Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

ДОСЛІДЖЕННЯ ТА РЕАЛІЗАЦІЯ МЕТОДУ ВІДСЛІДКОВУВАННЯ КОРИСТУВАЧІВ НА ВЕБСАЙТІ
(тема)

Виконав:
студент 2 курсу, групи ІНФм-23-1
Білоус А.М.
(прізвище, ініціали)

Спеціальності 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник доц. Вечірська І.Д.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Кобилін О.А.
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра Інформатики
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Білоусу Андрію Миколайовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження та реалізація методу відслідковування користувачів на вебсайті

затверджена наказом по університету від 25 листопада 2024 року № 1246Ст

2. Термін подання студентом роботи до екзаменаційної комісії 31 грудня 2024 р.

3. Вихідні дані до роботи математичні моделі кластеризації: алгоритми K-means та DBSCAN, теоретичні основи метрики силуету для оцінки кластеризації, використані програмні засоби: Jupyter Notebook, мова Python, інструменти для роботи з даними: бібліотеки scikit-learn, Pandas та Matplotlib, джерела даних: API Google Analytics для збору поведінкових метрик, методи попередньої обробки даних: нормалізація, видалення дублюючих записів, візуалізація результатів: двовимірні графіки та теплові карти.

4. Перелік питань, що потрібно опрацювати в роботі
1. Застосування алгоритмів K-means та DBSCAN для кластеризації.
2. Оцінка якості кластеризації за допомогою метрики силуету.
3. Аналіз поведінкових характеристик користувачів.
4. Практичне впровадження моделей класифікації.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) двовимірні графіки кластеризації методами K-means та DBSCAN, теплові карти розподілу користувачів між кластерами, таблиця порівняння характеристик кластерів.

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	25.11.2024	
2	Аналіз завдання, підбір літератури	26.11.24-30.11.24	
3	Аналіз літератури з досліджуваної проблеми	01.11.24-03.11.24	
4	Аналіз технічних засобів	04.11.24-05.11.24	
5	Розробка методу	05.12.24-08.12.24	
6	Програмна реалізація	08.12.24-10.12.24	
7	Оформлення пояснювальної записки	10.12.24-12.12.24	
8	Перевірка на плагіат	22.12.2024	
9	Рецензування	23.12.2024	
10	Підготовка презентації та доповіді	28.12.2024	
11	Занесення роботи в електронний архів	09.01.2025	
12	Попередній захист кваліфікаційної роботи	09.01.2025	

Дата видачі завдання 25 листопада 2024 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Вечірська І.Д.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до кваліфікаційної роботи: 69 с., 12 рис., 2 дод., 41 джерело.

ВІДСТЕЖЕННЯ КОРИСТУВАЧІВ, АНАЛІЗ ПОВЕДІНКИ, МЕТОДИ КЛАСТЕРИЗАЦІЇ, АНОМАЛЬНА ПОВЕДІНКА, АНАЛІТИКА ДАНИХ, МОДЕЛЮВАННЯ ПОВЕДІНКИ.

Об’єктом дослідження є поведінкові дані користувачів, зібрані під час їхньої взаємодії з вебсайтом. Метою дослідження є розробка ефективних методів відстеження, аналізу та кластеризації поведінкових даних користувачів для виявлення аномальних патернів та оптимізації структури вебсторінки.

У роботі використано методи збору даних, їх очищення, нормалізації та застосування алгоритмів кластеризації й виявлення аномалій. Проведено аналіз сучасних методів відстеження та їх недоліків, що дозволило розробити новий алгоритм для підвищення точності кластеризації та релевантності аналітики. В рамках дослідження запропоновано концептуальну логіку процесу відстеження, розроблено блок-схеми алгоритмів та програмну реалізацію.

Результатом дослідження є створення системи збору, очищення та кластеризації поведінкових даних користувачів для виявлення аномалій у поведінці, що сприяє покращенню аналітики вебсайту та підвищенню задоволеності користувачів.

USER TRACKING, BEHAVIOR ANALYSIS, CLUSTERING METHODS, ANOMALOUS BEHAVIOR, DATA ANALYTICS, BEHAVIORAL MODELING.

The object of the research is user behavior data collected during interaction with the website. The aim of the research is to develop effective methods for tracking, analyzing, and clustering user behavior data to identify anomalous patterns and optimize website structure.

The work involves methods for data collection, cleaning, normalization, and the application of clustering and anomaly detection algorithms. An analysis of current tracking methods and their limitations was conducted, leading to the development of a new algorithm to enhance clustering accuracy and analytic relevance. The study proposes a conceptual tracking logic, algorithmic flowcharts, and a software implementation.

The result of the research is a system for collecting, cleaning, and clustering user behavior data to detect anomalies in behavior, supporting improved website analytics and enhancing user satisfaction.

ЗМІСТ

Вступ 7
1. Аналіз предметної області 10
1.1 Опис та аналіз предметної області 10
1.2 Огляд існуючих систем 12
1.3 Постановка задачі 16
2 Математична модель відслідковування користувачів 17
2.1 Отримання даних 17
2.2 Кластеризація користувачів 18
2.3 Аналіз метода K-means 19
2.4 Аналіз метода DBSCAN 22
2.5 Застосування результатів кластеризації 25
2.6 Конфіденційність і безпека 27
3 Вибір підходів та технологій для розробки 30
3.1 Обґрунтування вибору технологій і методів реалізації 30
3.2 Середовище розробки Jupyter Notebook 32
3.3 Основні відомості про мову програмування Python 35
3.4 Бібліотеки для розробки: Scikit-learn, Pandas, Matplotlib, Seaborn 37
3.5 Google Analytics API 40
3.6 Результати та огляд створеного алгоритму 42
3.7 Перспективи дослідження 56
Висновки 58
Перелік джерел посилання 60
Додаток А 64
Додаток Б 68

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

GDPR – General Data Protection Regulation (загальний регламент захисту даних)

VPN – Virtual Private Network (віртуальна приватна мережа)

Boxplot – стандартизований спосіб відображення набору даних на основі підсумку

API – Application Programming Interface (спосіб взаємодії комп'ютерних програм між собою)

K-means – алгоритм машинног навчання, що вирішує задачу кластеризації

DBSCAN – Density-based spatial clustering of applications with noise (просторова кластеризація додатків із шумом на основі щільності)

PCA – Principal component analysis (аналіз головних компонент)

JSON – JavaScript Object Notation (нотація об’єктів JavaScript)

Cookies – невеликий набір даних, що надсилається веб-сервером і зберігається на комп'ютері користувача

ВСТУП

У сучасному цифровому середовищі ефективність роботи вебсайтів є однією з ключових умов їхнього успіху. Із зростанням конкуренції в інтернет-просторі та швидкими темпами розвитку технологій компанії все більше орієнтуються на пошук нових, інноваційних підходів для покращення користувацького досвіду, підвищення залученості відвідувачів та утримання їх на своїх платформах. За умов постійно змінюваного попиту та швидкої адаптації користувачів до нових цифрових інструментів, необхідно постійно вдосконалювати не лише функціональність сайту, а й способи взаємодії з користувачами. Одним із найбільш ефективних підходів у досягненні цієї мети є аналіз та розуміння поведінки користувачів, що дає можливість не тільки оцінити їхні переваги та вподобання, але й спрогнозувати їхні подальші дії.

Аналіз поведінки користувачів допомагає власникам сайтів оптимізувати навігацію, адаптувати контент до конкретних потреб різних сегментів аудиторії, а також підвищити рівень задоволеності та зручності користування сайтом. Крім того, це дозволяє значно знижувати відсоток відмов і забезпечувати більш ефективне використання рекламних бюджетів, оскільки отримані дані можна використовувати для більш точного таргетування реклами і створення персоналізованих пропозицій.

Одним із основних інструментів досягнення цих цілей є системи відстеження дій користувачів, які дозволяють збирати та аналізувати дані про їхню поведінку на сайті. Такі системи не лише фіксують елементарні дії, такі як кліки, перегляд сторінок чи заповнення форм, а й здатні збирати більш комплексні дані про час перебування на сайті, маршрут користувача, зміни в його поведінці під час сесії, а також інші важливі метрики. Збір таких даних дає змогу не лише оцінити ефективність окремих елементів інтерфейсу, а й здійснювати глибший аналіз взаємодії з контентом, що дозволяє

відстежувати такі важливі фактори, як інтерес до різних тем або потенційні бар'єри в процесі конверсії.

Аналіз поведінкових характеристик користувачів відкриває широкі можливості для персоналізації контенту, покращення маркетингових стратегій та підвищення ефективності цільових рекламних кампаній. Наприклад, виявлення патернів поведінки дозволяє виділити найбільш популярні сторінки та елементи інтерфейсу, що привертають увагу, і виявити найбільш перспективні напрямки для покращення контенту. Це також дає змогу ідентифікувати потенційні проблеми на сайті, такі як проблеми з навігацією, довгий час завантаження або непотрібні елементи, що можуть відволікати користувачів. В результаті це дозволяє власникам сайту вчасно реагувати на негативні фактори, що знижують ефективність роботи сайту.

Проте процеси збору та аналізу даних не є статичними, вони постійно еволюціонують, що зумовлено швидким розвитком технологій, змінними вимогами до безпеки та конфіденційності даних, а також новими трендами в поведінці користувачів. Існуючі підходи до збору та аналізу даних мають свої обмеження, і тому необхідно розробляти нові методи та інструменти для отримання більш детальної, точнішої та кориснішої інформації про взаємодію користувачів з вебсайтами. Так, з’являється потреба в застосуванні більш сучасних алгоритмів обробки даних, таких як машинне навчання, для покращення точності та гнучкості аналітики.

Ця робота присвячена дослідженню та реалізації методів відстеження користувачів на вебсайтах, зокрема розробці системи збору й аналізу даних, що охоплюють ключові етапи взаємодії користувачів із вебсторінкою. У межах дослідження було проведено комплексний аналіз існуючих підходів до відстеження користувачів і виявлено основні недоліки існуючих систем. Ці результати дозволяють запропонувати вдосконалений метод для підвищення точності, релевантності та масштабованості зібраної інформації. Особливу увагу було приділено інтеграції алгоритмів машинного навчання, зокрема методів кластеризації, для групування користувачів за їхніми

поведінковими характеристиками, що дозволяє будувати точніші профілі користувачів.

Актуальність дослідження полягає в розробці концептуальної логіки та математичних моделей, які забезпечують коректне та ефективне відстеження, обробку та кластеризацію користувацьких даних. Інтеграція сучасних методів машинного навчання дозволяє не лише підвищити точність збору даних, а й здійснити глибший аналіз поведінки користувачів на сайті. Також особлива увага приділяється виявленню аномалій у поведінці користувачів, що дозволяє швидко виявляти проблеми та оптимізувати процеси, знижуючи відсоток відмов і покращуючи взаємодію з потенційними клієнтами.

Таким чином, результати дослідження мають важливе практичне значення як для бізнесу, так і для розвитку інформаційних технологій у сфері цифрового маркетингу. Вони сприяють розробці ефективних рішень для аналізу та управління поведінкою користувачів на вебсайтах, що дозволить компаніям розробляти гнучкі маркетингові стратегії, ефективно керувати взаємодією з аудиторією та підвищити загальну ефективність їхніх онлайн-ресурсів.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Опис та аналіз предметної області

Зважаючи на стрімкий розвиток технологій, зростаючі вимоги до персоналізації та безпеки даних, вебсайти стають невід'ємною частиною сучасного бізнесу. Вони відіграють ключову роль у залученні та утриманні клієнтів, оскільки дозволяють компаніям адаптувати свої стратегії до індивідуальних потреб користувачів. Вебресурси дозволяють здійснювати глибокий аналіз взаємодії користувачів з контентом, продуктами та послугами, що пропонуються на сайті. Це надає бізнесу важливу інформацію для поліпшення користувацького досвіду, підвищення ефективності маркетингових кампаній та вдосконалення продуктів. За допомогою відстеження активності користувачів можна виявити патерни поведінки, що дозволяє створювати персоналізовані стратегії, які сприяють збільшенню рівня задоволення від взаємодії з вебсайтом і, як наслідок, покращенню конверсії.

Персоналізація стала однією з основних тенденцій сучасних онлайн-платформ. Користувачі все більше очікують, що інтерфейс сайту буде адаптований до їхніх уподобань, що дозволяє компаніям не лише залучати нових клієнтів, але й утримувати існуючих. Це включає в себе персоналізовані рекомендації товарів, адаптацію контенту під інтереси користувача, а також інші функції, які покращують взаємодію з сайтом. Проте, незважаючи на безліч переваг, методи відстеження поведінки користувачів стикаються з численними труднощами та викликами.

Однією з найбільших проблем є постійна еволюція технологій, спрямованих на блокування відстеження. Використання cookies, яке ще кілька років тому було стандартом для збору даних, нині стикається з труднощами через активну блокування цієї технології браузерами та антивірусними програмами. Це обмежує можливості для збору даних про

активність користувачів, що негативно позначається на ефективності маркетингових кампаній. Для подолання цих обмежень з'явилися більш складні методи, такі як browser fingerprinting, який дозволяє створити унікальний відбиток кожного користувача на основі характеристик його браузера та пристрою. Однак ці технології викликають серйозні етичні питання, оскільки часто вони реалізуються без явної згоди користувача, що може порушувати його право на конфіденційність.

Інша велика проблема – це дотримання вимог законодавства, що стосуються захисту персональних даних. У світлі введення нових стандартів, таких як Загальний регламент щодо захисту даних (GDPR) в Європейському Союзі, організації змушені переглядати свої методи збору інформації. Компанії повинні гарантувати, що будь-який збір даних здійснюється прозоро і відповідно до законодавчих вимог. Крім того, відповідність до таких норм як GDPR передбачає, що користувачі повинні мати чітке розуміння того, які дані збираються, як вони використовуються, а також можливість відмовитися від їхнього збору або видалення.

Це ставить бізнеси перед необхідністю перегляду своїх підходів до збору даних, забезпечуючи високий рівень прозорості та етики в процесі взаємодії з користувачами. Оскільки захист приватності став пріоритетом для споживачів, компанії повинні знайти баланс між використанням даних для оптимізації маркетингових кампаній і забезпеченням дотримання стандартів конфіденційності. Більше того, етичне використання даних стало важливим аспектом розвитку довіри між користувачами та бізнесами, що є необхідною умовою для успішної взаємодії в довгостроковій перспективі.

У результаті, тема дослідження та вдосконалення методів відстеження користувачів на вебсайтах є не лише актуальною, а й надзвичайно важливою для розвитку сучасного онлайн-бізнесу. Цей процес дозволяє не тільки покращити користувацький досвід, але й допомагає бізнесам адаптувати свої стратегії до індивідуальних потреб клієнтів, що має безпосередній вплив на їхню конкурентоспроможність на ринку. Водночас, важливим аспектом є

постійне вдосконалення технологій збору даних, що відповідають сучасним вимогам до конфіденційності та етики. Сучасні дослідження в цій галузі повинні шукати нові способи збору, обробки та зберігання даних, щоб не тільки підвищити ефективність бізнесу, але й задовольнити потреби користувачів і забезпечити дотримання вимог законодавства, що регулює приватність і захист персональних даних.

1.2 Огляд існуючих систем

Існуючі дослідження у сфері відстеження користувачів на вебсайтах висвітлюють різноманітні методи, які використовуються для збору інформації про поведінку відвідувачів. Одним з перших та найбільш поширених методів стало використання файлів cookies, що дозволяє зберігати певну інформацію про користувача прямо в його браузері. Cookies широко застосовуються для ідентифікації користувачів при повторних відвідуваннях, зберігання налаштувань та історії покупок. Проте, у зв’язку з посиленням вимог щодо конфіденційності та підвищенням обізнаності користувачів, дедалі частіше зустрічаються випадки блокування cookies або їх автоматичного видалення, що обмежує ефективність цього методу.

Іншим методом, який почав набирати популярності, є використання browser fingerprinting – технології, яка дозволяє збирати унікальні технічні параметри пристрою, такі як тип браузера, розширення екрану, версія операційної системи, встановлені шрифти та інші характеристики. Комбінація цих параметрів утворює своєрідний «відбиток» пристрою, що дозволяє ідентифікувати користувача навіть без використання cookies. Fingerprinting здатен забезпечити більш стійке і точне відстеження, проте викликає суттєві питання стосовно приватності, оскільки користувачі часто не усвідомлюють, що їх можна відстежувати без їхньої згоди.

Третім поширеним підходом є використання IP-ідентифікації, що базується на відстеженні IP-адреси відвідувача. Цей метод є простим і дозволяє загалом зрозуміти географічне розташування користувачів, але не може забезпечити високу точність для повторних відвідувачів, оскільки багато користувачів мають динамічні IP-адреси або використовують VPN.

У сучасних умовах розвитку законодавства щодо захисту персональних даних, таких як GDPR у Європейському Союзі, вимоги до методів збору та обробки даних значно посилилися. Зокрема, від компаній вимагається отримання згоди користувачів на збір даних і забезпечення права на забуття. Це суттєво обмежує застосування багатьох традиційних методів відстеження, що зумовлює потребу у пошуку нових рішень, які були б ефективними та водночас відповідали б вимогам конфіденційності.

Таким чином, аналіз існуючих підходів показує, що кожен із методів має як переваги, так і обмеження. Cookies є легкими у впровадженні, але залежними від згоди користувача та підлягають блокуванню; fingerprinting забезпечує більш надійне відстеження, проте порушує приватність; IP-ідентифікація є обмеженою через технічні особливості. Відповідно, виникає потреба у вдосконаленні існуючих методів або розробці нових, які забезпечували б належний баланс між ефективністю відстеження та захистом приватності.

За приклад можна узяти системи Google Analytics та Hotjar. Кожна з систем надає унікальні можливості для аналізу. Їх розгляд також дозволяє отримати уявлення про різні підходи до відстеження та збору даних.

Google Analytics є одним з найпоширеніших інструментів для вебаналітики. Він дозволяє власникам сайтів збирати та аналізувати великий обсяг даних про взаємодію користувачів із сайтом. Google Analytics надає статистику про:

– основні показники трафіку: кількість відвідувачів, сесій, джерела трафіку (органічний, платний, прямий, реферальний) тощо;

– демографічні та географічні дані: інформація про вік, стать користувачів, їх місцезнаходження;

– дії користувачів: поведінка на різних сторінках, переходи між сторінками, глибина прокрутки, кліки на кнопки та інші елементи;

– показники взаємодії: середня тривалість сесії, частота повернень, показник відмов (bounce rate) тощо.

Google Analytics ефективний для відстеження макрорівневих патернів поведінки користувачів і може використовуватися як для невеликих сайтів, так і для великих вебплатформ. Однак, хоча цей інструмент є надзвичайно потужним у сфері аналітики, він надає лише обмежені можливості для відстеження індивідуальних сесій або дій, які детальніше розглядає Hotjar.

Hotjar є інструментом аналітики, який фокусується на поведінкових аспектах користувачів і допомагає зрозуміти їхні дії на мікрорівні. Hotjar пропонує:

– теплові карти (heatmaps) – візуалізують, де саме користувачі клікають, як пересуваються сторінкою, які частини сторінки привертають найбільшу увагу;

– записи сесій користувачів – дозволяють відтворити процес взаємодії користувача з сайтом, що дає детальніше уявлення про його поведінку;

– опитування та форми зворотного зв’язку – збір безпосередніх відгуків від користувачів, який допомагає виявити проблеми у функціоналі чи структурі сторінки;

– аналіз поведінки в режимі реального часу – Hotjar пропонує інтерактивний підхід до відстеження, який можна налаштувати для кожної сторінки окремо.

Hotjar доповнює Google Analytics, дозволяючи не тільки зрозуміти, що користувачі роблять на сайті, але й чому вони це роблять. Його інструменти дозволяють зосередитися на зручності та ефективності користувацького

інтерфейсу, що особливо корисно для оптимізації дизайну і покращення користувацького досвіду.

Після огляду обох інструментів можна зазначити, що ці системи розставляють різні акценти та пріоритети. Google Analytics підходить для аналізу великих обсягів загальних статистичних даних, що дозволяє виявляти ключові демографічні та поведінкові патерни. Hotjar, натомість, орієнтований на детальний аналіз поведінки користувачів, надаючи візуальні інструменти для розуміння, як саме відвідувачі взаємодіють із сайтом.

Розуміння функцій обох систем допомагає власникам сайтів інтегрувати їхні здібності щодо аналізу для створення комплексної картини поведінки користувачів, що є важливим аспектом у розробці ефективної системи відстеження.

1.3 Постановка задачі

Метою цієї роботи є дослідження та розробка ефективного методу відстеження користувачів на вебсайті, який би забезпечував високу точність і надійність збору даних, одночасно відповідаючи сучасним вимогам щодо конфіденційності та відповідності законодавчим нормам. У контексті швидкого розвитку технологій збору інформації та зростаючої стурбованості користувачів щодо приватності, ця робота прагне знайти баланс між інформативністю та етичною обґрунтованістю методів відстеження.

Для досягнення цієї мети у роботі поставлено такі завдання:

– проаналізувати існуючі методи відстеження користувачів;

– визначити основні проблеми та обмеження поточних рішень;

– розробити метод або підхід до відстеження, що враховує сучасні вимоги конфіденційності;

– реалізувати запропонований метод та перевірити його ефективність на практиці;

– оцінити ефективність розробленого методу у порівнянні з існуючими.

Таким чином, ця робота спрямована на вирішення актуальної проблеми сучасного веб-аналітики – розробку методу, який дозволяє отримувати корисну інформацію про поведінку користувачів без порушення їхніх прав на приватність та у відповідності до етичних і законодавчих норм.

2 МАТЕМАТИЧНА МОДЕЛЬ ВІДСЛІДКОВУВАННЯ КОРИСТУВАЧІВ

2.1 Отримання даних

Отримання даних є одним з найважливіших етапів процесу відстеження та кластеризації користувачів на веб-сайті. Цей метод дозволяє глибше зрозуміти поведінкові характеристики аудиторії, виділяючи ключові шаблони, які сприяють подальшій оптимізації ресурсу та його адаптації до потреб відвідувачів. Основний підхід полягає у зборі та аналізі дій користувачів з наступною класифікацією їх у групи за схожими поведінковими ознаками.

Процес збору даних починається з реєстрації взаємодій користувачів із сайтом. Фіксуються такі параметри, як кількість кліків, перегляди сторінок, час перебування на кожній сторінці, переміщення миші та загальний час активності. Усі ці дані збираються в анонімному форматі без використання ідентифікаторів, що можуть порушити конфіденційність. Таким чином, процес збору відповідає вимогам Загального регламенту захисту даних (GDPR), гарантуючи безпеку та прозорість обробки інформації.

Зібрані дані проходять кілька етапів обробки для забезпечення їх якості та підготовки до подальшого аналізу. Першим кроком є виявлення аномалій, що можуть виникнути внаслідок технічних помилок або атипової поведінки користувачів. Після цього дані нормалізуються, щоб привести всі характеристики до єдиного масштабу. Такий підхід дозволяє зменшити вплив екстремальних значень і підвищити точність результатів аналізу.

Наступним важливим кроком є формування векторів характеристик для кожного користувача. Ці вектори включають широкий набір параметрів, таких як загальна кількість відвіданих сторінок, частота взаємодії з контентом, тривалість сесій, час активності та інші показники. Використання

векторів дозволяє легко порівнювати поведінку різних користувачів і виявляти схожості або відмінності в їхніх діях. Такі підходи є основою для кластеризації, оскільки вони дозволяють структурувати інформацію таким чином, щоб алгоритми могли ефективно групувати користувачів за схожими ознаками.

Інформація, отримана в результаті цього етапу, стає основою для подальшого аналізу, включаючи застосування алгоритмів кластеризації, таких як K-means та DBSCAN. Вона також використовується для побудови моделей прогнозування поведінки користувачів, що допомагає вдосконалити механізми персоналізації контенту та підвищити рівень залученості аудиторії. Завдяки цьому підходу дослідження поведінкових характеристик набуває комплексного характеру, забезпечуючи глибоке розуміння потреб і вподобань користувачів веб-сайту.

2.2 Кластеризація користувачів

Для кластеризації користувачів вебсайту можна застосовувати різні методи машинного навчання, які дозволяють групувати користувачів за схожістю їх поведінкових характеристик. Одними з найбільш поширених алгоритмів для цього є K-means, DBSCAN та інші. Кластеризація має на меті розподіл користувачів на групи, що мають схожі патерни поведінки або рівень взаємодії з контентом. У цьому контексті користувачів можна поділити на кілька категорій в залежності від їх активності на сайті.

Активні користувачі – це користувачі, які регулярно відвідують сайт, проводять значний час на сторінках та активно взаємодіють з контентом, наприклад, переглядають кілька сторінок, коментують або здійснюють покупки. Вони можуть бути постійними клієнтами або потенційно зацікавленими в нових пропозиціях. Для них можна створювати персоналізовані рекомендації та спеціальні акції.

Пасивні користувачі – користувачі, які рідко відвідують сайт, їх сесії короткі, а кількість взаємодій мінімальна. Таких користувачів можна класифікувати як потенційно неактивних. Для них варто впроваджувати стратегії залучення, наприклад, за допомогою персоналізованих email-розсилок або акцій.

Нові користувачі – люди, які вперше відвідують сайт і, ймовірно, потребують спеціальних інструкцій або гідів для ознайомлення з контентом. Їх поведінка може бути менш передбачуваною, оскільки вони ще не сформували постійні патерни взаємодії з вебсайтом. Для цієї групи корисно впроваджувати інтерфейси, які спрощують навігацію та пропонують орієнтуючі матеріали.

Постійні клієнти – користувачі, які регулярно взаємодіють з вебсайтом і мають певні вподобання або звички, що робить їх поведінку більш передбачуваною. Такі користувачі можуть мати специфічні інтереси або потреби, які можна виявити через їхні попередні взаємодії з сайтом. Для них можна створювати індивідуальні пропозиції, спеціальні знижки чи інші стимули для підтримки лояльності.

За допомогою кластеризації можна створити різні сегменти користувачів, що дозволяє сайтам не лише краще розуміти свою аудиторію, але й ефективніше взаємодіяти з кожним сегментом через персоналізовані маркетингові стратегії. Використовуючи алгоритми, такі як K-means, можна автоматично класифікувати користувачів у ці групи, що спрощує подальше прийняття рішень щодо маркетингових кампаній або змін у контенті сайту.

2.3 Аналіз метода K-means

Метод K-means є одним із найпопулярніших і найефективніших алгоритмів для кластеризації даних. Кластеризація, у свою чергу, є важливим процесом поділу набору даних на кілька груп або кластерів, кожен з яких

містить елементи, що демонструють подібні характеристики. Основна мета цього підходу полягає у формуванні кластерів таким чином, щоб елементи в межах одного кластеру були максимально схожими, тоді як елементи різних кластерів мали суттєві відмінності між собою.

Алгоритм K-means працює за кілька етапів. Спочатку необхідно визначити кількість кластерів, що позначається як K. Це число є гіперпараметром, який встановлюється перед початком роботи алгоритму. Після цього випадковим чином обираються початкові центри кластерів, які називають центроїдами. Існують також методи вдосконаленої ініціалізації, такі як K-means++, що покращують вибір початкових центроїдів і зменшують імовірність отримання локальних мінімумів. Центроїди є тимчасовими представниками кластерів, навколо яких групуються дані під час виконання алгоритму.

Наступним етапом є призначення кожного елемента до одного з кластерів на основі того, який центр кластера найближчий до елемента. Для цього зазвичай використовують евклідову відстань (2.1).

$$d(x, c) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2} \quad , \tag{2.1}$$

де x – точка;

c – центр кластера.

Кожен елемент призначається до кластера, центр якого знаходиться на мінімальній відстані.

Після призначення всіх елементів до кластерів відбувається етап оновлення центроїдів. Новий центр кожного кластеру обчислюється як середнє значення всіх точок, що належать цьому кластеру (2.2).

$$c_k = \frac{1}{|C_k|} \sum_{x \in C_k} x \quad , \tag{2.2}$$

де C_k – множина точок, що належать кластеру k;

c_k – новий центроїд.

Процес призначення елементів до кластерів і оновлення центрів повторюється до тих пір, поки центроїди не перестануть змінювати свої позиції, що вказує на досягнення збіжності. Завдяки своїй простоті та ефективності метод K-means широко використовується в різних галузях, таких як аналіз поведінки користувачів на вебсайтах, сегментація клієнтів, класифікація текстів, а також машинне навчання.

Попри свою популярність, метод K-means має певні обмеження. Одним із головних недоліків є необхідність заздалегідь визначити кількість кластерів. У ситуаціях, коли структура даних є невідомою, вибір оптимального значення K може бути складним завданням. Для вирішення цієї проблеми часто використовують додаткові методи, такі як метод "ліквідного коліна" (Elbow Method). Цей метод допомагає знайти оптимальне значення K шляхом аналізу зміни варіативності в даних. На графіку залежності дисперсії від кількості кластерів можна виявити точку, після якої подальше збільшення K призводить до незначного зменшення внутрішньокластерної варіації.

Ще одним недоліком методу є його чутливість до початкової ініціалізації центрів. У разі невдалого вибору початкових центроїдів алгоритм може застрягти в локальному мінімумі, не досягнувши глобально оптимального рішення. Для пом'якшення цього ефекту часто застосовують кілька запусків алгоритму з різними початковими умовами та обирають найкращий результат. Крім того, K-means може мати труднощі з обробкою даних, які мають складні форми кластерів, такі як витягнуті еліптичні або

перекриті структури. В таких випадках результати можуть бути неточними, оскільки алгоритм припускає сферичну форму кластерів.

Попри ці обмеження, метод K-means залишається надзвичайно корисним для аналізу великих обсягів даних і пошуку закономірностей у них. Він знайшов широке застосування в економіці, маркетингу, біології та навіть у соціальних науках. Наприклад, у сфері цифрового маркетингу K-means допомагає сегментувати клієнтів на основі їхньої поведінки, що дозволяє створювати персоналізовані рекламні кампанії. В аналізі текстів алгоритм використовується для кластеризації документів за темами, а в біології – для ідентифікації генетичних груп і класифікації біологічних видів.

Таким чином, метод K-means є потужним інструментом кластеризації, який забезпечує ефективний аналіз і групування даних. Завдяки своїй простоті, швидкості та універсальності, він продовжує залишатися одним із основних алгоритмів у сфері машинного навчання та аналізу даних, попри деякі обмеження, які можна подолати за допомогою вдосконалених методів ініціалізації та оптимізації кількості кластерів.

2.4 Аналіз метода DBSCAN

Алгоритм DBSCAN є одним із найпопулярніших методів кластеризації, який заснований на аналізі щільності даних. Його основна перевага полягає в тому, що він не вимагає попереднього визначення кількості кластерів, як це потрібно в методах на кшталт K-means. Це робить його особливо корисним для роботи з даними, структура яких невідома заздалегідь або має складні форми. DBSCAN здатний виділяти кластери неправильної форми та різної щільності, а також ефективно обробляти шумові дані, які не належать до жодного кластера. Така гнучкість робить алгоритм придатним для широкого спектра завдань, пов’язаних з аналізом даних.

Основою роботи DBSCAN є використання двох параметрів: ε (епсилон), який визначає радіус пошуку сусідів для кожної точки, та MinPts, що вказує мінімальну кількість точок, необхідних для утворення кластера. Вибір цих параметрів є критично важливим, оскільки вони безпосередньо впливають на результати кластеризації. Параметр ε визначає, наскільки близько точки повинні знаходитися одна до одної, щоб вважатися сусідами, тоді як MinPts визначає, скільки сусідів повинна мати точка, щоб стати ядром кластера. Завдяки цим параметрам алгоритм може адаптуватися до різних структур даних і знаходити кластери навіть у наборах із нерівномірним розподілом.

Процес роботи DBSCAN починається з аналізу кожної точки в наборі даних. Алгоритм перевіряє, чи є точка основною, тобто чи має вона достатньо сусідів у заданому радіусі ε. Якщо точка задовольняє цю умову, вона стає ядром кластера, навколо якого починається групування інших точок. Ті точки, що знаходяться в межах ε від основної, додаються до кластеру, і процес триває, поки не буде розширено всі можливі групи. Таким чином, кластер розростається, поки не будуть включені всі точки, які відповідають критеріям щільності.

Окрім основних точок, DBSCAN також працює з прикордонними точками. Це такі точки, які самі не є ядрами, але розташовані поблизу кластерів і можуть бути до них приєднані. Прикордонні точки сприяють збільшенню кластерів, однак самі не формують нові групи. Якщо ж точка не має достатньої кількості сусідів і не належить до жодного кластера, вона вважається шумовою. Обробка шуму є важливою перевагою DBSCAN, оскільки дозволяє автоматично ідентифікувати аномальні або неструктуровані дані, що не відповідають жодному кластеру.

Для обчислення відстаней між точками алгоритм найчастіше використовує евклідову відстань, проте залежно від специфіки даних можуть застосовуватися й інші метрики відстані, такі як мангеттенська чи косинусна подібність. Гнучкість у виборі метрики дозволяє адаптувати DBSCAN для

різноманітних наборів даних, включаючи географічні або текстові дані, де стандартна евклідова відстань може бути менш ефективною.

Незважаючи на численні переваги, алгоритм DBSCAN має і певні обмеження. Одним із них є складність вибору оптимальних параметрів ε та MinPts, особливо для наборів даних зі змінною щільністю. Якщо параметри обрані неправильно, алгоритм може створити надмірну кількість кластерів або ж, навпаки, не виявити жодного. Іншою проблемою є зниження продуктивності при роботі з високовимірними даними, де обчислення відстаней стає більш складним і менш точним через ефект розрідження простору. Проте ці обмеження можна частково обійти за допомогою попереднього аналізу даних і тестування різних параметрів алгоритму.

Завдяки своїй гнучкості та здатності обробляти складні структури даних, DBSCAN широко застосовується в різних галузях. У географічному аналізі він використовується для виявлення густонаселених регіонів або аномальних точок у просторі. У маркетингових дослідженнях цей метод допомагає сегментувати клієнтів за схожими поведінковими характеристиками, навіть якщо групи мають нечіткі межі. Крім того, DBSCAN активно використовується в комп’ютерному зорі для ідентифікації об’єктів на зображеннях, у фінансових аналітичних системах для виявлення шахрайських транзакцій і в системах виявлення аномалій у різних процесах.

Загалом, DBSCAN залишається одним із найефективніших алгоритмів для аналізу даних, де структура кластерів є складною або деякі дані містять шуми. Його здатність автоматично визначати кількість кластерів, працювати з даними нерівномірної щільності та виявляти аномалії робить його потужним інструментом у сучасній науці про дані та машинному навчанні. У нашому дослідженні DBSCAN дозволив не лише ефективно розділити користувачів за рівнем активності та поведінковими характеристиками, але й виявити відхилення, які можуть свідчити про специфічні тренди чи аномальні шаблони поведінки. Це підкреслює важливість даного методу для досліджень, орієнтованих на аналіз складних і нерегулярних наборів даних.

2.5 Застосування результатів кластеризації

Результати кластеризації є потужним інструментом для створення персоналізованого досвіду користувачів і вдосконалення взаємодії з сайтом. Після поділу аудиторії на окремі групи з подібними характеристиками стає можливим налаштовувати вміст і пропозиції відповідно до потреб кожного сегмента. Наприклад, для різних кластерів можна розробити персоналізовані рекомендації, які враховують поведінку користувачів. Одним групам пропонується контент, пов'язаний з їхніми інтересами, таким чином підвищуючи залученість і конверсії. Інші кластери, що включають нових відвідувачів, можуть отримувати ознайомлювальні повідомлення, які допомагають швидко орієнтуватися в структурі сайту та ефективно користуватися його можливостями.

Особливу увагу можна приділити активним користувачам, надаючи їм бонуси, знижки або ексклюзивний доступ до нових товарів. Це сприяє формуванню лояльності та мотивує до подальшого використання ресурсу. Наприклад, для користувачів, які часто здійснюють покупки або переглядають сторінки, можна запропонувати програми лояльності з накопичувальними знижками. Натомість менш активних клієнтів можна повернути за допомогою нагадувань про незавершені покупки або спеціальних пропозицій із вигідними умовами.

Новачки, які ще не мали достатньо часу для вивчення функціоналу сайту, також заслуговують на увагу. Для них доцільно підготувати інструкції, відео-огляди та вітальні повідомлення, які допоможуть їм швидше адаптуватися та скористатися всіма перевагами сервісу. Додатково можна стимулювати перші покупки за допомогою стартових бонусів або безкоштовної доставки.

Результати кластеризації є корисними і для оптимізації дизайну та навігації сайту. Аналіз поведінки користувачів дозволяє виявляти слабкі місця, такі як незручна структура меню або повільне завантаження сторінок, і

вчасно усувати їх. Наприклад, якщо група користувачів демонструє високий рівень відмов на певному етапі взаємодії, це сигналізує про необхідність тестування нових підходів до дизайну чи поліпшення юзабіліті.

Розробка рекламних кампаній також виграє від кластеризації. Завдяки точному сегментуванню користувачів можна створювати таргетовані оголошення для окремих груп із різними інтересами та рівнем активності. Наприклад, рекламні матеріали для частих покупців будуть відрізнятися від оголошень для тих, хто лише розглядає товари, але не здійснює покупок. Крім того, інтеграція поведінкових даних у рекламні системи, такі як Google Ads, дозволяє налаштовувати ефективні кампанії ремаркетингу.

Значний потенціал відкривається і для прогнозування майбутньої поведінки користувачів. Аналіз отриманих кластерів допомагає виявити тенденції та передбачити ймовірність майбутніх покупок, а також вчасно ідентифікувати користувачів, які можуть втратити інтерес до сайту. Це дозволяє будувати стратегії утримання клієнтів і вчасно реагувати на можливі ризики відтоку. Наприклад, для користувачів, які почали проводити менше часу на сайті, можна підготувати спеціальні пропозиції або персоналізовані нагадування.

Таким чином, кластеризація не лише допомагає зрозуміти поведінкові патерни аудиторії, а й відкриває нові можливості для розвитку бізнесу. Вона дозволяє налаштовувати індивідуальний підхід до кожної групи користувачів, підвищувати рівень задоволеності клієнтів, покращувати дизайн і функціонал сайту, а також оптимізувати маркетингові стратегії. У поєднанні з прогнозуванням поведінки це забезпечує стабільний ріст продажів і зміцнює позиції бізнесу на ринку.

2.6 Конфіденційність і безпека

Забезпечення анонімності та захисту персональних даних є ключовим аспектом методу. Відстеження здійснюється без використання особистих ідентифікаторів чи збору чутливої інформації, що дозволяє уникнути ризиків порушення конфіденційності. Усі дані, що використовуються для кластеризації та аналізу, зберігаються в анонімному вигляді, що відповідає сучасним стандартам захисту конфіденційності, зокрема GDPR.

Загальний регламент захисту даних, відомий як GDPR, набув чинності в Європейському Союзі 25 травня 2018 року. Це один із найбільш впливових законів у сфері захисту даних, що встановлює високі стандарти безпеки та конфіденційності для обробки особистої інформації користувачів. Основна мета GDPR – захист особистих даних громадян ЄС і підвищення прозорості щодо того, як компанії використовують ці дані (рис 2.1).

Рисунок 2.1 – Основні принципи GDPR

Одним з основних принципів GDPR є вимога прозорості. Це означає, що компанії зобов’язані чітко і відкрито інформувати користувачів про збір і використання їхніх даних. Користувачі повинні знати, які саме дані

збираються, як вони будуть використані, з ким поділятимуться та на який термін зберігатимуться. Така прозорість дозволяє користувачам ухвалювати обґрунтовані рішення та підвищує їхню обізнаність у питаннях конфіденційності.

GDPR також закликає компанії до мінімізації обробки даних, збираючи тільки ті дані, які є необхідними для досягнення конкретної мети, та уникаючи накопичення надмірної інформації. Цей принцип спрямований на зменшення ризиків втрати або неналежного використання даних. Ще один важливий аспект регламенту – обмеження терміну зберігання даних, який означає, що особисті дані не повинні зберігатися довше, ніж це необхідно для досягнення заявленої мети.

Особлива увага в GDPR приділяється правам користувачів на управління своїми даними. Громадяни ЄС мають право отримувати доступ до своїх даних, вимагати їх зміни, обмежувати обробку, а також видаляти інформацію в разі, якщо вона більше не потрібна. Наприклад, право на забуття дає користувачам можливість вимагати повного видалення своїх даних, а право на перенесення даних дозволяє передати їх іншій компанії.

Для забезпечення належного рівня безпеки GDPR запроваджує концепцію захисту даних за замовчуванням та захисту даних на етапі розробки. Компанії повинні враховувати захист приватності на всіх етапах розробки своїх продуктів, що включає застосування шифрування, багаторівневих систем безпеки та інших технічних заходів.

GDPR також вимагає швидкого реагування на порушення безпеки: у разі витоку даних компанії зобов’язані повідомити відповідні органи протягом 72 годин і, за необхідності, інформувати постраждалих користувачів. Це підвищує рівень відповідальності компаній перед клієнтами, оскільки вони повинні негайно реагувати на загрози для захисту особистої інформації.

У сфері веб-відстеження та кластеризації користувачів GDPR визначає чіткі обмеження, що стосуються збору та обробки поведінкових даних. Всі

дані повинні бути анонімізованими, щоб виключити можливість ідентифікації конкретної особи. Більше того, для кожної операції з обробки даних має бути отримана згода користувача. GDPR також зобов’язує компанії надавати користувачам доступ до зібраних даних та можливість управляти ними, забезпечуючи більш етичний підхід до роботи з інформацією.

GDPR сприяє розвитку нових технологій у сфері безпеки даних, що вимагає від компаній інвестувати в сучасні системи для забезпечення захисту та прозорості обробки даних. Такі інвестиції включають розробку алгоритмів анонімізації, шифрування даних на рівні користувачів, а також впровадження протоколів контролю доступу. Зокрема, технології блокчейну стають важливим інструментом для забезпечення незмінності даних і прозорості обробки.

Крім того, регламент вимагає документування всіх процесів, пов’язаних із обробкою персональних даних, щоб компанії могли підтвердити відповідність вимогам у разі перевірок. Ця вимога підвищує рівень підзвітності компаній і стимулює впровадження політик регулярного аудиту та моніторингу процесів обробки інформації.

Таким чином, GDPR не тільки створює суворі правила захисту даних, а й заохочує використання новітніх технологій і підходів для забезпечення безпеки інформації, підвищуючи довіру користувачів до цифрових сервісів і платформ.

3 ВИБІР ПІДХОДІВ ТА ТЕХНОЛОГІЙ ДЛЯ РОЗРОБКИ

3.1Обґрунтування вибору технологій і методів реалізації

Для вирішення задачі відстеження та кластеризації користувачів на вебсайті було обрано ряд технологій і методів, що забезпечують ефективну обробку великих обсягів даних, простоту інтеграції та високу продуктивність у процесі реалізації. Вибір цих інструментів був обумовлений необхідністю досягнення точних результатів аналізу поведінки користувачів з урахуванням вимог до масштабованості та зручності для подальших досліджень і застосувань.

Однією з основних технологій, обраних для реалізації, стала мова програмування Python. Її універсальність і велика кількість бібліотек для обробки даних і машинного навчання роблять Python ідеальним вибором для реалізації складних аналітичних завдань. Мова підтримує інтеграцію з популярними інструментами для збору даних, зокрема з Google Analytics API, що дозволяє отримувати різноманітні дані про взаємодію користувачів з вебсайтом в реальному часі. Це дає змогу здійснювати глибокий аналіз поведінкових патернів та оцінювати ефективність різних методів кластеризації.

Для організації та проведення досліджень було вибрано середовище Jupyter Notebook, яке забезпечує інтерактивність і зручність для покрокового виконання коду та наочної демонстрації результатів. Це середовище дозволяє швидко виконувати код, аналізувати дані та створювати візуалізації, що особливо важливо при роботі з великими наборами даних. Інтерфейс Jupyter Notebook дозволяє не лише тестувати алгоритми, але й створювати графіки і теплові карти для подальшого аналізу отриманих результатів.

Одним із ключових елементів для вирішення задачі кластеризації користувачів став вибір двох потужних алгоритмів – K-means та DBSCAN.

Алгоритм K-means є одним з найбільш популярних і ефективних методів для класифікації даних на основі поведінкових патернів. Його швидка робота і здатність справлятися з великими наборами даних дозволяють отримати точні результати при класифікації користувачів. Однією з головних переваг K-means є його простота та швидкість у порівнянні з іншими методами кластеризації, що робить його оптимальним для задач з великими обсягами інформації.

Для вирішення більш складних завдань кластеризації, зокрема для виявлення кластерів з довільною формою і обробки шумових даних, був обраний алгоритм DBSCAN. Цей метод дозволяє ефективно групувати дані, навіть якщо класи не є лінійно роздільними або мають складну геометрію. Однією з переваг DBSCAN є здатність ігнорувати шумові дані та виявляти аномальні точки, що робить його корисним інструментом для складніших типів кластеризації, де інші алгоритми можуть не дати задовільних результатів.

Для реалізації алгоритмів машинного навчання та кластеризації було використано бібліотеку scikit-learn, яка є однією з найпоширеніших і найпотужніших для Python. Ця бібліотека містить великий набір інструментів для кластеризації, масштабування даних та оцінки якості кластерів. Вона значно спрощує процес розробки та тестування різних алгоритмів, забезпечуючи високу гнучкість і продуктивність у роботі з даними.

Підготовка і обробка даних для кластеризації здійснювалась за допомогою бібліотеки Pandas. Завдяки своїм потужним інструментам для роботи з табличними даними та можливості обробляти великі обсяги інформації, Pandas дозволяє ефективно здійснювати попередню обробку даних, зокрема видаляти дублікати, заповнювати пропуски та нормалізувати показники. Це необхідно для забезпечення коректності кластеризації і досягнення точних результатів при аналізі користувацької поведінки.

Для візуалізації результатів кластеризації була використана бібліотека Matplotlib, яка дозволяє створювати різноманітні графіки, теплові карти та

інші візуальні елементи для демонстрації результатів досліджень. Візуалізація є важливим етапом у процесі аналізу, оскільки вона дозволяє наочно побачити закономірності, що виникають в результаті кластеризації, а також дає можливість зробити висновки щодо ефективності обраних методів.

Для збору даних про поведінку користувачів на вебсайті було використано Google Analytics API, що дозволило отримати точну інформацію в реальному часі. Цей інструмент надає доступ до різноманітних показників, таких як кількість переглядів сторінок, тривалість сеансів, коефіцієнти відмов та інші важливі метрики, які стали основою для кластеризації та аналізу поведінкових патернів користувачів.

Вибірка даних для дослідження містила інформацію про відвідування вебсайту, включаючи кількість переглядів сторінок, тривалість сеансів, джерела трафіку, показники відмов та кліки користувачів. Ці дані були підготовлені за допомогою Pandas, а для подальшого аналізу та кластеризації використовувалися алгоритми K-means і DBSCAN. Візуалізація результатів кластеризації була проведена за допомогою Matplotlib у середовищі Jupyter Notebook. Ця вибірка дозволила виявити закономірності поведінки користувачів та сформувати кластери за подібними ознаками, що дозволило оцінити ефективність використаних методів аналізу.

3.2 Середовище розробки Jupyter Notebook

Jupyter Notebook є надзвичайно потужним та універсальним інструментом для виконання наукових і технічних розробок, який надає можливість інтерактивно працювати з кодом, текстом, візуалізаціями та математичними формулами. Це середовище стало незамінним у багатьох сферах, включаючи науку про дані, машинне навчання, статистику та обробку великих даних. Його основною перевагою є інтерактивний характер

роботи, що дозволяє виконувати Python-код поетапно, покроково тестуючи алгоритми та відразу оцінюючи їх результати.

Однією з ключових переваг Jupyter Notebook є можливість комбінувати код з коментарями, візуалізаціями та математичними формулами в одному документі. Це дозволяє не лише програмувати, але й зберігати чітке пояснення до кожного етапу роботи. Для дослідників та розробників це створює ідеальні умови для створення документації, де код, результати та пояснення об’єднуються в одному інтерактивному середовищі. Завдяки такій інтеграції користувач може негайно отримувати візуальний зворотний зв'язок про виконану роботу. У контексті нашого дослідження це означає, що ми можемо детально документувати процес підготовки даних, налаштування алгоритмів кластеризації та аналізу результатів, що робить дослідження більш прозорим і зрозумілим.

Додатково, інтерактивність Jupyter Notebook дозволяє зручно виконувати код і одразу перевіряти результат, що є незамінним при розробці складних алгоритмів. Наприклад, при тестуванні різних варіантів кластеризації, ми можемо в реальному часі коригувати параметри алгоритмів і бачити, як змінюється результат. Це значно прискорює процес налагодження і дозволяє миттєво реагувати на будь-які помилки чи невідповідності в результатах. Також, завдяки відсутності необхідності перезавантажувати середовище після кожної зміни, це значно зменшує час на виконання експериментів.

З точки зору інтеграції з іншими бібліотеками Python, Jupyter Notebook є не менш потужним інструментом. Завдяки прямій інтеграції з популярними бібліотеками, такими як Pandas, scikit-learn, NumPy, Matplotlib та Seaborn, ми маємо змогу працювати з даними, будувати моделі та здійснювати візуалізацію прямо в одному середовищі. У процесі нашого дослідження ми використовували Pandas для попередньої обробки даних, включаючи очищення, нормалізацію та трансформацію змінних, що дозволило підготувати дані до кластеризації. Бібліотека scikit-learn використовувалася

для безпосереднього виконання алгоритмів кластеризації K-means і DBSCAN, а також для оцінки їх ефективності.

Ще однією перевагою Jupyter Notebook є його здатність працювати з великими обсягами даних завдяки підтримці інтеграції з бібліотеками, оптимізованими для обробки великих наборів даних. За допомогою таких бібліотек, як Dask, PySpark або Vaex, можна працювати з даними, які не вміщуються в оперативну пам'ять, використовуючи при цьому принципи, схожі на звичайну роботу з Pandas.

Візуалізація є важливим елементом в процесі аналізу даних, і Jupyter Notebook пропонує зручний інтерфейс для створення інтерактивних графіків та візуалізацій. За допомогою Matplotlib, Seaborn та Plotly ми змогли створювати детальні графіки для аналізу кластеризації, що дозволяло наочно оцінювати якість розподілу даних по кластерах. Така візуалізація не тільки полегшує розуміння складних результатів, але й дає можливість швидко адаптувати алгоритми, коригуючи їх на основі отриманих візуальних даних.

Jupyter Notebook також має функціональність для легкого експорту результатів у різні формати, такі як HTML, PDF, LaTeX та інші. Це дозволяє зберігати результати досліджень у зручному вигляді для подальшого аналізу, обміну результатами або публікацій. Експорт у формат HTML чи PDF робить процес документування простим і швидким, зберігаючи всі необхідні графіки, коди і коментарі.

Загалом, Jupyter Notebook є надзвичайно потужним та універсальним інструментом, що значно покращує процес розробки та аналізу в галузях, пов'язаних з даними та машинним навчанням. У нашому дослідженні цей інструмент став незамінним завдяки своїй зручності та здатності інтегрувати різні елементи роботи в одне середовище. Використання Jupyter Notebook дозволило ефективно реалізувати всі етапи дослідження, від підготовки даних до візуалізації результатів кластеризації, що в свою чергу підвищило продуктивність і точність аналізу.

3.3 Основні відомості про мову програмування Python

Python є однією з найбільш популярних мов програмування, що здобула визнання завдяки своїй простоті, гнучкості та широкому спектру можливостей. Завдяки зрозумілому синтаксису та високій читабельності, Python ідеально підходить для різних рівнів досвіду – від новачків до професіоналів. У контексті нашого дослідження Python став основним інструментом для збору, обробки та аналізу даних, а також для реалізації алгоритмів кластеризації користувачів. Це дозволило зосередитись на вирішенні наукових задач, не витрачаючи часу на складні аспекти програмування, що супроводжують багато інших мов.

Однією з основних переваг Python є його величезна екосистема бібліотек і модулів, що охоплюють всі аспекти роботи з даними – від збору і зберігання до аналізу та візуалізації. Ці бібліотеки не тільки значно спрощують процес розробки, а й дають можливість зосередитись на аналізі даних і вдосконаленні алгоритмів, а не на низькорівневих деталях програмування. У нашому дослідженні Python став незамінним інструментом завдяки таким бібліотекам, як Pandas, Scikit-learn і Matplotlib, які суттєво спростили та прискорили процес роботи.

Pandas, зокрема, стала основною бібліотекою для обробки та аналізу даних. Завдяки її потужним інструментам для маніпулювання великими масивами даних, обробки відсутніх значень і виконання операцій з різними форматами даних, ми змогли ефективно підготувати і очистити наші дані для подальшого аналізу. Бібліотека також забезпечила легкий доступ до можливості виконувати попередню обробку даних, що дозволило адаптувати їх до потреб алгоритмів кластеризації.

Scikit-learn, з іншого боку, стала ключовою бібліотекою для реалізації алгоритмів кластеризації. За допомогою цієї бібліотеки ми змогли реалізувати методи кластеризації, такі як K-means і DBSCAN, для групування користувачів за схожими поведінковими характеристиками. Крім того, Scikit-

learn надав інструменти для розрахунку метрики силуету, яка дозволила оцінити якість кластеризації та визначити, наскільки вдало була здійснена сегментація користувачів.

Для візуалізації результатів кластеризації ми використовували Matplotlib. Ця бібліотека допомогла нам створити графіки, які наочно демонструють розподіл користувачів між різними кластерами, а також теплові карти для вивчення залежностей між різними характеристиками поведінки. Завдяки зручному інтерфейсу і широким можливостям для налаштування графіків, Matplotlib дозволила зробити результати дослідження більш зрозумілими і наочними.

Окрім того, Python також має чудову інтеграцію з іншими технологіями, що значно полегшує роботу з даними. Зокрема, Google Analytics API став важливим інструментом для збору даних про активність користувачів на вебсайті. Цей API дозволив нам автоматизувати процес отримання даних та організувати їх у зручний формат JSON для подальшої обробки і аналізу. Завдяки гнучкості API, ми змогли налаштувати запити таким чином, щоб отримувати тільки найбільш релевантні метрики для нашого дослідження, що дозволило уникнути надлишку непотрібної інформації.

Ще однією важливою перевагою Python є підтримка інтерактивних середовищ розробки, таких як Jupyter Notebook. Цей інструмент надав нам можливість поєднувати програмний код, пояснення, а також графіки та візуалізації в одному інтерактивному документі, що дозволило зручно тестувати та відлагоджувати код без необхідності створювати окремі файли чи застосовувати додаткові середовища. Завдяки цьому процес розробки став набагато простішим і ефективнішим.

Таким чином, Python виявився ідеальним вибором для реалізації нашого дослідження. Його простота у використанні, потужність і велика кількість бібліотек дозволили не тільки реалізувати складні алгоритми кластеризації, але й інтегрувати їх із іншими інструментами для збору даних.

Крім того, Python допоміг автоматизувати численні етапи обробки та аналізу даних, що значно підвищило ефективність і точність роботи. В кінцевому результаті, Python став незамінним інструментом для досягнення мети дослідження – створення ефективної системи кластеризації користувачів на основі їх поведінкових характеристик, що може бути корисним для подальшої персоналізації маркетингових стратегій та покращення користувацького досвіду на вебсайті.

3.4 Бібліотеки для розробки: Scikit-learn, Pandas, Matplotlib, Seaborn

У рамках виконання роботи для обробки даних, реалізації алгоритмів кластеризації та візуалізації результатів використовувалися три ключові бібліотеки Python: Scikit-learn, Pandas, Matplotlib та Seaborn. Вони стали основними інструментами для реалізації всіх етапів дослідження, забезпечивши ефективність, зручність та швидкість виконання завдань.

Scikit-learn є однією з найпопулярніших бібліотек для машинного навчання в Python. Вона використовувалася в цьому дослідженні для реалізації алгоритмів кластеризації, таких як K-means та DBSCAN. Ці методи дозволяють ефективно групувати користувачів за схожістю їх поведінкових характеристик, таких як кількість переглянутих сторінок, час, проведений на сайті, та кількість взаємодій із контентом.

Зокрема, алгоритм K-means застосовувався для розбиття користувачів на заздалегідь визначену кількість кластерів, що дозволило визначити основні групи користувачів (активні, пасивні, нові та постійні). Алгоритм DBSCAN дозволив виявити кластери, які не мають фіксованої кількості, що є корисним для виявлення аномальних або рідкісних груп користувачів.

Однією з переваг використання Scikit-learn є наявність великої кількості функцій для оцінки якості кластеризації, зокрема метрики силуету, що дозволило детально оцінити згортання груп і чіткість розподілу

користувачів у кластерах. Це дозволяє не лише отримати результати, але й впевнитися в їх доцільності та коректності.

Pandas виступала основним інструментом для обробки та підготовки даних. Інформація про активність користувачів, що надходила через Google Analytics API, зберігалася у форматі JSON і підлягала обробці в Pandas. Бібліотека забезпечує потужні можливості для маніпуляцій з табличними даними, що є незамінним на етапі попередньої обробки великих обсягів даних.

На етапі підготовки даних виконувалися такі операції, як:

– очищення даних: видалення пропущених значень, записів, що повторюються та обробка аномалій;

– нормалізація даних: масштабування числових ознак до одного діапазону для уникнення впливу великих значень на результат кластеризації;

– структурування даних: перетворення сирих даних у форму, зручну для подальшої обробки (наприклад, створення векторів ознак, необхідних для алгоритмів кластеризації).

Завдяки гнучкості Pandas процес роботи з великими наборами даних став значно швидшим і ефективнішим. Це дозволило зосередитися на глибинному аналізі даних та безпомилковій підготовці інформації до кластеризації.

Для візуалізації результатів кластеризації та аналізу даних були використані дві основні бібліотеки, як Matplotlib та Seaborn. Візуалізація є важливою частиною роботи з даними, оскільки дозволяє наочно уявити структуру даних і зрозуміти взаємозв'язки між різними характеристиками користувачів.

Matplotlib є однією з найбільш поширених бібліотек для побудови графіків у Python. У рамках цього дослідження вона була використана для створення основних візуалізацій, які допомогли представити результати кластеризації. Наприклад, були побудовані двовимірні графіки, що показують, як користувачі розподіляються по різних кластерах. Це дозволило

чітко продемонструвати, які користувачі мають схожі поведінкові патерни, наприклад, за кількістю переглянутих сторінок чи часом, проведеним на сайті. Кожен кластер був представлений певним кольором, що полегшило сприйняття та інтерпретацію даних.

Також за допомогою Matplotlib були створені теплові карти, що показують взаємозв'язки між різними характеристиками користувачів і центрами кластерів. Це дозволило виявити важливі кореляції, такі як, наприклад, залежність між часом на сайті і кількістю здійснених взаємодій. Теплові карти допомогли наочно зрозуміти, які параметри поведінки користувачів визначають їх приналежність до певних кластерів.

У свою чергу, Seaborn є бібліотекою, що розширює можливості Matplotlib і надає ще більш елегантні та інформативні графіки. Seaborn використовувалась для побудови складніших візуалізацій, таких як паралельні координати або кореляційні матриці, що дозволили більш детально дослідити взаємозв'язки між різними ознаками користувачів. Завдяки Seaborn, графіки стали більш зручними для інтерпретації, оскільки вони надають більше статистичних і візуальних підказок, наприклад, використання відтінків кольору для позначення сильних або слабких кореляцій між ознаками.

Комбінація Matplotlib та Seaborn дозволила створити зрозумілі та естетично привабливі візуалізації, що значно полегшило процес аналізу та інтерпретації результатів кластеризації. Вони не лише допомогли виявити основні патерни у поведінці користувачів, але й стали важливим інструментом для представлення результатів у зручному для сприйняття форматі, що значно підвищує ефективність подальшої роботи з даними.

Завдяки цим бібліотекам, результати кластеризації не лише стали більш зрозумілими та доступними для аналізу, але й дозволили створити наочні презентації для представлення результатів дослідження.

Поєднання Scikit-learn, Pandas, Matplotlib та Seaborn стало потужним інструментом для виконання всіх етапів цього дослідження. Від підготовки

та обробки даних до кластеризації та візуалізації результатів, ці бібліотеки забезпечили надійний і ефективний інструментарій для досягнення поставлених цілей. Спільне використання цих інструментів дозволило не лише здійснити кластеризацію користувачів за їх поведінковими характеристиками, а й надати детальний аналіз отриманих груп користувачів для подальших маркетингових стратегій.

3.5 Google Analytics API

Google Analytics API відіграє ключову роль у цьому дослідженні, оскільки надає можливість автоматизованого доступу до великого обсягу поведінкових даних користувачів вебсайту. Цей інтерфейс дозволяє отримати широкий спектр метрик, які детально описують взаємодію відвідувачів з вебресурсом. Серед таких показників варто виділити тривалість сесій, кількість переглянутих сторінок, частоту повернень, відсоток відмов і багато інших параметрів. Ці метрики є надзвичайно важливими, оскільки вони надають цінну інформацію про рівень залученості користувачів, їхню активність та загальну поведінку на сайті. Зібрані дані формують основу для проведення подальшого аналізу та кластеризації, допомагаючи виявити різні патерни взаємодії і сегментувати аудиторію за поведінковими характеристиками.

Процес збирання інформації через Google Analytics API був налаштований для отримання даних у форматі JSON. Такий вибір пояснюється тим, що формат JSON є легким для зберігання, передачі та обробки даних, особливо коли йдеться про великі обсяги інформації. Ця особливість значно спрощує інтеграцію з програмним забезпеченням і дозволяє швидко імпортувати дані для подальшої аналітики. Крім того, гнучкість інтерфейсу API дозволила точніше налаштувати запити відповідно до специфічних потреб дослідження, сфокусувавшись на ключових метриках,

необхідних для кластеризації користувачів. Це, в свою чергу, допомогло уникнути надмірного обсягу зайвої інформації і зробило обробку даних більш ефективною.

Після збору даних їхня первинна обробка була виконана за допомогою бібліотеки Pandas, що є потужним інструментом для роботи з табличними даними. На цьому етапі було проведено очищення набору даних від некоректних або відсутніх значень, а також нормалізацію числових показників. Нормалізація відіграла важливу роль, оскільки дозволила привести всі параметри до єдиного масштабу, що забезпечило коректність подальших обчислень і покращило точність алгоритмів кластеризації. Крім цього, завдяки можливостям Pandas вдалося зручно структурувати інформацію та підготувати її для подальшого аналізу.

Інтеграція Google Analytics API з середовищем програмування Python стала важливим кроком у процесі автоматизації збору та обробки даних. Це дозволило значно підвищити швидкість і ефективність роботи, а також забезпечити високу якість і достовірність отриманих результатів. Використання цього API сприяло оптимізації всіх етапів аналізу, починаючи від збору інформації і закінчуючи її обробкою та кластеризацією. Такий підхід мінімізував ризики втрат або спотворень даних і гарантував точність фінальних висновків.

Дані, зібрані за допомогою Google Analytics API, стали основою для подальшого застосування алгоритмів кластеризації, таких як K-means і DBSCAN. Висока якість вихідних показників сприяла отриманню точних і репрезентативних результатів аналізу поведінкових характеристик користувачів. Таким чином, використання Google Analytics API дозволило не лише ефективно сегментувати аудиторію, а й виявити ключові закономірності в її поведінці, що є важливим кроком у вдосконаленні методів аналізу користувацької активності.

Впровадження цього інструменту продемонструвало, як сучасні технології можуть забезпечити детальний і всебічний аналіз взаємодії

користувачів із вебсайтом. Це дослідження показало, що комбінація Google Analytics API з сучасними методами обробки даних дозволяє створювати точні та ефективні алгоритми кластеризації. Завдяки цьому підходу вдалося не тільки досягти високої точності результатів, а й створити основу для подальших досліджень у сфері поведінкового аналізу.

3.6 Результати та огляд створеного алгоритму

При моделюванні відстеження та кластеризації користувачів на веб-сайті, використовуючи середовище Jupyter notebook. Для цього застосовували мови програмування Python та різноманітні бібліотеки, зокрема scikit-learn, Pandas та Matplotlib. Основною метою роботи було зібрати дані про активність користувачів, зокрема про їхні кліки, відвідані сторінки, час перебування на сайті, сесії та інші важливі події.

Процес збору даних реалізувався за допомогою Google Analytics API та власних інструментів логування подій. Вся отримана інформація про користувачів була збережена у форматі JSON і піддана обробці за допомогою бібліотеки Pandas.

За основу було узято вибірку Website Traffic. Вона містить детальну інформацію про трафік на вебсайті, зокрема щодо поведінки користувачів під час їхніх сесій. Дані охоплюють кілька важливих метрик, таких як кількість переглянутих сторінок, тривалість сесії, показник "Bounce Rate" (відсоток користувачів, які залишають сайт після перегляду тільки однієї сторінки), джерела трафіку, час, проведений на сторінці, кількість попередніх відвідувань, а також конверсії – відсоток користувачів, які виконали бажану дію, таку як покупка або підписка.

Ця вибірка дає змогу аналізувати, як користувачі взаємодіють з вебсайтом, і допомагає виявити патерни поведінки серед відвідувачів. Наприклад, метрики дозволяють оцінити рівень залученості користувачів, ефективність різних джерел трафіку, а також визначити, які сторінки або

секції сайту найкраще утримують увагу користувачів. Вибірка включає 2000 записів, що дає змогу проводити статистичні та аналітичні дослідження, виявляючи якості, які можуть допомогти у поліпшенні вебсайту та маркетингових стратегій.

Метод відстеження користувачів починається з підготовки даних, що включає завантаження інформації про поведінку користувачів із файлу та попередню обробку. Категоріальні дані кодуються у числовий формат для подальшої обробки. Потім дані нормалізуються для забезпечення рівномірного масштабу, що є важливим етапом перед застосуванням алгоритмів кластеризації. Після підготовки даних метод пропонує два підходи до групування користувачів – алгоритм K-means та алгоритм DBSCAN. K-means класифікує користувачів, розподіляючи їх у фіксовану кількість кластерів, обчислюючи відстань до центрів кластерів і оновлюючи їх, поки не буде досягнуто оптимального поділу. DBSCAN, у свою чергу, визначає групи на основі щільності даних і виявляє аномалії або «шум», що можуть свідчити про незвичайну поведінку. Для кожного з підходів результати візуалізуються за допомогою методу головних компонент (PCA), який зменшує вимірність даних для побудови двовимірних графіків. Крім того, теплові карти середніх значень по кластерах дозволяють виявити ключові характеристики груп користувачів і порівняти їх між собою. Такий підхід забезпечує глибокий аналіз поведінки користувачів та допомагає ідентифікувати тенденції у взаємодії з вебсайтом.

Метою цих алгоритмів було поділити користувачів на групи відповідно до їх поведінкових характеристик. Результати кластеризації були візуалізовані за допомогою бібліотеки Matplotlib, де було отримано двовимірні графіки, теплові карти, діаграми та боксплоти, які показували розподіл користувачів між різними кластерами.

Графік на рисунку 3.1 показує результати кластеризації K-means після застосування методу головних компонент (PCA) для зниження розмірності даних до двох компонент. На осі X розташована перша головна компонента

(PCA Component 1), а на осі Y – друга головна компонента (PCA Component 2). Точки на графіку мають різні кольори в залежності від належності до одного з трьох кластерів, що визначені алгоритмом K-means: фіолетовий (Cluster 0), бірюзовий (Cluster 1) і жовтий (Cluster 2). Кольори відображають інтенсивність кластерної належності, де яскравіші кольори вказують на більш високу ймовірність належності до відповідного кластеру. Графік дає змогу оцінити, як кластери розподіляються в просторі з двома основними компонентами, що дозволяє візуалізувати структуру даних після кластеризації.

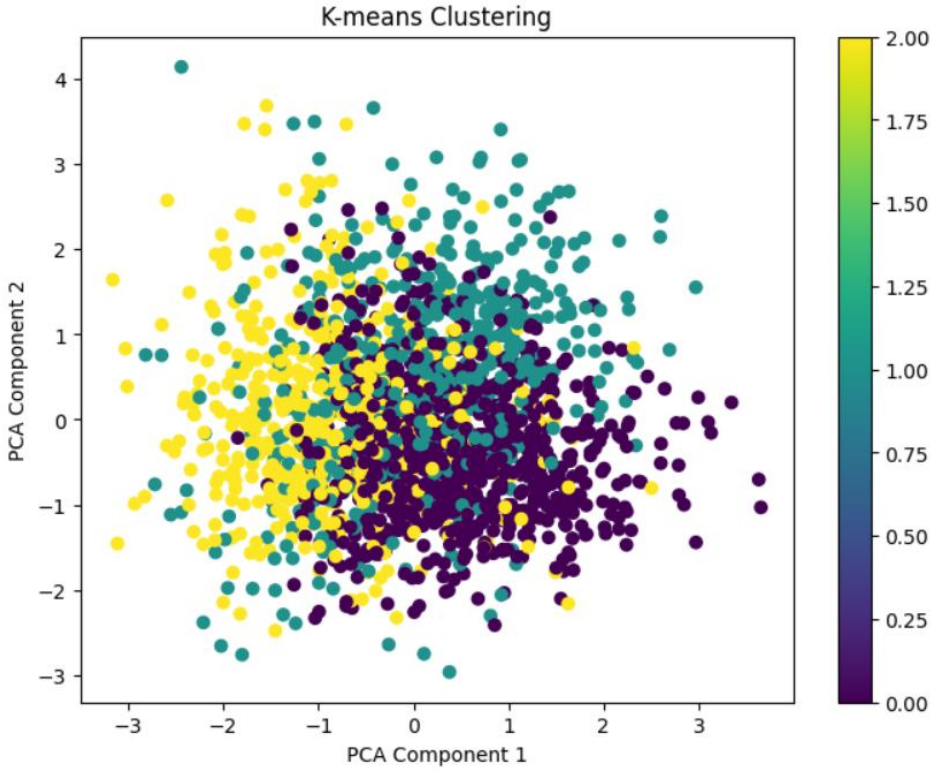

Рисунок 3.1 – Двовимірний графік метода K-means

Графік на рисунку 3.2 демонструє результати кластеризації методом DBSCAN, також з використанням зниження розмірності за допомогою методу головних компонент (PCA). На осі X представлена перша головна компонента (PCA Component 1), а на осі Y – друга головна компонента (PCA Component 2). Точки на графіку розфарбовані відповідно до результатів кластеризації, де кольори відображають різні кластери, визначені DBSCAN.

Найбільша частина точок має темно-фіолетовий колір, що свідчить про те, що більшість даних не належать до жодного окремого кластера і можуть бути визначені як шум (позначені як -1). Менші групи точок різних кольорів (жовтий, зелений, бірюзовий) вказують на окремі кластери. Графік надає змогу оцінити, як метод DBSCAN групує дані, виділяючи як щільні області, так і шумові точки.

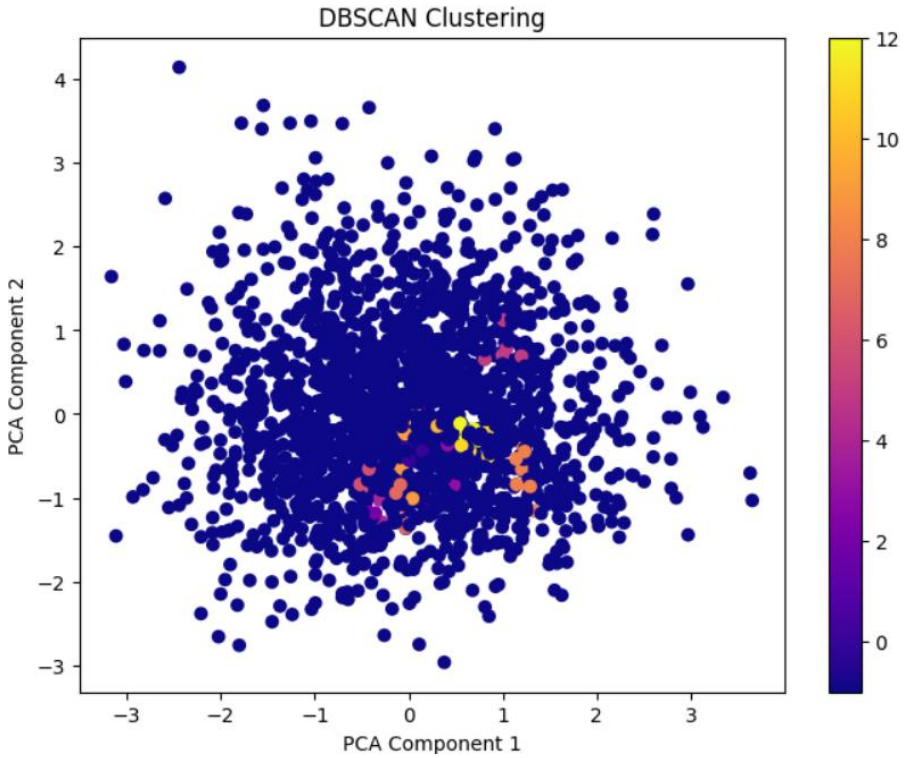

Рисунок 3.2 – Двовимірний графік метода DBSCAN

Графік на рисунку 3.3 являє собою теплову карту, яка ілюструє середні значення для різних показників у кожному з кластерів, отриманих за допомогою алгоритму K-means. На карті представлені показники, як-от "Page Views", "Session Duration", "Bounce Rate", "Traffic Source", "Time on Page", "Previous Visits" та "Conversion Rate". Кожен стовпець відповідає окремому кластеру (Cluster 0, Cluster 1, Cluster 2), а кожен рядок – певному показнику. Колір на карті варіюється від червоного до синього, де червоний позначає високе значення, а синій – низьке. Відзначається, що кластери 0 і 1 мають схожі середні значення для більшості показників, зокрема для "Page Views",

"Session Duration" і "Time on Page", тоді як кластер 2 демонструє знижені значення для цих показників, що може свідчити про менш активну взаємодію користувачів з сайтом. Ця теплова карта допомагає порівняти середні значення характеристик серед різних кластерів та виявити їхні відмінності.

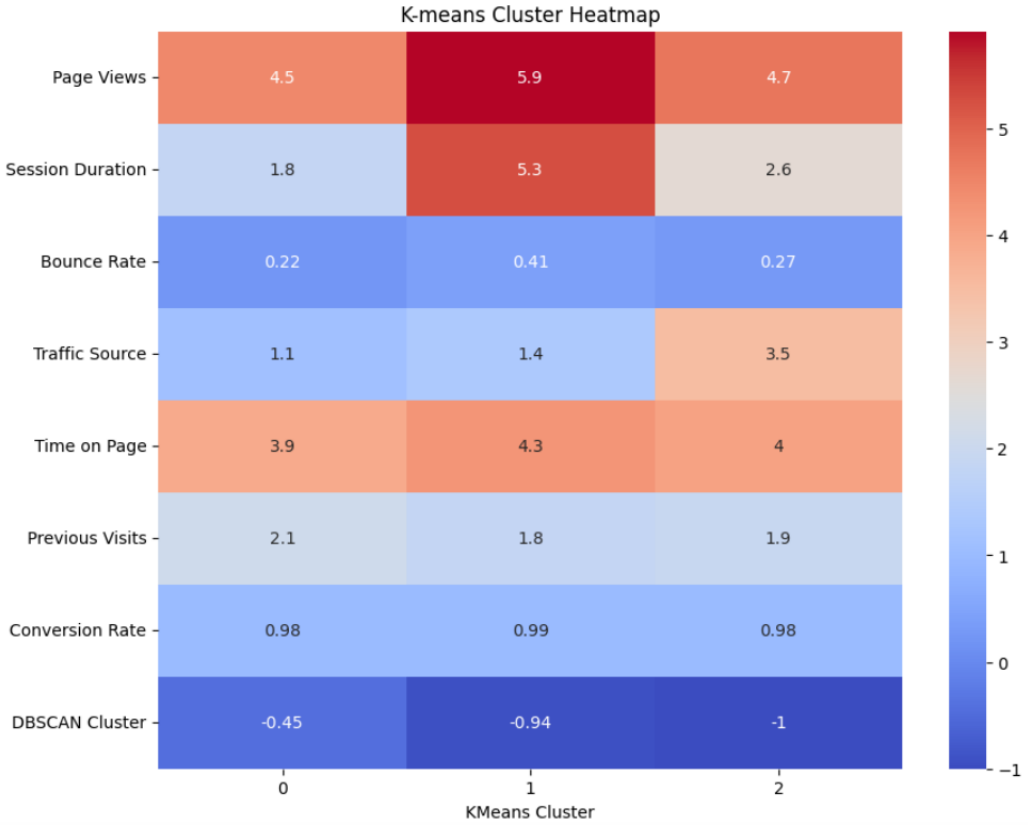

Рисунок 3.3 – Теплова карта метода K-means

Графік, зображений на рисунку 3.4, є ще однією тепловою картою, яка відображає середні значення для різних показників у кожному з кластерів, отриманих за допомогою алгоритму DBSCAN. Кожен стовпець представляє один з кластерів, а кожен рядок – це конкретний показник, такий як "Page Views", "Session Duration", "Bounce Rate", "Traffic Source", "Time on Page", "Previous Visits", "Conversion Rate", та "KMeans Cluster". Колір теплової карти змінюється від червоного до синього, де червоний колір вказує на високі значення, а синій – на низькі.

Кластери з номерами 5, 6, 8, 12 мають найбільші значення для "Page Views", що свідчить про значну кількість переглядів сторінок у цих

кластерах. Водночас для показників "Bounce Rate" і "Session Duration" найнижчі значення спостерігаються в кластерах 7, 10 та 12, що може свідчити про зменшену активність користувачів або коротші сеанси. Теплова карта також показує, що для показників, таких як "Conversion Rate", більшість кластерів мають значення близькі до 1, що вказує на високу ймовірність конверсії серед користувачів.

Водночас останній рядок вказує на належність кожного кластеру до конкретного кластеру K-means, де значення вказують на те, чи входять користувачі з певного кластера DBSCAN до одного з кластерів K-means. Цей графік допомагає порівняти дані по кластеризації для різних алгоритмів та їх вплив на характеристики користувачів.

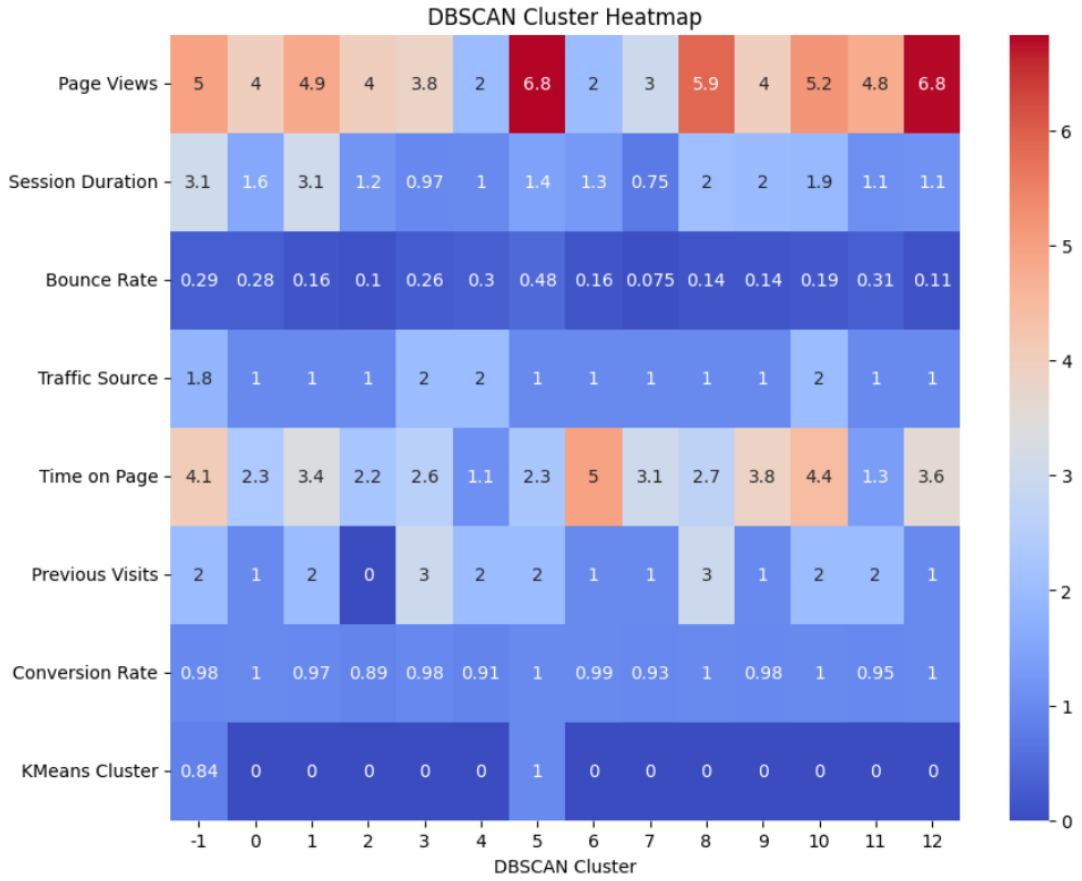

Рисунок 3.4 – Теплова карта метода DBSCAN

Графік на рисунку 3.5 являє собою гістограму показує розподіл кількості переглядів сторінок для кожного з кластерів, визначених за

допомогою алгоритму K-means. На осі X представлений показник "Page Views" (кількість переглядів сторінок), а на осі Y – кількість спостережень (Count). Кожен колір на графіку відповідає окремому кластеру: темно-фіолетовий (Cluster 0), бірюзовий (Cluster 1) та жовтий (Cluster 2). Видно, що кластер 0 має найбільшу кількість користувачів з кількістю переглядів від 3 до 6, тоді як кластери 1 та 2 мають більш виражений розподіл у діапазоні низьких переглядів (1–3 для кластера 1 та 7–12 для кластера 2). У графіку також присутні лінії розподілу ймовірності (ядерні оцінки щільності), що допомагають краще побачити форму розподілу даних у кожному кластері.

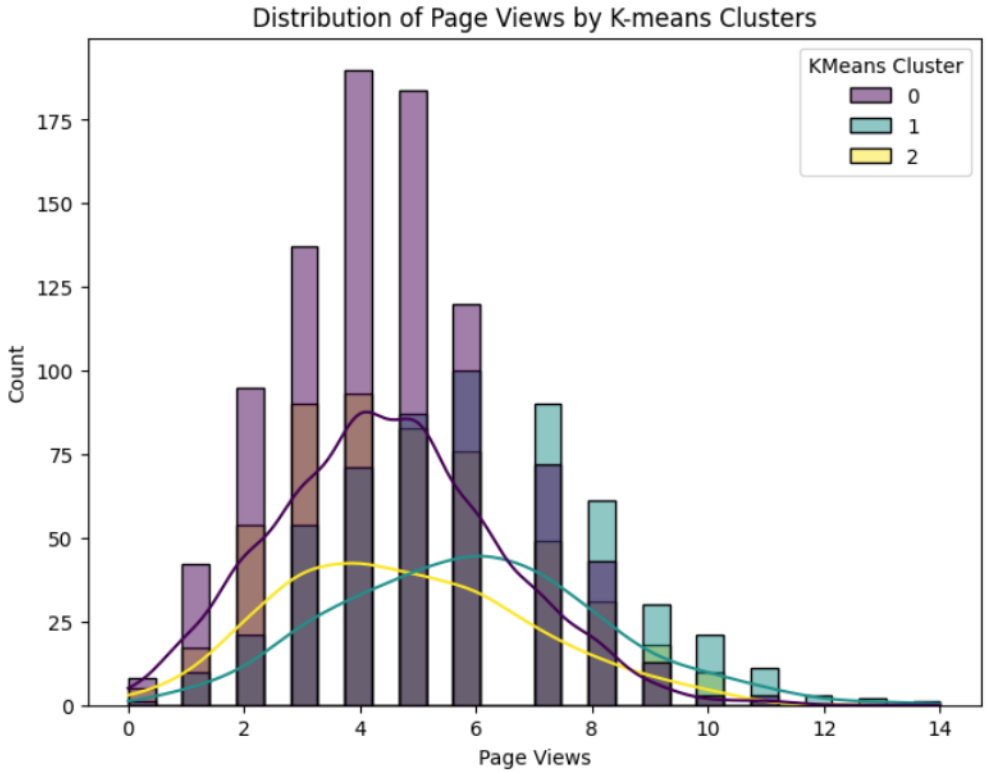

Рисунок 3.5 – Гістограма розподілу "Page Views" кластерами за допомогою метода K-means

Гістограма на рисунку 3.6 відображає розподіл тривалості сеансів ("Session Duration") для кожного з кластерів, визначених алгоритмом K-means. На осі X показано тривалість сеансу, а на осі Y – кількість користувачів. Розподіли для кожного кластера представлені різними

кольорами: фіолетовий для кластера 0, бірюзовий для кластера 1 і жовтий для кластера 2.

Видно, що кластер 0 має найбільшу кількість користувачів з короткими сеансами, приблизно від 0 до 3 одиниць часу, з більшістю значень на рівні 1-2 одиниці часу. Кластер 1, навпаки, має більш рівномірний розподіл, із помітним зростанням числа користувачів, що мають більш довгі сеанси, від 3 до 8 одиниць часу. Кластер 2 представлений на графіку жовтим кольором, і його розподіл також зосереджений на коротких сеансах, але з незначною присутністю користувачів, чия тривалість сеансу перевищує 10 одиниць часу.

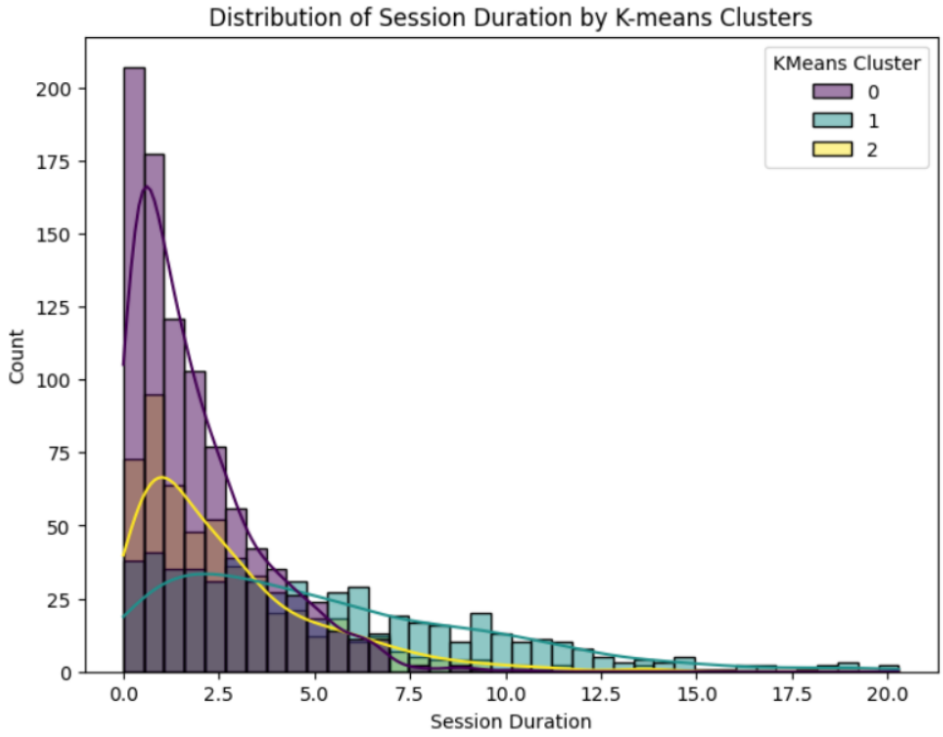

Рисунок 3.6 – Гістограма розподілу "Session Duration" кластерами за допомогою метода K-means

Гістограма на рисунку 3.7 показує розподіл показника "Bounce Rate" для кожного з кластерів, отриманих за допомогою алгоритму K-means. На осі X зображений показник "Bounce Rate" (відсоток відмов), а на осі Y – кількість користувачів у кожному бінові. Кожен колір відповідає окремому кластеру: фіолетовий – кластер 0, бірюзовий – кластер 1 і жовтий – кластер 2.

Видно, що кластер 0 характеризується високою концентрацією користувачів з низьким рівнем відмов, з найбільшою кількістю спостережень на рівні близько 0.2, що вказує на те, що користувачі з цього кластеру схильні залишати сайт після перегляду декількох сторінок. Кластер 1 має більш рівномірний розподіл по всьому діапазону значень, з тенденцією до більш високих значень Bounce Rate, що свідчить про більшу кількість користувачів, які покидають сайт після перегляду однієї сторінки. Кластер 2, що зображений жовтим кольором, має більшу концентрацію на високих значеннях Bounce Rate, зокрема в діапазоні від 0.6 до 0.8, що свідчить про більший рівень відмов у цьому кластері.

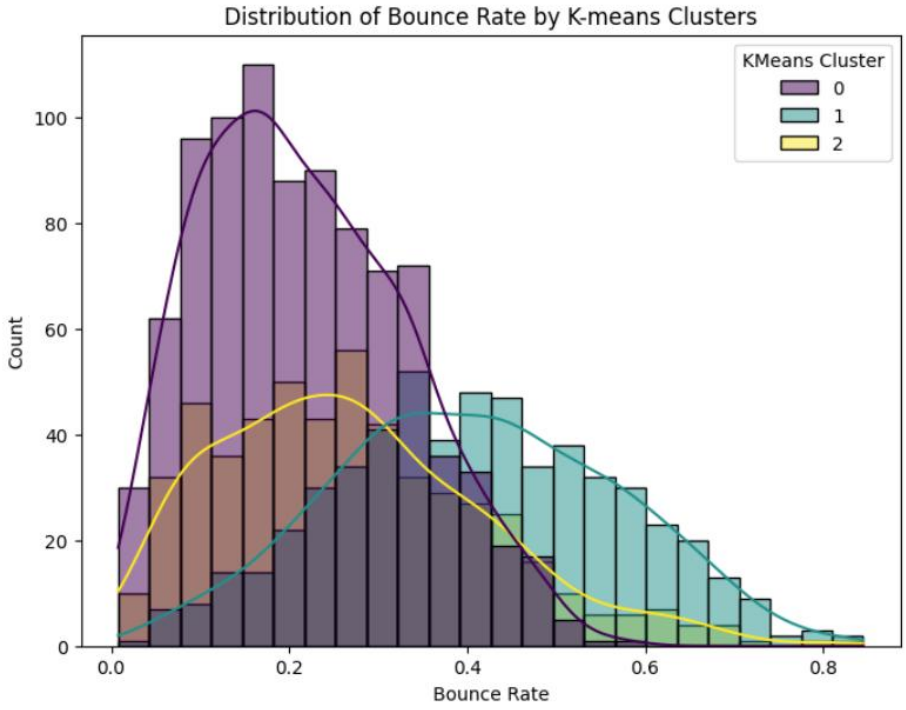

Рисунок 3.7 – Гістограма розподілу "Bounce Rate" кластерами за допомогою метода K-means

Boxplot, який зображений на рисунку 3.8, порівнює розподіл значень "Page Views" для трьох кластерів, визначених за допомогою алгоритму K-means. На осі X відображаються кластери (0, 1, 2), а на осі Y – кількість переглядів сторінок.

Для кожного кластера зображено коробку, яка представляє інтерквартильний діапазон (між першим та третім квартилем), а середня лінія в коробці вказує на медіану значень. Верхня та нижня лінії коробки – це межі діапазону, що охоплюють основну масу даних, а точки, що знаходяться поза межами цих ліній, є викидами, тобто аномальними значеннями.

Кластер 0 має найменший інтерквартильний діапазон, з основною масою значень близько до 4 переглядів, але з наявністю кількох викидів, що мають великі значення. Кластер 1 має більш широкий діапазон значень з основною масою даних між 4 і 6 переглядами. Кластер 2 демонструє більший діапазон значень, включаючи кілька високих викидів, що свідчить про наявність деяких користувачів з великими показниками переглядів.

Цей графік допомагає оцінити, як різняться значення "Page Views" для користувачів у кожному кластері.

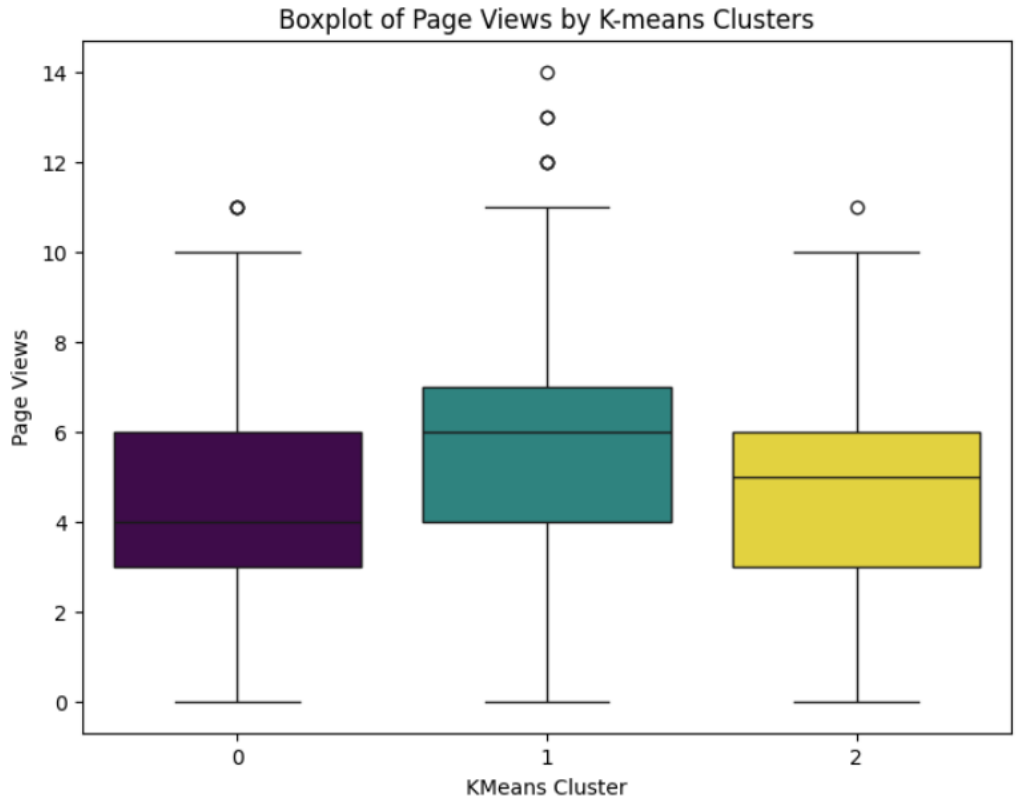

Рисунок 3.8 – Boxplot для порівняння кластерів "Page Views" за допомогою метода K-means

Boxplot, який зображений на рисунку 3.9, порівнює розподіл значень для трьох кластерів "Session Duration", визначених алгоритмом K-means. На осі X розташовані кластери (0, 1, 2), а на осі Y – тривалість сеансу.

Кластер 0 (фіолетовий) має значення тривалості сеансу, зосереджені в основному в діапазоні до 5 одиниць часу, з деякими викидами, що мають тривалість понад 10 одиниць часу. Кластер 1 має ширший інтерквартильний діапазон, з більшою кількістю користувачів, чия тривалість сеансу варіюється від 4 до 8 одиниць часу, але також з деякими викидами, що вказують на значно більші тривалості сеансів. Кластер 2 демонструє найбільший діапазон, що включає значну кількість користувачів з короткими сеансами, але також містить дуже великі значення тривалості сеансів, зокрема понад 15 одиниць часу, що вказує на присутність користувачів, які проводять значно більше часу на сайті.

Цей графік допомагає зрозуміти, як розподіляється тривалість сеансів серед різних кластерів і показує, що користувачі в кожному кластері мають різний рівень залученості до вебсайту.

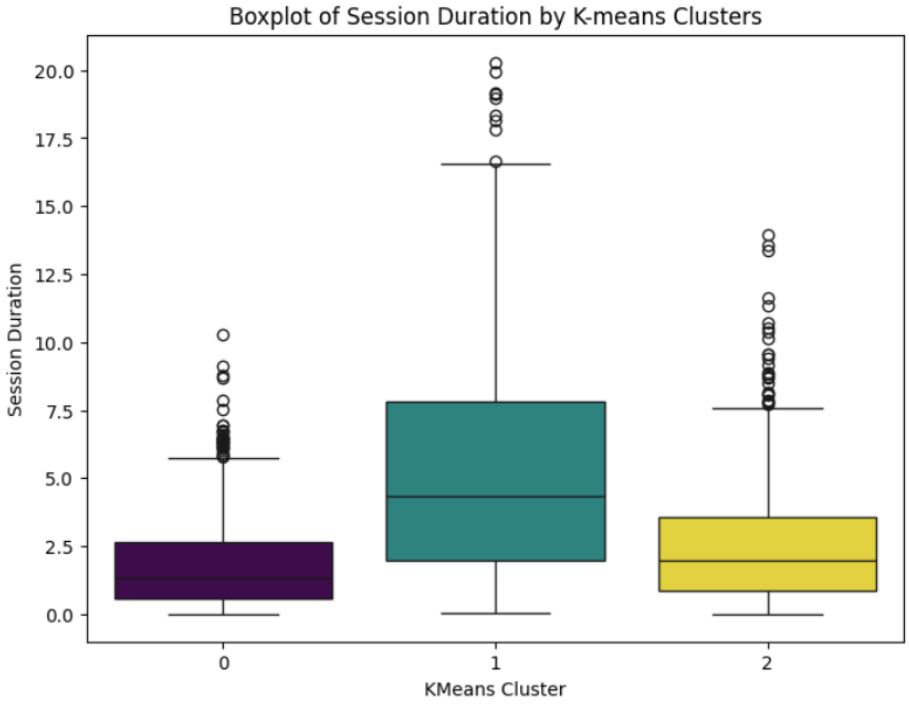

Рисунок 3.9 – Boxplot для порівняння кластерів "Session Duration" за допомогою метода K-means

Boxplot, який зображений на рисунку 3.10, порівнює розподіл показника "Bounce Rate" (відсоток відмов) для кожного з трьох кластерів, визначених за допомогою алгоритму K-means. На осі X розташовані кластери (0, 1, 2), а на осі Y – Bounce Rate.

Кластер 0 має низький середній Bounce Rate, з основною масою даних між 0.2 і 0.3, що свідчить про те, що користувачі з цього кластеру залишають сайт після перегляду кількох сторінок. Кластер 1 має більший діапазон значень, з основною масою даних у межах 0.2–0.5, що вказує на середній рівень відмов серед користувачів цього кластера. Кластер 2 показує середні значення для Bounce Rate в діапазоні близько до 0.3, але з деякими викидами, що мають високі значення відмов, близько 0.6.

Цей графік допомагає оцінити, наскільки високий рівень відмов у користувачів різних кластерів, а також показує, як варіюється цей показник між кластерами.

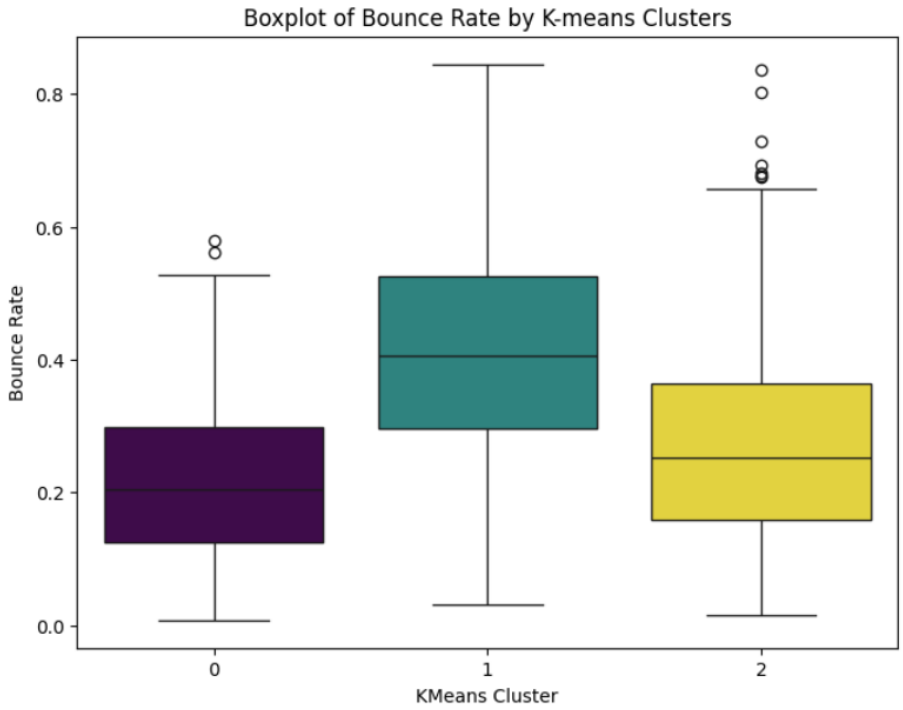

Рисунок 3.10 – Boxplot для порівняння кластерів "Bounce Rate" за допомогою метода K-means

Графік на рисунку 3.11 показує результати кластеризації методом DBSCAN, де для візуалізації виділені шуми. На осі X та Y представлені дві головні компоненти (PCA), що дозволяє зобразити дані в двовимірному просторі. Точки на графіку відображають різні кластери, визначені DBSCAN. Кластери позначені різними кольорами: фіолетовий для кластера 0, пурпурний для кластера 2, жовтий, оранжевий та рожевий для інших кластерів (5, 7, 10, 12), а сірі точки позначають шуми, які не належать до жодного з кластерів.

Графік дає змогу оцінити, як DBSCAN визначає різні групи даних, а також виділяє шуми, що знаходяться поза межами будь-яких кластерів. Це дозволяє побачити, як алгоритм обробляє щільні області та розподіляє дані на кластери, а також як відрізняються ці області по своїй щільності.

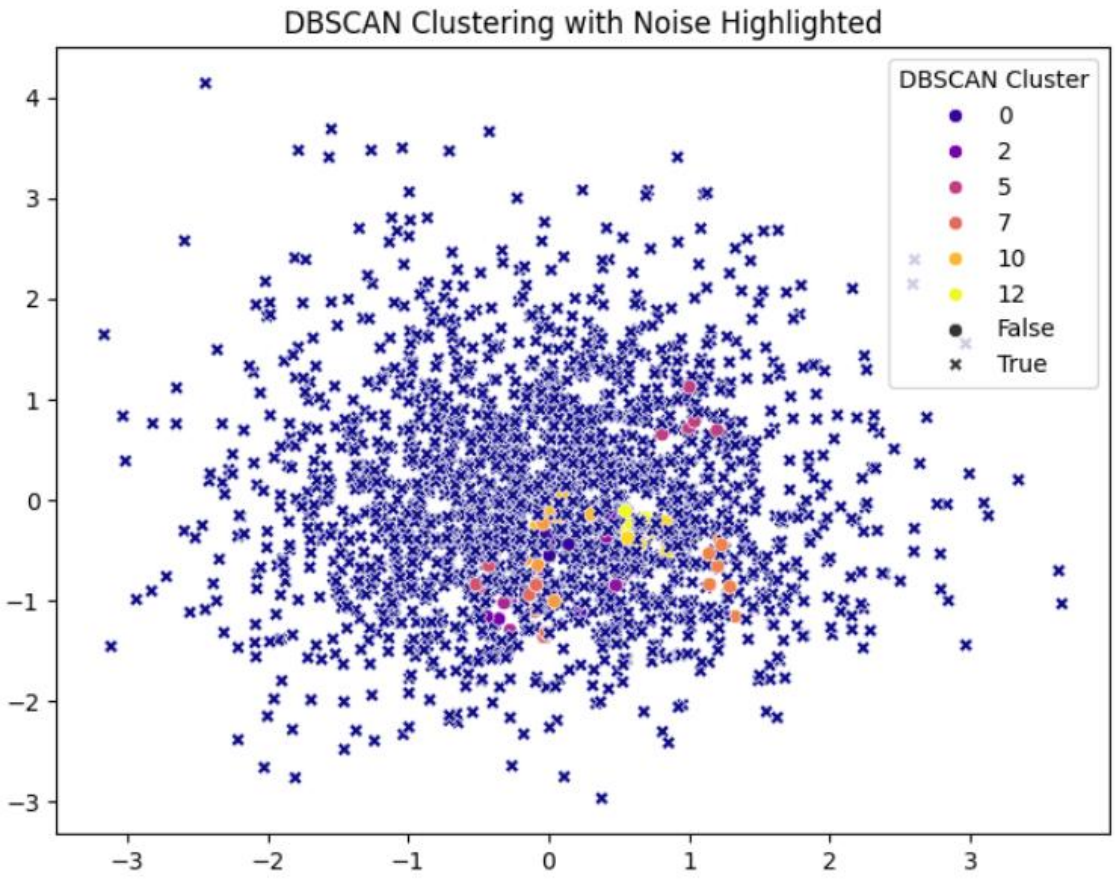

Рисунок 3.11 – Графік результатів кластеризації методом DBSCAN

За результатами кластеризації можна виділити чотири основних групи користувачів за їх поведінковими характеристиками.

Активні користувачі з високим рівнем взаємодії та конверсії (Кластер 0 за K-means і DBSCAN) – користувачі мають низький "Bounce Rate" і високий "Conversion Rate", що свідчить про їх високу зацікавленість і ефективність у виконанні бажаних дій на сайті. Вони мають середній час перебування на сайті і відвідують кілька сторінок. Це, ймовірно, ті користувачі, які знаходять потрібну інформацію або здійснюють покупки на сайті.

Користувачі з високою зацікавленістю, але з високим "Bounce Rate" (Кластер 1 за K-means) – користувачі проводять більше часу на сайті, але також мають вищий "Bounce Rate", що може свідчити про наявність проблем у контенті або навігації сайту, через які вони залишають сторінки після перегляду лише кількох з них. Вони можуть бути зацікавлені у певному контенті, але не досягають конверсії.

Малозалучені користувачі (Кластер 2 за K-means і DBSCAN) – група характеризується середніми показниками у більшості метрик. Вони мають помірний час перебування на сайті, відвідують кілька сторінок, але їх "Bounce Rate" може бути трохи вищим, ніж у більш зацікавлених груп. Ці користувачі можуть бути менш активними або менш зацікавленими в контенті.

Низькоактивні користувачі з низьким рівнем взаємодії (Кластери 3-12 за DBSCAN) – користувачі мають низький "Session Duration" та "Page Views", що свідчить про їх мінімальну активність на сайті. Вони можуть відвідувати сайт на короткий час або лише для перегляду кількох сторінок, не досягаючи значної взаємодії або конверсії. Ці користувачі можуть бути менш залученими до контенту або не знайти того, що шукали, що веде до їх швидкого покидання сайту.

Ці групи можуть допомогти визначити, які користувачі є найбільш цінними для бізнесу баланс між інформативністю та етичною обґрунтованістю методів відстеження, а також виявити проблеми з

утримання користувачів або виявити шляхи для оптимізації сайту та маркетингових стратегій для покращення взаємодії з користувачами.

3.7 Перспективи дослідження

Перспективи розвитку даного дослідження охоплюють кілька ключових напрямків, які дозволяють не лише поглибити існуючі підходи до кластеризації користувачів, але й значно покращити точність та ефективність аналізу. Одним із основних напрямків є вивчення впливу різних факторів на результативність класифікації. Особливо важливо дослідити, як зміна обсягів даних, якість вхідної інформації та вибір характеристик для класифікації можуть впливати на ефективність алгоритмів. Це дозволить розробити більш адаптивні стратегії кластеризації, які враховують специфіку даних і підвищують точність результатів. Наприклад, оптимізація вибору характеристик може істотно покращити класифікацію користувачів, якщо правильно визначити найбільш значущі фактори для кожного конкретного випадку.

Ще одним важливим напрямком є дослідження застосування методів глибокого навчання для покращення класифікації користувачів. Використання нейронних мереж для виявлення більш складних патернів у даних може суттєво підвищити точність кластеризації. Нейронні мережі, зокрема рекурентні та згорткові, можуть ефективно обробляти великі обсяги даних, виявляючи приховані зв’язки, які можуть бути неочевидними для традиційних алгоритмів машинного навчання. Вони дозволяють не лише класифікувати користувачів за явними ознаками, але й розпізнавати складніші взаємозв’язки в поведінці, що відкриває нові можливості для персоналізації.

Додатковим напрямком є розробка адаптивних методів кластеризації, які зможуть динамічно змінювати свої алгоритми в залежності від змін у

поведінці користувачів. В умовах постійних змін в інтернет-середовищі, коли користувачі можуть змінювати свої патерни взаємодії з вебсайтами, важливо мати можливість швидко адаптувати алгоритми кластеризації до нових умов. Це дозволить забезпечити більш гнучкий та актуальний підхід до аналізу даних, що є надзвичайно важливим для підтримки високої ефективності маркетингових кампаній та покращення взаємодії з користувачами.

Нарешті, важливим етапом є проведення оцінки результативності розроблених методів на великих обсягах даних. Вивчення їхньої стабільності та точності в умовах змінних факторів дозволить отримати більш точні результати та знизити ймовірність помилок класифікації. Це, у свою чергу, підвищить стабільність класифікації і дозволить краще передбачати поведінку користувачів, що позитивно вплине на персоналізацію контенту та оптимізацію маркетингових стратегій. Крім того, проведення таких досліджень дозволить зібрати важливі інсайти, що можуть бути використані для вдосконалення не лише технічних методів, але й загальної стратегії розвитку вебресурсів.

Загалом, результати дослідження відкривають широкі перспективи для подальшого розвитку технологій відслідковування користувачів та кластеризації, які можуть суттєво покращити персоналізацію контенту на вебсайтах. Це дозволить компаніям значно підвищити ефективність своїх онлайн-сервісів, сприяючи їхній конкурентоспроможності на ринку та забезпечуючи кращу взаємодію з користувачами.

ВИСНОВКИ

В результаті проведеного дослідження та реалізації методу відстеження та кластеризації користувачів на вебсайті було досягнуто кількох важливих результатів.

Було виявлено, що використання методів аналізу поведінки користувачів дозволяє суттєво покращити взаємодію з відвідувачами сайту. Кластеризація користувачів на основі їхніх поведінкових патернів дає змогу сегментувати аудиторію на різні групи, що дозволяє персоналізувати досвід користувачів та запропонувати їм більш релевантний контент чи функціональність. Це, в свою чергу, сприяє підвищенню ефективності взаємодії та покращенню конверсії.

Метод K-means був обраний для кластеризації, оскільки він є простим у застосуванні і ефективним для обробки великих обсягів даних. Алгоритм дозволяє швидко і точно поділити користувачів на групи за їхньою поведінкою на сайті. Хоча метод має певні обмеження, зокрема потребу в попередньому визначенні кількості кластерів, його застосування довело свою ефективність для задачі сегментації користувачів.

Важливим аспектом роботи було забезпечення конфіденційності даних. Оскільки всі дані, що збираються під час відстеження, є анонімними і не включають особисту інформацію користувачів, метод відповідає вимогам сучасних стандартів захисту персональних даних, таких як GDPR. Це дозволяє безпечно використовувати методи відстеження користувачів, зберігаючи при цьому довіру аудиторії.

Наукова новизна роботи полягає у впровадженні комплексного підходу до кластеризації користувачів, який базується на використанні сучасних бібліотек Python, таких як scikit-learn, Pandas та Matplotlib, у поєднанні з даними, отриманими через Google Analytics API. Вперше було створено детальну модель аналізу поведінкових даних, яка дозволяє визначати групи

користувачів із високою точністю, враховуючи такі параметри, як час сесії, кількість переглядів сторінок та відсоток відмов. Крім того, запропоновані візуалізації результатів кластеризації, такі як теплові карти та двовимірні графіки, значно спрощують аналіз отриманих даних і забезпечують їх наочне представлення для подальшого використання.

Результати роботи підтверджують доцільність використання методу відстеження та кластеризації для аналізу поведінки користувачів на вебсайтах. Цей підхід дозволяє оптимізувати інтерфейс і контент сайту, що сприяє поліпшенню досвіду користувачів та підвищенню ефективності вебресурсів.

Результати дослідження апробовано у вигляді тез доповідей під час XII Міжнародній науково-практичній конференції «Prospective directions of modern science and education in the world» [41].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Гороховатський, В. О., & Творошенко, І. С. (2021) Методи інтелектуального аналізу та оброблення даних: навч. Посібник, 92 с.

2. Кравець, Р. А. (2020) Дослідження і реалізація методу інтелектуального аналізу даних для вирішення маркетинговиїх завдань, 81 с.

3. Толкачов М.А. Розробка та дослідження методу шифрування даних в сучасних Web-застосунках : кваліфікаційна робота другого (магістерського) рівня вищої освіти: 122 Комп'ютерні науки. Харків, 2019, 60 с.

4. Толмачова Т.В. (2018) Аналіз даних активності веб-користувачів: кваліфікаційна робота другого (магістерського) рівня вищої освіти: 122 Комп'ютерні науки, 62 с.

5. Шафроненко А.Ю. (2017) Комплекс навчально-методичного забезпечення навчальної дисципліни "Математична логіка та теорія алгоритмів" підготовки бакалавра спеціальності. Харків: ХНУРЕ, 202 с.

6. Шевченко, С. М., Жданова, Ю. Д., Спасітєлєва, С. О., Мазур, Н. П., Складанний, П. М., & Негоденко, В. П. (2024) Математичні методи в кібербезпеці: кластерний аналіз та його застосування в інформаційній та кібернетичній безпеці. *Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка», 23*(3), 258-273.

7. Булгар, М. М. (2018). Кластеризація користувачів за їх інтересами, 392 с.

8. Скрит, І. П. (2024) Розробка та дослідження рекомендаційної системи на основі кластеризації користувачів, 69 с.

9. Mayer-Schönberger, V., & Cukier, K. (2016) *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin.

10. Qi, J., Yang, P., Fan, D., & Deng, Z. (2015, October). A survey of physical activity monitoring and assessment using internet of things technology. *International Conference on Computer and Information Technology; Ubiquitous*

Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing , pp. 2353-2358.

11. Pham, L. (2020) Real user monitoring for internal web application, 50 c.

12. Atterer, R., Wnuk, M., & Schmidt, A. (2015, May) Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *Proceedings of the 15th international conference on World Wide Web* (pp. 203-212).

13. González-Serrano, L., Talón-Ballestero, P., Muñoz-Romero, S., Soguero-Ruiz, C., & Rojo-Álvarez, J. L. (2020). A big data approach to customer relationship management strategy in hospitality using multiple correspondence domain description. *Applied Sciences*, *11*(1), 256.

14. Pan, R., & Ruiz-Martínez, A. (2023). Evolution of web tracking protection in Chrome. *Journal of Information Security and Applications*, *79 c*.

15. Bielova, N. (2017, October). Web tracking technologies and protection mechanisms. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 2607-2609).

16. Li, H., Yu, L., & He, W. (2019). The impact of GDPR on global technology development. *Journal of Global Information Technology Management*, *22*(1), 1-6.

17. Chasovskyi, D. Platform for a neural machine translation system demo and user data collection, 41c.

18. Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

19. Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014, February). DBSCAN: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pp. 232-238.

20. Zhang, L., Deng, S., & Li, S. (2017) Analysis of power consumer behavior based on the complementation of K-means and DBSCAN. *Conference on energy internet and energy system integration (EI2)*, pp. 1-5.

21. Silaparasetty, N., & Silaparasetty, N. (2020) Python programming in Jupyter notebook. *Machine Learning Concepts with Python and the Jupyter Notebook Environment: Using Tensorflow 2.0*, 119-145.

22. McKinney, W., & Team, P. D. (2015) Pandas-Powerful python data analysis toolkit. *Pandas–Powerful Python Data Analysis Toolkit*, 297c.

23. Hao, J., & Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348-361.

24. Yim, A., Chung, C., & Yu, A. (2018) *Matplotlib for Python Developers: Effective techniques for data visualization with Python*, 284 c.

25. Englehardt, S., & Narayanan, A. (2016) Online tracking: A 1-million-site measurement and analysis Draft: July 11th, 2016. *Technical Report*, 14 c.

26. Binns, R., Lyngs, U., Van Kleek, M., Zhao, J., Libert, T., & Shadbolt, N. (2018) Third party tracking in the mobile ecosystem. In *Proceedings of the 10th ACM Conference on Web Science*, pp. 23-31.

27. The daily traffic on a website "Website Traffic Data". URL: https://www.kaggle.com/datasets/sandeepkumar69/website-traffic-data (дата звернення 17.09.2024).

28. Binns, R., Lyngs, U., Van Kleek, M., Zhao, J., Libert, T., & Shadbolt, N. (2018). Third party tracking in the mobile ecosystem. In *Proceedings of the 10th ACM Conference on Web Science*, pp. 23-31.

29. Doroshenko, I., Knihnitska, T., & Kreshtanovych, M. (2024, January). COMPARISON OF DATA CLUSTERING ALGORITHMS. In *Sworld-Us Conference proceedings* (No. usc22-01, pp. 32-38).

30. Zakharov, K. (2016). Application of k-means clustering in psychological studies. *Tutorials in Quantitative Methods for Psychology*, *12*(2), 87-100.

31. Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 1-21.

32. General Data Protection Regulation (GDPR) – Legal Text. URL:

https://gdpr-info.eu/ (дата звернення 28.09.2024).

33. Kuner, C., Bygrave, L., Docksey, C., & Drechsler, L. (2020). *The EU general data protection regulation: a commentary*, 332 с.

34. Binns, R. "Fair processing in the age of big data". International Journal of Information Management. Feb 2020. URL: https://www.researchgate.net/publication/350408099 (дата звернення 29.09.2024).

35. Jernejcic, T., & Kettani, H. (2019, October). On the Intersection of Big Data and Privacy. In *Proceedings of the 4th International Conference on Big Data and Internet of Things*, pp. 1-4.

36. Miller, K. M., Lukic, K., & Skiera, B. (2024) The Impact of the General Data Protection Regulation (GDPR) on Online Tracking, 98 с.

37. Jahani, H., Jain, R., & Ivanov, D. (2023) Data science and big data analytics: A systematic review of methodologies used in the supply chain and logistics research, 58с.

38. Bartolini, C., Lenzini, G., & Robaldo, L. (2019) The DAta Protection REgulation COmpliance Model, 45 с.

39. Bekavac, I., & Garbin Praničević, D. (2015) Web analytics tools and web metrics tools: An overview and comparative analysis. *Croatian Operational Research Review*, *6*(2), 373-386.

40. Padmaja, S., & Sheshasaayee, A. (2016) Clustering of user behaviour based on web log data using improved K-means clustering algorithm. *International Journal of Engineering and Technology (IJET)*, *8*(1), 305-310.

41. Білоус А.М. (2024) ДОСЛІДЖЕННЯ ТА РОЗРОБКА МЕТОДІВ ВІДСТЕЖЕННЯ ТА КЛАСТЕРИЗАЦІЇ КОРИСТУВАЧІВ НА ВЕБСАЙТІ. In *The XII International Scientific and Practical Conference «Prospective directions of modern science and education in the world», November 19-22, 2024, Rotterdam, Netherlands*, p. 35.