Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Методи проєктування та розробки програмних компонентів для системи класифікації клієнтів компанії
(тема)

Виконав:
здобувач 2 року навчання,
групи СПзм-23-1
Дар'я МАЛИШЕНКО
(власне ім'я, прізвище)

Спеціальність
123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма
Системне програмування
(повна назва освітньої програми)

Керівник: доц. Олександр ШМАТКО
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ЕОМ _____ Андрій КОВАЛЕНКО
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Малишенко Дар'ї Олександрівні
(прізвище, ім'я, по батькові)

1. Тема роботи Методи проєктування та розробки програмних компонентів для системи класифікації клієнтів компанії
затверджена наказом по університету від " 07 " квітня 2025 р. № 53 Стз
2. Термін подання здобувачем роботи до екзаменаційної комісії 16 червня 2025 р.
3. Вхідні дані до роботи
1)Нормативні документи щодо класифікації клієнтів компанії;
2) Літературні джерела за темою дослідження;

4. Перелік питань, що потрібно опрацювати у роботі
1) аналіз проблеми та огляд існуючих рішень;
2) огляд методів та підходів до розробки систем класифікації клієнтів компанії
3) розробка опису програмних компонентів;
4) розробка програмних модулів;
5) висновки.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій Слайд-презентація – 19 слайдів

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

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

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Аналіз проблеми та огляд існуючих рішень	22.04.25-29.04.25	
2	Дослідження методів класифікації клієнтів	30.04.25-05.05.25	
3	Вибір технологій та інструментів для розробки	06.05.25-09.05.25	
4	Розробка програмних модулів	10.05.25-21.05.25	
5	Запуск та тестування програмних модулів	22.05.25-02.06.25	
6	Оформлення матеріалів кваліфікаційної роботи	3.06.25-05.06.25	
7	Подання кваліфікаційної роботи керівникові	06.06.25-09.06.25	
	та її попередній захист		
8	Подання кваліфікаційної роботи на рецензу-	10.06.25-12.06.25	
	вання		

Дата видачі завдання " 07 " квітня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ Доц. Олександр ШМАТКО

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 84 с., 15 рис., 12 табл., 3 дод., 26 джерел.

МАШИННЕ НАВЧАННЯ, КЛАСИФІКАЦІЯ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, КЛІЄНТСЬКА БАЗА, ПРОГРАМНІ КОМПОНЕНТИ.

Робота присвячена дослідженню методів проєктування та розробки програмних компонентів для систем класифікації клієнтів компанії з використанням сучасних підходів машинного навчання.

Об’єктом дослідження є методи та системи класифікації клієнтів компанії.

Предметом дослідження є методи та інструментальні засоби розробки програмного забезпечення для побудови ефективних моделей класифікації клієнтів на основі аналізу даних.

Метою роботи є підвищення точності класифікації клієнтів шляхом впровадження алгоритмів машинного навчання у процес проєктування програмних компонентів.

У результаті реалізовано програмний прототип системи класифікації, що забезпечує автоматизований аналіз клієнтських даних і підтримку процесу прийняття управлінських рішень у компанії. Отримані результати свідчать про високу ефективність застосування технологій машинного навчання для обробки великих обсягів інформації, оптимізації бізнес-процесів та підвищення якості взаємодії з клієнтами.

Запропоноване рішення має практичну цінність і може бути використане в різних галузях бізнесу для покращення систем аналітики та управління клієнтською базою.

ABSTRACT

Master’s thesis: 84 pages, 15 figures, 12 tables, 3 appendices, 26 sources.

MACHINE LEARNING, CLASSIFICATION, DATA MINING, CUSTOMER BASE, SOFTWARE COMPONENTS.

This thesis is devoted to the study of methods for designing and developing software components for customer classification systems using modern machine learning approaches.

The object of the research is the methods and systems for customer classification within a company.

The subject of the research is the methods and tools for software development aimed at building effective customer classification models based on data analysis.

The goal of the thesis is to improve the accuracy of customer classification by integrating machine learning algorithms into the design process of software components.

As a result, a software prototype of the classification system has been developed, which enables automated analysis of customer data and supports decision-making processes within a company. The obtained results demonstrate the high efficiency of applying machine learning technologies for processing large volumes of information, optimizing business processes, and enhancing the quality of customer interaction.

The proposed solution has practical value and can be applied across various business domains to improve analytics systems and customer base management.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

AI – штучний інтелект (англ., Artificial Intelligence)

ANN – штучна нейронна мережа (англ., Artificial Neural Network)

API – інтерфейс прикладного програмування (англ., Application Programming Interface)

AUC – площа під ROC-кривою (англ., Area Under Curve)

CPU – центральний процесор (англ., Central Processing Unit)

CRUD – основні операції з базою даних: створення, читання, оновлення, видалення (англ., Create, Read, Update, Delete)

CSV – формат табличних даних із розділювачем-комою (англ., Comma-Separated Values)

DB – база даних (англ., Database)

DL – глибинне навчання (англ., Deep Learning)

DT – дерево рішень (англ., Decision Tree)

EDA – розвідувальний аналіз даних (англ., Exploratory Data Analysis)

GPU – графічний процесор (англ., Graphics Processing Unit)

JSON – формат обміну даними (англ., JavaScript Object Notation)

KNN – метод k-ближчих сусідів (англ., K-Nearest Neighbors)

ML – машинне навчання (англ., Machine Learning)

RF – випадковий ліс (англ., Random Forest)

ROC – характеристика роботи класифікатора (англ., Receiver Operating Characteristic)

SVM – метод опорних векторів (англ., Support Vector Machine)

UI – інтерфейс користувача (англ., User Interface)

ВСТУП

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

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

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

У цьому контексті робота спрямована на розробку та реалізацію програмних компонентів для системи класифікації клієнтів компанії з використанням методів машинного навчання. Основна увага приділяється

аналізу ключових факторів, які впливають на якість сегментації, вибору відповідних алгоритмів, а також впровадженню архітектурних рішень, здатних забезпечити масштабованість, точність і зручність застосування в реальних умовах бізнесу.

Об’єктом дослідження є методи та системи класифікації клієнтів компанії.

Предметом дослідження є методи та інструментальні засоби розробки програмного забезпечення для побудови ефективних моделей класифікації клієнтів на основі аналізу даних.

Метою роботи є підвищення точності класифікації клієнтів шляхом впровадження алгоритмів машинного навчання у процес проєктування програмних компонентів.

Для досягнення поставленої мети необхідно виконати такі основні завдання:

- проаналізувати існуючі підходи до класифікації та сегментації клієнтів у бізнес-середовищі, зокрема методи, що базуються на інтелектуальному аналізі даних та алгоритмах машинного навчання;

- обґрунтувати вибір найбільш доцільних алгоритмів машинного навчання для задач класифікації клієнтів;

- розробити архітектуру програмної системи класифікації клієнтів, визначити основні компоненти та їх функціональну взаємодію;

- реалізувати програмний прототип системи, що здійснює обробку вхідних даних, навчання моделі та автоматичну класифікацію клієнтів за визначеними ознаками;

- провести тестування створеної системи, оцінити точність класифікації за допомогою відповідних метрик (точність, повнота, F1-оцінка, AUC тощо).

У процесі виконання дослідження було застосовано комплекс наукових методів, які забезпечили всебічне вивчення предметної області, обґрунтування вибору технічних рішень та ефективну реалізацію

програмного прототипу, зокрема:

- аналіз та синтез – для вивчення наукових джерел, сучасних підходів до класифікації клієнтів і машинного навчання, а також формування узагальненої концепції побудови системи;
- порівняльний аналіз алгоритмів машинного навчання – для визначення найбільш ефективних моделей класифікації, таких як дерева рішень, випадкові ліси, метод опорних векторів, нейронні мережі тощо;
- методи математичної статистики та теорії ймовірностей – для обробки даних, валідації моделей та оцінювання точності класифікації;
- методи програмної інженерії – для проєктування архітектури системи, моделювання її компонентів та впровадження відповідного програмного забезпечення;
- експериментальне моделювання – для побудови, навчання та тестування моделей класифікації на основі реальних або синтетичних даних;
- візуалізація даних та результатів моделювання – для представлення структури клієнтської бази, класифікаційних рішень і загальної ефективності побудованої системи.

Наукова новизна роботи полягає в розробці підходу до автоматизованої класифікації клієнтів компанії із застосуванням сучасних алгоритмів машинного навчання. У межах дослідження обґрунтовано методику побудови ефективної моделі класифікації на основі аналізу комбінованих характеристик клієнтів. Проведено порівняльне дослідження поширених алгоритмів машинного навчання. Результатом дослідження стала розробка функціонального програмного прототипу, який поєднує алгоритмічні рішення з інтуїтивно зрозумілим користувацьким інтерфейсом, орієнтованим на потреби бізнес-аналітиків та менеджерів.

Практична значимість роботи визначається можливістю безпосереднього використання результатів дослідження для автоматизації процесу класифікації клієнтів у компаніях, що працюють з великими обсягами клієнтських даних. Реалізоване рішення дозволяє суттєво

підвищити ефективність аналізу клієнтської бази, виявляти приховані закономірності у поведінці споживачів та формувати персоналізовані підходи до обслуговування. Використання запропонованої системи сприяє оптимізації маркетингових кампаній, покращенню якості взаємодії з клієнтами та підтримці процесів прийняття управлінських рішень. Розроблена система може бути адаптована до потреб компаній у різних галузях, включно з ритейлом, фінансовими послугами, телекомунікаціями та електронною комерцією.

1 ОГЛЯД МЕТОДІВ КЛАСИФІКАЦІЇ ТА СЕГМЕНТАЦІЇ КЛІЄНТІВ У БІЗНЕС-СЕРЕДОВИЩІ

1.1 Загальні відомості

В умовах високої конкуренції сучасного ринку формування ефективної маркетингової стратегії неможливе без глибокого розуміння поведінки споживачів. Одним із ключових інструментів знаннєорієнтованого маркетингу є сегментація клієнтів — процес поділу великого та різноманітного ринку на менші, однорідні групи на основі спільних характеристик, поведінкових ознак або уподобань. Такий підхід дає змогу бізнесу адаптувати маркетингові кампанії відповідно до специфічних потреб і вподобань кожного сегмента, що в результаті сприяє підвищенню задоволеності клієнтів і зниженню рівня їх відтоку [1].

Традиційно сегментація клієнтів ґрунтувалася на демографічних та географічних даних, зокрема таких як вік, стать, рівень доходу, освіта та місце проживання. Однак ці параметри не завжди повною мірою відображають реальні поведінкові особливості споживачів. Із появою алгоритмів машинного навчання сегментація на основі поведінкових шаблонів набула нового значення, перетворившись на потужний інструмент глибинного аналізу клієнтської бази [2].

Одним із найбільш поширених методів сегментації є алгоритм кластеризації K-means, який належить до неконтрольованого машинного навчання. Цей алгоритм поділяє клієнтів на K кластерів на основі схожості вибраних змінних, таких як історія покупок або активність на вебсайті. Метод K-means широко застосовується в таких галузях, як електронна комерція, банківська справа та телекомунікації, де дозволяє виокремити групи клієнтів із подібними потребами та перевагами. Водночас одним із ключових недоліків цього методу є складність у визначенні оптимальної

кількості кластерів (K), що часто залежить від суб'єктивного досвіду аналітика. Крім того, алгоритм чутливий до викидів у даних, а результати кластеризації можуть змінюватися залежно від початкового розміщення центрів кластерів [3].

З метою усунення цих обмежень були запропоновані модифіковані варіанти алгоритму K-means. Зокрема, у дослідженні [4] було розроблено підхід, який поєднує K-means з методом головних компонент (PCA) для автоматичного визначення оптимальної кількості кластерів. Запропонований алгоритм продемонстрував кращі результати за точністю та стабільністю в порівнянні з класичним K-means.

Ще одним популярним методом сегментації є ієрархічна кластеризація, яка формує групи клієнтів на основі схожості або відмінності між ними з використанням ієрархічної структури. Результати кластеризації представляються у вигляді дендрограми, яку можна обрізати на будь-якому рівні для отримання бажаної кількості кластерів. Ієрархічна кластеризація має низку переваг над методом K-means, зокрема здатність працювати з неповними даними та викидами, а також забезпечення більш інформативної візуалізації структури даних [5]. Водночас цей підхід має і певні недоліки: складність у визначенні кількості кластерів, висока обчислювальна складність при роботі з великими наборами даних, а також залежність результатів від вибору метрик відстані та методів зв'язування.

Для подолання вказаних обмежень дослідники пропонують удосконалення алгоритмів ієрархічної кластеризації. Наприклад, у роботі [6] описано масштабований алгоритм, заснований на стратегії «розділяй і володарюй», який показав високу точність і ефективність при роботі з великими обсягами даних.

У контексті динамічно змінюваних уподобань споживачів аналітика клієнтів стає критично важливою для підтримання високого рівня задоволеності та лояльності. Однією з основних причин відтоку клієнтів є негативний користувацький досвід, що може бути спричинений низькою

якістю обслуговування, нерелевантністю пропозицій або високими цінами. Тому компаніям необхідно постійно моніторити поведінку споживачів і адаптувати свої стратегії до змін у потребах цільової аудиторії [1; 2].

Дана робота спрямована на узагальнення та аналіз сучасних підходів до класифікації клієнтів із використанням алгоритмів машинного навчання. Основна увага приділяється перевагам і недолікам основних методів кластеризації, порівнянню їх ефективності, а також виявленню напрямів подальших досліджень у цій сфері.

1.2 Огляд публікацій за темою дослідження

Сегментація клієнтів є важливою складовою сучасної маркетингової стратегії, що дозволяє підвищити ефективність взаємодії з клієнтською базою за рахунок персоналізованого підходу. Із розвитком цифрових технологій і машинного навчання з’явилися нові можливості для глибшого аналізу клієнтської поведінки, що й обумовило активізацію наукових досліджень у цій сфері.

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

Найбільш поширеним методом кластеризації є K-means, який дозволяє ефективно групувати споживачів на основі подібності поведінкових ознак. Його використання описано в дослідженнях Jain [3], Balaji і Muruganantham [4], Shahriari і Razavi [5], де продемонстровано ефективність цього алгоритму в електронній комерції, банківській справі та сфері обслуговування.

Однак K-means має низку обмежень, зокрема чутливість до вибору початкових центрів кластерів і необхідність заздалегідь визначати кількість

груп. У зв’язку з цим у науковій літературі запропоновано кілька модифікацій. Так, Biswas [6] поєднує K-means із методом головних компонент (PCA), а Karimzadehgan [7] застосовує варіант K-medoids для підвищення стійкості кластеризації до викидів.

Ієрархічна кластеризація є ще одним популярним методом, що дозволяє побудову дерева кластерів із можливістю їх динамічного виділення. Дослідження Chen [8] і Martínez-De-Pisón [9] ілюструють застосування цього методу у телекомунікаціях та роздрібній торгівлі. Luo і Tan [10] представили модифікований варіант ієрархічної кластеризації на основі бінарного розщеплення.

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

Метод опорних векторів (SVM) продемонстрував хороші результати у задачах класифікації клієнтів у сфері e-commerce, як показано у дослідженні Fazlollahtabar [13]. Lee і Kim [14] запропонували гібридний підхід, який об’єднує переваги K-means та ієрархічної кластеризації. Tang [15] застосував fuzzy clustering у поєднанні з глибокими нейронними мережами, що дозволило досягти високої точності.

Zhang і Liu [16] використали autoencoder-мережі для зменшення розмірності даних перед застосуванням DBSCAN кластеризації. У дослідженні Wang [17] розглянуто вплив попередньої обробки даних на точність сегментації. Zhou [18] порівнює результати традиційних і гібридних підходів до кластеризації. Yu [19] розглядає сегментацію в реальному часі з використанням потокових даних.

Li [20] застосовує методи глибинного навчання до аналізу поведінки клієнтів у банківській сфері. Водночас Huang [21] аналізує застосування нейронних мереж у прогнозуванні відтоку клієнтів. Tsai [22] проводить

аналіз гібридних систем для сегментації в страхових компаніях. Xu [23] пропонує адаптивну модель кластеризації, що самостійно визначає кількість сегментів. Kim [24] досліджує інтерпретованість моделей класифікації, що важливо для прийняття рішень у бізнесі. В останніх роботах, зокрема Liu [25], розглядаються перспективи інтеграції класифікації клієнтів із CRM-системами.

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

1.3 Постановка задачі досліджень

У сучасних умовах цифрової трансформації бізнесу компанії стикаються з необхідністю ефективного управління великими обсягами клієнтських даних. Сегментація клієнтів на основі їхніх характеристик, поведінки та вподобань є одним із ключових інструментів для підвищення ефективності маркетингових стратегій, формування персоналізованих пропозицій та підвищення рівня лояльності споживачів. Проте традиційні методи сегментації, засновані на демографічних або географічних критеріях, часто не дозволяють досягти достатнього рівня точності й адаптивності. У цьому контексті застосування алгоритмів машинного навчання відкриває нові можливості для автоматизованої, гнучкої та високоточнішої класифікації клієнтів.

З огляду на результати аналізу наукової літератури та актуальність досліджуваної проблематики, у межах роботи ставиться мета — розробити методичний підхід і програмний засіб для класифікації клієнтів компанії з використанням алгоритмів машинного навчання, що дозволить покращити

якість сегментації та забезпечити підтримку управлінських рішень.

Відповідно до поставленої мети, у межах роботи формулюються такі основні наукові та прикладні задачі:

- дослідити існуючі підходи до сегментації клієнтів у бізнес-середовищі, включаючи класичні та інтелектуальні методи класифікації;

- провести аналіз та порівняння алгоритмів машинного навчання, придатних для задач кластеризації;

- визначити релевантні ознаки для сегментації клієнтів на основі поведінкових, транзакційних, соціально-демографічних та інших даних;

- побудувати архітектуру системи класифікації клієнтів з урахуванням вимог масштабованості, адаптивності та зручності використання;

- реалізувати прототип програмного забезпечення, що здійснює попередню обробку даних, навчання класифікаційної моделі та візуалізацію результатів;

- провести експериментальну перевірку ефективності обраних алгоритмів на тестовій вибірці, використовуючи відповідні метрики оцінювання якості класифікації;

- проаналізувати отримані результати, сформулювати рекомендації щодо практичного застосування системи та напрямів подальших досліджень.

Реалізація зазначених задач дозволить досягти поставленої мети дослідження та зробити внесок у вдосконалення підходів до інтелектуальної обробки клієнтських даних, що має важливе значення для прийняття рішень у сфері маркетингу, обслуговування та управління взаємовідносинами з клієнтами.

2 ОГЛЯД ТА ВИБІР МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ТА СЕГМЕНТАЦІЇ КЛІЄНТІВ У БІЗНЕС-СЕРЕДОВИЩІ

2.1 Огляд методів вирішення задачі класифікації та сегментації клієнтів у бізнес-середовищі

Сегментація клієнтів передбачає поділ клієнтської бази на маркетингові групи, представники яких мають спільні характеристики. Інакше кажучи, це процес групування клієнтів на основі подібних ознак, що дозволяє сформувати найбільш ефективну маркетингову стратегію. Сегментація включає збір інформації про кожного клієнта та подальший аналіз цих даних з метою виявлення різних закономірностей і патернів, на основі яких формуються сегменти.

Найпоширенішими методами збору інформації є особисті інтерв'ю, телефонні опитування, анкетування або дослідження, засновані на вже опублікованих даних, що стосуються відповідних ринкових категорій. До базової інформації, яка використовується для сегментації, належать: дані про оплату та доставку, історія покупок, використані промокоди, спосіб оплати тощо. Окрім цього, деякі компанії також збирають розширену інформацію, таку як причина придбання товару, джерело реклами, яке спонукало до купівлі, вік, стать та інші соціально-демографічні характеристики.

У сфері B2B (business-to-business) маркетингу клієнти сегментуються за низкою факторів, зокрема: галузь діяльності, кількість працівників, історія попередніх покупок, географічне розташування тощо. У B2C (business-to-consumer) маркетингу компанії зазвичай враховують вік, стать, сімейний стан, життєвий етап клієнта (наприклад, неодружений, одружений, розлучений, на пенсії) та місце проживання (сільська, приміська або міська місцевість).

Сегментація клієнтів є доцільною для бізнесу будь-якого розміру чи

галузі. Найпоширенішими типами сегментації є: демографічна, RFM-аналіз (частота, нещодавність та грошова вартість покупок), сегментація високоцінних клієнтів (HVC – High Value Customers), сегментація за статусом клієнта, поведінкова та психографічна.

Основними перевагами використання сегментації є вдосконалення маркетингової стратегії, оптимізація промоційної діяльності, ефективне управління бюджетом, а також розробка нових продуктів. У межах цієї роботи було застосовано базову функціональність аналітики для надання особам, що приймають рішення (у нашому випадку – бізнес-інвесторам), необхідної інформації для прийняття обґрунтованих управлінських рішень. Розроблено підхід, який дозволяє зменшити ризики та підтримати процес ухвалення рішень щодо нових інвестиційних проєктів.

Сегментація клієнтів є важливим напрямом у сфері інтелектуального аналізу даних, що дозволяє виділяти групи споживачів зі схожими характеристиками для реалізації персоналізованих маркетингових стратегій. У цьому розділі розглянуто найбільш поширені методи кластеризації клієнтів, які активно використовуються у сучасних інформаційних системах, орієнтованих на обробку великих даних.

Алгоритм K-means є одним із найпоширеніших методів неконтрольованого машинного навчання, що дозволяє розділити n об'єктів на k кластерів таким чином, щоб кожен об'єкт належав до кластеру з найближчим центром (центроїдом).

Цільова функція, яку мінімізує алгоритм, має вигляд:

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

де: C_i — множина об’єктів, що належать кластеру;

μ_i — центроїд кластеру;

$||x - \mu_i||^2$— евклідова відстань між об’єктом та центроїдом.

Ієрархічна кластеризація створює вкладену ієрархію кластерів. В основі методу лежить обчислення матриці відстаней між усіма парами об’єктів і поступове об’єднання найближчих елементів. Застосовуються різні методи зв’язування: одиничне, повне, середнє або метод Варда.

Для прикладу, у методі Варда обирається така пара кластерів A та B, яка мінімізує збільшення дисперсії в результаті об'єднання:

$$\Delta E = \frac{n_A n_B}{n_A + n_B} ||\mu_A - \mu_B||^2$$

де: n_A, n_B — кількість елементів у кластерах,

μ_A, μ_B — центроїди відповідних кластерів.

Метод DBSCAN (Density-Based Spatial Clustering of Applications with Noise) базується на щільності розміщення точок. Основні поняття — це ε-околи та мінімальна кількість точок у кластері (MinPts). Алгоритм об'єднує точки, які мають достатню щільність у просторі, та відокремлює викиди.

Формально, точка p належить до кластеру, якщо:

$$|N_\varepsilon(p)| \geq MinPts$$

де: $N_\varepsilon(p) = \{q \in D \mid dist(p, q) \leq \varepsilon\}$ — ε-окол точки

$dist(p, q)$ — відстань між точками.

Алгоритм Fuzzy C-means (FCM) дозволяє кожному об’єкту належати до кількох кластерів із певним ступенем приналежності u_{ij}, де i — номер об’єкта, j — номер кластеру.

Цільова функція має вигляд:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m ||x_i - c_j||^2$$

де: $u_{ij} \in [0,1]$— ступінь приналежності об’єкта i до кластеру j,

$m > 1$ — коефіцієнт нечіткості,

c_j — центроїд кластеру j.

Principal Component Analysis (PCA) є методом зменшення розмірності, який дозволяє виявити головні напрямки зміни даних у просторі. Це особливо корисно перед застосуванням кластеризації, оскільки зменшує навантаження на обчислення.

Коваріаційна матриця для даних:

$$C = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X}),$$

де: X — матриця вхідних даних,

$\bar{X}$ — матриця середніх значень.

Власні вектори цієї матриці формують простір головних компонент. Проєкція даних у новий простір дає змогу виконувати кластеризацію за меншою кількістю змінних.

Самоорганізовувані карти Кохонена SOM (Self-Organizing Maps) — це нейронна мережа без учителя, яка створює двовимірну мапу ознак, де просторово близькі об’єкти потрапляють у сусідні зони. SOM дозволяє зберігати топологію даних і є дуже зручним для візуалізації та початкової сегментації.

Вектор ваг w для нейронів оновлюється за формулою:

$$w_j(t+1) = w_j(t) + \alpha(t)h_{bj}(t)(x - w_j(t)),$$

де: x — вхідний вектор,

$\alpha(t)$ — коефіцієнт навчання,

$h_{bj}(t)$ — функція сусідства,

j — індекс нейрона.

Порівняльний аналіз методів наведено в таблиці Б.1 Додатку Б.

2.2 Огляд методів машинного навчання для вирішення задачі класифікації та сегментації клієнтів у бізнес-середовищі

У цьому розділі розглянуто ключові методи машинного навчання, які активно застосовуються для задач класифікації клієнтів. Увагу зосереджено на чотирьох популярних підходах: дерева рішень (Decision Tree, DT), метод опорних векторів (Support Vector Machines, SVM), метод k-ближчих сусідів (K-Nearest Neighbors, KNN) та випадковий ліс (Random Forest, RF).

Дерева рішень (Decision Tree, DT). Метод дерев рішень є одним із найдавніших підходів у машинному навчанні, що активно використовується з 1970-х років. Він реалізується у вигляді графічної структури, яка імітує процес прийняття рішень на основі умов і результатів. Для побудови дерева дані попередньо очищуються та трансформуються. Обираються цільові змінні для сегментації — категоріальні або числові, після чого будується дерево з використанням алгоритму пошуку оптимальних розділень.

Ключовим елементом алгоритму є визначення нечистоти вузлів. Для цього використовуються показники Gini-індексу або ентропії:

Gini-індекс:

$$Gini = 1 - \sum p_i^2 = 1 - (\frac{n_g}{n})^2 - (\frac{n_b}{n})^2$$

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

Ентропія:

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

У проведеному експерименті було протестовано два критерії (Gini та Entropy) з різними параметрами розгалуження, що дозволило досягти точності класифікації 83,24%. Метод є інтерпретованим, не потребує складної підготовки даних та підходить як для числових, так і для категоріальних змінних. Основними недоліками є можливість перенавчання та чутливість до викидів.

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

Рівняння гіперплощини:

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

де $\vec{w}$ — вектор ваг,

b — зміщення,

u — вектор ознак.

Метод показав точність 80,75%. Перевагами є висока ефективність у високорозмірних просторах і стійкість до перенавчання. Основний недолік — висока обчислювальна складність при великих обсягах даних і складність інтерпретації моделей.

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

Формула обчислення відстані у n-вимірному просторі:

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

У проведеному дослідженні із застосуванням 5-кратної перехресної перевірки при різних значеннях k досягнуто точності 79,78%. Перевагою є відсутність етапу навчання — модель будується безпосередньо на основі даних. Основним недоліком є погана масштабованість: при великій кількості об’єктів обчислення відстаней стає дорогим за ресурсами, особливо у високорозмірних просторах.

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

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

У експерименті з різними гіперпараметрами було досягнуто точності 89,61%, що свідчить про високу надійність і точність методу. Основною перевагою є низький ризик перенавчання, однак метод потребує значних обчислювальних ресурсів. Порівняльний аналіз методів наведено в таблиці Б.2 Додатку Б.

2.3 Формування функціональних та нефункціональних вимог до системи класифікації клієнтів у бізнес-середовищі

У рамках цієї роботи розглядається проєктування інформаційної системи, яка реалізує автоматизовану класифікацію клієнтів компанії з

використанням методів машинного навчання. Метою системи є підтримка прийняття управлінських рішень у сфері маркетингу, аналітики та стратегічного планування за рахунок виявлення груп клієнтів зі схожими характеристиками, поведінковими моделями або потенціалом до взаємодії.

Для ефективного проєктування системи необхідно визначити ключових користувачів, які взаємодіятимуть із платформою, описати типові сценарії використання, а також сформулювати функціональні та нефункціональні вимоги до її роботи.

У межах запропонованої системи передбачається три основні категорії користувачів:

- маркетолог — відповідальний за планування й реалізацію рекламних кампаній на основі результатів класифікації;
- бізнес-аналітик — використовує систему для виявлення тенденцій у клієнтській базі, аналізу активності та поведінки;
- системний адміністратор — забезпечує технічну підтримку платформи, моніторинг роботи системи та конфігурацію алгоритмів.

На основі сценаріїв взаємодії визначено перелік функціональних вимог, які мають бути реалізовані в системі. Вони охоплюють процеси імпорту даних, запуску класифікації, аналізу результатів та взаємодії користувача з результатами сегментації.

Таблиця 2.1 – Взаємозв’язок акторів, сценаріїв використання та функціональних вимог

Актор	Сценарій використання	Функціональні вимоги
1	2	3
Маркетолог	Перегляд результатів класифікації	F4 – Візуалізація сегментів
	Формування таргетованих кампаній	F6 – Експорт класифікованих даних

Продовження таблиці 2.1

1	2	3
Бізнес-аналітик	Завантаження набору даних	F1 – Імпорт вхідних даних
	Аналіз активності клієнтів у сегментах	F5 – Виведення метрик класифікації
	Побудова нової моделі сегментації	F2 – Навчання моделі класифікації
Системний адміністратор	Налаштування алгоритмів класифікації	F3 – Вибір параметрів моделі
	Моніторинг продуктивності системи	F7 – Журналізація дій користувачів

Нефункціональні вимоги визначають якісні характеристики системи класифікації клієнтів, які мають бути дотримані для забезпечення надійності, продуктивності та масштабованості рішення. Ці вимоги є критичними для впровадження платформи в умовах реального бізнес-середовища, де обсяги даних можуть постійно зростати, а вимоги до безпеки та швидкодії – посилюватися.

Таблиця 2.2 – Нефункціональні вимоги, критерії вимірювання та цільові значення

№	Нефункціональна вимога	Критерій вимірювання	Цільове значення
1	2	3	4
1	Масштабованість	Кількість клієнтів у базі	Підтримка понад 1 млн записів без зниження продуктивності

Продовження таблиці 2.2

1	2	3	4
2	Продуктивність класифікації	Час виконання класифікації	Не більше 2 хвилин для 100 тис. записів
3	Безпека даних	Рівень доступу користувача	Розмежування доступу за ролями, логування дій
4	Інтеграція	Підтримка стандартів обміну	REST API, CSV, JSON
5	Надійність	Uptime системи	Не менше 99.5% на місяць
6	Візуалізація результатів	Наявність інтерфейсу аналітики	Дашборди, графіки, сегментні таблиці
7	Гнучкість налаштувань	Можливість зміни алгоритму класифікації	Підтримка змін без перезапуску системи

Проєктування ефективної системи класифікації клієнтів вимагає не лише впровадження сучасних алгоритмів машинного навчання, але й чіткого формулювання функціональних та нефункціональних вимог. Функціональні вимоги окреслюють ключову функціональність платформи, тоді як нефункціональні — гарантують якісну роботу системи в умовах великого навантаження, необхідності масштабування та інтеграції в існуючу бізнес-екосистему. Дотримання цих вимог створює передумови для надійної, гнучкої та ефективної роботи системи в умовах реального бізнесу.

3 ПРОЄКТУВАННЯ АРХІТЕКТУРИ ТА ПРОГРАМНИХ КОМПОНЕНТІВ СИСТЕМИ КЛАСИФІКАЦІЇ ТА СЕГМЕНТАЦІЇ КЛІЄНТІВ У БІЗНЕС-СЕРЕДОВИЩІ

3.1 Проєктування архітектури системи

Архітектура системи кластеризації клієнтів у бізнес-середовищі являє собою багаторівневу структурну модель, яка забезпечує послідовну обробку, трансформацію, аналіз та інтерпретацію клієнтських даних з метою отримання релевантної інформації для прийняття управлінських рішень. Побудова такої архітектури ґрунтується на принципах виявлення знань у базах даних (knowledge discovery in databases, KDD), де кожен етап слугує логічним продовженням попереднього та забезпечує підготовку даних для подальшої аналітичної обробки. Узагальнене представлення цієї архітектури наведено на рисунку 3.1, що ілюструє основні функціональні блоки системи та напрямки їх взаємодії.

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

Після очищення дані проходять через фазу трансформації, де здійснюється приведення множинних джерел до уніфікованого формату ознак. На цьому етапі можуть виконуватись операції злиття датасетів,

розрахунку нових змінних, а також відбір релевантних ознак на основі статистичних або евристичних критеріїв. Сформований у такий спосіб вектор ознак подається на вхід сегментаційного модуля системи, де реалізується кластеризація клієнтів.

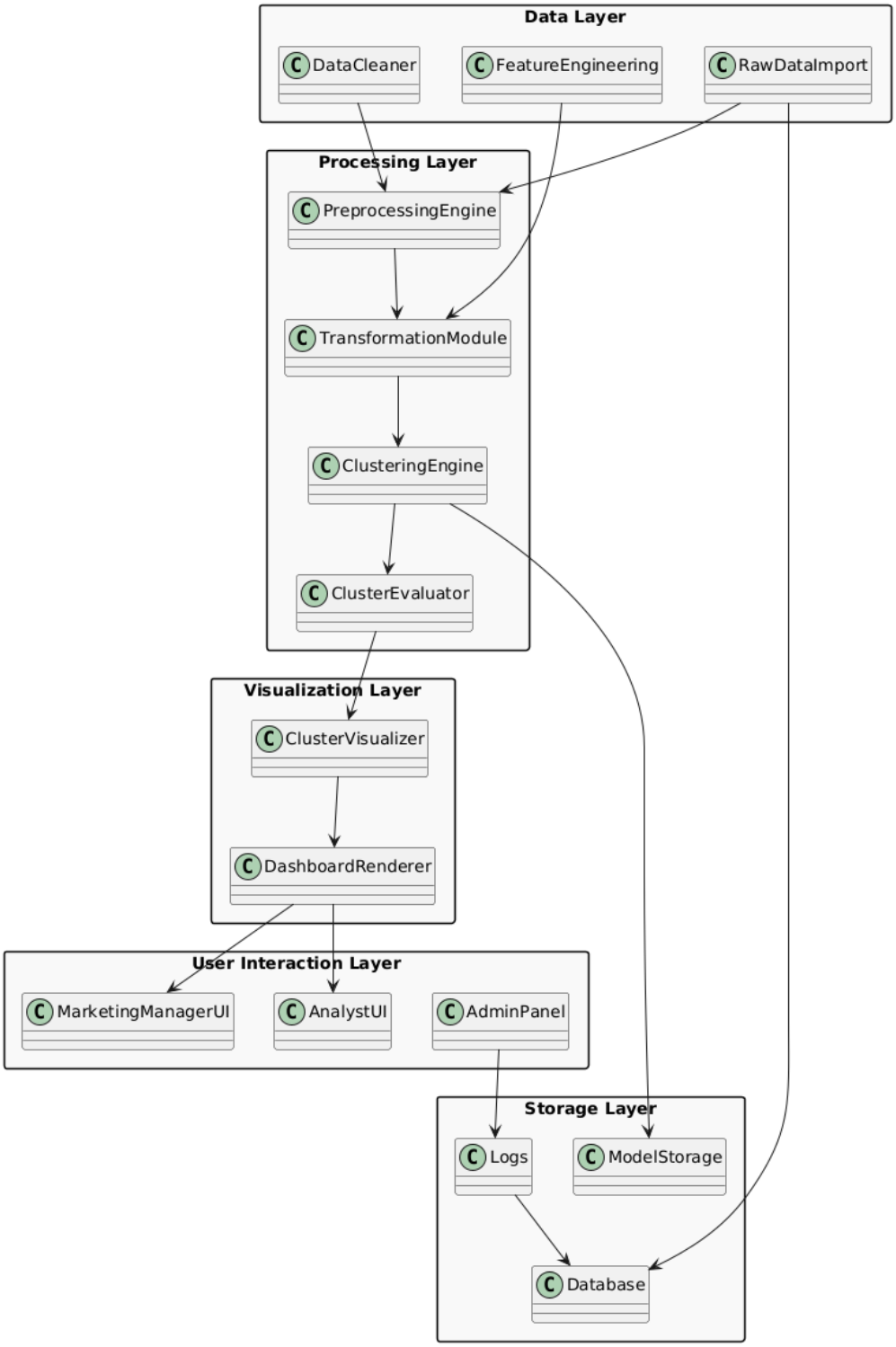

Рисунок 3.1 – Архітектура системи

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

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

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

Фінальним етапом архітектури є інтерпретація результатів та здобуття знань (manual investigation). На цьому рівні бізнес-аналітики, маркетологи або інші фахівці аналізують класифіковані сегменти, формують висновки, генерують гіпотези та приймають управлінські рішення. Отримані знання

можуть бути використані для розробки персоналізованих маркетингових кампаній, прогнозування поведінки клієнтів, покращення клієнтського досвіду або оптимізації роботи з окремими сегментами.

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

3.2 Проєктування програмних компонентів системи

3.2.1 Діаграма варіантів використання

Діаграма варіантів використання відображає основні сценарії взаємодії користувачів із системою класифікації клієнтів за сегментами. У системі визначено три ключові актори: маркетолог, бізнес-аналітик та системний адміністратор, кожен з яких взаємодіє з функціональністю, що відповідає його ролі.

Маркетолог має змогу переглядати результати класифікації клієнтів, що реалізується через функціональну вимогу F4 – Візуалізація сегментів, а також формувати таргетовані маркетингові кампанії на основі класифікованих даних (F6 – Експорт класифікованих даних).

Бізнес-аналітик виконує завантаження початкових наборів даних (F1), проводить аналіз активності клієнтів у різних сегментах з використанням метрик класифікації (F5), а також ініціює побудову нових моделей класифікації клієнтів (F2).

Системний адміністратор відповідає за налаштування параметрів

алгоритмів класифікації (F3) та контроль продуктивності системи, що супроводжується журналізацією дій користувачів (F7).

Таким чином, діаграма демонструє чіткий розподіл відповідальності та взаємозв’язок між сценаріями використання та функціональними вимогами, забезпечуючи повне уявлення про функціональну структуру системи. Діаграма використання системи представлена на рисунку 3.1.

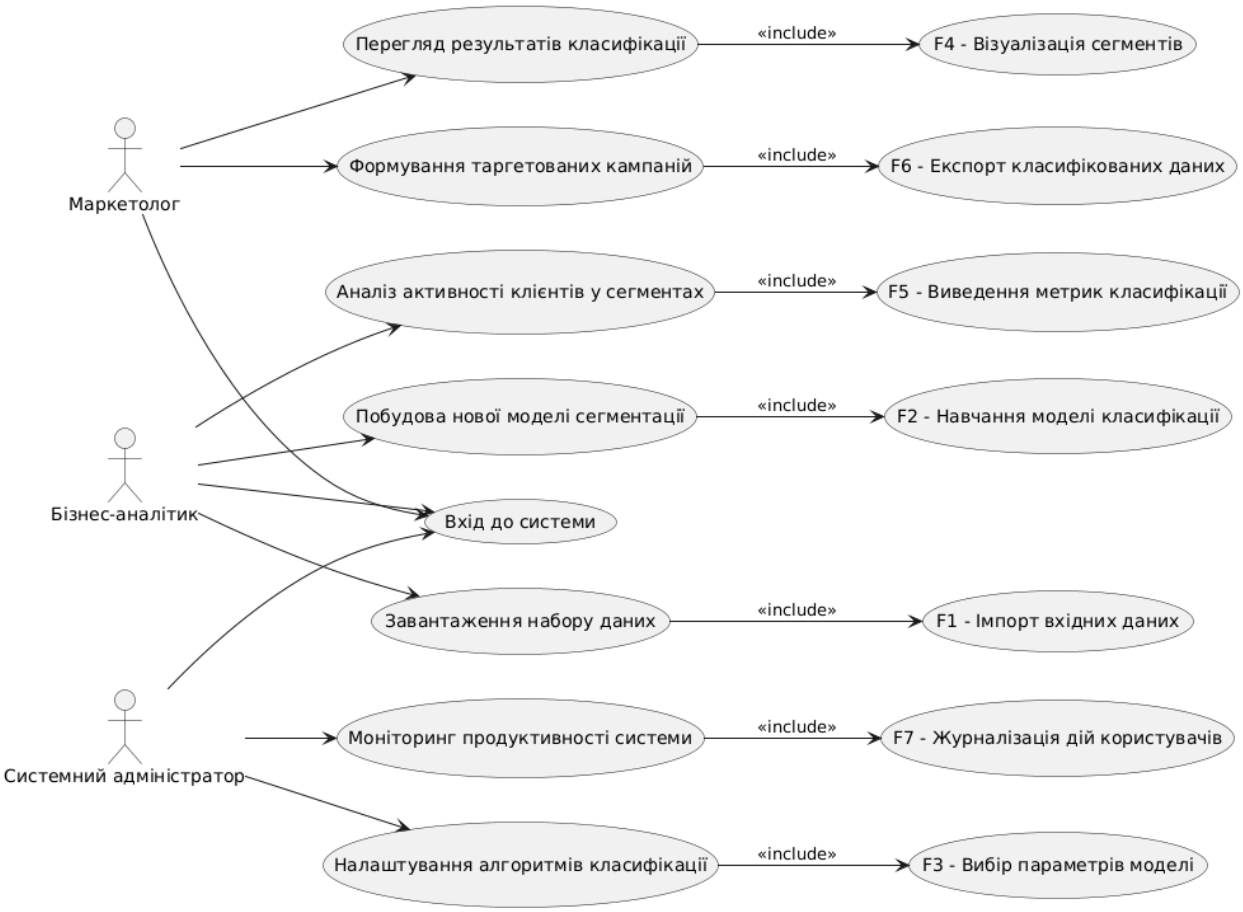

Рисунок 3.2 – Діаграмма варіантів використання

3.2.2 Діаграма компонентів

Діаграма компонентів (рисунок 3.3) описує структуру архітектури системи сегментації клієнтів на рівні її логічних блоків. Центральне місце посідають чотири основні компоненти: модуль попередньої обробки даних, модуль класифікації, механізм сегментації та модуль візуалізації. Всі вони

взаємодіють із інтерфейсом користувача, що забезпечує запуск процесів та перегляд результатів. Компонент зберігання даних (CSV або база даних) виступає джерелом та приймачем даних протягом усього процесу. Така архітектура забезпечує чіткий розподіл обов'язків і дозволяє незалежно вдосконалювати кожен модуль системи.

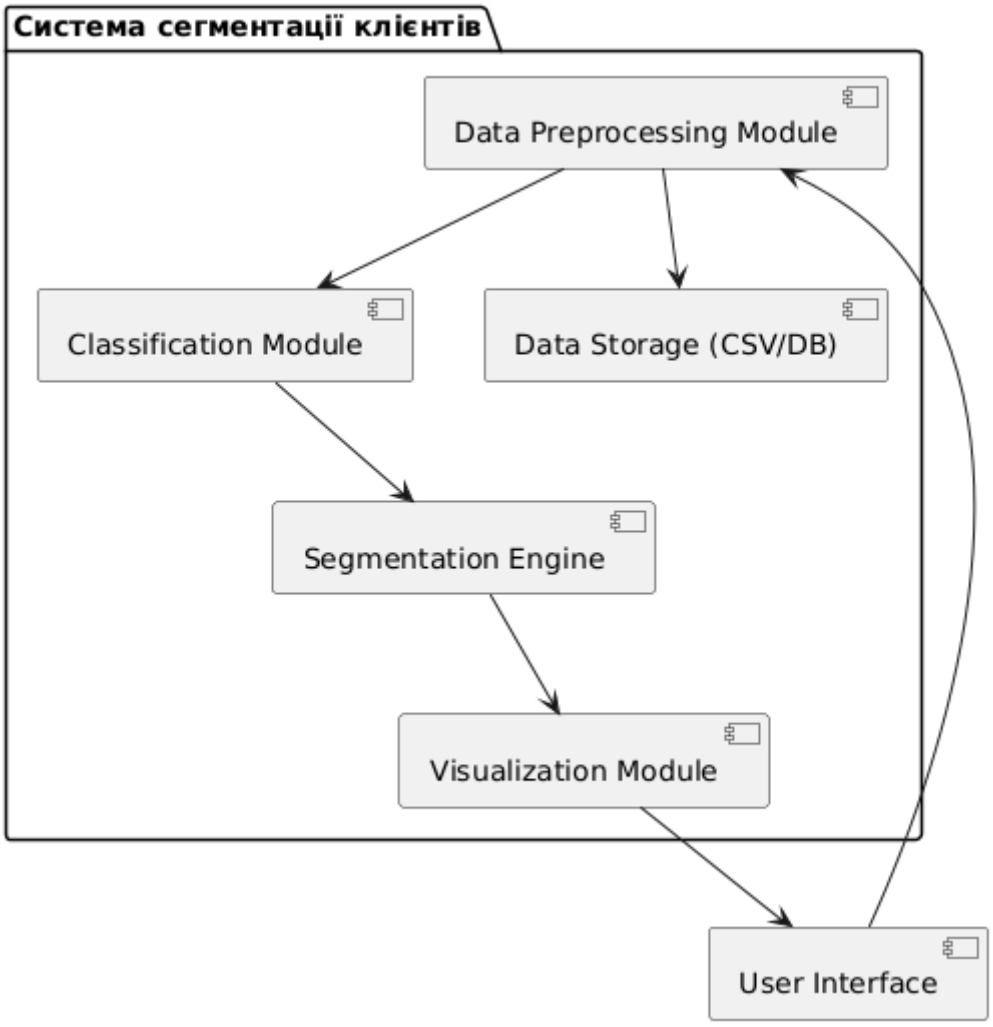

Рисунок 3.3 – Діаграма компонентів

3.2.3 Діаграма послідовності

Діаграма послідовності (рисунок 3.4) ілюструє взаємодію між компонентами системи сегментації клієнтів під час виконання повного сценарію роботи — від завантаження даних до перегляду результатів.

Процес починається з ініціативи користувача, який через веб-інтерфейс завантажує початковий набір даних. Ці дані передаються на сервер додатків, де зберігаються у відповідному сховищі. Далі користувач ініціює запуск процесу сегментації. Інтерфейс передає команду серверу, який витягує дані зі сховища, виконує попередню обробку, класифікацію та сегментацію клієнтів. Результати класифікації повертаються в сховище.

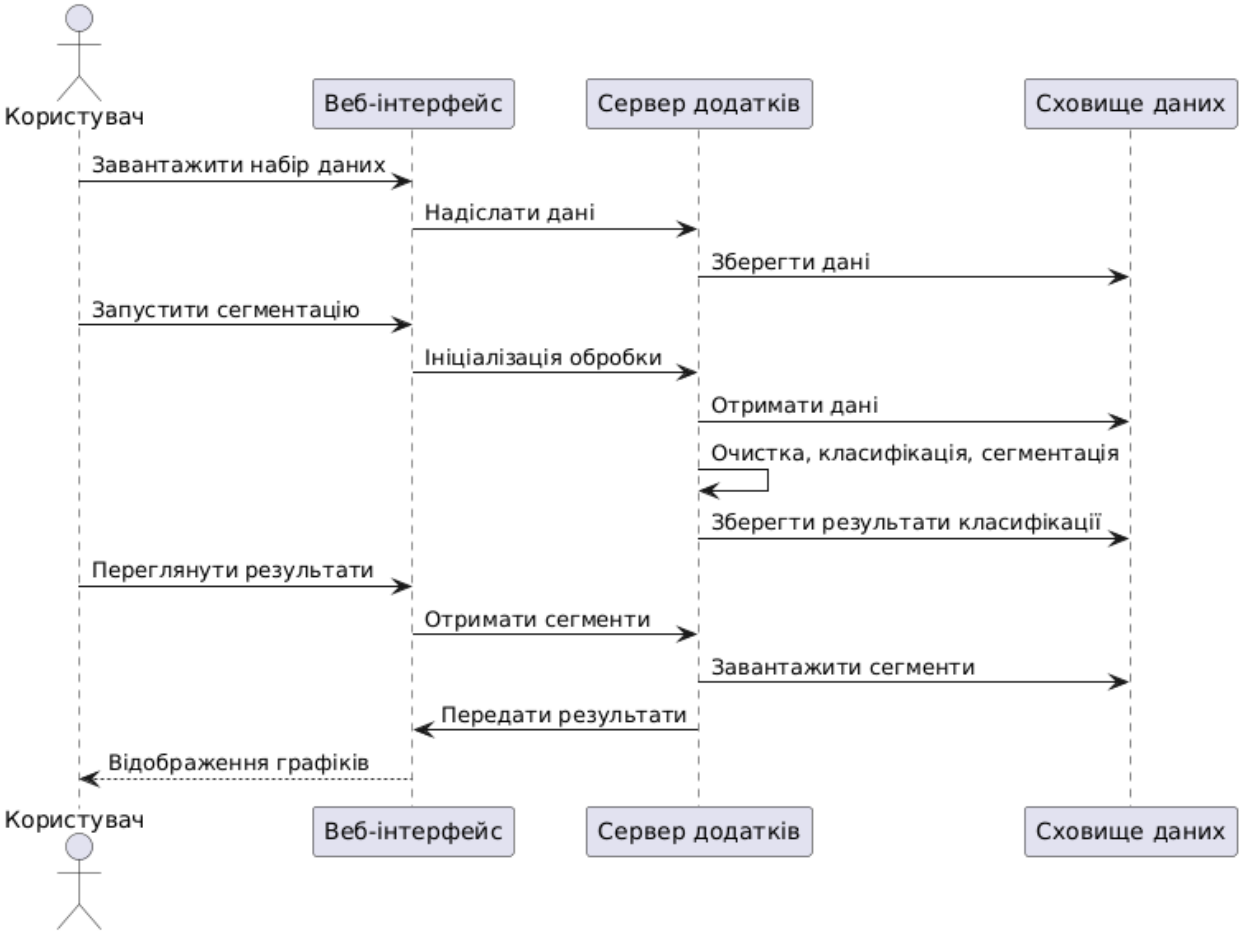

Рисунок 3.4 – Діаграма послідовності

Після завершення обчислень користувач звертається до інтерфейсу з метою перегляду результатів. Інтерфейс ініціює запит до серверу додатків, який у свою чергу завантажує сегментовані дані зі сховища і повертає їх у вигляді графіків та аналітичної візуалізації.

Таким чином, діаграма послідовності демонструє чіткий потік команд

та даних між користувачем, інтерфейсом, прикладною логікою та сховищем, забезпечуючи послідовність операцій і узгодженість результатів на кожному етапі.

3.2.4 Діаграма розгортання

Діаграма розгортання (рисунок 3.5) відображає фізичну структуру системи сегментації клієнтів, яка функціонує як розподілене програмне середовище з чітким розмежуванням відповідальності між вузлами. У даній архітектурі клієнтська частина розгортається на користувацькому комп’ютері, де виконується веб-інтерфейс, що забезпечує інтерактивну взаємодію з користувачем. Саме через цей інтерфейс ініціюються всі ключові процеси: завантаження даних, запуск класифікації, перегляд результатів сегментації та візуалізація даних.

На сервері додатків розміщено функціональні модулі, що реалізують логіку системи. Тут здійснюється попередня обробка даних, яка включає очищення та приведення до єдиного формату, після чого дані передаються до модуля класифікації, де застосовується алгоритм кластеризації, зокрема k-середніх. Результати класифікації передаються до модуля сегментації, який формує групи клієнтів на основі подібних характеристик. Далі активується модуль візуалізації, який генерує графічне представлення отриманих сегментів, що відображається через веб-інтерфейс користувача.

Зберігання вхідних і оброблених даних забезпечується окремим сервером, де функціонує файлове або реляційне сховище. Сховище використовується всіма модулями додатків для читання, запису й оновлення даних протягом усього життєвого циклу сегментації.

Рисунок 3.5 – Діаграма розгортання

Таким чином, архітектура системи поєднує інтерактивність клієнтської частини з потужною серверною обробкою та надійним зберіганням даних, що забезпечує її продуктивність, масштабованість і ефективну взаємодію з користувачем.

4 ДОСЛІДЖЕННЯ ПРОТОТИПУ СИСТЕМИ КЛАСИФІКАЦІЇ ТА СЕГМЕНТАЦІЇ КЛІЄНТІВ У БІЗНЕС-СЕРЕДОВИЩІ

4.1 Експериментальна платформа

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

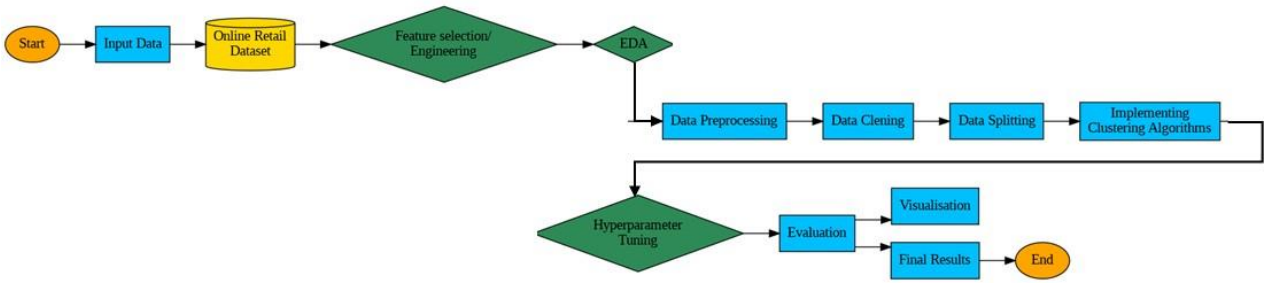

Рисунок 4.1 – Архітектура експериментального стенду системи кластеризації клієнтів

Архітектура експериментального середовища представлена на рисунку 4.1. Вона охоплює послідовність етапів, необхідних для обробки вхідних даних, навчання моделей кластеризації, їхньої валідації та подальшої інтерпретації. Початковим етапом є завантаження вхідного набору даних — у цьому випадку використано набір онлайн-рітейл-даних, який містить інформацію про клієнтські транзакції. Вхідні дані подаються у систему за допомогою модуля pandas, що використовується для статистичного аналізу

та обробки структурованих таблиць.

Після завантаження даних виконується первинне дослідження даних (EDA — Exploratory Data Analysis), у процесі якого аналізуються розподіли, виявляються пропущені значення, аномалії та кореляційні зв’язки. Для цього використовуються бібліотеки pandas, numpy, matplotlib та seaborn, які забезпечують статистичну обробку та графічне представлення розподілів. На основі отриманих результатів виконується відбір ознак і побудова нових змінних (feature engineering), що сприяє покращенню якості класифікації клієнтів.

Далі дані проходять етап попередньої обробки, який включає масштабування числових ознак (MinMaxScaler, StandardScaler), кодування категоріальних змінних (LabelEncoder), а також очищення та нормалізацію значень. Після очищення відбувається розділення даних на навчальну та тестову вибірки, що дозволяє здійснити незалежну перевірку якості кластеризації. Модулі train_test_split із sklearn дозволяють провести це розділення із заданою часткою валідаційної вибірки.

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

Оцінювання точності та стабільності моделей здійснюється на основі порівняльного аналізу результатів кластеризації. Результати класифікації відображаються у вигляді дво- та тривимірних графіків, побудованих за допомогою matplotlib та seaborn, що дозволяє візуалізувати просторове розташування кластерів і зрозуміти логіку розподілу клієнтів за сегментами.

На завершальному етапі формуються підсумкові результати, які

можуть бути інтегровані у бізнес-аналітику для створення персоналізованих пропозицій або адаптації маркетингових стратегій. Отримані знання можуть бути збережені у базу даних або передані до зовнішніх інформаційних систем.

У такий спосіб створений експериментальний стенд відображає повний цикл класифікації клієнтів — від імпорту даних і підготовки вибірки до кластеризації, оцінювання результатів і побудови висновків, що є базою для прийняття рішень у бізнес-середовищі. Застосування широкого спектра інструментів Python уможливлює гнучке налаштування системи та її масштабування під різні задачі класифікації.

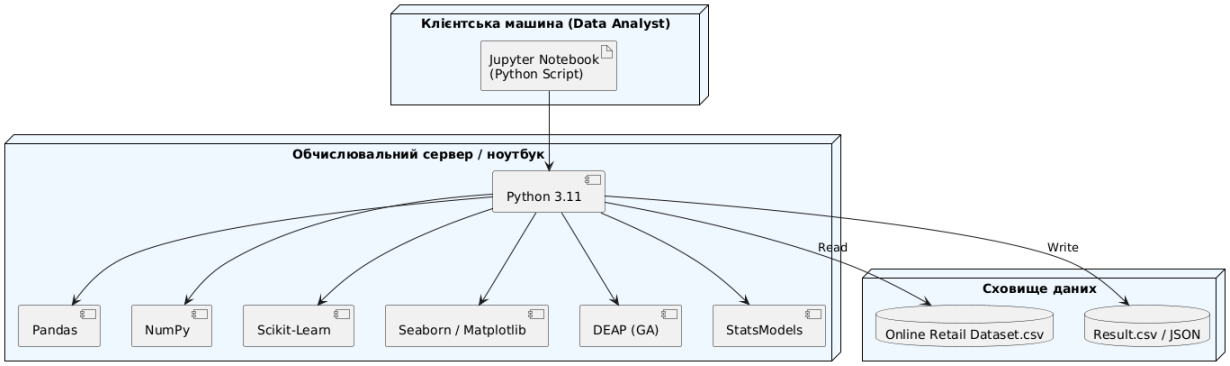

Рисунок 4.2 – Діаграма розгортання експериментальної платформи

З метою реалізації та тестування алгоритмів кластеризації клієнтів у межах експериментального дослідження було розгорнуто програмне середовище на основі мови програмування Python, що функціонує в рамках клієнт-серверної архітектури з доступом до локального або віддаленого сховища даних. На рисунку 4.2 представлено діаграму розгортання, яка ілюструє взаємозв’язок між основними програмними компонентами системи, їх розміщенням на апаратних або віртуалізованих вузлах, а також логіку обміну даними між ними.

Основним компонентом на клієнтському боці виступає Jupyter Notebook, який використовується як інтерактивне середовище для розробки та запуску Python-скриптів. У даному середовищі реалізовано основну логіку

обробки даних, побудови моделей кластеризації, а також візуалізації результатів. Jupyter Notebook функціонує на клієнтській машині, яка може бути як локальним комп’ютером аналітика, так і віддаленим сервером у хмарному середовищі (наприклад, Google Colab або JupyterHub).

Обробка даних здійснюється за допомогою інтерпретатора Python 3.11, у межах якого підключено низку спеціалізованих бібліотек. Бібліотека Pandas забезпечує завантаження та попередню обробку табличних даних. NumPy використовується для виконання числових обчислень та маніпуляцій з масивами. Scikit-Learn виконує функції реалізації алгоритмів кластеризації (K-means, DBSCAN, Gaussian Mixture) та оцінювання їхньої ефективності. Бібліотеки Seaborn та Matplotlib відповідають за побудову графіків і візуалізацію розподілу кластерів. Додатково використовується DEAP — бібліотека для реалізації еволюційних алгоритмів, зокрема для гіперпараметричної оптимізації кластеризаційних моделей, а також StatsModels для поглибленої статистичної аналітики.

Система взаємодіє з локальним сховищем даних, у якому розміщено вхідний набір — Online Retail Dataset, що використовується як джерело інформації про транзакційну поведінку клієнтів. У процесі виконання алгоритмів також створюються файли з результатами класифікації у форматах CSV або JSON, які зберігаються у відповідному каталозі для подальшого використання в аналітичній звітності.

Узаємозв’язок між компонентами представлено у вигляді спрямованих зв’язків, які позначають потік даних між модулями системи. Наприклад, Jupyter Notebook ініціює завантаження даних з джерела, здійснює виклик функцій бібліотек для аналізу, та передає результати візуалізації або збереження у файл. Кожен компонент системи виконує чітко визначену функцію у межах загальної архітектури, що забезпечує модульність, масштабованість та гнучкість експериментального середовища.

4.2 Експериментальні данні

У рамках даного дослідження було використано відкритий набір даних зі сфери роздрібної торгівлі, який отримано з UCI Machine Learning Repository — одного з найавторитетніших джерел у сфері машинного навчання. Конкретно було використано Online Retail Dataset, який знаходиться у вільному доступі за посиланням: https://archive.ics.uci.edu/dataset/352/online+retail. Вибір цього набору даних зумовлений його обсягом, структурованістю, а також широким застосуванням у наукових публікаціях, присвячених задачам кластеризації клієнтів та поведінкової аналітики у сфері електронної комерції.

Набір містить 541 909 записів, що охоплюють інформацію про транзакції, здійснені в період з грудня 2010 по грудень 2011 року однією з онлайн-компаній, що базується у Великобританії та займається продажем подарункових товарів. Кожен запис у наборі відповідає окремій товарній позиції у певному замовленні. Таким чином, набір відображає не лише покупки, але й поведінкові патерни клієнтів, що є критично важливими для задач класифікації та побудови клієнтських сегментів.

Набір даних складається з восьми основних змінних, кожна з яких має конкретну роль у формуванні ознак для подальшого машинного навчання:

- InvoiceNo — унікальний номер рахунку-фактури. Це номінативна змінна, яка представлена у вигляді шестизначного цілого числа. У випадках, коли номер починається з літери «C», мається на увазі скасування транзакції;
- StockCode — код товарної позиції. Також номінативна змінна у форматі п’ятизначного цілого числа, що ідентифікує кожний унікальний товар у базі;
- Description — назва товару. Номінативна текстова змінна, що містить словесний опис кожної товарної одиниці;
- Quantity — кількість одиниць товару, придбаних у межах однієї транзакції. Це числова змінна, яка може використовуватись для розрахунку інтенсивності покупок або обсягу витрат;

- InvoiceDate — дата та час здійснення замовлення. Змінна представлена у числовому форматі та дозволяє визначати як часові патерни купівельної активності (день, година), так і сезонні тенденції;

- UnitPrice — ціна одиниці товару у фунтах стерлінгів. Ця числова змінна дозволяє обчислювати загальну суму витрат на рівні позиції або замовлення;

- CustomerID — унікальний ідентифікатор клієнта. Це п’ятизначне номінативне число, яке забезпечує можливість відслідковувати поведінку окремих клієнтів протягом усього періоду дослідження;

- Country — країна проживання клієнта. Це номінативна змінна, яка дозволяє аналізувати географічну структуру клієнтської бази та сегментувати клієнтів за регіональною ознакою.

Загалом структура датасету (Таблиця 4.1) дозволяє провести як транзакційний аналіз, так і побудову агрегованих ознак на рівні клієнта. Саме такий підхід використано у даному дослідженні: на основі первинних даних було сформовано агреговані метрики за ключовими параметрами (частота покупок, середній чек, кількість повернень, середня кількість товарів на замовлення тощо), які стали основою для подальшої кластеризації клієнтів.

Таблиця 4.1 – Характеристики змінних датасету Online Retail

№	Назва змінної	Тип даних	Формат / Тип значень	Опис
1	2	3	4	5
1	InvoiceNo	Номінативна	Ціле число (6 цифр), текст	Унікальний номер рахунку-фактури. Початок з "C" означає скасування.

Продовження таблиці 4.1

1	2	3	4	5
2	StockCode	Номінативна	Ціле число (5 цифр), текст	Ідентифікатор товарної позиції.
3	Description	Номінативна	Текстовий рядок	Назва або опис товару.
4	Quantity	Кількісна	Ціле число	Кількість придбаних одиниць товару.
5	InvoiceDate	Часова	Дата і час (YYYY-MM-DD hh:mm:ss)	Час створення рахунку-фактури.
6	UnitPrice	Кількісна	Дійсне число (float)	Ціна за одиницю товару (у фунтах стерлінгів).
7	CustomerID	Номінативна	Ціле число (5 цифр), текст	Унікальний ідентифікатор клієнта.
8	Country	Номінативна	Текстовий рядок	Назва країни проживання клієнта.

Приклад даних наведено в таблиці 4.2.

Таблиця 4.2 - Приклади записів із датасету Online Retail

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	2	3	4	5	6	7	8
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
536366	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:28:00	2.75	13047	United Kingdom
536367	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:34:00	3.39	12583	France

Попередня обробка даних є одним із ключових етапів у процесі аналізу даних, що безпосередньо впливає на якість кластеризації та достовірність отриманих результатів. Метою цього етапу є приведення вихідного датасету до уніфікованого, повного та очищеного стану, який уможливлює коректну роботу алгоритмів машинного навчання без викривлень, спричинених шумами, відсутніми значеннями чи дублюваннями.

На початковому етапі обробки було виявлено наявність пропущених значень у двох критично важливих стовпцях — Description (опис товару) та CustomerID (ідентифікатор клієнта). Враховуючи, що відсутність ідентифікатора клієнта унеможливлює зв’язування транзакцій з конкретним користувачем, усі записи з порожнім значенням у полі CustomerID були видалені. У результаті цього кроку розмір вибірки зменшився з початкових 541 909 рядків до 406 829, зберігши при цьому усі вісім початкових ознак (InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country).

Наступним кроком було виявлення і видалення аномальних значень. Було встановлено, що 8905 записів містять від’ємні значення у полі Quantity, що є неприйнятним з точки зору бізнес-логіки (кількість товару не може бути негативною). Ці записи здебільшого відповідали скасованим транзакціям або поверненням і були виключені зі структурованого набору, що використовувався для подальшого аналізу.

Додатково проведено перевірку на дублікати — записи, що повністю збігаються за всіма ознаками (включаючи дату, кількість, назву товару, ідентифікатор клієнта та інші поля). Такі рядки не несуть нової інформації і можуть негативно впливати на баланс кластерів, тому були повністю видалені з датасету.

Після завершення вищезазначених етапів попередньої обробки, остаточний датасет містив 397 924 унікальні записи, збережених у структурі з вісьмома ознаками. Таким чином, було досягнуто повного очищення від шумів, пропущених значень і дублювань, що дозволяє перейти до етапу

побудови ознак і кластеризації.

Для ілюстрації структури очищених даних у таблиці нижче представлено приклади транзакцій, згрупованих за номером рахунку-фактури. Для кожного рахунку зазначено кількість товарних позицій та загальну вартість замовлення.

Таблиця 4.3 – Приклади агрегованих значень за номером рахунку

Номер рахунку-фактури	Кількість товарних позицій	Сумарна вартість (GBP)
1	2	3
536365	7	139.12
536366	2	22.20
536367	12	278.73
536368	4	70.05
536369	1	17.85

Ці приклади демонструють узагальнені показники на рівні транзакцій, які можуть бути використані як додаткові ознаки у процесі побудови класифікаційної моделі або у рамках RFM-аналізу. Попередня обробка забезпечила якісну основу для формування ознак, нормалізації, масштабування та подальшого застосування алгоритмів машинного навчання з метою сегментації клієнтів компанії.

4.3 Метрики оцінювання

Оцінювання якості кластеризації є критично важливим етапом у процесі сегментації клієнтів, оскільки дозволяє кількісно визначити, наскільки добре сформовані кластери відображають реальні закономірності у даних. У межах цього дослідження для валідації кластеризаційних моделей було використано три основні метрики: Silhouette Score, індекс Калінські-

Харабаза (Calinski-Harabasz Index) та індекс Девіса-Боулдіна (Davies-Bouldin Index). Кожна з них забезпечує специфічну оцінку когезійності (внутрішньої узгодженості) кластерів та ступеня їх відокремлення один від одного в багатовимірному просторі ознак.

Silhouette Score вимірює ступінь схожості об'єкта до свого кластеру порівняно зі схожістю до інших кластерів. Значення цієї метрики варіюється від -1 до 1, де значення, близькі до 1, свідчать про чітке відокремлення об’єкта у межах свого кластеру, а негативні значення — про можливе неправильне кластерне віднесення.

Для кожного об'єкта i розраховуються два значення:

- $a(i)$ — середня відстань від точки до всіх інших точок свого кластеру (міра згуртованості);

- $b(i)$ — найменша середня відстань до всіх точок іншого (найближчого) кластера (міра відокремлення).

Тоді:

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

Якщо $s(i)) \approx 1$, то точка добре згрупована у свій кластер.

Якщо $s(i) \approx 0$, то точка знаходиться на межі між двома кластерами.

Якщо $s(i)) < 0$, то точка, ймовірно, помилково класифікована.

У межах цього дослідження Silhouette Score використовувався як основний критерій для вибору оптимальної кількості кластерів при застосуванні алгоритму K-Means.

Індекс Калінські-Харабаза (Calinski-Harabasz Index), відомий також як індекс дисперсії, обчислює співвідношення між міжкластерною дисперсією та внутрішньокластерною дисперсією. Вищі значення метрики свідчать про краще групування, де кластери є водночас компактними і добре розділеними. Цей індекс особливо корисний для моделей, які формують кластери з чіткою геометричною структурою, що спостерігається, зокрема, при використанні

PCA або T-SNE.

Індекс Девіса-Боулдіна (Davies-Bouldin Index) обчислює середнє значення подібності між кожним кластером та найсхожішим до нього сусіднім кластером, де подібність визначається як сума внутрішніх дисперсій, поділена на відстань між центроїдами кластерів. На відміну від попередніх метрик, у цьому випадку нижчі значення індексу свідчать про кращу якість кластеризації. Його застосування дозволило уникнути ситуацій, коли кластери були надто близькими один до одного або мали розмиті межі.

Для підтримки числових метрик і кращого розуміння геометрії кластерів було використано інструменти візуалізації високовимірних даних, зокрема PCA (Principal Component Analysis) та T-SNE (t-distributed Stochastic Neighbor Embedding). Обидва методи дозволяють проєктувати багатовимірні дані у двовимірний простір, зберігаючи при цьому відносні відстані між об’єктами. PCA забезпечив лінійне зниження розмірності, корисне для формального аналізу структури вибірки, тоді як T-SNE дав змогу візуалізувати локальні структури та чіткі межі між кластерами, що особливо корисно у випадках перекриття кластерів.

Після остаточного визначення оптимальної кластерної структури було виконано профілювання клієнтів, що входять до кожного з кластерів. Для цього проводився аналіз середніх значень ключових характеристик, таких як загальні витрати клієнта, середній чек, частота покупок, географічне розташування, а також специфічні патерни транзакційної поведінки. На основі виявлених закономірностей кожен сегмент було описано за допомогою інтуїтивно зрозумілих назв, наприклад, «Цінні постійні покупці», «Клієнти з високим потенціалом», або «Рідкісні відвідувачі». Такі профілі дозволяють не лише зрозуміти характер кожної групи, але й використовувати ці знання для формування персоналізованих маркетингових стратегій або оптимізації цінової політики.

У межах дослідження було приділено особливу увагу дотриманню етичних стандартів обробки персональних даних. Усі записи, що містили

клієнтські ідентифікатори, були анонімізовані, а дані — оброблені відповідно до вимог GDPR (Загального регламенту захисту даних ЄС) та CCPA (Каліфорнійського закону про захист приватності споживачів). Забезпечено прозорість і відповідальність на кожному етапі обробки даних, що є критично важливим при роботі з потенційно чутливою фінансовою інформацією. Таким чином, поєднання формальних метрик, візуальних інструментів та відповідального ставлення до конфіденційності дозволило отримати надійні й практично корисні результати кластеризації.

4.4 Аналіз результатів дослідження

У цьому розділі представлено результати аналізу клієнтів на основі моделі RFM (Recency, Frequency, Monetary), а також результати порівняння ефективності алгоритмів кластеризації, використаних у дослідженні. Модель RFM дозволяє оцінити цінність клієнтів, ґрунтуючись на трьох ключових метриках: давності останньої покупки (Recency), частоті покупок (Frequency) та загальній грошовій вартості покупок (Monetary). Такий підхід широко використовується у маркетинговій аналітиці для сегментації клієнтської бази, пріоритезації комунікацій та формування персоналізованих стратегій взаємодії.

На першому етапі для кожного клієнта було обчислено значення RFM-показників. Зокрема, Recency визначалась як кількість днів з моменту останньої покупки до дати завершення спостереження, Frequency – як кількість унікальних транзакцій, здійснених клієнтом, а Monetary – як загальна сума витрат за всі замовлення. Таблиця 4.4 демонструє приклади перших п’яти клієнтів із відповідними значеннями RFM-компонентів.

Таблиця 4.4 – Значення компонентів RFM для окремих клієнтів

Customer ID	Recency	Frequency	Monetary (£)
1	2	3	4
12346	325	1	77,183.60
12347	1	182	4,310.00
12348	74	31	1,797.24
12349	18	73	1,757.55
12350	309	17	334.40

Для подальшого порівняння між клієнтами всі значення були нормалізовані, а потім трансформовані в оцінки (скори), які відображають відносну позицію клієнта у межах усієї бази. Кожна метрика R, F і M ранжувалась, а потім масштабувалась за формулою:

$$\text{Оцінка} = (\frac{\text{Поточне значення}}{\text{Максимальне значення}}) \times 100$$

Після обчислення індивідуальних R, F та M скорів, було проведено розрахунок загального RFM-скорингу за формулою:

$$RFM\ Score\ = R_{score}\ \times\ w_R\ +\ F_{score}\ \times\ w_F\ +\ M_{score}\ \times\ w_M$$

де w_R, w_F, w_M — ваги, що відповідають важливості кожної з компонент (у даному дослідженні вони були рівноважними або адаптовані згідно з бізнес-цілями).

Таблиця 4.5 демонструє приклади обчислених скорів для окремих клієнтів.

Таблиця 4.5 – RFM скоринг клієнтів

Customer ID	RFM Score
1	2
12346	0.06
12347	4.48
12348	2.09
12349	3.41
12350	1.10

На основі отриманих RFM скорів було реалізовано поділ клієнтів на п’ять основних категорій:

- Top customers – клієнти з RFM - скором вище 4.5;
- High-value customers – більше 4.0;
- Medium-value customers – більше 3.0;
- Low-value customers – більше 1.6;
- Lost customers – нижче або дорівнює 1.6.

Таблиця 4.6 демонструє сегментацію перших десяти клієнтів за результатами скорингу.

Таблиця 4.6 – Сегментація клієнтів на основі RFM

Customer ID	RFM Score	Customer Segment
1	2	3
12346	0.06	Lost customer
12347	4.48	High-value customer
12348	2.09	Low-value customer
12349	3.41	Medium-value customer
12350	1.10	Lost customer
12352	3.46	Medium-value customer

Продовження таблиці 4.6

1	2	3
12353	0.03	Lost customer
12354	2.67	Low-value customer
12355	0.93	Lost customer
12356	3.11	Medium-value customer

Для візуалізації отриманих результатів було побудовано кругову діаграму (рисунок 4.3), яка демонструє розподіл клієнтів між сегментами. Як видно з діаграми, найбільшу частку складають "втрачені клієнти" (31%), за ними слідують "клієнти з низькою цінністю" (30%), "середньоцінні клієнти" (21%), "цінні клієнти" (10%) і лише 8% клієнтів були ідентифіковані як "топові".

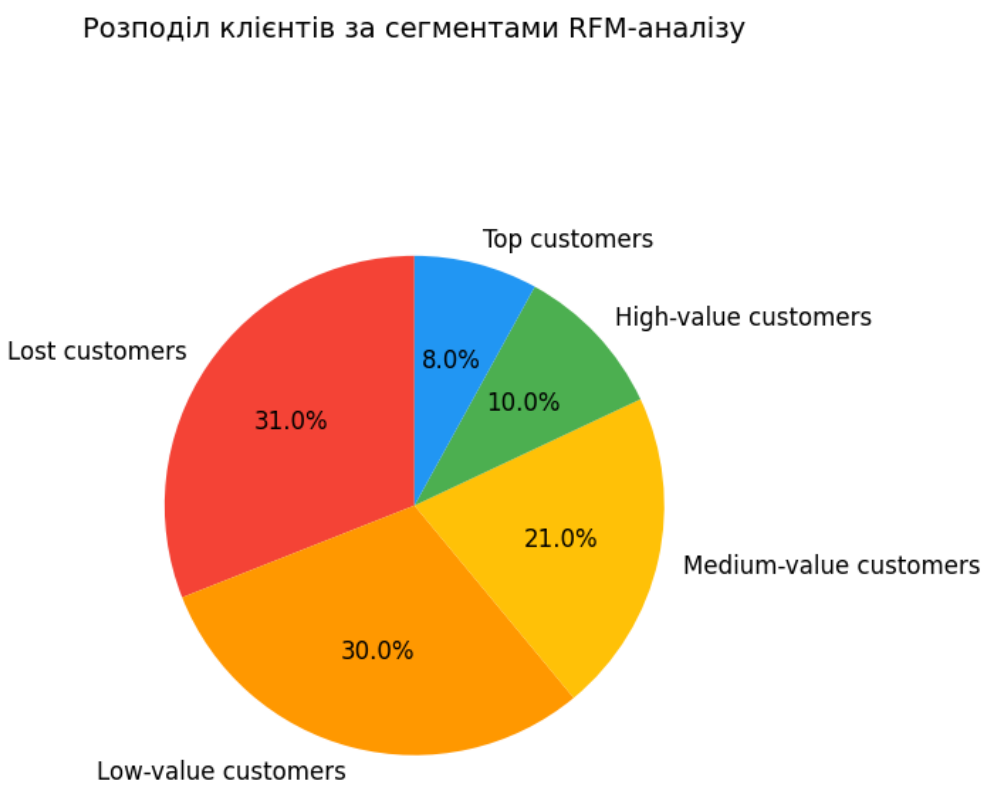

Рисунок 4.3 – Розподіл клієнтів за сегментами RFM-аналізу

Результати RFM-аналізу показали, що лише близько п’ятої частини клієнтів здійснюють регулярні та високовартісні покупки. Це свідчить про потенціал до оптимізації взаємодії з менш активними клієнтами шляхом впровадження персоналізованих кампаній, систем лояльності та рекомендаційних сервісів.

У ході дослідження було застосовано кілька алгоритмів кластеризації без учителя (unsupervised machine learning) з метою виявлення прихованих закономірностей у поведінці клієнтів на основі агрегованих даних. Якість кластеризації оцінювалася з використанням метрики Silhouette Score, яка дозволяє кількісно визначити ступінь згуртованості елементів всередині кластерів та їх розділення між собою. Кожен алгоритм було перевірено шляхом аналізу просторового розподілу об’єктів після кластеризації, візуалізації результатів та порівняння отриманих кластерів за формальними критеріями.

У якості одного з основних алгоритмів кластеризації було застосовано K-Means, який широко використовується завдяки своїй простоті реалізації та ефективності для великих обсягів даних. Для визначення оптимальної кількості кластерів (k) було використано метод "лікоть" (elbow method). Цей підхід передбачає побудову графіка залежності значення інерції (inertia) — тобто суми квадратів відстаней зразків до найближчих центрів кластерів — від кількості кластерів у діапазоні від 2 до 10. На графіку шукається точка з переломом кривої, яка нагадує "лікоть", і яка вважається найкращим компромісом між складністю моделі та якістю кластеризації.

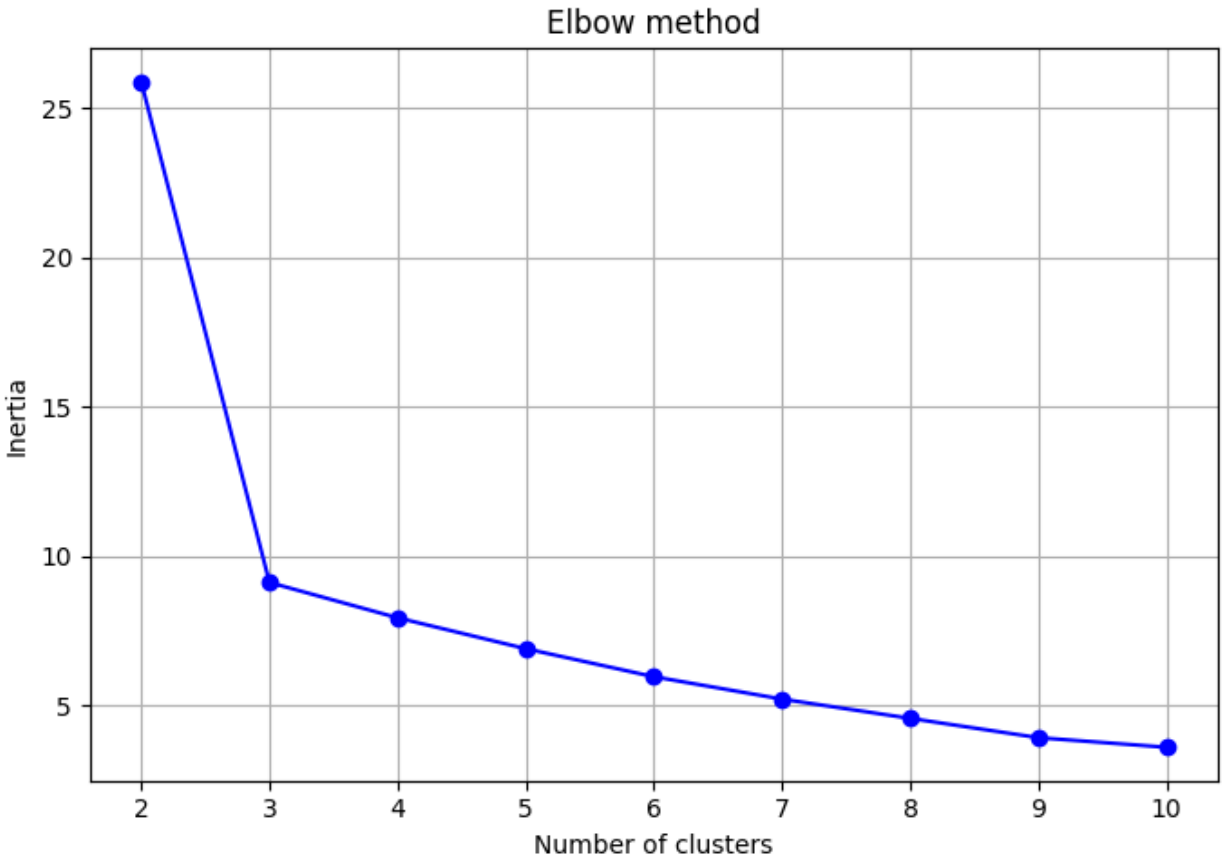

Рисунок 4.4 – Метод Elbow

На основі візуального аналізу графіка (рисунок 4.4) було встановлено, що оптимальна кількість кластерів дорівнює 3, оскільки саме при цьому значенні спостерігається найбільш виражений злам інерції. Після визначення параметра k було проведено масштабування ознак за допомогою Min–Max Normalization, щоб привести всі числові значення до одного діапазону. Це дозволяє уникнути зміщення в результатах кластеризації, яке може виникнути через дисбаланс у масштабах окремих ознак, особливо з огляду на те, що K-Means використовує евклідову відстань як метрику подібності.

Після масштабування було побудовано кластеризаційну модель методом K-Means з трьома кластерами. Результати моделі візуалізовано у вигляді діаграми розсіювання (scatter plot), де кожна точка представляє клієнта, а колір позначає кластер, до якого його віднесено (рисунок 4.5). Центроїди кластерів відображено як центри гравітації, які характеризують типові значення ознак для кожної групи клієнтів. Значення інерції вказує на

компактність кластерів, а візуальний аналіз дозволяє оцінити рівномірність розподілу об’єктів між ними.

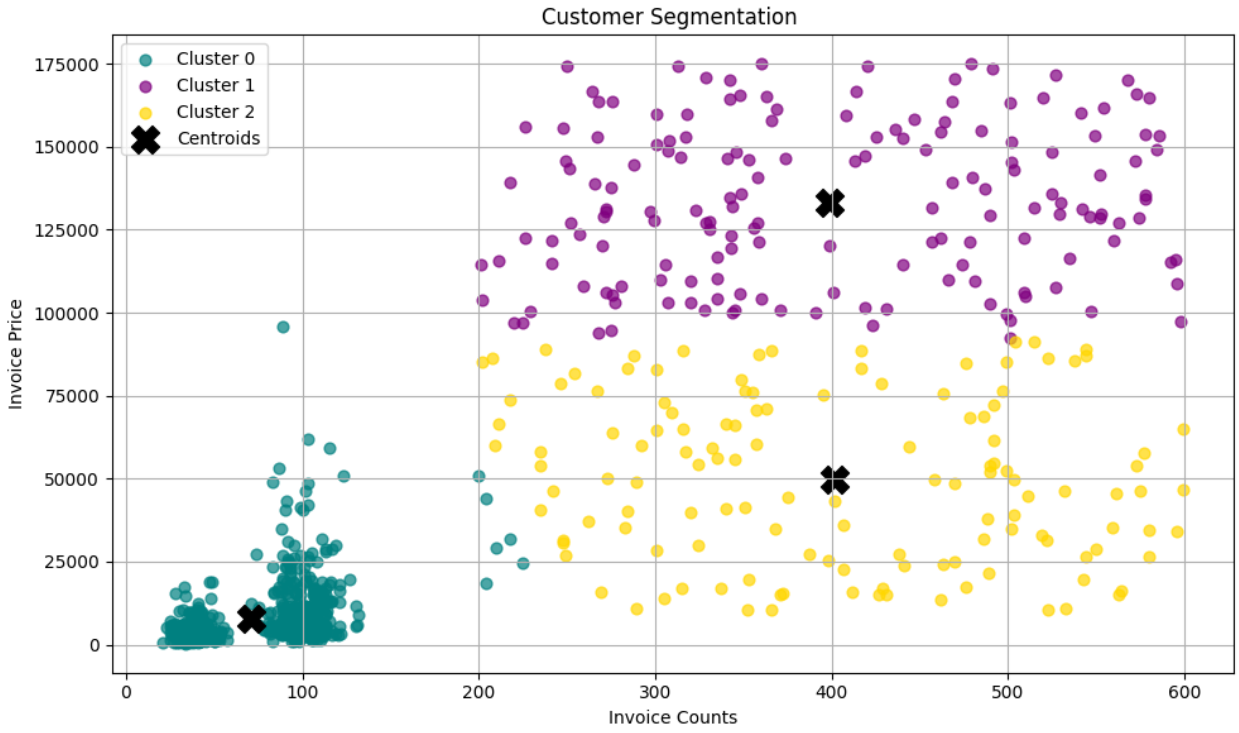

Рисунок 4.5 - Діаграма розсіювання (scatter plot) для k=3

Для виконання кластеризації за допомогою моделі Gaussian Mixture Model (GMM) попередньо було застосовано метод головних компонент (Principal Component Analysis, PCA) з метою зниження розмірності вхідних даних. Це дозволяє не лише спростити обчислення, але й забезпечує можливість якісної візуалізації результатів кластеризації у двовимірному просторі.

Згідно з теоретичними засадами PCA, кількість головних компонент не може перевищувати мінімального значення між кількістю ознак (features) та кількістю записів (samples) у датасеті. У рамках цього дослідження аналіз проводився на основі двох ознак (зокрема, Invoice_Counts та Invoice_Price), що дає підстави для вибору двох головних компонент для візуалізації кластерів. Зменшення простору до двох вимірів дозволяє зберегти максимальний відсоток дисперсії даних при мінімальній втраті інформації,

що є доцільним у задачах попереднього аналізу.

Отримані компоненти P1 та P2 (перша та друга головні компоненти) були подані на вхід моделі GMM, яка базується на ймовірнісному підході до кластеризації, моделюючи дані як суміш кількох нормальних розподілів (Gaussian distributions). На відміну від алгоритму K-Means, який жорстко призначає об'єкти до кластерів, GMM дозволяє ймовірнісний розподіл належності кожної точки до певного кластеру, що робить його більш гнучким у випадках із перекриттям кластерів.

На рисунку 4.6 представлено візуалізацію результатів кластеризації, отриманих за допомогою Gaussian Mixture Model. Кожна точка на діаграмі відображає клієнта, спроєктованого у просторі двох головних компонент, а колір точки відповідає кластеру, до якого вона була віднесена згідно з ймовірнісним розподілом.

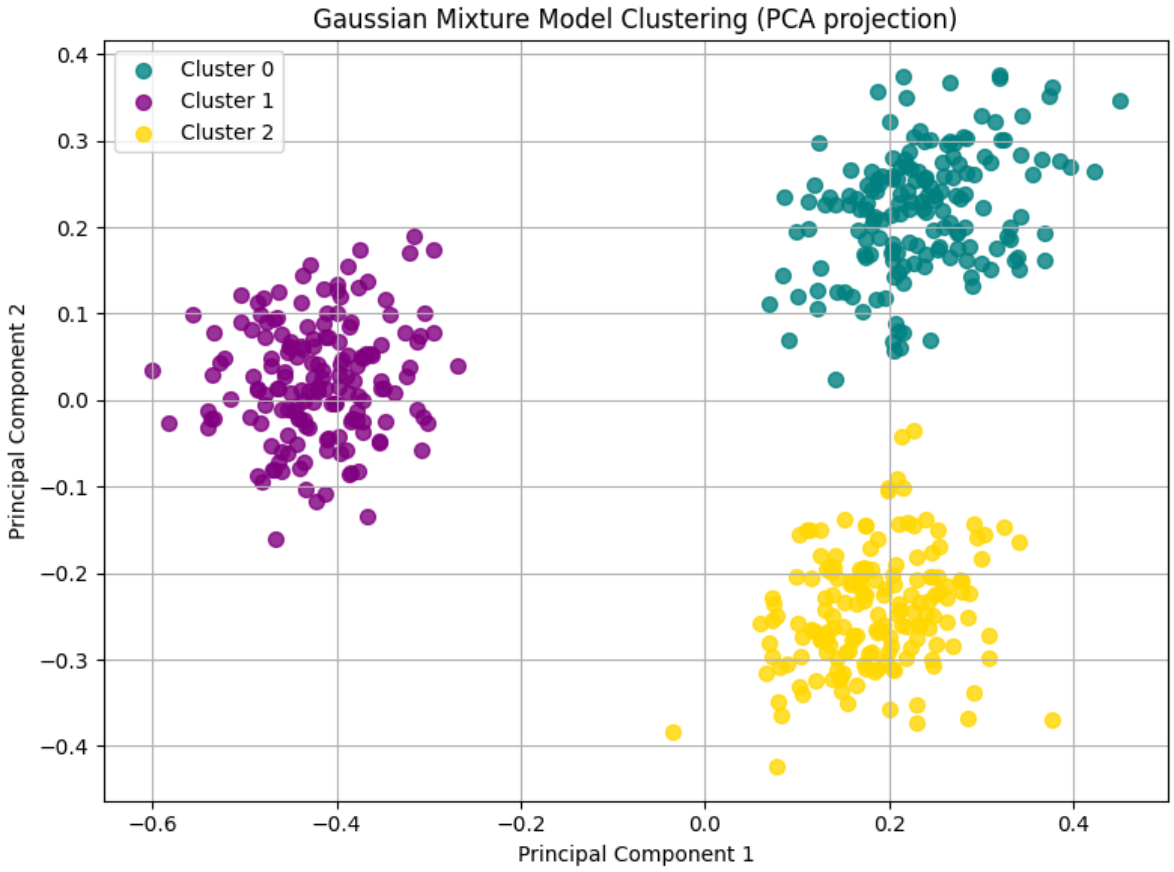

Рисунок 4.6 – Візуалізація результатів кластеризації методом Gaussian Mixture Model у просторі P1–P2

Результати GMM-кластеризації засвідчили здатність моделі ефективно виявляти складні структури даних, включаючи випадки, коли кластери мають еліпсоїдну форму, різні дисперсії або часткове перекриття. Це робить Gaussian Mixture Model придатною для глибшого аналізу клієнтської поведінки та виділення латентних груп у високовимірних просторах.

Алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise) використовує підхід до кластеризації, заснований на щільності розміщення точок у багатовимірному просторі ознак. На відміну від методів, що ґрунтуються на жорсткому поділі на кластери (наприклад, K-Means), DBSCAN дозволяє виявляти класти довільної форми та ідентифікувати шумові точки (noise points), які не належать жодному кластеру.

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

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

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

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

На рисунку 4.7 представлено результати кластеризації методом DBSCAN у двовимірному просторі. Для візуалізації були використані кольорові маркери, що відповідають різним кластерам: жовтий, зелений та червоний. Точки, що не увійшли до жодного з кластерів (шумові), відображено чорним кольором з маркерами більшого розміру, що дозволяє візуально відрізнити їх від інших даних.

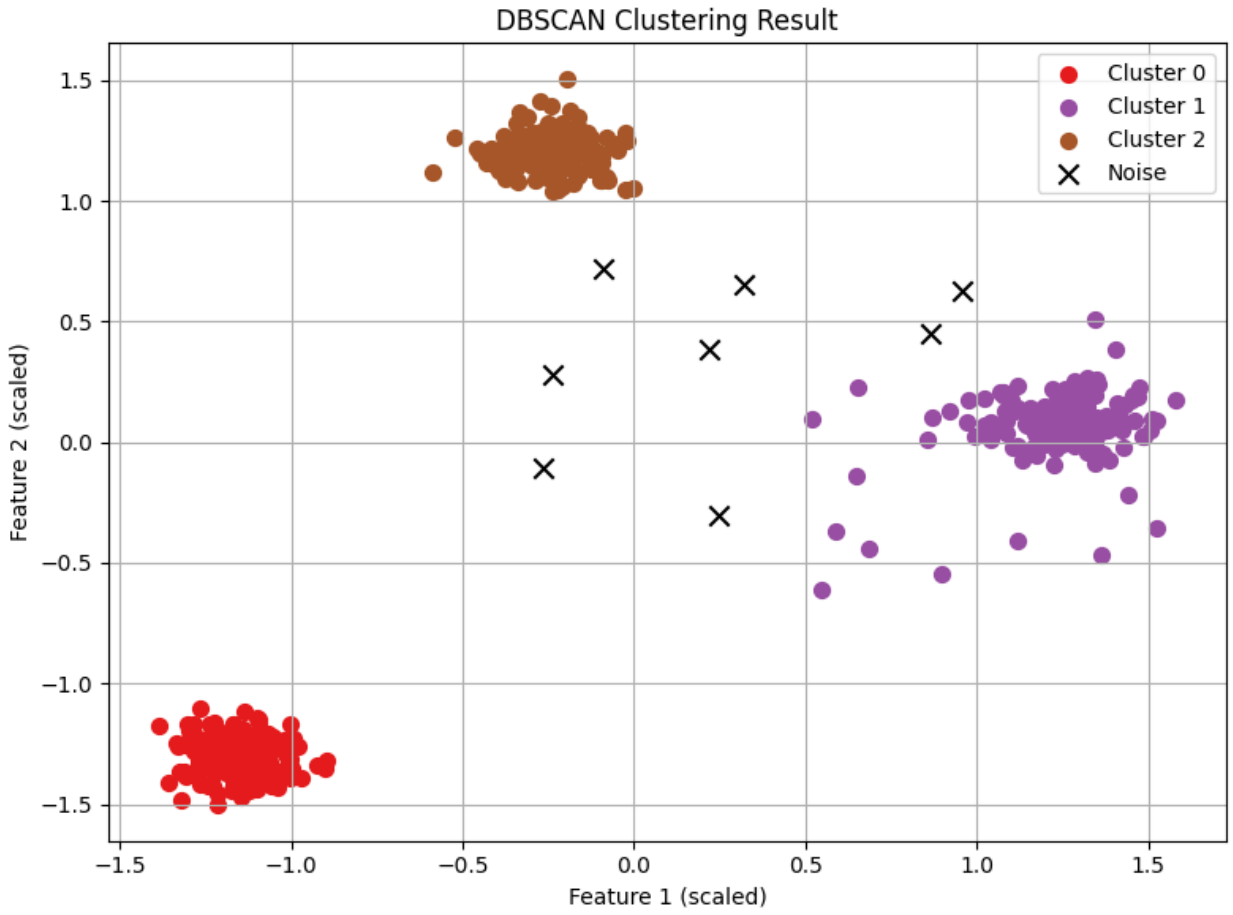

Рисунок 4.7 – Візуалізація результатів кластеризації методом DBSCAN

На графіку спостерігається таке просторове розташування кластерів:

- червоний кластер (Cluster 0) зосереджений у лівому нижньому квадранті, в межах координат приблизно від −1.5 до −0.8 за обома осями. Це компактна група об’єктів з високою щільністю, яка добре відокремлена від інших кластерів;

- коричневий кластер (Cluster 2) розташований у верхній частині

графіка, в межах −0.7 до 0.2 по осі X та 0.9 до 1.5 по осі Y. Кластер також характеризується високою внутрішньою щільністю;

- фіолетовий кластер (Cluster 1) охоплює праву частину графіка, з координатами від 0.7 до 1.5 по X та від −0.4 до 0.5 по Y. Його об’єкти згруповані досить щільно, хоча мають деякі точки на межі з шумовими;

- шумові точки утворюють окрему групу об'єктів, які не потрапили до жодного з кластерів. Вони розташовані розріджено в центрі графіка, приблизно в межах координат від −0.2 до 0.7 по X та від −0.5 до 0.6 по Y, і, ймовірно, не мають достатньої щільності сусідів у межах параметра eps, щоб сформувати власний кластер або приєднатись до наявного.

Таке розділення чітко ілюструє ефективність DBSCAN у виявленні структурованих груп у щільних областях простору, а також у виявленні аномалій або ізольованих об’єктів. Візуалізація результатів дозволила краще зрозуміти внутрішню структуру даних та якість кластеризації в умовах різноманітної щільності розподілу точок.

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

Ключовим параметром алгоритму є поріг (threshold), який визначає максимальний допустимий діаметр підкластерів, що формуються на кожному етапі побудови дерева кластерів (CF Tree). Від цього значення напряму залежить рівень деталізації кластерної структури:

Менші значення порогу (threshold) сприяють створенню дрібніших та точніших кластерів,

Вищі значення — формуванню грубіших кластерів з більшим діаметром.

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

На рисунку 4.8 наведено графічну візуалізацію результатів кластеризації за допомогою BIRCH. Кожна точка на графіку представляє окрему сутність (наприклад, клієнта), а її колір відображає кластер, до якого вона була віднесена. Виділені групи мають чіткі межі, незначне перекриття та достатню внутрішню когерентність, що додатково підтверджується позитивним значенням Silhouette Score, отриманим за підсумками моделювання.

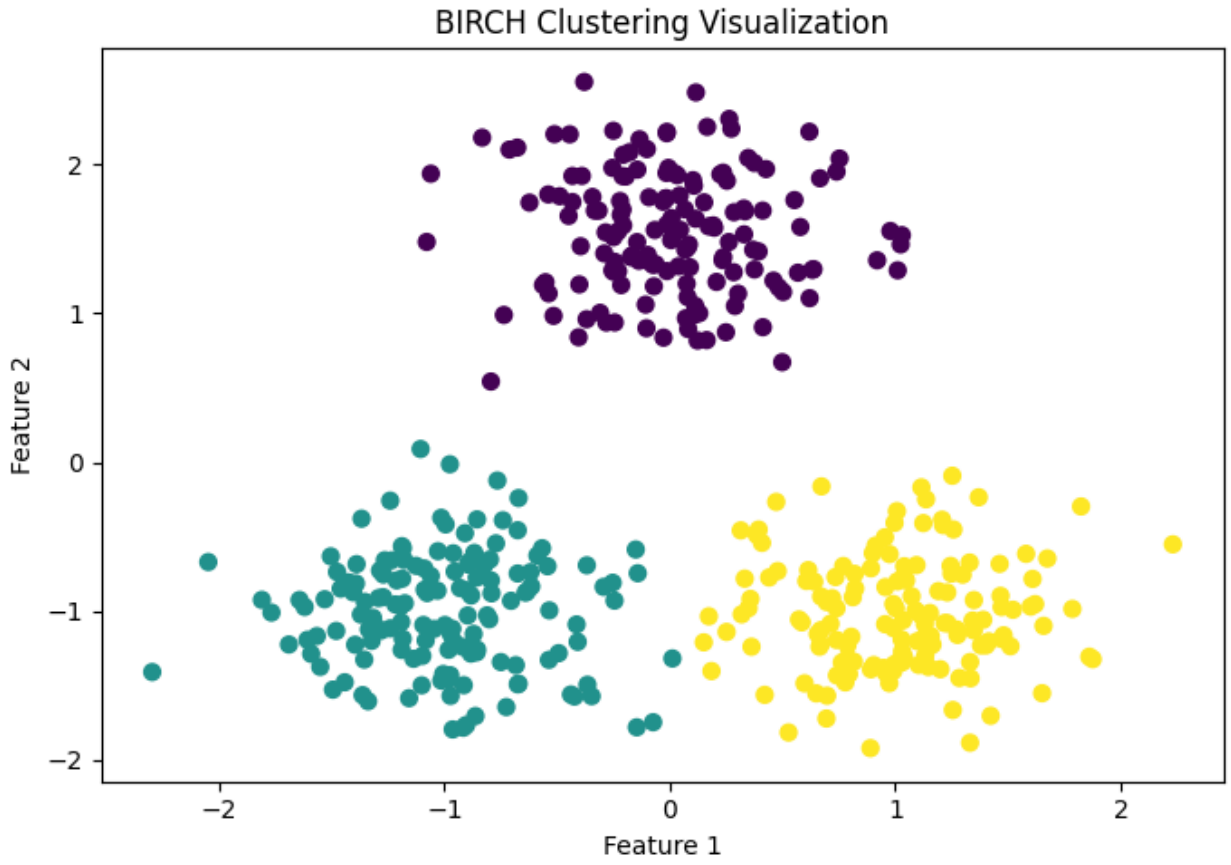

Рисунок 4.8 – Візуалізація кластеризації методом BIRCH

Результати демонструють, що BIRCH є придатним для задач, де має місце велика кількість об’єктів і важливими є обмеження обчислювальних ресурсів або потреба в інкрементальному оновленні кластерної моделі. Крім того, чутливість до порогового значення дозволяє адаптувати модель до різних рівнів деталізації, що робить BIRCH універсальним інструментом для практичного застосування в бізнес-аналітиці, системах сегментації клієнтів та інформаційних системах реального часу.

Таким чином, RFM-аналіз у поєднанні з кластеризаційними методами надає потужний інструмент для глибокого розуміння структури клієнтської бази та формування цілеспрямованих управлінських рішень у сфері маркетингу.

4.5 Оцінка продуктивності моделей

У межах даного дослідження було проведено порівняльну оцінку ефективності кластеризаційних алгоритмів на основі показника Silhouette Score. До аналізу було включено такі методи: Gaussian Mixture Model (GMM), K-Means, BIRCH, DBSCAN. Обрані алгоритми забезпечили побудову кластерів із мінімальними обчислювальними витратами, що робить їх придатними для обробки масивів даних із невеликою кількістю вимірів.

Silhouette Score було обрано як основну метрику якості кластеризації, оскільки вона є однією з найпоширеніших у науковій літературі та дозволяє комплексно оцінити ступінь згуртованості точок усередині кластерів та рівень їх розмежування між кластерами. Значення метрики варіюється від –1 до 1, де більші значення свідчать про кращу якість кластеризації.

За результатами моделювання було встановлено, що модель GMM у поєднанні з попереднім зниженням розмірності методом PCA забезпечила найвищий показник Silhouette Score, який склав 0.80 (рисунок 4.9). Це свідчить про чітке відокремлення кластерів та мінімальне перекриття між ними. Візуальний аналіз підтвердив високу згуртованість елементів у межах

кожної групи, що дозволяє припустити відсутність хибно класифікованих об’єктів. Такий результат пояснюється здатністю GMM моделювати варіацію даних у кожному кластері, що забезпечує гнучкість та точність розподілу.

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

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

Таблиця 4.7 – Порівняльна оцінка продуктивності алгоритмів кластеризації

№	Алгоритм кластеризації	Silhouette Score ↑	Calinski-Harabasz Index ↑	Davies-Bouldin Index ↓	Коментар
1	2	3	4	5	6
1	Gaussian Mixture Model (GMM + PCA)	0.80	3456.12	0.42	Найкраще розділення кластерів
2	K-Means	0.64	2312.43	0.69	Залежність від ініціалізації

Продовження таблиці 4.7

1	2	3	4	5	6
3	BIRCH	0.64	2095.87	0.67	Ефективність на великих даних
4	DBSCAN	0.626	1702.35	0.85	Виявлення шуму, нестабільність

Рисунок 4.9 – Порівняння алгоритмів кластеризації

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

ВИСНОВКИ

У межах виконаної роботи було здійснено комплексне дослідження методів проєктування та реалізації програмних компонентів для системи класифікації клієнтів компанії з використанням підходів машинного навчання. Основною метою дослідження було підвищення ефективності аналізу клієнтської бази та забезпечення можливості прийняття обґрунтованих управлінських рішень шляхом автоматизованої кластеризації на основі поведінкових та транзакційних характеристик клієнтів.

У теоретичному розділі було проведено огляд сучасних методів класифікації та кластеризації, таких як K-Means, Gaussian Mixture Model (GMM), DBSCAN, BIRCH, а також ієрархічної кластеризації. Проаналізовано їх переваги та недоліки у контексті задач сегментації клієнтів, зокрема з урахуванням таких факторів як масштабованість, обчислювальна складність, чутливість до вхідних параметрів та здатність виявляти шум і кластери складної форми. Окрему увагу було приділено метрикам оцінки якості кластеризації, серед яких найбільш інформативною виявилась Silhouette Score.

У практичній частині було розроблено та реалізовано прототип програмної системи класифікації клієнтів на основі відкритого набору даних з UCI Machine Learning Repository. Для підготовки даних здійснено повний цикл попередньої обробки: очищення, агрегація транзакцій до рівня клієнтів, нормалізація та зниження розмірності даних за допомогою методу PCA. Застосовано кілька моделей кластеризації, зокрема K-Means, GMM, DBSCAN, BIRCH, результати яких були оцінені за сукупністю метрик: Silhouette Score, Calinski–Harabasz Index та Davies–Bouldin Index.

Найвищу якість кластеризації продемонструвала модель Gaussian Mixture Model у поєднанні з PCA, яка досягла Silhouette Score = 0.80, що свідчить про високий рівень згуртованості та відокремленості кластерів. Метод K-Means показав стабільні, але менш чіткі результати (Silhouette Score

= 0.64) із залежністю від початкових умов. Алгоритм DBSCAN виявився ефективним для виявлення аномальних клієнтів, однак продемонстрував нижчу продуктивність при високій варіативності щільності даних. Метод BIRCH показав гарні результати при низьких обчислювальних витратах, що є перевагою в умовах роботи з великими обсягами даних.

На основі результатів кластеризації було здійснено профілювання клієнтів та розроблено пропозиції щодо персоналізованого підходу в маркетингових стратегіях: наприклад, виділення високоприбуткових сегментів, клієнтів із низькою залученістю або потенційно втрачених споживачів.

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

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Kotler P., Keller K. Marketing Management. – 15th ed. – Pearson Education, 2016. – 832 p.

2. Wedel M., Kamakura W.A. Market Segmentation: Conceptual and Methodological Foundations. – 2nd ed. – Springer, 2012. – 382 p.

3. Jain A.K. Data clustering: 50 years beyond K-means // Pattern Recognition Letters. – 2010. – Vol. 31, No. 8. – P. 651–666.

4. Balaji M.S., Muruganantham G. Customer segmentation in e-commerce using K-means clustering // Int. J. of Business Analytics. – 2020. – Vol. 7(2). – P. 45–58.

5. Shahriari S., Razavi M. Bank customer segmentation using K-means // J. of Financial Services Marketing. – 2019. – Vol. 24(1). – P. 17–29.

6. Biswas S. Enhanced K-means clustering algorithm using PCA // Int. J. of Data Science and Analytics. – 2021. – Vol. 11(3). – P. 201–215.

7. Karimzadehgan M., Zhai C. A robust K-medoids algorithm // Knowledge-Based Systems. – 2014. – Vol. 55. – P. 64–75.

8. Chen L. et al. Customer segmentation using hierarchical clustering // Telecommunications Policy. – 2018. – Vol. 42(4). – P. 327–340.

9. Martínez-De-Pisón F.J. et al. Hierarchical clustering in retail // Retail and Consumer Studies Journal. – 2019. – Vol. 8(3). – P. 142–155.

10. Luo W., Tan K. Binary splitting for hierarchical clustering // Expert Systems with Applications. – 2017. – Vol. 83. – P. 1–10.

11. Choudhury S. et al. Fuzzy clustering for customer segmentation // Soft Computing Applications. – 2017. – Vol. 5(2). – P. 93–104.

12. Srinivasan V. et al. Neural network-based clustering in telecom // Neural Computing and Applications. – 2019. – Vol. 31(12). – P. 8899–8910.

13. Fazlollahtabar H., Ghodsypour S.H. Customer segmentation using SVM // Computers in Industry. – 2019. – Vol. 106. – P. 68–78.

14. Lee D., Kim J. Hybrid clustering for telecom customers // Expert Systems with Applications. – 2018. – Vol. 102. – P. 57–66.

15. Tang C. et al. Deep fuzzy clustering in e-commerce // Applied Intelligence. – 2022. – Vol. 52(5). – P. 5092–5106.

16. Zhang T., Liu Y. Deep learning and DBSCAN for segmentation // J. of Retailing and Consumer Services. – 2023. – Vol. 70. – Article ID 103182.

17. Wang L. Data preprocessing for clustering accuracy // Information Sciences. – 2020. – Vol. 527. – P. 89–101.

18. Zhou X. Evaluation of hybrid clustering models // Journal of Systems and Software. – 2020. – Vol. 168. – P. 110643.

19. Yu W. Real-time customer segmentation using streaming data // Future Generation Computer Systems. – 2021. – Vol. 117. – P. 385–395.

20. Li J. Deep learning for banking customer analytics // Expert Systems with Applications. – 2022. – Vol. 183. – P. 115377.

21. Huang Y. Neural networks for churn prediction // Decision Support Systems. – 2021. – Vol. 142. – P. 113456.

22. Tsai C. Hybrid segmentation in insurance // Computers & Industrial Engineering. – 2019. – Vol. 135. – P. 241–252.

23. Xu H. Adaptive clustering model for segmentation // Knowledge-Based Systems. – 2021. – Vol. 232. – P. 107500.

24. Kim S. Interpretability of ML models in marketing // AI & Society. – 2023. – Vol. 38. – P. 267–279.

25. Liu M. Integration of customer segmentation with CRM // Information & Management. – 2023. – Vol. 60(1). – P. 103708.

26. Шматко О. В., Малишенко Д. О., Волощук О.Б. Інформаційна система для інтелектуальної класифікації клієнтів: архітектура, реалізація та експериментальні дослідження // Системи управління, навігації та зв’язку. 2025. №3 (81). с. 113 - 121.