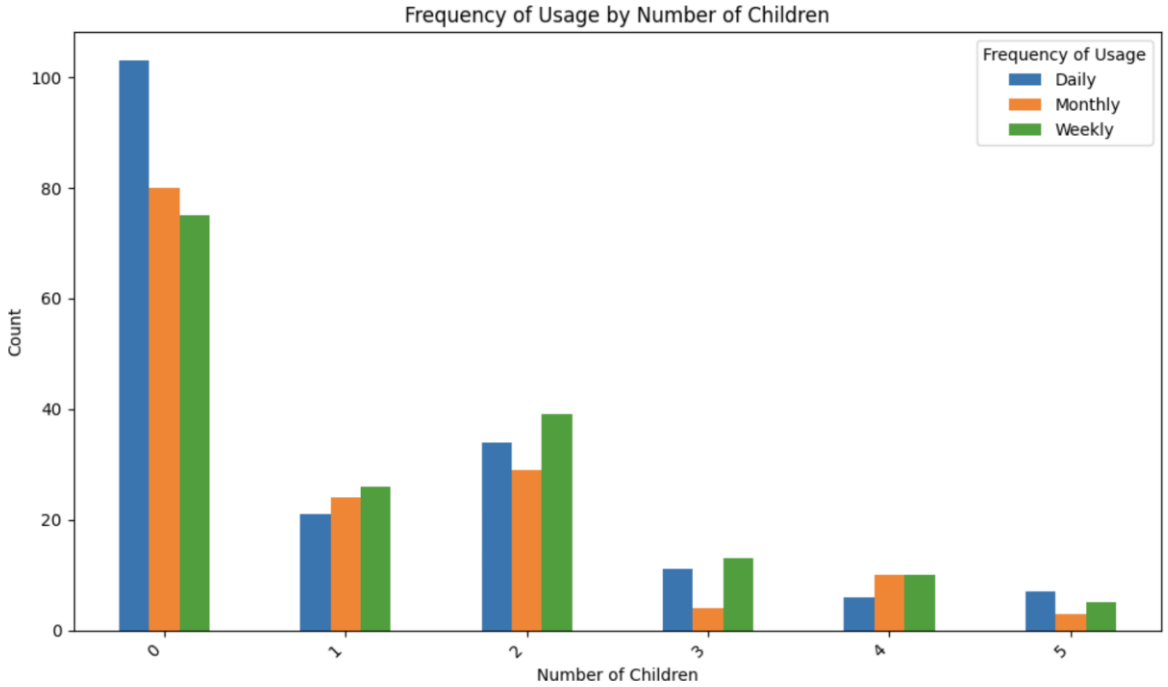

Рис. 6. Розподіл по частоті активності на платформі за кількістю дітей у користувача

Джерело: складено автором на основі даних [27]

Рис. 6. відображає взаємозв'язок між кількістю дітей у користувачів і частотою їхнього використання платформи. Найбільшу активність демонструють люди без дітей — у цій групі показники щоденного, щотижневого та щомісячного

використання є найвищими. У міру збільшення кількості дітей активність поступово спадає, особливо це помітно для щоденного використання. Винятком є користувачі з двома дітьми, серед яких частота щотижневого користування є відносно високою. Це може свідчити про обмеженість вільного часу у батьків і, як наслідок, рідше звернення до платформи.

На рисунку 7 показано розподіл кількості заходів у застосунок серед користувачів із різною частотою використання: щоденною, щомісячною та щотижневою. Для щоденних користувачів спостерігається відносно рівномірний розподіл, з деяким зростанням кількості входів ближче до високих значень (80–100 входів), що свідчить про групу інтенсивно активних користувачів. Щомісячні користувачі демонструють більш хвилеподібний розподіл: помітні піки активності в середньому та високому діапазонах входів, проте значна частина все ще залишається у нижніх значеннях. Щотижневе використання виглядає найбільш варіативним: активність розкидана по всьому діапазону, але з кількома локальними максимумами, що вказує на гетерогенність у поведінці цієї групи користувачів.

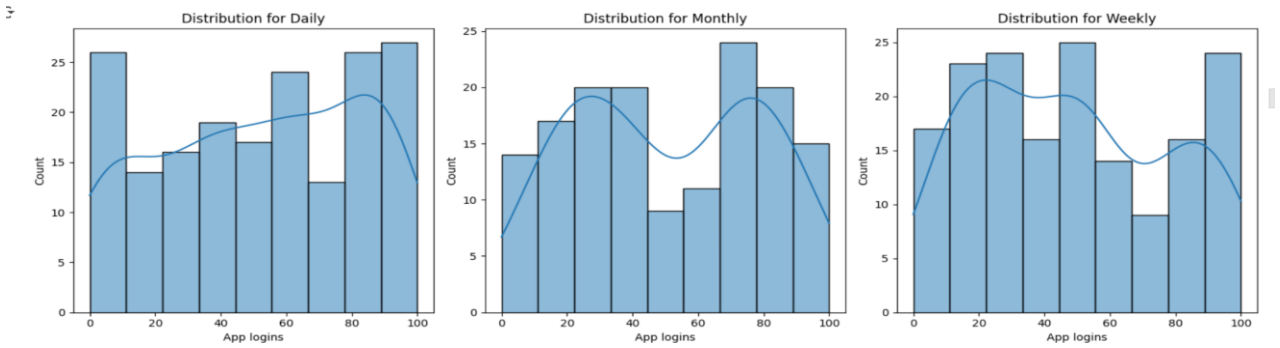

Рис. 7. Розподіл кількості заходів на платформу для кожної групи за частотою використання

Джерело: складено автором на основі даних [27]

Рис. 8. відображає найчастіше згадувані уподобання серед користувачів комунікаційної платформи. На першому місці за популярністю стоїть «Travel», що може вказувати на прагнення до нових вражень та змін обстановки. Решта інтересів — від музики та спорту до кулінарії, читання й фільмів — демонструють доволі

збалансоване розподілення, без явних відривів. Це свідчить про те, що аудиторія є доволі різноплановою, а платформа потенційно об’єднує людей із широким спектром захоплень.

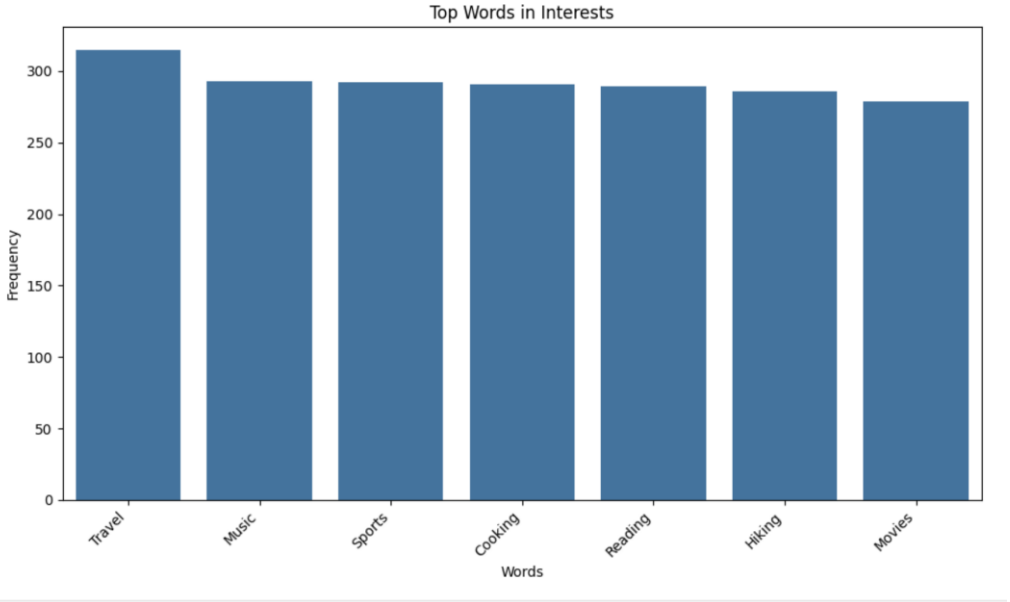

Рис. 8. Топ інтересів серед аудиторії платформи для спілкування

Джерело: складено автором на основі даних [27]

Рис. 9. демонструє кореляційні зв’язки між віком користувачів, кількістю їхніх інтересів та частотою входів у застосунок. Усі кореляційні коефіцієнти залишаються на дуже низькому рівні, що свідчить про відсутність виражених лінійних залежностей між цими характеристиками. Зокрема, кількість інтересів має лише незначний позитивний зв’язок із кількістю входів, а інші пари змінних практично не взаємопов’язані. Це може означати, що користувацька активність та зацікавленість платформою формується під впливом інших, більш суттєвих факторів.

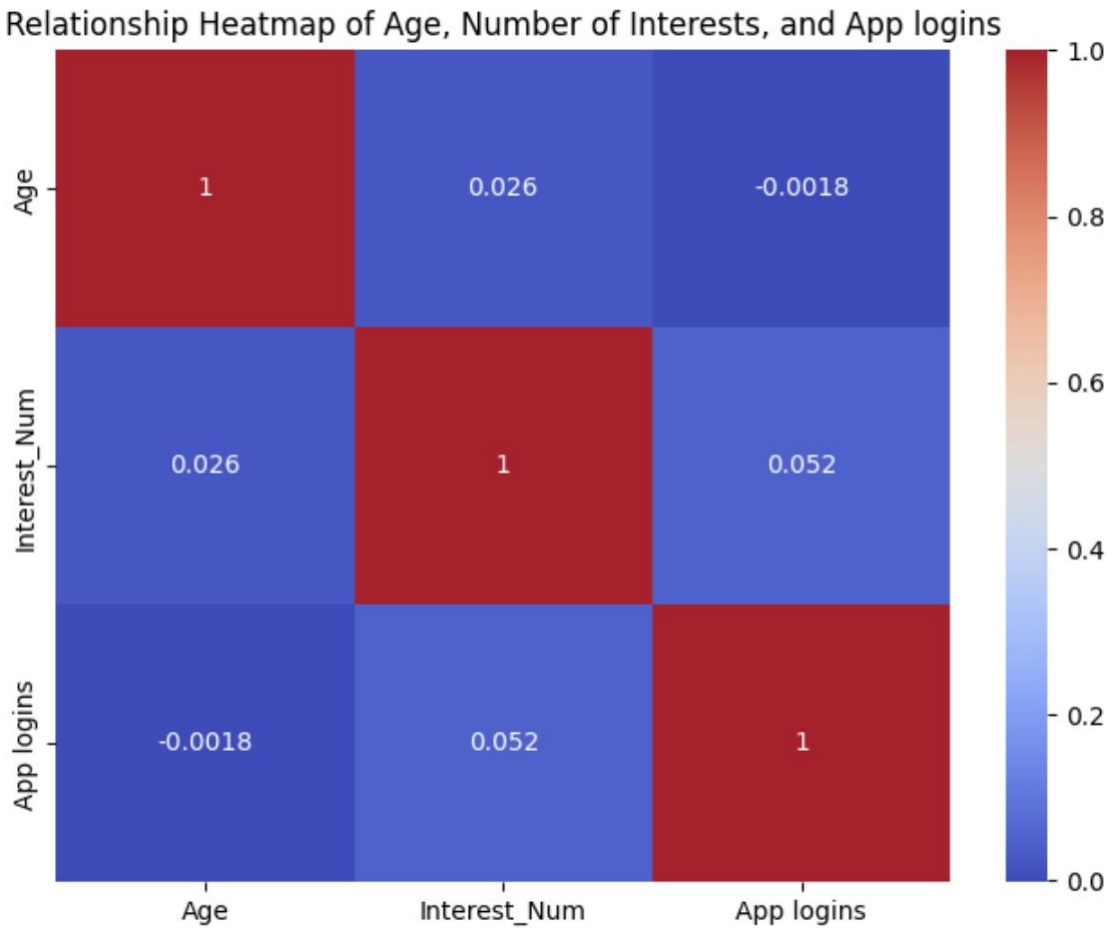

Рис. 9. Теплова карта взаємозв’язку віку, кількості інтересів та активності в додатку

Джерело: складено автором на основі даних [27]

Рис. 9. демонструє кореляційні зв’язки між віком користувачів, кількістю їхніх інтересів та частотою входів у застосунок. Усі кореляційні коефіцієнти залишаються на дуже низькому рівні, що свідчить про відсутність виражених лінійних залежностей між цими характеристиками. Зокрема, кількість інтересів має лише незначний позитивний зв’язок із кількістю входів, а інші пари змінних практично не взаємопов’язані. Це може означати, що користувацька активність та зацікавленість платформою формується під впливом інших, більш суттєвих факторів.

3.2. Побудова та реалізація алгоритму кластеризації

Перед початком кластеризації важливим етапом є визначення оптимальної кількості кластерів (параметр k), яка дозволяє найкраще сегментувати аудиторію відповідно до обраних характеристик. Один з найпоширеніших методів, що використовується з цією метою, — метод «лікоть» (Elbow Method). Метод «лікоть» (Elbow Method) використовується для визначення оптимальної кількості кластерів (k) в алгоритмі кластеризації, зокрема, в K-середніх (K-Means). Суть методу полягає у побудові графіка залежності інерції (внутрішньокластерної дисперсії) від кількості кластерів. Інерція зменшується при зміні кількості кластерів, після чого приріст стає менш значним. Ця точка перегину нагадує «лікоть» і вважається оптимальним значенням k [28].

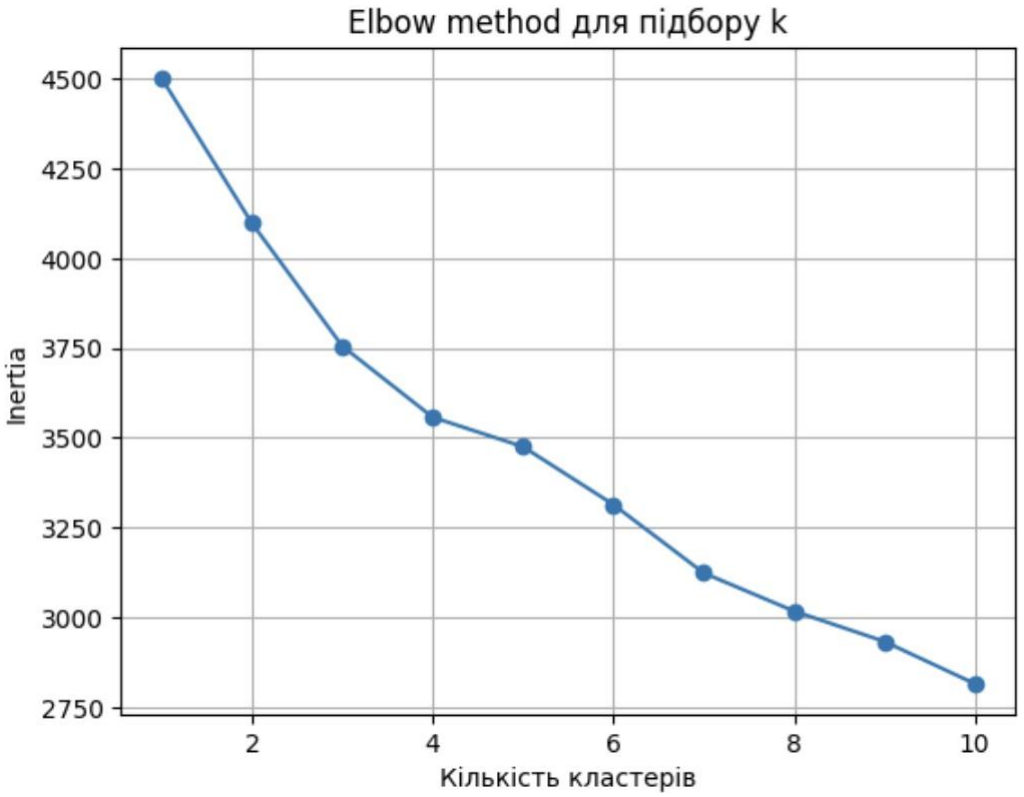

Рис. 10. Графік вибору оптимальної кількості кластерів

Джерело: складено автором на основі даних [27]

На рис. 10 представлено графік, побудований за методом «лікоть», де по осі абсцис відображено кількість кластерів, а по осі ординат — значення інерції. Як

видно з графіка, найбільш виражений перегин спостерігається при значенні k = 4, що свідчить про доцільність обрання чотирьох кластерів для подальшого аналізу аудиторії. Це значення забезпечує баланс між точністю сегментації та уникненням надмірного поділу на кластери, що могли б ускладнити інтерпретацію результатів.

Для поглибленої сегментації аудиторії було використано алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise), який відноситься до методів кластеризації на основі щільності. На відміну від алгоритму k-середніх, DBSCAN не потребує попереднього визначення кількості кластерів — замість цього формує групи на основі локальної густини точок. Такий підхід є ефективним для виявлення кластерів довільної форми та виділення викидів, що не належать до жодної з груп.

Ключовими гіперпараметрами DBSCAN є:

- **ε (epsilon)** — максимальна відстань між точками для їх визнання сусідніми. Надто мале значення призводить до великої кількості шумових точок, тоді як надто велике — до злиття кластерів.
- **min_samples** — мінімальна кількість точок у межах ε, необхідна для того, щоб точка вважалась ядром кластера. Цей параметр визначає поріг густини, при якому область простору розпізнається як кластер.

У межах дослідження було здійснено підбір оптимального поєднання параметрів ε та min_samples з метою досягнення чіткого структурування даних. Оцінка ефективності здійснювалась за кількістю виявлених кластерів та часткою шумових точок та візуалізацією результатів кластеризації.

Рис. 11. Розподіл кількості кластерів в залежності від параметрів ε та min_samples

Джерело: складено автором на основі даних [27]

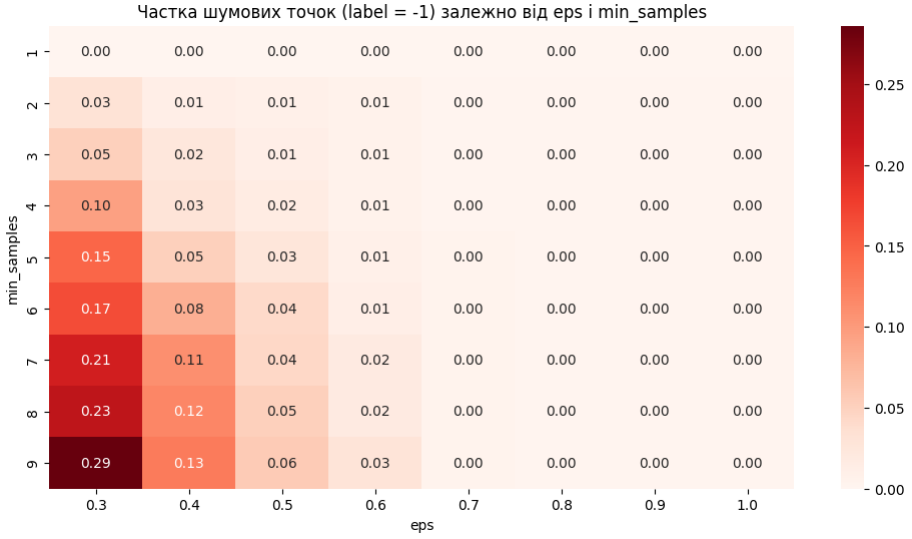

Рис. 12. Розподіл частки шумових точок в залежності від параметрів ε та min_samples

Джерело: складено автором на основі даних [27]

На основі побудованих розподілів кількості кластерів та частки шумових точок, наведених на рис. 11 та 12., було обрано оптимальні параметри для методу кластеризації DBSCAN. Налаштування обирались за такими критеріями:

- 2 < кількість кластерів < 6;
- частка шумових точок складає не більше 5%.

В результаті було обрано параметри ε = 0.4 та min_samples = 5, які в комбінації забезпечують кількість кластерів на рівні 3 та частку шумових точок 5%.

Оскільки DBSCAN працює в багатовимірному просторі ознак, для візуалізації кластерної структури було застосовано метод головних компонент (PCA) — статистичний метод для зниження розмірності даних, який допомагає перетворити великі набори даних з багатьма ознаками (вимірами) у більш компактну форму [29]. Завдяки цьому методу вдалось знизити розмірність до двох вимірів, при цьому не втративши забагато інформації – рівень поясненої дисперсії двома компонентами становить 76,28% (з яких PCA 1 пояснює 38,51%, а PCA 2 – 37,76%). Внесок кожної змінної в компоненти наведено у додатку А.

3.3. Опис результатів моделювання та якості кластеризації

У межах дослідження було проведено кластеризацію даних користувачів із застосуванням трьох алгоритмів: K-Means, DBSCAN та Hierarchical Agglomerative Clustering. Для порівняння ефективності моделей здійснено оцінювання якості кластеризації за низкою метрик, що відображають ступінь внутрішньої згуртованості кластерів, відокремленість груп та відповідність сегментації реальній структурі даних, а саме:

- Silhouette score (коефіцієнт силуету) – метрика, яка показує наскільки близька точка до інших точок у своєму кластері, і наскільки далеко вона від точок інших кластерів. Даний показник може набувати значень від -1 до 1 і чим ближчий він до 1, тим більш якісно було поділено дані на кластери [30].
- Davies-Bouldin Index (DBI) – оцінює середню схожість кожного кластера з найбільш схожим на нього. Тобто, чи є кластери компактними всередині й розділеними між собою. Значення, ближче до 0, відповідає більш якісному розподілу даних по кластерах [31]
- Calinski-Harabasz Index (CHI) – оцінює наскільки кластери віддалені один від одного, порівняно з тим, наскільки щільно точки зібрані всередині кластерів. Рахується як співвідношення відстані між кластерами та розсіювання

всередині кластерів. Більш якісні кластеризації зазвичай мають вищий CHI. [32]

У табл. 1 наведено порівняння результатів використаних методів кластеризації за даними показниками.

Таблиця 1

Порівняння результатів кластеризації за метриками якості

	Silhouette score	DBI	CHI
K-Means	0,612	1,868	43,834
DBSCAN	0,0864	1,566	11,608
Hierarchical	0,4608	2,372	34,233

Джерело: складено автором на основі даних [27]

Аналіз якості кластеризації демонструє перевагу методу K-Means. Він досяг найвищого значення коефіцієнта силуету (0.612), що свідчить про чітке відокремлення кластерів, та показав найкраще співвідношення між міжкластерною та внутрішньокластерною дисперсією згідно з індексом CHI (43.834). У той час як DBSCAN продемонстрував найнижчий DBI (1.566), що вказує на компактність кластерів, його загальна якість кластеризації залишалася низькою через дуже низьке значення Silhouette Score (0.0864) і невисоку міжкластерну віддаленість (CHI = 11.608), ймовірно через значну частку шумових точок. Ієрархічна кластеризація показала помірну якість за всіма показниками, зокрема Silhouette Score 0.4608 та CHI 34.233, однак її високий DBI (2.372) свідчить про нечітке розділення кластерів. Загалом, KMeans забезпечив найстабільніше і найінтерпретованіше розділення користувачів, тоді як DBSCAN та ієрархічні методи можуть розглядатися як допоміжні або альтернативні залежно від цілей подальшого аналізу.

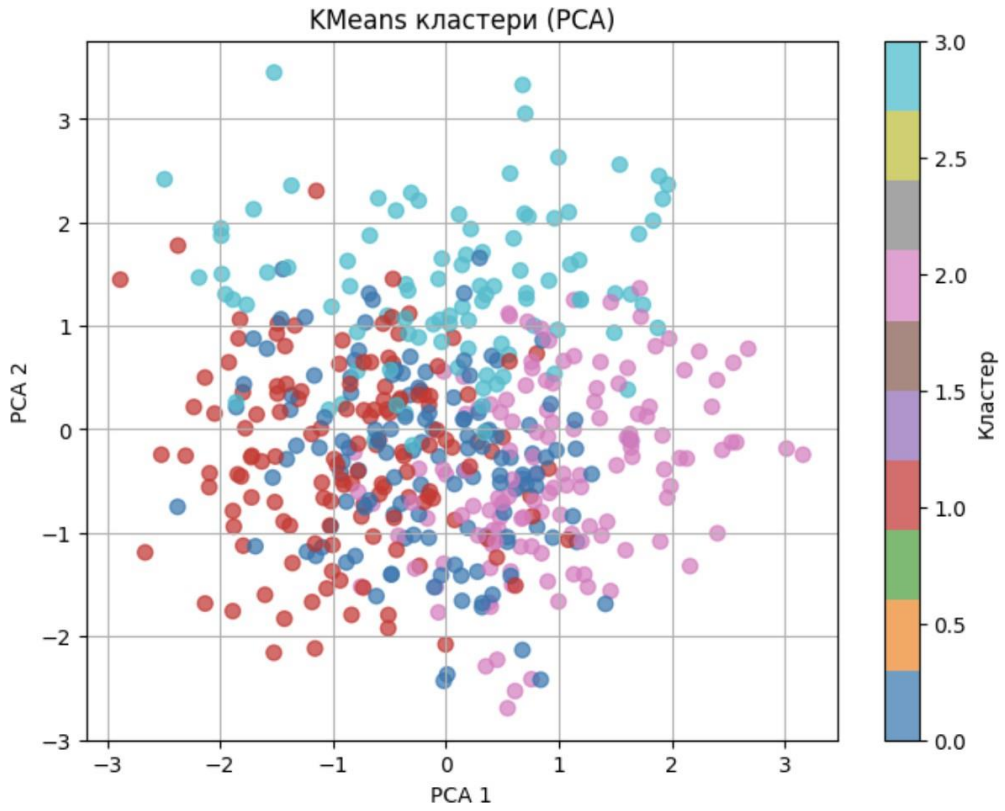

Рис. 13. Розподіл даних користувачів по кластерах за допомогою методу K-means
Джерело: складено автором на основі даних [27]

На рис. 13 представлено візуалізацію чотирьох кластерів за допомогою методу KMeans, що позначені різними кольорами:

1. **Червоний кластер** складається з найбільшої кількості точок. Це означає, що найбільша група користувачів має схожі патерни поведінки (з високою частотою взаємодії та середньою активністю).
2. **Рожевий кластер** на другому місці за чисельністю. Може відображати активних, але нішевих користувачів — наприклад, тих, хто надає перевагу текстовому спілкуванню.
3. **Синій кластер** трохи менше за рожевий і червоний, але все ще охоплює значну частину аудиторії. Цей кластер відповідає за «пасивних» користувачі, які реєструються на платформі, але спілкуються не так часто.
4. **Блакитний кластер** - найменший сегмент. Це група нішевих або нових користувачів із особливою моделлю взаємодії (наприклад, переважно голосові дзвінки чи рідкісні візити).

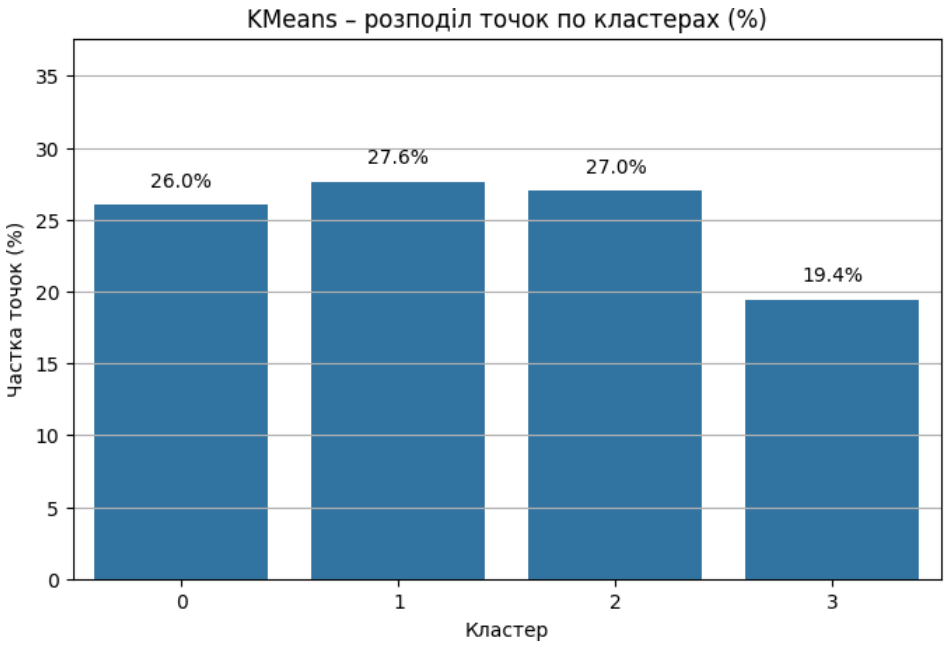

Рис. 14. Відсоткова структура кластерів, визначених методом K-Means

Джерело: складено автором на основі даних [27]

Як видно з рисунку 14, метод K-means розподілив користувачів на 4 кластери майже порівну. Це свідчить про те, що характеристики більшості користувачів є досить схожими і відсутня група користувачів, яка виділяється за певними ознаками.

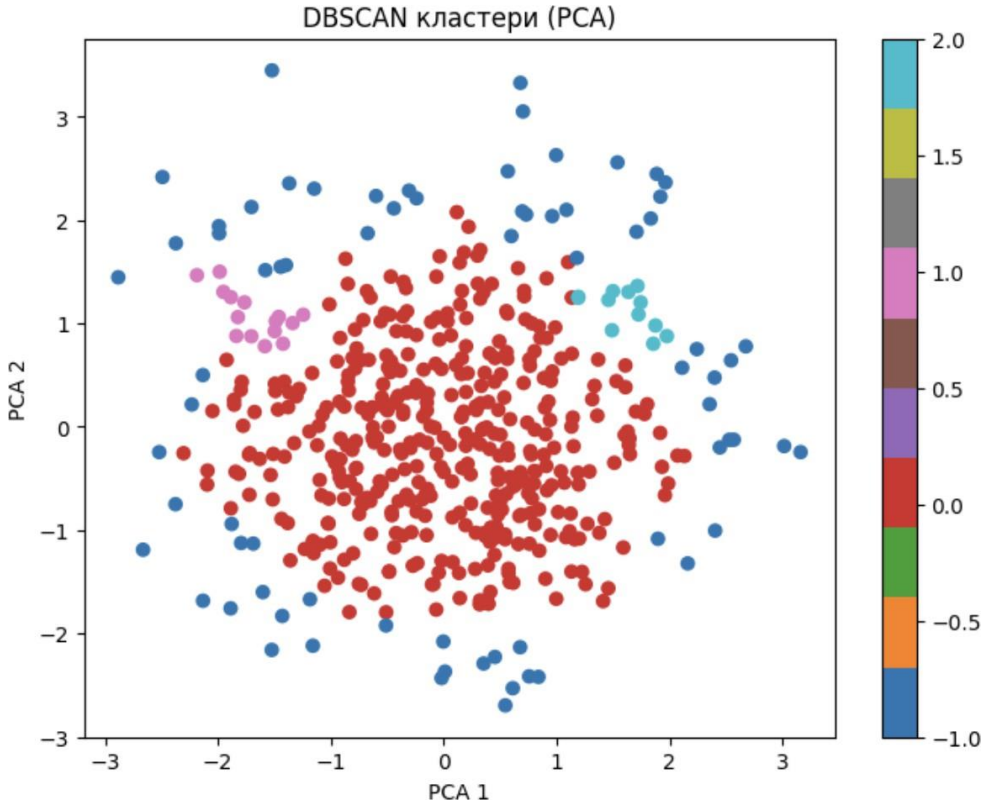

Рис. 15. Розподіл по кластерах за допомогою методу DBSCAN

Джерело: складено автором на основі даних [27]

З рисунку 15 видно, що є найбільший кластер (0), що позначений червоним кольором. Цей кластер об’єднує основну масу регулярних користувачів із стабільною поведінкою. Його чисельна перевага свідчить про чітко виражене ядро аудиторії платформи, яке має усталені патерни взаємодії та становить основну базу для монетизації, утримання та довгострокової сегментації.

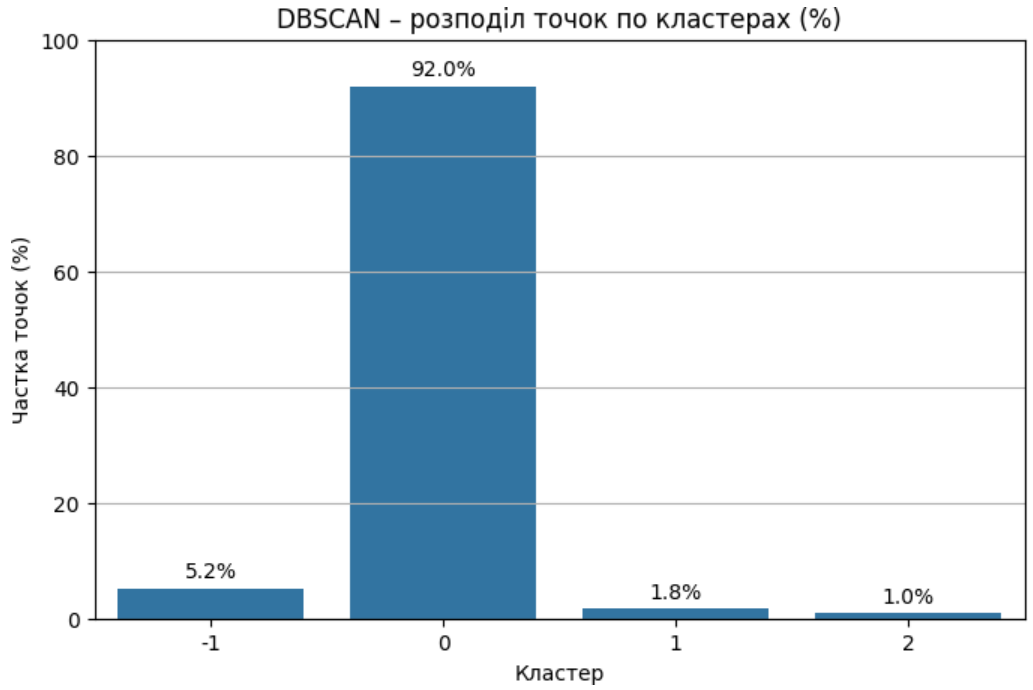

Рис. 16. Відсоткова структура кластерів, визначених методом DBSCAN

Джерело: складено автором на основі даних [27]

Рисунок 16 демонструє частотний розподіл користувачів за кластерами, виділеними алгоритмом DBSCAN. Найбільшу частку — 92% — становить 0 кластер. Цей кластер об’єднує основну масу регулярних користувачів із стабільною поведінкою. Його чисельна перевага свідчить про чітко виражене ядро аудиторії платформи, яке має усталені патерни взаємодії та становить основну базу для монетизації, утримання та довгострокової сегментації.

Кластер (-1) містить шумові точки, які не належать до жодного з основних кластерів. Хоча ця група є менш чисельною, її частка не є незначною — понад 5% — що вказує на наявність помітного прошарку користувачів з нестандартною або разовою активністю. Це може включати як технічні помилки, так і реальні сценарії нетипової взаємодії, що потребують проведення додаткового дослідження або A/B тестування.

Малі кластери 1 та 2 охоплюють близько 3% аудиторії, але мають стратегічну цінність. Кластер 1, що відповідає нішевим сценаріям (наприклад, переважне використання текстових повідомлень або специфічні часові вікна активності), може свідчити про сегмент із потенційно високим рівнем лояльності за умови точного

налаштування стратегії просування продукту. Кластер 2 вказує на нових або експериментальних користувачів, чия поведінка ще нестабільна. Цей сегмент є критичним для процесів онбордингу (процесу ознайомлення користувача з продуктом або сервісом, щоб допомогти йому швидко зрозуміти цінність і почати користування) та подальшого перетворення в активних постійних користувачів.

Комбінація аналізу просторової структури кластерів та їх кількісного складу підтверджує ефективність застосування алгоритму DBSCAN для поведінкової сегментації користувачів. Метод дозволив виявити один домінантний сегмент типових користувачів із стабільною активністю, дві невеликі нішеві групи з нетиповими сценаріями використання, а також чітко відокремити шумові точки, що представляють аномальну або разову активність. Такий підхід забезпечує глибше розуміння різних рівнів залученості аудиторії, формує основу для персоналізованого розвитку функціоналу та дає змогу здійснювати більш гнучкий і точний таргетинг — за критеріями типу активності, поведінкових патернів, місцезнаходження чи інтересів. У порівнянні з кластеризацією методом KMeans, DBSCAN виявився ефективнішим для ідентифікації як структурованих сегментів, так і аномальних випадків, що дозволяє застосовувати його для складних і динамічних сценаріїв аналізу користувацьких даних.

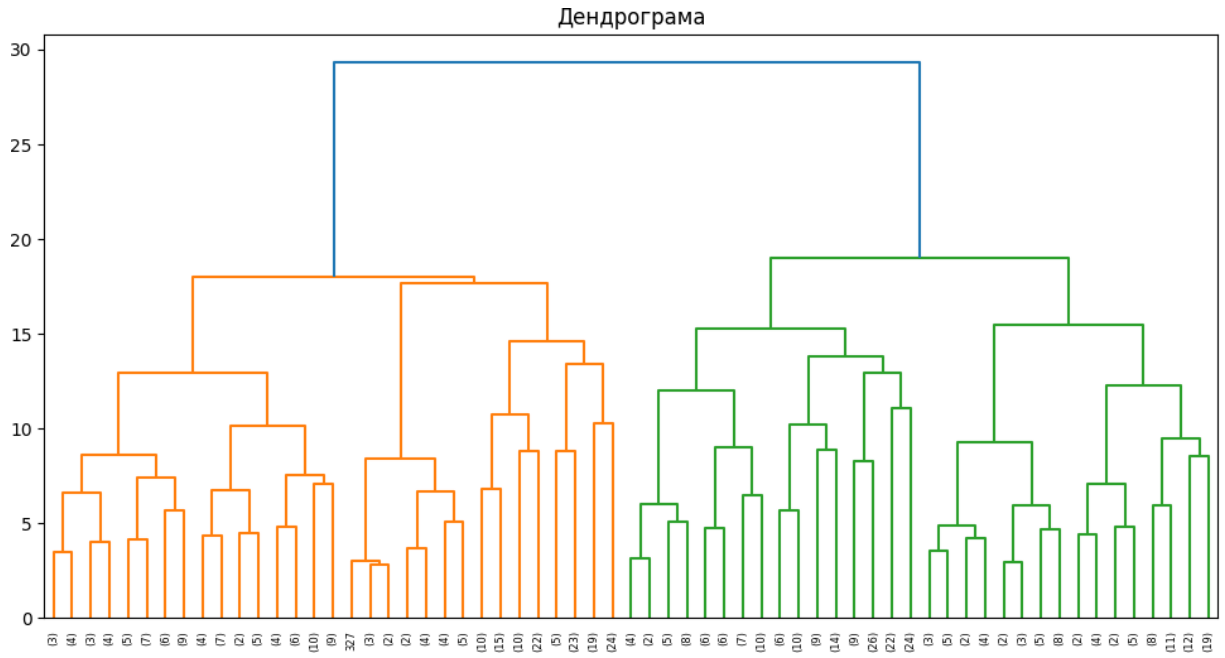

Рис. 17. Дендрограма кластерів користувачів, отримана методом ієрархічної агломеративної кластеризації (Hierarchical Agglomerative Clustering)

Джерело: розрахунки автора на основі даних [27]

На рис. 17 представлено дендрограму кластерів користувачів, отримана методом ієрархічної агломеративної кластеризації. На дендрограмі показано, як користувачі розподіляються по групах на основі схожості своїх характеристик. Така візуалізація дозволяє оцінити не лише наявність кластерної структури, а й глибше зрозуміти, наскільки користувачі різняться між собою за поведінковими або демографічними ознаками. Кожна гілка дерева відповідає окремому об’єднанню користувачів, а висота точок злиття демонструє рівень їхньої несхожості. Чим вище злиття — тим менш подібні об'єднані між собою елементи. Вертикальна вісь (Y) таким чином відображає відстань між об’єктами або групами, а горизонтальна вісь (X) показує лише порядок їхнього злиття без числового значення, що використовується для зручності читання графіка.

На побудованій дендрограмі добре видно дві великі гілки, які зливаються лише на значній висоті. Це свідчить про те, що користувачі поділяються на дві принципово різні групи, які формують незалежні сегменти. Такі відмінності можуть стосуватися, зокрема, рівня активності на платформі, частоти входів, кількості

інтересів, професійного статусу чи освітнього рівня. Наприклад, одна група може представляти більш досвідчених, постійно активних користувачів, тоді як інша — менш залучених або нових учасників. Це злиття на високому рівні означає, що між цими групами існує суттєва поведінкова або демографічна різниця, а отже, до них доцільно застосовувати різні маркетингові або комунікаційні підходи.

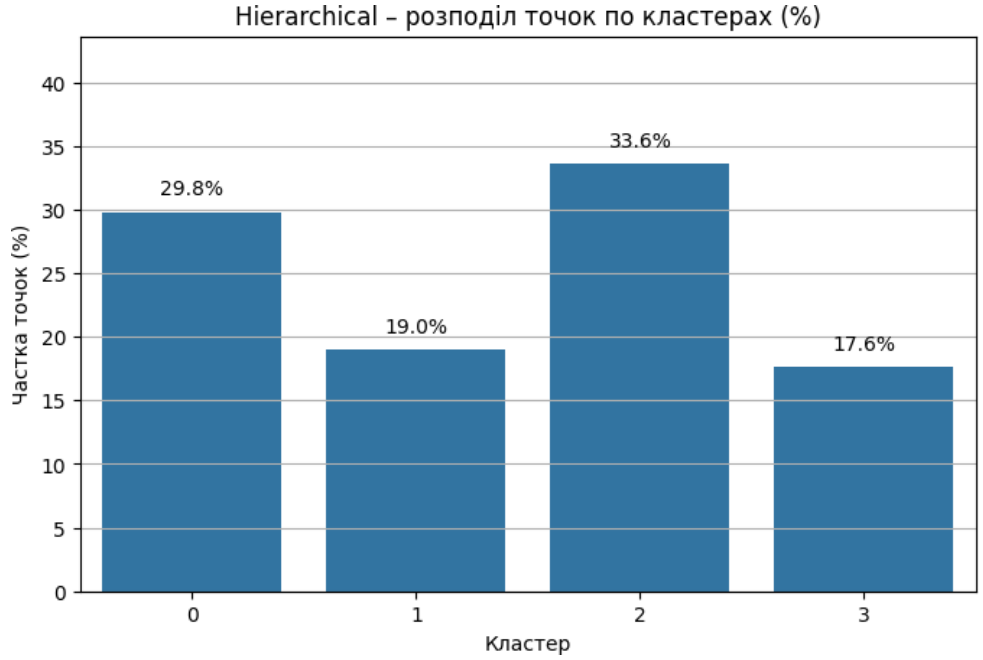

Рис. 18. Відсоткова структура кластерів, визначених методом ієрархічної кластеризації

Джерело: складено автором на основі даних [27]

На рисунку 18 представлено відсотковий розподіл точок за кластерами, отриманими в результаті ієрархічної кластеризації. Видно, що всі чотири кластери мають відносно рівномірне наповнення: найбільший кластер №2 охоплює 33,6% вибірки, тоді як найменший №3 — 17,6%. Такий баланс між групами свідчить про здатність алгоритму ефективно виявляти структурні особливості аудиторії без домінування одного сегмента. Таким чином підхід дає змогу формувати більш різнорідні поведінкові профілі, що корисно для адаптивної персоналізації сервісу та визначення стратегічно важливих підгруп користувачів.

Усередині кожної з основних груп помітно дрібніші підгілки. Зліва компактні, щільні злиття, які вказують на високу однорідність користувачів: ймовірно, вони

мають подібну активність, інтереси або шаблони використання платформи. Такий кластер зручно охоплювати уніфікованою рекламною стратегією. Натомість у правій частині дендрограми спостерігається більша вертикальна розтягнутість деяких гілок — це ознака внутрішньої різнорідності. В таких випадках доцільно дослідити підгрупи окремо або врахувати варіативність у потребах користувачів під час створення рекламного контенту, пропозицій або персоналізованих повідомлень.

Дендрограма також дозволяє оцінити оптимальну кількість кластерів для поділу аудиторії. Провівши горизонтальну лінію на певному рівні графіка, можна візуально визначити кількість кластерів як кількість гілок, які ця лінія перетинає. У цьому випадку, якщо провести лінію на середній висоті дерева, можна виокремити приблизно 4–5 підгруп — це може бути зручний і осмислений рівень деталізації для цільового сегментування користувачів.

Загалом, дендрограма демонструє як загальний поділ користувачів, так і глибшу внутрішню структуру кожного сегмента, дозволяючи виявити як чітко сформовані, однорідні групи, так і більш комплексні кластери з потенційними підсегментами. Це надає потужний інструмент для розробки точних рекламних стратегій, персоналізованих кампаній і більш гнучкого управління комунікацією з різними типами користувачів.

ВИСНОВКИ

У межах проведеного дослідження було здійснено всебічний аналіз можливостей кластеризації користувачів цифрової платформи. Виявлено, що цифрові канали поступово витіснили традиційні завдяки своїй гнучкості, широким можливостям охоплення та адаптації під конкретні аудиторії. Ефективна маркетингова стратегія потребує мультиканального підходу, який забезпечує релевантну комунікацію відповідно до поведінки та уподобань користувачів. Персоналізація стала ключовим інструментом підвищення ефективності — вона дозволяє створювати повідомлення, що враховують стиль життя, цінності та технічні характеристики аудиторії. Сегментація та поділ користувачів на кластери, у свою чергу, є основою такої персоналізації, оскільки допомагає зосередити ресурси на найперспективніших групах. Поєднання демографічних, психографічних і поведінкових ознак відкриває можливості для створення глибоко індивідуалізованого контенту та конкурентоспроможності в бізнес-середовищі.

Теоретичний аналіз засвідчує, що кластеризація та сегментація користувачів є фундаментальними елементами сучасної аналітики, особливо у сфері маркетингу, де зростає попит на персоналізований підхід до клієнтів. Застосування методів машинного навчання, таких як K-means та DBSCAN, відкриває нові можливості для виявлення прихованих патернів поведінки користувачів, що недоступні при традиційних методах аналізу. Водночас ефективне використання цих алгоритмів вимагає критичної оцінки вхідних даних, правильного підбору параметрів, оцінки якості кластеризації через метрики (Silhouette Score, DBI) і врахування обмежень кожного методу. Таким чином, кластеризація не лише є інструментом технічного аналізу, а й важливою ланкою у формуванні стратегічних рішень, спрямованих на покращення взаємодії з користувачами та підвищення ефективності бізнесу.

Також особливу увагу приділено практичному аспекту застосування кластерного аналізу, зокрема, побудові ефективних моделей сегментації користувачів для цілей персоналізації, маркетингу та стратегічного розвитку платформи. На основі вхідного набору даних, який включав соціально-демографічні

характеристики, поведінкові ознаки та інтереси користувачів, обґрунтовано доцільність включення та виключення окремих змінних, а також здійснено попередню обробку, масштабування і перетворення категоріальних даних для забезпечення коректності подальшого аналізу. Для зменшення розмірності даних і покращення візуального представлення було застосовано метод головних компонент (PCA), який дозволив зберегти понад 76% дисперсії всього набору ознак. У процесі реалізації кластерного аналізу досліджено три ключові методи: алгоритм K-Means, ієрархічну кластеризацію та алгоритм DBSCAN. Для кожного з них було здійснено підбір оптимальних параметрів з урахуванням як внутрішніх метрик якості, так і візуального аналізу простору кластерів. Окрім числових метрик, виконано візуальний аналіз кластерного простору після зниження розмірності даних, що дозволяє оцінити чіткість меж між групами та їх внутрішню структуру.. За результатами моделювання метод K-Means показав найкращу сукупність результатів: він продемонстрував найвищий коефіцієнт силуету (0.612), що свідчить про чітке відокремлення груп, а також найвищий індекс Калінські-Харабаша (43.834), який підтверджує хорошу міжкластерну віддаленість. DBSCAN виявився корисним у виявленні шуму та аномальних патернів, однак за основними метриками поступився іншим методам. Ієрархічна кластеризація дала помірні результати та добре зарекомендувала себе з точки зору інтерпретованості структури поділу, що було підтверджено побудованою дендрограмою.

Оцінка кластеризації була здійснена на основі трьох внутрішніх метрик — Silhouette Score, Davies-Bouldin Index та Calinski-Harabasz Index — що дозволило об’єктивно порівняти методи не лише за відстанями в кластерному просторі, але й за ступенем компактності та відокремленості груп. Водночас, дослідження підкреслило, що навіть хороші числові результати не гарантують практичної цінності кластеризації без якісного візуального та змістового аналізу. Саме тому результати були оцінені комплексно, з урахуванням стабільності кластерної структури, можливості її інтерпретації, а також відповідності цільовим сценаріям використання.

Важливим доповненням до числового аналізу стала візуалізація результатів кластеризації. Використовуючи PCA, було створено двовимірні проєкції, які дозволили оцінити щільність, форму та розподіл кластерів у просторі. Такі графіки дали змогу виявити як щільно згруповані сегменти, так і наявність перекриттів або шуму, що стало основою для перевірки гіпотез та подальшого профілювання користувачів.

Після побудови кластерної моделі було проведено кількісну оцінку структури кожного з отриманих кластерів. Зокрема, аналіз показав, що в методі K-Means виділено чотири основні сегменти користувачів із різним рівнем активності: найбільший кластер об’єднує активних користувачів з високою частотою взаємодії, тоді як найменший — більш пасивну або нішеву аудиторію. Ієрархічна кластеризація дала схожий результат, але з дещо більш рівномірним розподілом об’єктів між кластерами, що вказує на її здатність виявляти ширшу різноманітність поведінкових типів. У DBSCAN було виділено ядро стабільних користувачів, яке охоплює понад 90% усієї вибірки, а також ідентифіковано окрему групу з аномальними або нетиповими патернами активності, які позначені як шум. Такий підхід дозволив не лише сформувати сегменти, але й надати їм прикладну інтерпретацію на основі поведінкових характеристик аудиторії.

Отже, завдяки кластерному аналізу можна виділити, які канали комунікації доцільно використовувати для кожного сегменту, враховуючи їх поведінкові характеристики. Активним користувачам із високою частотою взаємодії рекомендовано застосовувати показувати рекламу в соціальних мережах (Facebook, Instagram, TikTok), push-повідомлення та in-app рекламу для підтримки постійного залучення. Нішевим, активним користувачам, які віддають перевагу текстовому спілкуванню, слід пропонувати комунікацію через месенджери (Telegram, Viber, WhatsApp) та email-маркетинг із таргетованою рекламою у текстових каналах. Пасивні користувачі найкраще реагують на email-розсилки, ремаркетинг у соціальних мережах і Google Ads, що стимулюють повторні візити. Для нішевої або нової аудиторії з унікальними моделями взаємодії найбільш ефективними є нативні

рекламні сітки, такі як Taboola, Outbrain та Criteo. Ці платформи дозволяють органічно інтегрувати рекламний контент у релевантні медіа, що сприяє кращій взаємодії з користувачами та підвищенню довіри до бренду. Такий диференційований підхід забезпечує оптимізацію рекламних зусиль та підвищення релевантності маркетингових повідомлень для кожного сегмента.

Проведений кластерний аналіз підтвердив практичну ефективність застосування алгоритмів машинного навчання для виявлення змістовних сегментів серед користувачів цифрової платформи. Отримані результати сформували надійну аналітичну базу для подальших рішень, пов’язаних з персоналізацією, оптимізацією користувацького досвіду та розвитком функціональності сервісу. З огляду на динамічне зростання обсягів даних та актуальність індивідуалізованих підходів у цифровому середовищі, дослідження підтверджує високу практичну цінність та перспективність використання інструментів Data Science у контексті сучасної економіки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Осипенко Н. О. Роль цифрових платформ у розвитку інтегрованих маркетингових комунікацій. *Scientific Bulletin of Kherson State University. Series Economic Sciences*. 2024. № 51. С. 63–67. URL: https://doi.org/10.32999/ksu2307-8030/2024-51-8 (дата звернення: 21.04.2025).

2. Monnappa A. The History and Evolution of Digital Marketing. Simplilearn.com. URL: https://www.simplilearn.com/history-and-evolution-of-digital-marketing-article (дата звернення: 21.04.2025).

3. Котлер Ф. Маркетинг від А до Я. 80 концепцій, які повинен знати кожен менеджер: навчальний посібник. Київ : Паблішер, 2018. 160 с.

4. Вдовічена О. Г., Дюгованець О. М., Чернова І. В. Digital-маркетинг як інструмент ефективності та конкурентоспроможності сучасного бізнесу: особливості та перспективи впровадження в Україні. *Інвестиції: практика та досвід*. 2022. Т. 2. С. 81–87.

5. Стамат В. Сегментація цільової аудиторії як важливий етап маркетингу на ринку готельно-ресторанного бізнесу. *Modern Economic*. 2022. Т. 35. С. 112–117. URL: https://doi.org/10.31521/modecon.v35(2022)-17 (дата звернення: 02.05.2025).

6. Маланчук І. І. Система управління маркетинговими кампаніями : Магістерська дисертація : 126. Київ, 2024. 102 с. URL: https://ela.kpi.ua/server/api/core/bitstreams/ccc281f8-be1e-4aaa-835f-6beb508275f6/content (дата звернення: 05.05.2025).

7. Meta Business Help Center. About Meta Pixel – Meta Platforms, Inc., 2024. URL: https://www.facebook.com/business/help/742478679120153 (дата звернення: 06.05.2025).

8. Regulation - 2016/679 - EN - gdpr - EUR-Lex. EUR-Lex – Access to European Union law. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016R0679 (дата звернення: 06.05.2025).

9. Балук Н. Р., Бойчук І. В. Аналіз маркетингової та рекламної діяльносту конкурентів засобами сервісів веб-аналітики. *Herald of Lviv University of Trade and*

Economics Economic sciences. 2024. № 74. С. 62–71. URL: https://doi.org/10.32782/2522-1205-2023-74-08 (дата звернення: 09.05.2025).

10. McGuirk M. Performing web analytics with Google Analytics 4: a platform review. *Journal of Marketing Analytics*. 2023. URL: https://doi.org/10.1057/s41270-023-00244-4 (дата звернення: 09.05.2025).

11. Wiechetek Ł., Mędrek M. Improving the university recruitment process with web analytics. *Scientific Papers of Silesian University of Technology. Organization and Management Series*. 2022. Vol. 2022, no. 158. P. 679–695. URL: https://doi.org/10.29119/1641-3466.2022.158.44 (дата звернення: 01.05.2025).

12. The analytics solution for unified customer measurement across data, content, and journeys. URL: https://business.adobe.com/products/adobe-analytics.html (дата звернення: 10.05.2025).

13. Maxpanel. URL: https://mixpanel.com/ (дата звернення: 10.05.2025).

14. Matomo. URL: https://matomo.org/ (дата звернення: 10.05.2025).

15. Федорченко А. В., Пономаренко І. В. Основні способи збору електронних адрес клієнтів у рамках реалізації email-маркетингу. *Проблеми інноваційно-інвестиційного розвитку*. 2019.

16. Rababah K. Customer Relationship Management (CRM) Processesfrom Theory to Practice: The Pre-implementation Plan ofCRM System. *International Journal of e-Education, e-Business, e-Management and e-Learning*. 2011. URL: https://doi.org/10.7763/ijeeee.2011.v1.4 (дата звернення: 11.05.2025).

17. Huang J.-J., Tzeng G.-H., Ong C.-S. Marketing segmentation using support vector clustering. Expert Systems with Applications. 2007. Vol. 32, no. 2. P. 313–317. URL: https://doi.org/10.1016/j.eswa.2005.11.028 (дата звернення: 11.05.2025).

18. Dolnicar S., Grün B. Methods in Segmentation. *Segmentation in Social Marketing*. Singapore, 2016. P. 93–107. URL: https://doi.org/10.1007/978-981-10-1835-0_7 (дата звернення: 15.05.2025).

19. Колодчак О. М. Інтелектуальний аналіз даних. Львів : Нац. ун-т «Львів. політехніка», 2013.

20. Пістунов І. М., Антонюк О. П., Турчанінова І. Ю. Кластерний аналіз в економіці. Дніпро : Нац. гірн. ун-т, 2008. 84 с.

21. Application Of K-Means Clustering For Customer Segmentation In Grocery Stores In Kenya / E. Omol et al. *International Journal of Science, Technology & Management*. 2024. Vol. 5, no. 1. P. 192–200. URL: https://doi.org/10.46729/ijstm.v5i1.1024 (дата звернення: 15.05.2025).

22. Omotosho M. Analysis of Customer Purchasing Behaviour using K-Means Clustering. *Medium.* URL: https://medium.com/@Temaydat/analysis-of-customer-purchasing-behaviour-using-k-means-clustering-5132d572b391 (дата звернення: 15.05.2025).

23. Shereenwalid. The Starbucks Effect: A Deep Dive into Customer Behaviour and Offer Preferences. *Medium.* URL: https://medium.com/@shereenwalid2003/the-starbucks-effect-a-deep-dive-into-customer-behaviour-and-offer-preferences-da6e1b84054c (дата звернення: 16.05.2025).

24. López O. G. Customer Segmentation Analysis Using K-Means: A Practical Guide. *Medium.* URL: https://medium.com/@oriolgilabertlopez/customer-segmentation-analysis-using-k-means-a-practical-guide-98f9bdf63317 (дата звернення: 16.05.2025).

25. Rahaman S. U. Data-Driven Customer Segmentation: Advancing Precision Marketing through Analytics and Machine Learning Techniques. *Journal of Artificial Intelligence, Machine Learning and Data Science*. 2022. Vol. 1, no. 1. P. 1356–1362. URL: https://doi.org/10.51219/jaimld/shafeeq-ur-rahaman/309 (дата звернення: 16.05.2025).

26. An efficient k-means clustering algorithm: analysis and implementation / T. Kanungo et al. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2002. Vol. 24, no. 7. P. 881–892. URL: https://doi.org/10.1109/tpami.2002.1017616 (дата звернення: 16.05.2025).

27. Kaggle. URL: http://kaggle.com/ (дата звернення: 03.05.2025).

28. Cui M. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*. 2020. Vol. 1, no. 1. P. 5–8.

29. Paul L. C., Suman A. A., Sultan N. Methodological analysis of principal component analysis (PCA) method. International Journal of Computational Engineering & Management. 2013. Vol. 16, no. 2. P. 32–38.

30. Shahapure K. R., Nicholas C. Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 6–9 October 2020*. 2020. URL: https://doi.org/10.1109/dsaa49011.2020.00096 (дата звернення: 16.05.2025).

31. Distance Analysis Measuring for Clustering using K-Means and Davies Bouldin Index Algorithm / A. Idrus et al. *TEM Journal*. 2022. P. 1871–1876. URL: https://doi.org/10.18421/tem114-55 (дата звернення: 17.05.2025).

32. Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta / I. F. Ashari et al. *Journal of Applied Informatics and Computing*. 2023. Vol. 7, no. 1. P. 89–97. URL: https://doi.org/10.30871/jaic.v7i1.4947 (дата звернення: 17.05.2025).

Додатки

Додаток А

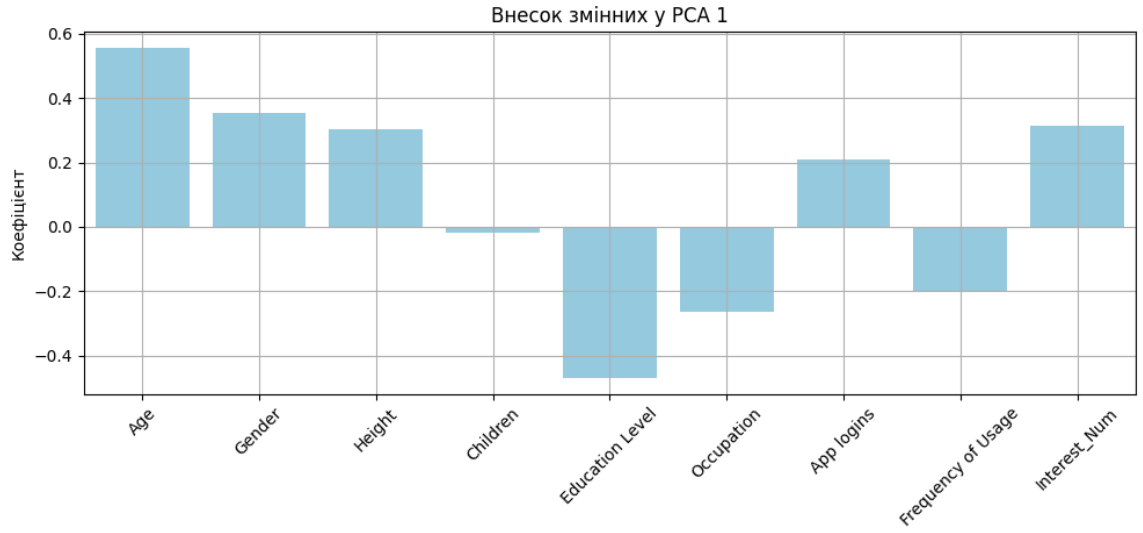

Рис. 19. Внесок змінних у компоненту PCA 1

Джерело: складено автором на основі [27]

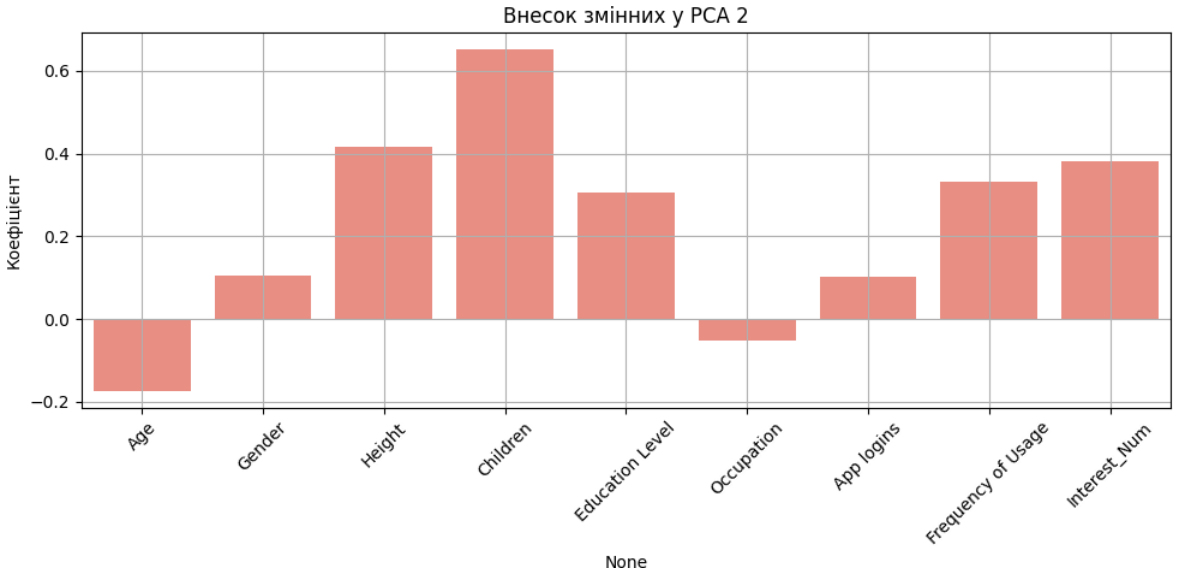

Рис. 20. Внесок змінних у компоненту PCA 2

Джерело: складено автором на основі [27]