Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя
(повне найменування вищого навчального закладу)
Факультет комп'ютерно-інформаційних систем і програмної інженерії
(назва факультету)
Кафедра комп'ютерних систем та мереж
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістра
(освітній ступінь))

на тему: **Методи та засоби сегментації множини користувачів при проектуванні та експлуатації комп'ютерних маркетингових систем**

Виконав: студент (ка) 6 курсу, групи СІм-61
спеціальності 123 «Комп'ютерна інженерія»
(шифр і назва спеціальності)

	(підпис)	Шуптарський В.В. (прізвище та ініціали)
Керівник	(підпис)	Яцишин В.В. (прізвище та ініціали)
Нормоконтроль	(підпис)	Тиш Є.В. (прізвище та ініціали)
Завідувач кафедри	(підпис)	Осухівська Г.М. (прізвище та ініціали)
Рецензент	(підпис)	(прізвище та ініціали)

Тернопіль
2021

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя
(повне найменування вищого навчального закладу)
Факультет комп'ютерно-інформаційних систем і програмної інженерії
Кафедра комп'ютерних систем та мереж

ЗАТВЕРДЖУЮ
Завідувач кафедри *Осухівська Г.М.*
«_____» _____ 2020 р.

З А В Д А Н Н Я

НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня *магістр*
(назва освітнього ступеня)

за спеціальністю *123 «Комп'ютерна інженерія»*
(шифр і назва спеціальності)

студенту *Шуптарському Володимиру Вікторовичу*
(прізвище, ім'я, по-батькові)

1. Тема проекту (роботи) *Методи та засоби сегментації множини користувачів при проектуванні та експлуатації комп'ютерних маркетингових систем*
Керівник проекту (роботи) *Яцишин Василь Володимирович, к.т.н., доц.*
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
Затверджені наказом ректора від «*28*» *жовтня* 2021 року № *4/7-916*
2. Термін подання студентом завершеної роботи
3. Вихідні дані до роботи *Тип комп'ютерних систем, дані про користувачів, математичні методи і моделі сегментації, мова програмування Python*
4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ. 1. Аналіз сучасних досліджень у галузі проектування маркетингових комп'ютерних систем 2. Проектування архітектури та розробка алгоритму сегментації користувачів 3. Програмна реалізація алгоритму сегментації користувачів і товарів при проектуванні маркетингових комп'ютерних систем. 4. Охорона праці та безпека в надзвичайних ситуаціях. Висновки
5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1. Актуальність і мета дослідження. 2. Задачі дослідження, , об'єкт і предмет, наукова новизна і практична цінність дослідження. 3. Напрями досліджень щодо автоматизації процесів у сфері маркетингу. 4. Структура процесу сегментації при проектуванні маркетингових комп'ютерних систем. 5. Методи сегментації з використанням машинного навчання 6. Архітектура комп'ютерної системи сегментації множини покупців. 7. Результати сегментації множини покупців у маркетингових комп'ютерних системах 8. Висновки

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці та	*Осухівська Г.М.*		
безпека в надзвичайних	*Стадник І.Я.*		
ситуаціях			

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	*Аналіз сучасних досліджень у галузі проектування*	*28.10.2021-*	*виконано*
	маркетингових комп'ютерних систем	*13.11.2021*	
2.	*Проектування архітектури та розробка алгоритму*	*13.11.2021 –*	*виконано*
	сегментації користувачів	*20.11.2021*	
3.	*Програмна реалізація алгоритму сегментації*	*21.11.2021 –*	*виконано*
	користувачів і товарів	*28.11.2021*	
4.	*Охорона праці та безпека в надзвичайних ситуаціях*	*28.11.2021 – 02.12.2021*	*виконано*
5.	*Оформлення пояснювальної записки*	*03.12.2021- 06.12.2021*	*виконано*
6.	*Оформлення графічного матеріалу*	*07.12.2021- 11.12.2021*	*виконано*
7.	*Попередній захист кваліфікаційної роботи магістра*	*15.12.2021*	*виконано*
8.	*Захист кваліфікаційної роботи магістра*		

Студент _____ *Шуптарський В.В.*
(підпис) (прізвище та ініціали)

Керівник проекту (роботи) _____ *Яцишин В.В.*
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Тема кваліфікаційної роботи: “Методи та засоби сегментації множини користувачів при проектуванні та експлуатації комп’ютерних маркетингових систем ” // Кваліфікаційна робота // Шуптарський Володимир Вікторович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп’ютерно-інформаційних систем та програмної інженерії, група СІм-61 // Тернопіль, 2021 // с. – 92, рис. – 45, табл. – 5, аркушів А1 – 8, додат. – 1, бібліогр. – 25.

Ключові слова: метод, сегментація, користувач, комп’ютерна система, маркетинг, експлуатація.

Мета кваліфікаційної роботи магістра полягає у дослідженні методів і засобів сегментації користувачів при проектуванні та експлуатації маркетингових систем.

У роботі запропоновано метод сегментації користувачів електронної комерції, що враховує особливості товарів і послуг, а також споживчу поведінку покупців, що дало змогу автоматизувати та інтегрувати модуль кластеризації у маркетингові комп’ютерні системи та використовувати дані з CRM-систем і систем обліку показників ефективності бізнес-діяльності..

Застосовано алгоритм кластеризації k-means, що враховує критерії опису товарів при виявленні груп подібних продуктів, а також дозволяє проводити сегментацію користувачів, кластери яких враховують тип продуктів, кількість відвідувань сайту електронної комерції і витрачені суми протягом 10 місяців, що в подальшому дає змогу проводити прогнозування і класифікацію нових покупців на основі вказаних критеріїв та визначених кластерів.

За допомогою мови програмування Python реалізовано процедури препроцесингу даних та алгоритм кластеризації k-means, який дозволяє проводити сегментацію користувачів і товарів без міток категорій та приймати ефективні управлінські рішення при маркетингових дослідженнях.

ABSTRACT

The theme of the thesis: " Methods and tools for multiple users segmentation in the design and use of computer marketing systems " /Master thesis / Shuptarskyi Volodymyr Viktorovych/ Ternopil Ivan Pul'uj National Technical University, Faculty of Computer Information Systems and software engineering, group CIm -61 // Ternopil, 2021// p. - 92, fig. – 45, table. – 5, Sheets A1 – 8, Add – 1, Ref. – 25.

Keywords: method, segmentation, user, computer system, marketing, operation.

The purpose of the master's thesis is to study the methods and means of user segmentation in the design and operation of marketing systems.

The paper proposes a method of segmentation of e-commerce users, which takes into account the characteristics of goods and services, as well as consumer behavior of customers, which allowed to automate and integrate the clustering module into marketing computer systems and use data from CRM-systems and business performance indicators. activities .

The k-means clustering algorithm is used, which takes into account the criteria of product description when identifying groups of similar products, and also allows segmentation of users whose clusters take into account product type, number of e-commerce site visits and amounts spent within 10 months, which allows forecasting and classification of new customers based on specified criteria and defined clusters.

Using the Python, data preprocessing procedures and the k-means clustering algorithm are implemented, which allows segmenting users and products without category labels and making effective management decisions in marketing research.

ЗМІСТ

ПЕРЕЛІК ОСНОВНИХ УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ І СКОРОЧЕНЬ

ACF	AutoCorellation Function
ANN	Artificial Neuro Network
CRM	Customer Relationship Management
PCA	Principal Component Analysis
UML	Unified Modeling Language
БД	База Даних
КС	Комп’ютерна Система
ПЗ	Програмне забезпечення

ВСТУП

Актуальність теми. Методи і засоби конкуренції на ринку товарів і послуг, а відповідно і боротьба за користувачів на сьогодні стає все гострішою і вимагає залучення додаткових фінансових, часових та інформаційних ресурсів. При цьому важливим є інтеграція маркетингових систем або надбудов в існуючі системи обліку та аналізу комерційної діяльності підприємства.

Сьогодні на ринку існує багато готових коробкових рішень, які дозволяють одночасно керувати як бізнес-процесами підприємства, так і додатковими й організаційними, до яких належать задачі маркетингу. Оскільки, такі рішення є загальними і практично універсальними для усіх гравців ринку, то виникає необхідність щодо гнучкості налаштування під потреби конкретної бізнес-структури. А це в свою чергу вимагає залучення додаткових фінансових ресурсів, або придбання додаткових сервісів. Тому актуальною задачею сьогоднішнього часу є створення окремих ізольованих сервісів, які б надавали можливість гнучко, прозоро та з використанням мінімальних ресурсів вирішувати задачі проектування комп’ютерних маркетингових метасистем. Особливою рисою таких систем, на відміну від існуючих, повинна бути можливість їхнього використання за потребою і незалежно від типу облікових чи CRM-систем.

Важливою задачею при проектуванні та експлуатації маркетингових систем є забезпечення можливості сегментації користувачів або споживачів продукції для формування цільових пропозицій і плану проведення рекламних та інших видів промоакцій. Також наявність такої інформації є доцільною у процесі аналізу ризиків, при залученні нових клієнтів, розширенні сегменту товарів і послуг. Тому актуальною задачею у сфері побудови комп’ютерних маркетингових систем є обґрунтування методів і засобів сегментації користувачів з використання елементів машинного навчання, які б забезпечували необхідну точність та стійкість результатів виявлення груп подібних споживачів, а також ефективність і продуктивність їх формування.

Мета і задачі дослідження. Мета роботи полягає у дослідженні методів і засобів сегментації користувачів при проектуванні та експлуатації маркетингових систем.

Для досягнення вказаної мети в роботі поставлено наступні **задачі**:

– проведення аналізу наукових публікацій і практик щодо побудови маркетингових комп'ютерних систем;

– обґрунтування та формалізація критеріїв сегментації покупців і товарів;

– обґрунтування методів штучного інтелекту для виявлення груп подібних товарів і користувачів;

– проектування архітектури маркетингової комп'ютерної системи для вирішення задач сегментації ринку;

– імплементація програмної моделі сегментації користувачів і товарів із застосуванням методів кластерного аналізу;

– верифікація і валідація результатів сегментації при проектуванні та експлуатації маркетингових комп'ютерних систем.

Об'єкт дослідження: процеси проектування та експлуатації маркетингових систем.

Предмет дослідження: критерії, моделі, методи і засоби сегментації користувачів.

Методи дослідження: При розв'язанні задач дипломного проектування використано наступні методи:

– аналіз та узагальнення – при дослідженні методів і засобів проектування та експлуатації маркетингових комп'ютерних систем;

– машинного навчання – при побудові та верифікації алгоритмів кластеризації товарів і користувачів;

– проектування та програмування – при побудові архітектури маркетингових комп'ютерних систем та програмуванні алгоритму кластеризації на основі k-means;

– експеримент – при верифікації результатів сегментації товарів і користувачів.

Наукова новизна отриманих результатів. Наукова новизна результатів дослідження полягає в наступному:

– уперше запропоновано метод сегментації користувачів електронної комерції, що враховує особливості товарів і послуг, а також споживчу поведінку покупців, що дало змогу автоматизувати та інтегрувати модуль кластеризації у маркетингові комп'ютерні системи та використовувати дані з CRM-систем і систем обліку показників ефективності бізнес-діяльності.

– набув подальшого розвитку алгоритм кластеризації k-means, що враховує критерії опису товарів при виявленні груп подібних продуктів, а також проводити сегментацію користувачів, кластери яких враховують тип продуктів, кількість відвідувань сайту електронної комерції і витрачені суми протягом 10 місяців, що в подальшому дає змогу проводити прогнозування і класифікацію нових покупців на основі вказаних критеріїв та визначених кластерів.

Практичне значення одержаних результатів. Практичне значення одержаних результатів полягає у реалізації процедур препроцесингу даних та алгоритму кластеризації k-means, який дозволяє проводити сегментацію користувачів і товарів без міток категорій та приймати ефективні управлінські рішення при маркетингових дослідженнях.

Публікації. Результати кваліфікаційної роботи апробовані на X міжнародній науково - технічній конференції молодих учених і студентів «Актуальні задачі сучасних технологій» (24-25 листопада 2021 р.) Тернопільського національного технічного університету імені Івана Пулюя та на IX науково-технічній конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні моделі, системи та технології» (8-9 грудня 2021 року) як тези конференцій.

1. Яцишин В.В., Шуптарський В.В., Цісарук Д.А. Алгоритми машинного навчання для сегментації користувачів у маркетингових комп'ютерних систем. Матеріали X міжнародної науково - технічної конференції молодих учених і студентів «Актуальні задачі сучасних технологій» (24-25 листопада 2021 р.)

Тернопільського національного технічного університету імені Івана Пулюя. Тернопіль: ТНТУ. 2021. С. 145.

2. Луцків А.М., Цісарук Д.А., Шуптарський В.В. Аналіз життєвого циклу процесу тестування програмного забезпечення комп'ютерних систем. Матеріали IX науково-технічної конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні моделі, системи та технології» (8-9 грудня 2021 року). Тернопіль: ТНТУ. 2021. С. 142.

Структура роботи. Кваліфікаційна робота містить розрахунково-пояснювальну записку та графічний матеріал. До складу записки входить вступу, 4 розділи, загальні висновки, список використаних джерел і додатки. Обсяг роботи: розрахунково-пояснювальна записка – 92 арк. формату А4, графічна частина – 8 аркушів формату А1.

РОЗДІЛ 1
АНАЛІЗ СУЧАСНИХ ДОСЛІДЖЕНЬ У ГАЛУЗІ ПРОЕКТУВАННЯ МАРКЕТИНГОВИХ КОМП'ЮТЕРНИХ СИСТЕМ

1.1. Аналіз задач у сфері проектування маркетингових комп'ютерних систем

Однією з найбільш глобальних і важливих проблем при побудові маркетингових комп'ютерних систем є автоматизоване дослідження ринків товарів і послуг. Вирішення цієї проблеми потребує розв'язання наведених нижче задач:

1. Аналіз поточного стану і визначення динаміки зміни ринку у певній галузі.
2. Оцінювання потужності і перспектив ринку для конкретної групи товарів і послуг.
3. Встановлення структури і виявлення сегментів ринку, їхній детальний опис.
4. Визначення рівня свободи ринкових відносин.
5. Виявлення факторів та особливостей формування збуту і пропозицій.
6. Характеристика каналів розподілу на ринку.

Об'єктами ринкових досліджень є процеси дослідження тенденцій розвитку ринку. До предметів дослідження входять економічні, науково-технічні, демографічні, екологічні, законодавчі й інші фактори впливу.

Окрім цього варто означити структуру і локалізацію ринку, потужність, динаміку продажів, бар'єри ринку, рівень конкуренції, складність кон'юнктури, потенційні можливості та ризики.

Основні результати дослідження ринку передбачають його розвиток, оцінку кон'юнктурних тенденцій, виявлення ключових факторів успіху. Визначаються найбільш ефективні способи продажу, управління конкурентоздатністю та здатність охопити нові ринки, здійснюються вибір цільових ринків та ринкових галузей. Найбільш типові види маркетингових досліджень наведені у табл. 1.1.

Таблиця 1.1

Типові види маркетингових досліджень

Напрям дослідження	Мета	Типова тематика досліджень
Дослідження споживчого ринку	Сегментація ринку продаж, вибір цільового ринку	Дослідження реакції споживачів на різні маркетингові заходи, аналіз поведінки споживачів, мотивація та надання переваги
Дослідження властивостей ринку	Оцінювання потужності ринку	Аналіз локації і структури ринку, його потенціалу і тенденцій розвитку
Дослідження макросередовища	Оцінювання зовнішніх можливостей та ризиків	Дослідження факторів зовнішнього впливу, які найбільше впливають на підприємство (правові, економічні і т.п.)
Дослідження внутрішньої структури підприємства	Формування номенклатури товарів і послуг	Аналіз слабких і сильних сторін фірми, профілю продукції
Аналіз конкурентів	Забезпечення підприємству конкурентних переваг	Аналіз конкурентних товарів, визначення стану конкурентів на ринку, пошук шляхів «мирного співіснування»

Продовження табл. 2.1

Напрям дослідження	Мета	Типова тематика досліджень
Вивчення збуту	Побудова ефективної мережі збуту	Дослідження різних методів збуту товарів, особливості типів різних посередників
Дослідження методів просування товарів	Підвищення рівня поінформованості і лояльності до підприємства і його товарів	Дослідження сприйняття бренду споживачами, рекламних акцій, методів стимулювання збуту
Дослідження цін	Оптимізація цін	Дослідження гнучкості попиту, структури собівартості товару, динаміки і тренду змін цін товарів на ринку
Дослідження товарів	Підвищення конкурентоздатності товару	Аналіз задоволеності споживачів товаром, тестування товарів

Достовірність маркетингового дослідження повинно бути підтвердженим за допомогою аналізу наступних параметрів:

– розмір вибірки;
– відносна похибка дослідження, %;
– репрезентативність вибірки;
– відносна стійкість результатів дослідження, %.

У результаті проведення маркетингового дослідження формується звіт до складу якого входять наступні частини:

– вступ – містить опис ситуацій, проблеми, імовірних робочих гіпотез, цілі маркетингового дослідження;

– опис методів одержання інформації та формування цільової вибірки, терміни проведення дослідження;

– представлення одержаних результатів із застосуванням порівняльних методів аналізу інформації;

– формування рекомендацій щодо розв'язку проблеми (підтвердження або відкидання гіпотез);

– прикладне вирішення маркетингової проблеми шляхом створення комп'ютерної системи автоматизації.

1.2. Аналіз основних понять сегментації користувачів і ринків при побудові маркетингових систем

Загальне дослідження ринку товарів і послуг передбачає необхідність його розгляду як диференційованої структури в залежності від груп користувачів та споживачів власних товарів підприємства, що у широкому масштабі визначає поняття ринкової сегментації.

Розвиток ринку на окремих сегментах за ознаками видів товару, територіальним розташуванням, типом найбільш представлених на цій частині ринку покупців за соціальними характеристиками.

Термін «сегментація ринку» вперше запропонував У. Сміт ще у 50-х роках XX ст. Сегментація дозволяє забезпечити уточнення попиту, структурувати його, виявити відповідні умови для вибору оптимальної стратегії та тактики на ринку.

Структура сегментації ринку умовно може бути розділена на три види:

– однорідну;

– розподілену;

– групову.

При однорідній структурі вподобання споживачів ринку товарів і послуг приблизно одинакові.

Розподілена або дифузорна структура передбачає наявність зворотних до однорідної структури вподобань. Точки, що характеризують вподобання потенційної кількості користувачів розташовані по всьому ринковому просторі. Якщо на ринку кілька торгових марок, то вони, швидше за все, будуть розкидані по всьому ринковому простору. При цьому значення вподобань значно відрізнятимуться між собою і тим самим задовольнятимуть різні потреби покупців.

При груповій (кластерній) структурі вподобань, споживачі на ринку формують чіткі групи покупців з однаковими побажаннями. Компанія, яка першою входить на ринок, може вибрати різні стратегії: намагатися привернути увагу всіх груп покупців, орієнтуватися на найбільший сегмент ринку або розробити кілька торгових марок, орієнтованих на кожен ринковий сегмент.

Сегмент ринку – це сукупність споживачів, які приблизно однаково реагують на властивості товару та фактори, які стимулюють маркетинг. Сегменти ринку диференціюються за типами споживачів та відповідними характеристиками їхньої поведінки і способу мислення. У даному випадку, об'єктами сегментації є споживачі послуг і товарів. В якості критеріїв сегментації можуть виступати конкурентоспроможність, ціна, вид товару та ін. Загальна схема сегментації ринків при маркетингових дослідженнях, а відповідно і при побудові автоматизованих комп’ютерних систем підтримки таких процесів представлено на рис. 1.1.

Види діяльності щодо сегментації включають:

1. Визначення принципів сегментації, критеріїв, оцінок, формування профілів, діаграм, матриць, тобто виконується розбиття ринку на основі деякої моделі.

2. Оцінювання рівня привабливості одержаних сегментів на основі одного або декількох критеріїв.

3. Вибір одного або кількох сегментів для формування пропозиції власних товарів.

4. Прийняття рішення щодо позиціонування товарів і послуг у кожному з обраних сегментів з урахуванням порівняння і перспектив.

5. Створення стратегії маркетингу для кожного цільового сегмента.

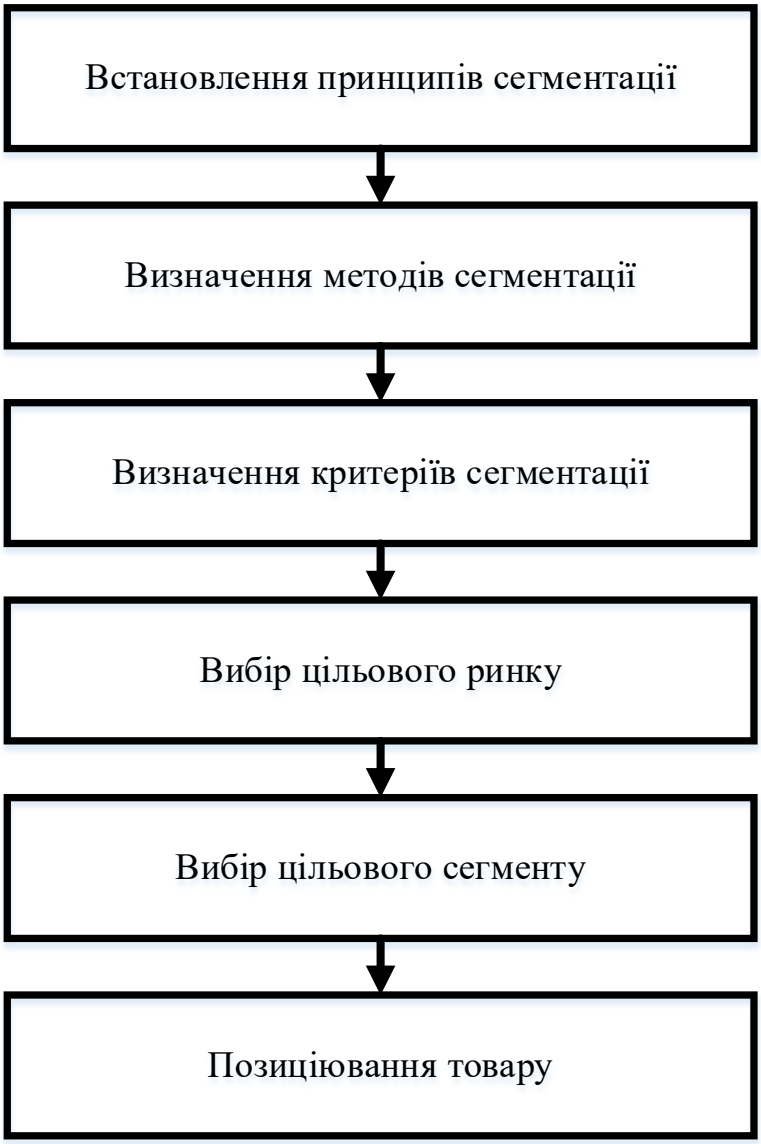


Рис. 1.1. Загальна схема сегментації

Передумовою для сегментування ринку є ідея, що покупці мають різні потреби і конкретне підприємство не в силах охопити всіх клієнтів в цілому. Для ідентифікації цільових ринків компанія звертається до цільового маркетингу, тобто виявляє основні сегменти ринку, вибирає один або кілька і розробляє відповідну маркетингову стратегію.

У філософії маркетингу можна виділити три стадії:

1) масовий маркетинг – вид маркетингової діяльності, який орієнтований на повне коло споживачів. Цей підхід є найменш витратним і створює великий потенційний ринок;

2) товарно-диференційований маркетинг має місце у випадку виробництва компанією різних видів товарів, які відрізняються якісними показниками, властивостями, різним стильовим спрямуванням;

3) цільовий маркетинг передбачає ситуацію, коли продавець виробляє розмежування між сегментами ринку, вибирає з них один або декілька і розробляє маркетингову програму для кожного з них.

Сегментування ринку представляє собою компромісний варіант між масовим маркетингом, який базується на однорідності покупців, і обмеженістю ресурсів компанії, що не має можливості (за винятком окремих випадків) розробляти маркетингові програми для кожного споживача окремо.

1.3. Аналіз вимог і принципів сегментації ринків при проектуванні комп'ютерних маркетингових систем

Для визначення потужності сегменту, який наявний на ринку, можна скористатися наступними показниками:

– сумарна кількість продукції, що реалізована протягом деякого інтервалу часу;

– сумарний обсяг реалізації товарів і послуг протягом випадково обраного періоду часу або за весь виробничий цикл продукції;

– загальна кількість можливих споживачів продукції і т.д.

Найбільш ефективним при виборі цільової продукції є сегмент ринку, що володіє найвищими кількісними показниками щодо реалізації продукції. Для забезпечення доступності сегменту на ринку, необхідним для підприємства є наявність можливостей щодо розподілу товарів, їхнього збуту, створення сприятливої транспортної логістики, наявність складських приміщень для розміщення товарів чи послуг у конкретному географічному сегменті ринку. Перспективним для розвитку вважається той сегмент ринку, який показує зростання протягом довгого періоду часу.

Важливими показниками перебування або входу підприємства на ринок є те, що його товари і послуги справді займатимуть велику частину ринку, а сам сегмент є стійким і буде залишатися таким у перспективі.

Ефективність функціонування підприємства у деякому сегменті ринку може бути визначена шляхом порівняння показників його рентабельності на протязі визначеного інтервалу часу, або з показниками іншого ринку, де присутня дана компанія. Проте, для різних підприємств та організацій показники рентабельності можуть бути різними в залежності від того, яка стратегія компанії та яка її кінцева мета.

Для визначення цільового сегменту проводиться аналіз найвищих фінансових показників, які з точки зору керівництва, є найбільш важливими для їхнього бізнесу. При вході на ринок, або при перебуванні на ньому варто оперувати інформацією щодо кількісних і якісних характеристик його сегментів, а також вимог до них. Крім того, діяльність у сегменті ринку повинна передбачати наявність засобів і каналів комунікації з кінцевими споживачами. Такі канали можуть бути як особистими, так і масовими. Окрім цього, потрібно проводити оцінювання захищеності конкретно взятого сегменту щодо поточної та імовірної конкуренції, виявити переваги і недоліки конкуруючих структур, а також встановити власні кількісні та якісні характеристики і переваги у сегменті.

Таким чином, провівши аналіз шляхом оцінювання власних потужностей виробництва та відповівши на поставлені вище запитання можна приймати зважене рішення щодо формування сегменту ринку і можливості бути гравцем на ньому.

На сьогодні використовують 5 принципів, які дають змогу сегментувати ринок, товари і послуги, а також групи кінцевих споживачів. До них належать:

1. «Відмінність між сегментами» – принцип, який передбачає, що при сегментації ринку повинні бути одержані групи різних кінцевих споживачів. Якщо результат не призвів до утворення таких кластерів, то відповідно не відбулось розбиття і як наслідок працює масовий маркетинг.

2. «Принцип схожості споживачів у рамках одного сегменту» – визначає подібність та однорідність можливих споживачів продукції з врахуванням їх

купівельної спроможності щодо одного виду товару, або групи товарів. На основі високої міри схожості покупців продукції формується план, наприклад для проведення промоакцій для усього цільового сегменту ринку.

3. «Вимога масштабності сегменту» – формує необхідність цільового сегменту бути доволі об’ємним і потужним з метою задоволення потреб підприємства у реалізації і відповідно погашенні затрат виробництва. Факторами, які впливають на величину масштабності сегменту ринку є вид і категорія товару чи послуги, а також розмір можливого ринку. Для прикладу, кількість кінцевих споживачів на одному ринку або його сегменті може становити мільйони кінцевих користувачів, як у випадку, наприклад, абонентів мобільного зв’язку, товарів широкого вжитку, а на іншому може бути на рівні сотень або десятків.

4. «Вимірюваність показників кінцевого споживача» – принцип, який орієнтований на проведення маркетингового аналізу з метою виявлення потреб існуючих та можливих споживачів, а також дослідження впливу і реакції цільового сегменту ринку на такі дії фірми. Важливість цього принципу проявляється в тому, що без взаємодії з кінцевими користувачами трудові і фінансові ресурси використовуються неефективно, оскільки просування товарів і послуг виконується наосліп.

5. «Принцип доступності кінцевого споживача» – визначає необхідність застосування каналів зв’язку підприємства з цільовою групою споживачів. В якості каналів зв’язку із споживачами можуть виступати засоби масової інформації, різного виду месенджери, веб-платформи, соціальні мережі, біг борди, сітілайти та ін. Даний принцип є досить важливим в контексті формування акційних пропозицій, підвищення продаж одного і того ж виду товару (Up-sell) та крос-продаж (Cross-selling). Доступність, або по-іншому досяжність споживачів, важлива також з точки зору поінформованості можливих споживачів щодо характеристик товарів, актуальності ціни та ін.

1.4. Аналіз підходів до сегментації товарів і послуг

Побудова маркетингових комп'ютерних систем вимагає застосування сучасних методів і засобів, які повинні бути орієнтованими на вирішення задач аналізу поведінки користувачів, формування промоакцій, сегментації товарів, послуг і користувачів. Одними із таких трендових технологій для проведення маркетингових досліджень є методи машинного навчання, зокрема, кластеризація і класифікація користувачів і товарів (рис. 1.2).

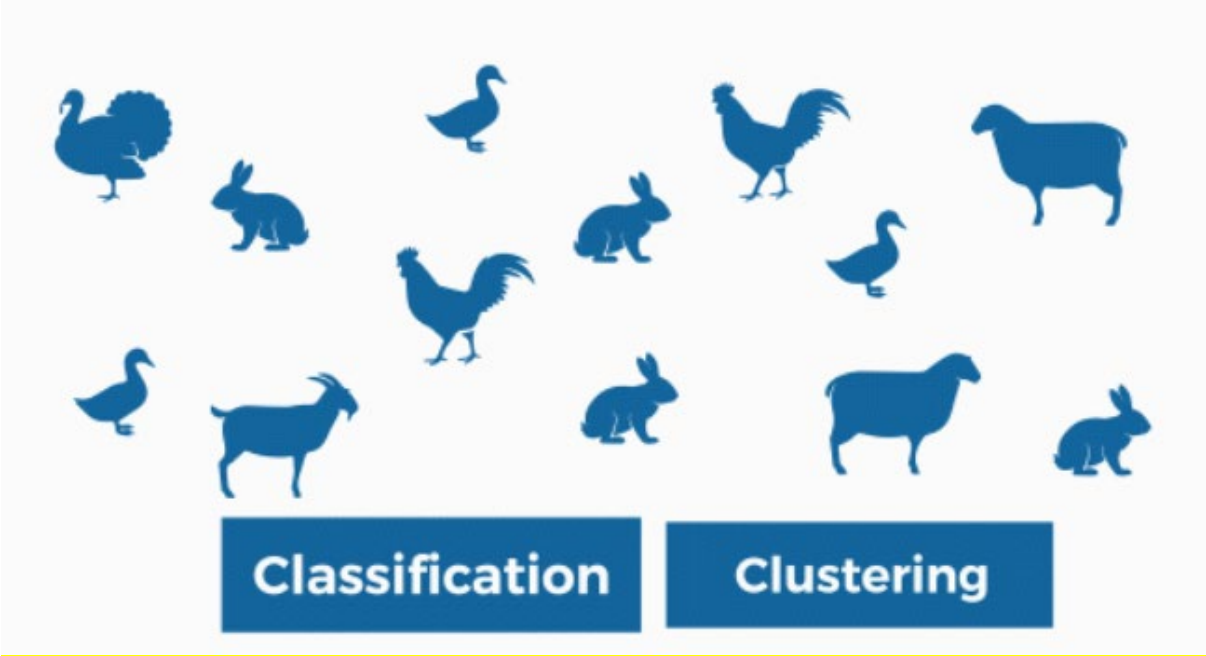


Рис. 1.2. Методи сегментації у машинному навчанні

У сфері машинного навчання кластеризація передбачає навчання без вчителя, тобто для цього типу алгоритму існує лише один набір вхідних даних без міток. Тому потрібно розв'язати задачу одержання інформації, не знаючи попередньо, яким буде вихід.

Кластеризація використовується в проектах для компаній, які хочуть виявити спільні властивості у своїх клієнтів, щоб застосувати сегментацію клієнтів, створити карти подорожей клієнтів або знайти групи та сформувати рекомендовані набори товарів чи послуг. Приклад кластерів показано на рис. 1.3.

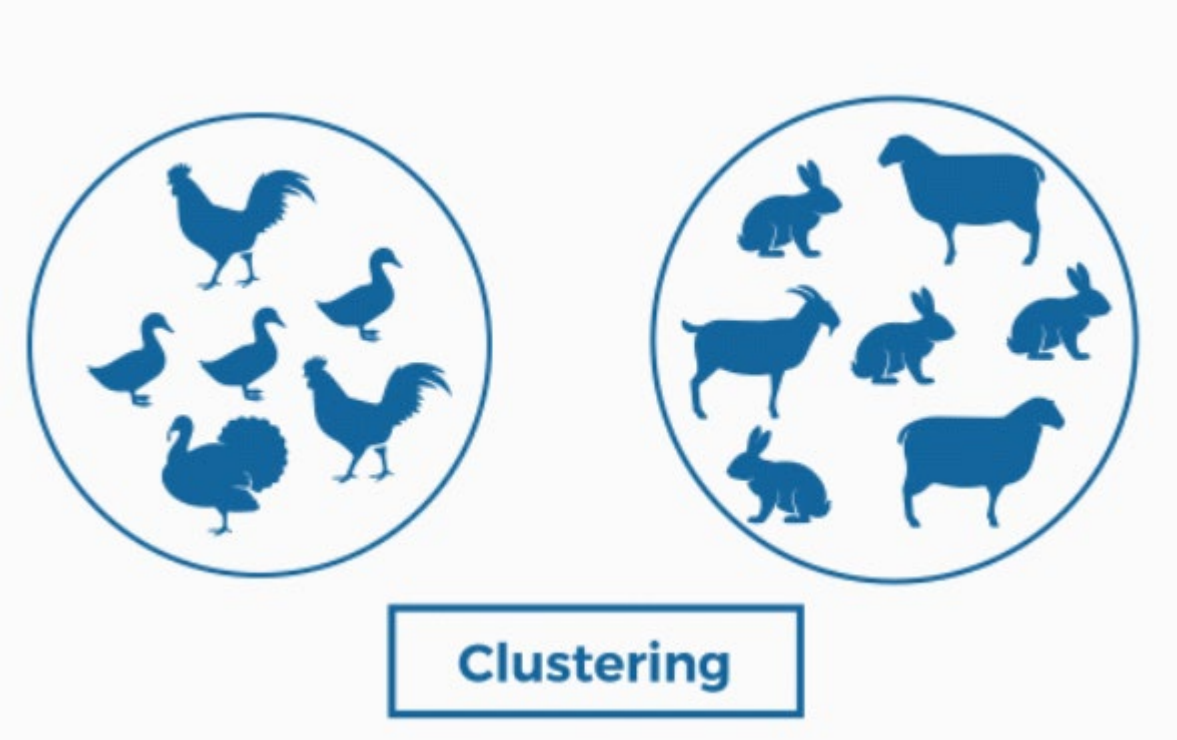


Рис. 1.3. Кластеризація об’єктів

Таким чином, якщо значний відсоток клієнтів мають певні спільні риси (вік, тип сім’ї тощо), компанія може рекомендувати проведення певної кампанії, послуги чи товару. Кластеризація також корисна для одержання загальних уявлень про інформацію якою оперує компанія.

З іншого боку, класифікація належить до алгоритмів навчання з вчителем, тобто наявний контроль за навчанням. Це означає, що вхідні дані мають мітки класів і наперед відомий можливий вихід алгоритму.

Розрізняють бінарну класифікацію, яка розв’язує задачі з категорійними відповідями (наприклад, "так" і "ні"), і мультикласифікація, для задач, де потрібно знайти більше двох класів, відповідаючи на більш відкриті відповіді, такі як «чудово », «звичайний» і «недостатній».

Класифікація використовується в багатьох галузях, наприклад у біології або в десятковій класифікації Дьюї для книг, при виявленні спаму в електронних листах. Приклад реалізації класифікації показано на рис. 1.4.

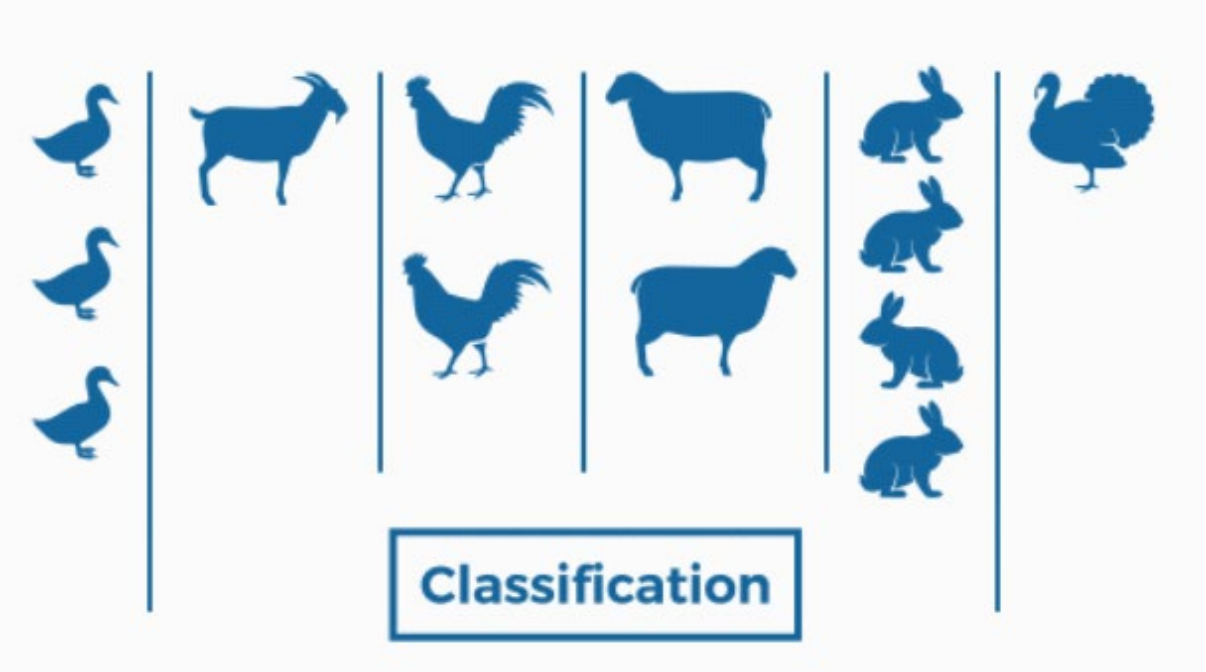


Рис. 1.4. Приклад класифікації об’єктів

У Bismart використовується класифікація і кластеризацію, які охоплені різними секторами. Наприклад, у галузі соціальних послуг використано кластеризацію для визначення груп населення, які користуються конкретними соціальними послугами. На основі даних соціальних служб можна визначити або об’єднати групи людей, які користуються подібними послугами, відповідно до їхніх атрибутів (кількість осіб, які підпорядковані їм, ступінь залежності, сімейний стан...). Таким чином, можна заздалегідь визначити, який тип послуг знадобиться новому користувачеві соціальних послуг, порівнявши їх атрибути з характеристиками кластерів.

Класифікація використовується, коли потрібно знати користувачів або клієнтів, щоб вирішити, які продукти чи кампанії будуть запущені в майбутньому.

Наприклад, у Bismart розроблено проект для страхової галузі, в якому потрібно було класифікувати клієнтів за претензіями від нещасних випадків, щоб поліс можна було класифікувати за кількістю передбачених претензій. Таким чином, компанія може вибрати споживачів з найменшою кількістю претензій.

Добре відомим застосуванням алгоритмів кластеризації є рекомендаційні системи Netflix. Хоча компанія досить стримана зі своїми алгоритмами,

підтверджено, що існує близько 2000 кластерів або спільнот, які мають спільні аудіовізуальні смаки. У кластер 290 входять люди, яким подобаються серіали «Загублені», «Чорне дзеркало» та «День бабака».

Netflix використовує ці кластери, щоб удосконалити свої знання про смаки глядачів і таким чином приймати кращі рішення при створенні нових оригінальних серіалів.

Класифікація зазвичай використовується у фінансовому секторі. В епоху онлайн-транзакцій, коли використання готівки помітно зменшилося, необхідно визначити, чи безпечні переміщення за допомогою карток. Суб'єкти можуть класифікувати операції як правильні або шахрайські, використовуючи історичні дані про поведінку клієнтів, щоб дуже точно виявляти шахраїв.

Задача, що розв’язується у кваліфікаційній роботі магістра відносно сегментації користувачів і товарів, повинна використовувати методи кластеризації, оскільки вхідні дані зазвичай не мають міток.

1.5. Висновки до розділу

1. Проведено аналіз задач у сфері проектування маркетингових комп’ютерних систем, що дало змогу виявити основні з них, зокрема, це стосується визначення трендів і динаміки зміни ринку у певній галузі, встановлення факторів впливу на розширення та стимулювання ринку, сегментації груп товарів і користувачів.

2. Проаналізовано основні поняття щодо сегментації товарів і послуг, їх позиціонування з урахуванням порівняння і перспектив, що дало змогу встановити стратегії маркетингу для кожного цільового сегмента.

3. Проведено аналіз вимог, принципів та підходів до сегментації, які стосуються маркетингових досліджень та автоматизації відповідних процесів, що дало змогу обґрунтувати необхідність впровадження алгоритмів і методів машинного навчання при сегментації покупців і товарів, а також ринків в цілому.

РОЗДІЛ 2
ПРОЕКТУВАННЯ АРХІТЕКТУРИ ТА РОЗРОБКА АЛГОРИТМУ СЕГМЕНТАЦІЇ КОРИСТУВАЧІВ

2.1. Аналіз методів і критеріїв сегментації

Найбільш поширеними і широко використовуваними методами сегментації, які застосовуються при побудові маркетингових систем є групування за ознаками і методи статистичного аналізу.

В основі методу групування лежить принцип послідовного розбиття множини об'єктів на підмножини (групи) з врахуванням найбільш важливих властивостей чи ознак. Одна з таких властивостей об’єкту виділяється як критерій на основі якого можна сформувати деяку систему показників (тип продукції, виробник товару, потенційний споживач конкретного виду товару).

Наступний крок полягає у побудові підмножин, у яких важливість даного критерію перевищує значення у порівнянні із значенням на множині можливих покупців визначеної продукції. Таким чином, внаслідок застосування принципу послідовного розбиття на два фрагменти на вибірці формуються підмножини або сегменти.

Для розв’язання задачі сегментації можуть бути застосованими методи багатокритеріальної класифікації. У цьому випадку декомпозиція виконується одночасно за сукупністю ознак, які підлягають аналізу. Представниками одного і того ж класу будуть кінцеві користувачі, які володіють схожими властивостями. Міра схожості покупців одного і того ж класу є значно більшою, ніж схожість екземплярів з інших класів.

На основі методу подібності об’єктів можна розв’язувати задачі типізації з можливістю одночасного врахування демографічних, соціальних, економічних і психографічних показників.

В якості прикладу можна навести розв’язок задачі щодо сегментації ринку шляхом формування типології користувачів, яка передбачає розбиття споживачів

на підмножини, які володіють однаковою або схожою «споживчою поведінкою». Формування типології представляє собою процес декомпозиції досліджуваної множини сутностей на однорідні та робастні у просторі і часі підмножини.

Споживчий ринок – це ринок кінцевих споживачів, які купують товари для особистого, домашнього або сімейного використання [3]. Критерії, які можуть використовуватися при формуванні сегментів ринку наведені у табл. 2.1.

Таблиця 2.1

Критерії сегментації ринку

Комплексний критерій	Елементарний критерій	Примітка
Географічний	Регіон	На рівні довготи і широти
	Адміністративний поділ	В залежності від країни, може бути: область, район, громада і т.п.
	Кількість населення	Градація може бути обрана кратною 5 тис.
	Щільність населення	Визначається типом населеного пункту
	Клімат	Визначається географічним розташуванням
Психографічний	Соціальний прошарок	
	Спосіб життя	Активний, пасивний і т.п.
	Особисті якості	Амбіційність, пасивність, імпульсивність

Продовження табл. 2.1

Комплексний критерій	Елементарний критерій	Примітка
Демографічний	Вік	Згідно класифікації прийнятою у державі, наприклад: діти, підлітки, молодь, середній вік, старший вік, літні люди
	Стать	Чоловіки та жінки
	Кількість членів сім'ї	
	Сімейний стан	
	Види професій	Наукові співробітники, технічні спеціалісти, підприємці, с/г працівники, працівники промисловості і т.п.
	Рівень освіти	Початкова, середня, вища
	Релігія	
	Раса	
	Національність	
Поведінковий	Рівень випадковості здійснення покупки	Товар куплено випадково, рівень купівлі товару носить не випадковий характер і т.п.
	Пошук вигідної пропозиції	Співвідношення ціна-якість товару

Продовження табл. 2.1

Комплексний критерій	Елементарний критерій	Примітка
Поведінковий	Статус постійного клієнта	Наявність картки постійного покупця, відсутність такої картки та ряд інших
	Рівень потрібності товару	Товар першої необхідності, купується лише раз у житті, товар популярний у певному сезоні
	Рівень лояльності до підприємства	Завжди купує товар даного підприємства, ніколи не купую продукцію його виробника
	Ступінь готовності придбати товар	Не готовність придбати продукт, погана поінформованість і т.п.
	Емоційне ставлення до товару	Позитивне ставлення до товару, негативне, нейтральне
	Соціально-економічний критерій	Соціальна приналежність до певної групи, професійна приналежність, соціальна група

Таким чином, визначивши основні фактори, які можуть впливати на сегментацію користувачів при проектуванні та експлуатації маркетингових комп’ютерних систем, необхідно побудувати архітектуру такої системи. Це дозволить визначити джерела збору та опрацювання необхідної для сегментації інформації.

2.2. Проектування архітектури маркетингової системи сегментації користувачів і товарів

Аналізуючи сучасні технології проведення маркетингових досліджень, доцільним є використання спроектованої архітектури збору та аналізу даних, яка показана на рис. 2.1.

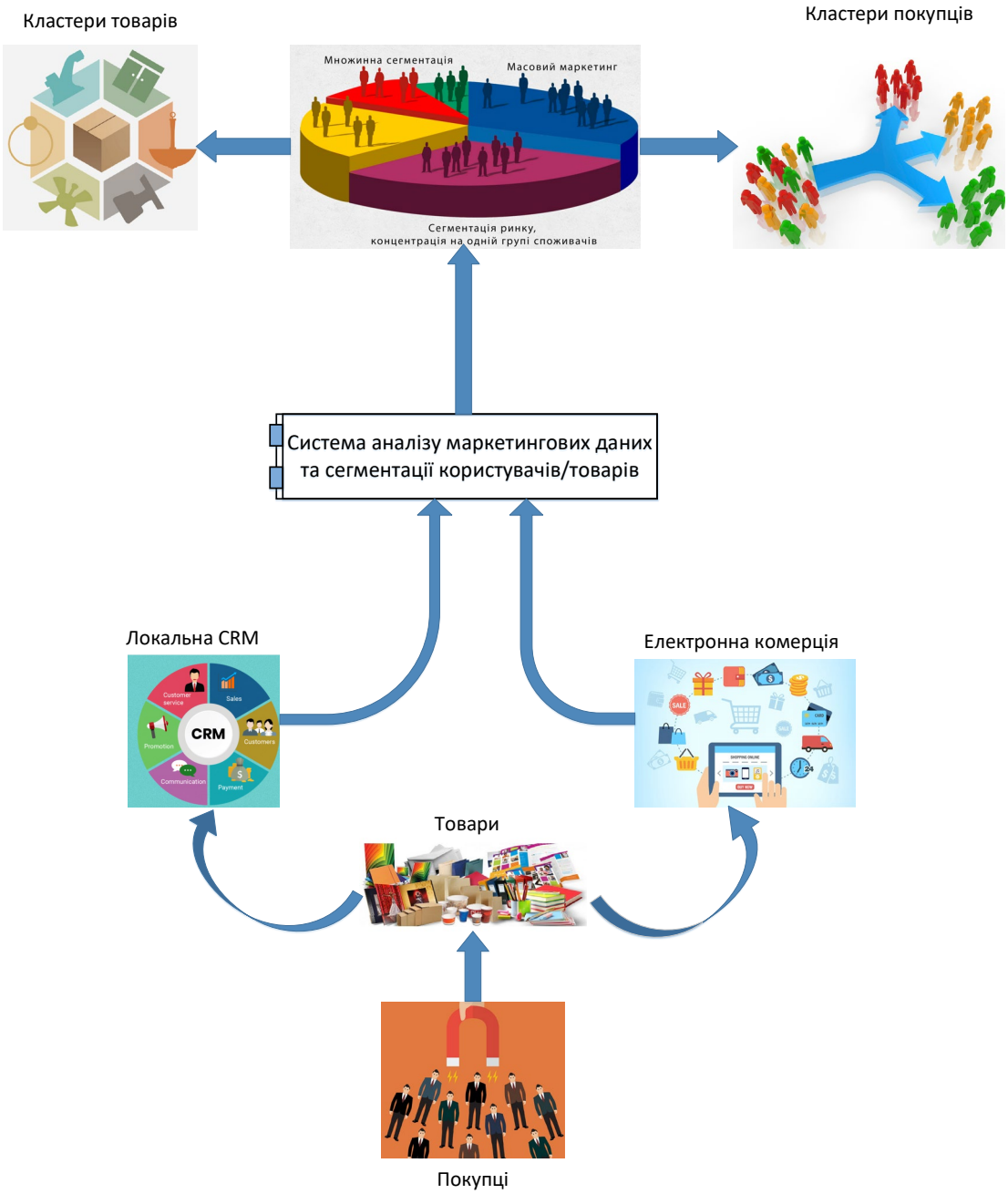


Рис. 2.1. Архітектура маркетингової системи сегментації користувачів

Як видно з рис. 2.1 до складу компонентів комп’ютерної маркетингової системи можуть входити:

– системи управління бізнес процесами на локальному рівні, що передбачають взаємодію з іншими комп’ютерними системами та містять інформацію про покупців і зведені показники ефективності діяльності підприємства;

– системи електронної комерції, які по суті реалізують продаж товарів або надання послуг кінцевим користувачам і володіють практично усіма властивостями необхідними для проведення макертингових досліджень;

– компонент, що відповідає за реалізацію інтелектуального аналізу даних та вирішує задачу сегментації користувачів або товарів.

Користувачі, купуючи різні товари або один товар можуть розраховуватись шляхом використання безконтактних носіїв, платіжних карток і т.п. Інформація про карту (ідентифікатор користувач) входить у транзакцію з продажу певних товарів і фіксується у відповідній базі даних. Доступ до бази даних транзакцій мають CRM системи, які дозволяють аналізувати певного типу дані на предмет зрізу певних показників продажів, наявності чи відсутність певного товару чи їхніх груп.

Аналогічним чином можна збирати інформацію про покупки користувачів при використанні засобів онлайн-торгівлі. Окрім цього, електронні магазини дозволяють збирати значно більше даних про користувачів, що дозволяє проводити більш якісну і точну їх класифікацію або кластеризацію.

Маркетингові дослідження на предмет формування кластерів користувачів чи товарів можна проводити лише після організації систем збору інформації, які в даному випадку, утворені на основі CRM-систем або електронних магазинів.

Модуль сегментації та кластеризації, який проектується у даній роботі, як частина маркетингової комп’ютерної системи, повинен реалізовувати алгоритми машинного навчання, наприклад k-means або EM. При цьому необхідним є виконання препроцесингу даних, виявлення важливих ознак користувачів і товарів, а також аналіз безпосередньо наявних у фреймі даних.

2.3. Методи алгоритми сегментації користувачів і товарів при проектуванні маркетингових комп'ютерних систем

Оскільки у вхідному дата фреймі відсутні мітки категорій користувачів і товарів, то доцільним є використання алгоритмів кластеризації для виявлення груп подібних покупців або товарів. Одним з ефективним алгоритмів кластеризації, з точки зору використання апаратних ресурсів і точності, є алгоритм k-means. Постановки задачі кластеризації, яка досліджується у кваліфікаційній роботі магістра показана на рис. 2.2.

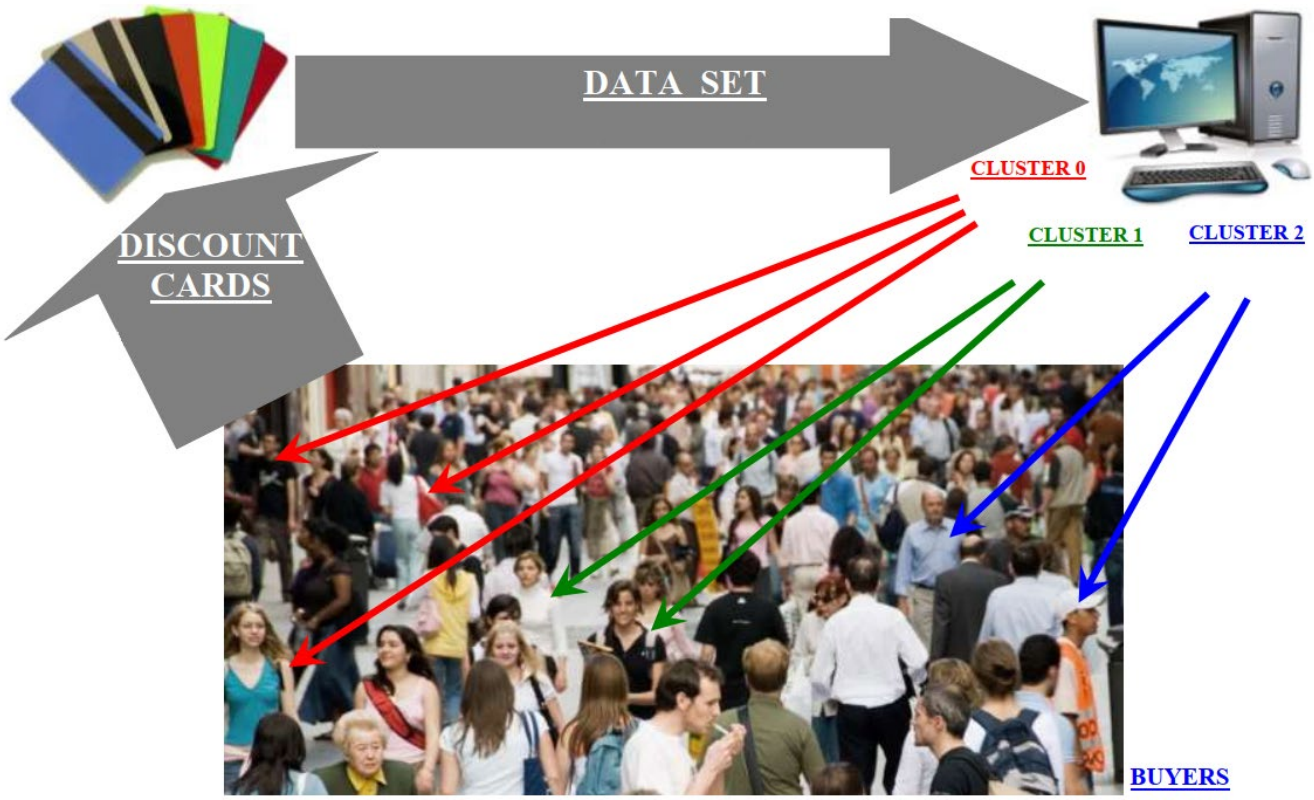


Рис. 2.2. Постановка задачі кластеризації

Даний алгоритм виконує поділ даних на найбільш придатні для аналізу групи з врахуванням існуючої інформації у дата фреймі. Дані розбивають на різні кластери, які зазвичай вибираються так, щоб вони були достатньо віддалені один від одного просторово. У такому випадку може бути використана Евклідова відстань, що забезпечує одержання ефективних результатів аналізу даних.

Кожен кластер має центр, який називають центроїдом, і точка даних об’єднується в певний кластер на основі того, наскільки близькі об’єкти до центроїда.

У загальному випадку, формально представити процес кластеризації можна, як показано на рис. 2.3 та визначити формулою 2.1.

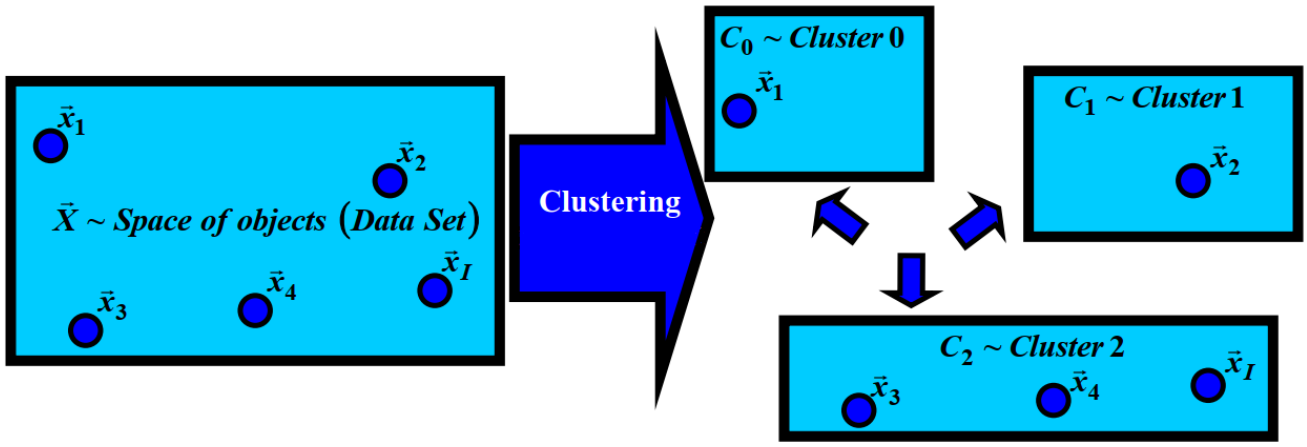


Рис. 2.3. Процес кластеризації об’єктів

$$\vec{X} = C_0 \cup C_1 \cup C_2 \tag{2.1}$$

де $\vec{X}$ – загальна сукупність об’єктів (простір об’єктів);

C_0, C_1, C_2 – кластери об’єктів;

Розрізняють два типи кластеризації:

– чітка кластеризація (hard clasterization);

– «м’яка» кластеризація (soft clasterization)

Формально чітку кластеризацію можна описати наступним чином:

$$I_{cluster_k}(\vec{x}) = I_{C_k}(\vec{x}) : C_k \to \{0,1\} \tag{2.2}$$

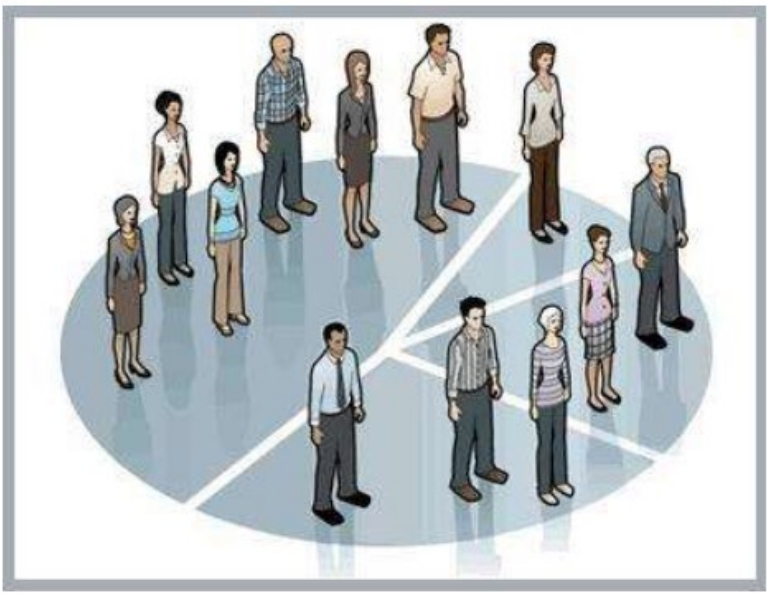
$I_{C_k}(\vec{x})$ – індикаторна (характеристична) функція кластеру C_k.

М’яку кластеризацію означають таким чином:

$$I_{fuzzycluster_k}(\vec{x}) = I_{fC_k}(\vec{x}): fC_k \rightarrow [0,1] \quad (2.3)$$

$I_{fC_k}(\vec{x})$ – індикаторна (характеристична) функція нечіткого кластеру fC_k.

Приклади чіткої та м’якої кластеризації показано на рис. 2.4 і рис. 2.5 відповідно.



soft) кластеризація

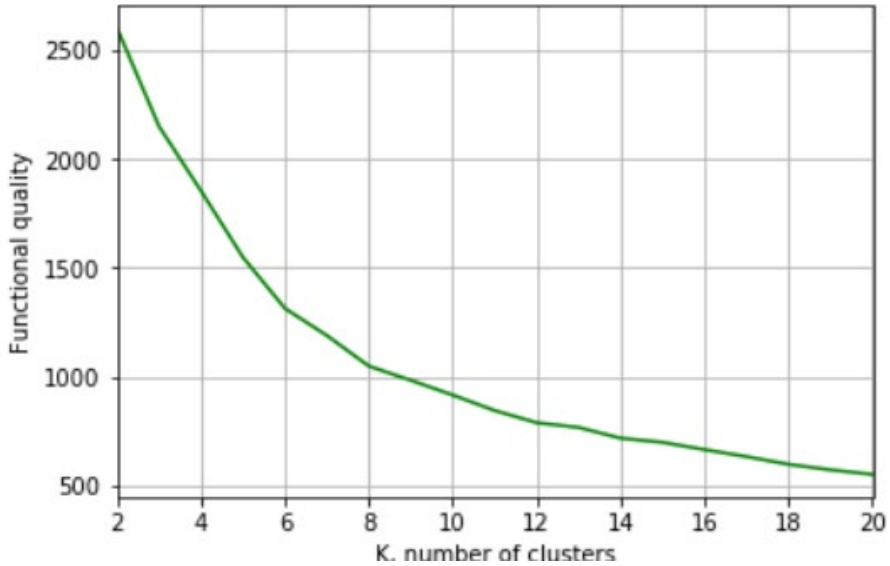
Для вибору кращої кількості кластерів на які ділиться дата сет, можна скористатися наступним виразом:

$$K_{best} = \arg\max_{K} |E(K+1) - E(K)|, \tag{2.4}$$

де E – функціонал якості;

K – кількість кластерів.

Як приклад, на рис. 2.6 показано залежність функціоналу якості від кількості кластерів. Найкращим варіантом згідно (2.4) буде, коли кількість кластерів рівна 3.



Алгоритм K-means ітераційно мінімізує відстані між кожною точкою даних та її центроїдом, щоб знайти найбільш оптимальне рішення для всіх точок даних. Алгоритм визначення кластерів з використанням k-means передбачає виконання наступних кроків:

– вибираються декілька випадкових точок набору даних як центроїди;

– розраховуються та зберігаються відстані між кожною точкою даних та центроїдними точками;

– на основі обчислення відстані, кожна точка отримує мітку до найближчого кластера;

Оновлення нових позицій центроїдів кластера виконується подібно до пошуку середнього значення в розташуваннях точок.

Якщо розташування центроїдів змінилося, процес повторюється з кроку 2, поки обчислений новий центр не залишиться незмінним, що означає, що члени кластерів і центроїди тепер встановлені.

Знаходження мінімальних відстаней між усіма точками означає, що точки даних були розділені, щоб утворити якомога компактніші кластери з найменшою дисперсією всередині них. Іншими словами, жодна інша ітерація не може мати нижчу середню відстань між центроїдами та точками даних, знайденими в них.

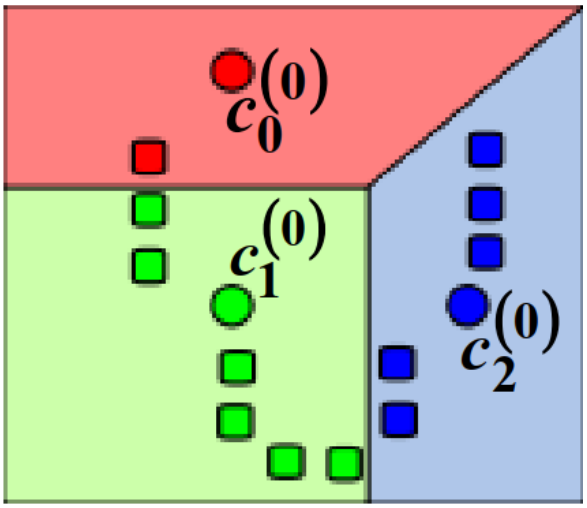
Формально для вказання того, що об’єкт відноситься до найближчого центру кластеру c_K можна записати наступним виразом:

$$k = \arg\min_k dist(\vec{x}, c_k), \overline{k = 0, K-1}, \tag{2.5}$$

де K – кількість кластерів;

c_k – центроїди (центри кластерів).

Графічно представлення формули (2.5) показано на рис. 2.7.



Обчислення відстаней від центру кластера до інших точок (об’єктів) може забезпечуватись рядом метрик:

– Евклідова відстань (2.6);
– Квадрат Евклідової відстані (2.7);
– Манхетнеська відстань (2.8);
– відстань Чебишева (2.9) та ін.

$$dist_{Euclidiean}(\vec{x}_{i_1}, \vec{x}_{i_2}) = \sqrt{\sum_{j=0}^{J-1}(f_j(\vec{x}_{i_1}) - f_j(\vec{x}_{i_2}))^2} \quad (2.6)$$

$$dist_{SQEuclidiean}(\vec{x}_{i_1}, \vec{x}_{i_2}) = \sum_{j=0}^{J-1}(f_j(\vec{x}_{i_1}) - f_j(\vec{x}_{i_2}))^2 \quad (2.7)$$

$$dist_{Manhatten}(\vec{x}_{i_1}, \vec{x}_{i_2}) = \sum_{j=0}^{J-1}|f_j(\vec{x}_{i_1}) - f_j(\vec{x}_{i_2})| \quad (2.8)$$

$$dist_{Chebyshev}(\vec{x}_{i_1}, \vec{x}_{i_2}) = \max_{j=0,1,2,..,J-1}|f_j(\vec{x}_{i_1}) - f_j(\vec{x}_{i_2})| \quad (2.9)$$

Для розрахунку нових центрів кластерів використовується ітераційний підхід, в основі якого наступна формула:

$$c_k^{(t+1)} = \frac{1}{\left|C_k^{(t)}\right|}\sum_{\vec{x}_i \in C_k^{(t)}} \vec{x}_i, \quad (2.10)$$

де C_k – кластер k;

$c_k^{(t+1)}$ – новий центр кластеру C_k ;

$t = 0, T-1$ – крок ітерації.

До основних базових метрик якості кластеризації у випадку метричного простору належать:

– середня внутрішньо кластерна відстань;
– середня міжкластерна відстань;

– відношення внутрішньо кластерної відстані до міжкластерної відстані.

Середня внутрішньо кластерна відстань обчислюється за формулою:

$$E_0 = \frac{\sum_{i_1<i_2}[y_{i_1}=y_{i_2}]\cdot dist(\vec{x}_{i_1},\vec{x}_{i_2})}{\sum_{i_1<i_2}[y_{i_1}=y_{i_2}]} \to min \quad (2.11)$$

Значення середньої міжкластерної відстані можна знайти за допомогою наступного виразу:

$$E_1 = \frac{\sum_{i_1<i_2}[y_{i_1}\neq y_{i_2}]\cdot dist(\vec{x}_{i_1},\vec{x}_{i_2})}{\sum_{i_1<i_2}[y_{i_1}\neq y_{i_2}]} \to max \quad (2.12)$$

Відношення внутрішньо кластерної відстані до міжкластерної відстані формально матиме вигляд:

$$E_2 = \frac{E_0}{E_1} \to min \quad (2.13)$$

У випадку, коли простір даних представлений у лінійному векторному просторі, то можна використовувати наступні метрики якості:

– сума середніх внутрішньо кластерних відстаней;
– сума міжкластерних відстаней;
– відношення двох попередніх показників.

Сума середніх внутрішньо кластерних відстаней обчислюється за формулою:

$$E_3 = \sum_{k\in K} \frac{1}{|C_k|} \sum_{\vec{x}_i\in C_k} dist^2(\vec{x}_i, c_k) \to min \quad (2.14)$$

Суму міжкластерних відстаней знаходять за формулою:

$$E_4 = \sum_{k\in K} dist^2(c_k, M_{\vec{X}^l}) \to max \quad (2.15)$$

Таким чином, формально обгрунтовано застосування алгоритму k-means та відповідних метрик оцінювання результатів його роботи для проведення сегментації користувачів при побудові та експлуатації маркетиногових комп’ютерних систем.

2.4. Аналіз вхідних даних для сегментації користувачів і товарів

Вхідними даними для сегментації користувачів обрано вибірку даних, що опублікована на платформі kaggle [20]. Структура датасету містить дані, які розподілені між вісьмома колонками:

– «InvoiceNo» – номінально представляє собою 6-значний цілісний номер, що ідентифікує транзакції (якщо номер починається з «c» – скасування транзакції);

– «StockCode» – номер замовлення, що виступає в якості ідентифікатора придбаного товару;

– «Description» – опис придбаного товару;

– «Quantity» – кількість придбаного товару;

– «InvoiceDate» – дата замовлення;

– «UnitPrice» – ціна за одиницю товару;

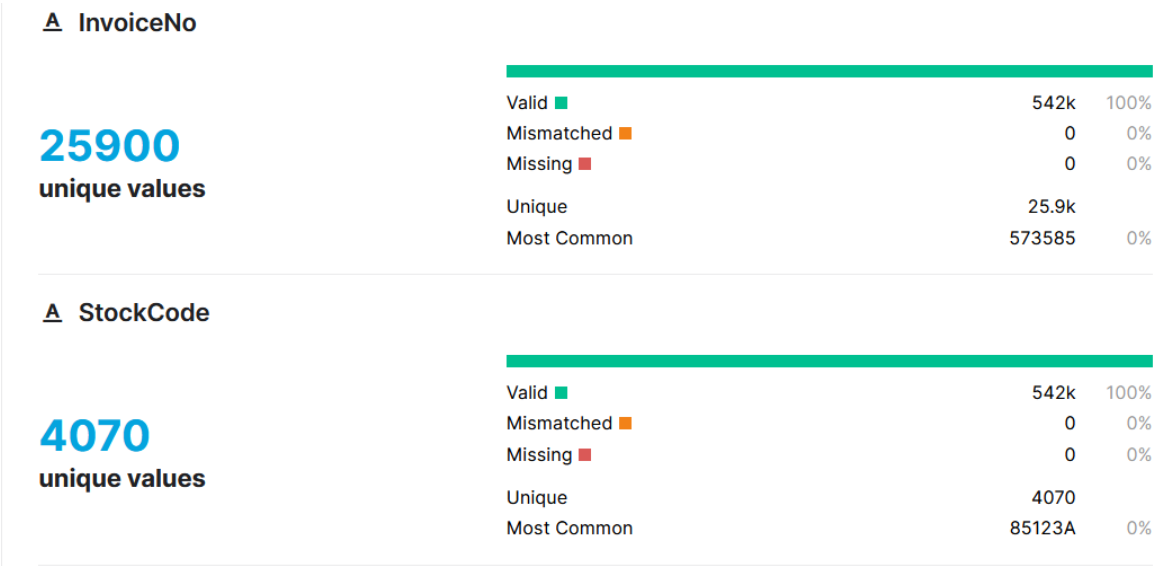
– «CustomerID» – ідентифікатор користувача (покупця);

– «Country» – країна походження користувача.

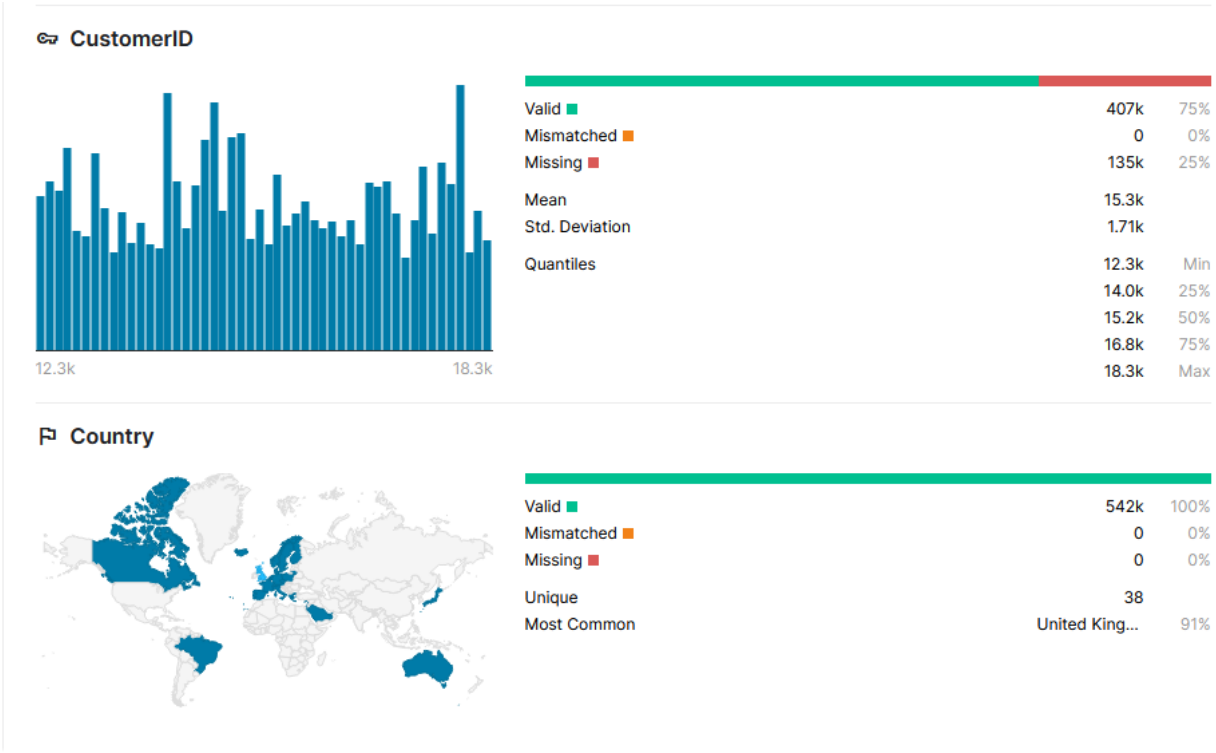
Фрагмент вхідного набору даних показано на рис. 2.8.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25

Статистичні показники розподілу даних за колонками «InvoiceNo» та «StockCode» показано на рис. 2.9.



Розподіл даних за користувачами та країнами показано на рис. 2.10.



Для того, щоб проводити сегментацію покупців першим крок необхідно імпортувати потрібні бібліотеки. У лістингу 2.1 наведено імпорт бібліотек, необхідних для проведення досліджень щодо формування груп подібних користувачів.

Лістинг 2.1. Імпорт бібліотек

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
import datetime, nltk, warnings
import matplotlib.cm as cm
import itertools
from pathlib import Path
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples,
silhouette_score
from sklearn import preprocessing, model_selection, metrics,
feature_selection

from sklearn.model_selection import GridSearchCV,
learning_curve
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix
from sklearn import neighbors, linear_model, svm, tree,
ensemble
from wordcloud import WordCloud, STOPWORDS
from sklearn.ensemble import AdaBoostClassifier
from sklearn.decomposition import PCA
from IPython.display import display, HTML
import plotly.graph_objs as go
from plotly.offline import init_notebook_mode,iplot
init_notebook_mode(connected=True)
warnings.filterwarnings("ignore")
plt.rcParams["patch.force_edgecolor"] = True
plt.style.use('fivethirtyeight')
mpl.rc('patch', edgecolor = 'dimgray', linewidth=1)
%matplotlib inline
```

Як видно з лістингу 2.1. основними використовуваними бібліотеками є: pandas, numpy,matplotlib, seaborn – використовуються для опрацювання даних та

візуалізації результатів. Бібліотеки sklearn, nltk в перспективі будуть застосовані до розв'язання сегментації користувачів і товарів.

Зчитування даних для подальшого опрацювання виконано за допомогою програмного коду, що представлений у лістингу 2.2.

Лістинг 2.2. Зчитування даних з вибірки

```
#_____________________
# read the datafile
df_initial = pd.read_csv('../input/data.csv',encoding="ISO-
8859-1",
                         dtype={'CustomerID': str,'InvoiceID':
str})
print('Dataframe dimensions:', df_initial.shape)
#_______
df_initial['InvoiceDate'] =
pd.to_datetime(df_initial['InvoiceDate'])
#_____________________________________________________________
# gives some infos on columns types and numer of null values
tab_info=pd.DataFrame(df_initial.dtypes).T.rename(index={0:'co
lumn type'})

tab_info=tab_info.append(pd.DataFrame(df_initial.isnull().sum(
)).T.rename(index={0:'null values (nb)'}))
tab_info=tab_info.append(pd.DataFrame(df_initial.isnull().sum(
)/df_initial.shape[0]*100).T.
                         rename(index={0:'null values (%)'}))
display(tab_info)
#_____________________
# show first lines
display(df_initial[:5])
```

У результаті виконання лістингу 2.1 одержано результат, як показано на рис. 2.11.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
column type	object	object	object	int64	datetime64[ns]	float64	object	object
null values (nb)	0	0	1454	0	0	0	135080	0
null values (%)	0	0	0.268311	0	0	0	24.9267	0

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom

Рис. 2.11. Результат зчитування даних

Аналізуючи кількість нульових значень у фреймі даних, цікаво відзначити, що ~25% записів не пов’язані з жодним клієнтом. З наявними даними неможливо встановити відповідність значення для користувача, тому ці записи не є інформативними для задачі сегментації і потребують видалення.

При аналізі та формуванні груп користувачів варто знати частоту замовлення товарів з різних країн (лістинг 2.3).

Лістинг 2.3. Визначення частоти замовлення з різних країн

```
temp = df_initial[['CustomerID', 'InvoiceNo',
'Country']].groupby(['CustomerID', 'InvoiceNo',
'Country']).count()
temp = temp.reset_index(drop = False)
countries = temp['Country'].value_counts()
print('Number of countries: {}'.format(len(countries)))
```

Результат виявлення кількості країн у вигляді карти показано на рис. 2.12.

Рис. 2.12. Візуалізація замовлень товарів за країнами

2.5. Визначення та аналіз інформації про користувачів

Важливим з точки зору сегментації користувачів є виявлення груп товарів і відповідно самих покупців. Для початку потрібно визначити загальну кількість споживачів і товарів, які були придбані, а також транзакції які при цьому були успішно проведеними. Для розв'язання цієї задачі необхідно з датасету вибрати інформацію про транзакції, які не були відхиленими. Фрейм даних містить ~400000 записів. Визначимо кількість користувачів і продуктів у цих записах за допомогою програмного коду, який наведений у лістингу 2.4.

Лістинг 2.4. Визначення кількості товарів і покупців

```
pd.DataFrame([{'products':
len(df_initial['StockCode'].value_counts()),
                'transactions':
len(df_initial['InvoiceNo'].value_counts()),
                'customers':
len(df_initial['CustomerID'].value_counts()),
              }], columns = ['products', 'transactions',
'customers'], index = ['quantity'])
```

Результат щодо загальної кількості транзакцій, кількості покупців і товарів показано на рис. 2.13.

	products	transactions	customers
quantity	3684	22190	4372

Рис. 2.13. Інформація щодо кількості товарів, транзакцій та споживачів

З рис. 2.6 видно, що у фреймі даних наявні 4372 користувачів, які іпридюали 3684 різних товарів. Загальна кількість здійснених транзакцій становить близько ~ 22 000. Наступний крок полягає у визначенні кількості товарів, які наявні у кожній транзакції (лістинг 2.5).

Лістинг 2.5. Кількість товарів у транзакції

```
temp = df_initial.groupby(by=['CustomerID', 'InvoiceNo'],
as_index=False)['InvoiceDate'].count()
nb_products_per_basket = temp.rename(columns =
{'InvoiceDate':'Number of products'})
nb_products_per_basket[:10].sort_values('CustomerID')
```

На рис. 2.14 показано результати виявленої кількості товарів, які входять в кожну окрему транзакцію.

	CustomerID	InvoiceNo	Number of products
0	12346	541431	1
1	12346	C541433	1
2	12347	537626	31
3	12347	542237	29
4	12347	549222	24
5	12347	556201	18
6	12347	562032	22
7	12347	573511	47
8	12347	581180	11
9	12348	539318	17

Рис. 2.14. Кількість входження товарів у транзакцію

Перші рядки таблиці, показаної на рис. 2.7, відображають важливі речі, які необхідно врахувати при проведенні сегментації покупців:

– наявність записів із префіксом «C» для змінної «InvoiceNo», що вказує на транзакції, які були скасовані;

– наявність користувачів, які фігурують у вибірці даних лише один раз і придбали тільки один товар (наприклад, користувач з CustomerID=12346);

– наявність користувачів, які часто купують велику кількість товарів при кожному замовленні.

Далі потрібно виявити та проаналізувати інформацію щодо транзакцій, які були скасованими. У лістингу 2.6 реалізовано визначення кількості таких транзакцій.

Лістинг 2.6. Виявлення скасованих транзакцій

```
nb_products_per_basket['order_canceled'] =
nb_products_per_basket['InvoiceNo'].apply(lambda x:int('C' in
x))
display(nb_products_per_basket[:5])
#_____________________________________________________________
n1 = nb_products_per_basket['order_canceled'].sum()
n2 = nb_products_per_basket.shape[0]
print('Number of orders canceled: {}/{} ({:.2f}%) '.format(n1,
n2, n1/n2*100))
```

Результат виконання лістингу 2.6 представлено на рис. 2.15.

	CustomerID	InvoiceNo	Number of products	order_canceled
0	12346	541431	1	0
1	12346	C541433	1	1
2	12347	537626	31	0
3	12347	542237	29	0
4	12347	549222	24	0

```
Number of orders canceled: 3654/22190 (16.47%)
```

Рис. 2.15. Кількість скасованих транзакцій

Виходячи з одержаних результатів (рис. 2.15) зауважимо, що кількість скасованих транзакцій досить велика (~ 16% від загальної кількості транзакцій), тому нехтувати таким великим об'ємом даних не варто. Потрібно провести додатковий аналіз щодо скасування транзакцій. Для цього розглянемо дані, які характерні для такого типу операцій (рис. 2.16).

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
61619	541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	2011-01-18 10:01:00	1.04	12346	United Kingdom
61624	C541433	23166	MEDIUM CERAMIC TOP STORAGE JAR	-74215	2011-01-18 10:17:00	1.04	12346	United Kingdom
286623	562032	22375	AIRLINE BAG VINTAGE JET SET BROWN	4	2011-08-02 08:48:00	4.25	12347	Iceland
72260	542237	84991	60 TEATIME FAIRY CAKE CASES	24	2011-01-26 14:30:00	0.55	12347	Iceland
14943	537626	22772	PINK DRAWER KNOB ACRYLIC EDWARDIAN	12	2010-12-07 14:57:00	1.25	12347	Iceland

Рис. 2.16. Особливості даних скасованих транзакцій

З даних рис. 2.16 видно, що у випадку скасування замовлення наявними є інші, практично ідентичні транзакції, які відрізняються лише значенням змінних Quantity та InvoiceDate. Тому доцільним є проведення перевірки істинності такого твердження для усіх наявних у фреймі даних. Для цього потрібно знайти записи, які вказують на від'ємну кількість товарів у замовленні, і перевірити, чи є подібне замовлення, що вказує ту саму кількість (але додатну) з таким же описом даних «CustomerID», «Description» та «UnitPrice» (лістинг 2.7, рис. 2.17).

Лістинг 2.7. Перевірка подібних транзакцій

```
df_check = df_initial[df_initial['Quantity'] <
0][['CustomerID','Quantity',
'StockCode','Description','UnitPrice']]
for index, col in  df_check.iterrows():
    if df_initial[(df_initial['CustomerID'] == col[0]) &
(df_initial['Quantity'] == -col[1])
```

```
& (df_initial['Description'] == col[2])].shape[0] == 0:
        print(df_check.loc[index])
        print(15*'-'+'>'+' HYPOTHESIS NOT FULFILLED')
        break
```

Як видно з результатів (рис. 2.17), що початкова гіпотеза не виконується через наявність запису «Знижка». Виконаємо повторну перевірки гіпотезу, відкинувши записи "Знижка" (рис. 2.18).

```
CustomerID        14527
Quantity             -1
StockCode             D
Description    Discount
UnitPrice          27.5
Name: 141, dtype: object
---------------> HYPOTHESIS NOT FULFILLED
```

Рис. 2.17. Результат виконання лістингу 2.7

```
154 CustomerID                                15311
Quantity                                     -1
StockCode                                 35004C
Description    SET OF 3 COLOURED  FLYING DUCKS
UnitPrice                                  4.65
Name: 154, dtype: object
---------------> HYPOTHESIS NOT FULFILLED
```

Рис. 2.18. Результат вибірки даних скасованих транзакцій без знижок

Як і у випадку (рис.2.17), одержаний результат (рис. 2.18) не підтверджує початкової гіпотези. Отже, скасування не обов’язково відповідає замовленням, які були зроблені заздалегідь. Далі варто створити нову змінну у фреймі даних, яка буде вказувати на те, чи була частина транзакції скасованою. Що стосується скасувань без аналогів, то деякі з них, ймовірно, пов’язані з тим, що замовлення на купівлю були виконані до дати формування фрейму даних (точка входу в базу даних). У реалізованій функції виконано перевірку двох можливих випадків:

– скасоване замовлення не містить відповідника у фреймі даних;

– існує принаймні один аналог запису із точною кількістю параметрів;.

Індекс відповідного замовлення про скасування зберігається у змінних-списках, розміри яких показано на рис. 2.19.

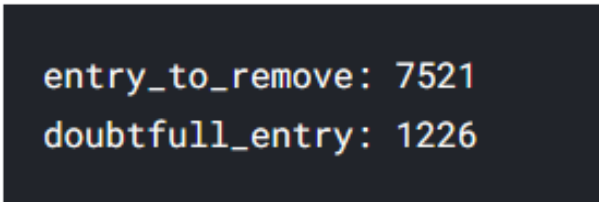


Рис. 2.19. Кількість записів, що відповідають встановленим вимогам

Серед цих записів (рис. 2.19), рядки, наведені у списку сумнівних записів, відповідають записам, які вказують на скасування, але для яких попередньо не вказано команду.

На практиці потрібно видалити всі дані, що відповідають встановленим вимогам. Вони становлять приблизно 1,4% і 0,2% відповідно від усіх даних, які наявні у вибірці даних.

Вище було помічено, що деякі значення стовпця «StockCode» вказують на певну транзакцію (тобто «D «для «Discount»). Встановимо область визначення змінної «StockCode» та визначимо можливі літери, які входять в ідентифікатор товару (лістинг 2.8).

Лістинг 2.8. Виявлення літер у значеннях колонки «StockCode»

```
list_special_codes =
df_cleaned[df_cleaned['StockCode'].str.contains('^[a-zA-Z]+',
regex=True)]['StockCode'].unique()
list_special_codes
for code in list_special_codes:
    print("{:<15} -> {:<30}".format(code,
df_cleaned[df_cleaned['StockCode'] ==
code]['Description'].unique()[0]))
```

У результаті (рис. 2.20) видно, що існує кілька типів своєрідних транзакцій, пов’язаних наприклад на портові збори або банківські комісії.

```
POST            -> POSTAGE
D               -> Discount
C2              -> CARRIAGE
M               -> Manual
BANK CHARGES    -> Bank Charges
PADS            -> PADS TO MATCH ALL CUSHIONS
DOT             -> DOTCOM POSTAGE
```

Рис. 2.20. Типи транзакцій

Наступний крок полягає в аналізі вартості кожної покупки і для цього створено змінну, яка буде представляти споживчий кошик покупця (лістинг 2.9).

Лістинг 2.9. Формування споживчого кошика покупця

```
df_cleaned['TotalPrice'] = df_cleaned['UnitPrice'] *
(df_cleaned['Quantity'] - df_cleaned['QuantityCanceled'])
df_cleaned.sort_values('CustomerID')[:5]
```

Виконавши лістинг 2.9, одержуємо результат, фрагмент якого показано на рис. 2.21.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	QuantityCanceled	TotalPrice
541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	2011-01-18 10:01:00	1.04	12346	United Kingdom	74215	0.0
549222	22375	AIRLINE BAG VINTAGE JET SET BROWN	4	2011-04-07 10:43:00	4.25	12347	Iceland	0	17.0
573511	22698	PINK REGENCY TEACUP AND SAUCER	12	2011-10-31 12:25:00	2.95	12347	Iceland	0	35.4

Рис. 2.21. Результат формування споживчого кошика покупця

Кожен запис у фреймі даних (рис. 2.21) вказує на вартість за один вид товару. Отже, замовлення розбиваються на кілька рядків. Споживчий кошик покупця формується шляхом сумарного підбиття усіх замовлень та відповідної вартості. Для того, щоб уявляти інформацію про тип замовлень, що існують в цьому наборі даних, побудуємо кругову діаграму, що відображає розподіл за вартістю товарів.

На рис. 2.22 показано розподіл замовлень за вартістю товарів, що дозволяє говорити про утворення певних кластерів за характеристикою ціни замовлень.

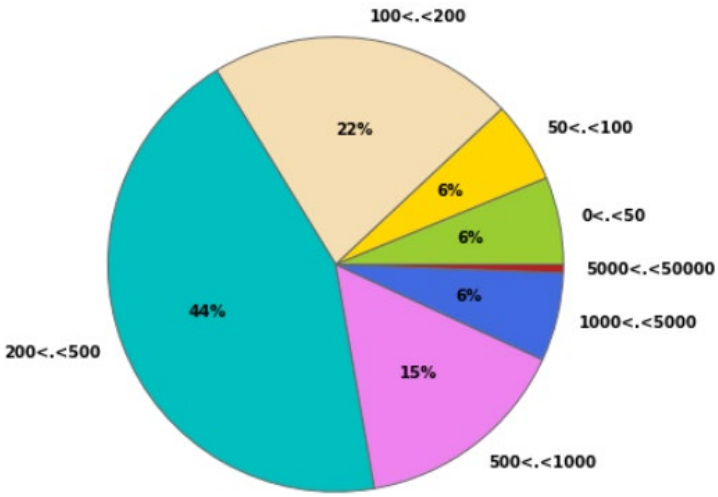


Рис. 2.22. Розподіл замовлень за вартістю

З діаграми рис. 2.22 видно, що переважна більшість замовлень стосується відносно великих покупок, враховуючи це ~ 65% покупок дають відносяться до кластеру на суму понад 200 фунтів стерлінгів.

2.6. Висновки до розділу

У даному розділі одержано наступні наукові та практичні результати:

1. Проаналізовано методи і критерії сегментації користувачів і товарів, що дало змогу встановити комплексні та атомарні показники, які необхідно враховувати при автоматизованому визначенні груп товарів і покупців, що не містять цільових міток сегментів.

2. Запропоновано, спроектовано та проаналізовано базову архітектуру маркетингової комп’ютерної системи сегментації користувачів і товарів, що використовує дані з CRM-систем і систем обліку показників бізнес-діяльності підприємства, а також передбачає необхідність реалізації підсистеми автоматизованого визначення груп товарів і покупців на основі методів кластеризації, що дає змогу знизити та оптимізувати ресурси на аналіз маркетингової інформації і дозволяє приймати оптимальні рішення при зміні кон’юнктури ринку.

3. Запропоновано та обгрунтовано застосування методу сегментації користувачів, що використовує алгоритм k-means і дає змогу автоматизувати процес групування об’єктів ринку без відповідних міток, а також використовувати метрики якості процесу сегментації.

4. Проведено аналіз експериментального фрейму даних до складу якого входить інформація про товари, користувачів, країни, транзакції та опис товару засобами мови програмування Python, що дало можливість виявити існуючі властивості груп товарів, а також сформувати нові емерджентні властивості, які стосуються споживчих кошиків, опису товарів та ін.

РОЗДІЛ 3
ПРОГРАМНА РЕАЛІЗАЦІЯ АЛГОРИТМУ СЕГМЕНТАЦІЇ КОРИСТУВАЧІВ І ТОВАРІВ ПРИ ПРОЕКТУВАННІ МАРКЕТИНГОВИХ КОМП'ЮТЕРНИХ СИСТЕМ

3.1. Формування сегментів товарів

У фреймі даних товари однозначно ідентифікуються за допомогою змінної «StockCode». Їхній короткий опис забезпечує змінна «Description». Для формування категорій товарів, пропонується провести аналіз значення вмістимого комірок у стовпці опису, що дозволить згрупувати продукти за відповідними класами. Для цього використано функцію добування даних зі стовпця «Description», яка наведена у лістингу 3.1.

Лістинг 3.1. Функція добування даних зі стовпця «Description»

```
is_noun = lambda pos: pos[:2] == 'NN'

def keywords_inventory(dataframe, colonne = 'Description'):
    stemmer = nltk.stem.SnowballStemmer("english")
    keywords_roots  = dict()  # collect the words / root
    keywords_select = dict()  # association: root <-> keyword
    category_keys   = []
    count_keywords  = dict()
    icount = 0
    for s in dataframe[colonne]:
        if pd.isnull(s): continue
        lines = s.lower()
        tokenized = nltk.word_tokenize(lines)
        nouns = [word for (word, pos) in
nltk.pos_tag(tokenized) if is_noun(pos)]

        for t in nouns:
            t = t.lower() ; racine = stemmer.stem(t)
            if racine in keywords_roots:
                keywords_roots[racine].add(t)
                count_keywords[racine] += 1
            else:
                keywords_roots[racine] = {t}
                count_keywords[racine] = 1
```

```
    for s in keywords_roots.keys():
        if len(keywords_roots[s]) > 1:
            min_length = 1000
            for k in keywords_roots[s]:
                if len(k) < min_length:
                    clef = k ; min_length = len(k)
            category_keys.append(clef)
            keywords_select[s] = clef
        else:
            category_keys.append(list(keywords_roots[s])[0])
            keywords_select[s] = list(keywords_roots[s])[0]

    print("Nb of keywords in variable '{}':
{}".format(colonne,len(category_keys)))
    return category_keys, keywords_roots, keywords_select,
count_keywords
```

Функція, представлена у лістингу 3.1 приймає в якості вхідних параметрів фрейм даних і аналізує вміст стовпця «Description», виконуючи такі операції:

– визначити назви (власні, поширені), що містяться в описі товарів;

– для кожного імені визначити корінь слова та сформувати колекцію імен, пов’язаних із конкретним коренем;

– підрахувати, скільки разів кожен корінь з’являється у фреймі даних

– коли для одного кореня вказано кілька слів, вважається, що ключове слово, пов’язане з цим коренем, є найкоротшим ім’ям (це дозволяє вибирати однину, коли є варіанти однини/множини)

Далі можна одержати список товарів, які належать до одного класу шляхом реалізації відповідного програмного коду (лістинг 3.2).

Лістинг 3.2. Формування категорій товарів

```
df_produits =
pd.DataFrame(df_initial['Description'].unique()).rename(column
s = {0:'Description'})
keywords, keywords_roots, keywords_select, count_keywords =
keywords_inventory(df_produits)
list_products = []
for k,v in count_keywords.items():
    list_products.append([keywords_select[k],v])
list_products.sort(key = lambda x:x[1], reverse = True)
```

У лістингу 3.2, після створення списку товарів, використано функцію з лістингу 3.1 для аналізу опису товарів. Виконання цієї функції повертає три змінні:

– «keywords» – список визначених ключових слів;

– keywords_roots – словник, де ключі – це корені ключових слів, а значення – це списки слів, пов’язаних з цими коренями;

– count_keywords – словник із зазначенням кількості випадків використання кожного слова.

Далі виконується перетворення словника «count_keywords» у список, щоб мати можливість відсортувати ключові слова відповідно до їх появи. Внаслідок виконання таких операцій та лістингу 3.3 одержуємо результат, який показано на рис. 3.1.

Лістинг 3.3. Формування

```
liste = sorted(list_products,
key = lambda x:x[1], reverse = True)
#_____________________________________
plt.rc('font', weight='normal')
fig, ax = plt.subplots(figsize=(7, 25))
y_axis = [i[1] for i in liste[:125]]
x_axis = [k for k,i in enumerate(liste[:125])]
x_label = [i[0] for i in liste[:125]]
plt.xticks(fontsize = 15)
plt.yticks(fontsize = 13)
plt.yticks(x_axis, x_label)
plt.xlabel("Nb. of occurences", fontsize = 18, labelpad = 10)
ax.barh(x_axis, y_axis, align = 'center')
ax = plt.gca()
ax.invert_yaxis()
#_____________________________________________________________
_____
plt.title("Words occurence",bbox={'facecolor':'k', 'pad':5},
color='w',fontsize = 25)
plt.show()
```

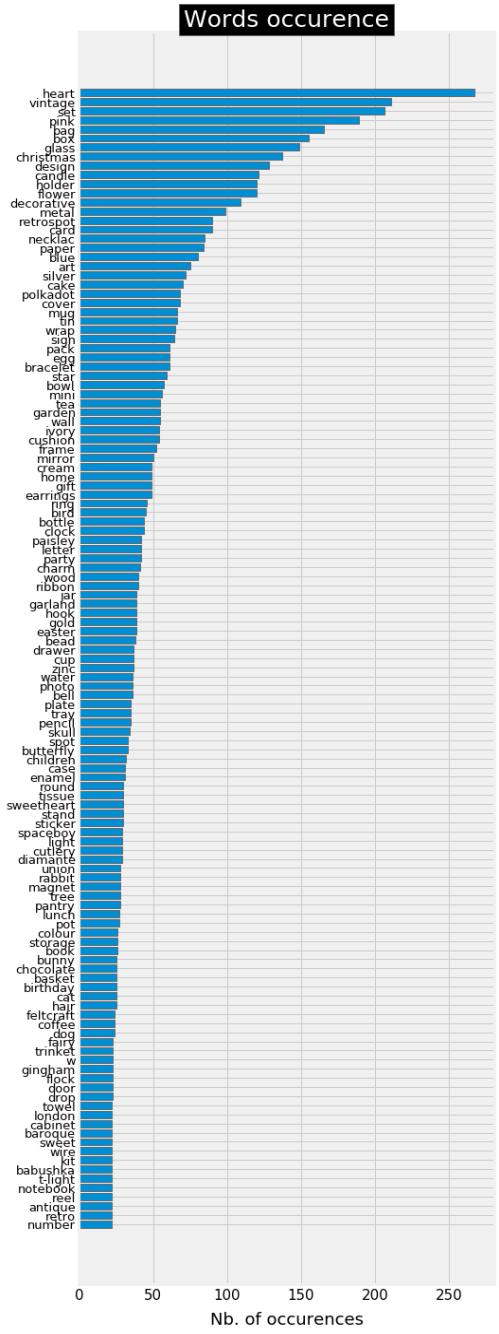


Рис. 3.1. Частота входження слів в опис товарів

Отриманий список містить понад 1400 ключових слів, а найпоширеніші з них зустрічаються у понад 200 товарах. Проте, вивчаючи зміст списку, слід зауважити, що деякі назви використовувати не доцільно. Інші не несуть інформації, як от для прикладу кольори. Тому варто відкинути ці слова з подальшого аналізу та задати обмеження щодо частоти повтору слів не менше, наприклад, як 13 разів. Для цього реалізовано лістинг 3.4.

Лістинг 3.4. Фільтрація списку товарів

```
list_products = []
for k,v in count_keywords.items():
    word = keywords_select[k]
    if word in ['pink', 'blue', 'tag', 'green', 'orange']:
continue
    if len(word) < 3 or v < 13: continue
    if ('+' in word) or ('/' in word): continue
    list_products.append([word, v])
#_____________________________________________________
list_products.sort(key = lambda x:x[1], reverse = True)
print('mots conservés:', len(list_products))
```

Тепер відфільтровані ключові слова можна використати для створення груп товарів. По-перше, потрібно визначити деяку матрицю X як показано у табл. 3.1.

Таблиця 3.1

Відношення між товарами і категоріями

	Слово 1	….	Слово j	…	Слово N
Товар 1	a_{11}	…	…	…	a_{1N}
…	…	…	…	…	…
Товар i	a_{i1}	…	a_{ij}	…	…
….	…	…	…	…	…
Товар M	a_{M1}	…	…	…	a_{MN}

У даному випадку, коефіцієнти a_{ij} приймають значення 1, якщо слово міститься в описі товару і 0 – за іншої умови.

Матриця X містить слова, які входять до опису товарів за принципом одноразового кодування. На практиці встановлено, що введення діапазону цін призводить до більш збалансованих груп з точки зору кількості елементів. Тому доцільно додати до цієї матриці 6 додаткових стовпців, які відображають ціни на товари (лістинг 3.5).

Лістинг 3.5. Формування діапазонів цін на товари

```
threshold = [0, 1, 2, 3, 5, 10]
label_col = []
for i in range(len(threshold)):
    if i == len(threshold)-1:
        col = '.>{}'.format(threshold[i])
    else:
        col = '{}<.<{}'.format(threshold[i],threshold[i+1])
    label_col.append(col)
    X.loc[:, col] = 0

for i, prod in enumerate(liste_produits):
    prix = df_cleaned[ df_cleaned['Description'] ==
prod]['UnitPrice'].mean()
    j = 0
    while prix > threshold[j]:
        j+=1
        if j == len(threshold): break
    X.loc[i, label_col[j-1]] = 1
```

Для того, щоб вибрати відповідні асортименти товару, виконано перевірку їх кількості продуктів у різних групах та одержано результат, який приведено на рис. 3.2.

gamme	nb. produits
0<.<1	964
1<.<2	1009
2<.<3	673
3<.<5	606
5<.<10	470
.>10	156

Рис. 3.2. Кількість товарів у визначених діапазонах цін

У випадку матриці із бінарним кодуванням найбільш ефективною для обчислення відстаней є метрика Хеммінга. Слід відмітити, що метод k-means бібліотеки sklearn використовує евклідову відстань, але це не найкращий вибір у випадку категорійних змінних.

Однак, щоб використовувати метрику Хеммінга, потрібно використовувати пакет kmodes, який недоступний на поточній платформі. Тому скористаємось методом k-means. Щоб визначити хоч приблизну кількість кластерів, застосовується метод силуетної оцінки, реалізацію якого мовою Python наведено у лістингу 3.6.

Лістинг 3.6. Визначення кількості кластерів

```
matrix = X.as_matrix()
for n_clusters in range(3,10):
    kmeans = KMeans(init='k-means++', n_clusters = n_clusters,
n_init=30)
    kmeans.fit(matrix)
    clusters = kmeans.predict(matrix)
    silhouette_avg = silhouette_score(matrix, clusters)
    print("For n_clusters =", n_clusters, "The average
silhouette_score is :", silhouette_avg)
```

У результаті використання методу силуетної оцінки, одержано результат, який показано на рис. 3.3.

```
For n_clusters = 3 The average silhouette_score is : 0.10071681758064248
For n_clusters = 4 The average silhouette_score is : 0.12208239761153944
For n_clusters = 5 The average silhouette_score is : 0.1470081849157512
For n_clusters = 6 The average silhouette_score is : 0.14389841472426354
For n_clusters = 7 The average silhouette_score is : 0.15212220110144017
For n_clusters = 8 The average silhouette_score is : 0.1558201267218184
For n_clusters = 9 The average silhouette_score is : 0.11656173409117862
```

Рис. 3.3. Результати силуетної оцінки кластерів

На практиці отримані вище оцінки можна вважати еквівалентними, оскільки, залежно від запуску, значення 0,1±0,05 будуть одержані для всіх кластерів з n_clusters > 3.

З іншого боку, встановлено, що крім визначених кластерів деякі кластери містять дуже мало елементів, тому варто розділити набір даних на 5 кластерів.

Щоб забезпечити хорошу сегментацію при кожному запуску програми, необхідно виконувати повторення до тих пір, поки не буде одержана найкраща можлива оцінка силуету, яка в даному випадку становить близько 0,15 (лістинг 3.7).

Лістинг 3.7. Навчання кластеризатора

```
n_clusters = 5
silhouette_avg = -1
while silhouette_avg < 0.145:
    kmeans = KMeans(init='k-means++', n_clusters = n_clusters,
n_init=30)
    kmeans.fit(matrix)
    clusters = kmeans.predict(matrix)
    silhouette_avg = silhouette_score(matrix, clusters)

    #km = kmodes.KModes(n_clusters = n_clusters, init='Huang',
n_init=2, verbose=0)
    #clusters = km.fit_predict(matrix)
    #silhouette_avg = silhouette_score(matrix, clusters)
    print("For n_clusters =", n_clusters, "The average
silhouette_score is :", silhouette_avg)
```

Найкраще середнє значення оцінки силуету при кількості кластерів 5 показано на рис. 3.4.

```
For n_clusters = 5 The average silhouette_score is : 0.14631355248870398
```

Рис. 3.4. Середня оцінка силуету

За допомогою програмного коду: «`pd.Series(clusters).value_counts()`» відображено кількість елементів у кожному кластері (рис. 3.5).

```
0    1009
2     964
3     829
1     606
4     470
dtype: int64
```

Рис. 3.5. Розподіл товарів за кластерами

Для того, щоб мати уявлення про якість кластеризації, можна представити силуетні оцінки кожного елемента різних кластерів. На рис. 3.6 показано силуети внутрішньокластерних оцінок.

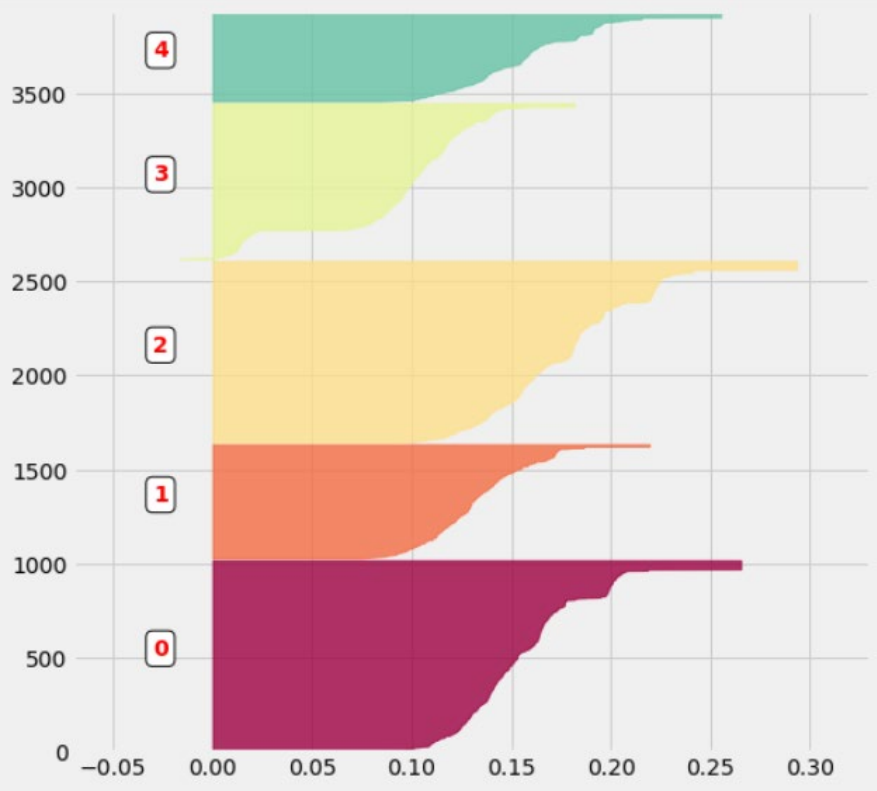


Рис. 3.6. Силуети внутрішньокластерної оцінки

Тепер можна відобразити тип об’єктів, які представляє кожен кластер. Щоб отримати загальне уявлення про їхній вміст, потрібно встановити, які ключові слова є найпоширенішими в кожному з них (лістинг 3.8 та рис. 3.7).

Лістинг 3.8. Представлення типів об’єктів у кластері

```
liste = pd.DataFrame(liste_produits)
liste_words = [word for (word, occurence) in list_products]

occurence = [dict() for _ in range(n_clusters)]

for i in range(n_clusters):
    liste_cluster = liste.loc[clusters == i]
    for word in liste_words:
        if word in ['art', 'set', 'heart', 'pink', 'blue',
'tag']: continue
        occurence[i][word] = sum(liste_cluster.loc[:,
0].str.contains(word.upper()))
```

Рис. 3.7. Тип об’єктів, представлених у кластері

З такого представлення видно, що, наприклад, один із кластерів містить об’єкти, які можна асоціювати з подарунками (ключові слова: Різдво, упаковка, листівка, ...).

Інший кластер містить предмети розкоші та ювелірні вироби (ключові слова: намисто, браслет, мереживо, срібло, ...). Тим не менш, можна також помітити, що багато слів з’являються в різних кластерах, і тому їх важко чітко розрізнити.

Щоб переконатися, що кластери дійсно відрізняються, потрібно проаналізувати їхню структуру. Враховуючи велику кількість змінних початкової матриці, спочатку виконується матрична декомпозиція PCA, яка представлена у лістингу 3.9.

Лістинг 3.9. Реалізація Principal Component Analysis

```
pca = PCA()
pca.fit(matrix)
pca_samples = pca.transform(matrix)
```

Після виконання лістингу 3.18 здійснюється перевірка величини дисперсії для кожного компонента. Візуалізація одержаних результатів показана на рис. 3.8.

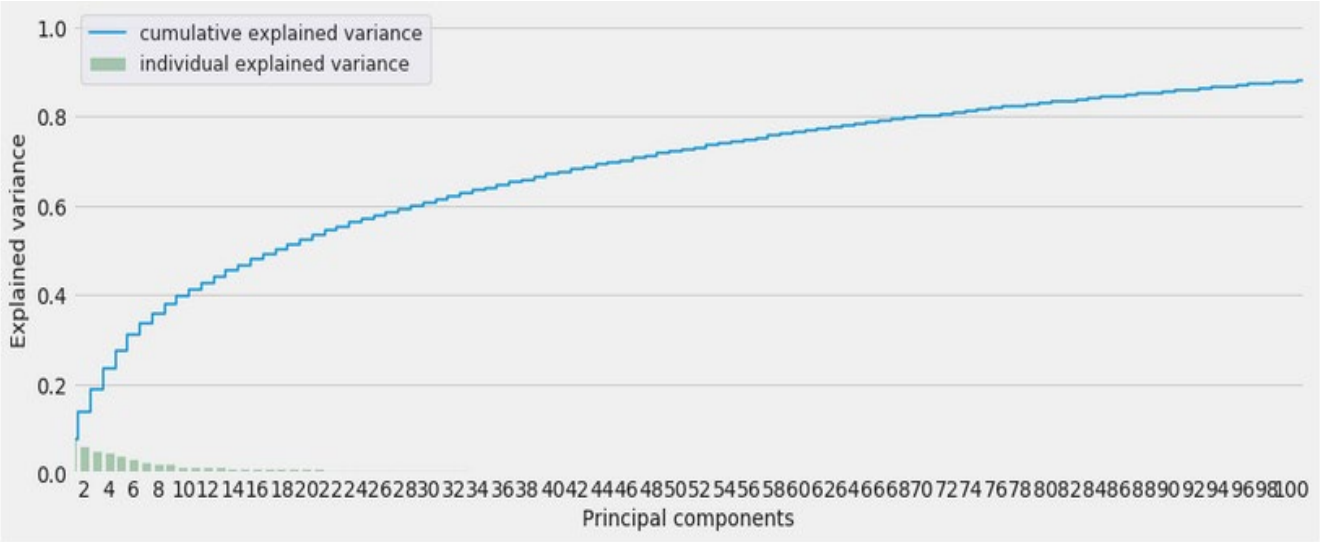


Рис. 3.8. Дисперсія компонентів

Як видно з рис. 3.8, кількість компонентів, необхідних для пояснення даних, надзвичайно важлива: потрібно більше 100 компонентів, щоб пояснити 90% дисперсії даних. На практиці зберігається лише обмежена кількість компонентів, оскільки ця декомпозиція виконується лише для візуалізації даних. У лістингу 3.10 наведено реалізацію декомпозиції матриці, що містить 50 компонентів.

Лістинг 3.10. PCA для 50 компонентів

```
pca = PCA(n_components=50)
matrix_9D = pca.fit_transform(matrix)
mat = pd.DataFrame(matrix_9D)
mat['cluster'] = pd.Series(clusters)
```

Візуалізація результатів матричної декомпозиції товарів за кластерами показано на рис. 3.9.

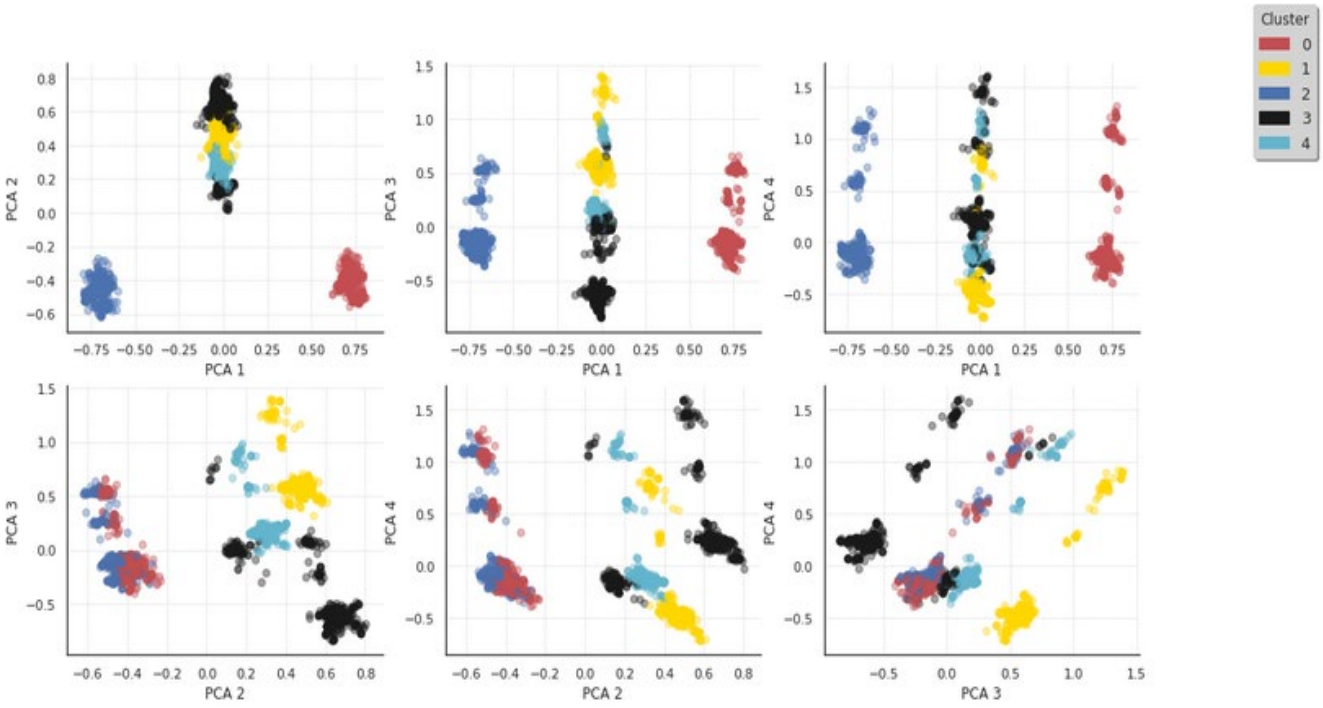


Рис. 3.9. Візуалізація розподілу елементів за кластерами

3.2. Сегментація користувачів

У попередньому підрозділі різні товари були згруповані у п’ять кластерів. Для продовження аналізу та виявлення сегментів покупців необхідно сформувати новий фрейм даних та забезпечити запис цієї інформації. Для цього створюється категоріальна змінна «categ_product», що зберігає номер кластера кожного товару (лістинг 3.11).

Лістинг 3.11. Створення категоріальної змінної «categ_product»

```
corresp = dict()
for key, val in zip (liste_produits, clusters):
    corresp[key] = val
#_____________________________________________________________________
_______
df_cleaned['categ_product'] = df_cleaned.loc[:,
'Description'].map(corresp)
```

Далі необхідно створити змінну «categ_N» ($N \in [0,4]$), що містить суму, витрачену на кожну категорію товарів (лістинг 3.12).

Лістинг 3.12. Реалізація змінної, що містить загальну вартість за категорію товарів

```
for i in range(5):
    col = 'categ_{}'.format(i)
    df_temp = df_cleaned[df_cleaned['categ_product'] == i]
    price_temp = df_temp['UnitPrice'] * (df_temp['Quantity'] -
df_temp['QuantityCanceled'])
    price_temp = price_temp.apply(lambda x:x if x > 0 else 0)
    df_cleaned.loc[:, col] = price_temp
    df_cleaned[col].fillna(0, inplace = True)
#_____________________________________________________________________
_____________________________
df_cleaned[['InvoiceNo', 'Description', 'categ_product', 'categ_0',
'categ_1', 'categ_2', 'categ_3','categ_4']][:5]
```

На рис. 3.10 показано результат виконання коду лістингів 3.11 і 3.12.

	InvoiceNo	Description	categ_product	categ_0	categ_1	categ_2	categ_3	categ_4
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	3	0.0	0.00	0.0	15.3	0.0
1	536365	WHITE METAL LANTERN	1	0.0	20.34	0.0	0.0	0.0
2	536365	CREAM CUPID HEARTS COAT HANGER	1	0.0	22.00	0.0	0.0	0.0
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	1	0.0	20.34	0.0	0.0	0.0
4	536365	RED WOOLLY HOTTIE WHITE HEART.	1	0.0	20.34	0.0	0.0	0.0

Рис. 3.10. Трансформований фрейм даних

До цього часу інформація, пов’язана з одним замовленням, була розділена на кілька рядків фрейму даних (по одному рядку на товар). Тому варто зібрати дані, пов’язані з певним замовленням, і ввести її в один запис. Для цього виконується програмний код, наведений у лістингу 3.13, що в результаті дає змогу зберігати для кожного замовлення кількість товарів у кошику, а також спосіб його розподілу на 5 кластерів.

Лістинг 3.13. Формування замовлень за споживчими кошиками

```
temp = df_cleaned.groupby(by=['CustomerID', 'InvoiceNo'],
as_index=False)['TotalPrice'].sum()
basket_price = temp.rename(columns = {'TotalPrice':'Basket
Price'})
for i in range(5):
    col = 'categ_{}'.format(i)
    temp = df_cleaned.groupby(by=['CustomerID',
'InvoiceNo'], as_index=False)[col].sum()
    basket_price.loc[:, col] = temp
df_cleaned['InvoiceDate_int'] =
df_cleaned['InvoiceDate'].astype('int64')
temp = df_cleaned.groupby(by=['CustomerID', 'InvoiceNo'],
as_index=False)['InvoiceDate_int'].mean()
df_cleaned.drop('InvoiceDate_int', axis = 1, inplace =
True)
basket_price.loc[:, 'InvoiceDate'] =
pd.to_datetime(temp['InvoiceDate_int'])
basket_price = basket_price[basket_price['Basket Price'] >
0]
basket_price.sort_values('CustomerID', ascending =
True)[:5]
```

На рис. 3.11 наведено модифікований дата фрейм з категоріальними змінними та сформованим кошиком товарів.

	CustomerID	InvoiceNo	Basket Price	categ_0	categ_1	categ_2	categ_3	categ_4	InvoiceDate
1	12347	537626	711.79	187.2	293.35	23.40	83.40	124.44	2010-12-07 14:57:00.000001024
2	12347	542237	475.39	130.5	169.20	84.34	91.35	0.00	2011-01-26 14:29:59.999999744
3	12347	549222	636.25	330.9	115.00	81.00	109.35	0.00	2011-04-07 10:42:59.999999232
4	12347	556201	382.52	74.4	168.76	41.40	78.06	19.90	2011-06-09 13:01:00.000000256
5	12347	562032	584.91	109.7	158.16	61.30	157.95	97.80	2011-08-02 08:48:00.000000000

Рис. 3.11. Фрейм даних зі сформованими споживчими кошиками

Колонка даних «basket_price» містить інформацію за період 12 місяців. Пізніше однією з цілей може бути побудова моделі, здатної характеризувати та передбачати звички клієнтів, які відвідують сайт електронної комерції, включно з першим відвідування. Щоб мати можливість протестувати модель реалістичним способом, виконано декомпозицію набору даних на навчальний, що містить транзакції за перших 10 місців і тестовий – решта даних дата фрейму.

Далі виконується формування корзини користувача, тобто групування різних записів, які відповідають одному і тому ж користувачу. Таким чином визначається кількість покупок, зроблених користувачем, а також мінімальна, максимальна, середня суми та загальна вартість, витрачена за всі відвідування сайту електронної комерції. У лістингу 3.14 наведено формування користувацьких споживчих корзин.

Лістинг 3.14. Формування корзин користувачів

```
transactions_per_user=basket_price.groupby(by=['CustomerID'])[
'Basket Price'].agg(['count','min','max','mean','sum'])
for i in range(5):
    col = 'categ_{}'.format(i)
    transactions_per_user.loc[:,col] =
basket_price.groupby(by=['CustomerID'])[col].sum() /\

transactions_per_user['sum']*100

transactions_per_user.reset_index(drop = False, inplace =
True)
basket_price.groupby(by=['CustomerID'])['categ_0'].sum()
transactions_per_user.sort_values('CustomerID', ascending =
True)[:5]
```

Результат сформованих користувацьких корзин показано на рис. 3.12.

	CustomerID	count	min	max	mean	sum	categ_0	categ_1	categ_2	categ_3	categ_
0	12347	5	382.52	711.79	558.172000	2790.86	29.836681	32.408290	10.442659	18.636191	8.676
1	12348	4	227.44	892.80	449.310000	1797.24	41.953217	0.000000	38.016069	20.030714	0.000
2	12350	1	334.40	334.40	334.400000	334.40	48.444976	0.000000	11.692584	39.862440	0.000
3	12352	6	144.35	840.30	345.663333	2073.98	12.892120	15.711338	0.491808	56.603728	14.30
4	12353	1	89.00	89.00	89.000000	89.00	13.033708	0.000000	0.000000	64.606742	22.35

Рис. 3.12. Сформовані користувацькі корзини

Додатково у фреймі рис. 3.12 створено дві додаткові змінні, які визначають кількість днів, що минули з моменту першої покупки і кількість днів з моменту останньої покупки.

Особливий інтерес становить категорія клієнтів, які здійснюють лише одну покупку. Однією з цілей може бути, наприклад, орієнтація на цих клієнтів, щоб утримати їх. Зазвичай цей тип клієнтів становить приблизно 1/3 від загальної кількості клієнтів.

Стовбець даних «Transactions_per_user» містить суму всіх виконаних замовлень. Кожен запис у цьому фреймі даних відповідає конкретному клієнту. Цю інформацію можна використати для визначення характеристики різних типів клієнтів. Програмно реалізується, як показано у лістингу 3.15.

Лістинг 3.15. Формування матриці транзакцій за користувачами

```
list_cols =
['count','min','max','mean','categ_0','categ_1','categ_2'
,'categ_3','categ_4']
#_____________________________________________________
selected_customers = transactions_per_user.copy(deep =
True)
matrix = selected_customers[list_cols].as_matrix()
```

Оскільки різні змінні мають досить різні діапазони області визначення, то перед продовженням аналізу потрібно стандартизувати значення утвореної у лістингу 3.24 матриці. Для цього використовується лістинг 3.16.

Лістинг 3.16. Стандартизація значень дата фрейму

```
scaler = StandardScaler()
scaler.fit(matrix)
print('variables mean values: \n' + 90*'-' + '\n' ,
scaler.mean_)
scaled_matrix = scaler.transform(matrix)
```

На рис .3.13 показано середні значення за стовпцями матриці.

```
variables mean values:
-----------------------------------------------------------------------
[  3.62305987 259.93189634 556.26687999 377.06036244  25.22916919
  16.37327913  13.98907929  28.73795868  15.67936332]
```

Рис. 3.13. Стандартизований дата фрейм

Перш ніж створювати кластери користувачів сайту електронної комерції, важливо визначити базис меншої розмірності, що дозволяє описати матрицю scaled_matrix. У цьому випадку, базис використовується для створення представлення різних кластерів і таким чином виконується перевірка якості поділу різних груп (лістинг 3.17).

Лістинг 3.17. Виконання PCA перетворення над матрицею користувачів

```
pca = PCA()
pca.fit(scaled_matrix)
pca_samples = pca.transform(scaled_matrix)
```

Величина дисперсії, що пояснює кожен із компонентів матриці показана на рис. 3.14.

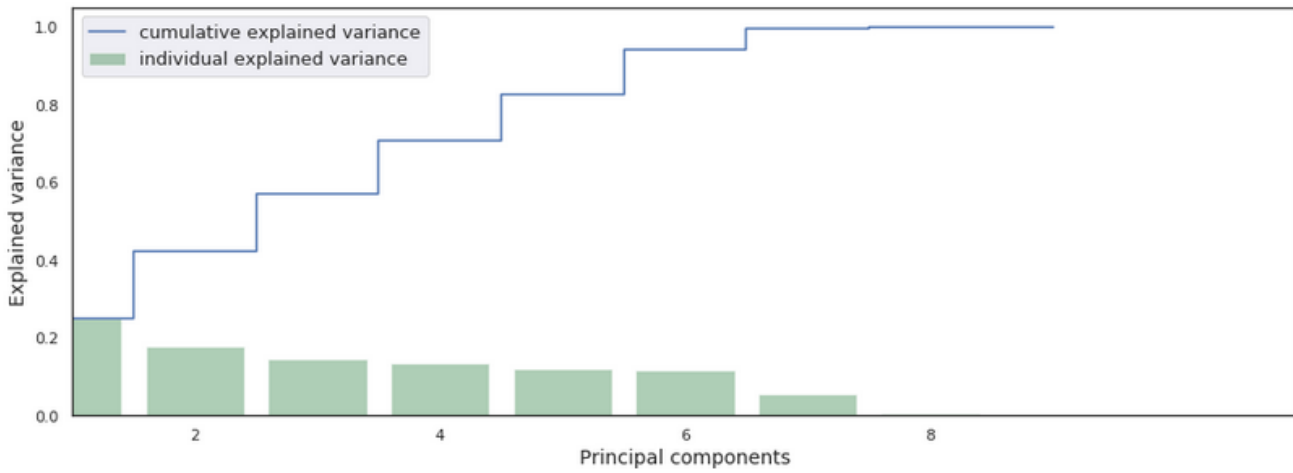


Рис. 3.14. Величина дисперсії за компонентами матриці

Далі необхідно визначити кластери клієнтів із стандартизованої матриці, яка була визначена раніше, за допомогою алгоритму k-means від scikit-learn. Кількість кластерів обирається на основі оцінки силуету, і найкращу оцінку отримано з 11 кластерами (лістинг 3.18).

Лістинг 3.18. Реалізація алгоритму k-means для визначення кластерів покупців

```
n_clusters = 11
kmeans = KMeans(init='k-means++', n_clusters =
n_clusters, n_init=100)
kmeans.fit(scaled_matrix)
clusters_clients = kmeans.predict(scaled_matrix)
silhouette_avg = silhouette_score(scaled_matrix,
clusters_clients)
print('score de silhouette:
{:<.3f}'.format(silhouette_avg))
```

Кількість покупців у кожному кластері показано на рис. 3.15.

	7	3	5	8	0	6	4	1	2	9	10
nb. clients	1476	520	376	335	287	243	191	151	12	10	7

Рис. 3.15. Розподіл клієнтів по кластерах

Як видно з рис. 3.15, існує певна нерівність у розмірах різних створених груп. Тому варто зрозуміти зміст цих кластерів, щоб підтвердити або спростувати цю конкретну декомпозицію. Спочатку використовується результат PCA (лістинг 3.19):

Лістинг 3.19. PCA над матрицею «scaled_matrix»

```
pca = PCA(n_components=6)
matrix_3D = pca.fit_transform(scaled_matrix)
mat = pd.DataFrame(matrix_3D)
mat['cluster'] = pd.Series(clusters_clients)
```

У результаті виконання лістингу 3.19, одержано та візуалізовано кластери користувачів, які представлено на рис. 3.16.

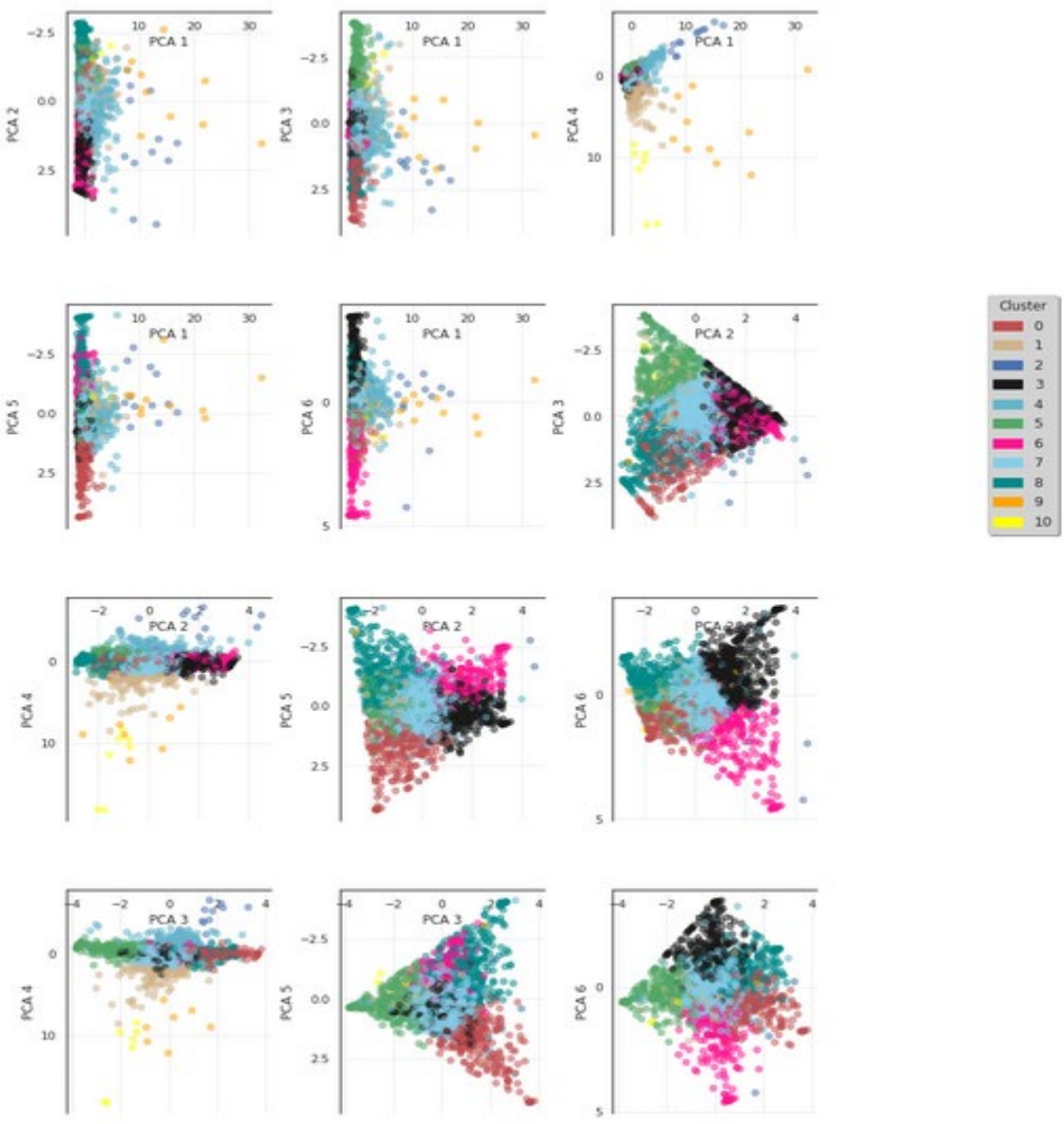


Рис. 3.16. Кластери користувачів на основі виконання PCA

З представлення (рис. 3.16) видно, наприклад, що перший головний компонент дозволяє відокремити найдрібніші кластери від решти. У більш загальному вигляді видно, що завжди існує представлення, в якому два кластери виглядають різними.

Як і у випадку з категоріями товарів, інший спосіб проаналізувати якість поділу – це проаналізувати показники силуетів у різних кластерах (лістинг 3.20).

Лістинг 3.20. Аналіз поділу кластерів на основі силуетів у різних кластерах

```
sample_silhouette_values = silhouette_samples(scaled_matrix,
clusters_clients)
#_____________________________________________
# define individual silouhette scores
sample_silhouette_values = silhouette_samples(scaled_matrix,
clusters_clients)
#_____________________
# and do the graph
graph_component_silhouette(n_clusters, [-0.15, 0.55],
len(scaled_matrix), sample_silhouette_values,
clusters_clients)
```

На рис. 3.17 візуалізовано результати аналізу, одержані у лістингу 3.29.

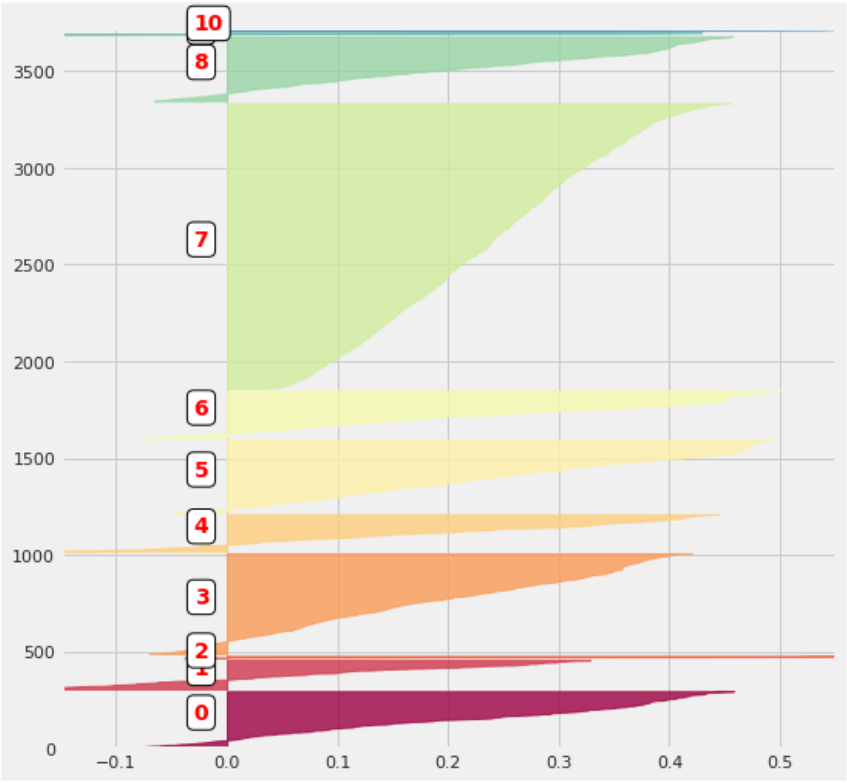


Рис. 3.17. Результати аналізу силуетів у різних кластерах

Аналізуючи всі попередні маніпуляції над даними, можна сказати, що різні кластери дійсно не перетинаються (принаймні, у глобальному плані). Залишилося

зрозуміти природу клієнтів у кожному кластері. Для цього потрібно додати до фрейму даних «selected_customers» змінну, яка визначає кластер, до якого належить кожен клієнт. Після цього необхідно усереднити вміст цього фрейму, спочатку вибравши різні групи клієнтів. Це дає доступ, наприклад, до середньої ціни кошиків, кількості відвідувань або загальних сум, витрачених клієнтами різних кластерів.

Далі виконується реорганізація вмісту фрейму шляхом впорядкування різних кластерів: спочатку по відношенню до кількості, надісланої в кожній категорії товарів, а потім, відповідно до загальної витраченої суми.

Результат наведених вище операцій показано на рис. 3.18.

	cluster	count	min	max	mean	sum	categ_0	categ_1	categ_2
0	3.0	2.446154	216.358404	330.827060	270.429001	672.531521	55.246825	8.370952	13.43995
1	0.0	2.160279	200.139826	333.617596	260.539379	657.522683	13.770524	51.530571	6.635269
2	6.0	2.201646	194.651235	319.070453	248.466047	587.264650	18.288488	6.308356	56.47669
3	5.0	2.353723	203.287606	349.455426	270.570636	696.040479	10.827895	6.655407	5.213568
4	8.0	2.465672	194.230597	309.968269	246.106019	622.394060	11.368548	13.067899	5.221641
5	7.0	3.346206	217.525984	466.041620	331.824516	1117.975591	24.913742	17.185747	13.27564
6	4.0	1.712042	1040.133298	1367.670424	1191.267217	2094.760529	26.142761	17.032076	11.85062
7	2.0	1.666667	3480.920833	3966.812500	3700.139306	5949.600000	20.102624	15.171169	22.89073
8	1.0	18.357616	88.416556	1623.636821	580.211471	10054.910066	23.931098	16.139653	12.24869
9	10.0	92.000000	10.985714	1858.250000	374.601553	34845.105714	25.832531	13.402971	13.11758
10	9.0	23.200000	415.148000	17158.271000	4853.774161	87883.059000	22.089309	18.726071	7.172572

Рис. 3.18. Результат виконання маніпуляцій у фреймі

Таким чином, було створено представлення різних морфотипів користувачів. Завершальним етапом кластеризації покупців є візуалізація кластерів за допомогою «Радарних діаграм» (рис. 3.19 та рис. 3.20).

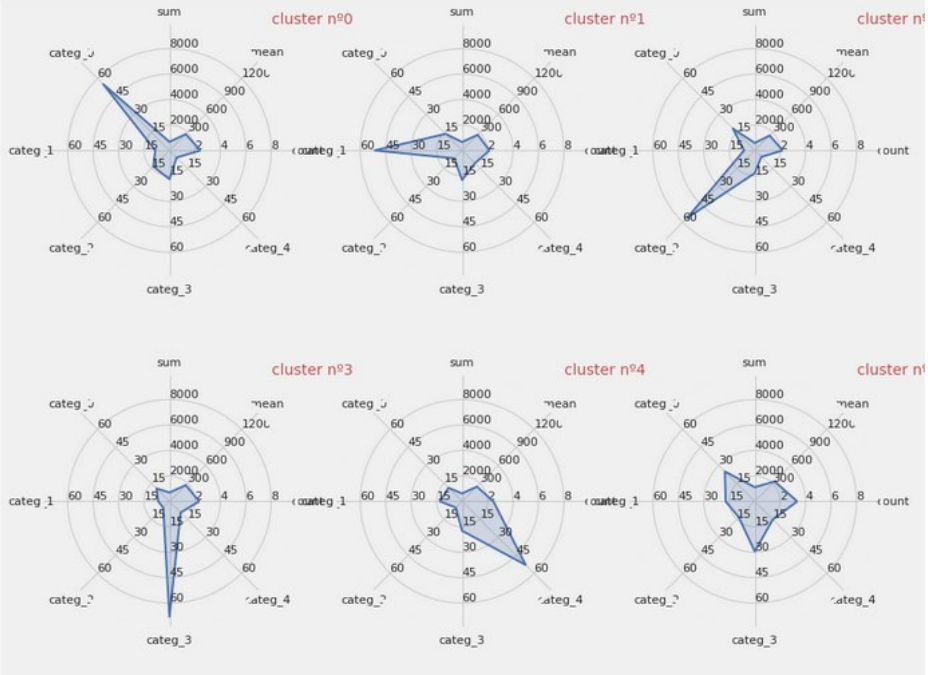


Рис. 3.19. Візуалізація перших 6-ти кластерів покупців

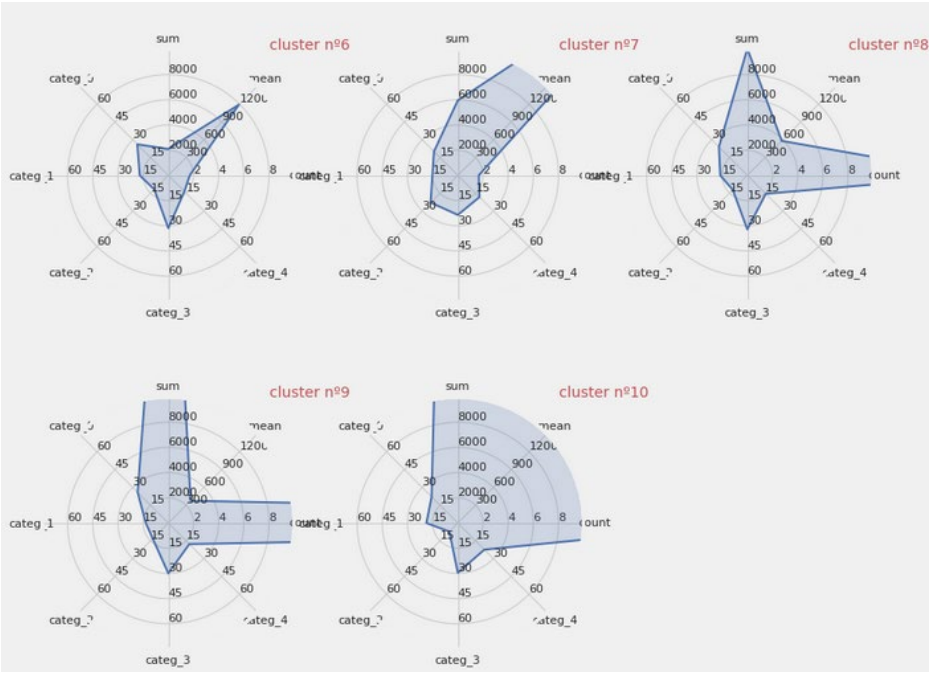


Рис. 3.20. Візуалізація останніх 5-ти кластерів покупців

Можна помітити, наприклад, що перші 5 кластерів відповідають значній перевазі покупок у певній категорії товарів. Інші кластери відрізнятимуться від середніх показників кошика, загальної суми, витраченої клієнтами або загальної кількості здійснених відвідувань.

3.3. Висновки до розділу

1. Реалізовано модель сегментації товарів, що враховує властивості товарів на основі аналізу їх опису, що дало змогу виділити 5 кластерів товарів, які з точки зору метрики внутрішньокластерного силуету є найбільшим оптимальною і становить приблизно 0,15, а кількість товарів у кластерах коливається від 470 до 1009 найменувань.

2. Проведено сегментацію користувачів на 11 основних кластерів, що враховують тип продуктів, які вони зазвичай купують, кількість відвідувань сайту електронної комерції і суми, які вони витратили протягом 10 місяців, що в подальшому дає змогу проводити прогнозування і класифікацію нових покупців на основі вказаних критеріїв та визначених кластерів.

РОЗДІЛ 4
ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1. Охорона праці

У кваліфікаційній роботі магістра досліджено методи та засоби сегментації множини користувачів при проектуванні та експлуатації комп'ютерних маркетингових систем. Обов'язковим елементом дослідження є визначення та аналіз вимог з охорони праці і техніки безпеки при розробці програмного засобу і проведенні експериментальних досліджень, що супроводжується використанням комп'ютерної техніки. Дотримання норм і правил охорони праці є важливим аспектом у контексті дотримання норм організації робочого місця, забезпечення комфортних та зручних умов праці осіб, які беруть участь у процесі, а це вимагає дослідження та дотримання вимог з охорони праці.

В Україні розроблено й діють ряд нормативних документів, які визначають вимоги і правила щодо використання комп'ютерної техніки, приміщень з екранними пристроями та ін. Основним нормативним документом при використанні комп'ютерної техніки є НПАОП 0.00-7.15-18 «Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями». Він регламентує, що приміщення для експлуатації комп'ютерної техніки повинно розміщуватися в північній або північно-східній частині будівлі. Площа одного робочого місця повинна становити щонайменше 6 м^2, об'єм — щонайменше 20 м^3, відстань між робочими столами — щонайменше 2,5 м у ряду і 1,2 м між рядами. Стіни приміщень потрібно фарбувати у пастельні тони з коефіцієнтом відбиття 0,5-0,6 [26].

З метою зменшення напруження очей потрібно, щоб відстань між краями сусідніх точок зображення на моніторі не перевищувала гранично оптимальний розмір літеро-цифрових знаків – 16-20, складних знаків – 35-40. Оптимальні співвідношення параметрів літер і цифр такі: ширина знака – 0,75 їх висоти,

товщина ліній при зворотному контрасті – 1/6-1/8, відстань між знаками — 0,25-0,5 висоти знака, між словами – 0,75-1, між рядками – 0,5-1 [26].

Для профілактики загальної втоми і особливо зорового аналізатора важливе значення має організація режиму праці та відпочинку. Загальна тривалість робочого дня не повинна перевищувати 8 год. Частота і тривалість перерв залежать від типу та інтенсивності виконуваних робіт. Під час робіт, які виконуються з великим навантаженням, рекомендуються перерви на 10-15 хв. через кожну годину, а при неінтенсивній і монотонній роботі — на 10-15 хв. через кожні дві години. Кількість мікропауз (тривалістю до хвилини) потрібно регулювати індивідуально.

Зміст регламентованих перерв може бути різний: виробнича гімнастика (вправи для очей, гімнастика, спрямована на корекцію вимушеної робочої пози, поліпшення венозного кровообігу, часткову дисфункцію рухової активності), альтернативна допоміжна робота, приймання їжі тощо.

Для того, щоб особи, які займаються проектуванням та оцінюванням якості людино-машинної взаємодії меншою мірою втомлювались і зберігали високий рівень працездатності, потрібно раціонально організовувати їхні робочі місця. Зокрема, робоче місце має відповідати основним антропометричним даним людини. Крісло або стілець на робочому місці повинні мати висоту сидіння 40-50 см від рівня підлоги, а також відповідний кут нахилу спинки.

Монітори потрібно розміщувати на висоті рівня очей (висота від підлоги до нижнього краю екрана має становити 95-100 см) на відстані 60-70 см від оператора (відстань від краю столу — 50-70 см). Кут зору працюючого щодо екрану має дорівнювати 10-20°, але не більше 40°, кут між верхнім краєм монітора і рівнем очей користувача має становити менш як 10°. Найдоцільніше розміщувати екран перпендикулярно до лінії погляду користувача. Кут нахилу екрана по вертикалі має становити 0-30° [26]. З цією метою сучасні монітори комплектують підставкою з поворотним кронштейном, що дає змогу регулювати кут нахилу монітора і горизонтально обертати його навколо вертикальної осі. Висоту екрана від поверхні підлоги регулюють змінюючи висоту робочої поверхні столу. Іноді монітори

встановлюють на спеціальні підставки, що уможливлює його переміщення у просторі у вертикальному та горизонтальному напрямах.

У приміщеннях, де виконуються роботи на ПК, повинно бути передбачене природне і загальне штучне освітлення. Робочі місця користувачів потрібно розміщувати так, щоб у поле зору не потрапляли вікна і освітлювальні прилади (монітори потрібно розміщувати під кутом 90-105° до вікон і на відстані 2,5-4 м від стін і віконних прорізів). У поле зору користувача не повинні потрапляти поверхні, що відбивають світло. Покриття столу має бути матовим з коефіцієнтом відбиття 0,25-0,4.

Для штучного освітлення приміщення рекомендується застосовувати світильники матового світла з розсіювачами, а спектральний склад ламп має наближатися до спектру сонячного світла (наприклад, люмінесцентні типу ЛБ). Оптимальна освітленість робочих місць — 400-500 лк.

У разі ураження електричним струмом необхідно терміново звільнити потерпілого від дії електричного струму (через відключення електроживлення в кімнаті, загального електроживлення на розподільному щиті або іншим способом). Викликати швидку медичну допомогу (подзвонивши за міським телефоном 103). Надати першу медичну допомогу потерпілому, враховуючи наступне:

– якщо потерпілий знепритомнів, але дихає, його необхідно рівно і зручно вкласти, розстебнути одяг, створити приплив свіжого повітря і забезпечити повний спокій;

– при відсутності ознак життя до прибуття лікарів потерпілому необхідно робити штучне дихання.

Дизайнери та архітектори інтерфейсів людино-машинної взаємодії при виконанні відповідних робіт несуть відповідальність за порушення вимог з охорони праці і правил техніки безпеки.

При дослідженні та розробці методу і засобу сегментації множини користувачів при проектуванні та експлуатації комп’ютерних маркетингових систем було дотримано усіх вище наведених вимог нормативних документів щодо охорони праці і техніки безпеки при експлуатації комп’ютерної техніки.

4.2. Здоровий спосіб життя людини та його вплив на професійну діяльність

Здоров'я людини ґрунтується на основі генетичних факторів, способу життя та екологічних умов. Однак певною мірою воно залежить також від свідомого ставлення людини до себе та оточуючого середовища [23].

Здоров'я людини – стан повного соціально-біологічного комфорту коли функція всіх органів і систем організму виважені з природним і соціальним середовищем, відсутні будь-які хвилювання, хворобливі стани та фізичні дефекти.

Критерій здоров'я визначається комплексом показників. Однак за найзагальнішими рисами здоров'я індивідуума можна визначити як природний стан організму, що характеризується повною зрівноваженістю будь-яких виражених хворобливих змін [24].

Здоров'я залежить від багатьох факторів, які об'єднуються в одне інтегральне поняття – здоровий спосіб життя. Його метою є навчити людину розумно ставитися до свого здоров'я, фізичної та психічної культури, загартовувати свій організм, вміло організовувати працю і відпочинок.

До основних складових здорового способу життя, згідно [24], належать :

1. Спосіб життя. Має велике значення для здоров'я людини і складається з чотирьох категорій:
   - економічної (рівень життя),
   - соціологічної (якість життя),
   - соціально-психологічної (стиль життя),
   - соціально-економічної (устрій життя)
2. Рівень культури. Культура – це самосвідоме ставлення до самого себе. Однак люди дуже часто нехтують своїм здоров'ям, ведуть неправильний спосіб життя, не дотримуються режиму, переїдають, курять. Тому для здоров'я потрібні знання, які увійшли б у повсякденну звичку людини [24].
3. Здоров'я в ієрархії потреб. Не завжди в житті людини здоров'я займає перше місце порівняно з речами та іншими матеріальними благами. У результаті

це призводить до шкоди не лише своєму здоров'ю, а й здоров'ю майбутніх поколінь. Отже, здоров'я повинно займати перше місце в ієрархії потреб людини [24].

4. Мотивування. На превеликий жаль, ціну здоров'я більшість людей усвідомлює лише тоді, коли воно значно втрачено. Тільки тоді виникає прагнення вилікувати захворювання, стати здоровим [24].

5. Зворотні зв'язки – нерозумне і довге випробовування стійкості свого організму нездоровим способом життя (алкоголь, нікотин). Тільки через певний час спрацьовують зворотні зв'язки людини, коли вона кидає шкідливі звички, проте це вже доволі часто надто запізно.

6. Навчання здоровому способу життя. Джерелом навичок з цього питання є передусім приклад батьків, допомагає також і медична освіта. Важливим фактором, що визначає реакцію людини на екстремальну ситуацію, є її психофізичні якості та загальний стан. Вони проявляються через чутливість людини до виявлення сигналів небезпеки, перед реакцією на ці сигнали. Показники, які зумовлюють можливості людини виявити небезпечну ситуацію та адекватно реагувати на неї, залежать від індивідуальних особливостей, зокрема від її нервової системи. На поведінку людини у небезпечній ситуації впливає й її психічний та фізичний стан.

7. Психічний стан. Сучасна людина зустрічається з багатьма факторами ризику, що негативно впливають на стан нервової та серцево-судинної систем, знижує опірність організму. При цьому виникає стресова реакція організму. Так, наприклад, психічна травма, отримана внаслідок конфлікту, виводить людину з нормального психічного стану, що може призвести до суттєвих змін у виконанні професійних функцій і загального функціонального стану. У перекладі «стрес» означає «напруження», тобто відповідь організму на поставлену перед ним проблему.

Стрес – це сукупність загальних неспецифічних біохімічних, фізіологічних і психологічних реакцій організму внаслідок дії надзвичайних подразників різної природи і характеру, які викликають порушення функцій органів.

Повне звільнення від стресу означає смерть, тому слабкий стрес є нормальним явищем у житті і потрібним для реалізації людської повноцінності. Однак якщо він інтенсивний і довготривалий, то може стати основою розвитку захворювань або зумовити смерть.

Медичні та соціологічні дослідження серед різних категорій населення показують, що люди по-різному реагують на надзвичайні ситуації. Є люди, стресостійкі до побутових негараздів, але дуже стресореактивні до сімейних проблем та невдач у коханні, інші боляче сприймають невдачі на роботі, ще інші – втрату соціального статусу.

Відомо, що в осіб до 30 років життєві потреби значно більші, ніж у людей старшого віку, а відтак стресові стани у них переважають.

Велике значення для розвитку стресового стану має поведінка в екстремальних умовах (аварія, кримінальна ситуація, стихійне лихо). Неправильна поведінка у таких ситуаціях найчастіше є причиною шкідливих наслідків стресу. Вона зумовлює результат стресу більше, ніж фактори зовнішнього середовища. У цих випадках стрес може виявитись у вигляді паніки, суєти, істерики.

Стійкість організму до різноманітних стресових станів є дуже індивідуальною. Деякі люди без усіляких наслідків переносять надзвичайно складні екстремальні ситуації, ніколи не непритомніють, не втрачають сили волі, психологічної рівноваги. Інші вже при незначних екстремальних ситуаціях втрачають витримку і віру в себе.

Для загартованості психічного стану людині треба використовувати фізичну працю, заняття спортом, прогулянки на свіжому повітрі та інші природні фактори.

По-друге, уміння володіти собою, керувати емоціями, психоемоційним напруженням. Це значить постійно контролювати свої дії, вчинки, залишатися врівноваженим навіть у найбільш напружених обставинах

Слід зазначити, що існують різноманітні психологічні засоби зняття нервового напруження для відновлення працездатності, до яких належать: психотерапія, психопрофілактика, психогігієна.

Для більш швидкого відновлення сил після втоми рекомендується використовувати навіюваний сон, тобто навчитися вводити себе на певний час у сон і самостійно виходити з нього бадьорим. Тривалість навіюваного сну 30 – 40 хв.

Найважливішим для людини є її фізичний стан здоров'я, який залежить як від біологічних факторів (спадковості), так і від складного комплексу соціальних, економічних, гігієнічних, кліматогеографічних та інших умов навколишнього середовища.

Під впливом несприятливих факторів рівень фізичного стану здоров'я знижується, а поліпшення умов сприяє його підвищенню.

Виходячи із концепції фізичного здоров'я, основним його критерієм слід вважати енергопотенціал біосистеми, оскільки життєдіяльність будь-якого живого організму залежить від акумуляції і мобілізації енергії для забезпечення фізичних функцій.

Здоров'я людини, опірність її організму до несприятливих умов навколишнього середовища, працездатність значною мірою залежать від харчування. Правильне і раціональне харчування є важливим фактором забезпечення життєдіяльності людини, росту та розвитку організму, запобігання та лікування хвороб, у тому числі й тих, які сталися внаслідок надзвичайних ситуацій.

Важливим фактором фізичного здоров'я є загартування організму. В основі загартування лежить властивість організму людини пристосовуватись до зміни умов навколишнього середовища. У людини відбувається процес пристосування організму до нових умов існування – виникає адаптація.

Висновок.

Отже, здоровий спосіб життя людини безпосередньо впливає на професійну діяльність. Сучасна професійна діяльність негативно позначається на здоров’ї і якості життя людини, однак з дотримання визначених правил і піклування про своє здоров’я, працівники можуть зменшити негативний вплив стресів та психологічного тиску, фізичного перенавантаження, а також підвищити продуктивність праці.

ВИСНОВКИ

Основні наукові та практичні результати полягають в наступному.

1. Проведено аналіз задач у сфері проектування маркетингових комп’ютерних систем, що дало змогу виявити основні з них, зокрема, це стосується визначення трендів і динаміки зміни ринку у певній галузі, встановлення факторів впливу на розширення та стимулювання ринку, сегментації груп товарів і користувачів.

2. Проаналізовано вимоги, принципи та підходи до сегментації, які стосуються маркетингових досліджень та автоматизації відповідних процесів, що дало змогу обґрунтувати необхідність впровадження алгоритмів і методів машинного навчання при виявленні груп покупців і товарів, а також ринків в цілому.

3. Проаналізовано методи і критерії сегментації користувачів і товарів, що дало змогу встановити комплексні та атомарні показники, які необхідно враховувати при автоматизованому визначенні груп товарів і покупців, що не містять цільових міток сегментів.

4. Запропоновано, спроектовано та проаналізовано базову архітектуру маркетингової комп’ютерної системи сегментації користувачів і товарів, що використовує дані з CRM-систем і систем обліку показників бізнес-діяльності підприємства, а також передбачає необхідність реалізації підсистеми автоматизованого визначення груп товарів і покупців на основі методів кластеризації, що дає змогу знизити та оптимізувати ресурси на аналіз маркетингової інформації і дозволяє приймати оптимальні рішення при зміні кон’юнктури ринку.

5. Запропоновано та обґрунтовано застосування методу сегментації користувачів, що використовує алгоритм k-means і дає змогу автоматизувати процес групування об’єктів ринку без відповідних міток, а також використовувати метрики якості процесу сегментації.

6. Реалізовано модель сегментації товарів, що враховує властивості товарів на основі аналізу їх опису, що дало змогу виділити 5 кластерів товарів, які з точки зору метрики внутрішньокластерного силуету є найбільшим оптимальною і становить приблизно 0,15, а кількість товарів у кластерах коливається від 470 до 1009 найменувань.

7. Проведено сегментацію користувачів на 11 основних кластерів, що враховують тип продуктів, які вони зазвичай купують, кількість відвідувань сайту електронної комерції і суми, які вони витратили протягом 10 місяців, що в подальшому дає змогу проводити прогнозування і класифікацію нових покупців на основі вказаних критеріїв та визначених кластерів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Філановський О. Головна маркетингова книга. Фабула. 2018 р. 304 с.

2. Гладуелл М. Поворотний момент. Як дрібні зміни спричиняють великі зрушення. Family Leisure Club. 2017 р. - 163 с.

3. Карада Ю. Прибыльный маркетинг. Litres. 2019 р. 187 с.

4. Python-recsys on Github. URL: https://github.com/ocelma/python-recsys (дата звернення 22.09.2021 р).

5. Preprocessing data. URL: https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing (дата звернення 02.05.2021 р.).

6. API reference. URL: https://pandas.pydata.org/docs/reference/index.html (дата звернення 10.09.2021 р.).

7. NumPy Reference. URL: https://numpy.org/doc/stable/reference/index.html (дата звернення 12.10.2021 р.)

8. Барсегян А., Куприянов М.,Степаненко В., Холод И. Технологии анализа данных. СПб. : Изд-во " БХВ-Петербург". 2008. 384 с.

9. Савчук Т.О. Застосування кластерного аналізу для колаборативної фільтрації. Вісник Хмельницького національного університету. №1. 2011. С. 186-192.

10. Лексин В.А. Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов. Математические методы распознавания образов-13. М. МАКС Пресс. 2007. С. 488-491.

11. Garbade M. J. Understanding K-means Clustering in Machine Learning. URL: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1 (дата звернення 10.11.2021 р.).

12. Pulkit Sharma. The Most Comprehensive Guide to K-Means Clustering You'll Ever Need. URL: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/ (дата звернення 12.11.2021 р.)

13. Shirkhorshidi A. S. Big data clustering: a review. International Conference on Computational Science and Its Applications. Springer, Cham, 2014. pp. 707-720.

14. Kurasova O. Strategies for big data clustering/ O. Kurasova et al. // 2014 IEEE 26th International Conference on Tools with Artificial Intelligence. IEEE, 2014. pp. 740-747.

15. B. Panda at al. MapReduce and its application to massively parallel learning of decision tree ensembles // Scaling Up Machine Learning: Parallel and Distributed Approaches. 2012. P.

16. Li Deng and Dong Yu. Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing, Vol. 7. N. 3-4. 2014. pp. 197–387.

17. Bengio Y. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35. N. 8. August 2013. pp. 1798–1828.

18. LeCun, Y., Bengio, Y., Hinton, G. Deep learning. Nature.2015. 521 (7553). pp. 436–444.

19. Golovko, V.A. Deep learning: an overview and main paradigms. Optical Memory and Neural Networks (Information Optics). Vol. 26, № 1. 2017. pp. 1–17.

20. Oliveira, T.P., Barbar, J.S., Soares, A.S.: Multilayer Perceptron and Stacked Autoencoder for Internet Traffic Prediction. In: IFIP International Conference on Network and Parallel Computing. Springer, Heidelberg. 2014. pp. 61–71.

21. Микитишин А.Г., Митник М.М., Стухляк П.Д., Пасічник В.В. Комп'ютерні мережі. Книга 2. Львів, «Магнолія 2006», 2014. 312 с.

22. Микитишин А.Г., Митник М.М., Стухляк П.Д. Телекомунікаційні системи та мережі. Тернопіль: Вид-во ТНТУ імені Івана Пулюя, 2016. 384 с.

23. НПАОП 0.00-7.15-18 «Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями». Київ. 2018.

24. Катренко Л.А., Катренко А.В. Охорона праці в галузі комп'ютингу. Львів: Магнолія-2006. 2012. 544 с.

25. Желібо Е.Н. Безпека життєдіяльності: Навчальний посібник. Київ: «Каравела», Львів: «Новий світ - 2000». 2001. 320с.

Додаток А

Тези конференцій

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Тернопільський національний технічний університет імені Івана Пулюя (Україна)
Університет імені П'єра і Марії Кюрі (Франція)
Маріборський університет (Словенія)
Технічний університет у Кошице (Словаччина)
Вільнюський технічний університет ім. Гедимінаса (Литва)
Білоруський національний технічний університет (Республіка Білорусь)
Міжнародний університет цивільної авіації (Марокко)
Наукове товариство ім. Т.Шевченка

АКТУАЛЬНІ ЗАДАЧІ СУЧАСНИХ ТЕХНОЛОГІЙ

Збірник
тез доповідей
Том І

X Міжнародної науково-практичної конференції молодих учених та студентів
24-25 листопада 2021 року

УКРАЇНА
ТЕРНОПІЛЬ – 2021

Матеріали X Міжнародної науково-практичної конференції молодих учених та студентів «АКТУАЛЬНІ ЗАДАЧІ СУЧАСНИХ ТЕХНОЛОГІЙ» – Тернопіль 24-25 листопада 2021 року

УДК 004.89

Яцишин В.В. канд. техн. наук, доцент, Шуптарський В.В., Цісарук Д.А.
Тернопільський національний технічний університет імені Івана Пулюя

АЛГОРИТМИ МАШИННОГО НАВЧАННЯ ДЛЯ СЕГМЕНТАЦІЇ КОРИСТУВАЧІВ У МАРКЕТИНГОВИХ КОМП'ЮТЕРНИХ СИСТЕМ

Yatsyshyn V.V. PhD, Assoc. Prof., Shuptarskyi V.V., Tsisaruk D.A.

MACHINE LEARNING ALGORITHMS FOR USER SEGMENTATION IN MARKETING COMPUTER SYSTEMS

Найбільш поширеними і широко використовуваними методами сегментації, які застосовуються при побудові маркетингових систем є групування за ознаками і методи статистичного аналізу.

В основі методу групування лежить принцип послідовного розбиття множини об'єктів на підмножини (групи) з врахуванням найбільш важливих властивостей чи ознак. Одна з таких властивостей об'єкту виділяється як критерій на основі якого можна сформувати деяку систему показників (тип продукції, виробник товару, потенційний споживач конкретного виду товару).

Побудова маркетингових комп'ютерних систем вимагає застосування сучасних методів і засобів, які повинні бути орієнтованими на вирішення задач аналізу поведінки користувачів, формування промоакцій, сегментації товарів, послуг і користувачів. Одними із таких трендових технологій для проведення маркетингових досліджень є методи машинного навчання, зокрема, кластеризація і класифікація користувачів і товарів.

У сфері машинного навчання кластеризація передбачає навчання без вчителя, тобто для цього типу алгоритму існує лише один набір вхідних даних без міток. Тому потрібно розв'язати задачу одержання інформації, не знаючи попередньо, яким буде вихід.

Кластеризація використовується в проектах для компаній, які хочуть виявити спільні властивості у своїх клієнтів, щоб застосувати сегментацію клієнтів, створити карти подорожей клієнтів або знайти групи та сформувати рекомендовані набори товарів чи послуг.

Таким чином, якщо значний відсоток клієнтів мають певні спільні риси (вік, тип сім'ї тощо), компанія може рекомендувати проведення певної кампанії, послуги чи товару. Кластеризація також корисна для одержання загальних уявлень про інформацію якою оперує компанія.

З іншого боку, класифікація належить до алгоритмів навчання з вчителем, тобто наявний контроль за навчанням. Це означає, що вхідні дані мають мітки класів і наперед відомий можливий вихід алгоритму. Розрізняють бінарну класифікацію, яка розв'язує задачі з категорійними відповідями (наприклад, "так" і "ні"), і мультикласифікація, для задач, де потрібно знайти більше двох класів, відповідаючи на більш відкриті відповіді, такі як «чудово », «звичайний» і «недостатній». Класифікація використовується в багатьох галузях, наприклад у біології або в десятковій класифікації Дьюї для книг, при виявленні спаму в електронних листах та ін. Класифікація зазвичай використовується у фінансовому секторі. В епоху онлайн-транзакцій, коли використання готівки помітно зменшилося, необхідно визначити, чи безпечні переміщення за допомогою карток. Суб'єкти можуть класифікувати операції як правильні або шахрайські, використовуючи історичні дані про поведінку клієнтів, щоб дуже точно виявляти шахраїв.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ

М А Т Е Р І А Л И

ІХ НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

8–9 грудня 2021 року

ТЕРНОПІЛЬ
2021

УДК 004.4

А.М. Луцків канд. техн. наук, доцент, Д.А. Цісарук, В.В. Шуптарський
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

АНАЛІЗ ЖИТТЄВОГО ЦИКЛУ ПРОЦЕСУ ТЕСТУВАННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ КОМП'ЮТЕРНИХ СИСТЕМ

UDC 004.4

A.M. Lutskiv PhD, Assoc. Prof., D.A. Tsisaruk, V.V. Shuptarskyi

ANALYSIS OF SOFTWARE TESTING LIFE CYCLE PROCESS IN COMPUTER SYSTEMS

Життєвий цикл тестування програмного забезпечення комп'ютерних систем передбачає виконання різних комплексних заходів та видів діяльності, які формують деяку послідовність. Основна ціль підпроцесів життєвого циклу тестування полягає у забезпеченні та контролі якості прототипу системи та кінцевого програмного продукту. Структуру життєвого циклу процесу тестування програмного забезпечення показано на рис. 1.

Рисунок 1. Життєвий цикл тестування програмного забезпечення комп'ютерних систем

Фаза тестування при реалізації програмного забезпечення відіграє важливу роль у загальному життєвому циклі комп'ютерної системи. Команді тестувальників потрібно спланувати багато заходів і кожна діяльність має бути орієнтована на забезпечення якості.

Характерною і необхідною ознакою кваліфікованої команди тестувальників є володіння інформацією щодо масштабу і повноти виконання подальших кроків щодо виявлення дефектів. Аналогічно, на ранній стадії тестування повинна бути можливість аналізу обсягу тестування продукту. Кожен документ щодо підготовки і проведення різних видів діяльності у процесі тестування повинен бути підготовлений належним чином і однаково трактуватись усіма учасниками проекту.

Якщо всі тестові сценарії охоплені належним чином у відповідності до вимог, то виконання тестового сценарію займатиме менше часу. Це допоможе виявляти помилки на ранній стадії.